Inference in generative models using the Wasserstein distance: supplementary materials

Espen Bernton*, Pierre E. Jacob*, Mathieu Gerber†, Christian P. Robert‡

Contents

*Department of Statistics, Harvard University, USA, ebernton@g.harvard.edu and pjacob@g.harvard.edu
†School of Mathematics, University of Bristol, UK
‡CEREMADE, Université Paris-Dauphine, PSL Research University, France, and Department of Statistics, University of Warwick, UK

1 Sequential Monte Carlo sampler for ABC posteriors

To approximate ABC posterior distributions, we apply an SMC algorithm which leads to an automatic choice of decreasing thresholds (ε_t), is parallelizable over the particles, and can leverage any MCMC kernel within the rejuvenation step. Choices of Markov kernel and adaptation rule for the thresholds are discussed in Sections 1.1 and 1.2. For dependent data, the algorithm can be readily modified to include reconstructions or curve-matching. We write $\mathfrak{D}$ for a generic distance between data sets in this section.

Let N be the number of particles, and $\varepsilon_0 = \infty$. The first step goes as follows.

1. Sample $\theta_0^1, \ldots, \theta_0^N$ from the prior π, set $w_0^k = N^{-1}$ for all $k \in 1:N$.
2. For each $k \in 1:N$, sample $z_{1:n}^k$ from $\mu_{\theta_0^k}^{(n)}$, and compute the distance $d_0^k = \mathfrak{D}(y_{1:n}, z_{1:n}^k)$.
3. Based on $(\theta_0^k)_{k=1}^N$ and $(d_0^k)_{k=1}^N$, compute the next threshold ε_1; see Section 1.2.

To approximate π^{ε_t} based on the samples approximating $\pi^{\varepsilon_{t-1}}$, perform the following for $t \geq 1$.

1. Define the weight $w_t^k \propto \mathbb{1}(d_{t-1}^k \leq \varepsilon_t)$ for all $k \in 1:N$, and normalize the weights.
2. Sample ancestor indices $a_{t-1}^1, \ldots, a_{t-1}^N$ such that $\mathbb{P}(a_{t-1}^k = j) = w_t^j$ for all $j, k \in 1:N$.
3. Sample θ_t^k from $K^{\varepsilon_t}(\theta_{t-1}^{a_{t-1}^k}, \cdot)$, a Markov kernel leaving π^{ε_t} invariant, and store the associated distance d_t^k and synthetic data set $z_{1:n}^k$, for all $k \in 1:N$; this is the rejuvenation step, see Section 1.1.
4. Based on $(\theta_t^k)_{k=1}^N$ and $(d_t^k)_{k=1}^N$, compute the next threshold ε_{t+1}; see Section 1.2.

We use systematic resampling to obtain $a_{t-1}^{1:N}$ at step t, noting that other resampling schemes exist with varying degrees of parallelism (Murray et al., 2016).

1.1 Choice of Markov kernels

Any choice of kernel K^ε leaving π^ε invariant would be valid, the simplest arguably being the kernel of Marjoram et al. (2003), i.e. a standard pseudo-marginal Metropolis–Hastings kernel. This kernel tends to be less and less effective at diversifying the particles as ε goes to zero, which can be compensated for by iterating the kernel more and more times. Instead, we use the r-hit kernel of Lee (2012), and Algorithm 6 of that paper in particular, which automatically adapts the computational budget of the kernel to the value of ε; see Lee and Łatuszyński (2014) for a theoretical comparison of MCMC kernels for ABC-type targets. Starting from a value θ and an associated distance d, the kernel works as follows, based on a desired number of hits $r \geq 2$ and a proposal distribution $g(\cdot|\theta)$ on the parameter space.

1. For $i \geq 1$, sample $\theta'^i \sim g(\cdot|\theta)$ and $z_{1:n}^i \sim \mu_{\theta'^i}^{(n)}$, until $\sum_{j=1}^i \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}^j) \leq \varepsilon) = r$. Let $K' = i$.
2. Sample an index L uniformly among $\{j \in \{1, \ldots, K'-1\} : \mathbb{1}(\mathfrak{D}(y_{1:n}, z_{1:n}^j) \leq \varepsilon)\}$.
3. For $i \geq 1$, sample $\theta^i \sim g(\cdot|\theta'^L)$ and $x_{1:n}^i \sim \mu_{\theta^i}^{(n)}$, until $\sum_{j=1}^i \mathbb{1}(\mathfrak{D}(y_{1:n}, x_{1:n}^j) \leq \varepsilon) = r-1$. Let $K = i$.
4. With probability
$$\min\left(1, \frac{\pi(\theta'^L)}{\pi(\theta)} \frac{g(\theta|\theta'^L)}{g(\theta'^L|\theta)} \frac{K}{K'-1}\right),$$
output θ'^L and the distance $\mathfrak{D}(y_{1:n}, z_{1:n}^L)$; otherwise output θ and d.

We use $r = 2$ as a default choice. The proposal distribution $g(\cdot|\theta)$ can be adapted based on the particles $(w_t^k, \theta_{t-1}^k)_{k=1}^N$ at step t, which approximate π^{ε_t}. By default, we can fit a parametric distribution on the particles, such as a Normal distribution using the empirical mean and covariance matrix of the particles. In the numerical experiments, we define $g(\cdot|\theta)$ as a mixture of 5 multivariate Normal distributions, fitted on the particles available at step t, to partially accommodate non-Gaussian features of the target.

1.2 Adaptation of thresholds

The SMC framework allows the automatic adaptation of thresholds (see e.g. Del Moral et al., 2012; Silk et al., 2013). A simple approach consists in choosing ε_{t+1} such that the effective sample size computed on the resulting weights $(w_{t+1}^k)_{k=1}^N$, defined as $\text{ESS}_{t+1} = (\sum_{k=1}^N (w_{t+1}^k)^2)^{-1}$, is above a certain value, e.g. $N/2$. We use a slightly different approach, motivated by the fact that multiple particles might be identical at any step $t \geq 1$ of the SMC sampler. Indeed, the resampling step leads to duplicate values, which the MCMC steps only partially diversify. Criteria that are based solely on the weights fail to account for this potential lack of diversity.

We define a diversity parameter α, set to 0.5 by default, which indicates the desired minimum proportion of unique particles within our sample, at all steps. The idea is to find ε_{t+1} such that, upon resampling the particles using weights $w_{t+1}^k \propto \mathbb{1}(d_t^k \leq \varepsilon_{t+1})$, we obtain at least a proportion α of unique particles. This is complicated by the randomness of the resampling step. Therefore, we clamp that randomness by drawing the uniform variables of the resampling step and keeping them fixed during the calculation of ε_{t+1}. The proportion of unique particles after resampling is then a deterministic function of ε, denoted by $f(\varepsilon)$. We could define ε_{t+1} as the value ε in the interval $[0, \varepsilon_t]$ such that $f(\varepsilon) = \alpha$. Since that equality might not be achievable, we use a numerical optimizer to minimize $|f(\varepsilon) - \alpha|$ over $\varepsilon \in [0, \varepsilon_t]$.

As a result, the sequence of thresholds (ε_t) decreases, but only if the MCMC steps have managed to diversify the particles. We can let the SMC sampler run indefinitely, saving the particles and distances obtained at every step t. We can decide to stop the algorithm if the sequence of thresholds (ε_t) does not decrease anymore, or if the cost of each step becomes prohibitive. We can otherwise run the algorithm for a given wall-clock time, or a given number of model simulations.

2 Theoretical properties: proofs and more

2.1 Preliminary results

A sequence of probability measures $(\mu_n)_{n\geq 1}$ is said to converge weakly in $\mathcal{P}_p(\mathcal{Y})$ to μ as $n \to \infty$ if $\mu_n \Rightarrow \mu$, i.e. converges weakly in the usual sense, and there exists $y_0 \in \mathcal{Y}$ such that $\int_{\mathcal{Y}} \rho(y, y_0)^p d\mu_n(y) \to \int_{\mathcal{Y}} \rho(y, y_0)^p d\mu(y)$.

Theorem 2.1. *The p-Wasserstein distance metrizes weak convergence in $\mathcal{P}_p(\mathcal{Y})$: a sequence μ_n converges weakly to μ in $\mathcal{P}_p(\mathcal{Y})$ if and only if $\mathfrak{W}_p(\mu_n, \mu) \to 0$.*

For a proof, see Villani (2008, Theorem 6.9). From this result one can deduce the continuity of the p-Wasserstein distance. If μ_n and ν_n converge weakly in $\mathcal{P}_p(\mathcal{Y})$ to μ and ν, then $\mathfrak{W}_p(\mu_n, \nu_n) \to \mathfrak{W}_p(\mu, \nu)$. On the other hand, if μ_n and ν_n converge weakly in the usual sense, the Wasserstein distance is only lower semicontinuous: $\liminf_{n\to\infty} \mathfrak{W}_p(\mu_n, \nu_n) \geq \mathfrak{W}_p(\mu, \nu)$. The following lemma is extended from Bassetti et al. (2006). Its second condition corresponds to Assumption 2.2, and is implied by the first condition.

Lemma 2.1. *Let $(\theta_n)_{n\geq 1}$ be a sequence in $\mathcal{H}$ and $\theta \in \mathcal{H}$. Suppose either of the following conditions holds.*

1. $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$ *implies that* $\mathfrak{W}_p(\mu_{\theta_n}, \mu_\theta) \to 0$.

2. $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$ *implies that* $\mu_{\theta_n} \Rightarrow \mu_\theta$.

Then, respectively,

1. $\mathcal{H} \times \mathcal{P}_p(\mathcal{Y}) \ni (\theta, \mu) \mapsto \mathfrak{W}_p(\mu_\theta, \mu)$ *is continuous.*

2. $\mathcal{H} \times \mathcal{P}(\mathcal{Y}) \ni (\theta, \mu) \mapsto \mathfrak{W}_p(\mu_\theta, \mu)$ *is lower semicontinuous.*

Proof. This follows directly from the two assumptions and the continuity/lower semicontinuity derived from Theorem 2.1. □

Lemma 2.2. *Suppose* $\mu_k^{(m)} \Rightarrow \mu^{(m)}$. *There exists a probability space and versions of* $\hat{\mu}_{k,m}$ *and* $\hat{\mu}_m$, *denoting empirical distributions based on draws from* $\mu_k^{(m)}$ *and* $\mu^{(m)}$, *such that* $\hat{\mu}_{k,m} \Rightarrow \hat{\mu}_m$ *almost surely.*

Proof. By Skorokhod's representation theorem, there exists a probability space $(\tilde{\mathbb{P}}, \tilde{\Omega}, \tilde{\Sigma})$ and random variables $\tilde{X}_k^{1:m} \sim \mu_k^{(m)}$ and $\tilde{X}^{1:m} \sim \mu^{(m)}$ such that $\tilde{X}_k^{1:m} \to \tilde{X}^{1:m}$ $\tilde{\mathbb{P}}$-almost surely. Let $\hat{\mu}_{k,m}$ and $\hat{\mu}_m$ be the empirical distributions of these samples. By Varadarajan (1958b) and since $\mathcal{Y}$ is separable, there exists a fixed countable subset $C^\star$ of continuous and bounded functions on $\mathcal{Y}$, such that for any sequence of measures $\mu_n \in \mathcal{P}(\mathcal{Y})$, μ_n converges weakly to μ if and only if $\int f d\mu_n \to \int f d\mu$ for all $f \in C^\star$. Fix one such f. Then,

$$\int f d\hat{\mu}_{k,m} = \frac{1}{m}\sum_{i=1}^{m} f(\tilde{X}_k^i) \to \frac{1}{m}\sum_{i=1}^{m} f(\tilde{X}^i) = \int f d\hat{\mu}_m,$$

on a set of $\tilde{\mathbb{P}}$-probability one, by the continuous mapping theorem. Taking the countable intersection of these sets over $f \in C^\star$, we get that $\hat{\mu}_{k,m} \Rightarrow \hat{\mu}_m$ $\tilde{\mathbb{P}}$-almost surely. □

Lemma 2.3. *The function* $(\nu, \mu^{(m)}) \mapsto \mathbb{E}\mathfrak{W}_p(\nu, \hat{\mu}_m)$ *is lower semicontinuous with respect to weak convergence. Furthermore, if* $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$ *implies that* $\mu_{\theta_n}^{(m)} \Rightarrow \mu_\theta^{(m)}$, *then the map* $(\nu, \theta) \mapsto \mathbb{E}\mathfrak{W}_p(\nu, \hat{\mu}_{\theta,m})$ *is lower semicontinuous.*

Proof. Let $\mu_k^{(m)} \Rightarrow \mu^{(m)}$ and $\nu_k \Rightarrow \nu$. Then there exist versions of the corresponding empirical measures such that $\hat{\mu}_{k,m} \Rightarrow \hat{\mu}_m$ almost surely, by Lemma 2.2. By the lower semicontinuity of the p-Wasserstein distance and Fatou's lemma,

$$\mathbb{E}\mathfrak{W}_p(\nu, \hat{\mu}_m) \le \mathbb{E}\liminf_{k\to\infty} \mathfrak{W}_p(\nu_k, \hat{\mu}_{k,m}) \le \liminf_{k\to\infty} \mathbb{E}\mathfrak{W}_p(\nu_k, \hat{\mu}_{k,m}).$$

The lower semicontinuity of $(\nu, \theta) \mapsto \mathbb{E}\mathfrak{W}_p(\nu, \hat{\mu}_{\theta,m})$ is proved analogously to Lemma 2.1. □

2.2 MWE

2.2.1 Existence, measurability and consistency

Assumption 2.1. *The data-generating process is such that* $\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \to 0$, $\mathbb{P}$-*almost surely.*

Assumption 2.2. *The map* $\theta \mapsto \mu_\theta$ *is continuous in the sense that* $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$ *implies* $\mu_{\theta_n} \Rightarrow \mu_\theta$.

For the next assumption, define $\varepsilon_\star$ as $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$; we will use this definition throughout.

Assumption 2.3. *For some* $\varepsilon > 0$, *the set* $B_\star(\varepsilon) = \{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) \le \varepsilon_\star + \varepsilon\}$ *is bounded.*

Theorem 2.2 (Existence and consistency of the MWE)**.** *Under Assumptions 2.1-2.3, there exists a set* $E \subset \Omega$ *with* $\mathbb{P}(E) = 1$ *such that, for all* $\omega \in E$, $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$*, and there exists* $n(\omega)$ *such that, for all* $n \geq n(\omega)$*, the sets* $\text{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ *are non-empty and form a bounded sequence with*

$$\limsup_{n\to\infty} \underset{\theta\in\mathcal{H}}{\text{argmin}}\, \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \subset \underset{\theta\in\mathcal{H}}{\text{argmin}}\, \mathfrak{W}_p(\mu_\star, \mu_\theta).$$

Before giving the proof, we recall a definition and a proposition.

Definition 2.1. A sequence of functions $f_n : \mathcal{H} \to \mathbb{R}$ is said to epi-converge to $f : \mathcal{H} \to \mathbb{R}$ if for all $\theta \in \mathcal{H}$,

$$\begin{cases} \liminf_{n\to\infty} f_n(\theta_n) \geq f(\theta) & \text{for every sequence } \theta_n \to \theta, \\ \limsup_{n\to\infty} f_n(\theta_n) \leq f(\theta) & \text{for some sequence } \theta_n \to \theta. \end{cases}$$

A useful equivalent formulation of epi-convergence can be found in Proposition 7.29 of Rockafellar and Wets (2009), paraphrased here.

