CONTROL, OPTIMAL TRANSPORT AND NEURAL DIFFERENTIAL EQUATIONS IN SUPERVISED LEARNING

MINH-NHAT PHUNG AND MINH-BINH TRAN

Abstract. From the perspective of control theory, neural differential equations (neural ODEs) have become an important tool for supervised learning. In the fundamental work of Ruiz-Balet and Zuazua [17], the authors pose an open problem regarding the connection between control theory, optimal transport theory, and neural differential equations. More precisely, they inquire how one can quantify the closeness of the optimal flows in neural transport equations to the true dynamic optimal transport. In this work, we propose a construction of neural differential equations that converge to the true dynamic optimal transport in the limit, providing a significant step in solving the formerly mentioned open problem.

Contents

Date: March 25, 2025.
M.-N. Phung and M.-B. T are funded in part by a Humboldt Fellowship, NSF CAREER DMS-2303146, and NSF Grants DMS-2204795, DMS-2305523, DMS-2306379.

1. Introduction

Given a dataset $(x_j, y_j)_{j=1}^M$, in which $x_j \in \mathbb{R}^d$ is the input and $y_j \in \mathbb{R}^d$ is the corresponding output. Supervised Learning (SL) aims to identify that transport $x = (x_j)_{j=1}^M$ to $y = (y_j)_{j=1}^M$. This task is normally realized via the use of neural networks. A neural network is a system of multiple layers, in which each layer contains many nodes representing neurons. The first layer is input layer and the last is output layer, others are hidden layers. One layer will receive data from the previous layer, perform transformations and then push the data to the next. A compound neural network can be written as [17, Equation (1.1)] and [3, Equation (5)]

$$\sum_{i=1}^N W_i \sigma(A_i x + b_i) \tag{1}$$

where $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ the so-called activation function, $A_i, W_i \in \mathbb{R}^d \times \mathbb{R}^d$ and $b_i \in \mathbb{R}^d$.

Residual neural networks (ResNets) are generally formulated as follows [17]

$$x_{l+1} = \sum_{i=1}^N W_{i,l} \sigma(A_{i,l} x_l + b_{i,l}), \tag{2}$$

which can be recast as a Neural Differential Equations [11, 17]

$$\dot{x}(t) = \sum_{i=1}^N W_i(t) \sigma(A_i(t) x(t) + b_i(t)). \tag{3}$$

The above mentioned SL problem of transporting $x = (x_j)_{j=1}^M$ to $y = (y_j)_{j=1}^M$ now becomes a control problem:

Control Problem (A): *Given $T > 0$, x_0 x_T, can we construct $\big(W_i\big)_{i=1}^N$, $\big(A_i\big)_{i=1}^N$, $\big(b_i\big)_{i=1}^N$ such that we can drive the solution of the Neural ODE* (3) *from $x(0) = x_0$ to $x(T) = x_T$.*

Control Problem (A) has been first studied in the ground breaking work of Ruiz-Balet and Zuazua [17], as well as several later works of Zuazua and coauthors [1, 18, 24, 29, 41, 42, 43]. Writing the continuity equation of (3), we find

$$\partial_t \mu_t + \nabla_x \left(\sum_{i=1}^N W_i \sigma(A_i x + b_i) \mu_t \right) = 0, \tag{4}$$

in which $\mu_t(x) = \delta_{x=x(t)}$. Control Problem (A) can now be turned into the following equivalent control problem.

Control Problem (B): *Given $T > 0$, x_0, x_T, can we construct $\big(W_i\big)_{i=1}^N$, $\big(A_i\big)_{i=1}^N$, $\big(b_i\big)_{i=1}^N$ such that we can drive the solution of the transport equation* (4) *from $\mu_0(x) = \delta_{x=x_0}$ to $\mu_T(x) = \delta_{x=x_T}$.*

The connection between the neural transport equation (4) and Optimal Transport Theory has been first discussed in [17, 18]. While trying to solve Control Problem (A) and Control Problem (B), in [17, Section 7], the authors also posed the open problem of "how close the optimal flows of neural transport equations are from the actual optimal solution of the dynamic optimal transport". *In this work, we aim to explore this unexplored direction of research to make connection between Control Theory, Optimal Transport Theory and Neural ODEs by solving the problem raised by the authors of* [17].

From the theory of transport equations, letting $\zeta_t(x)$ be a fixed function from $\mathbb{R}_+ \times \mathbb{R}^d$ to $\mathbb{R}$, and supposing that a is the solution of

$$\zeta_t(x(t)) \ = \ \frac{\dot{a}(t)}{a(t)}, \tag{5}$$

then the quantity $\mu_t(x) = \delta_{x=x(t)}a(t)$ is the solution of the inhomogeneous transport equation [28]

$$\partial_t\mu_t + \nabla_x\left(\sum_{i=1}^{N} W_i\sigma(A_ix + b_i)\mu_t\right) = \zeta_t\mu_t. \tag{6}$$

While (4) is mass conserving, the transport equation (6) is certainly not mass conserving. It transports the measure $\mu_0(x) = \delta_{x=x_0}a(0)$ at time $t = 0$ to the measure $\mu_T(x) = \delta_{x=x_T}a(T)$ at time $t = T$. Due to the appearance of $a(0)$ and $a(T)$ and the fact that the mass is not conserved in the process, (6) is referred to as an unbalanced transport. We now turn Control Problem (B) into a new problem.

Control and Optimal Transport Problem (C): *Given $T > 0$, x_0, x_T, can we construct $\zeta_t(x)$, $\left(W_i\right)_{i=1}^N$, $\left(A_i\right)_{i=1}^N$, $\left(b_i\right)_{i=1}^N$ such that we can drive the solution of the transport equation* (6) *from $\mu_0(x) = \delta_{x=x_0}a(0)$ to $\mu_T(x) = \delta_{x=x_T}a(T)$, where a is the solution of* (5)*. Moreover, the optimal flows of the neural transport equation* (6) *also converge the actual optimal solution of the dynamic optimal transport in a certain limit.*

Note that if one can solve the Control and Optimal Transport Problem (C), both Control Problem (A) and Control Problem (B) are solved at the same time. However, by adding the new functions $\zeta_t(x)$, a, we have more degrees of freedom to deal with our transport problem and the problem of examining how closely the optimal flows of neural transport equations (4) align with the true optimal solution of dynamic optimal transport becomes much easier. On the contrary, (5)-(6) can be associated to an Unbalanced Optimal Transport process, in which the total mass of the density is controlled by an additional vector field ζ.

The formulation for optimal transportation had been known first under the name of Monge's problem. The problem has been revived in 1942 by Kantorovich [31], and it is now referred to as Monge-Kantorovich's problem. For many years, various authors have utilized the solutions of primal and dual optimal transportation problems to study many features of partial differential equations [7, 15, 20, 26, 35]. In the recent years, optimal transportation has found important applications in computer science and machine learning. In supervised learning, the Wasserstein distance is used to determine the loss in classification [21], ensuring the fairness in the system under investigation [23]. In generative Artificial Intelligence, optimal transportation has been used for Wasserstein Generative Adversarial Networks [2] and Wassenstein Auto-encoders [39]. In general, there are two types of Optimal Transport problems. Balanced Optimal Transport normally imposes the condition that the input measures have the same masses same with the output measures, while in the Unbalanced Optimal Transport problem, the two measures do not neccessarily have the same total masses. Therefore, the Unbalanced Optimal Transport framework eases the many constraints imposed by Balanced Optimal Transport and takes Balanced Optimal Transport as a special case. The framework is quite commonly used in Machine Learning and Artificial Intelligence since it provides more flexibility, robustness, and computational efficiency, making it better suited for real-world applications [10, 19, 37, 40].

Our strategy of solving Control and Optimal Transport Problem (C) is as follows.

- First, we associate (5)-(6) to an unbalanced optimal transportation problem, which is an optimization problem of minimizing a total transport cost (see Subsection 2.1). We show that minimization problem has a solution (see Subsection 2.2). However, since we want an explicit solution of the minimization problem to build our neural networks, we then design a numerical algorithm that gives the solution explicitly, using the Sinkhorn algorithm [12] and its variations (see Subsection 2.3). To be more precise, the construction of our algorithm has an inspiration from the work [34]. Since the work [34] has built for a discrete unbalanced optimal transport problem on a finite lattice, one main challenge is to generalize this algorithm to the unbalanced optimal transport problem (5)-(6), which is defined in the continuum setting. It is well know that generalizing a discrete problem posed on a finite lattice to the continuum problem is a very challenging problem. As a result, several new ideas have to be introduced in our work, to deal with the problem. For instance, instead of the Kullback-Leiber divergence used in [34], the Pearson divergence needs to be used. Since our theory is based on the Pearson divergence, numerous advanced techniques in optimal transport theory (for instance, [26, 33]) need to be incorporated into our framework.
- Second, based on the solution of the minimization problem, obtained from our numerical algorithm, we construct the transport equation (or gradient flows) and the convex solution of Monge-Ampère equation (see Subsection 2.4).
- Finally, the neural network can constructed using the above gradient flows, in combination with ideas from approximation theory, leading to the solution of the Control and Optimal Transport Problem (C) (see Subsection 2.5). As a result, we can construct $\big(W_i\big)_{i=1}^N$, $\big(A_i\big)_{i=1}^N$, $\big(b_i\big)_{i=1}^N$, such that the neural transport equation (6) converges the actual optimal solution of the dynamic optimal transport in certain limits, providing a significant step toward the solution of the open problem of " how close the optimal flows of neural transport equations are from the actual optimal solution of the dynamic optimal transport" posed in [17] (see Theorem 11 and Remark 12). Note that a solution to the Control and Optimal Transport Problem (C) is also a solution to both Control Problem (A) and Control Problem (B).

Acknowledgement The authors would like to thank Prof. Enrique Zuazua for enlightening discussions on the topic. They are also indebted to Prof. Nhan-Phu Chung and Prof Nam Q. Le for several suggestions and remarks to improve the quality of the paper.

2. Settings and main results

2.1. **The problem of minimizing the total transport cost.** As discussed in the introduction, the optimal transport problem can be viewed as the problem of minimizing the total transport cost. In this subsection, we will formulate the minimization problem.

Let d be an integer, we denote $\mathcal{M}(\mathbb{R}^d)$ as the set of all finite Radon measure of $\mathbb{R}^d$.

For a map $T:\mathbb{R}^d\to\mathbb{R}^d$ and $\mu\in\mathcal{M}(\mathbb{R}^d)$, the pushforward $T_\#\mu$ is defined by the identity

$$\int_{\mathbb{R}^d} f(x)dT_\#\mu(x) = \int_{\mathbb{R}^d} f(T(x))d\mu(x)$$

for all Borel measurable function f.

We denote $\mathcal{L}^d$ as the Lebesgue measure of $\mathbb{R}^d$. For convenience, we usually only write $\mathcal{L}$ and omit the dimension. We also consider the space $\mathcal{M}^\infty_{\mathcal{L},c}$ consisting of $\mu = f\mathcal{L} \in \mathcal{M}(\mathbb{R}^d)$ such that f has a compact support, $f \geq 0$ almost everywhere (a.e.) and $f \in L^\infty$.

For two measures $\mu, \nu \in \mathcal{M}(\mathbb{R}^d)$, if μ is absolutely continuous with respect to ν, we will denote this by $\mu \ll \nu$ and the Radon-Nykodim derivative is denoted by $\frac{d\mu}{d\nu}$. We define the marginal entropic cost as follows

$$F(\mu|\nu) := \begin{cases} \int_{\mathbb{R}^d} \left(\frac{d\mu}{d\nu} - 1\right)^2 d\nu & \text{in the case } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

In the optimal transport problem we always consider a cost function $C : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty]$ satisfying:

(C1) C is lower semi-continuous under the Euclidean topology;

(C2) For any compact sets $X, Y \subset \mathbb{R}^d$, $\|C\|_{L^\infty(X\times Y)} < +\infty$.

Given the 2 measures $\mu_0(x) = f\mathcal{L}(x), \mu_T(x) = g\mathcal{L}(x) \in \mathcal{M}^\infty_{\mathcal{L},c}(\mathbb{R}^d)$, the problem of optimally transporting μ_0 to μ_T is equivalent to the optimization problem of minimizing the total transport cost

$$d_C(f,g) := \inf_{\gamma\in\mathcal{M}(\Omega)} \left\{ \int_\Omega C(x,y)d\gamma(x,y) + \tau F(\gamma_x|f\mathcal{L}) + \tau F(\gamma_y|g\mathcal{L}) \right\}, \tag{7}$$

where $\Omega := \operatorname{supp} f \times \operatorname{supp} g$, $\pi_x : (x,y) \mapsto x, \pi_y : (x,y) \mapsto y, \gamma_x(dx) = (\pi_x)_\# \gamma(dx,dy), \gamma_y(dy) = (\pi_y)_\#\gamma(dx,dy)$ and τ is a given scaling factor. For convenience, we also use the notations $\Omega_f := \operatorname{supp} f$. In most cases, we will assume that $|\Omega_f|, |\Omega_g| > 0$, where $|\Omega_f|$ is the volume of the set Ω_f.

We denote the Euclidean norm on $\mathbb{R}^d$ by $\|\cdot\|$. Let us recall the 2-Wasserstein distance. If $F(\mu|\nu) = \begin{cases} 0, & \text{if } \mu = \nu \\ +\infty, & \text{if } \mu \neq \nu \end{cases}$ and $C(x,y) = \|x-y\|^2$, the infimum value in (7) is $W_2^2(\mu,\nu)$.

We back to our entropic cost, if γ is not absolutely continuous with respect to Lebesgue measure, we see that either $F(\gamma_x|f\mathcal{L})$ or $F(\gamma_y|g\mathcal{L})$ is infinity. To be more specific in our situation, γ must be absolutely continuous with respect to Lebesgue measure and the supports of γ_x, γ_y are Ω_f, Ω_g respectively. Hence, it suffices to consider $\gamma = k\mathcal{L}$ with $k \geq 0$ and $k \in L^1(\Omega)$. For simplification, we write

$$F(k_x|f) = F(\gamma_x|f\mathcal{L}), \quad F(k_y|g) = F(\gamma_y|g\mathcal{L}),$$

where $k_x = \int_{\Omega_g} kdy, k_y = \int_{\Omega_f} kdx$. We also define

$$L^2_+(\Omega) := \{k \in L^2(\Omega)|k \geq 0\}. \tag{8}$$

Lemma 1. *For $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^\infty_{\mathcal{L},c}(\mathbb{R}^d)$, we have*

$$d_C(f,g) = \inf_{k\in L^2_+(\Omega)} \left\{ \int_\Omega C(x,y)k(x,y)dxdy + \tau F(k_x|f) + \tau F(k_y|g) \right\}.$$

As a consequence of this lemma, we only need to solve the optimal problem on L^2 space. The proof of the lemma can be found in Section A.1.

For $\varepsilon > 0$, we define

$$f_\varepsilon(x) := \begin{cases} \max\{\varepsilon, f(x)\} & \text{if } x \in \Omega_f, \\ 0 & \text{otherwise} \end{cases}. \tag{9}$$

Now, we also have an approximated optimization problem

$$d_C^\varepsilon(f,g) := d_C(f_\varepsilon, g_\varepsilon) = \inf_{k(x,y)\in L^2_+(\Omega)} \left\{ \int_\Omega C(x,y)k(x,y)dxdy + \tau F(k_x|f_\varepsilon) + \tau F(k_y|g_\varepsilon) \right\}, \tag{10}$$

where $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^\infty_{\mathcal{L},c}(\mathbb{R}^d)$.

The following lemma shows that (10) tends to (7) as $\varepsilon \to 0^+$.

Lemma 2. *For $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^\infty_{\mathcal{L},c}(\mathbb{R}^d)$ and $\varepsilon > 0$, d_C^ε and d_C satisfy*

$$0 \le d_C^\varepsilon(f,g) - d_C(f,g) \le O(\varepsilon),$$

where the term $O(\varepsilon)$ means $O(\varepsilon) \to 0$ as $\varepsilon \to 0^+$.

Readers can see the proof for this lemma in Section A.2.

We add a small regulator $\eta\|k\|^2_{L^2}$ into the minimization problem. Instead of (10), our minimization problem becomes

$$d_C^{\varepsilon,\eta}(f,g) := \inf_{k(x,y)\in L^2_+(\Omega)} \left\{ \int_\Omega C(x,y)k(x,y)dxdy + \eta\|k(x,y)\|^2_{L^2} + \tau F(k_x|f_\varepsilon) + \tau F(k_y|g_\varepsilon) \right\}, \tag{11}$$

where $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^\infty_{\mathcal{L},c}(\mathbb{R}^d)$.

The following lemma shows that (11) converges (10) as $\eta \to 0^+$.

Lemma 3. *For $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^\infty_{\mathcal{L},c}(\mathbb{R}^d)$ and $\varepsilon, N > 0$, there exists $\eta_N > 0$ such that*

$$\left| d_C^{\varepsilon,\eta}(f,g) - d_C^\varepsilon(f,g) \right| < \frac{1}{N}, \quad \forall 0 < \eta < \eta_N.$$

In another word, $d_C^{\varepsilon,\eta}(f,g)$ is continuous at 0 with respect to η.

We will prove the lemma in Section A.3.

The next lemma says that the infinum in (11) is indeed the minimum.

Lemma 4. *Problem (11) admits a minimizer.*

We also show the proof of this lemma in Section A.4.

We next define

$$L^2_\delta(\Omega) = \{k \in L^2(\Omega) | k(x,y) \ge \delta, \text{a.e. } (x,y) \in \Omega\}. \tag{12}$$

We observe that if $k \in L^2_\delta(\Omega)$ then $k_x \in L^2_{\delta|\Omega_g|}(\Omega_f)$ and $k_y \in L^2_{\delta|\Omega_f|}(\Omega_g)$.

Next, we consider the optimization problem

$$d_{\delta,C}^{\varepsilon,\eta}(f,g) := \inf_{k(x,y)\in L^2_\delta(\Omega)} \left\{ \int_\Omega C(x,y)k(x,y)dxdy + \eta\|k(x,y)\|^2_{L^2} + \tau F(k_x|f_\varepsilon) + \tau F(k_y|g_\varepsilon) \right\}, \tag{13}$$

where $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^\infty_{\mathcal{L},c}(\mathbb{R}^d)$. The following lemma says that $d_{\delta,C}^{\varepsilon,\eta}(f,g)$ converges to $d_C^{\varepsilon,\eta}(f,g)$ as $\delta \to 0$.

Lemma 5. *For $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^{\infty}_{\mathcal{L},c}(\mathbb{R}^d)$ and $\varepsilon, \eta, \delta > 0$, $d^{\varepsilon,\eta}_{\delta,C}$ and $d^{\varepsilon,\eta}_{C}$ satisfy*

$$\left| d^{\varepsilon,\eta}_{\delta,C}(f,g) - d^{\varepsilon,\eta}_{C}(f,g) \right| \leq O(\delta),$$

where the term $O(\delta)$ means $O(\delta) \to 0$ as $\delta \to 0^+$.

Reader can find the proof of Lemma 5 in Section A.5.

As a summary, our main goal is to study problem (7), which can be approximated by problems (10), (11), (13). Therefore, we will mainly investigate the minimization problem (13).

2.2. **Dual problem.** In solving a minimization problem, a common strategy is to investigate the dual problem. For $v\mathcal{L} \in \mathcal{M}^{\infty}_{\mathcal{L},c}(\mathbb{R}^d)$ and $u^* \in L^2(\operatorname{supp} v)$, we define

$$F^*_{\delta}(u^*|v) := \sup_{u \in L^2_{\delta}(\operatorname{supp} v)} \left\{ \int_{\operatorname{supp} v} uu^* dx - F(u|v) \right\}.$$

For $w^* \in L^2(\Omega)$ we define

$$\bar{C}^*_{\eta,\delta}(w^*) := \sup_{k \in L^2_{\delta}(\Omega)} \left\{ \int_{\Omega} kw^* dxdy - \int_{\Omega} Ckdxdy - \eta \|k\|^2_{L^2} \right\}.$$

We then have the dual problem

$$D^{\varepsilon,\eta}_{\delta,C}(f,g) := \inf_{\substack{k_1^* \in L^2(\Omega_f) \\ k_2^* \in L^2(\Omega_g)}} \left\{ \bar{C}^*_{\eta,\delta}(k_1^* + k_2^*) + \tau F^*_{\delta|\Omega_g|}(-k_1^*/\tau | f_\varepsilon) + \tau F^*_{\delta|\Omega_f|}(-k_2^*/\tau | g_\varepsilon) \right\}. \tag{14}$$

Lemma 6. *Problem (13) and problem (14) admit minimizers. Furthermore, k is a minimizer of (13) and (k_1^*, k_2^*) is a minimizer of (14) if and only if*

$$k(x,y) = \max\left\{ \delta, \frac{k_1^*(x) + k_2^*(y) - C(x,y)}{2\eta} \right\}, \tag{15}$$

$$k_x(x) = f_\varepsilon(x) \max\left\{ \frac{\delta|\Omega_g|}{f_\varepsilon(x)}, 1 - \frac{k_1^*(x)}{2\tau} \right\}, \tag{16}$$

$$k_y(y) = g_\varepsilon(y) \max\left\{ \frac{\delta|\Omega_f|}{g_\varepsilon(y)}, 1 - \frac{k_2^*(y)}{2\tau} \right\}, \tag{17}$$

for almost everywhere $x \in \Omega_f, y \in \Omega_g$ with respect to Lebesgue measure.

