Bayesian Learning with Wasserstein Barycenters

Gonzalo Rios[1,2], **Julio Backhoff-Veraguas**[3], **Joaquin Fontbona**[1,2], **Felipe Tobar**[1]

[1]Center for Mathematical Modeling, University of Chile
[2]Dept. of Mathematical Engineering, University of Chile
[3]Inst. of Stats. and Math. Methods in Economics, TU Vienna

Abstract

In this work we introduce a novel paradigm for Bayesian learning based on optimal transport theory. Namely, we propose to use the Wasserstein barycenter of the posterior law on models, as an alternative to the maximum a posteriori estimator and Bayesian model average. We exhibit conditions granting the existence and consistency of this estimator, discuss some of its basic and specific properties, and provide insight for practical implementations relying on standard sampling in finite-dimensional parameter spaces. We thus contribute to the recent blooming of applications of optimal transport theory in machine learning, beyond the discrete setting so far considered. The advantages of the proposed estimator are presented in theoretical terms and through analytical and numeral examples.

1 Model Selection

A probabilistic generative model is a distribution over the data space $\mathcal{X}$, typically $\mathcal{X} \subseteq \mathbb{R}^d$. In this sense, learning a model m from samples $D = \{x_1, ...x_n\} \subset \mathcal{X}$ consists in choosing a distribution over $\mathcal{X}$ from a model space $\mathcal{M}$ that *best* explains the data. Here and in the sequel, we assume that $\mathcal{M} \subseteq \mathcal{P}_{ac}(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$, where $\mathcal{P}_{ac}(\mathcal{X})$ is the set of measures absolutely continuous with respect to a common reference measure λ on $\mathcal{X}$, and $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$.

We say that $\mathcal{M}$ is finitely parametrized if there is integer k, a set $\Theta \subseteq \mathbb{R}^k$ termed parameter space and a (measurable) function $\mathcal{T} : \Theta \mapsto \mathcal{P}_{ac}(\mathcal{X})$, called parametrization mapping, such that $\mathcal{M} = \mathcal{T}(\Theta)$; in such case we denote the model as $m_\theta := \mathcal{T}(\theta)$. In the parametric case, the learning task explained in abstract terms in the previous paragraph boils down to finding the *best* model parameters θ, which are usually found in a frequentist fashion through the maximum likelihood estimator (MLE). We next illustrate the role of the above-introduced objects in a standard learning application.

Example: In linear regression, data consist of input (z_i) and output (y_i) pairs, that is, $x_i = (z_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, ..., n$, and the model space is given by the set of joint distributions $p(z, y) = p(y|z)p(z)$ with linear relationship between y and z. If we moreover assume that $y|z$ is normally distributed, then $p(y|z) = \mathcal{N}(y; z^\top \beta, \sigma^2)$ for some fixed $\beta \in \mathbb{R}^d$ and $\sigma^2 > 0$. In this setting we need to choose the parameters β, σ^2 and $p(z)$ to obtain the joint distribution $p(z, y)$, aka the generative model, though one often needs to deal with the conditional distribution $p(y|z)$, aka discriminative model. Hence, for each fixed $p_0 \in \mathcal{P}_{ac}(\mathbb{R}^d)$, the parameter space $\Theta = \mathbb{R}^d \times \mathbb{R}^+$ induces a model space $\mathcal{M}$ through the mapping $(\beta, \sigma) \mapsto \mathcal{T}(\beta, \sigma)$, where $\mathcal{T}(\beta, \sigma)$ has the density $\mathcal{N}(y; z^\top \beta, \sigma^2)p_0(z)$, $(z, y) \in \mathbb{R}^d \times \mathbb{R}$. Conditioning this joint density $p(y, z)$ wrt a new input $z_\star$, we obtain the predictive distribution addressing the regression problem $p(y|z_\star)$. In particular, denoting $\mathbf{y} = (y_1, ..., y_n)^\top \in \mathbb{R}^n$ and $Z = (z_1, ..., z_n)^\top \in \mathbb{R}^{n \times d}$, the MLE parameters are then given by

Preprint. Work in progress.

$\hat{\beta} = (Z^\top Z)^{-1} Z^\top \mathbf{y}$ and $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - Z\hat{\beta})^\top (\mathbf{y} - Z\hat{\beta})$. A similar treatment can be given to other regression (non-linear, non-normal and/or multivariate), classification and general generative models.

Although the finitely-parametrized case may seem illustrative, the proposed methodology is conceived for both the parametric and non-parametric cases. Additionally, we will adopt the Bayesian viewpoint, which provides a probabilistic framework to deal with model uncertainty, in terms of a *prior distribution* Π on the space $\mathcal{M}$ of models; we refer the reader to Ghosal and van der Vaart [2017], Murphy [2012] and references therein for mathematical background on Bayesian statistics and methods. A critical challenge in the Bayesian perspective, is that of calculating a predictive law on $\mathcal{X}$ from the posterior distribution on $\mathcal{M}$, usually referred to as the *predictive posterior*. This shall be the learning task to which this work is devoted, where our motivation is to find alternative learning strategies which can cope with drawbacks of standard approaches such as maximum a posteriori (MAP) or Bayesian model average.

As a convention, we shall use the same notation for an element $m(dx) \in \mathcal{M}$ and its density $m(x)$. Moreover, we assume that the *true model* $m_0 \in \mathcal{P}_{ac}(\mathcal{X})$ —such that $x_1, ...x_n$ are i.i.d. according to m_0— does exist, although in general m_0 may not be an element of $\mathcal{M}$. We now summarize the main contributions of this article.

2 Main Contribution and Outline of the Article

The main contribution of our work is to introduce the concept of *Bayesian Wasserstein barycenter estimator* as a novel model-selection criterion based on optimal transport theory. See Ambrosio and Gigli [2013], Ambrosio et al. [2004], Villani [2003, 2008], McCann and Guillen [2011] for a thorough introduction to optimal transport. The contribution of the article is structured as follows.

In Section 3 we propose a general framework for Bayesian estimation based on loss functions over probability distributions. This allows us to cover both finitely-parametrized and parameter-free model spaces, and also to retrieve classical selection criteria including MAP, Bayes estimators, and Bayesian model average estimators (and generalizations), as particular instances of *Fréchet means* [Panaretos and Zemel, 2017] with respect to suitable metrics on spaces of probability distributions.

Then, in Sections 4 and 5, we recall the notions of Wasserstein distances and, relying on the previously developed framework, we introduce the Bayesian Wasserstein barycenter estimator. At a theoretical level, we study existence, uniqueness and statistical consistency for this estimator. At a practical level, Section 6 provides illustrative examples and numerical evidence supporting the potential of the proposed estimator, highlighting the computationally-appealing Gaussian case and the use of real-world data.

3 Bayesian Learning in Model Space

Consider a fixed *prior* probability measure $\Pi \in \mathcal{P}(\mathcal{M})$ on the model space $\mathcal{M}$. We assume as customary that, conditionally on the choice of model m, the data $x_1, ..., x_n \in \mathcal{X}$ are distributed as i.i.d. observations from the common law m. As a consequence we can write

$$\Pi(dx_1, ..., dx_n|m) = m(x_1) \cdots m(x_n)\, \lambda(dx_1) \cdots \lambda(dx_n). \tag{1}$$

By virtue of the Bayes rule, the posterior distribution on models given the data, $\Pi(dm|x_1, ..., x_n)$, denoted for simplicity $\Pi_n(dm)$, is given by

$$\Pi_n(dm) := \frac{\Pi(x_1, ..., x_n|m)\,\Pi(dm)}{\Pi(x_1, ..., x_n)} = \frac{m(x_1) ...m(x_n)\,\Pi(dm)}{\int_{\mathcal{M}} \bar{m}(x_1) ...\bar{m}(x_n)\,\Pi(d\bar{m})}. \tag{2}$$

The density $\Lambda_n(m)$ of $\Pi_n(dm)$ with respect the prior $\Pi(dm)$ is called the likelihood function.

