


# STOCHASTIC GRADIENT DESCENT FOR BARYCENTERS IN WASSERSTEIN SPACE

JULIO BACKHOFF , \* *Universität Wien*  
JOAQUIN FONTBONA, \*\* \*\*\* *Universidad de Chile*  
GONZALO RIOS, \*\*\*\* *NoiseGrasp SpA*  
FELIPE TOBAR, \*\* \*\*\*\*\* *Universidad de Chile*

## Abstract

We present and study a novel algorithm for the computation of 2-Wasserstein population barycenters of absolutely continuous probability measures on Euclidean space. The proposed method can be seen as a stochastic gradient descent procedure in the 2-Wasserstein space, as well as a manifestation of a law of large numbers therein. The algorithm aims to find a Karcher mean or critical point in this setting, and can be implemented ‘online’, sequentially using independent and identically distributed random measures sampled from the population law. We provide natural sufficient conditions for this algorithm to almost surely converge in the Wasserstein space towards the population barycenter, and we introduce a novel, general condition which ensures uniqueness of Karcher means and, moreover, allows us to obtain explicit, parametric convergence rates for the expected optimality gap. We also study the mini-batch version of this algorithm, and discuss examples of families of population laws to which our method and results can be applied. This work expands and deepens ideas and results introduced in an early version of Backhoff-Veraguas *et al.* (2022), in which a statistical application (and numerical implementation) of this method is developed in the context of Bayesian learning.

**Keywords:** Wasserstein distance; Wasserstein barycenter; Fréchet mean; Karcher mean; gradient descent; stochastic gradient descent

2020 Mathematics Subject Classification: Primary 62L20  
Secondary 60F15; 65C35

## 1. Introduction

Let  $\mathcal{P}(\mathcal{X})$  denote the space of Borel probability measures over a Polish space  $\mathcal{X}$ . Given  $\mu, \nu \in \mathcal{P}(\mathcal{X})$ , denote by  $\Gamma(\mu, \nu) := \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \gamma(\mathrm{d}x, \mathcal{X}) = \mu(\mathrm{d}x), \gamma(\mathcal{X}, \mathrm{d}y) = \nu(\mathrm{d}y)\}$  the set of couplings (transport plans) with marginals  $\mu$  and  $\nu$ . For a fixed, compatible, complete metric  $d$ , and given a real number  $p \geq 1$ , we define the  $p$ -Wasserstein space by  $\mathcal{W}_p(\mathcal{X}) := \{\eta \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} d(x_0, x)^p \eta(\mathrm{d}x) < \infty, \text{ for some } x_0\}$ . Accordingly, the  $p$ -Wasserstein distance

---

Received 3 March 2022; accepted 16 April 2024.

\* Postal address: Oskar-Morgenstern-Platz 1, Vienna 1090, Austria. Email: [julio.backhoff@univie.ac.at](mailto:julio.backhoff@univie.ac.at)

\*\* Postal address: Beauchef 851, Santiago, Chile.

\*\*\* Email: [fontbona@dim.uchile.cl](mailto:fontbona@dim.uchile.cl)

\*\*\*\* Email: [grios@noisegrasp.com](mailto:grios@noisegrasp.com)

\*\*\*\*\* Email: [ftobar@uchile.cl](mailto:ftobar@uchile.cl)

© The Author(s), 2024. Published by Cambridge University Press on behalf of Applied Probability Trust. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

between measures  $\mu, \nu \in \mathcal{W}_p(\mathcal{X})$  is given by

$$W_p(\mu, \nu) = \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/p}. \quad (1)$$

When  $p = 2$ ,  $\mathcal{X} = \mathbb{R}^q$ ,  $d$  is the Euclidean distance, and  $\mu$  is absolutely continuous, which is the setting we will soon adopt for the remainder of the paper, Brenier's theorem [50, Theorem 2.12(ii)] establishes the uniqueness of a minimizer for the right-hand side of (1). Furthermore, this optimizer is supported on the graph of the gradient of a convex function. See [6, 51] for further general background on optimal transport.

We recall now the definition of the Wasserstein population barycenter.

**Definition 1.** Given  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ , write  $V_p(\bar{m}) := \int_{\mathcal{P}(\mathcal{X})} W_p(m, \bar{m})^p \Pi(\mathrm{d}m)$ . Any measure  $\hat{m} \in \mathcal{W}_p(\mathcal{X})$  that is a minimizer of the problem  $\inf_{\bar{m} \in \mathcal{P}(\mathcal{X})} V_p(\bar{m})$  is called a  $p$ -Wasserstein population barycenter of  $\Pi$ .

Wasserstein barycenters were first introduced and analyzed in [1], in the case when the support of  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  is finite and  $\mathcal{X} = \mathbb{R}^q$ . More generally, [12, 38] considered the so-called *population barycenter*, i.e. the general case in Definition 1 where  $\Pi$  may have infinite support. These works addressed, among others, the basic questions of the existence and uniqueness of solutions. See also [35] for the Riemannian case. The concept of the Wasserstein barycenter has been extensively studied from both theoretical and practical perspectives over the last decade: we refer the reader to the overview in [43] for statistical applications, and to [17, 26, 27, 44] for computational aspects of optimal transport and applications in machine learning.

In this article we develop a stochastic gradient descent (SGD) algorithm for the computation of 2-Wasserstein population barycenters. The method inherits features of the SGD rationale, in particular:

- It exhibits a reduced computational cost compared to methods based on the direct formulation of the barycenter problem using all the available data, as SGD only considers a limited number of samples of  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  per iteration.
- It is *online* (or *continual*, as referred to in the machine learning community), meaning that it can incorporate additional datapoints sequentially to update the barycenter estimate whenever new observations becomes available. This is relevant even when  $\Pi$  is finitely supported.
- Conditions ensuring convergence towards the Wasserstein barycenter as well as finite-dimensional convergence rates for the algorithm can be provided. Moreover, the variance of the gradient estimators can be reduced by using mini-batches.

From now on we make the following assumption.

**Assumption 1.**  $\mathcal{X} = \mathbb{R}^q$ ,  $d$  is the squared Euclidean metric, and  $p = 2$ .

We denote by  $\mathcal{W}_{2,\mathrm{ac}}(\mathcal{X})$  the subspace of  $\mathcal{W}_2(\mathcal{X})$  of absolutely continuous measures with finite second moment and, for any  $\mu \in \mathcal{W}_{2,\mathrm{ac}}(\mathcal{X})$  and  $\nu \in \mathcal{W}_2(\mathcal{X})$  we write  $T_\mu^\nu$  for the ( $\mu$ -almost sure, unique) gradient of a convex function such that  $T_\mu^\nu(\mu) = \nu$ . Notice that  $x \mapsto T_\mu^\nu(x)$  is a  $\mu$ -almost sure (a.s.) defined map and that we use throughout the notation  $T_\mu^\nu(\rho)$  for the image measure/law of this map under the measure  $\rho$ .

Recall that a set  $B \subset \mathcal{W}_{2,\text{ac}}(\mathcal{X})$  is *geodesically convex* if, for every  $\mu, \nu \in B$  and  $t \in [0, 1]$  we have  $((1-t)I + tT_\mu^\nu)(\mu) \in B$ , with  $I$  denoting the identity operator. We will also assume the following condition for most of the article.

**Assumption 2.**  $\Pi$  gives full measure to a geodesically convex  $W_2$ -compact set  $K_\Pi \subset \mathcal{W}_{2,\text{ac}}(\mathcal{X})$ .

In particular, under Assumption 2 we have  $\int_{\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} d(x, x_0)^2 m(dx) \Pi(dm) < \infty$  for all  $x_0$ . Moreover, for each  $\nu \in \mathcal{W}_2(\mathcal{X})$  and  $\Pi(dm)$  for almost every (a.e.)  $m$ , there is a unique optimal transport map  $T_m^\nu$  from  $m$  to  $\nu$  and, by [38, Proposition 6], the 2-Wasserstein population barycenter is unique.

**Definition 2.** Let  $\mu_0 \in K_\Pi$ ,  $m_k \stackrel{\text{i.i.d.}}{\sim} \Pi$ , and  $\gamma_k > 0$  for  $k \geq 0$ . We define the stochastic gradient descent (SGD) sequence by

$$\mu_{k+1} := [(1 - \gamma_k)I + \gamma_k T_{\mu_k}^{m_k}](\mu_k) \quad \text{for } k \geq 0. \quad (2)$$

The reasons why we can truthfully refer to the above sequence as stochastic gradient descent will become apparent in Sections 2 and 3. We stress that the sequence is a.s. well defined, as we can show by induction that  $\mu_k \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$  a.s. thanks to Assumption 2. We also refer to Section 3 for remarks on the measurability of the random maps  $\{T_{\mu_k}^{m_k}\}_k$  and sequence  $\{\mu_k\}_k$ .

Throughout the article, we assume the following conditions on the steps  $\gamma_k$  in (2), commonly required for the convergence of SGD methods:

$$\sum_{k=1}^{\infty} \gamma_k^2 < \infty, \quad (3)$$

$$\sum_{k=1}^{\infty} \gamma_k = \infty. \quad (4)$$

In addition to barycenters, we will need the concept of *Karcher means* (cf. [55]), which, in the setting of the optimization problem in  $\mathcal{W}_{2,\text{ac}}(\mathcal{X})$  considered here, can be intuitively understood as an analogue of a critical point of a smooth function in Euclidean space (see the discussion in Section 2).

**Definition 3.** Given  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ , we say that  $\mu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$  is a Karcher mean of  $\Pi$  if  $\mu(\{x: x = \int_{m \in \mathcal{P}(\mathcal{X})} T_\mu^m(x) \Pi(dm)\}) = 1$ .

It is known that any 2-Wasserstein barycenter is a Karcher mean, though the latter is in general a strictly larger class; see [4] or Example 1. However, if there is a unique Karcher mean, then it must coincide with the unique barycenter. We can now state the main result of the article.

**Theorem 1.** We assume Assumptions 1 and 2, conditions (3) and (4), and that the 2-Wasserstein barycenter  $\hat{\mu}$  of  $\Pi$  is the unique Karcher mean. Then, the SGD sequence  $\{\mu_k\}_k$  in (2) is a.s.  $W_2$ -convergent to  $\hat{\mu} \in K_\Pi$ .

An interesting aspect of Theorem 1 is that it hints at a law of large numbers (LLN) on the 2-Wasserstein space. Indeed, in the conventional LLN, for i.i.d. samples  $X_i$  the summation  $S_k := (1/k) \sum_{i \leq k} X_i$  can be expressed as

$$S_{k+1} = \frac{1}{k+1} X_{k+1} + \left(1 - \frac{1}{k+1}\right) S_k.$$

Therefore, if we rather think of sample  $X_k$  as a measure  $m_k$  and of  $S_k$  as  $\mu_k$ , we immediately see the connection with

$$\mu_{k+1} := \left[ \frac{1}{k+1} T_{\mu_k}^{m_k} + \left( 1 - \frac{1}{k+1} \right) I \right] (\mu_k),$$

obtained from (2) when we take  $\gamma_k = 1/(k+1)$ . The convergence in Theorem 1 can thus be interpreted as an analogy to the convergence of  $S_k$  to the mean of  $X_1$  (we thank Stefan Schrott for this observation).

In order to state our second main result, Theorem 2, we first introduce a new concept.

**Definition 4.** Given  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  we say that a Karcher mean  $\mu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$  of  $\Pi$  is pseudo-associative if there exists  $C_\mu > 0$  such that, for all  $\nu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$ ,

$$W_2^2(\mu, \nu) \leq C_\mu \int_{\mathcal{X}} \left| \int_{\mathcal{P}(\mathcal{X})} (T_\nu^m(x) - x) \Pi(dm) \right|^2 \nu(dx). \quad (5)$$

Since the term on the right-hand side of (5) vanishes for any Karcher mean  $\nu$ , the existence of a pseudo-associative Karcher mean implies  $\mu = \nu$ , hence uniqueness of Karcher means. We will see that, moreover, the existence of a pseudo-associative Karcher mean implies a Polyak–Łojasiewicz inequality for the functional minimized by the barycenter, see (20). This in turn can be utilized to obtain convergence rates for the expected optimality gap in a similar way to the Euclidean case. While the pseudo-associativity condition is a strong requirement, in the following result we are able to weaken Assumption 2 into the following milder assumption.

**Assumption 2'.**  $\Pi$  gives full measure to a geodesically convex set  $K_\Pi \subset \mathcal{W}_{2,\text{ac}}(\mathcal{X})$  that is  $\mathcal{W}_2$ -bounded (i.e.  $\sup_{m \in K_\Pi} \int |x|^2 m(dx) < \infty$ ).

Clearly this condition is equivalent to requiring that the support of  $\Pi$  be  $\mathcal{W}_2$ -bounded and consist of absolutely continuous measures.

We now present our second main result.

**Theorem 2.** We assume Assumptions 1 and 2', and that the 2-Wasserstein barycenter  $\hat{\mu}$  of  $\Pi$  is a pseudo-associative Karcher mean. Then,  $\hat{\mu}$  is the unique barycenter of  $\Pi$ , and the SGD sequence  $\{\mu_k\}_k$  in (2) is a.s. weakly convergent to  $\hat{\mu} \in K_\Pi$  as soon as (3) and (4) hold. Moreover, for every  $a > C_{\hat{\mu}}^{-1}$  and  $b \geq a$  there exists an explicit constant  $C_{a,b} > 0$  such that, if  $\gamma_k = a/(b+k)$  for all  $k \in \mathbb{N}$ , the expected optimality gap satisfies

$$\mathbb{E}[F(\mu_k) - F(\hat{\mu})] \leq \frac{C_{a,b}}{b+k}.$$

Let us explain the reason for the terminology ‘pseudo-associative’ we have chosen. This comes from the fact that the inequality in Definition 4 holds true (with equality and  $C_\mu = 1$ ) as soon as the associativity property  $T_\mu^m(x) = T_\nu^m \circ T_\mu^\nu(x)$  holds  $\mu(dx)$ -a.s. for each pair  $\nu, m \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$ ; see Remark 2. The previous identity was assumed to hold for the results proved in [12], and is always valid in  $\mathbb{R}$ , since the composition of monotone functions is monotone. Further examples where the associativity property holds are discussed in Section 6. Thus, in all those settings, (5) and Theorem 2 hold as soon as Assumption 2' is granted, and, further, we have the explicit LLN-like expression  $\mu_{k+1} = ((1/k) \sum_{i=1}^k T_{\mu_0}^{m_i})(\mu_0)$ . As regards the more general pseudo-associativity property, we will see that it holds, for instance, in the Gaussian framework studied in [19], and in certain classes of scatter-location families, which we discuss in Section 6.

The remainder of the paper is organized as follows:

- In Section 2 we recall basic ideas and results on gradient descent algorithms for Wasserstein barycenters.
- In Section 3 we prove Theorem 1, providing the technical elements required to that end.
- In Section 4 we discuss the notion of pseudo-associative Karcher means and their relation to the so-called Polyak–Łojasiewicz and variance inequalities, and we prove Theorem 2.
- In Section 5 we introduce the mini-batch version of our algorithm, discuss how this improves the variance of gradient-type estimators, and state extensions of our previous results to that setting.
- In Section 6 we consider closed-form examples and explain how in these cases the existence of pseudo-associative Karcher means required in Theorem 2 can be guaranteed. We also explore certain properties of probability distributions that are ‘stable’ under the operation of taking their barycenters.