Proposition 2.1 (Proposition 7.29 of Rockafellar and Wets (2009))**.** *The sequence* $f_n : \mathcal{H} \to \mathbb{R}$ *epi-converges to* $f : \mathcal{H} \to \mathbb{R}$ *if and only if*

$$\begin{cases} \liminf_{n\to\infty} \inf_{\theta\in\mathcal{K}} f_n(\theta) \geq \inf_{\theta\in\mathcal{K}} f(\theta) & \textit{for every compact set } \mathcal{K} \subset \mathcal{H}, \\ \limsup_{n\to\infty} \inf_{\theta\in\mathcal{O}} f_n(\theta) \leq \inf_{\theta\in\mathcal{O}} f(\theta) & \textit{for every open set } \mathcal{O} \subset \mathcal{H}. \end{cases}$$

In an colloquial sense, epi-convergence is the weakest notion of convergence for which the minimizer of f_n converges to the minimizer of f. Showing that the function $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$ epi-converges to $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$ almost surely is the key step in the proof of Theorem 2.2.

Proof of Theorem 2.2. First note that, for any $\nu \in \mathcal{P}(\mathcal{Y})$, the lower semicontinuity of the map $\theta \mapsto \mathfrak{W}_p(\nu, \mu_\theta)$ follows from Lemma 2.1, via Assumption 2.2. Since $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta) = \varepsilon_\star$, the set $B_\star(\varepsilon)$ with the ε of Assumption 2.3 is non-empty, by definition of the infimum. Moreover, since $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$ is lower semicontinuous, the set $B_\star(\varepsilon)$ is closed. By Assumption 2.3, $B_\star(\varepsilon)$ is therefore compact. In other words, again by lower semicontinuity, the set $\text{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$ is non-empty.

We now show that $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$ epi-converges to $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$ $\mathbb{P}$-almost surely. Let E denote the set of probability one from Assumption 2.1 and let $\omega \in E$. Fix $\mathcal{K} \subset \mathcal{H}$ compact. By lower semicontinuity of $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$, we know that $\inf_{\theta\in\mathcal{K}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) = \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n})$, for some sequence $\theta_n = \theta_n(\omega) \in \mathcal{K}$. Hence,

$$\begin{aligned} \liminf_{n\to\infty} \inf_{\theta\in\mathcal{K}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) &= \liminf_{n\to\infty} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n}) \\ &= \lim_{k\to\infty} \mathfrak{W}_p(\hat{\mu}_{n_k}(\omega), \mu_{\theta_{n_k}}) \quad \exists \text{ subsequence converging to the } \liminf, \\ &= \lim_{m\to\infty} \mathfrak{W}_p(\hat{\mu}_{n_{k_m}}(\omega), \mu_{\theta_{n_{k_m}}}) \quad \exists \text{ subsequence } \theta_{n_{k_m}} \to \bar{\theta} \in \mathcal{K} \text{ by compactness}, \\ &= \liminf_{m\to\infty} \mathfrak{W}_p(\hat{\mu}_{n_{k_m}}(\omega), \mu_{\theta_{n_{k_m}}}) \\ &\geq \mathfrak{W}_p(\mu_\star, \mu_{\bar{\theta}}) \quad \text{by l.s.c., Assumptions 2.1 and 2.2 }, \omega \in E, \\ &\geq \inf_{\theta\in\mathcal{K}} \mathfrak{W}_p(\mu_\star, \mu_\theta). \end{aligned}$$

Fix $\mathcal{O} \subset \mathcal{H}$ open. By definition of the infimum, there exists a sequence $\theta_n \in \mathcal{O}$ such that $\mathfrak{W}_p(\mu_\star, \mu_{\theta_n}) \to \inf_{\theta\in\mathcal{O}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$. Now, $\inf_{\theta\in\mathcal{O}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n})$. Hence,

$$\begin{aligned}
\limsup_{n\to\infty} \inf_{\theta\in\mathcal{O}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) &\leq \limsup_{n\to\infty} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n}) \\
&\leq \limsup_{n\to\infty} \left(\mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathfrak{W}_p(\mu_\star, \mu_{\theta_n})\right) \quad \text{by the triangle inequality,} \\
&\leq \limsup_{n\to\infty} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\star) + \limsup_{n\to\infty} \mathfrak{W}_p(\mu_\star, \mu_{\theta_n}) \quad \text{by positivity,} \\
&= \limsup_{n\to\infty} \mathfrak{W}_p(\mu_\star, \mu_{\theta_n}) \quad \text{by Assumption 2.1, } \omega \in E, \\
&= \inf_{\theta\in\mathcal{O}} \mathfrak{W}_p(\mu_\star, \mu_\theta) \quad \text{by definition of } \theta_n.
\end{aligned}$$

Using Proposition 2.1, the sequence of functions $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ epi-converges to $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$.

Theorem 7.29b) of Rockafellar and Wets (2009) implies that $\limsup_{n\to\infty}(\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)) \leq \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta) = \varepsilon_\star$. So, for all $\alpha > 0$, there exists $n_\alpha(\omega)$, such that for $n \geq n_\alpha(\omega)$, $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\star + \alpha$. Let $\alpha \in (0, \varepsilon/2)$. The set $\{\theta \in \mathcal{H} : \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\star + \varepsilon/2\}$ is non-empty for $n \geq n_\alpha(\omega)$, by definition of the infimum. Let θ belong to this set. Then, by the triangle inequality, $\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$. By Assumption 2.1, there exists an $n_\varepsilon(\omega)$ such that for $n \geq n_\varepsilon(\omega)$, $\mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\star) \leq \varepsilon/2$. So, if $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega)\}$, we have that $\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_\star + \varepsilon$. This means that $\{\theta \in \mathcal{H} : \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\star + \varepsilon/2\} \subset B_\star(\varepsilon)$. As a consequence, for $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega)\}$, $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta\in B_\star(\varepsilon)} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$.

By Theorem 7.31a) of Rockafellar and Wets (2009), this implies $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$. Also, for $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega)\}$ and by the same reasoning as for the map $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$, the sets $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ are non-empty. By Theorem 7.31b) of Rockafellar and Wets (2009), the result follows. The same argument holds for $\varepsilon_n\text{-}\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ with $\varepsilon_n \to 0$, since, eventually, $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta) + \varepsilon_n \leq \varepsilon_\star + \alpha$.

□

Theorem 2.3 (Measurability of the MWE). *Suppose that $\mathcal{H}$ is a σ-compact Borel measurable subset of $\mathbb{R}^{d_\theta}$. Under Assumption 2.2, for any $n \geq 1$ and $\varepsilon > 0$, there exists a Borel measurable function $\hat{\theta}_n : \Omega \to \mathcal{H}$ that satisfies*

$$\hat{\theta}_n(\omega) \in \begin{cases} \operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta), & \textit{if this set is non-empty,} \\ \varepsilon\text{-}\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\theta), & \textit{otherwise.} \end{cases}$$

Before the proof, we first recall a useful result from Brown and Purves (1973), also used in Bassetti et al. (2006).

Theorem 2.4 (Corollary 1 in Brown and Purves (1973)). *Let X, Y be Polish, $D \subset Y \times X$ be Borel, and $f : D \to \mathbb{R}$ be Borel measurable. Suppose that for all $y \in proj(D)$, the section $D_y = \{x : (y, x) \in D\}$ is σ-compact and that $f_y = f(y, \cdot)$ is lower semicontinuous with respect to the relative topology on D_y. Then*

1. *The sets $G = proj(D)$ and $I = \{y \in G : \text{for some } x \in D_y, f(y, x) = \inf f_y\}$ are Borel.*

2. *For each $\varepsilon > 0$, there exists a Borel measurable function ϕ_ε such that for $y \in G$,*

$$f(y, \phi_\varepsilon(y)) \begin{cases} = \inf f_y & \text{if } y \in I. \\ \leq \varepsilon + \inf f_y & \text{if } y \notin I, \inf f_y \neq -\infty. \\ \leq -\varepsilon^{-1} & \text{if } y \notin I, \inf f_y = -\infty. \end{cases}$$

Proof of Theorem 2.3. First note that $\mathcal{Y}^\infty$ endowed with the product topology is Polish since $(\mathcal{Y}, \rho)$ is Polish. Also, $\hat{\mu}_n(\omega)$ depends on ω only through $y = Y(\omega)$, where $Y = (Y_t)_{t\in\mathbb{Z}}$. We can therefore write $\hat{\mu}_n(\omega) = \hat{\mu}_n(y)$, where $y \in \mathcal{Y}^\infty$, and consider the empirical measure a function on $\mathcal{Y}^\infty$. The map $y \mapsto \hat{\mu}_n(y)$ is measurable with respect to the Borel σ-algebra on $\mathcal{P}_p(\mathcal{Y})$ with respect to weak convergence. Recall also that $(\mathcal{P}_p(\mathcal{Y}), \mathfrak{W}_p)$ is Polish since $\mathcal{Y}$ is Polish by Theorem 6.18 of Villani (2008).

Let $\mathcal{D} = \mathcal{Y}^\infty \times \mathcal{H}$. By Lemma 2.1 and Assumption 2.2, the map $(\mu, \theta) \mapsto \mathfrak{W}_p(\mu, \mu_\theta)$ is lower semicontinuous (and therefore measurable). Hence the map $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n(y), \mu_\theta)$ is also lower semicontinuous on $\mathcal{H}$ for any $y \in \mathcal{Y}^\infty$. Being the composition of measurable functions, $(y, \theta) \mapsto \mathfrak{W}_p(\hat{\mu}_n(y), \mu_\theta)$ is measurable on $\mathcal{D}$. In light of this, the result follows by a direct application of Theorem 2.4.

□

2.2.2 Asymptotic distribution

Let $p = 1$, $\mathcal{Y} = \mathbb{R}$, and $\rho(x, y) = |x - y|$. This implies that $\mathfrak{W}_1(\mu, \nu) = \int_0^1 |F_\mu^{-1}(s) - F_\nu^{-1}(s)| ds = \int_\mathbb{R} |F_\mu(t) - F_\nu(t)| dt$, where F_μ and F_ν denote the cumulative distribution functions (CDFs) of μ and ν respectively (see e.g. Ambrosio et al., 2005, Theorem 6.0.2). Also assume that $\mathcal{H}$ is endowed with a norm: $\rho_\mathcal{H}(\theta, \theta') = \|\theta - \theta'\|_\mathcal{H}$, and that $\mu_\star = \mu_{\theta_\star}$ for some $\theta_\star$ in the interior of $\mathcal{H}$. Recall the following assumption, stated in a more general form than what we require here.

Assumption 2.4. *For all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\inf_{\theta \in \mathcal{H} : \rho_\mathcal{H}(\theta, \theta_\star) \geq \varepsilon} \mathfrak{W}_p(\mu_\star, \mu_\theta) > \mathfrak{W}_p(\mu_\star, \mu_{\theta_\star}) + \delta.$$

Definition 2.2. Suppose the sequence $\Omega \times \mathbb{R} \ni (\omega, t) \mapsto X_n(\omega, t)$ and $\Omega \times \mathbb{R} \ni (\omega, t) \mapsto X(\omega, t)$ are stochastic processes with almost all their sample paths in $L_1(\mathbb{R})$. Then X_n converges weakly to X in $L_1(\mathbb{R})$ if $\mathbb{E} f(X_n) \to \mathbb{E} f(X)$ as $n \to \infty$ for all bounded continuous functions $f : L_1(\mathbb{R}) \to \mathbb{R}$.

Definition 2.3. The stochastic process $\Omega \times \mathbb{R} \ni (\omega, t) \mapsto G_\mu(\omega, t)$ is a μ-Brownian bridge if it is a zero mean Gaussian process with covariance function $\mathbb{E} G_\mu(s) G_\mu(t) = \min\{F_\mu(s), F_\mu(t)\} - F_\mu(s) F_\mu(t)$.

Theorem 2.5 (Theorem 2.1a in del Barrio et al. (1999))**.** *Let $Y = (Y_t)_{t\in\mathbb{Z}} \sim \mu_\star^\infty$, and define $F_n(\omega, t) = \hat{\mu}_n(\omega)(-\infty, t]$ and $F_\star(t) = \mu_\star(-\infty, t]$. The stochastic process $\sqrt{n}(F_n - F_\star)$ converges weakly in $L_1(\mathbb{R})$ to $G_\star$, where $G_\star$ is a $\mu_\star$-Brownian bridge, if and only if $\int_0^\infty \sqrt{\mathbb{P}(|Y_0| > t)} dt < \infty$.*

For a stationary sequence, let $\tilde{\alpha}_t = \sup_{u \in \mathbb{R}} \mathbb{E} |\mathbb{P}(Y_t \leq u | \mathcal{F}_{-\infty}^0) - \mathbb{P}(Y_t \leq u)|$. Note that for stationary sequences, $\tilde{\alpha}$-mixing is weaker than α-mixing (defined in Section 2.6).

Theorem 2.6 (Proposition 3.5 in Dede (2009))**.** *Suppose $Y = (Y_t)_{t\in\mathbb{Z}}$ is ergodic and stationary, and that*

$$\sum_{k \geq 1} \frac{1}{\sqrt{k}} \int_0^\infty \min\{\sqrt{\tilde{\alpha}_k}, \sqrt{\mathbb{P}(|Y_0| > t)}\} dt < \infty.$$

Then $\sqrt{n}(F_n - F_\star)$ converges weakly in $L_1(\mathbb{R})$ to a zero mean Gaussian process $G_\star$ with sample paths in $L_1(\mathbb{R})$ and covariance satisfying: for every $f, g \in L_\infty(\mathbb{R})$,

$$\mathbb{E}f(G_\star)g(G_\star) = \int_{\mathbb{R}^2} f(s)g(u)C(s,u)dsdu,$$

where

$$C(s,u) = \sum_{t \in \mathbb{Z}} \left\{ \mathbb{P}(X_0 \leq s, X_t \leq u) - F_\star(s)F_\star(u) \right\}.$$

Remark 2.1. Dede (2009) also provides other conditions, e.g. on ϕ-mixing coefficients, for which the convergence above holds.

The following results contain Theorem 5.3 of the main text as a special case.

Theorem 2.7. *Suppose $\mu_\star = \mu_{\theta_\star}$ for some $\theta_\star$ in the interior of $\mathcal{H}$, and that either the conditions of Theorem 2.5 or Theorem 2.6 are satisfied. Suppose that there exists a non-singular $D_\star \in (L_1(\mathbb{R}))^{d_\theta}$ such that*

$$\int_{\mathbb{R}} |F_\theta(t) - F_\star(t) - \langle \theta - \theta_\star, D_\star(t) \rangle| dt = o(\|\theta - \theta_\star\|_{\mathcal{H}}),$$

as $\|\theta - \theta_\star\|_{\mathcal{H}} \to 0$. Under Assumptions 2.1-2.4, the goodness-of-fit statistic satisfies

$$\sqrt{n} \inf_{\theta \in \mathcal{H}} \mathfrak{W}_1(\hat{\mu}_n, \mu_\theta) \Rightarrow \inf_{u \in \mathcal{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_\star(t) \rangle| dt,$$

as $n \to \infty$, where $G_\star$ is given as in Theorem 2.5 or Theorem 2.6 respectively.

Theorem 2.8. *Suppose that the conditions in Theorem 2.7 hold. Suppose also that the random map $\mathcal{H} \ni u \mapsto \int_{\mathbb{R}} |G_\star(t) - \langle u, D_\star(t) \rangle| dt$ has an almost surely unique infimum. Then the MWE of order 1 satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \Rightarrow \operatorname*{argmin}_{u \in \mathcal{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_\star(t) \rangle| dt,$$

as $n \to \infty$, where $G_\star$ is given as in Theorem 2.5 or Theorem 2.6.