Given the model space $\mathcal{M}$, a loss function $L : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is a non-negative functional. We interpret $L(m_0, \bar{m})$ 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, we define the Bayes risk (or expected loss) $R(\bar{m}|D)$ and the Bayes estimator $\hat{m}_L$ as follows:

$$R_L(\bar{m}|D) := \int_{\mathcal{M}} L(m, \bar{m})\Pi_n(dm)\,, \tag{3}$$

$$\hat{m}_L \in \arg\min_{\bar{m}\in\mathcal{M}} R_L(\bar{m}|D). \tag{4}$$

Since both L and Π_n operate directly on the model space, model learning according to the above eqs. does not depend on geometric aspects of parameter spaces. Moreover, the above point of view allows us to define loss functions in terms of various metrics/divergences directly on the space $\mathcal{P}_{ac}(\mathcal{X})$, and therefore to enhance the classical Bayesian estimation framework, by using transportation distances. Before further developing these ideas, we briefly describe how this general framework includes finitely-parametrized model spaces, and discuss standard choices in that setting, together with their appealing features and drawbacks.

3.1 Parametric setting

Given $p \in \mathcal{P}(\Theta)$ a prior distribution over a parameter space Θ, its *push forward* through the map $\mathcal{T}$ is the probability measure $\Pi = \mathcal{T}(p)$ given by $\Pi(A) = p(\mathcal{T}^{-1}(A))$. Expressing the likelihood function $\Lambda_n(m)$ in terms of the parameter θ such that $\mathcal{T}(\theta) = m$, we then easily recover from (2) the standard posterior distribution over the parameter space, $p(d\theta|x_1, \dots, x_n)$. Moreover, any loss function L induces a functional ℓ defined on $\Theta \times \mathbb{R}^k$ (or a subset) by $\ell(\theta_0, \hat{\theta}) = L(m_{\theta_0}, m_{\hat{\theta}})$, interpreted as the cost of choosing parameter $\hat{\theta}$ when the actual true parameter is θ_0. The Bayes risk Berger [2013] of $\bar{\theta} \in \Theta$ is then defined by

$$R_\ell(\bar{\theta}|D) = \int_\Theta \ell(\theta, \bar{\theta}) p(d\theta|x_1, ..., x_n) = \int_{\mathcal{M}} L(m, \bar{m}) \Pi_n(dm), \tag{5}$$

where $\Pi_n(dm) = \Lambda_n(m)\Pi(dm)$, with the prior distribution $\Pi = \mathcal{T}(p)$. The associated Bayes estimator is of course given by $\hat{\theta}_\ell \in \arg\min_{\bar{\theta}\in\Theta} R_\ell(\bar{\theta}|D)$.

For instance, the 0-1 loss defined as $\ell_{0-1}(\theta, \bar{\theta}) = 1 - \delta_{\bar{\theta}}(\theta)$ yields $R_{\ell_{0-1}}(\bar{\theta}|D) = 1 - p(\bar{\theta}|D)$, that is, the corresponding Bayes estimator is the posterior mode, also referred to as Maximum a Posteriori Estimator (MAP), $\hat{\theta}_{\ell_{0-1}} = \hat{\theta}_{MAP}$. The 0-1 loss penalizes all incorrect parameters in the same way, so there is no *partial credit* under this loss function. For continuous-valued quantities the use of a quadratic loss $\ell_2(\theta, \bar{\theta}) = \|\theta - \bar{\theta}\|^2$ is often preferred, as it corresponds to ordinary Euclidean distance in the parameter space. The corresponding Bayes estimator is the posterior mean $\hat{\theta}_{\ell_2} = \int_\Theta \theta p(d\theta|D)$. In one dimensional parameter space, the absolute loss $\ell_1(\theta, \bar{\theta}) = |\theta - \bar{\theta}|$ yields the posterior median estimator, and is sometimes preferred to the ℓ_2 one for being more robust in presence of outliers.

The MAP approach is computationally appealing as it reduces to an optimization problem in a finite dimensional space. The performance of this method might however be highly sensitive to the choice of the initial condition used in the optimization algorithm [Wright and Nocedal, 1999]. This is a critical drawback of MAP estimation, since likelihood functions of expressive models may be populated with numerous local optima. A second drawback of this method is that it fails to capture global information of the model space or to provide a measure of uncertainty of the parameters, which might result in an overfit of the predictive distribution. Indeed, the mode can often be a very poor summary or untypical choice of the posterior distribution (e.g. the mode of an exponential density is 0, irrespective of its parameter). Yet another serious failure of MAP estimation is its dependence on the parameterization. Indeed, for instance, in the case of a Bernoulli distribution on $\{0, 1\}$ with $p(y = 1) = \mu$ and an uniform prior on $[0, 1]$ for μ, the mode can be anything in $[0, 1]$. On the other hand, parameterizing the model by $\theta = \mu^{1/2}$ yields the mode 1, while parametrizing it by $\theta = 1 - (1 - \mu)^{1/2}$ yields 0 as mode.

Using general Bayes estimators on parametrized models of course enables for a richer choice of criteria for model selection (by integrating global information of the parameter space) while providing a measure of uncertainty (through the Bayes risk value). However, the approach might also neglect parameterization related issues, such as overparametrization of the model space (we say that $\mathcal{T}$ overparametrizes $\mathcal{M}$ if it is not one-to-one). The latter might result in a multi-modal posterior distribution over parameters. For example, take $\mathcal{X} = \Theta = \mathbb{R}$, $m_0 = \mathcal{N}(x; \mu, 1)$ and $\mathcal{T}(\theta) = \mathcal{N}(x|\theta^2, 1)$. If we choose a symmetric prior $p(\theta)$, e.g. $p(\theta) = \mathcal{N}(\theta|0, 1)$, then with enough data, the posterior distribution is symmetric with modes near $\{\mu, -\mu\}$, so both ℓ_1 and ℓ_2 estimators are close to 0.

3.2 Posterior average estimators

The next result, proved in Appendix A.1 illustrates the fact that many Bayesian estimators, including the classic *model average estimator*, correspond to finding a so-called Fréchet mean [Panaretos and Zemel, 2017], or barycenter, under a suitable metric/divergence on probability measures.

Proposition 1. *Let $\mathcal{M} = \mathcal{P}_{ac}(\mathcal{X})$ and consider the loss functions $L(m, \bar{m})$ given by:*

i) The L_2-distance: $L_2(m, \bar{m}) = \frac{1}{2}\int_{\mathcal{X}} (m(x) - \bar{m}(x))^2 \lambda(dx)$,

ii) The reverse Kullback-Leibler divergence: $D_{KL}(m||\bar{m}) = \int_{\mathcal{X}} m(x) \ln \frac{m(x)}{\bar{m}(x)} \lambda(dx)$,

iii) The forward Kullback-Leibler divergence $D_{KL}(\bar{m}||m) = \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{m(x)} \lambda(dx)$,

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

Then, in cases i) and ii) the corresponding Bayes estimators of Equation (4) *coincide with the* Bayesian model average*:*

$$\bar{m}(x) := \mathbb{E}^{\Pi_n}[m] = \int_{\mathcal{M}} m(x)\Pi_n(dm). \tag{6}$$

Furthermore, with Z denoting a normalizing constant, the Bayes estimators corresponding to the cases iii) and iv) are given by the exponential model average *and the* square model average*, respectively:*

$$\hat{m}_{exp}(x) = \frac{1}{Z}\exp\int_{\mathcal{M}} \ln m(x)\Pi_n(dm)\ , \quad \hat{m}_2(x) = \frac{1}{Z}\left(\int_{\mathcal{M}} \sqrt{m(x)}\Pi_n(dm)\right)^2. \tag{7}$$

All the above described Bayesian estimators (eqs. (6) and (7)) share a common feature: 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 all the above distances are *vertical* [Santambrogio, 2015], in the sense that computing the distance between m and $\bar{m}$ involves the integral of vertical displacements between the graphs of these two densities. An undesirable fact about *vertical averages* is that they do not preserve properties of the original model space. For example, if the posterior distribution is equally concentrated on two different models $m_1 = \mathcal{N}(\mu_1, 1)$ and $m_2 = \mathcal{N}(\mu_2, 1)$ with $\mu_1 \neq \mu_2$, that is, both models are unimodal (Gaussian) with unit variance, the model average is in turn a bimodal (non-Gaussian) distribution with variance strictly greater than 1. More generally, model averages might yield intractable representations or be hardly interpretable in terms of the prior and parameters.