In the remainder of this introduction we provide independent discussions on various aspects of our results (some of them suggested by the referees) and on related literature.

### 1.1. On the assumptions

Although Assumption 2 might appear strong at first sight, it can be guaranteed in suitable parametric situations (e.g. Gaussian, or the scatter-location setting recalled in Section 6.4) or, more generally, under moment and density constraints on the measures in  $K_\Pi$ . For instance, if  $\Pi$  is supported on a finite set of measures with finite Boltzmann entropy, then Assumption 2 is guaranteed. More generally, if the support of  $\Pi$  is a 2-Wasserstein compact set and the Boltzmann entropy is uniformly bounded on it, then Assumption 2 is fulfilled too: see Lemma 3.

The conditions of Assumption 2 and the uniqueness of a Karcher mean in Theorem 1 are natural substitutes for, respectively, the compactness of SGD sequences and the uniqueness of critical points, which hold under their usual sets of assumptions ensuring the convergence of SGD in Euclidean space to a minimizer, e.g. some growth control and certain strict convexity-type conditions at a minimum. The usual reasoning underlying the convergence analysis of SGD in Euclidean spaces, however, seem not to be applicable in our context, as the functional  $\mathcal{W}_{2,\text{ac}}(\mathcal{X}) \ni \mu \mapsto F(\mu) := \int W_2(m, \bar{m})^2 \Pi(dm)$  is not convex, in fact not even  $\alpha$ -convex for some  $\alpha \in \mathbb{R}$ , when  $\mathcal{W}_{2,\text{ac}}(\mathcal{X})$  is endowed with its almost Riemannian structure induced by optimal transport (see [6, Chapter 7.2]). The function  $F$  is also not  $\alpha$ -convex for some  $\alpha \in \mathbb{R}$  when we use generalized geodesics (see [6, Chapter 9.2]). In fact, SGD in finite-dimensional Riemannian manifolds could provide a more suitable framework to draw inspiration from; see, e.g., [13]. Ideas useful in that setting seem not straightforward to leverage since that work either assumes negative curvature (while  $\mathcal{W}_{2,\text{ac}}(\mathcal{X})$  is positively curved), or that the functional to be minimized be rather smooth and have bounded derivatives of first and second order.

The assumption that  $K_\Pi$  is contained in a subset of  $\mathcal{W}_2(\mathcal{X})$  of absolutely continuous probability measures, and hence the existence of optimal transport maps between elements of  $K_\Pi$ , appears as a more structural requirement, as it is needed to construct the iterations in (2). Indeed, by dealing with an extended notion of Karcher mean, it is in principle also possible to

define an analogous iterative scheme in a general setting, including in particular the case of discrete laws, and thus to relax to some extent the absolute continuity requirement. However, this introduces additional technicalities, and we unfortunately were not able to provide conditions ensuring the convergence of the method in reasonably general situations. For completeness of the discussion, we sketch the main ideas of this possible extension in the Appendix.

## 1.2. Uniqueness of Karcher means

Regarding situations where uniqueness of Karcher means can be granted, we refer to [43, 55] for sufficient conditions when the support of  $\Pi$  is finite, based on the regularity theory of optimal transport, and to [12] for the case of an infinite support, under rather strong assumptions. We remark also that in one dimension, the uniqueness of Karcher means is known to hold without further assumptions. A first counterexample where this uniqueness is not guaranteed is given in [4]; see also the simplified Example 1. A general understanding of the uniqueness of Karcher means remains an open and interesting challenge, however. This is not only relevant for the present work, but also for the (non-stochastic) gradient descent method of [55] and the fixed-point iterations of [4]. The notion of pseudo-associativity introduced in Definition 4 provides an alternative viewpoint on this question, complementary to the aforementioned ones, which might deserve being further explored.

## 1.3. Gradient descents in Wasserstein space

Gradient descent (GD) in Wasserstein space was introduced as a method to compute barycenters in [4, 55]. The SGD method we develop here was introduced in an early version of [9] ([arXiv:1805.10833](https://arxiv.org/abs/1805.10833)) as a way to compute Bayesian estimators based on 2-Wasserstein barycenters. In view of the independent, theoretical interest of this SGD method, we decided to separately present this algorithm here, along with a deeper analysis and more complete results on it, and devote [9] exclusively to its statistical application and implementation. More recently, [19] obtained convergence rates for the expected optimality gap for these GD and SGD methods in the case of Gaussian families of distributions with uniformly bounded, uniformly positive-definite covariance matrices. This relied on proving a Polyak–Łojasiewicz inequality, and also derived quantitative convergence bounds in  $W_2$  for the SGD sequence, relying on a variance inequality, which was shown to hold under general, though strong, conditions on the dual potential of the barycenter problem (verified under the assumptions in that paper). We refer to [16] for more on the variance inequality.

The Riemannian-like structure of the Wasserstein space and its associated gradient have been utilized in various other ways with statistical or machine learning motivations in recent years; see, e.g., [21] for particle-like approximations, [34] for a sequential first-order method, and [18, 39] for information geometry perspectives.

## 1.4. Computational aspects

Implementing our SGD algorithm requires, in general, computing or approximating optimal transport maps between two given absolutely continuous distributions (some exceptions where explicit closed-form maps are available are given in Section 6). During the last two decades, considerable progress has been made on the numerical resolution of the latter problem through partial differential equation methods [7, 10, 14, 40] and, more recently, through entropic regularization approaches [25, 27, 29, 41, 44, 49] crucially relying on the Sinkhorn algorithm [47, 48], which significantly speeds up the approximate resolution of the problem.

It is also possible to approximate optimal transport maps via estimators built from samples (based on the Sinkhorn algorithm, plug-in estimators, or using stochastic approximations) [11, 28, 33, 42, 45]. We also refer to [37] for an overview and comparison of sample-free methods for continuous distributions, for instance based on neural networks (NNs).

### 1.5. Applications

The proposed method to compute Wasserstein barycenters is well suited to situations where a population law  $\Pi$  on infinitely many probability measures is considered. The Bayesian setting addressed in [9] provides a good example of such a situation, i.e. where we need to compute the Wasserstein barycenter of a prior/posterior distribution  $\Pi$  on models that has a possibly infinite support.

Further instances of population laws  $\Pi$  with infinite support arise in the context of Bayesian deep learning [54], in which an NN's weights are sampled from a given law (e.g. Gaussian, uniform, or Laplace); in the case that these NNs parametrize probability distributions, the collection of resulting probability laws is distributed according to a law  $\Pi$  with possibly infinite support. Another example is variational autoencoders [36], which model the parameters of a law by a simple random variable (usually Gaussian) that is then passed to a *decoder* NN. In both cases, the support of  $\Pi$  is infinite naturally. Furthermore, sampling from  $\Pi$  is straightforward in these cases, which eases the implementation of the SGD algorithm.

### 1.6. Possible extensions

It is in principle possible to define and study SGD methods similar to (2) for the minimization on the space of measures of other functionals than the barycentric objective function, or with respect to other geometries than the 2-Wasserstein one. For instance, [17] introduces the notion of barycenters of probability measures based on *weak optimal transport* [8, 31, 32] and extends the ideas and algorithm developed here to that setting. A further, natural, example of functionals to consider are the entropy-regularized barycenters dealt with for numerical purposes in some of the aforementioned works and most recently in [20, 22].

In a different vein, it would be interesting to study conditions ensuring the convergence of the algorithm in (2) to stationary points (Karcher means) when the latter is a class that strictly contains the minimum (barycenter). Under suitable conditions, such convergence can be expected to hold by analogy with the behavior of the Euclidean SGD algorithm in general (not necessarily convex) settings [15]. In fact, we believe this question can also be linked to the pseudo-associativity of Karcher means. A deeper study is left for future work.

## 2. Gradient descent in Wasserstein space: A review

We first survey the gradient descent method for the computation of 2-Wasserstein barycenters. This method will serve as motivation for the subsequent development of the SGD in Section 3. For simplicity, we take  $\Pi$  to be finitely supported. Concretely, we suppose in this section that  $\Pi = \sum_{i \leq L} \lambda_i \delta_{m_i}$ , with  $L \in \mathbb{N}$ ,  $\lambda_i \geq 0$ , and  $\sum_{i \leq L} \lambda_i = 1$ . We define the operator over absolutely continuous measures

$$G(m) := \left( \sum_{i=1}^L \lambda_i T_m^{m_i} \right)(m). \quad (6)$$

Notice that the fixed points of  $G$  are precisely the Karcher means of  $\Pi$  presented in the introduction. Thanks to [4] the operator  $G$  is continuous for the  $W_2$  distance. Also, if at least one of



the  $L$  measures  $m_i$  has a bounded density, then the unique Wasserstein barycenter  $\hat{m}$  of  $\Pi$  has a bounded density as well and satisfies  $G(\hat{m}) = \hat{m}$ . This suggests defining, starting from  $\mu_0$ , the sequence

$$\mu_{n+1} := G(\mu_n) \quad \text{for } n \geq 0. \quad (7)$$

The next result was proven in [4, Theorem 3.6] and independently in [55, Theorem 3, Corollary 2]:

**Proposition 1.** *The sequence  $\{\mu_n\}_{n \geq 0}$  in (7) is tight, and every weakly convergent subsequence of  $\{\mu_n\}_{n \geq 0}$  converges in  $W_2$  to an absolutely continuous measure in  $\mathcal{W}_2(\mathbb{R}^q)$  that is also a Karcher mean. If some  $m_i$  has a bounded density, and if there exists a unique Karcher mean, then  $\hat{m}$  is the Wasserstein barycenter of  $\Pi$  and  $W_2(\mu_n, \hat{m}) \rightarrow 0$ .*

For the reader's convenience, we present next a counterexample to the uniqueness of Karcher means.

**Example 1.** In  $\mathbb{R}^2$  we take  $\mu_1$  as the uniform measure on  $B((-1, M), \varepsilon) \cup B((1, -M), \varepsilon)$  and  $\mu_2$  the uniform measure on  $B((-1, -M), \varepsilon) \cup B((1, M), \varepsilon)$ , with  $\varepsilon$  a small radius and  $M \gg \varepsilon$ . Then, if  $\Pi = \frac{1}{2}(\delta_{\mu_1} + \delta_{\mu_2})$ , the uniform measure on  $B((-1, 0), \varepsilon) \cup B((1, 0), \varepsilon)$  and the uniform measure on  $B((0, M), \varepsilon) \cup B((0, -M), \varepsilon)$  are two distinct Karcher means.

Thanks to the *Riemann-like* geometry of  $\mathcal{W}_{2,\text{ac}}(\mathbb{R}^q)$  we can reinterpret the iterations in (7) as a gradient descent step. This was discovered in [43, 55]. In fact, [55, Theorem 1] shows that the functional

$$F(m) := \frac{1}{2} \sum_{i=1}^L \lambda_i W_2^2(m_i, m) \quad (8)$$

defined on  $\mathcal{W}_2(\mathbb{R}^q)$  has a Fréchet derivative at each point  $m \in \mathcal{W}_{2,\text{ac}}(\mathbb{R}^q)$  given by

$$F'(m) = - \sum_{i=1}^L \lambda_i (T_m^{m_i} - I) = I - \sum_{i=1}^L \lambda_i T_m^{m_i} \in L^2(m), \quad (9)$$

where  $I$  is the identity map in  $\mathbb{R}^q$ . More precisely, for such  $m$ , when  $W_2(\hat{m}, m)$  goes to zero,

$$\frac{F(\hat{m}) - F(m) - \int_{\mathbb{R}^q} \langle F'(m)(x), T_m^{\hat{m}}(x) - x \rangle m(dx)}{W_2(\hat{m}, m)} \rightarrow 0$$

thanks to [6, Corollary 10.2.7]. It follows from Brenier's theorem [50, Theorem 2.12(ii)] that  $\hat{m}$  is a fixed point of  $G$  defined in (6) if and only if  $F'(\hat{m}) = 0$ . The gradient descent sequence with step  $\gamma$  starting from  $\mu_0 \in \mathcal{W}_{2,\text{ac}}(\mathbb{R}^q)$  is then defined by (cf. [55])

$$\mu_{n+1} := G_\gamma(\mu_n) \quad \text{for } n \geq 0, \quad (10)$$

where

$$G_\gamma(m) := [I + \gamma F'(m)](m) = \left[ (1 - \gamma)I + \gamma \sum_{i=1}^L \lambda_i T_m^{m_i} \right](m) = \left[ I + \gamma \sum_{i=1}^L \lambda_i (T_m^{m_i} - I) \right](m).$$

Note that, by (9), the iterations in (10) truly correspond to a gradient descent in  $\mathcal{W}_2(\mathbb{R}^q)$  for the function in (8). We remark also that, if  $\gamma = 1$ , the sequence in (10) coincides with that in (7), i.e.  $G_1 = G$ . These ideas serve as inspiration for the stochastic gradient descent iteration in the next part.



### 3. Stochastic gradient descent for barycenters in Wasserstein space

The method presented in Section 2 is well suited to calculating the empirical barycenter. For the estimation of a population barycenter (i.e. when  $\Pi$  does not have finite support) we would need to construct a convergent sequence of empirical barycenters, which can be computationally expensive. Furthermore, if a new sample from  $\Pi$  arrives, the previous method would need to recalculate the barycenter from scratch. To address these challenges, we follow the ideas of stochastic algorithms [46], widely adopted in machine learning [15], and define a *stochastic* version of the gradient descent sequence for the barycenter of  $\Pi$ .

Recall that for  $\mu_0 \in K_\Pi$  (in particular,  $\mu_0$  absolutely continuous),  $m_k \stackrel{\text{i.i.d.}}{\sim} \Pi$  defined in some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , and  $\gamma_k > 0$  for  $k \geq 0$ , we constructed the SGD sequence as  $\mu_{k+1} := [(1 - \gamma_k)I + \gamma_k T_{\mu_k}^{m_k}](\mu_k)$  for  $k \geq 0$ . The key ingredients for the convergence analysis of the above SGD iterations are the functions

$$F(\mu) := \frac{1}{2} \int_{\mathcal{P}(\mathcal{X})} W_2^2(\mu, m) \Pi(dm), \quad F'(\mu)(x) := - \int_{\mathcal{P}(\mathcal{X})} (T_\mu^m - I) \Pi(dm)(x),$$

the natural analogues to the eponymous objects in (8) and (9). In this setting, we can formally (or rigorously, under additional assumptions) check that  $F'$  is the actual Frechet derivative of  $F$ , and this justifies naming  $\{\mu_k\}_k$  a SGD sequence. In the following, the notation  $F$  and  $F'$  always refers to the functions just defined. Observe that the population barycenter  $\hat{\mu}$  is the unique minimizer of  $F$ . The following lemma justifies that  $F'$  is well defined and that  $\|F'(\hat{\mu})\|_{L^2(\hat{\mu})} = 0$ , and in particular that  $\hat{\mu}$  is a Karcher mean. This is a generalization of the corresponding result in [4] where only the case  $|\text{supp}(\Pi)| < \infty$  is covered.

