A novel notion of barycenter for probability distributions based on optimal weak mass transport

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce weak barycenters of a family of probability distributions, based on the recently developed notion of optimal weak transport of mass [24], [8]. We provide a theoretical analysis of this object and discuss its interpretation in the light of convex ordering between probability measures. In particular, we show that, rather than averaging in a geometric way the input distributions, as the Wasserstein barycenter based on classic optimal transport does, weak barycenters extract common geometric information shared by all the input distributions, encoded as a latent random variable that underlies all of them. We also provide an iterative algorithm to compute a weak barycenter for a finite family of input distributions, and a stochastic algorithm that computes them for arbitrary populations of laws. The latter approach is particularly well suited for the *streaming setting*, i.e., when distributions are observed sequentially. The notion of weak barycenter and our computation approaches are illustrated on synthetic examples, validated on 2D real-world data and compared to standard Wasserstein barycenters.

1 Introduction

Optimal transport (OT) [38] has had a tremendous impact in the machine learning (ML) community recently, as it provides meaningful and implementable distances between probability distributions [32], thus advancing many aspects in the field, see e.g. [6, 40, 31]. The space of probability measures on $\mathbb{R}^d$ with finite 2-nd moment can be *metrised* with the Wasserstein-2 distance, the computation of which amounts to finding a transport plan that minimises the quadratic average cost of transporting mass from a source probability measure onto a target one. In this context, a natural method for averaging a finite family of probability measures is to compute their Fréchet mean, with respect to the Wasserstein-2 distance, which corresponds to the *Wasserstein barycenter* introduced in [1].

The goal of the present work is to explore theoretical features and potential applications to ML of barycenters of probability measures analogously defined in terms of *optimal weak transport* (OWT, see [24]) or more precisely quadratic barycentric transport costs. In a nutshell, for a source measure μ and a target measure ν, the OWT problem aims to transport mass so that the conditional spatial mean of target support points y, given their source support points x, is close to x in average. This amounts to finding an intermediate measure η, possibly *more concentrated* than ν, in the sense of convex ordering of probability measures, which is *close* to μ with respect to the Wasserstein-2 distance.

The main motivation of our work is to investigate the effect and meaning of combining a family of probability measures using OWT instead of OT. To that end, we will define the *weak barycenter* of this family through an optimisation problem, and discuss some of its properties. Importantly, we will see that, rather than averaging the input distributions in a metric sense, solving a weak barycenter problem corresponds to finding probability measures that encode geometric or shape information *shared across* all of them. In fact, the weak barycenter problem will be interpreted as finding a

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

latent random variable, common to all the input distributions. Implications of this latent variable interpretation, in terms of robustness to outliers, will also be drawn.

A second motivation for our work is to develop and implement computational methods for weak barycenters, capitalising on the fact that the optimal weak coupling between *any* pair of distributions, with finite second moments, is always realised by a unique optimal *map*. This property is in sharp contrast to standard OT, where the absolute continuity with respect to the Lebesgue measure of the source or target measure is typically needed to grant the existence and uniqueness of a map—the so-called Monge map— realising the optimal coupling between them. This map is often required in different ways to compute Wasserstein barycenters (see [4], [41] or [33]).

Similarly to the Wasserstein barycenter problem, we will develop a fixed-point formulation of the weak barycenter problem, based on OWT plans. This allows us to, following [4, 41], construct an iterative procedure to compute a weak barycenter for a finite family of distributions and analyse its convergence properties. We will also define and study the so-called *weak population barycenters*, that deal with a population of probability measures distributed according to a given law $\mathbb{Q}$ supported on the Wasserstein-2 space, as in [28] for the OT case. Extending ideas from [33], we will then propose an iterative stochastic algorithm for online computation of the weak population barycenter, from a *stream* of probability measures sampled from $\mathbb{Q}$. We will then provide numerical simulations using this proposed method, in order to illustrate the geometric meaning of the weak barycenter, and we will compare it with related objects obtained with standard OT or its entropy regularised counterpart.

Organisation of the paper. Sec. 2 documents the background on OWT and the assumptions underlying our work. Sec. 3 analyses the weak barycenter problem, and provides an interpretation in the light of the convex ordering of probability measures and a latent variable model. Sec. 4 extends the previous concepts from a finite family of measures to an infinite population of distributions. Sec. 5 introduces two algorithms, specifically adapted to the finite and the population frameworks. Sec. 6 and 7 present the computational setting and the experimental validation of our proposal respectively. Lastly, Sec. 8 discusses our findings and future research questions, and the Appendix contains all the proofs of technical results as well as additional details. A toolbox for implementing the algorithms and the OWT problem is also provided.

2 Background: optimal weak transport

The optimal transport (OT) problem [39] aims to find the lowest cost to transfer the mass from one probability measure onto another. Therefore, OT is a natural way to compare two probability distributions while incorporating geometric information on them and on their supports. In particular, the Wasserstein-p distance W_p, associated with the Euclidean cost in $\mathbb{R}^d$, metrises the space $\mathcal{P}_p(\mathbb{R}^d)$ of probability measures on $\mathbb{R}^d$ with finite p-moment. Precisely, for $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$,

$$W_p(\mu,\nu) = \left(\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p \mathrm{d}\pi(x,y) \right)^{1/p}, \tag{1}$$

where π is a *transport plan* between μ and ν, that is, an element of the set $\Pi(\mu,\nu)$ of probability measures on the product space $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. For $p = 2$ and μ absolutely continuous (*a.c.*), the unique optimal plan is concentrated on the graph of a measurable map called *M*onge map such that $\nu = T\#\mu$, see (14) in Appendix A.1.

