Linear Convergence of Gradient Descent for Quadratically Regularized Optimal Transport

Alberto González-Sanz[*] Marcel Nutz[†] Andrés Riveros Valdevenito[‡]

September 16, 2025

Abstract

In optimal transport, quadratic regularization is an alternative to entropic regularization when sparse couplings or small regularization parameters are desired. Here quadratic regularization means that transport couplings are penalized by the squared L^2 norm, or equivalently the χ^2 divergence. While a number of computational approaches have been shown to work in practice, quadratic regularization is analytically less tractable than entropic, and we are not aware of a previous theoretical convergence rate analysis. We focus on the gradient descent algorithm for the dual transport problem in continuous and semi-discrete settings. This problem is convex but not strongly convex; its solutions are the potential functions that approximate the Kantorovich potentials of unregularized optimal transport. The gradient descent steps are straightforward to implement, and stable for small regularization parameter—in contrast to Sinkhorn's algorithm in the entropic setting. Our main result is that gradient descent converges linearly; that is, the L^2 distance between the iterates and the limiting potentials decreases exponentially fast. Our analysis centers on the linearization of the gradient descent operator at the optimum and uses functional-analytic arguments to bound its spectrum. These techniques seem to be novel in this area and are substantially different from the approaches familiar in entropic optimal transport.

Keywords Gradient Descent; Optimal Transport; Quadratic Regularization
AMS 2020 Subject Classification 49N10; 49N05; 90C25

1 Introduction

Optimal transport has become ubiquitous in many areas where distributions or data sets need to be compared, such as statistics, machine learning and image processing. Given compactly supported probability distributions P and Q on $\mathbb{R}^d$, the optimal transport problem with quadratic cost is

$$\mathrm{OT}(P,Q) = \inf_{\pi\in\Pi(P,Q)} \int \frac{1}{2}\|x-y\|^2 d\pi(x,y), \tag{1}$$

[*] Department of Statistics, Columbia University, ag4855@columbia.edu.

[†] Departments of Mathematics and Statistics, Columbia University, mnutz@columbia.edu. Research supported by NSF Grants DMS-2106056, DMS-2407074.

[‡] Department of Statistics, Columbia University, ar4151@columbia.edu.

where $\Pi(P,Q)$ denotes the set of couplings; that is, probability distributions on $\mathbb{R}^d\times\mathbb{R}^d$ with marginals (P,Q). The optimal value $\mathrm{OT}(P,Q)$ defines the Wasserstein distance between P and Q and hence is the target of numerous computational approaches (see [25] for a recent monograph). Following [6], entropic regularization and the corresponding Sinkhorn algorithm are likely the most popular as of now. The entropically regularized optimal transport problem is

$$\mathrm{EOT}_\varepsilon(P,Q)=\inf_{\pi\in\Pi(P,Q)}\int\frac{1}{2}\|x-y\|^2d\pi(x,y)+\varepsilon\,\mathrm{KL}\big(\pi|P\otimes Q\big), \tag{2}$$

where $\varepsilon>0$ is a parameter determining the strength of regularization and $\mathrm{KL}(\pi|P\otimes Q)$ is the Kullback–Leibler divergence between π and the product $P\otimes Q$. Sinkhorn's iteration can be described as the coordinate-ascent algorithm for the dual problem of (2) which seeks a pair of functions $f,g:\mathbb{R}^d\to\mathbb{R}$ maximizing

$$\int f(x)+g(y)-\varepsilon e^{\frac{f(x)+g(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}}d(P\otimes Q)(x,y). \tag{3}$$

Thanks to the algebraic properties of the exponential function in (3), the coordinate-wise maximization (i.e., optimizing separately f or g) has a closed-form solution, leading to the iteration

$$g_n(y)=-\varepsilon\log\left(\int e^{\frac{f_{n-1}(x)-\frac{1}{2}\|x-y\|^2}{\varepsilon}}dP(x)\right),\quad f_n(x)=-\varepsilon\log\left(\int e^{\frac{g_n(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}}dQ(y)\right). \tag{4}$$

The exponential also entails that the dual objective (3) is strongly concave (when restricted to uniformly bounded $f(x)+g(y)$). This implies (see [4]) that the iteration converges linearly to the maximizer of (3), which in turn approximates the dual solution of the optimal transport problem (1) in the limit $\varepsilon\to0$ of vanishing regularization. Several other proofs of linear convergence are known, starting with [10] for the discrete case, and a recent body of literature analyzes the corresponding constants in detail (see [5] and the references therein).

While Sinkhorn's algorithm has been very successful in many applications, it has limitations. Using KL divergence entails that the optimal coupling of (2) always has full support (equal to the support of $P\otimes Q$). This is known as overspreading in applications, as the true optimal transport for (1) is typically sparse (even given by a deterministic map). For example, overspreading can correspond to blurring in an image processing task [2] or bias in a manifold learning task [29]. Separately, a computational limitation is faced when small regularization parameters ε are desired to closely approximate (1): since exponentially large and small values occur in (4), the algorithm tends to become unstable for small values of ε, an issue that can be mitigated only to some extent [26]. See also [17], where a Dijkstra-type search algorithm is proposed as a replacement.

Starting with [22, 2, 9], an alternate approach is to regularize with a different divergence. The most tractable choice is the χ^2 divergence, or equivalently, penalization by the squared L^2 norm of the density, leading to the quadratically regularized optimal transport problem,

$$\mathrm{QOT}_\epsilon(P,Q)=\inf_{\pi\in\Pi(P,Q)}\int\frac{1}{2}\|x-y\|^2d\pi(x,y)+\frac{\varepsilon}{2}\left\|\frac{d\pi}{d(P\otimes Q)}\right\|^2_{L^2(P\otimes Q)}. \tag{5}$$

It is known that $\mathrm{QOT}_\epsilon(P,Q)$ approximates the unregularized optimal transport cost $\mathrm{OT}(P,Q)$ at rate $\epsilon^{2/(d+2)}$ as $\varepsilon \to 0$ (see [8], and [11] for the leading constant). In contrast to entropic regularization, the optimal coupling π_* of (5) has sparse support for small ε; this has been observed empirically since the initial works (e.g., [2, 9, 21, 1]) and established theoretically more recently in [28, 15]. See Figure 1 for an illustration. A number of computational approaches have been considered, mostly targeting the dual problem of (5),

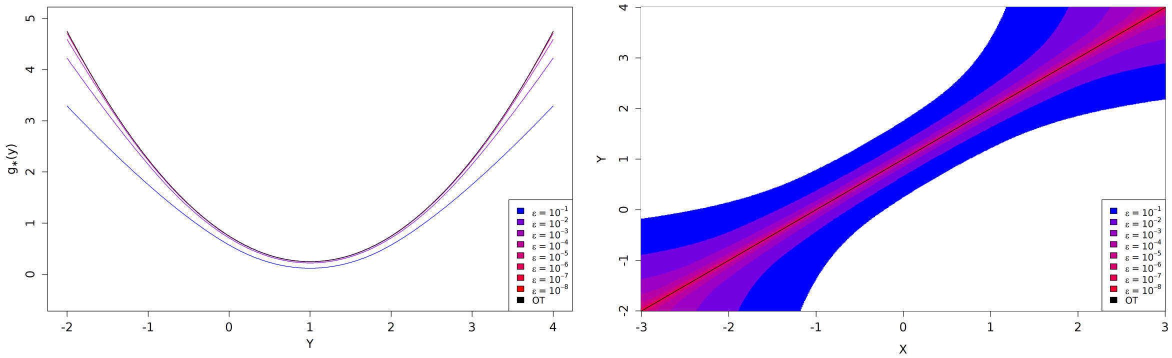

Figure 1: Transporting $P = N(0,1)$ to $Q = N(1,1)$; the measures are truncated to [-3,3] and [-2,4], respectively. Left: Dual solutions of QOT_ϵ and OT. Right: Supports of the optimal couplings are sparse and converge to the optimal transport map.

$$\sup_{f,g:\mathbb{R}^d\to\mathbb{R}} \int f(x)+g(y) - \frac{1}{2\varepsilon}\left(f(x)+g(y)-\frac{1}{2}\|x-y\|^2\right)_+^2 d(P\otimes Q)(x,y) \tag{6}$$

or its first-order condition of optimality,

$$\begin{cases} \varepsilon = \int \left(f(x)+g(y)-\frac{1}{2}\|x-y\|^2\right)_+ dP(x), \\ \varepsilon = \int \left(f(x)+g(y)-\frac{1}{2}\|x-y\|^2\right)_+ dQ(y). \end{cases} \tag{7}$$

Here $(t)_+ := \max\{t,0\}$ denotes the positive part function; it appears as a consequence of the constraint that couplings are nonnegative measures. In the realm of discrete marginals, [22] proposed a mirror gradient method while [9] used a Newton-type algorithm to solve a minimum-cost flow problem on a graph. In [2], the authors leveraged a generic L-BFGS solver and also introduced an alternating minimization scheme. A similar Gauss–Seidel method was suggested in [21]; the idea is to alternately solve the equations in (7). While Sinkhorn's algorithm implements the analogous iteration for entropic regularization, the equations in (7) cannot be solved in closed form. To implement this implicit method, [21] explored both direct search strategies and a semi-smooth Newton method. These approaches were further examined in [20] where the authors noted good empirical convergence but a high computational cost per iteration in large-scale settings. To mitigate this, they proposed several alternative methods, including cyclic projections, dual gradient descent, and its accelerated variant. While no theoretical analysis is given, their numerical experiments suggest that all three approaches are efficient and consistent. (Note that the numerical methods and experiments can be found in the preprint [20]; they are omitted in the journal version [19].) In the continuous setting, several works including [7, 12, 16, 18, 27] apply neural networks to the dual problem. For instance, [18] uses neural networks and gradient descent to compute regularized Wasserstein barycenters.

While numerous algorithms have been used successfully, we are not aware of any convergence rates in the literature. Note that rates are not obvious given the lack of strong concavity in (6)—while the early work of [22] mentions examining linear convergence as an "important avenue for future work," not much progress has been made in that direction until the present paper. More generally, conventional wisdom is that quadratically regularized optimal transport "works" in practice but is difficult to analyze theoretically. The present work provides not only a convergence analysis, but also develops techniques that will be useful in other theoretical studies on regularized optimal transport.

Specifically, we consider the gradient descent algorithm for the dual problem (6). With $\Gamma(f,g)$ denoting the objective function in (6), the gradient descent (or ascent, to be precise) in $L^2(P) \times L^2(Q)$ with step size $\eta > 0$ is

$$\begin{pmatrix} f_{n+1} \\ g_{n+1} \end{pmatrix} = \begin{pmatrix} f_n \\ g_n \end{pmatrix} + \eta \cdot \mathrm{D}\Gamma \begin{pmatrix} f_n \\ g_n \end{pmatrix} \tag{8}$$

for $n \geq 0$, where the initial values (f_0, g_0) are given and the explicit form of the gradient is

$$\mathrm{D}\Gamma \begin{pmatrix} f \\ g \end{pmatrix}(x,y) = \begin{pmatrix} 1 - \frac{1}{\varepsilon} \int \left(f(x) + g(\tilde{y}) - \frac{1}{2}\|x - \tilde{y}\|^2\right)_+ dQ(\tilde{y}) \\ 1 - \frac{1}{\varepsilon} \int \left(f(\tilde{x}) + g(y) - \frac{1}{2}\|\tilde{x} - y\|^2\right)_+ dP(\tilde{x}) \end{pmatrix}. \tag{9}$$

We observe that the iteration (8) is fairly straightforward to implement, similarly as the Sinkhorn iteration, however it is free from exponentially large or small values. Indeed, even a naive implementation is stable for very small parameters ε, a key reason for the growing interest in quadratic regularization. We observe that evaluating the integrals in (9) may require replacing P and Q by empirical samples, especially in high-dimensional problems. In that respect, it is useful to know that the quadratically regularized optimal transport problem has parametric sample complexity [13]—it does not suffer from the same curse of dimensionality as optimal transport.

