

## BAYESIAN LEARNING WITH WASSERSTEIN BARYCENTERS\*

JULIO BACKHOFF-VERAGUAS<sup>1</sup>, JOAQUIN FONTBONA<sup>2,\*\*</sup>, GONZALO RIOS<sup>3</sup>  
AND FELIPE TOBAR<sup>4</sup>

**Abstract.** We introduce and study a novel model-selection strategy for Bayesian learning, based on optimal transport, along with its associated predictive posterior law: the Wasserstein population barycenter of the posterior law over models. We first show how this estimator, termed Bayesian Wasserstein barycenter (BWB), arises naturally in a general, parameter-free Bayesian model-selection framework, when the considered Bayesian risk is the Wasserstein distance. Examples are given, illustrating how the BWB extends some classic parametric and non-parametric selection strategies. Furthermore, we also provide explicit conditions granting the existence and statistical consistency of the BWB, and discuss some of its general and specific properties, providing insights into its advantages compared to usual choices, such as the model average estimator. Finally, we illustrate how this estimator can be computed using the stochastic gradient descent (SGD) algorithm in Wasserstein space introduced in a companion paper, and provide a numerical example for experimental validation of the proposed method.

**Mathematics Subject Classification.** 62G05, 62F15, 62H10, 62G20.

Received March 3, 2022. Accepted November 6, 2022.

### 1. INTRODUCTION

Given a set  $\mathcal{M}$  of probability distributions on some data space  $\mathcal{X}$ , learning a model  $m \in \mathcal{M}$  from data points  $D = (x_1, \dots, x_n) \in \mathcal{X}^n$  consists in choosing, under a given criterion, an element of  $\mathcal{M}$  that *best* explains  $D$  as data sampled from it. The Bayesian paradigm provides a probabilistic approach to deal with model uncertainty in terms of a *prior distribution* on models  $\mathcal{M}$ , and also furnishes strategies to address the problem of model selection, based on the *posterior distribution* on  $\mathcal{M}$  given  $D$ . This type of estimators, usually called *predictive posterior laws*, include classical Bayesian estimators such as the maximum a posteriori estimator (MAP), the posterior mean, the Bayesian model average estimator (BMA) and generalizations thereof. Predictive posterior estimators typically result from selection criteria consisting in optimizing some loss function, averaged with

---

\*Partially funded by 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).

*Keywords and phrases:* Bayesian learning, non-parametric estimation, Wasserstein distance and barycenter, consistency, MCMC, stochastic gradient descent in Wasserstein space.

<sup>1</sup> Faculty of Mathematics, University of Vienna, Vienna, Austria.

<sup>2</sup> Center for Mathematical Modeling and Department of Mathematical Engineering, Universidad de Chile, Santiago, Chile.

<sup>3</sup> NoiseGrasp SpA, Santiago, Chile.

<sup>4</sup> Initiative for Data & Artificial Intelligence and Center for Mathematical Modeling, Universidad de Chile, Santiago, Chile.

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

respect to the posterior law over models, or Bayesian risk function. We refer the reader to [23, 37] and references therein for mathematical background on Bayesian statistics and their use in the machine learning community.

In Section 2, we will formulate the general problem of Bayesian model selection directly on the space of probability measures (or models) on the data space, and show how this abstract framework covers both classic finitely-parametrized settings and parameter-free model spaces, allowing us to retrieve classical selection criteria as particular cases. An eye-opening observation that will follow from adopting this viewpoint is that many classical predictive posteriors can be seen as instances of *Fréchet means* [22], or barycenters in the space of probability measures, with respect to specific metrics or divergences between them, that play the role of abstract loss functions defined on the model space.

Building upon this general framework, the main goals of this work are to introduce a novel Bayesian model-selection criterion by proposing a loss function on models coming from the theory of optimal transport, and to study some of the distinctive features of the predictive posterior law that results from it. More precisely, let us consider observations  $D = (x_1, \dots, x_n)$  in a metric space  $(\mathcal{X}, d)$  and a set of candidate models  $\mathcal{M}$  that generated these observations. Equipping the set  $\mathcal{M}$  with a prior distribution  $\Pi$ , and denoting the corresponding posterior distribution over  $\mathcal{M}$  by  $\Pi_n$ , we will define the *Bayesian Wasserstein barycenter estimator* (BWB) as a minimizer  $\hat{m}_p^n \in \mathcal{M}$  of the “risk function”

$$\mathcal{M} \ni m \mapsto \int_{\mathcal{P}(\mathcal{X})} W_p(m, \bar{m})^p \Pi_n(d\bar{m}), \quad (1.1)$$

where  $\mathcal{P}(\mathcal{X})$  is the set of probability measures on  $\mathcal{X}$ ,  $p \geq 1$  and  $W_p$  is the celebrated  $p$ -Wasserstein distance on probability measures on  $\mathcal{X}$  associated with  $d$ , see [45, 46].

Wasserstein barycenters were initially studied in [1] and, since then, the concept has been extensively explored from both theoretical and practical perspectives. We refer the reader to the overview [38] for statistical applications, to the works [12, 15, 16, 41] for applications in machine learning, and to [3, 4, 10, 33] for a presentation of recent developments and further references. As a cautionary tale, we mention that the problem of computing Wasserstein barycenters is known to be NP-hard, see [2]. To cope with this, fast methods aiming at learning and generating approximate Wasserstein barycenters on the basis of neural networks techniques, have also been proposed, see e.g. [31, 32].

In Section 3 we will recall Wasserstein distances and revisit Wasserstein barycenters together with their basic properties such as existence, uniqueness and absolute continuity. In Section 4, we will rigorously introduce the BWB estimator  $\hat{m}_p^n$ , which will correspond to the so-called *population Wasserstein barycenter* [33] for the posterior distribution on models  $\Pi_n$ , and we will state some of its main properties. Specifically, we will show that the BWB has less variance than the BMA and we will study its statistical consistency. In particular, we will address the question of “posterior consistency”, or asymptotic concentration of the posteriors  $\Pi_n$  around the Dirac mass on a model  $m_0$  as  $n \rightarrow \infty$ , whenever the data consists of i.i.d. observations following the law  $m_0$ , alongside the question of convergence of the BWB to the “true” distribution  $m_0$  of the data in that setting. We refer the reader to [17, 23], and references therein, for detailed accounts on posterior consistency, a highly desirable feature of a Bayesian estimation procedure, both from a semi-frequentist perspective as well as from the “merging of opinions” point of view on Bayesian statistics (cf. [23], Chap. 6). After reviewing central notions and tools from the framework of posterior consistency, namely the celebrated Schwartz’ theorem [44], Theorem 6.17 of [23] and the notion of Kullback-Leibler support of  $\Pi$ , we will provide in that section equivalent and (verifiable) sufficient conditions for both the posterior consistency in the Wasserstein topology and the a.s. convergence

$$\lim_{n \rightarrow \infty} W_p(\hat{m}_p^n, m_0) = 0 \text{ a.s.}$$

to hold, when  $\hat{m}_p^n$  is the BWB computed with  $n$  i.i.d. observations sampled from  $m_0$ .

Additionally, Sections 1–3 present a series of examples illustrating the main concepts of our work, their relationship to standard objects in Bayesian statistics and the applicability of our theoretical results.

Lastly, we will show how the BWB estimator can be calculated using a novel stochastic algorithm, introduced in the companion paper [7], to compute population Wasserstein barycenters in a general setting. This algorithm, presented in Section 5, can be seen as an abstract *stochastic gradient descent method in the Wasserstein space* and is advantageous compared to gradient or fixed-point algorithms developed in [3, 4, 48], whose application is restricted to barycenters of model spaces  $\mathcal{M}$  comprised of finitely-many elements only. Moreover, our algorithm has theoretical guarantees of convergence under suitable conditions, it can be easily implemented for some families of regular models for which optimal transport maps are explicit or easily computed (we recall one such family in Sect. 6.1), and its convergence rate can be studied and established in some cases, see [13] for the Gaussian setting. A comprehensive numerical experiment illustrating this method, and its natural “batch” variants, will be presented in Section 6 and compared to (more conventional) empirical barycenter estimators.

### Notation:

- We denote  $\mathcal{P}(\mathcal{X})$  the set of (Borel) probability measures on  $\mathcal{X}$  endowed with the weak topology, and  $\mathcal{P}_{ac}(\mathcal{X})$  the (measurable) subset of absolutely continuous probability measures, with respect to a common reference  $\sigma$ -finite measure  $\lambda$  on  $\mathcal{X}$ .
- As a convention, we shall use the same notation for an element  $m(dx) \in \mathcal{P}_{ac}(\mathcal{X})$  and its density  $m(x)$  with respect to  $\lambda$ .
- We denote by  $\text{supp}(\nu)$  the support of a measure  $\nu$  and by  $|\text{supp}(\nu)|$  its cardinality.
- Given  $\Gamma \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  and a measurable subset  $\mathcal{M} \subseteq \mathcal{P}(\mathcal{X})$ , we say that  $\mathcal{M}$  is a *model space for*  $\Gamma$  if  $\Gamma(\mathcal{M}) = 1$ .
- Last, given a measurable map  $T : \mathcal{Y} \rightarrow \mathcal{Z}$  and a measure  $\nu$  on  $\mathcal{Y}$  we denote by  $\nu_{\#}T$  the image measure (or push-forward), that is, the measure on  $\mathcal{Z}$  given by  $\nu_{\#}T(\cdot) = \nu(T^{-1}(\cdot))$ .

## 2. BAYESIAN LEARNING IN MODEL SPACE

We start by setting a general framework for Bayesian learning which covers both finitely-parametrized settings (including hierarchical models) and parameter-free models. We consider a probability measure  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  understood as a *prior* distribution on the model space  $\mathcal{M} \subseteq \mathcal{P}_{ac}(\mathcal{X})$ . In particular we have

$$\Pi(\mathcal{M}) = \Pi(\mathcal{P}_{ac}(\mathcal{X})) = 1.$$

For each  $n \in \mathbb{N} \setminus \{0\}$ ,  $\Pi$  canonically induces a law  $\mathbf{\Pi}$  on  $\mathcal{X}^n \times \mathcal{M}$ , representing the joint law of a random model  $m$  chosen according to  $\Pi$  and a sample

$$D = (x_1, \dots, x_n) \subseteq \mathcal{X}^n$$

of i.i.d. observations drawn from it. That is,

$$\mathbf{\Pi}(dx_1, \dots, dx_n, dm) := m(dx_1) \cdots m(dx_n) \Pi(dm) = m(x_1) \cdots m(x_n) \lambda(dx_1) \cdots \lambda(dx_n) \Pi(dm).$$

Note that, in the above equation and throughout,  $\Pi(dm)$  denotes integration over  $m \in \mathcal{M}$  w.r.t.  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ , whereas integration over  $x \in \mathcal{X}$  w.r.t.  $m \in \mathcal{P}(\mathcal{X})$  is denoted  $m(dx) = m(x)\lambda(dx)$ . The law on  $\mathcal{X}^n$  of the data  $D$ , conditionally on a model  $m$ , is thus given by

$$\mathbf{\Pi}(dx_1, \dots, dx_n | m) := m(x_1) \cdots m(x_n) \lambda(dx_1) \cdots \lambda(dx_n), \quad (2.1)$$

with density  $\mathbf{\Pi}(x_1, \dots, x_n | m) = m(x_1) \cdots m(x_n)$  with respect to  $\lambda^{\otimes n}$ , and the marginal density of  $D$  with respect to  $\lambda^{\otimes n}$  is  $\mathbf{\Pi}(x_1, \dots, x_n) := \int_{\mathcal{M}} \bar{m}(x_1) \cdots \bar{m}(x_n) \Pi(d\bar{m})$ .

The *posterior distribution*  $\Pi(dm|x_1, \dots, x_n)$  given the data  $D$  is also an element of  $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ , which we denote  $\Pi_n$  for simplicity and which, by virtue of the Bayes rule, is given in this setting by

$$\Pi_n(dm) := \frac{\Pi(x_1, \dots, x_n|m) \Pi(dm)}{\Pi(x_1, \dots, x_n)} = \frac{m(x_1) \cdots m(x_n) \Pi(dm)}{\int_{\mathcal{M}} \bar{m}(x_1) \cdots \bar{m}(x_n) \Pi(d\bar{m})}. \quad (2.2)$$

Notice that  $(x_1, \dots, x_n) \mapsto \Pi_n = \Pi(\cdot|x_1, \dots, x_n)$  defines  $\lambda^{\otimes n}$  a.e. a measurable function from  $\mathcal{X}^n$  to  $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ . The density  $\Lambda_n(m)$  of  $\Pi_n(dm)$  with respect the prior  $\Pi(dm)$  is called the *likelihood* function. The fact that  $\Pi_n \ll \Pi$  implies that a model space  $\mathcal{M}$  for  $\Pi$  is a model space for  $\Pi_n$  too.

We call *loss function* a non-negative functional on models  $L : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$ , interpreting  $L(m_0, \bar{m})$  as the cost of selecting model  $\bar{m} \in \mathcal{M}$  when the true model is  $m_0 \in \mathcal{M}$ . With a loss function and the posterior distribution over models  $\Pi_n$ , the Bayes risk (or expected loss)  $R_L(\bar{m}|D)$  and the corresponding Bayes estimator  $\hat{m}_L$  (or predictive posterior law) are respectively defined as follows:

$$R_L(\bar{m}|D) := \int_{\mathcal{M}} L(m, \bar{m}) \Pi_n(dm), \quad (2.3)$$

$$\hat{m}_L \in \operatorname{argmin}_{\bar{m} \in \mathcal{M}} R_L(\bar{m}|D). \quad (2.4)$$

See [8] for further background on Bayes risk and statistical decision theory. A key consequence of defining both  $L$  and  $\Pi_n$  directly on the model space  $\mathcal{M}$  (rather than on parameter space), is that learning according to equations (2.3)–(2.4) does not depend on the chosen parametrization or the geometry of the parameter space. Moreover, this point of view will allow us to define loss functions in terms of various metrics/divergences directly on the space  $\mathcal{P}(\mathcal{X})$ , and therefore to enhance the classical Bayesian estimation framework through the use of optimal transportation distances on that space. Before further developing these ideas, we discuss how this general framework includes model spaces which are finitely parametrized, and recall some standard choices in that setting. We will also discuss the advantages of formulating the problem of model selection directly on the model space, even when this space can be finitely parametrized.

## 2.1. Parametric setting

We say that  $\mathcal{M}$  is finitely parametrized if there is an integer  $k$ , a measurable set  $\Theta \subseteq \mathbb{R}^k$  termed parameter space, and a measurable function  $\mathcal{T} : \Theta \mapsto \mathcal{P}_{ac}(\mathcal{X})$ , called parametrization, such that  $\mathcal{M} = \mathcal{T}(\Theta)$ . In other words,  $m = \mathcal{T}(\theta)$  is the model corresponding to parameter  $\theta$ , which is classically denoted  $p(\cdot|\theta)$  or  $p_\theta(\cdot)$ . In general,  $\mathcal{T}$  is a one-to-one function (the model space is otherwise said to be *over-parametrized*).

In the standard parametric Bayesian framework, a prior distribution is a probability law  $p \in \mathcal{P}(\Theta)$  over  $\Theta$ , typically assumed to have a (equally denoted) density  $p(\theta)$  with respect to the Lebesgue measure. A function  $\ell : \Theta \times \Theta \rightarrow \mathbb{R}_+$  is called a loss function (on parameters), whereby  $\ell(\theta_0, \bar{\theta})$  is interpreted as the cost of choosing parameter  $\bar{\theta}$  when the true parameter is  $\theta_0$ . The parametric Bayes risk [8] of  $\bar{\theta} \in \Theta$  is then given by

$$R_\ell(\bar{\theta}|D) = \int_{\Theta} \ell(\theta, \bar{\theta}) p(\theta|x_1, \dots, x_n) d\theta, \quad (2.5)$$

where  $p(\theta|x_1, \dots, x_n)$  is the posterior density of  $\theta$  given observations  $x_1, \dots, x_n$ . The associated Bayes estimator is defined as  $\hat{\theta}_\ell \in \operatorname{argmin}_{\bar{\theta} \in \Theta} R_\ell(\bar{\theta}|D)$ . Hence, if the model space  $\mathcal{M}$  is finitely parametrized, learning a model boils down to finding the *best* model parameter  $\theta \in \Theta$  under a given criterion, quantified by the parametric risk function  $R_\ell(\cdot|D)$ . Among continuous-valued losses on the parameter space, the *de facto* choice is the quadratic one  $\ell_2(\theta, \bar{\theta}) = \|\theta - \bar{\theta}\|^2$ , whose associated Bayes estimator is the posterior mean  $\hat{\theta}_{\ell_2} = \int_{\Theta} \theta p(\theta|D) d\theta$ . For one-dimensional parameter spaces, the absolute loss  $\ell_1(\theta, \bar{\theta}) = |\theta - \bar{\theta}|$  yields the posterior median(s) estimator(s). The 0-1 loss formally given by  $\ell_{0-1}(\theta, \bar{\theta}) = 1 - \delta_{\bar{\theta}}(\theta)$ , with  $\delta_{\bar{\theta}}$  the Dirac mass at  $\bar{\theta}$ , yields the risk  $R_{\ell_{0-1}}(\bar{\theta}|D) = 1 - p(\bar{\theta}|D)$  and its corresponding Bayes estimator is the posterior mode or maximum a posteriori estimator (MAP),  $\hat{\theta}_{\ell_{0-1}} = \hat{\theta}_{MAP}$ .



The parametric case is embedded into the considered model-space setting as follows. The push-forward of  $p$  through  $\mathcal{T}$  defines a prior  $\Pi = p_{\#}\mathcal{T}$  over the model space  $\mathcal{M}$  in the sense discussed at the beginning of this section. If  $\mathcal{T}$  is one to one, a loss function  $\ell : \Theta \times \Theta \rightarrow \mathbb{R}_+$  induces a loss function  $L$  on the model space  $\mathcal{M} = \mathcal{T}(\Theta)$ , such that  $L(\mathcal{T}(\theta_0), \mathcal{T}(\bar{\theta})) = \ell(\theta_0, \bar{\theta})$ . More generally, any loss function  $L$  on  $\mathcal{M} \times \mathcal{M}$  induces a loss functional  $\ell$  on  $\Theta \times \Theta$  defined as  $\ell(\theta_0, \bar{\theta}) := L(\mathcal{T}(\theta_0), \mathcal{T}(\bar{\theta}))$ . Moreover, when the data  $x_1, \dots, x_n$  under the model parameterized by  $\theta$  consists of an i.i.d. sample from  $p(\cdot|\theta) = \mathcal{T}(\theta)$ , one can verify that  $\Pi_n$  given in equation (2.2) corresponds precisely to the the push-forward through  $\mathcal{T}$  of  $p(\theta|x_1, \dots, x_n)$  and that  $R_\ell(\bar{\theta}|D)$  in equation (2.5) is given by

$$R_\ell(\bar{\theta}|D) = \int_{\mathcal{M}} L(m, \bar{m}) \Pi_n(dm),$$

with  $\Pi_n(dm) = \Lambda_n(m) \Pi(dm)$  associated with the prior on model space  $\Pi = p_{\#}\mathcal{T}$ , and  $\bar{m} = \mathcal{T}(\bar{\theta})$ .

**Example 2.1.** Consider the parametric Bayesian model with parameter space  $\Theta$  and sample space  $\mathcal{X}$  both equal to  $\mathbb{R}^d$  and Gaussian parametrized models  $\mathcal{N}(\theta, \Sigma)$ , with  $\Sigma \in \mathbb{R}^{d \times d}$  a fixed covariance matrix, and  $\theta$  a random mean with prior  $\mathcal{N}(\mu_0, \Sigma_0)$ . The mean  $\mu_0 \in \mathbb{R}^d$  and covariance matrix  $\Sigma_0$  are fixed hyperparameters. This is classically denoted:

$$p(x|\theta) = \mathcal{N}(x; \theta, \Sigma), \quad p(\theta) = p(\theta|\mu_0, \Sigma_0) = \mathcal{N}(\theta; \mu_0, \Sigma_0).$$

From now on,  $\mathcal{N}(y; \nu, \mathcal{K})$  stands for the density of the Gaussian law  $\mathcal{N}(\nu, \mathcal{K})$  evaluated on the value  $y$ . Following from the introduced notation, the parametrization  $\mathcal{T} : \Theta \rightarrow \mathcal{P}(\mathcal{X})$  is thus given by  $\mathcal{T}(\theta) = \mathcal{N}(\theta, \Sigma)$ , the model space is  $\mathcal{M} = \{\mathcal{N}(\theta, \Sigma) : \theta \in \mathbb{R}^d\}$  and a measure  $m$  sampled from the prior  $\Pi(dm) = p_{\#}\mathcal{T}(dm)$  is a Gaussian distribution on  $\mathcal{X} = \mathbb{R}^d$ , with fixed covariance matrix  $\Sigma$  and random mean  $\theta$  distributed according to  $\mathcal{N}(\mu_0, \Sigma_0)$ . In this case, if both  $\Sigma$  and  $\Sigma_0$  are nonsingular, the posterior of  $\theta$  given  $D = (x_1, \dots, x_n)$  is

$$p(\theta|D) = p(\theta|D, \mu_0, \Sigma_0) = \mathcal{N}(\theta; (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}_n), (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}), \quad (2.6)$$

with  $\bar{x}_n$  the sample mean of the observations  $D$ . A model  $m$  sampled from the posterior  $\Pi_n(dm)$  given  $D$  is thus obtained by sampling  $\theta$  distributed according to  $p(\theta|D)$  in equation (2.6) and then setting  $m = \mathcal{N}(\theta, \Sigma)$ . The posterior mean estimator of the parameter  $\theta$  is therefore given by

$$\hat{\theta}_{\ell_2} = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}_n), \quad (2.7)$$

and the corresponding predictive posterior model is the Gaussian law  $\mathcal{N}(\hat{\theta}_{\ell_2}, \Sigma)$ . Observe that, in this case, one could equivalently obtain  $\hat{\theta}_{\ell_2}$  as the mean of the predictive posterior law  $\hat{m}_L$  associated with the loss function  $L$  on models, defined by  $L(m, \bar{m}) = \|\int_{\mathbb{R}^d} x m(dx) - \int_{\mathbb{R}^d} x \bar{m}(dx)\|^2$ . This illustrates the equivalence of (some) pairs of losses in the model and parameter spaces that result on the same Bayes estimators. Additionally, observe that Bayesian inference on the mean and the covariance matrix could be similarly formulated in terms of a loss functions on models too, take *e.g.*  $L'(m, \bar{m}) = L(m, \bar{m}) + \|\Sigma_m^{1/2} - \Sigma_{\bar{m}}^{1/2}\|_{\text{Fro}}^2$  with  $L(m, \bar{m})$  as before,  $\Sigma_m$  the covariance matrix of a r.v. with law  $m$  and  $\|\cdot\|_{\text{Fro}}$  the Frobenius norm on matrices.