We shall next introduce the analogous objects in the case of Wasserstein distances, which are *horizontal* distances [Santambrogio, 2015], in the sense that they involve integrating horizontal displacements between the graphs of the densities. We will further develop the theory of the corresponding Bayes estimators, which will correspond to *Wasserstein barycenters* [Agueh and Carlier, 2011, Pass, 2013, Kim and Pass, 2017, Le Gouic and Loubes, 2017] arising in optimal transport theory. In Fig. 1 we illustrate a vertical (left) and a horizontal (right) interpolation between two Gaussian densities, for the reader's convenience.

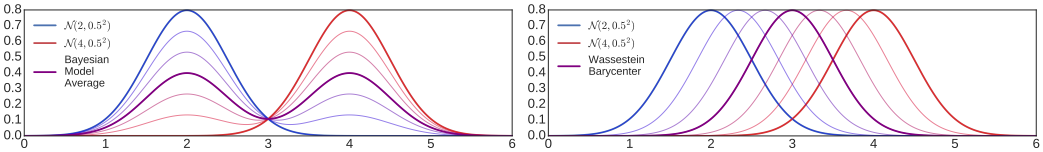

Figure 1: Vertical interpolation (left) and horizontal interpolation (right) of two Gaussian densities.

4 Optimal Transport and Wasserstein Distances: A Summary

Optimal transport has become increasingly popular within the machine learning community [Kolouri et al., 2017], though most of the published works have focused on the discrete setting (e.g. comparing

histograms in Cuturi [2013], Cuturi and Doucet [2014], classification in Frogner et al. [2015] and images in Courty et al. [2017], Arjovsky et al. [2017], among others). Since we will focus on continuous distributions, we next review definitions and results needed to present our approach.

Let $(\mathcal{X}, d)$ be a complete and separable metric space. Given two measures μ, υ over $\mathcal{X}$ we denote $\Gamma(\mu, \upsilon)$ the set of couplings with marginals μ and υ, i.e. $\gamma \in \Gamma(\mu, \upsilon)$ if $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ and we have that $\gamma(dx, \mathcal{X}) = \mu(dx)$ and $\gamma(\mathcal{X}, dy) = \upsilon(dy)$. We say that γ is a deterministic coupling if $\gamma = (Id \times T)(\mu)$ where $T : \mathcal{X} \to \mathcal{X}$ is a (measurable) so-called transport map satisfying $T(\mu) = \upsilon$.

Given a real $p \geq 1$ we define the p-Wasserstein transportation metric between measures μ and υ by

$$W_p(\mu, \upsilon) = \left(\inf_{\gamma \in \Gamma(\mu, \upsilon)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \gamma(dx, dy) \right)^{\frac{1}{p}}, \tag{8}$$

which is a complete metric in $\mathcal{W}_p(\mathcal{X})$, the space of probability measure with finite p-th moment. Since $\Gamma(\mu, \upsilon)$ is not empty (it contains $\mu \otimes \upsilon$), convex and compact with respect to the weak (also called weak-*) topology, minima always exist under mild assumptions on d (joint lower semicontinuity). If μ is non-atomic, then $\{(Id \times T)(\mu) : T \text{ measurable}, T(\mu) = \upsilon\}$ are the extreme points of $\Gamma(\mu, \upsilon)$. In the case $p = 2$, $\mathcal{X} = \mathbb{R}^d$, and if μ has a density, then the minimzer of (8) is unique and is induced by and optimal map which is characterized as $T = \nabla\varphi$ for φ a lower semicontinuous convex function. In some cases there exists an explicit form for the optimal transport and the Wasserstein distance, e.g. for the generic one-dimensional case, and for the multivariate Gaussian case with $p = 2$ (see Cuesta-Albertos et al. [1993]).

5 Bayesian Wasserstein Barycenter Estimator

In this section we propose a novel Bayesian estimator obtained by using the Wasserstein distance as a loss function. This approach will yield an estimator given by a Fréchet mean in the Wasserstein metric, termed Wasserstein barycenter [Agueh and Carlier, 2011]. We state conditions for the statistical *consistency* of our estimator which ensure that it has clear advantages over the Bayesian model average. Finally we show that our estimator has a consistent finite approximation, which can be computed by an iterative method. In the Gaussian case we provide an explicit algorithm.

Throughout we assume that $p \geq 1$ and that $(\mathcal{X}, d)$ is a separable locally-compact geodesic space. For a more detailed summary of the notion of Wasserstein barycenters we refer to Appendix A.2.

5.1 Wasserstein population barycenter

We assume that the support of the prior $\Pi(dm)$ is contained in the subspace $\mathcal{P}_{ac}(\mathcal{X}) \subseteq \mathcal{W}_p(\mathcal{X})$ of probability measures with a density with respect to a fixed reference measure λ. Moreover, we assume that for some (and then all) $\tilde{m} \in \mathcal{W}_p(\mathcal{X})$ it satisfies

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

These conditions are surmised into the statement that $\Pi \in \mathcal{W}_p(\mathcal{P}_{p,ac}(\mathcal{X}))$. We now propose to use the Wasserstein barycenter as the model selection criterion from the posterior Π_n, namely:

Definition 1. *Given a prior $\Pi \in \mathcal{W}_p(\mathcal{P}_{p,ac}(\mathcal{X}))$ and a set $D = \{x_1, ..., x_n\}$ which determines Π_n as in* (2)*, the p-Wasserstein Bayes risk of $\bar{m} \in \mathcal{P}_{p,ac}(\mathcal{X})$, and a Bayes estimator $\hat{m}_p$ of W_p over a model space $\mathcal{M} \subseteq \mathcal{W}_p(\mathcal{X})$, are defined respectively by:*

$$V_p^n(\bar{m}|D) = \int_{\mathcal{P}(\mathcal{X})} W_p(m, \bar{m})^p \Pi_n(dm), \tag{9}$$

$$\hat{m}_p^n \in \arg\min_{\bar{m} \in \mathcal{M}} V_p^n(\bar{m}|D). \tag{10}$$

In the case $\mathcal{M} = \mathcal{W}_p(\mathcal{X})$, any $\hat{m} \in \mathcal{W}_p(\mathcal{X})$ that is a minimum of $V_p^n(\bar{m}|D)$ is referred to as a p-Wasserstein population barycenter of Π_n [Bigot et al., 2012, Bigot and Klein, 2012].

Remark 1. *In the case that $\mathcal{X}$ is a locally compact separable geodesic space (in particular if $\mathcal{X} = \mathbb{R}^d$), the existence of a population barycenter is granted if $\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$, see [Le Gouic*

and Loubes, 2017, Theorem 2] and Appendix A.2 for our own argument. The latter condition is equivalent to the model average $\bar{m}^n(dx) = \mathbb{E}^{\Pi_n}[m](dx)$ *having a finite p-moment, since*

$$\int_{\mathcal{W}_p(\mathcal{X})} W_p(\delta_y, m)^p \Pi_n(dm) = \int_{\mathcal{W}_p(\mathcal{X})} \int_{\mathcal{X}} d(y,x)^p m(dx) \Pi_n(dm) \quad (11)$$

$$= \int_{\mathcal{X}} d(y,x)^p \int_{\mathcal{W}_p(\mathcal{X})} m(dx) \Pi_n(dm), \quad (12)$$

for any $y \in \mathcal{X}$.

