Diffeomorphic Registration using Sinkhorn Divergences

Lucas de Lara, Alberto González-Sanz, Jean-Michel Loubes

Diffeomorphic Registration using Sinkhorn Divergences

Lucas De Lara*, Alberto González-Sanz*, and Jean-Michel Loubes*

*Institut de Mathématiques de Toulouse, Université Paul Sabatier

Abstract

The diffeomorphic registration framework enables to define an optimal matching function between two probability measures with respect to a data-fidelity loss function. The non-convexity of the optimization problem renders the choice of this loss function crucial to avoid poor local minima. Recent work showed experimentally the efficiency of entropy-regularized optimal transportation costs, as they are computationally fast and differentiable while having few minima. Following this approach, we provide in this paper a new framework based on Sinkhorn divergences, unbiased entropic optimal transportation costs, and prove the statistical consistency with rate of the empirical optimal deformations.

Keywords: Diffeomorphic Registration, Entropic Optimal Transport

1 Introduction

Diffeomorphic deformations describe a large class of computational frameworks whose goal is to find optimal deformations of the ambient space, defined as a diffeomorphisms generated through flow equations [Joshi and Miller, 2000, Beg et al., 2005, Younes, 2010]. They amount to solving an optimization problem involving two terms: an objective loss function characterizing in which sense the deformation should be optimal; a penalization over the kinetic energy spent by the transformation. The versatility of the problem formulation along with the appealing mathematical properties of diffeomorphisms made diffeomorphic deformations widely used in various application fields. In particular, they have been popularized for *diffeomorphic registration* in medical image analysis. This task consists of constructing diffeomorphic matching functions between shapes in order to establish spatial correspondences [Sotiras et al., 2013]. More recently, Younes [2020] proposed to apply flows of diffeomorphisms in a machine-learning context, where the optimal deformation is designed to render the data classes linearly separable.

This paper focuses on the diffeomorphic registration problem between two shapes. More specifically, we address the setting where the shapes are represented by probability measures: a formulation that has received a growing interest over the past few years to address unlabeled landmarks [Glaunes, 2005, Bauer et al., 2015, Feydy et al., 2017, Feydy and Trouvé, 2018]. In this case, the objective loss function, referred as the *data-fidelity loss*, is defined as a metric between probability measures. Squares of *maximum mean discrepancies* (MMD), which are well-known kernel-based distances, became the canonical choice for such settings. In particular, their use for diffeomorphic registration enjoys a well-established theory [Glaunes et al., 2004, Glaunes, 2005, Younes, 2010]. However, they also suffer from important practical drawbacks.

As pointed out by Feydy et al. [2017], the non-convexity of the optimization problem on the diffeomorphic deformation renders the choice of the loss function crucial to avoid poor local minima, whereas an MMD possesses

many. This is why they proposed to use optimal transport metrics as an alternative. More precisely, they define the data-fidelity loss as the entropy-regularized optimal transportation cost between unbalanced measures, which has two critical advantages. Firstly, it benefits from the non-locality of optimal transport metrics, leading to few local minima. Secondly, entropic regularization alleviates the computational burden of standard optimal transport: it allows for fast computation and and differentiation of the cost through the celebrated Sinkhorn's algorithm [Cuturi, 2013]. Nevertheless, while this alternative loss for diffeomorphic registration performs better experimentally, it lacks the statistical theory that was proven for squares of MMDs. Moreover, the entropic regularization induces a well-known bias making the loss not minimal between two identical measures. The latter issue motivates the employment of a *Sinkhorn divergence*: a symmetric unbiased version of the standard entropy-regularized optimal transportation cost. In [Feydy et al., 2019], the authors showed that Sinkhorn divergences performed significantly better than their biased counterparts for registration purpose. However, their carried out their analysis using flows of gradients (an approach reviewed in [Santambrogio, 2017]) instead of flows of diffeomorphisms. To the best of our knowledge, Sinkhorn divergences have never been studied and implemented for *diffeomorphic* registration.

This paper addresses diffeomorphic registration for Sinkhorn-divergence-based fidelity losses from both a theoretical and practical viewpoint. By leveraging some recent advances on these divergences [Feydy et al., 2019, Genevay et al., 2019, del Barrio et al., 2022], we show in a statistically-driven approach that the deformation obtained by solving the optimization problem between empirical measures converges with the parametric rate $\sqrt{n}$ to its population counterpart, where n is the sample size. Additionally, we illustrate the practicality of our method through numerical experiments. This furnishes a new theoretically and practically grounded framework for diffeomorphic matching of probability measures.

Outline The rest of the paper is organized as follows. In Section 2, we specify the basic mathematical notations that will be used throughout the paper. In Section 3, we set up the general problem we address by introducing the diffeomorphic registration framework for arbitrary data-fidelity losses. In Section 4, we present the necessary background on optimal transport and entropic regularization, in order to properly define Sinkhorn divergences. Additionally, we study some indispensable regularity properties of entropic optimal transport. In Section 5, we state our main results, that is the existence and statistical consistency of the optimal deformations. In Section 6, we recall the implementation of diffeomorphic registration, and present the numerical experiments where we benchmark Sinkhorn divergences with other losses. All the proofs are deferred to Section B, while Section A recalls key mathematical tools from empirical process theory and Frechet differentiability.

2 Preliminaries and notations

In this section, we introduce the definitions and notations that will be used throughout the paper. The first part is dedicated to classes of smooth functions; the second one addresses probability measures.

2.1 Smooth functions

Let $d_1 \geq 1$ and $\mathcal{X}$ be an arbitrary subset of $\mathbb{R}^{d_1}$ with non-empty interior denoted by $\mathring{\mathcal{X}}$. For $p \geq 1$ and $d_2 \geq 1$, we define $\mathcal{C}^p(\mathcal{X}, \mathbb{R}^{d_2})$ as the set of p-continuously Frechet differentiable functions from $\mathcal{X}$ to $\mathbb{R}^{d_2}$. We also define $\mathcal{L}^p(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$ the set of symmetric p-multilinear operators from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$. The p-th derivative of some $F \in \mathcal{C}^p(\mathcal{X}, \mathbb{R}^{d_2})$ is denoted by $F^{(p)}$. It maps any point $x \in \mathring{\mathcal{X}}$ to $F^{(p)}(x)[\cdot] \in \mathcal{L}^p(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$. By convention we set $F^{(0)} = F$. For any $L \in \mathcal{L}^p(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$, we define the operator norm as

$$\|L\|_{op} := \sup\{\|L[\delta_1, \dots, \delta_k]\| \mid \delta_i \in \mathbb{R}^{d_1},\ \|\delta_i\| \leq 1\}$$

where $\|\cdot\|$ is the Euclidean norm. For example, if $F \in \mathcal{C}^1(\mathcal{X}, \mathbb{R})$, then $\|F'(x)\|_{op} = \|\nabla F(x)\|$ where ∇F is the gradient of F. This enables to define, for any $F \in \mathcal{C}^p(\mathcal{X}, \mathbb{R}^{d_2})$, the functional norm,

$$\|F\|_{p,\infty} := \max_{0 \leq k \leq p} \left\|F^{(k)}\right\|_{\infty},$$

where $\|F\|_{\infty} := \sup_{x \in \mathcal{X}} \|F(x)\|$, and $\left\|F^{(k)}\right\|_{\infty} := \sup_{x \in \mathring{\mathcal{X}}} \left\|F^{(k)}(x)\right\|_{op}$ for $k \geq 1$. In addition, for any $R > 0$ we denote by $\mathcal{C}^p_R(\mathcal{X}, \mathbb{R}^{d_2})$ the class of functions $F \in \mathcal{C}^p(\mathcal{X}, \mathbb{R}^{d_2})$ such that $\|F\|_{p,\infty} \leq R$, and write B_R for the centered Euclidean ball of radius R.

2.2 Actions on probability measures

We write $\mathbb{E}[X]$ for the expectation of any random variable X. The symbol $\otimes$ denotes the product of measures. For two measures μ and ν on $\mathbb{R}^d$, the relation $\mu \ll \nu$ means that μ is absolutely continuous with respect to ν, that is $(\nu(E) = 0 \implies \mu(E) = 0)$ for every measurable set $E \subseteq \mathbb{R}^d$.

We define two kinds of actions involving probability measures. Let μ be a probability measure on $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. The action of μ on f defines the real number:

$$\mu(f) := \int f \mathrm{d}\mu = \mathbb{E}_{X \sim \mu}[f(X)].$$

Now, consider a measurable function $F : \mathbb{R}^d \to \mathbb{R}^d$. The action of F on μ defines a probability measure called the *push-forward* measure, defined as:

$$F_\sharp \mu := \mu \circ F^{-1}(\cdot).$$

If a random variable X follows the law μ, then the image variable $F(X)$ follows the law $F_\sharp \mu$. The push-forward operation enables to write changes of variables. Formally,

$$\int f \mathrm{d}(F_\sharp \mu) = \int (f \circ F) \mathrm{d}\mu.$$

3 Diffeomorphic measure transportation

In this section we present the necessary background on diffeomorphic registration of probability measures. We refer to [Younes, 2010] for a complete and precise treatment of this topic. Firstly, we recall how to define diffeomorphisms through flow equations. Secondly, we introduce the diffeomorphic measure transportation problem for arbitrary data-fidelity losses.

3.1 Generating diffeomorphic deformations

The diffeomorphic deformation framework can be framed as a fluid mechanics problem, where points in $\mathbb{R}^d$ are transported by a vector field representing a stream varying across time in the ambient space. We begin by reviewing the corresponding formalism and theory.

For an integer $p \geq 1$ let $\mathcal{B}_p$ be the space of functions in $\mathcal{C}^p(\mathbb{R}^d, \mathbb{R}^d)$ whose derivatives up to order p vanish to zero at infinity. This together with the norm $\|\cdot\|_{p,\infty}$ is a Banach space. Next, denote by V a Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$ and norm $\|\cdot\|_V$, and assume that V is *continuously embedded* in $\mathcal{B}_p$. This corresponds to the hypothesis below.

Assumption 3.1. *The space V is included in $\mathcal{B}_p$, and there exists a constant $c_V > 0$ such that for any $v \in V$,*

$$\|v\|_{p,\infty} \leq c_V \|v\|_V.$$

Physically, a function $v \in V$ represents a stationary vector field in the ambient space, specifying the speed vector $v(x) \in \mathbb{R}^d$ of the stream running at every position $x \in \mathbb{R}^d$. Then, define the class L^2_V of vector fields $t \in [0,1] \mapsto v_t \in V$ indexed by time and space satisfying $\int_0^1 \|v_t\|_V^2 \mathrm{d}t \leq \infty$, which is a Hilbert space endowed with the inner product,

$$\langle v, u \rangle_{L^2_V} := \int_0^1 \langle v_t, u_t \rangle_V \mathrm{d}t.$$

We recall that a sequence $\{v^n\}_{n\in\mathbb{N}}$ in L^2_V *converges weakly* to v if for any $u \in L^2_V$,

$$\langle v^n, u \rangle_{L^2_V} \xrightarrow[n\to+\infty]{} \langle v^n, u \rangle_{L^2_V}. \tag{1}$$

The associated norm in L^2_V is given by

$$\|v\|_{L^2_V} := \sqrt{\int_0^1 \|v_t\|_V^2 \mathrm{d}t},$$

and we use the notation

$$L^2_{V,M} := \{v \in L^2_V \mid \|v\|_{L^2_V} \leq M\}$$

for the centered ball of radius $M > 0$ in L^2_V.

