A Further Results on the MPRW and MEPRW Estimators

In this section, we discuss the measurability of the MPRW and MEPRW estimators. For a generic function f on the domain $\mathcal{X}$, we define $\delta\text{-argmin}_{x\in\mathcal{X}} f = \{x \in \mathcal{X} : f(x) \le \inf_{x\in\mathcal{X}} f + \delta\}$. Our results are summarized in the following two theorems.

Theorem A.1 *Under Assumption 3.1, for any $n \ge 1$ and $\delta > 0$, there exists a Borel measurable function $\widehat{\theta}_n : \Omega \to \Theta$ such that*

$$\widehat{\theta}_n(\omega) \in \begin{cases} \operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) & \text{if this set is nonempty,} \\ \delta\text{-}\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) & \text{otherwise.} \end{cases}$$

Theorem A.2 *Under Assumption 3.1, for any $n \ge 1$, $m \ge 1$ and $\delta > 0$, there exists a Borel measurable function $\widehat{\theta}_{n,m} : \Omega \to \Theta$ such that*

$$\widehat{\theta}_{n,m}(\omega) \in \begin{cases} \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m}) \mid X_{1:n}] & \text{if this set is nonempty,} \\ \delta\text{-}\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m}) \mid X_{1:n}] & \text{otherwise.} \end{cases}$$

We also present the asymptotic distribution of the goodness-of-fit statistics as well as the MPRW estimator in the well-specified setting and establish the rate of convergence. For this we require the well separability of the model in Assumption A.1 and the non-singularity of $D_\star$ in Assumption A.2 to take place of the local strong identifiability in Assumption 3.8.

Assumption A.1 *For any $\epsilon > 0$, there exists $\delta > 0$ so that $\inf_{\theta\in\Theta:\|\theta-\theta_\star\|_\Theta\ge\epsilon} \mathcal{PW}_{1,1}(\mu_{\theta_\star}, \mu_\theta) > \delta$.*

Assumption A.2 *There exists a non-singular $D_\star$ such that Assumption 3.6 holds true.*

Theorem A.3 *Suppose that $\mu_\star = \mu_{\theta_\star}$ for some $\theta_\star$ in the interior of Θ. Under Assumption 3.1-3.3, 3.6-3.7 and A.1-A.2, the goodness-of-fit statistics satisfies*

$$\sqrt{n} \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_\theta) \Rightarrow \inf_{\theta\in\Theta} \max_{u\in\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t)\rangle| \, dt, \quad \text{as } n \to +\infty.$$

Suppose also that the random map $\theta \to \max_{u\in\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t)\rangle| \, dt$ has a unique infimum almost surely. Then the MPRW estimator of order 1 satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_\star) \Rightarrow \operatorname*{argmin}_{\theta\in\Theta} \max_{u\in\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t)\rangle| \, dt, \quad \text{as } n \to +\infty.$$

Both the weak convergence results are valid for the metric induced by the norm $\|\cdot\|_L$.

B Postponed Proofs in Subsection 3.1

This section lays out the detailed proofs for Lemma 3.1, Theorem 3.2 and 3.3.

B.1 Preliminary technical results

For completeness, we collect several preliminary technical results[6] which will be used in the proofs.

Theorem B.1 (Prokhorov's theorem) *Let $\mathscr{P}(\mathbb{R}^d)$ denote the collection of all probability measures defined on $\mathbb{R}^d$ with the Borel σ-algebra and $\{\mu_i\}_{i\in\mathbb{N}}$ is a tight sequence in $\mathscr{P}(\mathbb{R}^d)$. Then every subsequence of $\{\mu_i\}_{i\in\mathbb{N}}$ has a subsequence that converges weakly in $\mathscr{P}(\mathbb{R}^d)$. Moreover, if every weakly convergent subsequence has the same limit, the whole sequence converges weakly to this limit.*

[6]For the Prokhorov's theorem, we only present the results on the Euclidean space. For more results on general separable metric space, we refer the interested readers to Billingsley (2013).