The discussion in Remark 1 yields the following useful result:

Proposition 2. *Assume* $\mathcal{X} = \mathbb{R}^d$ *and that the model average* $\bar{m}^n(dx) = \mathbb{E}^{\Pi_n}[m](dx)$ *has finite* p*-moment. Then, the* p*-Wasserstein barycenter of* Π_n *exists.*

In Appendix A.3 we provide conditions on the prior Π ensuring that

$$\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X})) \text{ for all n,}$$

and therefore the existence of the barycenter estimator. From now on throughout this section, we assume implicitly that $\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ for all n.

Remark 2. *In the case* $1 < p < \infty$, $\mathcal{X} = \mathbb{R}^d$ *and that* $\Pi \in \mathcal{W}_p(\mathcal{P}_{p,ac}(\mathcal{X}))$*, the population barycenter of* Π_n *is unique; the argument is almost identical to [Le Gouic and Loubes, 2017, Proposition 6] so we skip it.*

5.2 Statistical Consistency

A natural question is whether our estimator is *consistent* in the statistical sense (see Schwartz [1965], Diaconis and Freedman [1986], Ghosal and van der Vaart [2017]). In a nutshell, consistency corresponds to the convergence of our estimator towards the true model m_0, as we observe more data.

Definition 2. *The prior* Π *is said to be strongly consistent at* m_0 *if for each open neighborhood* U *of* m_0 *in the weak topology of* $\mathcal{M}$*, we have*

$$\Pi_n(U^c) \to 0\,, \quad m_0^{(\infty)} - a.s.$$

Here and in the sequel, $m_0^{(\infty)}$ denotes the product probability measure corresponding to the infinite sample $\{x_1, ..., x_n, ...\}$ of i.i.d. data distributed according to m_0.

Remark 3. *As mentioned in [Ghosal and van der Vaart, 2017, Proposition 6.2], strong consistency is equivalent to the* $m_0^{(\infty)}$*-almost sure weak convergence of* Π_n *to* δ_{m_0}*. For simplicity of the presentation, we do not examine the notion of weak consistency in this work.*

The celebrated Schwartz's theorem provides sufficient conditions for strong consistency. See Schwartz [1965] or [Ghosal and van der Vaart, 2017, Proposition 6.16] for more modern treatment. A key ingredient is:

Definition 3. *A density* m_0 *belongs to the Kullback-Leibler support of* Π*, denoted* $m_0 \in KL(\Pi)$*, if* $\Pi(m : D_{KL}(m_0||m) < \varepsilon) > 0$ *for every* $\varepsilon > 0$*, where* $D_{KL}(m_0||m) = \int m_0 \log \frac{m_0}{m} d\lambda$.

Remark 4. *As can be derived from [Ghosal and van der Vaart, 2017, Example 6.20], in our particular setting, we have*

$$\Pi \textit{ is strongly consistent at } m_0 \textit{ with respect to the weak topology } \iff m_0 \in KL(\Pi).$$

We now come to the important question, of whether our Wasserstein barycenter estimator converges to the model m_0. We assume from now on that

$$m_0 \in KL(\Pi) \text{ and } m_0 \in \mathcal{W}_p(\mathcal{X}).$$

Note that the first condition implies that the model is "correct" or "well specified", in the sense that $m_0 \in \mathcal{M}$, as discussed in Berk et al. [1966], Grendár and Judge [2009], Kleijn et al. [2004, 2012]. This setting could be slightly relaxed in the "misspecified" setting dealt with in those works.

We first provide a rather direct sufficient condition for the convergence of $\hat{m}_p^n$ to m_0. We use W_p to denote throughout the Wasserstein distance both on $\mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))$ and on $\mathcal{W}_p(\mathcal{X})$, not to make the notation heavier.

Proposition 3. *If $W_p(\Pi_n, \delta_{m_0}) \to 0$ $m_0^{(\infty)}$-almost surely, then $W_p(\hat{m}_p^n, m_0) \to 0$ $m_0^{(\infty)}$-almost surely.*

Proof. We have, by minimality of the barycenter

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

It follows that the last terms goes to 0 too. On the other hand,

$$W_p(m_0, \hat{m}_p^n)^p \ \leq c \, W_p(m, \hat{m}_p^n)^p + c \, W_p(m, m_0)^p \, , \ \forall m,$$

where the constant c only depends on p. Thus

$$\begin{aligned} W_p(m_0, \hat{m}_p^n)^p \ &\leq c \int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \, \Pi_n(dm) + c \int_{\mathcal{M}} W_p(m, m_0)^p \, \Pi_n(dm) \\ &= c \int_{\mathcal{M}} W_p(m, \hat{m}_p^n)^p \, \Pi_n(dm) + c \, W_p(\Pi_n, \delta_{m_0})^p, \end{aligned}$$

so we conclude. □

The next result states that if the posterior distribution is consistent in the Wasserstein sense (a weaker condition than above), then our estimator is consistent (under some conditions) for models in the Kullback-Leibler support $KL(\Pi)$ of Π.

Proposition 4. *If the posterior distribution Π_n is p-Wasserstein strongly consistent at $m_0 \in KL(\Pi)$, then the barycenter is strongly consistent at m_0 with some regularity conditions, like e.g. $\int W_p^q(m, m_0)\Pi_n(dm) < C$ for some $q > p$ and $C > 0$. It is possible to relax the condition to $\int W_p^p(m, m_0)\Pi(dm) < \infty$ if the likelihood function $\Lambda_n(m)$ converge uniformly to 0 as $n \to \infty$, on the sets $B^c(m_0, \epsilon) = \{\nu | W_p(\nu, m_0) > \epsilon\}$ for every $\epsilon > 0$.*

We finally specialize to the consistency of the 2-Wasserstein barycenter in the case of (sub-)Gaussian models:

Proposition 5. *In the case that $p = 2$, $\mathcal{X} = \mathbb{R}^d$, $\lambda =$ Lebesgue, and assuming that supp$(\Pi) \subset \mathcal{P}_{2,ac}(\mathbb{R}^d)$ consists of measures which have uniformly sub-Gaussian tails (i.e. with decay $O(e^{-x^2})$ or lighter), then the posterior distribution Π_n is strongly 2-Wasserstein consistent for all $m_0 \in KL(\Pi)$.*

The proofs of the previous two propositions will be provided in the next version of this work. In any case we shall not need these results in the sequel

5.3 Bayesian Wasserstein barycenter vs Bayesian model average

It is illustrative to compare the Bayesian model average with our barycenter estimators. First we show that if the Bayesian model converges, then our estimator converges too. We even have the stronger condition that the posterior distribution converges in $\mathcal{W}_p$. Recall that the model average is given by $\bar{m}^n(dx) = \mathbb{E}_{\Pi_n}[m](dx)$.