We consider here the optimal weak transport (OWT) problem introduced in [24] and in particular the special case of barycentric transport costs. The OWT problem is then defined for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$V(\mu|\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d} \|x - \int_{\mathbb{R}^d} y \mathrm{d}\pi_x(y)\|^2 \mathrm{d}\mu(x), \tag{2}$$

where π_x is the *disintegration* of the transport plan π with respect to the first marginal μ, *i.e.* $\pi(\mathrm{d}x\mathrm{d}y) = \pi_x(\mathrm{d}y)\mu(\mathrm{d}x)$. As our work strongly leans on OWT theory, we recall in Appendix A.2, Th. 6, that V is continuous with respect to the Wasserstein metric [8]. Additionally, the two following results from [7] (stated for our specific setting) lay the ground for our proposed weak barycenters.

Theorem 1 ([7], Theorem 1.2)**.** *The problem (2) admits a unique minimiser.*

This first result strongly differs from the classical OT setting, for which the uniqueness of an optimal transport plan is not guaranteed for arbitrary measures. The optimisation problem in eq. (2) can

also be reformulated thanks to the Brenier-Strassen theorem [23], [7], through the notion of convex ordering. We denote by $\eta \leq_c \nu$ the *convex order of measures*, meaning that $\int \phi \mathrm{d}\eta \leq \int \phi \mathrm{d}\nu$ for any convex function ϕ that is nonnegative or integrable with respect to $\eta + \nu$. By Strassen's theorem [37], two distributions are in convex order if and only if there exists a martingale coupling between them. The following result is a generalisation of the result originally proved in [23], Th. 1.2.

Theorem 2 ([7], Theorem 1.4)**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_1(\mathbb{R}^d)$. There exists a unique $\eta^* \leq_c \nu$ such that*

$$W_2^2(\mu, \eta^*) = \inf_{\eta \leq_c \nu} W_2^2(\mu, \eta) = V(\mu|\nu). \tag{3}$$

Moreover, there exists a convex function $\psi : \mathbb{R}^d \to \mathbb{R}$ of class C^1 with $\nabla\psi$ being 1-Lipschitz, such that $\nabla\psi \# \mu = \eta^*$. Finally, the optimal coupling $\pi^{\mu,\nu} \in \Pi(\mu, \nu)$ verifies $\int y d\pi_x^{\mu,\nu}(y) = \nabla\psi(x)$ μ-a.s.

The measurable map, or barycentric projection, $S_\mu^\nu(x) := \int_{\mathbb{R}^d} y \mathrm{d}\pi_x^{\mu,\nu}(y)$ associated to the plan $\pi^{\mu,\nu}$ achieving the minimum in eq. (2) is consequently uniquely defined and will be called *optimal barycentric projection*. From this notation, we can write the OWT cost in terms of an OT cost according to $V(\mu|\nu) = W_2^2(\mu, S_\mu^\nu \# \mu)$. We emphasise that S_μ^ν is directly related to the optimisation problem (2), whereas applied works such as [35, 34] make use of a barycentric projection constructed from a transport plan solving an OT problem between μ and ν (often regularised) as a substitute for the Monge map, which may not exist (more details on S_μ^ν are displayed in Appendix A.3).

Last, let us note that OWT is somehow also related to the martingale OT problem developed in the stochastic finance community [11, 2, 25], which puts the focus on the optimal transfer of mass between distributions assumed to be in convex order themselves.

3 Optimal weak transport barycenters and latent variable interpretation

3.1 Definition and basic properties

The classical Wasserstein barycenter problem for a set of probability measures $\nu_1, \ldots, \nu_n \in \mathcal{P}_2(\mathbb{R}^d)$ with weights $\lambda_1, \ldots, \lambda_n$ in the simplex (i.e. $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$) is defined [1] by

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \sum_{i=1}^n \lambda_i W_2^2(\mu, \nu_i). \tag{4}$$

In a similar fashion, based on the weak transport cost in eq. (2), we propose the following variant:

Definition 1. *The set of weak barycenters of a finite family of measures $\{\nu_i\}_{i=1,\ldots,n} \in \mathcal{P}_2(\mathbb{R}^d)$ with weights $\{\lambda_i\}_{i=1,\ldots,n}$ in the simplex is defined as*

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min} \sum_{i=1}^n \lambda_i V(\mu|\nu_i). \tag{5}$$

Thus, a weak barycenter averages, with respect to the Wasserstein metric, an optimally chosen set of probability measures $\{\eta_1, \ldots, \eta_n\}$ which are *more concentrated* than the corresponding ν_i, in the sense that $\eta_i \leq_c \nu_i$ for each $1 \leq i \leq n$. The existence of a solution is established in the following:

Proposition 1. *The weak barycenter problem in eq. (5) admits a minimiser $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.*

See Sec. B of the Appendix for the proof (which relies on Prokhorov's theorem), as well as for proofs of this Section. Uniqueness is in general not granted: the next result shows that the set of solutions is indeed an interval, with respect to the partial order of convex ordering of probability measures.

