Supplementary materials for "On point estimation with the Wasserstein distance"

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

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 $\int_{\mathcal{Y}} \rho(y, y_0)^p d\mu_n(y) \to \int_{\mathcal{Y}} \rho(y, y_0)^p d\mu(y)$. Recall that $\mathcal{Y}$ is a subset of $\mathbb{R}^d$ for $d \geq 1$.

Theorem 1.1. *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 $\mathcal{W}_p(\mu_n, \mu) \to 0$.*

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 $\mathcal{W}_p(\mu_n, \nu_n) \to \mathcal{W}_p(\mu, \nu)$. On the other hand, if μ_n and ν_n converge weakly in the usual sense, the Wasserstein distance is only lower semicontinuous: $\liminf_{n\to\infty} \mathcal{W}_p(\mu_n, \nu_n) \geq \mathcal{W}_p(\mu, \nu)$. 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. All limits in the lemma are understood to be as $n \to \infty$.

Lemma 1.1. *Let $(\theta_n)_{n\geq 1}$ be a sequence in $\mathcal{H}$ and $\theta \in \mathcal{H}$. Suppose that either of the following conditions holds.*

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

Then, respectively,

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

Proof. This follows directly from the two assumptions and the continuity/lower semicontinuity derived from Theorem 1.1. □

Lemma 1.2. *The function $(\nu, \mu^{(m)}) \mapsto \mathbb{E}\mathcal{W}_p(\nu, \hat{\mu}_m)$ 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 $(\nu, \theta) \mapsto \mathbb{E}\mathcal{W}_p(\nu, \hat{\mu}_{\theta,m})$ is lower semicontinuous.*

*Department of Statistics, Harvard University, USA, ebernton@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

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.

Indeed, 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.

By the lower semicontinuity of the p-Wasserstein distance and Fatou's lemma,

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

The lower semicontinuity of $(\nu, \theta) \mapsto \mathbb{E}\mathcal{W}_p(\nu, \hat{\mu}_{\theta,m})$ is proved analogously to Lemma 1.1. □

2 MWE

2.1 Existence, measurability, and consistency

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

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

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

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

Theorem 2.1 (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 $\omega \in E$, $\inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta)$, and there exists $n(\omega)$ such that, for all $n \geq n(\omega)$, the sets $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ are non-empty and form a bounded sequence with*

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

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*

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

In an colloquial sense, epi-convergence is a weak notion of convergence for which the minimizer of f_n converges to the minimizer of f. Showing that the function $\theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_\theta)$ epi-converges to $\theta \mapsto \mathcal{W}_p(\mu_\star, \mu_\theta)$ almost surely is the key step in the proof of Theorem 2.1.

Proof of Theorem 2.1. First note that, for any $\nu \in \mathcal{P}(\mathcal{Y})$, the lower semicontinuity of the map $\theta \mapsto \mathcal{W}_p(\nu, \mu_\theta)$ follows from Lemma 1.1, via Assumption 2.2. Since $\inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta) = \varepsilon_\star$, the set $B_\star(\varepsilon)$ with the ε of Assumption 2.3 is non-empty, by definition of the infimum. Moreover, since $\theta \mapsto \mathcal{W}_p(\mu_\star, \mu_\theta)$ 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 $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta)$ is non-empty.

We now show that $\theta \mapsto \mathcal{W}_p(\hat{\mu}_n, \mu_\theta)$ epi-converges to $\theta \mapsto \mathcal{W}_p(\mu_\star, \mu_\theta)$ $\mathbb{P}$-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 $\theta \mapsto \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$, we know that $\inf_{\theta\in\mathcal{K}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) = \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_{\theta_n})$, for some sequence $\theta_n = \theta_n(\omega) \in \mathcal{K}$. Hence,

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

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

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

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

Theorem 7.29b) of Rockafellar and Wets (2009) implies that

$$\limsup_{n\to\infty} \inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta) = \varepsilon_\star.$$

So, for all $\alpha > 0$, there exists $n_\alpha(\omega)$, such that for $n \geq n_\alpha(\omega)$, $\inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\star + \alpha$. Let $\alpha \in (0, \varepsilon/2)$. The set $\{\theta \in \mathcal{H} : \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\star + \varepsilon/2\}$ 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, $\mathcal{W}_p(\mu_\star, \mu_\theta) \leq \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\star) + \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$. By Assumption 2.1, there exists an $n_\varepsilon(\omega)$ such that for $n \geq n_\varepsilon(\omega)$, $\mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\star) \leq \varepsilon/2$. So, if $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega)\}$, we have that $\mathcal{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_\star + \varepsilon$. This means that $\{\theta \in \mathcal{H} : \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \leq \varepsilon_\star + \varepsilon/2\} \subset B_\star(\varepsilon)$. As a consequence, for $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega)\}$, $\inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta\in B_\star(\varepsilon)} \mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$.

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

□

Theorem 2.2 (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*

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

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

Theorem 2.3 (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 $I = \{y \in G :$ 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 $y \in G$,*

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

Proof of Theorem 2.2. 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 $(\mathcal{P}_p(\mathcal{Y}), \mathcal{W}_p)$ 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 1.1 and Assumption 2.2, the map $(\mu, \theta) \mapsto \mathcal{W}_p(\mu, \mu_\theta)$ is lower semicontinuous (and therefore measurable). Hence the map $\theta \mapsto \mathcal{W}_p(\hat{\mu}_n(y), \mu_\theta)$ is also lower semicontinuous on $\mathcal{H}$ for any $y \in \mathcal{Y}^\infty$. Being the composition of measurable functions, $(y, \theta) \mapsto \mathcal{W}_p(\hat{\mu}_n(y), \mu_\theta)$ is measurable on $\mathcal{D}$. In light of this, the result follows by a direct application of Theorem 2.3.