Proposition 6. *If $m_0^{(\infty)}$-a.s. the p-moments of the model average converge to those of $m_0 \in KL(\Pi)$, then $W_p(\Pi_n, \delta_{m_0}) \to 0$ $m_0^{(\infty)}$-a.s. In particular, also $W_p(\hat{m}_p^n, m_0) \to 0$ $m_0^{(\infty)}$-almost surely.*

Proof. We already know that the prior is strongly consistent at m_0 with respect to the weak topology. Notice that

$$\int W_p(m, \delta_x)^p \Pi_n(dm) = \int \int d(x, z)^p m(dz) \Pi_n(dm) = \int d(x, z)^p \bar{m}^n(dz),$$

from which it follows that $\Pi_n \to \delta_0$ not only weakly but in W_p, almost surely. We conclude by Proposition 3. □

Remark 5. *Since by Remark 4, the prior is strongly consistent at $m_0 \in KL(\Pi)$ with respect to the weak topology, [Ghosal and van der Vaart, 2017, Theorem 6.8] and the discussion thereafter imply that the model average is consistent at m_0 too.*

We now briefly consider the case of $\Pi \in \mathcal{W}_2(\mathcal{P}_{2,ac}(\mathbb{R}^d))$, $\mathcal{M} = \mathcal{P}_{2,ac}(\mathbb{R}^d)$, and $\lambda =$ Lebesgue. Let $\hat{m}$ be its unique population barycenter, 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 $m \in \mathcal{W}_2(\mathcal{X})$ (the existence of which is proved in Fontbona et al. [2010]). We will prove in Lemma 3 in Appendix A.2 that $\hat{m} = (\int T^m \Pi(dm))(\hat{m})$. Thanks to this fixed-point property, for all convex functions ϕ we have

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

where $\bar{m} = \mathbb{E}^{\Pi}[m]$ is the Bayesian model average. We have used here Jensen's inequality and Fubini. All in all we established that the 2-Wasserstein barycenter estimator $\hat{m}$ is less dispersed than the Bayesian model average: namely, in the convex-order sense. It follows too that $\hat{m}$ has the same mean as $\bar{m}$. We can replace Π by Π_n in this discussion, and obtain:

Proposition 7. *Given a prior $\Pi \in \mathcal{W}_2(\mathcal{P}_{2,ac}(\mathbb{R}^d))$, let $\bar{m}^n$ be the Bayesian model average and $\hat{m}^n$ the 2-Wasserstein barycenter of the posterior Π_n. Then we have*

$$\mathbb{E}_{\bar{m}^n}[x] = \mathbb{E}_{\hat{m}^n}[x] \quad \textit{and} \quad \mathbb{E}_{\bar{m}^n}[\|x\|^2] \geq \mathbb{E}_{\hat{m}^n}[\|x\|^2],$$

so in particular the 2-Wasserstein barycenter estimator has less variance than the model average.

5.4 Wasserstein Empirical Barycenter

In practice, except in special cases, we cannot calculate integrals over the whole model space $\mathcal{M}$. Thus we must approximate such integrals by e.g. Monte Carlo methods. For this reason, we now discuss the *empirical Wasserstein barycenter* and its usefulness as an estimator.

Definition 4. *Given $m_i \overset{iid}{\sim} \Pi_n$ for $i = 1, ..., k$, we define the empirical distribution as*

$$\Pi_n^{(k)} = \frac{1}{k}\sum_{i=1}^{k} \delta_{m_i} \ \in \mathcal{P}(\mathcal{M})\,. \tag{13}$$

Using $\Pi_n^{(k)}$ instead of Π_n, we define the p-Wasserstein empirical Bayes risk $V_p^{(n,k)}(\bar{m}|D)$, as well as a corresponding empirical Bayes estimator $\hat{m}_p^{(n,k)}$, which in the case $\mathcal{M} = \mathcal{W}_p$ is referred to as a p-Wasserstein empirical barycenter of Π_n (see Bigot et al. [2012], Bigot and Klein [2012]).

Remark 6. *To compute a empirical distribution, we can use efficient Markov Chain Monte Carlo (MCMC) techniques [Goodman and Weare, 2010] or transport sampling procedures [El Moselhy and Marzouk, 2012, Parno, 2015, Kim et al., 2015, Marzouk et al., 2016] to generate models m_i from Π_n.*

Remark 7. *It is known that a.s. $\Pi_n^{(k)}$ converges weakly to Π_n as $k \to \infty$. If Π_n has finite p-th moments, then by the strong law of large numbers we have convergence of p-th moments too*

$$\int W_p(m, m_0)^p \Pi_n^{(k)}(dm) = \frac{1}{k}\sum_{i=1}^{k} W_p(m_i, m_0)^p \to \int W_p(m, m_0)^p \Pi_n(dm) \text{ a.s.} \tag{14}$$

Thus we have that a.s. $\Pi_n^{(k)} \to \Pi_n$ in $\mathcal{W}_p$. Thanks to [Le Gouic and Loubes, 2017, Theorem 3], any sequence of empirical barycenters $(\hat{m}_n^k)_{k\geq 1}$ of $(\Pi_n^k)_{k\geq 1}$ converges (up to selection of a subsequence) in p-Wasserstein distance to a (population) barycenter $\hat{m}_n$ of Π_n. Combining these facts, the following result is immediate:

Proposition 8. *If $W_p(\Pi_n, \delta_{m_0}) \to 0$, $m_0^{(\infty)}$-a.s., there exists a data-dependent sequence k_n such that the empirical barycenters $(\hat{m}_n^{k_n})_{n\geq 1}$ satisfy $W_p(\hat{m}_n^{k_n}, m_0) \to 0$, $m_0^{(\infty)}$-a.s.*

6 The Gaussian Case: Examples, Methods and Experiments

In this section, we consider $\mathcal{X} = \mathbb{R}^d$, $p = 2$ and $\mathcal{M}$ consists of Gaussian distributions, since in this case exact results are available [Álvarez-Esteban et al., 2015, 2016]. As a matter of fact, for two

models given by the univariate Gaussian distributions $m_0 = \mathcal{N}(\mu_0, \sigma_0^2)$ and $m_1 = \mathcal{N}(\mu_1, \sigma_1^2)$, we have that

- The optimal transport T from model m_0 to model m_1 is given by $T(x) = \mu_1 + \frac{\sigma_1}{\sigma_0}(x - \mu_0)$,
- the Wasserstein interpolation map is $T_t(x) = (1-t)x + t(\mu_1 + \frac{\sigma_1}{\sigma_0}(x - \mu_0))$, $t \in [0, 1]$,
- the Wasserstein-interpolated model is $m_t = \mathcal{N}((1-t)\mu_0 + t\mu_1, ((1-t)\sigma_0 + t\sigma_1)^2)$,
- the Wasserstein-barycenter model is given by $\hat{m} = m_{1/2} = \mathcal{N}(\frac{\mu_0+\mu_1}{2}, (\frac{\sigma_0+\sigma_1}{2})^2)$.

Relying on the tractability of the Gaussian case, we nex tvalidate the proposed approach experimentally in three aspects. We first compare the variance of the proposed estimator against those of model average and MAP in an univariate example; we then equip the proposed method with a fast computation of barycenters in a multivariate case; lastly, we present an example with real-world data.

6.1 A conjugate prior for closed-form Bayesian Wasserstein learning

We now present a consistent continuous prior for Gaussian distributions, which allows us to calculate and compare the MAP estimator, the model average and the 2-Wassestein barycenter in closed form.

Consider the observations $D = \{x_1, ..., x_n\}$ generated by the true model $m_0 = \mathcal{N}(\bar{\mu}, \bar{\sigma}^2) \in \mathcal{M}$, where $\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2) | \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$. Let us also choose the prior over models by placing a Normal-inverse-gamma distribution (NIG) over the parameters (μ, σ^2), given by $\mathcal{NIG}(\mu, \sigma^2 | \mu_0, \lambda_0, \alpha_0, \beta_0) = \mathcal{N}(\mu | \mu_0, \sigma^2/\lambda_0)\mathcal{IG}(\sigma^2 | \alpha_0, \beta_0)$, $\mu_0 \in \mathbb{R}$ and $\lambda_0, \alpha_0, \beta_0 \in \mathbb{R}^+$, which induces a prior Π over models $\mathcal{M}$. As the NIG distribution is conjugate to the Gaussian likelihood, the posterior distribution of the model parameters is given by $(\mu, \sigma^2 | x_1, ..., x_n) \sim \mathcal{NIG}(\mu_n, \lambda_n, \alpha_n, \beta_n)$ with $\mu_n = \frac{\lambda_0\mu_0 + n\bar{x}_n}{\lambda_0 + n}$, $\lambda_n = \lambda_0 + n$, $\alpha_n = \alpha_0 + \frac{n}{2}$ and $\beta_n = \beta_0 + \frac{1}{2}\left(n\bar{s}_n + \frac{n\lambda_0(\bar{x}_n - \mu_0)^2}{\lambda_0 + n}\right)$, where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{s}_n = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. See more details in Murphy [2007].