In the following, we denote by X and Y_i random variables with respective laws μ and ν_i, for $1 \leq i \leq n$, and δ_a the Dirac measure supported on $a \in \mathbb{R}^d$.

Lemma 1. *If μ is a weak barycenter of $\{\nu_i\}_{i=1\ldots,n}$ and $\mu' \leq_c \mu$, then μ' also is a weak barycenter. In particular, the Dirac measure supported on $\mathbb{E}_\mu(X)$ is always a weak barycenter. Moreover, a Dirac distribution $\delta_{\bar{\omega}}$ is a weak barycenter if and only if $\bar{\omega} = \sum_{i=1}^n \lambda_i \mathbb{E}_{\nu_i} Y_i$.*

A consequence of the above lemma is that for any weak barycenter μ,

$$\mathbb{E}_\mu(X) = \sum_{i=1}^n \lambda_i \mathbb{E}_{\nu_i}(Y_i)\,, \tag{6}$$

and the value of the weak barycenter problem is given by

$$\inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i V(\mu|\nu_i) = \sum_{i=1}^n \lambda_i \|\mathbb{E}(Y_i)\|^2 - \|\sum_{i=1}^n \lambda_i \mathbb{E}(Y_i)\|^2. \tag{7}$$

We can also derive the following characterisation on the set of weak barycenters:

Proposition 2. *A measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is a weak barycenter of $\{\nu_i\}_{i=1\dots,n}$ if and only if its mean satisfies (6) and $\hat{\mu} \leq_c \hat{\nu}_i$ holds for all $1 \leq i \leq n$, where $\hat{\nu}$ denotes the centered version of a law ν.*

For instance, in the case of one dimensional Gaussian distributions $\nu_i = \mathcal{N}(m, \sigma_i^2)$, the set of weak barycenters includes $\{\mu = \mathcal{N}(m, \sigma^2) \mid 0 \leq \sigma^2 \leq \min_{1\leq i\leq n} \sigma_i^2\}$.

A natural question is whether a "maximal" weak barycenter exists, in the sense of convex ordering (up to translation by the mean). For $d = 1$, the answer is affirmative. When the means $\mathbb{E}(Y_i)$ are equal, this follows from the complete lattice property of the set of probability measures with respect to the convex ordering (see [27]); the general case can then be reduced to the latter using Proposition 2. For $d \geq 2$, this property is in general not true and the answer depends on the family $\{\nu_i\}_{i=1\dots,n}$.

In the particular case of *a.c.* measures $(\nu_i)_{1\leq i\leq n}$, we can obtain a bound on the distance between the Wasserstein and weak barycenters.

Lemma 2. *Let $\nu_1, \dots, \nu_n \in \mathcal{P}_2(\mathbb{R}^d)$ be a.c., at least one of them with bounded density. Let $\bar{\mu}$ and $\tilde{\mu}$ respectively denote the weak and the Wasserstein barycenters. Then*

$$W_2^2(\bar{\mu}, \tilde{\mu}) \leq 2 \sum_{i=1}^n \lambda_i \left(\mathbb{E}\|Y_i\|^2 - \|\mathbb{E}Y_i\|^2\right).$$

Consequently, the variances of the distributions $(\nu_i)_{1\leq i\leq n}$ control the distance between the Wasserstein and weak barycenters, which are closer the more concentrated each ν_i is.

3.2 Weak barycenters as latent variables

We now show that the weak barycenter problem can be reinterpreted as that of finding a latent variable underlying the realisations of random variables of laws ν_i for all $1 \leq i \leq n$:

Theorem 3. *Let μ be a weak barycenter of $\{\nu_i\}_{i=1\dots,n}$. Then, for each $1 \leq i \leq n$, a random variable $Y_i \sim \nu_i$ can be realised as*

$$Y_i = X + (\mathbb{E}Y_i - \mathbb{E}X) + \bar{Y}_i$$

where $X \sim \mu$ and $\bar{Y}_i = Y_i - \mathbb{E}(Y_i|X)$ is centered conditionally on X. Moreover, one has $S_\mu^\nu(X) = X + (\mathbb{E}Y_i - \mathbb{E}X)$ for all $i = 1, \dots, n$. Finally, we have $\mathbb{E}(Y_i - \mathbb{E}Y_i|X - \mathbb{E}X) = X - \mathbb{E}X$ or, equivalently, $\hat{\mu} \leq_c \hat{\nu}_i$, with $\hat{\mu}$ and $\hat{\nu}_i$ the laws of $X - \mathbb{E}X$ and $Y_i - \mathbb{E}Y_i$ respectively.

That is to say, each $Y_i \sim \nu_i$ can be realised by sampling a random variable X common to all $i = 1, \dots, n$ and distributed according to the weak barycenter μ, translating that value by $\mathbb{E}Y_i - \mathbb{E}X$ and adding at last a cluster specific component $\bar{Y}_i$ or idiosyncratic noise, centered conditionally on X.

4 Extension for the population barycenter

The population Wasserstein barycenter introduced in [28] and [3] extends the definition of Wasserstein barycenter for an infinite number of measures. This formulation is particularly relevant for the construction of an iterative algorithm to compute the barycenter for the streaming case, that is, when the measures are received *online*. The proofs are reported in Section C of the Appendix.