**Lemma 1.** *Let  $\tilde{\Pi}$  be a probability measure concentrated on  $\mathcal{W}_{2,\text{ac}}(\mathbb{R}^q)$ . There exists a jointly measurable function  $\mathcal{W}_{2,\text{ac}}(\mathbb{R}^q) \times \mathcal{W}_2(\mathbb{R}^q) \times \mathbb{R}^q \ni (\mu, m, x) \mapsto T_\mu^m(x)$  that is  $\mu(dx) \Pi(dm) \tilde{\Pi}(d\mu)$ -a.s. equal to the unique optimal transport map from  $\mu$  to  $m$  at  $x$ . Furthermore, letting  $\hat{\mu}$  be a barycenter of  $\Pi$ ,  $x = \int T_{\hat{\mu}}^m(x) \Pi(dm)$ ,  $\hat{\mu}(dx)$ -a.s.*

*Proof.* The existence of a jointly measurable version of the unique optimal maps is proved in [30]. Let us prove the last assertion. Letting  $T_{\hat{\mu}}^m =: T^m$ , we have, by Brenier's theorem [50, Theorem 2.12(ii)],

$$\begin{aligned} \int W_2(\hat{\mu}, m)^2 \Pi(dm) &= \int \int |x - T^m(x)|^2 \hat{\mu}(dx) \Pi(dm) \\ &= \int \int \left| x - \int T^{\bar{m}}(x) \Pi(d\bar{m}) + \int T^{\bar{m}}(x) \Pi(d\bar{m}) - T^m(x) \right|^2 \hat{\mu}(dx) \Pi(dm) \\ &= \int \left| x - \int T^{\bar{m}}(x) \Pi(d\bar{m}) \right|^2 \hat{\mu}(dx) \\ &\quad + \int \int \left| \int T^{\bar{m}}(x) \Pi(d\bar{m}) - T^m(x) \right|^2 \hat{\mu}(dx) \Pi(dm), \end{aligned}$$

where we used the fact that

$$2 \int \int \left\langle x - \int T^{\bar{m}}(x) \Pi(d\bar{m}), \int T^{\bar{m}}(x) \Pi(d\bar{m}) - T^m(x) \right\rangle \Pi(dm) \hat{\mu}(dx) = 0.$$

The term in the last line is an upper bound for

$$\int W_2\left(\left(\int T^{\bar{m}} \Pi(d\bar{m})\right)(\mu), m\right)^2 \Pi(dm) \geq \int W_2(\hat{\mu}, m)^2 \Pi(dm).$$

We conclude that  $\int |x - \int T^{\bar{m}}(x) \Pi(d\bar{m})|^2 \hat{\mu}(dx)$ , as required.  $\square$

Notice that Lemma 1 ensures that the SGD sequence  $\{\mu_k\}_k$  is well defined as a sequence of (measurable)  $\mathcal{W}_2$ -valued random variables. More precisely, denoting by  $\mathcal{F}_0$  the trivial sigma-algebra and  $\mathcal{F}_{k+1}$ ,  $k \geq 0$ , the sigma-algebra generated by  $m_0, \dots, m_k$ , we can inductively apply the first part of Lemma 1 with  $\bar{\Pi} = \text{Law}(\mu_k)$  to check that  $T_{\mu_k}^{m_k}(x)$  is measurable with respect to  $\mathcal{F}_{k+1} \otimes \mathcal{B}(\mathbb{R}^q)$ , where  $\mathcal{B}$  stands for the Borel sigma-field. This implies that both  $\mu_k$  and  $\|F'(\mu_k)\|_{L^2(\mu_k)}^2$  are measurable with respect to  $\mathcal{F}_k$ .

The next proposition suggests that, in expectation, the sequence  $\{F(\mu_k)\}_k$  is essentially decreasing. This is a first insight into the behavior of the sequence  $\{\mu_k\}_k$ .

**Proposition 2.** *The SGD sequence in (2) satisfies, almost surely,*

$$\mathbb{E}[F(\mu_{k+1}) - F(\mu_k) \mid \mathcal{F}_k] \leq \gamma_k^2 F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2. \quad (11)$$

*Proof.* Let  $\nu \in \text{supp}(\Pi)$ . Clearly,  $([(1 - \gamma_k)I + \gamma_k T_{\mu_k}^{m_k}], T_{\mu_k}^\nu)(\mu_k)$  is a feasible (not necessarily optimal) coupling with first and second marginals  $\mu_{k+1}$  and  $\nu$  respectively. Writing  $O_m := T_{\mu_k}^m - I$ , we have

$$\begin{aligned} W_2^2(\mu_{k+1}, \nu) &\leq \|(1 - \gamma_k)I + \gamma_k T_{\mu_k}^{m_k} - T_{\mu_k}^\nu\|_{L^2(\mu_k)}^2 \\ &= \|-O_\nu + \gamma_k O_{m_k}\|_{L^2(\mu_k)}^2 = \|O_\nu\|_{L^2(\mu_k)}^2 - 2\gamma_k \langle O_\nu, O_{m_k} \rangle_{L^2(\mu_k)} + \gamma_k^2 \|O_{m_k}\|_{L^2(\mu_k)}^2. \end{aligned}$$

Evaluating  $\mu_{k+1}$  on the functional  $F$ , thanks to the previous inequality we have

$$\begin{aligned} F(\mu_{k+1}) &= \frac{1}{2} \int W_2^2(\mu_{k+1}, \nu) \Pi(d\nu) \\ &\leq \frac{1}{2} \int \|O_\nu\|_{L^2(\mu_k)}^2 \Pi(d\nu) - \gamma_k \left\langle \int O_\nu \Pi(d\nu), O_{m_k} \right\rangle_{L^2(\mu_k)} + \frac{\gamma_k^2}{2} \|O_{m_k}\|_{L^2(\mu_k)}^2 \\ &= F(\mu_k) + \gamma_k \langle F'(\mu_k), O_{m_k} \rangle_{L^2(\mu_k)} + \frac{\gamma_k^2}{2} \|O_{m_k}\|_{L^2(\mu_k)}^2. \end{aligned}$$

Taking conditional expectation with respect to  $\mathcal{F}_k$ , and as  $m_k$  is independently sampled from this sigma-algebra, we conclude that

$$\begin{aligned} \mathbb{E}[F(\mu_{k+1}) \mid \mathcal{F}_k] &\leq F(\mu_k) + \gamma_k \left\langle F'(\mu_k), \int O_m \Pi(dm) \right\rangle_{L^2(\mu_k)} + \frac{\gamma_k^2}{2} \int \|O_m\|_{L^2(\mu_k)}^2 \Pi(dm) \\ &= (1 + \gamma_k^2) F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2. \end{aligned} \quad \square$$

The next lemma states some key continuity properties of the functions  $F$  and  $F'$ . For this result, Assumption 2 can be dropped and it is only required that  $\int \int \|x\|^2 m(dx) \Pi(dm) < \infty$ .

**Lemma 2.** Let  $(\rho_n)_n \subset \mathcal{W}_{2,\text{ac}}(\mathbb{R}^d)$  be a sequence converging with respect to  $W_2$  to  $\rho \in \mathcal{W}_{2,\text{ac}}(\mathbb{R}^d)$ . Then, as  $n \rightarrow \infty$ ,

- (i)  $F(\rho_n) \rightarrow F(\rho)$ ;
- (ii)  $\|F'(\rho_n)\|_{L^2(\rho_n)} \rightarrow \|F'(\rho)\|_{L^2(\rho)}$ .

*Proof.* We prove both convergence claims using Skorokhod's representation theorem. Thanks to that result, in a given probability space  $(\Omega, \mathcal{G}, \mathbb{P})$  we can simultaneously construct a sequence of random vectors  $(X_n)_n$  of laws  $(\rho_n)_n$  and a random variable  $X$  of law  $\rho$  such that  $(X_n)_n$  converges  $\mathbb{P}$ -a.s. to  $X$ . Moreover, by [23, Theorem 3.4], the sequence  $(T_{\rho_n}^m(X_n))_n$  converges  $\mathbb{P}$ -a.s. to  $T_\rho^m(X)$ . Notice that, for all  $n \in \mathbb{N}$ ,  $T_{\rho_n}^m(X_n)$  distributes according to the law  $m$ , and the same holds true for  $T_\rho^m(X)$ .

We now enlarge the probability space  $(\Omega, \mathcal{G}, \mathbb{P})$  (maintaining the same notation for simplicity) with an independent random variable  $\mathbf{m}$  in  $\mathcal{W}_2(\mathbb{R}^d)$  with law  $\Pi$  (thus independent of  $(X_n)_n$  and  $X$ ). Applying Lemma 1 with  $\tilde{\Pi} = \delta_{\rho_n}$  for each  $n$ , or with  $\tilde{\Pi} = \delta_\rho$ , we can show that  $(X_n, T_{\rho_n}^{\mathbf{m}}(X_n))$  and  $(X, T_\rho^{\mathbf{m}}(X))$  are random variables in  $(\Omega, \mathcal{G}, \mathbb{P})$ . By conditioning on  $\{\mathbf{m} = m\}$ , we further obtain that  $(X_n, T_{\rho_n}^{\mathbf{m}}(X_n))_{n \in \mathbb{N}}$  converges  $\mathbb{P}$ -a.s. to  $(X, T_\rho^{\mathbf{m}}(X))$ .

Notice that  $\sup_n \mathbb{E}(\|X_n\|^2 \mathbf{1}_{\|X_n\|^2 \geq M}) = \sup_n \int_{\|x\|^2 \geq M} \|x\|^2 \rho_n(dx) \rightarrow 0$  as  $M \rightarrow \infty$  since  $(\rho_n)_n$  converges in  $\mathcal{W}_2(\mathcal{X})$ , while, upon conditioning on  $\mathbf{m}$ ,

$$\sup_n \mathbb{E}(\|T_{\rho_n}^{\mathbf{m}}(X_n)\|^2 \mathbf{1}_{\|T_{\rho_n}^{\mathbf{m}}(X_n)\|^2 \geq M}) = \int_{\mathcal{W}_2(\mathcal{X})} \left( \int_{\|y\|^2 \geq M} \|y\|^2 m(dy) \right) \Pi(dm) \rightarrow 0$$

by dominated convergence, since  $\int_{\|x\|^2 \geq M} \|x\|^2 m(dx) \leq \int \|x\|^2 m(dx) = W_2^2(m, \delta_0)$  and  $\Pi \in \mathcal{W}_2(\mathcal{W}_2(\mathcal{X}))$ . By Vitali's convergence theorem, we deduce that  $(X_n, T_{\rho_n}^{\mathbf{m}}(X_n))_{n \in \mathbb{N}}$  converges to  $(X, T_\rho^{\mathbf{m}}(X))$  in  $L^2(\Omega, \mathcal{G}, \mathbb{P})$ . In particular, as  $n \rightarrow \infty$ ,

$$\int_{\mathcal{P}(\mathcal{X})} W_2^2(\rho_n, m) \Pi(dm) = \mathbb{E}|X_n - T_{\rho_n}^{\mathbf{m}}(X_n)|^2 \rightarrow \mathbb{E}|X - T_\rho^{\mathbf{m}}(X)|^2 = \int_{\mathcal{P}(\mathcal{X})} W_2^2(\rho, m) \Pi(dm),$$

which proves the convergence in (i). Now denoting by  $\mathcal{G}_\infty$  the sigma-field generated by  $(X_1, X_2, \dots)$ , we also obtain that

$$\mathbb{E}(X_n - T_{\rho_n}^{\mathbf{m}}(X_n) | \mathcal{G}_\infty) \rightarrow \mathbb{E}(X - T_\rho^{\mathbf{m}}(X) | \mathcal{G}_\infty) \quad \text{in } L^2(\Omega, \mathcal{G}, \mathbb{P}). \quad (12)$$

Observe now that the following identities hold:

$$\begin{aligned} F'(\rho_n)(X_n) &= \mathbb{E}(X_n - T_{\rho_n}^{\mathbf{m}}(X_n) | X_n) = \mathbb{E}(X_n - T_{\rho_n}^{\mathbf{m}}(X_n) | \mathcal{G}_\infty) \\ F'(\rho)(X) &= \mathbb{E}(X - T_\rho^{\mathbf{m}}(X) | X) = \mathbb{E}(X - T_\rho^{\mathbf{m}}(X) | \mathcal{G}_\infty). \end{aligned}$$

The convergence in (ii) follows from (12),  $\mathbb{E}(F'(\rho_n)(X_n))^2 = \|F'(\rho_n)\|_{L^2(\rho_n)}^2$ , and  $\mathbb{E}(F'(\rho)(X))^2 = \|F'(\rho)\|_{L^2(\rho)}^2$ .

We now proceed to prove the first of our two main results.

*Proof of Theorem 1.* Let us denote by  $\hat{\mu}$  the unique barycenter, write  $\hat{F} := F(\hat{\mu})$ , and introduce  $h_t := F(\mu_t) - \hat{F} \geq 0$  and  $\alpha_t := \prod_{i=1}^{t-1} 1/(1 + \gamma_i^2)$ . Thanks to the condition in (3),

the sequence  $(\alpha_t)$  converges to some finite  $\alpha_\infty > 0$ , as can be verified simply by applying logarithms. By Proposition 2,

$$\mathbb{E}[h_{t+1} - (1 + \gamma_t^2)h_t \mid \mathcal{F}_t] \leq \gamma_t^2 \hat{F} - \gamma_t \|F'(\mu_t)\|_{L^2(\mu_t)}^2 \leq \gamma_t^2 \hat{F},$$

from which, after multiplying by  $\alpha_{t+1}$ , the following bound is derived:

$$\mathbb{E}[\alpha_{t+1}h_{t+1} - \alpha_t h_t \mid \mathcal{F}_t] \leq \alpha_{t+1} \gamma_t^2 \hat{F} - \alpha_{t+1} \gamma_t \|F'(\mu_t)\|_{L^2(\mu_t)}^2 \leq \alpha_{t+1} \gamma_t^2 \hat{F}. \quad (13)$$

Now defining  $\hat{h}_t := \alpha_t h_t - \sum_{i=1}^t \alpha_i \gamma_{i-1}^2 \hat{F}$ , we deduce from (13) that  $\mathbb{E}[\hat{h}_{t+1} - \hat{h}_t \mid \mathcal{F}_t] \leq 0$ , namely that  $(\hat{h}_t)_{t \geq 0}$  is a supermartingale with respect to  $(\mathcal{F}_t)$ . The fact that  $(\alpha_t)$  is convergent, together with the condition in (3), ensures that  $\sum_{i=1}^\infty \alpha_i \gamma_{i-1}^2 \hat{F} < \infty$ , and thus  $\hat{h}_t$  is uniformly lower-bounded by a constant. Therefore, the supermartingale convergence theorem [53, Corollary 11.7] implies the existence of  $\hat{h}_\infty \in L^1$  such that  $\hat{h}_t \rightarrow \hat{h}_\infty$  a.s. But then, necessarily,  $h_t \rightarrow h_\infty$  a.s. for some non-negative random variable  $h_\infty \in L^1$ .