We can now turn to the definition of diffeomorphic deformations. Any vector field $v \in L^2_V$ generates a deformation $\phi^v := (\phi^v_t)_{t\in[0,1]}$, function of both time and space variables, defined as the unique solution to the following *flow equation*,

$$\forall x \in \mathbb{R}^d,\ \forall t \in [0,1], \quad \phi_t(x) = x + \int_0^t v_s\big(\phi_s(x)\big)\mathrm{d}s. \tag{2}$$

The parametric curve $(\phi^v_t(x))_{t\in[0,1]}$ represents the trajectory across time of a point initially located at $\phi_0(x) = x \in \mathbb{R}^d$. Remarkably, for every $t \in [0,1]$ the transformation ϕ^v_t is a p-continuously differentiable diffeomorphism. Moreover, as a direct consequence of Theorem 5 in [Glaunes, 2005], these diffeomorphic transformations are smooth over compact sets.

Lemma 3.1 (Smoothness of diffeomorphic deformations)**.** *Assume that Assumption 3.1 holds. Then for any radius $M > 0$ and any compact set $K \subset \mathbb{R}^d$, there exists a constant $R = R((K,d);(V,p);M) > 0$ such that for any $v \in L^2_{V,M}$,*

$$\max_{0\leq k\leq p} \left\{ \sup_{t\in[0,1], x\in B_R} \left\| \left(\phi^v_t\right)^{(k)}(x) \right\|_{op} \right\} \leq R.$$

In particular, $\sup_{x\in K} \|x\| \leq R$.

In practice, the space of vector fields V is constructed through the choice of a kernel function. This is enabled by Assumption 3.1 which entails that V is a reproducing kernel Hilbert space (RKHS), characterized by a unique non-negative symmetric matrix-valued kernel function $\mathrm{Ker} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$. In particular, the choice of the kernel function sets the order of regularity p of the vector fields. For instance, the typical choice of a Gaussian kernel, that is

$$\mathrm{Ker}(x,y) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) I_d \tag{3}$$

where $\sigma > 0$ is the bandwidth parameter and I_d the identity matrix, leads to $p = +\infty$.

3.2 Diffeomorphic matching of distributions

In general, diffeomorphic deformation frameworks amount to finding solutions to Equation 2 that are optimal in some sense. In this work, we focus on the diffeomorphic measure transportation framework, which aims at matching two probability measures.

Formally, let Λ be a positive loss function between probability measures, and set α and β two probabilities on the ambient space $\mathbb{R}^d$. For a given regularization weight $\lambda > 0$, an optimal matching function between α and β is a diffeomorphism ϕ^v solution to (2) where v minimizes

$$J_\lambda(v) := \Lambda(\phi^v_{1\sharp}\alpha, \beta) + \lambda \|v\|^2_{L^2_V}. \tag{4}$$

The first term of the objective function (4) is the *data-fidelity loss*, which tends to match $\phi^v_{1\sharp}\alpha$ with β, while the second term is the regularizer, which penalizes the kinetic energy spent by the trajectories $(\phi^v_t)_{t\in[0,1]}$, keeping them as close as possible to the identity function. The parameter λ governs the trade-off between the two contributions. The objective J_λ always admits minimizers, provided that the term $v \in L^2_V \mapsto \Lambda(\phi^v_{1\sharp}\alpha, \beta) \in \mathbb{R}^+$ is weakly continuous. For a minimizer v^*, the function $\phi^{v^*}_1$ is an optimal matching between α and β, and the family $(\phi^{v^*}_t)_{t\in[0,1]}$ provides an interpolation between the two measures.

In practical settings, one typically does not have access to the full probability measures α and β but to empirical observations. This naturally raises the question of estimating an optimal matching function between α and β on the basis of independent samples. Concretely, let $x_1, \dots, x_n \sim \alpha$ and $y_1, \dots, y_n \sim \beta$ be independent samples, and define the empirical probability measures $\alpha_n := n^{-1}\sum_{i=1}^n \delta_{x_i}$ and $\beta_n := n^{-1}\sum_{j=1}^n \delta_{y_j}$. Plugging these discrete measures in the original objective function (4) leads to the following empirical objective function:

$$J_{\lambda,n}(v) := \Lambda(\phi^v_{1\sharp}\alpha_n, \beta_n) + \lambda \|v\|^2_{L^2_V}. \tag{5}$$

In Theorem 5.1 we prove under some assumptions that if the data-fidelity loss Λ is a *Sinkhorn divergence*, a divergence derived from entropic optimal transport, then any sequence of minimizers $\{v^n\}_{n\in\mathbb{N}}$ of the empirical problem (5) converges up to the extraction of a subsequence to a minimizer of the population problem (4) as the sample size n increases to infinity.

4 Entropic optimal transport

In this section, we first briefly present the necessary background on optimal transport and entropic regularization, in order to properly define Sinkhorn divergences. We refer to [Villani, 2003, 2008, Peyré et al., 2019] for further insight on these topics. Then, we introduce some properties of these divergences, which will be useful to later demonstrate the main results of this paper.

4.1 Transportation costs and Sinkhorn divergences

Let α and β be two probability measures on $\mathcal{X}$ a subset of $\mathbb{R}^d$, and $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ a positive ground cost function. Typically, $C(x, y) := \|x - y\|^2$. The optimal transportation cost with respect to C between α and β is defined as,

$$\mathcal{T}_C(\alpha, \beta) := \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} C(x, y)\mathrm{d}\pi(x, y), \tag{6}$$

where $\Pi(\alpha, \beta)$ is the set of couplings admitting α as first marginal and β as second marginal. In particular, for an integer $k \geq 1$ and D a distance on $\mathcal{X}$, the quantity $(\mathcal{T}_{D^k})^{\frac{1}{k}}$ yields a distance between measures referred as

the *Wasserstein distance* of order k. Transportation costs and optimal transport distances became popular in many machine-learning-related problems for their appealing geometric properties, but suffer from being computationally challenging in practice. This triggered a growing literature on fast approximations of (6), the most popular being entropy-regularized versions, which can be computed through the Sinkhorn algorithm [Cuturi, 2013]. For $\varepsilon > 0$, the *entropy-regularized* transportation cost w.r.t. C is defined as

$$\mathcal{T}_{C,\varepsilon}(\alpha, \beta) := \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} C(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\alpha \otimes \beta), \tag{7}$$

where $\mathrm{KL}(\mu|\nu)$ denotes the *Kullback-Leibler* divergence between probability measures μ and ν given by $\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(z)\right)\mathrm{d}\mu(z)$ if $\mu \ll \nu$, and $+\infty$ otherwise.

Critically, the entropic transportation cost $\mathcal{T}_{C,\varepsilon}$ suffers from the so-called *entropic bias*, that is $\mathcal{T}_{C,\varepsilon}(\alpha, \alpha) \neq 0$ in general. As illustrated in [Feydy et al., 2019], this entails that the minimum of $\mathcal{T}_{C,\varepsilon}(\alpha, \cdot)$ is not reached at α but at a shrunken version of α with smaller support, making the entropic cost an unreliable loss function. The Sinkhorn divergence was originally introduced to fix this undesirable effect. It is formally defined as

$$S_{C,\varepsilon}(\alpha, \beta) := \mathcal{T}_{C,\varepsilon}(\alpha, \beta) - \frac{1}{2}\mathcal{T}_{C,\varepsilon}(\alpha, \alpha) - \frac{1}{2}\mathcal{T}_{C,\varepsilon}(\beta, \beta).$$

As aforementioned, using a non-local similarity measure such as an entropic-optimal-transport cost instead of a local similarity measure such as a squared MMD leads to fewer local solutions when minimizing (5). Moreover, it does not suffer from the computational burden of standard optimal transport. This is why Feydy et al. [2017] advocated the use of the entropy-regularized transportation cost (7) for diffeomorphic registration, providing empirical evidences of the benefits of this approach. However, they did not rely on the unbiased Sinkhorn divergences, for which little was known until [Feydy et al., 2019] that demonstrated several key properties. In particular, if C is continuous and $\mathcal{X}$ is compact, then $S_{C,\varepsilon}$ is symmetric positive definite, smooth and convex in each of its input distributions. Additionally, in contrast to the standard regularized transportation cost, it metrizes the convergence in law. The goal of this paper is precisely to use a Sinkhorn divergence for the data-fidelity loss, while providing statistical guarantees. The demonstrations are based on the dual formulation of entropic optimal transport for which we derive some important results next.

4.2 Regularity of the dual formulation

The minimization problem (7) has the following dual formulation,

$$\mathcal{T}_{C,\varepsilon}(\alpha, \beta) = \sup_{f,g \in \mathcal{C}(\mathcal{X},\mathbb{R})} \int_{\mathcal{X}} f(x)\mathrm{d}\alpha(x) + \int_{\mathcal{X}} g(y)\mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{X}} e^{\frac{f(x)+g(y)-C(x,y)}{\varepsilon}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) + \varepsilon. \tag{8}$$

The functions f and g are referred as *potentials*. Note that Equation 8 can also be compactly written as,

$$\mathcal{T}_{C,\varepsilon}(\alpha, \beta) = \sup_{f,g \in \mathcal{C}(\mathcal{X},\mathbb{R})} (\alpha \otimes \beta)\left(h^{f,g}_{C,\varepsilon}\right),$$

where

$$h^{f,g}_{C,\varepsilon}(x, y) := f(x) + g(y) - \varepsilon e^{\frac{f(x)+g(y)-C(x,y)}{\varepsilon}} + \varepsilon. \tag{9}$$

We call the function $h^{f,g}_{C,\varepsilon}$ the *global potential*. It will play a key role in the proofs.

A remarkable property of entropic optimal transport, investigated in [Genevay et al., 2019, Feydy et al., 2019], is that the potentials of the dual formulation inherit the regularity of the ground cost function C if the measures α and β are compactly supported. This setting will be useful to derive statistical guarantees. More

specifically, it allows to restrict the set of feasible potentials to smooth functions regardless of the involved probability measures, as stated in the next lemma which readily follows from Proposition 1 in [Genevay et al., 2019] (see also Lemma 4.1. in [del Barrio et al., 2022] for the particular case of the quadratic ground cost).