In addition, we have

$$d^{\varepsilon,\eta}_{\delta,C}(f,g) + D^{\varepsilon,\eta}_{\delta,C}(f,g) = 0. \tag{18}$$

Lemma 6 is the foundation of the construction of the optimization algorithm and the neural network. We divided the proof of this lemma into four smaller sections, which are included in Section A.6.

Remark 7. The minimizer (k_1^*, k_2^*) of (14) can be taken so that

$$k_1^* \in \mathcal{L}^f_g := \left\{ k^* \in L^2(\Omega_f) | k^* \leq 2\tau \left(1 - \frac{\delta|\Omega_g|}{f_\varepsilon}\right) a.e. \right\} \quad \text{and} \quad k_2^* \in \mathcal{L}^g_f.$$

2.3. **Optimization algorithm.** The minimizer (k_1^*, k_2^*) of (14) is essential in our construction of our neural networks. Below, we will show that the minimizer (k_1^*, k_2^*) of (14) exists and can be computed explicitly via a numerical algorithm. Therefore, (k_1^*, k_2^*) is explicit and can be used to formulate our neural network.

Our algorithm has an inspiration by the one developed in [34], which is based on a previous work [27].

Given functions f, g and four positive constants $\varepsilon, \delta, \eta, \tau$ such that $\delta \max\{|\Omega_f|, |\Omega_g|\} \le \varepsilon$, we will design an algorithm to approximate the solution (k_1^*, k_2^*) of (14). Let $E := \max\{\|f_\varepsilon\|_{L^\infty}, \|g_\varepsilon\|_{L^\infty}\}$, we compute

$$\alpha := \sqrt{\frac{1}{2}\left(\max\left\{\frac{|\Omega_f|}{\eta}, \frac{|\Omega_g|}{\eta}\right\}^2 + \left(\frac{E-\varepsilon/2}{\tau}\right)^2\right)}. \tag{19}$$

We then compute $q := 2\sqrt{\alpha\tau/\varepsilon}$, $s := \sqrt{\alpha\varepsilon/(\tau)}/2$ and $r := \frac{q}{1+q}$. We also performs computations on $f_\varepsilon, g_\varepsilon$, which will be used to design the transformation between steps.

For $n \in \mathbb{N}$, let $X^n, X_0^n, X_*^n, Y^n, Y_0^n, Y_*^n$ be the six L^2 data functions after n steps of the algorithm. We start the algorithm with $X^0 = X_0^0 = X_*^0 = Y^0 = Y_1^0 = Y_*^0 = 0$. For convenience, we always suppose that the data marked by X are functions supported by Ω_f and the data marked by Y are functions supported by Ω_g.

We have the following algorithm:

(1) X^{n+1} can be updated by X_*^n, Y_*^n

$$X^{n+1}(x) = \frac{1}{2\eta}\int_{\Omega_g} \max\{2\eta\delta, X_*^n(x) + Y_*^n(y) - C(x,y)\}\,dy + \left(\frac{X_*^n(x)}{2\tau} - 1\right) f_\varepsilon(x).$$

(2) X_0^{n+1} can be updated by X_0^n, X^n, X^{n+1}

$$X_0^{n+1}(x) = 4\tau \min\left\{\frac{sX_0^n(x)}{\varepsilon + 4\tau s} - \frac{(1+r)X^{n+1}(x) - rX^n(x)}{\varepsilon + 4\tau s}, \frac{1}{2} - \frac{|\Omega_g|}{2f_\varepsilon(x)}\right\}.$$

(3) X_*^{n+1} can be updated by X_*^n, X_0^{n+1}

$$X_*^{n+1}(x) = \frac{qX_*^n(x)}{1+q} + \frac{X_0^{n+1}(x)}{1+q}.$$

Nodes Y can be built the same way.

(4) Y^{n+1} can be updated by X_*^n, Y_*^n

$$Y^{n+1}(x) = \frac{1}{2\eta}\int_{\Omega_f} \max\{2\eta\delta, X_*^n(x) + Y_*^n(y) - C(x,y)\}\,dx + \left(\frac{Y_*^n(y)}{2\tau} - 1\right) g_\varepsilon(x).$$

(5) Y_0^{n+1} can be updated by Y_0^n, Y^n, Y^{n+1}

$$Y_0^{n+1}(x) = 4\tau \min\left\{\frac{sY_0^n(y)}{\varepsilon + 4\tau s} - \frac{(1+r)Y^{n+1}(y) - rY^n(y)}{\varepsilon + 4\tau s}, \frac{1}{2} - \frac{|\Omega_f|}{2g_\varepsilon(y)}\right\}.$$

(6) Y_*^{n+1} can be updated by Y_*^n, Y_0^{n+1}

$$Y_*^{n+1}(x) = \frac{qY_*^n(x)}{1+q} + \frac{Y_0^{n+1}(x)}{1+q}.$$

After $L+1$ steps, we obtain $\bar{k_1^*} := X_0^{L+1}$ and $\bar{k_2^*} := Y_0^{L+1}$. These functions indeed approximate the minimizer of (14).

Proposition 8. *Given the functions f, g so that $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^\infty_{\mathcal{L},c}$, C is the cost satisfies (C1) and (C2), and the constants $\varepsilon, \eta, \delta, \tau > 0$ such that $\delta \max\{|\Omega_f|, |\Omega_g|\} \le \varepsilon$, the above algorithm gives L^2 functions $(\bar{k_1^*}, \bar{k_2^*})$ such that*

$$\|k_j^* - \bar{k_j^*}\|_{L^2} \lesssim r^{L/2} \ = \ \left(\frac{q}{q+1}\right)^{L/2}, \quad j = 1, 2,$$

where (k_1^*, k_2^*) is a minimizer of (14).

The Proposition 8 shows the convergence of the algorithm. Section 3.3 contains the proof for this proposition.

2.4. **Transport equation.** We want to apply the output from the optimization algorithm to build a continuity equation transporting f to g, i.e., for $T > 0$, we wish to obtain the continuity equation

$$\begin{cases} \partial_t \mu_t + \nabla_x(\xi_t \mu_t) = \zeta_t \mu_t, \\ \mu_0 = f\mathcal{L}, \mu_T = g\mathcal{L}. \end{cases}$$

Our strategy is as follows. First, we construct $\bar{f}, \bar{g}$ that approximate f, g in a certain limit (see the statement of Proposition 9). Next, we construct ξ, ζ such that

$$\begin{cases} \partial_t \mu_t + \nabla_x(\xi_t \mu_t) = \zeta_t \mu_t, \\ \mu_0 = \bar{f}\mathcal{L}, \mu_T = \bar{g}\mathcal{L}. \end{cases}$$

To this end, we need to recall some results concerning the associated Monge-Ampère equation. We first construct $\widetilde{k_1^*}, \widetilde{k_2^*}$ supported by Ω_f, Ω_g such that $\|\widetilde{k_i^*} - \bar{k_1^*}\|_{L^\infty} < \mathrm{err}_0$ for some given $\mathrm{err}_0 > 0$. Then, we define $\bar{k} = \max\left\{\delta, \frac{\widetilde{k_1^*} + \widetilde{k_2^*} - C}{2\eta}\right\}$ on Ω. We need to find a map $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ such that $\varphi_\# \bar{k}_x = \bar{k}_y$, that is

$$\int_{\mathbb{R}^d} h(\varphi(x)) \bar{k}_x dx = \int_{\mathbb{R}^d} h(y) \bar{k}_y(y) dy \tag{20}$$

for all measurable h. When we consider $\varphi = \nabla\phi$ for some convex function ϕ and Ω_{g_ε} is convex, we obtain the Monge-Ampère equation

$$\bar{k}_y(\nabla\phi)\det(D^2\phi) = \bar{k}_x. \tag{21}$$

The existence of a convex function ϕ such that $\varphi = \nabla\phi$ solving (20) is proved in the work of McCann (see [35]). The solution $\nabla\phi$ is uniquely determined almost everywhere. We note that $\nabla\phi$ solving (20) does not mean ϕ solves (21). The convexity of Ω_g ensures that ϕ is a solution of (21) in the Aleksandrov sense (see [13, Chapter 4]). Since the Monge-Ampère equation is given in its explicit form, we assume that the solution is given explicitly. In practice, the solution of the Monge-Ampére equation can be computed using well studied numerical schemes (see, for instance [4, 5, 6, 8]).

Proposition 9. *We assume that $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^{\infty}_{\mathcal{L},c}$ such that f, g are continuous, and we also assume C is a Lipschitz function on Ω. For some given small number* $\mathrm{err}_0 > 0$, *we can construct $\widetilde{k_1^*}, \widetilde{k_2^*}$ supported by Ω_f, Ω_g satisfying*

$$\|\bar{k_i^*} - \widetilde{k_i^*}\|_{L^\infty} < \mathrm{err}_0$$

and $\bar{k_1^*}, \bar{k_2^*}$ are obtained in Proposition 8 with a large number L. Let Ω_f, Ω_g convex and their boundaries are smooth, and let ϕ be a convex function solving (21)*, which depends on L, err_0. Note that $f_\varepsilon, g_\varepsilon$ are defined in* (9)*. We can construct $\overline{f} \in L^2(\Omega_f), \overline{g} \in L^2(\Omega_g)$ satisfying*

$$\|f_\varepsilon - \overline{f}\|_{L^2}, \|g_\varepsilon - \overline{g}\|_{L^2} \lesssim r^{L/2} + \mathrm{err}_0.$$

Moreover, transport the equation can be built as follows

$$\begin{cases} \partial_t \mu_t + \nabla_x(\xi_t \mu_t) = \zeta_t \mu_t, \\ \mu_0 = \overline{f}\mathcal{L}, \mu_T = \overline{g}\mathcal{L}, \end{cases} \tag{22}$$

where the vector fields are given by

$$\xi_t(x) = \frac{1}{T}\left(\nabla\phi(\mathbb{T}_t^{-1}(x)) - \mathbb{T}_t^{-1}(x)\right), \tag{23}$$

$$\zeta_t(x) = \xi_t \cdot \left(-\frac{\nabla \widetilde{k_1^*}(-T\xi_t(x) - x)}{2\tau - \widetilde{k_1^*}(-T\xi_t(x) - x)} + \frac{\nabla \widetilde{k_2^*}(x)}{2\tau - \widetilde{k_2^*}(x)} \right), \tag{24}$$

$$\mathbb{T}_t(x) = \frac{T-t}{T}x + \frac{t}{T}\nabla\phi(x). \tag{25}$$

The proof of the proposition is given in Section 3.4.

Remark 10. Suppose we want to transport $f\mathcal{L}$ to $g\mathcal{L}$, the proposition gives a construction of an inhomogeneous transport equation (22) that transports $\overline{f}\mathcal{L}$ to $\overline{g}\mathcal{L}$, where $\overline{f}\mathcal{L}$ and $\overline{g}\mathcal{L}$ converge to $f\mathcal{L}$ and $g\mathcal{L}$ as $L \to \infty$ and $\mathrm{err}_0, \varepsilon \to 0$. The inhomogeneous transport equation is built upon the values $\bar{k_1^*}, \bar{k_2^*}$ are obtained in Proposition 8. Proposition 9 establishes an approximation in L^2 sense for the transport equation, we will use this transport equation to make a neural transport equation.

2.5. **Solution of the Control and Optimal Transport Problem (C).** We will build a compound neural network with controlled using the the push forward map $\nabla\phi$ and the minimizer (k_1^*, k_2^*) of (14).

As discussed above, since the Monge-Ampère equation has an explicit form, we make the assumption that that the solution is given explicitly. In practice, the solution of the Monge-Ampère equation can be numerically obtained using well established numerical schemes (see, as an example, those in [4, 5, 6, 8]).

Theorem 11. *We assume that $f\mathcal{L}, g\mathcal{L} \in \mathcal{M}^{\infty}_{\mathcal{L},c}$ such that f, g are continuous and we also assume C is a Lipschitz function on Ω. Let Ω_f, Ω_g be convex and their boundaries are smooth, we recall the definitions of $f_\varepsilon, g_\varepsilon$ in* (9) *and definitions of $\mathbb{T}, \xi, \zeta$ defined in* (23) - (25) *in Proposition 9, $L+1$ is the number of steps in our algorithm in Section 2.3,* $\mathrm{err}_0 > 0$ *be a given number. Let ϕ be a convex function and $\nabla\phi$ be the solution of* (21) *depending on L, err_0.*

Let $\mathrm{Lip}([0,T],\mathbb{R}^d)$ *be the space of Lipschitz functions from* $[0,T]$ *to* $\mathbb{R}^d$*, and let* $\sigma = \begin{pmatrix} e^{-\|x\|^2} \\ e^{-\|x\|^2} \\ \vdots \\ e^{-\|x\|^2} \end{pmatrix}$ *be the Gaussian activation function, we can build* $\widetilde{f},\widetilde{g} \in L^2(\mathbb{R}^d)$ *and construct a finite set* $M^{\mathrm{err}_1}_{L,\mathrm{err}_0} \subset \mathbb{N}^d$ *with triplets* $(W_{\vec{m}}, A_{\vec{m}}, b_{\vec{m}}) \in \mathrm{Lip}([0,T],\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^{d\times d} \times \mathbb{R}^d, \vec{m} \in M^{\mathrm{err}_1}_{L,\mathrm{err}_0}$ *such that the controlled compound neural equation*

$$\begin{cases} \partial_t\mu_t + \nabla_x\left(\sum_{\vec{m}\in M^{\mathrm{err}_1}_{L,\mathrm{err}_0}} W_{\vec{m}}\sigma(A_{\vec{m}}x + b_{\vec{m}})\mu_t\right) = \zeta_t\mu_t, \\ \mu_0 = \widetilde{f}\mathcal{L}, \mu_T = \widetilde{g}\mathcal{L}, \end{cases} \tag{26}$$

satisfies

$$\left\| \sum_{\vec{m}\in M^{\mathrm{err}_1}_{L,\mathrm{err}_0}} W_{\vec{m}}\sigma(A_{\vec{m}}x + b_{\vec{m}}) - \xi_t(x) \right\|_{L^2} \le \mathrm{err}_1;$$

and the measures $\widetilde{f}\mathcal{L}, \widetilde{g}\mathcal{L}$ *converge weakly to* $f_\varepsilon\mathcal{L}, g_\varepsilon\mathcal{L}$ *as* $\mathrm{err}_0, \mathrm{err}_1 \to 0^+$ *and* $L \to +\infty$*.*

The proof of the theorem is given in Section 3.5.

Remark 12. Suppose we want to transport $f\mathcal{L}$ to $g\mathcal{L}$, the theorem gives a construction of a neural differential equation (26) that transports $\widetilde{f}\mathcal{L}$ to $\widetilde{g}\mathcal{L}$, where $\widetilde{f}\mathcal{L}$ and $\widetilde{g}\mathcal{L}$ converge to $f\mathcal{L}$ and $g\mathcal{L}$ as $\mathrm{err}_0, \mathrm{err}_1 \to 0^+$, $L \to \infty$ and $\varepsilon \to 0$. That means flows of the neural transport equation (6) also converge the actual optimal solution of the dynamic optimal transport in the limit of $\mathrm{err}_0, \mathrm{err}_1 \to 0^+$ and $L \to \infty$ and $\varepsilon \to 0$.

In this first work, we choose the Gaussian activation function [30, 36] since it shortens some parts of the proof. In the upcoming second paper, we will extend the results to other activation functions such as ReLU, sigmoids, etc.

Remark 13. In Theorem 11, instead of using $\mathbb{T}, \xi, \zeta$ defined by $\nabla\phi$ solving (21), we can consider a sequence $\nabla\bar{\phi}_n \xrightarrow{L^2} \nabla\phi$, where $\bar{\phi}_n$ is strongly convex, supported on Ω_f and $\bar{\phi}_n \in C^2$. We suppose that the sequence $\nabla\bar{\phi}_n$ is obtained by numerically solving (21) using well studied numerical schemes (see, for instance [4, 5, 6, 8]). We then define new vector fields

$$\begin{aligned} \bar{\mathbb{T}}_t(x) &= \frac{T-t}{T}x + \frac{t}{T}\nabla\bar{\phi}_n(x), \\ \bar{\xi}_t(x) &= \frac{1}{T}\left(\nabla\bar{\phi}_n(\bar{\mathbb{T}}_t^{-1}(x)) - \bar{\mathbb{T}}_t^{-1}(x)\right), \\ \bar{\zeta}_t(x) &= \bar{\xi}_t(x)\cdot\left(-\frac{\nabla\widetilde{k}_1^*(-T\bar{\xi}_t(x)-x)}{2\tau - \widetilde{k}_1^*(-T\bar{\xi}_t(x)-x)} + \frac{\nabla\widetilde{k}_2^*(x)}{2\tau - \widetilde{k}_2^*(x)}\right). \end{aligned}$$

Then we can construct $\widetilde{f},\widetilde{g} \in L^2(\mathbb{R}^d)$, a finite subset $M^{\mathrm{err}_1,\mathrm{n}}_{L,\mathrm{err}_0}$ of $\mathbb{N}^d$ and triplets $(W_{\vec{m}}, A_{\vec{m}}, b_{\vec{m}}) \in \mathrm{Lip}([0,T],\mathbb{R}^d\times\mathbb{R}^d)\times\mathbb{R}^{d\times d}\times\mathbb{R}^d, \vec{m} \in M^{\mathrm{err}_1,\mathrm{n}}_{L,\mathrm{err}_0}$ such that the controlled compound neural equation

$$\begin{cases} \partial_t\mu_t + \nabla_x\left(\sum_{\vec{m}\in M^{\mathrm{err}_1,\mathrm{n}}_{L,\mathrm{err}_0}} W_{\vec{m}}\sigma(A_{\vec{m}}x + b_{\vec{m}})\mu_t\right) = \bar{\zeta}_t\mu_t, \\ \mu_0 = \widetilde{f}\mathcal{L}, \mu_T = \widetilde{g}\mathcal{L}, \end{cases}$$

satisfies

$$\left\| \sum_{\vec{m} \in M_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n}}} W_{\vec{m}} \sigma(A_{\vec{m}} x + b_{\vec{m}}) - \bar{\xi}_t(x) \right\|_{L^2} \le \mathrm{err};$$

and the measures $\widetilde{f}\mathcal{L}, \widetilde{g}\mathcal{L}$ converge weakly to $f_\varepsilon \mathcal{L}, g_\varepsilon \mathcal{L}$ as $\mathrm{err}_0, \mathrm{err}_1 \to 0^+$ and $n, L \to +\infty$. We will give the proof for this scenario, that we have the sequence $\nabla \bar{\phi}_n \xrightarrow{L^2} \nabla \phi$, since this proof also covers the proof of Theorem 11.

3. Proofs of the main results

3.1. **Convex analysis.** Let us define the function $G : L^2(\Omega_f) \times L^2(\Omega_g) \to [0, \infty]$ by

$$\begin{aligned} G(u,v) := & \frac{1}{4\eta} \int_\Omega \max\{2\eta\delta, R_{u,v}^C(x,y)\}(2R_{u,v}^C(x,y) - \max\{2\eta\delta, R_{u,v}^C(x,y)\})dxdy \\ & + \tau \int_{\Omega_f} \left(1 - \frac{u(x)}{2\tau}\right)^2 f_\varepsilon(x)dx + \tau \int_{\Omega_g} \left(1 - \frac{v(y)}{\tau}\right)^2 g_\varepsilon(y)dy, \end{aligned}$$

where $R_{u,v}^C(x,y) := u(x) + v(y) - C(x,y)$.