Proof. The proofs of these two results follow the steps outlined in Pollard (1980)'s Theorem 4.2 and 7.2 respectively, which also generalize to the setting where the map $\mathcal{H} \ni u \mapsto \int_{\mathbb{R}} |G_\star(t) - \langle u, D_\star(t) \rangle| dt$ does not necessarily have a unique minimum. The delta methods employed therein hold for the 1-Wasserstein distance after noticing that the representation in Theorem 6.0.2 of Ambrosio et al. (2005) embeds the 1-Wasserstein space isometrically in $L_1(\mathbb{R})$ under its canonical norm. Moreover, the well-separation of $\theta_\star$ provided by Assumption 2.4, the consistency and measurability of the MWE proved earlier, and Theorems 2.5 and 2.6 proved in del Barrio et al. (1999) and Dede (2009) respectively, guarantee that Pollard's conditions are satisfied. Take note that the measurability concerns outlined in his Section 3 do not apply to here, as $L_1(\mathbb{R})$ is separable. Now, with only minor changes in notation (most importantly, apply Pollard's steps to $\tilde{F}_\theta = F_\theta - F_\star$ and $\tilde{F}_n = F_n - F_\star$), the proof follows. □

The differentiability condition can sometimes be established from more familiar concepts of differentiability, such as differentiability in quadratic mean (Le Cam, 1970). The following proposition gives such a result. Suppose that the model family is absolutely continuous with respect to the Lebesgue measure λ on

$\mathbb{R}$, and denote the density $d\mu_\theta/d\lambda$ of μ_θ by f_θ. Let $\xi_\theta(y) = \sqrt{f_\theta(y)}$ for all $y \in \mathbb{R}$. Le Cam (1970) introduced the concept of differentiability in quadratic mean, which we define below.

Definition 2.4. The model $\mathcal{M}$ is differentiable in quadratic mean at $\theta_\star$ if there exists $\dot{\xi}_{\theta_\star} \in (L_2(\mathbb{R}))^{d_\theta}$ and $R_{\theta-\theta_\star} \in (L_2(\mathbb{R}))^{d_\theta}$ such that $\xi_\theta = \xi_{\theta_\star} + \langle \theta - \theta_\star,\ \dot{\xi}_{\theta_\star} \rangle + R_{\theta-\theta_\star}$, where $[\int_{\mathbb{R}} R^2_{\theta-\theta_\star}(y)dy]^{1/2} = o(\|\theta - \theta_\star\|_{\mathcal{H}})$ as $\|\theta - \theta_\star\|_{\mathcal{H}} \to 0$.

Differentiability in quadratic mean holds for many classical models, such as exponential families and many location-scale families (see e.g. Section 12.2 in Lehmann and Romano (2005)).

Proposition 2.2. *Suppose that the model family is supported on a set $S \subset \mathbb{R}$ of bounded Lebesgue measure, and that it is differentiable in quadratic mean at $\theta_\star$. Let $D_\star(t) = \int_{-\infty}^t 2\xi_{\theta_\star}(y)\dot{\xi}_{\theta_\star}(y)dy$ for $t \in S$ and zero elsewhere. Then $\int_{\mathbb{R}} |F_\theta(t) - F_\star(t) - \langle \theta - \theta_\star, D_\star(t)\rangle| dt = o(\|\theta - \theta_\star\|_{\mathcal{H}})$, as $\|\theta - \theta_\star\|_{\mathcal{H}} \to 0$.*

Proof.

$$
\begin{aligned}
&\int_{\mathbb{R}} |F_\theta(t) - F_\star(t) - \langle \theta - \theta_\star, D_\star(t)\rangle| dt \\
&= \int_S \left| \int_{-\infty}^t \xi_\theta^2(y) - \xi_{\theta_\star}^2(y) - 2\xi_{\theta_\star}(y)\langle \theta - \theta_\star, \dot{\xi}_{\theta_\star}(y)\rangle dy \right| dt \\
&\le \int_S \int_{\mathbb{R}} |\xi_\theta^2(y) - \xi_{\theta_\star}^2(y) - 2\xi_{\theta_\star}(y)\langle \theta - \theta_\star, \dot{\xi}_{\theta_\star}(y)\rangle| dy dt \\
&\le c \int_{\mathbb{R}} \langle \theta - \theta_\star, \dot{\xi}_{\theta_\star}(y)\rangle^2 + R^2_{\theta-\theta_\star}(y) + 2|\xi_{\theta_\star}(y) R_{\theta-\theta_\star}(y)| + 2|\langle \theta - \theta_\star, \dot{\xi}_{\theta_\star}(y)\rangle R_{\theta-\theta_\star}(y)| dy \\
&= o(\|\theta - \theta_\star\|_{\mathcal{H}}),
\end{aligned}
$$

where c is some constant and the last equality follows by applying the Cauchy-Schwarz inequality to the two last terms of the integrand. □

2.3 MEWE

In order to show similar results for the MEWE, we introduce the following assumptions.

Assumption 2.5. *For any $m \ge 1$, if $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$, then $\mu_{\theta_n}^{(m)} \Rightarrow \mu_\theta^{(m)}$.*

Assumption 2.6. *If $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$, then $\mathbb{E}\mathfrak{W}_p(\mu_{\theta_n}, \hat{\mu}_{\theta_n,n}) \to 0$.*

Assumption 2.5 is a slightly stronger version of Assumption 2.2, stating that we not only need weak convergence of the "model" distributions μ_θ, but also of the sample distributions $\mu_\theta^{(m)}$ for any $m \ge 1$. Assumption 2.6 is implied by $\sup_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) \to 0$, which in turn might hold when $\mathcal{H}$ is compact and the inequalities in Fournier and Guillin (2015) hold. In the next result, we prove an analogous version of Theorem 2.2 for the MEWE as $\min\{n, m\} \to \infty$. For simplicity, we write m as a function of n and require that $m(n) \to \infty$ as $n \to \infty$.

Theorem 2.9. *Under Assumptions 2.1-2.3 and 2.5-2.6, there exists a set $E \subset \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \to \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$, and there exists $n(\omega)$ such that, for all $n \ge n(\omega)$, the sets $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$ are non-empty and form a bounded sequence with*

$$
\limsup_{n\to\infty} \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \subset \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta).
$$

Proof of Theorem 2.9. As before, for any $\nu \in \mathcal{P}(\mathcal{Y})$, lower semicontinuity of the map $\theta \mapsto \mathfrak{W}_p(\nu, \mu_\theta)$ follows from Lemma 2.1, via Assumption 2.2. Since $\inf_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta) = \varepsilon_\star$, $B_\star(\varepsilon)$ with the ε of Assumption 2.3 is non-empty, by definition of the infimum. Moreover, since $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$ is lower semicontinuous, the set $B_\star(\varepsilon)$ is closed. By Assumption 2.3, $B_\star(\varepsilon)$ is therefore compact. In other words, again by lower semicontinuity, the set $\operatorname{argmin}_{\theta \in \mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$ is non-empty.

We show that $\theta \mapsto \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m(n)})$ epi-converges to $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$ $\mathbb{P}$-almost surely. Let E denote the set of probability one from Assumption 2.1 and let $\omega \in E$. Fix $\mathcal{K} \subset \mathcal{H}$ compact. By lower semicontinuity of $\theta \mapsto \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$, ensured by Lemma 2.3 and Assumption 2.5, we know that $\inf_{\theta \in \mathcal{K}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) = \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)})$, for some sequence $\theta_n = \theta_n(\omega) \in \mathcal{K}$. Hence,

$$
\begin{aligned}
&\liminf_{n\to\infty} \inf_{\theta\in\mathcal{K}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \\
&\quad = \liminf_{n\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)}) \\
&\quad = \lim_{k\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_{n_k}(\omega), \hat{\mu}_{\theta_{n_k},m(n_k)}) \quad \exists \text{ subsequence converging to the } \liminf, \\
&\quad = \lim_{\ell\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_{n_{k_\ell}}(\omega), \hat{\mu}_{\theta_{n_{k_\ell}},m(n_{k_\ell})}) \quad \exists \text{ subsequence } \theta_{n_{k_\ell}} \to \bar{\theta} \in \mathcal{K} \text{ by compactness}, \\
&\quad = \liminf_{\ell\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_{n_{k_\ell}}(\omega), \hat{\mu}_{\theta_{n_{k_\ell}},m(n_{k_\ell})}) \\
&\quad \geq \liminf_{\ell\to\infty} [\mathfrak{W}_p(\hat{\mu}_{n_{k_\ell}}(\omega), \mu_{\theta_{n_{k_\ell}}}) - \mathbb{E}\mathfrak{W}_p(\mu_{\theta_{n_{k_\ell}}}, \hat{\mu}_{\theta_{n_{k_\ell}},m(n_{k_\ell})})] \quad \text{by the triangle inequality}, \\
&\quad \geq \liminf_{\ell\to\infty} \mathfrak{W}_p(\hat{\mu}_{n_{k_\ell}}(\omega), \mu_{\theta_{n_{k_\ell}}}) - \limsup_{\ell\to\infty} \mathbb{E}\mathfrak{W}_p(\mu_{\theta_{n_{k_\ell}}}, \hat{\mu}_{\theta_{n_{k_\ell}},m(n_{k_\ell})}) \\
&\quad \geq \mathfrak{W}_p(\mu_\star, \mu_{\bar{\theta}}) \quad \text{by l.s.c., Assumptions 2.1, 2.2 and 2.6, } \omega \in E, \\
&\quad \geq \inf_{\theta\in\mathcal{K}} \mathfrak{W}_p(\mu_\star, \mu_\theta).
\end{aligned}
$$

Fix $\mathcal{O} \subset \mathcal{H}$ open. By definition of the infimum, there exists a sequence $\theta_n \in \mathcal{O}$ such that $\mathfrak{W}_p(\mu_\star, \mu_{\theta_n}) \to \inf_{\theta\in\mathcal{O}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$. Now, $\inf_{\theta\in\mathcal{O}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \leq \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)})$. Hence,

$$
\begin{aligned}
&\limsup_{n\to\infty} \inf_{\theta\in\mathcal{O}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \\
&\quad \leq \limsup_{n\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)}) \\
&\quad \leq \limsup_{n\to\infty} [\mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathfrak{W}_p(\mu_\star, \mu_{\theta_n}) + \mathbb{E}\mathfrak{W}_p(\mu_{\theta_n}, \hat{\mu}_{\theta_n,m(n)})] \quad \text{by the triangle inequality}, \\
&\quad = \limsup_{n\to\infty} \mathfrak{W}_p(\mu_\star, \mu_{\theta_n}) \quad \text{by Assumptions 2.1 and 2.6, } \omega \in E, \\
&\quad = \inf_{\theta\in\mathcal{O}} \mathfrak{W}_p(\mu_\star, \mu_\theta) \quad \text{by definition of } \theta_n.
\end{aligned}
$$

Theorem 7.29b) of Rockafellar and Wets (2009) implies that $\limsup_{n\to\infty}(\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})) \leq \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta) = \varepsilon_\star$. Hence, for all $\alpha > 0$, there exists $n_\alpha(\omega)$, such that $n \geq n_\alpha(\omega)$ implies that $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \leq \varepsilon_\star + \alpha$. Let $\alpha \in (0, \varepsilon/3)$. The set $\{\theta \in \mathcal{H} : \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \leq \varepsilon_\star + \varepsilon/3\}$ is non-empty for $n \geq n_\alpha(\omega)$, by definition of the infimum. Let θ belong to this set. Then, by the triangle inequality, $\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) + \mathbb{E}\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,m(n)})$. By Assumption 2.1, there exists an $n_\varepsilon(\omega)$ such that for $n \geq n_\varepsilon(\omega)$, $\mathfrak{W}_p(\hat{\mu}_n(\omega), \mu_\star) \leq \varepsilon/3$. By Assumption 2.6, there exists an $\hat{n}(\omega)$ such that for $n \geq \hat{n}(\omega)$, $\mathbb{E}\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,m(n)}) \leq \varepsilon/3$. So, if $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega), \hat{n}(\omega)\}$, we have that $\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_\star + \varepsilon$. This means that $\{\theta \in \mathcal{H} : \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \leq \varepsilon_\star + \varepsilon/3\} \subset B_\star(\varepsilon)$. As a consequence, for $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega), \hat{n}(\omega)\}$, $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) = \inf_{\theta\in B_\star(\varepsilon)} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$.

By Theorem 7.31a) of Rockafellar and Wets (2009), this implies that $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \to \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\mu_\star, \mu_\theta)$. Also, for $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega), \hat{n}(\omega)\}$ and by the same reasoning as for the map $\theta \mapsto \mathfrak{W}_p(\mu_\star, \mu_\theta)$, the sets $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$ are non-empty. By Theorem 7.31b) of Rockafellar and Wets (2009), the result follows. The same argument holds for ε_n- $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$ with $\varepsilon_n \to 0$, since, eventually, $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) + \varepsilon_n \leq \varepsilon_\star + \alpha$.

□

The next result considers the case where the data and n is fixed, while $m \to \infty$. It shows that the MEWE converges to the MWE, assuming the latter exists. We summarize this condition in the following assumption, in which the observed empirical distribution is kept fixed and $\varepsilon_n = \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$.

Assumption 2.7. *For some $\varepsilon > 0$, the set $B_n(\varepsilon) = \{\theta \in \mathcal{H} : \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta) \leq \varepsilon_n + \varepsilon\}$ is bounded.*

Theorem 2.10. *Under Assumptions 2.2 and 2.5-2.7, $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \to \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$, and there exists an $\hat{m}$ such that, for all $m \geq \hat{m}$, the sets $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$ are non-empty and form a bounded sequence with*

$$\limsup_{m\to\infty} \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \subset \operatorname*{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta).$$

Proof of Theorem 2.10. Lower semicontinuity of the map $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$ follows from Lemma 2.1, via Assumption 2.2. Since $\inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta) = \varepsilon_n$, $B_n(\varepsilon)$ with the ε of Assumption 2.3 is non-empty, by definition of the infimum. Moreover, since $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$ is lower semicontinuous, the set $B_n(\varepsilon)$ is closed. By Assumption 2.7, $B_n(\varepsilon)$ is therefore compact. In other words, again by lower semicontinuity, the set $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$ is non-empty.