The general model space framework applies equally to models with parameter and data spaces that might be of different nature:

**Example 2.2.** Assume  $\Theta = \mathbb{R}_+$  and  $\mathcal{X} = \mathbb{N}$ , with  $p(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$ ,  $x \in \mathcal{X}$ , and  $p(\lambda) \propto e^{-\beta\lambda} \lambda^{\alpha-1}$ ,  $\lambda \in \Theta$ . In this case,  $\mathcal{T} : \Theta \rightarrow \mathcal{P}(\mathcal{X})$  is given by  $\mathcal{T}(\lambda) = \text{Pois}(\lambda)$ , the Poisson distribution of parameter  $\lambda$ , and a measure  $m$  sampled from the prior  $\Pi(dm) = p_{\#}\mathcal{T}(dm)$  is a Poisson law with random parameter  $\lambda$  distributed according to the Gamma( $\alpha, \beta$ ) law. The latter is a conjugate prior for the Poisson distribution, and  $m$  sampled from  $\Pi_n(dm)$ , the posterior on models given  $D = (x_1, \dots, x_n)$ , is again a Poisson law  $\text{Pois}(\lambda)$ , with random parameter  $\lambda$  distributed according to Gamma( $n\bar{x}_n + \alpha, n + \beta$ ).

**Remark 2.3.** Hierarchical models are also catered for in the proposed setting. For instance, if in Example 2.1 the hyperparameter  $\mu_0$  is random with known density  $\pi_0$  on  $\mathbb{R}^d$ , it can be integrated out and the prior  $p$  on parameters becomes an infinite Gaussian mixture:

$$p(\theta) = p(\theta|\Sigma_0) = \int \mathcal{N}(\theta; \mu_0, \Sigma) \pi_0(\mu_0) d\mu_0.$$

The parametrization mapping  $\mathcal{T}$  is in this case the same as before, and the corresponding prior  $\Pi(dm) = p_{\sharp}\mathcal{T}(dm)$  and posteriors  $\Pi_n(dm)$  on models  $m \in \mathcal{P}(\mathcal{X})$  follow the same rationale as above.

As a cautionary note, the following example illustrates how defining estimators directly in terms of parameters might result in non-intrinsic criteria for model selection.

**Example 2.4.** On the parameter space  $\Theta = [0, 1]$  consider the priors  $p(d\theta) = d\theta$  and  $\hat{p}(d\theta) = 2\theta d\theta$ , and their associated parametrization maps  $\mathcal{T}(\theta) := B(\theta)$  and  $\hat{\mathcal{T}} := B(\theta^2)$  respectively. Here  $B(\xi)$  denotes the law of a  $\{0, 1\}$ -valued Bernoulli r.v. with  $\xi$  the probability of it being equal to 1. Notice that we have  $\Pi := p_{\sharp}\mathcal{T} = \hat{p}_{\sharp}\hat{\mathcal{T}}$ , as the law of  $\theta^2$  under  $\hat{p}$  is uniform on  $\Theta$ . Starting from the prior  $p$ , the posterior density of  $\theta$  given observations  $x_1, \dots, x_N$  is proportional to  $\theta^{S_N}(1-\theta)^{N-S_N}$ , with  $S_N = |\{i \leq N : x_i = 1\}|$ . Thus the MAP estimator for  $\theta$  is in this case  $\theta_N := \frac{S_N}{N}$ . On the other hand, starting from the prior  $\hat{p}$ , the posterior density of  $\theta$  is proportional to  $\theta^{2S_N+1}(1-\theta)^{N-S_N}$  and now the associated MAP estimator for  $\theta$  is  $\hat{\theta}_N := \sqrt{\frac{2S_N+1}{2N+1}}$ . Hence  $\mathcal{T}(\theta_N) \neq \hat{\mathcal{T}}(\hat{\theta}_N)$  in general (although their discrepancy vanishes in the limit as  $N \rightarrow \infty$ ). To summarize, although the same prior and posteriors at the level of models can arise by considering different parametrizations, the latter may easily define very different estimated models even if we agree on the estimation method (here the MAP).

## 2.2. Non parametric setting: posterior average estimators

The general “learning in model space” approach relies on loss functions that compare directly distributions (instead of their parameters), and thus allows us to define selection criteria based on intrinsic features of the models. It also allows for a wider choice of model-selection criteria, which can account for geometric or information-theoretic properties of the models. The next result illustrates the fact that many examples of Bayesian estimators or predictive posterior, including the classical *model average estimator*, correspond to finding an instance of *Fréchet mean* or barycenter [22, 48] under a suitable metric/divergence on probability measures. See Appendix A for the proof.

**Proposition 2.5.** Consider on the model  $\mathcal{M} = \mathcal{P}_{ac}(\mathcal{X})$  the loss functions  $L(m, \bar{m})$  given by:

- i) The  $L_2$ -distance:  $L(m, \bar{m}) = L_2(m, \bar{m}) := \frac{1}{2} \int_{\mathcal{X}} (m(x) - \bar{m}(x))^2 \lambda(dx)$ ,
- ii) The squared Hellinger distance  $L(m, \bar{m}) = H^2(m, \bar{m}) := \frac{1}{2} \int_{\mathcal{X}} \left( \sqrt{m(x)} - \sqrt{\bar{m}(x)} \right)^2 \lambda(dx)$ .
- iii) The forward Kullback-Leibler divergence:  $L(m, \bar{m}) = D_{KL}(m||\bar{m}) := \int_{\mathcal{X}} m(x) \ln \frac{m(x)}{\bar{m}(x)} \lambda(dx)$ ,
- iv) The reverse Kullback-Leibler divergence  $L(m, \bar{m}) = D_{RKL}(m||\bar{m}) = D_{KL}(\bar{m}||m) := \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{m(x)} \lambda(dx)$ .

Assume in each case that the infimum of the corresponding Bayes risk  $\bar{m} \mapsto R_L(\bar{m}|D)$  defined in equation (2.3) is attained with finite value in  $\mathcal{M}$ . Then, in cases i) and iii) the corresponding Bayes estimators (2.4) coincide with the standard Bayesian model average:

$$\hat{m}_{BMA}(x) := \mathbb{E}_{\Pi_n}[m](x) = \int_{\mathcal{M}} m(x) \Pi_n(dm). \quad (2.8)$$

Furthermore, the Bayes estimators corresponding to the cases ii) and iv) are given by the square model average and the exponential model average, respectively:

$$\hat{m}_{sq\tau}(x) = \frac{1}{Z_{sq\tau}} \left( \int_{\mathcal{M}} \sqrt{m(x)} \Pi_n(dm) \right)^2, \quad \hat{m}_{exp}(x) = \frac{1}{Z_{exp}} \exp \int_{\mathcal{M}} \ln m(x) \Pi_n(dm), \quad (2.9)$$

where  $Z_{sq\tau}$  and  $Z_{exp}$  denote the corresponding normalizing constants.

**Example 2.6.** If the posterior distribution was approximately equally concentrated on the models  $m_0 = \mathcal{N}(\mu_0, 1)$  and  $m_1 = \mathcal{N}(\mu_1, 1)$  with  $\mu_0 \neq \mu_1$ , that is, two (unimodal) Gaussian distributions with unit variance, then the standard model average is a bimodal non-Gaussian distribution with variance strictly larger than 1.

**Example 2.7.** In the parametric Bayesian model discussed in Example 2.1, the Bayesian model average estimator is the convolution of distributions on  $\mathbb{R}^d$ :

$$\begin{aligned} \hat{m}_{BMA}(x) &= \int_{\mathbb{R}^d} \mathcal{N}(x; \theta, \Sigma) \mathcal{N}(\theta; (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}_n), (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}) d\theta \\ &= \int_{\mathbb{R}^d} \mathcal{N}(x - \theta; 0, \Sigma) \mathcal{N}(\theta; (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}_n), (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}) d\theta \\ &= \mathcal{N}(x; (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}_n), \Sigma + (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}), \end{aligned} \quad (2.10)$$

that is, a Gaussian density with mean equal to the posterior mean estimator  $\hat{\theta}_{\ell_2}$ —see equation (2.7)—and a covariance matrix that is strictly larger (in the usual order on nonnegative definite symmetric matrices) than that of the predictive posterior law  $\mathcal{N}(\hat{\theta}_{\ell_2}, \Sigma)$  associated with it.

The Bayesian estimators considered Proposition 2.5, equations (2.8)–(2.9), share the following characteristic: their values at each point  $x \in \mathcal{X}$  are computed in terms of some posterior *average* of the values of certain functions evaluated at  $x$ . This is due to the fact that the corresponding distances/divergences on probability distributions are “vertical” [43]: computing the distance between distributions  $m$  and  $\bar{m}$  involves the integration of vertical displacements between the graphs of their densities across their domain. An undesirable fact about *vertical averages* is that they are not well suited to incorporate *geometric* properties into the model space (as illustrated by Exam. 2.6). More generally, model averages might yield solutions that can be hardly interpretable in terms of the prior and parameters, or even be intractable. This motivates us to explore the use of “horizontal” distances between probability distributions, thus extending the concept of Bayes estimator by making use of the geometric features of the model space. The next section presents the main ideas we will rely on to build such Bayes estimators.

### 3. WASSERSTEIN DISTANCES AND BARYCENTERS: A QUICK REVIEW

We shall now introduce objects analogous to those in Proposition 2.5 but suited to *Wasserstein distances* on the space of probability measures. The following framework is adopted in the sequel:

**Assumption 3.1.** The metric space  $(\mathcal{X}, d)$  is a separable locally-compact geodesic space endowed with a  $\sigma$ -finite Borel measure  $\lambda$  and  $p \geq 1$ .

By *geodesic* we mean that the space  $(\mathcal{X}, d)$  is complete and any pair of points admit a mid-point with respect to  $d$ . Next, we briefly recall some basic elements of optimal transport theory and Wasserstein distances, referring to [45, 46] for general background.

### 3.1. Optimal transport and the Wasserstein distance

Given two measures  $\mu, \nu$  over  $\mathcal{X}$  we denote by  $\text{Cpl}(\mu, \nu)$  the set of transport plans or couplings with marginals  $\mu$  and  $\nu$ , i.e.,  $\gamma \in \text{Cpl}(\mu, \nu)$  if and only if  $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ ,  $\gamma(dx, \mathcal{X}) = \mu(dx)$  and  $\gamma(\mathcal{X}, dy) = \nu(dy)$ . Given a real number  $p \geq 1$  we define the  $p$ -Wasserstein space  $\mathcal{W}_p(\mathcal{X})$  by

$$\mathcal{W}_p(\mathcal{X}) := \left\{ \eta \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} d(x_0, x)^p \eta(dx) < \infty, \text{ some } x_0 \right\}.$$

The  $p$ -Wasserstein distance between measures  $\mu$  and  $\nu$  is given by

$$W_p(\mu, \nu) = \left( \inf_{\gamma \in \text{Cpl}(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \gamma(dx, dy) \right)^{\frac{1}{p}}. \quad (3.1)$$

An optimizer of the right-hand side of equation (3.1) always exists and is called an *optimal transport*. The distance  $W_p$  turns  $\mathcal{W}_p(\mathcal{X})$  into a complete metric space. If in equation (3.1) we assume that  $p = 2$ ,  $\mathcal{X}$  is the Euclidean space, and if  $\mu$  is absolutely continuous, then Brenier's theorem ([45], Thm. 2.12(ii)) establishes the uniqueness of a minimizer, and guarantees that it is supported on the graph of the subdifferential of a convex function. The corresponding gradient is thus called an *optimal transport map*. Explicit formulae for such optimal transport maps do exist in some cases, e.g., for generic one-dimensional distributions and multivariate Gaussians when  $p = 2$  (see [14]). Contrary to the distances / divergences considered in Proposition 2.5, Wasserstein distances are *horizontal* [43], in the sense that they involve integrating horizontal displacements between the graphs of probability densities.

**Example 3.2.** The squared 2-Wasserstein distance between multivariate Gaussian distributions  $m = \mathcal{N}(\theta, \Sigma)$ , and  $\bar{m} = \mathcal{N}(\bar{\theta}, \bar{\Sigma})$  is given by  $W_2^2(m, \bar{m}) = \|\theta - \bar{\theta}\|^2 + \text{Tr} \left( \Sigma + \bar{\Sigma} - 2(\Sigma^{1/2} \bar{\Sigma} \Sigma^{1/2})^{1/2} \right)$ —see e.g. [19, 24]. Furthermore, if  $\Sigma$  and  $\bar{\Sigma}$  commute, we have  $W_2^2(m, \bar{m}) = \|\theta - \bar{\theta}\|^2 + \|\Sigma^{1/2} - \bar{\Sigma}^{1/2}\|_{\text{Fro}}^2$ .

**Example 3.3.** The 1-Wasserstein distance between Poisson distributions  $m = \text{Pois}(\lambda)$  and  $\bar{m} = \text{Pois}(\bar{\lambda})$  is  $|\lambda - \bar{\lambda}|$ , which can be verified from the well-known expression  $W_1(m, \bar{m}) = \int |m(-\infty, t] - \bar{m}(-\infty, t]| dt$  (valid for general one-dimensional distributions  $m, \bar{m}$ ), and the so-called “Poisson-Gamma dual relation”:  $e^{-\lambda} \sum_{k=0}^n \frac{\lambda^k}{k!} = \int_{\lambda}^{\infty} \frac{t^n}{n!} e^{-t} dt$ . When  $\lambda > \bar{\lambda}$ , the optimal coupling between  $X \sim \text{Pois}(\lambda)$  and  $Y \sim \text{Pois}(\bar{\lambda})$  is obtained taking  $X \sim \text{Pois}(\lambda)$  and  $Y$  binomial with parameters  $(X, \frac{\bar{\lambda}}{\lambda})$  conditionally on  $X$ . Alternatively, one could consider the coupling  $(X, Y) := (N_{\lambda}, N_{\bar{\lambda}})$ , where  $\{N_t\}_{t \geq 0}$  is a Poisson process with intensity 1.

### 3.2. Wasserstein barycenter

Let us now recall the Wasserstein barycenter, introduced in [1] and further studied in [28, 33, 40], among others. Our definition slightly extends the ones in those works in that the optimization problem is posed in a possibly strict subset of the usual one.

**Definition 3.4.** Let  $\Gamma \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ . The  $p$ -Wasserstein risk of  $\nu \in \mathcal{P}(\mathcal{X})$  is

$$V_p(\nu) := \int_{\mathcal{P}(\mathcal{X})} W_p(m, \nu)^p \Gamma(dm) \leq +\infty.$$

Given a measurable set  $\mathcal{M} \subseteq \mathcal{P}(\mathcal{X})$ , any measure  $\hat{m}_p \in \mathcal{M}$  which attains the quantity

$$\inf_{\nu \in \mathcal{M}} V_p(\nu)$$

with finite value, is called a  $p$ -Wasserstein barycenter of  $\Gamma$  over  $\mathcal{M}$ .

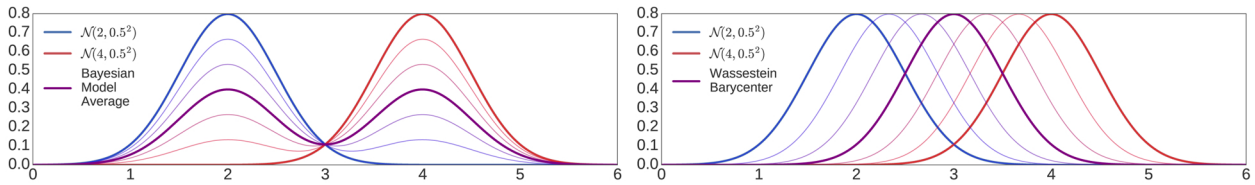


FIGURE 1. Vertical interpolation (*left*) and horizontal interpolation (*right*) of two Gaussian densities.

Notice that in principle we are not assuming  $\mathcal{M}$  to be a model space for  $\Gamma$ , but this will often be the case. When the support of  $\Gamma$  is infinite and  $\mathcal{M} = \mathcal{W}_p(\mathcal{X})$ , this object is termed *p-Wasserstein population barycenter* of  $\Gamma$  as introduced in [33]; see [10].

**Example 3.5.** Given two univariate Gaussian distributions  $m_0 = \mathcal{N}(\mu_0, \sigma_0^2)$  and  $m_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ , one can verify—using the expression in Example 3.2—that the 2-Wasserstein barycenter for  $\Gamma(dm) = \frac{1}{2}\delta_{m_0}(dm) + \frac{1}{2}\delta_{m_1}(dm)$  is given by  $\hat{m} = \mathcal{N}(\frac{\mu_0 + \mu_1}{2}, (\frac{\sigma_0 + \sigma_1}{2})^2)$ . This should be compared to Example 2.6. Figure 1 illustrates the corresponding vertical and a horizontal interpolations between two Gaussian densities with different means and the same variance.

Let us introduce additional notation for the sequel and review then some basic properties of Wasserstein barycenters. Considering  $\mathcal{W}_p(\mathcal{X})$  with the complete metric  $W_p$  as a base Polish metric space, we define  $\mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  in the natural way:  $\Gamma \in \mathcal{P}(\mathcal{W}_p(\mathcal{X}))$  is an element of  $\mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  if it is concentrated on a set of measures with finite moments of order  $p$ , and moreover for some (and then all)  $\tilde{m} \in \mathcal{W}_p(\mathcal{X})$  it satisfies

$$\int_{\mathcal{P}(\mathcal{X})} W_p(m, \tilde{m})^p \Gamma(dm) < \infty.$$

We endow  $\mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  with the corresponding  $p$ -Wasserstein distance, which we also denote  $W_p$  for simplicity. Also, if  $\Gamma$  is concentrated on measures with finite moments of order  $p$  which have densities with respect to  $\lambda$ , then we write  $\Gamma \in \mathcal{P}(\mathcal{W}_{p,ac}(\mathcal{X}))$  and use the notation  $\Gamma \in \mathcal{W}_p(\mathcal{W}_{p,ac}(\mathcal{X}))$  if, as before,  $\int_{\mathcal{P}(\mathcal{X})} W_p(m, \tilde{m})^p \Gamma(dm) < \infty$  for some  $\tilde{m}$ .

**Remark 3.6.** If  $\Gamma \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  has a  $p$ -Wasserstein barycenter  $\hat{m}_p$  over  $\mathcal{M}$ , then

$$\infty > \int_{\mathcal{P}(\mathcal{X})} W_p(m, \hat{m}_p)^p \Gamma(dm) = \int_{\mathcal{W}_p(\mathcal{X})} W_p(m, \hat{m}_p)^p \Gamma(dm),$$

hence  $\Gamma \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ . Moreover,  $\Gamma \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  is equivalent to the corresponding *model average*  $\bar{m}(dx) := \mathbb{E}_\Gamma[m](dx)$  having a finite  $p$ -moment, since for any  $y \in \mathcal{X}$ ,

$$\int_{\mathcal{W}_p(\mathcal{X})} W_p(\delta_y, m)^p \Gamma(dm) = \int_{\mathcal{W}_p(\mathcal{X})} \int_{\mathcal{X}} d(y, x)^p m(dx) \Gamma(dm) = \int_{\mathcal{X}} d(y, x)^p \int_{\mathcal{W}_p(\mathcal{X})} m(dx) \Gamma(dm).$$

We next state an existence result first established in Theorem 2 of [33] for the case  $\mathcal{M} = \mathcal{W}_p$ . See Appendix B for a simpler, more direct, proof.

**Theorem 3.7.** Suppose Assumption 3.1 holds,  $\Gamma \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ , and  $\mathcal{M} \subseteq \mathcal{W}_p$  is a weakly closed set. There exists a  $p$ -Wasserstein barycenter of  $\Gamma$  over  $\mathcal{M}$  if and only if  $\Gamma \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ .

Regarding uniqueness, the following general result was proven in Proposition 6 of [33] for the case  $\mathcal{X} = \mathbb{R}^q$  with  $d$  the Euclidean distance and  $p = 2$  (observe that, in that situation, the previous result applies):

**Lemma 3.8.** Assume  $\Gamma \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^q))$  and that there exists a set  $A \subseteq \mathcal{W}_2(\mathbb{R}^q)$  of measures with

$$\mu \in A, B \in \mathcal{B}(\mathbb{R}^q), \dim(B) \leq q-1 \implies \mu(B) = 0,$$

and  $\Gamma(A) > 0$ . Then,  $\Gamma$  admits a unique 2-Wasserstein population barycenter over  $\mathcal{W}_2(\mathbb{R}^q)$ .