Let us consider a probability measure $\mathbb{Q} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$, meaning that $\mathbb{Q}$ is supported on a set of measures with finite moments of order 2, such that for some (and thus all) $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we have that $\int_{\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu, \nu) \mathrm{d}\mathbb{Q}(\nu) < \infty$.

Definition 2. *We define the set of weak population barycenters of a distribution $\mathbb{Q} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ as*

$$\arg\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathcal{P}_2(\mathbb{R}^d)} V(\mu|\nu) d\mathbb{Q}(\nu). \tag{8}$$

The following lemma guarantees that the map $(x, \nu) \mapsto S_\mu^\nu(x)$ appearing in eq. (8) through $V(\mu|\nu) = \int \|x - S_\mu^\nu(x)\|^2 \mathrm{d}\mu(x)$ is well defined.

Lemma 3. *The function $(\mu, \nu) \in (\mathcal{P}_2(\mathbb{R}^d))^2 \mapsto \pi^{\mu,\nu} \in \mathcal{P}_2(\mathbb{R}^{2d})$ mapping (μ, ν) to the unique optimal plan $\pi^{\mu,\nu}$ realising $V(\mu|\nu)$ in eq. (2) is continuous. As a consequence, for each $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ the function $(x, \nu) \in \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \mapsto S_\mu^\nu(x)$ is measurable.*

Using similar arguments as those of Proposition 1 and the fact that any probability measure can be approximated by a sequence of probability measures with finite support, the following proposition confirms that the weak population barycenter problem is also well defined.

Proposition 3. *The minimisation problem in eq. (8) admits a solution.*

5 Algorithms via fixed-point representations

5.1 Weak barycenter

For the Wasserstein barycenter problem in eq. (4), the authors in [1] proved that if at least one of the measures $\nu_1, \dots, \nu_n$ is *a.c.*, the Wasserstein barycenter is unique. Furthermore, if all the ν_i's are *a.c.*, and at least one of them has a bounded density, then the unique Wasserstein barycenter is also *a.c.* and verifies a fixed-point equation. This last property has been thoroughly studied by [4] and [41] and leveraged to compute an approximation of the barycenter via an iterative algorithm based on Monge maps, whose existence and uniqueness are guaranteed by the *a.c.* of the measures involved.

Akin to the fixed-point methodology in the classical Wasserstein scenario, we define an iterative procedure based on the barycentric projection computed in the optimal weak transport problem in eq. (2), that is valid for arbitrary distributions. Therefore, we consider the following iterative rule for probability measures $\nu_1, \dots, \nu_n \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\mu_{k+1} = G(\mu_k), \text{ with } G(\mu) = \left(\sum_{i=1}^n \lambda_i S_\mu^{\nu_i} \right) \#\mu, \tag{9}$$

where for each $i = 1, \dots, n$, the optimal barycentric projection is given by $S_\mu^{\nu_i} : x \mapsto \int y \mathrm{d}\pi_x^{\mu,\nu_i}(y)$, for $\pi^{\mu,\nu_i} \in \Pi(\mu, \nu_i)$ achieving the minimum in the OWT problem in eq. (2). The proposed iterative procedure is presented in Algorithm 1 in Appendix E.1.

A fundamental difference between the fixed-point computation of the Wasserstein barycenter [4] and a weak barycenter is that the optimal Monge map T_μ^ν in the OT problem verifies $T_\mu^\nu \#\mu = \nu$, whereas the pushforward measure $S_\mu^\nu \#\mu$ in the OWT setting still depends on μ. We will then prove that the iterative algorithm in eq. (9), based on the maps $S_\mu^{\nu_i}$, admits converging subsequences. A convenient result is the continuity of the functional G in eq. (9), which can be proven using Arzela-Ascoli theorem on a set of barycentric projections as well as the Skorohod's representation theorem.

Theorem 4. *The function $\mu \mapsto G(\mu)$ defined in eq. (9) is W_2-continuous from $\mathcal{P}_2(\mathbb{R}^d)$ to $\mathcal{P}_2(\mathbb{R}^d)$.*

Using an approach similar to [4] for the Wasserstein barycenter, we can state the following results for the proposed fixed-point procedure.

Proposition 4. *If μ is a weak-barycenter, i.e. a solution of problem (5), then $G(\mu) = \mu$ i.e. $x = \sum_{i=1}^n \lambda_i S_\mu^{\nu_i}(x)$, $\mu(x)$-a.s.*

The inverse implication of Proposition 4 is not necessarily true, that is, some fixed points may not be weak barycenters. However, a Dirac delta $\delta_\omega, \omega \in \mathbb{R}^d$, that meets the fixed-point condition $\delta_\omega = G(\delta_\omega)$, is a weak barycenter (see Lemma 1).