Lemma 4.1 (Smoothness of the optimal potentials)**.** *Let μ and ν be two measures on a compact set $K \subset \mathbb{R}^d$, and suppose that the ground cost function C belongs to $\mathcal{C}^q(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$ with $q \geq 1$. Then, there exists a constant $m = m((K,d);(C,q);\varepsilon) > 0$ such that*

$$\mathcal{T}_{C,\varepsilon}(\mu,\nu) = \sup_{f,g \in \mathcal{C}(K,\mathbb{R})} (\mu \otimes \nu)\left(h^{f,g}_{C,\varepsilon}\right) = \sup_{f,g \in \mathcal{C}^q_m(K,\mathbb{R})} (\mu \otimes \nu)\left(h^{f,g}_{C,\varepsilon}\right).$$

Naturally, the smoothness of f, g and C renders the global potential $h^{f,g}_{C,\varepsilon}$ smooth as well. Combining Lemma 4.1 with the following result ensures the smoothness of the optimal global potential under smooth data-processing transformations, such a for instance diffeomorphic transformations.

Proposition 4.1 (Smoothness of the optimal global potential)**.** *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, suppose that the ground cost function C belongs to $\mathcal{C}^q(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$ with $q \geq 1$, and set $p \geq 1$. Then for any $m > 0$ and $R > 0$, there exists a constant $H = H(m;R;(C,q);\varepsilon;p) > 0$ such that for any $f, g \in \mathcal{C}^q_m(B_R, \mathbb{R})$ and $T_1, T_2 \in \mathcal{C}^p_R(\mathcal{X}, \mathbb{R}^d)$,*

$$h^{f,g}_{C,\varepsilon} \circ (T_1, T_2) \in \mathcal{C}^\kappa_H(\mathcal{X} \times \mathcal{X}, \mathbb{R}),$$

where $\kappa := \min\{p,q\}$.

We are now ready to state and prove our main results.

5 Main results

This section focuses on the main theoretical contributions of the paper, namely the existence and statistical consistency of the empirical optimal matching function between α and β when using a Sinkhorn divergence.

Firstly, we show that the objective functions J_λ and $J_{\lambda,n}$ with $\Lambda = S_{C,\varepsilon}$ admit minimizers. We recall that a function $\Psi : L^2_V \to \mathbb{R}$ is *weakly continuous* if for any sequence $\{v^n\}_{n\in\mathbb{N}}$ weakly converging to some $v \in L^2_V$ (see Equation 1), we have $\Psi(v^n) \xrightarrow[n\to+\infty]{} \Psi(v)$. Theorem 7 in [Glaunes, 2005] states that J_λ admits a minimum if $v \in L^2_V \mapsto \Lambda(\phi^v_{1\sharp}\alpha, \beta)$ is weakly continuous. Therefore, existence directly follows from the proposition below.

Proposition 5.1 (Existence of the optimal vector fields)**.** *Let α and β be two probability measures on $\mathcal{X}$ a compact subset of $\mathbb{R}^d$, suppose that the ground cost function C belongs to $\mathcal{C}^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$, and assume that Assumption 3.1 holds. Then the function $v \in L^2_V \mapsto S_{C,\varepsilon}(\phi^v_{1\sharp}\alpha, \beta)$ is weakly continuous.*

The minimizer is not unique in general, due to the non-convexity of the data-fidelity loss with respect to v. Uniqueness could be artificially achieved by choosing λ very large, thereby rendering the objective function strictly convex, but this would make the purpose of the regularization meaningless.

We now turn to our main theorem, which is divided in two items. The first one ensures the convergences of the empirical solutions to their population counterparts; the second one specifies the speed of this convergence.

Theorem 5.1 (Consistency of the optimal vector fields)**.** *Let α_n and β_n be empirical measures corresponding respectively to α and β, two probability measures on $\mathcal{X}$ a compact subset of $\mathbb{R}^d$, suppose that the ground cost function C belongs to $\mathcal{C}^q(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$ with $q \geq 1$, and assume that Assumption 3.1 holds. If, for any $n \in \mathbb{N}^*$, v^n denotes a minimizer of $J_{\lambda,n}$ for $\Lambda = S_{C,\varepsilon}$, the following results hold.*

(i) There exists a minimizer of J_λ denoted by v^* such that up to the extraction of a subsequence

$$\|v^n - v^*\|_{L^2_V} \xrightarrow[n\to\infty]{a.s.} 0,$$

and

$$\sup_{t\in[0,1]} \left\{ \left\| \phi_t^{v^n} - \phi_t^{v^*} \right\|_\infty + \left\| (\phi_t^{v^n})^{-1} - (\phi_t^{v^*})^{-1} \right\|_\infty \right\} \xrightarrow[n\to\infty]{a.s.} 0.$$

(ii) If $\kappa := \min\{p, q\} > d$, then for any radius $M > 0$ there exists a constant $A = A(\lambda; (\mathcal{X}, d); (C, q); \varepsilon; (V, p))$ such that

$$\mathbb{E}\left[|J_\lambda(v^n) - J_\lambda(v^*)|\right] \leq \frac{A}{\sqrt{n}}.$$

Note that Glaunes et al. [2004] proved a similar consistency result when the data-fidelity loss is the square of an MMD, but did not determine the speed of convergence as in (ii). The demonstration of (i) follows the steps of their proof (see Theorem 16 in [Glaunes, 2005]). The idea is to show the convergence of $\sup_{v\in L^2_{V,M}} |J_{\lambda,n}(v) - J_\lambda(v)|$ as n increases to infinity, where $L^2_{V,M}$ contains all the minimizers independently of n. The main challenge when addressing an entropic optimal transport cost comes from the fact that it does not satisfy a triangle inequality, nor a data-processing inequality, and is hence harder to control. We remedy to this issue by proving and applying the following intermediary result:

Proposition 5.2 (Uniform consistency of entropic optimal transport up to smooth data-processing transformations)**.** *Let α_n and β_n be empirical measures corresponding respectively to α and β, two probability measures on $\mathcal{X}$ a compact subset of $\mathbb{R}^d$, and suppose that the ground cost function C belongs to $\mathcal{C}^q(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}^+)$ with $q \geq 1$. Set $p \geq 1$ and write $\kappa := \min\{p, q\}$. Then, the following results hold:*

(i) For any $R > 0$

$$\sup_{T_1,T_2\in\mathcal{C}^p_R(\mathcal{X},\mathbb{R}^d)} |\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - \mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)| \xrightarrow[n\to+\infty]{a.s.} 0.$$

(ii) If $\kappa > d$, then for any $R > 0$ there exists a constant $A = A(R; (C, q); \varepsilon; (\mathcal{X}, d); p) > 0$ such that

$$\mathbb{E}\left[\sup_{T_1,T_2\in\mathcal{C}^p_R(\mathcal{X},\mathbb{R}^d)} |\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - \mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)| \right] \leq \frac{A}{\sqrt{n}}.$$

Notice that as a direct consequence of the triangle inequality, a similar result holds for $S_{C,\varepsilon}$. Hence, as diffeomorphisms are smooth on compact sets according to Lemma 3.1, we can apply Proposition 5.2 to control $\sup_{v\in L^2_{V,M}} |J_{\lambda,n}(v) - J_\lambda(v)|$.

Although Proposition 5.2 is motivated by diffeomorphic registration, we believe it has further interest. Remark in particular that the objective (4) shares similarities with generative modelling [Goodfellow et al., 2014]; an input distribution α is passed through a parametric function ϕ_1^v meant to generate a target distribution β by minimizing a certain loss Λ. In particular, generative modelling using the Wasserstein-1 distance or a Sinkhorn divergence has proved to be efficient for diverse applications [Arjovsky et al., 2017, Genevay et al., 2018]. The main difference in (4) comes from the parameter v being infinitely dimensional, and characterizing a diffeomorphism instead of a neural network. However, Proposition 5.2 is general enough to be applied in the context of generative modelling with Sinkhorn divergences, in order to derive statistical guarantees for smooth generators.

Remark 5.1. *Proposition 5.1 and Proposition 5.1 still hold for $\mathcal{T}_{C,\varepsilon}$ instead of $\mathcal{S}_{C,\varepsilon}$. However, we emphasize that it is preferable to use a Sinkhorn divergence in practice, since it does not suffer from the aforementioned entropic bias. In particular, the experiments from the next section illustrate that debiasing leads to more accurate registrations.*

6 Implementation

This section addresses the practical aspects of diffeomorphic registration through Sinkhorn divergence. Firstly, we briefly recall how to compute a minimizer of $J_{\lambda,n}$ for an arbitrary loss Λ. Then, we illustrate the procedure for Sinkhorn divergences on numerical experiments.

6.1 Resolution procedure

This subsection introduces the basic knowledge for solving a diffeomorphic registration problem. It is meant to keep the paper as self-contained as possible. See [Younes, 2010] for a complete overview of the procedure.

To practically minimize $J_{\lambda,n}$, one must first write the optimal vector fields v in a finite parametric form, and then perform a gradient descent on the coefficients of this decomposition. Recall that Assumption 3.1 implies that V is a RKHS, thereby characterized by a unique matrix-valued symmetric positive kernel function $\mathrm{Ker} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$. For simplicity, we address the case of the Gaussian kernel defined in (3). Statistically, the bandwidth parameter σ represents the correlation between the morphed points; physically, it quantifies the fluid viscosity. When σ is small, the points have independent trajectories; when it is large, the points move as a whole.

