Quadratically Regularized Optimal Transport: Existence and Multiplicity of Potentials

Marcel Nutz*

February 10, 2024

Abstract

The optimal transport problem with quadratic regularization is useful when sparse couplings are desired. The density of the optimal coupling is described by two functions called potentials; equivalently, potentials can be defined as a solution of the dual problem. We prove the existence of potentials for a general square-integrable cost. Potentials are not necessarily unique, a phenomenon directly related to sparsity of the optimal support. For discrete problems, we describe the family of all potentials based on the connected components of the support, for a graph-theoretic notion of connectedness. On the other hand, we show that continuous problems have unique potentials under standard regularity assumptions, regardless of sparsity. Using potentials, we prove that the optimal support is indeed sparse for small regularization parameter in a continuous setting with quadratic cost, which seems to be the first theoretical guarantee for sparsity in this context.

Keywords Optimal Transport, Quadratic Regularization, Potentials
AMS 2020 Subject Classification 90C25; 49N05

1 Introduction

We are concerned with quadratically regularized optimal transport; that is,

$$\mathrm{QOT}_\varepsilon(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathsf{X}\times\mathsf{Y}} c(x,y)\,\pi(dx,dy) + \frac{\varepsilon}{2}\left\|\frac{d\pi}{dP}\right\|^2_{L^2(P)} \tag{1.1}$$

*Columbia University, Depts. of Statistics and Mathematics, mnutz@columbia.edu. Research supported by NSF Grants DMS-1812661, DMS-2106056. The author is grateful to Alberto González-Sanz, Gilles Mordant, Andrés Riveros Valdevenito and Johannes Wiesel for stimulating discussions.

where μ, ν are given marginal distributions on separable spaces X, Y and $\Pi(\mu, \nu)$ denotes the set of their couplings. Moreover, P is a (product) reference measure for computing the density $d\pi/dP$. The function $c : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ is a given "cost" and the regularization parameter $\varepsilon > 0$ controls the strength of the regularization. Quadratic regularization is also called Euclidean or χ^2 regularization. We shall prove for a general cost $c \in L^2(P)$ that the unique optimal coupling π_* for (1.1) has a density of the form

$$\frac{d\pi_*}{dP}(x, y) = \big(f(x) + g(y) - c(x, y)/\varepsilon\big)_+ \tag{1.2}$$

for certain functions $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$; cf. Theorem 2.2. These functions are called *potentials*. They are also described as solutions to a dual problem, but are in general non-unique. For discrete problems, non-uniqueness is the typical case, and we shall describe the family of all potentials (Theorem 3.5). Our description uses a decomposition of the support of π_* into "components" (for a certain notion of connectedness; cf. Definition 3.2). In particular, the multiplicity of the potentials is directly related to sparsity of the support. For continuous and semi-discrete problems, the situation is quite different: under standard regularity conditions we show that the potentials are unique (up to an additive constant); cf. Theorem 3.7. We use potentials to show that the support of π_ε is sparse for small $\varepsilon > 0$, in the continuous setting with quadratic Euclidean cost c. Specifically, we show that the support is contained in a neighborhood of the graph of Brenier's map (Theorem 4.1). To the best of our knowledge, this is the first theoretical result on sparsity of continuous regularized transport.

Regularization has many purposes in optimal transport—to facilitate computation, to obtain smoother couplings and dual potentials, to improve sampling complexity, and others. Two regularizations are primarily used: entropic regularization penalizes couplings by the Kullback–Leibler divergence while quadratic regularization penalizes by the L^2-norm of the density. Entropic regularization (EOT) is the most frequent choice, as it allows for Sinkhorn's algorithm (e.g., [7, 28]) and has strong smoothness properties. This smoothness is intimately linked to the full support property of the optimal coupling, which can be a blessing or a curse ("overspreading") depending on the application. While for small values of the regularization parameter ε, the actual weight of the EOT coupling might be quite small in large regions, the issue is aggravated by a second issue of EOT: its computation is difficult for small values of ε (e.g., [32]). By contrast, quadratic regularization is empirically known to give rise to couplings with *sparse support* for a range of ε. Moreover, as its computation does not involve logarithms and exponentials,

one can use regularization parameters that are several orders of magnitude smaller than in EOT without running into issues with machine precision. For those reasons, quadratic regularization is used in applications where sparsity and/or weak regularization are desired.

Quadratically regularized optimal transport was first addressed by [2, 12] in discrete settings. It is also a special case of optimal transport with convex regularization [8]; see also the predecessors referenced in [12]. The formulation of [2] is closer to ours; the authors present several experiments highlighting the sparsity of the optimal coupling and derive theoretical results regarding duality and convergence as the regularization parameter ε tends to zero. The authors further illustrate how entropic regularization can lead to blurrier results in image processing tasks. In [12], quadratic regularization is studied for a minimum-cost flow problem on a graph; this includes discrete optimal transport as a particular case. The authors introduce a Newton-type algorithm and discuss sparsity in several examples. In a continuous setting, several works including [10, 16, 18, 19, 33] have applied optimization techniques on the dual problem of regularized optimal transport. For instance, [19] applies neural networks and gradient descent to compute regularized Wasserstein barycenters. The authors compare entropic and quadratic regularization and highlight that the entropic penalty produces a blurrier image at the smallest computationally feasible regularization ($\varepsilon = 10^{-2}$ for entropic, $\varepsilon = 10^{-5}$ for quadratic). Recently, [35] uses quadratically regularized optimal transport in a manifold learning task related to single cell RNA sequencing; specifically, the optimal coupling is used to produce an adaptive affinity matrix. In this context, sparsity is crucial to avoid bias—a full support coupling would introduce shortcuts through ambient space instead of following the data manifold. In this application, the transport problem is of "self-transport" type: the marginals $\mu = \nu$ are identical and the cost c is symmetric. In that situation we will show that the potentials f, g can be chosen to be symmetric; i.e., $f = g$. While symmetry eliminates certain degrees of freedom, we shall see that non-uniqueness can still occur.

The first work rigorously addressing a continuous setting is [21]. The authors derive duality results and present two algorithms, a nonlinear Gauss–Seidel method and a semismooth Newton method. The theoretical results assume that the marginal spaces X, Y are compact subsets of $\mathbb{R}^d$ and that the marginal distributions μ, ν are absolutely continuous with densities uniformly bounded away from zero. The reference measure P is taken to be the Lebesgue measure. The authors apply weak* compactness in L^1 to solve the dual problem, and then this solution provides potentials. By contrast, our approach is not of topological nature. It covers in a unified way dis-

crete and continuous settings, and different reference measures P, avoiding technical restrictions almost entirely. The paper [20] generalizes some of the results of [21] to Orlicz space regularizations and shows Gamma convergence as $\varepsilon \to 0$ to the unregularized optimal transport problem. This convergence is studied quantitatively in [11], where a rate of convergence is derived based on quantization arguments, while [1] shows stability with respect to the marginal distributions; both cover quadratic regularization as a special case of more general f-divergences. The unpublished work [9] also considers optimal transport with regularization by an f-divergence, with quadratic regularization being a special case. The authors emphasize the analogy to c-convex conjugation in optimal transport (cf. the semi-smooth dual studied, e.g., in [2]) and use it to derive a priori estimates for the potentials. In a setting with uniformly bounded cost c and $P = \mu \otimes \nu$, these results are leveraged to obtain existence of the potentials and convergence of the nonlinear Gauss–Seidel algorithm. The paper also states results regarding the uniqueness of the potentials and differentiability of $\nu \mapsto \mathrm{QOT}_\varepsilon(\mu, \nu)$ which however are flawed. Specifically, uniqueness is asserted in a general setting including discrete problems, based on an assertion that the dual problem is strictly concave. We emphasize that the dual problem (2.6) is not strictly concave in the case of quadratic regularization ($x \mapsto -x_+^2$ is constant on $\mathbb{R}_-$) and uniqueness fails in simple situations such as $\mu = \nu = \frac{1}{2}(\delta_0 + \delta_1)$ with $c(x, y) = |x - y|^2$ and $\varepsilon = 1/3$ (see Example 3.1 for details). In cases where uniqueness holds, it does so for very different reasons.

To the best of our knowledge, apart from the aforementioned, we are the first to describe the multiplicity of the potentials; specifically, to describe the family of all potentials in the discrete case and prove uniqueness in a continuous (and semi-continuous) case. The connection between the family of all potentials and "components" of the support also seems to be novel. While sparsity of the support has been highlighted as an empirical finding, Theorem 4.1 seems to be the first theoretical result in its direction. As mentioned above, there is no analogue to this sparsity in EOT, where the support of the optimal coupling always equals the support of $\mu \otimes \nu$.

For the existence of the potentials, we pursue a novel path inspired by [13]: while the works cited above attack the dual problem, we leverage the (straightforward) fact that the primal problem has a solution given by a Hilbert space projection. To construct potentials, we introduce approximating problems with finitely many equality constraints instead of the marginal constraints. Their solutions have the form (1.2) and converge to the optimal density. To achieve the passage to the limit, we must show that a sequence of functions $f_n(x) + g_n(y)$ converges to a limit of the form $f(x) + g(y)$. This

problem is surprisingly subtle, but we can take advantage of insights found in the context of Schrödinger bridges [14, 31]. This line of argument avoids the conditions on the marginals and costs in previous approaches. It also allows us to cover different reference measures P in a unified way. Once the form (1.2) is obtained, the properties of the dual problem follow easily by standard arguments.

The remainder of this paper is organized as follows. Section 2 states the existence and duality results, and basic regularity properties of potentials. In Section 3, we characterize the family of all potentials in the discrete case and prove the uniqueness in the continuous case. Section 4 applies potentials to prove sparsity in the setting of quadratic cost. Section 5 contains the proof of the existence and duality result, while Section 6 proves the same result for the self-transport problem; i.e., with potentials $f = g$.

2 Problem Formulation and Existence

Consider two Polish[1] probability spaces $(\mathsf{X}, \mathcal{B}(\mathsf{X}), \mu)$ and $(\mathsf{Y}, \mathcal{B}(\mathsf{Y}), \nu)$. We endow $\mathsf{X} \times \mathsf{Y}$ with the product σ-field and denote by $\Pi(\mu, \nu)$ the set of couplings of (μ, ν); that is, measures π on $\mathsf{X} \times \mathsf{Y}$ satisfying $\pi(A \times \mathsf{Y}) = \mu(A)$ for $A \in \mathcal{B}(\mathsf{X})$ and $\pi(\mathsf{X} \times B) = \nu(B)$ for $B \in \mathcal{B}(\mathsf{Y})$. We also use the standard notation $(f \oplus g)(x, y) := f(x) + g(y)$ for functions $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$, and $\mathcal{P}(\mathsf{X})$ for the set of probability measures on X.

We further consider measures $(\tilde{\mu}, \tilde{\nu}) \in \mathcal{P}(\mathsf{X}) \times \mathcal{P}(\mathsf{Y})$ satisfying $\mu \sim \tilde{\mu}$ and $\nu \sim \tilde{\nu}$ (where $\sim$ denotes mutual absolute continuity) and

$$\frac{d\mu}{d\tilde{\mu}} \in L^2(\tilde{\mu}), \quad \left(\frac{d\mu}{d\tilde{\mu}}\right)^{-1} \in L^\infty(\tilde{\mu}), \quad \frac{d\nu}{d\tilde{\nu}} \in L^2(\tilde{\nu}), \quad \left(\frac{d\nu}{d\tilde{\nu}}\right)^{-1} \in L^\infty(\tilde{\nu}), \tag{2.1}$$

and denote their product

$$P := \tilde{\mu} \otimes \tilde{\nu}. \tag{2.2}$$

Finally, we are given a cost function

$$\begin{aligned} &c \in L^2(P) \quad \text{satisfying} \quad c \geq c_1 \oplus c_2 \\ &\text{for some} \qquad c_1 \in L^1(\mu) \cap L^1(\tilde{\mu}), \quad c_2 \in L^1(\nu) \cap L^1(\tilde{\nu}); \end{aligned} \tag{2.3}$$

[1] More generally, our results hold for all separable probability spaces; cf. Remark B.1.

the lower bound ensures in particular that for any $\pi \in \Pi(\mu,\nu)$, the integral $\int c\,d\pi$ is well defined with values in $(-\infty,\infty]$. With this notation in place, the quadratically regularized optimal transport problem (with $\varepsilon = 1$) is

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c\,d\pi + \frac{1}{2}\left\|\frac{d\pi}{dP}\right\|^2_{L^2(P)} \tag{2.4}$$

where (by convention) any coupling $\pi \not\ll P$ has infinite cost. The extension to general regularization parameter $\varepsilon > 0$ is straightforward; see Remark 2.3.

Remark 2.1. (a) For the reference measure P in (2.2), our default is $(\tilde\mu,\tilde\nu) := (\mu,\nu)$. This choice leads to a meaningful problem (2.4) in discrete and continuous settings, and a consistent scaling for discrete-to-continuous limits. However, numerous works use other choices for $\tilde\mu,\tilde\nu$, especially uniform measures on a certain domain (usually discrete or in $\mathbb{R}^d$). While the choice of reference measure is often not highlighted in the literature, we shall see that it can be quite crucial.[2] For the entropic optimal transport problem, it is known that changing the measures $\tilde\mu,\tilde\nu$ does not affect the optimal coupling (e.g., [24]). That fact has no analogue for the present problem. Consequently, we provide the results for general P. Example A.1 and Proposition A.2 show that the optimal coupling and even the optimal support can depend on $\frac{d\mu}{d\tilde\mu}$ and $\frac{d\nu}{d\tilde\nu}$, even for the straightforward cost $c \equiv 0$.

(b) The integrability condition on $(\frac{d\mu}{d\tilde\mu})^{-1}, (\frac{d\nu}{d\tilde\nu})^{-1}$ in (2.1) is used only to obtain a convenient lower bound for potentials (Lemma 5.1). No condition is needed for the existence results in Propositions 5.8 and 6.4.

(c) While our problem (2.4) makes sense if merely $\mu \ll \tilde\mu$ and $\nu \ll \tilde\nu$, there is no real gain in generality for (2.4) from this weaker condition. Hence, we use equivalent measures.