We provide a theoretical analysis in a general setting covering continuous and semi-discrete marginals; in fact, we will only assume that one of the two marginal measures has connected support and does not charge small sets. The main result (Theorem 2.3) shows that for step size $\eta < \varepsilon$, the iterates (f_n, g_n) converge linearly in L^2 to the solution (f_*, g_*) of the dual problem (6); that is, there exist $\delta_* < 1$ and $n \geq n_0$ such that

$$\|(f_n, g_n) - (f_*, g_*)\|_{L^2(P) \times L^2(Q)} \leq \delta_*^n$$

for all $n \geq n_0$. Our numerical experiments suggest that this bound accurately captures the behavior of the algorithm: convergence is approximately geometric after a burn-in period.

Our mathematical analysis centers on the linearization $\mathbb{L}$ of the gradient descent operator $\mathbb{I} + \eta \mathrm{D}\Gamma$ at the optimum (f_*, g_*). Indeed, we show that the linear operator $\mathbb{L}_n$ mapping $(f_n, g_n) - (f_*, g_*)$ to $(f_{n+1}, g_{n+1}) - (f_*, g_*)$ converges in operator norm to $\mathbb{L}$. The main result in Theorem 2.3 then boils down to showing that $\mathbb{L}$ is a strict contraction. Thus, we establish that the spectrum of the self-adjoint operator $\mathbb{L}$ is contained in $(-1,1)$. We mention that a similar proof strategy would likely apply to a class of f-divergence regularizations that includes L^p regularization for $1 < p < 2$; more specifically, the class detailed in [14, Assumption 2.1].

While our analysis does not proceed through a Polyak–Łojasiewicz (PL) inequality or quadratic growth condition, the following remarks may give additional intuition for the geometry underlying linear convergence—and also illustrate that the techniques developed in this paper are useful beyond the proof of Theorem 2.3. Under an additional regularity condition (that P does not charge small sets, like Q in Assumption 2.1), one can check that Γ is twice differentiable at its maximizer (f_*, g_*). Adapting the proof techniques in Section 4, one can then show strict positive definiteness,

$$\delta := \inf_{\|(u,v)\|_{L^2}=1} \langle [-\mathrm{D}^2\Gamma(f_*, g_*)](u,v), (u,v)\rangle > 0.$$

This result is related to a local quadratic growth condition around (f_*, g_*). However, it is not clear how to ensure existence of $\mathrm{D}^2\Gamma(f,g)$ at points $(f,g) \neq (f_*, g_*)$ or even the continuity of $(f,g) \mapsto \mathrm{D}^2\Gamma(f,g)$, so that a rigorous statement is not immediate. Under additional smoothness assumptions on (P,Q), exploiting the proof techniques in Section 5 and with additional work, we hope to show in future work that a quadratic growth condition holds in an L^∞ neighborhood:

$$\Gamma(f,g) \leq \Gamma(f_*, g_*) - \left\{\delta/2 + \omega(\|(f,g) - (f_*, g_*)\|_\infty)\right\} \|(f,g) - (f_*, g_*)\|_{L^2}^2$$

for a function ω with $\lim_{t\to 0} \omega(t) = 0$. We note that the above inequality is still weaker than the quadratic growth condition in L^2 which would be needed to directly infer linear convergence of gradient descent from standard results.

The remainder of the paper is organized as follows. Section 2 details the setting, the gradient descent algorithm and our main result on its convergence. Section 3 gathers preliminary results for the proof. Section 4 studies the linearized gradient descent operator $\mathbb{L}$ and forms the core of our analysis. Section 5 completes the proof of the main result by showing the convergence of $\mathbb{L}_n$ to $\mathbb{L}$. Section 6 concludes with numerical experiments.

2 Problem statement and main result

Let P, Q be probability measures on $\mathbb{R}^d$. The following is a standing assumption throughout the paper.

Assumption 2.1 (Marginals)**.** The topological supports of $P, Q \in \mathcal{P}(\mathbb{R}^d)$, denoted

$$\Omega := \operatorname{spt} P \quad \text{and} \quad \Omega' := \operatorname{spt} Q, \qquad \text{are compact.}$$

Moreover, Q does not charge the boundary of any convex subset of $\mathbb{R}^d$, and its support Ω' is connected.

Roughly speaking, Assumption 2.1 imposes that one of the two marginals be continuous. As there is no structural condition on the other marginal, it covers both continuous transport problems (where both marginals are continuous) and semi-discrete problems (where one marginal is discrete and the other is continuous).

The quadratically regularized optimal transport (QOT) problem with regularization parameter $\varepsilon > 0$ is

$$\mathrm{QOT}_\epsilon(P, Q) := \inf_{\pi \in \Pi(P,Q)} \int \frac{1}{2}\|x-y\|^2 d\pi(x,y) + \frac{\varepsilon}{2} \left\| \frac{d\pi}{d(P\otimes Q)} \right\|^2_{L^2(P\otimes Q)} \tag{10}$$

with the convention that the last term is $+\infty$ if $\pi \not\ll P \otimes Q$. This problem has a unique solution $\pi_* \in \Pi(P, Q)$, and π_* is characterized within $\Pi(P, Q)$ by having a density of the form

$$\frac{d\pi_*}{d(P\otimes Q)}(x,y) = \frac{1}{\varepsilon}\left(f_*(x) + g_*(y) - \frac{1}{2}\|x-y\|^2\right)_+ \tag{11}$$

for a pair $(f_*, g_*) \in L^2(P) \times L^2(Q)$ called the potentials. If (f_*, g_*) are potentials, then $(f_* - c, g_* + c)$ are also potentials for any $c \in \mathbb{R}$. To remove this ambiguity, we work with the subspace

$$L^2_\oplus = \left\{(f,g) \in L^2(P) \times L^2(Q) : \int f dP = \int g dQ\right\} = \{(c, -c) : c \in \mathbb{R}\}^\perp \subset L^2(P) \times L^2(Q).$$

As a consequence of the connectedness in Assumption 2.1, the potentials are unique in $L^2_\oplus$ (see Lemma 3.1 below for all these assertions). The orthogonal projection onto $L^2_\oplus$ is

$$\mathrm{proj}_\oplus : L^2(P) \times L^2(Q) \to L^2_\oplus, \qquad \begin{pmatrix} f \\ g \end{pmatrix} \mapsto \begin{pmatrix} f + \frac{1}{2}\left(\int g dQ - \int f dP\right) \\ g - \frac{1}{2}\left(\int g dQ - \int f dP\right) \end{pmatrix} \tag{12}$$

and $L^2_\oplus$ is naturally a Hilbert space with the induced inner product

$$\langle (f,g), (u,v)\rangle_{L^2_\oplus} = \langle (f,g), (u,v)\rangle_{L^2(P)\times L^2(Q)} = \langle f, u\rangle_{L^2(P)} + \langle g, v\rangle_{L^2(Q)}.$$

The potentials are also characterized as the unique solution of the dual problem of (10),

$$\sup_{(f,g)\in L^2_\oplus} \Gamma(f,g), \tag{13}$$

where the dual objective function is

$$\Gamma(f,g) = \int f(x)dP(x) + \int g(y)dQ(y) - \frac{1}{2\varepsilon}\int \left(f(x) + g(y) - \frac{1}{2}\|x-y\|^2\right)^2_+ d(P\otimes Q)(x,y).$$

The gradient of Γ at $(f,g) \in L^2_\oplus$ is

$$\mathrm{D}\Gamma\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} 1 - \frac{1}{\varepsilon}\int \left(f(\cdot) + g(y) - \frac{1}{2}\|\cdot - y\|^2\right)_+ dQ(y) \\ 1 - \frac{1}{\varepsilon}\int \left(f(x) + g(\cdot) - \frac{1}{2}\|x - \cdot\|^2\right)_+ dP(x) \end{pmatrix} \in L^2_\oplus. \tag{14}$$

Thus, the gradient descent algorithm with step size $\eta > 0$ is

$$\begin{pmatrix} f_{n+1} \\ g_{n+1} \end{pmatrix} = \begin{pmatrix} f_n \\ g_n \end{pmatrix} + \eta \cdot \mathrm{D}\Gamma \begin{pmatrix} f_n \\ g_n \end{pmatrix} \tag{15}$$

for $n \geq 0$, where the initial values $(f_0, g_0) \in L^2_\oplus$ are given inputs. Note that $(f_n, g_n) \in L^2_\oplus$ implies $(f_{n+1}, g_{n+1}) \in L^2_\oplus$.

Assumption 2.2. The initial values $f_0 : \Omega \to \mathbb{R}$ and $g_0 : \Omega' \to \mathbb{R}$ are Lipschitz continuous functions normalized such that $\int f dP = \int g dQ$. The step size η either satisfies

(i) $\eta \in (0, \varepsilon)$, or

(ii) $\eta = \varepsilon$ and $\operatorname{spt} \pi_* \neq \Omega \times \Omega'$.

We remark that the condition $\operatorname{spt} \pi_* \neq \Omega \times \Omega'$ is harmless. Indeed, Lemma 3.3 shows that any algorithm should first check the nonnegativity condition (22) which is equivalent to $\operatorname{spt} \pi_* = \Omega \times \Omega'$; if it holds, the potentials are given by the explicit expression (23) and there is no need for gradient descent in the first place. In practice, $\operatorname{spt} \pi_* = \Omega \times \Omega'$ only occurs for large values of ε that are rarely of interest.

Theorem 2.3 (Linear convergence of gradient descent)**.** *Under Assumption 2.2, there exist constants $\delta_* \in (0, 1)$ and $n_0 \in \mathbb{N}$ such that the iterates (f_n, g_n) of (15) satisfy*

$$\|(f_n, g_n) - (f_*, g_*)\|_{L^2_\oplus} \leq \delta_*^n \quad \text{for all } n \geq n_0.$$

The proof, which occupies the rest of the paper, has the following structure. We observe that the iterates (f_n, g_n) satisfy

$$\begin{pmatrix} f_{n+1} - f_* \\ g_{n+1} - g_* \end{pmatrix} = \mathbb{L}_n \begin{pmatrix} f_n - f_* \\ g_n - g_* \end{pmatrix}$$

for an operator $\mathbb{L}_n$ (see Lemma 5.1) which converges in operator norm to a limit $\mathbb{L}$ as $n \to \infty$ (Proposition 5.5). The limiting operator $\mathbb{L}$ is the linearization of the gradient descent operator at the optimum (f_*, g_*). The key step is to show that the operator norm of $\mathbb{L}$ is strictly smaller than one (Proposition 4.1). It then follows that $\mathbb{L}_n$ is a strict contraction for $n \geq n_0$, which is the assertion of Theorem 2.3.