Proposition 5. *Let $(\mu_k)_k$ be the sequence defined by the iterative procedure $\mu_{k+1} = G(\mu_k)$ and starting from $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then $(\mu_k)_k$ is tight and every converging subsequence must converge to a fixed point of G.*

We observe that these results also hold for the classical Wasserstein barycenter of *a.c.* measures $\{\nu_i\}_{i=1\dots,n}$ such that at least one of them has a bounded density. Moreover, the inverse implication, namely if μ is a fixed-point then it is a barycenter, is not straightforward even in the Wasserstein barycenter case, for which one considers the fixed-point equation given by

$\mu = (\sum_{i=1}^n \lambda_i T_\mu^{\nu_i})\#\mu$, with $T_\mu^{\nu_i}$ the Monge map verifying $\nu_i = T_\mu^{\nu_i}\#\mu$. Indeed, [1] prove that if μ checks $x = \sum_{i=1}^n \lambda_i T_\mu^{\nu_i}(x)$ for every $x \in \mathbb{R}^d$, not only μ-almost everywhere, then μ is a Wasserstein barycenter. Also, [41, Theorem 2] provide additional conditions for this to be true by essentially invoking more smoothness on the distributions $\{\nu_i\}_{i=1\dots,n}$. Additionally, they only conjecture that under the same assumptions, the fixed-point is unique. Our method, however, includes arbitrary probability measures. Therefore, we do not expect to obtain similar results as in the Wasserstein barycenter case, for which smoothness is required.

5.2 Weak population barycenter

Based on [33], we construct a stochastic iterative algorithm for computing the weak population barycenter in eq. (8). We clarify that [33] is constrained to probability measures $\mathbb{Q}$ supported on distributions that are *a.c.*, whereas in our setting these distributions only need to belong to $\mathcal{P}_2(\mathbb{R}^d)$. Let us notice that our algorithms can be interpreted as geodesic gradient descent as in [33] and [15], however, OWT is not a metric and its potential geodesic structure is so far unknown. Therefore, the proposed algorithm only aim to mimic Riemmanian gradient descent. Our fixed-point result for the weak population barycenter problem is stated in the following Lemma:

Lemma 4. *If μ is a weak population barycenter of $\mathbb{Q}$, then $x = \int S_\mu^\nu(x) d\mathbb{Q}(\nu)$, $\mu(x)$-a.s.*

As in the finite case, the inverse implication is difficult to obtain. In particular, this has not been proven for the classical population Wasserstein barycenter in [33], where it boils down to prove the uniqueness of an absolutely continuous fixed point of $\mu \mapsto (\int T_\mu^\nu \mathrm{d}\mathbb{Q}(\nu))\#\mu$, where T_μ^ν is the Monge map between μ and ν. As explained in [33], the uniqueness of such fixed points has also been studied under some strong assumptions in [13] by considering parametric classes of random probability measures with compact support. Note that this result is expected to be true by again invoking more smoothness on the distributions at hand. As our method focuses (in particular) on discrete probability measures, the conditions under which the inverse equality holds are beyond the scope of our work. However, from the experimental results in Section 7, we believe our method presents practical advantages.

We next develop an iterative procedure that converge towards a distribution μ verifying the fixed-point equation $x = \int S_\mu^\nu(x)\mathrm{d}\mathbb{Q}(\nu)$, $\mu(x)$-a.s. Our method is presented in the following definition and illustrated in Algorithm 2 in Appendix E.1.

Definition 3. *Let $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, $\nu^k \overset{i.i.d.}{\sim} \mathbb{Q}$ and $\gamma_k > 0$. We define the following iterative procedure for $k \geq 0$:*

$$\mu_{k+1} = \left[(1-\gamma_k)\mathrm{id} + \gamma_k S_{\mu_k}^{\nu^k}\right]\#\mu_k, \tag{10}$$

where $S_{\mu_k}^{\nu^k}$ is the optimal barycentric projection between μ_k and ν^k and id *is the identity operator.*

Assuming the following conditions on the steps γ_k:

$$\sum_{k=1}^\infty \gamma_k^2 < \infty \qquad \text{and} \qquad \sum_{k=1}^\infty \gamma_k = \infty, \tag{11}$$

we are able to prove the convergence of the iterative scheme in a similar fashion as Theorem 4.7 in [33]. The proof is to a large extent very similar to that of the classical Wasserstein.

Theorem 5. *Under the conditions in eq. (11), the sequence $(\mu_k)_k$ defined in eq. (10) is a.s. relatively compact in W_q for all $q < 2$ (in particular it is tight). Moreover, a limit point μ verifies $x = \int S_\mu^\nu(x) d\mathbb{Q}(\nu)$, $\mu(x)$-a.s.*

Note that we prove in Prop. 7 in Appendix that the function $\mu \mapsto \| \int S_\mu^\nu \mathrm{d}\mathbb{Q} - \mathrm{id}\|^2_{\mathbb{L}^2(\mu)}$ is continuous w.r.t. W_2. This result, however, was obtained under additional constraints for the OT case in [33].

6 Computational aspects

6.1 Related methods for computing Wasserstein barycenters

Recall that a barycentric map can be constructed from a transport plan solving an OT problem (possibly regularised) as in [17], where the support of the barycenter updated by a barycentric projection

corresponds to a gradient step (for a non-convex function). These barycentric maps, however, do not provide a solution of the OWT problem. [36] focuses on the computation of Wasserstein barycenters for a fixed number of measures and a stream of observations per measure; additionally, [30] proposed an entropy-regularised alternative via stochastic optimisation for computing the Wasserstein barycenter of *a.c.* distributions only from observations. Constrained by their assumption of *a.c.*, [41] compute the Wasserstein barycenter by smoothing the observed empirical distributions. Finally, [19] compares the complexity of both the *sample* Wasserstein barycenter and a stochastic approximation to estimate a population barycenter (discrete measures and entropic regularisation).