In view of the second term in (2.4), the minimization is equivalently restricted to couplings with square-integrable density. Indeed, let

$$\mathcal{Z} := \{Z \in L^2(P) : Z\,dP \in \Pi(\mu,\nu)\},$$

which is nonempty as it contains $d(\mu\otimes\nu)/dP$ due to (2.1). Then we can rephrase (2.4) as

$$\mathscr{P} := \inf_{Z\in\mathcal{Z}} \int cZ\,dP + \frac{1}{2}\|Z\|^2_{L^2(P)}, \tag{2.5}$$

[2]The function `ot.stochastic.loss_dual_quadratic` in the `Python Optimal Transport` package follows [33] and assumes $(\tilde\mu,\tilde\nu) = (\mu,\nu)$; see [29]. The function `OptimalTransport.quadreg` in the `OptimalTransport.jl` package follows [21] and assumes that $\tilde\mu,\tilde\nu$ are uniform; see [27].

called the *primal problem* below. Note that $\mathscr{P} \in \mathbb{R}$ due to (2.3) and the Cauchy–Schwarz inequality. The *dual problem* is

$$\mathscr{D} := \sup_{f \in L^1(\mu),\, g \in L^1(\nu)} \int f \, d\mu + \int g \, d\nu - \frac{1}{2} \int (f \oplus g - c)_+^2 \, dP. \tag{2.6}$$

The two problems are in strong duality, and both admit optimizers.

Theorem 2.2. *(i) Strong duality holds: $\mathscr{P} = \mathscr{D}$.*

(ii) The primal problem (2.5) *has a unique solution $Z_* \in \mathcal{Z}$ given by the $L^2(P)$-projection $Z_* = \arg\min_{Z \in \mathcal{Z}} \|Z + c\|^2_{L^2(P)}$ of $-c$ onto $\mathcal{Z}$. In particular, Z_* is characterized by*

$$\langle Z_* + c, Z - Z_* \rangle_{L^2(P)} \geq 0 \quad \textit{for all} \quad Z \in \mathcal{Z}.$$

The coupling $\pi_* \in \Pi(\mu, \nu)$ given by $d\pi_* = Z_* \, dP$ is the unique solution of the regularized transport problem (2.4).

(iii) There exist measurable functions $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$ satisfying the following conditions, and these conditions are equivalent:

(a) $(f \oplus g - c)_+$ is the density $Z_* \in \mathcal{Z}$ of the optimal coupling π_*,

(b) $(f \oplus g - c)_+$ is the density of some coupling $\pi \in \Pi(\mu, \nu)$,

(c) $(f, g) \in L^1(\mu) \times L^1(\nu)$ is a solution of the dual problem,

(d) (f, g) satisfies the system

$$\int_{\mathsf{X}} (f(x) + g(y) - c(x,y))_+ \, \tilde{\mu}(dx) = \frac{d\nu}{d\tilde{\nu}}(y) \quad \textit{for } \tilde{\nu}\textit{-a.e. } y \in \mathsf{Y}, \tag{2.7}$$

$$\int_{\mathsf{Y}} (f(x) + g(y) - c(x,y))_+ \, \tilde{\nu}(dy) = \frac{d\mu}{d\tilde{\mu}}(x) \quad \textit{for } \tilde{\mu}\textit{-a.e. } x \in \mathsf{X}. \tag{2.8}$$

Any such (f, g), necessarily in $L^1(\mu) \times L^1(\nu)$, are called potentials.

(iv) Suppose that $(\mathsf{X}, \mu, \tilde{\mu}) = (\mathsf{Y}, \nu, \tilde{\nu})$ and $c(x, y) = c(y, x)$. Then the existence in (iii) also holds with the additional requirement that $f = g$, and the dual problem (2.6) *has the same value $\mathscr{D}$ if restricted to $f = g$.*

For ease of reference, the following remark states the corresponding formulas for general regularization parameter $\varepsilon > 0$.

Remark 2.3 (General $\varepsilon > 0$)**.** Often the regularized transport problem is considered with a parameter $\varepsilon > 0$ for the quadratic penalty:

$$\mathscr{P}_\varepsilon = \inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi + \frac{\varepsilon}{2} \|d\pi/dP\|^2_{L^2(P)}.$$

Note that $\mathscr{P}_\varepsilon = \varepsilon \mathscr{P}(\bar{c})$ if $\mathscr{P}(\bar{c})$ is the primal problem (2.5) for the cost $\bar{c} = c/\varepsilon$. The corresponding dual is

$$\mathscr{D}_\varepsilon = \sup_{f \in L^1(\mu),\, g \in L^1(\nu)} \varepsilon \int f \, d\mu + \varepsilon \int g \, d\nu - \frac{\varepsilon}{2} \int (f \oplus g - c/\varepsilon)^2_+ \, dP \tag{2.9}$$

and the optimal density then takes the form $Z_\varepsilon = (f \oplus g - c/\varepsilon)^2_+$ for the optimizers (f, g) of (2.9). In situations where ε is varied, it is often convenient to consider the *rescaled* potentials $(f_\varepsilon, g_\varepsilon) := (\varepsilon f, \varepsilon g)$. After this change of variables, the dual problem reads

$$\mathscr{D}_\varepsilon = \sup_{f \in L^1(\mu),\, g \in L^1(\nu)} \int f \, d\mu + \int g \, d\nu - \frac{\varepsilon}{2} \int \left(\frac{f \oplus g - c}{\varepsilon} \right)^2_+ \, dP \tag{2.10}$$

and the optimal density takes the form

$$Z_\varepsilon = \left(\frac{f_\varepsilon \oplus g_\varepsilon - c}{\varepsilon} \right)_+$$

for the optimizers $(f_\varepsilon, g_\varepsilon)$ of (2.10). The rescaled potentials incorporate the correct scaling in particular for the limit $\varepsilon \to 0$.

For brevity, and without loss of generality, we use $\varepsilon = 1$ in the remainder of the paper, except in Section 4 where we consider the limit $\varepsilon \to 0$.

The proof of Theorem 2.2 consists of two parts. The main part is to prove (iii)(a); i.e., that the optimal density Z_* is of the form $(f \oplus g - c)_+$. The other assertions in (i)–(iii) follow from that fact and elementary arguments. To show the main part, we construct approximating problems whose solutions have the form $(f_n \oplus g_n - c)_+$ and converge to the optimal density Z_* as $n \to \infty$. Then, we argue that $f_n \oplus g_n$ converges to a function $f \oplus g$ on a sufficiently large set. The details of the proof are deferred to Section 5. In the symmetric setting of self-transport assumed in (iv), the construction of Section 5 generally yields potentials $f \neq g$. Section 6 presents a more refined construction guaranteeing $f = g$.

As we want to use (d) in the derivation below, let us observe that the equivalence of (b) and (d) in Theorem 2.2 is straightforward and independent of all other claims.

Proof of (b)⇔(d). The left-hand side in (2.7) is the formula for the density of the second marginal of the measure $d\pi := (f \oplus g - c)_+ \, dP$ wrt. $\tilde{\nu}$. If $\pi \in \Pi(\mu, \nu)$, the second marginal is ν, giving the right-hand side. Similarly for (2.8). Conversely, if the marginal densities equal $(d\mu/d\tilde{\mu}, d\nu/d\tilde{\nu})$, then π is a coupling. □

In the remainder of this section we gather some properties of potentials to be used in Section 3. (Additional bounds and integrability properties are stated in Section 5.2.) Let (f, g) be potentials. By (2.8), there is a set X_0 of full $\tilde{\mu}$-measure such that for $x \in \mathsf{X}_0$,

$$F_x(t) := \int_{\mathsf{Y}} (t + g(y) - c(x,y))_+ \, \tilde{\nu}(dy) = \frac{d\mu}{d\tilde{\mu}}(x) \quad \text{for} \quad t = f(x). \tag{2.11}$$

As $c \in L^2(P)$ and $\mu \sim \tilde{\mu}$, we may further choose X_0 such that $c(x, \cdot) \in L^1(\tilde{\nu})$ and $\frac{d\mu}{d\tilde{\mu}}(x) > 0$ for $x \in \mathsf{X}_0$. We observe that for all $x \in \mathsf{X}_0$, the function $t \mapsto F_x(t)$ is continuous, nondecreasing, strictly increasing on the set where is it positive, $\lim_{t\to-\infty} = 0$ and $\lim_{t\to\infty} = \infty$. For $x \in \mathsf{X}_0$, we conclude that there exists a unique t with $F_x(t) = \frac{d\mu}{d\tilde{\mu}}(x)$. In particular, the value of $f(x)$ is uniquely determined by g and $c(x, \cdot)$, μ-a.s. We record this fact for ease of reference.

Lemma 2.4. *One potential uniquely determines the other: if (f, g) and (f', g) are potentials, then $f = f'$ μ-a.s.*

More generally, $g^c(x) := \{t : F_x(t) = \frac{d\mu}{d\tilde{\mu}}(x)\}$ yields a version of the potential f that is defined everywhere on X_0. In fact, we can choose versions of c and $d\mu/d\tilde{\mu}$ such that $\mathsf{X}_0 = \mathsf{X}$, and then g^c is defined everywhere on X. Following standard arguments in optimal transport, we may think of g^c as a conjugate of f. This point of view was emphasized in [9] to show how potentials inherit properties from the cost function. Specifically, the setting of [9] assumes that $(\tilde{\mu}, \tilde{\nu}) = (\mu, \nu)$. We state (a generalization of) that result in the next lemma, before discussing how it breaks down when $\mu \neq \tilde{\mu}$.

Lemma 2.5 (Oscillation)**.** *Suppose that $\mu = \tilde{\mu}$ and define*

$$\Delta(x, x') := \sup_{y \in \mathsf{Y}} |c(x,y) - c(x',y)| \in [0, \infty].$$

If (f, g) are potentials, then f satisfies $|f(x) - f(x')| \leq \Delta(x, x')$ for all x, x' outside a μ-nullset. In particular:

(i) If the oscillation $\operatorname{osc} c(\cdot, y) \leq C$ *for all y, then* $\operatorname{osc} f \leq C$ *μ-a.s.*

(ii) If c is bounded, then f is bounded. More precisely, $\|f+\alpha\|_{L^\infty(\mu)} \le 2\|c\|_\infty$ after choosing the centering α such that $\|(f+\alpha)_+\|_{L^\infty(\mu)} = \|(f+\alpha)_-\|_{L^\infty(\mu)}$.

(iii) If a metric is given on X and $x \mapsto c(x,y)$ is uniformly continuous with modulus of continuity ω, then f admits a version that is ω-continuous on $\operatorname{spt}\mu$. If $x \mapsto c(x,y)$ is L-Lipschitz and X is a Hilbert space, then f admits a version that is L-Lipschitz on X.

Proof. For all x, x' in a set X_0 of full μ-measure, writing $\Delta = \Delta(x,x')$, we use (2.11) at x and $\frac{d\mu}{d\tilde\mu} \equiv 1$ to find

$$\begin{aligned}\int (f(x)+g(y)-c(x',y) \mp \Delta)_+ \,\tilde\nu(dy) &\lesseqgtr \int (f(x)+g(y)-c(x,y))_+ \,\tilde\nu(dy)\\ &= \frac{d\mu}{d\tilde\mu}(x) = 1 = \frac{d\mu}{d\tilde\mu}(x').\end{aligned}$$

Now using (2.11) at x' and the strict monotonicity of $F_{x'}$ yield $f(x') \in [f(x)-\Delta, f(x)+\Delta]$. This immediately implies (i) which in turn implies (ii). If c is ω-continuous, the above shows that f is ω-continuous on X_0. Thus f can be uniquely extended to a ω-continuous function on the closure of X_0, which is a superset of $\operatorname{spt}\mu$ as $\mu(\mathsf{X}_0) = 1$. If f is Lipschitz on $\operatorname{spt}\mu$ and X is a Hilbert space, we can further use Kirszbraun's theorem and extend to a Lipschitz function on the whole space. □

Lemma 2.5 showcases the idea that the potential f inherits regularity from the cost c.[3] This breaks down when $\mu \neq \tilde\mu$. For instance, Proposition A.2 shows in particular that $f = \frac{d\mu}{d\tilde\mu}$ (up to additive constant) when $c \equiv 0$ and $\nu = \tilde\nu$. Hence, in general, the regularity of the potentials depends on the regularity of the marginal densities. The next lemma exemplifies how (2.11) can still be used to obtain regularity results.

Lemma 2.6 (Lipschitz potentials)**.** *Let X be endowed with a metric. Let c be bounded and let $x \mapsto c(x,y)$ be Lipschitz uniformly in y. Moreover, let $\frac{d\mu}{d\tilde\mu}$ and $(\frac{d\mu}{d\tilde\mu})^{-1}$ be bounded and Lipschitz, and let $\frac{d\nu}{d\tilde\nu}$ be bounded. Then f admits a version that is bounded and Lipschitz on $\operatorname{spt}\mu$. If X is a Hilbert space, then f admits a version that is bounded and Lipschitz on X.*

Proof. For brevity, we write $\xi(x) = (\frac{d\mu}{d\tilde\mu}(x))^{-1}$ and

$$\tilde f(x) = \xi(x) f(x), \qquad \tilde g(x,y) = \xi(x) g(y), \qquad \tilde c(x,y) = \xi(x) c(x,y).$$

[3] In contrast to EOT, this principle does not extend to higher-order regularity in general: the potential need not have a C^1 version even for a C^∞ cost.