The RKHS viewpoint enables to parametrize the optimal vector fields through a kernel trick. Firstly, note that the minimization of $J_{\lambda,n}$ can be formulated as an optimal control problem. It amounts to solving

$$\min_{v\in L^2_V} \Lambda(\alpha_n(1),\beta_n) + \lambda\|v\|^2_{L^2_V};\ \text{subject to } \alpha_n(t) = \phi^v_{t\,\sharp}\alpha_n \text{ for any } t\in[0,1]. \tag{10}$$

Then, since the constraint involves a finite number n of trajectories, the so-called *reduction principle* (see Theorem 14 in [Glaunes, 2005]) entails that any solution to problem (10), that is any minimizer of $J_{\lambda,n}$, can be written as,

$$v^n_t(x) = \sum_{i=1}^n \mathrm{Ker}\left(x, z^a_i(t)\right) a_i(t),$$

where $a := (a_1,\dots,a_n)$ denotes n unspecified time functions of $L^2([0,1],\mathbb{R}^d)$, and the control trajectories $z^a := (z^a_1,\dots,z^a_n)$ are defined by

$$z^a_i(t) = x_i + \int_0^t \sum_{j=1}^n \mathrm{Ker}(z^a_i(s), z^a_j(s)) a_j(s)\mathrm{d}s. \tag{11}$$

This enables to recast Problem (10) as minimizing,

$$E_{\lambda,n}(a) := \Lambda\left(\frac{1}{n}\sum_{k=1}^n \delta_{z^a_k(1)},\beta\right) + \lambda\int_0^1 \sum_{i,j=1}^n a_i(t)\cdot\mathrm{Ker}\left(z^a_i(t),z^a_j(t)\right)a_j(t)\mathrm{d}t, \tag{12}$$

where $\cdot$ denotes the Euclidean inner product. The gradient of $E_{\lambda,n}$ was originally derived in [Glaunes et al., 2004] for the MMD case, and re-expressed in [Glaunes, 2005, Younes, 2020] for more general settings. It can be written as $\nabla E_{\lambda,n}(a) = 2\gamma a - p^a$ where $p^a := (p^a_1,\dots,p^a_n)$ denotes n functions of $L^2([0,1],\mathbb{R}^d)$ satisfying for any $i\in\{1,\dots,n\}$ and $t\in[0,1]$,

$$\begin{aligned} p^a_i(t) := {} & \nabla_{z^a_i(1)}\Lambda\left(\frac{1}{n}\sum_{k=1}^n \delta_{z^a_k(1)},\beta\right) \\ & - \frac{1}{\sigma^2}\int_t^1 \sum_{j=1}^n \mathrm{Ker}\left(z^a_i(t),z^a_j(t)\right)\left[a_i(t)\cdot p^a_j(t) + a_j(t)\cdot p^a_i(t) - 2\lambda a_i(t)\cdot a_j(t)\right]\left(z^a_i(t) - z^a_j(t)\right). \end{aligned} \tag{13}$$

In order to practically track all the functions of the continuous time variable, one must discretize the time scale $[0, 1]$ into τ sub-intervals of equal sizes, which recasts a, z^a and p^a as $(\tau + 1) \times n \times d$ tensors. Then, equations (11) and (13) are successively solved at each iteration of the gradient descent by solving the associated discrete dynamical systems. By plugging the solutions z^a and p^a into the formula of $\nabla E_{\lambda,n}(a)$ one can update the variable a with $a \leftarrow a - \xi \times (2\lambda a - p^a)$ where ξ denotes the step size. The computational complexity of an iteration is in $O(n^2 d\tau)$. However, the dynamical systems can be parallelized in the number of points and the dimension. In practice, we employed recent work on the automatic differentiation of geometrical losses [Feydy et al., 2017] to implement our algorithm computing the optimal deformation. We refer to the code[1] for more details. At the end of the process, we obtain the following deformation,

$$\phi_t^{a,\tau}(x) := x + \frac{1}{\tau} \sum_{s=0}^{t-1} \sum_{j=1}^{n} \mathrm{Ker}(x, z_j^a(s)) a_j(s).$$

Note that this approach handles any data-fidelity loss Λ as long as it is differentiable with respect to the data points of the discrete distributions. Both Sinkhorn divergences and squares of MMDs satisfy this property. In the numerical experiments, we use the GeomLoss package [Feydy et al., 2019] to compute the gradient of the losses by automatic differentiation.

6.2 Numerical experiments

In [Feydy et al., 2019], the authors proposed an alternative measure registration framework based on the gradient flow of the data-fidelity loss. It amounts to updating the source distribution $\alpha_n := n^{-1} \sum_{i=1}^{n} \delta_{x_i}$ by carrying out a gradient descent on $\Lambda(\alpha_n, \beta_n)$ with respect to the positions $x_1, \ldots, x_n$. This model-free method enables to faithfully match one distribution to another, even when the supports have irregularities such as holes. In this section, we firstly adapt their experiments, more precisely the ones from the example section of the GeomLoss package website,[2] by using diffeomorphic deformations instead of gradient flows.

[1] `https://github.com/lucasdelara/lddmm-sinkhorn/`

[2] `https://www.kernel-operations.io/geomloss/`

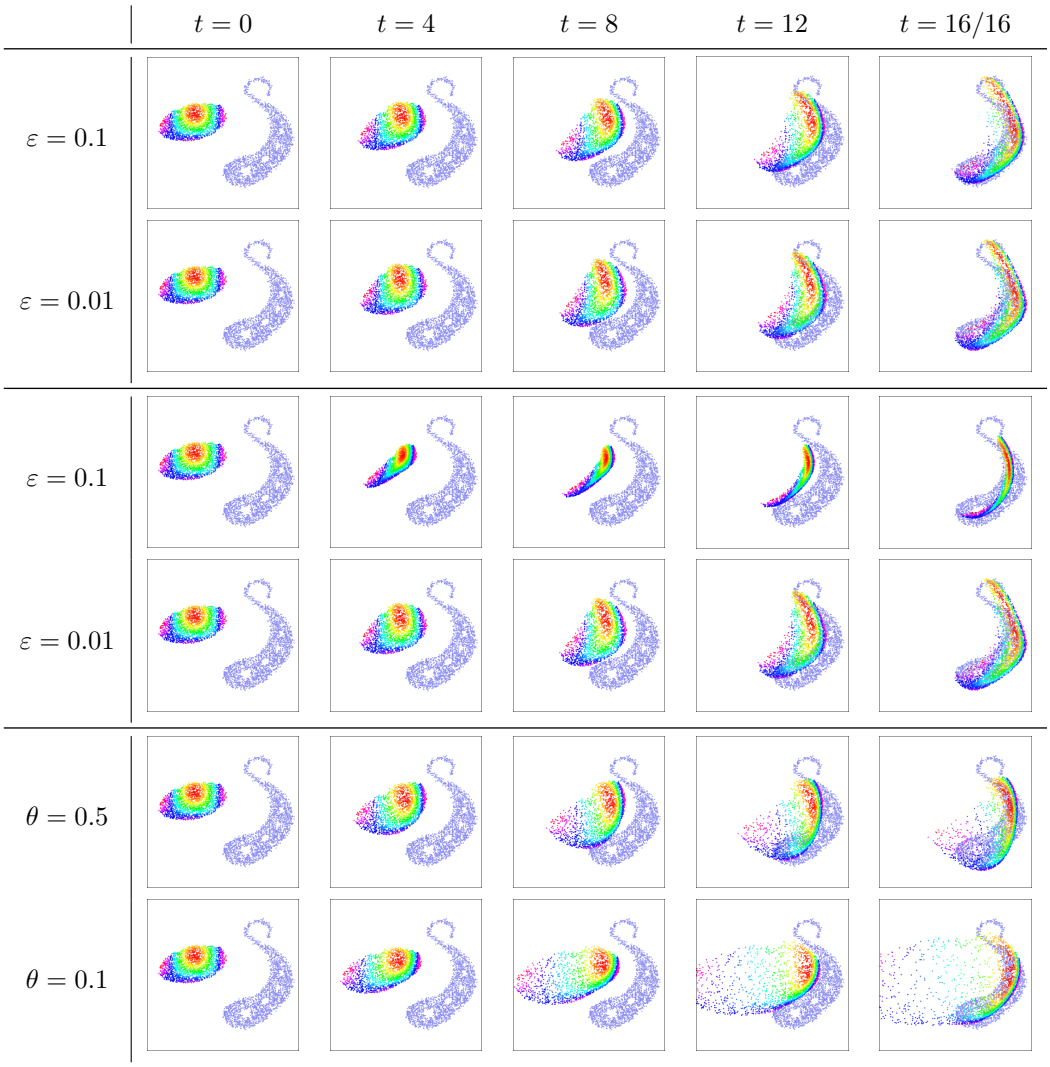


Table 1: 2-D diffeomorphic registration. The colored distribution is $\alpha_m(t)$, while the blue distribution is β_m. The first row is for the Sinkhorn divergence $S_{C,\varepsilon}$, the second row for the biased entropic transportation cost $\mathcal{T}_{C,\varepsilon}$, the third row for the squared Gaussian maximum mean discrepancy MMD^2_θ.

The objective is matching two blob-like point clouds in dimension 2. We proceed as follows. Firstly, we learn the optimal matching between two samples of size $n = 1,000$ using the previously described procedure. Secondly, we display the obtained time interpolation between two new independent samples of size $m = 2,000$. In order to benchmark the influence of the data-fidelity loss, we consider a fixed setting where V is defined through a Gaussian kernel with bandwidth $\sigma = 0.175$, the regularization has weight $\lambda = 10^{-8}$, and the time scale is uniformly divided into $\tau = 16$ intervals. Then, we compare the results obtained with different losses: (unbiased) Sinkhorn divergences, biased entropic transportation costs, and squared Gaussian maximum mean discrepancies. Recall that the squared Gaussian MMD with bandwidth parameter $\theta > 0$ is defined as,

$$\mathrm{MMD}^2_\theta(\mu,\nu) := \int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(-\frac{\|x-y\|^2}{2\theta^2}\right) \mathrm{d}(\mu-\nu)(x)\mathrm{d}(\mu-\nu)(y).$$

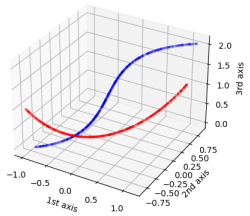
The ground cost function for the Sinkhorn divergences is always $C(x, y) := \|x - y\|^2$. Figure 1 shows the results for different values of the parameters ε and θ. Overall, the Sinkhorn divergence (first row) provides a sharper matching than the Gaussian MMD (third row), and the accuracy increases as the entropic parameter ε decreases. Moreover, the figures emphasize the importance of debiasing when ε is big: in contrast to the Sinkhorn divergence, the standard entropic transportation cost (second row) shrinks the morphed distribution.

All in all, our experimental observations about the role of the losses are similar to the ones made by Feydy et al. [2019] in the context of gradient flows. Critically, compared to their approach, we work with a transformation that is smooth at any time. This regularity constraint reduces the flexibility of the matching, which leads to a less accurate fitting than gradient flows. This affects particularly the anomalous parts of the targeted support, namely the holes and the tail. In contrast, regularity enables the deformation to generalize to any new out-of-sample observations. Additionally, it prevents from tearing the mass apart. The color map on the distribution $\alpha_m(t)$ enables to track the location of the moved points through time. Notice that, as a direct consequence of the smoothness, the chromatic continuity between morphed points is preserved throughout the process.

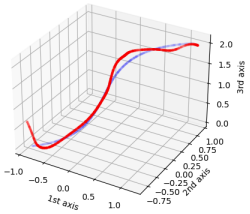
In a second time, we implement the diffeomorphic matching of two curves in $\mathbb{R}^3$. We define the distributions α and β as the uniform distributions over respectively $\{(x, \sin(x), x^2) \mid x \in [-1, 1]\}$ and $\{(x^3, \sin(x), x) \mid x \in [-0.95, 1.05]\}$. As before, the optimal matching functions are firstly trained for various losses on a testing set of size $n = 1,000$, and then applied on an independent testing dataset of size $m = 2,000$. We set $\sigma = 0.2$, $\lambda = 10^{-5}$, and $\tau = 16$. The results are reported in Figure 1. As anticipated, we observe that the Sinkhorn divergence and the entropic transportation cost for a small ε provide a very similar, faithful registration. When ε increases, the quality of the matching decreases for both losses. As for the 2-D experiments, the contrast between $S_{C,\varepsilon}$ and $\mathcal{T}_{C,\varepsilon}$ sharpens for larger values of ε, since the entropic bias intensifies. Moreover, these simulations also illustrate the shrinkage effect of this bias: the length of the red curve in Figure 1e is smaller than the others. Finally, we note that both optimal-transport fidelity losses outperform the Gaussian maximum mean discrepancy, siding with Feydy et al. [2017].