□

2.2 Asymptotic distribution

Let $p = 1$, $\mathcal{Y} = \mathbb{R}$, and $\rho(x, y) = |x - y|$. In this case we have $\mathcal{W}_1(\mu, \nu) = \int_0^1 |F_\mu^{-1}(s) - F_\nu^{-1}(s)| ds = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)| dt$, 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). For this reason, we will occasionally use the notation $\mathcal{W}_1(\mu, \nu) = \|F_\mu - F_\nu\|_{L_1}$. Assume also that $\mathcal{H}$ is endowed with a norm: $\rho_{\mathcal{H}}(\theta, \theta') = \|\theta - \theta'\|_{\mathcal{H}}$. We recall results from del Barrio et al. (1999) and Dede (2009), after a few definitions.

Definition 2.2. Suppose that the sequence $\Omega \times \mathbb{R} \ni (\omega, t) \mapsto X_n(\omega, t)$ for all n, 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 is said to converge 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 2.4 (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 $\tilde{\alpha}_t = \sup_{u \in \mathbb{R}} \mathbb{E} |\mathbb{P}(Y_t \le u | \mathcal{F}_{-\infty}^0) - \mathbb{P}(Y_t \le u)|$. Note that for stationary sequences, $\tilde{\alpha}$-mixing is weaker than α-mixing, as defined later in Section 4.

Theorem 2.5 (Proposition 3.5 in Dede (2009)). *Suppose that $Y = (Y_t)_{t \in \mathbb{Z}}$ is ergodic and stationary, and that*

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

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 $f, g \in L_\infty(\mathbb{R})$,

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

where

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

Dede (2009) also provides other conditions, e.g. on ϕ-mixing coefficients, for which the convergence above holds. We first consider the well-specified setting, in which our results follow directly from Pollard (1980).

2.2.1 Well-specified setting

Suppose that $\mu_\star = \mu_{\theta_\star}$ for some $\theta_\star$ in the interior of $\mathcal{H}$, and consider the following assumptions:

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

$$\inf_{\theta \in \mathcal{H} : \|\theta - \theta'\|_{\mathcal{H}} \ge \varepsilon} \mathcal{W}_1(\mu_{\theta_\star}, \mu_\theta) > \delta.$$

Assumption 2.5. *There exists a non-singular $D_{\theta_\star} \in (L_1(\mathbb{R}))^{d_\theta}$ such that*

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

as $\|\theta - \theta_\star\|_{\mathcal{H}} \to 0$.

The following results contain Theorem 2.3 of the main text as a special case.

Theorem 2.6. *Suppose that* $\mu_\star = \mu_{\theta_\star}$ *for some* $\theta_\star$ *in the interior of* $\mathcal{H}$*, and that the conditions of either Theorem 2.4 or Theorem 2.5 are satisfied. Under Assumptions 2.1-2.5, the goodness-of-fit statistic satisfies*

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

as $n \to \infty$*, where* $G_\star$ *is given as in Theorem 2.4 or Theorem 2.5 respectively.*

Theorem 2.7. *Suppose that the conditions in Theorem 2.6 hold. Suppose also that the random map* $\mathcal{H} \ni u \mapsto \int_{\mathbb{R}} |G_\star(t) - \langle u, D_{\theta_\star}(t) \rangle| dt$ *has an almost surely unique infimum. Then the MWE of order 1 satisfies*

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

as $n \to \infty$*, where* $G_\star$ *is given as in Theorem 2.4 or Theorem 2.5.*

Proof. The proofs of these two results follow the steps outlined in Pollard (1980)'s Theorems 4.2 and 7.2 respectively, which also generalize to the setting where the map $\mathcal{H} \ni u \mapsto \int_{\mathbb{R}} |G_\star(t) - \langle u, D_{\theta_\star}(t) \rangle| dt$ does not necessarily have a unique minimum (see also Section 2.2.2 below). The delta methods employed therein hold for the 1-Wasserstein distance due to the representation $\mathcal{W}_1(\mu, \nu) = \|F_\mu - F_\nu\|_{L_1}$. Moreover, the well-separation of $\theta_\star$ provided by Assumption 2.4, the consistency and measurability of the MWE proved earlier, and Theorems 2.4 and 2.5 proved in del Barrio et al. (1999) and Dede (2009) respectively, guarantee that Pollard's conditions are satisfied. Note that the measurability concerns outlined in his Section 3 do not apply to here, as $L_1(\mathbb{R})$ is separable.