As c and $\frac{d\nu}{d\tilde{\nu}}$ are bounded, Lemma 5.1 shows that $\|g\|_\infty < \infty$. As ξ is Lipschitz and g is bounded, we see that $\tilde{g}(x,y)$ is Lipschitz in x, uniformly in y. Similarly, $x \mapsto \tilde{c}(x,y)$ is bounded and Lipschitz (uniformly in y) as a product of bounded Lipschitz functions. Thus

$$\Delta(x,x') := \sup_{y\in\mathsf{Y}} |\tilde{c}(x,y) - \tilde{c}(x',y)| + \sup_{y\in\mathsf{Y}} |\tilde{g}(x,y) - \tilde{g}(x',y)| \le L d_{\mathsf{X}}(x,x')$$

for some $L > 0$, where d_{X} denotes the metric on X. Multiplying (2.11) with $\xi(x) > 0$ leads to

$$\int_{\mathsf{Y}} (\xi(x)t + \tilde{g}(x,y) - \tilde{c}(x,y))_+ \, \tilde{\nu}(dy) = 1 \quad \text{for} \quad t = f(x)$$

and then arguing as in the proof of Lemma 2.5 yields

$$\tilde{f}(x') \in [\tilde{f}(x) - \Delta, \tilde{f}(x) + \Delta] \qquad \text{for} \quad \Delta = \Delta(x,x');$$

that is, $|\tilde{f}(x) - \tilde{f}(x')| \le L d_{\mathsf{X}}(x,x')$. Thus $\tilde{f}$ is bounded Lipschitz. As $\xi^{-1} = \frac{d\mu}{d\tilde{\mu}}$ is also bounded Lipschitz, the product $f = \xi^{-1}\tilde{f}$ is again bounded Lipschitz. It follows that f admits a Lipschitz version on X_0 and hence on $\operatorname{spt}\mu$. If X is Hilbert, that version can again be extended to a Lipschitz function on X. □

3 Multiplicity of Potentials

In this section we study the multiplicity of the potentials (f,g). There is always a trivial non-uniqueness, as $(f+\alpha, g-\alpha)$ have the same sum $(f+\alpha)\oplus(g-\alpha) = f \oplus g$ for any $\alpha \in \mathbb{R}$. The main question is whether the potentials are unique up to this additive constant, or if there are *further* degrees of freedom in choosing the potentials.

3.1 Discrete Case

We shall describe the full family of potentials based on the geometry of the support $\operatorname{spt}\pi_*$ of the optimal coupling. Uniqueness typically fails. Let us first study a minimal example to obtain some guidance.

Example 3.1. Let $\mathsf{X} = \mathsf{Y} = \{0,1\}$ and $\mu = \tilde{\mu} = \nu = \tilde{\nu} = \frac{1}{2}(\delta_0 + \delta_1)$. Let $c(x,y) = (2+\gamma)\mathbf{1}_{x\neq y}$ where $\gamma \ge 0$. We claim that the optimal density is

$$Z_*(x,y) = 2\mathbf{1}_{x=y},$$

meaning that the optimal coupling π_* is the uniform measure on the diagonal, and that (f, g) are potentials if and only if

$$\begin{cases} f(0) = \alpha, \quad g(0) = 2 - \alpha, \\ f(1) = \beta, \quad g(1) = 2 - \beta \\ \text{for some } \alpha, \beta \in \mathbb{R} \text{ with } |\alpha - \beta| \le \gamma. \end{cases} \tag{3.1}$$

Indeed, for (f, g) as in (3.1), we have $(f \oplus g)(x, y) = 2$ when $x = y$, whereas $(f \oplus g)(0, 1) = 2 + \alpha - \beta$ and $(f \oplus g)(1, 0) = 2 - \alpha + \beta$. In particular, $(f \oplus g)(x, y) \le 2 + \gamma$ when $x \ne y$. As a result, $(f \oplus g - c)_+ = 2\mathbf{1}_{x=y} =: Z \in \mathcal{Z}$. Now Theorem 2.2 shows that Z is the primal optimizer and that (f, g) are potentials. Conversely, let (f, g) be potentials. Define $\alpha := f(0)$ and $\beta := f(1)$. Then $(f \oplus g - c)_+ = 2\mathbf{1}_{x=y}$ implies that g and α, β must satisfy the conditions in (3.1). In summary, (3.1) describes the family of all potentials when $\gamma \ge 0$.

Next, consider $\gamma \in [-2, 0)$, or equivalently $c(x, y) = \eta \mathbf{1}_{x \ne y}$ with $\eta \in [0, 2)$. Direct calculation shows that $Z_* = (1 + \eta/2)\mathbf{1}_{x=y} + (1 - \eta/2)\mathbf{1}_{x \ne y}$ with constant potentials $(f, g) \equiv (\alpha, 1 + \eta/2 - \alpha)$ for any $\alpha \in \mathbb{R}$. Here the optimal support is the full space, $\operatorname{spt} \pi_* = \mathsf{X} \times \mathsf{Y}$, and correspondingly, the identity $Z_* = (f \oplus g - c)_+ = f \oplus g - c$ determines $f \oplus g$ everywhere. Let us summarize:

(i) For $\gamma > 0$, the potentials are non-unique beyond the trivial shift by a constant. A second degree of freedom arises because the two points $(0, 0)$ and $(1, 1)$ of the support $\operatorname{spt} \pi_*$ do not overlap in terms of X or Y coordinates. In the language introduced below, the singletons $\{(0, 0)\}$ and $\{(1, 1)\}$ are the two components of $\operatorname{spt} \pi_*$, and they are related to the fact that the potentials are given by a two-parameter family (indexed by α, β).

(ii) For $-2 \le \gamma < 0$, we have $\operatorname{spt} \pi_* = \mathsf{X} \times \mathsf{Y}$ which has a single component. The potentials span a one-parameter family $(f + \alpha, g - \alpha)_{\alpha \in \mathbb{R}}$; i.e., are as unique as can be.

(iii) In the boundary case $\gamma = 0$, the support still has two components as in (i), but the two-parameter family degenerates to a one-parameter family since the constraint $|\alpha - \beta| \le 0$ pins down the second parameter to a single value.

The following notion of connectedness is the key to generalizing the above observations to arbitrary discrete problems. It was first introduced by [5] in a different context.

Definition 3.2. Let $E \subset \mathsf{X}\times\mathsf{Y}$ be any subset. Two points $(x,y),(x',y') \in E$ are *connected*, denoted $(x,y) \sim (x',y')$, if there exist $k \in \mathbb{N}_0$ and $(x_i,y_i)_{i=1}^k \in E^k$ such that the points

$$(x,y),(x_1,y),(x_1,y_1),(x_2,y_1),\ldots,(x_k,y_k),(x',y_k),(x',y') \tag{3.2}$$

all belong to E. In that case, $(x_i,y_i)_{i=1}^k$ is called a *path* (in E) from (x,y) to (x',y'). The relation $\sim$ is an equivalence relation on E. The corresponding equivalence classes C are called the *components* of E, and we denote by $\mathcal{C}$ the collection of all components. A set $B \subset E$ is connected (in E) if any two points in B are connected; thus, the components are the maximal connected subsets of E.

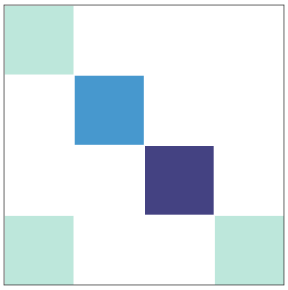

Figure 1: Illustration of a subset E (colored area) of the square $\mathsf{X}\times\mathsf{Y} = [0,1]^2$ with three components (color coding).

For the list (3.2), the crucial property is that only one coordinate is changed in each step. In our notation, the first coordinate changes first, but because a point can be repeated in the list, this entails no loss of generality.

The present notion of connectedness is graph-theoretic and quite different from the topological one. For instance, two connected subsets are connected to one another as soon as they have points with a common X-coordinate (or Y-coordinate); cf. Fig. 1.

Remark 3.3. Denote by $\mathrm{proj}_{\mathsf{X}}$ the canonical projection $(x,y) \mapsto x$. It follows from Definition 3.2 that for any $(x,y) \in E$, there is exactly one component C with $x \in \mathrm{proj}_{\mathsf{X}}(C)$. That is, $\{\mathrm{proj}_{\mathsf{X}}(C)\}_{C\in\mathcal{C}}$ is a partition of $\mathrm{proj}_{\mathsf{X}}(E)$. The analogue holds for Y.

Lemma 3.4. *Let $C \subset E$ be connected. Let $f, f' : \mathrm{proj}_{\mathsf{X}}(C) \to \mathbb{R}$ and $g, g' : \mathrm{proj}_{\mathsf{Y}}(C) \to \mathbb{R}$ be functions such that $f \oplus g = f' \oplus g'$ on C. Then there*

exists $\alpha \in \mathbb{R}$ *such that* $f = f' + \alpha$ *on* $\mathrm{proj}_{\mathsf{X}}(C)$ *and* $g = g' - \alpha$ *on* $\mathrm{proj}_{\mathsf{Y}}(C)$. *In particular, the set of all functions* f', g' *such that* $f \oplus g = f' \oplus g'$ *on* C, *is the one-parameter family* $(f + \alpha, g - \alpha)_{\alpha \in \mathbb{R}}$.

Proof. Fix $(x_0, y_0) \in C$ and define $\alpha := f(x_0) - f'(x_0)$. Let $(x', y') \in C$. By connectedness, there exists a path

$$(x_0, y_0), (x_1, y_0), (x_1, y_1), (x_2, y_1), \ldots, (x_k, y_k), (x', y_k), (x', y')$$

in E. We have $f(x_0) = f'(x_0) + \alpha$. As $f \oplus g = f' \oplus g'$ holds at (x_0, y_0), it follows that $g(y_0) = g'(y_0) - \alpha$. Similarly, $f \oplus g = f' \oplus g'$ holds at (x_1, y_0), hence it follows that $f(x_1) = f'(x_1) + \alpha$. Continuing inductively, we obtain that $f(x') = f'(x') + \alpha$ and $g(y') = g'(y') - \alpha$. □

We can now state the main result of this subsection.

Theorem 3.5. *Let* μ, ν *be finitely supported. Without loss of generality,* $\mathsf{X} = \operatorname{spt} \mu$ *and* $\mathsf{Y} = \operatorname{spt} \nu$. *Let* $C_1, \ldots, C_N$ *be the components (cf. Definition 3.2) of the optimal support* $\operatorname{spt} \pi_*$ *and fix arbitrary potentials* (f, g). *The family of all potentials is given by*

$$\left(f + \sum_{i=1}^{N} \alpha_i \mathbf{1}_{\mathrm{proj}_{\mathsf{X}}(C_i)},\ g - \sum_{i=1}^{N} \alpha_i \mathbf{1}_{\mathrm{proj}_{\mathsf{Y}}(C_i)} \right), \qquad (\alpha_1, \cdots, \alpha_N) \in T$$

where $\{0\} \in T \subset \mathbb{R}^N$ *is the closed convex polytope*

$$T = \{(\alpha_1, \cdots, \alpha_N) \in \mathbb{R}^N : \alpha_i - \alpha_j \leq a_{ij}\}$$

with $a_{ij} \in \mathbb{R}_+$ *given by*

$$a_{ij} := \inf_{(x,y) \in [\mathrm{proj}_{\mathsf{X}}(C_i) \times \mathrm{proj}_{\mathsf{Y}}(C_j)] \setminus \operatorname{spt} \pi_*} c(x, y) - f(x) - g(y) \geq 0. \tag{3.3}$$

Proof. By definition, (f', g') are potentials iff $(f' \oplus g' - c)_+ = (f \oplus g - c)_+$ on $\operatorname{spt} \mu \times \operatorname{spt} \nu$ (which is the whole space $\mathsf{X} \times \mathsf{Y}$ by our assumption). This identity amounts to two requirements:

(a) As $\operatorname{spt} \pi_*$ is the subset where the density is strictly positive, we see that $(f' \oplus g' - c)_+ = (f \oplus g - c)_+$ on $\operatorname{spt} \pi_*$ is equivalent to $f' \oplus g' = f \oplus g$ on $\operatorname{spt} \pi_*$. Using Lemma 3.4 and Remark 3.3, the family of all functions (f', g') satisfying $f' \oplus g' = f \oplus g$ on $\operatorname{spt} \pi_*$ is

$$\left(f + \sum_{i=1}^{N} \alpha_i \mathbf{1}_{\mathrm{proj}_{\mathsf{X}}(C_i)},\ g - \sum_{i=1}^{N} \alpha_i \mathbf{1}_{\mathrm{proj}_{\mathsf{Y}}(C_i)} \right), \qquad \alpha_i \in \mathbb{R},\ 1 \leq i \leq N.$$

(b) The complement of $\operatorname{spt}\pi_*$ is the set where the density is zero, hence on that set, $(f' \oplus g' - c)_+ = (f \oplus g - c)_+$ is equivalent to $f' \oplus g' \le c$. As the sets $\operatorname{proj}_{\mathsf{X}}(C_i) \times \operatorname{proj}_{\mathsf{Y}}(C_j)$ form a partition of $\mathsf{X} \times \mathsf{Y}$, that is further equivalent to $(f + \alpha_i) \oplus (g - \alpha_j) \le c$ on $[\operatorname{proj}_{\mathsf{X}}(C_i) \times \operatorname{proj}_{\mathsf{Y}}(C_j)] \setminus \operatorname{spt}\pi_*$ for all $1 \le i, j, \le N$. This is, in turn, equivalent to $\alpha_i - \alpha_j \le a_{ij}$ where a_{ij} is given by (3.3). We have $a_{ij} \ge 0$ because $(f \oplus g - c)_+ > 0$ on $\operatorname{spt}\pi_*$. □

Theorem 3.5 shows that the potentials form at most an N-parameter family, where N is the number of components of the optimal support. We expect in typical cases that they form a nondegenerate N-parameter family. However, the boundary case $\gamma = 2$ in Example 3.1 illustrates that the polytope can degenerate to a subset of dimension $< N$.

In the continuous case, on the other hand, it is rather intuitive that the constraint $f' \oplus g' \le c$ outside $\operatorname{spt}\pi_*$ becomes a severe restriction: if the involved functions are continuous, then as $f' \oplus g' = c$ on the boundary of $\operatorname{spt}\pi_*$, $|f' \oplus g' - c|$ can be arbitrarily small close to the boundary, leaving no space to vary the α_i (except for the trivial shift). This may serve as an intuition for the situation in Section 3.2 below, where potentials are unique up to the unavoidable additive constant.

3.1.1 Counterexample for Self-Transport

In Example 3.1, the potentials are unique if we impose the constraint $f = g$. However, the following example shows that potentials need not be unique even if the constraint $f = g$ is imposed in a symmetric setting.