Our strategy is to replace the minimization problem (14) *by minimizing G. In this section, we will study some important features of the new function G.*

We compute the Fréchet derivatives of G as follows

$$\begin{aligned} D_1 G(u,v)(x) &= \frac{1}{2\eta} \int_{\Omega_g} \max\{2\eta\delta, R_{u,v}^C(x,y)\}dy + \left(\frac{u(x)}{2\tau} - 1\right) f_\varepsilon(x), \\ D_2 G(u,v)(y) &= \frac{1}{2\eta} \int_{\Omega_f} \max\{2\eta\delta, R_{u,v}^C(x,y)\}dx + \left(\frac{v(y)}{2\tau} - 1\right) g_\varepsilon(y), \end{aligned}$$

where D_1, D_2 are differentiation with respect to the variables u and v.

We will now check that $D_1 G$ is indeed Fréchet derivative of G with respect to variable u. Then, by a similar proof, $D_2 G$ is the derivative of G with respect to v. For $u, h \in L^2(\Omega_f), g \in L^2(\Omega_g)$, we compute

$$\begin{aligned} \Delta_u G(u,v,h) = & \frac{\left| G(u+h,v) - G(u,v) - \int_{\Omega_f} D_1 G(u,v) h dx \right|}{\|h\|_{L^2}} \\ = & \frac{1}{4\eta \|h\|_{L^2}} \left| \int_\Omega \max\{2\eta\delta, R_{u,v}^C + h\}(2R_{u,v}^C + 2h - \max\{2\eta\delta, R_{u,v}^C + h\})dxdy \right. \\ & \left. - \int_\Omega \max\{2\eta\delta, R_{u,v}^C\}(2R_{u,v}^C + 2h - \max\{2\eta\delta, R_{u,v}^C\})dxdy + \frac{\eta}{\tau} \int_{\Omega_f} h^2 f_\varepsilon dx \right| \\ = & \frac{1}{4\eta \|h\|_{L^2}} \left| \int_\Omega \big(\max\{2\eta\delta, R_{u,v}^C + h\} - \max\{2\eta\delta, R_{u,v}^C\}\big) \right. \\ & \left. \times (2R_{u,v}^C + 2h - \max\{2\eta\delta, R_{u,v}^C + h\} - \max\{2\eta\delta, R_{u,v}^C\})dxdy + \frac{\eta}{\tau} \int_{\Omega_f} h^2 f_\varepsilon dx \right|. \end{aligned}$$

If $R_{u,v}^C + h, R_{u,v}^C < 2\eta\delta$ then

$$\left(\max\{2\eta\delta, R_{u,v}^C + h\} - \max\{2\eta\delta, R_{u,v}^C\}\right)\left(2R_{u,v}^C + 2h - \max\{2\eta\delta, R_{u,v}^C + h\} - \max\{2\eta\delta, R_{u,v}^C\}\right) = 0.$$

Otherwise, we have $|\max\{2\eta\delta, R_{u,v}^C + h\} - \max\{2\eta\delta, R_{u,v}^C\}| < |h|$ and $|2R_{u,v}^C + 2h - \max\{2\eta\delta, R_{u,v}^C + h\} - \max\{2\eta\delta, R_{u,v}^C\}| < |h|$. Hence,

$$\Delta_u G(u, v, h) \lesssim \|h\|_{L^2}.$$

The limit of $\Delta_u G(u, v, h)$ when $h \to 0$ is 0, which means $D_1 G$ is the Fréchet derivative of G with respect to the variable u.

For $u \in L^2(\Omega_f), v \in L^2(\Omega_g)$, we define

$$w(u, v) := \frac{1}{2}\left(\|u\|_{L^2}^2 + \|v\|_{L^2}^2\right). \tag{27}$$

Now, we will prove that G is strongly convex, i.e., there is $m > 0$ such that $G(u, v) - mw(u, v)$ is a convex function. For $u, h \in L^2(\Omega_f), v \in L^2(\Omega_g)$, we first compute and estimate

$$\begin{aligned}\int_{\Omega_f} (D_1 G(u + h, v) - D_1 G(u, v)) h dx &= \frac{1}{2\eta}\int_{\Omega} (\max\{2\eta\delta, R_{u,v}^C + h\} - \max\{2\eta\delta, R_{u,v}^C\}) h dx dy \\ &\quad + \frac{1}{2\tau}\int_{\Omega_f} h^2 f_\varepsilon dx \\ &\geq \frac{\varepsilon}{2\tau}\int_{\Omega_f} h^2 dx.\end{aligned}$$

Thus,

$$\int_{\Omega_f}\left(D_1\left(G(u+h) - \frac{\varepsilon}{4\tau}\|u+h\|_{L^2}^2\right) - D_1\left(G(u, v) - \frac{\varepsilon}{4\tau}\|u\|_{L^2}^2\right)\right) h dx \geq 0.$$

By [25, (2.6) Chapter 2 Part 1], $G(u, v) - \frac{\varepsilon}{4\tau}\|u\|_{L^2}^2$ is convex with respect to the variable u. A similar computation for the variable v can also be done. Hence, G is $\frac{\varepsilon}{2\tau}$-strongly convex. Next, we define

$$G_w(u, v) := G(u, v) - \frac{\varepsilon}{4\tau} w(u, v),$$

which is $\frac{\varepsilon}{4\tau}$-strongly convex. The derivatives $D_1 G_w, D_2 G_w$ of G_w are Lipschitz continuous in L^2 with respect to the variables u, v, respectively. Indeed, for u, v, h as before and α defined in (19), we have the estimate

$$\begin{aligned}&\|D_1 G_w(u + h, v) - D_1 G_w(u, v)\|_{L^2}^2 \\ &\quad = \left\| D_1 G(u + h, v) - D_1 G(u, v) - \frac{\varepsilon}{4\tau} h\right\|_{L^2}^2 \\ &\quad = \left\| \frac{1}{2\eta}\int_{\Omega_g} (\max\{2\eta\delta, R_{u,v}^C + h\} - \max\{2\eta\delta, R_{u,v}^C\}) dy - \left(\frac{f_\varepsilon}{2\tau} - \frac{\varepsilon}{4\tau}\right) h\right\|_{L^2}^2 \\ &\quad \leq \frac{|\Omega_g|^2}{2\eta^2}\int_{\Omega} |h|^2 dx + \frac{1}{2\tau^2}\left\| f_\varepsilon - \frac{\varepsilon}{2}\right\|_{L^\infty} \|h\|_{L^2}^2 \\ &\quad \leq \frac{1}{2}\left(\frac{|\Omega_g|^2}{\eta^2} + \frac{(E - \varepsilon/2)^2}{\tau^2}\right)\|h\|_{L^2}^2 = \alpha^2 \|h\|_{L^2}^2.\end{aligned}$$

The same estimate can also be done for D_2G_w.

We consider the convex conjugate of G_w

$$G_w^*(u^*,v^*) = \sup_{(u,v)\in L^2(\Omega_f)\times L^2(\Omega_g)} \left\{ \int_{\Omega_f} u^* u dx + \int_{\Omega_g} v^* v dy - G_w(u,v) \right\}, (u^*,v^*) \in L^2(\Omega_f)\times L^2(\Omega_g). \tag{28}$$

Since G_w is Fréchet differentiable, it is also Gâteaux differentiable. By [25, Proposition 5.3 Chapter 1 Part 1], for all $(u,v) \in L^2(\Omega_f) \times L^2(\Omega_g)$, we deduce

$$G_w^*(D_1G_w(u,v), D_2G_w(u,v)) + G_w(u,v) = \int_{\Omega_f} D_1G_w(u,v) u dx + \int_{\Omega_g} D_1G_w(u,v) v dy. \tag{29}$$

By the strong convexity of G_w, if $(\bar{u},\bar{v}) \in L^2(\Omega_f) \times L^2(\Omega_g)$ satisfies

$$G_w^*(D_1G_w(u,v), D_2G_w(u,v)) + G_w(\bar{u},\bar{v}) = \int_{\Omega_f} D_1G_w(u,v)\bar{u} dx + \int_{\Omega_g} D_1G_w(u,v)\bar{v} dy,$$

then $(u,v) = (\bar{u},\bar{v})$ in $L^2(\Omega_f) \times L^2(\Omega_g)$. By this uniqueness property, we define

$$L^* := \left\{ (u^*,v^*) \in L^2(\Omega_f) \times L^2(\Omega_g) | \exists (u,v) \in L^2(\Omega_f) \times L^2(\Omega_g) : (u^*,v^*) = (D_1G_w(u,v), D_2G_w(u,v)) \right\},$$

and $(D_1G^*, D_2G^*) : L^* \to L^2(\Omega_f) \times L^2(\Omega_g)$ as the inverse of (D_1G, D_2G), that is

$$(D_1G^*, D_2G^*)(D_1G_w(u,v), D_2G_w(u,v)) = (u,v).$$

For $(u^*,v^*), (u_0^*,v_0^*) \in L^*$, let

$$\begin{aligned} \Delta G_w^*(u_0^*,v_0^*,u^*,v^*) &:= G_w^*(u^*,v^*) - G_w^*(u_0^*,v_0^*) \\ &\quad - \int_{\Omega_f} (u^*-u_0^*) D_1G_w^*(u_0^*,v_0^*) dx - \int_{\Omega_g} (v^*-v_0^*) D_2G_w^*(u_0^*,v_0^*) dy. \end{aligned}$$

We want to show that

$$\Delta G_w^*(u_0^*,v_0^*,u^*,v^*) \geq \frac{1}{2\alpha} \left(\|u^*-u_0^*\|_{L^2}^2 + \|v^*-v_0^*\|_{L^2}^2 \right), \tag{30}$$

in which α is defined by (19).

For $u,h \in L^2(\Omega_f), v \in L^2(\Omega_g)$, we have

$$\begin{aligned} & \left| G_w(u+h,v) - G_w(u,v) - \int_{\Omega_f} D_1G_w(u,v) h dx \right| \\ = & \left| \int_0^1 \partial_t G_w(u+th,v) - \int_0^1 \int_{\Omega_f} D_1G_w(u,v) h dx dt \right| \\ = & \left| \int_0^1 \int_{\Omega_f} (D_1G_w(u+th,v) - D_1G_w(u,v)) h dx dt \right| \\ \leq & \int_0^1 \|D_1G_w(u+th,v) - D_1G_w(u,v)\|_{L^2} \|h\|_{L^2} dt \\ \leq & \int_0^1 t dt \alpha \|h\|_{L^2}^2 = \frac{\alpha}{2} \|h\|_{L^2}^2. \end{aligned}$$

We can also obtain similar estimates for the variable v. Thus, we get

$$\begin{aligned}
&\left|G_w(u,v) - G_w(u_0,v_0) - \int_{\Omega_f} D_1G_w(u_0,v_0)(u-u_0)dx - \int_{\Omega_g} D_2G_w(u_0,v_0)(v-v_0)dy\right| \\
&\quad \le \left|G_w(u,v) - G_w(u_0,v) - \int_{\Omega_f} D_1G_w(u_0,v_0)(u-u_0)dx\right| \\
&\quad\quad + \left|G_w(u_0,v) - G_w(u_0,v_0) - \int_{\Omega_f} D_1G_w(u_0,v_0)(v-v_0)dx\right| \\
&\quad \le \frac{\alpha}{2}\left(\|u-u_0\|_{L^2}^2 + \|v-v_0\|_{L^2}^2\right).
\end{aligned}$$

We take $(u_0,v_0) = (D_1G_w^*, D_2G_w^*)(u_0^*,v_0^*)$, then we have $(u_0^*,v_0^*) = (D_1G_w, D_2G_w)(u_0,v_0)$. Thus, we find

$$\begin{aligned}
G_w(u,v) &\le G_w(u_0,v_0) + \int_{\Omega_f} D_1G_w(u_0,v_0)(u-u_0)dx + \int_{\Omega_g} D_2G_w(u_0,v_0)(v-v_0)dy \\
&\quad + \frac{\alpha}{2}\left(\|u-u_0\|_{L^2}^2 + \|v-v_0\|_{L^2}^2\right) \\
&= -G^*(u_0^*,v_0^*) + \int_{\Omega_f} u_0^* u dx + \int_{\Omega_g} v_0^* v dy + \frac{\alpha}{2}\left(\|u-u_0\|_{L^2}^2 + \|v-v_0\|_{L^2}^2\right).
\end{aligned}$$

For $(u^*,v^*) \in L^*$, we have

$$\begin{aligned}
G_w^*(u^*,v^*) &= \sup_{(u,v)\in L^2(\Omega_f)\times L^2(\Omega_g)} \left\{\int_{\Omega_f} u^* u dx + \int_{\Omega_g} v^* v dy - G_w(u,v)\right\} \\
&\ge G_w^*(u_0^*,v_0^*) \\
&\quad + \sup_{(u,v)\in L^2(\Omega_f)\times L^2(\Omega_g)} \left\{\int_{\Omega_f}(u^*-u_0^*)u dx + \int_{\Omega_g}(v^*-v_0^*)v dy - \frac{\alpha}{2}\left(\|u-u_0\|_{L^2}^2 + \|v-v_0\|_{L^2}^2\right)\right\} \\
&= G_w^*(u_0^*,v_0^*) + \int_{\Omega_f}(u^*-u_0^*)u_0 dx + \int_{\Omega_g}(v^*-v_0^*)v_0 dy + \frac{1}{2\alpha}\left(\|u^*-u_0^*\|_{L^2}^2 + \|v^*-v_0^*\|_{L^2}^2\right).
\end{aligned}$$

Hence, $\Delta G_w^*(u_0^*,v_0^*,u^*,v^*) \ge \frac{1}{2\alpha}\left(\|u^*-u_0^*\|_{L^2}^2 + \|v^*-v_0^*\|_{L^2}^2\right)$.

3.2. **The minimization algorithm.** In this subsection, we will show that (X^{n+1}, Y^{n+1}) of the algorithm constructed in Subsection 2.3 can be computed via (X^n, Y^n) using the function G defined in Subsection 3.1. To this end, we recall $\mathcal{L}_g^f$ in Remark 7 and then consider two additional minimization problems:

$$\inf_{(u,v)\in\mathcal{L}_g^f\times\mathcal{L}_f^g} \left\{\int_{\Omega_f} u^* u dx + \int_{\Omega_g} v^* v dy + \frac{\varepsilon}{4\tau} w(u,v) + \gamma_1 w(u-u_0, v-v_0)\right\}, \tag{31}$$

where $(u^*,v^*) \in L^*, (u_0,v_0) \in \mathcal{L}_g^f \times \mathcal{L}_f^g$ and γ_1 is a constant; the second is

$$\inf_{(u^*,v^*)\in L^*} \left\{-\int_{\Omega_f} u^* u dx - \int_{\Omega_g} v^* v dy + G_w^*(u^*,v^*) + \gamma_2 \Delta G_w^*(u_0^*,v_0^*,u^*,v^*)\right\}, \tag{32}$$

where $(u_0^*, v_0^*) \in L^*, (u,v) \in \mathcal{L}_g^f \times \mathcal{L}_f^g$ and γ_2 is a constant. These minimization problems has the following important properties, that are summarized in the following lemma.

Lemma 14. *The problem* (31) *admits a unique minimizer (up to almost everywhere with respect to the Lebesgue measure). We denote the unique minimizer of problem* (31) *by*

$$(\mathbf{u}, \mathbf{v})(u^*, v^*, u_0, v_0, \gamma_1).$$

The problem (32) *admits a unique minimizer, which is equal to* $(D_1 G_w(\varphi_u, \varphi_v), D_2 G_w(\varphi_u, \varphi_v))$ *in* L^2*, where*

$$\varphi_u = \varphi_u(u_0^*, v_0^*, \gamma_2) := \frac{u + \gamma_2 D_1 G_w^*(u_0^*, v_0^*)}{1+\gamma_2}, \varphi_v = \varphi_v(u_0^*, v_0^*, \gamma_2) := \frac{v + \gamma_2 D_2 G_w^*(u_0^*, v_0^*)}{1+\gamma_2},$$

and G_w^* *is defined in* (28)*.*

Consider the algorithm constructed in Subsection 2.3 and $n \in \mathbb{N}$*, then*

$$(X_0^{n+1}, Y_0^{n+1}) = (\mathbf{u}, \mathbf{v})\left((r+1)X^{n+1} - rX^n, (r+1)Y^{n+1} - rY^n, X_0^n, Y_0^n, s\right), \tag{33}$$

and

$$(X^{n+1}, Y^{n+1}) = (D_1 G_w(X_*^n, Y_*^n), D_2 G_w(X_*^n, Y_*^n)). \tag{34}$$

Proof of Lemma 14. In (31), the term we want to minimize can be rewritten in the form

$$\int_{\Omega_f} \left\{\frac{1}{2}\left(\frac{\varepsilon}{4\tau} + \gamma_1\right) u^2 + u(u^* - \gamma_1 u_0) + \frac{\gamma_1}{2} u_0^2\right\} + \int_{\Omega_g} \left\{\frac{1}{2}\left(\frac{\varepsilon}{4\tau} + \gamma_1\right) v^2 + v(v^* - \gamma_1 v_0) + \frac{\gamma_1}{2} v_0^2\right\}.$$

The minimizer is

$$u = 4\tau \min\left\{\frac{\gamma_1 u_0 - u^*}{\varepsilon + 4\tau\gamma_1}, \frac{1}{2} - \frac{|\Omega_g|}{2 f_\varepsilon}\right\}, v = 4\tau \min\left\{\frac{\gamma_1 v_0 - v^*}{\varepsilon + 4\tau\gamma_1}, \frac{1}{2} - \frac{|\Omega_f|}{2 g_\varepsilon}\right\}.$$

We easily obtain (33).