Remark 2.4. Assumption 2.2 allows for step sizes $\eta \leq \varepsilon$. A classical result states that for a strongly convex function with L-Lipschitz gradient, the gradient descent algorithm converges linearly for any step size $\eta < 2/L$ (see [23, Theorem 2.1.15]). In our problem, the gradient $\mathrm{D}\Gamma$ is Lipschitz in $L^2_\oplus$ with Lipschitz constant bounded by $2/\varepsilon$, as can be seen directly by applying the inequality $(t)_+ - (s)_+ \leq \mathbb{I}_{\{t \geq 0\}}(t - s)$. In that sense, Assumption 2.2 is in line with the classical result. Our experiments in Section 6 suggest that convergence can break down for $\eta > \varepsilon$.

Remark 2.5. We focus on the transport cost $c(x, y) = \frac{1}{2}\|x - y\|^2$ which is the most important example in practice. While the continuity of this function is used throughout, the specific form is used only to infer the regularity properties in Lemma 3.2 (i) which, in turn, are used to argue that certain sets occurring in the proof of Proposition 5.5 are negligible. While we do not know how to rigorously guarantee the latter in general, it seems plausible that Theorem 2.3 could extend to more general transport costs. The analysis of the linearized gradient descent operator in Section 4 directly extends to general continuous costs $c(x, y)$.

3 Preliminaries

Let $\mathcal{C}(\Omega)$ denote the space of continuous functions $f : \Omega \to \mathbb{R}$. For $f \in \mathcal{C}(\Omega)$ and $g \in \mathcal{C}(\Omega')$, we denote $(f \oplus g)(x, y) := f(x) + g(y)$. Consider the quotient space

$$\mathcal{C}_\oplus := \big(\mathcal{C}(\Omega) \times \mathcal{C}(\Omega')\big) \,/ \sim_\oplus$$

where $(f, g) \sim_\oplus (u, v)$ if and only if $f \oplus g = u \oplus v$, and endow $\mathcal{C}_\oplus$ with the norm

$$\|(u, v)\|_{\mathcal{C}_\oplus} := \inf_{a\in\mathbb{R}} \{\|u + a\|_\infty + \|v - a\|_\infty\}\,.$$

Next, we detail some properties of the potentials (f_*, g_*) that will be used throughout our convergence analysis. The following system (16) can be understood as the first-order condition of optimality for the dual problem (6).

Lemma 3.1. *There exists a unique pair $(f_*, g_*) \in \mathcal{C}_\oplus$ solving the system*

$$\begin{cases} \varepsilon = \displaystyle\int \left(f_*(x) + g_*(y) - \tfrac{1}{2}\|x - y\|^2\right)_+ dP(x) & \text{for all } y \in \Omega', \\ \varepsilon = \displaystyle\int \left(f_*(x) + g_*(y) - \tfrac{1}{2}\|x - y\|^2\right)_+ dQ(y) & \text{for all } x \in \Omega. \end{cases} \tag{16}$$

The pair (f_*, g_*) is also characterized as the unique maximizer of the dual problem (6) and by the relation (11) with the primal solution π_*.

Proof. Existence of a solution $(f_*, g_*) \in \mathcal{C}_\oplus$ of (16), as well as the equivalence of (16) with solving the dual problem and with (11), are shown for instance in [24]. Uniqueness is also shown in [24], but only under additional conditions on (P, Q). Next, we provide a more general proof.

Let (f_*, g_*) and (f'_*, g'_*) be continuous solutions of (16). Fix $y \in \Omega'$, then (16) implies that

$$P\left\{x \in \Omega : f_*(x) + g_*(y) - \tfrac{1}{2}\|x - y\|^2 > 0\right\} > 0;$$

in particular, the latter set contains an element x_0. Continuity of g_* implies that

$$f_*(x_0) + g_*(\tilde{y}) - \tfrac{1}{2}\|x_0 - \tilde{y}\|^2 > 0 \quad \text{for all } \tilde{y} \text{ in a neighborhood } U_y \text{ of } y \text{ in } \Omega'. \tag{17}$$

The relation (11) shows that $\frac{d\pi_*}{d(P\otimes Q)}$ admits a continuous version, and it is a general fact that if the density of a measure admits a continuous version, then that version is uniquely determined at every point of the support. Since (11) holds both with (f_*, g_*) and (f'_*, g'_*), and both are continuous, we conclude from (17) that

$$f_*(x_0) + g_*(\tilde{y}) - \tfrac{1}{2}\|x_0 - \tilde{y}\|^2 = f'_*(x_0) + g'_*(\tilde{y}) - \tfrac{1}{2}\|x_0 - \tilde{y}\|^2 \quad \text{for all } \tilde{y} \in U_y.$$

Hence, $g_*(\tilde{y}) - g'_*(\tilde{y}) = f'_*(x_0) - f_*(x_0) =: c$ for all $\tilde{y} \in U_y$, showing that $g_* - g'_*$ is locally constant in Ω'. As Ω' is connected, it follows that $g_* - g'_* = c$ is constant in Ω. It now follows from (16) that $f_* - f'_* = -c$ (see, e.g., [24, Lemma 2.4]) and hence $(f_*, g_*) = (f'_*, g'_*)$ in $\mathcal{C}_\oplus$. □

Next, we detail two properties of the set

$$\mathcal{E} := \left\{(x,y) \in \Omega \times \Omega' : f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2 \geq 0\right\}. \tag{18}$$

We denote its sections by

$$\begin{aligned} \mathcal{S}_x &:= \left\{y \in \Omega' : f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2 \geq 0\right\}, \quad x \in \Omega, & (19)\\ \mathcal{T}_y &:= \left\{x \in \Omega : f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2 \geq 0\right\}, \quad y \in \Omega'. & (20) \end{aligned}$$

Lemma 3.2.

(i) For any $x \in \Omega$, there is a convex set $C_x \subset \mathbb{R}^d$ with nonempty interior such that

$$\begin{aligned} \mathcal{S}_x &= \left\{y \in \Omega' : f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2 \geq 0\right\} = C_x \cap \Omega', \\ \mathcal{N}_x &:= \left\{y \in \Omega' : f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2 = 0\right\} = \partial C_x \cap \Omega'. \end{aligned}$$

In particular, $\mathcal{N}_x$ is Q-negligible.

(ii) There exists a constant $\lambda > 0$ such that $Q(\mathcal{S}_x) \geq \lambda$ and $P(\mathcal{T}_y) \geq \lambda$ for all $(x,y) \in \Omega \times \Omega'$.

Proof. (i) This is due to a concavity property that was previously used in [15, 28]; we detail the proof for the sake of completeness. One first observes that f_*, g_* of (16) can be extended to continuous functions on $\mathbb{R}^d$ such that

$$\begin{cases} \varepsilon = \displaystyle\int \left(f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2\right)_+ dP(x) & \text{for all } y \in \mathbb{R}^d, \\ \varepsilon = \displaystyle\int \left(f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2\right)_+ dQ(y) & \text{for all } x \in \mathbb{R}^d. \end{cases} \tag{21}$$

Let f_*, g_* satisfy (21). Write $F(x) = f_*(x) - \|x\|^2/2$ and $G(y) = g_*(y) - \|y\|^2/2$. We show that $F : \mathbb{R}^d \to \mathbb{R}$ is concave. Indeed, let $x, x' \in \mathbb{R}^d$ and $\rho \in [0,1]$. Convexity of $t \mapsto (t)_+$ and using (21) at both x and x' yield

$$\begin{aligned} &\int \left[\rho F(x) + (1-\rho)F(x') + G(y) + \langle \rho x + (1-\rho)x', y\rangle\right]_+ dQ(y) \\ &\leq \rho \int \left[F(x) + G(y) + \langle x, y\rangle\right]_+ dQ(y) + (1-\rho)\int \left[F(x') + G(y) + \langle x', y\rangle\right]_+ dQ(y) = \varepsilon. \end{aligned}$$

On the other hand, (21) at $x'' := \rho x + (1-\rho)x'$ yields

$$\int \left[F(\rho x + (1-\rho)x') + G(y) + \langle \rho x + (1-\rho)x', y\rangle\right]_+ dQ(y) = \varepsilon.$$

Together, it follows that $\rho F(x) + (1-\rho)F(x') \leq F(\rho x + (1-\rho)x')$, as claimed.

Analogously, G is concave. In particular, for any $x \in \Omega$, the function

$$y \mapsto f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2$$

is concave, which implies that its super-level set $\hat{\mathcal{S}}_x := \{y \in \mathbb{R}^d : f_*(x) + g_*(y) - \frac{1}{2}\|x-y\|^2 \geq 0\}$ is convex. Moreover, (21) implies that the open set $\{y \in \mathbb{R}^d : f_*(x) + g_*(y) - \frac{1}{2}\|x-y\|^2 > 0\}$

has positive Q-measure and in particular is nonempty. As a consequence, $\hat{\mathcal{S}}_x$ is a convex set with nonempty interior and its boundary in $\mathbb{R}^d$ is the zero-level set $\{y \in \mathbb{R}^d : f_*(x) + g_*(y) - \frac{1}{2}\|x-y\|^2 = 0\}$. Since the boundary of a convex set is Q-negligible by Assumption 2.1, the proof is complete.

(ii) This follows from an argument given in the proof of [1, Proposition 5.1] for a class of divergences; however, for the present case, we can also give a straightforward proof: Set

$$C := \sup_{x\in\Omega,\, y\in\Omega'} f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2.$$

For any $x \in \Omega$, (16) yields $\varepsilon = \int \left(f_*(x) + g_*(y) - \frac{1}{2}\|x-y\|^2\right)_+ dQ(y) \le Q(\mathcal{S}_x)C$, and now the claim $Q(\mathcal{S}_x) \ge \lambda$ follows with $\lambda := \varepsilon/C$. The proof of $P(\mathcal{T}_y) \ge \lambda$ is analogous. □

The last lemma shows that the case $\operatorname{spt}\pi_* = \Omega \times \Omega'$ of full support is straightforward.

Lemma 3.3. *Set $p = \int x dP$ and $q = \int y dQ$. Then $\operatorname{spt}\pi_* = \Omega \times \Omega'$ if and only if*

$$\varepsilon + pq - qx - py + \langle x, y\rangle \ge 0 \quad \textit{for all } (x,y) \in \Omega \times \Omega'. \tag{22}$$

In that case, the potentials are given by

$$f_*(x) = \frac{1}{2}\|x\|^2 - qx + \frac{\varepsilon + pq}{2}, \qquad g_*(y) = \frac{1}{2}\|y\|^2 - py + \frac{\varepsilon + pq}{2}. \tag{23}$$

Proof. Let $f(x)$ and $g(y)$ denote the above right-hand sides and note that (22) is equivalent to $f(x) + g(y) - \frac{1}{2}\|x-y\|^2 \ge 0$ for all $(x,y) \in \Omega \times \Omega'$. Assuming $\operatorname{spt}\pi_* = \Omega \times \Omega'$, we see from (11) that the potentials satisfy $f_*(x) + g_*(y) - \frac{1}{2}\|x-y\|^2 \ge 0$ and hence (16) simplifies to

$$\varepsilon = \int f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2 \, dP(x),$$
$$\varepsilon = \int f_*(x) + g_*(y) - \tfrac{1}{2}\|x-y\|^2 \, dQ(y).$$

It is then straightforward to verify $(f_*, g_*) = (f, g)$. Conversely, if $f(x) + g(y) - \frac{1}{2}\|x-y\|^2 \ge 0$ for all $(x,y) \in \Omega \times \Omega'$, the fact that (f,g) solves the above system means that (f,g) also solves (16), and we conclude by the uniqueness in Lemma 3.1. □

4 Contractivity of linearized gradient descent

In this section, we study the (formal) linearization of the gradient descent operator

$$\begin{pmatrix} f \\ g \end{pmatrix} \mapsto \begin{pmatrix} f \\ g \end{pmatrix} + \eta \cdot \mathrm{D}\Gamma \begin{pmatrix} f \\ g \end{pmatrix} \tag{24}$$

at the dual optimum (f_*, g_*); namely, the operator $\mathbb{L} : L^2_\oplus \to L^2_\oplus$,

$$\mathbb{L}\begin{pmatrix} f \\ g \end{pmatrix} = \operatorname{proj}_\oplus \begin{pmatrix} f\left(1 - \frac{\eta}{\varepsilon} Q(\mathcal{S}_{(\cdot)})\right) - \frac{\eta}{\varepsilon}\int_{\mathcal{S}_{(\cdot)}} g \, dQ \\ g\left(1 - \frac{\eta}{\varepsilon} P(\mathcal{T}_{(\cdot)})\right) - \frac{\eta}{\varepsilon}\int_{\mathcal{T}_{(\cdot)}} f \, dP \end{pmatrix}, \tag{25}$$

where $x \mapsto \mathcal{S}_x$ and $y \mapsto \mathcal{T}_y$ were defined in (19) and (20).[1] We will show that $\mathbb{L}$ is self-adjoint on the Hilbert space $(L^2_\oplus, \langle\cdot,\cdot\rangle_{L^2_\oplus})$, so that its operator norm can be computed via

$$\|\mathbb{L}\|_{\mathrm{op}} = \sup_{\|(f,g)\|_{L^2_\oplus} \le 1} |\langle \mathbb{L}(f,g),(f,g)\rangle| \,.$$

The main result of this section is the following.