□

2.2.2 Misspecified setting

To study the asymptotic distribution of the MWE in the misspecified setting, we adapt the arguments outlined in Section 7 of Pollard (1980). Define $f(x, u) = \|x - \langle u, D_{\theta_\star} \rangle\|_{L_1}$ and $m(x) = \inf_u f(x, u)$. Let $\mathcal{K}$ be the class of all compact, convex, non-empty subsets of a set $L_1(\mathbb{R})$ equipped with its canonical distance. The corresponding Hausdorff metric on $\mathcal{K}$ is defined by $d_H(K_1, K_2) = \inf\{\delta > 0 : K_1 \subset K_2^\delta, K_2 \subset K_1^\delta\}$, where $K^\delta = \cup_{x \in K}\{z \in M : \|z - x\|_{L_1} \leq \delta\}$. Let $K(x, \beta) = \{u : f(x, u) \leq m(x) + \beta\}$. The function $x \mapsto K(x, \beta)$ maps into $\mathcal{K}$ and, by Pollard (1980, Lemma 7.1), is measurable. Let also $H_n = \sqrt{n}(F_n - F_{\theta_\star}) = \sqrt{n}(F_n - F_\star) + \sqrt{n}(F_\star - F_{\theta_\star})$ and $H_n^\star = G_\star + \sqrt{n}(F_\star - F_{\theta_\star})$. Let $M_n = \{\theta \in \mathcal{H} : \mathcal{W}_1(\hat{\mu}_n, \mu_\theta) \leq \inf_\theta \mathcal{W}_1(\hat{\mu}_n, \mu_\theta) + n^{-1/2}\eta_n\}$, where $\eta_n > 0$ is any sequence such that $\eta_n = o_{\mathbb{P}}(1)$ and M_n is non-empty. That is, M_n is a set of approximate MWEs.

Consider the following assumption:

Assumption 2.6. *There exists a neighborhood* N *of* $\theta_\star$ *and a constant* $c_\star > 0$ *such that for any* $\theta \in N$*,*

$$\mathcal{W}_1(\mu_\theta, \mu_\star) \geq \mathcal{W}_1(\mu_{\theta_\star}, \mu_\star) + c_\star \|\theta - \theta_\star\|_{\mathcal{H}}.$$

In the well-specified setting, this condition follows from Assumption 2.5. The next result concerns the distribution of the set M_n as n becomes large.

Theorem 2.8. *Suppose Assumptions 2.1-2.6 hold for some* $\theta_\star$ *in the interior of* $\mathcal{H}$*, and that the conditions of either Theorem 2.4 or Theorem 2.5 are satisfied. Then, there exist positive real numbers* $\beta_n \to 0$ *such that*

1. $\mathbb{P}_\star\left(\{M_n \subset \theta_\star + n^{-1/2}K(H_n, \beta_n)\}\right) \to 1$ *as* $n \to \infty$, *where* $\mathbb{P}_\star$ *denotes inner probability, and*
2. *if* F_n *and* $G_\star$ *are versions of the processes such that* $\sqrt{n}(F_n - F_\star) \to G_\star$ *in* $L_1(\mathbb{R})$ *almost surely, then* $d_H\left(K(H_n^\star, 0), K(H_n, \beta_n)\right) = o_{\mathbb{P}}(1)$.

Since $K(H_n^\star, 0) = \operatorname{argmin}_u \|G_\star + \sqrt{n}(F_\star - F_{\theta_\star}) - \langle u, D_{\theta_\star}\rangle\|_{L_1}$, one can interpret this result as saying that the limit of the set of approximate MWEs M_n behaves distributionally like the limit of the sets $\theta_\star + n^{-1/2} \operatorname{argmin}_u \|G_\star + \sqrt{n}(F_\star - F_{\theta_\star}) - \langle u, D_{\theta_\star}\rangle\|_{L_1}$ in the Hausdorff metric sense. Note that the latter sequence does not depend on the data. Since the assumptions guarantee that $\sqrt{n}(F_n - F_\star) \to G_\star$ weakly in $L_1(\mathbb{R})$, there exist versions of these variables that converge almost surely. For simplicity, we assume without loss of generality that these are the variable we work with. As noted by Pollard (1980), establishing the measurability of the sets $\{M_n \subset \theta_\star + n^{-1/2}K(H_n, \beta_n)\} \subset \Omega$ is hard, which is why the result is stated in terms of inner probability. See also Pollard (1980, pp. 67) for further comments on the sequence β_n.

Proof of Theorem 2.8. Let $\theta \in N$, where N is the set from Assumption 2.6. By Assumption 2.4 and Pollard (1980, Lemma 4.1) or the proof of Theorem 2.1, we know that the minimizers of $\|F_n - F_\theta\|_{L_1}$ will be attained in N with probability going to one. For $\theta \in N$, we have that

$$\begin{aligned}\|F_n - F_\theta\|_{L_1} &\geq \|F_\star - F_\theta\|_{L_1} - \|F_n - F_\star\|_{L_1} \quad \text{by the triangle inequality,}\\ &\geq \|F_\star - F_{\theta_\star}\|_{L_1} + c_\star\|\theta - \theta_\star\|_{\mathcal{H}} - \|F_n - F_\star\|_{L_1} \quad \text{by Assumption 2.6,}\\ &\geq \|F_n - F_{\theta_\star}\|_{L_1} + c_\star\|\theta - \theta_\star\|_{\mathcal{H}} - 2\|F_n - F_\star\|_{L_1} \quad \text{by the triangle inequality.}\end{aligned}$$

Let $\xi_n = \sqrt{n}(4\|F_n - F_\star\|_{L_1} + 2\eta_n)/c_\star$ and $S_n = \{\theta : \sqrt{n}\|\theta - \theta_\star\|_{\mathcal{H}} \leq \xi_n\}$. Then, by the assumptions on η_n and $\sqrt{n}(F_n - F_\star)$, we know that $n^{-1/2}\xi_n = o_{\mathbb{P}}(1)$. If $\theta \in N \cap S_n^c$, then by the inequality derived above, $\|F_n - F_\theta\|_{L_1} > \|F_n - F_{\theta_\star}\|_{L_1} + 2(\|F_n - F_\star\|_{L_1} + \eta_n)$. Thus, with inner probability going to one, it has to be that $M_n \subset S_n$.