Example 3.6. Let $\mathsf{X} = \mathsf{Y} = \{0, 1\}$ and $\mu = \tilde{\mu} = \nu = \tilde{\nu} = \frac{1}{2}(\delta_0 + \delta_1)$. Let $c(x, y) = (2+\gamma)\mathbf{1}_{x=y}$ where $\gamma > 0$. This setting is similar to Example 3.1, but the roles of the diagonal and the off-diagonal are interchanged. In analogy to Example 3.1, the optimal density is $Z_*(x, y) = 2\mathbf{1}_{x\neq y}$, meaning that the optimal coupling π_* is the uniform measure on the *off*-diagonal, and (f, g) are potentials if and only if

$$\begin{cases} f(0) = \alpha, \quad g(0) = 2 - \beta, \\ f(1) = \beta, \quad g(1) = 2 - \alpha, \\ \text{for some } \alpha, \beta \in \mathbb{R} \text{ with } |\alpha - \beta| \le \gamma. \end{cases}$$

Next, we impose the symmetry constraint $f = g$. This is equivalent to the two equations $\alpha = 2 - \beta$ and $\beta = 2 - \alpha$, which are however redundant. Then $|\alpha - \beta| \le \gamma$ is equivalent to $\alpha \in [1 - \gamma/2, 1 + \gamma/2]$ and we conclude that the

family of all symmetric potentials (f, g) is

$$\begin{cases} f(0) = g(0) = \alpha, \\ f(1) = g(1) = 2 - \alpha, \\ \text{for some } \alpha \in [1 - \gamma/2, 1 + \gamma/2]. \end{cases}$$

In particular, f is not unique.

3.2 Continuous Case

Denote by $\mathcal{L}^d$ the Lebesgue measure on $\mathbb{R}^d$. We show that in a regular (yet very standard) setting, the potentials (f, g) are a.s. unique up to an additive constant.

Theorem 3.7. *Let $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$ and $\mu \sim \mathcal{L}^d$ on $\operatorname{spt}\mu$, and assume that $\operatorname{int}\operatorname{spt}\mu$ is connected.*[4] *Moreover, let c be Lipschitz and differentiable on a neighborhood of $\operatorname{spt}\mu \times \operatorname{spt}\nu$. Let either*[5]

(i) $(\tilde{\mu}, \tilde{\nu}) = (\mu, \nu)$ *or*

(ii) $\frac{d\mu}{d\tilde{\mu}}, (\frac{d\mu}{d\tilde{\mu}})^{-1}, \frac{d\nu}{d\tilde{\nu}}, (\frac{d\nu}{d\tilde{\nu}})^{-1}$ *be bounded and Lipschitz, and c be bounded.*

Then the potential f is uniquely determined μ-a.s., up to an additive constant.

We recall from Lemma 2.4 that g is then also uniquely determined ν-a.s., up to constant. In Theorem 3.7, the crucial connectedness assumption is imposed on μ, whereas ν can be quite general. In particular, the result applies in the setting of semi-discrete optimal transport where, typically, μ is given by a nice population density and ν is an empirical measure. The condition on μ is satisfied, for instance, if μ admits a density wrt. $\mathcal{L}^d$ and $\{d\mu/d\mathcal{L}^d > 0\}$ is convex.

Proof of Theorem 3.7. Let (f, g) be potentials. We can extend c to a Lipschitz function on $\mathsf{X} \times \mathsf{Y}$. Recall from Lemma 2.5 (or Lemma 2.6) that f, g admit versions that are defined everywhere on $\mathbb{R}^d$ and Lipschitz. In the remainder of the proof, f and g denote those versions. Note that $Z := (f \oplus g - c)_+$ is a Lipschitz version of the density of the optimal coupling π_*. In particular,

[4] This refers to the usual topological connectedness, not Definition 3.2.

[5] The purpose of this condition is to ensure that any potential is Lipschitz. We could instead assume the latter directly, or assert that uniqueness holds within the class of Lipschitz potentials.

the set $E = \{Z > 0\}$ is open and satisfies $\pi_*(E) = 1$. Moreover, $E \subset \operatorname{spt} \pi_*$ and hence $\operatorname{spt} \pi_* = \overline{E} = \{Z \geq 0\}$. Finally, it is not hard to check that the open set $\operatorname{proj}_{\mathsf{X}} E$ satisfies

$$\mu(\operatorname{proj}_{\mathsf{X}} E) = 1 \quad \text{and} \quad \operatorname{spt} \mu = \overline{\operatorname{proj}_{\mathsf{X}} E}. \tag{3.4}$$

Let $\operatorname{dom} \nabla f$ denote the set where f is differentiable. By Rademacher's theorem, the complement of $\operatorname{dom} \nabla f$ is $\mathcal{L}^d$-null. On the other hand, it follows from (3.4) and $\mu \sim \mathcal{L}^d$ on $\operatorname{spt} \mu$ that $\mathcal{L}^d(\operatorname{spt} \mu \setminus \operatorname{proj}_{\mathsf{X}} E) = 0$. In summary, $\Lambda := \operatorname{dom} \nabla f \cap \operatorname{proj}_{\mathsf{X}} E \subset \operatorname{spt} \mu$ has full $\mathcal{L}^d$-measure within $\operatorname{spt} \mu$

Next, we check that ∇f is uniquely determined on Λ. Indeed, let $x_0 \in \Lambda$, then $(x_0, y_0) \in E$ for some y_0. We have

$$Z(x, y) = f(x) + g(y) - c(x, y), \qquad (x, y) \in B_r(x_0, y_0)$$

for small $r > 0$, due to the definition of E. (Here $B_r(z)$ denotes the open ball of radius r around z.) Differentiation thus yields that $\nabla f(x_0) = \nabla_x Z(x_0, y_0) + \nabla_x c(x_0, y_0)$. The right-hand side is uniquely determined. In summary, we have shown that f is a Lipschitz function with ∇f uniquely determined $\mathcal{L}^d$-a.e. on the open and connected set $\operatorname{int} \operatorname{spt} \mu$. This implies that f is uniquely determined up to additive constant on $\operatorname{int} \operatorname{spt} \mu$ (see, e.g., [30, Formula 2]). As $\operatorname{proj}_{\mathsf{X}} E$ is open, (3.4) implies $\mu(\operatorname{int} \operatorname{spt} \mu) = 1$, completing the proof. □

Remark 3.8. In Theorem 3.7, suppose that we are in the symmetric setting of self-transport and we additionally impose that $f = g$. We readily see that this constraint pins down the additive constant in Theorem 3.7 and hence gives uniqueness for f μ-a.s.

Theorem 3.7 is a satisfactory result in a regular setting and covers most applications of interest. In the remainder of the section, we comment briefly on subtleties that occur in a possible extension of the analysis that we performed in the discrete case to non-discrete (and irregular) cases. First, while the definition of components applies to arbitrary sets, two almost-surely equal sets may have substantially different components. Second, some components may carry no mass, making them negligible. Next, we highlight the first aspect by detailing a continuous version of Example 3.1 and comparing with another version.

Example 3.9. Let $\mathsf{X} = \mathsf{Y} = [0, 1]$ and $\mu = \tilde{\mu} = \nu = \tilde{\nu} = \mathcal{L}^1|_{[0,1]}$ the uniform measure. Let $E := [0, 1/2)^2 \cup (1/2, 1]^2$ be the "block diagonal" and $E^c = [0, 1]^2 \setminus E$. Let $c(x, y) = (2 + \gamma)\mathbf{1}_{E^c}$ where $\gamma > 0$. We claim that

$$Z_*(x, y) = 2\mathbf{1}_E$$

is a version of the optimal density, meaning that the optimal coupling π_* is the uniform measure on E, and that (f, g) are potentials if and only if $\mathcal{L}^1$-a.s.,

$$\begin{cases} f(x) = \alpha, \quad g(y) = 2 - \alpha, & x, y \in [0, 1/2), \\ f(x) = \beta, \quad g(y) = 2 - \beta, & x, y \in (1/2, 1], \\ \text{for some } \alpha, \beta \in \mathbb{R} \text{ with } |\alpha - \beta| \leq \gamma. \end{cases} \tag{3.5}$$

Indeed, for such (f, g), we have $f \oplus g = 2$ on E, whereas $(f \oplus g) \in \{2 + \alpha - \beta, 2 - \alpha + \beta\}$ on E^c. In particular, $f \oplus g \leq 2 + \gamma$ on E^c. As a result, $(f \oplus g - c)_+ = 2\mathbf{1}_E = Z_*$. Now Theorem 2.2 shows that Z_* is the primal optimizer and that (f, g) are potentials. Conversely, let (f, g) be potentials. As $f(x) + g(y) = 2$ for $\mathcal{L}^2$-a.e. $(x, y) \in [0, 1/2)^2$, it follows that $f(x) = 2 - 2\int_0^{1/2} g(y)\, dy =: \alpha$ for μ-a.e. $x \in [0, 1/2)$ and similarly $g(y) = 2 - 2\int_0^{1/2} f(x)\, dx = 2 - \alpha$ for ν-a.e. $y \in [0, 1/2)$. Analogously, $f = \beta$ μ-a.s. on $(1/2, 1]$ and $g = 2 - \beta$ ν-a.s. on $(1/2, 1]$, for some $\beta \in \mathbb{R}$. We then have $f \oplus g = 2 + \alpha - \beta$ P-a.s. on $[0, 1/2) \times (1/2, 1]$ and $f \oplus g = 2 - \alpha + \beta$ P-a.s. on $(1/2, 1] \times [0, 1/2)$. In order to satisfy $(f \oplus g - c)_+ = Z_* = 2\mathbf{1}_E$ P-a.s., we must have $|\alpha - \beta| \leq \gamma$.

In Example 3.9, the number of components of $\{Z_* > 0\}$ correctly describes the degrees of freedom in choosing potentials. To achieve this, we picked a "good" version of Z_*—the following shows that the number of components of $\{Z_* > 0\}$ can depend on the version of the density Z_*. While the "support" of π_* has a canonical definition in the discrete case, that is not the case here (and the topological support is not a good choice).

Remark 3.10. In Example 3.9, the family in (3.5) are precisely the functions f, g such that $(f \oplus g - c)_+ = 2\mathbf{1}_E$ *everywhere* on $[0, 1]^2$ (without exceptional nullsets). Next, consider the same example but define $\tilde{E} = [0, 1/2]^2 \cup [1/2, 1]^2$, which is the topological support of π_*. Obviously $\tilde{E} = E$ P-a.s., hence $Z_*(x, y) = 2\mathbf{1}_{\tilde{E}}$ is another version of the optimal density. But by contrast with the above, $(f \oplus g - c)_+ = 2\mathbf{1}_{\tilde{E}}$ *everywhere* on $[0, 1]^2$ if and only if $f \equiv \alpha$ and $g \equiv 2 - \alpha$ for some $\alpha \in \mathbb{R}$. To wit, the extra degree of freedom represented by β has been eliminated, as $f(1/2) = \alpha$ and $f(1/2) = \beta$ now imply $\alpha = \beta$ (or similarly for g). The same happens if we replace $\tilde{E}$ by $[0, 1/2] \times [0, 1/2) \cup [1/2, 1] \times (1/2, 1]$ or the symmetric counterpart. Indeed, $\tilde{E}$ and the latter two sets have a single component.

Due to the closure in its definition, the topological support is a rather large set carrying the measure, whereas for the current purpose we require a

small set, or more precisely, a set with the *maximum number of non-negligible components.*

4 Sparsity for Small Regularization

Let μ, ν be compactly supported probability measures on $\mathbb{R}^d$. We specialize our general setting to $\mathsf{X} = \operatorname{spt}\mu$ and $\mathsf{Y} = \operatorname{spt}\nu$, with cost $c(x,y) = \|x-y\|^2$ the squared Euclidean distance. We further assume that $\mu \ll \mathcal{L}^d$ and that $\operatorname{int}\operatorname{spt}\mu$ is connected, and choose $(\tilde{\mu}, \tilde{\nu}) = (\mu, \nu)$ for simplicity (alternately, we can impose the conditions of Lemma 2.6).

In this setting of quadratic cost, it is well known that the unregularized transport problem

$$\mathscr{P}_0 = \inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi \tag{4.1}$$

has a unique optimal coupling π_0, given by Brenier's map. See, e.g., [34] for background. In particular, π_0 is concentrated on the graph of a function $T : \mathbb{R}^d \to \mathbb{R}^d$, and hence "sparse" (as sparse as a coupling can be). Regularity results for T are known under conditions on the marginals.

Let $\pi_\varepsilon \in \Pi(\mu,\nu)$ be the optimal coupling of the quadratically regularized problem with regularization parameter $\varepsilon > 0$ (cf. Remark 2.3),

$$\mathscr{P}_\varepsilon = \inf_{\pi \in \Pi(\mu,\nu)} \int c \, d\pi + \frac{\varepsilon}{2} \|d\pi/dP\|^2_{L^2(P)}.$$

The next result formalizes and establishes that π_ε is sparse for small ε, by showing that $\operatorname{spt}\pi_\varepsilon$ is contained in a small neighborhood of the sparse set $\operatorname{spt}\pi_0$. To the best of our knowledge, it is the first theoretical result showing sparsity of quadratically regularized transport in a continuous setting.

Theorem 4.1 (Sparsity for quadratic cost)**.** *Let U be an open neighborhood of $\operatorname{spt}\pi_0$. Then $\operatorname{spt}\pi_\varepsilon \subset U$ for all sufficiently small $\varepsilon > 0$.*

Proof. We recall that in the present setting, the unregularized transport problem (4.1) admits a unique optimal coupling π_0 and a unique (up to constant) Kantorovich potential $f : \mathsf{X} \to \mathbb{R}$ that is Lipschitz continuous. Fix $x_0 \in \mathsf{X}$; then we may normalize $f(x_0) = 0$ to have uniqueness. Let g be the c-concave conjugate of f, so that (f,g) is the solution of the dual optimal transport problem

$$\mathscr{D}_0 = \sup_{(f,g) \in C(\mathsf{X}) \times C(\mathsf{Y}) \colon f \oplus g \le c} \int f \, d\mu + \int g \, d\nu. \tag{4.2}$$

It is known that $\operatorname{spt}\pi_0 = \{f \oplus g = c\}$. Indeed, $\pi_0\{f \oplus g = c\} = 1$ holds for general costs. The inclusion $\operatorname{spt}\pi_0 \supset \{f \oplus g = c\}$, which is crucial below, follows because the section $\{f(x) + g(\cdot) = c(x,\cdot)\} \subset \mathsf{Y}$ is a singleton for μ-a.e. $x \in \mathsf{X}$ (as the subdifferential of an a.e. differentiable function).