7 Conclusion

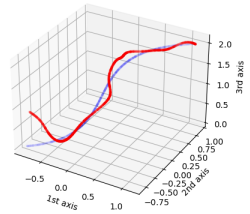
We proposed to use Sinkhorn divergences as the fidelity loss in diffeomorphic registration problems. We derived the statistical theory, and illustrated the efficiency of this method compared to past approaches based on MMDs or *biased* entropic transportation costs. As such, this paper paves way for accurate and smooth measure registration with certifiable asymptotic guarantees. Moreover, carrying out this work led us to further investigate the dual formulation of entropic optimal transport, complementing recent papers on the subject. An avenue for extension could be to consider the registration of *unbalanced* measures using Sinkhorn divergences, which would align with the work of Feydy et al. [2017].



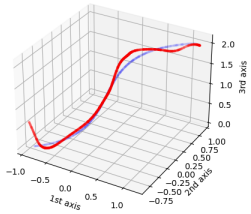
(a) $t = 0$



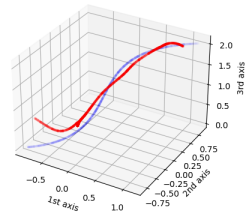
(b) $S_{C,\varepsilon}$: $\varepsilon = 0.1$



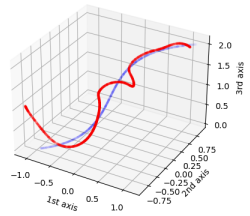
(c) $S_{C,\varepsilon}$: $\varepsilon = 0.5$



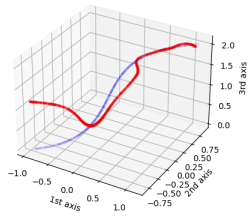
(d) $\mathcal{T}_{C,\varepsilon}$: $\varepsilon = 0.1$



(e) $\mathcal{T}_{C,\varepsilon}$: $\varepsilon = 0.5$



(f) MMD^2_θ : $\theta = 0.5$



(g) MMD^2_θ : $\theta = 1$

Figure 1: 3-D diffeomorphic registration. Figure (a) represents the dataset before registration, while Figures (b) to (g) represents the dataset at the end of the registration for different fidelity losses. The red distribution is $\alpha_m(0)$ or $\alpha_m(1)$; the blue distribution is β_m.

A Preliminary results

This section recalls some useful results. Section A.1 contains a brief reminder on entropy numbers of classes of functions, in order to derive an upper bound on empirical processes; Section A.2 focuses on the chain rule for composite Frechet derivatives up to arbitrary high orders.

A.1 Empirical processes

In the proof of Proposition 5.2, we will bound the sampling error between the empirical entropic transportation cost and its population counterpart by a centered empirical process indexed by a class of smooth functions. Recalling the theory introduced in [Van Der Vaart and Wellner, 1996, Koltchinskii, 2011], we present in this subsection intermediary results on such processes.

Let $\mathcal{X}$ be a compact convex subset of $\mathbb{R}^d$. For any probability measure μ on $\mathcal{X}$ and $r \geq 1$, we define the $L_r(\mu)$-norm on $\mathcal{C}(\mathcal{X}, \mathbb{R})$ as $\|h\|_{r,\mu} := \left(\int |h|^r \mathrm{d}\mu \right)^{1/r}$. In empirical process theory, the complexity of classes of functions is commonly evaluated through the so-called *covering* and *bracketing* numbers. Let $\mathcal{H}$ be a class of function included in $\mathcal{C}(\mathcal{X}, \mathbb{R})$, and $\epsilon > 0$ a constant. The covering number $N(\epsilon, \mathcal{H}, L_r(\mu))$ is defined as the minimal number of $L_r(\mu)$-balls of radius ϵ needed to cover the class of functions $\mathcal{H}$. The center of the balls need not belong to $\mathcal{H}$, but must have finite norm. Additionally, given two functions l and u with finite norm but not necessarily in $\mathcal{F}$, the *bracket* $[l, u]$ is the set of all functions f such that $l \leq h \leq u$. An $(\epsilon, L_r(\mu))$-bracket is a bracket $[l, u]$ such that $\|l - u\|_{r,\mu} \leq \epsilon$. Then, the bracketing number $N_{[\,]}(\epsilon, \mathcal{F}, L_r(\mu))$ is the minimal number of $(\epsilon, L_r(\mu))$-bracket needed to cover $\mathcal{F}$.

These numbers have essential applications in statistics. The supremum of a centered empirical process indexed by a class of functions with a finite bracketing number converges uniformly almost-surely to zero. Moreover, with a sharper control on the bracketing number, one can derive the following convergence rate:

Proposition A.1. *Let μ_n be an empirical measure of a probability measure μ corresponding to a compact convex subset $\mathcal{X}$ of $\mathbb{R}^d$, and set $H > 0$ a constant. Consider the class of functions $\mathcal{H} := \mathcal{C}^\kappa_H(\mathcal{X}, \mathbb{R})$ for some integer $\kappa \geq 0$. If $\kappa > d/2$, then there exists a constant $A = A((H, \kappa); (\mathcal{X}, d))$ such that,*

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} |\mu_n(h) - \mu(h)| \right] \leq \frac{A}{\sqrt{n}}.$$

Proof. Combining Theorem 2.1 with Theorem 3.11 in [Koltchinskii, 2011], we directly have that,

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} |\mu_n(h) - \mu(h)| \right] \leq 2 \times \frac{c}{\sqrt{n}} \mathbb{E} \int_0^{2\sigma_n} \sqrt{\log N(\epsilon, \mathcal{H}, L_2(\mu_n))} d\epsilon,$$

where $c > 0$ is some constant and $\sigma_n := \sup_{h \in \mathcal{H}} \mu_n(h^2)$. By definition of $\mathcal{H}$, it follows that $\sigma_n \leq H^2$. Besides, we can upper bound the covering number in the right term by the bracketing number $N_{[\,]}(2\epsilon, \mathcal{H}, L_2(\mu_n))$ (see p.84 in [Van Der Vaart and Wellner, 1996]). In addition, according to Corollary 2.7.2 in [Van Der Vaart and Wellner, 1996], there exists a constant $\rho = \rho((H, \kappa); (\mathcal{X}, d)) > 0$ such that,

$$\log N_{[\,]}(2\epsilon, \mathcal{H}, L_2(\mu_n)) \leq \rho (2\epsilon)^{-d/\kappa}.$$

Note that the right term does not depend on μ_n. All in all,

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} |\mu_n(h) - \mu(h)| \right] \leq \frac{2c}{\sqrt{n}} \mathbb{E} \int_0^{2H^2} \sqrt{\rho (2\epsilon)^{-d/\kappa}} d\epsilon.$$

The integral is finite as $\kappa > d/2$. Consequently, the upper bound defines a constant $A = A((H, \kappa); (\mathcal{X}, d))$. This concludes the proof. □

Remark that the convexity assumption on the compact domain $\mathcal{X}$ is not restrictive, as it suffices to extend the probability measure μ on the convex hull of $\mathcal{X}$.

A.2 Frechet derivative

The proof of Proposition 4.1 requires bounding the Frechet derivatives of arbitrary high orders of composite functions. We rely on the generalization of Faà di Bruno's formula proposed by Clark and Houssineau [2013] to carry out the computation.

Let $F : \mathbb{R}^{d_2} \to \mathbb{R}^{d_3}$ and $G : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ be two differentiable functions up to order $k \geq 1$. Denote by $\Omega(k)$ the set of partitions of $\{1, \ldots, k\}$, and write $|\cdot|$ for the cardinality of a set. For any $z := (\delta_1, \ldots, \delta_k) \in (\mathbb{R}^{d_1})^k$, $x \in \mathbb{R}^{d_1}$, and $\omega := \{\omega_1, \ldots, \omega_{|\omega|}\} \in \Omega(k)$, we define $z^G_{\omega_i}(x) := G^{(|\omega_i|)}\left[(\delta_j)_{j\in\omega_i}\right]$ for every $1 \leq i \leq |\omega|$. Then, according to Theorem 2 in [Clark and Houssineau, 2013],

$$(F \circ G)^{(k)}(x)[\delta_1, \ldots, \delta_k] = \sum_{\omega\in\Omega(k)} F^{(|\omega|)}(G(x)) \left[z^G_{\omega_1}(x), \ldots, z^G_{\omega_{|\omega|}}(x)\right]. \tag{14}$$

This results implies a chain rule on the operator norms of derivatives of composite functions, which will greatly simplify the computations of later proofs.

Proposition A.2. *Let $F : \mathbb{R}^{d_2} \to \mathbb{R}^{d_3}$ and $G : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ be two differentiable functions up to order $k \geq 1$. Then, for any $x \in \mathbb{R}^{d_1}$,*

$$\left\|\left(F \circ G\right)^{(k)}(x)\right\|_{op} \leq \sum_{\omega\in\Omega(k)} \left\|F^{(|\omega|)}\left(G(x)\right)\right\|_{op} \times \prod_{1\leq i\leq|\omega|} \left\|G^{(|\omega_i|)}(x)\right\|_{op}.$$

Proof. According to the triangle inequality and (14)

$$\left\|(F \circ G)^{(k)}(x)\right\|_{op} \leq \sum_{\omega\in\Omega(k)} \sup_{\|\delta_1\|,\ldots,\|\delta_k\|\leq 1} \left\|F^{(|\omega|)}(G(x)) \left[z^G_{\omega_1}(x), \ldots, z^G_{\omega_{|\omega|}}(x)\right]\right\|.$$

Using the linearity of the derivatives, we can write the right term of this inequality as,

$$\sum_{\omega\in\Omega(k)} \sup_{\|\delta_1\|,\ldots,\|\delta_k\|\leq 1} \left\|F^{(|\omega|)}(G(x)) \left[\frac{z^G_{\omega_1}(x)}{\left\|G^{(|\omega_1|)}(x)\right\|_{op}}, \ldots, \frac{z^G_{\omega_1}(x)}{\left\|G^{(|\omega_{|\omega|}|)}(x)\right\|_{op}}\right]\right\| \times \prod_{1\leq i\leq|\omega|} \left\|G^{(|\omega_i|)}(x)\right\|_{op}.$$

In addition, note that for any $x \in \mathbb{R}^{d_1}$,

$$\sup_{\|\delta_1\|,\ldots,\|\delta_k\|\leq 1} \left\|z^G_{\omega_i}(x)\right\| \leq \left\|G^{(|\omega_i|)}(x)\right\|_{op}.$$

Therefore,

$$\left\|\left(F \circ G\right)^{(k)}(x)\right\|_{op} \leq \sum_{\omega\in\Omega(k)} \left\|F^{(|\omega|)}\left(G(x)\right)\right\|_{op} \times \prod_{1\leq i\leq|\omega|} \left\|G^{(|\omega_i|)}(x)\right\|_{op}.$$

□

B Proofs of the main results

This sections details all the mathematical proofs of the paper.