Next, we approximate $\theta \mapsto \sqrt{n}\|F_n - F_\theta\|_{L_1}$ with the convex map $\theta \mapsto \sqrt{n}\|F_n - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle\|_{L_1}$ over the set S_n. First, note that Assumption 2.5 implies that the remainder $R_\theta = F_\theta - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle$ satisfies $\|R_\theta\|_{L_1} \leq \|\theta - \theta_\star\|_{\mathcal{H}} \cdot \Delta(\|\theta - \theta_\star\|_{\mathcal{H}})$, where Δ is an increasing function such that $\Delta(t) = o(1)$ as $t \to 0$. Define $\Gamma_n = \sup_{\theta \in S_n} |\sqrt{n}\|F_n - F_\theta\|_{L_1} - \sqrt{n}\|F_n - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle\|_{L_1}|$. We then have that

$$\begin{aligned}\Gamma_n &= \sup_{\theta \in S_n} \left|\sqrt{n}\|F_n - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle - R_\theta\|_{L_1} - \sqrt{n}\|F_n - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle\|_{L_1}\right|\\ &\leq \sup_{\theta \in S_n} \sqrt{n}\|R_\theta\|_{L_1} \quad \text{by the triangle inequality,}\\ &\leq \sup_{\theta \in S_n} \sqrt{n}\|\theta - \theta_\star\|_{\mathcal{H}} \cdot \Delta(\|\theta - \theta_\star\|_{\mathcal{H}}) \quad \text{by Assumption 2.5,}\\ &\leq \xi_n \Delta\left(\frac{\xi_n}{\sqrt{n}}\right) = o_{\mathbb{P}}(1) \quad \text{by the definitions of } S_n \text{ and } \xi_n.\end{aligned}$$

In other words, we have uniform control over the difference between $\theta \mapsto \sqrt{n}\|F_n - F_\theta\|_{L_1}$ and its convex approximation over S_n. Moreover, the map $\theta \mapsto \sqrt{n}\|F_n - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle\|_{L_1}$ also attains its minimum on S_n with probability going to one, since for $\theta \in N$ such that $\Delta(\|\theta - \theta_\star\|_{\mathcal{H}}) \leq c_\star/2$,

$$\begin{aligned}\|F_n - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle\|_{L_1} &= \|F_n - F_\theta + R_\theta\|_{L_1}\\ &\geq \|F_n - F_\theta\|_{L_1} - \|R_\theta\|_{L_1} \quad \text{by the triangle inequality,}\end{aligned}$$

$$\geq \|F_n - F_{\theta_\star}\|_{L_1} + c_\star\|\theta-\theta_\star\|_{\mathcal{H}} - 2\|F_n - F_\star\|_{L_1} - \|\theta-\theta_\star\|_{\mathcal{H}}\cdot\Delta(\|\theta-\theta_\star\|_{\mathcal{H}}) \quad \text{by Ass. 2.6, tri. ineq.,}$$
$$\geq \|F_n - F_{\theta_\star}\|_{L_1} + \frac{1}{2}c_\star\|\theta-\theta_\star\|_{\mathcal{H}} - 2\|F_n - F_\star\|_{L_1}.$$

Hence, if $\theta \in N \cap S_n^c$ and $\Delta(\|\theta-\theta_\star\|_{\mathcal{H}}) \leq c_\star/2$, then $\|F_n - F_\theta\|_{L_1} > \|F_n - F_{\theta_\star}\|_{L_1} + \eta_n = \|F_n - F_{\theta_\star} - \langle 0, D_{\theta_\star}\rangle\|_{L_1} + \eta_n$. In other words, $m(H_n) = \inf_{u\in L_n} f(H_n, u)$ with probability going to one, where we have used the reparameterization $L_n = \{u : u = \sqrt{n}(\theta-\theta_\star), \theta \in S_n\}$, or equivalently $S_n = \theta_\star + n^{-1/2}L_n$.

Now, since $\Gamma_n = o_{\mathbb{P}}(1)$, we can find a sequence of positive real numbers $\gamma_n \to 0$ such that $\mathbb{P}(\Gamma_n \leq \gamma_n) \to 1$. Similarly, we can find $\delta_n > 0$ and $\alpha_n > 0$ such that $\mathbb{P}(\eta_n \leq \delta_n) \to 1$ and $\mathbb{P}(\|H_n - H_n^\star\|_{L_1} \leq \alpha_n) \to 1$. Define $\beta_n = \max\{2\gamma_n + \delta_n, 2\alpha_n\}$. Let τ be such that $\tau \in L_n$ and $\theta_\star + n^{-1/2}\tau \in M_n$, and suppose that $\Gamma_n \leq \gamma_n$ and $\eta_n \leq \delta_n$. By combining the approximations developed above, we have that

$$\begin{aligned} m(H_n) &\geq \inf_{u\in L_n} \sqrt{n}\|F_n - F_{\theta_\star + n^{-1/2}u}\|_{L_1} - \gamma_n \\ &\geq \sqrt{n}\|F_n - F_{\theta_\star + n^{-1/2}\tau}\|_{L_1} - \gamma_n - \delta_n \\ &\geq f(H_n, \tau) - 2\gamma_n - \delta_n. \end{aligned}$$

Since $2\gamma_n + \delta_n \leq \beta_n$, we have that $\tau \in K(H_n, \beta_n)$. This proves the first part of the theorem, as the events considered above all hold with (inner) probability going to one.