Our method for barycentric computation differs from previous ones in the literature in that it i) can process a *stream* of an unknown number of measures, ii) does not require the measures to be a.c., and iii) does not appeal to additional regularisation of the measures or the Wasserstein metric.

6.2 Setting and computation of OWTs

Both Algorithms 1 and 2 require the computation of the optimal barycentric projection associated to the OWT problem in eq. (2). For two discrete measures $\mu = \sum_{i=1}^{r} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, this boils down to solving

$$\min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{r} a_i \left\| x_i - \sum_{j=1}^{m} \frac{y_j \pi_{ij}}{a_i} \right\|^2 = \min_{\pi \in \mathbb{R}^{r \times m}} \sum_{i=1}^{r} a_i \| x_i - \left(\frac{\pi \mathbf{y}}{\mathbf{a}} \right)_i \|^2 + 1_{\Pi(\mu,\nu)}(\pi), \tag{12}$$

where 1_C is the indicator function of the set C i.e. $1_C(\pi) = \pi$ if $\pi \in C$ and $1_C(\pi) = \infty$ otherwise. We propose to solve the OWT problem in eq. (12) with a proximal algorithm. The optimal barycentric projection is then constructed as $\frac{\pi \mathbf{y}}{\mathbf{a}}$. The details and examples are presented in Appendix E.2.

6.3 Comparison setting

In the next section, we will compare our proposed computation for weak barycenters in Definition 2 (Algorithm 2) to the classic Wasserstein barycenter for a stream of a.c. measures. Namely, we will run Algorithm 1 and 2 by, following [17, 34], replacing optimal barycentric projections by the barycentric projections associated either to i) an optimal plan in the Kantorovich problem referred to as *OT barycenter*, or ii) the optimal Sinkhorn plan in the entropy regularised OT [16], referred to as *OT Sinkhorn barycenter* given by the KL divergence as

$$\arg\min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \mathrm{d}\pi(x, y) + \varepsilon KL(\pi | \mu \otimes \nu). \tag{13}$$

These optimal plans for OT and regularised OT problem were computed using *POT toolbox* [21]. Notice that what we call OT barycenter (resp. OT Sinkhorn barycenter) is not solving a Wasserstein barycenter problem (resp. a regularised Wasserstein barycenter problem).

7 Experimental results

This section is devoted to the empirical validation of our proposal on both synthetic and real-world data. We first focused in Algorithm 2 since multiple algorithm to compute a Wasserstein barycenter for a fixed number of distributions are already available [17, 36]. We present a robustness to outliers experiment, then we validate our OWT barycenter on synthetic dataset and then on real-world one. Additional experiments are presented in Appendix E.3.

7.1 Robustness to outliers

OT's sensitivity to outliers is a well-known problem that can be addressed *e*.g. with unbalanced OT [9]. We observed that OWT also allows to deal with outliers, which is coherent with the latent variable interpretation of Section 3.2. We illustrate this by showing 50 sets of $20 - 30$ observations from different 2D Gaussian measures (Figure 1(left)), where each observation may be corrupted by random translations (Bernoulli $p = 0.05$) thus producing outliers. Figure 1(right) shows the resulting barycenters (dots), and barycenters without outliers (crosses) for Wasserstein barycenter (red) and weak barycenter (black), which shows robustness to outliers.

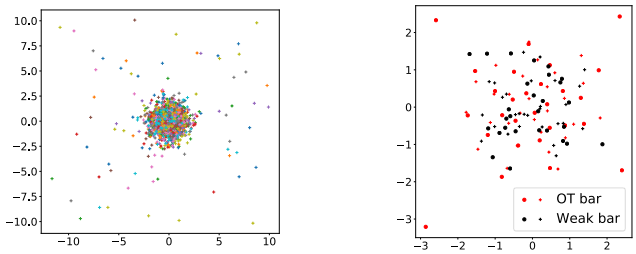

Figure 1: (left) Empirical Gaussian distributions. (right) Their OWT (black) and OT (red) barycenters for Gaussian observations (crosses) and corrupted observations (dots).

7.2 Synthetic distributions

We implemented the proposed sequential computation of weak barycenters (Algorithm 2) on two examples of synthetic distributions: Gaussians and spirals. In each case, we sampled r observations from a random distribution at each step, and considered K steps (and thus K measures for each case).

2D Gaussians ($r = 100$ & $K = 15$). We considered distributions $\mathcal{N}(m, I)$, with m uniformly distributed on $(-3, 3) \times (-5, -5)$ and I the identity matrix. Fig. 2 (top left) shows the empirical distributions together with the OWT and OT barycenters, the weak barycenter being the most concentrated as expected. The three remaining plots illustrate the behaviour of the barycenters constructed as stated in Sec. 6.3. For a small regularisation parameter ε in eq. (13), the OT and OT Sinkhorn barycenters are similar, however, as ε increases the OT Sinkhorn (OTS) barycenter becomes closer to the weak barycenter and thus even more concentrated. Critically, for a very large ε, as the entropy tends to spread the mass in the regularised optimal plan, the associated barycentric projection will roughly moves the mass at the spatial mean of the target distribution's support.