Proposition 4.1. *Under Assumption 2.2, the operator $\mathbb{L} : L^2_\oplus \to L^2_\oplus$ is a strict contraction; i.e., $\|\mathbb{L}\|_{\mathrm{op}} < 1$.*

The proof, which occupies the rest of the section, has the following structure. Since $\mathbb{L}$ is self-adjoint, either $\alpha := \|\mathbb{L}\|_{\mathrm{op}}$ or $\alpha := -\|\mathbb{L}\|_{\mathrm{op}}$ belongs to the spectrum of $\mathbb{L}$, and

$$\alpha = \sup_{\|(f,g)\|_{L^2_\oplus} = 1} \langle \mathbb{L}(f,g),(f,g)\rangle_{L^2_\oplus} \quad \text{or} \quad \alpha = \inf_{\|(f,g)\|_{L^2_\oplus} = 1} \langle \mathbb{L}(f,g),(f,g)\rangle_{L^2_\oplus}.$$

We deduce that

$$\left\| \begin{pmatrix} f_n((1-\alpha) - \frac{\eta}{\varepsilon} Q(\mathcal{S}_{(\cdot)})) - \frac{\eta}{\varepsilon}\int_{\mathcal{S}_{(\cdot)}} g_n dQ + b_n \\ g_n((1-\alpha) - \frac{\eta}{\varepsilon} P(\mathcal{T}_{(\cdot)})) - \frac{\eta}{\varepsilon}\int_{\mathcal{T}_{(\cdot)}} f_n dQ - b_n \end{pmatrix} \right\|_{L^2(P)\times L^2(Q)} \to 0$$

for a sequence $(f_n, g_n) \in L^2_\oplus$ with unit norm. Moreover, we establish that the sequence

$$\begin{pmatrix} \int_{\mathcal{S}_{(\cdot)}} g_n dQ \\ \int_{\mathcal{T}_{(\cdot)}} f_n dQ \end{pmatrix} \in L^2(P) \times L^2(Q)$$

is strongly pre-compact. We deduce that $(f_n, g_n)_n$ is strongly convergent along a subsequence, and then, that one of the numbers $\pm\|\mathbb{L}\|_{\mathrm{op}}$ is an eigenvalue of $\mathbb{L}$. We conclude by proving that all eigenvalues of $\mathbb{L}$ lie in the open interval $(-1,1)$, which is the main part of the argument.

We start with several auxiliary results and detail the proof of Proposition 4.1 at the end of the section. Denote by $\mathcal{B}_r(x)$ the open ball of radius r around x.

Lemma 4.2. *The following operators are compact:*

$$\mathbb{A}_1 : L^2(Q) \to L^2(P), \qquad \mathbb{A}_1(g) = \frac{\int_{\mathcal{S}_{(\cdot)}} g dQ}{Q(\mathcal{S}_{(\cdot)})},$$

$$\mathbb{A}_2 : L^2(P) \to L^2(Q), \qquad \mathbb{A}_2(f) = \frac{\int_{\mathcal{T}_{(\cdot)}} f dP}{P(\mathcal{T}_{(\cdot)})}.$$

Proof. We prove that $\mathbb{A}_1$ is compact, the second claim is analogous. It suffices to show that if $\{u_n\}_n \subset L^2(Q)$ converges weakly to zero, then $\|\mathbb{A}_1(u_n)\|_{L^2(P)} \to 0$. In view of Lemma 3.2 (ii),

$$\|\mathbb{A}_1(u_n)\|^2_{L^2(P)} = \int \left(\frac{\int_{\mathcal{S}_x} u_n(y)\, dQ(y)}{Q(\mathcal{S}_x)} \right)^2 dP(x) \le \lambda^{-2} \int \left(\int_{\mathcal{S}_x} u_n(y)\, dQ(y) \right)^2 dP(x).$$

[1] The expression in (25) serves as the definition of $\mathbb{L}$. While linearizing (24) is the intuition giving rise to (25), differentiability of (24) may not hold under the stated assumptions as $\partial\mathcal{T}_y$ need not be P-negligible.

Set $h_n(x) := \left(\int_{\mathcal{S}_x} u_n(y)\, dQ(y)\right)^2$. As $u_n \to 0$ weakly in $L^2(Q)$,

$$h_n(x) = \langle \mathbb{I}_{\mathcal{S}_x}, u_n \rangle^2_{L^2(Q)} \to 0 \quad \text{for all } x \in \Omega.$$

On the other hand, Jensen's inequality yields $h_n(x) \le \|u_n\|^2_{L^2(Q)}$. As weakly convergent sequences are norm-bounded, this shows that (h_n) is uniformly bounded by a constant. Using the dominated convergence theorem, we conclude that $\|\mathbb{A}_1(u_n)\|_{L^2(P)} \to 0$. □

Lemma 4.3. *For $\mathbb{A}_1, \mathbb{A}_2$ as defined in Lemma 4.2, the equation*

$$\begin{pmatrix} \mathbb{A}_1(g) \\ \mathbb{A}_2(f) \end{pmatrix} = -\begin{pmatrix} f \\ g \end{pmatrix} \tag{26}$$

on $L^2(P) \times L^2(Q)$ has the solution set $\{(f,g) = (c,-c) : c \in \mathbb{R}\}$.

Proof. Let $(f,g) \in L^2(P) \times L^2(Q)$ be any solution of (26), and recall the definition (18). Using the first row of (26) yields

$$\int_{\mathcal{E}} (f(x)+g(y))h(x)dQ(y)dP(x) = \int_{\Omega} \left(\int_{\mathcal{S}_x} (f(x)+g(y))h(x)dQ(y) \right) dP(x) = 0$$

for all $h \in L^2(P)$. In particular, choosing $h := f$ yields

$$0 = \int_{\mathcal{E}} (f^2(x) + f(x)g(y))dQ(y)dP(x).$$

Analogously,

$$\int_{\mathcal{E}} (g^2(y) + f(x)g(y))dQ(y)dP(x) = 0.$$

Adding the two displays shows that $\int_{\mathcal{E}} (g(y)+f(x))^2 dQ(y)dP(x) = 0$ and hence that $g \oplus f = 0$ holds $P \otimes Q$-a.s. in $\mathcal{E}$.

Next, we argue that f and g admit continuous versions. Recall (11) and that f_*, g_* are continuous (see [24, Lemma 2.6]). In particular, π_* is supported in $\mathcal{E}$ and its density is continuous. For any test function $h \in L^2(P)$, it follows from $g \oplus f = 0$ $P \otimes Q$-a.s. in $\mathcal{E}$ that

$$\begin{aligned} 0 &= \int h(x)(g(y)+f(x)) \frac{d\pi_*}{d(P \otimes Q)}(x,y)dP(x)dQ(y) \\ &= \int \left(h(x)f(x) + h(x) \int \frac{d\pi_*}{d(P \otimes Q)}(x,y)g(y)dQ(y) \right) dP(x). \end{aligned}$$

Therefore,

$$f(x) = -\int \frac{d\pi_*}{d(P \otimes Q)}(x,y)g(y)dQ(y) \quad P\text{-a.s.}, \tag{27}$$

and as the right-hand side is continuous in x, this shows that f admits a continuous version. The argument for g is analogous.

It remains to show that the continuous versions f, g with $g \oplus f = 0$ on $\mathcal{E}$ satisfy $(f,g) = (-c, c)$ on $\Omega \times \Omega'$, for some $c \in \mathbb{R}$. Let $y \in \Omega'$. As in (17), there exist $x_0 \in \Omega$ and a neighborhood U_y of y in Ω' such that $\{x_0\} \times U_y \subset \mathcal{E}$. Thus $g(\tilde{y}) = -f(x_0)$ for all $\tilde{y} \in U_y$. We have shown that g is locally constant. As Ω' is connected by Assumption 2.1, it follows that $g \equiv c$ is constant on Ω', and now (26) implies that $f \equiv -c$ on Ω. □

Lemma 4.4. *The operator $\mathbb{L}$ is self-adjoint in $(L^2_\oplus, \langle\cdot,\cdot\rangle_{L^2_\oplus})$.*

Proof. We first consider the auxiliary operator $\mathbb{M} : L^2(P) \times L^2(Q)$ defined by

$$\mathbb{M}\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} f \cdot Q(\mathcal{S}_{(\cdot)}) + \displaystyle\int_{\mathcal{S}_{(\cdot)}} g\, dQ \\ g \cdot P(\mathcal{T}_{(\cdot)}) + \displaystyle\int_{\mathcal{T}_{(\cdot)}} f\, dP \end{pmatrix}.$$

The following representation readily yields that $\mathbb{M}$ is self-adjoint:

$$\mathbb{M}\left(\begin{array}{c} f \\ g \end{array}\right)(x,y) = \left(\begin{array}{c} \int \mathbb{I}_{\mathcal{E}}(x,y')(f(x)+g(y'))dQ(y') \\ \int \mathbb{I}_{\mathcal{E}}(x',y)(f(x')+g(y))dP(x') \end{array}\right),$$

where

$$\mathcal{E} = \left\{(x,y) \in \Omega \times \Omega' : f_*(x) + g_*(y) \geq \frac{1}{2}\|x-y\|^2\right\}.$$

As a consequence, $\mathbb{M}$ induces a self-adjoint operator $\mathbb{M}_\oplus := \mathrm{proj}_\oplus \mathbb{M}$ on $L^2_\oplus$. Indeed, for every $(f,g),(u,v) \in L^2_\oplus$ we have

$$\begin{aligned}\langle (f,g), \mathrm{proj}_\oplus \mathbb{M}(u,v)\rangle_{L^2_\oplus} &= \langle (f,g), \mathbb{M}(u,v)\rangle_{L^2(P)\times L^2(Q)} \\ &= \langle \mathbb{M}(f,g),(u,v)\rangle_{L^2(P)\times L^2(Q)} = \langle \mathrm{proj}_\oplus \mathbb{M}(f,g),(u,v)\rangle_{L^2_\oplus}.\end{aligned}$$

Recalling from (25) that $\mathbb{L} = \mathbb{I} - \frac{\eta}{\varepsilon}\mathbb{M}_\oplus$, it follows that $\mathbb{L}$ is also self-adjoint. □

Proof of Proposition 4.1. To show that $\|\mathbb{L}\|_{\mathrm{op}} < 1$, recall (e.g., [3, Proposition 6.9]) that Lemma 4.4 implies $\|\mathbb{L}\|_{\mathrm{op}} = \max\{|\alpha^+|, |\alpha^-|\}$ where

$$\alpha^+ := \sup_{\|(f,g)\|_{L^2_\oplus}=1} \langle \mathbb{L}(f,g),(f,g)\rangle_{L^2_\oplus}, \qquad \alpha^- := \inf_{\|(f,g)\|_{L^2_\oplus}=1} \langle \mathbb{L}(f,g),(f,g)\rangle_{L^2_\oplus}.$$

We set

$$\alpha := \begin{cases} \alpha^+, & \text{if } \|\mathbb{L}\|_{\mathrm{op}} = |\alpha^+|, \\ \alpha^-, & \text{otherwise.} \end{cases} \tag{28}$$

Suppose for contradiction that $\|\mathbb{L}\|_{\mathrm{op}} \geq 1$. Note that this implies $\alpha \geq 1$ in the first case of (28) and $\alpha \leq -1$ in the second.