Proof of Lemma 3.2. Let us start with a preliminary remark. For any $v \in L^2_V$, it follows from Assumption 3.1 that $\int_0^1 \|v_t\|_{p,\infty} \mathrm{d}t \leq c_V \int_0^1 \|v_t\|_V \mathrm{d}t$. Besides, by Cauchy-Schwarz inequality $\int_0^1 \|v_t\|_V \mathrm{d}t \leq \|v\|_{L^2_V}$, leading to $\int_0^1 \|v_t\|_{p,\infty} \mathrm{d}t \leq c_V \|v\|_{L^2_V}$.

We now turn to the proof. Recall that by definition $\phi^v_t(x) = x + \int_0^t v_s \circ \phi^v_s(x) \mathrm{d}s$. Consequently, by the triangle inequality we have for any compact set $K \subset \mathbb{R}^d$ that

$$\sup_{t\in[0,1],x\in K} \|\phi^v_t(x)\| \leq \sup_{x\in K} \|x\| + \int_0^1 \|v_s\|_\infty \mathrm{d}s \leq \sup_{x\in K} \|x\| + c_V \|v\|_{L^2_V}.$$

Therefore,

$$\sup_{v\in L^2_{V,M},t\in[0,1],x\in K} \|\phi^v_t(x)\| \leq \sup_{x\in K} \|x\| + c_V M.$$

Moreover, combining to Theorem 5 in [Glaunes, 2005] with the preliminary remark, we know that for any $1 \leq k \leq p$, there exist two positive constants c_k and c'_k such that for any $v \in L^2_V$,

$$\sup_{t\in[0,1]} \left\| (\phi^v_t)^{(k)} \right\|_\infty \leq c_k \exp\left(c'_k \|v\|_{L^2_V} \right).$$

Hence,

$$\sup_{v\in L^2_{V,M},t\in[0,1]} \left\| (\phi^v_t)^{(k)} \right\|_\infty \leq c_k \exp\left(c'_k M \right).$$

Then, setting

$$R((K,d);(V,p);M) := \max\left\{ \max_{1\leq k\leq p} \{c_k \exp(c'_k M)\}, \sup_{x\in K} \|x\| + c_V M \right\}$$

concludes the proof. □

Proof of Lemma 4.2. Let μ and ν be probability measures on a compact set $K \subset \mathbb{R}^d$. In a first time, let us show that optimal potentials $(f,g) \in \mathcal{C}(K,\mathbb{R}) \times \mathcal{C}(K,\mathbb{R})$ for $\mathcal{T}_{C,\varepsilon}(\mu,\nu)$ can be chosen as universally-bounded Lipschitz functions. The optimality condition on the potentials (see for instance [Genevay, 2019]) can be written as,

$$\exp\left(-\frac{f(x)}{\varepsilon}\right) = \int_K \exp\left(\frac{g(y) - C(x,y)}{\varepsilon}\right) \mathrm{d}\nu(y).$$

Remark that since C is continuously differentiable, f is therefore continuously differentiable. Differentiating both sides of this expression leads to,

$$\nabla f(x) = \int_K \nabla_1 C(x,y) \exp\left(\frac{f(x) + g(y) - C(x,y)}{\varepsilon}\right) \mathrm{d}\nu(y),$$

where ∇_1 denotes the gradient with respect to x, the first variable of C. Let us define $\Gamma^{f,g}_{C,\varepsilon}(x,y) := \exp\left(\frac{f(x)+g(y)-C(x,y)}{\varepsilon}\right)$. According to the primal-dual relationship [Genevay, 2019], an optimal solution π to the primal problem has the expression,

$$\mathrm{d}\pi(x,y) = \Gamma^{f,g}_{C,\varepsilon}(x,y) \mathrm{d}\mu(x) \mathrm{d}\nu(y).$$

Since by definition $\pi \in \Pi(\mu,\nu)$, we consequently obtain that $\int_K \Gamma^{f,g}_{C,\varepsilon}(x,y) \mathrm{d}\nu(y) = 1$. Therefore,

$$\|\nabla f\|_\infty \leq \sup_{x,y\in K} \|\nabla_1 C(x,y)\|.$$

As similar argument can be made for g. This shows that f and g are ℓ-Lipschitz with $\ell = \ell((K,d);C) > 0$. Now, note that for any constant $c \in \mathbb{R}$, the pair $(f + c, g - c)$ is still a pair of optimal potentials. As a

consequence, they can be chosen without loss of generality such that $f(x_0) = 0$ for a given $x_0 \in K$. Thus, using the Lipschitz property we get $f(x) \leq \ell \|x - x_0\|$, hence $\|f\|_\infty \leq \ell \operatorname{diam}(K)$. To bound g, we use Proposition 1 in [Genevay et al., 2019] which states that $\inf_{x\in K}\{f(x) - C(x,y)\} \leq g(y) \leq \sup_{x\in K}\{f(x) - C(x,y)\}$. This entails that $\|g\|_\infty \leq \|f\|_\infty + \sup_{x,y\in K} |C(x,y)| \leq \ell \operatorname{diam}(K) + \sup_{x,y\in K} |C(x,y)|$. All in all, there exists a constant $\ell_1 = \ell_1((K,d);C)$ such that f and g are ℓ_1-bounded and ℓ_1-Lipschitz continuous.

Analogously, one can bound the successive derivatives of f and g up to order q, the maximum order or differentiability of C, using Proposition 1 in [Genevay et al., 2019]. In particular, this result ensures that for any $1 \leq k \leq q$, both $\left\|f^{(k)}\right\|_\infty$ and $\left\|g^{(k)}\right\|_\infty$ are bounded by a polynomial in ε^{-1} whose coefficients depend only on C and K. This implies that there exists a constant $m = m((K,d);(C,q);\varepsilon) > 0$ such that f and g belong to $\mathcal{C}^q_m(K,\mathbb{R})$. □

Proof of Proposition 4.3. Let $m > 0$ and $R > 0$. Set $f, g \in \mathcal{C}^q_m(B_R,\mathbb{R})$. Note that the function $h^{f,g}_{C,\varepsilon}$ belongs to $\mathcal{C}^q(B_R \times B_R,\mathbb{R})$. In a first time, we do not focus on any data processing operations, and show that $h^{f,g}_{C,\varepsilon}$ and its derivatives up to order q are uniformly bounded. By definition,

$$h^{f,g}_{C,\varepsilon}(x,y) = f(x) + g(y) - \varepsilon \exp\left(\frac{f(x) + g(y) - C\big(x,y\big)}{\varepsilon}\right) + \varepsilon.$$

Before going further, we define the constant

$$C_\infty(R) := \max_{0\leq k\leq q}\left\{\sup_{(x,y)\in B_R\times B_R}\left\|C^{(k)}(x,y)\right\|_{op}\right\} \tag{15}$$

Then, using the triangle inequality and the bounds on f, g and C we obtain,

$$\left\|h^{f,g}_{C,\varepsilon}\right\|_\infty \leq \|f\|_\infty + \|g\|_\infty + \varepsilon \exp\left(\frac{\|f\|_\infty + \|g\|_\infty + C_\infty}{\varepsilon}\right) + \varepsilon \leq 2m + \varepsilon \exp\left(\frac{2m + C_\infty}{\varepsilon}\right) + \varepsilon.$$

Notice that the upper bound does not depend on the choice of f and g. We prove similar bounds for arbitrary high orders of derivatives using the chain rule. We divide the problem by studying the function,

$$\Gamma^{f,g}_{C,\varepsilon} : (x,y) \in B_R \times B_R \mapsto \exp\frac{f(x) + g(y) - C(x,y)}{\varepsilon},$$

which is κ-continuously differentiable. Using Proposition A.2 with $F = \exp$, we obtain for any $1 \leq k \leq q$,

$$\left\|\left(\Gamma^{f,g}_{C,\varepsilon}\right)^{(k)}(x,y)\right\|_{op} \leq \left|\Gamma^{f,g}_{C,\varepsilon}(x,y)\right| \sum_{\omega\in\Omega(k)} \prod_{1\leq i\leq|\omega|} \varepsilon^{-1}\left\|f^{(|\omega_i|)}(x) + g^{(|\omega_i|)}(y) - C^{(|\omega_i|)}(x,y)\right\|_{op},$$

Then,

$$\left\|\left(\Gamma^{f,g}_{C,\varepsilon}\right)^{(k)}\right\|_\infty \leq \exp\left(\frac{2m + C_\infty(R)}{\varepsilon}\right) \sum_{\omega\in\Omega(k)} \varepsilon^{-|\omega|}(2m + C_\infty(R))^{|\omega|}. \tag{16}$$

We now turn back to $h^{f,g}_{C,\varepsilon}$. Since $\left(h^{f,g}_{C,\varepsilon}\right)^{(k)} = f^{(k)} + g^{(k)} - \varepsilon\left(\Gamma^{f,g}_{C,\varepsilon}\right)^{(k)}$ we finally have

$$\left\|\left(h^{f,g}_{C,\varepsilon}\right)^{(k)}\right\|_\infty \leq 2m + \exp\left(\frac{2m + C_\infty(R)}{\varepsilon}\right) \sum_{\omega\in\Omega(k)} \varepsilon^{-|\omega|+1}(2m + C_\infty(R))^{|\omega|}.$$

By defining,

$$H_0(m;R;(C,q);\varepsilon) := (2m + \varepsilon) + \varepsilon \exp\left(\frac{2m + C_\infty(R)}{\varepsilon}\right) \times \max_{0\leq k\leq q}\left\{\sum_{\omega\in\Omega(k)} \varepsilon^{-|\omega|}(2m + C_\infty(R))^{|\omega|}\right\},$$

we conclude that

$$\left\| \left(h_{C,\varepsilon}^{f,g} \right)^{(k)} \right\|_{q,\infty} \leq H_0.$$