**Remark 3.9.** Observe that the model space  $\mathcal{M} = \mathcal{W}_{p,ac}(\mathcal{X})$  is not weakly closed. Nevertheless, the existence and uniqueness of a population barycenter over that set can still be guaranteed when  $p = 2$ ,  $\mathcal{X} = \mathbb{R}^q$ ,  $d$  is the Euclidean distance,  $\lambda$  is the Lebesgue measure, and

$$\Gamma(\{m : \|\frac{dm}{d\lambda}\|_\infty < \infty\}) > 0. \quad (3.2)$$

This was proven in Theorem 6.2 of [28] for compact finite-dimensional manifolds with lower-bounded Ricci curvature (equipped with the volume measure), but one can read-off the (non-compact but flat) Euclidean case  $\mathcal{X} = \mathbb{R}^q$  from the proof therein, in order to establish the absolute continuity of a barycenter over  $\mathcal{W}_2(\mathbb{R}^q)$ , in the setting of Lemma 3.8. If  $|\text{supp}(\Gamma)| < \infty$  then equation (3.2) can be relaxed to the condition  $\Gamma(\{m : m \ll \lambda\}) > 0$ , as shown in [1] or Theorem 5.1 of [28].

The following statement, corresponding to Lemma 3.1 of [7], provides a useful description of barycenters which generalizes a result proven in [3] when  $|\text{supp}(\Gamma)| < \infty$ .

**Lemma 3.10.** Assume  $p = 2$ ,  $\mathcal{X} = \mathbb{R}^q$ ,  $d = \text{Euclidean distance}$ ,  $\lambda = \text{Lebesgue measure}$ . Let  $\Gamma \in \mathcal{W}_2(\mathcal{W}_2(\mathcal{X}))$  and  $\tilde{\Gamma} \in \mathcal{W}_2(\mathcal{W}_{2,ac}(\mathcal{X}))$ . There exists a jointly measurable function  $\mathcal{W}_{2,ac}(\mathbb{R}^q) \times \mathcal{W}_2(\mathbb{R}^q) \times \mathbb{R}^q \ni (\mu, m, x) \mapsto T_\mu^m(x)$  which is  $\mu(dx)\Gamma(dm)\tilde{\Gamma}(d\mu)$ -a.s. equal to the unique optimal transport map from  $\mu$  to  $m$  at  $x$ . Furthermore, letting  $\hat{\mu}$  be a barycenter of  $\Gamma$ , we have

$$x = \int T_{\hat{\mu}}^m(x) \Gamma(dm), \hat{\mu}(dx) - \text{a.s.}$$

#### 4. BAYESIAN WASSERSTEIN BARYCENTER AND STATISTICAL PROPERTIES

Building on the Wasserstein distance as a loss function on models, we arrive to the following central object of the article:

**Definition 4.1.** Let us consider a prior  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  with model space  $\mathcal{M} \subseteq \mathcal{W}_{p,ac}(\mathcal{X})$  and data  $D = (x_1, \dots, x_n)$  which determines  $\Pi_n$  as in equation (2.2). We define the  $p$ -Wasserstein Bayes risk of  $\bar{m} \in \mathcal{W}_p(\mathcal{X})$  and a Bayes Wasserstein barycenter (BWB) estimator  $\hat{m}_p^n$  over  $\mathcal{M}$  respectively as follows:

$$V_p^n(\bar{m}|D) := \int_{\mathcal{P}(\mathcal{X})} W_p(m, \bar{m})^p \Pi_n(dm), \text{ and} \quad (4.1)$$

$$\hat{m}_p^n \in \underset{\bar{m} \in \mathcal{M}}{\text{argmin}} V_p^n(\bar{m}|D), \quad (4.2)$$

if the corresponding minimum is finite.

**Example 4.2.** In the setting of Example 2.1, the 2-Wasserstein loss function on Gaussian models (see Example 3.2) induces the usual quadratic loss on the mean parameters:  $W_2^2(m, \bar{m}) = \ell(\theta, \bar{\theta}) := \|\theta - \bar{\theta}\|^2$ . Thus, in the notation of equation (2.5), we have

$$V_2^n(\bar{m}|D) = R_\ell(\bar{\theta}|D).$$

This implies that, for any tuple of data point  $D = (x_1, \dots, x_n)$  the BWB corresponds to the Gaussian distribution  $\mathcal{N}(\bar{\theta}, \Sigma) \in \mathcal{M}$  with mean  $\bar{\theta} = \hat{\theta}_{\ell_2} = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}_n)$ , that is, the posterior mean estimator of



the mean parameter. Moreover, in this particular case the barycentric cost or optimal 2-Wasserstein Bayes risk equals the trace of the covariance of a random vector with law given in equation (2.6), *i.e.*,  $\text{Tr}((\Sigma_0^{-1} + n\Sigma^{-1})^{-1})$ . Notice that the covariance  $\Sigma$  of the BWB is strictly smaller than that of the corresponding BMA estimator in equation (2.10) (in the usual order on symmetric positive semidefinite matrices). This is, in fact, a general property of the BWB as claimed in Proposition 4.9 below.

**Example 4.3.** Similarly, in the parametric setting of Example 2.2, the problem of finding BWB estimators can be written in terms of the parametric loss induced by the corresponding  $p$ -Wasserstein distance on Poisson distributions, computed in Example 3.3. We thus deduce for  $p = 1$  that the BWB estimator  $\hat{m}_1^n$  corresponds, in that setting, to the law  $\text{Pois}(\lambda_1)$  with  $\lambda_1$  the median of the posterior distribution  $\text{Gamma}(n\bar{x}_n + \alpha, n + \beta)$ ; for which an explicit expression is not available.

**Example 4.4.** Let us assume  $\mathcal{X} = \mathbb{R}$  and that  $\Pi(dm)$  is supported on continuous models  $m$  over  $\mathbb{R}$ . Let  $F_m$  and  $Q_m := F_m^{-1}$  denote, respectively, the cumulative distribution function and the right-continuous quantile function of such  $m$ . The coupling  $(x, T_{m_0}^m(x))$ , with  $x$  distributed like  $m_0$  and  $T_{m_0}^m$  the increasing map  $T_{m_0}^m(x) := Q_m(F_{m_0}(x))$ , is known to be optimal for the  $p$ -Wasserstein distance, for any  $p \geq 1$  (see [45], Rem. 2.19(iv)). The BWB  $\hat{m}_n = \hat{m}_n^p$  of the posterior  $\Pi_n$  is also independent of  $p$  and is characterized via its quantile, as follows:

$$Q_{\hat{m}_n}(\cdot) = \int_{\mathcal{M}} Q_m(\cdot) \Pi_n(dm).$$

Interestingly enough, the model average  $\bar{m}_n := \int m \Pi_n(dm)$  of  $\Pi_n$  is in turn characterized by its averaged cumulative distribution function:  $F_{\bar{m}_n}(\cdot) = \int_{\mathcal{M}} F_m(\cdot) \Pi_n(dm)$ . See Section 5 in the companion paper [7] for details and further discussion on one-dimensional Wasserstein barycenters, in particular, on geometric properties they inherit from the elements  $m$  of the support of the prior/posterior.

**Remark 4.5.** In the sequel, unless otherwise stated, an a.s. statement about  $\Pi_n$  is meant to hold almost surely with respect to the marginal law  $m^{\otimes n}(dx_1, \dots, dx_n) \int_{\mathcal{M}} \Pi(dm)$  of a data sample of size  $n$  (sometimes called prior predictive distribution). In particular, for  $\Pi(dm)$ -almost every  $m$ , such statement holds for  $m^{\otimes n}(dx_1, \dots, dx_n)$ -almost every sample  $(x_1, \dots, x_n)$ .

**Remark 4.6.** We observe that  $\Pi \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  implies for each  $n \geq 1$  that

$$\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X})) \text{ a.s.}$$

Indeed, for fixed  $\tilde{m} \in \mathcal{W}_p(\mathcal{X})$ , we have (we thank an anonymous referee for pointing this out):

$$\begin{aligned} \infty &> \int_{\mathcal{M}} W_p(m, \tilde{m})^p \Pi(dm) = \int_{\mathcal{X}^n} \int_{\mathcal{M}} W_p(m, \tilde{m})^p m(x_1) \cdots m(x_n) \Pi(dm) \lambda(dx_1) \cdots \lambda(dx_n) \\ &= \int_{\mathcal{X}^n} \int_{\mathcal{M}} W_p(m, \tilde{m})^p \Pi_n(dm) \int_{\mathcal{M}} \tilde{m}(x_1) \cdots \tilde{m}(x_n) \Pi(d\tilde{m}) \lambda(dx_1) \cdots \lambda(dx_n) \quad (4.3) \\ &= \int_{\mathcal{X}^n} \left[ \int_{\mathcal{M}} W_p(m, \tilde{m})^p \Pi_n(dm) \right] \tilde{m}^{\otimes n}(dx_1, \dots, dx_n) \int_{\mathcal{M}} \Pi(d\tilde{m}). \end{aligned}$$

However,  $\Pi \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  is in general not enough for  $\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  to hold a.s. for i.i.d. data points  $(x_1, \dots, x_n)$  sampled from a *fixed* law  $m_0$ , which would be the natural setting to formulate the question of Bayesian consistency (see next subsection). Corollary 4.19 below ensures this fact for suitable given laws  $m_0$ , in the framework of Bayesian consistency. In Appendix C we further provide a sufficient condition on the prior  $\Pi$  (termed *integrability after updates*) ensuring that  $\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  for *every* possible tuple of data points  $(x_1, \dots, x_n)$ .

The following statement gathers the discussion in Section 3.2 for the case  $\Gamma = \Pi_n$ , as well as the main point of Remark 4.6:

**Theorem 4.7.** *Suppose Assumption 3.1 holds and the model space  $\mathcal{M}$  is weakly closed. Let  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  be a prior with model space  $\mathcal{M} \subseteq \mathcal{W}_{p,ac}(\mathcal{X})$  and  $\Pi_n$  be the corresponding posterior given the data  $D = (x_1, \dots, x_n)$ . The following are equivalent:*

- a) *A  $p$ -Wasserstein barycenter estimator  $\hat{m}_p^n$  for  $\Pi_n$  over  $\mathcal{M}$  exists a.s.*
- b)  *$\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  a.s.*
- c) *The model average  $\bar{m}^n(dx) = \mathbb{E}_{\Pi_n}[m](dx)$  has a.s. a finite  $p$ -moment.*

Moreover, if  $\Pi \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ , a  $p$ -Wasserstein barycenter estimator  $\hat{m}_p^n$  over  $\mathcal{M}$  exists a.s. for every  $n \geq 1$ .

We thus make in all the sequel the following assumption:

**Assumption 4.8.**  $\Pi \in \mathcal{W}_p(\mathcal{W}_{p,ac}(\mathcal{X}))$  and there exists a weakly closed model space  $\mathcal{M} \subseteq \mathcal{W}_{p,ac}(\mathcal{X})$  for  $\Pi$ .

Since  $\Pi_n \ll \Pi$  a.s., Assumption 4.8 together with Remark 4.6 imply that  $\Pi_n \in \mathcal{W}_p(\mathcal{W}_{p,ac}(\mathcal{X}))$  a.s. for all  $n$ .

We will next study some basic statistical properties of the BWB estimator.

#### 4.1. Variance reduction with respect to BMA

In this subsection, we assume that  $\mathcal{X} = \mathbb{R}^d$ ,  $\lambda$  = Lebesgue measure,  $d$  = Euclidean distance and  $p = 2$ . Let  $\hat{m} := \hat{m}_2^n$  be the unique population barycenter of  $\Pi_n$  in that case, and denote by  $(m, x) \mapsto T^m(x)$  a measurable function equal  $\lambda(dx)\Pi(dm)$  a.e. to the unique optimal transport map from  $\hat{m}$  to  $m \in \mathcal{W}_2(\mathcal{X})$ . As a consequence of Lemma 3.10 we have the fixed-point property  $\hat{m} = (\int T^m \Pi(dm))(\hat{m})$ . Thus, for all convex functions  $\varphi$ , non negative or with at most quadratic growth, we have

$$\begin{aligned} \mathbb{E}_{\hat{m}}[\varphi(x)] &= \int_{\mathcal{X}} \varphi(x) \hat{m}(dx) = \int_{\mathcal{X}} \varphi \left( \int_{\mathcal{M}} T^m(x) \Pi_n(dm) \right) \hat{m}(dx) \\ &\leq \int_{\mathcal{X}} \int_{\mathcal{M}} \varphi(T^m(x)) \Pi_n(dm) \hat{m}(dx) = \int_{\mathcal{M}} \int_{\mathcal{X}} \varphi(T^m(x)) \hat{m}(dx) \Pi_n(dm) \\ &= \int_{\mathcal{M}} \int_{\mathcal{X}} \varphi(x) m(dx) \Pi_n(dm) = \int_{\mathcal{X}} \varphi(x) \int_{\mathcal{M}} m(dx) \Pi_n(dm) = \mathbb{E}_{\hat{m}_{BMA}}[\varphi(x)], \end{aligned}$$

where  $\hat{m}_{BMA} = \mathbb{E}_{\Pi_n}[m]$  is the Bayesian model average in equation (2.8). We have used Jensen's inequality and Fubini's theorem. This means that the BWB estimator is less spread, or *smaller*, in the sense of convex-order of probability measures, than the BMA. As a consequence, we have:

**Proposition 4.9.** *Consider  $p = 2$ , and let  $\hat{m}_{BMA}$  and  $\hat{m}_2^n$  respectively denote the BMA and the BWB estimators associated with  $\Pi_n$ . Then, we have  $\mathbb{E}_{\hat{m}_{BMA}}[x] = \mathbb{E}_{\hat{m}_2^n}[x]$  and  $\mathbb{E}_{\hat{m}_{BMA}}[\|x\|^2] \geq \mathbb{E}_{\hat{m}_2^n}[\|x\|^2]$ . In other words, the BWB estimator has less variance than the BMA. Furthermore, with  $\bar{x}$  denoting the mean w.r.t. the BMA or BWB, the corresponding covariances satisfy:  $\mathbb{E}_{\hat{m}_{BMA}}[(x - \bar{x})(x - \bar{x})^t] \geq \mathbb{E}_{\hat{m}_2^n}[(x - \bar{x})(x - \bar{x})^t]$  in the usual order for symmetric positive semidefinite matrices. The inequalities are strict unless  $\Pi_n$  is a Dirac mass.*

*Proof.* Given the previous discussion, the equality of means is obtained by taking  $\varphi(x)$  and  $-\varphi(x)$  equal to each coordinate of  $x$  and the inequality of variances by taking  $\varphi(x) = \|x\|^2$ . The inequality for the covariances follows by taking  $\varphi(x) = (y^t(x - \bar{x}))^2$  for arbitrary  $y \in \mathbb{R}^d$ .

For the last claim, we just need to make explicit the corresponding equality case of Jensen's inequality as used in the preceding discussion. Since  $\|x\|^2$  is strictly convex, the equality of second moments implies that  $\hat{m}_2^n(dx)$  a.e.  $x$ , that  $T^m(x)$  is  $\Pi_m(dm)$  a.s. constant. This entails that the map  $T := T^m$  does not depend on  $m$  and that for  $\Pi_m(dm)$  a.e.  $m$ , it holds that  $m = T(\hat{m}_2^n)$ . The equality case for the covariances is reduced to the previous one considering their traces using also the equality of means.  $\square$

## 4.2. Convergence to the true model and Bayesian consistency

A natural question in Bayesian statistics is whether a given predictive posterior estimator is *consistent* (see [17, 23, 44]). In short, this means the convergence of the predictive posterior law, in some specified sense, towards the *true* model  $m_0$ , as we observe more and more i.i.d. data sampled from  $m_0$ . We are specifically interested in the question of whether the BWB estimator converges to the model  $m_0$  and, more precisely, on conditions which guarantee that

$$W_p(\hat{m}_p^n, m_0) \rightarrow 0, \quad m_0^{(\infty)} \text{ a.s.} \quad (4.4)$$

as  $n \rightarrow \infty$ , where  $\hat{m}_p^n$  is for each  $n$  a BWB over a model space  $\mathcal{M}$ . Here and in the sequel,  $m_0^{(\infty)}$  denotes the product law of the infinite sample  $\{x_n\}_n$  of i.i.d. data distributed according to  $m_0$ . We will see that this question is linked to the notion of consistency (*cf.* [23], Def. 6.1) of the prior, introduced next:

**Definition 4.10.** The prior  $\Pi$  is said to be consistent at  $m_0$  in the weak topology (resp.  $p$ -Wasserstein topology) if for each open neighbourhood  $U$  of  $m_0$  in the weak topology of  $\mathcal{P}(\mathcal{X})$  (resp.  $p$ -Wasserstein topology of  $\mathcal{W}_p(\mathcal{X})$ ), one has  $\Pi_n(U^c) \rightarrow 0$ ,  $m_0^{(\infty)} - \text{a.s.}$

**Remark 4.11.** We notice that in the literature of Bayesian consistency, see *e.g.* [23, 44],  $\Pi$  satisfying the above property (in either topology) would be called “strongly consistent” at  $m_0$  in allusion to the  $m_0^{(\infty)}$ -a.s. convergences, whereas the term “weakly consistent” would be used when those convergences hold in probability. Since we will be only dealing with the almost sure notion, in order to avoid possible confusions with topological concepts, the adverb “strongly” is omitted throughout when referring to consistency of the prior, while the adverb “weakly” only refers to the weak topology on probability measures.

The celebrated Schwartz theorem [44] provides sufficient conditions for consistency w.r.t. a given topology, see also Chapter 6 of [23] for a modern treatment. A key ingredient is the notion of Kullback-Leibler support:

**Definition 4.12.** A measure  $m_0$  is an element the Kullback-Leibler support of  $\Pi$ , denoted

$$m_0 \in \text{KL}(\Pi),$$

if  $\Pi(m : D_{KL}(m_0||m) < \varepsilon) > 0$  for every  $\varepsilon > 0$ , with  $D_{KL}(m_0||m)$  the reverse Kullback-Leibler entropy defined as  $\int \log \frac{m_0}{m}(x) m_0(dx)$  if  $m_0 \ll m$  and as  $+\infty$  otherwise.

**Remark 4.13.** The statistical model is interpreted as being “correct” or *well specified*, if the data distribution  $m_0$  is an element of  $\text{supp}(\Pi)$ , the support of  $\Pi$  w.r.t. the weak topology, see [9, 23, 26, 29, 30]. The condition  $m_0 \in \text{KL}(\Pi)$  is stronger. Indeed, by the Csiszar-Pinsker inequality and the fact that the dual bounded-Lipschitz distance (metrizing the weak topology in  $\mathcal{P}(\mathcal{X})$ ) is majorized by the total variation distance, one can check that  $\text{KL}(\Pi) \subseteq \text{supp}(\Pi)$ . In particular, one has  $\text{KL}(\Pi) \subseteq \mathcal{M}$  for any weakly closed model space  $\mathcal{M}$  for  $\Pi$ . (The reader may consult the mentioned works for the *misspecified* framework too).

**Remark 4.14.** If  $m_0 \in \text{KL}(\Pi)$ , then  $m_0(dx) \ll m(dx) \int_{\mathcal{M}} \Pi(dm)$ , the marginal law of one data point  $x$  in the Bayesian model defined by  $\Pi$ . Indeed, for any measurable set  $A \subseteq \mathcal{X}$  such that  $\int_{\mathcal{M}} \int_A m(dx) \Pi(dm) = 0$  we have  $\Pi(m : m_0 \ll m, m(A) = 0) = \Pi(m : m_0 \ll m) \geq \Pi(m : D_{KL}(m_0||m) < +\infty) > 0$ . This will be useful later.

We recall (see Theorem 6.17 and Example 6.20 in [23]) the following result which concerns the weak topology:

**Theorem 4.15.** Assume (only) that  $(\mathcal{X}, d)$  is Polish endowed with a  $\sigma$ -finite Borel measure  $\lambda$ , that  $\Pi \in \mathcal{P}_{ac}(\mathcal{X})$  and that  $m_0 \in \text{KL}(\Pi)$ . Then,  $\Pi$  is consistent at  $m_0$  in the weak topology.

We now state a general result relating consistency of  $\Pi$  at  $m_0$  in the  $p$ -Wasserstein topology and the convergence (4.4) (or consistency of the predictive posterior  $\hat{m}_p^n$ ), with other asymptotic properties of the posterior

laws in the Wasserstein setting. Recall that the notation  $W_p$  throughout stands for the Wasserstein distance both in  $\mathcal{W}_p(\mathcal{X})$  and  $\mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ .

**Theorem 4.16.** *Suppose Assumptions 3.1 and 4.8 hold, that  $m_0 \in \mathcal{M}$ , and that*

$$\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X})), \quad m_0^{\otimes n} - \text{a.s. for all } n \geq 1. \quad (4.5)$$

The following are equivalent:

- a)  $W_p(\Pi_n, \delta_{m_0}) \rightarrow 0$ ,  $m_0^{(\infty)}$ -a.s. as  $n \rightarrow \infty$ .
- b)  $m_0^{(\infty)}$ -a.s. as  $n \rightarrow \infty$ , we have  $W_p(\hat{m}_p^n, m_0) \rightarrow 0$  and the barycentric cost (or optimal  $p$ -Wasserstein Bayes risk or)  $\int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \Pi_n(dm)$  goes to 0.
- c)  $\Pi$  is consistent at  $m_0$  in the  $p$ -Wasserstein topology and the  $p$ -moment of the BMA estimator (2.8) converge  $m_0^{(\infty)}$ -a.s. to that of  $m_0$  as  $n \rightarrow \infty$ , i.e. for some (and then all)  $x_0 \in \mathcal{X}$  we have

$$\int_{\mathcal{X}} d(x, x_0)^p \hat{m}_{BMA}^n(dx) = \int_{\mathcal{M}} \int_{\mathcal{X}} d(x, x_0)^p m(dx) \Pi_n(dm) \rightarrow \int_{\mathcal{X}} d(x, x_0)^p m_0(dx), \quad m_0^{(\infty)} \text{ a.s.} \quad (4.6)$$

- d)  $\Pi$  is consistent at  $m_0$  in the weak topology and for some (and then all)  $x_0 \in \mathcal{X}$  we have

$$\int_{\mathcal{X}} \left| \int_{\mathcal{X}} d(x, x_0)^p m(dx) - \int_{\mathcal{X}} d(x, x_0)^p m_0(dx) \right| \Pi_n(dm) \rightarrow 0, \quad m_0^{(\infty)} \text{ -a.s.} \quad (4.7)$$

*Proof.* By minimality of a barycenter over  $\mathcal{M}$ ,

$$\int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \Pi_n(dm) \leq \int_{\mathcal{M}} W_p(m, m_0)^p \Pi_n(dm) = W_p(\Pi_n, \delta_{m_0})^p.$$

Thus, for some  $c > 0$  depending only on  $p$ ,

$$W_p(m_0, \hat{m}_p^n)^p \leq c \int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \Pi_n(dm) + c \int_{\mathcal{M}} W_p(m, m_0)^p \Pi_n(dm) \quad (4.8)$$

$$\leq 2c W_p(\Pi_n, \delta_{m_0})^p \quad (4.9)$$

proving that a)  $\Rightarrow$  b). The converse b)  $\Rightarrow$  a) follows from

$$\int_{\mathcal{M}} W_p(m, m_0)^p \Pi_n(dm) \leq c \int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \Pi_n(dm) + c W_p(\hat{m}_p^n, m_0)^p.$$

Let us now show that a)  $\Rightarrow$  c). The convergence  $W_p(\Pi_n, \delta_{m_0}) \rightarrow 0$  implies (by the Portmanteau theorem) that  $\limsup_{n \rightarrow \infty} \Pi_n(F) \leq \delta_{m_0}(F)$  for all closed sets  $F$  of  $\mathcal{W}_p(\mathcal{X})$ . Taking  $F = U^c$  with  $U$  a neighborhood of  $m_0$  yields the consistency of  $\Pi$  at  $m_0$  in the  $p$ -Wasserstein topology. Moreover it implies that for each  $\tilde{m} \in \mathcal{W}_p(\mathcal{X})$ ,

$$\int_{\mathcal{M}} W_p^p(m, \tilde{m}) \Pi_n(dm) \rightarrow \int_{\mathcal{M}} W_p^p(m, \tilde{m}) \delta_{m_0}(dm) = W_p^p(m_0, \tilde{m}), \quad m_0^{(\infty)} \text{ -a.s.}$$

as  $n \rightarrow \infty$ . Choosing  $\tilde{m} = \delta_{x_0}$ , we have  $W_p^p(m, \tilde{m}) = \int_{\mathcal{X}} d(x, x_0)^p m(dx)$  for any  $m$ , from where (4.6) follows.

We next prove that c)  $\Rightarrow$  a). The space  $\mathcal{W}_p(\mathcal{X})$  being Polish, there is countable basis  $\mathcal{U}$  of open neighborhoods of  $m_0$  such that  $m_0^{(\infty)}$ -a.s.,  $\Pi_n(U^c) \rightarrow 0$  for all  $U \in \mathcal{U}$ . Thus, if  $G \subseteq \mathcal{W}_p(\mathcal{X})$  is any open set such that  $m_0 \in G$ , for some  $U \in \mathcal{U}$  we have  $U \subseteq G$  and therefore  $\liminf_{n \rightarrow \infty} \Pi_n(G) \geq \liminf_{n \rightarrow \infty} \Pi_n(U) = 1 - \lim_{n \rightarrow \infty} \Pi_n(U^c) = 1 = \delta_{m_0}(G)$ ,  $m_0^{(\infty)}$ -a.s. This implies, by the Portmanteau theorem, that the sequence  $(\Pi_n)_n$  weakly converges to  $\delta_{m_0}$ ,  $m_0^{(\infty)}$ -a.s., as probability measures on the metric space  $\mathcal{W}_p(\mathcal{X})$ . As in the previous (converse) implication,

we obtain from (4.6) the convergence of some moments of order  $p$  of  $\Pi_n$ , to the corresponding moment of  $\delta_{m_0}$ ,  $m_0^{(\infty)}$ -a.s. This plus the weak convergence just established imply that  $W_p(\Pi_n, \delta_{m_0}) \rightarrow 0$   $m_0^{(\infty)}$ -a.s.

We have thus established that a), b) and c) are equivalent. Notice now that the function

$$m \mapsto \Phi(m) := \left| \int_{\mathcal{X}} \left( d(x, x_0)^p - \int_{\mathcal{X}} d(y, x_0)^p m_0(dy) \right) m(dx) \right|, \quad (4.10)$$

is continuous on  $\mathcal{W}_p(\mathcal{X})$ , since  $x \mapsto d(x, x_0)^p - \int_{\mathcal{X}} d(y, x_0)^p m_0(dy)$  is continuous with polynomial growth of order  $p$  on  $\mathcal{X}$ . Moreover,  $|\Phi(m)| \leq \int_{\mathcal{X}} d(y, x_0)^p m_0(dy) + W_p^p(m, \delta_{x_0})$ , that is,  $\Phi$  has polynomial growth of order at most  $p$  on  $\mathcal{W}_p(\mathcal{X})$ . Therefore, if a) or equivalently c) holds, we have  $m_0^{(\infty)}$ -a.s. that

$$\int_{\mathcal{M}} \Phi(m) \Pi_n(dm) \rightarrow \int_{\mathcal{M}} \Phi(m) \delta_{m_0}(dm) = \Phi(m_0) = 0$$

which is tantamount to (4.7). Moreover if c) holds, consistency of  $\Pi$  at  $m_0$  in the weak topology is obvious. This shows that c)  $\Rightarrow$  d).

Last, if d) holds, we deduce with Markov's inequality that, for each rational  $\varepsilon > 0$ ,

$$\Pi_n \left\{ m \in \mathcal{M} : \left| \int_{\mathcal{X}} d(x, x_0)^p m(dx) - \int_{\mathcal{X}} d(x, x_0)^p m_0(dx) \right| \geq \varepsilon \right\} \rightarrow 0, m_0^{(\infty)} \text{ a.s.} \quad (4.11)$$

as  $n \rightarrow \infty$ . This, together with the consistency of  $\Pi$  at  $m_0$  w.r.t. weak topology, is equivalent to having that consistency w.r.t. the  $p$ -Wasserstein topology. Moreover, we have

$$\left| \int_{\mathcal{X}} d(x, x_0)^p \hat{m}_{BMA}^n(dx) - \int_{\mathcal{X}} d(x, x_0)^p m_0(dx) \right| \leq \int_{\mathcal{M}} \left| \int_{\mathcal{X}} d(x, x_0)^p m(dx) - \int_{\mathcal{X}} d(x, x_0)^p m_0(dx) \right| \Pi_n(dm)$$

and so the convergence (4.7) implies the convergence (4.6). This and the previous show that d) implies c), concluding the proof.  $\square$

**Remark 4.17.** The convergence (4.11) for each  $\varepsilon > 0$  is exactly what one must add to consistency at  $m_0$  of  $\Pi$  in the weak topology to obtain such consistency in the  $p$ -Wasserstein topology. However, the latter is only equivalent to the  $m_0^{(\infty)}$ -a.s. weak convergence of  $\Pi_n$  to  $\delta_{m_0}$  as measures on the metric space  $\mathcal{W}_p(\mathcal{X})$  as  $n \rightarrow \infty$ , which is not enough to grant the  $m_0^{(\infty)}$ -a.s. convergence in  $\mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$  of  $\Pi_n$  to  $\delta_{m_0}$ . Similarly, consistency at  $m_0$  of  $\Pi$  in the  $p$ -Wasserstein topology cannot in general be obtained by adding only the convergence of moments 4.6 to the consistency at  $m_0$  of  $\Pi$  in the weak topology. Of course, if the space  $\mathcal{X}$  is bounded, weak and  $p$ -Wasserstein topologies on it coincide, and consistency of  $\Pi$  at  $m_0$  in the weak topology *implies* in that case the convergence (4.7) (since the function  $\Phi$  in the previous proof is in that case continuous and bounded). Hence, in that case, all the equivalent properties in Theorem 4.16 are satisfied. The same is true if  $\Pi(dm)$ -a.e.  $m$  is supported on a fixed bounded (weakly) closed set  $\mathcal{Y} \subseteq \mathcal{X}$  (just replace  $\mathcal{X}$  by  $\mathcal{Y}$ ).

An immediate consequence of the proof of Theorem 4.16 (cf. the estimate (4.8)) is the following bound, which can be used to obtain quantitative estimates for the average rate of convergence of the BWB, if the posterior and the Wasserstein distance between models are explicit enough:

**Corollary 4.18.** *Under the assumptions of Theorem 4.16, for some constant  $c_p > 0$  we have*

$$\mathbb{E} \left( W_p(m_0, \hat{m}_p^n)^p \right) \leq c_p \mathbb{E} \left( W_p(\Pi_n, \delta_{m_0})^p \right).$$

The following result gathers sufficient conditions for consistency at  $m_0$  of  $\Pi$  in the  $p$ -Wasserstein topology and convergence of the BWB.

**Corollary 4.19.** *Suppose Assumptions 3.1 and 4.8 hold, and moreover that  $m_0 \in \text{KL}(\Pi)$ . Then  $\Pi$  is consistent at  $m_0$  in the weak topology and condition (4.5) holds. If moreover either condition (i) or (ii) below hold, then  $\Pi$  is consistent at  $m_0$  in the  $p$ -Wasserstein topology and  $W_p(\hat{m}_p^n, m_0) \rightarrow 0$ ,  $m_0^{(\infty)}$ -a.s. as  $n \rightarrow \infty$ :*

(i) *the convergence (4.7) holds;*

(ii) *for some  $q > p$ ,  $m_0^{(\infty)}$ -a.s. the  $q$  moments of the BMA estimator are bounded uniformly in  $n$ .*

*Proof.* In view of Theorem 4.15, to prove the first claim we just need to prove that (4.5) holds. To that end, let us first check that, whenever  $m_0 \in \text{KL}(\Pi)$ , we have  $m_0^{\otimes n}(dx_1, \dots, dx_n) \ll m^{\otimes n}(dx_1, \dots, dx_n) \int_{\mathcal{M}} \Pi(dm)$ . If  $A \subseteq \mathcal{X}^n$  measurable is such that  $\int_A m^{\otimes n}(dx_1, \dots, dx_n) \int_{\mathcal{M}} \Pi(dm) = 0$ , then for  $\Pi(dm)$ -a.e.  $m$  and every  $x_i \in \mathcal{X}$ ,  $i = 2, \dots, n$ , one has  $m(\{x \in \mathcal{X} : (x, x_2, \dots, x_n) \in A\}) = 0$ , from which we get  $m_0(\{x \in \mathcal{X} : (x, x_2, \dots, x_n) \in A\}) = 0$  by Remark 4.14, hence  $m_0^{\otimes n}(A) = 0$ . We conclude noting that the set  $A := \{(x_1, \dots, x_n) \in \mathcal{X}^n : \Pi_n \notin \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))\}$  has null  $m^{\otimes n}(dx_1, \dots, dx_n) \int_{\mathcal{M}} \Pi(dm)$ -measure, by Remark 4.6.

Given the previous and part d) of Theorem 4.16, the convergence (4.7) immediately implies that  $\Pi$  is consistent at  $m_0$  in the  $p$ -Wasserstein topology and that  $W_p(\hat{m}_p^n, m_0) \rightarrow 0$   $m_0^{(\infty)}$ -a.s. as  $n \rightarrow \infty$ . To conclude the proof it is enough to show that the  $m_0^{(\infty)}$ -a.s. uniform boundedness of the  $q$  moments of the BMA estimator for some  $q > p$ , implies, under the given assumptions, that said convergence (4.7) holds. Consider to that end the function  $\Phi$  defined by (4.10) in the proof of Theorem 4.16. For each  $\varepsilon > 0$  we have

$$\begin{aligned} \int_{\mathcal{M}} \Phi(m) \Pi_n(dm) &\leq \varepsilon + \int_{\mathcal{M} \cap \{\Phi(m) \geq \varepsilon\}} \Phi(m) \Pi_n(dm) \\ &\leq \varepsilon + \left( \int_{\mathcal{M}} \Phi(m)^{q/p} \Pi_n(dm) \right)^{p/q} (\Pi_n(m : \Phi(m) \geq \varepsilon))^{1-p/q} \end{aligned}$$

The assumption on the  $q$ -moments of the BWA estimators imply that  $\sup_n \int_{\mathcal{M}} \Phi(m)^{q/p} \Pi_n(dm) < \infty$ ,  $m_0^{(\infty)}$ -a.s., following from applying Jensen's inequality to the convex function  $s \mapsto |s|^{q/p}$  on  $\mathbb{R}$  and the integral w.r.t.  $m(dx)$  defining  $\Phi(m)$ . Since  $\varepsilon > 0$  is arbitrary, it just remains to ensure that  $\Pi_n(m : \Phi(m) \geq \varepsilon) \rightarrow 0$   $m_0^{(\infty)}$ -a.s. as  $n \rightarrow \infty$  (that is, the convergence (4.11) holds). The elementary relations  $t = t \wedge R + (t - R)_+$  and  $(t - R)_+ \leq t \mathbf{1}_{\{t > R\}}$  for all  $t, R \geq 0$  yield the bound

$$\begin{aligned} \Pi_n(m : \Phi(m) \geq \varepsilon) &\leq \Pi_n \left( m : \left| \int_{\mathcal{X}} R \wedge d(x, x_0)^p m(dx) - \int_{\mathcal{X}} R \wedge d(x, x_0)^p m_0(dx) \right| \geq \varepsilon/2 \right) \\ &\quad + \Pi_n \left( m : \int_{\{x \in \mathcal{X} : d(x, x_0)^p > R\}} d(x, x_0)^p [m(dx) + m_0(dx)] \geq \varepsilon/2 \right), \end{aligned}$$

where the first term on the r.h.s. goes  $m_0^{(\infty)}$ -a.s. to 0 as  $n \rightarrow \infty$ , since  $x \mapsto R \wedge d(x, x_0)^p$  is a bounded continuous function and  $\Pi$  is consistent at  $m_0$  in the weak topology. The second term on the r.h.s. is bounded by

$$\frac{2}{\varepsilon} \left[ \sup_{n \in \mathbb{N}} \int_{\mathcal{M}} \int_{\mathcal{X}} \mathbf{1}_{\{d(x, x_0)^p > R\}} d(x, x_0)^p m(dx) \Pi_n(dm) + \int_{\mathcal{X}} \mathbf{1}_{\{d(x, x_0)^p > R\}} d(x, x_0)^p m_0(dx) \right].$$

For each  $\varepsilon > 0$ , this expression can be made arbitrarily small by taking  $R$  large enough, since  $m_0 \in \mathcal{W}_p(\mathcal{X})$  and because the uniform boundedness of the  $q$ -moments of the BWA estimators implies their  $p$ -moments are uniformly integrable. We conclude that  $\Pi_n(m : \Phi(m) \geq \varepsilon) \rightarrow 0$   $m_0^{(\infty)}$ -a.s. as  $n \rightarrow \infty$  as desired.  $\square$



**Example 4.20.** Given Gaussian distributions  $m_0 = \mathcal{N}(\theta_0, \Sigma)$  and  $m = \mathcal{N}(\theta, \Sigma)$  on  $\mathcal{X} = \mathbb{R}^d$ , one can check that  $D_{KL}(m_0 || m) = \frac{1}{2}(\theta - \theta_0)^t \Sigma^{-1}(\theta - \theta_0)$  and hence that  $\text{KL}(\Pi) = \text{supp}(\Pi)$  in the parametric Bayesian model dealt with in Examples 2.1 and 4.2. Therefore, by Theorem 4.15,  $\Pi$  in those examples is consistent w.r.t. the weak topology at  $m_0 = \mathcal{N}(\theta_0, \Sigma)$  for any  $\theta_0 \in \mathbb{R}^d$ . Moreover, the prior  $\Pi$  is easily seen to be in  $\mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^d))$ , hence Lemma 4.19 ensures that  $\Pi_n \in \mathcal{W}_2(\mathcal{W}_2(\mathbb{R}^d))$ ,  $m_0^{\otimes n}$ -a.s. for all  $n \geq 1$ . This fact alternatively follows from the existence of a 2-Wasserstein barycenter for  $\Pi_n$  for any data points  $D$ , verified in this case. To verify consistency of  $\Pi$  in the 2-Wasserstein topology at any  $m_0$  as well as convergence of the BWB, let us compute  $W_2^2(\Pi_n, \delta_{m_0})$  and apply directly Theorem 4.16. Denoting by  $\theta$  a r.v. with law  $\mathcal{N}(\hat{\theta}_{\ell_2}, (\Sigma_0^{-1} + n\Sigma^{-1})^{-1})$ , we find

$$\begin{aligned} W_2^2(\Pi_n, \delta_{m_0}) &= \int_{\mathcal{M}} W_2^2(m, m_0) \Pi_n(dm) \\ &= \mathbb{E}_{\theta}(\|\theta - \theta_0\|^2) \\ &= \|\theta_0 - \hat{\theta}_{\ell_2}\|^2 + \mathbb{E}_{\theta}(\|\hat{\theta}_{\ell_2} - \theta\|^2) \\ &= W_2^2(\hat{m}_2^n, m_0) + \int_{\mathcal{M}} W_2^2(m, \hat{m}_2^n) \Pi_n(dm), \end{aligned}$$

with  $\mathbb{E}_{\theta}(\|\hat{\theta}_{\ell_2} - \theta\|^2) = \int_{\mathcal{M}} W_2^2(m, \hat{m}_2^n) \Pi_n(dm) = \text{tr}((\Sigma_0^{-1} + n\Sigma^{-1})^{-1}) \rightarrow 0$  as  $n \rightarrow \infty$  and

$$\|\theta_0 - \hat{\theta}_{\ell_2}\|^2 = W_2^2(\hat{m}_2^n, m_0) = \|(\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x}_n) - \theta_0\|^2$$

which  $m_0^{(\infty)}$  a.s. goes to 0 as  $n \rightarrow \infty$  by the Law of Large Numbers, whenever the data  $(x_n)_{n \geq 1}$  consists of an i.i.d. sample from the true model  $m_0$ . Notice that the identity  $W_2^2(\Pi_n, \delta_{m_0}) = W_2^2(\hat{m}_2^n, m_0) + \int_{\mathcal{M}} W_2^2(m, \hat{m}_2^n) \Pi_n(dm)$  found in this case, can be seen as a *bias-variance type decomposition* for the posterior law  $\Pi_n$  on  $\mathcal{W}_2(\mathcal{X})$ . In this case, one can readily see that  $\mathbb{E}(W_2^2(\hat{m}_2^n, m_0)) \leq C/n$ .

**Example 4.21.** Although the BWB estimator  $\hat{m}_1^n$  in the Poisson parametric Bayesian model of Example 2.2 is not explicit (see Exam. 4.3), we can still apply Theorem 4.16 a) to prove that it converges to the true model generating the data. Indeed, if  $m_0 = \text{Pois}(\lambda_0)$ , using the expression for the 1-Wasserstein distance between Poisson laws computed in Example 3.3 we see that

$$W_1(\Pi_n, \delta_{m_0}) = \int W_1(m, m_0) \Pi_n(dm) = \mathbb{E}|\lambda - \lambda_0| \leq \mathbb{E}(|\lambda - \lambda_0|^2)^{1/2},$$

with  $\lambda$  a r.v. with law  $\text{Gamma}(n\bar{x}_n + \alpha, n + \beta)$ . An elementary computation using the Gamma distribution's mean and variance shows, in this case, that

$$\mathbb{E}(|\lambda - \lambda_0|^2) = \frac{n\bar{x}_n + \alpha}{(n + \beta)^2} + \left[ \frac{n\bar{x}_n + \alpha}{n + \beta} - \lambda_0 \right]^2,$$

which goes to zero  $m_0^{(\infty)}$  a.s. as  $n \rightarrow \infty$ , whenever the data  $(x_n)_{n \geq 1}$  is an i.i.d. sample from the true model  $m_0$ . We deduce with Corollary 4.19 that  $\mathbb{E}(W_1(m_0, \hat{m}_1^n)) \leq C/\sqrt{n}$ .