To prove (32), we rewrite the term that we want to minimize as

$$(1+\gamma_2)G_w^*(u^*, v^*) - \gamma_2 G_w^*(u_0^*, v_0^*) - \int_{\Omega_f} u^*(u + \gamma_2 D_1 G_w^*(u_0^*, v_0^*)) - \int_{\Omega_f} v^*(v + \gamma_2 D_2 G_w^*(u_0^*, v_0^*))$$
$$+ \int_{\Omega_f} u_0^* \gamma_2 D_1 G_w^*(u_0^*, v_0^*) + \int_{\Omega_f} v_0^* \gamma_2 D_2 G_w^*(u_0^*, v_0^*).$$

We obtain the formula for the minimizer using the definition of the convex conjugate and (29). □

3.3. **Proof of Proposition 8.** In this subsection, we will construct $(\bar{k_1^*}, \bar{k_2^*})$, mentioned in Proposition 8 inductively using the algorithm defined in Subsection 2.3. To this end, let us fix $u \in \mathcal{L}_g^f, v \in \mathcal{L}_f^g$ and an $n \in \mathbb{N}$. From (33) and the convexity of w, we apply the first derivative test for convex functions ([25, Proposition 5.5 Chapter 1 Part 1]), which gives

$$\int_{\Omega_f} \left((r+1)X^{n+1} - rX^n\right)(u - X_0^{n+1})dx + \int_{\Omega_g} \left((r+1)Y^{n+1} - rY^n\right)(v - Y_0^{n+1})dy$$
$$+ \frac{\varepsilon}{4\tau}\left(\int_{\Omega_f} X_0^{n+1}(u - X_0^{n+1})dx + \int_{\Omega_g} Y_0^{n+1}(v - Y_0^{n+1})dy\right)$$

$$+ s\left(\int_{\Omega_f}(X_0^{n+1}-X_0^n)(u-X_0^{n+1})dx+\int_{\Omega_g}(Y_0^{n+1}-Y_0^n)(v-Y_0^{n+1})dy\right)\geq 0.$$

We note that

$$w(u-u_0,v-v_0)=w(u,v)-w(u_0,v_0)-\int_{\Omega_f}u_0(u-u_0)dx-\int_{\Omega_g}v_0(v-v_0)dy,$$

which leads us to the following estimate

$$\begin{aligned}
&\int_{\Omega_f}\left((r+1)X^{n+1}-rX^n\right)(u-X_0^{n+1})dx+\int_{\Omega_g}\left((r+1)Y^{n+1}-rY^n\right)(v-Y_0^{n+1})dy\\
&+\frac{\varepsilon}{4\tau}\left(w(u,v)-w(X_0^{n+1},Y_0^{n+1})\right)+s(w(u-X_0^n,v-Y_0^n)-w(X_0^{n+1}-X_0^n,Y_0^{n+1}-Y_0^n))\\
=&\int_{\Omega_f}\left((r+1)X^{n+1}-rX^n\right)(u-X_0^{n+1})dx+\int_{\Omega_g}\left((r+1)Y^{n+1}-rY^n\right)(v-Y_0^{n+1})dy\\
&+s\left(\int_{\Omega_f}(X_0^{n+1}-X_0^n)(u-X_0^{n+1})dx+\int_{\Omega_g}(Y_0^{n+1}-Y_0^n)(v-Y_0^{n+1})dy\right)\\
&+\frac{\varepsilon}{4\tau}\left(\int_{\Omega_f}X_0^{n+1}(u-X_0^{n+1})dx+\int_{\Omega_g}Y_0^{n+1}(v-Y_0^{n+1})dy\right)\\
&+sw(u-X_0^{n+1},v-Y_0^{n+1})+\frac{\varepsilon}{4\tau}w(u-X_0^{n+1},v-Y_0^{n+1})\\
\geq&\left(s+\frac{\varepsilon}{4\tau}\right)w(u-X_0^{n+1},v-Y_0^{n+1}).
\end{aligned}$$

In short, we now have,

$$\begin{aligned}
&\int_{\Omega_f}\left((r+1)X^{n+1}-rX^n\right)(u-X_0^{n+1})dx+\int_{\Omega_g}\left((r+1)Y^{n+1}-rY^n\right)(v-Y_0^{n+1})dy\\
&+\frac{\varepsilon}{4\tau}\left(w(u,v)-w(X_0^{n+1},Y_0^{n+1})\right)\\
\geq&\left(s+\frac{\varepsilon}{4\tau}\right)w(u-X_0^{n+1},v-Y_0^{n+1})-sw(u-X_0^n,v-Y_0^n)+sw(X_0^{n+1}-X_0^n,Y_0^{n+1}-Y_0^n).
\end{aligned}\tag{35}$$

By (34) and [25, (5.2) Chapter 1 Part 1], we obtain

$$\begin{aligned}
A_{X,Y}&=(1+q)G_w(X_*^{n+1},Y_*^{n+1})+\frac{\varepsilon}{4\tau}w(X_0^{n+1},Y_0^{n+1})-G(u,v)\\
&\leq(1+q)G_w(X_*^{n+1},Y_*^{n+1})-G_w(X_*^{n+1},Y_*^{n+1})-\int_{\Omega_f}X^{n+2}(u-X_*^{n+1})dx-\int_{\Omega_g}Y^{n+2}(v-Y_*^{n+1})dy\\
&\quad+\frac{\varepsilon}{4\tau}\left(w(X_0^{n+1},Y_0^{n+1})-w(u,v)\right).
\end{aligned}$$

Since $X_*^{n+1} = \frac{q}{1+q} X_*^n + \frac{1}{1+q} X_0^{n+1}$, we get $q(X_*^{n+1} - X_*^n) - X_0^{n+1} = -X_*^{n+1}$. Similarly, we also get $q(Y_*^{n+1} - Y_*^n) - Y_0^{n+1} = -Y_*^{n+1}$. Then, we further estimate

$$A_{X,Y} \le q \left(\underbrace{G_w(X_*^{n+1}, Y_*^{n+1}) - \int_{\Omega_f} X^{n+2}(X_*^{n+1} - X_*^n) dx - \int_{\Omega_g} Y^{n+2}(Y_*^{n+1} - Y_*^n) dy}_{B_{X,Y}} \right)$$
$$- \int_{\Omega_f} X^{n+2}(u - X_0^{n+1}) dx - \int_{\Omega_g} Y^{n+2}(v - X_0^{n+1}) dy + \frac{\varepsilon}{4\tau} \left(w(X_0^{n+1}, Y_0^{n+1}) - w(u,v) \right).$$

We work on $B_{X,Y}$ first. By (29) and (34), we have

$$\int_{\Omega_f} X^{n+2} X_*^{n+1} dx + \int_{\Omega_g} Y^{n+2} Y_*^{n+1} dy = G_w(X_*^{n+1}, Y_*^{n+1}) + G_w^*(X^{n+2}, Y^{n+2}),$$
$$\int_{\Omega_f} X^{n+1} X_*^n dx + \int_{\Omega_g} Y^{n+1} Y_*^n dy = G_w(X_*^n, Y_*^n) + G_w^*(X^{n+1}, Y^{n+1}).$$

We rewrite $B_{X,Y}$ as

$$B_{X,Y} = G_w(X_*^n, Y_*^n) + G_w^*(X^{n+1}, Y^{n+1}) - G_w^*(X^{n+2}, Y^{n+2})$$
$$+ \int_{\Omega_f} (X^{n+2} - X^{n+1}) X_*^n dx + \int_{\Omega_g} (Y^{n+2} - Y^{n+1}) Y_*^n dy.$$

By the definition of $D_1 G^*, D_2 G^*$ and ΔG_w^*, we get

$$\Delta G_w^*(X^{n+1}, Y^{n+1}, X^{n+2}, Y^{n+2}) = G_w^*(X^{n+2}, Y^{n+2}) - G_w^*(X^{n+1}, Y^{n+1})$$
$$- \int_{\Omega_f} (X^{n+2} - X^{n+1}) X_*^n dx - \int_{\Omega_g} (Y^{n+2} - Y^{n+1}) Y_*^n dy.$$

Thus, $B_{X,Y} = G_w(X_*^n, Y_*^n) - \Delta G_w^*(X^{n+1}, Y^{n+1}, X^{n+2}, Y^{n+2})$, and hence it follows that

$$A_{X,Y} \le q \left(G_w(X_*^n, Y_*^n) - \Delta G_w^*(X^{n+1}, Y^{n+1}, X^{n+2}, Y^{n+2}) \right)$$
$$- \int_{\Omega_f} X^{n+2}(u - X_0^{n+1}) dx - \int_{\Omega_g} Y^{n+2}(v - X_0^{n+1}) dy + \frac{\varepsilon}{4\tau} \left(w(X_0^{n+1}, Y_0^{n+1}) - w(u,v) \right).$$

Applying (30) and (35) to the right hand side of the above inequality, we obtain

$$A_{X,Y} \le q G_w(X_*^n, Y_*^n) - \frac{q}{2\alpha} \left(\|X^{n+2} - X^{n+1}\|_{L^2}^2 + \|Y^{n+2} - Y^{n+1}\|_{L^2}^2 \right)$$
$$+ \int_{\Omega_f} \left(X^{n+2} - X^{n+1} - r(X^{n+1} - X^n) \right) (X_0^{n+1} - u) dx$$
$$+ \int_{\Omega_g} \left(Y^{n+2} - Y^{n+1} - r(Y^{n+1} - Y^n) \right) (Y_0^{n+1} - v) dy$$
$$- \left(s + \frac{\varepsilon}{4\tau} \right) w(u - X_0^{n+1}, v - Y_0^{n+1}) + s w(u - X_0^n, v - X_0^n) - s w(X_0^{n+1} - X_0^n, Y_0^{n+1} - Y_0^n).$$

Hence, we have proved

$$(1+q) \left(G_w(X_*^{n+1}, Y_*^{n+1}) - G(u,v) \right) + \frac{\varepsilon}{4\tau} w(X_0^{n+1}, Y_0^{n+1}) \le q \left(G_w(X_*^n, Y_*^n) - G(u,v) \right) + \widetilde{A}_n,$$

where

$$\begin{aligned}\widetilde{A}_n = & -\frac{q}{2\alpha}\left(\|X^{n+2}-X^{n+1}\|_{L^2}^2+\|Y^{n+2}-Y^{n+1}\|_{L^2}^2\right)\\ & +\int_{\Omega_f}\left(X^{n+2}-X^{n+1}-r(X^{n+1}-X^n)\right)(X_0^{n+1}-u)dx\\ & +\int_{\Omega_g}\left(Y^{n+2}-Y^{n+1}-r(Y^{n+1}-Y^n\right)(Y_0^{n+1}-v)dy\\ & -\left(s+\frac{\varepsilon}{4\tau}\right)w(u-X_0^{n+1},v-Y_0^{n+1})+sw(u-X_0^n,v-X_0^n)-sw(X_0^{n+1}-X_0^n,Y_0^{n+1}-Y_0^n).\end{aligned}$$

We will use the above inequality inductively. To this end, let β_m be a positive sequence satisfying

$$\beta_m q=\beta_{m-1}(1+q),\forall m\geq 1.$$

We can easily see that

$$\beta_m=\beta_0 r^{-m},\forall m\geq 0.$$

Choosing $\beta_0=1$, we get $\beta_m=r^{-m}$. Now, the inductive steps give

$$\begin{aligned}\beta_{m+1}q\left(G_w(X_1^{m+1},Y_1^{m+1})-G(u,v)\right)+\frac{\varepsilon}{4\tau}\sum_{j=0}^m\beta_j w(X_0^{j+1},Y_0^{j+1})\\ \leq q\left(G_w(0,0)-G(u,v)\right)+\sum_{j=0}^m\beta_j\widetilde{A}_j.\end{aligned}$$

We will next bound $\sum_{j=0}^m\beta_j\widetilde{A}_j$. Noticing that $\frac{q\varepsilon}{4\tau}=s$, we find

$$\beta_j\left(s+\frac{\varepsilon}{4\tau}\right)=r^{-j}\left(s+\frac{\varepsilon}{4\tau}\right)=r^{-j-1}\left(\frac{qs}{1+q}+\frac{q\varepsilon}{4\tau(1+q)}\right)=\beta_{j+1}s,$$

which implies

$$\begin{aligned}\sum_{j=0}^m\beta_j\left(\left(s+\frac{\varepsilon}{4\tau}\right)w(u-X_0^{j+1},v-Y_0^{j+1})-sw(u-X_0^j,v-X_0^j)\right)\\ =\beta_m\left(s+\frac{\varepsilon}{4\tau}\right)w(u-X_0^{m+1},v-Y_0^{m+1})-sw(u,v).\end{aligned}$$

We also have another identity

$$\begin{aligned}&\sum_{j=0}^m\beta_j\int_{\Omega_f}\left(X^{j+2}-X^{j+1}-r(X^{j+1}-X^j)\right)(X_0^{j+1}-u)dx\\ &=-\sum_{j=1}^m\beta_{j-1}\int_{\Omega_f}\left(X^{j+1}-X^j\right)(X_0^{j+1}-X_0^j)dx+\beta_m\int_{\Omega_f}\left(X^{m+2}-X^{m+1}\right)(X_0^{m+1}-u)dx.\end{aligned}$$

For $1 \le j \le m$, we have the follow estimate

$$-\frac{\beta_{j-1}q}{2\alpha}\|X^{j+1}-X^j\|_{L^2}^2+\beta_{j-1}\int_{\Omega_f}\left(X^{j+1}-X^j\right)(X_0^{j+1}-X_0^j)dx-\frac{\beta_j s}{2}\|X_0^{j+1}-X_0^j\|_{L^2}^2$$
$$=\frac{\beta_{j-1}}{2}\left(\frac{1}{s}-\frac{q}{\alpha}\right)\|X^{j+1}-X^j\|_{L^2}^2-\frac{\beta_{j-1}s}{2}\left\|\frac{X^{j+1}-X^j}{s}-X_0^{j+1}+X_0^j\right\|_{L^2}^2$$
$$-\frac{\beta_{j-1}\varepsilon}{8\tau}\|X_0^{j+1}-X_0^j\|_{L^2}^2 \;\le\; 0.$$

For β_m, we have a similar estimate

$$-\frac{\beta_m q}{2\alpha}\|X^{m+2}-X^{m+1}\|_{L^2}^2+\beta_m\int_{\Omega_f}\left(X^{m+2}-X^{m+1}\right)(X_0^{m+1}-u)dx-\frac{\beta_m s}{2}\|X^{m+1}-u\|_{L^2}^2\le 0.$$

We also repeat these estimates for the Y counterpart. Therefore, we obtain the following bound

$$\sum_{j=0}^m \beta_j\widetilde{A}_j \le sw(u,v)-\frac{\varepsilon\beta_m}{4\tau}w(u-X_0^{m+1},v-X_0^{m+1}).$$

For all m, we now obtain the estimate

$$\beta_{m+1}q\left(G_w(X_*^{m+1},Y_*^{m+1})-G(u,v)\right)+\frac{\varepsilon\beta_m}{4\tau}w(u-X_0^{m+1},v-X_0^{m+1})+\frac{\varepsilon}{4\tau}\sum_{j=0}^m\beta_j w(X_0^{j+1},Y_0^{j+1})$$
$$\le q\left(G_w(0,0)-G(u,v)\right)+sw(u,v).$$

Recalling that $X_*^1 = X_0^1/(1+q)$, we deduce $r^{-1}X_*^1 = X_0^1/q$. Similarly, we have $X_*^2 = rX_*^1+(1-r)X_0^2$ so $r^{-2}X_*^2 = r^{-1}X_*^1+(r^{-2}-r^{-1})X_0^2 = 1/q(X_0^1+r^{-1}X_0^2)$. Inductively,

$$\beta_{m+1}X_*^{m+1}=\frac{1}{q}\sum_{j=0}^m\beta_j X_0^{j+1}.$$

A similar equality also holds for Y. By the convexity and homogeneity of w,

$$\left(\frac{\beta_{m+1}q}{\sum_{j=0}^m\beta_j}\right)^2 w\left(X_*^{m+1},Y_*^{m+1}\right)\le\sum_{j=0}^m\frac{\beta_j}{\sum_{j'=0}^m\beta_{j'}}w\left(X_0^{j+1},Y_0^{j+1}\right).$$

We compute

$$\frac{\beta_{m+1}q}{\sum_{j=0}^m\beta_j}=\frac{r^{-m-1}q}{1+r^{-1}+\cdots+r^{-m}}=\frac{r^m(1-r)q}{(1-r^{m+1})r^{m+1}}=\frac{1}{1-r^{m+1}}>1.$$

This gives

$$\beta_{m+1}qw\left(X_*^{m+1},Y_*^{m+1}\right)<\sum_{j=0}^m\beta_j w(X_0^{j+1},Y_0^{j+1}).$$

As a consequence,

$$\beta_{m+1}q\left(G(X_*^{m+1},Y_*^{m+1})-G(u,v)\right)+\frac{\varepsilon\beta_m}{4\tau}w(u-X_0^{m+1},v-X_0^{m+1})\le q\left(G(0,0)-G(u,v)\right)+sw(u,v).$$

Let $(u,v) = (k_1^*, k_2^*)$ be the minimizer of (14) in $\mathcal{L}_g^f \times \mathcal{L}_f^g$. By (49) and (50) of the Appendix, (k_1^*, k_2^*) is also the minimizer of G. Therefore,

$$\left\|k_1^* - X_0^{m+1}\right\|_{L^2}^2 + \left\|k_2^* - Y_0^{m+1}\right\|_{L^2}^2 \leq \frac{4\tau r^m}{\varepsilon}\left(q\left(G(0,0) - G(k_1^*,k_2^*)\right) + sw(k_1^*,k_2^*)\right). \tag{36}$$

We then set $(\bar{k_1^*}, \bar{k_2^*}) = (X_0^{L+1}, Y_0^{L+1})$. Finally, as an additional note, the algorithm is built upon pointwise maximum, minimum and integration over fixed compact sets. Hence, when f, g, C are continuous, the output X_0^m, Y_0^m are also continuous.

3.4. **Proof of Proposition 9.** In this subsection, we will construct the optimal transport equation from the values (X_0^{L+1}, Y_0^{L+1}) obtained from the algorithm. To this end, we recall the relation of the minimizer k of (11) and the minimizer (k_1^*, k_2^*) of (14) in the system (15) - (17) and Remark 7.

From the approximation algorithm and (36), we have $(\bar{k_1^*}, \bar{k_2^*}) = (X_0^{L+1}, Y_0^{L+1})$ such that

$$\left\|k_1^* - \bar{k_1^*}\right\|_{L^2}^2 + \left\|k_2^* - \bar{k_2^*}\right\|_{L^2}^2 \leq \frac{4\tau r^L}{\varepsilon}\left(q\left(G(0,0) - G(k_1^*,k_2^*)\right) + sw(k_1^*,k_2^*)\right).$$

Since Ω_f is compact, Stone-Weierstrass Theorem says that the space $C^\infty(\Omega_f)$ is dense in $C(\Omega_f)$ under the uniform topology. In a similar way, $C^\infty(\Omega_g)$ is dense in $C(\Omega_g)$ under the uniform topology. We take $\widetilde{k_1^*}, \widetilde{k_2^*}$ as smooth functions satisfying

$$\|\bar{k_i^*} - \widetilde{k_i^*}\|_{L^\infty} \leq \mathrm{err}_0,$$

for any $0 < \mathrm{err}_0 < \frac{\tau\delta\max\{|\Omega_\mathrm{f}|,|\Omega_\mathrm{g}|\}}{\varepsilon}$.

For $\bar{k}(x,y) = \max\left\{\delta, \frac{\widetilde{k_1^*}(x) + \widetilde{k_2^*}(y) - C(x,y)}{2\eta}\right\}$, we get

$$\left\|k - \bar{k}\right\|_{L^2} \leq \frac{\sqrt{\tau} r^{L/2}}{\eta\sqrt{\varepsilon}}\sqrt{q\left(G(0,0) - G(k_1^*,k_2^*)\right) + sw(k_1^*,k_2^*)} + \frac{|\Omega|\mathrm{err}_0}{\eta}.$$

Recalling that $\bar{k}_x = \int_{\Omega_g}\bar{k}dy, \bar{k}_y = \int_{\Omega_f}\bar{k}dx$, we deduce

$$\begin{aligned}\left\|k_x - \bar{k}_x\right\|_{L^2} \leq& \frac{\sqrt{|\Omega_g|\tau} r^{L/2}}{\eta\sqrt{\varepsilon}}\sqrt{q\left(G(0,0) - G(k_1^*,k_2^*)\right) + sw(k_1^*,k_2^*)} + \frac{\sqrt{|\Omega_g|}|\Omega|\mathrm{err}_0}{\eta},\\ \left\|k_y - \bar{k}_y\right\|_{L^2} \leq& \frac{\sqrt{|\Omega_f|\tau} r^{L/2}}{\eta\sqrt{\varepsilon}}\sqrt{q\left(G(0,0) - G(k_1^*,k_2^*)\right) + sw(k_1^*,k_2^*)} + \frac{\sqrt{|\Omega_f|}|\Omega|\mathrm{err}_0}{\eta}.\end{aligned}$$

We define

$$\overline{f} = \frac{2\tau\bar{k}_x}{2\tau - \widetilde{k_1^*}}, \quad \overline{g} = \frac{2\tau\bar{k}_y}{2\tau - \widetilde{k_2^*}}.$$

Since $k_1^*, \widetilde{k_1^*} \leq 2\tau - \tau\frac{\delta|\Omega_g|}{f_\varepsilon}$ a.e., we will prove

$$\left\|\frac{1}{2\tau - k_1^*}\right\|_{L^\infty}, \left\|\frac{1}{2\tau - \widetilde{k_1^*}}\right\|_{L^\infty} \leq \frac{\|f_\varepsilon\|_{L^\infty}}{\tau\delta|\Omega_g|}, \tag{37}$$

$$\left\|\frac{k_x}{2\tau - k_1^*}\right\|_{L^\infty} = \frac{\|f_\varepsilon\|_{L^\infty}}{2\tau}, \tag{38}$$

$$\left\|\frac{k_1^*}{2\tau - k_1^*}\right\|_{L^\infty} \le 2\frac{\|f_\varepsilon\|_{L^\infty}}{\delta|\Omega_g|}. \tag{39}$$

The estimates in (37) are obvious. The equality in (38) comes from (16). The proof of (39) goes as follows. If $k_1^* \le 0$ then $-k_1^* \le 4\tau - 2k_1^*$, and since $\delta|\Omega_g| \le \varepsilon \le f_\varepsilon$, we find

$$\left|\frac{k_1^*}{2\tau - k_1^*}\right| \le 2 \le \frac{2f_\varepsilon}{\delta|\Omega_g|}.$$