Let $(f_\varepsilon, g_\varepsilon)$ be the rescaled potentials as defined in Remark 2.3. By Lemma 2.5 (iii), f_ε can be chosen to be L-Lipschitz, where L is the Lipschitz constant of c on the compact set $\mathsf{X} \times \mathsf{Y}$ (note that $f_\varepsilon/\varepsilon$ is the potential for c/ε without rescaling). We may normalize $f_\varepsilon(x_0) = 0$ and then f_ε is also bounded uniformly in ε. The Arzela–Ascoli theorem shows that given a sequence $\varepsilon_n \to 0$, a subsequence of (f_{ε_n}) converges uniformly to some limit f_*. After passing to another subsequence (still denoted ε_n), we also have uniform convergence $g_{\varepsilon_n} \to g_*$. We show below that $(f_*, g_*) = (f, g)$. In particular, the uniqueness of the Kantorovich potential then implies that $(f_\varepsilon, g_\varepsilon) \to (f, g)$ uniformly for $\varepsilon \to 0$.

Let U be an open neighborhood of $\operatorname{spt}\pi_0$. As $\operatorname{spt}\pi_0 = \{f \oplus g = c\}$ and f, g, c are continuous and $\mathsf{X} \times \mathsf{Y}$ is compact, there exists $\delta > 0$ such that $\{f \oplus g \geq c - \delta\} \subset U$. Recall that the density of π_ε has the form

$$Z_\varepsilon = \left(\frac{f_\varepsilon \oplus g_\varepsilon - c}{\varepsilon}\right)_+. \tag{4.3}$$

In view of the uniform convergence $(f_\varepsilon, g_\varepsilon) \to (f, g)$, there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$,

$$\operatorname{spt}\pi_\varepsilon = \overline{\{Z_\varepsilon > 0\}} = \overline{\{f_\varepsilon \oplus g_\varepsilon > c\}} \subset \{f \oplus g \geq c - \delta\} \subset U.$$

It remains to prove $(f_*, g_*) = (f, g)$. To that end, we show that (f_*, g_*) solves the dual problem (4.2). We first verify that (f_*, g_*) is in the dual domain; i.e., that $f_* \oplus g_* \leq c$. Suppose that $f_*(x) + g_*(y) > c(x,y)$ at some $(x,y) \in \mathsf{X} \times \mathsf{Y}$. Then by continuity, $\{f_* \oplus g_* > c\}$ includes a compact neighborhood B of (x,y). In view of (4.3), it follows that $Z_\varepsilon \to \infty$ uniformly on B as $\varepsilon \to 0$. As $(\mu \otimes \nu)(B) > 0$ due to $(x,y) \in \mathsf{X} \times \mathsf{Y} = \operatorname{spt}(\mu \otimes \nu)$, this contradicts the fact that $\int Z_\varepsilon \, d(\mu \otimes \nu) = 1$ for all $\varepsilon > 0$.

Second, we verify that (f_*, g_*) is optimal for (4.2). By duality we have $\mathscr{D}_\varepsilon = \mathscr{P}_\varepsilon$ and $\mathscr{P}_0 = \mathscr{D}_0$, and clearly $\mathscr{P}_0 \leq \mathscr{P}_\varepsilon$ as the quadratic penalty is nonnegative. Thus also $\mathscr{D}_0 \leq \mathscr{D}_\varepsilon$, which by (2.10) yields

$$\begin{aligned}\mathscr{D}_0 \leq \mathscr{D}_{\varepsilon_n} &= \int f_{\varepsilon_n} \, d\mu + \int g_{\varepsilon_n} \, d\nu - \frac{\varepsilon_n}{2} \int \left(\frac{f_{\varepsilon_n} \oplus g_{\varepsilon_n} - c}{\varepsilon_n}\right)_+^2 dP \\ &\leq \int f_{\varepsilon_n} \, d\mu + \int g_{\varepsilon_n} \, d\nu \to \int f_* \, d\mu + \int g_* \, d\nu.\end{aligned}$$

Hence (f_*, g_*) solves (4.2) and the proof is complete. □

Remark 4.2. (a) Theorem 4.1 has a trivial analogue in the discrete setting, where it is known that $\pi_\varepsilon = \pi_0$ for small $\epsilon > 0$, for a certain optimal transport π_0. See, e.g., [8]. This observation goes back to [23] for more general linear programs with quadratic regularization.

(b) As $\pi_\varepsilon \to \mu \otimes \nu$ for $\varepsilon \to \infty$, sparsity certainly requires ε to be small in some sense. One can hope for a quantitative version of Theorem 4.1, stating that $\operatorname{spt} \pi_\varepsilon$ is in a δ-neighborhood of $\operatorname{spt} \pi_0$, where $\delta = \delta(\varepsilon)$ has an explicit dependence on ε. This problem is left for future research. The proof given above merely uses the straightforward (qualitative) convergence of the potentials; see also [17, 25] for related results on the convergence of potentials for entropic regularization. The proof does extend to more general costs c: the key property is that the Kantorovich dual $f \oplus g$ "detaches" from the cost c outside the support of the optimal coupling π_0. See [6, 22] for recent developments in this direction.

5 Proof of Theorem 2.2(i)–(iii)

We first take care of generalities—primal existence, automatic integrability and weak duality—which will reduce the proof of Theorem 2.2 to the main task, namely to show that the optimal density is of the form $Z_* = (f \oplus g - c)_+$.

5.1 Primal Existence

The existence and uniqueness of the primal optimizer is straightforward and well known. We detail the argument for later use.

Proof of Theorem 2.2(ii). Recall (2.3). For any $Z \in \mathcal{Z}$, clearly

$$\begin{aligned} \int cZ\,dP + \frac{1}{2}\|Z\|^2_{L^2(P)} &= \int (cZ + \tfrac{1}{2}Z^2)\,dP \\ &= \tfrac{1}{2}\|c+Z\|^2_{L^2(P)} - \tfrac{1}{2}\|c\|^2_{L^2(P)}, \end{aligned}$$

where the last term is a finite constant independent of Z. In particular,

$$\mathscr{P} = \inf_{Z \in \mathcal{Z}} \tfrac{1}{2}\|c+Z\|^2_{L^2(P)} - \tfrac{1}{2}\|c\|^2_{L^2(P)}. \tag{5.1}$$

The subset $\mathcal{Z} \subset L^2(P)$ is closed, convex, and nonempty as $d(\mu \otimes \nu)/dP \in \mathcal{Z}$ due to (2.1). The claim thus follows from the existence, uniqueness and characterization of the Hilbert space projection onto $\mathcal{Z}$. □

5.2 Integrability Properties

This subsection establishes the automatic integrability of potentials. We first derive a simple lower bound similar to a result in [21]; see (2.3) for notation.

Lemma 5.1 (Lower bound)**.** *Let $(f \oplus g - c)_+$ be the P-density of a coupling $\pi \in \Pi(\mu, \nu)$, for some measurable functions $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$. Then*

$$(f - c_1)_- \in L^\infty(\tilde{\mu}), \qquad (g - c_2)_- \in L^\infty(\tilde{\nu}).$$

In particular, $f_- \in L^1(\tilde{\mu}) \cap L^1(\mu)$ and $g_- \in L^1(\tilde{\nu}) \cap L^1(\nu)$. If $c_- \in L^\infty(P)$, then $(f_-, g_-) \in L^\infty(\tilde{\mu}) \times L^\infty(\tilde{\nu})$.

Moreover, if $c \in L^\infty(P)$, then $(f - \frac{d\mu}{d\tilde{\mu}})_+ \in L^\infty(\tilde{\mu})$ and $(g - \frac{d\mu}{d\tilde{\mu}})_+ \in L^\infty(\tilde{\nu})$.

Proof. Recall (2.11) and (2.3). For $\tilde{\mu}$-a.e. $x \in \mathsf{X}$,

$$\begin{aligned} F_x(t) &= \int (t + g(y) - c(x,y))_+ \, \tilde{\nu}(dy) \\ &\le \int ((g(y) - c_2(y))_+ - (t - c_1(x)_-)_+ \, \tilde{\nu}(dy) \\ &\le \int h(y) \mathbf{1}_{h(y) > (t - c_1(x))_-} \, \tilde{\nu}(dy) = \Phi((t - c_1(x))_-) \end{aligned}$$

where $h(y) := (g(y) - c_2(y))_+$ and $\Phi(\alpha) := \int h(y) \mathbf{1}_{h(y) > \alpha} \, \tilde{\nu}(dy)$. As $c_2 \in L^1(\tilde{\nu})$ and, necessarily, $g_+ \in L^1(\tilde{\nu})$, we have $h \in L^1(\tilde{\nu})$. In particular, Φ is a nonincreasing function with $\lim_{\alpha \to \infty} \Phi(\alpha) = 0$. Defining the generalized inverse $\Phi^{-1}(v) := \sup\{u : \Phi(u) \ge v\}$, it follows with $\frac{d\mu}{d\tilde{\mu}}(x) = F_x(t)$ for $t = f(x)$ that

$$(f(x) - c_1(x))_- \le \Phi^{-1}\left(\frac{d\mu}{d\tilde{\mu}}(x)\right) \qquad \text{for } \tilde{\mu}\text{-a.e. } x \in \mathsf{X}.$$

As we have assumed in (2.1) that $\frac{d\mu}{d\tilde{\mu}}$ is a.s. uniformly bounded away from zero, this shows that $(f - c_1)_- \in L^\infty(\tilde{\mu})$. The assertion on g_- follows similarly. Suppose that $c \in L^\infty(P)$. Then

$$\frac{d\mu}{d\tilde{\mu}}(x) = F_x(t) = \int (t + g(y) - c(x,y))_+ \, \tilde{\nu}(dy) \ge t - \|g_-\|_{L^\infty(\tilde{\nu})} - \|c\|_{L^\infty(P)}$$

for $t = f(x)$ implies $f(x) - \frac{d\mu}{d\tilde{\mu}}(x) \le \|g_-\|_{L^\infty(\tilde{\nu})} + \|c\|_{L^\infty(P)}$, and it was shown above that $\|g_-\|_{L^\infty(\tilde{\nu})}$ is finite when $c \in L^\infty(P)$. □

The next lemma will entail in particular that the algebraic form $(f \oplus g - c)_+$ already identifies the optimal coupling, without any (a priori) integrability condition on (f, g). As a consequence, the system of equations in Theorem 2.2 (d) fully characterizes potentials, without a separate condition.

Lemma 5.2 (Automatic integrability)**.** *Let $Z = (f \oplus g - c)_+$ be the P-density of a coupling $\pi \in \Pi(\mu, \nu)$, for some measurable functions $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$. Then*

(i) $(f_+, g_+) \in L^2(\tilde{\mu}) \times L^2(\tilde{\nu})$,

(ii) $Z \in \mathcal{Z}$*; i.e.,* $Z \in L^2(P)$,

(iii) $(f, g) \in L^1(\mu) \times L^1(\nu)$.

Proof. Recall from (2.8) that for $\tilde{\mu}$-a.e. $x \in \mathsf{X}$,

$$\begin{aligned} \frac{d\mu}{d\tilde{\mu}}(x) &= \int (f(x) + g(y) - c(x,y))_+ \, \tilde{\nu}(dy) \\ &\geq f(x) - \int g_- \, d\tilde{\nu} - \int c(x,y) \, \tilde{\nu}(dy). \end{aligned} \tag{5.2}$$

We have $\int g_- \, d\tilde{\nu} < \infty$ by Lemma 5.1, and the last term is in $L^2(\tilde{\mu})$ by Jensen:

$$\int \left(\int c(x,y) \, \tilde{\nu}(dy) \right)^2 \tilde{\mu}(dx) \leq \iint c(x,y)^2 \, \tilde{\nu}(dy) \tilde{\mu}(dx) = \|c\|^2_{L^2(P)} < \infty.$$

Hence (5.2) and (2.1) establish that f is bounded from above by a function in $L^2(\tilde{\mu})$; that is, $f_+ \in L^2(\tilde{\mu})$. Analogously, $g_+ \in L^2(\tilde{\nu})$, completing (i).

As $c \in L^2(P)$, it follows that $Z = (f(x)+g(y)-c(x,y))_+ \in L^2(P)$, which is (ii). The Cauchy–Schwarz inequality and (2.1) yield $L^2(\tilde{\mu}) \subset L^1(\mu)$ and $L^2(\tilde{\nu}) \subset L^1(\nu)$. Hence (i) implies $(f_+, g_+) \in L^1(\mu) \times L^1(\nu)$. On the other hand, $(f_-, g_-) \in L^1(\mu) \times L^1(\nu)$ by Lemma 5.1, showing (iii). □

5.3 Weak Duality

The next lemma provides weak duality and reduces strong duality to existence of potentials.

Lemma 5.3. *(i) For all $Z \in \mathcal{Z}$ and $(f, g) \in L^1(\mu) \times L^1(\nu)$,*

$$\int \left(cZ + \tfrac{1}{2} Z^2 \right) \, dP \geq \int f \, d\mu + \int g \, d\nu - \frac{1}{2} \int (f \oplus g - c)_+^2 \, dP,$$

with equality holding iff $Z = (f \oplus g - c)_+$ P-a.s. In particular, $\mathscr{P} \geq \mathscr{D}$.

(ii) Let $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$ be measurable and suppose that $Z := (f \oplus g - c)_+$ is the density of a coupling. Then

(a) $Z \in \mathcal{Z}$ *and* $(f, g) \in L^1(\mu) \times L^1(\nu)$,

(b) *there is no duality gap:* $\mathscr{P} = \mathscr{D}$,

(c) Z *is optimal for the primal problem* (2.5),

(d) (f, g) *is optimal for the dual problem* (2.6).

(iii) *Conversely, suppose that* $\mathscr{P} = \mathscr{D}$. *If* $(f, g) \in L^1(\mu) \times L^1(\nu)$ *is optimal for the dual problem* (2.6), *then* $(f \oplus g - c)_+$ *is in* $\mathcal{Z}$ *and optimal for the primal problem* (2.5).