Thus, our goal now is to prove that  $h_\infty = 0$  a.s. Taking expectations in (13) and summing over  $t$  to obtain a telescopic summation, we obtain

$$\mathbb{E}[\alpha_{t+1}h_{t+1}] - \mathbb{E}[h_0\alpha_0] \leq \hat{F} \sum_{s=1}^t \alpha_{s+1} \gamma_s^2 - \sum_{s=1}^t \alpha_{s+1} \gamma_s \mathbb{E}[\|F'(\mu_s)\|_{L^2(\mu_s)}^2].$$

Then, taking  $\liminf$ , applying Fatou on the left-hand side and monotone convergence on the right-hand side, we obtain

$$-\infty < \mathbb{E}[\alpha_\infty h_\infty] - \mathbb{E}[h_0\alpha_0] \leq C - \mathbb{E}\left[\sum_{s=1}^\infty \alpha_{s+1} \gamma_s \|F'(\mu_s)\|_{L^2(\mu_s)}^2\right].$$

In particular, since  $(\alpha_t)$  is bounded away from 0, we have

$$\sum_{t=1}^\infty \gamma_t \|F'(\mu_t)\|_{L^2(\mu_t)}^2 < \infty \quad \text{a.s.} \quad (14)$$

Note that  $\mathbb{P}(\liminf_{t \rightarrow \infty} \|F'(\mu_t)\|_{L^2(\mu_t)}^2 > 0) > 0$  would be at odds with the conditions in (14) and (4), so

$$\liminf_{t \rightarrow \infty} \|F'(\mu_t)\|_{L^2(\mu_t)}^2 = 0 \quad \text{a.s.} \quad (15)$$

Observe also that, from Assumption 2 and Lemma 2, we have

$$\text{for all } \varepsilon > 0, \inf_{\{\rho: F(\rho) \geq \hat{F} + \varepsilon\} \cap K_\Pi} \|F'(\rho)\|_{L^2(\rho)}^2 > 0. \quad (16)$$

Indeed, we can see that the set  $\{\rho: F(\rho) \geq \hat{F} + \varepsilon\} \cap K_\Pi$  is  $W_2$ -compact, using Lemma 2(i), and then check that the function  $\rho \mapsto \|F'(\rho)\|_{L^2(\rho)}^2$  attains its minimum on it, using part (ii) of that result. That minimum cannot be zero, as otherwise we would have obtained a Karcher mean that is not equal to the barycenter (contradicting the uniqueness of the Karcher mean). Note also that, a.s.,  $\mu_t \in K_\Pi$  for each  $t$ , by the geodesic convexity part of Assumption 2. We deduce

the following a.s. relationships between events:

$$\begin{aligned} \{h_\infty \geq 2\varepsilon\} &\subset \{\mu_t \in \{\rho: F(\rho) \geq \hat{F} + \varepsilon\} \cap K_\Pi \text{ for all } t \text{ large enough}\} \\ &\subset \bigcup_{\ell \in \mathbb{N}} \{\|F'(\mu_t)\|_{L^2(\mu_t)}^2 > 1/\ell: \text{ for all } t \text{ large enough}\} \\ &\subset \left\{ \liminf_{t \rightarrow \infty} \|F'(\mu_t)\|_{L^2(\mu_t)}^2 > 0 \right\}, \end{aligned}$$

where (16) was used to obtain the second inclusion. It follows using (15) that  $\mathbb{P}(h_\infty \geq 2\varepsilon) = 0$  for every  $\varepsilon > 0$ , and hence  $h_\infty = 0$  as required.

To conclude, we use the fact that the sequence  $\{\mu_t\}_t$  is a.s. contained in the  $W_2$ -compact  $K_\Pi$  by Assumption 2, and the first convergence in Lemma 2 to deduce that the limit  $\tilde{\mu}$  of any convergent subsequence satisfies  $F(\tilde{\mu}) - \hat{F} = h_\infty = 0$ , and then  $F(\tilde{\mu}) = \hat{F}$ . Hence  $\tilde{\mu} = \hat{\mu}$  by the uniqueness of the barycenter. This implies that  $\mu_t \rightarrow \hat{\mu}$  in  $\mathcal{W}_2(\mathcal{X})$  a.s. as  $t \rightarrow \infty$ .  $\square$

**Remark 1.** Observe that, for a fixed  $\mu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$  and a random  $m \sim \Pi$ , the random variable  $-(T_\mu^m - I)(x)$  is an unbiased estimator of  $F'(\mu)(x)$  for  $\mu(\text{d}x)$  almost everywhere,  $x \in \mathbb{R}^d$ . A natural way to jointly quantify the pointwise variances of these estimators is through the *integrated variance*,

$$\mathbb{V}[-(T_\mu^m - I)] := \int \text{Var}_{m \sim \Pi}[T_\mu^m(x) - x] \mu(\text{d}x),$$

which is the equivalent (for unbiased estimators) of the mean integrated square error from non-parametric statistics [52]. Simple computations yield the following expression for it, which will be useful in the next two sections:

$$\begin{aligned} \mathbb{V}[-(T_\mu^m - I)] &= \mathbb{E}[\|-(T_\mu^m - I)\|_{L^2(\mu)}^2] - \|\mathbb{E}[-(T_\mu^m - I)]\|_{L^2(\mu)}^2 \\ &= 2F(\mu) - \|F'(\mu)\|_{L^2(\mu)}^2. \end{aligned} \quad (17)$$

We close this section with the promised statement of Lemma 3, referred to in the introduction, giving us a sufficient condition for Assumption 2.

**Lemma 3.** *If the support of  $\Pi$  is  $\mathcal{W}_2(\mathbb{R}^d)$ -compact and there is a constant  $C_1 < \infty$  such that  $\Pi(\{m \in \mathcal{W}_{2,\text{ac}}(\mathbb{R}^d): \int \log(\text{d}m/\text{d}x) m(\text{d}x) \leq C_1\}) = 1$ , then Assumption 2 is fulfilled.*

*Proof.* Let  $K$  be the support of  $\Pi$ , which is  $\mathcal{W}_2(\mathbb{R}^d)$ -compact. By the de la Vallée Poussin criterion, there is a  $V: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ , increasing, convex, and super-quadratic (i.e.  $\lim_{r \rightarrow +\infty} V(r)/r^2 = +\infty$ ), such that  $C_2 := \sup_{m \in K} \int V(\|x\|) m(\text{d}x) < \infty$ . Observe that  $p(\text{d}x) := \exp\{-V(\|x\|)\} \text{d}x$  is a finite measure (without loss of generality a probability measure). Moreover, for the relative entropy with respect to  $p$  we have  $H(m|p) := \int \log(\text{d}m/\text{d}p) \text{d}m = \int \log(\text{d}m/\text{d}x) m(\text{d}x) + \int V(\|x\|) m(\text{d}x)$  if  $m \ll \text{d}x$  and  $+\infty$  otherwise. Now define  $K_\Pi := \{m \in \mathcal{W}_{2,\text{ac}}(\mathbb{R}^d): H(m|p) \leq C_1 + C_2, \int V(\|x\|) m(\text{d}x) \leq C_2\}$  so that  $\Pi(K_\Pi) = 1$ . Clearly,  $K_\Pi$  is  $\mathcal{W}_2$ -closed, since the relative entropy is weakly lower-semicontinuous, and also  $\mathcal{W}_2$ -relatively compact, since  $V$  is super-quadratic. Finally,  $K_\Pi$  is geodesically convex by [50, Theorem 5.15].  $\square$

For instance, if all  $m$  in the support of  $\Pi$  is of the form

$$m(\text{d}x) = \frac{e^{-V_m(x)}}{\int_y e^{-V_m(y)} \text{d}y} \text{d}x$$

with  $V_m$  bounded from below, then one way to guarantee the conditions in Lemma 3, assuming without loss of generality that  $V_m \geq 0$ , is to ask that  $\int_{\mathcal{Y}} e^{-V_m(y)} dy \geq A$  and  $\int_{\mathcal{Y}} |y|^{2+\varepsilon} e^{-V_m(y)} dy \leq B$  for some fixed  $\varepsilon, A, B > 0$ . In words: tails are controlled and the measures cannot be too concentrated.

#### 4. A condition granting uniqueness of Karcher means and convergence rates

The aim of this section is to prove Theorem 2, a refinement of Theorem 1 under the additional assumption that the barycenter is a pseudo-associative Karcher mean. Notice that, with the notation introduced in Section 3, Definition 4 of pseudo-associative Karcher mean  $\mu$  simply reads as

$$W_2^2(\mu, \nu) \leq C_\mu \|F'(\nu)\|_{L^2(\nu)}^2 \quad (18)$$

for all  $\nu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$ .

**Remark 2.** Suppose  $\mu$  is a Karcher mean of  $\Pi$  such that, for all  $\nu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$  and  $\Pi(dm)$  for almost every  $m$ ,

$$T_\mu^m(x) = T_\nu^m \circ T_\mu^\nu(x), \quad \mu(dx) \text{ a.e. } x \in \mathbb{R}^q. \quad (19)$$

Then,  $\mu$  is pseudo-associative, with  $C_\mu = 1$  and equality holding in Definition 4. Indeed, in that case we have

$$\begin{aligned} \|F'(\nu)\|_{L^2(\nu)}^2 &= \int_{\mathcal{X}} \left| \int_{\mathcal{P}(\mathcal{X})} (T_\nu^m \circ T_\mu^\nu(x) - T_\mu^\nu(x)) \Pi(dm) \right|^2 \mu(dx) \\ &= \int_{\mathcal{X}} \left| \int_{\mathcal{P}(\mathcal{X})} T_\mu^m(x) \Pi(dm) - T_\mu^\nu(x) \right|^2 \mu(dx) \\ &= \int_{\mathcal{X}} |x - T_\mu^\nu(x)|^2 \mu(dx) = W_2^2(\mu, \nu). \end{aligned}$$

We will thus say that the Karcher mean  $\mu$  is *associative* simply if (19) holds.

The following is an analogue in Wasserstein space of a classical property implying convergence rates in gradient-type optimization algorithms.

**Definition 5.** We say that  $\mu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$  satisfies a Polyak–Łojasiewicz inequality if

$$F(\nu) - F(\mu) \leq \frac{\bar{C}_\mu}{2} \|F'(\nu)\|_{L^2(\nu)}^2 \quad (20)$$

for some  $\bar{C}_\mu > 0$  and every  $\nu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$ .

We next state some useful properties.

**Lemma 4.** Suppose  $\mu$  is a Karcher mean of  $\Pi$ . Then:

- (i) For all  $\nu \in \mathcal{W}_{2,\text{ac}}(\mathcal{X})$ ,  $F(\nu) - F(\mu) \leq \frac{1}{2} W_2^2(\mu, \nu)$ .
- (ii) If  $\mu$  is pseudo-associative, then it is the unique barycenter, and it satisfies the Polyak–Łojasiewicz inequality (20) with  $\bar{C}_\mu = C_\mu$ .
- (iii) If the associativity relation in (19) holds,  $F(\nu) - F(\mu) = \frac{1}{2} W_2^2(\mu, \nu) = \frac{1}{2} \|F'(\nu)\|_{L^2(\nu)}^2$ .

*Proof.* For (i), we notice that this is a particular case of [19, Theorem 7], which we can prove by an elementary argument based on the notion of Karcher mean. Indeed, by the definition of the function  $F$  and the fact that  $(T_\mu^\nu, T_\mu^m)(\mu)$  is a coupling of  $\nu$  and  $m$ ,

$$\begin{aligned} F(\nu) - F(\mu) &\leq \frac{1}{2} \int_{\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} \{|T_\mu^\nu(x) - T_\mu^m(x)|^2 - |x - T_\mu^m(x)|^2\} \mu(dx) \Pi(dm) \\ &= \frac{1}{2} \int_{\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} \{|T_\mu^\nu(x)|^2 - 2\langle T_\mu^m(x), T_\mu^\nu(x) \rangle - x^2 + 2\langle x, T_\mu^m(x) \rangle\} \mu(dx) \Pi(dm) \\ &= \frac{1}{2} \int_{\mathcal{X}} |T_\mu^\nu(x) - x|^2 \mu(dx), \end{aligned}$$

where in the second equality we twice used the fact that  $\int_{\mathcal{P} \Rightarrow (\mathcal{X})} T_\mu^m(x) \Pi(dm) = x$  for  $\mu(dx)$  a.e.  $x$ .

For (ii), the claim is obvious in view of the previous argument and (18).

For (iii), taking into account Remark 2, it is enough to notice that

$$F(\nu) = \int_{\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} |y - T_\nu^m(y)|^2 \nu(dy) \Pi(dm) = \int_{\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} |T_\mu^\nu(x) - T_\mu^m(x)|^2 \mu(dx) \Pi(dm)$$

under (19), in which case the inequality in the proof of (i) is an equality.  $\square$

**Remark 3.** Recall that  $\Pi$  is said to satisfy a variance inequality [2] if, for some constant  $C > 0$ ,  $F(\nu) - F(\hat{\mu}) \geq \frac{1}{2} CW_2^2(\hat{\mu}, \nu)$ , with  $\hat{\mu}$  the barycenter of  $\Pi$ . It readily follows that  $\hat{\mu}$  is a pseudo-associative Karcher mean as soon as a variance inequality and the Polyak–Łojasiewicz inequality 20 hold. That was the case in [19], and so the barycenter in the Gaussian setting considered therein is a pseudo-associative Karcher mean too. Notice that if the barycenter is associative, by Lemma 4(iii) the variance inequality holds; this is the case for the examples discussed in Sections 6.1–6.3. Notice also that, if a variance inequality holds, bounds on the optimality gap for the SDG sequence as in Theorem 2 immediately yield similar bounds for the expected squared Wasserstein distance  $\mathbb{E}[W_2^2(\mu_k, \hat{\mu})]$ .