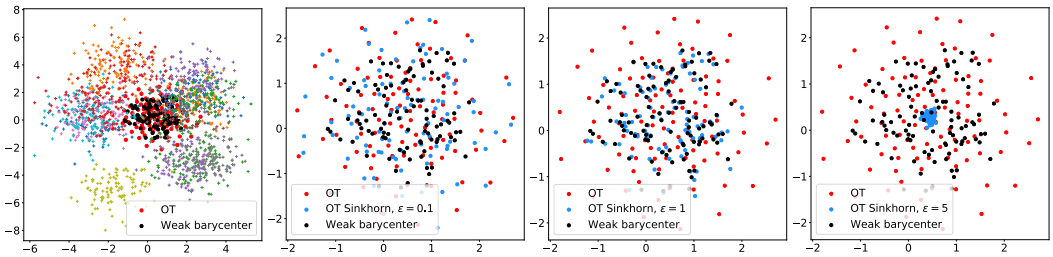

Figure 2: (left) Empirical Gaussian distributions and their OWT (black) and OT (red) barycenters computed with algorithm (10). Illustration of the weak (black), OT (red) and OT Sinkhorn (blue) barycenters for different values of $\varepsilon = 0.1, 1.5$.

Spiral distributions. ($p \in (200, 225)$ & $K = 10$). In this experiment, we considered distributions supported on a spiral—see Figure 3(left), with random ratio in $(0, 3)$. The OT and OWT barycenters are presented in Fig. 3 (right). Again, the weak barycenter seems to better preserve the shape of the spiral than the OT barycenter.

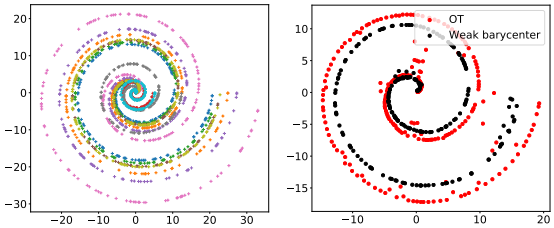

Figure 3: (left) Distributions supported on spiral. (right) OWT (black) and OT (red) barycenters computed with algorithm (10).

7.3 Real-world dataset: MNIST dataset

We considered the well-known MNIST dataset [29] of grayscale images of handwritten digits. The images, of size 28×28 pixels, can be normalised and thus be interpreted as discrete probability

measures supported on a two-dimensional grid of size 28×28. We computed the barycenters with 30 steps of Algorithm 1 between two digits "8", that are noisy versions of the same number with the aim to produce a more stable barycenter. To produce noisy data, we randomly (Bernoulli $p = 0.1$) move pixels of the prototype digit displayed in Fig. 4 (left). Fig. 5 shows the barycenters using the OWT, OT, and entropic-OT (for $\varepsilon = 1$). This example illustrates how OWT reduces dispersion, so that weak barycenter provides the best uniformly spread results among the barycenters considered, with the two loops of the "8" well shaped.

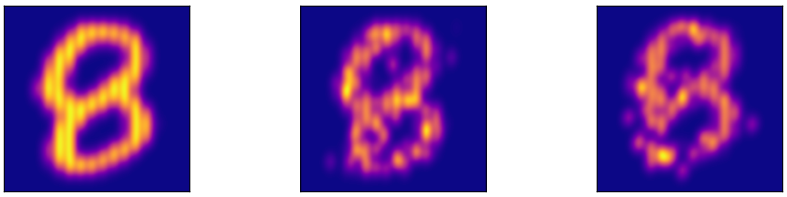

Figure 4: Digit "8" from MNIST dataset. (left) Prototype "8". (middle & right) Noisy versions of the prototype by randomly (Bernoulli $p = 0.1$) moving pixels.

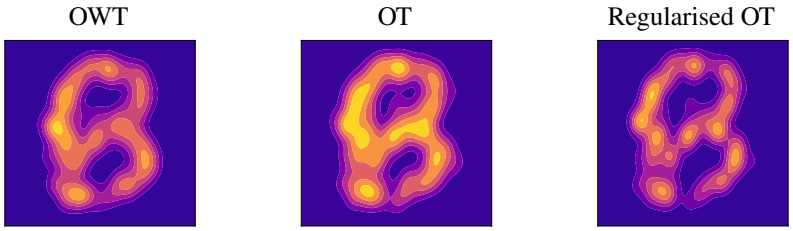

Figure 5: Digit "8" from MNIST dataset. Comparison of three barycenters constructed with Algorithm 1: OWT plan (left), OT plan (middle) and entropy regularised OT plan for $\varepsilon = 1$ (right).

8 Discussion

We have introduced the weak barycenter problem, that extracts common geometric information of probability measures on $\mathbb{R}^d$ based on optimal weak transport and we additionally showed that it can be interpreted as a latent variable model. From the fixed-point formulation defined in terms of optimal weak transport maps, irrespective of the regularity assumptions on the measures involved, we developed practical computation via an iterative algorithm with guaranteed convergence. In particular, the proposed algorithms do not require a common grid on the sample space, when processing either observed data or samples from distributions. We have also proposed weak barycenters of a possibly infinite population of measures sampled from a given distribution on the Wasserstein space. In this case, we developed a stochastic procedure for computing the weak barycenter for streaming data, where distributions are incorporated into the barycentric computation as they arrive. This has critical implications for continual-learning methods in the ML community.