As noticed earlier, consistency of  $\Pi$  at  $m_0$  w.r.t. the  $p$ -Wasserstein topology is not enough to grant that  $W_p(\Pi_n, \delta_{m_0}) \rightarrow 0$   $m_0^{(\infty)}$  - a.s.. But we will see next that this is true under a boundedness condition on the support of  $\Pi$ . Recall that  $\text{supp}(\Pi)$  is said to be bounded in  $\mathcal{W}_p(\mathcal{X})$  if

$$\text{diam}(\Pi) := \sup_{m, \bar{m} \in \text{supp}(\Pi)} W_p(m, \bar{m}) < \infty.$$

A typical example is the finitely parametrized case with compact parameter space  $\Theta$  and continuous parametrization mapping  $\mathcal{T} : \Theta \rightarrow \mathcal{W}_p(\mathcal{X})$ . More generally,  $\text{diam}(\Pi) < \infty$  amounts to  $\Pi$  being supported on a set of models with centered  $p$ -moments bounded by a constant. In particular,  $\mathcal{X}$  and the support of every  $m \in \text{supp}(\Pi)$  may be unbounded, and still  $\text{supp}(\Pi)$  be bounded. We have

**Lemma 4.22.** *If  $\Pi$  is consistent at  $m_0$  in the  $p$ -Wasserstein topology and  $\text{supp}(\Pi)$  is bounded in  $\mathcal{W}_p(\mathcal{X})$ , then  $W_p(\Pi_n, \delta_{m_0}) \rightarrow 0$   $m_0^{(\infty)}$ -a.s. as  $n \rightarrow \infty$ .*

*Proof.* Let  $\varepsilon > 0$  and  $B = \{m : W_p(m, m_0) < \varepsilon\}$ , then

$$\begin{aligned} W_p(\Pi_n, \delta_{m_0})^p &= \int_{\mathcal{M}} W_p(m, m_0)^p \Pi_n(dm) \\ &= \int_B W_p(m, m_0)^p \Pi_n(dm) + \int_{B^c} W_p(m, m_0)^p \Pi_n(dm) \\ &\leq \varepsilon^p + \int_{B^c} W_p(m, m_0)^p \Pi_n(dm). \end{aligned}$$

Since  $\varepsilon > 0$  is arbitrary, we only need to check that the second term in the last line goes  $m_0^{(\infty)}$ -a.s. to zero as  $n \rightarrow \infty$ . By consistency we have  $\Pi_n(B^c) \rightarrow 0$ ,  $m_0^{(\infty)}$ -a.s. as  $n \rightarrow \infty$ , and since  $\text{supp}(\Pi_n) \subseteq \text{supp}(\Pi)$ , we conclude that

$$\int_{B^c} W_p(m, m_0)^p \Pi_n(dm) \leq \text{diam}(\Pi)^p \Pi_n(B^c) \rightarrow 0, \quad m_0^{(\infty)} - \text{a.s.}$$

□

**Remark 4.23.**  $\Pi$  consistent at  $m_0$  in the weak topology and  $\text{supp}(\Pi)$  bounded in  $\mathcal{W}_p(\mathcal{X})$  is in general not enough to obtain the conclusion of Lemma 4.22. To see this, write

$$W_p(\Pi_n, \delta_{m_0})^p = \int_{B_w} W_p(m, m_0)^p \Pi_n(dm) + \int_{B_w^c} W_p(m, m_0)^p \Pi_n(dm), \quad (4.12)$$

where  $B_w$  is a fixed small weak neighborhood of  $m_0$  (e.g. one can use the bounded Lipschitz distance to build  $B_w$ ). As in the proof of Lemma 4.22, the second term in the r.h.s. of (4.12) goes  $m_0^{(\infty)}$ -a.s. to zero as  $n \rightarrow \infty$  in this case too. However there is no reason why the term  $\int_{B_w} W_p(m, m_0)^p \Pi_n(dm)$  should be small. The reason is that for the functional  $m \mapsto \Psi(m) := W_p(m, m_0)^p$ , even if bounded on  $\mathcal{M} := \text{supp}(\Pi)$ , there is no reason why  $\Psi|_{B_w \cap \mathcal{M}}$  should be small (no matter how small  $B_w$  may be). Indeed, the statement “ $\Psi|_{B_w \cap \mathcal{M}}$  is small if  $B_w$  is small” would mean mathematically that the weak and the  $p$ -Wasserstein topologies coincide on  $\mathcal{M}$  locally around  $m_0$ , but this is not true in general, even if  $\mathcal{M}$  is bounded<sup>1</sup>. This should not be confused with the fact that, if  $\mathcal{M}$  equipped with the  $p$ -Wasserstein topology and metric is bounded, then on  $\mathcal{W}_p(\mathcal{M})$  weak and  $p$ -Wasserstein convergence coincide. Consistency at  $m_0$  of  $\Pi$  in the weak topology, on the other hand, is equivalent to  $\Pi_n \rightarrow \delta_{m_0}$ ,  $m_0^{(\infty)}$ -a.s. in the weak topology of  $\mathcal{W}_p(\mathcal{M})$ , when  $\mathcal{M}$  is equipped with the weak topology. This does not imply  $\Pi_n \rightarrow \delta_{m_0}$ ,  $m_0^{(\infty)}$ -a.s. in the weak topology of  $\mathcal{W}_p(\mathcal{M})$ , when  $\mathcal{M}$  is equipped with the  $p$ -Wasserstein topology, and hence provides another point of view as to why the l.h.s. of (4.12) does not go to zero without stronger assumptions.

The next result based on Schwartz theorem provides a (rather strong) condition ensuring that the equivalent properties in Theorem 4.16 hold. Unfortunately, we have not been able to prove such a result under more general assumptions.

<sup>1</sup>For instance, for  $\mathcal{X} = \mathbb{R}$  and  $p = 1$  we have that  $\mathcal{M} := \{m_n := \frac{n-1}{n}\delta_0 + \frac{1}{n}\delta_n\}_{n \in \mathbb{N}} \cup \{\delta_0\}$  is 1-Wasserstein bounded, and yet  $m_n \rightarrow \delta_0$  weakly but not in 1-Wasserstein topology.

**Theorem 4.24.** *Under Assumptions 3.1 and 4.8, suppose moreover that  $m_0 \in KL(\Pi)$  and that, for some  $\lambda_0 > 0$  and  $x_0 \in \mathcal{X}$ , one has*

$$\sup_{m \in \text{supp}(\Pi)} \int_{\mathcal{X}} e^{\lambda_0 d^p(x, x_0)} m(dx) < +\infty.$$

*Then,  $\Pi$  is consistent at  $m_0$  in the  $p$ -Wasserstein topology. Moreover, we have  $W_p(\Pi_n, \delta_{m_0}) \rightarrow 0$ ,  $m_0^{(\infty)}$ -a.s. and the BWB estimator is consistent in the sense that*

$$W_p(\hat{m}_p^n, m_0) \rightarrow 0, \quad m_0^{(\infty)} - \text{a.s.}$$

Before proving Theorem 4.24 some remarks on its assumptions and proof are in order.

**Remark 4.25.** The uniform control assumed on  $p$ -exponential moments implies, by Jensen's inequality, that  $\sup_{m \in \text{supp}(\Pi)} W_p^p(m, \delta_{x_0}) = \sup_{m \in \text{supp}(\Pi)} \int_{\mathcal{X}} d^p(x, x_0) m(dx) < +\infty$ . By triangle inequality, this in turn implies that  $\text{supp}(\Pi)$  is bounded.

**Remark 4.26.** The general picture of Bayesian consistency (including the misspecified case) parallels in several aspects Sanov's large deviations theorem (see *e.g.* the Bayesian Sanov Theorem ([26], Thm. 2.1) and references therein), and it similarly relies on exponential controls of (posterior) integrals. The proof of Theorem 4.24 follows the argument of Example 6.20 in [23], where Theorem 4.15 above is obtained by combining Schwartz' theorem ([23], Thm. 6.17) with Hoeffding's concentration inequality, to get uniform exponential controls of the posterior mass of complements of weak neighborhoods of  $m_0$ , which are defined in terms of bounded random variables. A  $p$ -exponential moment control is what is needed to derive such concentration inequalities for unbounded random variables (*e.g.* moments), defining neighborhoods in the Wasserstein topology. Notice that finite  $p$ -exponential moments are also required for Sanov's theorem to hold in the  $p$ -Wasserstein topology [47]. The uniform bound on exponential moments appears however too strong an assumption (it does not hold *e.g.* in the setting of Example 4.20). The question of relaxing that condition is left for future work.

**Remark 4.27.** In the *misspecified* framework dealt with in [9, 23, 26, 29, 30], and paralleling results applicable to the weak topology in those works, we expect  $\hat{m}_p^n \rightarrow \text{argmin}_{m \in \mathcal{M}} D_{KL}(m_0 || m)$  w.r.t.  $W_p$  to hold, under suitable assumptions.

*Proof of Theorem 4.24.* First we show that if  $U$  is any  $\mathcal{W}_p(\mathcal{X})$ -neighbourhood of  $m_0$  then  $m_0^{(\infty)}$ -a.s. we have  $\liminf_n \Pi_n(U) \geq 1$ . According to Schwartz Theorem in the extended form ([23], Thm. 6.17), under the assumption that  $m_0 \in KL(\Pi)$ , it is enough to find for each such  $U$  a sequence of measurable functions or “tests”  $\varphi_n : \mathcal{X}^n \rightarrow [0, 1]$  such that

1.  $\varphi_n(x_1, \dots, x_n) \rightarrow 0$ ,  $m_0^{(\infty)} - \text{a.s.}$  and
2.  $\limsup_n \frac{1}{n} \log \left( \int_{U^c} m^{\otimes n} (1 - \varphi_n) \Pi(dm) \right) < 0$ .

First we will construct tests  $\{\varphi_n\}_n$  that satisfy Point 1 and Point 2 above, over an appropriate subbase of neighbourhoods  $U$ , to finally extend those properties to general neighborhoods.

Recall that  $\mu_k \rightarrow \mu$  in  $W_p(\mathcal{X})$  iff for all continuous functions  $\psi$  on  $\mathcal{X}$  with  $|\psi(x)| \leq K(1 + d^p(x, x_0))$  for some  $K \in \mathbb{R}_+$  it holds that  $\int_{\mathcal{X}} \psi(x) \mu_n(dx) \rightarrow \int_{\mathcal{X}} \psi(x) \mu(dx)$ ; see [46]. Given such  $\psi$  and  $\varepsilon > 0$  we define the open sets

$$U_{\psi, \varepsilon} := \left\{ m : \int_{\mathcal{X}} \psi(x) m(dx) < \int_{\mathcal{X}} \psi(x) m_0(dx) + \varepsilon \right\},$$

which form a sub-base for the  $p$ -Wasserstein neighborhood system at the distribution  $m_0$ . We can assume that  $K = 1$  by otherwise considering  $U_{\psi/K, \varepsilon/K}$  instead. Given a neighborhood  $U := U_{\psi, \varepsilon}$  of  $m_0$  as above, we define

the test functions

$$\varphi_n(x_1, \dots, x_n) = \begin{cases} 1 & \text{if } \frac{1}{n} \sum_{i=1}^n \psi(x_i) > \int_{\mathcal{X}} \psi(x) m_0(dx) + \frac{\varepsilon}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

By the law of large numbers,  $m_0^{(\infty)} - a.s : \varphi_n(x_1, \dots, x_n) \rightarrow 0$ , so Point 1 is verified. Point 2 is trivial if  $r := \Pi(U^c) = 0$ , so we assume from now on that  $r > 0$ . Thanks to the exponential moments control assumed on  $\text{supp}(\Pi)$ , for  $\Pi(dm)$  a.e.  $m$  the random variable  $Z = 1 + d^p(X, x_0)$  with  $X \sim m$  has a moment-generating function  $\mathcal{L}_m(t)$  which is finite for all  $t \in [0, \lambda_0]$ , namely

$$\mathcal{L}_m(t) := \mathbb{E} [e^{tZ}] = e^t \int_{\mathcal{X}} e^{td^p(x, x_0)} m(dx) < +\infty.$$

We thus have the bounds

$$\int_{\mathcal{X}} |\psi(x)|^k m(dx) \leq \mathbb{E} [Z^k] \leq k! \mathcal{L}_m(t) t^{-k}, \quad \forall \lambda_0 \geq t > 0,$$

for all  $k \in \mathbb{N}$ . Therefore, we may apply Bernstein's inequality in the form of Corollary 2.10 in [36] to the random variables  $\{-\psi(x_i)\}_i$  under the measure  $m^{(\infty)}$  on  $\mathcal{X}^{\mathbb{N}}$ , obtaining for any  $\alpha < 0$  that

$$m^{(\infty)} \left( \sum_{i=1}^n [\psi(x_i) - \int_{\mathcal{X}} \psi(x) m(dx)] \leq \alpha \right) \leq e^{-\frac{\alpha^2}{2(v-c\alpha)}},$$

with  $v := 2n\mathcal{L}_m(t)t^{-2}$ ,  $c := t^{-1}$ , and  $0 < t \leq \lambda_0$ . Going back to the tests  $\varphi_n$  and using the definition of  $U^c$  we deduce that

$$\begin{aligned} \int_{U^c} m^{\otimes n} (1 - \varphi_n) \Pi(dm) &= \int_{U^c} m^{\otimes n} \left( \frac{1}{n} \sum_{i=1}^n \psi(x_i) \leq \int_{\mathcal{X}} \psi(x) m_0(dx) + \frac{\varepsilon}{2} \right) \Pi(dm) \\ &\leq \int_{U^c} m^{\otimes n} \left( \frac{1}{n} \sum_{i=1}^n \psi(x_i) \leq \int_{\mathcal{X}} \psi(x) m(dx) - \frac{\varepsilon}{2} \right) \Pi(dm) \\ &= \int_{U^c} m^{\otimes n} \left( \sum_{i=1}^n [\psi(x_i) - \int_{\mathcal{X}} \psi(x) m(dx)] \leq -\frac{n\varepsilon}{2} \right) \Pi(dm) \\ &\leq \int_{U^c} \exp \left\{ -\frac{n\varepsilon^2}{2} \frac{t^2}{8\mathcal{L}_m(t) + t\varepsilon} \right\} \Pi(dm) \\ &\leq r \exp \left\{ -\frac{n\varepsilon^2}{2} \frac{t^2}{8\sup_{m \in \text{supp}(\Pi)} \mathcal{L}_m(t) + t\varepsilon} \right\}. \end{aligned}$$

Thanks to the uniform control of exponential moments on the support, we conclude as desired that

$$\limsup_n \frac{1}{n} \log \left( \int_{U^c} m^{\otimes n} (1 - \varphi_n) \Pi(dm) \right) \leq -\frac{t^2 \varepsilon^2}{16 \sup_{m \in \text{supp}(\Pi)} \mathcal{L}_m(t) + 2t\varepsilon} < 0.$$

Now, a general neighborhood  $U$  contains a finite intersection of, say,  $N \in \mathbb{N}$  elements of the sub-base, *i.e.*  $\bigcap_{i=1}^N U_{\psi_i, \varepsilon_i} \subseteq U$ , so

$$\int_{U^c} m^{\otimes n} (1 - \varphi_n) \Pi(dm) \leq \sum_{i=1}^N \int_{U_{\psi_i, \varepsilon_i}^c} m^{\otimes n} (1 - \varphi_n) \Pi(dm).$$

Therefore we can conclude as in the sub-base case that Point 2 is verified. All in all, we have established that  $\Pi$  is  $p$ -Wasserstein consistent at  $m_0$ . Thanks to Lemma 4.22 and the boundedness of  $\text{supp}(\Pi)$  (see Rem. 4.25) the last two claims are immediate.  $\square$

## 5. BWB CALCULATION THROUGH DESCENT ALGORITHMS IN WASSERSTEIN SPACE

In this section we review some of the available methods to compute Wasserstein barycenters, and then explain how they can be used to calculate the BWB estimator. We will first survey the method proposed by Álvarez-Esteban, Barrio, Cuesta-Albertos and Matrán in Theorem 3.6 of [3] and by Zemel and Panaretos in Theorem 3 and Corollary 2 of [48], applicable in the setting where the probability distribution  $\Gamma$  on the model space has finite support. This method is interpreted in [48] as a *gradient descent in the Wasserstein space*. We will then discuss the main aspects of the *stochastic gradient descent in Wasserstein space* (SGDW), introduced in our companion paper [7], building also upon the gradient descent idea. The latter method allows moreover the computation of population barycenters for a distribution  $\Gamma$  on the model space, using a streaming of random probability distributions sampled from it. In particular, we will recall the conditions established in [7] which ensure its convergence. The reader is referred to [16, 18, 34, 41] for alternative approaches to computing Wasserstein barycenters.

The two discussed methods can be easily implemented when explicit analytical expressions for optimal transport maps between distributions in the model space are available (see [7] or Section 6 for some examples). Moreover, in that case these methods can easily be coupled with sampling procedures (MCMC or others) for the posterior distribution  $\Pi_n$  in order to compute the BWB. We will see that the computation of the BWB through the SGDW has several advantageous features. In particular, when expressions for optimal transport maps are known, it can be done at nearly the same cost as the posterior sampling.

From now on we specialize Assumption 3.1, and make the following set of assumptions:

**Assumption 5.1.**  $p = 2$ ,  $\mathcal{X} = \mathbb{R}^q$ ,  $d =$  Euclidean metric,  $\lambda =$  Lebesgue measure. Furthermore,  $\Gamma \in \mathcal{W}_2(\mathcal{W}_{p,ac}(\mathbb{R}^q))$  and there is a model space  $\mathcal{M} \subseteq \mathcal{W}_{p,ac}(\mathbb{R}^q)$  for  $\Gamma$  which is weakly closed.

The following concept will be central in the “gradient-type” algorithms we consider:

**Definition 5.2.** We say that  $\mu \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$  is a Karcher mean of  $\Gamma \in \mathcal{W}_2(\mathcal{W}_{2,ac}(\mathbb{R}^q))$  if

$$\mu \left( \left\{ x : x = \int_{\mathcal{W}_2(\mathbb{R}^q)} T_\mu^m(x) \Gamma(dm) \right\} \right) = 1.$$

**Remark 5.3.** It is known that any 2-Wasserstein barycenter is a Karcher mean (c.f. [48]). However, the class of Karcher means is in general a strictly larger one, see [3]. For conditions ensuring uniqueness of Karcher means, see [38, 48] for the case when the support of  $\Gamma$  is finite and [10] for the case of an infinite support. In one dimension, the uniqueness of Karcher means holds without further assumptions. See [7] for a deeper discussion.

### 5.1. Gradient descent on Wasserstein space

Consider  $\Gamma \in \mathcal{W}_2(\mathcal{W}_{p,ac}(\mathbb{R}^q))$  finitely supported: for some  $m_i \in \mathcal{W}_{p,ac}(\mathbb{R}^q)$ ,  $i = 1, \dots, L \in \mathbb{N}$ , we have

$$\Gamma = \sum_{i \leq L} \lambda_i \delta_{m_i}.$$

Following [3] and [48], we define an operator over  $\mathcal{W}_{p,ac}(\mathcal{X})$  by

$$G(m) := \left( \sum_{i=1}^L \lambda_i T_m^{m_i} \right) (m). \quad (5.1)$$

Starting from  $\mu_0 \in \mathcal{W}_{p,ac}(\mathcal{X})$  one can then define the sequence

$$\mu_{n+1} := G(\mu_n), \text{ for } n \geq 0. \quad (5.2)$$

The next result proven in Theorem 3.6 of [3] and independently in Theorem 3 and Corollary 2 of [48], establishes the convergence of the above sequence to a fixed-point of  $G$ , which is nothing other than a Karcher mean for  $\Gamma$ :

**Proposition 5.4.** *The sequence  $\{\mu_n\}_{n \geq 0}$  in equation (5.2) is tight and every weakly convergent subsequence of  $\{\mu_n\}_{n \geq 0}$  converges in  $W_2$  to an absolutely continuous measure in  $\mathcal{W}_2(\mathbb{R}^q)$  which is a Karcher mean of  $\Gamma$ . If some  $m_i$  has a bounded density, and if there exists a unique Karcher mean  $\hat{m}$ , then  $\hat{m}$  is the Wasserstein barycenter of  $\Gamma$  and  $W_2(\mu_n, \hat{m}) \rightarrow 0$ .*

Panaretos and Zemel ([48], Thm. 1) discovered that the sequence (5.2) can indeed be interpreted as a gradient descent (GD) scheme with respect to the *Riemannian-like* structure of the Wasserstein space  $\mathcal{W}_2(\mathbb{R}^q)$ . In fact, the functional on  $\mathcal{W}_2(\mathbb{R}^q)$  given by

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

has a *Frechet derivative* at each point  $m \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$ , given by

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

where  $I$  is the identity map in  $\mathbb{R}^q$ . This means that for each such  $m$ , one has

$$\frac{F(\hat{m}) - F(m) - \int_{\mathbb{R}^q} \langle F'(m)(x), T_{\hat{m}}^m(x) - x \rangle m(dx)}{W_2(\hat{m}, m)} \rightarrow 0, \quad (5.3)$$

when  $W_2(\hat{m}, m)$  goes to zero, by virtue of Corollary 10.2.7 in [5]. It follows from Brenier's theorem ([45], Thm. 2.12(ii)) that  $\hat{m}$  is a fixed point of  $G$  defined in equation (5.1) if and only if  $F'(\hat{m}) = 0$ . The gradient descent sequence in Wasserstein space (GDW) with step  $\gamma$  starting from  $\mu_0 \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$  is defined by (c.f. [48])

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

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

and it coincides with the sequence in equation (5.2) if  $\gamma = 1$ . These ideas serve as inspiration for the stochastic gradient descent iteration in the next part.