*Proof of Theorem 2.* If we assume the pseudo-associativity condition (5) holds, the uniqueness of Karcher means (hence equal to the barycenter) is immediate, as noted in Section 1. Under that assumption, the inequality (20) holds thanks to Lemma 4(ii), and convergence estimates can then be deduced by adapting classic arguments of the SGD algorithm in an Euclidean setting; see, e.g., [15]. Indeed, by Proposition 2, (17), and (20),

$$\begin{aligned} \mathbb{E}[F(\mu_{k+1}) - F(\mu_k) \mid \mathcal{F}_k] &\leq \gamma_k^2 F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2 \\ &= \frac{\gamma_k^2}{2} \mathbb{V}[-(T_{\mu_k}^m - I)] - \left(1 - \frac{\gamma_k}{2}\right) \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2 \\ &\leq \gamma_k^2 \bar{F} - \left(1 - \frac{\gamma_k}{2}\right) \frac{2\gamma_k}{C_\mu} (F(\mu_k) - \hat{F}) \\ &\leq \gamma_k^2 \bar{F} - \gamma_k C_\mu^{-1} (F(\mu_k) - \hat{F}) \end{aligned}$$

for some finite (deterministic) upper bound  $\bar{F} > 0$  of the sequence  $(F(\mu_t))_{t \geq 1}$  under Assumption 2'. In the last line we used the fact that  $\gamma_k \leq \gamma_0 \leq 1$  for the choice  $\gamma_k = a/(b+k)$



with fixed  $b \geq a > 0$ . It follows that

$$\mathbb{E}[F(\mu_{k+1}) - \hat{F}] \leq \gamma_k^2 \bar{F} + (1 - \gamma_k C_\mu^{-1}) \mathbb{E}[F(\mu_k) - \hat{F}] \quad \text{for all } k \in \mathbb{N}. \quad (21)$$

Assume now that  $a > C_{\hat{\mu}}$  and that, for certain  $c \geq a^2 \bar{F} / (C_{\hat{\mu}}^{-1} a - 1) > 0$  and a given  $k \in \mathbb{N}$ ,

$$\mathbb{E}[F(\mu_k) - \hat{F}] \leq \frac{c}{b+k}. \quad (22)$$

Let us show that  $\mathbb{E}[F(\mu_{k+1}) - \hat{F}] \leq c/(b+k+1)$  too. By (21),

$$\mathbb{E}[F(\mu_{k+1}) - \hat{F}] \leq c \frac{(b+k-1)}{(b+k)^2} + \frac{c(1 - aC_\mu^{-1}) + a^2 \bar{F}}{(b+k)^2} \leq c \frac{(b+k-1)}{(b+k)^2} \leq \frac{c}{b+k+1},$$

where we used the choice of  $c$  in the second inequality. Taking

$$c = \max \{ b \mathbb{E}[F(\mu_0) - \hat{F}], a^2 \bar{F} / (C_{\hat{\mu}}^{-1} a - 1) \},$$

the inequality in (22) holds for  $k = 0$ , and hence for all  $k \in \mathbb{N}$  by induction.

Concerning the a.s. weak convergence of  $\mu_k$  to  $\hat{\mu}$ , we can follow the proof of Theorem 1 up to (15), and then derive in the present setting (16), i.e.

$$\text{for all } \varepsilon > 0, \inf_{\{\rho: F(\rho) \geq \hat{F} + \varepsilon\} \cap K_\Pi} \|F'(\rho)\|_{L^2(\rho)}^2 > 0,$$

from the Polyak–Łojasiewicz inequality (20), which holds thanks to Lemma 4(ii). As before, this is then used to obtain that  $F(\mu_k) \rightarrow \hat{F}$  almost surely. Since  $K_\Pi$  is  $\mathcal{W}_2$ -bounded, it is in particular tight. The fact that  $v \mapsto F(v)$  is weakly lower semicontinuous and the uniqueness of minimizers of  $F$  now entail that, almost surely,  $\mu_k \rightarrow \hat{\mu}$  weakly.  $\square$

**Remark 4.** In addition to concluding that, almost surely,  $\mu_k \rightarrow \hat{\mu}$  weakly, the above proof also establishes the a.s. existence of a subsequence  $n_k$  such that  $\mu_{n_k} \rightarrow \hat{\mu}$  in  $\mathcal{W}_2$ . Indeed, this follows from (15) together with (18).

The above proof shows that we may relax the definition of pseudo-associativity by requiring (5) to hold for  $v \in K_\Pi$  (or more precisely, for  $v \in \{\mu_k : k \in \mathbb{N}\}$ ) only.

## 5. Variance reduction via batch SGD

Paralleling the Euclidean setting, the function  $-(T_{\mu_k}^{m_k} - I)$  can be seen as an estimator of the gradient in the  $k$ th step of the SGD scheme. Hence, conditionally on  $m_0, \dots, m_{k-1}$ , its integrated variance given by  $2F(\mu_k) - \|F'(\mu_k)\|_{L^2(\mu_k)}^2$  could be large for some sampled  $\mu_k$ , which would yield a slow convergence if the steps  $\gamma_k$  are not small enough. To cope with this issue, as customary in stochastic algorithms, we are led to consider alternative estimators of  $F'(\mu)$  with less (integrated) variance.

**Definition 6.** Let  $\mu_0 \in K_\Pi$ ,  $m_k^i \stackrel{\text{i.i.d.}}{\sim} \Pi$ , and  $\gamma_k > 0$  for  $k \geq 0$  and  $i = 1, \dots, S_k$ . The batch stochastic gradient descent (BSGD) sequence is given by

$$\mu_{k+1} := \left[ (1 - \gamma_k)I + \gamma_k \frac{1}{S_k} \sum_{i=1}^{S_k} T_{\mu_k}^{m_k^i} \right] (\mu_k). \quad (23)$$

Notice that  $(1/S_k) \sum_{i=1}^{S_k} T_{\mu_k}^{m_i} - I$  is an unbiased estimator of  $-F'(\mu_k)$ . Proceeding as in Proposition 2, with  $\mathcal{F}_{k+1}$  now denoting the sigma-algebra generated by  $\{m_\ell^i: \ell \leq k, i \leq S_k\}$ , and writing  $\Pi$  also for the law of an i.i.d. sample of size  $S_k$ , we now have

$$\begin{aligned} \mathbb{E}[F(\mu_{k+1}) | \mathcal{F}_k] &= F(\mu_k) + \gamma_k \left\langle F'(\mu_k), \int \left( \frac{1}{S_k} \sum_{i=1}^{S_k} T_{\mu_k}^{m_i} - I \right) \Pi(dm_k^1 \cdots dm_k^{S_k}) \right\rangle_{L^2(\mu_k)} \\ &\quad + \frac{\gamma_k^2}{2} \int \left\| \frac{1}{S_k} \sum_{i=1}^{S_k} T_{\mu_k}^{m_i} - I \right\|_{L^2(\mu_k)}^2 \Pi(dm_k^1 \cdots dm_k^{S_k}) \\ &= F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2 + \frac{\gamma_k^2}{2} \int \left\| \frac{1}{S_k} \sum_{i=1}^{S_k} T_{\mu_k}^{m_i} - I \right\|_{L^2(\mu_k)}^2 \Pi(dm_k^1 \cdots dm_k^{S_k}) \\ &\leq F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2 + \frac{\gamma_k^2}{2} \frac{1}{S_k} \sum_{i=1}^{S_k} \int \|T_{\mu_k}^{m_i} - I\|_{L^2(\mu_k)}^2 \Pi(dm_k^i) \\ &= (1 + \gamma_k^2) F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2. \end{aligned}$$

The arguments in the proof of Theorem 1 can then be easily adapted to get the following result.

**Theorem 3.** *Under the assumptions of Theorem 1, the BSGD sequence  $\{\mu_t\}_{t \geq 0}$  in (23) converges almost surely to the 2-Wasserstein barycenter of  $\Pi$ .*

The supporting idea for using mini-batches is *reducing the noise* of the estimator of  $F'(\mu)$ .

**Proposition 3.** *The integrated variance of the mini-batch estimator of fixed batch size  $S$  for  $F'(\mu)$ , given by  $-(1/S) \sum_{i=1}^S (T_\mu^{m_i} - I)$ , decreases linearly in the sample size. More precisely,*

$$\mathbb{V} \left[ -\frac{1}{S} \sum_{i=1}^S (T_\mu^{m_i} - I) \right] = \frac{1}{S} \mathbb{V} [-(T_\mu^{m_1} - I)] = \frac{1}{S} [2F(\mu) - \|F'(\mu)\|_{L^2(\mu)}^2].$$

*Proof.* In a similar way to (17), the integrated variance of the mini-batch estimator is

$$\begin{aligned} \mathbb{V} \left[ -\frac{1}{S} \sum_{i=1}^S (T_\mu^{m_i} - I) \right] &= \mathbb{E} \left[ \left\| -\frac{1}{S} \sum_{i=1}^S (T_\mu^{m_i} - I) \right\|_{L^2(\mu)}^2 \right] - \left\| \mathbb{E} \left[ -\frac{1}{S} \sum_{i=1}^S (T_\mu^{m_i} - I) \right] \right\|_{L^2(\mu)}^2 \\ &= \mathbb{E} \left[ \left\| -\frac{1}{S} \sum_{i=1}^S (T_\mu^{m_i} - I) \right\|_{L^2(\mu)}^2 \right] - \|F'(\mu)\|_{L^2(\mu)}^2. \end{aligned}$$

The first term can be expanded as

$$\begin{aligned} \left\| -\frac{1}{S} \sum_{i=1}^S (T_{\mu}^{m_i} - I) \right\|_{L^2(\mu)}^2 &= \frac{1}{S^2} \left\langle \sum_{i=1}^S (T_{\mu}^{m_i} - I), \sum_{j=1}^S (T_{\mu}^{m_j} - I) \right\rangle_{L^2(\mu)} \\ &= \frac{1}{S^2} \sum_{i=1}^S \sum_{j=1}^S \langle T_{\mu}^{m_i} - I, T_{\mu}^{m_j} - I \rangle_{L^2(\mu)} \\ &= \frac{1}{S^2} \sum_{i=1}^S \|-(T_{\mu}^{m_i} - I)\|_{L^2(\mu)}^2 + \frac{1}{S^2} \sum_{j \neq i}^S \langle T_{\mu}^{m_i} - I, T_{\mu}^{m_j} - I \rangle_{L^2(\mu)}. \end{aligned}$$

By taking expectation, as the samples  $m_i \sim \Pi$  are independent, we get

$$\begin{aligned} \mathbb{E} \left[ \left\| -\frac{1}{S} \sum_{i=1}^S (T_{\mu}^{m_i} - I) \right\|_{L^2(\mu)}^2 \right] &= \frac{1}{S^2} \sum_{i=1}^S \mathbb{E}[W_2^2(\mu, m_i)] + \frac{1}{S^2} \sum_{j \neq i}^S \langle \mathbb{E}[T_{\mu}^{m_i} - I], \mathbb{E}[T_{\mu}^{m_j} - I] \rangle_{L^2(\mu)} \\ &= \frac{2}{S^2} \sum_{i=1}^S F(\mu) + \frac{1}{S^2} \sum_{j \neq i}^S \langle F'(\mu), F'(\mu) \rangle_{L^2(\mu)} \\ &= \frac{2}{S} F(\mu) + \frac{S-1}{S} \|F'(\mu)\|_{L^2(\mu)}^2. \end{aligned}$$

Subtracting  $\|F'(\mu)\|_{L^2(\mu)}^2$  yields the asserted identities.  $\square$

The mini-batch implementation is easily seen to inherit the convergence estimates established in Theorem 2.

**Theorem 4.** *Under the assumptions of Theorem 2, the BSGD sequence  $\{\mu_k\}_k$  in (23) is almost surely convergent to  $\hat{\mu} \in K_{\Pi}$  as soon as (3) and (4) hold. Moreover, for every  $a > C_{\hat{\mu}}^{-1}$  and  $b \geq a$  there exists an explicit constant  $C_{a,b} > 0$  such that, if  $\gamma_k = a/(b+k)$  for all  $k \in \mathbb{N}$ , the expected optimality gap satisfies*

$$\mathbb{E}[F(\mu_k) - F(\hat{\mu})] \leq \frac{C_{a,b}}{b+k}.$$

## 6. SGD for closed-form Wasserstein barycenters

As discussed in the introduction, recent developments have enabled the approximate computation of optimal transport maps between two given absolutely continuous distributions, and hence the practical implementation of the SGD in fairly general cases is feasible in principle. We will not address this issue in the present work, but rather content ourselves with analyzing some families of models considered in [5, 24] for which this additional algorithmic aspect can be avoided, their optimal transport maps being explicit and easy to evaluate. Further, we will examine some of their closure properties that are preserved under the operation of *taking barycenter*. This is important, for instance, in the context of the statistical application in [9], wherein  $\Pi$  really represents a posterior distribution on models and its barycenter is postulated as a useful representative of the posterior distribution on models. It is thus desirable that the representative model share some of the properties of all the models charged by the posterior.

In the settings that will be presented, the pseudo-associativity condition (5) is either satisfied or conditions ensuring it can be given explicitly. Convergence bounds as in Theorem 2 can be easily deduced in those cases for each specification of  $\Pi$  satisfying the required assumptions.

### 6.1. Univariate distributions

We assume that  $m$  is a continuous distribution over  $\mathbb{R}$ , and denote respectively by  $F_m$  and  $Q_m := F_m^{-1}$  its cumulative distribution function and its right-continuous quantile function. The increasing transport map from a continuous  $m_0$  to  $m$ , also known as the monotone rearrangement, is given by  $T_{m_0}^m(x) = Q_m(F_{m_0}(x))$ , and is known to be  $p$ -Wasserstein optimal for  $p \geq 1$  (see [50, Remark 2.19(iv)]). Given  $\Pi$ , the barycenter  $\hat{m}$  is also independent of  $p$  and characterized via its quantile, i.e.

$$Q_{\hat{m}}(\cdot) = \int Q_m(\cdot) \Pi(dm). \quad (24)$$

In words: the quantile function of the barycenter is equal to the average quantile function. Our SGD iteration, starting from a distribution function  $F_\mu(x)$ , sampling some  $m \sim \Pi$ , and with step  $\gamma$ , produces the measure  $\nu = ((1 - \gamma)I + \gamma T_\mu^m)(\mu)$ . This is characterized by its quantile function,  $Q_\nu(\cdot) = (1 - \gamma)Q_\mu(\cdot) + \gamma Q_m(\cdot)$ . The BSGD iteration is

$$Q_\nu(\cdot) = (1 - \gamma)Q_\mu(\cdot) + \frac{\gamma}{S} \sum_{i=1}^S Q_{m^i}(\cdot).$$

As explained in Remark 2, the barycenter is automatically pseudo-associative, since transport maps (i.e. increasing functions) are in this case associative in the sense discussed therein, and hence Theorem 2 applies. Moreover, by Remark 3, it entails convergence bounds in  $W_2^2$  for the SGD sequence itself.

Interestingly, the model average  $\bar{m} := \int m \Pi(dm)$  is characterized by the *averaged cumulative distribution function*, i.e.  $F_{\bar{m}}(\cdot) = \int F_m(\cdot) \Pi(dm)$ . The model average does not preserve intrinsic *shape* properties from the distributions such as symmetry or unimodality. For example, if  $\Pi = 0.3 \times \delta_{m_1} + 0.7 \times \delta_{m_2}$  with  $m_1 = \mathcal{N}(1, 1)$  and  $m_2 = \mathcal{N}(3, 1)$ , the *model average* is an asymmetric bimodal distribution with modes on 1 and 3, while the Wasserstein barycenter is the Gaussian distribution  $\hat{m} = \mathcal{N}(2.4, 1)$ .