By the definition of α, there exists a sequence $(f_n,g_n)_n$ with $\|(f_n,g_n)\|_{L^2_\oplus} = 1$ such that $\langle \mathbb{L}(f_n,g_n),(f_n,g_n)\rangle_{L^2_\oplus} \to \alpha$. Recall that (f_n,g_n) can be considered as elements of $L^2(P) \times L^2(Q)$ with $\int f_n dP = \int g_n dQ$. Note that $\|\mathbb{L}\|_{\mathrm{op}} = |\alpha|$ implies $\|\mathbb{L}(f_n,g_n)\|^2_{L^2_\oplus} \leq \alpha^2$ and hence

$$\begin{aligned}\|(\mathbb{L}-\alpha\mathbb{I})(f_n,g_n)\|^2_{L^2_\oplus} &= -2\alpha\langle \mathbb{L}(f_n,g_n),(f_n,g_n)\rangle_{L^2_\oplus} + \alpha^2 + \|\mathbb{L}(f_n,g_n)\|^2_{L^2_\oplus} \\ &\leq 2\alpha(\alpha - \langle \mathbb{L}(f_n,g_n),(f_n,g_n)\rangle_{L^2_\oplus}) \to 0.\end{aligned}$$

Hence, there exists a sequence $\{b_n\}_n \subset \mathbb{R}$ such that

$$\left\|\left(\begin{array}{c} f_n((1-\alpha) - \frac{\eta}{\varepsilon}Q(\mathcal{S}_{(\cdot)})) - \frac{\eta}{\varepsilon}\int_{\mathcal{S}_{(\cdot)}} g_n dQ + b_n \\ g_n((1-\alpha) - \frac{\eta}{\varepsilon}P(\mathcal{T}_{(\cdot)})) - \frac{\eta}{\varepsilon}\int_{\mathcal{T}_{(\cdot)}} f_n dQ - b_n \end{array}\right)\right\|_{L^2(P)\times L^2(Q)} \to 0. \tag{29}$$

In fact, the sequence $\{b_n\}_n$ is bounded by the choice of (f_n, g_n), hence converges to a limit b after passing to a subsequence. After passing to another subsequence, the Banach–Alaoglu theorem yields that the bounded sequence $(f_n, g_n)_n$ has a weak limit (f, g) in $L^2(P) \times L^2(Q)$. Recalling Lemma 4.2 and the fact that compact operators map weakly convergent to strongly convergent sequences, it follows that

$$\left\| \begin{pmatrix} \mathbb{A}_1(g_n) \\ \mathbb{A}_2(f_n) \end{pmatrix} - \begin{pmatrix} \mathbb{A}_1(g) \\ \mathbb{A}_2(f) \end{pmatrix} \right\|_{L^2(P)\times L^2(Q)} \to 0.$$

Together with (29), we deduce that

$$\begin{pmatrix} f_n \\ g_n \end{pmatrix} \to \begin{pmatrix} \frac{\frac{\eta}{\varepsilon}\int_{\mathcal{S}_{(\cdot)}} g dQ + b}{(1-\alpha) - \frac{\eta}{\varepsilon}Q(\mathcal{S}_{(\cdot)})} \\ \frac{\frac{\eta}{\varepsilon}\int_{\mathcal{T}_{(\cdot)}} f dP - b}{(1-\alpha) - \frac{\eta}{\varepsilon}P(\mathcal{T}_{(\cdot)})} \end{pmatrix} \quad \text{in } L^2(P) \times L^2(Q),$$

where we have used that the denominator is bounded away from zero thanks to $|\alpha| \geq 1$ and $\frac{\eta}{\varepsilon}P(\mathcal{T}_{(\cdot)}), \frac{\eta}{\varepsilon}Q(\mathcal{S}_{(\cdot)}) \in (\frac{\lambda\eta}{\varepsilon}, 1]$, where λ is as in Lemma 3.2 (ii). In particular, the sequence (f_n, g_n) converges strongly. It follows that (f_n, g_n) converges strongly to its weak limit (f, g), and in particular that $\|(f, g)\|_{L^2_\oplus} = 1$ and $\int f dP = \int g dQ$. Thus, the equation

$$\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} \frac{\frac{\eta}{\varepsilon}\int_{\mathcal{S}_{(\cdot)}} g dQ + b}{(1-\alpha) - \frac{\eta}{\varepsilon}Q(\mathcal{S}_{(\cdot)})} \\ \frac{\frac{\eta}{\varepsilon}\int_{\mathcal{T}_{(\cdot)}} f dP - b}{(1-\alpha) - \frac{\eta}{\varepsilon}P(\mathcal{T}_{(\cdot)})} \end{pmatrix}$$

admits the solution $(f, g) \in L^2_\oplus$, or

$$\begin{cases} f((1-\alpha) - \frac{\eta}{\varepsilon}Q(\mathcal{S}_{(\cdot)})) = \frac{\eta}{\varepsilon}\int_{\mathcal{S}_{(\cdot)}} g dQ + b, \\ g((1-\alpha) - \frac{\eta}{\varepsilon}P(\mathcal{T}_{(\cdot)})) = \frac{\eta}{\varepsilon}\int_{\mathcal{T}_{(\cdot)}} f dP - b. \end{cases} \tag{30}$$

Next, we show that $b = 0$. Integrating the first and second equation of (30) with respect to P and Q, respectively, and applying Fubini's theorem, we obtain

$$(1-\alpha)\int f(x)\, dP(x) = \frac{\eta}{\varepsilon}\int_{\mathcal{E}} (f(x) + g(y))\, d(P \otimes Q)(x, y) + b, \tag{31}$$

$$(1-\alpha)\int g(y)\, dQ(y) = \frac{\eta}{\varepsilon}\int_{\mathcal{E}} (f(x) + g(y))\, d(P \otimes Q)(x, y) - b. \tag{32}$$

Subtracting (32) from (31), we find $(1-\alpha)\left(\int f\, dP - \int g\, dQ\right) = 2b$, and recalling that $\int f dP = \int g dQ$, we conclude $b = 0$.

In summary, (f, g) satisfies $\|(f, g)\|_{L^2_\oplus} = 1$ and $\int f dP = \int g dQ$ and solves the system

$$\begin{cases} f\left((1-\alpha) - \dfrac{\eta}{\varepsilon}Q(\mathcal{S}_{(\cdot)})\right) = \dfrac{\eta}{\varepsilon}\int_{\mathcal{S}_{(\cdot)}} g\, dQ, \\ g\left((1-\alpha) - \dfrac{\eta}{\varepsilon}P(\mathcal{T}_{(\cdot)})\right) = \dfrac{\eta}{\varepsilon}\int_{\mathcal{T}_{(\cdot)}} f\, dP. \end{cases} \tag{33}$$

Case $\alpha > 1$ *or* $\alpha < -1$. Then, (33) implies

$$\|f\|_\infty \le \begin{cases} \|f\|_\infty \sup_{x,y} \frac{\frac{\eta}{\varepsilon}Q(\mathcal{S}_x)}{|1-\alpha|+\frac{\eta}{\varepsilon}Q(\mathcal{S}_x)} \frac{\frac{\eta}{\varepsilon}P(\mathcal{T}_y)}{|1-\alpha|+\frac{\eta}{\varepsilon}P(\mathcal{T}_y)} & \text{if } \alpha > 1, \\ \|f\|_\infty \sup_{x,y} \frac{\frac{\eta}{\varepsilon}Q(\mathcal{S}_x)}{1-\alpha-\frac{\eta}{\varepsilon}Q(\mathcal{S}_x)} \frac{\frac{\eta}{\varepsilon}P(\mathcal{T}_y)}{1-\alpha-\frac{\eta}{\varepsilon}P(\mathcal{T}_y)} & \text{if } \alpha < -1. \end{cases}$$

We observe that the above supremum is < 1 in either case, and thus that $f = 0$. Similarly, $g = 0$, contradicting that $\|(f,g)\|_{L^2_\oplus} = 1$.

Case $\alpha = 1$. In this case, (33) specializes to

$$\begin{cases} fQ(\mathcal{S}_{(\cdot)}) = -\int_{\mathcal{S}_{(\cdot)}} g dQ, \\ gP(\mathcal{T}_{(\cdot)}) = -\int_{\mathcal{T}_{(\cdot)}} f dP, \end{cases}$$

which by Lemma 4.3 means that $(f,g) = (0,0)$ in $L^2_\oplus$, again contradicting $\|(f,g)\|_{L^2_\oplus} = 1$.