Proof. Consider $Z \in \mathcal{Z}$ and $(f, g) \in L^1(\mu) \times L^1(\nu)$. Then

$$\int \left(cZ + \tfrac{1}{2}Z^2\right) dP = \int f \oplus g \, d(\mu \otimes \nu) - \int \left((f \oplus g - c)Z - \tfrac{1}{2}Z^2\right) dP$$

as $Z\,dP \in \Pi(\mu, \nu)$. Note that $[0, \infty) \ni z \mapsto az - z^2/2$ has a unique maximum at $z = a_+$ with maximum value $a_+^2/2$. Using this pointwise with $a = f \oplus g - c$, we deduce

$$\int \left(cZ + \tfrac{1}{2}Z^2\right) dP \geq \int f \oplus g \, d(\mu \otimes \nu) - \frac{1}{2} \int (f \oplus g - c)_+^2 \, dP$$

with equality holding if and only if $Z = (f \oplus g - c)_+$ P-a.s. This shows (i). In view of the automatic integrability shown in Lemma 5.2, (ii) follows from (i). To see (iii), consider $Z := Z_* \in \mathcal{Z}$ in the left-hand side of in (i) and dual optimizers (f, g) on the right-hand side. As $\mathscr{P} = \mathscr{D}$ was assumed, the assertion of (i) on equality implies $Z_* = (f \oplus g - c)_+$ P-a.s. □

5.4 Construction of Potentials

In view of Lemma 5.3, our main task is to construct measurable functions $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$ such that $(f \oplus g - c)_+ \in \mathcal{Z}$. More precisely, we shall show directly that the optimal density Z_* is of that form.

From a convex programming point of view, the marginal constraints μ, ν in the primal problem (2.5) correspond to infinitely many equality constraints; namely, $\int \phi \, d\pi = \int \phi \, d\mu$ whenever ϕ is a bounded measurable function on X, and similarly for ν. As the spaces X, Y are separable, countably many test functions ϕ are sufficient to encode the marginals. Our plan is to approximate the primal problem (2.5) with auxiliary problems having finitely many constraints (that can be solved by elementary arguments) and then pass to the limit (which is more delicate).

The problems with finitely many constraints are described in the next lemma. We emphasize that $\mathcal{Z}_n$ consists of densities of measures that are not necessarily probability measures.

Lemma 5.4. *Fix $n \in \mathbb{N}$ and bounded measurable functions $\phi_i : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$, $1 \le i \le n$ with $\int \phi_i \, d(\mu \otimes \nu) = 0$. Let*

$$\mathcal{Z}_n = \left\{ Z \in L^2(P) : Z \ge 0, \int \phi_i Z \, dP = 0, \, 1 \le i \le n \right\}.$$

There is a unique solution $Z_n \in \mathcal{Z}_n$ of

$$\inf_{Z \in \mathcal{Z}_n} \int cZ \, dP + \frac{1}{2} \|Z\|^2_{L^2(P)} \tag{5.3}$$

and Z_n is characterized within $\mathcal{Z}_n$ by being of the form

$$Z_n = (b_1\phi_1 + \cdots + b_n\phi_n - c)_+ \qquad \text{for some } b_i \in \mathbb{R}. \tag{5.4}$$

Proof. As $\mathcal{Z}_n \subset L^2(P)$ is convex, closed and nonempty, existence and uniqueness of the minimizer $Z_n = \arg\min_{Z \in \mathcal{Z}_n} \|Z + c\|_{L^2(P)}$ follow by Hilbert space projection as in the proof of Theorem 2.2 (ii). Next, we argue as in the proof of Lemma 5.3 (i), with $f \oplus g$ replaced by $b \cdot \Phi$ where $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$ and $\Phi = (\phi_1, \ldots, \phi_n)$. Noting that $\int b \cdot \Phi \, d(\mu \otimes \nu) = 0$, we obtain

$$\int \left(cZ + \tfrac{1}{2} Z^2 \right) dP \ge \frac{1}{2} \int (b \cdot \Phi - c)_+^2 \, dP \quad \text{for all} \quad b \in \mathbb{R}^n,$$

with equality holding if and only if $Z = (b \cdot \Phi - c)_+$ P-a.s. for some $b \in \mathbb{R}^n$. As a result, we only need to prove that there exists $Z \in \mathcal{Z}_n$ of the form $Z = (b \cdot \Phi - c)_+$. To that end, we first show that the problem

$$\inf_{b \in \mathbb{R}^n} G(b), \qquad G(b) := \int (b \cdot \Phi - c)_+^2 \, dP$$

admits a minimizer b_*. Note that $G : \mathbb{R}^n \to \mathbb{R}$ is convex and continuous. By projecting onto the orthogonal complement of $\{b \in \mathbb{R}^n : b \cdot \Phi = 0 \ P\text{-a.s.}\}$, we can reduce to a situation where $b \cdot \Phi = 0$ P-a.s. only for $b = 0$. For $b \ne 0$ we then have $P\{b \cdot \Phi \ne 0\} > 0$.

We claim that this implies $P\{b \cdot \Phi > 0\} > 0$. Indeed, if not, then $b \cdot \Phi \le 0$ P-a.s. and $P\{b \cdot \Phi < 0\} > 0$. As $\mu \otimes \nu \sim P$, it follows that $b \cdot \Phi \le 0$ $(\mu \otimes \nu)$-a.s. and $(\mu \otimes \nu)\{b \cdot \Phi < 0\} > 0$. Thus $\int b \cdot \Phi \, d(\mu \otimes \nu) < 0$, contradicting that $\int b \cdot \Phi \, d(\mu \otimes \nu) = 0$.

Clearly $P\{b \cdot \Phi > 0\} > 0$ implies the radial coercivity condition

$$\lim_{\alpha \to \infty} G(\alpha b) = \infty, \qquad 0 \ne b \in \mathbb{R}^n.$$

Any convex, lower semicontinuous function $G : \mathbb{R}^n \to \mathbb{R}$ satisfying this condition attains its minimum [15, Lemma 3.5, p. 126].

Let b_* be a minimizer. Note that $z \mapsto z_+^2$ is continuously differentiable with derivative $2z_+$, and recall that Φ is bounded. Differentiation under the integral yields the first-order condition

$$\int \phi_i (b_* \cdot \Phi - c)_+ \, dP = 0, \qquad 1 \le i \le n.$$

This shows that $(b_* \cdot \Phi - c)_+ \in \mathcal{Z}_n$, as desired. □

Lemma 5.5. *Let $Z_* \in \mathcal{Z}$ be the primal optimizer. There exist bounded measurable functions $f_n : \mathsf{X} \to \mathbb{R}$ and $g_n : \mathsf{Y} \to \mathbb{R}$, $n \ge 1$ such that*

$$Z_* = \lim_{n\to\infty} (f_n \oplus g_n - c)_+ \qquad P\text{-a.s. and in } L^2(P).$$

Proof. As $L^1(\mu)$ is separable, there are bounded measurable functions $\phi_k^\mu : \mathsf{X} \to \mathbb{R}$, $k \ge 1$ such that $\rho \in \mathcal{P}(\mathsf{X} \times \mathsf{Y})$ satisfies $\int \phi_k^\mu(x)\, \rho(dx, dy) = 0$ for all $k \ge 1$ if and only if the first marginal of ρ is μ. Similarly, there are functions $\phi_k^\nu : \mathsf{Y} \to \mathbb{R}$ for ν. Let $\phi_{2i-1}(x, y) := \phi_i^\mu(x)$ and $\phi_{2i}(x, y) := \phi_i^\nu(x)$, so that $\rho \in \Pi(\mu, \nu)$ if and only if $\int \phi_i \, d\rho = 0$ for all $i \ge 1$. Define $Z_n, \mathcal{Z}_n$ as in Lemma 5.4 and note that Z_n is of the desired form $Z_n = (f_n \oplus g_n - c)_+$; namely, f_n is a linear combination of $(\phi_k^\mu)_{k\le n}$ and g_n is a linear combination of $(\phi_k^\nu)_{k\le n}$. Thus, it suffices to show that $Z_n \to Z_*$ in $L^2(P)$.

Note that $\mathcal{Z}_n, \mathcal{Z} \subset L^2(P)$ are closed and convex, $\mathcal{Z}_n \supset \mathcal{Z}_{n+1}$, and $\mathcal{Z} = \cap_{n\ge1} \mathcal{Z}_n$. Moreover, Z_n and Z_* are the projections of $-c$ onto $\mathcal{Z}_n$ and $\mathcal{Z}$ in $L^2(P)$, respectively:

$$Z_n = \operatorname*{arg\,min}_{Z\in\mathcal{Z}_n} \|Z + c\|, \qquad Z_* = \operatorname*{arg\,min}_{Z\in\mathcal{Z}} \|Z + c\|,$$

where $\|\cdot\| = \|\cdot\|_{L^2(P)}$. It is a general fact of Hilbert spaces that such nested projections converge; i.e., $Z_n \to Z_*$ in $L^2(P)$. One way of obtaining that fact is to use the parallelogram law for $Z_m + c$ and $Z_n + c$, giving

$$\tfrac{1}{4}\|Z_m - Z_n\|^2 = \frac{\|Z_m + c\|^2}{2} + \frac{\|Z_n + c\|^2}{2} - \left\| \frac{Z_m + Z_n}{2} + c \right\|^2 =: a_{m,n}.$$

As $(Z_m + Z_n)/2 \in \mathcal{Z}_{m\wedge n}$, we see that $\limsup_{m,n\to\infty} a_{m,n} \le 0$, hence (Z_n) is a Cauchy sequence in $L^2(P)$. Its limit Z must lie in $\mathcal{Z}$ and then $\|Z + c\| = \lim_n \|Z_n + c\| \le \|Z_* + c\|$ shows $Z = Z_*$, where the inequality is due to $\mathcal{Z}_n \supset \mathcal{Z}$. □

To pass to the limit of $f_n \oplus g_n$, we shall use the following result.

Lemma 5.6. *Consider two sequences $f_n : \mathsf{X} \to \mathbb{R}$ and $g_n : \mathsf{Y} \to \mathbb{R}$, $n \geq 1$ of measurable functions, and $\pi \in \mathcal{P}(\mu, \nu)$. Suppose that*

$$\lim_{n\to\infty} f_n \oplus g_n = h \quad \pi\text{-a.s.}$$

for a measurable function $h : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$. Then there are functions $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$ such that $h = f \oplus g$ π-a.s. If $\pi \ll P$, the functions f, g can be chosen to be measurable.

The first part of Lemma 5.6 is [31, Proposition 2.1], the second is [14, Proposition 3.19]. The latter assumes $\pi \ll \pi_{\mathsf{X}} \otimes \pi_{\mathsf{Y}}$, where $\pi_{\mathsf{X}}, \pi_{\mathsf{Y}}$ are the marginals of π, which is equivalent to $\pi \ll P$.[6] We comment briefly on the proof of Lemma 5.6 in the subsequent remark, but give a detailed proof of its (more difficult) symmetric version in Lemma 6.3 below.

Remark 5.7. Results like Lemma 5.6 are surprisingly subtle; the difficulty is that the convergence of $f_n \oplus g_n$ only holds on a subset $E \subset \mathsf{X} \times \mathsf{Y}$. First of all, note that convergence of $f_n \oplus g_n$ does not imply a separate convergence of f_n and g_n. If $E = \mathsf{X} \times \mathsf{Y}$, the limits of f_n and g_n do exist after a single normalization (e.g., $f_n(x_0) = 0$ for some fixed $x_0 \in \mathsf{X}$) to pin down the familiar additive constant. But if E is sparse (as will typically be the case in our application), we have seen in Section 3.1 that f_n can be shifted by a different constant on each component of E. While [31, 14] do not use the concept of connectedness, their proofs roughly boil down to choosing a normalization for each component of E. Because there are uncountably many components in general, the measurability of the resulting function is not guaranteed. In fact, a counterexample due to N. Gantert (reported in [31]) shows that even when a limit $f \oplus g$ exists, it can happen that there is no measurable choice of f and g. (See also [14] for further examples.) In the last part of Lemma 5.6, the condition $\pi \ll P$ ensures that, after removing certain nullsets from X and Y, only countably many normalizations are necessary. See the proof of Lemma 6.3 for details.

We can now conclude the desired result on the shape of the optimal density $Z_* \in \mathcal{Z}$, completing the proof of Theorem 2.2 (i)–(iii).

Proposition 5.8. *There exist measurable $f : \mathsf{X} \to \mathbb{R}$ and $g : \mathsf{Y} \to \mathbb{R}$ such that $Z_* = (f \oplus g - c)_+$ P-a.s.*

[6] In fact, it suffices to assume that π is absolutely continuous wrt. *any* product probability measure. See Step 4 in the proof of Lemma 6.3 below.

Proof. By Lemma 5.5 we have $Z_* = \lim_{n\to\infty}(f_n \oplus g_n - c)_+$ P-a.s. for some bounded measurable functions f_n, g_n. As π_* is concentrated on $\{Z_* > 0\}$, we also have

$$Z_* = \lim_{n\to\infty}(f_n \oplus g_n - c) = \Big(\lim_{n\to\infty} f_n \oplus g_n\Big) - c \qquad \pi_*\text{-a.s.},$$

showing that the limit $h := Z_* + c = \lim_{n\to\infty} f_n \oplus g_n$ exists and is finite π_*-a.s. Clearly $\pi_* \ll P$, hence the condition of Lemma 5.6 is satisfied and the claim follows. □

6 Self-Transport: Proof of Theorem 2.2(iv)

In this section we adapt the above arguments to the case of self-transport; i.e., the marginals coincide and the cost is symmetric:

$$(\mathsf{X}, \mu, \tilde{\mu}) = (\mathsf{Y}, \nu, \tilde{\nu}) \qquad \text{and} \qquad c(x,y) = c(y,x). \tag{6.1}$$

Note that $P = \tilde{\mu} \otimes \tilde{\mu}$ is then also symmetric.

In this setting we may expect that there exist symmetric potentials (f, g); i.e., with $f = g$. (Of course, not all potentials will be symmetric; see, e.g., Example 3.6.) However, the proof of Lemma 5.6 above in general does not produce symmetric potentials, *even if* the approximations f_n, g_n are symmetric. This is due to the normalizations for f_n that are made in the proof. These normalizations are key to obtain convergence, but break the symmetry between f_n and g_n.