We show that $\theta \mapsto \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$ epi-converges to $\theta \mapsto \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$ as $m \to \infty$. Fix $\mathcal{K} \subset \mathcal{H}$ compact. By lower semicontinuity of $\theta \mapsto \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$, ensured by Lemma 2.3 and Assumption 2.5, we know that $\inf_{\theta\in\mathcal{K}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) = \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_m,m})$, for some sequence $\theta_m \in \mathcal{K}$. Hence,

$$\begin{aligned}
&\liminf_{m\to\infty} \inf_{\theta\in\mathcal{K}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \\
&\quad = \liminf_{m\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta_m,m}) \\
&\quad = \lim_{k\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta_{m_k},m_k}) \quad \exists \text{ subsequence converging to the } \liminf, \\
&\quad = \lim_{\ell\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta_{m_{k_\ell}},m_{k_\ell}}) \quad \exists \text{ subsequence } \theta_{m_{k_\ell}} \to \bar{\theta} \in \mathcal{K} \text{ by compactness,} \\
&\quad = \liminf_{\ell\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta_{m_{k_\ell}},m_{k_\ell}}) \\
&\quad \geq \liminf_{\ell\to\infty} [\mathfrak{W}_p(\hat{\mu}_n, \mu_{\theta_{m_{k_\ell}}}) - \mathbb{E}\mathfrak{W}_p(\mu_{\theta_{m_{k_\ell}}}, \hat{\mu}_{\theta_{m_{k_\ell}},m_{k_\ell}})] \quad \text{by the triangle inequality,} \\
&\quad \geq \liminf_{\ell\to\infty} \mathfrak{W}_p(\hat{\mu}_n, \mu_{\theta_{m_{k_\ell}}}) - \limsup_{\ell\to\infty} \mathbb{E}\mathfrak{W}_p(\mu_{\theta_{m_{k_\ell}}}, \hat{\mu}_{\theta_{m_{k_\ell}},m_{k_\ell}}) \\
&\quad \geq \mathfrak{W}_p(\hat{\mu}_n, \mu_{\bar{\theta}}) \quad \text{by l.s.c., Assumptions 2.2 and 2.6,} \\
&\quad \geq \inf_{\theta\in\mathcal{K}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta).
\end{aligned}$$

Fix $\mathcal{O} \subset \mathcal{H}$ open. By definition of the infimum, there exists a sequence $\theta_m \in \mathcal{O}$ such that $\mathfrak{W}_p(\hat{\mu}_n, \mu_{\theta_m}) \to \inf_{\theta\in\mathcal{O}} \mathfrak{W}_p(\hat{\mu}_n, \mu_\theta)$. Now, $\inf_{\theta\in\mathcal{O}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \leq \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta_m,m})$. Hence,

$$\limsup_{m\to\infty} \inf_{\theta\in\mathcal{O}} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \leq \limsup_{m\to\infty} \mathbb{E}\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta_m,m})$$

$$\leq \limsup_{m\to\infty} [\mathfrak{W}_p(\hat\mu_n, \mu_{\theta_m}) + \mathbb{E}\mathfrak{W}_p(\mu_{\theta_m}, \hat\mu_{\theta_m,m})] \quad \text{by the triangle inequality,}$$
$$= \limsup_{m\to\infty} \mathfrak{W}_p(\hat\mu_n, \mu_{\theta_m}) \quad \text{by Assumption 2.6,}$$
$$= \inf_{\theta\in\mathcal{O}} \mathfrak{W}_p(\mu_\star, \mu_\theta) \quad \text{by definition of } \theta_m.$$

Theorem 7.29b) of Rockafellar and Wets (2009) implies that $\limsup_{m\to\infty}(\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m})) \leq \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat\mu_n, \mu_\theta) = \varepsilon_n$. Hence, for all $\alpha > 0$, there exists m_α, such that for $m \geq m_\alpha$, $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) \leq \varepsilon_n + \alpha$. Let $\alpha \in (0, \varepsilon/2)$. The set $\{\theta \in \mathcal{H} : \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) \leq \varepsilon_n + \varepsilon/2\}$ is non-empty for $m \geq m_\alpha$, by definition of the infimum. Let θ belong to this set. Then, by the triangle inequality, $\mathfrak{W}_p(\hat\mu_n, \mu_\theta) \leq \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) + \mathbb{E}\mathfrak{W}_p(\mu_\theta, \hat\mu_{\theta,m})$. By Assumption 2.6, there exists an $\hat m$ such that for $m \geq \hat m$, $\mathbb{E}\mathfrak{W}_p(\mu_\theta, \hat\mu_{\theta,m}) \leq \varepsilon/2$. So, if $m \geq \max\{m_\alpha, \hat m\}$, we have that $\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_n + \varepsilon$. This means that $\{\theta \in \mathcal{H} : \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) \leq \varepsilon_n + \varepsilon/2\} \subset B_n(\varepsilon)$. As a consequence, for $m \geq \max\{m_\alpha, \hat m\}$, $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) = \inf_{\theta\in B_n(\varepsilon)} \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m})$.

By Theorem 7.31a) of Rockafellar and Wets (2009), we know that $\inf_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) \to \inf_{\theta\in\mathcal{H}} \mathfrak{W}_p(\hat\mu_n, \mu_\theta)$ as $m \to \infty$. Also, for $m \geq \max\{m_\alpha, \hat m\}$ and by the same reasoning as for the map $\theta \mapsto \mathfrak{W}_p(\hat\mu_n, \mu_\theta)$, the sets $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat\mu_n, \hat\mu_{\theta,m})$ are non-empty. By Theorem 7.31b) of Rockafellar and Wets (2009), the result follows. □

Theorem 2.11 (Measurability of the MEWE)**.** *Suppose that $\mathcal{H}$ is a σ-compact Borel measurable subset of $\mathbb{R}^{d_\theta}$. Under Assumption 2.5, for any $n \geq 1$ and $m \geq 1$ and $\varepsilon > 0$, there exists a Borel measurable function $\hat\theta_{n,m} : \Omega \to \mathcal{H}$ that satisfies*

$$\hat\theta_{n,m}(\omega) \in \begin{cases} \operatorname{argmin}_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m}), & \textit{if this set is non-empty,} \\ \varepsilon\text{-}\operatorname{argmin}_{\theta\in\mathcal{H}} \mathbb{E}\mathfrak{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m}), & \textit{otherwise.} \end{cases}$$

Proof. The proof is identical to that of Theorem 2.3, applying Lemma 2.3 instead of 2.1. □

2.4 WABC posterior

2.4.1 Behavior as $\varepsilon \to 0$ for fixed observations

The following result establishes some conditions under which a non-negative measure of discrepancy between data sets $\mathfrak{D}$ yields an ABC posterior that converges to the true posterior as $\varepsilon \to 0$, while the observations are kept fixed.

Proposition 2.3. *Assume that the posterior distribution is well-defined, and suppose that $\mu_\theta^{(n)}$ has a continuous density $f_\theta^{(n)}$, satisfying*

$$\sup_{y_{1:n}\in\mathcal{Y}^n, \theta\in\mathcal{H}} f_\theta^{(n)}(y_{1:n}) < \infty.$$

Suppose also that $\mathfrak{D}$ is continuous in the sense that, for any $y_{1:n}$, $\mathfrak{D}(y_{1:n}, z_{1:n}) \to \mathfrak{D}(y_{1:n}, x_{1:n})$ whenever $z_{1:n} \to x_{1:n}$ component-wise in the ground metric ρ. Suppose that either

1. *$f_\theta^{(n)}$ is n-exchangeable, such that $f_\theta^{(n)}(y_{1:n}) = f_\theta^{(n)}(y_{\sigma(1:n)})$ for any $\sigma \in \mathcal{S}_n$, and $\mathfrak{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{\sigma(1:n)}$ for some $\sigma \in \mathcal{S}_n$, or*
2. *$\mathfrak{D}(y_{1:n}, z_{1:n}) = 0$ if and only if $z_{1:n} = y_{1:n}$.*

Then, keeping $y_{1:n}$ fixed, the ABC posterior converges strongly to the posterior as $\varepsilon \to 0$.

Proof of Proposition 2.3. We follow a similar approach to that in Proposition 1 of Rubio and Johansen (2013). Fix $y_{1:n}$ and let $q^{\varepsilon}(\theta)$ denote the normalized quasi-likelihood induced by the ABC procedure, i.e.

$$q^{\varepsilon}(\theta)=\frac{\int_{\mathcal{Y}^n}\mathbb{1}\left(\mathfrak{D}(y_{1:n},z_{1:n})\leq\varepsilon\right)f_{\theta}^{(n)}(z_{1:n})dz_{1:n}}{\int_{\mathcal{Y}^n}\mathbb{1}\left(\mathfrak{D}(y_{1:n},z'_{1:n})\leq\varepsilon\right)dz'_{1:n}}=\int_{\mathcal{Y}^n}K^{\varepsilon}(y_{1:n},z_{1:n})f_{\theta}^{(n)}(z_{1:n})dz_{1:n},$$

where $K^{\varepsilon}(y_{1:n},z_{1:n})$ denotes the density of the uniform distribution on $\{z_{1:n}:\mathfrak{D}(y_{1:n},z_{1:n})\leq\varepsilon\}$, evaluated at some $z_{1:n}$. Note that the sets $A^{\varepsilon}=\{z_{1:n}:\mathfrak{D}(y_{1:n},z_{1:n})\leq\varepsilon\}$ are compact, due to the continuity of $\mathfrak{D}$. Now,

$$\begin{aligned}\left|q^{\varepsilon}(\theta)-f_{\theta}^{(n)}(y_{1:n})\right| &\leq \int_{\mathcal{Y}^n}K^{\varepsilon}(y_{1:n},z_{1:n})\left|f_{\theta}^{(n)}(z_{1:n})-f_{\theta}^{(n)}(y_{1:n})\right|dz_{1:n}\\ &\leq \sup_{z_{1:n}\in A^{\varepsilon},\theta\in\mathcal{H}}\left|f_{\theta}^{(n)}(z_{1:n})-f_{\theta}^{(n)}(y_{1:n})\right|\\ &= \left|f_{\theta}^{(n)}(z_{1:n}^{\varepsilon})-f_{\theta}^{(n)}(y_{1:n})\right|\end{aligned}$$

for some $z_{1:n}^{\varepsilon}\in A^{\varepsilon}$, where the second inequality holds since $\int_{\mathcal{Y}^n}K^{\varepsilon}(y_{1:n},z_{1:n})dz_{1:n}=1$, and the last equality holds by compactness of A^{ε} and continuity of $f_{\theta}^{(n)}$. Since $z_{1:n}^{\varepsilon}\in A^{\varepsilon}$, $\lim_{\varepsilon\to 0}z_{1:n}^{\varepsilon}\in\cap_{\varepsilon\in\mathbb{Q}^+}A^{\varepsilon}$. Under condition 1, $\cap_{\varepsilon\in\mathbb{Q}^+}A^{\varepsilon}=\{y_{\sigma(1:n)}:\sigma\in\mathcal{S}_n\}$, by continuity of $\mathfrak{D}$. Similarly, under condition 2, $\cap_{\varepsilon\in\mathbb{Q}^+}A^{\varepsilon}=\{y_{1:n}\}$. In both cases, taking the limit $\varepsilon\to 0$ yields $|q^{\varepsilon}(\theta)-f_{\theta}^{(n)}(y_{1:n})|\to 0$, due to the continuity of $f_{\theta}^{(n)}$ (and n-exchangeability under condition 1).

Since $\sup_{y_{1:n}\in\mathcal{Y}^n,\theta\in\mathcal{H}}f_{\theta}^{(n)}(y_{1:n})<M$ for some $0<M<\infty$, then $\sup_{y_{1:n}\in\mathcal{Y}^n,\theta\in\mathcal{H}}q^{\varepsilon}(\theta)<M$ also, for any ε. By the bounded convergence theorem, $\int_{\mathcal{H}}\pi(d\theta)q^{\varepsilon}(\theta)\to\int_{\mathcal{H}}\pi(d\theta)f_{\theta}^{(n)}(y_{1:n})$ as $\varepsilon\to 0$. Hence

$$\lim_{\varepsilon\to 0}\pi^{\varepsilon}\left(d\theta|y_{1:n}\right)=\frac{\lim_{\varepsilon\to 0}\pi\left(d\theta\right)q^{\varepsilon}(\theta)}{\lim_{\varepsilon\to 0}\int_{\mathcal{H}}\pi\left(d\vartheta\right)q^{\varepsilon}(\vartheta)}=\frac{\pi\left(d\theta\right)f_{\theta}^{(n)}(y_{1:n})}{\int_{\mathcal{H}}\pi\left(d\vartheta\right)f_{\vartheta}^{(n)}(y_{1:n})},$$

which is the true posterior distribution. □

2.4.2 Concentration as n increases and ε decreases

Recall the assumptions made in the main text:

Assumption 2.8. *The data-generating process is such that $\mathfrak{W}_p(\hat{\mu}_n,\mu_\star)\to 0$, in $\mathbb{P}$-probability, as $n\to\infty$.*

Assumption 2.9. *For any $\varepsilon>0$, $\mu_{\theta}^{(n)}(\mathfrak{W}_p(\mu_\theta,\hat{\mu}_{\theta,n})>\varepsilon)\leq c(\theta)f_n(\varepsilon)$, where $f_n(\varepsilon)$ is a sequence of functions that are strictly decreasing in ε for fixed n and $f_n(\varepsilon)\to 0$ for fixed ε as $n\to\infty$. The function $c:\mathcal{H}\to\mathbb{R}^+$ is π-integrable, and satisfies $c(\theta)\leq c_0$ for some $c_0>0$, for all θ such that, for some $\delta_0>0$, $\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\delta_0+\epsilon_\star$.*

Assumption 2.10. *There exist $L>0$ and $c_\pi>0$ such that, for all ε small enough,*

$$\pi\left(\{\theta\in\mathcal{H}:\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon+\varepsilon_\star\}\right)\geq c_\pi\varepsilon^L.$$

Theorem 2.12. *Under Assumptions 2.8-2.10, consider a sequence $(\varepsilon_n)_{n\geq 0}$ such that, as $n\to\infty$, $\varepsilon_n\to 0$, $f_n(\varepsilon_n)\to 0$, and $\mathbb{P}(\mathfrak{W}_p(\hat{\mu}_n,\mu_\star)\leq\varepsilon_n)\to 1$. Then, the WABC posterior with threshold $\varepsilon_n+\varepsilon_\star$ satisfies, for* *some* $0 < C < \infty$ *and any* $0 < R < \infty$,

$$\pi^{\varepsilon_n+\varepsilon_\star}\left(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) > \varepsilon_\star + 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R)\} \,|y_{1:n}\right) \leq \frac{C}{R},$$

with $\mathbb{P}$*-probability going to* 1 *as* $n \to \infty$.

Proof of Theorem 2.12. We first look at the WABC posterior probability of the sets $\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta\}$. Note that, using Bayes' formula, for all $\varepsilon, \delta > 0$,

$$\pi^{\varepsilon+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta | y_{1:n}\right) = \frac{\mathbb{P}_\theta(\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta,\ \mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)}{\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)},$$

where $\mathbb{P}_\theta$ denotes the distribution of $\theta \sim \pi$ and of the synthetic data $z_{1:n} \sim \mu_\theta^{(n)}$, keeping the observed data $y_{1:n}$ and hence $\hat{\mu}_n$ fixed. We aim to upper bound this expression, and proceed by upper bounding the numerator and lower bounding the denominator.