Additional studies should be carried out to gain further insight on the latent variable model interpretation of weak barycenters, and to highlight how they complement the aggregate information represented by the Wasserstein barycenter. Moreover, we believe at least two theoretical aspects which have not been treated in detail here should be addressed in the future. The first one is to exhibit general conditions on the family of input measures (or on the law of the population) for the existence of weak barycenters that are not Dirac masses. The second one, is to provide conditions on those input measures for a “maximal” weak barycenter (in terms of convex ordering) to exist when $d \geq 2$, among all the solutions of the weak barycenter problem (and, if possible, a way of constructing it by regularisation most probably). The statistical behaviour of the weak barycenter can also be investigated, in particular when constructed from large empirical random samples of given distributions. Lastly, the weak population barycenter could also be used to construct a predictive posterior in the context of Bayesian learning, as was done for Wasserstein barycenters in [33].

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] A. Alfonsi, J. Corbetta, and B. Jourdain. Sampling of probability measures in the convex order and approximation of martingale optimal transport problems. *Available at SSRN 3072356*, 2017.

[3] P.C. Alvarez-Esteban, E Del Barrio, J.A. Cuesta-Albertos, and C. Matrán. Wide consensus for parallelized inference. *arXiv: 1511.05350*, 2015.

[4] P.C. Álvarez-Esteban, E. Del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[5] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[6] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. PMLR, 2017.

[7] J. Backhoff-Veraguas, M. Beiglböck, and G. Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):203, 2019.

[8] J. Backhoff-Veraguas, M. Beiglböck, and G. Pammer. Weak monotone rearrangement on the line. *Electronic Communications in Probability*, 25, 2020.

[9] Y. Balaji, R. Chellappa, and S. Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *arXiv:2010.05862*, 2020.

[10] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[11] M. Beiglböck, P. Henry-Labordère, and F. Penkner. Model-independent bounds for option prices—a mass transport approach. *Finance and Stochastics*, 17(3):477–501, 2013.

[12] J. Bigot, E. Cazelles, and N. Papadakis. Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8(4):719–755, 2019.

[13] J. Bigot and T. Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

[14] M.W. Botsko. An elementary proof of Lebesgue's differentiation theorem. *The American Mathematical Monthly*, 110(9):834–838, 2003.

[15] S. Chewi, T. Maunu, P. Rigollet, and A.J. Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. *arXiv: 2001.01700*, 2020.

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

[17] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. *Proceedings of the 31st International Conference on Machine Learning*, 32(2):685–693, 2014.

[18] A. Dessein, N. Papadakis, and J.-L. Rouas. Regularized optimal transport and the ROT mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

[19] D. Dvinskikh. Stochastic approximation versus sample average approximation for population Wasserstein barycenters. *arXiv e-prints*, pages arXiv–2001, 2020.

[20] E.A. Feinberg, P.O. Kasyanov, and Y. Liang. Fatou's lemma in its classical form and Lebesgue's convergence theorems for varying measures with applications to Markov decision processes. *Theory of Probability & Its Applications*, 65(2):270–291, 2020.

[21] R. Flamary, N. Courty, A. Gramfort, M.Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T.H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D.J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[22] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[23] N. Gozlan and N. Juillet. On a mixture of Brenier and Strassen theorems. *Proceedings of the London Mathematical Society*, 120(3):434–463, 2020.

[24] N. Gozlan, C. Roberto, P.-M. Samson, and P. Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11):3327–3405, 2017.

[25] J. Guyon and P. Henry-Labordere. *Nonlinear option pricing*. CRC Press, 2013.

[26] F. Hahne, A.H. Khodabakhshi, A. Bashashati, C.-J. Wong, R.D. Gascoyne, A.P. Weng, V. Seyfert-Margolis, K. Bourcier, A. Asare, T. Lumley, et al. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry*, 77(2):121–131, 2010.

[27] R.P. Kertz and U. Rösler. Complete lattices of probability measures with applications to martingale theory. *Lecture Notes-Monograph Series*, pages 153–177, 2000.

[28] T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168(3-4):901–917, 2017.

[29] Y. LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[30] L. Li, A. Genevay, M. Yurochkin, and J. Solomon. Continuous regularized wasserstein barycenters. *arXiv preprint arXiv:2008.12534*, 2020.

[31] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

[32] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[33] G. Rios, J. Backhoff-Veraguas, J. Fontbona, and F. Tobar. Bayesian learning with Wasserstein barycenters. *arXiv preprint arXiv:1805.10833v3*, 2018.

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

[35] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

[36] M. Staib, S. Claici, J. Solomon, and S. Jegelka. Parallel streaming wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 2647–2658, 2017.

[37] V. Strassen. The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, 36(2):423–439, 1965.

[38] C. Villani. *Topics in Optimal Transportation*. Number 58. American Mathematical Soc., 2003.

[39] C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

[40] J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.

[41] Y. Zemel and V. Panaretos. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5 and Discussion 8.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Each Theorem, Proposition and Lemma is carefully proved in Appendices B, C and D.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The toolbox for implementing the algorithms and the OWT problem and replicate the experiments of this paper is included in the Supplement and will be made publicly available after anonymity is lifted.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix Sections E.1 and E.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See [21, 18, 10, 34].
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] A toolbox to reproduce our proposed model is provided in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]