We will show that the above prior is strongly consistent at m_0. The mean of the posterior is $(\mu_n, \frac{\beta_n}{\alpha_n - 1/2})$, which converges to $(\bar{x}, \bar{s}) = \lim_{n\to\infty}(\bar{x}_n, \bar{s}_n)$ and are respectively the mean and variance of m_0, due to the strong law of large numbers. Since the variance of the posterior is $\mathcal{O}(\frac{1}{n})$ in both variables (μ, σ^2), the posterior converges a.s. in the weak topology to the point mass at $(\bar{\mu}, \bar{\sigma}^2)$, therefore, NIG prior is strongly consistent at m_0 in the weak topology.

Additionally, we know [Murphy, 2012] that the MAP estimator is $\mathcal{N}(x | \mu_n, \frac{\beta_n}{\alpha_n + \frac{3}{2}})$, while the model average is the Student's $t-$distribution $t_{2\alpha_n}(x | \mu_n, \frac{\beta_n(1+\lambda_n)}{\alpha_n\lambda_n})$ with variance is $\frac{\beta_n(1+\lambda_n)}{(\alpha_n - 1)\lambda_n}$. This reveals the non-Gaussianity of the model average, despite the prior (and all posteriors) being Gaussian.

The second moment of the model average is given by $\mu_n^2 + \frac{\beta_n(1+\lambda_n)}{(\alpha_n - 1)\lambda_n} = \mu_n^2 + \frac{\bar{s}_n}{1 + \mathcal{O}(\frac{1}{n})} + \mathcal{O}(\frac{1}{n})$, which converges to the second moment of m_0. By Prop. 6 the 2-Wasserstein barycenter of the posterior (which exists) converges a.s. to m_0 and is given by $\mathcal{N}(x | \mu_n, \hat{\sigma}^2)$. From [Álvarez-Esteban et al., 2018, Thm. 3.10], denoting σ_m^2 the variance for a model $m \in \mathcal{M}$, the barycenter variance $\hat{\sigma}^2$ satisfies

$$\hat{\sigma}^2 = \int \left(\hat{\sigma}\sigma_m^2\hat{\sigma}\right)^{1/2} \Pi_n(dm) = \hat{\sigma} \int \sigma_m \Pi_n(dm). \tag{15}$$

Furthermore, using the variance posterior $\mathcal{IG}(\sigma^2 | \alpha_n, \beta_n)$ and the change of variable $z = \sigma^2$ we have

$$\hat{\sigma} = \int \sigma_m \Pi_n(dm) = \int z^{1/2} \mathcal{IG}(z | \alpha_n, \beta_n) dz = \frac{\beta_n^{1/2}\Gamma(\alpha_n - \frac{1}{2})}{\Gamma(\alpha_n)}. \tag{16}$$

Thus, the MAP, average and barycenter models have the same mean μ_n but different variance. Fig. 2 (left) compares the variances in a numerical example with $a_0 = 2$, $\lambda_0 = 1$ and $\beta_n = 1$. Note that our estimator has higher variance than MAP, but less than the model average.

6.2 Empirical barycenter of multivariate Gaussian distributions

Let us consider: (i) a set of k Gaussian models $\{m_i = \mathcal{N}(0, \Sigma_i)\}_{i=1}^k$, where $\Sigma_1, ..., \Sigma_k$ are $d \times d$ symmetric positive semidefinite matrices, with at least one of them being symmetric and positive

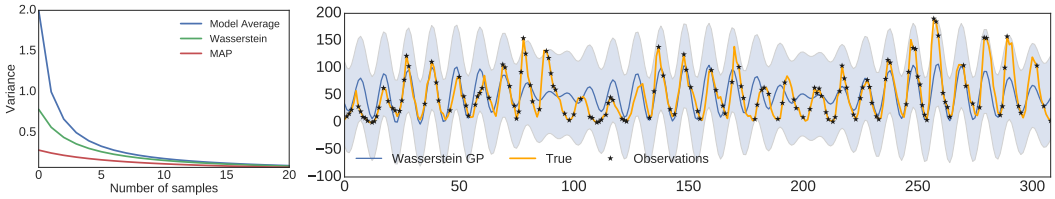

Figure 2: Left: Variance of the selected model under three criterion. Right: A Gaussian process regression with a cosine kernel, learning with Wasserstein barycenter.

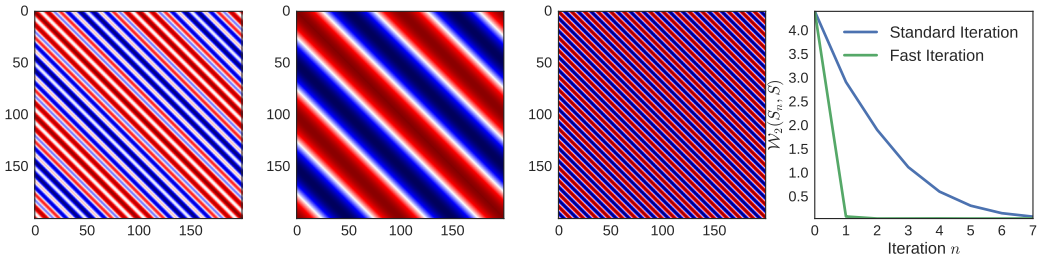

Figure 3: Barycenter (first), covariance matrices (second, third) and convergence speeds (fourth).

definite (SPD); and also (ii) the normalized weights $\lambda_1, ..., \lambda_k \in \mathbb{R}^+$, $\sum_{i=1}^k \lambda_i = 1$. These k Gaussian models can be regarded as learned from minibatches or different parameter-setting criteria.

Owing to Álvarez-Esteban et al. [2015], the Wasserstein barycenter of the discrete measure $\Pi^{(k)} = \sum_{i=1}^k \lambda_i \delta_{m_i}$ is given by $\hat{m} = \mathcal{N}(0, \bar{\Sigma})$, where $\bar{\Sigma}$ is the unique SPD solution of the equation $H(\bar{\Sigma}) = \sum_{i=1}^k \lambda_i \left(\bar{\Sigma}^{\frac{1}{2}} \Sigma_i \bar{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}} = \bar{\Sigma}$. For a computationally-appealing solution to $H(\bar{\Sigma}) = \bar{\Sigma}$, Álvarez-Esteban et al. [2016] define the *fast* sequence that, starting from some SPD matrix S_0, the succession $S_{n+1} = S_n^{-1/2} \left(\sum_{i=1}^k \lambda_i \left(S_n^{1/2} \Sigma_i S_n^{1/2}\right)^{1/2}\right)^2 S_n^{-1/2}$, for $n \geq 0$ converge in 2-Wasserstein to the barycenter, i.e. $W_2\left(\mathcal{N}(0, S_n), \mathcal{N}(0, \bar{\Sigma})\right) \to 0$ as $n \to \infty$. Fig. 3 shows the 2-Wasserstein distance of the fast and fixed-point sequences $\tilde{S}_{n+1} = H(\tilde{S}_n)$ to the barycenter as a function of the iteration index, using $k = 2$ Gaussians of dimension 200×200.