If $k_1^* > 0$, then $f_\varepsilon > \delta|\Omega_g|$ and we get

$$\left(1 + \frac{\delta|\Omega_g|}{f_\varepsilon - \delta|\Omega_g|}\right) k_1^* \le 2\tau,$$

which is equivalent to

$$\frac{k_1^*}{2\tau - k_1^*} \le \frac{f_\varepsilon}{\delta|\Omega_g|} - 1.$$

Hence, we obtain (39). As a consequence, we have the estimate

$$\begin{aligned}
\left\|f_\varepsilon - \overline{f}\right\|_{L^2} &\le 4\tau^2 \left\|\frac{k_x - \bar{k}_x}{(2\tau - k_1^*)(2\tau - \widetilde{k_1^*})}\right\|_{L^2} + 2\tau \left\|\frac{k_x(k_1^* - \widetilde{k_1^*})}{(2\tau - k_1^*)(2\tau - \widetilde{k_1^*})}\right\|_{L^2} \\
&\quad + 2\tau \left\|\frac{k_1^*(k_x - \bar{k}_x)}{(2\tau - k_1^*)(2\tau - \widetilde{k_1^*})}\right\|_{L^2} \\
&\le \frac{E^2}{\delta|\Omega_g|}\left((4 + \frac{1}{\tau})\frac{1}{\delta|\Omega_g|}\left\|k_x - \bar{k}_x\right\|_{L^2} + \frac{1}{2\tau}\left\|k_1^* - \widetilde{k_1^*}\right\|_{L^2}\right) \\
&\le \frac{E^2 r^{L/2}}{\delta|\Omega_g|\sqrt{\varepsilon}}\left((4 + \frac{1}{\tau})\frac{\sqrt{\tau}}{\eta\delta\sqrt{|\Omega_g|}} + \frac{1}{\sqrt{\tau}}\right)\sqrt{q\left(G(0,0) - G(k_1^*, k_2^*)\right) + sw(k_1^*, k_2^*)} \\
&\quad + \frac{E^2}{\delta|\Omega_g|}\left((4 + \frac{1}{\tau})\frac{|\Omega|}{\eta\delta\sqrt{|\Omega_g|}} + \frac{|\Omega_f|}{2\tau}\right)\mathrm{err}_0.
\end{aligned}$$

Similarly, we obtain the following estimate for g_ε

$$\begin{aligned}
\|g_\varepsilon - \overline{g}\|_{L^2} &\le \frac{E^2 r^{L/2}}{\delta|\Omega_f|\sqrt{\varepsilon}}\left((4 + \frac{1}{\tau})\frac{\sqrt{\tau}}{\eta\delta\sqrt{|\Omega_f|}} + \frac{1}{\sqrt{\tau}}\right)\sqrt{q\left(G(0,0) - G(k_1^*, k_2^*)\right) + sw(k_1^*, k_2^*)} \\
&\quad + \frac{E^2}{\delta|\Omega_f|}\left((4 + \frac{1}{\tau})\frac{|\Omega|}{\eta\delta\sqrt{|\Omega_f|}} + \frac{|\Omega_g|}{2\tau}\right)\mathrm{err}_0.
\end{aligned}$$

We now investigate the convex solution ϕ of (21). By [26], ϕ is strongly convex and $\phi \in C^2$. For each fixed $t \in [0, T]$, the strong convexity implies $\mathbb{T}_t$ is injective and $\mathrm{Hess}(\mathbb{T}_t)$ is strictly positive. This means the map $\mathbb{T}_t^{-1} : \mathbb{T}_t(\Omega_f) \to \Omega_f$ is well-defined for each $t \in [0, T]$ and $\mathbb{T}^{-1} \in C^1$. As a consequence, ξ_t is well-defined and $\xi_t \in C^1$.

We see that $\xi_t(\mathbb{T}_t(x)) = \frac{1}{T}(\nabla\phi(x) - x) = \partial_t \mathbb{T}_t$. By [13, Theorem 5.34], the map $\mu_t = (\mathbb{T}_t)_\#(\bar{k}_x \mathcal{L})$ is the solution of the homogeneous transport equation

$$\begin{cases} \partial_t \mu_t + \nabla \cdot (\xi_t \mu_t) = 0, \\ \mu_0 = \bar{k}_x \mathcal{L}. \end{cases}$$

Of course, for this equation, we have $\mu_T = \bar{k}_y \mathcal{L}$ since $\nabla\phi$ solves (21).

For the inhomogeneous part, we note that since $\widetilde{k}_1^*, \widetilde{k}_2^*$ are smooth then ζ_t is well-defined and is in C^1. We compute

$$\begin{aligned} \zeta_t(\mathbb{T}_t(x)) &= \frac{1}{T}(\nabla\phi(x) - x) \cdot \left(-\frac{\nabla \widetilde{k}_1^*(x - \nabla\phi(x) - \mathbb{T}_t(x))}{2\tau - \widetilde{k}_1^*(x - \nabla\phi(x) - \mathbb{T}_t(x))} + \frac{\nabla \widetilde{k}_2^*(\mathbb{T}_t(x))}{2\tau - \widetilde{k}_2^*(\mathbb{T}_t(x))} \right) \\ &= \frac{\partial_t(-\widetilde{k}_1^*(x - \nabla\phi(x) - \mathbb{T}_t(x)))}{2\tau - \widetilde{k}_1^*(x - \nabla\phi(x) - \mathbb{T}_t(x))} - \frac{\partial_t(-\widetilde{k}_2^*(\mathbb{T}_t))}{2\tau - \widetilde{k}_2^*(\mathbb{T}_t)} \\ &= \partial_t \left(\log\left(2\tau - \widetilde{k}_1^*(x - \nabla\phi(x) - \mathbb{T}_t(x))\right) - \log\left(2\tau - \widetilde{k}_2^*(\mathbb{T}_t(x))\right)\right) \end{aligned}$$

Hence, we get

$$e^{\int_0^t \zeta_{t'}(\mathbb{T}_{t'}(x))dt'} = \frac{2\tau - \widetilde{k}_1^*(x - \nabla\phi(x) - \mathbb{T}_t(x))}{2\tau - \widetilde{k}_2^*(\mathbb{T}_t(x))}.$$

We take $\mu_0 = \overline{f}\mathcal{L}$, then $\mu_T = (\mathbb{T}_T)_\#(\mu_0 e^{\int_0^T \zeta_{t'}(\mathbb{T}_t(x))dt'}) = \overline{g}\mathcal{L}$ since

$$\begin{aligned} \int_{\mathbb{R}^d} \varphi(x) d(\mathbb{T}_T)_\#(\mu_0 e^{\int_0^T \zeta_{t'}(\mathbb{T}_{t'}(x))dt'}) &= \int_{\mathbb{R}^d} \varphi(\mathbb{T}_T(x)) e^{\int_0^T \zeta_{t'}(\mathbb{T}_{t'}(x))dt'} \frac{2\tau \bar{k}_x}{2\tau - \widetilde{k}_1^*} dx \\ &= \int_{\mathbb{R}^d} \varphi(\mathbb{T}_T(x)) \frac{2\tau \bar{k}_x}{2\tau - \widetilde{k}_2^*(\mathbb{T}_T)} dx \\ &= \int_{\mathbb{R}^d} \varphi \frac{2\tau \bar{k}_y}{2\tau - \widetilde{k}_2^*} dy = \int_{\mathbb{R}^d} \varphi \overline{g} dy, \end{aligned}$$

for all Borel function φ.

For $\mu_t = (\mathbb{T}_t)_\#(\mu_0 e^{\int_0^t \zeta_{t'}(\mathbb{T}_{t'}(x))dt'})$, we will show that μ_t solves

$$\partial_t \mu_t + \nabla \cdot (\xi_t \mu_t) = \zeta_t \mu_t.$$

To this end, for any test function $\varphi \in C^\infty$, we compute

$$\begin{aligned} \partial_t \int_{\mathbb{R}^d} \varphi(x) d\mu_t(x) &= \partial_t \int_{\mathbb{R}^d} \varphi(\mathbb{T}_t(x)) e^{\int_0^t \zeta_{t'}(\mathbb{T}_{t'}(x))dt} d\mu_0 \\ &= \int_{\mathbb{R}^d} \partial_t \mathbb{T}_t(x) \nabla\varphi(\mathbb{T}_t(x)) e^{\int_0^t \zeta_{t'}(\mathbb{T}_{t'}(x))dt} d\mu_0 + \int_{\mathbb{R}^d} \varphi(\mathbb{T}_t) \zeta_t(\mathbb{T}_t(x)) e^{\int_0^t \zeta_{t'}(\mathbb{T}_{t'}(x))dt} d\mu_0 \\ &= \int_{\mathbb{R}^d} \xi_t(\mathbb{T}_t(x)) \nabla\varphi(\mathbb{T}_t(x)) e^{\int_0^t \zeta_{t'}(\mathbb{T}_{t'}(x))dt} d\mu_0 + \int_{\mathbb{R}^d} \varphi(\mathbb{T}_t) \zeta_t(\mathbb{T}_t(x)) e^{\int_0^t \zeta_{t'}(\mathbb{T}_{t'}(x))dt} d\mu_0 \\ &= \int_{\mathbb{R}^d} \xi_t(x) \nabla\varphi(x) d\mu_t + \int_{\mathbb{R}} \varphi(x) \zeta_t(x) d\mu_t. \end{aligned}$$

Thus, μ_t is a solution of the inhomogeneous transport equation.

We also notice that μ_t is the unique solution and μ_t is continuous with respect to t according to [32, Proposition 3.6].

3.5. **Proof of Theorem 11.** In this subsection, we will construct the desired Neural ODE. One of the main ingredient of the proof is to use mollifiers and Hermite polynomials to approximate $\bar{\xi}_t$, that can be seen later in (40) and (44). As thus, the activation function is later chosen to be (46), which comes from Hermite polynomials.

As discussed in Remark 13, we will present the proof for the numerical solution $\nabla\bar{\phi}_n$ of the Monge-Ampère equation (21). Theorem 11 is the special case with $\bar{\phi}_n = \phi$. Hence, the proof for Remark 13 is sufficient.

First, observing that $\nabla\bar{\phi}_n \in C^1$, we find

$$\int_{\bar{\mathbb{T}}_t(\Omega_f)} \|\bar{\xi}_t(x)\|^2 dx \leq \frac{1}{T\min_{x\in\Omega_f}\det(D\bar{\mathbb{T}}_t(x))}\int_{\Omega_f}\left\|\nabla\bar{\phi}_n(x) - x\right\|^2 dx < \infty.$$

This means $\bar{\xi}_t \in L^2(\mathbb{R}^d)$ for each n and we also have the convention that $\bar{\xi}_t(x) = 0$ for $x \notin \bar{\mathbb{T}}_t(\Omega_f)$. We recall that if A, B are positive definite matrices then

$$\det(tA + (1-t)B) \geq \det(A)^t \det(B)^{1-t}.$$

As $\bar{\phi}_n$ is strongly convex, $\det(\text{Hess}\,\bar{\phi}_n(x)) > 0$ for all $x \in \Omega_f$. Since $\bar{\phi}_n \in C^2$ and Ω_f is compact $\min_{x\in\Omega_f}\det(\text{Hess}\,\bar{\phi}_n) > 0$. We then have

$$\det(D\bar{\mathbb{T}}_t) \geq \det(\text{Hess}\,\bar{\phi}_n)^{t/T} \geq \min\left\{1, \min_{\Omega_f}\det(\text{Hess}\,\bar{\phi}_n)\right\} > 0.$$

We then deduce that for each n, $\bar{\xi}_t$ is bounded in L^2 uniformly for all $t \in [0, T]$.

Let

$$\Gamma(x) = \begin{cases} e^{-1/(1-\|x\|^2)}A_1, & \text{if } \|x\| < 1 \\ 0 & \text{if } \|x\| \geq 1 \end{cases},$$

where $1/A_1 = \int_{B(0,1)} e^{-1/(1-\|x\|^2)}$. We consider the mollifier $\Gamma_\rho = \frac{1}{\rho^d}\Gamma(\frac{x}{\rho})$ and $(\bar{\xi}_t^\rho)_j(x) = \int_{\mathbb{R}^d}\Gamma_\rho(y)(\bar{\xi}_t)_j(x-y)dy$. There is $\rho_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n}} > 0$ such that for $0 < \rho < \rho_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n}}$

$$\left\|(\bar{\xi}_t)_j - (\bar{\xi}_t)_j^\rho\right\|_{L^2} < \mathrm{err}_1/3, \forall \mathrm{t} \in [0,\mathrm{T}]. \tag{40}$$

Recalling that $\bar{\xi}_t \in C^1(\bar{\mathbb{T}}_t(\Omega_f))$, hence $\bar{\xi}_t$ is Lipschitz with constant $\max_{x\in\bar{T}_t(\Omega_f)}\|D\bar{\xi}_t(x)\|_{HM}$, where $\|\cdot\|_{HM}$ is the Hilbert-Schmidt norm. Let us recall that for a maxtrix $E \in \mathbb{R}^{d\times d}$ the norm is defined by $\|E\|_{HM} = \sup_{v\neq 0}\frac{\|Ev\|}{\|v\|}$.

We have

$$D\bar{\xi}_t = D(\bar{\mathbb{T}}_t^{-1})(D\nabla\bar{\phi}_n(\bar{\mathbb{T}}_t^{-1}) - Id) = (t\,\text{Hess}\,\bar{\phi}_n(\bar{\mathbb{T}}_t^{-1}) + (1-t)Id)^{-1}(\text{Hess}\,\bar{\phi}_n(\bar{\mathbb{T}}_t^{-1}) - Id),$$

whose norm is uniformly bounded over t, due to the fact that $\bar{\phi}_n \in C^2$, $\bar{\phi}_n$ is strongly convex and its support is compact. We denote $\mathfrak{L}_n$ to be the bound of its norm.

We introduce several notations

$$\bar{\mathbb{T}}_t(\Omega_f)^\rho = \{x \in \bar{\mathbb{T}}_t(\Omega_f) | B(x,\rho) \subset \bar{\mathbb{T}}_t(\Omega_f)\}, \tag{41}$$

$$\bar{\mathbb{T}}_t(\Omega_f)_\rho = \{x \in \mathbb{R}^d | B(x,\rho) \cap \bar{\mathbb{T}}_t(\Omega_f) \neq \emptyset\}, \tag{42}$$

$$\overline{(U,V)} = \{tx + (1-t)y | x \in U, y \in V, t \in [0,1]\}. \tag{43}$$

Next, we estimate

$$\begin{aligned}\left\|(\bar{\xi}_t)_j-(\bar{\xi}_t)_j^{\rho}\right\|_{L^2}^2 &\lesssim \int_{\bar{\mathbb{T}}_t(\Omega_f)^{\rho}}\int_{B(0,\rho)}\mathfrak{L}_n\|y\|^2\Gamma_\rho(y)dydx \\ &\quad+\int_{\bar{\mathbb{T}}_t(\Omega_f)_{\rho}\setminus\bar{\mathbb{T}}_t(\Omega_f)^{\rho}}\int_{B(0,\rho)}\|\bar{\xi}_t(x)-\bar{\xi}_t(x-y)\|^2\Gamma_\rho(y)dydx \\ &\lesssim \rho^2\mathfrak{L}_n|\bar{\mathbb{T}}_t(\Omega_f)|+|\bar{\mathbb{T}}_t(\Omega_f)_\rho\setminus\bar{\mathbb{T}}_t(\Omega_f)^\rho|\max_x\|\bar{\xi}_t\|^2 \\ &\lesssim \rho^2\mathfrak{L}_n|\overline{(\Omega_f,\Omega_g)}|+\rho^d\int_{\partial\bar{\mathbb{T}}_t(\Omega_f)}dS\max_{x\in\Omega_f}(\|\nabla\bar{\phi}(x)\|+\|x\|)/T.\end{aligned}$$

We can see that the above norm uniformly converges to 0 in t as ρ goes to 0.

For a fixed $\rho<\rho_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n}}$, we approximate $(\bar{\xi}_t)_j^{\rho}$ using the Hermite polynomials defined by

$$H_j(x)=(-1)^j e^{x^2}\frac{\partial^j}{\partial x^j}e^{-x^2}$$

for the one dimensional case $d=1$. For the higher dimension case $d>1$, we denote $\vec{n}\in\mathbb{N}^d$ and the Hermite polynomials are defined as $H_{\vec{n}}:\mathbb{R}^d\to\mathbb{R}$

$$H_{\vec{n}}(x)=\prod_{j=1}^{d}H_{n_j}(x_j)=(-1)^{|\vec{n}|}e^{\|x\|^2}\partial^{\vec{n}}e^{-\|x\|^2},$$

where $|\vec{n}|=n_1+n_2+\cdots+n_d$ and $\partial^{\vec{n}}=\partial_{x_1}^{n_1}\dots\partial_{x_d}^{n_d}$. The set $\left\{\frac{H_{\vec{n}}(x)}{\sqrt{\vec{n}!2^{|\vec{n}|}\pi^{d/2}}}e^{-\|x\|^2/2}\right\}$ is then a complete orthonormal set in $L^2(\overline{\bar{\mathbb{T}}_t(\Omega_f)_\rho})$. This means we can write

$$(\bar{\xi}_t(x))_j^{\rho}e^{\|x\|^2/2}=\sum_{\vec{n}\in\mathbb{N}^d}(\widetilde{\xi}_{\vec{n},t})_j^{\rho}\frac{H_{\vec{n}}(x)}{\sqrt{\vec{n}!2^{|\vec{n}|}\pi^{d/2}}}e^{-\|x\|^2/2},$$

where $\vec{n}!=n_1!\dots n_d!$ and the coeffients are defined by

$$(\widetilde{\xi}_{\vec{n},t})_j^{\rho}=\int_{\mathbb{R}^d}(\bar{\xi}_t(x))_j^{\rho}\frac{H_{\vec{n}}(x)}{\sqrt{\vec{n}!2^{|\vec{n}|}\pi^{d/2}}}dx.$$

Thus, we get

$$(\bar{\xi}_t(x))_j^{\rho}=\sum_{\vec{n}\in\mathbb{N}^d}(-1)^{|\vec{n}|}(\widetilde{\xi}_{\vec{n},t})_j^{\rho}\frac{\partial^{\vec{n}}e^{-\|x\|^2}}{\sqrt{\vec{n}!2^{|\vec{n}|}\pi^{d/2}}}.$$

We have $\|\partial^{|\vec{n}|}(\bar{\xi}_t)_j^{\rho}\|_{L^2}, |\vec{n}|\le d+1$, which is bounded uniformly in t. By [16, Theorem 3.9], for $\mathrm{err}_2>0$ we can choose $\mathbf{n}=\mathbf{n}_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n},\mathrm{err}_2}$ uniformly with respect to t such that

$$\left\|(\bar{\xi}_t(x))_j^{\rho}-\sum_{\vec{n}:n_l\le\mathbf{n}}(-1)^{|\vec{n}|}(\widetilde{\xi}_{\vec{n},t})_j^{\rho}\frac{\partial^{\vec{n}}e^{-\|x\|^2}}{\sqrt{\vec{n}!2^{|\vec{n}|}\pi^{d/2}}}\right\|_{L^2}<\mathrm{err}_2.$$

Therefore, there exist a finite subset $N=N_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n}}$ of $\mathbb{N}^d$ such that for all t

$$\left\|(\bar{\xi}_t(x))_j-\sum_{\vec{n}\in N}(-1)^{|\vec{n}|}(\widetilde{\xi}_{\vec{n},t})_j^{\rho}\frac{\partial^{\vec{n}}e^{-\|x\|^2}}{\sqrt{\vec{n}!2^{|\vec{n}|}\pi^{d/2}}}\right\|_{L^2}<2\mathrm{err}_1/3.$$

In the one dimensional case $d = 1$, for each $j \in \mathbb{N}$, by [9, Proposition 2], one has

$$\left\| \frac{\partial^j}{\partial x^j} e^{-x^2} - \frac{1}{\varsigma^j} \sum_{l=0}^{j} (-1)^l \frac{j!}{j!(j-l)!} e^{-(x-l\varsigma)^2} \right\|_{L^2} = O(\varsigma).$$

Let $|N|_\infty = \max\{\|\vec{n}\|_{L^\infty} | \vec{n} \in N\}$ and $|N|$ be the cardinality of N. Since $\frac{\partial^j}{\partial x^j} e^{-x^2}$ is bounded in L^2, there is $\varsigma_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n}}$ such that for $\varsigma < \varsigma_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n}}$

$$\max_{j \le |N|_\infty} \left\| \frac{\partial^j}{\partial x^j} e^{-x^2} - \frac{1}{\varsigma^j} \sum_{l=0}^{j} (-1)^l \frac{j!}{j!(j-l)!} e^{-(x-l\varsigma)^2} \right\|_{L^2} \left(\sum_{j=0}^{|N|_\infty} \left\| \frac{\partial^j}{\partial x^j} e^{-x^2} \right\|_{L^2} \right)^{d-1} < \frac{\mathrm{err}_1}{3d|N| \max_{\vec{n} \in N, t} \|(\widetilde{\xi}_{\vec{n},t})^\rho\|}.$$