Case $\alpha = -1$. In this case, (33) can be written as

$$\begin{cases} 2f(x) = \frac{\eta}{\varepsilon} \int_{\mathcal{S}_x} (f(x) + g(y'))dQ(y'), \\ 2g(y) = \frac{\eta}{\varepsilon} \int_{\mathcal{T}_y} (f(x') + g(y))dP(x'). \end{cases}$$

Adding these equations, taking squares, applying the inequality $(a+b)^2 \le 2(a^2+b^2)$ and then Jensen's inequality, we deduce

$$\begin{aligned} 4\left(f(x)+g(y)\right)^2 &= \left(\frac{\eta}{\varepsilon}\right)^2 \left(\int_{\mathcal{S}_x} (f(x)+g(y'))dQ(y') + \int_{\mathcal{T}_y} (f(x')+g(y))dP(x') \right)^2 \\ &\le 2\left(\frac{\eta}{\varepsilon}\right)^2 \left(\left(\int_{\mathcal{S}_x} (f(x)+g(y'))dQ(y') \right)^2 + \left(\int_{\mathcal{T}_y} (f(x')+g(y))dP(x') \right)^2 \right) \\ &\le 2\left(\frac{\eta}{\varepsilon}\right)^2 \left(Q(\mathcal{S}_x) \int (f(x)+g(y'))^2 dQ(y') + P(\mathcal{T}_y) \int (f(x')+g(y))^2 dP(x') \right) \end{aligned}$$

and now integrating w.r.t. $P \otimes Q$ yields

$$\int \left(f(x)+g(y)\right)^2 dP(x)dQ(y) \le \left(\frac{\eta}{\varepsilon}\right)^2 (P \otimes Q)(\mathcal{E}) \int (f(x)+g(y))^2 dP(x)dQ(y).$$

Recalling Assumption 2.2, we have either $\eta < \varepsilon$, or $\eta = \varepsilon$ and $(P \otimes Q)(\mathcal{E}) < 1$. As a consequence, the factor $\left(\frac{\eta}{\varepsilon}\right)^2 (P \otimes Q)(\mathcal{E})$ is strictly smaller than one and we obtain the desired contradiction to $\|(f,g)\|_{L^2_\oplus} = 1$. □

5 Proof of linear convergence

Recall from (15) that the gradient descent iterates satisfy

$$\begin{pmatrix} f_{n+1} - f_n \\ g_{n+1} - g_n \end{pmatrix} = \frac{\eta}{\varepsilon} \cdot \operatorname{proj}_\oplus \begin{pmatrix} \varepsilon - \int \left(f_n(\cdot) + g_n(y) - \frac{1}{2}\|\cdot - y\|^2\right)_+ dQ(y) \\ \varepsilon - \int \left(f_n(x) + g_n(\cdot) - \frac{1}{2}\|x - \cdot\|^2\right)_+ dP(x) \end{pmatrix}. \qquad (34)$$

We first represent the iterates in a form convenient for our analysis, with the operator $\mathbb{L}_n$ introduced in Lemma 5.1. To prove the main result, it then remains to show that $\mathbb{L}_n$ converges to $\mathbb{L}$ in operator norm. We first show that $(f_n, g_n) \to (f_*, g_*)$; this is a straightforward Arzelà–Ascoli argument (Lemma 5.4). Next, the proof of $\mathbb{L}_n \to \mathbb{L}$ is given in Proposition 5.5. Combining $\|\mathbb{L}_n - \mathbb{L}\|_{\mathrm{op}} \to 0$ with the fact that $\mathbb{L}$ is a contraction (Proposition 4.1), we complete the proof of the main result in Corollary 5.6.

Let $\mathcal{L}_1(\lambda)$ denote one-dimensional Lebesgue measure.

Lemma 5.1. *The gradient descent iterates (f_n, g_n) satisfy*

$$\begin{pmatrix} f_{n+1} - f_* \\ g_{n+1} - g_* \end{pmatrix} = \mathbb{L}_n \begin{pmatrix} f_n - f_* \\ g_n - g_* \end{pmatrix} \tag{35}$$

for the operator

$$\mathbb{L}_n \begin{pmatrix} f \\ g \end{pmatrix} := \mathrm{proj}_\oplus \begin{pmatrix} f(1 - \frac{\eta}{\varepsilon} \cdot [\mathcal{L}_1 \otimes Q](\mathcal{S}_{n,(\cdot)})) - \frac{\eta}{\varepsilon} \cdot \int_{\mathcal{S}_{n,(\cdot)}} g(y) d[\mathcal{L}_1 \otimes Q](\lambda, y) \\ g(1 - \frac{\eta}{\varepsilon} \cdot [\mathcal{L}_1 \otimes P](\mathcal{T}_{n,(\cdot)})) - \frac{\eta}{\varepsilon} \cdot \int_{\mathcal{T}_{n,(\cdot)}} f(x) d[\mathcal{L}_1 \otimes P](\lambda, x) \end{pmatrix}, \tag{36}$$

where

$$\mathcal{S}_{n,x} := \left\{ (\lambda, y) \in [0,1] \times \Omega' : \ \lambda(f_*(x) + g_*(y)) + (1-\lambda)(f_n(x) + g_n(y)) \geq \frac{1}{2}\|x - y\|^2 \right\},$$
$$\mathcal{T}_{n,y} := \left\{ (\lambda, x) \in [0,1] \times \Omega : \ \lambda(f_*(x) + g_*(y)) + (1-\lambda)(f_n(x) + g_n(y)) \geq \frac{1}{2}\|x - y\|^2 \right\}.$$

Proof. In view of (16), (34) implies that

$$\begin{aligned} &\begin{pmatrix} f_{n+1} - f_* \\ g_{n+1} - g_* \end{pmatrix} - \begin{pmatrix} f_n - f_* \\ g_n - g_* \end{pmatrix} \\ &= \frac{\eta}{\varepsilon} \cdot \mathrm{proj}_\oplus \begin{pmatrix} \int \left(f_*(\cdot) + g_*(y) - \frac{1}{2}\| \cdot - y\|^2\right)_+ - \left(f_n(\cdot) + g_n(y) - \frac{1}{2}\| \cdot - y\|^2\right)_+ dQ(y) \\ \int \left(f_*(x) + g_*(\cdot) - \frac{1}{2}\|x - \cdot\|^2\right)_+ - \left(f_n(x) + g_n(\cdot) - \frac{1}{2}\|x - \cdot\|^2\right)_+ dP(x) \end{pmatrix}. \end{aligned} \tag{37}$$

After applying the fundamental theorem of calculus in the form

$$\phi(1)_+ - \phi(0)_+ = \int_0^1 \frac{d}{d\lambda}[\phi(\lambda)_+] \, d\mathcal{L}_1(\lambda)$$

to the functions

$$\begin{aligned} \phi(\lambda) &= \lambda(f_*(x) + g_*(y)) + (1-\lambda)(f_n(x) + g_n(y)) - \frac{1}{2}\|x - y\|^2, \\ \frac{d}{d\lambda}[\phi(\lambda)_+] &= \mathbb{I}_{\phi(\lambda) \geq 0} \cdot (f_*(x) + g_*(y) - f_n(x) - g_n(y)), \end{aligned}$$

we get

$$\begin{aligned} &\begin{pmatrix} f_{n+1} - f_* \\ g_{n+1} - g_* \end{pmatrix} - \begin{pmatrix} f_n - f_* \\ g_n - g_* \end{pmatrix} \\ &\qquad = -\frac{\eta}{\varepsilon} \cdot \mathrm{proj}_\oplus \begin{pmatrix} (f_n - f_*)[\mathcal{L}_1 \otimes Q](\mathcal{S}_{n,(\cdot)}) + \int_{\mathcal{S}_{n,(\cdot)}} (g_n(y) - g_*(y)) d[\mathcal{L}_1 \otimes Q](\lambda, y) \\ (g_n - g_*)[\mathcal{L}_1 \otimes P](\mathcal{T}_{n,(\cdot)}) + \int_{\mathcal{T}_{n,(\cdot)}} (f_n(x) - f_*(x)) d[\mathcal{L}_1 \otimes P](\lambda, x) \end{pmatrix} \end{aligned}$$

and the claim follows. □

The next two lemmas establish that the gradient descent iterates are uniformly bounded and equicontinuous.

Lemma 5.2. *Let $\eta \in (0, \varepsilon]$, then*

$$\left\| \begin{pmatrix} f_{n+1} - f_* \\ g_{n+1} - g_* \end{pmatrix} \right\|_{\mathcal{C}_\oplus} \leq 2 \cdot \left\| \begin{pmatrix} f_0 - f_* \\ g_0 - g_* \end{pmatrix} \right\|_{\mathcal{C}_\oplus}. \tag{38}$$

Proof. Up to taking equivalence classes in $\mathcal{C}_\oplus$, (37) states that

$$\begin{aligned} \begin{pmatrix} f_{n+1} - f_* \\ g_{n+1} - g_* \end{pmatrix} &= \begin{pmatrix} f_n - f_* \\ g_n - g_* \end{pmatrix} \\ &+ \frac{\eta}{\varepsilon} \cdot \begin{pmatrix} \int \left(f_*(\cdot) + g_*(y) - \frac{1}{2}\| \cdot - y\|^2\right)_+ - \left(f_n(\cdot) + g_n(y) - \frac{1}{2}\| \cdot - y\|^2\right)_+ dQ(y) \\ \int \left(f_*(x) + g_*(\cdot) - \frac{1}{2}\|x - \cdot\|^2\right)_+ - \left(f_n(x) + g_n(\cdot) - \frac{1}{2}\|x - \cdot\|^2\right)_+ dP(x) \end{pmatrix}. \end{aligned}$$

Using the inequality

$$(t)_+ - (s)_+ \leq \mathbb{I}_{\{t \geq 0\}}(t - s) \tag{39}$$

with $t = f_*(x) + g_*(y) - \frac{1}{2}\|x - y\|^2$ and $s = f_n(x) + g_n(y) - \frac{1}{2}\|x - y\|^2$, we infer

$$\begin{aligned} f_{n+1}(x) - f_*(x) &\leq (f_n(x) - f_*(x)) + \frac{\eta}{\varepsilon} \cdot \int_{\mathcal{S}_x} (f_*(x) - f_n(x)) + (g_*(y) - g_n(y)) dQ(y) \\ &= (f_n(x) - f_*(x)) \left(1 - \frac{\eta}{\varepsilon} \cdot Q(\mathcal{S}_x)\right) - \frac{\eta}{\varepsilon} \cdot \int_{\mathcal{S}_x} (g_n(y) - g_*(y)) dQ(y) \\ &\leq \|f_n - f_*\|_\infty \left(1 - \frac{\eta}{\varepsilon} \cdot Q(\mathcal{S}_x)\right) + \frac{\eta}{\varepsilon} \cdot Q(\mathcal{S}_x) \|g_n - g_*\|_\infty. \end{aligned}$$

Note that the right-hand side is a convex combination of $\|f_n - f_*\|_\infty$ and $\|g_n - g_*\|_\infty$ as $\eta \in (0, \varepsilon]$. Writing

$$\tilde{\mathcal{S}}_{n,x} := \left\{ y \in \Omega' : \ f_n(x) + g_n(y) \geq \frac{1}{2}\|x - y\|^2 \right\}, \quad x \in \Omega, \ n \in \mathbb{N}, \tag{40}$$

we analogously get

$$f_*(x) - f_{n+1}(x) \leq \|f_n - f_*\|_\infty \left(1 - \frac{\eta}{\varepsilon} \cdot Q(\tilde{\mathcal{S}}_{n,x})\right) + \frac{\eta}{\varepsilon} \cdot Q(\tilde{\mathcal{S}}_{n,x}) \|g_n - g_*\|_\infty,$$

a convex combination of the same quantities. Together, it follows that

$$\|f_{n+1} - f_*\|_\infty \leq \max\left(\|f_n - f_*\|_\infty, \|g_n - g_*\|_\infty\right).$$

Repeating the same argument for $\|g_{n+1} - g_*\|_\infty$, we conclude

$$\max\left(\|f_{n+1} - f_*\|_\infty, \|g_{n+1} - g_*\|_\infty\right) \leq \max\left(\|f_n - f_*\|_\infty, \|g_n - g_*\|_\infty\right)$$

for all n and hence

$$\max\left(\|f_{n+1} - f_*\|_\infty, \|g_{n+1} - g_*\|_\infty\right) \leq \max\left(\|f_0 - f_*\|_\infty, \|g_0 - g_*\|_\infty\right).$$

The claim follows after recalling the definition of the norm $\| \cdot \|_{\mathcal{C}_\oplus}$. □

Lemma 5.3. *Let $\eta \in (0, \varepsilon]$ and let (f_0, g_0) be Lipschitz with constant L_0. Then for every $n \geq 1$, the gradient descent iterates (f_n, g_n) are Lipschitz with constant L, where*

$$L = \max\{L_0, C\}, \qquad C := \max\{\|x - y\| : x \in \Omega, y \in \Omega\}.$$