6.3 Bayesian Wasserstein learning for Gaussian processes using a real-world data

We considered the Sunspots time series (available from Pedregosa et al. [2011]) between 1700 and 2008, and used half of the data (154 points) for training and the rest for testing. Setting a flat prior [Gelman et al., 2008] over the hyperparameters, we trained a Gaussian process (GP) [Rasmussen and Williams, 2006] with constant mean function and cosine covariance kernel. Using MCMC we generated k independent mean vectors and covariance matrices. We then found the barycenter GP by averaging the mean vectors and applying the fast sequence to the covariances. According to Prop. 8, the number of sampled models k is data-dependent, thus we searched for k based on empirical convergence of the barycenter. We remind the reader that in this case $\mathcal{M}$ is the space of all Gaussian processes [Mallasto and Feragen, 2017] and that the true model m_0 is unknown.

Fig. 2 (right) shows the posterior predictive mean and the 95%-confidence interval of the barycenter model. Note that our model was able to recover a varying-waveform, close-to-periodic, signal using a prior with support only for perfectly-periodic time series. This validates the proposed methodology to handle model mismatch, and in this case recover the signal frequency. Table 1 shows that the model selected with Wasserstein barycenter has a superio performance than MAP and model average in mean absolute error (MAE) and square root mean error (RMSE) on observed and test data.

Score	MAE			RMSE		
Model / Dataset	Obs	Test	Total	Obs	Test	Total
MAP	29.380	29.067	29.223	37.515	36.483	37.001
Model Average	27.631	25.057	26.340	35.460	31.648	33.602
Wasserstein	23.143	22.552	22.846	30.874	28.977	29.937

Table 1: Result of model selection with Sunspots dataset.

7 Discussion and Future Work

We have proposed an unifying framework for Bayesian model selection, covering standard selection criteria, to then introduce the novel *Bayesian Wasserstein barycenter estimator*. We have also illustrated the appealing statistical properties of the proposed estimator, and shown implementation examples in parametric and nonparametric Gaussian cases, where the desired performance of the proposed method was validated experimentally. Future research will include extensions to more-expressive families of distributions to better explain complex real-world data, as well as the development of computationally-efficient methods.

References

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

P. C. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. A note on the computation of Wasserstein barycenters. 2015.

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

P. C. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Wide consensus aggregation in the Wasserstein space. application to location-scatter families. 2018.

L. Ambrosio and N. Gigli. *A user's guide to optimal transport*. Springer, 2013.

L. Ambrosio, N. Gigli, and G. Savaré. Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl*, 15(3-4):327–343, 2004.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

R. H. Berk et al. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

J. Bigot and T. Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *arXiv preprint arXiv:1212.2562*, 2012.

J. Bigot, T. Klein, et al. Consistent estimation of a population barycenter in the Wasserstein space. *ArXiv e-prints*, 2012.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

J. Cuesta-Albertos, L. Ruschendorf, and A. Tuero-Diaz. Optimal coupling of multivariate distributions and stochastic processes. *Journal of Multivariate Analysis*, 46(2):335–361, 1993.

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

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

P. Diaconis and D. Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

T. A. El Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

J. Fontbona, H. Guérin, and S. Méléard. Measurability of optimal transportation and strong coupling of martingale measures. *Electron. Commun. Probab.*, 15:124–133, 2010. ISSN 1083-589X. doi: 10.1214/ECP.v15-1534. URL https://doi.org/10.1214/ECP.v15-1534.

C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

A. Gelman, A. Jakulin, M. G. Pittau, Y.-S. Su, et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5(1):65–80, 2010.

M. Grendár and G. Judge. Asymptotic equivalence of empirical likelihood and Bayesian map. *Ann. Statist.*, 37(5A):2445–2457, 10 2009. doi: 10.1214/08-AOS645. URL https://doi.org/10.1214/08-AOS645.

S. Kim, D. Mesa, R. Ma, and T. P. Coleman. Tractable fully Bayesian inference via convex optimization and optimal transport theory. *arXiv preprint arXiv:1509.08582*, 2015.

Y.-H. Kim and B. Pass. Wasserstein barycenters over riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017.

B. J. K. Kleijn, A. W. Van der Vaart, et al. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

B. J. K. Kleijn et al. *Bayesian asymptotics under misspecification*. PhD thesis, Vrije Universiteit Amsterdam, 2004.

S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

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

A. Mallasto and A. Feragen. Learning from uncertain curves: The 2-Wasserstein metric for gaussian processes. In *Advances in Neural Information Processing Systems*, pages 5665–5674, 2017.

Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*, 2016.

R. J. McCann and N. Guillen. Five lectures on optimal transportation: geometry, regularity and applications. *Analysis and geometry of metric measure spaces: lecture notes of the séminaire de Mathématiques Supérieure (SMS) Montréal*, pages 145–180, 2011.

K. P. Murphy. Conjugate Bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16, 2007.

K. P. Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

V. Panaretos and Y. Zemel. Fréchet means and procrustes analysis in Wasserstein space. *Bernoulli*, 2017.

M. Parno. *Transport maps for accelerated Bayesian computation*. PhD thesis, Massachusetts Institute of Technology, 2015.

B. Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264 (4):947–963, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT, 2006.

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.

L. Schwartz. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

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

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

S. J. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

A Appendix

A.1 Bayes Estimators as Generalized Model Averages

Let $\mathcal{M} = \mathcal{P}_{2,ac}$ and consider the squared L_2-distance between densities as loss function

$$L_2(m, \bar{m}) = \frac{1}{2} \int_{\mathcal{X}} \left(m(x) - \bar{m}(x)\right)^2 \lambda(dx),$$

by Fubini's theorem we have that

$$R_L(\bar{m}|D) = \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{M}} \left(m(x) - \bar{m}(x)\right)^2 \Pi(dm|D)\lambda(dx),$$

where, by the fundamental lemma of calculus of variations, denoting

$$\mathcal{L}\left(x, \bar{m}, \bar{m}'\right) = \frac{1}{2} \int_{\mathcal{M}} \left(m(x) - \bar{m}(x)\right)^2 \Pi(dm|D)$$

the extrema of $R_L(\bar{m}|D)$ are weak solutions of the Euler-Lagrange equation

$$\begin{array}{rcl} \frac{\partial \mathcal{L}\left(x,\bar{m},\bar{m}'\right)}{\partial \bar{m}} & = & \frac{d}{dx} \frac{\partial \mathcal{L}\left(x,\bar{m},\bar{m}'\right)}{\partial \bar{m}'} \\ \int_{\mathcal{M}} \left(m(x) - \bar{m}(x)\right) \Pi(dm|D) & = & 0, \end{array}$$

so we have that the optimal is reached on the Bayesian model average

$$\int_{\mathcal{M}} m(x)\Pi(dm|D).$$

Similarly, if we take the loss function as the reverse Kullback-Leibler divergence as loss function

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

we have that the associate Bayes risk can be written as

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

and changing the constant C by the entropy of $\mathbb{E}[m]$ we have that

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

so the extremum of $R_{D_{RKL}}(\bar{m}|D)$ is given by the Bayesian model average. Instead if we take the forward Kullback-Leibler divergence as loss function

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

we have

$$\begin{array}{rcl} R_{D_{KL}}(\bar{m}|D) & = & \int_{\mathcal{M}} \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{m(x)} \lambda(dx) \Pi(dm|x_1, ..., x_n) \\ & = & \int_{\mathcal{X}} \bar{m}(x) \ln \bar{m}(x) \lambda(dx) - \int_{\mathcal{X}} \bar{m}(x) \int_{\mathcal{M}} \ln m(x) \Pi(dm|x_1, ..., x_n) \lambda(dx) \\ & = & \int_{\mathcal{X}} \bar{m}(x) \ln \bar{m}(x) \lambda(dx) - \int_{\mathcal{X}} \bar{m}(x) \ln \exp \mathbb{E}[\ln m] \lambda(dx) \\ & = & \int_{\mathcal{X}} \bar{m}(x) \ln \frac{\bar{m}(x)}{\exp \mathbb{E}[\ln m]} \lambda(dx). \end{array}$$