For $\vec{m}, \vec{n}$, we denote $\vec{m} \le \vec{n}$ when $m_1 \le n_1, \ldots, m_d \le n_d$ and consider $M = M_{L,\mathrm{err}_0}^{\mathrm{err}_1,\mathrm{n}} = \{\vec{m} \in \mathbb{N}^d | \exists \vec{n} \in N : \vec{m} \le \vec{n}\}$. For $\vec{n} \in \mathbb{N}^d$,

$$\left\| \partial^{\vec{n}} e^{-\|x\|^2} - \frac{1}{\varsigma^{|\vec{n}|}} \sum_{\vec{m} \le \vec{n}} (-1)^{|\vec{m}|} \frac{\vec{n}!}{\vec{m}!(\vec{n}-\vec{m})!} e^{-\|x - \vec{m}\varsigma\|^2} \right\|_{L^2} < \frac{\mathrm{err}_1}{3|N| \max(\bar{\xi}_{\vec{n},t})},$$

Therefore, we obtain the estimate

$$\left\| (\bar{\xi}_t(x))_j - \sum_{\vec{m} \in M} (\xi_{\vec{m},t})_j^{\rho,N} e^{-\|x - \vec{m}\varsigma\|^2} \right\|_{L^2} < \mathrm{err}_1, \tag{44}$$

where

$$(\xi_{\vec{m},t})_j^{\rho,N} = \frac{(-1)^{|\vec{m}|}}{\vec{m}! \pi^{d/2}} \sum_{\vec{n} \in N : \vec{m} \le \vec{n}} \frac{1}{(\vec{n}-\vec{m})!(2\varsigma)^{|\vec{n}|}} \int_{\mathbb{R}^d} (\bar{\xi}_t(x))_j^\rho e^{\|x\|^2} \partial^{\vec{n}} e^{-\|x\|^2} dx.$$

We next show that $(\xi_{\vec{m},t})_j^{\rho,N}$ is Lipschitz continuous in the variable t. We estimate

$$\begin{aligned} \left| \int_{\mathbb{R}^d} ((\bar{\xi}_t)_j^\rho - (\bar{\xi}_{t'})_j^\rho) e^{\|x\|^2} \partial^{\vec{n}} e^{-\|x\|^2} dx \right| \le \frac{1}{T} \int_{\mathbb{R}^d} \int_{\Omega_f} \|\nabla \bar{\phi}_n(y) - y\| e^{\|x\|^2} \partial^{\vec{n}} e^{-\|x\|^2} \\ \times \left| \Gamma_\rho(x - \bar{\mathbb{T}}_t(y)) |\det(D\bar{\mathbb{T}}_t(y))|^{-1} - \Gamma_\rho(x - \bar{\mathbb{T}}_{t'}(y)) |\det(D\bar{\mathbb{T}}_{t'}(y))|^{-1} \right| dy dx. \end{aligned} \tag{45}$$

We recall that $|\det(D\bar{\mathbb{T}}_t(y))|^{-1}$ is uniformly bounded in t. We also have

$$\begin{aligned} &\left| \Gamma_\rho(x - \bar{\mathbb{T}}_t(y)) |\det(D\bar{\mathbb{T}}_{t'}(y))| - \Gamma_\rho(x - \bar{\mathbb{T}}_{t'}(y)) |\det(D\bar{\mathbb{T}}_t(y))| \right| \\ &\le \left| \Gamma_\rho(x - \bar{\mathbb{T}}_t(y)) (|\det(D\bar{\mathbb{T}}_{t'}(y))| - |\det(D\bar{\mathbb{T}}_t(y))|) \right| \\ &\quad + \left| (\Gamma_\rho(x - \bar{\mathbb{T}}_t(y)) - \Gamma_\rho(x - \bar{\mathbb{T}}_{t'}(y))) |\det(D\bar{\mathbb{T}}_t(y))| \right|. \end{aligned}$$

As $\bar{\mathbb{T}}_t(y) - \bar{\mathbb{T}}_{t'}(y) = (t - t')(\nabla \bar{\phi}_n(y) - y)$, we get

$$\left| |\det(D\bar{\mathbb{T}}_{t'}(y))| - |\det(D\bar{\mathbb{T}}_t(y))| \right| \lesssim |t - t'| (\max_{i,j} |(D^2 \bar{\phi}_n(y) - Id)_{i,j}|).$$

Since Γ_ρ is a smooth function,

$$\left| (\Gamma_\rho(x - \bar{\mathbb{T}}_t(y)) - \Gamma_\rho(x - \bar{\mathbb{T}}_{t'}(y))) \right| \lesssim \left\| \bar{\mathbb{T}}_t(y) - \bar{\mathbb{T}}_{t'}(y) \right\| \lesssim |t - t'| \left\| \nabla \bar{\phi}_n(y) - y \right\|.$$

Now, the integrand in (45) vanishes when x is outside of both $\bar{\mathbb{T}}_t(\Omega_f)_\rho$ and $\bar{\mathbb{T}}_{t'}(\Omega_f)_\rho$. If the integrand does not vanish, then $B(x,\rho)\cap\overline{(\Omega_f,\Omega_g)}\neq\emptyset$, which means x is in a bounded set. The domain of integration in (45) can be considered as a compact set and the Lipschitz continuity follows.

Finally, we construct our controlled compound neural network. The activation function in our network is chosen to be

$$\sigma=\begin{pmatrix} e^{-\|x\|^2}\\ e^{-\|x\|^2}\\ \vdots\\ e^{-\|x\|^2}\end{pmatrix} \tag{46}$$

and we define the neural differential equation as

$$\begin{cases}\partial_t\mu_t+\nabla_x\left(\sum_{\vec{m}\in M}W_{\vec{m}}(t)\sigma(A_{\vec{m}}x+b_{\vec{m}})\mu_t\right)=\bar{\zeta}_t\mu_t,\\ \mu_0=\overline{f}\mathcal{L},\end{cases} \tag{47}$$

where $(W_{\vec{m}}(t))_{ij}=(\xi_{\vec{m},t})_j^{\rho,N}\delta_{i,j}, A_{\vec{m}}=Id, b_{\vec{m}}=-\vec{m}\varsigma$. The term with $\bar{\zeta}$ is the feedback control.

For $\widetilde{\xi}_t(x)=\sum_{\vec{m}\in N}W_{\vec{m}}(t)\sigma(A_{\vec{m}}x+b_{\vec{m}}), x\in\bar{\mathbb{T}}_t(\Omega_f)_\rho$, there exists a unique $\widetilde{\mathbb{T}}.(x)$ solving

$$\begin{cases}\partial_t\widetilde{\mathbb{T}}_t(x)=\widetilde{\xi}_t(\widetilde{\mathbb{T}}_t(x)) \quad \text{on } [0,T],\\ \widetilde{T}_0(x)=x.\end{cases}$$

The solution $\widetilde{\mathbb{T}}_t$ is Lipschitz with respect to x. We estimate

$$\begin{aligned}\partial_t\|\widetilde{\mathbb{T}}_t(y)-\widetilde{\mathbb{T}}_t(x)\|^2&=2\langle\widetilde{\xi}_t(\widetilde{\mathbb{T}}_t(y))-\widetilde{\xi}_t(\widetilde{\mathbb{T}}_t(x)),\widetilde{\mathbb{T}}_t(y)-\widetilde{\mathbb{T}}_t(x)\rangle\\ &\leq 2\max_{z\in\bar{\mathbb{T}}_t(\Omega_f)_\rho}\|D\widetilde{\xi}_t(z)\|_{HM}\|\widetilde{\mathbb{T}}_t(y)-\widetilde{\mathbb{T}}_t(x)\|^2.\end{aligned}$$

Then, Gronwall's Inequality gives

$$\|\widetilde{\mathbb{T}}_t(y)-\widetilde{\mathbb{T}}_t(x)\|\leq e^{\max_{z\in\bar{\mathbb{T}}_t(\Omega_f)_\rho}\|D\widetilde{\xi}_t(z)\|_{HM}}\|y-x\|.$$

We estimate the difference between $\bar{\mathbb{T}}$ and $\widetilde{\mathbb{T}}$. To this end, we notice that

$$\begin{aligned}\partial_t\|\bar{\mathbb{T}}_t(x)-\widetilde{\mathbb{T}}_t(x)\|_{L^2}^2&=2\langle\bar{\xi}_t(\bar{\mathbb{T}}_t(x))-\bar{\xi}_t(\widetilde{\mathbb{T}}_t(x))+\bar{\xi}_t(\widetilde{\mathbb{T}}_t(x))-\widetilde{\xi}_t(\widetilde{\mathbb{T}}_t(x)),\bar{\mathbb{T}}_t(x)-\widetilde{\mathbb{T}}_t(x)\rangle_{L^2}\\ &\leq 2\mathfrak{L}_n\|\bar{\mathbb{T}}_t(x)-\widetilde{\mathbb{T}}_t(x)\|_{L^2}^2+2\mathrm{err}_1\|\bar{\mathbb{T}}_t(\mathrm{x})-\widetilde{\mathbb{T}}_t(\mathrm{x})\|_{\mathrm{L}^2}.\end{aligned}$$

Hence,

$$\|\bar{\mathbb{T}}_t(x)-\widetilde{\mathbb{T}}_t(x)\|_{L^2}\leq\frac{\mathrm{err}_1}{\mathfrak{L}_n}\left(e^{\mathfrak{L}_nt}-1\right).$$

The solution of (47) has the explicit form $(\widetilde{\mathbb{T}}_t)_\#(\overline{f}e^{\int_0^t\bar{\zeta}_{t'}(\widetilde{\mathbb{T}}_{t'})dt'}\mathcal{L})$. Meanwhile, the solution of

$$\begin{cases}\partial_t\mu_t+\nabla_x(\bar{\xi}_t\mu_t)=\bar{\zeta}_t\mu_t,\\ \mu_0=\overline{f}\mathcal{L},\end{cases} \tag{48}$$

has the explicit form $(\bar{\mathbb{T}}_t)_\#(\overline{f}e^{\int_0^t\bar{\zeta}_{t'}(\bar{\mathbb{T}}_{t'})dt'}\mathcal{L})$. We take $\widetilde{f}=\overline{f}$ and let $\widetilde{g}\mathcal{L}$ be the solution of (47) at time T, let $\underline{g}\mathcal{L}$ be the solution of (48) at time T. We wish to show that $\int_{\widetilde{\mathbb{T}}_T(\Omega_f)}\varphi\widetilde{g}dx\to$

$\int_{\mathbb{T}_T(\Omega_f)} \underline{g}dx$ as $\mathrm{err}_1 \to 0^+$ for all bounded continuous function φ. By Portmanteau's Theorem, it is sufficient to consider φ as a bounded Lipschitz function. In such case, we estimate

$$\begin{aligned}&\left|\int_{\Omega_f} \overline{f}\left(\varphi(\bar{\mathbb{T}}_T(x))e^{\int_0^T \bar{\zeta}_t(\bar{\mathbb{T}}_t(x))dt} - \varphi(\widetilde{\mathbb{T}}_T(x))e^{\int_0^T \bar{\zeta}_t(\widetilde{\mathbb{T}}_t(x))dt}\right)dx\right|\\ &\le \left|\int_{\Omega_f} \overline{f}(\varphi(\bar{\mathbb{T}}_T(x)) - \varphi(\widetilde{\mathbb{T}}_T(x)))e^{\int_0^T \bar{\zeta}_t(\bar{\mathbb{T}}_t(x))dt}dx\right|\\ &\quad + \left|\int_{\Omega_f} \overline{f}\varphi(\widetilde{\mathbb{T}}_T(x))\left(e^{\int_0^T \bar{\zeta}_t(\bar{\mathbb{T}}_t(x))dt} - e^{\int_0^T \bar{\zeta}_t(\widetilde{\mathbb{T}}_t(x))dt}\right)dx\right|.\end{aligned}$$

Performing a similar computation with the proof of Theorem 9, we find

$$e^{\int_0^T \bar{\zeta}_t(\bar{\mathbb{T}}_t(x))dt} = \frac{2\tau - \widetilde{k}_1^*(x)}{2\tau - \widetilde{k}_2^*(\bar{\mathbb{T}}_T(x))}$$

is a bounded function. Hence,

$$\left|\int_{\Omega_f} \overline{f}(\varphi(\bar{\mathbb{T}}_T(x)) - \varphi(\widetilde{\mathbb{T}}_T(x)))e^{\int_0^T \bar{\zeta}_t(\bar{\mathbb{T}}_t(x))dt}dx\right| \lesssim \|\overline{f}\|_{L^2} Lip(\varphi)\|\bar{\mathbb{T}}_T(x) - \widetilde{\mathbb{T}}_T(x)\|_{L^2}.$$

Since $\widetilde{\zeta} \in C^1$, we also have the bound

$$\left|e^{\int_0^T \bar{\zeta}_t(\bar{\mathbb{T}}_t(x))dt} - e^{\int_0^T \bar{\zeta}_t(\widetilde{\mathbb{T}}_t(x))dt}\right| \lesssim \max|\nabla\bar{\zeta}|\int_0^T \left|\bar{\mathbb{T}}_t(x) - \widetilde{\mathbb{T}}_t(x)\right|dt.$$

This leads to

$$\left|\int_{\Omega_f} \overline{f}\varphi(\widetilde{\mathbb{T}}_T(x))\left(e^{\int_0^T \bar{\zeta}_t(\bar{\mathbb{T}}_t(x))dt} - e^{\int_0^T \bar{\zeta}_t(\widetilde{\mathbb{T}}_t(x))dt}\right)dx\right| \lesssim \|\overline{f}\|_{L^2}\max|\varphi|\max|\nabla\widetilde{\zeta}|\|\bar{\mathbb{T}}_T(x) - \widetilde{\mathbb{T}}_T(x)\|_{L^2}.$$

We take $\widetilde{f} = \overline{f}$ and let $\widetilde{g}\mathcal{L}$ be the solution of (47) at time T, let $\underline{g}\mathcal{L}$ be the solution of (48) at time T. We have proved that $\widetilde{g}\mathcal{L}$ converges weakly to $\underline{g}\mathcal{L}$ when $\mathrm{err}_1 \to 0^+$.

When $n \to +\infty$, we have $\nabla\bar{\phi}_n \to \nabla\phi$, and hence for $\underline{k}_y\mathcal{L} = (\nabla\bar{\phi}_n)_\#\bar{k}_x\mathcal{L}$ we get

$$W_2^2(\underline{k}_y\mathcal{L}, \bar{k}_y\mathcal{L}) \le \int_{\Omega_f} \|\nabla\bar{\phi}_n - \nabla\phi\|_{L^2}^2 \bar{k}_x dx \to 0,$$

where W_2 is the Wasserstein distance. Using [14, Theorem 6.9], we deduce that $\underline{k}_y\mathcal{L}$ converges weakly to $\bar{k}_y\mathcal{L}$ when $n \to +\infty$. Then, $\underline{g}\mathcal{L} = \frac{2\tau\underline{k}_y}{2\tau - \widetilde{k}_2^*}\mathcal{L}$ converges weakly to $\overline{g}\mathcal{L} = \frac{2\tau\bar{k}_y}{2\tau - \widetilde{k}_2^*}\mathcal{L}$ when $n \to +\infty$. When $\mathrm{err}_0 \to 0^+$ and $L \to +\infty$, $\overline{f}\mathcal{L}, \overline{g}\mathcal{L}$ converge to $f_\varepsilon\mathcal{L}, g_\varepsilon\mathcal{L}$ in L^2 by Proposition 9. For any bounded and continuous function φ, we have

$$\left|\int_{\Omega_f} \varphi(\overline{f} - f_\varepsilon)dx\right| \le \max|\varphi||\Omega_f|\|\overline{f} - f_\varepsilon\|_{L^2} \to 0 \quad \text{as } \mathrm{err}_0 \to 0^+, \mathrm{L} \to +\infty.$$

It means $\overline{f}\mathcal{L}$ converges weakly to $f_\varepsilon\mathcal{L}$ in measure sense as $\mathrm{err}_0 \to 0^+, \mathrm{L} \to +\infty$. Similarly, $\overline{g}\mathcal{L}$ converges weakly to $g_\varepsilon\mathcal{L}$ in measure sense as $\mathrm{err}_0 \to 0^+, \mathrm{L} \to +\infty$. Therefore, $\widetilde{f}\mathcal{L}, \widetilde{g}\mathcal{L}$ converge weakly to $f_\varepsilon\mathcal{L}, g_\varepsilon\mathcal{L}$ as $\mathrm{err}_0, \mathrm{err}_1 \to 0^+$ and $n, L \to +\infty$.

Appendix A. Proofs of the lemmas

A.1. **Proof of Lemma 1.** For $k \geq 0, k \in L^1(\Omega)$ and $\kappa > 0$, we consider

$$k^\kappa(x,y) = \min\{k(x,y), \kappa\}.$$

The function k^κ is non-negative and bounded above by κ. Since Ω is compact, $k^\kappa \in L^2_+(\Omega)$. We observe that

$$\int_\Omega C(x,y)k^\kappa(x,y)dxdy - \int_\Omega C(x,y)k(x,y)dxdy \leq 0,$$

and

$$\begin{aligned} F((k^\kappa)_x|f) - F(k_x|f) &= \int_{\Omega_f} ((k^\kappa)_x - k_x)\left(\frac{(k^\kappa)_x + k_x}{f} - 2\right)dx \\ &\leq \int_{\substack{\Omega_f \\ 2(k^\kappa)_x < (k^\kappa)_x + k_x < 2f}} ((k^\kappa)_x - k_x)\left(\frac{(k^\kappa)_x + k_x}{f} - 2\right)dx \\ &\leq \int_{\Omega_f} 2(k_x - (k^\kappa)_x)dx \end{aligned}$$

The Monotone Convergence Theorem implies that $\int_{\Omega_f} 2(k_x - (k^\kappa)_x)dx \to 0$ as $\kappa \to +\infty$. The quantity $F(k_y|g)$ can be estimated similarly. Hence, the infimum in (7) can also be approximated by $k \in L^2_+(\Omega)$.

A.2. **Proof of Lemma 2.** We first consider the transportation from $(f_\varepsilon - f)|_{\{x|f(x)<\varepsilon\}}$ to 0. The only transport plan is $\gamma = 0$ and we find

$$d_C((f_\varepsilon - f)|_{\{x|f(x)<\varepsilon\}}, 0) = \int_{\{x|f(x)<\varepsilon\}} (\varepsilon - f)dx \leq \varepsilon|\{x|f(x) < \varepsilon\}| = O(\varepsilon),$$

The same estimate can also be obtained for g_ε.

By [33, Corollary 4.13], we obtain the subadditivity of d_C, and thus we get

$$d_C(f_\varepsilon, g_\varepsilon) - d_C(f,g) \leq d_C((f_\varepsilon - f)|_{\{x|f(x)<\varepsilon\}}, 0) + d_C(0, (g_\varepsilon - g)|_{\{y|g(y)<\varepsilon\}}) = O(\varepsilon).$$

For $k \in L^2_+(\Omega)$, we observe that

$$F(k_x|f) - F(k_x|f_\varepsilon) = \int_{\Omega_f}\left(\frac{k_x^2}{f_\varepsilon f} + 1\right)(f - f_\varepsilon)\,dx \leq 0.$$

The quantity $F(k_y|g)$ can be similarly estimated. As a result, $d_C(f,g) \leq d_C(f_\varepsilon, g_\varepsilon)$. Therefore,

$$0 \leq d_C^\varepsilon(f,g) - d_C(f,g) \leq O(\varepsilon).$$

A.3. **Proof of Lemma 3.** Without loss of generality, we omit f, g in our proof. There exists $k^{[N]} \in L^2_+(\Omega)$ such that

$$d_C^\varepsilon + \frac{1}{2N} > \int_\Omega C(x,y)k^{[N]}(x,y)dxdy + \tau F(k_x^{[N]}|f_\varepsilon) + \tau F(k_y^{[N]}|g_\varepsilon).$$

For $\eta < \frac{1}{2N\|k^{[N]}\|_{L^2}^2}$ with the convenion that $1/\|k^{[N]}\|_{L^2}^2 = +\infty$ if $\|k^{[N]}\|_{L^2} = 0$, we have

$$d_C^\varepsilon + \frac{1}{N} > \int_\Omega C(x,y)k^{[N]}(x,y)dxdy + \eta\|k^{[N]}\|_{L^2}^2 + \tau F(k_x^{[N]}|f_\varepsilon) + \tau F(k_y^{[N]}|g_\varepsilon) \geq d_C^{\varepsilon,\eta}.$$

Note that for any $\eta > 0$,

$$\begin{aligned}\int_\Omega C(x,y)k(x,y)dxdy + \eta\|k\|_{L^2}^2 + \tau F(k_x|f_\varepsilon) + \tau F(k_y|g_\varepsilon) \\ \geq \int_\Omega C(x,y)k(x,y)dxdy + \tau F(k_x|f_\varepsilon) + \tau F(k_y|g_\varepsilon) \\ \geq d_C^\varepsilon, \qquad \forall k \in L^2_+(\Omega).\end{aligned}$$

Thus, $d_C^{\varepsilon,\eta} \geq d_C^\varepsilon$. Therefore, when we take $\eta_N = \frac{1}{2N\|k^{[N]}\|_{L^2}^2}$, we obtain the result of the lemma.