The following reasoning illustrates the fact that Wasserstein barycenters preserve *geometric properties* in a way that, e.g., the model average does not. A continuous distribution  $m$  on  $\mathbb{R}$  is unimodal with a mode on  $\tilde{x}_m$  if its quantile function  $Q(y)$  is concave for  $y < \tilde{y}_m$  and convex for  $y > \tilde{y}_m$ , where  $Q(\tilde{y}_m) = \tilde{x}_m$ . Likewise,  $m$  is symmetric with respect to  $x_m \in \mathbb{R}$  if  $Q(\frac{1}{2} + y) = 2x_m - Q(\frac{1}{2} - y)$  for  $y \in [0, \frac{1}{2}]$  (note that  $x_m = Q_m(\frac{1}{2})$ ). These properties can be analogously described in terms of the function  $F_m$ . Let us show that the barycenter preserves unimodality/symmetry.

**Proposition 4.** *If  $\Pi$  is concentrated on continuous symmetric (resp. symmetric unimodal) univariate distributions, then the barycenter  $\hat{m}$  is symmetric (resp. symmetric unimodal).*

*Proof.* Using the quantile function characterization in (24) we have, for  $y \in [0, \frac{1}{2}]$ ,

$$Q_{\hat{m}}\left(\frac{1}{2} + y\right) = \int Q_m\left(\frac{1}{2} + y\right) \Pi(dm) = 2x_{\hat{m}} - Q_{\hat{m}}\left(\frac{1}{2} - y\right),$$

where  $x_{\hat{m}} := \int x_m \Pi(dm)$ . In other words,  $\hat{m}$  is symmetric with respect to  $x_{\hat{m}}$ . Now, if each  $m$  is unimodal in addition to symmetric, its mode  $\tilde{x}_m$  coincides with its median  $Q_m(\frac{1}{2})$ , and

$Q_m(\lambda y + (1 - \lambda)z) \geq$  (resp.  $\leq$ )  $\lambda Q_m(y) + (1 - \lambda)Q_m(z)$  for all  $y, z <$  (resp.  $>$ )  $\frac{1}{2}$  and  $\lambda \in [0, 1]$ . Integrating these inequalities with respect to  $\Pi(dm)$  and using (24), we deduce that  $\hat{m}$  is unimodal (with mode  $\frac{1}{2}$ ) too.  $\square$

Unimodality is not preserved in general non-symmetric cases. Still, for some families of distributions unimodality is maintained after taking barycenter, as we show in the next result.

**Proposition 5.** *If  $\Pi \in \mathcal{W}_p(\mathcal{W}_{ac}(\mathbb{R}))$  is concentrated on log-concave univariate distributions, then the barycenter  $\hat{m}$  is unimodal.*

*Proof.* If  $f(x)$  is a log-concave density, then  $-\log(f(x))$  is convex and so is  $e^{-\log(f(x))} = 1/f(x)$ . Some computation reveals that  $f(x) dx$  is unimodal for some  $\tilde{x} \in \mathbb{R}$ , so its quantile  $Q(y)$  is concave for  $y < \tilde{y}$  and convex for  $y > \tilde{y}$ , where  $Q(\tilde{y}) = \tilde{x}$ . Since  $1/f(x)$  is convex decreasing for  $x < \tilde{x}$  and convex increasing for  $x > \tilde{x}$ ,  $1/f(Q(y))$  is convex. So  $(dQ/dy)(y) = 1/f(Q(y))$  is convex, positive, and with a minimum at  $\tilde{y}$ . Given  $\Pi$ , its barycenter  $\hat{m}$  satisfies

$$\frac{dQ_{\hat{m}}}{dy} = \int \frac{dQ_m}{dy} \Pi(dm),$$

so if all the  $dQ_m/dy$  are convex, then  $dQ_{\hat{m}}/dy$  is convex and positive, with a minimum at some  $\hat{y}$ . Thus,  $Q_{\hat{m}}(y)$  is concave for  $y < \hat{y}$  and convex for  $y > \hat{y}$ , and  $\hat{m}$  is unimodal with a mode at  $Q_{\hat{m}}(\hat{y})$ .  $\square$

Useful examples of log-concave distribution families include the general normal, exponential, logistic, Gumbel, chi-square, and Laplace laws, as well as the Weibull, power, gamma, and beta families when their shape parameters are larger than one.

## 6.2. Distributions sharing a common copula

If two multivariate distributions  $P$  and  $Q$  over  $\mathbb{R}^q$  share the same copula, then their  $W_p$  distance to the  $p$ th power is the sum of the  $W_p(\mathbb{R})$  distances between their marginals raised to the  $p$ th power. Furthermore, if the marginals of  $P$  have no atoms, then an optimal map is given by the coordinate-wise transformation  $T(x) = (T^1(x_1), \dots, T^q(x_q))$ , where  $T^i(x_i)$  is the monotone rearrangement between the marginals  $P^i$  and  $Q^i$  for  $i = 1, \dots, q$ . This setting allows us to easily extend the results from the univariate case to the multidimensional case.

**Lemma 5.** *If  $\Pi \in \mathcal{W}_p(\mathcal{W}_{ac}(\mathbb{R}^q))$  is concentrated on a set of measures sharing the same copula  $C$ , then the  $p$ -Wasserstein barycenter  $\hat{m}$  of  $\Pi$  has copula  $C$  as well, and its  $i$ th marginal  $\hat{m}^i$  is the barycenter of the  $i$ th marginal measures of  $\Pi$ . In particular, the barycenter does not depend on  $p$ .*

*Proof.* It is known [3, 24] that for two distributions  $m$  and  $\mu$  with respective  $i$ th marginals  $m^i$  and  $\mu^i$  for  $i = 1, \dots, q$ , the  $p$ -Wasserstein metric satisfies  $W_p^p(m, \mu) \geq \sum_{i=1}^q W_p^p(m^i, \mu^i)$ , where equality is reached if  $m$  and  $\mu$  share the same copula  $C$  (we have abused the notation, denoting by  $W_p$  the  $p$ -Wasserstein distance on  $\mathbb{R}^q$  as well as on  $\mathbb{R}$ ). Thus,

$$\int W_p^p(m, \mu) \Pi(dm) \geq \int \sum_{i=1}^q W_p^p(m^i, \mu^i) \Pi(dm) = \sum_{i=1}^q \int W_p^p(\nu, \mu^i) \Pi^i(d\nu),$$

where  $\Pi^i$  is defined via the identity  $\int_{\mathcal{P}(\mathbb{R})} f(\nu) \Pi^i(d\nu) = \int_{\mathcal{P}(\mathbb{R}^q)} f(m^i) \Pi(dm)$ . The infimum for the lower bound is reached on the univariate measures  $\hat{m}^1, \dots, \hat{m}^q$  where  $\hat{m}^i$  is the  $p$ -barycenter

of  $\Pi^i$ , which means that  $\hat{m}^i = \operatorname{argmin} \int W_p^p(v, \mu^i) \Pi^i(dv)$ . It is plain that the infimum is reached on the distribution  $\hat{m}$  with copula  $C$  and  $i$ th marginal  $\hat{m}^i$  for  $i = 1, \dots, q$ , which then has to be the barycenter of  $\Pi$  and is independent of  $p$ .  $\square$

A Wasserstein SGD iteration, starting from a distribution  $\mu$ , sampling  $m \sim \Pi$ , and with step  $\gamma$ , both  $\mu$  and  $m$  having copula  $C$ , produces the measure  $\nu = ((1 - \gamma)I + \gamma T_\mu^m)(\mu)$  characterized by having copula  $C$  and the  $i$ th marginal quantile functions  $Q_{\nu^i}(\cdot) = (1 - \gamma)Q_{\mu^i}(\cdot) + \gamma Q_{m^i}(\cdot)$  for  $i = 1, \dots, q$ . The BSGD iteration works analogously. Alternatively, we can perform (batch) stochastic gradient descent component-wise (with respect to the marginals  $\Pi^i$  of  $\Pi$ ) and then make use of the copula  $C$ . As in the one-dimensional case, the barycenter is in this case automatically pseudo-associative since it is associative. Bounds for the expected optimality gap and in  $W_2^2$  for the SGD sequence can be similarly deduced in this case too.

### 6.3. Spherically equivalent distributions

We denote here by  $\mathcal{L}(\cdot)$  the law of a random vector, so  $m = \mathcal{L}(x)$  and  $x \sim m$  are synonyms. Following [24], another multidimensional case is constructed as follows: Given a fixed measure  $\tilde{m} \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$ , its associated family of spherically equivalent distributions is

$$\mathcal{S}_0 := \mathcal{S}(\tilde{m}) = \left\{ \mathcal{L} \left( \frac{\alpha(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2} \tilde{x} \right) \mid \alpha \in \mathcal{ND}(\mathbb{R}), \tilde{x} \sim \tilde{m} \right\},$$

where  $\|\cdot\|_2$  is the Euclidean norm and  $\mathcal{ND}(\mathbb{R})$  is the set of non-decreasing non-negative functions of  $\mathbb{R}_+$ . These types of distributions include the simplicially contoured distributions, and also elliptical distributions with the same correlation structure.

If  $y \sim m \in \mathcal{S}_0$ , then  $\alpha(r) = Q_{\|y\|_2}(F_{\|\tilde{x}\|_2}(r))$ , where  $Q_{\|y\|_2}$  is the quantile function of the norm of  $y$ ,  $F_{\|\tilde{x}\|_2}$  is the distribution function of the norm of  $\tilde{x}$ , and

$$y \sim \frac{\alpha(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2} \tilde{x}.$$

More generally, if

$$m_1 = \mathcal{L} \left( \frac{\alpha_1(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2} \tilde{x} \right), \quad m_2 = \mathcal{L} \left( \frac{\alpha_2(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2} \tilde{x} \right),$$

then the optimal transport from  $m_1$  to  $m_2$  is given by

$$T_{m_1}^{m_2}(x) = \frac{\alpha(\|x\|_2)}{\|x\|_2} x,$$

where  $\alpha(r) = Q_{\|x_2\|_2}(F_{\|x_1\|_2}(r))$ . Since  $F_{\|x_1\|_2}(r) = F_{\|\tilde{x}\|_2}(\alpha_1^{-1}(r))$  and  $Q_{\|x_2\|_2}(r) = \alpha_2(Q_{\|\tilde{x}\|_2}(r))$ , we see that  $\alpha(r) = \alpha_2(Q_{\|\tilde{x}\|_2}(F_{\|\tilde{x}\|_2}(\alpha_1^{-1}(r)))) = \alpha_2(\alpha_1^{-1}(r))$ , so finally

$$T_{m_1}^{m_2}(x) = \frac{\alpha_2(\alpha_1^{-1}(\|x\|_2))}{\|x\|_2} x.$$

Note that these kinds of transports are closed under composition and convex combination, and contain the identity. An SGD iteration, starting from a distribution

$$\mu = \mathcal{L} \left( \frac{\alpha_0(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2} \tilde{x} \right),$$

sampling

$$m = \mathcal{L}\left(\frac{\alpha(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2} \tilde{x}\right) \sim \Pi$$

with step  $\gamma$ , produces  $m_1 = T_0^{\gamma,m}(\mu) := ((1 - \gamma)I + \gamma T_\mu^m)(\mu)$ . Since

$$T_0^{\gamma,m}(x) = \frac{(\gamma\alpha + (1 - \gamma)\alpha_0)(\alpha_0^{-1}(\|x\|_2))}{\|x\|_2} x,$$

we have

$$m_1 = \mathcal{L}\left(\frac{\alpha_1(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2} \tilde{x}\right)$$

with  $\alpha_1 = \gamma\alpha + (1 - \gamma)\alpha_0$ . Analogously, the batch stochastic gradient iteration produces  $\alpha_1 = (1 - \gamma)\alpha_0 + (\gamma/S) \sum_{i=1}^S \alpha_{m^i}$ . Note that these iterations live in  $\mathcal{S}_0$ , and thus the barycenter  $\hat{m} \in \mathcal{S}_0$ .

For the barycenter

$$\hat{m} = \mathcal{L}\left(\frac{\hat{\alpha}(\|\tilde{x}\|_2)}{\|\tilde{x}\|_2} \tilde{x}\right),$$

the equation  $\int T_{\hat{m}}^m(x) \Pi(dm) = x$  can be expressed as  $\hat{\alpha}(r) = \int \alpha_m(r) \Pi(dm)$ , or equivalently  $\mathcal{Q}_{\|\hat{y}\|_2}^{\hat{m}}(p) = \int \mathcal{Q}_{\|y\|_2}^m(p) \Pi(dm)$ , where  $\mathcal{Q}_{\|y\|_2}^m$  is the quantile function of the norm of  $y \sim m$ . This is similar to the univariate setting and, as in that case, the barycenter is associative, and convergence bounds for the optimality gap and the SDG sequence can be established.

#### 6.4. Scatter-location family

We borrow here the setting of [5], where another useful multidimensional case is defined as follows: Given a fixed distribution  $\tilde{m} \in \mathcal{W}_{2,\text{ac}}(\mathbb{R}^q)$ , referred to as the *generator*, the generated scatter-location family is given by

$$\mathcal{F}(\tilde{m}) := \{\mathcal{L}(A\tilde{x} + b) \mid A \in \mathcal{M}_+^{q \times q}, b \in \mathbb{R}^q, \tilde{x} \sim \tilde{m}\},$$

where  $\mathcal{M}_+^{q \times q}$  is the set of symmetric positive-definite matrices of size  $q \times q$ . Without loss of generality we can assume that  $\tilde{m}$  has zero mean and identity covariance. A clear example is the multivariate Gaussian family  $\mathcal{F}(\tilde{m})$ , with  $\tilde{m}$  a standard multivariate normal distribution.

The optimal transport between two members of  $\mathcal{F}(\tilde{m})$  is explicit. If  $m_1 = \mathcal{L}(A_1\tilde{x} + b_1)$  and  $m_2 = \mathcal{L}(A_2\tilde{x} + b_2)$  then the optimal map from  $m_1$  to  $m_2$  is given by  $T_{m_1}^{m_2}(x) = A(x - b_1) + b_2$ , where  $A = A_1^{-1}(A_1A_2^2A_1)^{1/2}A_1^{-1} \in \mathcal{M}_+^{q \times q}$ . The family  $\mathcal{F}(\tilde{m})$  is therefore geodesically closed.

If  $\Pi$  is supported on  $\mathcal{F}(\tilde{m})$ , then its 2-Wasserstein barycenter  $\hat{m}$  belongs to  $\mathcal{F}(\tilde{m})$ . Call its mean  $\hat{b}$  and its covariance matrix  $\hat{\Sigma}$ . Since the optimal map from  $\hat{m}$  to  $m$  is  $T_{\hat{m}}^m(x) = A_m^m(x - \hat{b}) + b_m$  where  $A_m^m = \hat{\Sigma}^{-1/2}(\hat{\Sigma}^{1/2}\Sigma_m\hat{\Sigma}^{1/2})^{1/2}\hat{\Sigma}^{-1/2}$ , and we know that  $\hat{m}$ -almost surely  $\int T_{\hat{m}}^m(x) \Pi(dm) = x$ , then we must have  $\int A_m^m \Pi(dm) = I$ , since clearly  $\hat{b} = \int b_m \Pi(dm)$ . As a consequence, we have  $\hat{\Sigma} = \int (\hat{\Sigma}^{1/2}\Sigma_m\hat{\Sigma}^{1/2})^{1/2} \Pi(dm)$ .