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

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

So the extremum of $R_{D_{KL}}(\bar{m}|D)$ is the Bayesian *exponential* model average given by

$$\hat{m}(x) = \tfrac{1}{Z} \exp \int_{\mathcal{M}} \ln m(x) \Pi(dm|x_1, ..., x_n).$$

Finally, if we take the squared Hellinger distance as loss function

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

we check analogously that the extremum of $R_{H^2}(\bar{m}|D)$ is the Bayesian *square* model average, namely

$$\begin{aligned} \hat{m}(x) &= \tfrac{1}{Z} \left(\int_{\mathcal{M}} \sqrt{m(x)} \Pi(dm|x_1, ..., x_n) \right)^2 \\ Z &= \int_{\mathcal{X}} \left(\int_{\mathcal{M}} \sqrt{m(x)} \Pi(dm|x_1, ..., x_n) \right)^2 \lambda(dx). \end{aligned}$$

A.2 Wasserstein Barycenters

We start following the presentation in Le Gouic and Loubes [2017]. Let $\mathcal{X}$ be a locally compact separable geodesic space with associated metric d. The latter means that any pair of points admit a mid-point with respect to d. As before $W_p(\cdot, \cdot)$ denotes the Wasserstein distance of order p based on d; see (8). This distance is defined on $\mathcal{P}_p(\mathcal{X})$, the set of probability measures which integrate $d(\cdot, x)^p$ for some $x \in \mathcal{X}$.

We can now consider $\mathcal{P}_p(\mathcal{X})$ with the complete metric W_p as a base Polish space, and define $\mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$ analogously, with an associated Wasserstein distance of order p which for simplicity we still call W_p.

Let $\Pi \in \mathcal{P}_p(\mathcal{P}_p(\mathcal{X}))$. We observe that by definition its *model-average* belongs to $\mathcal{P}_p(X)$, since

$$\infty > \int W_p(m, \delta_x)^p \Pi(dm) = \int\int d(x, y)^p m(dy) \Pi(dm) = \int d(x, y)^p \int m(dy) \Pi(dm).$$

A p-Wasserstein barycenter of $\Pi \in \mathcal{P}_p(\mathcal{P}_p(X))$ is an optimizer of

$$V(\Pi) := \inf \left\{ \int_{\mathcal{P}_p(\mathcal{X})} W_p^p(\nu, m) \Pi(dm) : \nu \in \mathcal{P}_p(\mathcal{X}) \right\},$$

as the Def. 1 in the main text. We state an existence result first obtained in [Le Gouic and Loubes, 2017, Theorem 2]; our argument here seems more elementary.

Lemma 1. *There exists a minimizer for $V(\Pi)$, i.e. a p-Wasserstein barycenter.*

Proof. Taking $\nu = \delta_x$ we get that $V(\Pi)$ is finite. Now, let $\{\nu_n\} \subset \mathcal{P}_p(\mathcal{X})$ such that

$$\int_{\mathcal{P}_p(\mathcal{X})} W_p(\nu_n, m)^p \Pi(dm) \searrow V(\Pi).$$

For n large enough we have

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

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 $\epsilon > 0$, that there is ℓ large enough such that $\sup_n \nu_n(\{y \in \mathcal{X} : d(x, y) > \ell\}) \leq \epsilon$. As explained in Le Gouic and Loubes [2017], the assumptions made 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 $\nu \in \mathcal{P}_p(\mathcal{X})$ which is its weak limit. We can conclude by Fatou's lemma:

$$V(\Pi) = \lim \int W_p(\nu_n, m)^p \Pi(dm) \geq \int \liminf W_p(\nu_n, m)^p \Pi(dm) \geq \int W_p(\nu, m)^p \Pi(dm).$$

□

Let us now consider the relevant case of $\mathcal{X} = \mathbb{R}^q$ with d the euclidean distance and $p = 2$. We take

$$\Pi \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^q)),$$

observing that in such situation the previous lemma applies. We recall now the uniqueness result stated in [Le Gouic and Loubes, 2017, Proposition 6]:

Lemma 2. *Assume that there exists a set $A \subset \mathcal{P}_2(\mathbb{R}^q)$ of measures with*

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

and $\Pi(A) > 0$. Then Π admits a unique 2-barycenter.

We now further assume that all measures in the support of Π are absolutely continuous with respect to Lebesgue measure. The previous lemma guarantees the uniqueness of the barycenter. We last provide a useful characterization of barycenters, which is a generalization of the corresponding result in Álvarez-Esteban et al. [2016] where only the case $|\text{supp}(\Pi)| < \infty$ is covered:

Lemma 3. *Let ν be the unique barycenter of Π in this setting. Let $(m, x) \mapsto T^m(x)$ denote a measurable function equal $\lambda(dx)\Pi(dm)$ a.e. to the unique optimal transport map from $\hat{m}$ to $m \in \mathcal{W}_2(\mathcal{X})$ (the existence of which is proved in Fontbona et al. [2010]). Then $x = \int T^m(x)\Pi(dm),\ \nu(dx)-a.s.$*

Proof. Assume the assertion is not true, so in particular

$$\begin{aligned}0 &< \int \left(x - \int T^m(x)\Pi(dm)\right)^2 \nu(dx)\\ &= \int |x|^2\nu(dx) - 2\int\int xT^m(x)\Pi(dm)\nu(dx) + \int \left(\int T^m(x)\Pi(dm)\right)^2 \nu(dx).\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\int W_2\left(\left(\int T^m\Pi(dm)\right)(\nu)\,,\,\bar{m}\right)^2\Pi(d\bar{m}) &\leq \int \left[T^{\bar{m}}(x) - \int T^m(x)\Pi(dm)\right]^2\nu(dx)\Pi(d\bar{m})\\ &= \int\int \left[T^m(x)\right]^2\nu(dx)\Pi(dm) - \int\left(\int T^m(x)\Pi(dm)\right)^2\nu(dx),\end{aligned}$$

after a few computations. But, by Brenier's theorem of optimal transport we know that

$$\int\int (x - T^m(x))^2\nu(dx)\Pi(dm) = \int W_2(\nu, m)^2\Pi(dm).$$

Bringing together these three observations, we deduce

$$\int W_2\left(\left(\int T^m\Pi(dm)\right)(\nu)\,,\,\bar{m}\right)^2\Pi(d\bar{m}) < \int W_2(\nu, m)^2\Pi(dm),$$

and in particular ν cannot be the barycenter. □

A.3 A Condition for Existence of Barycenters of Bayesian Posteriors

We last provide a general condition on the prior Π ensuring that

$$\Pi_n \in \mathcal{W}_p(\mathcal{W}_p(\mathcal{X}))\ \text{ for all n,}$$

and therefore the existence of the barycenter estimator.

Definition 5. *We say that $\Pi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ is* 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 $y \in \mathcal{X}, \varepsilon > 0$:*
$$\int_{\mathcal{M}}\left(\int_{\mathcal{X}} d(y, z)^p m(dz)\right)^{1+\varepsilon}\Pi(dm) < \infty.$$

Condition (2) above could be intuitively summarized with the notation $\Pi \in \mathcal{W}_{p+}(\mathcal{W}_p(\mathcal{X}))$.

Lemma 4. *Suppose that Π is 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 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 Π fulfills 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 ε small enough and t close enough to 1. We conclude. □

Lemma 5. *Suppose that Π is integrable after updates. Then for all $n \in \mathbb{N}$ and $\{x_1, \ldots, x_n\} \in \mathcal{X}^n$, the posterior Π_n is also integrable after updates.*

Proof. By Lemma 4, we obtain that Π_1 is integrable after updates. By induction, suppose Π_{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 Π_n is integrable after updates. □