By the triangle inequality, $u \in K(H_n^\star, 0)$ implies that $u \in K(H_n, 2\|H_n - H_n^\star\|_{L_1})$. Hence, with probability going to one, $K(H_n^\star, 0) \subset K(H_n, \beta_n)$. Similarly, $u \in K(H_n, \beta_n)$ implies that $u \in K(H_n^\star, \beta_n + 2\|H_n - H_n^\star\|_{L_1})$. Recall that $\beta_n + 2\|H_n - H_n^\star\|_{L_1} \to 0$ almost surely. Let $E \subset \Omega$ denote the set on which this occurs. Then, for every every $\delta > 0$, there exists $n(\omega)$ such that for $n \geq n(\omega)$, $K(H_n(\omega), \beta_n) \subset K(H_n^\star(\omega), 0)^\delta$. By the definition of the Hausdorff metric, these set inclusions imply that $d_H(K(H_n^\star, 0), K(H_n, \beta_n)) = o_{\mathbb{P}}(1)$. □

2.2.3 Differentiability condition

The differentiability condition in Assumption 2.5 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 $D_{\theta_\star}(t) = \int_{-\infty}^t 2\xi_{\theta_\star}(y)\dot{\xi}_{\theta_\star}(y)dy$ for $t \in S$ and zero elsewhere. Then $\int_{\mathbb{R}}|F_\theta(t) - F_{\theta_\star}(t) - \langle\theta-\theta_\star, D_{\theta_\star}(t)\rangle|dt = o(\|\theta-\theta_\star\|_{\mathcal{H}})$, as $\|\theta-\theta_\star\|_{\mathcal{H}} \to 0$.*

Proof.

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

$$\leq \int_S \int_{\mathbb{R}} |\xi_\theta^2(y) - \xi_{\theta_\star}^2(y) - 2\xi_{\theta_\star}(y)\langle \theta - \theta_\star, \dot{\xi}_{\theta_\star}(y)\rangle| dydt$$
$$\leq c \int_{\mathbb{R}} \langle \theta - \theta_\star, \dot{\xi}_{\theta_\star}(y)\rangle^2 + R^2_{\theta-\theta_\star}(y) + 2|\xi_{\theta_\star}(y) R_{\theta-\theta_\star}(y)| + 2|\langle \theta - \theta_\star, \dot{\xi}_{\theta_\star}(y)\rangle R_{\theta-\theta_\star}(y)| dy$$
$$= o(\|\theta - \theta_\star\|_{\mathcal{H}}),$$

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

3 MEWE

3.1 Existence, measurability, and consistency

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

Assumption 3.1. *For any $m \geq 1$, if $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$, then $\mu_{\theta_n}^{(m)} \Rightarrow \mu_\theta^{(m)}$ as $n \to \infty$.*

Assumption 3.2. *If $\rho_{\mathcal{H}}(\theta_n, \theta) \to 0$, then $\mathbb{E}_n \mathcal{W}_p(\mu_{\theta_n}, \hat{\mu}_{\theta_n,n}) \to 0$ as $n \to \infty$.*

Assumption 3.1 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 $m \geq 1$. Assumption 3.2 is implied by $\sup_{\theta \in \mathcal{H}} \mathbb{E}_n \mathcal{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) \to 0$, 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 2.1 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 3.1. *Under Assumptions 2.1-2.3 and 3.1-3.2, there exists a set $E \subset \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, $\inf_{\theta \in \mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \to \inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta)$, and there exists $n(\omega)$ such that, for all $n \geq n(\omega)$, the sets $\operatorname{argmin}_{\theta \in \mathcal{H}} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$ are non-empty and form a bounded sequence with*

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

Proof of Theorem 3.1. As before, for any $\nu \in \mathcal{P}(\mathcal{Y})$, lower semicontinuity of the map $\theta \mapsto \mathcal{W}_p(\nu, \mu_\theta)$ follows from Lemma 1.1, via Assumption 2.2. Since $\inf_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta) = \varepsilon_\star$, $B_\star(\varepsilon)$ with the ε of Assumption 2.3 is non-empty, by definition of the infimum. Moreover, since $\theta \mapsto \mathcal{W}_p(\mu_\star, \mu_\theta)$ 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 $\operatorname{argmin}_{\theta \in \mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta)$ is non-empty.