By the triangle inequality,

$$\mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \mathfrak{W}_p(\mu_\star, \hat{\mu}_n) + \mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) + \mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta).$$

On the events $\{\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta,\ \mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star\}$, we have

$$\delta < \mathfrak{W}_p(\mu_\star, \mu_\theta) \leq \mathfrak{W}_p(\mu_\star, \hat{\mu}_n) + \mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta) + \varepsilon + \varepsilon_\star.$$

Let $A(n, \varepsilon) = \{y_{1:n} : \mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \leq \varepsilon/3\}$. Assuming $y_{1:n} \in A(n, \varepsilon)$ implies that

$$\delta < \mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta) + \frac{4\varepsilon}{3} + \varepsilon_\star.$$

Using this to bound the numerator, we get by a simple reparametrization that for any $\zeta > 0$,

$$\pi^{\varepsilon+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star, \mu_\theta) > 4\varepsilon/3 + \varepsilon_\star + \zeta | y_{1:n}\right) \leq \frac{\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta) > \zeta)}{\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)}.$$

The remainder of the proof follows from further bounding this fraction using the assumptions we made on the convergence rate of empirical measures in the Wasserstein distance. Focusing first on the numerator, for any $\zeta > 0$ we have by Assumption 2.9 that

$$\begin{aligned}\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_{\theta,n}, \mu_\theta) > \zeta) &= \int_{\mathcal{H}} \mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \zeta)\pi(d\theta) \\ &\leq \int_{\mathcal{H}} c(\theta) f_n(\zeta)\pi(d\theta) \leq c_1 f_n(\zeta),\end{aligned}$$

for some constant $c_1 < +\infty$. For the denominator,

$$\begin{aligned}&\mathbb{P}_\theta(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star) = \int_{\mathcal{H}} \mu_\theta^{(n)}(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)\pi(d\theta) \\ &\geq \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta) \leq \varepsilon/3+\varepsilon_\star} \mu_\theta^{(n)}(\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)\pi(d\theta) \quad \text{(by non-negativity of integrand)} \\ &\geq \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta) \leq \varepsilon/3+\varepsilon_\star} \mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\star, \mu_\theta) + \mathfrak{W}_p(\hat{\mu}_n, \mu_\star) + \mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) \leq \varepsilon + \varepsilon_\star)\pi(d\theta)\end{aligned}$$

(by the triangle inequality)

$$\geq \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon/3+\varepsilon_\star} \mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta,\hat{\mu}_{\theta,n}) \leq \varepsilon/3)\pi(d\theta)$$

(since $\mathfrak{W}_p(\mu_\star,\mu_\theta) \leq \varepsilon/3+\varepsilon_\star$ and $\mathfrak{W}_p(\hat{\mu}_n,\mu_\star) \leq \varepsilon/3$)

$$= \pi(\mathfrak{W}_p(\mu_\star,\mu_\theta) \leq \varepsilon/3+\varepsilon_\star) - \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon/3+\varepsilon_\star} \mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta,\hat{\mu}_{\theta,n}) > \varepsilon/3)\pi(d\theta)$$

$$\geq \pi(\mathfrak{W}_p(\mu_\star,\mu_\theta) \leq \varepsilon/3+\varepsilon_\star) - \int_{\mathfrak{W}_p(\mu_\star,\mu_\theta)\leq\varepsilon/3+\varepsilon_\star} c(\theta)f_n(\varepsilon/3)\pi(d\theta) \quad \text{(by Assumption 2.9).}$$

We now make more specific choices for ε and ζ, starting with assuming that $\varepsilon/3 \leq \delta_0$, such that $c(\theta) \leq c_0$ for some constant $c_0 > 0$ in the last integrand above, by Assumption 2.9. The last line above is then greater than or equal to $\pi(\mathfrak{W}_p(\mu_\star,\mu_\theta) \leq \varepsilon/3+\varepsilon_\star)\,(1 - c_0 f_n(\varepsilon/3))$. Replacing ε with ε_n such that $f_n(\varepsilon_n/3) \to 0$ implies that $c_0 f_n(\varepsilon_n/3) \leq 1/2$ for sufficiently large n. Hence,

$$\pi(\mathfrak{W}_p(\mu_\star,\mu_\theta) \leq \varepsilon_n/3+\varepsilon_\star)\,(1 - c_0 f_n(\varepsilon_n/3)) \geq \frac{1}{2}\pi(\mathfrak{W}_p(\mu_\star,\mu_\theta) \leq \varepsilon_n/3+\varepsilon_\star) \geq c_\pi \varepsilon_n^L,$$

for sufficiently large n, by Assumption 2.10. We can summarize the bounds derived above as follows,

$$\pi^{\varepsilon_n+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star,\mu_\theta) > 4\varepsilon_n/3+\varepsilon_\star+\zeta|y_{1:n}\right) \leq C f_n(\zeta)\varepsilon_n^{-L},$$

where $C = c_1/c_\pi$.

Set some $R > 0$ and note that for any $n \geq 1$, because the function f_n is strictly decreasing under Assumption 2.9, $f_n^{-1}(\varepsilon_n^L/R)$ is well-defined in the sense that f_n^{-1} is defined at ε_n^L/R. Choosing $\zeta_n = f_n^{-1}(\varepsilon_n^L/R)$ leads to

$$\pi^{\varepsilon_n+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star,\mu_\theta) > 4\varepsilon_n/3+\varepsilon_\star+f_n^{-1}(\varepsilon_n^L/R)|y_{1:n}\right) \leq \frac{C}{R}.$$

Since we assumed that $\mathbb{P}\left(\{\omega : y_{1:n}(\omega) \in A(n,\varepsilon_n)\right) \to 1$ as $n \to \infty$, the statement above holds with probability going to one. □

Assumption 2.11. *There exist $K > 0$, $\alpha > 0$ and an open neighborhood $U \subset \mathcal{H}$ of $\theta_\star$, such that, for all $\theta \in U$,*

$$\rho_{\mathcal{H}}(\theta,\theta_\star) \leq K(\mathfrak{W}_p(\mu_\theta,\mu_\star) - \varepsilon_\star)^\alpha.$$

Corollary 2.1. *Under Assumptions 2.4 and 2.8-2.11, consider a sequence $(\varepsilon_n)_{n\geq 0}$ such that, as $n \to \infty$, $\varepsilon_n \to 0$, $f_n(\varepsilon_n) \to 0$, $f_n^{-1}(\varepsilon_n^L) \to 0$ and $\mathbb{P}(\mathfrak{W}_p(\hat{\mu}_n,\mu_\star) \leq \varepsilon_n) \to 1$. Then the WABC posterior with threshold $\varepsilon_n+\varepsilon_\star$ satisfies, for some $0 < C < \infty$ and any $0 < R < \infty$,*

$$\pi^{\varepsilon_n+\varepsilon_\star}\left(\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta,\theta_\star) > K(4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R))^\alpha\}\,|y_{1:n}\right) \leq \frac{C}{R},$$

with $\mathbb{P}$-probability going to 1.

Proof of Corollary 2.1. Let $\delta > 0$ be such that $\{\theta \in \mathcal{H} : \rho_{\mathcal{H}}(\theta,\theta_\star) \leq \delta\} \subset U$, where U is the set in Assumption 2.11. By Assumption 2.4, there exists a $\delta' > 0$ such that $\rho_{\mathcal{H}}(\theta,\theta_\star) > \delta$ implies $\mathfrak{W}_p(\mu_\theta,\mu_\star) - \varepsilon_\star > \delta'$. Let n be large enough such that $4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R) < \delta'$, which implies $\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star,\mu_\theta) - \varepsilon_\star \leq 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R)\} \subset U$.

From Theorem 2.12, we know that

$$\pi^{\varepsilon_n+\varepsilon_\star}\left(\mathfrak{W}_p(\mu_\star,\mu_\theta)-\varepsilon_\star \le 4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R)|y_{1:n}\right) \ge 1-\frac{C}{R},$$

with probability going to one. Applying the inequality in Assumption 2.11 gives

$$\pi^{\varepsilon_n+\varepsilon_\star}\left(\rho_{\mathcal{H}}(\theta,\theta_\star) \le K(4\varepsilon_n/3 + f_n^{-1}(\varepsilon_n^L/R))^\alpha|y_{1:n}\right) \ge 1-\frac{C}{R},$$

with probability going to one. □

2.5 Explicit rates of WABC concentration

In this section, we elaborate on the details of how we derive explicit rates of concentration of the WABC posterior for certain models.

2.5.1 Well-specified location model, i.i.d data

Suppose the distributions μ_θ on $\mathcal{Y}$ have densities such that $f_\theta(y) = f_0(y-\theta)$ for all $y \in \mathcal{Y}$ and $\theta \in \mathcal{H} = \mathcal{Y} = \mathbb{R}^{d_y}$. Let $\rho(x,y) = \rho_{\mathcal{H}}(x,y) = \|x-y\|$ be the Euclidean distance. Then, using the definition of the Wasserstein distance, the convexity of the function $x \mapsto \|x+\theta-\theta_\star\|^p$, and Jensen's inequality, one gets $\|\theta-\theta_\star\| \le \mathfrak{W}_p(\mu_\theta,\mu_{\theta_\star})$ for all $p \ge 1$. Therefore, with $K=1$ and $\alpha=1$, Assumption 2.11 holds. For location models, the rate of convergence of $\hat{\mu}_{\theta,n}$ to μ_θ is the same for all θ, so that we can indeed take $C(\theta)$ and $c(\theta)$ to be constants. The assumption $\mathcal{E}_{\beta,\gamma}(\mu_\theta) < \infty$ for some $\beta > p$ and $\gamma > 0$ is, for instance, satisfied in the Normal case. Note also that

$$\begin{aligned}\mathfrak{W}_p^p(\mu_\theta,\mu_{\theta_\star}) &= \inf_{\gamma\in\Gamma(\mu_\theta,\mu_{\theta_\star})}\int_{\mathcal{Y}\times\mathcal{Y}}\|x-y\|^p d\gamma(x,y)\\ &= \inf_{\gamma\in\Gamma(\mu_0,\mu_0)}\int_{\mathcal{Y}\times\mathcal{Y}}\|\theta+x-(\theta_\star+y)\|^p d\gamma(x,y) \le \|\theta-\theta_\star\|^p\end{aligned}$$

where the inequality is obtained by considering the maximal coupling between μ_0 and μ_0. Combined with the reverse inequality obtained via Jensen, we have $\mathfrak{W}_p(\mu_\theta,\mu_{\theta_\star}) = \|\theta-\theta_\star\|$ for all $p \ge 1$. This formula shows that Assumption 2.4 is satisfied.

To satisfy Assumption 2.10, we need to lower-bound the prior mass of balls $\{\theta \in \mathcal{H} : \|\theta-\theta_\star\| \le \varepsilon\}$. These have a volume $v\varepsilon^{d_y}$ for some $v > 0$ independent of ε, thus the assumption holds with $L = d_y$ for any prior with strictly positive density on $\mathcal{H} = \mathcal{Y} = \mathbb{R}^{d_y}$.

2.5.2 Misspecified location model, i.i.d. data

Consider a misspecified location model, in which θ denotes the mean of μ_θ and $\theta_\star$ the mean of $\mu_\star$. Assume that both $\mathcal{E}_{\beta,\gamma}(\mu_0) < \infty$ and $\mathcal{E}_{\beta,\gamma}(\mu_\star) < \infty$, and let $p=2$. Let $\varepsilon_n = c_\varepsilon(\log(n)/n)^{1/k}$ for some constant c_ε, where $k = 2p$ if $p > d_y/2$ and $k = d_y$ if $p < d_y/2$. In the case $p=2$, Bickel and Freedman (1981) showed that $\mathfrak{W}_2^2(\mu_\theta,\mu_\star) = \|\theta-\theta_\star\|^2 + \mathfrak{W}_2^2(\mu_0,\mu_\star^0) = \|\theta-\theta_\star\|^2 + \varepsilon_\star^2$, where $\mu_\star^0$ refers to $\mu_\star$ centered at zero. This formula shows that Assumption 2.4 holds. Now,

$$K(\mathfrak{W}_2(\mu_\theta,\mu_\star)-\varepsilon_\star))^\alpha = K(\sqrt{\|\theta-\theta_\star\|^2+\varepsilon_\star^2}-\varepsilon_\star)^\alpha$$

$$
\begin{aligned}
&\geq K\left\{\frac{\|\theta-\theta_\star\|^2}{2\varepsilon_\star}-\frac{\|\theta-\theta_\star\|^4}{8\varepsilon_\star^3}\right\}^\alpha \quad (\text{for } \|\theta-\theta_\star\|^2\leq 8\varepsilon_\star^2)\\
&\geq K\left\{\frac{\|\theta-\theta_\star\|^2}{2\varepsilon_\star}\left(1-\frac{\|\theta-\theta_\star\|^2}{4\varepsilon_\star^2}\right)\right\}^\alpha \quad (\text{for } \|\theta-\theta_\star\|^2\leq 4\varepsilon_\star^2)\\
&\geq K\left(\frac{\|\theta-\theta_\star\|^2}{4\varepsilon_\star}\right)^\alpha \quad (\text{for } \|\theta-\theta_\star\|^2\leq 2\varepsilon_\star^2),
\end{aligned}
$$

where the first inequality uses the fact that $\sqrt{a^2+b^2}-a\geq\frac{b^2}{2a}-\frac{b^4}{8a^3}$ whenever $0\leq b^2\leq 8a^2$. Hence, Assumption 2.11 is satisfied with $\alpha=1/2$ and $K=2\sqrt{\varepsilon_\star}$, leading to concentration rates that are the square root of those derived for the well-specified case. As illustrated in the main text, this is not simply due to the lack of sharpness in the above approximation.

2.5.3 Dependent data: AR(1)

As described in Fournier and Guillin (2015), deviation inequalities for the convergence of the empirical distribution in the Wasserstein sense have also been derived for certain classes of dependent data. Their inequalities are in the form of moment inequalities, which we convert to concentration inequalities via Markov's inequality. This does not yield sharp bounds, and is one reason for why the rates found for this example will be slower than those obtained for i.i.d. data.

A stationary stochastic process $Y=(Y_t)_{t\in\mathbb{Z}}$ with marginal distribution μ is ρ-mixing with mixing coefficients ρ_t for $t\geq 0$ if $\rho_t\to 0$ and for all $f,g\in L_2(\mu)$ and $i,j\geq 1$ we have

$$
\mathrm{Cov}(f(Y_i),g(Y_j))\leq\rho_{|i-j|}\sqrt{\mathrm{Var}(f(X_i))\mathrm{Var}(g(X_j))}.
$$

From the proof of Theorem 14 of Fournier and Guillin (2015) together with Markov's inequality, it can be seen that a ρ-mixing process with marginal distribution $\mu_\star$ and $s=\sum_{t=0}^\infty\rho_t<\infty$ satisfies

$$
\mathbb{P}(\mathfrak{W}_p(\mu_\star,\hat\mu_n)>\varepsilon)\leq c(s,p,q,d_y)M_q^{p/q}(\mu_\star)f_n(\varepsilon)=c(s,p,q,d_y)M_q^{p/q}(\mu)\frac{1}{\varepsilon^p v_n(p,q,d_y)},
$$

where $v_n=v_n(p,q,d_y)\to\infty$ at a rate depending of p,q and d_y, $c(s,p,q,d_y)$ is function depending on s,p,q and d_y, and $M_q(\mu)=\int_{\mathcal{Y}}\|x\|^q d\mu(x)$.

We now focus on the autoregressive model in Examples 4.1 and 4.3 of the main text. Recall that we defined the delay reconstructed time series $\tilde y_t=(y_t,y_{t-1})$ for $t\geq 2$, which can be written

$$
\tilde y_{t+1}=\begin{pmatrix}\phi & 0\\ 1 & 0\end{pmatrix}\tilde y_t+\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}\sigma\omega_t\\ \nu_t\end{pmatrix},
$$

where ν_t is an independent Gaussian noise process. Expressed in this way, the process $\tilde y_t$ satisfies the definition of an ARMA process given in Pham and Tran (1985). Their Theorem 3.1 can then be shown to hold, with the conclusion that $\tilde y_t$ satisfies the weaker notion of α-mixing, with coefficients $\alpha_t=O(|\phi|^t)$ (for a definition, see e.g. Bradley et al., 2005). However, Kolmogorov and Rozanov (1960) show in their Theorem 2 that stationary Gaussian processes that are α-mixing are also ρ-mixing, with coefficients satisfying $\rho_t\leq 2\pi\alpha_t$, implying that also $\rho_t=O(|\phi|^t)$.