A stochastic gradient descent iteration, starting from a distribution  $\mu = \mathcal{L}(A_0\tilde{x} + b_0)$ , sampling some  $m = \mathcal{L}(A_m\tilde{x} + b_m) \sim \Pi$ , and with step  $\gamma$ , produces the measure  $\nu = T_0^{\gamma,m}(\mu) := ((1 - \gamma)I + \gamma T_\mu^m)(\mu)$ . If  $\tilde{x}$  has a multivariate distribution  $\tilde{F}(x)$ , then  $\mu$  has distribution  $F_0(x) = \tilde{F}(A_0^{-1}(x - b_0))$  with mean  $b_0$  and covariance  $\Sigma_0 = A_0^2$ . We have  $T_0^{\gamma,m}(x) = ((1 - \gamma)I + \gamma A_\mu^m)(x - b_0) + \gamma b_m + (1 - \gamma)b_0$  with  $A_\mu^m := A_0^{-1}(A_0A_m^2A_0)^{1/2}A_0^{-1}$ . Then

$$F_\nu(x) =: F_1(x) = F_0([T_0^{\gamma,m}]^{-1}(x)) = \tilde{F}([(1 - \gamma)A_0 + \gamma A_\mu^m A_0]^{-1}(x - \gamma b_m - (1 - \gamma)b_0)),$$



with mean  $b_1 = (1 - \gamma)b_0 + \gamma b_m$  and covariance

$$\begin{aligned}\Sigma_1 &= A_1^2 = [(1 - \gamma)A_0 + \gamma A_0^{-1}(A_0 A_m^2 A_0)^{1/2}][(1 - \gamma)A_0 + \gamma (A_0 A_m^2 A_0)^{1/2} A_0^{-1}] \\ &= A_0^{-1}[(1 - \gamma)A_0^2 + \gamma (A_0 A_m^2 A_0)^{1/2}][(1 - \gamma)A_0^2 + \gamma (A_0 A_m^2 A_0)^{1/2}]A_0^{-1} \\ &= A_0^{-1}[(1 - \gamma)A_0^2 + \gamma (A_0 A_m^2 A_0)^{1/2}]^2 A_0^{-1}.\end{aligned}$$

The batch stochastic gradient descent iteration is characterized by

$$b_1 = (1 - \gamma)b_0 + \frac{\gamma}{S} \sum_{i=1}^S b_{m^i}, \quad A_1^2 = A_0^{-1} \left[ (1 - \gamma)A_0^2 + \frac{\gamma}{S} \sum_{i=1}^S (A_0 A_{m^i}^2 A_0)^{1/2} \right]^2 A_0^{-1}.$$

Notice that if all the matrices  $\{A_m : m \in \text{supp}(\Pi)\}$  share the same eigenspaces, then the barycenter is pseudo-associative, as it is in fact associative. In the case that these matrices do not necessarily share the same eigenspaces, and  $\tilde{m}$  is Gaussian, it is still possible to give conditions under which the barycenter is pseudo-associative: as we have already mentioned, this property holds in the setting of [19] under conditions of uniform boundedness and uniform positivity of covariance matrices.

### Appendix A. Some remarks in the case of general measures

We discuss a natural counterpart to the SGD sequence in the case where  $\Pi$  is not necessarily supported on absolutely continuous measures. However, in this case we neither have a verifiable result guaranteeing the convergence of the SGD method, nor a verifiable result saying that the accumulation points of the SGD sequence are necessarily Karcher means. The goal of this section is to introduce the objects needed for an analysis of the general case, to advance the convergence analysis as much as possible, and to illustrate the difficulties that we encounter that make us stop short of obtaining general convergence results.

While we retain Assumption 1, we modify Assumption 2 into the following.

**Assumption 2''.**  $\Pi = \sum_{j=1}^{\ell} \lambda_j \delta_{v_j}$ , with each  $v_j \in \mathcal{W}_2(\mathcal{X})$  and where the  $\lambda_j \in (0, 1)$  sum to 1.

**Remark 5.** We can find  $V : \mathcal{X} \rightarrow [0, \infty)$  convex, continuous, and super-quadratic (i.e.  $\lim_{|y| \rightarrow \infty} V(y)/|y|^2 = +\infty$ ) and  $C \in (0, \infty)$  such that  $\int V dv_j \leq C$  for each  $j$ . Defining

$$K_{\Pi} = \left\{ \mu : \int V d\mu \leq C \right\}, \quad (25)$$

it follows that  $K_{\Pi}$  is compact in  $\mathcal{W}_2(\mathcal{X})$  and  $\Pi(K_{\Pi}) = 1$ . Moreover, if  $(X, Y)$  is any optimal coupling with marginals  $\mu$  and  $\nu$ , with  $\mu$  and  $\nu$  in  $K_{\Pi}$ , then also  $\text{Law}(tX + (1 - t)Y) \in K_{\Pi}$  for each  $t \in (0, 1)$ . In the following, we denote by  $K_{\Pi}$  a set with these properties which may or may not be given explicitly as in (25).

Throughout, we denote by  $\text{Opt}(\mu, \nu)$  the set of optimal couplings attaining the infimum defining  $W_2(\mu, \nu)$ , which is non-empty and may contain multiple elements, and we fix

$$\mathcal{W}_2(\mathcal{X})^2 \ni (\mu, \nu) \mapsto G(\mu, \nu) \in \text{Opt}(\mu, \nu),$$

a measurable selection of optimal couplings. In practical terms an algorithm for computing or approximating optimal couplings would be automatically measurable.

**Definition 7.** Let  $\mu_0 \in K_\Pi$ ,  $m_k \stackrel{\text{i.i.d.}}{\sim} \Pi$ , and  $\gamma_k > 0$  for  $k \geq 0$ . We define the SGD sequence by

$$\mu_{k+1} := \text{Law}[(1 - \gamma_k)X + \gamma_k Y], \quad (26)$$

where  $\text{Law}(X, Y) = G(\mu_k, m_k)$ .

By Remark 5 we have  $\{\mu_k : k \in \mathbb{N}\} \subset K_\Pi$  (almost surely).

**Definition 8.**  $\mu \in \mathcal{W}_2(\mathcal{X})$  is a Karcher mean of  $\Pi$  if,  $\mu(\text{d}x)$ -almost surely,  $x = \sum_{j=1}^\ell \lambda_j \mathbb{E}[Y_j | X = x]$ , where (for each  $j$ )  $\text{Law}(X, Y_j)$  is some coupling in  $\text{Opt}(\mu, \nu_j)$ .

We first observe that if  $\mu$  is absolutely continuous, then it is a Karcher mean according to Definition 8 if and only if it is a Karcher mean according to Definition 3. On the other hand, if  $\mu$  is a barycenter of  $\Pi$  then it is also a Karcher mean according to Definition 8. Indeed, if  $\text{Law}(X, Y_j) \in \text{Opt}(\mu, \nu_j)$  and we couple all these random variables in the same probability space, then

$$\sum_{i=1}^\ell \lambda_i W_2^2(\mu, \nu_i) = \mathbb{E} \left[ \sum_{i=1}^\ell \lambda_i |X - Y_i|^2 \right] \geq \mathbb{E} \left[ \sum_{j=1}^\ell \lambda_j \left| Y_j - \sum_{i=1}^\ell \lambda_i Y_i \right|^2 \right] \geq \sum_{i=1}^\ell \lambda_i W_2^2(\tilde{\mu}, \nu_i),$$

so the inequalities above must be actual equalities and we have  $\mu = \text{Law}(\sum_i \lambda_i Y_i) := \tilde{\mu}$  as well as  $X = \sum_i \lambda_i Y_i$  (almost surely). Taking conditional expectation with respect to  $X$  in the latter, we conclude.

In direct analogy to the absolutely continuous case, we write  $F(\mu) := \frac{1}{2} \sum_j \lambda_j W_2^2(\mu, \nu_j)$ , and we introduce  $F'(\mu)(x) := x - \sum_{j=1}^\ell \lambda_j \mathbb{E}[Y_j | X = x]$ , where  $(X, Y_j) \sim G(\mu, \nu_j)$ . With this notation we have the implication that  $\|F'(\mu)\|_{L^2(\mu)} = 0 \Rightarrow \mu$  is a Karcher mean. As before, we denote by  $\mathcal{F}_0$  the trivial sigma-algebra and  $\mathcal{F}_{k+1}$ ,  $k \geq 0$ , the sigma-algebra generated by  $m_0, \dots, m_k$ .

**Proposition 6.** The SGD sequence in (26) satisfies, almost surely,

$$\mathbb{E}[F(\mu_{k+1}) - F(\mu_k) | \mathcal{F}_k] \leq \gamma_k^2 F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2.$$

*Proof.* We can define a coupling  $(X_k, Y_k, Z_i)$  such that  $\text{Law}(X_k, Y_k) \in G(\mu_k, m_k)$ ,  $\text{Law}(X_k, Z_i) \in G(\mu_k, \nu_i)$ , and where  $Y_k$  and  $Z_i$  are independent conditionally on  $\mathcal{F}_k$ . Then the coupling  $((1 - \gamma_k)X_k + \gamma_k Y_k, Z_i)$  has first marginal  $\mu_{k+1}$  and second marginal  $\nu_i$ . We compute

$$\begin{aligned} W_2^2(\mu_{k+1}, \nu_i) &\leq \mathbb{E}[|(1 - \gamma_k)X_k + \gamma_k Y_k - Z_i|^2] \\ &= \mathbb{E}[|X_k - Z_i|^2] - 2\gamma_k \mathbb{E}[\langle X_k - Z_i, X_k - Y_k \rangle] + \gamma_k^2 \mathbb{E}[|X_k - Y_k|^2]. \end{aligned}$$

Hence,

$$\begin{aligned} F(\mu_{k+1}) &= \frac{1}{2} \int W_2^2(\mu_{k+1}, \nu) \Pi(d\nu) \\ &\leq \frac{1}{2} \sum_{i=1}^{\ell} \lambda_i \mathbb{E}[|X_k - Z_i|^2] - \gamma_k \mathbb{E} \left[ \left\langle X_k - \sum_{i=1}^{\ell} \lambda_i Z_i, X_k - Y_k \right\rangle \right] + \frac{1}{2} \gamma_k^2 \mathbb{E}[|X_k - Y_k|^2] \\ &= F(\mu_k) - \gamma_k \mathbb{E} \left[ \left\langle X_k - \sum_{i=1}^{\ell} \lambda_i Z_i, X_k - Y_k \right\rangle \right] + \frac{1}{2} \gamma_k^2 W_2^2(\mu_k, m_k). \end{aligned}$$

Taking conditional expectation with respect to  $\mathcal{F}_k$ , and as  $m_k$  is independently sampled from this sigma-algebra, we see that

$$\mathbb{E}[W_2^2(\mu_k, m_k) \mid \mathcal{F}_k] = \sum_{i=1}^{\ell} \lambda_i W_2^2(\mu_k, \nu_i) = 2F(\mu_k).$$

Similarly, we have

$$\mathbb{E} \left[ \mathbb{E} \left[ \left\langle X_k - \sum_{i=1}^{\ell} \lambda_i Z_i, X_k - Y_k \right\rangle \mid \mathcal{F}_k \right] \right] = \mathbb{E} \left[ \left\langle X_k - \sum_{i=1}^{\ell} \lambda_i \mathbb{E}[Z_i \mid X_k], X_k - \sum_{i=1}^{\ell} \lambda_i \mathbb{E}[Y_i \mid X_k] \right\rangle \right].$$

We conclude that  $\mathbb{E}[F(\mu_{k+1}) \mid \mathcal{F}_k] \leq (1 + \gamma_k^2)F(\mu_k) - \gamma_k \|F'(\mu_k)\|_{L^2(\mu_k)}^2$ .  $\square$

The next lemma is the only part where we use that  $\Pi$  is supported in finitely many measures. With more refined arguments we could surely avoid this limitation.

**Lemma 6.** *Let  $(\rho_n)_n \subset \mathcal{W}_2(\mathbb{R}^q)$  be a sequence converging with respect to  $W_2$  to  $\rho \in \mathcal{W}_2(\mathbb{R}^q)$ . Then, as  $n \rightarrow \infty$ ,*

- (i)  $F(\rho_n) \rightarrow F(\rho)$ ;
- (ii)  $\|F'(\rho_n)\|_{L^2(\rho_n)} \rightarrow 0$  implies that  $\rho$  is a Karcher mean.

*Proof.* Part (i) is immediate since  $W_2(\rho_n, \nu_j) \rightarrow W_2(\rho, \nu_j)$  for each  $j$ . For part (ii), we first observe that  $G(\rho_n, \nu_j)$  is tight for each  $j$ . Passing to a subsequence if necessary, the Skorokhod representation theorem yields, on some probability space, a sequence of random variables  $X^n \sim \rho_n$  and  $Y_j^n \sim \nu_j$ , as well as  $X \sim \rho$  and  $Y_j \sim \nu_j$ , such that  $X^n \rightarrow X$  and also, for each  $j$ ,  $Y_j^n \rightarrow Y_j$  (convergence is a.s. and in  $L^2$ ). The sequence  $\{\mathbb{E}[Y_j^n \mid X^n] : n \in \mathbb{N}\}$  is  $L^2$ -bounded, and hence by repeatedly applying Komlos' theorem, we find yet another subsequence (re-labeled so it is indexed by  $\mathbb{N}$ ) such that  $(1/n) \sum_{r \leq n} \mathbb{E}[Y_j^r \mid X^r] \rightarrow Z_j$  almost surely and in  $L^2$ , for each  $j$ .

We check that  $\mathbb{E}[Z_j \mid X] = \mathbb{E}[Y_j \mid X]$ . Indeed, if  $g$  is bounded and continuous,

$$\begin{aligned}
\mathbb{E}[g(X)Z_j] &= \lim_n \mathbb{E} \left[ g(X) \left( \frac{1}{n} \sum_{r \leq n} E[Y_j^r | X^r] \right) \right] \\
&= \lim_n \frac{1}{n} \sum_{r \leq n} \mathbb{E}[g(X^r)E[Y_j^r | X^r]] + \frac{1}{n} \sum_{r \leq n} \mathbb{E}[\{g(X) - g(X^r)\}E[Y_j^r | X^r]] \\
&= \lim_n \frac{1}{n} \sum_{r \leq n} \mathbb{E}[g(X^r)Y_j^r] = \mathbb{E}[g(X)Y_j]
\end{aligned}$$

by dominated convergence. Also,

$$\begin{aligned}
0 &= \lim_n \|F'(\rho_n)\|_{L^2(\rho_n)}^2 \\
&= \lim_n \mathbb{E} \left[ \left| X^n - \sum_{j=1}^{\ell} \lambda_j \mathbb{E}[Y_j^n | X^n] \right|^2 \right] \\
&= \lim_n \frac{1}{n} \sum_{r \leq n} \mathbb{E} \left[ \left| X^r - \sum_{j=1}^{\ell} \lambda_j \mathbb{E}[Y_j^r | X^r] \right|^2 \right] \\
&\geq \lim_n \mathbb{E} \left[ \left| \frac{1}{n} \sum_{r \leq n} X^r - \sum_{j=1}^{\ell} \lambda_j \frac{1}{n} \sum_{r \leq n} \mathbb{E}[Y_j^r | X^r] \right|^2 \right] \\
&\geq \mathbb{E} \left[ \left| X - \sum_{j=1}^{\ell} \lambda_j Z_j \right|^2 \right] \geq \mathbb{E} \left[ \left| X - \sum_{j=1}^{\ell} \lambda_j \mathbb{E}[Z_j | X] \right|^2 \right],
\end{aligned}$$

and we conclude that  $X = \sum_{j=1}^{\ell} \lambda_j \mathbb{E}[Y_j | X]$ , almost surely. Finally, we remark that  $\text{Law}(X, Y_j) \in \text{Opt}(\rho, \nu_j)$ , and hence we conclude that  $\rho$  is a Karcher mean.  $\square$

We can now provide a convergence result. The hypotheses of this result are implied by the hypotheses of Theorem 1 in the case that  $\Pi$  is concentrated on absolutely continuous measures. However, we stress that we do not know of any example where Theorem 5 is applicable and  $\Pi$  is concentrated on non-absolutely continuous measures. For this reason we rather think that this result and its proof are a template of what *could happen* or *could be done* in the general case, and use it really to illustrate the difficulties present in the general case.