Proof. Arguing by inductively, let L_{n-1} denote the Lipschitz constant of f_{n-1}. Fix $x, x' \in \Omega$ with $x \neq x'$. By (34),

$$\begin{aligned} f_n(x) - f_n(x') = (f_{n-1}(x) - f_{n-1}(x')) &- \frac{\eta}{\varepsilon} \cdot \int \left(f_{n-1}(x) + g_{n-1}(y) - \frac{1}{2}\|x - y\|^2 \right)_+ dQ(y) \\ &+ \frac{\eta}{\varepsilon} \cdot \int \left(f_{n-1}(x') + g_{n-1}(y) - \frac{1}{2}\|x' - y\|^2 \right)_+ dQ(y). \end{aligned}$$

Recalling the definition of $\tilde{\mathcal{S}}_{n-1,x'}$ from (40) and using the inequality (39), we get

$$\begin{aligned} &f_n(x) - f_n(x') \\ &\leq (f_{n-1}(x) - f_{n-1}(x')) + \frac{\eta}{\varepsilon} \cdot \int_{\tilde{\mathcal{S}}_{n-1,x'}} (f_{n-1}(x') - f_{n-1}(x)) + \frac{\|x - y\|^2 - \|x' - y\|^2}{2} dQ(y) \\ &\leq (f_{n-1}(x) - f_{n-1}(x'))(1 - \frac{\eta}{\varepsilon} \cdot Q(\tilde{\mathcal{S}}_{n-1,x'})) + \frac{\eta}{\varepsilon} \cdot \int_{\tilde{\mathcal{S}}_{n-1,x'}} \frac{\|x - y\|^2 - \|x' - y\|^2}{2} dQ(y) \\ &\leq (f_{n-1}(x) - f_{n-1}(x'))(1 - \frac{\eta}{\varepsilon} \cdot Q(\tilde{\mathcal{S}}_{n-1,x'})) + \frac{\eta}{\varepsilon} \cdot C\|x - x'\| Q(\tilde{\mathcal{S}}_{n-1,x'}), \end{aligned}$$

where $C = \max\{\|x - y\| : x \in \Omega, y \in \Omega\}$. As a consequence,

$$\frac{f_n(x) - f_n(x')}{\|x - x'\|} \leq L_{n-1}(1 - \frac{\eta}{\varepsilon} \cdot Q(\tilde{\mathcal{S}}_{n-1,x'})) + \frac{\eta}{\varepsilon} \cdot Q(\tilde{\mathcal{S}}_{n-1,x'})C.$$

Noting that the right-hand side is a convex combination of L_{n-1} and C, we have

$$\frac{f_n(x) - f_n(x')}{\|x - x'\|} \leq \max\left(L_{n-1}, C\right)$$

and the claim for f_n follows. The proof for g_n is analogous. □

We can now conclude the convergence of (f_n, g_n) to (f_*, g_*) in $\mathcal{C}_\oplus$.

Lemma 5.4. *Let $\eta \in (0, \varepsilon]$, then $\|(f_n, g_n) - (f_*, g_*)\|_{\mathcal{C}_\oplus} \to 0$.*

Proof. In view of Lemmas 5.2 and 5.3, the Arzelà–Ascoli theorem shows that after passing to a subsequence, (f_n, g_n) converges to a limit (f_∞, g_∞) in $\mathcal{C}_\oplus$. By the continuity of $\mathrm{D}\Gamma$, it follows from (15) that

$$\begin{pmatrix} f_\infty \\ g_\infty \end{pmatrix} = \begin{pmatrix} f_\infty \\ g_\infty \end{pmatrix} + \eta \cdot \mathrm{D}\Gamma \begin{pmatrix} f_\infty \\ g_\infty \end{pmatrix} \quad \text{and hence} \quad \mathrm{D}\Gamma \begin{pmatrix} f_\infty \\ g_\infty \end{pmatrix} = 0$$

in $L^2_\oplus$, meaning that (f_∞, g_∞) solves (16). Recalling from Lemma 3.1 that (f_*, g_*) is the unique solution of (16), the claim follows. □

We can now prove the main technical result of this section.

Proposition 5.5. *Let Assumption 2.2 hold. We have* $\|\mathbb{L}_n - \mathbb{L}\|_{\mathrm{op}} \to 0$.

Proof. Comparing the definitions of $\mathbb{L}$ and $\mathbb{L}_n$ in (25) and (36), we see that $\|\mathbb{L}_n - \mathbb{L}\|_{\mathrm{op}} \to 0$ is implied by the two limits

$$\|[\mathcal{L}_1 \otimes Q](\mathcal{S}_{n,(\cdot)}) - Q(\mathcal{S}_{(\cdot)})\|_{L^2(P)} \to 0, \tag{41}$$

$$\sup_{\|h\|_{L^2(Q)\leq 1}} \left\| \int_{\mathcal{S}_{n,(\cdot)}} h(y) d[\mathcal{L}_1 \otimes Q](\lambda, y) - \int_{\mathcal{S}_{(\cdot)}} h(y) dQ(y) \right\|_{L^2(P)} \to 0 \tag{42}$$

and the symmetric results for the second component. Clearly (42) implies (41) by specializing to $h = 1$, hence we focus on (42). Separating $h = (h)_+ - (h)_-$, it even suffices to prove

$$\sup_{\|h\|_{L^2(Q)\leq 1},\, h\geq 0} \left\| \int_{\mathcal{S}_{n,(\cdot)}} h(y) d[\mathcal{L}_1 \otimes Q](\lambda, y) - \int_{\mathcal{S}_{(\cdot)}} h(y) dQ(y) \right\|_{L^2(P)} \to 0. \tag{43}$$

Consider $x \in \Omega$ and $h \geq 0$ with $\|h\|_{L^2(Q)\leq 1}$. Note that the sets

$$\mathcal{S}^{\pm}_{n,x} := \left\{ y \in \Omega' : \; f_*(x) + g_*(y) - \frac{1}{2}\|x - y\|^2 \geq \mp \|(f_n, g_n) - (f_*, g_*)\|_{\mathcal{C}_\oplus} \right\}$$

satisfy

$$[0,1] \times \mathcal{S}^{-}_{n,x} \;\subset\; \mathcal{S}_{n,x} \;\subset\; [0,1] \times \mathcal{S}^{+}_{n,x}$$

and thus

$$0 \leq \int_{\mathcal{S}^{-}_{n,x}} h(y) dQ(y) \leq \int_{\mathcal{S}_{n,x}} h(y) d[\mathcal{L}_1 \otimes Q](\lambda, y) \leq \int_{\mathcal{S}^{+}_{n,x}} h(y) dQ(y).$$

Hence, it suffices to show that

$$\mathcal{E}^{\pm}_n = \int \left(\int_{\mathcal{S}^{\pm}_{n,x}} h(y) dQ(y) - \int_{\mathcal{S}_x} h(y) dQ(y) \right)^2 dP(x) \to 0 \tag{44}$$

uniformly in h. We show this for $\mathcal{E}^{+}_n$, the proof for $\mathcal{E}^{-}_n$ is similar. In view of $\|h\|_{L^2(Q)\leq 1}$, the Cauchy–Schwarz inequality yields

$$\begin{aligned}\mathcal{E}^{+}_n &= \int \left(\int (\mathbb{I}_{\mathcal{S}^{+}_{n,x}}(y) - \mathbb{I}_{\mathcal{S}_x}(y)) h(y) dQ(y) \right)^2 dP(x) \\ &\leq \|h\|^2_{L^2(Q)} \int \|\mathbb{I}_{\mathcal{S}^{+}_{n,x}} - \mathbb{I}_{\mathcal{S}_x}\|^2_{L^2(Q)} dP(x) \leq \int \|\mathbb{I}_{\mathcal{S}^{+}_{n,x}} - \mathbb{I}_{\mathcal{S}_x}\|^2_{L^2(Q)} dP(x),\end{aligned}$$

and we note that the right-hand side is independent of h. As $|\mathbb{I}_{\mathcal{S}^{+}_{n,x}}(y) - \mathbb{I}_{\mathcal{S}_x}(y)| \leq 1$, it suffices to show that $|\mathbb{I}_{\mathcal{S}^{+}_{n,x}}(y) - \mathbb{I}_{\mathcal{S}_x}(y)| \to 0$ for $(P \otimes Q)$-almost all (x, y). Write $\xi_*(x, y) := f_*(x) + g_*(y) - \frac{1}{2}\|x - y\|^2$. Since the set $\{(x, y) \in \Omega \times \Omega' : \; \xi_*(x, y) = 0\}$ is $(P \otimes Q)$-negligible by Lemma 3.2 (i) and Fubini's theorem, it suffices to show $|\mathbb{I}_{\mathcal{S}^{+}_{n,x}}(y) - \mathbb{I}_{\mathcal{S}_x}(y)| \to 0$

for $(x, y) \in \Omega \times \Omega'$ with $\xi_*(x, y) \neq 0$. Fix such a pair (x, y). By Lemma 5.4, there exists n_0 such that

$$\|(f_n, g_n) - (f_*, g_*)\|_{\mathcal{C}_\oplus} \leq \frac{1}{2}|\xi_*(x, y)| \quad \text{for all } n \geq n_0.$$

If $\xi_*(x, y) > 0$, it follows that $\mathbb{I}_{\mathcal{S}_x}(y) = \mathbb{I}_{\mathcal{S}^+_{n,x}}(y) = 1$ for $n \geq n_0$, whereas if $\xi_*(x, y) < 0$, then $\mathbb{I}_{\mathcal{S}_x}(y) = \mathbb{I}_{\mathcal{S}^+_{n,x}}(y) = 0$ for $n \geq n_0$. In either case, $|\mathbb{I}_{\mathcal{S}^+_{n,x}}(y) - \mathbb{I}_{\mathcal{S}_x}(y)| = 0$ for $n \geq n_0$, completing the proof. □

Combining Propositions 4.1 and 5.5, we deduce that $\mathbb{L}_n$ a uniform contraction for $n \geq n_0$.

Corollary 5.6. *Let Assumption 2.2 hold. There exist $\delta_* \in (0, 1)$ and $n_0 \in \mathbb{N}$ such that $\|\mathbb{L}_n\|_{\mathrm{op}} \leq \delta_*$ for all $n \geq n_0$.*

In view of (35), Corollary 5.6 implies Theorem 2.3.

6 Numerical experiments

In this section, we provide numerical experiments for the gradient descent algorithm. The key quantity of interest is

$$\Delta_n := \|(f_n, g_n) - (f_{n-1}, g_{n-1})\|_{L^2_\oplus}.$$

On the one hand, we see from (34) that

$$\Delta_n = \frac{\eta}{\varepsilon} \left\| \begin{pmatrix} \varepsilon - \int \left(f_n(\cdot) + g_n(y) - \frac{1}{2}\|\cdot - y\|^2\right)_+ dQ(y) \\ \varepsilon - \int \left(f_n(x) + g_n(\cdot) - \frac{1}{2}\|x - \cdot\|^2\right)_+ dP(x) \end{pmatrix} \right\|_{L^2_\oplus}$$

and hence Δ_n has a direct interpretation as η/ε times the L^2 norm of the constraint violation in (16). This captures how far the measure with density $\left(f_n(x) + g_n(y) - \frac{1}{2}\|x - y\|^2\right)_+$ is from being a coupling of P and Q. On the other hand, we will observe that $\Delta_n \approx \delta_*^n$ for large n, for some $\delta_* \in (0, 1)$, implying that also $\|(f_n, g_n) - (f_*, g_*)\|_{L^2_\oplus} \approx \delta_*^n$. This will confirm that Theorem 2.3 accurately captures the convergence behavior.