In particular, note that $s\leq 1/(1-|\phi|)$ and $d_y=2$. Restricting $\mathcal{H}$ to be the compact set $\{\theta=(\phi,\sigma):\theta\in[-1+\delta_1,1-\delta_1]\times[\delta_2,M]\}$, for some small $\delta_1,\delta_2>0$ and some large $M<\infty$, implies that both $c(s,p,q,d_y)$

and $M_q^{p/q}(\mu_\theta)$ are bounded functions of θ. Letting $p = 2$ and $q > 4$, the concentration property can then be written, for all $\theta \in \mathcal{H}$,

$$\mathbb{P}(\mathfrak{W}_2(\mu_\theta, \hat{\mu}_{\theta,n}) > \varepsilon) \le c\frac{n^{-1/2} + n^{-(q-2)/2}}{\varepsilon^2} \le 2c\frac{n^{-1/2}}{\varepsilon^2},$$

for some constant $c > 0$. Letting $\varepsilon_n = c_\varepsilon n^{-1/(2L+4)}$ and doing the same calculation of $f_n^{-1}(\varepsilon_n^L/R)$ and $f_n(\varepsilon_n)$ as in the earlier examples, we can verify the conditions of Theorem 2.12.

We now look for a value of α such that Assumption 2.11 is satisfied. Let Σ be the covariance of (y_t, y_{t-1}) from Example 4.3 of the main text,

$$\Sigma = \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi \\ \phi & 1 \end{pmatrix},$$

and define $\Sigma_\star$ analogously for some $\theta_\star = (\phi_\star, \sigma_\star)$ in the interior of $\mathcal{H}$. Then

$$\mathfrak{W}_2^2(\mu_\theta, \mu_{\theta_\star}) = \mathrm{tr}(\Sigma + \Sigma_\star - 2(\Sigma^{1/2}\Sigma_\star\Sigma^{1/2})^{1/2}) = \left\|\Sigma^{1/2} - \Sigma_\star^{1/2}\right\|_{\mathrm{Frob}}^2,$$

where $\|\cdot\|_{\mathrm{Frob}}$ denotes the Frobenius matrix norm, the first equality follows from the closed form 2-Wasserstein distance between two multivariate Normal distributions, and the second follows because $\Sigma_\star\Sigma = \Sigma\Sigma_\star$. This formula shows that Assumption 2.4 holds. By Cholesky decomposition,

$$\begin{aligned}
\left\|\Sigma^{1/2} - \Sigma_\star^{1/2}\right\|_{\mathrm{Frob}}^2 &= \left\|\frac{\sigma}{\sqrt{1-\phi^2}}\begin{pmatrix} 1 & \phi \\ 0 & \sqrt{1-\phi^2} \end{pmatrix} - \frac{\sigma_\star}{\sqrt{1-\phi_\star^2}}\begin{pmatrix} 1 & \phi_\star \\ 0 & \sqrt{1-\phi_\star^2} \end{pmatrix}\right\|_{\mathrm{Frob}}^2 \\
&= \left(\frac{\sigma}{\sqrt{1-\phi^2}} - \frac{\sigma_\star}{\sqrt{1-\phi_\star^2}}\right)^2 + \left(\frac{\sigma\phi}{\sqrt{1-\phi^2}} - \frac{\sigma_\star\phi_\star}{\sqrt{1-\phi_\star^2}}\right)^2 + (\sigma - \sigma_\star)^2 \\
&\ge \left(\frac{\sigma\phi}{\sqrt{1-\phi^2}} - \frac{\sigma_\star\phi_\star}{\sqrt{1-\phi_\star^2}}\right)^2 + (\sigma - \sigma_\star)^2 \\
&= (\eta - \eta_\star)^2 + (\sigma - \sigma_\star)^2,
\end{aligned}$$

where $\eta = \sigma\phi/\sqrt{1-\phi^2}$. Note that the reparametrization $(\phi, \sigma) \mapsto (\eta, \sigma)$ is one-to-one. Choosing $\rho_{\mathcal{H}}(\theta, \theta_\star) = \sqrt{(\eta - \eta_\star)^2 + (\sigma - \sigma_\star)^2}$ yields $\rho_{\mathcal{H}}(\theta, \theta_\star) \le \mathfrak{W}_2(\mu_\theta, \mu_{\theta_\star})$. Therefore, Assumption 2.11 holds with $K = 1$ and $\alpha = 1$. Using Corollary 2.1, we obtain a concentration rate of $n^{-1/(2L+4)}$. $\mathcal{H}$ is compact also in the $\rho_{\mathcal{H}}$ distance, and with a uniform prior on (η, σ) it follows that $L = 2$. We can therefore only bound the concentration rate by $n^{-1/8}$.

2.6 Some results for checking assumptions

The following proposition gives three data-generating mechanisms for which $\mathfrak{W}_p(\hat{\mu}_n, \mu_\star) \to 0$ $\mathbb{P}$-almost surely, which is Assumption 2.1. The three conditions worked with here are mainly chosen for illustrative purposes, and are by no means exhaustive. We first definitions of properties used in the conditions.

Definition 2.5. The stochastic process $Y = (Y_t)_{t\in\mathbb{Z}}$ is stationary if for any $k \in \mathbb{N}$ and $\tau, t_1, \ldots t_k \in \mathbb{Z}$ we have that $(Y_{t_1}, \ldots, Y_{t_k}) \sim (Y_{t_1+\tau}, \ldots, Y_{t_k+\tau})$.

Definition 2.6. The map $T : \Omega \to \Omega$ is $\mathbb{P}$-measure preserving if $\mathbb{P}(T^{-1}(A)) = \mathbb{P}(A)$ for all $A \in \mathcal{F}$.

Definition 2.7. The map $T : \Omega \to \Omega$ is $\mathbb{P}$-ergodic if it is $\mathbb{P}$-measure preserving, and such that for all $A \in \mathcal{F}$

with $T^{-1}(A) = A$ we have that $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$. The stochastic process $Y = (Y_t)_{t\in\mathbb{Z}}$ is ergodic if it can be represented by $Y_t = Y_0 \circ T^t$ for some ergodic T and some random variable Y_0.

Definition 2.8. The stochastic process $Y = (Y_t)_{t\in\mathbb{Z}}$ is α-mixing with mixing coefficients

$$\alpha_t = \sup_{k\in\mathbb{Z}} \sup_{A\in\mathcal{F}_{-\infty}^k, B\in\mathcal{F}_{k+t}^\infty} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

if $\alpha_t \to 0$ as $t\to\infty$, where $\mathcal{F}_{-\infty}^k = \sigma(Y_i : i \le k)$ and $\mathcal{F}_k^\infty = \sigma(Y_i : i \ge k)$.

Proposition 2.4. *Suppose $Y = (Y_t)_{t\in\mathbb{Z}}$ is a stochastic process such that either*

1. *$Y \sim \mu_\star^\infty$, for some $\mu_\star \in \mathcal{P}_p(\mathcal{Y})$, i.e. the observations are i.i.d, or*
2. *$(Y_t)_{t\in\mathbb{Z}}$ is ergodic and stationary, represented by $Y_t = Y_0 \circ T^t$, where $Y_0 \sim \mu_\star \in \mathcal{P}_p(\mathcal{Y})$ and T is an ergodic, measure preserving map, or*
3. *$(Y_t)_{t\in\mathbb{Z}}$ is $\alpha-$mixing with mixing coefficients α_t such that $\sum_{t=1}^\infty \alpha_t^{1-1/2r} < \infty$, with $Y_t \sim \mu_t$ such that μ_t converges weakly to $\mu_\star$ in $\mathcal{P}_p(\mathcal{Y})$ and satisfies $\sup_t \mathbb{E}\|Y_t\|_{\mathcal{Y}}^q < \infty$ for some $1 \le \max(r,p) < q < 2r$ (where it is assumed $\rho(x,y) = \|x-y\|_{\mathcal{Y}}$ for simplicity).*

Then there exists a set $E \in \mathcal{F}$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, $\mathfrak{W}_p(\hat\mu_n(\omega), \mu_\star) \to 0$.

Proof. Under condition 1., Theorem 3 in Varadarajan (1958a) establishes that there exists a set E_1 with $\mathbb{P}(E_1) = 1$ such that for all $\omega \in E_1$, $\hat\mu_n(\omega)$ converges weakly to $\mu_\star$. By the strong law of large numbers, there exist a set E_2 with $\mathbb{P}(E_2) = 1$ and an $x_0 \in \mathcal{X}$ such that $\int_{\mathcal{X}} \rho(x,x_0)^p d\hat\mu_n(\omega)(x) \to \int_{\mathcal{X}} \rho(x,x_0)^p d\mu_\star(x)$ for all $\omega \in E_2$. Then, in light of Theorem 2.1, the claim holds on $E = E_1 \cap E_2$.

Consider condition 2. By Varadarajan (1958b), there exists a fixed countable set $C^\star$ of continuous and bounded functions on $\mathcal{Y}$, such that for any sequence of measures μ_n on $\mathcal{Y}$, μ_n converges weakly to μ if and only if $\int f d\mu_n \to \int f d\mu$ for all $f \in C^\star$. Fix $f \in C^\star$. We know that $f \circ Y_0$ is measurable and that $\mathbb{E}|f\circ Y_0| < \infty$ since f is bounded, so by Birkhoff's ergodic theorem there exists a set E_f such that $\mathbb{P}(E_f) = 1$ and

$$\int_{\mathcal{Y}} f d\hat\mu_n(\omega) = \frac{1}{n}\sum_{t=1}^n f(Y_t(\omega)) = \frac{1}{n}\sum_{t=1}^n f\circ Y_0 \circ T^t(\omega) \to \int_{\mathcal{Y}} f d\mu_\star,$$

for all $\omega \in E_f$. Moreover, since $\mu_\star \in \mathcal{P}_p(\mathcal{Y})$ we know $\int_{\mathcal{Y}} \rho(y,y_0)^p d\mu_\star(y) < \infty$ and that there exists a set E_0 with $\mathbb{P}(E_0) = 1$ such that

$$\int \rho(y,y_0)^p d\hat\mu_n(y)(\omega) \to \int \rho(y,y_0)^p d\mu_\star(y),$$

for all $\omega \in E_0$. Since $C^\star$ is countable we know that $\mathbb{P}(\cap_{f\in C^\star} E_f \cap E_0) = 1$. In other words, this means that $\mathfrak{W}_p(\hat\mu_n(\omega), \mu_\star) \to 0$ for all $\omega \in E = \cap_{f\in C^\star} E_f \cap E_0$.

Under condition 3., we first note that since $(Y_t)_{t\in\mathbb{Z}}$ is $\alpha-$mixing, then so is $(f\circ Y_t)_{t\in\mathbb{Z}}$ for any measurable f, with mixing coefficients bounded above by α_t since $\sigma(f(Y_i) : i \le k) \subset \sigma(Y_i : i \le k)$. Also, since μ_t converges weakly to $\mu_\star$ in $\mathcal{P}_p(\mathcal{Y})$ we have that for all $f \in C^\star$,

$$\frac{1}{n}\sum_{t=1}^n \int_{\mathcal{Y}} f d\mu_t \to \int_{\mathcal{Y}} f d\mu_\star,$$

and

$$\frac{1}{n}\sum_{t=1}^{n}\int_{\mathcal{Y}}\|y\|_{\mathcal{Y}}^{p}d\mu_t(y)\to\int_{\mathcal{Y}}\|y\|_{\mathcal{Y}}^{p}d\mu_\star(y).$$

By Hansen (1991) Corollary 4, we know that for all $f\in C^\star$ we have that the zero-mean, α-mixing sequence $f(Y_t)-\int_{\mathcal{Y}} f d\mu_t$ satisfies

$$\frac{1}{n}\sum_{t=1}^{n}\left\{f(Y_t)-\int_{\mathcal{Y}} f d\mu_t\right\}\to 0 \quad \mathbb{P}\text{-almost surely.}$$

Similarly,

$$\frac{1}{n}\sum_{t=1}^{n}\left\{\|Y_t\|_{\mathcal{Y}}^{p}-\int_{\mathcal{Y}}\|y\|_{\mathcal{Y}}^{p}d\mu_t(y)\right\}\to 0 \quad \mathbb{P}\text{-almost surely.}$$

Together this gives us that

$$\int_{\mathcal{Y}} f d\hat{\mu}_n=\frac{1}{n}\sum_{t=1}^{n}f(Y_t)\to\int_{\mathcal{Y}} f d\mu_\star \quad \mathbb{P}\text{-almost surely.}$$

and

$$\int_{\mathcal{Y}}\|y\|_{\mathcal{Y}}^{p}d\hat{\mu}_n(y)=\frac{1}{n}\sum_{t=1}^{n}\|Y_t\|_{\mathcal{Y}}^{p}\to\int_{\mathcal{Y}}\|y\|_{\mathcal{Y}}^{p}d\mu_\star(y) \quad \mathbb{P}\text{-almost surely.}$$

Then, again by the countability of $C^\star$, we can conclude that $\mathfrak{W}_p(\hat{\mu}_n(\omega),\mu_\star)\to 0$ for all ω in a set E defined analogously to the one for the second set of conditions. □

Proposition 2.5. *Suppose either of the conditions of Lemma 2.1 holds. Suppose there exists a proper, connected and compact subset $\mathcal{S}\subset\mathcal{H}$ with positive Lebesgue measure such that $\inf_{\theta\in\mathcal{H}\setminus\mathcal{S}}\mathfrak{W}_p(\mu_\star,\mu_\theta)>\inf_{\theta\in\mathcal{H}}\mathfrak{W}_p(\mu_\star,\mu_\theta)$. Then there exists a $\theta_\star$ attaining the infimum of $\theta\mapsto\mathfrak{W}_p(\mu_\star,\mu_\theta)$. If $\theta_\star$ is unique, then it is well-separated.*

Proof. Since $\theta\mapsto\mathfrak{W}_p(\mu_\star,\mu_\theta)$ is continuous/lower semicontinuous, it attains a minimum $\theta_\star$ on $\mathcal{S}$. This is also the global minimum by the assumption on $\mathcal{S}$. If $\theta_\star$ is unique, it is well-separated in the sense of Assumption 2.4, for all $\varepsilon>0$, there exists $\delta>0$ such that

$$\inf_{\theta\in\mathcal{H}:\rho_{\mathcal{H}}(\theta,\theta_\star)\geq\varepsilon}\mathfrak{W}_p(\mu_\star,\mu_\theta)>\mathfrak{W}_p(\mu_\star,\mu_{\theta_\star})+\delta.$$

Indeed, let $\varepsilon>0$, and consider $\{\theta\in\mathcal{H}:\rho_{\mathcal{H}}(\theta,\theta_\star)\geq\varepsilon\}$. Either the set is contained in $\mathcal{H}\setminus\mathcal{S}$, and thus well-separation follows, or, $\{\theta\in\mathcal{H}:\rho_{\mathcal{H}}(\theta,\theta_\star)\geq\varepsilon\}\cap\mathcal{S}$ is not empty. Then we show that it is compact. Since $\mathcal{S}$ is compact, there exists $\bar{\varepsilon}\geq\varepsilon$ such that $\mathcal{S}\subset\{\theta\in\mathcal{H}:\rho_{\mathcal{H}}(\theta,\theta_\star)\leq\bar{\varepsilon}\}$. Therefore $\{\theta\in\mathcal{H}:\rho_{\mathcal{H}}(\theta,\theta_\star)\geq\varepsilon\}\cap\mathcal{S}=\{\theta\in\mathcal{H}:\bar{\varepsilon}\geq\rho_{\mathcal{H}}(\theta,\theta_\star)\geq\varepsilon\}\cap\mathcal{S}$. Now $\{\theta\in\mathcal{H}:\bar{\varepsilon}\geq\rho_{\mathcal{H}}(\theta,\theta_\star)\geq\varepsilon\}$ is compact. An intersection of compact sets is compact. Therefore, $\theta\mapsto\mathfrak{W}_p(\mu_\star,\mu_\theta)$ being continuous/lower semicontinuous, an infimum is attained on $\{\theta\in\mathcal{H}:\rho_{\mathcal{H}}(\theta,\theta_\star)\geq\varepsilon\}\cap\mathcal{S}$, and by uniqueness of $\theta_\star$, well-separation follows. □

3 ABC with other transport distances

In this section we establish a version of Theorem 5.4 of the main text that holds for the ABC posterior based on the Hilbert distance (Theorem 3.1) as well as a version of this result that holds for the ABC posterior based on the swapping distance (Corollary 3.1).