Below, we first argue that the approximations can indeed be chosen to be symmetric (Lemma 6.1 and Lemma 6.2). Then, we guarantee a symmetric limit with a precise construction that avoids normalizations on certain components and coordinates the normalizations between others (Lemma 6.3).

Lemma 6.1. *Let* (6.1) *hold. Fix* $n \in \mathbb{N}$ *and bounded measurable functions* $\phi_i : \mathsf{X} \to \mathbb{R}$, $1 \le i \le n$ *with* $\int \phi_i \, d\mu = 0$. *Let*

$$\mathcal{Z}_n^{\mathrm{sym}} = \Big\{ Z \in L^2(P) : Z \ge 0,\ Z(x,y) = Z(y,x), \\ \int \phi_i(x) Z(x,y)\, P(dx,dy) = 0,\ 1 \le i \le n \Big\}.$$

There is a unique solution $Z_n \in \mathcal{Z}_n^{\mathrm{sym}}$ *of*

$$\inf_{Z \in \mathcal{Z}_n^{\mathrm{sym}}} \int cZ \, dP + \frac{1}{2}\|Z\|^2_{L^2(P)} \tag{6.2}$$

and Z_n is characterized within $\mathcal{Z}_n^{\mathrm{sym}}$ by being of the form

$$Z_n(x,y) = (b \cdot \Phi(x) + b \cdot \Phi(y) - c(x,y))_+ \qquad \text{for some } b \in \mathbb{R}^n, \tag{6.3}$$

where $\Phi := (\phi_1, \dots, \phi_n)$.

Proof. Following the same arguments as in the proof of Lemma 5.4, it suffices to show that there exists $Z \in \mathcal{Z}_n^{\mathrm{sym}}$ of the form (6.3). To that end, we now show that the symmetric problem

$$\inf_{b \in \mathbb{R}^n} G(b), \qquad G(b) := \int (b \cdot \Phi(x) + b \cdot \Phi(y) - c(x,y))_+^2 \, P(dx,dy)$$

admits a minimizer b_*. As in the proof of Lemma 5.4, we may assume that $b \cdot \Phi = 0$ $\tilde{\mu}$-a.s. only for $b = 0$, and then existence of an optimizer b^* follows by the same coercivity argument. Set $Z_n(x,y) := (b^* \cdot \Phi(x) + b^* \cdot \Phi(y) - c(x,y))_+$. The first-order condition at b^* now gives

$$\int \phi_i(x) Z_n(x,y) \, P(dx,dy) + \int \phi_i(y) Z_n(x,y) \, P(dx,dy) = 0, \quad 1 \le i \le n.$$

Because Z_n and P are symmetric, both integrals must have the same value; i.e., both vanish. This shows that $Z_n \in \mathcal{Z}_n$, as desired. □

Lemma 6.2. *Let (6.1) hold. Let $Z_* \in \mathcal{Z}$ be the primal optimizer. There exist bounded measurable functions $f_n : \mathsf{X} \to \mathbb{R}$, $n \ge 1$ such that*

$$Z_* = \lim_{n \to \infty} (f_n \oplus f_n - c)_+ \qquad P\text{-a.s. and in } L^2(P).$$

Proof. Using Lemma 6.1 instead of Lemma 5.4, the argument is analogous to Lemma 5.5. □

The following passage to the limit $n \to \infty$ is the main step.

Lemma 6.3. *Let $(\mathsf{X}, \mu, \tilde{\mu}) = (\mathsf{Y}, \nu, \tilde{\nu})$ and let $\pi \in \Pi(\mu, \mu)$ be symmetric; i.e., $\pi(dx,dy) = \pi(dy,dx)$. Consider a sequence $f_n : \mathsf{X} \to \mathbb{R}$, $n \ge 1$ of measurable functions such that*

$$\lim_{n \to \infty} f_n \oplus f_n = h \quad \pi\text{-a.s.}$$

for a measurable function $h : \mathsf{X} \times \mathsf{X} \to \mathbb{R}$. Then there is a function $f : \mathsf{X} \to \mathbb{R}$ such that $h = f \oplus f$ π-a.s. If $\pi \ll P$, then f can be chosen to be measurable.

Proof. We give the proof in four steps. Steps 1 and 4 follow [14, 31] whereas Steps 2 and 3 deal with the particular issue of constructing a symmetric limit. In a quite different context, issues with a similar flavor recently appeared in financial mathematics [26].

Step 1. Consider the measurable set

$$S = \Big\{ \lim_{n\to\infty} (f_n \oplus f_n) \text{ exists in } \mathbb{R} \Big\} \subset \mathsf{X} \times \mathsf{X}.$$

We denote by $S_x = \{y \in \mathsf{X} : (x,y) \in S\}$ and $S^y = \{x \in \mathsf{X} : (x,y) \in S\}$ its sections, which are also measurable. For any $x, x' \in \mathsf{X}$, either

$$S_x = S_{x'} \quad \text{or} \quad S_x \cap S_{x'} = \emptyset.$$

Indeed, suppose that there is a point $z \in S_x \cap S_{x'}$, and consider any $y \in S_x$. Then $y \in S_{x'}$ as

$$f_n(x') + f_n(y) = [f_n(x') + f_n(z)] - [f_n(x) + f_n(z)] + [f_n(x) + f_n(y)]$$

and the terms on the right-hand side converge.

Fix a disintegration $\pi(dx,dy) = \mu(dx) \otimes \kappa(x,dy)$. The set $\mathsf{X}_0 = \{x \in \mathsf{X} : \kappa(x, S_x) = 1\}$ is measurable and has full μ-measure. For $x \in \mathsf{X}_0$ we have in particular that $S_x \neq \emptyset$. Moreover, any $y \in \mathsf{X}_0$ is contained in S_x for some $x \in \mathsf{X}_0$. This follows by the symmetry of π, as otherwise $\mu(S^y) = 0$. Define the equivalence relation $\sim$ on X_0 via

$$x \sim x' \quad \text{if} \quad S_x = S_{x'}.$$

For any $x \in \mathsf{X}_0$, let $A(x) \subset \mathsf{X}_0$ be the equivalence class of x. This set is measurable as it has the representation $A(x) = \{x' \in \mathsf{X}_0 : \lim(f_n(x') + f_n(y)) \text{ exists in } \mathbb{R}\}$ for any $y \in S_x$. Let $(x_\lambda)_{\lambda\in\Lambda}$ be a system of representatives containing exactly one member of each equivalence class.

Step 2. Define $C(x) := A(x) \times (S_x \cap \mathsf{X}_0)$ for $x \in \mathsf{X}_0$, and write $C_\lambda := C(x_\lambda)$. Then $(C_\lambda)_{\lambda\in\Lambda}$ is a measurable partition of $S \cap (\mathsf{X}_0 \times \mathsf{X}_0)$. By definition, each C_λ is a measurable rectangle, denoted

$$C_\lambda = A_\lambda \times B_\lambda.$$

Both $(A_\lambda)_{\lambda\in\Lambda}$ and $(B_\lambda)_{\lambda\in\Lambda}$ are measurable partitions of X_0. In fact, by symmetry, $\{A_\lambda : \lambda \in \Lambda\} = \{B_\lambda : \lambda \in \Lambda\}$. In the language of Definition 3.2, $(C_\lambda)_{\lambda\in\Lambda}$ are the connected components of $S \cap (\mathsf{X}_0 \times \mathsf{X}_0)$, seen as a subset of $\mathsf{X}_0 \times \mathsf{X}_0$. The geometry is special here as any two connected points can be joined by a path with length $k = 1$.

Define the reflection $\hat{C} = \{(y,x) : (x,y) \in C\}$ for any $C \subset \mathsf{X} \times \mathsf{X}$. By symmetry, $\hat{C}_\lambda = C_{\lambda'}$ for some λ'. We need to distinguish two types of components. On the one hand, we define

$$\Lambda_{\mathrm{diag}} = \{\lambda \in \Lambda : C_\lambda = \hat{C}_\lambda\}.$$

For $\lambda \in \Lambda_{\mathrm{diag}}$, we observe that $A_\lambda = B_\lambda$; i.e., $C_\lambda = A_\lambda \times A_\lambda$ is a square.

It is useful to visualize $S \cap (\mathsf{X}_0 \times \mathsf{X}_0)$ as a symmetric block matrix (Fig. 2). Then C_λ, $\lambda \in \Lambda_{\mathrm{diag}}$ are square blocks along the diagonal. Next, we describe the off-diagonal blocks, for which there is a symmetry between the lower and upper triangle matrices.

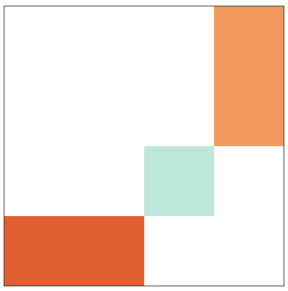

Figure 2: Example for $\mathsf{X} = [0,1]$ with three components (the diagonal is displayed top-left to bottom-right, as for matrices). Here Λ_{low}, Λ_{diag}, Λ_{upp} each have one element.

Indeed, for each $\lambda \in \Lambda \setminus \Lambda_{\mathrm{diag}}$, there is exactly one $\lambda' \in \Lambda \setminus \Lambda_{\mathrm{diag}}$ such that $C_{\lambda'} = \hat{C}_\lambda$. To avoid redundancy, we partition $\Lambda \setminus \Lambda_{\mathrm{diag}}$ into $\Lambda_{\mathrm{low}} \cup \Lambda_{\mathrm{upp}}$ so that for each such unordered pair $\{\lambda, \lambda'\}$, one index (say λ) is in Λ_{low} and the other is in Λ_{upp}. We note that the collections $(A_\lambda)_{\lambda \in \Lambda_{\mathrm{low}}}$ and $(B_\lambda)_{\lambda \in \Lambda_{\mathrm{upp}}}$ coincide. For the subsequent construction, it will be important that

$$(A_\lambda)_{\lambda \in \Lambda_{\mathrm{diag}}} \cup (A_\lambda)_{\lambda \in \Lambda_{\mathrm{low}}} \cup (B_\lambda)_{\lambda \in \Lambda_{\mathrm{low}}} \qquad \text{is a partition of } \mathsf{X}_0. \tag{6.4}$$

Step 3. We define the function f separately on each set of the partition (6.4).

(i) Let $\lambda \in \Lambda_{\mathrm{diag}}$. As $C_\lambda = A_\lambda \times A_\lambda$ is a square, we see that $(x,y) \in C_\lambda$ implies $(x,x) \in C_\lambda$ (and similarly for y). The property $(x,x) \in C_\lambda \subset S$ means that $f_n(x) + f_n(x)$ is convergent, which of course means that $f_n(x)$ is convergent. In brief, $f_n(x)$ is convergent for all $x \in A_\lambda$. We thus define

$$f(x) := \lim_n f_n(x) \quad \text{for} \quad x \in A_\lambda.$$

Clearly $f \oplus f = \lim_n (f_n \oplus f_n)$ on C_λ.

(ii) Let $\lambda \in \Lambda_{\text{low}}$ and $(x, y) \in C_\lambda$. In this case, $f_n(x_\lambda)$ need not be convergent. Define

$$f_n'(x) := f_n(x) - f_n(x_\lambda), \qquad f_n''(y) := f_n(y) + f_n(x_\lambda). \tag{6.5}$$

For $(x, y) \in C_\lambda$, we also have $(x_\lambda, y) \in C_\lambda$, implying that $f_n''(y)$ and $f_n'(x) = [f_n(x) + f_n(y)] - [f_n(y) + f_n(x_\lambda)]$ are both convergent. Define

$$f(x) := \lim_n f_n'(x) \quad \text{for} \quad x \in A_\lambda, \qquad f(y) := \lim_n f_n''(y) \quad \text{for} \quad y \in B_\lambda.$$

These are well defined as $A_\lambda \cap B_\lambda = \emptyset$ for $\lambda \in \Lambda_{\text{low}}$. Moreover,

$$f \oplus f = \lim_n (f_n' \oplus f_n'') = \lim_n (f_n \oplus f_n) \quad \text{on} \quad C_\lambda.$$

By (6.4), the combination of (i) and (ii) defines $f : \mathsf{X}_0 \to \mathbb{R}$. Crucially, disjointness of the unions in (6.4) ensures that there is no contradiction between our definitions of f for different λ within (i) and (ii), and also not between (i) and (ii).

We have $f \oplus f = \lim_n (f_n \oplus f_n)$ on C_λ for $\lambda \in \Lambda_{\text{diag}} \cup \Lambda_{\text{low}}$. If $f \oplus f = \lim_n (f_n \oplus f_n)$ on C_λ, the same holds on $\hat{C}_\lambda$. Thus, we also have $f \oplus f = \lim_n (f_n \oplus f_n)$ on C_λ for $\lambda \in \Lambda_{\text{upp}}$. In summary, $f \oplus f = \lim_n (f_n \oplus f_n)$ on $\mathsf{X}_0 \times \mathsf{X}_0$. Finally, we set $f := 0$ on the μ-nullset $\mathsf{X} \setminus \mathsf{X}_0$ and note that $\mu(\mathsf{X}_0) = 1$ implies $\pi(\mathsf{X}_0 \times \mathsf{X}_0) = 1$ as $\pi \in \Pi(\mu, \mu)$. This gives the desired conclusion $f \oplus f = \lim_n (f_n \oplus f_n)$ π-a.s. We remark that in general, the function f need not be measurable, as it may incorporate *uncountably* many normalizations (6.5) with arbitrary choice of x_λ. The next step removes that issue.