## 5.2. Stochastic gradient descent for population barycenters

We recall next the stochastic gradient descent sequence introduced in the companion paper [7], where the reader is referred to for details. Additionally to the assumptions in the previous part, we will also make use of an extra one introduced in [7]:

**Assumption 5.5.**  $\Gamma$  has a  $W_2$ -compact model space  $K_\Gamma \subseteq \mathcal{W}_{2,ac}(\mathbb{R}^q)$ . Moreover this set is *geodesically convex*: for every  $\mu, \nu \in K_\Gamma$  and  $t \in [0, 1]$ ,  $((1 - t)I + tT_\mu^\nu)(\mu) \in K_\Gamma$ , with  $I$  the identity operator.

In particular, under these assumptions, for each  $\nu \in \mathcal{W}_2(\mathbb{R}^q)$  and  $\Gamma(dm)$  a.e.  $m$ , there is a unique optimal transport map  $T_m^\nu$  from  $m$  to  $\nu$  and, by Proposition 6 of [33], the 2-Wasserstein population barycenter is unique. We notice that, although strong at first sight, assumption 5.5 can be guaranteed in suitable parametric situations (*e.g.*, Gaussian, or even the location scatter setting recalled in Sect. 6.1), or under moment and density constraints on the measures in  $K_\Gamma$  (*e.g.*, under uniform bounds on their moments of order  $2 + \varepsilon$  and their Boltzmann entropy, which are geodesically convex functionals, see [5]).



**Definition 5.6.** Let  $\mu_0 \in K_\Gamma$ ,  $m_k \stackrel{\text{iid}}{\sim} \Gamma$ , and  $\gamma_k > 0$  for  $k \geq 0$ . We define the stochastic gradient descent in Wasserstein space sequence (SGDW) by

$$\mu_{k+1} := [(1 - \gamma_k)I + \gamma_k T_{\mu_k}^{m_k}] (\mu_k), \text{ for } k \geq 0. \quad (5.4)$$

The sequence is a.s. well-defined, as one can show by induction that  $\mu_k \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$  a.s. thanks to Assumption 5.5. The rationale for definition 5.4 is similar to that of Section 5.1, though now we wish to emphasize the population case: If we call now

$$F(\mu) := \frac{1}{2} \int_{\mathcal{W}_2(\mathbb{R}^q)} W_2^2(\mu, m) \Gamma(dm) \quad (5.5)$$

the functional minimized by a 2-Wasserstein barycenter, then we (formally at least) expect

$$F'(\mu)(x) = - \int_{\mathcal{W}_2(\mathbb{R}^q)} (T_\mu^m - I) \Gamma(dm)(x). \quad (5.6)$$

Hence,  $(I - T_\mu^m)$  with  $m \sim \Gamma$ , is an unbiased estimator of  $F'(\mu)$ . This immediately suggests the stochastic descent sequence (5.4) introduced in [7], drawing inspiration from the classic SGD ideas [42].

Clearly  $\mu$  is a Karcher mean for  $\Gamma$  iff  $\|F'(\mu)\|_{L^2(\mu)} = 0$ . Just like for the GD sequence, the SGD sequence is typically expected to converge to stationary points, or Karcher means in the present setting, rather than to minimisers. Next theorem provides sufficient conditions for the SGDW sequence to a.s. converge to a Wasserstein barycenter, and is the main result of [7]. The following assumption on the steps  $\gamma_k$ , standard in the framework of SGD methods, is needed:

$$\sum_{k=1}^{\infty} \gamma_k^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \gamma_k = \infty. \quad (5.7)$$

**Theorem 5.7.** Suppose Assumptions 5.1 and 5.5, as well as conditions (5.7) hold. Furthermore, suppose that  $\Gamma$  admits a unique Karcher mean. Then, the SGD sequence  $\{\mu_k\}_k$  in equation (5.4) is a.s. convergent to the unique 2-Wasserstein barycenter  $\hat{\mu}$  of  $\Gamma$ . Moreover, we have  $\hat{\mu} \in K_\Gamma$ .

### 5.3. Batch stochastic gradient descent on Wasserstein space

We briefly recall how the variance of the SGDW sequence can be reduced by using batches:

**Definition 5.8.** Let  $\mu_0 \in K_\Gamma$ ,  $m_k^i \stackrel{\text{iid}}{\sim} \Gamma$ , and  $\gamma_k > 0$  for  $k \geq 0$  and  $i = 1, \dots, S_k$ . The batch stochastic gradient descent (BSGD) sequence is given by

$$\mu_{k+1} := \left[ (1 - \gamma_k)I + \gamma_k \frac{1}{S_k} \sum_{i=1}^{S_k} T_{\mu_k}^{m_k^i} \right] (\mu_k). \quad (5.8)$$

The following two results, extracted from [7], justify the above definition: The first result states that this sequence is still converging, while the second one states that batches help *reducing noise*:

**Proposition 5.9.** Under the assumptions of Theorem 5.7 the BSGD sequence  $\{\mu_t\}_{t \geq 0}$  in equation (5.8) converges a.s. to the 2-Wasserstein barycenter of  $\Gamma$ .

**Proposition 5.10.** The batch estimator for  $F'(\mu)$  of batch size  $S$ , given by  $-\frac{1}{S} \sum_{i=1}^S (T_\mu^{m_i} - I)$ , has integrated variance

$$\mathbb{V}[-\frac{1}{S} \sum_{i=1}^S (T_\mu^{m_i} - I)] := \int \text{Var}_{(m_i) \sim \Pi^{\otimes S}} \left[ \frac{1}{S} \sum_{i=1}^S (T_\mu^{m_i}(x) - x) \right] \mu(dx) = \mathcal{O}(\frac{1}{S}),$$

i.e.  $\mathbb{V}[-\frac{1}{S} \sum_{i=1}^S (T_{\mu}^{m_i} - I)]$  decreasing linearly in the batch size.

#### 5.4. Computation of the BWB

It is immediate to deduce a simple methodology based on the SGDW algorithm, to compute the BWB estimator for general (finitely or infinitely supported) posterior laws  $\Pi_n \in \mathcal{W}_2(\mathcal{W}_{2,ac}(\mathbb{R}^q))$ .

We make the practical assumption that we are capable of generating independent models  $m_i$  from the posteriors  $\Pi_n$  (in the parametric setting, this can be done through efficient Markov Chain Monte Carlo (MCMC) techniques [6, 11, 25] or transport sampling procedures [20, 27, 35, 39]). On the theoretical side, we assume conditions 5.1 and 5.5 are satisfied by  $\Gamma = \Pi$ , the prior law on models, implying the posterior  $\Pi_n$  a.s. satisfies those conditions for all  $n$  too.

The proposed method can be sketched as follows:

1. Given a prior on models  $\Pi$  and data  $x_1, \dots, x_n$ , sample  $\mu_0^{(n)} \sim \Pi_n$  and set  $k = 0$ .
2. Sample  $m_k^{(n)}$  independent from  $\mu_0^{(n)}, m_0^{(n)}, \dots, m_{k-1}^{(n)}$
3. Set

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

4. Increase  $k$  by 1 and go to (2).

The algorithm can be run until  $k \in \mathbb{N}$  large enough such that the squared 2-Wasserstein distance

$$W_2^2(\mu_{k+1}^{(n)}, \mu_k^{(n)}) = \gamma_k^2 \int_{\mathbb{R}^q} |x - T_{\mu_k^{(n)}}^{m_k^{(n)}}(x)|^2 d\mu_k^{(n)}(x)$$

between  $\mu_{k+1}^{(n)}$  and  $\mu_k^{(n)}$  is repeatedly smaller than some given positive threshold.

Moreover, under the assumptions of Theorem 4.24,  $\Pi$  is consistent in the  $W_2$ -topology at the law  $m_0$  of the data  $x_1, \dots, x_0$  and then, for some (random) large enough  $n, k \in \mathbb{N}$  and given  $\varepsilon > 0$  one has  $W_2(\mu_k^{(n)}, m_0) \leq W_2(\mu_k^{(n)}, \hat{m}_2^n) + W_2(\hat{m}_2^n, m_0) \leq \varepsilon$ , where  $\hat{m}_2^n$  is the BWB. Notice that, besides the sequential generation of a finite i.i.d. sequence  $\mu_0^{(n)}, m_0^{(n)}, \dots, m_k^{(n)} \sim \Pi_n$ , at each step  $k + 1$  one only needs to compose a new transport map  $\left[ (1 - \gamma_k)I + \gamma_k T_{\mu_k^{(n)}}^{m_k^{(n)}} \right]$  with the transport map pushing forward  $\mu_0^{(n)}$  to  $\mu_k^{(n)}$ , cumulatively constructed in the previous iterations. If an expression for each map  $T_{\mu_k^{(n)}}^{m_k^{(n)}}$  is available, this can easily be done (and stored), specifying (only) the values of the lastly computed map on a pre-fixed grid.

The batch version SGWD can be implemented in a similar way to compute the BWB estimator.

**Remark 5.11.** An alternative, natural approach would be to sample, for given  $n$  a fixed number  $k$  of realizations  $m_i \stackrel{\text{iid}}{\sim} \Pi_n$ ,  $i = 1, \dots, k$ , and compute, using the GDW algorithm, the Wasserstein barycenter  $\hat{m}_2^{(n,k)}$  of the (finitely supported) empirical measure

$$\Pi_n^{(k)} := \frac{1}{k} \sum_{i=1}^k \delta_{m_i} \in \mathcal{W}_2(\mathcal{W}_{2,ac}(\mathbb{R}^q)),$$

or 2-Wasserstein empirical barycenter of  $\Pi_n$  (see [10]). Indeed, by Varadarajan's theorem, conditionally on  $\Pi_n$ , a.s.  $\Pi_n^{(k)}$  converges weakly as  $k \rightarrow \infty$  to  $\Pi_n$ , and in the  $W_2$  metric as soon as  $\Pi_n \in \mathcal{W}_2(\mathcal{W}_{2,ac}(\mathbb{R}^q))$ . Since under our assumptions  $\Pi_n$  a.s. has a unique 2-Wasserstein barycenter  $\hat{m}_2^n$  and, by Theorem 3 of [33],  $\hat{m}_2^{(n,k)}$  converges

with respect to  $W_2$  a.s. as  $k \rightarrow \infty$  to it, we also get through this approach that

$$W_2(\hat{m}_2^{(n,k)}, m_0) \leq W_2(\hat{m}_2^{(n,k)}, \hat{m}_2^n) + W_2(\hat{m}_2^n, m_0) \leq \varepsilon$$

for some (random) large enough  $n, k \in \mathbb{N}$ . The clear disadvantage of this method is computational: besides generating  $k$  samples of  $\Pi_n$ , we need to additionally run a possibly large number of GDW steps to approximate  $\hat{m}_2^{(n,k)}$ , and we need to evaluate at each step  $k$  new transport maps (instead of one, for the SGDW). Moreover, if additional  $k'$  new samples from  $\Pi_n$  become available, we need to run the whole scheme again to take advantage of this new information. On the contrary, the online nature of the SGDW method allows one to refine the already computed estimator by only performing  $k'$  new steps of the algorithm.

## 6. NUMERICAL EXPERIMENTS

Before presenting the experimental validation of the proposed methods we give a brief presentation of the scatter-location family of distributions. Our experiments consider this family because the optimal transport maps between two laws in the scatter-location family can be described explicitly. This property facilitates the numerical computation/approximation of barycenters, as the various iterative algorithms described so far take a more amenable form. See [7] for further examples.

### 6.1. Location-Scatter family

We follow the setting of [4]: Given a fixed distribution  $\tilde{m} \in \mathcal{W}_{2,ac}(\mathbb{R}^q)$ , referred to as *generator*, the associated location-scatter 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. Note that  $\mathcal{F}(\tilde{m})$  is the multivariate normal family if  $\tilde{m}$  is the standard multivariate normal distribution.

The optimal map between two members  $m_1 = \mathcal{L}(A_1\tilde{x} + b_1)$  and  $m_2 = \mathcal{L}(A_2\tilde{x} + b_2)$  of  $\mathcal{F}(\tilde{m})$  is explicit, given by  $T_{m_1}^{m_2}(x) = A(x - b_1) + b_2$  where  $A = A_1^{-1}(A_1A_2^2A_1)^{1/2}A_1^{-1} \in \mathcal{M}_+^{q \times q}$ . This family of optimal maps contains the identity and is closed under convex combination.

If  $\Gamma$  is supported on  $\mathcal{F}(\tilde{m})$ , then its 2-Wasserstein barycenter  $\hat{m}$  belongs to  $\mathcal{F}(\tilde{m})$ . In fact, 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  $\int T_{\hat{m}}^m(x)\Gamma(dm) = x$ ,  $\hat{m}$ -almost surely, then we must have that  $\int A_m^m\Gamma(dm) = I$ , since clearly  $\hat{b} = \int b_m\Gamma(dm)$ , and as a consequence  $\hat{\Sigma} = \int (\hat{\Sigma}^{1/2}\Sigma_m\hat{\Sigma}^{1/2})^{1/2}\Gamma(dm)$ .

A stochastic gradient descent iteration, starting from a distribution  $\mu = \mathcal{L}(A_0\tilde{x} + b_0)$ , sampling some  $m = \mathcal{L}(A_m\tilde{x} + b_m) \sim \Gamma$ , 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 that  $T_0^{\gamma,m}(x) = ((1 - \gamma)I + \gamma A_\mu^m)(x - b_0) + \gamma b_m + (1 - \gamma)b_0$  with  $A_\mu^m := A_0^{-1}(A_0A_m^2A_0)^{1/2}A_0^{-1}$ . Then  $\nu$  has distribution

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

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

$$\begin{aligned} \Sigma_1 &= A_1^2 = [(1 - \gamma)A_0 + \gamma A_0^{-1}(A_0A_m^2A_0)^{1/2}][(1 - \gamma)A_0 + \gamma(A_0A_m^2A_0)^{1/2}A_0^{-1}] \\ &= A_0^{-1}[(1 - \gamma)A_0^2 + \gamma(A_0A_m^2A_0)^{1/2}][(1 - \gamma)A_0^2 + \gamma(A_0A_m^2A_0)^{1/2}]A_0^{-1} \\ &= A_0^{-1}[(1 - \gamma)A_0^2 + \gamma(A_0A_m^2A_0)^{1/2}]^2A_0^{-1}. \end{aligned}$$

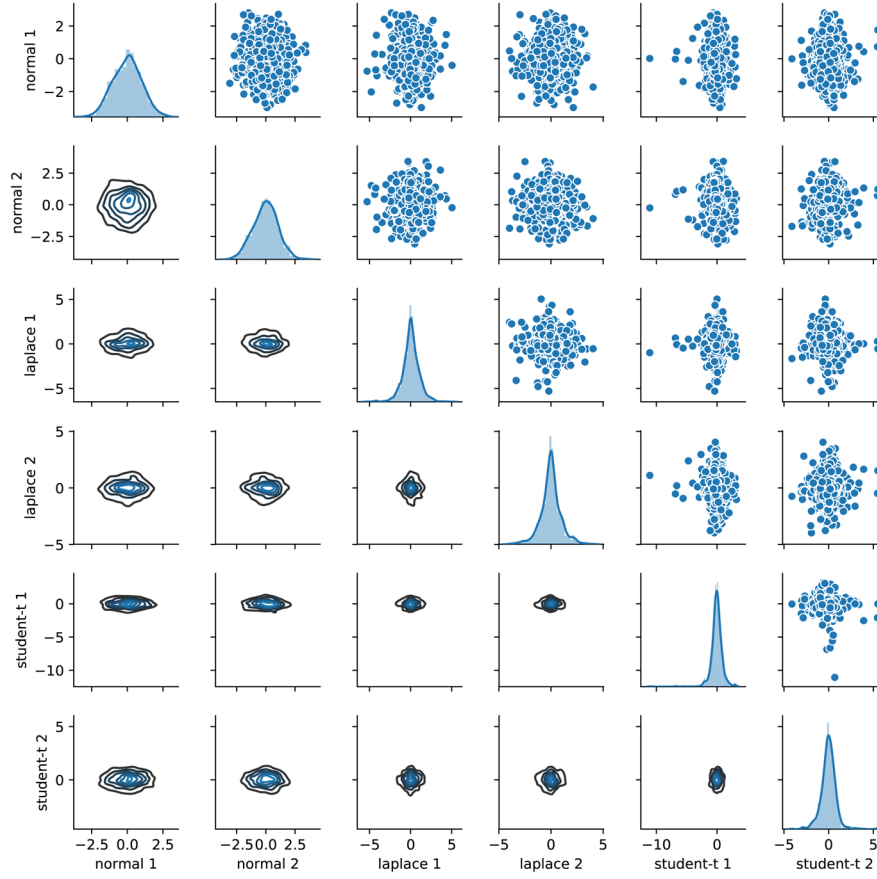


FIGURE 2. Samples from the univariate (diagonal) and bivariate (off-diagonal) marginals for 6 coordinates from the generator distribution  $\tilde{m}$ . The diagonal and lower triangular plots are smoothed histograms, whereas the upper-diagonal ones are collections of samples.

The batch stochastic gradient descent iteration is characterized by

$$b_1 = (1 - \gamma)b_0 + \frac{\gamma}{S} \sum_{i=1}^S b_{m^i}$$

$$A_1^2 = A_0^{-1}[(1 - \gamma)A_0^2 + \frac{\gamma}{S} \sum_{i=1}^S (A_0 A_{m^i}^2 A_0)^{1/2}]^2 A_0^{-1}.$$

## 6.2. Experiment

We considered a model within a location-scatter family (LS), with generator  $\tilde{m}$  on  $\mathbb{R}^{15}$  with independent coordinates, as follows:

- coordinates 1–5 are standard Normal distributions
- coordinates 6–10 are standard Laplace distributions, and
- coordinates 11–15 are standard Student's  $t$ -distributions (3 degrees of freedom).

Figure 2 shows samples (uni- and bi-variate marginals) from coordinates  $\{1, 2, 6, 7, 11, 12\}$  of  $\tilde{m}$ .

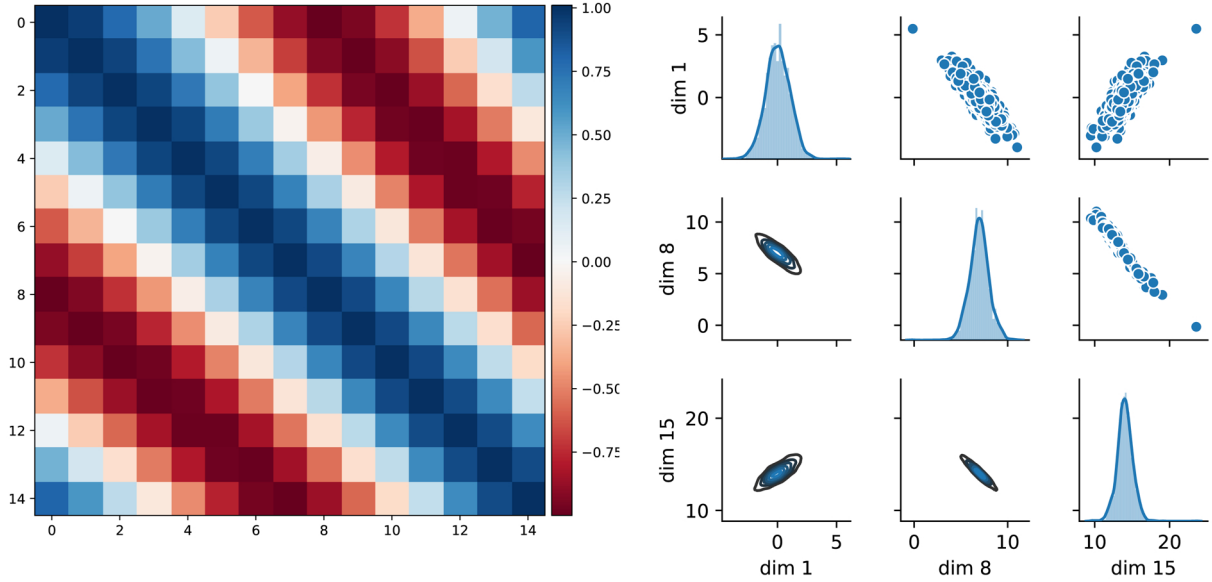


FIGURE 3. True model  $m_0$ : covariance matrix (left), and univariate and bivariate marginals for dimensions 1, 8 and 15 (right). Notice that some coordinates are positively or negatively correlated, and some may even be close to uncorrelated.

In this LS family, we chose the true model  $m_0$  with location vector  $b \in \mathbb{R}^{15}$  defined as  $b_i = i - 1$  for  $i = 1, \dots, 15$ , and scatter matrix  $A = \Sigma^{1/2}$  with  $\Sigma_{i,j} = K\left(\left(\frac{i-1}{14}\right)^{1.1}, \left(\frac{j-1}{14}\right)^{1.1}\right)$  for  $i, j = 1, \dots, 15$ <sup>2</sup>, with the covariance (kernel) function  $K(i, j) = \varepsilon \delta_{ij} + \sigma \cos(\omega(i - j))$ . Given the parameters  $\varepsilon, \sigma$  and  $\omega$ , the so constructed covariance matrix will be denoted  $\Sigma_{\varepsilon, \sigma, \omega}$ . We chose the parameters  $\varepsilon = 0.01$ ,  $\sigma = 1$  and  $\omega = 5.652 \approx 1.8\pi$  for  $m_0$ . Thus, under the true model  $m_0$  the coordinates can be negatively/positively correlated and there is also a coordinate-independent noise component due to the Kronecker delta  $\delta_{ij}$ . Figure 3 shows the covariance matrix and three coordinates of the *true* model  $m_0$ .