3.1 Notation and assumptions

Let $H : [0,1] \to [0,1]^{d_y}$ be the Hilbert space-filling curve such that $H(0) = 0 \in \mathbb{R}^{d_y}$. Although we refer to H as *the* Hilbert curve in what follows, there exist in fact several Hilbert curves. We recall that the mapping H is Hölder continuous: there exists a constant $C_{d_y} < \infty$ such that $\rho\big(H(x), H(y)\big) \leq C_{d_y} |x-y|^{1/d_y}$, $\forall x, y \in [0,1]$. In what follows, let $\tilde{\varepsilon} = C_{d_y}^{-1} \varepsilon^{d_y}$. The mapping H is also surjective, which implies that H admits a Borel measurable pseudo-inverse $h : [0,1]^{d_y} \to [0,1]$ such that $H(h(x)) = x$ for all $x \in [0,1]^{d_y}$ (Gerber et al., 2017, Proposition 2).

Below we denote by μ^h the image of probability measure $\mu \in \mathcal{P}((0,1)^{d_y})$ by h and by $\mathcal{P}_b(\mathcal{Y}) \subset \mathcal{P}(\mathcal{Y})$ the set of probability measures on $\mathcal{Y}$ that admit a continuous and bounded density w.r.t. λ_{d_y}, the Lebesgue measure on $\mathbb{R}^{d_y}$. Henceforth, we fix $p \geq 1$ and a metric ρ on $(0,1)^{d_y}$ and denote by $\mathfrak{W}_p^*$ the p-Wasserstein distance on $\mathcal{P}(\mathbb{R})$ obtained for $\rho(x,y) = |x - y|$.

To simplify the presentation, the results presented below assume that the model is well-specified and the following assumptions are considered.

Assumption 3.1. *The model $\mathcal{M} = \{\mu_\theta : \theta \in \mathcal{H}\}$ is well-specified; that is, there exists a $\theta_\star \in \mathcal{H}$ such that $\mu_{\theta_\star} = \mu_\star$. In addition, $\mathcal{M} \subset \mathcal{P}_p(\mathcal{Y}) \cap \mathcal{P}_b(\mathcal{Y})$.*

Assumption 3.2. *For any $\varepsilon > 0$, $\mu_\theta^{(n)}(\mathfrak{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) > \varepsilon) \leq c(\theta) f_n(\varepsilon)$, where $(f_n)_{n\geq 1}$ is a sequence of strictly decreasing functions and such that $f_n(\varepsilon) \to 0$ as $n \to \infty$, and the function $c : \mathcal{H} \to \mathbb{R}^+$ is π-integrable.*

Assumption 3.3. *For any $\varepsilon > 0$, $\mu_\theta^{(n)}(\mathfrak{W}_1^*(\mu_\theta^h, \hat{\mu}_{\theta,n}^h) > \varepsilon) \leq \tilde{c}(\theta) \tilde{f}_n(\varepsilon)$, where $(\tilde{f}_n)_{n\geq 1}$ is a sequence of strictly decreasing functions such that $\tilde{f}_n(\varepsilon_n) \to 0$ for any sequence $(\varepsilon_n)_{n\geq 1}$ that converges to 0 sufficiently slowly as $n \to \infty$. The function $\tilde{c} : \mathcal{H} \to \mathbb{R}^+$ satisfies $\tilde{c}(\theta) \leq \tilde{c}_0$ for some $\tilde{c}_0 > 0$, for all θ such that, for some $\tilde{\delta}_0 > 0$, $\mathfrak{W}_1^*(\mu_\star^h, \mu_\theta^h) \leq \tilde{\delta}_0$.*

Note that Assumption 3.3 tends to hold under weaker conditions than the analogous Assumption 3.2, since μ_θ^h and $\mu_\star^h$ are distributions on the bounded interval $[0,1]$. The moment conditions in Fournier and Guillin (2015) are therefore satisfied. For instance, in the case of i.i.d. data, Assumption 3.3 is automatically satisfied by their Theorem 1 and Markov's inequality. However, it may be that other functions $\tilde{f}_n$ and $\tilde{c}$ exist to give sharper bounds.

Assumption 3.4. *The function $r(\varepsilon) := \pi\left(\{\theta \in \mathcal{H} : \mathfrak{W}_1^*(\mu_\star^h, \mu_\theta^h) \leq \varepsilon\}\right)$ satisfies $r(\varepsilon) > 0$ for any $\varepsilon > 0$.*

In Section 3.4, we illustrate that this assumption holds under relatively weak conditions.

3.2 Hilbert distance ABC

We denote below by $\pi_{\mathfrak{H}_p}^\varepsilon$ the ABC posterior based on the Hilbert distance $\mathfrak{H}_p$; that is,

$$\pi_{\mathfrak{H}_p}^\varepsilon (d\theta | y_{1:n}) = \frac{\pi(d\theta) \int_{\mathcal{Y}^n} \mu_\theta^{(n)}(dz_{1:n}) \mathbb{1}\left(\mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) < \varepsilon\right)}{\int_{\mathcal{H}} \pi(d\vartheta) \int_{\mathcal{Y}^n} \mu_\vartheta^{(n)}(dz'_{1:n}) \mathbb{1}\left(\mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}'_{\vartheta,n})) < \varepsilon\right)}.$$

Before stating the consistency result (Theorem 3.1) we first show that $\mathfrak{H}_p$ defines a distance between empirical distributions of the same size.

Proposition 3.1. *For any integer $n \geq 1$ and real number $p \geq 1$, $\mathfrak{H}_p$ defines a distance on the space of empirical distributions of size n.*

Proof of Proposition 3.1. Let $x_{1:n}$, $y_{1:n}$ and $z_{1:n}$ be three vectors in $\mathcal{Y}^n$ and denote by $\hat{\mu}_n^x$, $\hat{\mu}_n^y$ and $\hat{\mu}_n^z$ the corresponding empirical distributions of size n. Since ρ is a metric on $\mathcal{Y}$,

$$\mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) \geq 0, \quad \mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) = \mathfrak{H}_p(\hat{\mu}_n^z, \hat{\mu}_n^x)$$

and $\mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) = 0$ if and only if $\hat{\mu}_n^x = \hat{\mu}_n^z$. To conclude the proof it therefore remains to show that

$$\mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) \leq \mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^y) + \mathfrak{H}_p(\hat{\mu}_n^y, \hat{\mu}_n^z).$$

To this end, we define

$$\begin{aligned}
&\rho_{xy} = \big(\rho\big(x_{\sigma_x(1)}, y_{\sigma_y(1)}\big), \dots, \rho\big(x_{\sigma_x(n)}, y_{\sigma_y(n)}\big)\big), \quad \rho_{xz} = \big(\rho\big(x_{\sigma_x(1)}, z_{\sigma_z(1)}\big), \dots, \rho\big(x_{\sigma_x(n)}, z_{\sigma_z(n)}\big)\big) \\
&\rho_{yz} = \big(\rho\big(y_{\sigma_y(1)}, z_{\sigma_z(1)}\big), \dots, \rho\big(y_{\sigma_y(n)}, z_{\sigma_z(n)}\big)\big)
\end{aligned}$$

and denote by $\|\cdot\|_p$ the L_p-norm on $\mathbb{R}^n$. Then,

$$\begin{aligned}
\mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^z) &= n^{-1/p}\|\rho_{xz}\|_p \\
&\leq n^{-1/p}\|\rho_{xy}\|_p + n^{-1/p}\|\rho_{xz} - \rho_{xy}\|_p \\
&= \mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^y) + n^{-1/p}\left(\sum_{i=1}^n \left|\rho\big(x_{\sigma_x(i)}, z_{\sigma_z(i)}\big) - \rho\big(x_{\sigma_x(i)}, y_{\sigma_y(i)}\big)\right|^p\right)^{1/p} \\
&\leq \mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^y) + n^{-1/p}\left(\sum_{i=1}^n \rho\big(y_{\sigma_y(i)}, z_{\sigma_z(i)}\big)^p\right)^{1/p} \\
&= \mathfrak{H}_p(\hat{\mu}_n^x, \hat{\mu}_n^y) + \mathfrak{H}_p(\hat{\mu}_n^y, \hat{\mu}_n^z),
\end{aligned}$$

where the first inequality uses the triangle inequality, and the last uses the reverse triangle inequality. ☐

Theorem 3.1. *Let $\mathcal{Y} = (0,1)^{d_y}$, $h_{\mathcal{Y}} = h$, and $1 \leq p \leq d_y$. Under Assumptions 3.1-3.4, consider a sequence $(\varepsilon_n)_{n\geq 0}$ such that, as $n \to \infty$, $\varepsilon_n \to 0$, $\tilde{f}_n(\tilde{\varepsilon}_n) \to 0$, and $f_n(\varepsilon_n) \to 0$. Then, the ABC posterior based on the Hilbert distance with threshold ε_n satisfies, for some $0 < C < \infty$ and any $0 < R < \infty$,*

$$\pi_{\mathfrak{H}_p}^{\varepsilon_n}\left(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) > 4\varepsilon_n/3 + f_n^{-1}(r(\tilde{\varepsilon}_n/3)/R)\} \,|y_{1:n}\right) \leq \frac{C}{R},$$

with $\mathbb{P}$-probability going to 1 as $n \to \infty$.

Proof. We first look at the ABC posterior probability of the set $\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta\}$. Note that, using Bayes' formula, for all $\varepsilon, \delta > 0$,

$$\pi_{\mathfrak{H}_p}^{\varepsilon}\left(\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta | y_{1:n}\right) = \frac{\mathbb{P}_\theta(\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta,\ \mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon)}{\mathbb{P}_\theta(\mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon)},$$

where $\mathbb{P}_\theta$ denotes the distribution of $\theta \sim \pi$ and of the synthetic data $z_{1:n} \sim \mu_\theta^{(n)}$, keeping the observed data $y_{1:n}$ and hence $\hat{\mu}_n$ fixed.

We first study the numerator. Using the fact that for any $\theta \in \mathcal{H}$,

$$\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}), \quad \mu_\theta^{(n)}\text{-almost surely},$$

it follows that

$$\mathbb{P}_\theta(\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta,\ \mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon) \leq \mathbb{P}_\theta(\mathfrak{W}_p(\mu_\star, \mu_\theta) > \delta,\ \mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon),$$

where, under Assumption 3.2, the right-hand side is bounded as in the proof of Theorem 2.12. Hence, for any $\zeta > 0$ and

$$y_{1:n} \in A(n, \varepsilon) := \Big\{ y'_{1:n} : \mathfrak{W}_p\Big(n^{-1} \sum_{i=1}^n \delta_{y'_i}, \mu_\star\Big) \leq \varepsilon/3 \Big\},$$

we have

$$\pi^\varepsilon_{\mathfrak{H}_p}\Big(\mathfrak{W}_p(\mu_\star, \mu_\theta) > 4\varepsilon/3 + \zeta | y_{1:n}\Big) \leq \frac{c_1 f_n(\zeta)}{\mathbb{P}_\theta(\mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon)} \tag{1}$$

for a constant $c_1 < \infty$ and with the sequence of functions $(f_n)_{n\geq 1}$ as in Assumption 3.2.

We now lower bound the denominator in (1). Let $y_i^h = h(y_i)$ and $z_i^h = h(z_i)$, for $i \in 1:n$. Then, for any $\theta \in \mathcal{H}$ and $\mu_\theta^{(n)}$-almost surely,

$$\begin{aligned}
\mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) &\leq \mathfrak{H}_{d_y}(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \qquad (\text{since } \|x\|_p \leq \|x\|_{d_y}, \text{ for any } x \in \mathbb{R}^n \text{ and } 1 \leq p \leq d_y) \\
&= \Big(\frac{1}{n}\sum_{i=1}^n \rho\big(y_{\sigma_y(i)}, z_{\sigma_z(i)}\big)^{d_y}\Big)^{1/d_y} \qquad (\text{by definition of } \mathfrak{H}_{d_y}) \\
&= \Big(\frac{1}{n}\sum_{i=1}^n \rho\Big(H(y^h_{(i)}), H(z^h_{(i)})\Big)^{d_y}\Big)^{1/d_y} \qquad (\text{since } H(h(x)) = x, \forall\, x \in (0,1)^{d_y}) \\
&\leq \Big(\frac{C_{d_y}}{n}\sum_{i=1}^n |y^h_{(i)} - z^h_{(i)}|\Big)^{1/d_y} \qquad (\text{by the Hölder property of } H) \\
&= \Big(C_{d_y}\, \mathfrak{W}_1^*(\hat{\mu}^h_n, \hat{\mu}^h_{\theta,n})\Big)^{1/d_y} \qquad (\text{by definition of } \mathfrak{W}_1^*),
\end{aligned}$$

so that

$$\mathbb{P}_\theta\big(\mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \varepsilon\big) \geq \mathbb{P}_\theta\big(C_{d_y}^{1/d_y} \mathfrak{W}_1^*(\hat{\mu}^h_n, \hat{\mu}^h_{\theta,n})^{1/d_y} \leq \varepsilon\big) = \mathbb{P}_\theta\big(\mathfrak{W}_1^*(\hat{\mu}^h_n, \hat{\mu}^h_{\theta,n}) \leq C_{d_y}^{-1}\,\varepsilon^{d_y}\big).$$

To bound the right-hand side, let $\tilde{\varepsilon} = C_{d_y}^{-1}\varepsilon^{d_y}$ and assume henceforth that

$$y_{1:n} \in A^h(n, \varepsilon) := \Big\{ y'_{1:n} : \mathfrak{W}_1^*\big(n^{-1}\sum_{i=1}^n \delta_{h(y'_i)}, \mu^h_\star\big) \leq C_{d_y}^{-1}\,\varepsilon^{d_y}/3 \Big\}. \tag{2}$$