Step 4. Under the condition $\pi \ll P = \tilde{\mu} \otimes \tilde{\mu}$, we must have $\kappa(x, dy) \ll \tilde{\mu}$ for μ-a.e. $x \in \mathsf{X}$ by Fubini's theorem for kernels (or by Remark B.2), and we may choose κ so that this holds without exceptional set. Then $\kappa(x_\lambda, S_{x_\lambda}) = 1$ implies $\tilde{\mu}(S_{x_\lambda}) > 0$. As the sets S_{x_λ} are disjoint and $\tilde{\mu}$ is a finite measure, $\tilde{\mu}(S_{x_\lambda}) > 0$ can hold for at most countably many λ. Thus there is a *countable* set $\Lambda_* \subset \Lambda$ such that $P(\cup_{\lambda \in \Lambda_*} C_\lambda) = 1$ and $P(C_\lambda) > 0$ for $\lambda \in \Lambda_*$. The set $\mathsf{X}_1 := \cup_{\lambda \in \Lambda_*} A_\lambda$ is measurable and satisfies $\tilde{\mu}(\mathsf{X}_1) = 1$. We may redefine $f := 0$ outside X_1 and still have $f \oplus f = \lim_n (f_n \oplus f_n)$ π-a.s. This modified function f is then a countable sum of the form $f = \sum_{\lambda \in \Lambda_*} f^\lambda \mathbf{1}_{D_\lambda}$ where D_λ runs over elements of (6.4) and each f^λ is defined explicitly in (i) or (ii). In particular, f is measurable. □

In the symmetric setting (6.1), the optimal coupling π_* must be symmetric: if $\pi_*(dx, dy)$ is an optimizer, then so is $\pi_*(dy, dx)$, hence both coincide,

by uniqueness. Combining Lemmas 6.2 and 6.3, we can now conclude as in Proposition 5.8, completing the proof of Theorem 2.2.

Proposition 6.4. *Let* (6.1) *hold. There exists a measurable* $f : \mathsf{X} \to \mathbb{R}$ *such that* $Z_* = (f \oplus f - c)_+$ *P-a.s.*

A Examples with Constant Cost

If $c \equiv 0$ and $(\mu, \nu) = (\tilde{\mu}, \tilde{\nu})$, then clearly $\pi_* = P$. In particular, π_* has full support. The next example illustrates that when $(\mu, \nu) \neq (\tilde{\mu}, \tilde{\nu})$ are different (but still equivalent), the optimal coupling and even the optimal support can change.

Example A.1 (Reference can change optimal support)**.** Let $\mathsf{X} = \mathsf{Y} = \{0, 1\}$ and $c \equiv 0$. Let

$$\tilde{\mu} = \tilde{\nu} = \frac{1}{2}(\delta_0 + \delta_1), \qquad \mu = \nu = (1 - \lambda)\delta_0 + \lambda\delta_1, \qquad \lambda \in (0, 1/4].$$

We claim that the optimal density is

$$Z_* := (f \oplus f)_+, \qquad f(0) = 2 - 4\lambda, \quad f(1) = -2 + 8\lambda$$

and in particular that the optimal support is $\operatorname{spt} \pi_* = \{(0, 0), (0, 1), (1, 0)\}$.

By Theorem 2.2, it suffices to check that $(f \oplus f)_+$ is the P-density of a coupling. Indeed, $d\pi := (f \oplus f)_+ \, dP$ has weights

$$\begin{aligned} \pi\{0, 0\} &= [f(0) + f(0)]/4 = 1 - 2\lambda, \\ \pi\{0, 1\} = \pi\{1, 0\} &= [f(0) + f(1)]/4 = \lambda, \\ \pi\{1, 1\} &= [f(1) + f(1)]_+/4 = 0, \end{aligned}$$

showing that $\pi \in \Pi(\mu, \nu)$.

Stated for general marginal spaces, the next proposition elaborates on Example A.1 by giving a sharp condition for π_* to have (or not have) full support, or more precisely, to be equivalent to P.

Proposition A.2. *Let* $c \equiv 0$*. Then*

$$\pi_* \sim P \qquad \Longleftrightarrow \qquad \frac{d\mu}{d\tilde{\mu}} + \frac{d\nu}{d\tilde{\nu}} > 1 \quad P\text{-a.s.}$$

In that case, $d\pi_*/dP = \frac{d\mu}{d\tilde{\mu}} + \frac{d\nu}{d\tilde{\nu}} - 1$ *and* $(f, g) = (\frac{d\mu}{d\tilde{\mu}} - \frac{1}{2}, \frac{d\nu}{d\tilde{\nu}} - \frac{1}{2})$ *are potentials.*

In particular, if $\nu = \tilde{\nu}$*, the optimal coupling is* $\pi_* = \mu \otimes \nu$ *for any* $\mu \sim \tilde{\mu}$*.*

Proof. Suppose that $\pi_* \sim P$. By Theorem 2.2, $0 < Z_* = (f \oplus g - c)_+ = f \oplus g$ P-a.s. for some potentials f, g. Then (2.7) and (2.8) become $f = \frac{d\mu}{d\tilde{\mu}} - \int g \, d\tilde{\nu}$ and $g = \frac{d\nu}{d\tilde{\nu}} - \int f \, d\tilde{\mu}$, which amounts to

$$f \oplus g = \frac{d\mu}{d\tilde{\mu}} \oplus \frac{d\nu}{d\tilde{\nu}} - 1 \quad P\text{-a.s.}$$

In particular, $Z_* > 0$ yields $\frac{d\mu}{d\tilde{\mu}} + \frac{d\nu}{d\tilde{\nu}} > 1$ P-a.s.

Conversely, let $\frac{d\mu}{d\tilde{\mu}} + \frac{d\nu}{d\tilde{\nu}} > 1$ P-a.s. and define $(f, g) := (\frac{d\mu}{d\tilde{\mu}} - \frac{1}{2}, \frac{d\nu}{d\tilde{\nu}} - \frac{1}{2})$. Going backwards, (f, g) solve (2.7) and (2.8), hence Theorem 2.2 shows that $Z_* = (f \oplus g - c)_+ = f \oplus g > 0$. In particular, $\pi_* \sim P$.

If $\nu = \tilde{\nu}$, then $\frac{d\mu}{d\tilde{\mu}} + \frac{d\nu}{d\tilde{\nu}} = \frac{d\mu}{d\tilde{\mu}} + 1 > 1$ P-a.s. and the claim follows. □

B Technical Remarks

For simplicity, we have assumed in the main text that the spaces X, Y are Polish. In fact, we do not make direct use of the topology, and we can allow much more general spaces.

Remark B.1 (General spaces)**.** Our results on regularized optimal transport hold for arbitrary *separable* probability spaces $(\mathsf{X}, \mathcal{F}_\mathsf{X}, \mu)$ and $(\mathsf{Y}, \mathcal{F}_\mathsf{Y}, \nu)$, not necessarily Polish. Separability is used for the finite-dimensional approximations in the proofs of Lemmas 5.5 and 6.2.

A probability space $(\mathsf{X}, \mathcal{F}_\mathsf{X}, \mu)$ is called separable if there is a countable family $(A_n) \subset \mathcal{F}_\mathsf{X}$ such that for every $A \in \mathcal{F}_\mathsf{X}$ and $\varepsilon > 0$, there exists n with $\mu(A \triangle A_n) < \varepsilon$. This property holds if and only if $L^1(\mathsf{X}, \mathcal{F}_\mathsf{X}, \mu)$ is separable; i.e., has a dense countable subset (consider simple functions based on (A_n) or see [3, Exercise 4.7.63, p. 306]). See [4, Section 7.14(iv), pp. 132–133] for some very general sufficient conditions for separability; for instance, any Radon measure on a metric space is separable.

The proofs of Lemmas 5.6 and 6.3 use the existence of disintegrations; i.e., regular conditional distributions. However, we only apply those results in the absolutely continuous case $\pi \ll P$, and then that existence holds on general spaces (Remark B.2 below).

The last remark recalls how to construct disintegrations from joint densities, without need for Polish or Blackwell spaces.

Remark B.2 (Disintegration from density)**.** Let $\mu, \tilde{\mu} \in \mathcal{P}(\mathsf{X})$ and $\nu, \tilde{\nu} \in \mathcal{P}(\mathsf{Y})$ be probability measures on arbitrary measurable spaces $(\mathsf{X}, \mathcal{F}_\mathsf{X})$ and

$(\mathsf{Y}, \mathcal{F}_\mathsf{Y})$, and let $\pi \in \Pi(\mu, \nu)$ satisfy $\pi \ll \tilde{\mu} \otimes \tilde{\nu}$. Then there is a disintegration $\pi(dx, dy) = \mu(dx) \otimes \kappa(x, dy)$ with $\kappa(x, dy) \ll \tilde{\nu}(dy)$ for all $x \in \mathsf{X}$.

This is a standard fact from probability theory. Indeed, note that $\pi \ll \tilde{\mu} \otimes \tilde{\nu}$ immediately implies $\mu \ll \tilde{\mu}$. Let $D_\mu = \frac{d\mu}{d\tilde{\mu}}$ and $D_\pi = \frac{d\pi}{d(\tilde{\mu} \otimes \tilde{\nu})}$, and define $\kappa(x, dy) := \frac{D(x,y)}{D_\mu(x)} \mathbf{1}_{D_\mu(x) \neq 0} \, \tilde{\nu}(dy)$. Then κ is a Markov kernel and for any $A \in \mathcal{F}_\mathsf{X}$ and $B \in \mathcal{F}_\mathsf{Y}$,

$$(\mu \otimes \kappa)(A \times B) = \int_A \left(\int_B \frac{D_\pi(x, y)}{D_\mu(x)} \mathbf{1}_{D_\mu(x) \neq 0} \, \tilde{\nu}(dy) \right) \mu(dx) = \pi(A \times B).$$

References

[1] E. Bayraktar, S. Eckstein, and X. Zhang. Stability and sample complexity of divergence regularized optimal transport. *Preprint arXiv:2212.00367v1*, 2022.

[2] M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. volume 84 of *Proceedings of Machine Learning Research*, pages 880–889, 2018.

[3] V. I. Bogachev. *Measure theory. Vol. I.* Springer-Verlag, Berlin, 2007.

[4] V. I. Bogachev. *Measure Theory. Vol. II.* Springer-Verlag, Berlin, 2007.

[5] J. M. Borwein and A. S. Lewis. Decomposition of multivariate functions. *Canad. J. Math.*, 44(3):463–482, 1992.

[6] G. Carlier, P. Pegon, and L. Tamanini. Convergence rate of general entropic optimal transport costs. *Calc. Var. Partial Differential Equations*, 62(4):Paper No. 116, 28, 2023.

[7] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300. 2013.

[8] A. Dessein, N. Papadakis, and J.-L. Rouas. Regularized optimal transport and the rot mover's distance. *J. Mach. Learn. Res.*, 19(15):1–53, 2018.

[9] S. Di Marino and A. Gerolin. Optimal transport losses and Sinkhorn algorithm with general convex regularization. *Preprint arXiv:2007.00976v1*, 2020.

[10] S. Eckstein and M. Kupper. Computation of optimal transport and related hedging problems via penalization and neural networks. *Appl. Math. Optim.*, 83(2):639–667, 2021.

[11] S. Eckstein and M. Nutz. Convergence rates for regularized optimal transport via quantization. *To appear in Math. Oper. Res.*, 2022. Preprint arXiv:2208.14391v2.

[12] M. Essid and J. Solomon. Quadratically regularized optimal transport on graphs. *SIAM J. Sci. Comput.*, 40(4):A1961–A1986, 2018.

[13] H. Föllmer. Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Math.*, pages 101–203. Springer, Berlin, 1988.

[14] H. Föllmer and N. Gantert. Entropy minimization and Schrödinger processes in infinite dimensions. *Ann. Probab.*, 25(2):901–926, 1997.

[15] H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time*. W. de Gruyter, Berlin, 3rd edition, 2011.

[16] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29*, pages 3440–3448, 2016.

[17] N. Gigli and L. Tamanini. Second order differentiation formula on $\mathrm{RCD}^*(K, N)$ spaces. *J. Eur. Math. Soc. (JEMS)*, 23(5):1727–1795, 2021.

[18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5769–5779, 2017.

[19] L. Li, A. Genevay, M. Yurochkin, and J. Solomon. Continuous regularized Wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17755–17765. Curran Associates, Inc., 2020.

[20] D. Lorenz and H. Mahler. Orlicz space regularization of continuous optimal transport problems. *Appl. Math. Optim.*, 85(2):Paper No. 14, 33, 2022.

[21] D. A. Lorenz, P. Manns, and C. Meyer. Quadratically regularized optimal transport. *Appl. Math. Optim.*, 83(3):1919–1949, 2021.

[22] H. Malamut and M. Sylvestre. Convergence rates of the regularized optimal transport: Disentangling suboptimality and entropy. *Preprint arXiv:2306.06940v1*, 2023.

[23] O. L. Mangasarian and R. R. Meyer. Nonlinear perturbation of linear programs. *SIAM J. Control Optim.*, 17(6):745–752, 1979.

[24] M. Nutz. *Introduction to Entropic Optimal Transport*. Lecture notes, Columbia University, 2021. `https://www.math.columbia.edu/~mnutz/docs/EOT_lecture_notes.pdf`.

[25] M. Nutz and J. Wiesel. Entropic optimal transport: convergence of potentials. *Probab. Theory Related Fields*, 184(1-2):401–424, 2022.

[26] M. Nutz, J. Wiesel, and L. Zhao. Limits of semistatic trading strategies. *Math. Finance*, 33(1):185–205, 2023.

[27] OptimalTransport.jl. Optimaltransport.quadreg, 2023. `https://juliaoptimaltransport.github.io/OptimalTransport.jl/dev/#Quadratically-regularised-optimal-transport` [Accessed 12/01/2023].

[28] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[29] Python Optimal Transport. ot.stochastic.loss_dual_quadratic, 2023. `https://pythonot.github.io/gen_modules/ot.stochastic.html#id22` [Accessed 12/01/2023].

[30] L. Qi. The maximal normal operator space and integration of subdifferentials of nonconvex functions. *Nonlinear Anal.*, 13(9):1003–1011, 1989.

[31] L. Rüschendorf and W. Thomsen. Closedness of sum spaces and the generalized Schrödinger problem. *Teor. Veroyatnost. i Primenen.*, 42(3):576–590, 1997.

[32] B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM J. Sci. Comput.*, 41(3):A1443–A1481, 2019.

[33] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. Large scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, 2018.

[34] C. Villani. *Optimal transport, old and new*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

[35] S. Zhang, G. Mordant, T. Matsumoto, and G. Schiebinger. Manifold learning with sparse regularised optimal transport. *Preprint arXiv:2307.09816v1*, 2023.