We show that $\theta \mapsto \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m(n)})$ epi-converges to $\theta \mapsto \mathcal{W}_p(\mu_\star, \mu_\theta)$ $\mathbb{P}$-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 $\theta \mapsto \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$, ensured by Lemma 1.2 and Assumption 3.1, we know that $\inf_{\theta \in \mathcal{K}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) = \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)})$, for some sequence $\theta_n = \theta_n(\omega) \in \mathcal{K}$. Hence,

$$\liminf_{n\to\infty} \inf_{\theta \in \mathcal{K}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$$
$$= \liminf_{n\to\infty} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta_n,m(n)})$$
$$= \lim_{k\to\infty} \mathbb{E}_{m(n_k)} \mathcal{W}_p(\hat{\mu}_{n_k}(\omega), \hat{\mu}_{\theta_{n_k},m(n_k)}) \quad \exists \text{ subsequence converging to the } \liminf,$$
$$= \lim_{\ell\to\infty} \mathbb{E}_{m(n_{k_\ell})} \mathcal{W}_p(\hat{\mu}_{n_{k_\ell}}(\omega), \hat{\mu}_{\theta_{n_{k_\ell}},m(n_{k_\ell})}) \quad \exists \text{ subsequence } \theta_{n_{k_\ell}} \to \bar{\theta} \in \mathcal{K} \text{ by compactness},$$

$$
\begin{aligned}
&= \liminf_{\ell\to\infty} \mathbb{E}_{m(n_{k_\ell})} \mathcal{W}_p(\hat\mu_{n_{k_\ell}}(\omega), \hat\mu_{\theta_{n_{k_\ell}}, m(n_{k_\ell})}) \\
&\geq \liminf_{\ell\to\infty} [\mathcal{W}_p(\hat\mu_{n_{k_\ell}}(\omega), \mu_{\theta_{n_{k_\ell}}}) - \mathbb{E}_{m(n_{k_\ell})} \mathcal{W}_p(\mu_{\theta_{n_{k_\ell}}}, \hat\mu_{\theta_{n_{k_\ell}}, m(n_{k_\ell})})] \quad \text{by the triangle inequality,} \\
&\geq \liminf_{\ell\to\infty} \mathcal{W}_p(\hat\mu_{n_{k_\ell}}(\omega), \mu_{\theta_{n_{k_\ell}}}) - \limsup_{\ell\to\infty} \mathbb{E}_{m(n_{k_\ell})} \mathcal{W}_p(\mu_{\theta_{n_{k_\ell}}}, \hat\mu_{\theta_{n_{k_\ell}}, m(n_{k_\ell})}) \\
&\geq \mathcal{W}_p(\mu_\star, \mu_{\bar\theta}) \quad \text{by l.s.c., Assumptions 2.1, 2.2 and 3.2, } \omega \in E, \\
&\geq \inf_{\theta\in\mathcal{K}} \mathcal{W}_p(\mu_\star, \mu_\theta).
\end{aligned}
$$

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

$$
\begin{aligned}
&\limsup_{n\to\infty} \inf_{\theta\in\mathcal{O}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m(n)}) \\
&\qquad \leq \limsup_{n\to\infty} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta_n,m(n)}) \\
&\qquad \leq \limsup_{n\to\infty} [\mathcal{W}_p(\hat\mu_n(\omega), \mu_\star) + \mathcal{W}_p(\mu_\star, \mu_{\theta_n}) + \mathbb{E}_{m(n)} \mathcal{W}_p(\mu_{\theta_n}, \hat\mu_{\theta_n,m(n)})] \quad \text{by the triangle inequality,} \\
&\qquad = \limsup_{n\to\infty} \mathcal{W}_p(\mu_\star, \mu_{\theta_n}) \quad \text{by Assumptions 2.1 and 3.2, } \omega \in E, \\
&\qquad = \inf_{\theta\in\mathcal{O}} \mathcal{W}_p(\mu_\star, \mu_\theta) \quad \text{by definition of } \theta_n.
\end{aligned}
$$

Theorem 7.29b) of Rockafellar and Wets (2009) implies that

$$
\limsup_{n\to\infty} (\inf_{\theta\in\mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m(n)})) \leq \inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta) = \varepsilon_\star.
$$

Hence, for all $\alpha > 0$, there exists $n_\alpha(\omega)$, such that $n \geq n_\alpha(\omega)$ implies that $\inf_{\theta\in\mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m(n)}) \leq \varepsilon_\star + \alpha$. Let $\alpha \in (0, \varepsilon/3)$. The set $\{\theta \in \mathcal{H} : \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m(n)}) \leq \varepsilon_\star + \varepsilon/3\}$ 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, $\mathcal{W}_p(\mu_\star, \mu_\theta) \leq \mathcal{W}_p(\hat\mu_n(\omega), \mu_\star) + \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m(n)}) + \mathbb{E}_{m(n)} \mathcal{W}_p(\mu_\theta, \hat\mu_{\theta,m(n)})$. By Assumption 2.1, there exists an $n_\varepsilon(\omega)$ such that for $n \geq n_\varepsilon(\omega)$, $\mathcal{W}_p(\hat\mu_n(\omega), \mu_\star) \leq \varepsilon/3$. By Assumption 3.2, there exists an $\hat n(\omega)$ such that for $n \geq \hat n(\omega)$, $\mathbb{E}_{m(n)} \mathcal{W}_p(\mu_\theta, \hat\mu_{\theta,m(n)}) \leq \varepsilon/3$. So, if $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega), \hat n(\omega)\}$, we have that $\mathcal{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_\star + \varepsilon$. This means that $\{\theta \in \mathcal{H} : \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m(n)}) \leq \varepsilon_\star + \varepsilon/3\} \subset B_\star(\varepsilon)$. As a consequence, for $n \geq \max\{n_\alpha(\omega), n_\varepsilon(\omega), \hat n(\omega)\}$, $\inf_{\theta\in\mathcal{H}} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m(n)}) = \inf_{\theta\in B_\star(\varepsilon)} \mathbb{E}_{m(n)} \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta,m(n)})$.