A.4. **Proof of Lemma 4.** We consider

$$\mathcal{O}(k) = \int_\Omega C(x,y)k(x,y)dxdy + \eta\|k(x,y)\|_{L^2}^2 + \tau F(k_x|f_\varepsilon) + \tau F(k_y|g_\varepsilon).$$

It is clear that $\mathcal{O}(k) \geq 0$. We also have $\mathcal{O}(0) = \|f_\varepsilon\|_{L^1} + \|g_\varepsilon\|_{L^1} < \infty$, so the infimum value of $\mathcal{O}$ is not $+\infty$. We take $\left\{k^{[n]}\right\}_{n\in\mathbb{N}} \subset L^2_+(\Omega)$ such that $\mathcal{O}(k^{[n]}) \to d_C^{\varepsilon,\eta}(f,g)$. Then, $\|k^{[n]}\|_{L^2}$ is bounded. Because $L^2(\Omega)$ is reflexive and $\{k^{[n]}\}$ bounded, by [22, Theorem 5.18], there exists a convergent subsequence $\left\{k^{[n_j]}\right\}_{j\in\mathbb{N}}$ under weak topology of $L^2(\Omega)$. We denote the limit by $k^{[0]} \in L^2_+(\Omega)$.

To show that $k^{[0]}$ is a minimizer, it suffices to show that $\mathcal{O}(k)$ is lower semi-continuous. With a bit of abusing the notations we assume $k^{[n]}$ weakly converges to $k^{[0]}$ in $L^2(\Omega)$. We take

$$C_m(x,y) = \inf_{(x_0,y_0)\in\Omega} \left\{C(x_0,y_0) + m\left(\|x-x_0\|_2 + \|y-y_0\|_2\right)\right\}.$$

It is clear that C_m is an increasing sequence in the pointwise sense and $C_m(x,y) \leq C(x,y)$. Since C is lower semi-continuous and bounded from below, for each (x,y) and ϵ, there exists ρ_ϵ such that $C(x_1,y_1) > C(x,y) - \epsilon$ for all $\|x-x_1\|_2 + \|y-y_1\|_2 < \rho_\epsilon$. For $m > \frac{C(x,y)+\epsilon}{\rho_\epsilon}$, we consider (x_2,y_2) such that $C_m(x,y) + \epsilon > C(x_2,y_2) + m\left(\|x-x_2\|_2 + \|y-y_2\|_2\right)$, then $\|x-x_2\|_2 + \|y-y_2\|_2 < \rho_\epsilon$. Now, we see that

$$C_m(x,y) \geq C(x_2,y_2) + n\left(\|x-x_2\|_2 + \|y-y_2\|_2\right) > C(x,y) - \varepsilon + m\left(\|x-x_2\|_2 + \|y-y_2\|_2\right).$$

If $x_2 = x, y_2 = y$, then $C_m(x,y) > C(x,y) - \epsilon$. If $\|x-x_2\|_2 + \|y-y_2\|_2 > 0$ then we can pick m sufficiently large and we get $C_n(x,y) > C(x,y)$, which is a contradiction. This means that $C_m(x,y) \to C(x,y)$ as $m \to +\infty$. For $x, z \in \Omega_f, y, w \in \Omega_g$, we have

$$\begin{aligned}C(x_0,y_0) + m\left(\|x-x_0\|_2 + \|y-y_0\|_2\right) \leq C(x_0,y_0) + m\left(\|z-x_0\|_2 + \|w-y_0\|_2\right) \\ + m\left(\|x-z\|_2 + \|y-w\|_2\right).\end{aligned}$$

This leads to $C_m(x,y) - C_m(z,w) \leq m\left(\|x-z\|_2 + \|y-w\|_2\right)$. By switching (x,y) and (z,w) we get $|C_m(x,y) - C_m(z,w)| \leq m\left(\|x-z\|_2 + \|y-w\|_2\right)$. Therefore, C_m is continous. By the properties of the weak topology, the monotonicity of C_m in m, and the Monotone Convergence Theorem, we have

$$\liminf_n \int_\Omega Ck^{[n]}dxdy \geq \liminf_{n\to\infty} \int_\Omega C_m k^{[n]}dxdy = \int_\Omega C_m k^{[0]}dxdy \overset{m\to\infty}{\longrightarrow} \int_\Omega Ck^{[0]}dxdy.$$

Hence, $\int_\Omega Ckdxdy$ is lower semi-continuous.

It also follows that the norm $\|k\|_{L^2}$ is lower semi-continuous under the weak topology. Thus,

$$\liminf_n \|k^{[n]}\|_{L^2}^2 = (\liminf_n \|k^{[n]}\|_{L^2})^2 \geq \|k^{[0]}\|_{L^2}^2.$$

Finally, we will study the entropic marginal cost. Since $k^{[n]}$ weakly converges to $k^{[0]}$ in $L^2(\Omega)$, for $\psi \in L^2(\Omega_f)$, we have

$$\lim_n \int_{\Omega_f} \psi \frac{k_x^{[n]}}{\sqrt{f_\varepsilon}} dx = \lim_n \int_\Omega \frac{\psi}{\sqrt{f_\varepsilon}} k_x^{[n]} dxdy = \int_\Omega \frac{\psi}{\sqrt{f_\varepsilon}} k_x^0 dxdy = \int_{\Omega_f} \psi \frac{k_x^{[0]}}{\sqrt{f_\varepsilon}} dx.$$

This means $\frac{k_x^{[n]}}{\sqrt{f_\varepsilon}}$ weakly converges to $\frac{k_x^{[0]}}{\sqrt{f_\varepsilon}}$ in $L^2(\Omega_f)$. Similarly, we also have $k_x^{[n]}$ weakly converges to k_x^0 in $L^2(\Omega_f)$. Then, we find

$$\liminf_n \left\| \frac{k_x^{[n]}}{\sqrt{f_\varepsilon}} \right\|_{L^2}^2 = \left(\liminf_n \left\| \frac{k_x^{[n]}}{\sqrt{f_\varepsilon}} \right\|_{L^2} \right)^2 \geq \left\| \frac{k_n^{[0]}}{\sqrt{f_\varepsilon}} \right\|_{L^2}^2.$$

As a consequence, we get

$$\begin{aligned} \liminf_n F(k_x^{[n]}|f_\varepsilon) &= \liminf_n \int_\Omega \left(\frac{(k_2^{[n]})^2}{f_\varepsilon} - 2k_x^{[n]} + f_\varepsilon \right) dx \\ &\geq \int_\Omega \left(\frac{(k_x^{[0]})^2}{f_\varepsilon} - 2k_x^{[0]} + f_\varepsilon \right) dx \;=\; F(k_x^{[0]}|f_\varepsilon). \end{aligned}$$

The fact that marginal entropic cost $F(k_y|g_\varepsilon)$ is lower semi-continuous can be proved in a similar way. The function $\mathcal{O}$ is then lower semi-continuous as it is the sum of lower semi-continuous functions.

Remark 15. We want to be careful with the topology being used here. The statement "$\int_\Omega Ckdxdy$ is lower semi-continuous" is similar to a part of [14, Lemma 4.3], and "$F(k_x|f_\varepsilon)$ is lower semi-continuous" is similar to a part of [33, Corollary 2.9]. But, we are working with the weak topology of L^2 and the quoted statements are for the weak topology of measures.

A.5. **Proof of Lemma 5.** Let $k \in L_+^2(\Omega)$ be the minimizer of (11). For $\delta > 0$, we consider

$$k_{(\delta)}(x,y) = \begin{cases} \max\{k(x,y), \delta\} & \text{if } (x,y) \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

On Ω, C is bounded and hence $0 \leq \int_\Omega Ck_{(\delta)}dxdy - \int_\Omega Ckdxdy \leq (\sup_{(x,y)\in\Omega} C)\delta|\Omega| = O(\delta)$.

We estimate the difference

$$\begin{aligned} \big|\|k_{(\delta)}(x,y)\|_{L^2}^2 - \|k(x,y)\|_{L^2}^2\big| &\leq \|k_{(\delta)}(x,y) - k(x,y)\|_{L^2}(\|k_{(\delta)}(x,y) - k(x,y)\|_{L^2} + 2\|k(x,y)\|_{L^2}) \\ &\leq \delta|\Omega|(\delta|\Omega| + 2\|k(x,y)\|_{L^2}) = O(\delta). \end{aligned}$$

We also estimate the difference of the entropic cost

$$\big|F((k_{(\delta)})_x|f_\varepsilon) - F(k_x|f_\varepsilon)\big| = \left| \int_{\Omega_f} ((k_{(\delta)})_x - k_x) \left(\frac{1}{f_\varepsilon} - 2 \right) dx \right| \leq \delta|\Omega| \left(2 + \frac{1}{\varepsilon} \right) = O(\delta).$$

The conclusion of the lemma can be obtained by summing up the differences.

A.6. **Proof of Lemma 6.**

A.6.1. *Minimizer of* (13). The proofs for the claim that Problem (13) admits a minimizer is the same as the proof of Lemma 4 with L^2_δ replacing L^2_+.

A.6.2. *Proof of* (18). For $v\mathcal{L} \in \mathcal{M}^{\mathcal{L}}_{l,c}(\mathbb{R}^d)$, $v(x) > 0, \forall x \in \operatorname{supp} v$, and $u^* \in L^2(\operatorname{supp} v)$, we compute

$$\begin{aligned}
F^*_\theta(u^*|v) &= \sup_{u\in L^2_\theta(\operatorname{supp} v)} \left\{ \int_{\operatorname{supp} v} uu^* dx - F(u|v) \right\} \\
&= \sup_{\substack{u\in L^2_\theta(\operatorname{supp} v)\\ u(x)=0 \text{ if } v(x)=0}} \left\{ \int_{\operatorname{supp} v} \left[\frac{-u^2}{v} + u\,(u^*+2) - v \right] dx \right\} \\
&= \int_{\operatorname{supp} v} v \left[\max\left\{ \frac{\theta}{v}, \frac{u^*}{2} + 1 \right\} \left(u^* + 2 - \max\left\{ \frac{\theta}{v}, \frac{u^*}{2} + 1 \right\} \right) - 1 \right] dx.
\end{aligned}$$

Combining with the regularized parameter, we get

$$\begin{aligned}
(\tau F)^*_\theta(u^*|v) &= \tau F^*_\theta(u^*/\tau|v) \\
&= \tau \int_{\operatorname{supp} v} v \left[\max\left\{ \frac{\theta}{v}, \frac{u^*}{2\tau} + 1 \right\} \left(\frac{u^*}{\tau} + 2 - \max\left\{ \frac{\theta}{v}, \frac{u^*}{2\tau} + 1 \right\} \right) - 1 \right] dx.
\end{aligned} \tag{49}$$

The maximum solution u_0 in the convex conjugate is given by

$$u_0(x) = v \max\left\{ \frac{\theta}{v}, \frac{u^*(x)}{2\tau} + 1 \right\}$$

for almost everywhere x in $\operatorname{supp} v$.

Next, we compute

$$\begin{aligned}
\bar{C}^*_{\eta,\delta}(k^*) &= \sup_{k\in L^2_\delta(\Omega)} \int_\Omega \left[k\,(k^* - C) - \eta k^2 \right] dxdy \\
&= \eta \int_\Omega \max\left\{ \delta, \frac{k^* - C}{2\eta} \right\} \left(\frac{k^* - C}{\eta} - \max\left\{ \delta, \frac{k^* - C}{2\eta} \right\} \right) dxdy.
\end{aligned} \tag{50}$$

The maximum solution k_0 in the convex conjugate is given by

$$k_0 = \max\left\{ \delta, \frac{k^* - C}{2\eta} \right\},$$

for almost everywhere (x, y) in Ω.

From (14), we immediately get

$$-D^{\varepsilon,\eta}_{\delta,C}(f,g) = \bar{D}^{\varepsilon,\eta}_{\delta,C}(f,g) :== \sup_{\substack{k_1^*\in L^2(\Omega_f)\\ k_2^*\in L^2(\Omega_g)}} \left\{ -\bar{C}^*_{\eta,\delta}(k_1^*(x) + k_2^*(y)) - \tau F^*_{\delta|\Omega_g|}(-k_1^*/\tau|f_\varepsilon) - \tau F^*_{\delta|\Omega_f|}(-k_2^*/\tau|g_\varepsilon) \right\}.$$

We want to show that $\bar{D}^{\varepsilon,\eta}_{\delta,C}(f,g) = d^{\varepsilon,\eta}_{\delta,C}(f,g)$.

We consider the following quantity

$$\begin{aligned}
L(k, k_1^*, k_2^*) = \int_\Omega C(x,y)k(x,y)dxdy + \eta\|k\|^2_{L^2} - \int_\Omega k(x,y)(k_1^*(x) + k_2^*(y))dxdy \\
- \tau F^*_{\delta|\Omega_g|}(-k_1^*/\tau|f_\varepsilon) - \tau F^*_{\delta|\Omega_f|}(-k_2^*/\tau|g_\varepsilon).
\end{aligned}$$

For a domain $\mho \subset \mathbb{R}^d$, we define the convex and lower semi-continuous indicator $I_{L^2_\theta(\mho)} : L^2(\mho) \to \mathbb{R} \cup \{\infty\}$ as

$$I_{L^2_\theta(\mho)}(u) := \begin{cases} 0 & \text{if } u \in L^2_\theta(\mho), \\ \infty & \text{otherwise.} \end{cases}$$

By [33, Corollary 2.9], $F(\cdot|v)$ is convex, and by the proof of Lemma 4, $F(\cdot|v)$ is lower semi-continuous. Applying [25, Proposition 4.1 Chapter 1 Part 1], we obtain

$$\begin{aligned} I_{L^2_\theta(\operatorname{supp} v)} + \tau F(u|v) &= (I_{L^2_\theta(\operatorname{supp} v)} + \tau F(u|v))^{**} \\ &= \sup_{u^* \in L^2(\operatorname{supp} v)} \left\{ \int_{\operatorname{supp} v} u u^* dx - \tau F^*_\theta(u^*/\tau|v) \right\}. \end{aligned}$$

We substitute $\theta = \delta|\Omega_g|, \delta|\Omega_f|$, $u = k_x, k_y$ and $v = f_\varepsilon, g_\varepsilon$, respectively and find

$$d^{\varepsilon,\eta}_{\delta,C}(f,g) = \inf_{k \in L^2_\delta(\Omega)} \sup_{\substack{k_1^* \in L^2(\Omega_f) \\ k_2^* \in L^2(\Omega_g)}} L(k, k_1^*, k_2^*).$$

On the other hand, by definition, we have

$$\bar{D}^{\varepsilon,\eta}_{\delta,C}(f,g) = \sup_{\substack{k_1^* \in L^2(\Omega_f) \\ k_2^* \in L^2(\Omega_g)}} \inf_{k \in L^2_\delta(\Omega)} L(k, k_1^*, k_2^*).$$

We can see that $d^{\varepsilon,\eta}_{\delta,C}(f,g) \geq \bar{D}^{\varepsilon,\eta}_{\delta,C}(f,g)$. To obtain the desired equality, we apply the Minimax Duality Theorem (see [33, Theorem 2.4] for the statement, or [38, Theorem 3.1] for the proof). To this end, we need to check that $L(\cdot,\cdot,\cdot)$ satisfies all the needed conditions.

(1) For fixed $k_1^* \in L^2(\Omega_f), k_2^* \in L^2(\Omega_g)$, the function $L(\cdot, k_1^*, k_2^*)$ is convex and lower semi-continuous under weak topology in $L^2(\Omega)$. In Lemma 4, we already proved that $\int_\Omega Ck dxdy + \eta \|k\|^2_{L^2}$ is lower semi-continuous. We also have $\int_\Omega k(k_1^* + k_2^*) dxdy$ is continuous under the weak topology. The convexity of $L(\cdot, k_1^*, k_2^*)$ follows from the convexity of $\|k\|^2_{L^2}$ and the linearity of $\int_\Omega Ck dxdy, \int_\Omega k(k_1^* + k_2^*) dxdy$.

(2) For fixed $k \in L^2_\delta(\Omega)$, the function $L(k, \cdot)$ is concave in $L^2(\Omega_f) \times L^2(\Omega_g)$. This follows from the linearity of $\int_\Omega k(k_1^* + k_2^*) dxdy$ and the convexity of $F^*_\theta(\cdot|v)$ (see [25, Definition 4.1 Chapter 1 Part 1]).

(3) There are $M > \bar{D}^{\varepsilon,\eta}_{\delta,C}(f,g), k_1^\star \in L^2(\Omega_f), k_2^\star \in L^2(\Omega_g)$ such that

$$\left\{ k \in L^2_\delta(\Omega) | L(k, k_1^\star, k_2^\star) \leq M \right\} \text{ is compact under weak topology of } L^2(\Omega).$$

By choosing $M = d^{\varepsilon,\eta}_{\delta,C}(f,g) + 1$ and $k_1^\star = k_2^\star = 0$, for $k \in L^2_\delta(\Omega)$ and $L(k, k_1^\star, k_2^\star) \leq M$, we have $\|k\|_{L^2} \leq \sqrt{\frac{M}{\eta}}$. By Kakutani's Theorem, $\{k \in L^2_\delta(\Omega) | L(k, k_1^\star, k_2^\star) \leq M\}$ is a subset of a compact set. The set is also a closed set as it is a sublevel set of the lower semi-continuous function $L(\cdot, k_1^\star, k_2^\star)$. Hence, the set $\{k \in L^2_\delta(\Omega) | L(k, k_1^\star, k_2^\star) \leq M\}$ is compact.

Thus, we are done proving $d^{\varepsilon,\eta}_{\delta,C}(f,g) + D^{\varepsilon,\eta}_{\delta,C}(f,g) = 0$.