**Theorem 5.** Assume Assumptions 1 and 2'', conditions (3) and (4), that  $\Pi$  admits a unique Karcher mean (in the sense of Definition 8) in  $K_{\Pi}$ , and that  $K_{\Pi}$  contains a (hence, exactly one) 2-Wasserstein barycenter  $\hat{\mu}$ . Then, the SGD sequence  $\{\mu_k\}_k$  in (26) is almost surely convergent to  $\hat{\mu}$ .

*Proof.* The proof of Theorem 1 can be followed verbatim up to (15), namely that  $\liminf_{t \rightarrow \infty} \|F'(\mu_t)\|_{L^2(\mu_t)}^2 = 0$  almost surely.

We observe that if  $\rho_n \rightarrow \rho$  with  $\{\rho_n\}_n \subset K_{\Pi}$  is such that  $F(\rho_n) \geq F(\hat{\mu}) + \varepsilon$ , for  $\varepsilon > 0$ , and  $\|F'(\rho_n)\|_{L^2(\rho_n)}^2 \rightarrow 0$ , then by Lemma 6  $\rho$  is a Karcher mean in  $K_{\Pi}$  that is necessarily different from  $\hat{\mu}$ . This would contradict the uniqueness of Karcher means in  $K_{\Pi}$ . Thus we establish that,

for all  $\varepsilon > 0$ ,  $\inf_{\{\rho: F(\rho) \geq \hat{F} + \varepsilon\} \cap K_\Pi} \|F'(\rho)\|_{L^2(\rho)}^2 > 0$ . From here on we can again follow the proof of Theorem 1.  $\square$

In the absolutely continuous case covered in the rest of this paper, we required the Karcher mean to be absolutely continuous. Moreover, in this case there is a unique barycenter (automatically absolutely continuous). Once we leave the absolutely continuous world we have to be more careful, as illustrated by the following example.

**Example 2.** Let  $Z$  be an  $\mathcal{X}$ -valued random variable with finite second moment, and let  $b_j \in \mathcal{X}$  with  $\sum_j \lambda_j b_j = 0$ . Define  $\nu_j := \text{Law}(Z + b_j)$ . Then  $X \sim \mu$  is a Karcher mean if and only if  $X = \mathbb{E}[Z | X]$  for some optimal coupling of  $X$  and  $Z$ . Hence,  $X := Z$  and  $X := \mathbb{E}[Z]$  are both Karcher means per Definition 8, even if  $Z$  (and so each  $\nu_j$ ) is absolutely continuous. On the other hand, we can check that the barycenter of the  $\nu_j$  is unique and given by  $\text{Law}(Z)$ .

**Example 3.** We continue in the setting of the previous example and choose  $\mathcal{X} = \mathbb{R}^2$ . Here,  $Z = (B, 0)$  with  $B \in \{-1, 1\}$  with equal probabilities, and  $\nu_1 = \text{Law}((B, 1))$ ,  $\nu_2 = \text{Law}((B, -1))$ . In this case we check that  $\mu^p := p(\delta_{(1,0)} + \delta_{(-1,0)}) + (1-2p)\delta_{(0,0)}$  is a Karcher mean for each  $p \in [0, \frac{1}{2}]$ . In particular, the barycenter, given by  $\mu^1$ , is not isolated in the sense that any ball around  $\mu^1$  contains other Karcher means (taking  $p$  close to 0.5), and moreover those Karcher means might be more ‘regular’ than the barycenter in the sense of having a strictly larger support. Also note that if we define  $K := \{\text{Law}((B, r)) : r \in I\}$  with  $I$  some interval containing  $\pm 1$ , then  $K$  is compact, contains the  $\nu_i$ , and transport maps between elements in  $K$  are associative.

Examples 2 and 3 show that uniqueness of Karcher means (in the sense of Definition 8), or even their uniqueness within a nice set  $K_\Pi$ , is an assumption that is difficult to verify in the general case. Furthermore, we encounter the same problem with a localized version of this assumption (for example: ‘uniqueness of Karcher means in a small ball intersected with  $K_\Pi$ ’). To complicate the matter further, it is known that in the general case multiple barycenters may exist. For all these reasons we are inclined to believe that the correct object of study in the general case is the regularized barycenter problem (or, more ambitiously, the limit of such regularized problems as the regularization parameter goes to zero at a suitable speed). We refer the reader to [20, 22], and the references therein, for promising results in this direction.

### Acknowledgements

We thank two anonymous referees for their valuable comments and questions that motivated some improvements of our results and of the presentation of the paper.

### Funding information

This research was funded in whole or in part by the by the Austrian Science Fund (FWF) DOI 10.55776/P36835 (JB) and the ANID-Chile grants Fondecyt-Regular 1201948 (JF) and 1210606 (FT); Center for Mathematical Modeling ACE210010 and FB210005 (JF, FT); and Advanced Center for Electrical and Electronic Engineering FB0008 (FT).

### Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

## References

- [1] AGUEH, M. AND CARLIER, G. (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43**, 904–924.
- [2] AHIDAR-COUTRIX, A., LE GOUIC, T. AND PARIS, Q. (2020). Convergence rates for empirical barycenters in metric spaces: Curvature, convexity and extendable geodesics. *Prob. Theory Relat. Fields* **177**, 323–368.
- [3] ALFONSI, A. AND JOURDAIN, B. (2014). A remark on the optimal transport between two probability measures sharing the same copula. *Statist. Prob. Lett.* **84**, 131–134.
- [4] ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. AND MATRÁN, C. (2016). A fixed-point approach to barycenters in Wasserstein space. *J. Math. Anal. Appl.* **441**, 744–762.
- [5] ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. AND MATRÁN, C. (2018). Wide consensus aggregation in the Wasserstein space. Application to location-scatter families. *Bernoulli* **24**, 3147–3179.
- [6] AMBROSIO, L., GIGLI, N. AND SAVARÉ, G. (2008). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd edn. Birkhäuser, Basel.
- [7] ANGENT, S., HAKER, S. AND TANNENBAUM, A. (2003). Minimizing flows for the Monge–Kantorovich problem. *SIAM J. Math. Anal.* **35**, 61–97.
- [8] BACKHOFF-VERAGUAS, J., BEIGLBÖCK, M. AND PAMMER, G. (2019). Existence, duality, and cyclical monotonicity for weak transport costs. *Calc. Var. Partial Differ. Equ.* **58**, 203.
- [9] BACKHOFF-VERAGUAS, J., FONTBONA, J., RIOS, G. AND TOBAR, F. (2022). Bayesian learning with Wasserstein barycenters. *ESAIM: Prob. Statist.* **26**, 436–472.
- [10] BENAMOU, J.-D. AND BRENIER, Y. (2000). A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik* **84**, 375–393.
- [11] BERCU, B. AND BIGOT, J. (2021). Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *Ann. Statist.* **49**, 968–987.
- [12] BIGOT, J. AND KLEIN, T. (2018). Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Prob. Statist.* **22**, 35–57.
- [13] BONNABEL, S. (2013). Stochastic gradient descent on Riemannian manifolds. *IEEE Trans. Automatic Control* **58**, 2217–2229.
- [14] BONNEEL, N., VAN DE PANNE, M., PARIS, S. AND HEIDRICH, W. (2011). Displacement interpolation using Lagrangian mass transport. *ACM Trans. Graph.* **30**, 1–12.
- [15] BOTTOU, L., CURTIS, F. E. AND NOCEDAL, J. (2018). Optimization methods for large-scale machine learning. *SIAM Rev.* **60**, 223–311.
- [16] CARLIER, G., DELALANDE, A. AND MERIGOT, Q. (2022). Quantitative stability of barycenters in the Wasserstein space. Preprint, arXiv:2209.10217.
- [17] CAZELLES, E., TOBAR, F. AND FONTBONA, J. (2021). A novel notion of barycenter for probability distributions based on optimal weak mass transport. In *Advances in Neural Information Processing Systems*, Vol. **34**, eds M. Ranzato et al.
- [18] CHEN, Y. AND LI, W. (2020). Optimal transport natural gradient for statistical manifolds with continuous sample space. *Inf. Geom.* **3**, 1–32.
- [19] CHEWI, S., MAUNU, T., RIGOLLET, P. AND STROMME, A. J. (2020). Gradient descent algorithms for Bures–Wasserstein barycenters. *Proc. Mach. Learn. Res.* **125**, 1276–1304.
- [20] CHIZAT, L. (2023). Doubly regularized entropic Wasserstein barycenters. Preprint, arXiv:2303.11844.
- [21] CHIZAT, L. AND BACH, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proc. 32nd Conf. Neural Inf. Proc. Syst.*, pp. 3036–3046.
- [22] CHIZAT, L. AND VAŠKEVICIUS, T. (2023). Computational guarantees for doubly entropic Wasserstein barycenters via damped Sinkhorn iterations. Preprint, arXiv:2307.13370.
- [23] CUESTA-ALBERTOS, J., MATRÁN, C. AND TUERO-DÍAZ, A. (1997). Optimal transportation plans and convergence in distribution. *J. Multivar. Anal.* **60**, 72–83.
- [24] CUESTA-ALBERTOS, J., RUSCHENDORF, L. AND TUERO-DÍAZ, A. (1993). Optimal coupling of multivariate distributions and stochastic processes. *J. Multivar. Anal.* **46**, 335–361.
- [25] CUTURI, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. 26th Int. Conf. Adv. Neural Inf. Proc. Syst.*, pp. 2292–2300.
- [26] CUTURI, M. AND DOUCET, A. (2014). Fast computation of Wasserstein barycenters. *Proc. Mach. Learn. Res.* **32**, 685–693.
- [27] CUTURI, M. AND PEYRÉ, G. (2016). A smoothed dual approach for variational Wasserstein problems. *SIAM J. Imag. Sci.* **9**, 320–343.
- [28] DEB, N., GHOSAL, P. AND SEN, B. (2021). Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. In *Advances in Neural Information Processing Systems*, Vol. **34**, eds M. Ranzato et al., pp. 29736–29753.

- [29] DOGNIN, P. *et al.* (2018). Wasserstein barycenter model ensembling. In *Proc. 7th Int. Conf. Learning Representations*.
- [30] FONTBONA, J., GUÉRIN, H. AND MÉLÉARD, S. (2010). Measurability of optimal transportation and strong coupling of martingale measures. *Electron. Commun. Probab.* **15**, 124–133.
- [31] GOZLAN, N. AND JUILLET, N. (2020). On a mixture of Brenier and Strassen theorems. *Proc. London Math. Soc.* **120**, 434–463.
- [32] GOZLAN, N., ROBERTO, C., SAMSON, P.-M. AND TETALI, P. (2017). Kantorovich duality for general transport costs and applications. *J. Funct. Anal.* **273**, 3327–3405.
- [33] HÜTTER, J.-C. AND RIGOLLET, P. (2021). Minimax estimation of smooth optimal transport maps. *Ann. Statist.* **49**, 1166–1194.
- [34] KENT, C., LI, J., BLANCHET, J. AND GLYNN, P. W. (2021). Modified Frank Wolfe in probability space. In *Advances in Neural Information Processing Systems*, Vol. **34**, eds M. Ranzato *et al.*, pp. 14448–14462.
- [35] KIM, Y.-H. AND PASS, B. (2017). Wasserstein barycenters over Riemannian manifolds. *Adv. Math.* **307**, 640–683.
- [36] KINGMA, D. P. AND WELLING, M. (2019). An introduction to variational autoencoders. *Found. Trends Mach. Learn.* **12**, 307–392.
- [37] KOROTIN, A. *et al.* (2021). Do neural optimal transport solvers work? A continuous Wasserstein-2 benchmark. In *Advances in Neural Information Processing Systems*, Vol. **34**, eds M. Ranzato *et al.*, pp. 14593–14605.
- [38] LE GOUIC, T. AND LOUBES, J.-M. (2017). Existence and consistency of Wasserstein barycenters. *Prob. Theory Relat. Fields* **168**, 901–917.
- [39] LI, W. AND MONTÚFAR, G. (2018). Natural gradient via optimal transport. *Inf. Geom.* **1**, 181–214.
- [40] LOEPER, G. AND RAPETTI, F. (2005). Numerical solution of the Monge–Ampère equation by a Newton’s algorithm. *Comptes Rendus Math.* **340**, 319–324.
- [41] MALLASTO, A., GEROLIN, A. AND MINH, H. Q. (2021). Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Inf. Geom.* **5**, 289–323.
- [42] MANOLE, T., BALAKRISHNAN, S., NILES-WEED, J. AND WASSERMAN, L. (2021). Plugin estimation of smooth optimal transport maps. Preprint, arXiv:2107.12364.
- [43] PANARETOS, V. M. AND ZEMEL, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer, New York.
- [44] PEYRÉ, G. AND CUTURI, M. (2019). Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **11**, 355–607.
- [45] POOLADIAN, A.-A. AND NILES-WEED, J. (2021). Entropic estimation of optimal transport maps. Preprint, arXiv:2109.12004.
- [46] ROBBINS, H. AND MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22**, 400–407.
- [47] SINKHORN, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.* **35**, 876–879.
- [48] SINKHORN, R. AND KNOPP, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* **21**, 343–348.
- [49] SOLOMON, J. *et al.* (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graphics* **34**, 1–11.
- [50] VILLANI, C. (2003). *Topics in Optimal Transportation* (Graduate Studies Math. 58). American Mathematical Society, Providence, RI.
- [51] VILLANI, C. (2008). *Optimal Transport: Old and New*. Springer, New York.
- [52] WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York.
- [53] WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge University Press.
- [54] WILSON, A. G. AND IZMAILOV, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, Vol. **33**, eds H. Larochelle *et al.*, pp. 4697–4708.
- [55] ZEMEL, Y. AND PANARETOS, V. M. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli* **25**, 932–976.