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

□

3.2 Convergence to the MWE

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 $\varepsilon_n = \inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat\mu_n, \mu_\theta)$.

Assumption 3.3. *For some $\varepsilon > 0$, the set $B_n(\varepsilon) = \{\theta \in \mathcal{H} : \mathcal{W}_p(\hat\mu_n, \mu_\theta) \leq \varepsilon_n + \varepsilon\}$ is bounded.*

Theorem 3.2. *Under Assumptions 2.2 and 3.1-3.3,* $\inf_{\theta\in\mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) \to \inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat\mu_n, \mu_\theta)$*, and there exists an* $\hat m$ *such that, for all* $m \geq \hat m$*, the sets* $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m})$ *are non-empty and form a bounded sequence with*

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

Proof of Theorem 3.2. Lower semicontinuity of the map $\theta \mapsto \mathcal{W}_p(\hat\mu_n, \mu_\theta)$ follows from Lemma 1.1, via Assumption 2.2. Since $\inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat\mu_n, \mu_\theta) = \varepsilon_n$, $B_n(\varepsilon)$ with the ε of Assumption 2.3 is non-empty, by definition of the infimum. Moreover, since $\theta \mapsto \mathcal{W}_p(\hat\mu_n, \mu_\theta)$ is lower semicontinuous, the set $B_n(\varepsilon)$ is closed. By Assumption 3.3, $B_n(\varepsilon)$ is therefore compact. In other words, again by lower semicontinuity, the set $\operatorname{argmin}_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat\mu_n, \mu_\theta)$ is non-empty.

We show that $\theta \mapsto \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m})$ epi-converges to $\theta \mapsto \mathcal{W}_p(\hat\mu_n, \mu_\theta)$ as $m \to \infty$. Fix $\mathcal{K} \subset \mathcal{H}$ compact. By lower semicontinuity of $\theta \mapsto \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m})$, ensured by Lemma 1.2 and Assumption 3.1, we know that $\inf_{\theta\in\mathcal{K}} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) = \mathbb{E}_m \mathcal{W}_p(\hat\mu_n(\omega), \hat\mu_{\theta_m,m})$, for some sequence $\theta_m \in \mathcal{K}$. Hence,

$$\begin{aligned}
&\liminf_{m\to\infty} \inf_{\theta\in\mathcal{K}} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) \\
&\quad = \liminf_{m\to\infty} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta_m,m}) \\
&\quad = \lim_{k\to\infty} \mathbb{E}_{m_k} \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta_{m_k},m_k}) \quad \exists \text{ subsequence converging to the } \liminf, \\
&\quad = \lim_{\ell\to\infty} \mathbb{E}_{m_{k_\ell}} \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta_{m_{k_\ell}},m_{k_\ell}}) \quad \exists \text{ subsequence } \theta_{m_{k_\ell}} \to \bar\theta \in \mathcal{K} \text{ by compactness}, \\
&\quad = \liminf_{\ell\to\infty} \mathbb{E}_{m_{k_\ell}} \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta_{m_{k_\ell}},m_{k_\ell}}) \\
&\quad \geq \liminf_{\ell\to\infty} [\mathcal{W}_p(\hat\mu_n, \mu_{\theta_{m_{k_\ell}}}) - \mathbb{E}_{m_{k_\ell}} \mathcal{W}_p(\mu_{\theta_{m_{k_\ell}}}, \hat\mu_{\theta_{m_{k_\ell}},m_{k_\ell}})] \quad \text{by the triangle inequality}, \\
&\quad \geq \liminf_{\ell\to\infty} \mathcal{W}_p(\hat\mu_n, \mu_{\theta_{m_{k_\ell}}}) - \limsup_{\ell\to\infty} \mathbb{E}_{m_{k_\ell}} \mathcal{W}_p(\mu_{\theta_{m_{k_\ell}}}, \hat\mu_{\theta_{m_{k_\ell}},m_{k_\ell}}) \\
&\quad \geq \mathcal{W}_p(\hat\mu_n, \mu_{\bar\theta}) \quad \text{by l.s.c., Assumptions 2.2 and 3.2}, \\
&\quad \geq \inf_{\theta\in\mathcal{K}} \mathcal{W}_p(\hat\mu_n, \mu_\theta).
\end{aligned}$$

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

$$\begin{aligned}
\limsup_{m\to\infty} \inf_{\theta\in\mathcal{O}} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) &\leq \limsup_{m\to\infty} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta_m,m}) \\
&\leq \limsup_{m\to\infty} [\mathcal{W}_p(\hat\mu_n, \mu_{\theta_m}) + \mathbb{E}_m \mathcal{W}_p(\mu_{\theta_m}, \hat\mu_{\theta_m,m})] \quad \text{by the triangle inequality}, \\
&= \limsup_{m\to\infty} \mathcal{W}_p(\hat\mu_n, \mu_{\theta_m}) \quad \text{by Assumption 3.2}, \\
&= \inf_{\theta\in\mathcal{O}} \mathcal{W}_p(\mu_\star, \mu_\theta) \quad \text{by definition of } \theta_m.
\end{aligned}$$

Theorem 7.29b) of Rockafellar and Wets (2009) implies that

$$\limsup_{m\to\infty} (\inf_{\theta\in\mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m})) \leq \inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\hat\mu_n, \mu_\theta) = \varepsilon_n.$$