Then,

$$\mathbb{P}_\theta\big(\mathfrak{W}_1^*(\hat{\mu}^h_n, \hat{\mu}^h_{\theta,n}) \leq \tilde{\varepsilon}\big) = \int_{\mathcal{H}} \mu_\theta^{(n)}(\mathfrak{W}_1^*(\hat{\mu}^h_n, \hat{\mu}^h_{\theta,n}) \leq \tilde{\varepsilon})\pi(d\theta)$$

$$
\begin{aligned}
&\geq \int_{\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3} \mu_\theta^{(n)}(\mathfrak{W}_1^*(\hat{\mu}_n^h,\hat{\mu}_{\theta,n}^h)\leq\tilde{\varepsilon})\pi(d\theta) \quad \text{(by non-negativity of integrand)}\\
&\geq \int_{\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3} \mu_\theta^{(n)}(\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)+\mathfrak{W}_1^*(\hat{\mu}_n^h,\mu_\star^h)+\mathfrak{W}_1^*(\mu_\theta^h,\hat{\mu}_{\theta,n}^h)\leq\tilde{\varepsilon})\pi(d\theta)\\
&\hspace{20em}\text{(by the triangle inequality)}\\
&\geq \int_{\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3} \mu_\theta^{(n)}(\mathfrak{W}_1^*(\mu_\theta^h,\hat{\mu}_{\theta,n}^h)\leq\tilde{\varepsilon}/3)\pi(d\theta)\\
&\hspace{10em}\text{(since } \mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3 \text{ and } \mathfrak{W}_1^*(\hat{\mu}_n,\mu_\star)\leq\tilde{\varepsilon}/3)\\
&= \pi(\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3) - \int_{\mathfrak{W}^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3} \mu_\theta^{(n)}(\mathfrak{W}_1^*(\mu_\theta^h,\hat{\mu}_{\theta,n}^h)>\tilde{\varepsilon}/3)\pi(d\theta)\\
&\geq \pi(\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3) - \int_{\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3} \tilde{c}(\theta)\tilde{f}_n(\tilde{\varepsilon}/3)\pi(d\theta) \quad \text{(by Assumption 3.3)},
\end{aligned}
$$

where the third inequality uses (2). We now make specific choices for ε (and thus $\tilde{\varepsilon}$) and ζ, starting by assuming that $\tilde{\varepsilon}/3 \leq \tilde{\delta}_0$, such that $\tilde{c}(\theta) \leq \tilde{c}_0$ for some constant $\tilde{c}_0 > 0$ in the last integrand above, by Assumption 3.3. The last line above is then greater than or equal to $\pi(\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}/3)(1-\tilde{c}_0\tilde{f}_n(\tilde{\varepsilon}/3))$. Replacing $\tilde{\varepsilon}$ with $\tilde{\varepsilon}_n$ such that $\tilde{f}_n(\tilde{\varepsilon}_n/3)\to 0$ implies that $\tilde{c}_0\tilde{f}_n(\tilde{\varepsilon}_n/3)\leq 1/2$ for sufficiently large n. Hence,

$$
\pi(\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}_n/3)\left(1-\tilde{c}_0\tilde{f}_n(\tilde{\varepsilon}_n/3)\right) \geq \frac{1}{2}\pi(\mathfrak{W}_1^*(\mu_\star^h,\mu_\theta^h)\leq\tilde{\varepsilon}_n/3) = \frac{1}{2}r(\tilde{\varepsilon}_n/3)
$$

for sufficiently large n, by Assumption 3.4. Summarizing the bounds derived above, we get

$$
\pi_{\mathcal{H}_p}^{\varepsilon_n}(\mathfrak{W}_p(\mu_\star,\mu_\theta)>4\varepsilon_n/3+\zeta|y_{1:n}) \leq \frac{Cf_n(\zeta)}{r(\tilde{\varepsilon}_n/3)},
$$

where $C = 2c_1$. Set some $R > 0$ and assume that $f_n^{-1}(r(\tilde{\varepsilon}_n/3)/R)$ is well-defined, in the sense that f_n^{-1} is defined at $r(\tilde{\varepsilon}_n/3)/R$, which can be achieved by taking n large enough. Choosing $\zeta_n = f_n^{-1}(r(\tilde{\varepsilon}_n/3)/R)$ leads to

$$
\pi_{\mathfrak{H}_p}^{\varepsilon_n}\left(\mathfrak{W}_p(\mu_\star,\mu_\theta)>4\varepsilon_n/3+f_n^{-1}(r(\tilde{\varepsilon}_n/3)/R)|y_{1:n}\right) \leq \frac{C}{R}. \tag{3}
$$

Under well-specification, $\mathbb{P}\left(\{\omega : y_{1:n}(\omega)\in A(n,\varepsilon_n)\right)\to 1$ and $\mathbb{P}\left(\{\omega : y_{1:n}(\omega)\in A^h(n,\varepsilon_n)\}\right)\to 1$, as $n\to\infty$, is implied by $f_n(\varepsilon_n/3)\to 0$ and $\tilde{f}_n(\tilde{\varepsilon}_n/3)\to 0$ respectively. As a consequence, the statement in (3) holds with probability going to one.

□

3.3 Swapping distance ABC

Let $\mathfrak{S}_p(\hat{\mu}_n,\hat{\mu}_{\theta,n})$ be the swapping distance. Then, the result for the ABC posterior based on the this distance, defined by

$$
\pi_{\mathfrak{S}_p}^{\varepsilon}(d\theta|y_{1:n}) = \frac{\pi(d\theta)\int_{\mathcal{Y}^n}\mu_\theta^{(n)}(dz_{1:n})\mathbb{1}\left(\mathfrak{S}_p(\hat{\mu}_n,\hat{\mu}_{\theta,n})<\varepsilon\right)}{\int_{\mathcal{H}}\pi(d\vartheta)\int_{\mathcal{Y}^n}\mu_\vartheta^{(n)}(dz'_{1:n})\mathbb{1}\left(\mathfrak{S}_p(\hat{\mu}_n,\hat{\mu}'_{\vartheta,n}))<\varepsilon\right)},
$$

is given in the next corollary.

Corollary 3.1. *Let $\mathcal{Y}=(0,1)^{d_y}$, $h_{\mathcal{Y}}=h$, and $1\leq p\leq d_y$. Under Assumptions 3.1-3.4, consider a sequence $(\varepsilon_n)_{n\geq 0}$ such that, as $n\to\infty$, $\varepsilon_n\to 0$, $\tilde{f}_n(\tilde{\varepsilon}_n)\to 0$ and $f_n(\varepsilon_n)\to 0$. Then, the ABC posterior based on the* *swapping distance with threshold ε_n satisfies, for some $0 < C < \infty$ and any $0 < R < \infty$,*

$$\pi^{\varepsilon_n}_{\mathfrak{H}_p}\left(\{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu_\star, \mu_\theta) > 4\varepsilon_n/3 + f_n^{-1}(r(\tilde{\varepsilon}_n/3)/R)\}\,|y_{1:n}\right) \leq \frac{C}{R},$$

with $\mathbb{P}$-probability going to 1 as $n \to \infty$.

Proof. The result follows from the proof of Theorem 3.1 and from the fact that, for all $n \geq 1$ and $\theta \in \mathcal{H}$,

$$\mathfrak{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \mathfrak{S}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}) \leq \mathfrak{H}_p(\hat{\mu}_n, \hat{\mu}_{\theta,n}), \quad \mathbb{P} - a.s.$$

□

3.4 Some results for checking Assumption 3.4

The main text and the earlier discussion in this document contain remarks and references to results that are useful for establishing Assumptions 3.1-3.3. In this section, we develop conditions under which Assumption 3.4 holds. Consider the following assumptions and technical lemmas.

Assumption 3.5. *The map $\theta \mapsto \mu_\theta$ is injective.*

The above assumption is made implicitly in the main text, but we state it explicitly here, for simplicity.

Assumption 3.6. *For any sequence $(\theta_n)_{n\geq 1}$ and θ in $\mathcal{H}$ such that $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$, we have that $\mathfrak{W}_p(\mu_{\theta_n}, \mu_\theta) \to 0$.*

Lemma 3.1. *Let $\mu \in \mathcal{P}_b((0,1)^{d_y})$ and $(\mu_n)_{n\geq 1}$ be a sequence of probability measures on $(0,1)^{d_y}$ such that $\mu_n \Rightarrow \mu$. Then, as $n \to \infty$, $\mathfrak{W}_p(\mu_n^h, \mu^h) \to 0$.*

Proof. By Theorem 2.1, $\mathfrak{W}_p(\mu_n^h, \mu^h) \to 0$ if and only if (i) $\mu_n^h \Rightarrow \mu^h$ and (ii) there exists a $y_0 \in (0,1)$ such that $\int_{(0,1)} \rho(y, y_0)^p \mu_n^h(dy) \to \int_{(0,1)} \rho(y, y_0)^p \mu^h(dy)$. By Gerber et al. (2017, Theorem 9), $\mu_n^h \Rightarrow \mu^h$ and thus the result follows from the fact that, for any fixed $y_0 \in (0,1)$, the mapping $y \mapsto \rho(y, y_0)^p$ is continuous and bounded on $(0,1)$. □

Lemma 3.2. *Let $\mathcal{Y} = (0,1)^{d_y}$. Under Assumption 3.5, the map $\theta \mapsto \mu_\theta^h$ is injective.*

Proof. Let θ and θ' be two distinct points in $\mathcal{H}$. Then, under Assumption 3.5, there exists a measurable set $A \subset \mathcal{Y}$ such that $\mu_\theta(A) \neq \mu_{\theta'}(A)$. To conclude the proof let $I_A = h(A)$ and note that, since the mapping $h : (0,1)^{d_y} \to [0,1]$ is one-to-one, we have $\mu_\theta^h(I_A) = \mu_\theta(A) \neq \mu_{\theta'}(A) = \mu_{\theta'}^h(I_A)$. □

Lemma 3.3. *Let $\mathcal{Y} = (0,1)^{d_y}$. Under Assumptions 3.1 and 3.6, for any sequence $(\theta_n)_{n\geq 1}$ and θ in $\mathcal{H}$ such that $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$, we have that $\mathfrak{W}_p(\mu_{\theta_n}^h, \mu_\theta^h) \to 0$.*

Proof. The result is a direct consequence of Lemma 3.1 and Assumption 3.6. □

Lemma 3.4. *Let $\mathcal{Y} = (0,1)^{d_y}$. Under Assumptions 3.1, 3.5 and 3.6, the mapping $\theta \mapsto \mathfrak{W}_p(\mu_\star^h, \mu_\theta^h)$ has a unique minimum on $\mathcal{H}$ at $\theta = \theta_\star$. In addition, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$\inf_{\{\theta \in \mathcal{H} :\, \rho(\theta, \theta_\star) \geq \varepsilon\}} \mathfrak{W}_p(\mu_\star^h, \mu_\theta^h) > \delta.$$

Proof. Under Assumptions 3.5 and 3.1, there exists a unique $\theta_\star \in \mathcal{H}$ such that $\mu_{\theta_\star} = \mu_\star$, and thus $\mu^h_{\theta_\star} = \mu^h_\star$. Then, the first part of the result follows from the fact that, under Assumption 3.5, the map $\theta \mapsto \mu^h_\theta$ is injective by Lemma 3.2. The second part of lemma follows from Proposition 2.5, using the fact that, under Assumptions 3.6 and 3.1, the map $\theta \mapsto \mathfrak{W}_p(\mu^h_\theta, \mu^h_{\theta_\star})$ is continuous by Lemmas 2.1 and 3.3. □

Lemma 3.5. *Let $\mathcal{Y} = (0,1)^{d_y}$. Under Assumptions 3.1, 3.5 and 3.6, for any $\varepsilon > 0$ the set $\tilde{B}_{\theta_\star}(\varepsilon) = \{\theta \in \mathcal{H} : \mathfrak{W}_p(\mu^h_{\theta_\star}, \mu^h_\theta) \leq \varepsilon\}$ is compact and has strictly positive Lebesgue measure.*

Proof. Under Assumptions 3.5 and 3.6, by Lemmas 2.1 and 3.3, the map $\theta \mapsto \mathfrak{W}_p(\mu^h_\theta, \mu^h_{\theta_\star})$ is continuous. Therefore, for any $\varepsilon > 0$ the set $\tilde{B}_{\theta_\star}(\varepsilon)$ is compact since it is the pre-image of the compact set $[0, \varepsilon]$ by the continuous mapping $\theta \mapsto \mathfrak{W}_p(\mu^h_\theta, \mu^h_{\theta_\star})$. Lastly, for any $\varepsilon > 0$ the set $\tilde{B}_{\theta_\star}(\varepsilon)$ has strictly positive measure since, under Assumption 3.5, the map $\theta \mapsto \mu^h_\theta$ is injective by Lemma 3.2. □

Proposition 3.2. *Let $\mathcal{Y} = (0,1)^{d_y}$, $h_\mathcal{Y} = h$, and $1 \leq p \leq d_y$. Assume that the prior distribution $\pi(d\theta)$ admits a density (with respect to the Lebesgue measure) which is continuous and strictly positive on a neighborhood of $\theta_\star$. Then, under Assumptions 3.1,3.5 and 3.6, Assumption 3.4 holds.*

Proof. By Lemma 3.5, the set $\tilde{B}_{\theta_\star}(\varepsilon) := \{\theta \in \mathcal{H} : \mathfrak{W}^*_1(\mu^h_\star, \mu^h_\theta) \leq \varepsilon\}$ is compact and has strictly positive Lebesgue measure for any $\varepsilon > 0$. In addition, because the mapping $\theta \mapsto \mathfrak{W}^*_1(\mu^h_\star, \mu^h_\theta)$ has a unique minimum at $\theta = \theta_\star$ (Lemma 3.4) and is continuous (Lemma 3.3), it follows that $\cap_{n=1}^\infty \big(\tilde{B}_{\theta_\star}(\tilde{\varepsilon}_n/3)\big) = \{\theta_\star\}$. Since, by assumption, the prior distribution $\pi(d\theta)$ admits a density (w.r.t. Lebesgue measure) which is continuous and strictly positive on a neighborhood of $\theta_\star$, it follows that $\pi\left(\{\theta \in \mathcal{H} : \mathfrak{W}^*_1(\mu^h_\star, \mu^h_\theta) \leq \varepsilon\}\right) > 0$ for any $\varepsilon > 0$. □

References

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser Verlag AG, Basel, second edition. 7, 8

Bassetti, F., Bodini, A., and Regazzini, E. (2006). On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302. 3, 6

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6):1196–1217. 16

Bradley, R. C. et al. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(2):107–144. 17

Brown, L. D. and Purves, R. (1973). Measurable selections of extrema. *Annals of Statistics*, 1(5):902–912. 6

Dede, S. (2009). An empirical central limit theorem in l^1 for stationary sequences. *Stochastic Processes and their Applications*, 119:3494 – 3515. 7, 8

del Barrio, E., Giné, E., and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, pages 1009–1071. 7, 8

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020. 3

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738. 9, 17, 21

Gerber, M., Chopin, N., and Whiteley, N. (2017). Negative association, ordering and convergence of resampling methods. *arXiv preprint arXiv:1707.01845*. 21, 25

Hansen, B. E. (1991). Strong laws for dependent heterogeneous processes. *Econometric Theory*, 7(2):213 – 221. 20

Kolmogorov, A. N. and Rozanov, Y. A. (1960). On strong mixing conditions for stationary gaussian processes. *Theory of Probability & Its Applications*, 5(2):204–208. 17

Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimators. *Annals of Mathematical Statistics*, 41:802–828. 8, 9

Lee, A. (2012). On the choice of MCMC kernels for approximate Bayesian computation with SMC samplers. In *Proceedings of the 2012 Winter Simulation Conference*, pages 304–315. 2

Lee, A. and Łatuszyński, K. (2014). Variance bounding and geometric ergodicity of Markov chain Monte Carlo kernels for approximate Bayesian computation. *Biometrika*, 101(3):655–671. 2

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, third edition. 9

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328. 2

Murray, L. M., Lee, A., and Jacob, P. E. (2016). Parallel resampling in the particle filter. *Journal of Computational and Graphical Statistics*, 25(3):789–805. 2

Pham, T. D. and Tran, L. T. (1985). Some mixing properties of time series models. *Stochastic processes and their applications*, 19(2):297–303. 17

Pollard, D. (1980). The minimum distance method of testing. *Metrika*, 27:43–70. 8

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media. 5, 6, 10, 11, 12

Rubio, F. J. and Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654. 13

Silk, D., Filippi, S., and Stumpf, M. P. (2013). Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems. *Statistical applications in genetics and molecular biology*, 12(5):603–618. 3

Varadarajan, V. S. (1958a). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics*, 19(1):23–26. 19

Varadarajan, V. S. (1958b). Weak convergence of measures in separable metric spaces. *Sankhyā: The Indian Journal of Statistics*, 19(1):15–22. 4, 19

Villani, C. (2008). *Optimal transport, old and new*. Springer-Verlag New York. 3, 7