A.6.3. *Equations* (15) - (17). We take a sequence $\left\{((k_1^*)^{[n]}, (k_2^*)^{[n]})\right\}_{n\in\mathbb{N}} \subset L^2(\Omega_f)\times L^2(\Omega_g)$ such that

$$d^{\varepsilon,\eta}_{\delta,C}(f,g)+\bar{C}^*_{\eta,\delta}((k_1^*)^{[n]},(k_2^*)^{[n]})+\tau F^*_{\delta|\Omega_g|}(-(k_1^*)^{[n]}/\tau|f_\varepsilon)+\tau F^*_{\delta|\Omega_f|}(-(k_2^*)^{[n]}/\tau|g_\varepsilon)\to 0^+,$$

as $n\to\infty$. Let $k\in L^2_\delta(\Omega)$ be the minimizer of (13), we estimate

$$\begin{aligned}
&\int_\Omega Ckdxdy+\eta\|k\|^2_{L^2}-\int_\Omega k((k_1^*)^{[n]}+(k_2^*)^{[n]})dxdy+\bar{C}^*_{\eta,\delta}((k_1^*)^{[n]}+(k_2^*)^{[n]})\\
&=\eta\int_\Omega\left(k-\max\left\{\delta,\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}\right\}\right)\\
&\times\left(k+\max\left\{\delta,\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}\right\}-\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{\eta}\right)dxdy\\
&=\eta\int_\Omega\left(k-\max\left\{\delta,\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}\right\}\right)^2dxdy\\
&+2\eta\int_\Omega\left(k-\max\left\{\delta,\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}\right\}\right)\\
&\times\left(\max\left\{\delta,\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}\right\}-\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}\right)dxdy\\
&\geq\eta\left\|k-\max\left\{\delta,\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}\right\}\right\|^2_{L^2}.
\end{aligned}$$

We obtain the desired inequality simply by considering $\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}\geq\delta$ or $\frac{(k_1^*)^{[n]}+(k_2^*)^{[n]}-C}{2\eta}<\delta\leq k$. Similarly, we also have

$$\begin{aligned}
&\tau F(k_x|f_\varepsilon)+\int_{\Omega_f}k_x(k_1^*)^{[n]}dx+\tau F^*_{\delta|\Omega_g|}(-(k_1^*)^{[n]}/\tau|f_\varepsilon)\\
&\quad=\tau\int_{\Omega_f}f_\varepsilon\left(\frac{k_x}{f_\varepsilon}-\max\left\{\frac{\delta|\Omega_g|}{f_\varepsilon},1-\frac{(k_1^*)^{[n]}}{2\tau}\right\}\right)^2dx\\
&\quad+\tau\int_{\Omega_f}f_\varepsilon\left(\frac{k_x}{f_\varepsilon}-\max\left\{\frac{\delta|\Omega_g|}{f_\varepsilon},1-\frac{(k_1^*)^{[n]}}{2\tau}\right\}\right)\\
&\quad\times\left(\max\left\{\frac{\delta|\Omega_g|}{f_\varepsilon},1-\frac{(k_1^*)^{[n]}}{2\tau}\right\}-\left(1-\frac{(k_1^*)^{[n]}}{2\tau}\right)\right)dx\\
&\quad\geq\tau\varepsilon\left\|\frac{k_x}{f_\varepsilon}-\max\left\{\frac{\delta|\Omega_g|}{f_\varepsilon},1-\frac{(k_1^*)^{[n]}}{2\tau}\right\}\right\|^2_{L^2},
\end{aligned}$$

and

$$\begin{aligned}
&\tau F(k_y|g_\varepsilon) + \int_{\Omega_g} k_y (k_2^*)^{[n]} dy + \tau F^*_{\delta|\Omega_f|}(-(k_2^*)^{[n]}/\tau | g_\varepsilon) \\
&= \tau \int_{\Omega_g} g_\varepsilon \left(\frac{k_y}{g_\varepsilon} - \max\left\{ \frac{\delta|\Omega_f|}{g_\varepsilon}, 1 - \frac{(k_2^*)^{[n]}}{2\tau} \right\} \right)^2 dy \\
&+ \tau \int_{\Omega_g} g_\varepsilon \left(\frac{k_y}{g_\varepsilon} - \max\left\{ \frac{\delta|\Omega_f|}{g_\varepsilon}, 1 - \frac{(k_2^*)^{[n]}}{2\tau} \right\} \right) \\
&\times \left(\max\left\{ \frac{\delta|\Omega_f|}{g_\varepsilon}, 1 - \frac{(k_2^*)^{[n]}}{2\tau} \right\} - \left(1 - \frac{(k_2^*)^{[n]}}{2\tau} \right) \right) dy \\
&\geq \tau\varepsilon \left\| \frac{k_y}{g_\varepsilon} - \max\left\{ \frac{\delta|\Omega_f|}{g_\varepsilon}, 1 - \frac{(k_2^*)^{[n]}}{2\tau} \right\} \right\|_{L^2}^2 .
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\max\left\{ \delta, \frac{(k_1^*)^{[n]} + (k_2^*)^{[n]} - C}{2\eta} \right\} &\xrightarrow{L^2} k, \\
\max\left\{ \frac{\delta|\Omega_g|}{f_\varepsilon}, 1 - \frac{(k_1^*)^{[n]}}{2\tau} \right\} &\xrightarrow{L^2} \frac{k_x}{f_\varepsilon}, \\
\max\left\{ \frac{\delta|\Omega_f|}{g_\varepsilon}, 1 - \frac{(k_2^*)^{[n]}}{2\tau} \right\} &\xrightarrow{L^2} \frac{k_y}{g_\varepsilon}.
\end{aligned}$$

If a minimizer of (14) exists, then (15), (16), (17) will follow immediately.

A.6.4. *Minimizer of* (14). We set

$$(k_1^\star)^{[n]} = \min\left\{ (k_1^*)^{[n]}, 2\tau\left(1 - \frac{\delta|\Omega_g|}{f_\varepsilon}\right) \right\}, (k_2^\star)^{[n]} = \min\left\{ (k_2^*)^{[n]}, 2\tau\left(1 - \frac{\delta|\Omega_f|}{g_\varepsilon}\right) \right\}.$$

We get

$$1 - \frac{(k_1^\star)^{[n]}}{2\tau} = \max\left\{ \frac{\delta|\Omega_g|}{f_\varepsilon}, 1 - \frac{(k_1^*)^{[n]}}{2\tau} \right\}, 1 - \frac{(k_2^\star)^{[n]}}{2\tau} = \max\left\{ \frac{\delta|\Omega_f|}{g_\varepsilon}, 1 - \frac{(k_2^*)^{[n]}}{2\tau} \right\}.$$

Using the L^2 convergence, $(k_1^\star)^{[n]}, (k_2^\star)^{[n]}$ is bounded in L^2 and both

$$\begin{aligned}
&\tau F(k_x|f_\varepsilon) + \int_{\Omega_f} k_x (k_1^\star)^{[n]} dx + \tau F^*_{\delta|\Omega_g|}(-(k_1^\star)^{[n]}/\tau | f_\varepsilon), \\
&\tau F(k_y|g_\varepsilon) + \int_{\Omega_g} k_y (k_2^\star)^{[n]} dx + \tau F^*_{\delta|\Omega_f|}(-(k_2^\star)^{[n]}/\tau | g_\varepsilon)
\end{aligned}$$

converge to 0 as $n \to +\infty$. Also by the convergence, for fixed $M > 0$, there is N_M such that $\|(k_1^\star)^{[n]}\|_{L^2} \leq 2\tau\left(\varepsilon^{-1}\|k_x\|_{L^2} + 1 + M\right)$ and $\|(k_2^\star)^{[n]}\|_{L^2} \leq 2\tau\left(\varepsilon^{-1}\|k_y\|_{L^2} + 1 + M\right)$ for $n \geq N_M$.

For x, y satisfying $k(x,y) = \delta$, we have

$$0 \leq \max\left\{\delta, \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta}\right\} - k \leq \max\left\{\delta, \frac{(k_1^*)^{[n]} + (k_2^*)^{[n]} - C}{2\eta}\right\} - k \tag{51}$$

and

$$\left(k - \max\left\{\delta, \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta}\right\}\right)\left(\max\left\{\delta, \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta}\right\} - \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta}\right) = 0. \tag{52}$$

Thus, for $\int_\Omega Ck + \eta\|k\|^2 - \int_\Omega k((k_1^\star)^{[n]} + (k_2^\star)^{[n]}) + \bar{C}^*_{\eta,\delta}((k_1^\star)^{[n]} + (k_2^\star)^{[n]})$, we only need to consider the case $k(x,y) > \delta$. We set $\Omega^k = \{(x,y) \in \Omega : k(x,y) > \delta\}$. By Egoroff's Theorem, for $\lambda_1 \in \mathbb{Z}^+$, there exists a measurable set $\mathcal{N}_1^{\lambda_1} \subset \pi_x(\Omega^k)$ such that $|\mathcal{N}_1^{\lambda_1}| < 1/\lambda_1$ and

$$\max\left\{\delta|\Omega_g|, 1 - \frac{(k_1^*)^{[n]}}{2\tau}\right\} \longrightarrow k_x \text{ uniformly on } \pi_x(\Omega^k) \setminus \mathcal{N}_1^{\lambda_1}.$$

If $k_x(x) \geq \delta|\Omega_g| + 1/\lambda_0$ for some $\lambda_0 > 0$, for sufficiently large n,

$$1 - \frac{(k_1^*)^{[n]}}{2\tau} \geq \delta|\Omega_g| + \frac{1}{2\lambda_0} \text{ on } \pi_x(\Omega^k) \setminus \mathcal{N}_1^{\lambda_1}.$$

Hence, there is n_{λ_1} such that $(k_1^\star)^{[n]} = (k_1^*)^{[n]}$ for $n \geq n_{\lambda_1}$ on $\pi_x(\Omega^k) \setminus \mathcal{N}_1^{\lambda_1}$. Similarly, there exists m_{λ_2} such that $(k_2^\star)^{[n]} = (k_2^*)^{[n]}$ for $n \geq m_{\lambda_2}$ on $\pi_y(\Omega^k) \setminus \mathcal{N}_2^{\lambda_2}$ for $\lambda_2 \in \mathbb{Z}^+$ and a measurable set $\mathcal{N}_2^{\lambda_2} \subset \pi_y(\Omega^k) : |\mathcal{N}_2^{\lambda_2}| < \lambda_2$. We note that in [22, Theorem 2.33], the set $\mathcal{N}^\lambda$ canbe taken so that as λ increase, $\mathcal{N}^\lambda$ decrease in the inclusion sense.

We now study $\mathcal{N} = \{(x,y) \in \Omega^k : x \in \mathcal{N}_1^{\lambda_1} \text{ or } y \in \mathcal{N}_2^{\lambda_2} \text{ or } k_x(x) < \delta|\Omega_g| + \frac{1}{\lambda_0} \text{ or } k_y(y) < \delta|\Omega_f| + \frac{1}{\lambda_0}\}$. For $(x,y) \in \mathcal{N} : \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta} \geq k$, we come back to (51) and (52). For $(x,y) \in \mathcal{N}$ and $\frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta} < k$, we have

$$\left(k - \max\left\{\delta, \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta}\right\}\right)\left(k + \max\left\{\delta, \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta}\right\} - \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{\eta}\right)$$
$$\leq 2k\left(k - \frac{(k_1^\star)^{[n]} + (k_2^\star)^{[n]} - C}{2\eta}\right).$$

By Cauchy-Schwarz's inequality, the integration on $\mathcal{N}$ is bounded by

$$\int_{\mathcal{N}}\left(k - \max\left\{\delta, \frac{(k_1^*)^{[n]} + (k_2^*)^{[n]} - C}{2\eta}\right\}\right)^2$$
$$+ 4\sqrt{\int_{\mathcal{N}} k^2(\|k\|_{L^2}^2 + \frac{1}{2\eta}(8\tau^2(\varepsilon^{-1}\max\{\|k_x\|_{L^2}, \|k_y\|_{L^2}\} + 1 + M)^2 + (\|C\|_{L^\infty}|\Omega|)^2))}$$
$$= \int_{\mathcal{N}}\left(k - \max\left\{\delta, \frac{(k_1^*)^{[n]} + (k_2^*)^{[n]} - C}{2\eta}\right\}\right)^2 + 4\sqrt{\int_{\mathcal{N}} k^2\mathcal{K}}.$$

Summing the cases, we get

$$\int_\Omega Ckdxdy + \eta\|k\|_{L^2}^2 - \int_\Omega k((k_1^\star)^{[n]} + (k_2^\star)^{[n]})dxdy + \bar{C}^*_{\eta,\delta}((k_1^\star)^{[n]} + (k_2^\star)^{[n]})$$
$$\leq \int_\Omega Ckdxdy + \eta\|k\|_{L^2}^2 - \int_\Omega k((k_1^*)^{[n]} + (k_2^*)^{[n]}) + \bar{C}^*_{\eta,\delta}((k_1^*)^{[n]} + (k_2^*)^{[n]}) + 4\sqrt{\int_{\mathcal{N}} k^2\mathcal{K}}$$

for $n \geq \max\{N_M, n_{\lambda_1}, m_{\lambda_2}\}$. As we increase $\lambda_0, \lambda_1, \lambda_2$, the quantity $\mathcal{N}$ decreases in the inclusion sense and $|\mathcal{N}| \to 0^+$. By the Monotone Convergence Theorem, $\int_{\mathcal{N}} k^2 \to 0^+$, which implies $4\sqrt{\int_{\mathcal{N}} k^2\mathcal{K}} \to 0^+$. Thus, by choosing $\lambda_0, \lambda_1, \lambda_2$ sufficiently large and n also sufficiently large, we get

$$\int_\Omega Ckdxdy + \eta\|k\|_{L^2}^2 - \int_\Omega k((k_1^\star)^{[n]} + (k_2^\star)^{[n]})dxdy + \bar{C}^*_{\eta,\delta}((k_1^\star)^{[n]} + (k_2^\star)^{[n]}) \to 0^+.$$

This means $((k_1^\star)^{[n]}, (k_2^\star)^{[n]})$ is a bounded sequence minimizing (14). By [22, Theorem 18], $((k_1^\star)^{[n]}, (k_2^\star)^{[n]})$ has a convergent subsequence in weak topology of L^2. Since $\bar{C}^*_{\eta,\delta}, F^*_\theta$ are lower semi-continuous in the weak topology (see [25, Definition 4.1 Chapter 1 Part 1]), the problem (14) admits a minimizer.

To finish the proof, we check that $|\mathcal{N}|$, in fact, converges to 0 as $\lambda_0, \lambda_1, \lambda_2 \to \infty$. We have

$$\mathcal{N} \subset \{(x,y) \in \Omega^k : x \in \mathcal{N}_1^{\lambda_1}\} \cup \{(x,y) \in \Omega^k : y \in \mathcal{N}_2^{\lambda_2}\}$$
$$\cup \{(x,y) \in \Omega^k : k_x(x) < \delta|\Omega_g| + \frac{1}{\lambda_0}\} \cup \{(x,y) \in \Omega^k : k_y(y) < \delta|\Omega_f| + \frac{1}{\lambda_0}\}.$$

For the first and second sets in the union, we estimate

$$\left|\left\{(x,y) \in \Omega^k : x \in \mathcal{N}_1^{\lambda_1}\right\}\right| \leq |\mathcal{N}_1^{\lambda_1}||\Omega_g| \to 0^+,$$
$$\left|\left\{(x,y) \in \Omega^k : y \in \mathcal{N}_2^{\lambda_2}\right\}\right| \leq |\mathcal{N}_2^{\lambda_2}||\Omega_f| \to 0^+.$$

For the third one, we notice that

$$\left\{(x,y) \in \Omega^k : k_x(x) = \delta|\Omega_g|\right\} = \bigcap_{\lambda_0} \left\{(x,y) \in \Omega^k : k_x(x) < \delta|\Omega_g| + \frac{1}{\lambda_0}\right\}.$$

For $k_x(x) = \delta|\Omega_g|$, we have $k(x,y) = \delta$ for almost everywhere $y \in \Omega_g$. This implies the intesection is of measure 0. By [22, Theorem 1.8], we obtain

$$\left|\left\{(x,y) \in \Omega^k : k_x(x) < \delta|\Omega_g| + \frac{1}{\lambda_0}\right\}\right| \to 0^+.$$

The fourth set can be similarly estimated.

References

[1] A. Alcalde, G. Fantuzzi, and E. Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. *arXiv preprint arXiv:2407.01602*, 2024.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. *International conference on machine learning*, 2017.

[3] P. M. Baggenstoss. Trainable compound activation functions for machine learning. *EUSIPCO*, 2022.

[4] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Two numerical methods for the elliptic Monge-Ampère equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(4):737–758, 2010.
[5] J.-D. Benamou, B. D Froese, and A. M. Oberman. Numerical solution of the optimal transportation problem using the monge–ampère equation. *Journal of Computational Physics*, 260:107–126, 2014.
[6] R. J. Berman. The sinkhorn algorithm, parabolic optimal transport and geometric monge–ampère equations. *Numerische Mathematik*, 145(4):771–836, 2020.
[7] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.
[8] S. Brenner, L.-Y. Sung, Z. Tan, and H. Zhang. A nonlinear least-squares convexity enforcing co interior penalty method for the monge–ampère equation on strictly convex smooth planar domains. *Communications of the American Mathematical Society*, 4(14):607–640, 2024.
[9] C.Calcaterra and A.Boldt. Approximating with gaussians. *arXiv preprint arXiv:0805.3795*, 2008.
[10] L. Chapel, R. Flamary, H. Wu, C. Févotte, and G. Gasso. Unbalanced optimal transport through non-negative penalized linear regression. *Advances in Neural Information Processing Systems*, 34:23270–23282, 2021.
[11] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.
[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
[13] C.Villani. *Topics in Optimal Transportation*, volume 58. Graduate Studies in Mathematics, 2003.
[14] C.Villani. *Optimal Transport: Old and New*. Springer Berlin, Heidelberg, 2008.
[15] D.Cordero-Erausquin, W.Gangbo, and C.Houdré. Inequalities for generalized entropy and optimal transportation. *Contemp. Math.*, 353, 05 2003.
[16] D.Funaro and O.Kavian. Approximation of some diffusion evolution equations in unbounded domains by hermite functions. *Mathematics of Computation - Math. Comput.*, 57:597–619, 10 1991.
[17] D.Ruiz-Balet and E.Zuazua. Neural ode control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.
[18] D.Ruiz-Balet and E.Zuazua. Control of neural transport for normalising flows. *Journal de Mathématiques Pures et Appliquées*, 181:58–90, 2024.
[19] K. Fatras, T. Séjourné, R. Flamary, and N. Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.
[20] F.Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[21] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a wassenstein loss. *Advances in Neural Information Processing Systems*, 2015.
[22] G.B.Folland. *Real Analysis: Modern Techniques and Their Application, 2nd Edition*. John Wiley & Sons, 1999.
[23] P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes. Obtaining fairness using optimal transport theory. *International Conference on Machine Learning*, 2019.
[24] M. Hernández and E. Zuazua. Deep neural networks: Multi-classification and universal approximation. *arXiv preprint arXiv:2409.06555*, 2024.
[25] R. Témam I. Ekeland. *Convex analysis and variational problems*. Society for Industrial and Applied Mathematics, 1999.
[26] L.A.Caffarelli. Boundary regularity of maps with convex potentials–ii. *Annals of Mathematics*, 144(3):453–496, 1996.
[27] G. Lan and Y. Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018.
[28] M. Liero, A. Mielke, and G. Savaré. Optimal transport in competition with reaction: The hellinger–kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.
[29] K. Liu and E. Zuazua. Representation and regression problems in neural networks: Relaxation, generalization, and numerics. *arXiv preprint arXiv:2412.01619*, 2024.
[30] D. Lowe and D. Broomhead. Multivariable functional interpolation and adaptive networks. *Complex systems*, 2(3):321–355, 1988.

[31] L.V.Kantorovich. On the translocation of masses. *J.Math.Sci.*, 133:1381–1382, 2006.
[32] S. Maniglia. Probabilistic representation and uniqueness results for measure-valued solutions of transport equations. *Journal de Mathématiques Pures et Appliquées*, 87(6):601–626, 2007.
[33] G.Savaré M.Liero, A.Mielke. Optimal entropy-transport problems and a new hellinger - kantorovich distance between positive measures. *Invent. math.*, 211:969–1117, 03 2018.
[34] Q. M. Nguyen, H. H. Nguyen, Y. Zhou, and L. M. Nguyen. On unbalanced optimal transport: Gradient methods, sparsity and approximation error. *J. Mach. Learn. Res.*, 24:384:1–384:41, 2022.
[35] R.J.McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80-2:309–323, 11 1995.
[36] F. Schwenker, H. A. Kestler, and G. Palm. Three learning phases for radial-basis-function networks. *Neural networks*, 14(4-5):439–458, 2001.
[37] T. Séjourné, F.-X. Vialard, and G. Peyré. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4995–5021. PMLR, 2022.
[38] S. Simon. *Minimax and Mononicity*. Springer Berlin, Heidelberg, 1998.
[39] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. *International Conference on Learning Representations*, 2018.
[40] K. D. Yang and C. Uhler. Scalable unbalanced optimal transport using generative adversarial networks. *arXiv preprint arXiv:1810.11447*, 2018.
[41] E. Zuazua. Control and machine learning. *Collections*, 55(08), 2022.
[42] E. Zuazua. Progress and future directions in machine learning through control theory. In *FGS 2024 French-German-Spanish Conference on Optimization*, 2024.
[43] A. Álvarez López, A. H. Slimane, and E.Zuazua. Interplay between depth and width for interpolation in neural odes. *Neural Networks*, 180:106640, 2024.

Department of Mathematics, Texas A&M University, College Station, TX 77843, USA
Email address: pmnt1114@tamu.edu

Department of Mathematics, Texas A&M University, College Station, TX 77843, USA
Email address: minhbinh@tamu.edu