Hence, for all $\alpha > 0$, there exists m_α, such that for $m \geq m_\alpha$, $\inf_{\theta\in\mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) \leq \varepsilon_n + \alpha$. Let $\alpha \in (0, \varepsilon/2)$. The set $\{\theta \in \mathcal{H} : \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) \leq \varepsilon_n + \varepsilon/2\}$ is non-empty for $m \geq m_\alpha$, by definition of the infimum. Let θ belong to this set. Then, by the triangle inequality, $\mathcal{W}_p(\hat\mu_n, \mu_\theta) \leq \mathbb{E}_m \mathcal{W}_p(\hat\mu_n, \hat\mu_{\theta,m}) + \mathbb{E}_m \mathcal{W}_p(\mu_\theta, \hat\mu_{\theta,m})$. By Assumption 3.2, there exists an $\hat m$ such that for $m \geq \hat m$, $\mathbb{E}_m \mathcal{W}_p(\mu_\theta, \hat\mu_{\theta,m}) \leq \varepsilon/2$. So, if $m \geq \max\{m_\alpha, \hat m\}$, we have that $\mathcal{W}_p(\mu_\star, \mu_\theta) \leq \varepsilon_n + \varepsilon$. This means that $\{\theta \in \mathcal{H} : \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \leq \varepsilon_n + \varepsilon/2\} \subset B_n(\varepsilon)$. As a consequence, for $m \geq \max\{m_\alpha, \hat{m}\}$, $\inf_{\theta \in \mathcal{H}} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) = \inf_{\theta \in B_n(\varepsilon)} \mathbb{E}_m \mathcal{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$.

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

Theorem 3.3 (Measurability of the MEWE). *Suppose that $\mathcal{H}$ is a σ-compact Borel measurable subset of $\mathbb{R}^{d_\theta}$. Under Assumption 3.1, 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*

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

Proof. The proof is identical to that of Theorem 2.2, applying Lemma 1.2 instead of 1.1. □

4 Checking the assumptions

The following proposition gives three data-generating mechanisms for which $\mathcal{W}_p(\hat{\mu}_n, \mu_\star) \to 0$ $\mathbb{P}$-almost surely, which is Assumption 2.1. The three conditions below are mainly chosen for illustrative purposes, and are by no means exhaustive. We first give definitions that are used in the conditions. We denote by $\mathcal{F}$ the measurable sets of Ω.

Definition 4.1. 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 $(Y_{t_1}, \ldots, Y_{t_k})$ and $(Y_{t_1+\tau}, \ldots, Y_{t_k+\tau})$ have the same distribution.

Definition 4.2. 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 4.3. 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 4.4. 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 $\mathcal{F}_{-\infty}^k = \sigma(Y_i : i \leq k)$ and $\mathcal{F}_k^\infty = \sigma(Y_i : i \geq k)$.

Proposition 4.1. *Suppose that $Y = (Y_t)_{t \in \mathbb{Z}}$ 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 $1 \leq \max(r, p) < q < 2r$ (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 $\omega \in E$, $\mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\star) \to 0$.

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 1.1, 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 $\mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\star) \to 0$ 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 $\sigma(f(Y_i) : i \leq k) \subset \sigma(Y_i : i \leq k)$. 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 $\mathcal{W}_p(\hat{\mu}_n(\omega), \mu_\star) \to 0$ for all ω in a set E defined analogously to the one for the second set of conditions. ☐

The following proposition can be used to verify Assumption 2.4.

Proposition 4.2. *Suppose that either of the conditions of Lemma 1.1 holds. Suppose that there exists a proper, connected and compact subset $\mathcal{S} \subset \mathcal{H}$ with positive Lebesgue measure such that $\inf_{\theta\in\mathcal{H}\setminus\mathcal{S}} \mathcal{W}_p(\mu_\star, \mu_\theta) > \inf_{\theta\in\mathcal{H}} \mathcal{W}_p(\mu_\star, \mu_\theta)$. Then there exists a $\theta_\star$ attaining the infimum of $\theta \mapsto \mathcal{W}_p(\mu_\star, \mu_\theta)$. If $\theta_\star$ is unique, then it is well-separated.*

Proof. Since $\theta \mapsto \mathcal{W}_p(\mu_\star, \mu_\theta)$ 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

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

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, $\theta \mapsto \mathcal{W}_p(\mu_\star, \mu_\theta)$ 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. ☐

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. 5

Bassetti, F., Bodini, A., and Regazzini, E. (2006). On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302. 1, 4

Brown, L. D. and Purves, R. (1973). Measurable selections of extrema. *Annals of Statistics*, 1(5):902–912. 4

Dede, S. (2009). An empirical central limit theorem in l^1 for stationary sequences. *Stochastic Processes and their Applications*, 119:3494 – 3515. 5, 6

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. 5, 6

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

Hansen, B. E. (1991). Strong laws for dependent heterogeneous processes. *Econometric Theory*, 7(2):213 – 221. 13

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

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, third edition. 8

Pollard, D. (1980). The minimum distance method of testing. *Metrika*, 27:43–70. 5, 6, 7

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media. 2, 3, 4, 10, 11, 12

Varadarajan, V. S. (1958a). On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics*, 19(1):23–26. 13

Varadarajan, V. S. (1958b). Weak convergence of measures in separable metric spaces. *Sankhyā: The Indian Journal of Statistics*, 19(1):15–22. 2, 13

Villani, C. (2008). *Optimal transport, old and new*. Springer-Verlag New York. 1, 4