The model prior  $\Pi$  is the push-forward induced by a prior over the mean vector  $b$  and the parameters of the covariance  $\Sigma_{\varepsilon, \sigma, \omega}$ , chosen independently according to:

$$p(b, \Sigma_{\varepsilon, \sigma, \omega}) = \mathcal{N}(b|0, I) \text{Exp}(\varepsilon|20) \text{Exp}(\sigma|1) \text{Exp}(\omega^{-1}|15), \quad (6.1)$$

where  $\text{Exp}(\cdot|\lambda)$  is a exponential density with rate  $\lambda$ . Given  $n$  samples from the true model  $m_0$  (also referred to as *observations* or *data points*),  $k$  samples are produced from the posterior measure  $\Pi_n$  using Markov chain Monte Carlo (MCMC).

The numerical analysis presented in what follows focuses on the behavior of the BWB as a function of both the number of samples  $k$  and the number of data points  $n$ .

### 6.2.1. Numerical consistency of the empirical posterior

We first validate the empirical measure  $\Pi_n^{(k)}$  as a consistent sample version of the true posterior under the  $W_2$  distance, that is, we check that  $W_2(\Pi_n^{(k)}, \delta_{m_0}) \rightarrow W_2(\Pi_n, \delta_{m_0})$  for large  $k$ . We estimate  $W_2(\Pi_n^{(k)}, \delta_{m_0})$  10 times for each combination of the (number of) observations  $n$  and samples  $k$  in the following sets

- $k \in \{1, 5, 10, 20, 50, 100, 200, 500, 1000\}$

<sup>2</sup>We chose  $\left(\frac{j-1}{14}\right)^{1.1}$  for  $j = 1, \dots, 15$  because this defines a non-uniform grid over  $[0, 1]$ .

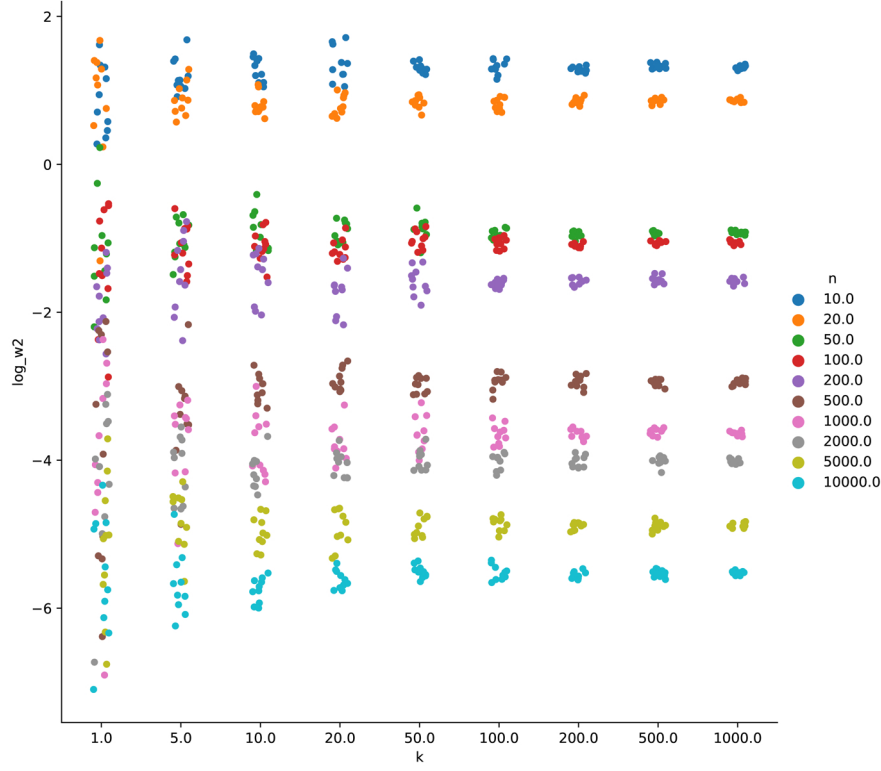


FIGURE 4. Wasserstein distance between the empirical measure  $\Pi_n^{(k)}$  and  $\delta_{m_0}$  in logarithmic scale for different number of observations  $n$  (color coded) and samples  $k$  ( $x$ -axis). For each pair  $(n, k)$ , 10 estimates of  $W_2(\Pi_n^{(k)}, \delta_{m_0})$  are shown.

TABLE 1. Standard deviation of  $W_2^2(\Pi_n^{(k)}, \delta_{m_0})$ , using 10 simulations, for different values of observations  $n$  and samples  $k$ .

$n \setminus k$	1	5	10	20	50	100	200	500	1000
10	1.2506	0.8681	0.5880	0.9690	0.2354	0.3440	0.1253	0.1330	0.0972
20	1.5168	0.5691	0.3524	0.3182	0.1850	0.1841	0.1049	0.0811	0.0509
50	0.3479	0.0948	0.1275	0.0572	0.0623	0.0229	0.0157	0.0085	0.0092
100	0.2003	0.1092	0.0712	0.0469	0.0431	0.0254	0.0087	0.0079	0.0084
200	0.0749	0.1249	0.0717	0.0533	0.0393	0.0101	0.0092	0.0109	0.0072
500	0.0478	0.0285	0.0093	0.0086	0.0053	0.0056	0.0045	0.0023	0.0022
1000	0.0299	0.0113	0.0113	0.0064	0.0067	0.0036	0.0016	0.0012	0.0007
2000	0.0145	0.0071	0.0040	0.0031	0.0027	0.0019	0.0014	0.0011	0.0006
5000	0.0072	0.0031	0.0015	0.0018	0.0010	0.0007	0.0004	0.0005	0.0002
10000	0.0038	0.0020	0.0005	0.0005	0.0004	0.0004	0.0002	0.0002	0.0001

- $n \in \{10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000\}$ .

Figure 4 shows the 10 estimates of  $W_2(\Pi_n^{(k)}, \delta_{m_0})$  for different values of  $k$  (in the  $x$ -axis) and of  $n$  (color coded). Notice how the estimates become more concentrated for larger  $k$  and that the Wasserstein distance between the empirical measure  $\Pi_n^{(k)}$  and the true model  $m_0$  decreases for larger  $n$ . Additionally, Table 1 shows that the standard deviation of the 10 estimates of  $W_2(\Pi_n^{(k)}, \delta_{m_0})$  decreases as either  $n$  or  $k$  increases.



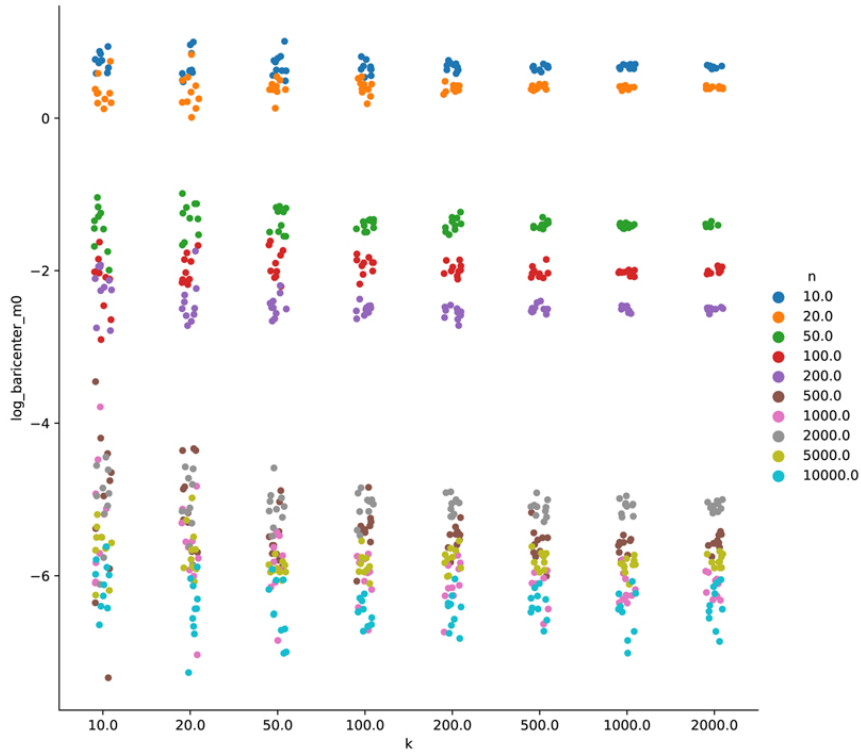


FIGURE 5.  $W_2$  distance between the empirical barycenters  $\hat{m}_2^{(n,k)}$  and the true model  $m_0$  in logarithmic scale for different number of observations  $n$  (color coded) and samples  $k$  ( $x$ -axis). For each pair  $(n, k)$ , 10 estimates of  $W_2(\hat{m}_2^{(n,k)}, m_0)$  are shown.

### 6.2.2. Wasserstein distance between the empirical barycenter and the true model

For each empirical posterior  $\Pi_n^{(k)}$ , we computed the empirical Wasserstein barycenter  $\hat{m}_2^{(n,k)}$  as suggested in Remark 5.11. Thus, we used the iterative GDW procedure in equation (5.2), namely the (deterministic) gradient descent method, and repeated this calculation 10 times. As a stopping criterion for gradient descent, we considered the relative variation of the  $W_2$  cost; the computation was terminated when this cost fell below  $10^{-4}$ . Figure 5 shows the  $W_2$  distances between the so computed barycenters and the true model, while Table 2 shows the average distance for each pair  $(n, k)$ . Notice that, in general, both the average and standard deviation of the barycenters decrease as either  $n$  or  $k$  increases, yet for large values (*e.g.*,  $n = 2000, 5000$ ) numerical issues appear.

### 6.2.3. Distance between the empirical barycenter and the Bayesian model average

We then compared the empirical Wasserstein barycenters  $\hat{m}_2^{(n,k)}$  to the standard Bayesian model averages, denoted here  $\bar{m}^{(n,k)}$ , in terms of their distance to the true model  $m_0$ , for  $n = 1000$  observations. To that end, we estimated the  $W_2$  distances via empirical approximations with 1000 samples for each model based on [21]. We simulated this procedure 10 times for  $k \in \{10, 20, 50, 100, 200, 500, 1000\}$ . Figure 6 shows the sample average and variance of the  $W_2$  distances of the Wasserstein barycenters and Bayesian model averages. The empirical barycenter is seen to be closer to the true model than the model average, regardless of the number of MCMC samples  $k$ .

TABLE 2. Sample average of  $W_2^2(\hat{m}_2^{(n,k)}, m_0)$ , using 10 simulations, for different values of observations  $n$  and samples  $k$ .

n / k	10	20	50	100	200	500	1000	2000
10	2.1294	2.0139	2.0384	1.9396	1.9608	1.9411	1.9699	1.9548
20	1.4382	1.4498	1.4826	1.4973	1.4785	1.4953	1.4955	1.4914
50	0.2455	0.2759	0.2639	0.2468	0.2499	0.2483	0.2443	0.2454
100	0.1211	0.1387	0.1509	0.1458	0.1379	0.1328	0.1318	0.1349
200	0.1116	0.0922	0.0859	0.0817	0.0777	0.0824	0.0820	0.0819
500	0.0094	0.0077	0.0043	0.0047	0.0041	0.0038	0.0037	0.0039
1000	0.0068	0.0039	0.0031	0.0025	0.0023	0.0022	0.0021	0.0021
2000	0.0072	0.0066	0.0063	0.0062	0.0063	0.0060	0.0062	0.0062
5000	0.0037	0.0037	0.0028	0.0029	0.0031	0.0031	0.0028	0.0030
10000	0.0023	0.0017	0.0017	0.0015	0.0016	0.0017	0.0016	0.0017

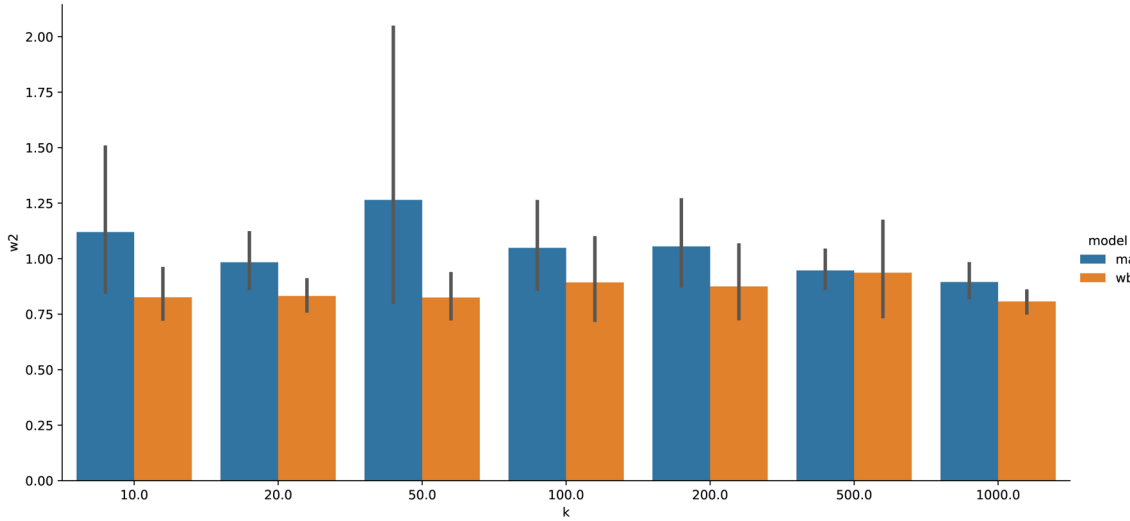


FIGURE 6. Averages (bars) and standard deviations (vertical lines) of  $W_2^2(\hat{m}_n^{(k)}, m_0)$  denoted as **wb** in orange, and  $W_2^2(\tilde{m}_n^{(k)}, m_0)$  denoted as **ma** in blue, for  $n = 1000$  and different numbers of samples  $k$ . We considered 10 simulations for each  $k$ .

#### 6.2.4. Computation of the Wasserstein barycenter with batch-SGDW

Lastly, we compared the empirical barycenters  $\hat{m}_2^{(m,k)}$  to the barycenter obtained by the batch SGDW method in equation (5.8) with different batch sizes. Here, we shall denote the latter by  $\hat{m}_2^{(n,t,s)}$  with  $n$  the number of observations,  $t$  the steps of the algorithm, and  $s$  the batch size. Fig. 7 shows the evolution of the  $W_2^2$  distance between the stochastic gradient descent sequences and the true model  $m_0$  for  $n \in \{10, 20, 50, 100, 200, 500, 1000\}$  observations and batches of sizes  $s \in \{1, 15\}$ , with step-size  $\gamma_t = \frac{1}{t}$  for  $t = 1, \dots, 200$ . Notice from Fig. 7 that the larger the batch, the more concentrated the trajectories of  $\hat{m}_{n,s}$  become. Additionally, Table 3 summarizes the means of the distance  $W_2^2$  to the true model  $m_0$ , using the sequences after  $t = 100$  against the empirical estimator using all the simulations with  $k \geq 100$ . Finally, Table 4 shows the standard deviation of the distance  $W_2^2$  to the true model  $m_0$ , which can be seen to decrease as the batch size grows. Critically, we observe that for batch sizes  $s \geq 5$  the stochastic estimation was *better* than its empirical counterpart, *i.e.*, it had smaller variance with similar (or even smaller) bias. This is noteworthy given the fact that computing our Wasserstein barycenter estimator via the batch stochastic gradient descent method is computationally less demanding than computing it via the empirical method.

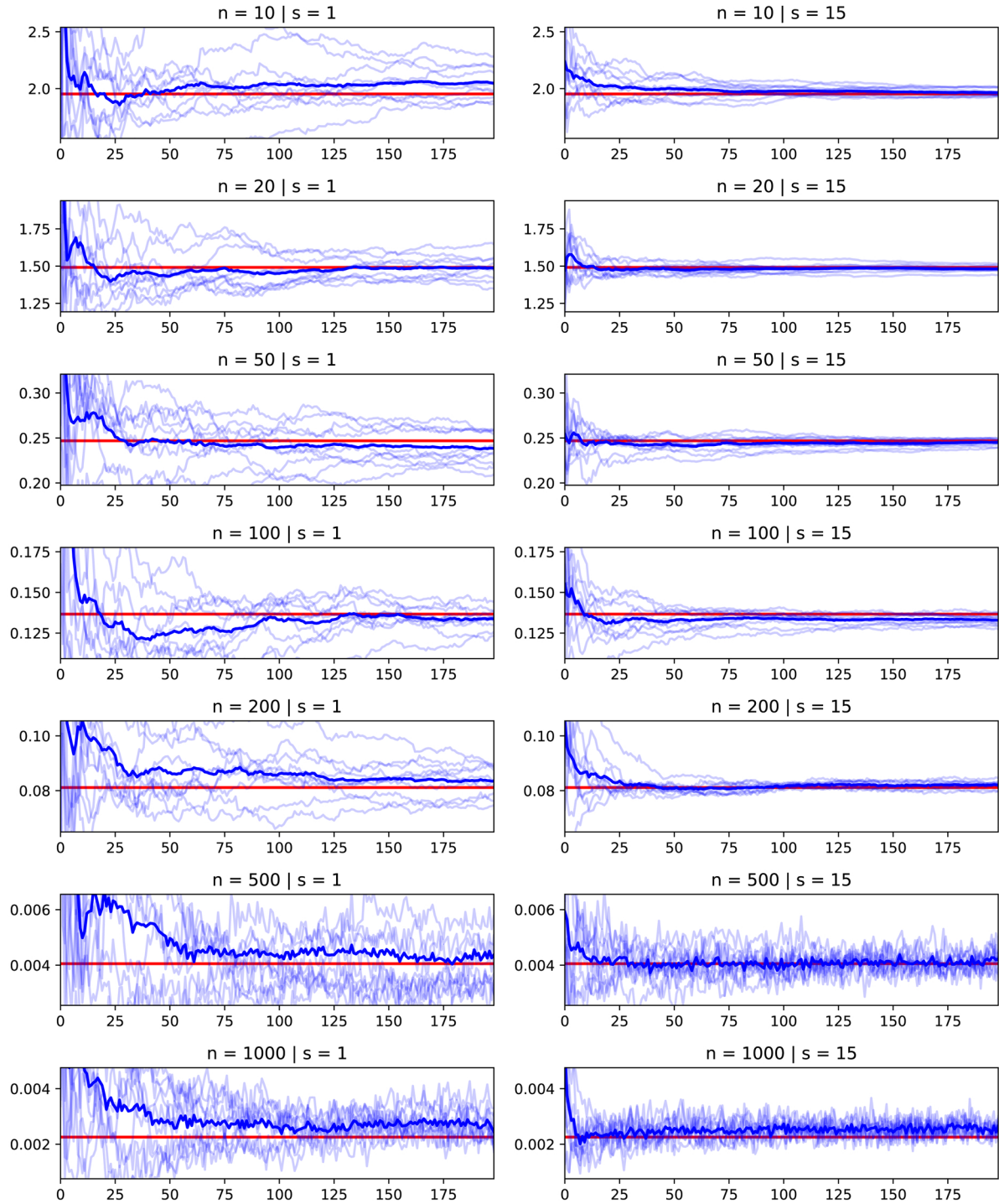


FIGURE 7. Evolution of the  $W_2^2$  cost for 10 realizations of the SGD sequence computing the BWB and their mean (blue), versus an empirical barycenter estimator (red), for  $n = 10, 20, 50, 100, 200, 500, 1000$  and batches sizes  $s = 1, 15$ .

TABLE 3. Means of  $W_2^2$  of the stochastic gradient estimations (using the sequences with  $t \geq 100$ ) and that of the empirical estimator (using the simulations with  $k \geq 100$ ), across different combinations of observations  $n$  and batch size  $s$ .

n / s	1	2	5	10	15	20	empirical
10	2.0421	2.0091	1.9549	1.9721	1.9732	1.9712	1.9532
20	1.4819	1.4868	1.5100	1.4852	1.4840	1.4891	1.4916
50	0.2406	0.2512	0.2465	0.2427	0.2444	0.2460	0.2469
100	0.1340	0.1392	0.1340	0.1349	0.1334	0.1338	0.1366
200	0.0843	0.0811	0.0819	0.0807	0.0820	0.0819	0.0811
500	0.0044	0.0042	0.0039	0.0039	0.0041	0.0040	0.0041

TABLE 4. Std. deviations of  $W_2^2$  of the stochastic gradient estimations (using the sequences with  $t \geq 100$ ) and that of empirical estimator (using the simulations with  $k \geq 100$ ), across different combinations of observations  $n$  and batch size  $s$ .

n / s	1	2	5	10	15	20	empirical
10	0.1836	0.1071	0.0526	0.0474	0.0397	0.0232	0.0916
20	0.0751	0.0565	0.0553	0.0189	0.0253	0.0186	0.0790
50	0.0210	0.0174	0.0072	0.0084	0.0050	0.0039	0.0138
100	0.0102	0.0076	0.0049	0.0048	0.0035	0.0023	0.0112
200	0.0074	0.0045	0.0021	0.0035	0.0013	0.0017	0.0047
500	0.0016	0.0007	0.0005	0.0004	0.0004	0.0004	0.0009
1000	0.0005	0.0006	0.0004	0.0004	0.0003	0.0003	0.0005

### 6.2.5. Conclusion

Based on this illustrative numerical example, we can conclude:

- the empirical posterior constructed using MCMC is consistent under the  $W_2$  distance and therefore it can be relied upon to compute Wasserstein barycenters,
- the empirical Wasserstein barycenter estimator tends to converge faster (and with lower variance) to the true model than the empirical Bayesian model average,
- computing the population Wasserstein barycenter estimator via batch stochastic gradient descent is promising as an alternative to computing the empirical barycenter (*i.e.*, to applying the deterministic gradient descent method to a finitely-sampled posterior).