The left panels in Figure 2 plot $\log(\Delta_n)$ against n for three different pairs of marginals (P, Q) and for numerous different step sizes η. The regularization parameter $\varepsilon = 10^{-1}$ and the initialization $f_0 \equiv g_0 \equiv 0.5$ are fixed. The gradient descent iteration is run until $\Delta_n \leq 10^{-10}$. The integrals in (14) are approximated using the regular trapezoid rule over a discretized mesh of [-0.1, 1.6], with step size 0.001, giving approximately 1700 discretization points for the one-dimensional marginals, while the same integration procedure is applied on a mesh of $[-0.1, 2] \times [-0.1, 2]$, with step size 0.05, to obtain approximately the same number of discretization points.

For all step sizes $\eta \leq \varepsilon$, we observe a linear behavior after a burn-in period, confirming that $\Delta_n \approx \delta_*^n$ where $\delta_* \in (0, 1)$. We also show some step sizes with $\eta/\varepsilon > 1$; specifically, we increase the step size until convergence breaks. In the three experiments, convergence breaks when η/ε reaches the values 1.13, 1.28, and 1.42. This suggests that the condition $\eta/\varepsilon \leq 1$ in Assumption 2.2 is fairly sharp; see also Remark 2.4.

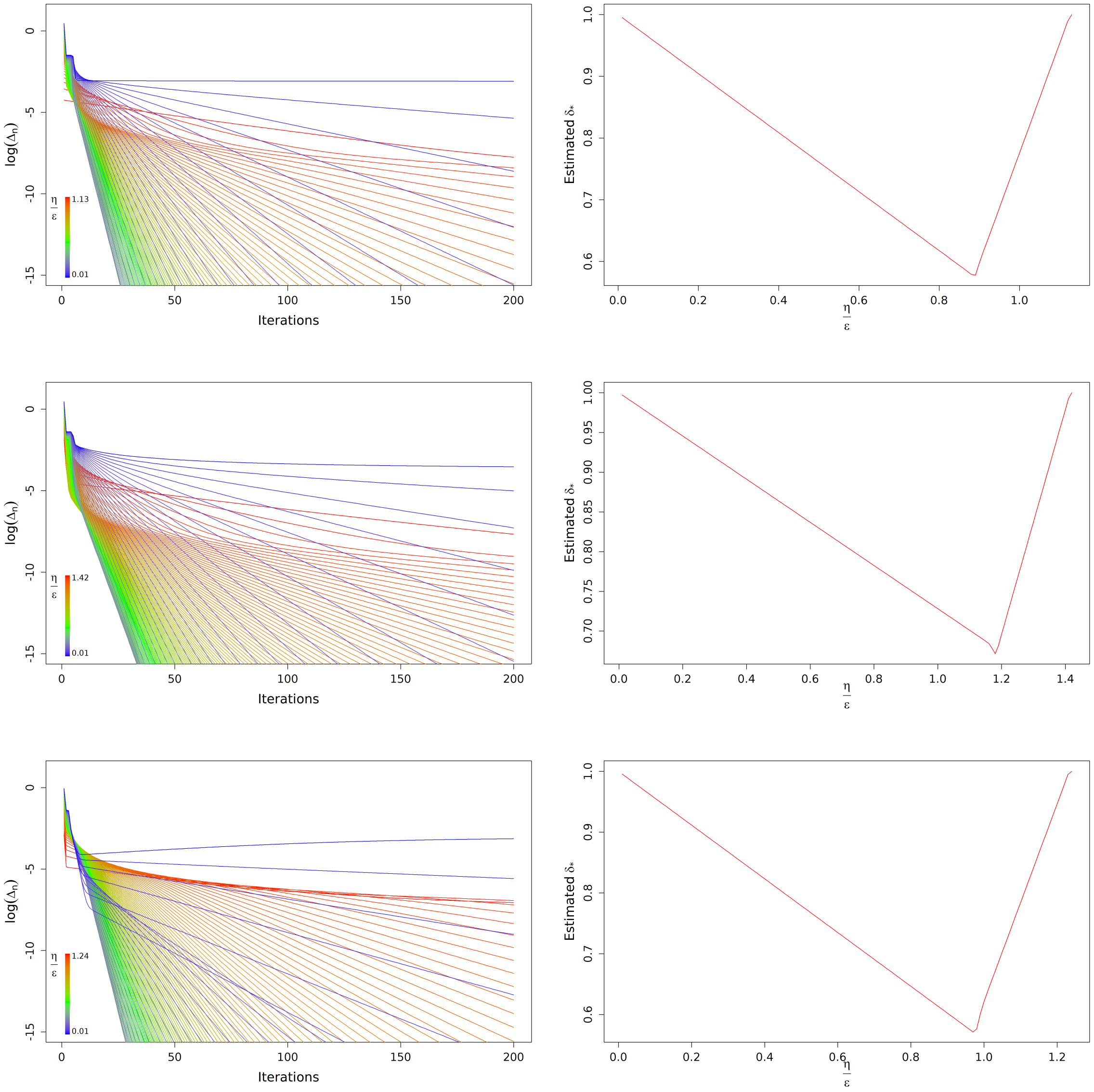

Figure 2: Gradient descent algorithm for three pairs of marginals: (top) $P = U[0,1]$ and $Q = U[0.5, 1.5]$, (middle) $P = U[0,1]$ and $Q = \beta(0.1, 0.2)$, (bottom) $P = U([0,1] \times [0,1])$ and $Q = U([1/\sqrt{2}, 1 + 1/\sqrt{2}] \times [1/\sqrt{2}, 1 + 1/\sqrt{2}])$. Left panels show the convergence for different step sizes η , right panels show the estimated δ_* for different η, both for fixed $\varepsilon = 10^{-1}$. Step size is increased until convergence breaks, which happens at $\eta/\varepsilon = 1.13$, 1.42, 1.24, respectively.

The right panels in Figure 2 show an estimate of δ_* for each step size. The estimate is obtained by a linear regression over $\log(\Delta_{N-k})$, $k = 0, \ldots, 9$, where N is the iteration where the convergence criterion is reached. We observe that δ_* depends substantially on η/ε, with a distinct V shape. Overall, the value of δ_* is quite small for a large range of step sizes (bounded away from zero and the break point), indicating fast convergence. As a caveat, we emphasize that the estimated δ_* pertains to the particular trajectory of the algorithm which depends on the chosen initialization, meaning that the constant in Theorem 2.3 could be worse.

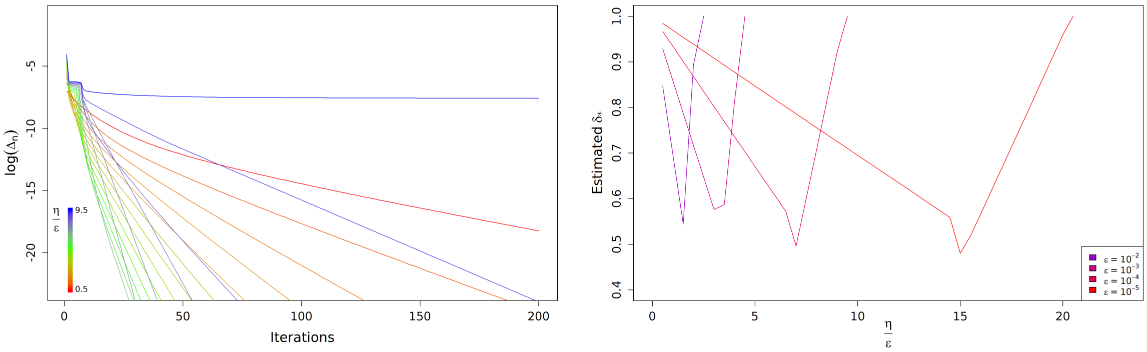

Figure 3: Repeating the first experiment (top row of Figure 2) with smaller values of ε and warm-start initialization. Left panel shows convergence for $\varepsilon = 10^{-4}$ and different step sizes η. Right panel shows estimated δ_* for several values of ε and varying step size η. Step size is increased until convergence breaks.

Figure 3 repeats the first experiment (i.e., the top row in Figure 2) but varies the regularization parameter ε. In addition, the experiment for $\varepsilon = 10^{-k}$ is initialized with the potentials found for $\varepsilon = 10^{-k+1}$. Even so, a longer burn-in period is observed for smaller values of ε (for constant initialization, the burn-in period would be even longer). We observe that the break point as well as the optimal ratio η/ε increase as ε decreases, extending beyond Assumption 2.2. Comparing with Remark 2.4, a possible explanation is that for small values of ε, the Lipschitz constant of the gradient $D\Gamma$ is substantially smaller than $2/\varepsilon$. Specifically, the positive part operator in (14) implies that the integral is approximately taken only over the support of the optimal coupling. This support is conjectured to be sparse, with sections of diameter $\sim \varepsilon^{\frac{1}{d+2}}$ (see [15, 28]), which suggests a Lipschitz constant that shrinks with ε.

References

[1] E. Bayraktar, S. Eckstein, and X. Zhang. Stability and sample complexity of divergence regularized optimal transport. *Bernoulli*, 31(1):213–239, 2025.

[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] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[4] G. Carlier. On the linear convergence of the multimarginal Sinkhorn algorithm. *SIAM J. Optim.*, 32(2):786–794, 2022.

[5] L. Chizat, A. Delalande, and T. Vaškevičius. Sharper exponential convergence rates for Sinkhorn's algorithm in continuous settings. *Math. Program.*, 2025.

[6] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[7] 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.

[8] S. Eckstein and M. Nutz. Convergence rates for regularized optimal transport via quantization. *Math. Oper. Res.*, 49(2):1223–1240, 2024.

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

[10] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra Appl.*, 114/115:717–735, 1989.

[11] A. Garriz-Molina, A. González-Sanz, and G. Mordant. Infinitesimal behavior of quadratically regularized optimal transport and its relation with the porous medium equation. *arXiv:2407.21528*, 2024.

[12] 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.

[13] A. González-Sanz, E. del Barrio, and M. Nutz. Sample complexity of quadratically regularized optimal transport. Forthcoming, 2025.

[14] A. González-Sanz, S. Eckstein, and M. Nutz. Sparse regularized optimal transport without curse of dimensionality, 2025.

[15] A. González-Sanz and M. Nutz. Sparsity of quadratically regularized optimal transport: Scalar case. *arXiv:2410.03353*, 2024.

[16] 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.

[17] N. Lahn, D. Mulchandani, and S. Raghvendra. A graph theoretic additive approximation of optimal transport. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13831–13841. Curran Associates, Inc., 2019.

[18] 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.

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

[20] D. A. Lorenz and H. Mahler. Orlicz-space regularization for optimal transport and algorithms for quadratic regularization. *arXiv:1909.06082*, 2019.

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

[22] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen. Tsallis regularized optimal transport and ecological inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017.

[23] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[24] M. Nutz. Quadratically regularized optimal transport: existence and multiplicity of potentials. *SIAM J. Math. Anal.*, 57(3):2622–2649, 2025.

[25] 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.

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

[27] 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.

[28] J. Wiesel and X. Xu. Sparsity of quadratically regularized optimal transport: Bounds on concentration and bias. *Preprint arXiv:2410.03425v1*, 2024.

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