We now include data processing transformations. Set $T_1, T_2 \in \mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)$. It follows from the regularity of $h_{C,\varepsilon}^{f,g}$ that $h_{C,\varepsilon}^{f,g} \circ (T_1, T_2) \in \mathcal{C}^\kappa(\mathcal{X} \times \mathcal{X}, \mathbb{R})$. Since $T_1(x), T_2(y) \in B_R$, and because $h_{C,\varepsilon}^{f,g}$ is bounded by H_0 on $B_R \times B_R$, the function $h_{C,\varepsilon}^{f,g} \circ (T_1, T_2)$ is bounded on $\mathcal{X} \times \mathcal{X}$ regardless of the choice of f, g, T_1 and T_2. Here again, we use the chain rule to build higher-order bounds. From Proposition A.2 applied with $F = h_{C,\varepsilon}^{f,g}$ and $G = (T_1, T_2)$ it follows that for any $1 \leq k \leq \kappa$,

$$\left\| \left(h_{C,\varepsilon}^{f,g} \circ (T_1, T_2) \right)^{(k)}(x,y) \right\|_{op} \leq \sum_{\omega \in \Omega(k)} \left\| \left(h_{C,\varepsilon}^{f,g} \right)^{(|\omega|)} \circ (T_1, T_2)(x,y) \right\|_{op} \times \prod_{1 \leq i \leq |\omega|} \left\| (T_1, T_2)^{(|\omega_i|)}(x,y) \right\|_{op}. \quad (17)$$

Then, remark that for any $1 \leq k \leq \kappa$,

$$\begin{aligned}
\left\| (T_1, T_2)^{(k)}(x,y) \right\|_{op}^2 &= \sup_{\|\delta_i\| \leq 1} \left\| (T_1, T_2)^{(k)}(x,y)(\delta_1, \ldots, \delta_k)) \right\|^2 \\
&\leq \sup_{\|\delta_i\| \leq 1} \left\| T_1^{(k)}(x)(\delta_1, \ldots, \delta_k) \right\|^2 + \sup_{\|\delta_i\| \leq 1} \left\| T_2^{(k)}(y)(\delta_1, \ldots, \delta_k) \right\|^2 \\
&= \left\| T_1^{(k)}(x) \right\|_{op}^2 + \left\| T_2^{(k)}(y) \right\|_{op}^2 \\
&\leq 2R^2.
\end{aligned}$$

We can therefore bound the right term of (17), leading to

$$\left\| \left(h_{C,\varepsilon}^{f,g} \circ (T_1, T_2) \right)^{(k)} \right\|_{\infty} \leq \sum_{\omega \in \Omega(k)} H_0 \times \prod_{1 \leq i \leq |\omega|} \sqrt{2}R = H_0 \sum_{\omega \in \Omega(k)} (\sqrt{2}R)^{|\omega|}.$$

We conclude by defining

$$H(m; R; (C, q); \varepsilon, p) := H_0(m; R; (C, q); \varepsilon) \times \max_{0 \leq k \leq \kappa} \left\{ \sum_{\omega \in \Omega(k)} (\sqrt{2}R)^{|\omega|} \right\},$$

which leads to,

$$\left\| h_{C,\varepsilon}^{f,g} \circ (T_1, T_2) \right\|_{\kappa,\infty} \leq H.$$

Proof of Proposition 5.1. Let $\{v^n\}_{n \in \mathbb{N}}$ be a sequence of vector fields in L_V^2 weakly converging to some $v \in L_V^2$. Proposition 4 in [Glaunes, 2005] implies that for every $x \in \mathcal{X}$,

$$\left| \phi_1^{v^n}(x) - \phi_1^v(x) \right| \xrightarrow[n \to +\infty]{} 0. \quad (18)$$

Next, we aim at showing that this entails $\phi_1^{v^n}{}_\sharp \alpha \xrightarrow[n \to +\infty]{w} \phi_{1\sharp}^v \alpha$, where w denotes the weak* convergence of *probability measures*. Firstly, note that as a consequence of the uniform-boundedness principle (see Theorem 2.5 in [Rudin, 1991]), the weak convergence of $\{v^n\}_{n \in \mathbb{N}}$ to v implies that there exists $M > 0$ such that $\{v^n\}_{n \in \mathbb{N}} \cup \{v\} \subset L_{V,M}^2$. Hence, according Lemma 3.1, there exists some $R = R((K, d); (V, p); M) > 0$ such that the measures $\{\phi_1^{v^n}{}_\sharp \alpha\}_{n \in \mathbb{N}}$ and $\phi_{1\sharp}^v \alpha$ are all probabilities on B_R. Secondly, recall that showing the weak* convergence of amounts to check that for any bounded test functions $h \in \mathcal{C}(B_R, \mathbb{R})$ we have that $\int h \mathrm{d}(\phi_1^{v^n}{}_\sharp \alpha) \xrightarrow[n \to +\infty]{} \int h \mathrm{d}(\phi_{1\sharp}^v \alpha)$. Let $h \in \mathcal{C}(B_R, \mathbb{R})$ be a bounded function and use the push-forward change-of-variable formula to write $\int h \mathrm{d}(\phi_1^{v^n}{}_\sharp \alpha) = \int (h \circ \phi_1^{v^n}) \mathrm{d}\alpha$. By continuity of h and according to (18), the sequence of

functions $\{h \circ \phi_1^{v^n}\}_{n\in\mathbb{N}}$ converges point-wise to $h \circ \phi_1^v$. In addition, as h is bounded, this sequence is dominated by a constant. We can therefore apply the dominated convergence theorem to obtain that ${\phi_1^{v^n}}_\sharp \alpha \xrightarrow[n\to+\infty]{w} {\phi_1^v}_\sharp \alpha$.

We conclude the proof using Proposition 13 in [Feydy et al., 2019], which states that $\mathcal{T}_{C,\varepsilon}$ (and consequently $S_{C,\varepsilon}$) is weak* continuous w.r.t. each of its input measures, provided that the ground cost function C is Lipschitz on their compact domains. This condition readily follows from the continuity of the derivative of C on the compact set $B_R \times B_R$. Therefore, $v \mapsto S_{C,\varepsilon}({\phi_1^v}_\sharp \alpha, \beta)$ is weakly continuous on L^2_V. □

□

Proof of Proposition 5.3. Let $R > 0$. In a first time, we demonstrate the following Glivenko-Cantelli theorem (i):

$$\sup_{T_1,T_2\in\mathcal{C}_R^p(\mathcal{X},\mathbb{R}^d)} \left|\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - \mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)\right| \xrightarrow[n\to+\infty]{a.s.} 0.$$

In a second time, when $\kappa = \min\{p, q\} \geq d$, we show the following rate of convergence (ii):

$$\mathbb{E} \sup_{T_1,T_2\in\mathcal{C}_R^p(\mathcal{X},\mathbb{R}^d)} \left|\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - \mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)\right| \leq \frac{A}{\sqrt{n}},$$

where $A > 0$ is a constant. In both cases, the key idea of the proof is to note that the quantity

$$\sup_{T_1,T_2\in\mathcal{C}_R^p(\mathcal{X},\mathbb{R}^d)} \left|\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - \mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)\right|$$

is the supremum of a centered empirical process indexed by a class of smooth functions, and as such can be controlled via classical results from empirical process theory (see Section A.1).

Let T_1 and T_2 be two arbitrary functions in $\mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)$. By definition, the image sets $T_1(\mathcal{X})$ and $T_2(\mathcal{X})$ are contained in B_R. Thus, using the dual formulation, the entropic transportation costs can be written as,

$$\begin{aligned}
\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) &= \sup_{f,g\in\mathcal{C}(B_R,\mathbb{R})} (T_{1\sharp}\alpha_n \otimes T_{2\sharp}\beta_n)(h_{C,\varepsilon}^{f,g}),\\
\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta) &= \sup_{f,g\in\mathcal{C}(B_R,\mathbb{R})} (T_{1\sharp}\alpha \otimes T_{2\sharp}\beta)(h_{C,\varepsilon}^{f,g}).
\end{aligned}$$

We apply Lemma 4.1 with $\mu = T_{1\sharp}\alpha$ and $\nu = T_{2\sharp}\beta$ which are probability measures on B_R. This implies that there exists a constant $m = m(B_R; (C, q), \varepsilon) > 0$ such that

$$\begin{aligned}
\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) &= \sup_{f,g\in\mathcal{C}_m^q(B_R,\mathbb{R})} (T_{1\sharp}\alpha_n \otimes T_{2\sharp}\beta_n)(h_{C,\varepsilon}^{f,g})\\
&= \sup_{f,g\in\mathcal{C}_m^q(B_R,\mathbb{R})} (\alpha_n \otimes \beta_n)(h_{C,\varepsilon}^{f,g} \circ (T_1, T_2)),
\end{aligned}$$

where we used the push-forward change-of-variable formula. Proceeding similarly with the empirical measures we get,

$$\begin{aligned}
\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta) &= \sup_{f,g\in\mathcal{C}_m^q(B_R,\mathbb{R})} (T_{1\sharp}\alpha \otimes T_{2\sharp}\beta)(h_{C,\varepsilon}^{f,g})\\
&= \sup_{f,g\in\mathcal{C}_m^q(B_R,\mathbb{R})} (\alpha \otimes \beta)(h_{C,\varepsilon}^{f,g} \circ (T_1, T_2)),
\end{aligned}$$

Then, by using a classical error decomposition, we can control the difference between these two terms as follows,

$$\begin{aligned}
&|\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - \mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)| \leq\\
&\qquad \sup_{f,g\in\mathcal{C}_m^q(B_R,\mathbb{R})} |(\alpha_n \otimes \beta_n)(h_{C,\varepsilon}^{f,g} \circ (T_1, T_2)) - (\alpha \otimes \beta)(h_{C,\varepsilon}^{f,g} \circ (T_1, T_2))|.
\end{aligned}$$

After taking the supremum in $\mathcal{C}_R^p(\mathcal{X}, \mathbb{R}^d)$ on both sides of this inequality we get,

$$\sup_{T_1,T_2\in\mathcal{C}_R^p(\mathcal{X},\mathbb{R}^d)} \left|\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - \mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)\right| \leq$$
$$\sup_{T_1,T_2\in\mathcal{C}_R^p(\mathcal{X},\mathbb{R}^d);f,g\in\mathcal{C}_m^q(B_R,\mathbb{R})} \left|(\alpha_n\otimes\beta_n)(h_{C,\varepsilon}^{f,g}\circ(T_1,T_2)) - (\alpha\otimes\beta)(h_{C,\varepsilon}^{f,g}\circ(T_1,T_2))\right|.$$

The right term of this inequality can be seen as a centered empirical process indexed by the class of functions $\{h_{C,\varepsilon}^{f,g}\circ(T_1,T_2) \mid T_1,T_2 \in \mathcal{C}_R^p(\mathcal{X},\mathbb{R}^d); f,g \in \mathcal{C}_m^q(B_R,\mathbb{R})\}$. Empirical process theory provides convergence guarantees when the index class is regular enough. Besides, we know from Proposition 4.1 that there exists a constant $H := H(R;(C,q);\varepsilon,p) > 0$ such that this class is included in $\mathcal{C}_H^\kappa(\mathcal{X}\times\mathcal{X},\mathbb{R})$. Therefore,

$$\sup_{T_1,T_2\in\mathcal{C}_R^p(\mathcal{X},\mathbb{R}^d)} |\mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - \mathcal{T}_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)| \leq \sup_{h\in\mathcal{C}_H^\kappa(\mathcal{X}\times\mathcal{X},\mathbb{R})} |(\alpha_n\otimes\beta_n)(h) - (\alpha\otimes\beta)(h)|.$$