*Acknowledgements.* We would like to thank two anonymous referees for suggesting us some of the parametric examples we have analyzed, and for their valuable questions and comments, which allowed us to clarify the presentation of the paper and refine our results. This work was partially funded by 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).

## APPENDIX A. BAYES ESTIMATORS AS GENERALIZED MODEL AVERAGES

We prove here Proposition 2.5. For notational simplicity we omit the subscripts of estimators and normalizing constants.

**i) Squared  $L_2$ -distance:** By Fubini's theorem, minimizing  $R_L(\bar{m}|D)$  in this case amounts to minimize

$$\bar{m} \mapsto \frac{1}{2} \int_{\mathcal{X}} \left\{ \int_{\mathcal{M}} (m(x) - \bar{m}(x))^2 \Pi_n(dm) \right\} \lambda(dx),$$

over the set of densities. But the optimal value cannot be better than if we minimize pointwise the term  $\bar{m}(x)$  inside the curly brackets, obtaining the candidate

$$\hat{m}_{BMA}(x) = \int_{\mathcal{M}} m(x) \Pi_n(dm) = \mathbb{E}[m](x).$$

As this pointwise minimizer is already a probability density, we conclude.

**ii) Squared Hellinger distance:** Writing

$$H^2(m, \bar{m}) = \frac{1}{2} \int_{\mathcal{X}} \left( \sqrt{m(x)} - \sqrt{\bar{m}(x)} \right)^2 \lambda(dx) = 1 - \int_{\mathcal{X}} \sqrt{m(x)\bar{m}(x)} \lambda(dx),$$

we see that optimizing the associated Bayes risk amounts to maximizing over  $\bar{m} \in L^1(\mathcal{X}, \lambda)$  the concave functional  $\bar{m} \mapsto \int_{\mathcal{X}} \sqrt{\bar{m}(x)} f(x) \lambda(dx)$  with  $\sqrt{y} = -\infty$  for  $y < 0$  and  $f(x) = \int_{\mathcal{M}} \sqrt{m(x)} \Pi_n(dm)$ , under the constraint  $\int_{\mathcal{X}} \bar{m}(x) \lambda(dx) = 1$ . Notice by the Cauchy-Schwarz inequality that this functional is finite if (and only if)  $\bar{m} \geq 0$ ,  $\lambda$ -a.e. Hence, introducing  $\gamma$  a real Lagrange multiplier for the constraint, we need to find a critical point  $\bar{m} \geq 0$  of the concave functional

$$\bar{m} \mapsto \int_{\mathcal{X}} \sqrt{\bar{m}(x)} f(x) \lambda(dx) + \gamma \left( \int_{\mathcal{X}} \bar{m}(x) \lambda(dx) - 1 \right),$$

which must thus be a maximum. Again, as in i), we cannot do better in this case than if we maximize for each  $x \in \mathcal{X}$  the concave functional  $y \mapsto \sqrt{y} f(x) + \gamma y$  over  $y \geq 0$ . Finding for each  $x$  the critical point  $y$  in terms of  $\gamma$  and integrating then w.r.t.  $\lambda(dx)$  to get rid of  $\gamma$ , we find the extremum of  $R_{H^2}(\bar{m}|D)$  is attained when  $\bar{m}$  equals the Bayesian *square* model average:

$$\begin{aligned} \hat{m}(x) &= \frac{1}{Z} \left( \int_{\mathcal{M}} \sqrt{m(x)} \Pi_n(dm) \right)^2 \text{ with} \\ Z &= \int_{\mathcal{X}} \left( \int_{\mathcal{M}} \sqrt{m(x)} \Pi_n(dm) \right)^2 \lambda(dx). \end{aligned}$$

**iii) Forward Kullback-Leibler divergence:** the loss function  $L(m, \bar{m})$  is now

$$D_{KL}(m||\bar{m}) = \int_{\mathcal{X}} m(x) \ln \frac{m(x)}{\bar{m}(x)} \lambda(dx),$$

and so the associated Bayes risk can be written as

$$\begin{aligned} R_{D_{KL}}(\bar{m}|D) &= \int_{\mathcal{M}} \int_{\mathcal{X}} m(x) \ln \frac{m(x)}{\bar{m}(x)} \lambda(dx) \Pi_n(dm) \\ &= \int_{\mathcal{X}} \int_{\mathcal{M}} m(x) \ln m(x) \Pi_n(dm) \lambda(dx) - \int_{\mathcal{X}} \int_{\mathcal{M}} m(x) \Pi_n(dm) \ln \bar{m}(x) \lambda(dx) \\ &= C - \int_{\mathcal{X}} \mathbb{E}[m](x) \ln \bar{m}(x) \lambda(dx). \end{aligned}$$

Introducing the Boltzmann entropy of  $\mathbb{E}[m]$  and adjusting the constant  $C$  we get that

$$\begin{aligned} R_{D_{KL}}(\bar{m}|D) &= C' + \int_{\mathcal{X}} \mathbb{E}[m](x) \ln \mathbb{E}[m](x) \lambda(dx) - \int_{\mathcal{X}} \mathbb{E}[m](x) \ln \bar{m}(x) \lambda(dx) \\ &= C' + D_{KL}(\mathbb{E}[m]||\bar{m}), \end{aligned}$$

so the optimizer of  $R_{D_{KL}}(\bar{m}|D)$  is the Bayesian model average.

iv) **Reverse Kullback-Leibler divergence:** since  $D_{RKL}(m||\bar{m}) = D_{KL}(\bar{m}||m)$ , we have

$$\begin{aligned} R_{D_{RKL}}(\bar{m}|D) &= \int_{\mathcal{M}} \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{m(x)} \lambda(dx) \Pi_n(dm) \\ &= \int_{\mathcal{X}} \bar{m}(x) \ln \bar{m}(x) \lambda(dx) - \int_{\mathcal{X}} \bar{m}(x) \int_{\mathcal{M}} \ln m(x) \Pi_n(dm) \lambda(dx) \\ &= \int_{\mathcal{X}} \bar{m}(x) \ln \bar{m}(x) \lambda(dx) - \int_{\mathcal{X}} \bar{m}(x) \ln \exp \mathbb{E}[\ln m(x)] \lambda(dx) \\ &= \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{\exp \mathbb{E}[\ln m(x)]} \lambda(dx). \end{aligned}$$

Denote by  $Z$  the normalization constant so that  $\frac{1}{Z} \int_{\mathcal{X}} \exp \mathbb{E}[\ln m](x) \lambda(dx) = 1$ . Then,

$$\begin{aligned} R_{D_{RKL}}(\bar{m}|D) + \ln Z &= \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{\exp \mathbb{E}[\ln m(x)]} \lambda(dx) + \int_{\mathcal{X}} \bar{m}(x) \ln Z \lambda(dx) \\ &= \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{\frac{1}{Z} \exp \mathbb{E}[\ln m(x)]} \lambda(dx) \\ &= D_{RKL} \left( \frac{1}{Z} \exp \mathbb{E}[\ln m(x)] || \bar{m} \right). \end{aligned}$$

Therefore, the extremum of  $R_{D_{RKL}}(\bar{m}|D)$  is attained when the last expression vanishes, in other words when  $\bar{m}$  is the Bayesian *exponential* model average given by

$$\hat{m}(x) = \frac{1}{Z} \exp \int_{\mathcal{M}} \ln m(x) \Pi_n(dm).$$

## APPENDIX B. WASSERSTEIN BARYCENTERS

*Proof of Theorem 3.7.* Assume  $\Gamma \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ . Then  $B_V := \inf_{\nu \in \mathcal{M}} V_p(\nu)$  is finite. Now, let  $\{\nu_n\} \subseteq \mathcal{M}$  be such that

$$\int_{\mathcal{W}_p(\mathcal{X})} W_p(\nu_n, m)^p \Gamma(dm) \searrow B_V \text{ as } n \rightarrow \infty.$$

For  $n$  large enough we have

$$W_p \left( \nu_n, \int_{\mathcal{W}_p(\mathcal{X})} m \Gamma(dm) \right)^p \leq \int_{\mathcal{W}_p(\mathcal{X})} W_p(\nu_n, m)^p \Gamma(dm) \leq B_V + 1 =: K,$$

by convexity of optimal transport costs. From this we derive that (for every  $x$ )

$$\sup_n \int_{\mathcal{X}} d(x, y)^p \nu_n(dy) < \infty.$$

By Markov inequality this shows, for each  $\varepsilon > 0$ , that there is  $\ell$  large enough such that  $\sup_n \nu_n(\{y \in \mathcal{X} : d(x, y) > \ell\}) \leq \varepsilon$ . As explained in [33], the assumptions on  $\mathcal{X}$  imply that  $\{y \in \mathcal{X} : d(x, y) \leq \ell\}$  is compact (Hopf-Rinow theorem), and so we deduce the tightness of  $\{\nu_n\}$ . By Prokhorov theorem, up to selection of a subsequence, there exists  $\nu \in \mathcal{M}$  which is its weak limit. By Fatou's lemma:

$$B_V = \lim \int W_p(\nu_n, m)^p \Gamma(dm) \geq \int \liminf W_p(\nu_n, m)^p \Gamma(dm) \geq \int W_p(\nu, m)^p \Gamma(dm),$$

hence  $\nu$  is a  $p$ -Wasserstein barycenter. For the converse implication, see Remark 3.6.  $\square$

## APPENDIX C. A CONDITION FOR EXISTENCE OF BARYCENTERS OF BAYESIAN POSTERIORIS

We last provide a general condition on the prior  $\Pi$  ensuring that, for given  $p \geq 1$ ,

$$\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X})) \text{ for all possible data points } (x_1, \dots, x_n) \in \mathcal{X}^n \text{ and all } n.$$

**Definition C.1.** We say that  $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$  is *p-integrable after updates* if it satisfies the conditions

1. For all  $x \in \mathcal{X}, \ell > 1$ :

$$\int_{\mathcal{M}} m(x)^\ell \Pi(dm) < \infty.$$

2. For some  $y \in \mathcal{X}$  and  $\varepsilon > 0$ :

$$\int_{\mathcal{M}} \left( \int_{\mathcal{X}} d(y, z)^p m(dz) \right)^{1+\varepsilon} \Pi(dm) < \infty.$$

**Remark C.2.** Condition (2) above can be intuitively denoted as  $\Pi \in \mathcal{W}_{p+}(\mathcal{W}_p(\mathcal{X}))$ . Of course, one has  $\mathcal{W}_{p+}(\mathcal{W}_p(\mathcal{X})) \subseteq \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ .

**Remark C.3.** If  $\Pi \in \mathcal{P}(\mathcal{W}_{p,ac}(\mathcal{X}))$  has finite support, then Conditions (1) and (2) are satisfied. On the other hand, if  $\Pi$  is supported on a scatter-location family (see Section 6.1) containing one element with a bounded density and a finite  $p$ -moment, then Conditions (1) and (2) are fulfilled if for example  $\text{supp}(\Pi)$  is tight. Conditions (1) and (2) are also satisfied in Example 4.2.

**Lemma C.4.** Suppose that  $\Pi$  is *p-integrable after updates*. Then, for each  $x \in \mathcal{X}$ , the measure

$$\tilde{\Pi}(dm) := \frac{m(x)\Pi(dm)}{\int_{\mathcal{M}} \bar{m}(x)\Pi(d\bar{m})},$$

is also *p-integrable after updates*.

*Proof.* We verify Property (1) first. Let  $\ell > 1$  and  $\bar{x} \in \mathcal{X}$  given. Then

$$\int_{\mathcal{M}} m(\bar{x})^\ell m(x) \Pi(dm) \leq \left( \int_{\mathcal{M}} m(x)^s \Pi(dm) \right)^{1/s} \left( \int_{\mathcal{M}} m(\bar{x})^{t\ell} \Pi(dm) \right)^{1/t},$$

with  $s, t$  conjugate Hölder exponents. This is finite since  $\Pi$  fulfils Property (1).

We now establish Property (2). Let  $y \in \mathcal{X}, \varepsilon > 0$ . Then

$$\begin{aligned} & \int_{\mathcal{M}} \left( \int_{\mathcal{X}} d(y, z)^p m(dz) \right)^{1+\varepsilon} m(x) \Pi(dm) \\ & \leq \left( \int_{\mathcal{M}} m(x)^s \Pi(dm) \right)^{1/s} \left( \int_{\mathcal{M}} \left( \int_{\mathcal{X}} d(y, z)^p m(dz) \right)^{(1+\varepsilon)t} \Pi(dm) \right)^{1/t}. \end{aligned}$$

The first term in the r.h.s. is finite by Property (1). The second term in the r.h.s. is finite by Property (2), if we take  $\varepsilon$  small enough and  $t$  close enough to 1. We conclude.  $\square$

**Lemma C.5.** Suppose that  $\Pi$  is *p-integrable after updates*. Then for all  $n \in \mathbb{N}$  and  $(x_1, \dots, x_n) \in \mathcal{X}^n$ , the posterior  $\Pi_n$  is also *p-integrable after updates*.

*Proof.* By Lemma C.4, we obtain that  $\Pi_1$  is integrable after updates. By induction, suppose  $\Pi_{n-1}$  has this property. Then as

$$\Pi_n(dm) = \frac{m(x_n)\Pi_{n-1}(dm)}{\int_{\mathcal{M}} \bar{m}(x_n)\Pi_{n-1}(d\bar{m})},$$



we likewise conclude that  $\Pi_n$  is  $p$ -integrable after updates.  $\square$

## REFERENCES

- [1] M. Agueh and G. Carlier, Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43** (2011) 904–924.
- [2] J. Alschuler and E. Boix-Adsera, Wasserstein barycenters are NP-hard to compute. Preprint [arXiv:2101.01100] (2021).
- [3] P.C. Álvarez-Esteban, E. del Barrio, J.A. Cuesta-Albertos and C. Matrán, A fixed-point approach to barycenters in Wasserstein space. *J. Math. Anal. Appl.* **441** (2016) 744–762.
- [4] P.C. Álvarez-Esteban, E. del Barrio, J.A. Cuesta-Albertos and C. Matrán, Wide Consensus aggregation in the Wasserstein Space. Application to location-scatter families. *Bernoulli* **24** (2018) 3147–3179.
- [5] L. Ambrosio, N. Gigli and G. Savaré, Gradient flows in metric spaces and in the space of probability measures. Lectures in Mathematics ETH Zürich, 2nd edn., Birkhäuser Verlag, Basel (2008).
- [6] C. Andrieu, N. De Freitas, A. Doucet and M.I. Jordan, An introduction to MCMC for machine learning. *Mach. Learn.* **50** (2003) 5–43.
- [7] J. Backhoff-Veraguas, J. Fontbona, G. Rios and F. Tobar, Stochastic gradient descent in Wasserstein space. Preprint [arXiv:2201.04232] (2022).
- [8] J.O. Berger, Statistical decision theory and Bayesian analysis. Springer Science & Business Media (2013).
- [9] R.H. Berk *et al.*, Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Stat.* **37** (1966) 51–58.
- [10] J. Bigot and T. Klein, Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probab. Stat.* **22** (2018) 35–57.
- [11] S. Brooks, A. Gelman, G. Jones and X.-L. Meng, Handbook of Markov chain Monte Carlo. CRC Press (2011).
- [12] E. Cazelles, F. Tobar and J. Fontbona, A novel notion of barycenter for probability distributions based on optimal weak mass transport, *2021 Conference on Neural Information Processing Systems NeurIPS* (2021) [arXiv:2102.13380].
- [13] S. Chewi, T. Maunu, P. Rigollet and A.J. Stromme, Gradient descent algorithms for Bures-Wasserstein barycenters, in Conference on Learning Theory, PMLR (2020) 1276–1304.
- [14] J. Cuesta-Albertos, L. Ruschendorf and A. Tuero-Díaz, Optimal coupling of multivariate distributions and stochastic processes. *J. Multivariate Anal.* **46** (1993) 335–361.
- [15] M. Cuturi and A. Doucet, Fast computation of Wasserstein barycenters, in International Conference on Machine Learning (2014) 685–693.
- [16] M. Cuturi and G. Peyré, A smoothed dual approach for variational Wasserstein problems. *SIAM J. Imag. Sci.* **9** (2016) 320–343.
- [17] P. Diaconis and D. Freedman, On the consistency of Bayes estimates. *Ann. Stat.* (1986) 1–26.
- [18] P. Dognin, I. Melnyk, Y. Mroueh, J. Ross, C.D. Santos and T. Sercu, Wasserstein barycenter model ensembling. Preprint [arXiv:1902.04999] (2019).
- [19] D.C. Dowson and B.V. Landau, The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.* **12** (1982) 450–455.
- [20] T.A. El Moselhy and Y.M. Marzouk, Bayesian inference with optimal maps. *J. Comput. Phys.* **231** (2012) 7815–7850.
- [21] R. Flamary and N. Courty, POT Python Optimal Transport library (2017).
- [22] M. Fréchet, Les éléments aléatoires de nature quelconque dans un espace distancié, in Annales de l’institut Henri Poincaré **10** (1948) 215–310.
- [23] S. Ghosal and A. van der Vaart, , vol. 44 of *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press (2017).
- [24] C.R. Givens and R.M. Shortt, A class of Wasserstein metrics for probability distributions. *Michigan Math. J.* **31** (1984) 231–240.
- [25] J. Goodman and J. Weare, Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci.* **5** (2010) 65–80.
- [26] M. Grendár and G. Judge, Asymptotic equivalence of empirical likelihood and Bayesian map. *Ann. Statist.* **37** (2009) 2445–2457.
- [27] S. Kim, D. Mesa, R. Ma and T.P. Coleman, Tractable fully Bayesian inference via convex optimization and optimal transport theory. Preprint [arXiv:1509.08582] (2015).
- [28] Y.-H. Kim and B. Pass, Wasserstein barycenters over Riemannian manifolds. *Adv. Math.* **307** (2017) 640–683.
- [29] B.J.K. Kleijn, Bayesian asymptotics under misspecification, Ph.D. thesis, Vrije Universiteit Amsterdam (2004).
- [30] B.J.K. Kleijn and A.W. Van der Vaart, The Bernstein-von-Mises theorem under misspecification. *Electr. J. Stat.* **6** (2012) 354–381.
- [31] A. Korotin, V. Egiazarian, L. Lingxiao and E. Burnaev, Wasserstein Iterative Networks for Barycenter Estimation. [arXiv:2201.12245] (2022).
- [32] J. Lacombe, J. Digne, N. Courty and N. Bonneel, Learning to generate Wasserstein barycenters. Preprint [arXiv:2102.12178] (2021).
- [33] T. Le Gouic and J.-M. Loubes, Existence and consistency of Wasserstein Barycenters. *Probab. Theory Related Fields* **168** (2017) 901–917.
- [34] A. Mallasto, A. Gerolin and H.Q. Minh, Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Inf. Geometry* (2021) 1–35.

- [35] Y. Marzouk, T. Moselhy, M. Parno and A. Spantini, Sampling via measure transport: An introduction. *Handbook of Uncertainty Quantification* (2016) 1–41.
- [36] P. Massart, Concentration Inequalities and Model Selection. Springer (2007).
- [37] K.P. Murphy, Machine learning: a probabilistic perspective. The MIT Press, Cambridge, MA (2012).
- [38] V.M. Panaretos and Y. Zemel, An invitation to statistics in Wasserstein space. SpringerBriefs in Probability and Mathematical Statistics, Springer, Cham (2020).
- [39] M. Parno, Transport maps for accelerated Bayesian computation, Ph.D. thesis, Massachusetts Institute of Technology (2015).
- [40] B. Pass, Optimal transportation with infinitely many marginals. *J. Funct. Anal.* **264** (2013) 947–963.
- [41] G. Peyré and M. Cuturi, Computational optimal transport: with applications to data science. *Found. Trends Mach. Learn.* **11** (2019) 355–607.
- [42] H. Robbins and S. Monro, A stochastic approximation method. *Ann. Math. Stat.* (1951) 400–407.
- [43] F. Santambrogio, Optimal transport for applied mathematicians. *Birkhäuser*, NY (2015) 99–102.
- [44] L. Schwartz, On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **4** (1965) 10–26.
- [45] C. Villani, Topics in optimal transportation. *Graduate Studies in Mathematics*, vol. 58, American Mathematical Society, Providence, RI (2003).
- [46] C. Villani, Optimal transport. Old and new, *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, vol. 338. Springer-Verlag, Berlin (2009).
- [47] R. Wang, X. Wang and L. Wu, Sanov’s theorem in the Wasserstein distance: a necessary and sufficient condition. *Stat. Probab. Lett.* **80** (2010) 505–512.
- [48] Y. Zemel and V. Panaretos, Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli* **25** (2019) 932–976.

## Subscribe to Open (S2O)

A fair and sustainable open access model



This journal is currently published in open access under a Subscribe-to-Open model (S2O). S2O is a transformative model that aims to move subscription journals to open access. Open access is the free, immediate, online availability of research articles combined with the rights to use these articles fully in the digital environment. We are thankful to our subscribers and sponsors for making it possible to publish this journal in open access, free of charge for authors.

### Please help to maintain this journal in open access!

Check that your library subscribes to the journal, or make a personal donation to the S2O programme, by contacting [subscribers@edpsciences.org](mailto:subscribers@edpsciences.org)

More information, including a list of sponsors and a financial transparency report, available at: <https://www.edpsciences.org/en/math-s2o-programme>