Let us set $\mathcal{H} := \mathcal{C}_H^\kappa(\mathcal{X}\times\mathcal{X},\mathbb{R})$. According to Corollary 2.7.2. and Theorem 2.4.1 in [Van Der Vaart and Wellner, 1996], $\mathcal{H}$ is a so-called $(\alpha\otimes\beta)$-*Glivenko-Cantelli* class of functions, meaning that

$$\sup_{h\in\mathcal{H}} |(\alpha_n\otimes\beta_n)(h) - (\alpha\otimes\beta)(h)| \xrightarrow[n\to+\infty]{a.s.} 0;$$

This implies (i). In addition, by Proposition A.1, if $\kappa \geq (2d)/2$ then there exists a positive constant $A := A(R;(C,q);\varepsilon;(\mathcal{X},d);p)$ such that,

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} |(\alpha_n\otimes\beta_n)(h) - (\alpha\otimes\beta)(h)|\right] \leq \frac{A}{\sqrt{n}}.$$

This proves (ii). □

Before proving Theorem 5.1, we need the next intermediary result:

Lemma B.1. *There exists $M = M(\lambda;(\mathcal{X},d);(C,q);\varepsilon) > 0$ such that*

$$\left\{\bigcup_{n\in\mathbb{N}} \operatorname*{arg\,min}_{v\in L_V^2} J_{\lambda,n}(v)\right\} \cup \operatorname*{arg\,min}_{v\in L_V^2} J_\lambda(v) \subseteq L_{V,M}^2.$$

Proof. The proof generalizes an argument made for a squared MMD in Theorem 16 from [Glaunes, 2005] to a Sinkhorn divergence. Let $n\in\mathbb{N}$ and set v^n a minimizer of $J_{\lambda,n}$. Notice that the vector flow uniformly equal to zero generates the identity function, that is $\phi_t^0 = I$ for any $t\in[0,1]$. Thus, by definition of a minimizer and by non-negativity of the Sinkhorn divergence, we readily have that

$$\lambda\|v^n\|_{L_V^2}^2 \leq J_{\lambda,n}(v^n) \leq J_{\lambda,n}(0) = S_{C,\varepsilon}(\alpha_n,\beta_n).$$

Therefore, $\|v^n\|_{L_V^2}^2 \leq \lambda^{-1} S_{C,\varepsilon}(\alpha_n,\beta_n)$. To conclude, let us bound uniformly the right-term of this inequality. According to Lemma 4.1 applied with α_n and β_n there exists a constant $m = m((\mathcal{X},d);(C,q);\varepsilon)$ such that,

$$\mathcal{T}_{C,\varepsilon}(\alpha_n,\beta_n) = \sup_{f,g\in\mathcal{C}_m^q(\mathcal{X},R)} (\alpha_n\otimes\beta_n)\left(h_{C,\varepsilon}^{f,g}\right).$$

Moreover, for any $x,y\in\mathcal{X}$,

$$h_{C,\varepsilon}^{f,g}(x,y) = f(x) + g(x) - \varepsilon e^{\frac{f(x)+g(y)-C(x,y)}{\varepsilon}} + \varepsilon \leq m + m + 0 + \varepsilon.$$

Thus,

$$\mathcal{T}_{C,\varepsilon}(\alpha_n, \beta_n) \leq 2m + \varepsilon.$$

The same bound holds for the two auto-correlation terms of the Sinkhorn divergence, namely $\mathcal{T}_{C,\varepsilon}(\alpha_n, \alpha_n)$ and $\mathcal{T}_{C,\varepsilon}(\beta_n, \beta_n)$. Therefore, the triangle inequality leads to

$$\mathcal{S}_{C,\varepsilon}(\alpha_n, \beta_n) \leq 4m + 2\varepsilon.$$

Consequently,

$$\|v^n\|^2_{L^2_V} \leq \frac{4m + 2\varepsilon}{\lambda}.$$

To conclude, we set $M(\lambda; (\mathcal{X}, d); (C, q); \varepsilon) := \sqrt{\frac{4m+2\varepsilon}{\lambda}}$. Note that this bound does not depend on n. As such, the minima $\{v^n\}_{n\in\mathbb{N}}$ all belong to $L^2_{V,M}$. A similar reasoning for v^* a minimizer of J_λ shows that all the minimizers of J_λ also belong to $L^2_{V,M}$. □

Proof of Theorem 5.4. Let $M > 0$ be arbitrary (for now). Set $v \in L^2_{V,M}$ and compute

$$|J_{\lambda,n}(v) - J_\lambda(v)| = |S_{C,\varepsilon}(\phi^v_{1\sharp}\alpha_n, \beta_n) - S_{C,\varepsilon}(\phi^v_{1\sharp}\alpha, \beta)|.$$

According to Lemma 3.1, there exists a constant $R = R((\mathcal{X}, d); (V, p); M)$ such that for any $\phi \in \{\phi^v_t \mid t \in [0, 1], v \in L^2_{V,M}\}$, the restriction $\phi|_\mathcal{X}$ and the identity function I both belong to $\mathcal{C}^p_R(\mathcal{X}, \mathbb{R}^d)$. This leads to

$$\sup_{v\in L^2_{V,M}} |J_{\lambda,n}(v) - J_\lambda(v)| \leq \sup_{T_1,T_2\in\mathcal{C}^p_R(\mathcal{X},\mathbb{R}^d)} |S_{C,\varepsilon}(T_{1\sharp}\alpha_n, T_{2\sharp}\beta_n) - S_{C,\varepsilon}(T_{1\sharp}\alpha, T_{2\sharp}\beta)|. \tag{19}$$

From here, let us demonstrate the convergence of the minima, that is item (i). According to Lemma B.1, there exists $M = M(\lambda; (\mathcal{X}, d); (C, q); \varepsilon) > 0$ such that all the minimizers of $J_{\lambda,n}$ belong to $L^2_{V,M}$. Next, we show that any weakly-converging subsequences of $\{v^n\}_{n\in\mathbb{N}}$ tend to a minimizer of J_λ. Set v^* a minimizer of J_λ, and let $\{u^n\}_{n\in\mathbb{N}}$ be a subsequence with limit u. First, let's show that $\lim_{n\to+\infty} J_{\lambda,n}(u^n) = J_\lambda(u)$. By the triangle inequality, $|J_{\lambda,n}(u^n) - J_\lambda(u)| \leq |J_{\lambda,n}(u^n) - J_\lambda(u^n)| + |J_\lambda(u^n) - J_\lambda(u)|$. The first term tends to zero by Proposition 5.2 and Equation 19 specified with $M(\lambda; (\mathcal{X}, d); (C, q); \varepsilon)$, while the second term tends to zero according to Proposition 5.1 which ensures the weak continuity of J_λ. Second, note that the optimality condition entails that $J_{\lambda,n}(u^n) \leq J_{\lambda,n}(v^*)$, and that $\lim_{n\to+\infty} J_{\lambda,n}(v^*) = J_\lambda(v^*)$. Then, at the limit $J_\lambda(u) \leq J_\lambda(v^*)$, meaning that u is a minimizer of J_λ. Therefore, any weakly-converging subsequence $\{u^n\}_{n\in\mathbb{N}}$ of $\{v^n\}_{n\in\mathbb{N}}$ tends to a minimizer u of J_λ.

To conclude on the convergence of the generated diffeomorphisms, we rely on Remark 1 in [Glaunes, 2005], stating that

$$\sup_{t\in[0,1]} \left\{ \left\| \phi^{u^n}_t - \phi^u_t \right\|_\infty + \left\| (\phi^{u^n}_t)^{-1} - (\phi^u_t)^{-1} \right\|_\infty \right\} \leq 2c_V \|u^n - u\|_{L^2_V} \exp\left(c_V \|u\|_{L^2_V} \right).$$

We showed that $\|u^n - u\|_{L^2_V} \xrightarrow[n\to\infty]{} 0$. Consequently, the upper bound tends to zero as n increases to infinity. This completes the proof of (i).

Item (ii) readily follows from Proposition 5.2 stating that if $\kappa \geq d$, then there exists for any $M > 0$ a constant $A = A(\lambda; (\mathcal{X}, d); (C, q); \varepsilon; (V, p), M) > 0$ such that

$$\mathbb{E}\left[\sup_{v\in L^2_{V,M}} |J_{\lambda,n}(v) - J_\lambda(v)| \right] \leq \frac{A}{\sqrt{n}}.$$

To conclude, recall that both $\{v^n\}_{n\in\mathbb{N}}$ and v^* belong to $L^2_{V,M}$ for the constant M from Lemma B.1, and apply the classical deviation inequality

$$J_\lambda(v^n) - J_\lambda(v^*) \leq 2 \sup_{v\in L^2_{V,M}} |J_{\lambda,n}(v) - J_\lambda(v)|.$$

□

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

M. Bauer, S. Joshi, and K. Modin. Diffeomorphic density matching by optimal information transport. *SIAM Journal on Imaging Sciences*, 8(3):1718–1751, 2015.

M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

D. E. Clark and J. Houssineau. Faa di bruno's formula for variational calculus, 2013.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

E. del Barrio, A. Gonzalez-Sanz, J.-M. Loubes, and J. Niles-Weed. An improved central limit theorem and fast convergence rates for entropic transportation costs, 2022. URL `https://arxiv.org/abs/2204.09105`.

J. Feydy and A. Trouvé. Global divergences between measures: from hausdorff distance to optimal transport. In *International Workshop on Shape in Medical Imaging*, pages 102–115. Springer, 2018.

J. Feydy, B. Charlier, F.-X. Vialard, and G. Peyré. Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 291–299. Springer, 2017.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

A. Genevay. *Entropy-Regularized Optimal Transport for Machine Learning. (Régularisation Entropique du Transport Optimal pour le Machine Learning).* PhD thesis, PSL Research University, Paris, France, 2019. URL `https://tel.archives-ouvertes.fr/tel-02319318`.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

J. Glaunes. *Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l'anatomie numérique.* PhD thesis, 2005.

J. Glaunes, A. Trouvé, and L. Younes. Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

S. C. Joshi and M. I. Miller. Landmark matching via large deformation diffeomorphisms. *IEEE transactions on image processing*, 9(8):1357–1370, 2000.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

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

W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1991. ISBN 9780070542365.

F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013.

A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2008. ISBN 978-3-540-71049-3. OCLC: ocn244421231.

L. Younes. *Shapes and diffeomorphisms*, volume 171. Springer, 2010.

L. Younes. Diffeomorphic learning. *Journal of Machine Learning Research*, 21(220):1–28, 2020. URL `http://jmlr.org/papers/v21/18-415.html`.