Theorem B.2 (Theorem 4.1 in Villani (2008)) *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two Polish probability spaces; let $a : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ and $b : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ be upper semi-continuous such that a and b are absolutely integrable with respect to the measures μ and ν respectively. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be lower semi-continuous, such that $c(x, y) \geq a(x) + b(y)$ for all x, y. Then there exists an optimal coupling $\pi \in \Pi(\mu, \nu)$ which minimizes the total cost $\mathbb{E}[c(X, Y)]$.*

Lemma B.3 (Lemma 4.4 in Villani (2008)) *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. Let $P \subseteq \mathscr{P}(\mathcal{X})$ and $Q \subseteq \mathscr{P}(\mathcal{Y})$ be tight subsets of $\mathscr{P}(\mathcal{X})$ and $\mathscr{P}(\mathcal{Y})$ respectively. Then the set of all transportation plans whose marginals lie in P and Q respectively, is itself tight in $\mathscr{P}(\mathcal{X} \times \mathcal{Y})$.*

Theorem B.4 (Theorem 6.9 in Villani (2008)) *Let $(\mathcal{X}, d)$ be a Polish space and $p \in [1, +\infty)$. The Wasserstein distance $\mathcal{W}_p$ metrizes the weak convergence in $\mathscr{P}_p(\mathcal{X})$. That is, if $\{\mu_i\}_{i \in \mathbb{N}_n}$ is a sequence of measures in $\mathscr{P}_p(\mathcal{X})$ and $\mu \in \mathscr{P}_p(\mathcal{X})$, then $\mu_i \Rightarrow \mu$ if and only if $\mathcal{W}_p(\mu_i, \mu) \to 0$.*

Definition B.1 (Lower semi-continuity) *We say that $f : \mathcal{X} \to \mathbb{R}$ is lower semi-continuous if for any $x_0 \in \mathcal{X}$ and any $y < f(x_0)$, there exists a neighborhood U of x_0 such that $f(x) > y$ for all x in U. In the case of a metric space, this is equivalent to $\liminf_{x \to x_0} f(x) \geq f(x_0)$ for any $x_0 \in \mathcal{X}$.*

B.2 Proof of Lemma 3.1

We first show that, for any $\mu \in \mathscr{P}_p(\mathbb{R}^d)$ and $\nu \in \mathscr{P}_p(\mathbb{R}^d)$, the following inequality holds true,

$$\underline{\mathcal{PW}}_{p,k}(\mu, \nu) \leq \overline{\mathcal{PW}}_{p,k}(\mu, \nu) \leq \mathcal{W}_p(\mu, \nu). \tag{6}$$

Indeed, by the definition of $\underline{\mathcal{PW}}_{p,k}$ and $\overline{\mathcal{PW}}_{p,k}$, the first inequality is trivial. For the second inequality, we derive from the definition of $\overline{\mathcal{PW}}_{p,k}$ that

$$\overline{\mathcal{PW}}^p_{p,k}(\mu, \nu) = \sup_{E \in \mathbb{S}_{d,k}} \mathcal{W}^p_p(E^\star_\# \mu, E^\star_\# \nu) = \sup_{E \in \mathbb{S}_{d,k}} \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|E^\top (x - y)\|^p \, d\pi(x, y).$$

Since $E \in \mathbb{S}_{d,k}$, we have $\|E^\top (x - y)\| \leq \|x - y\|$. Thus, we have $\overline{\mathcal{PW}}^p_{p,k}(\mu, \nu) \leq \mathcal{W}^p_p(\mu, \nu)$. Putting these pieces together yields Eq. (6). For any sequence $\{\mu_i\}_{i \in \mathbb{N}} \subseteq \mathscr{P}_p(\mathbb{R}^d)$ and $\mu \in \mathscr{P}_p(\mathbb{R}^d)$, we conclude from Eq. (6) that $\mathcal{W}_p(\mu_i, \mu) \to 0$ implies $\overline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ and $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$.

The remaining step is to show that $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ implies $\mathcal{W}_p(\mu_i, \mu) \to 0$. Indeed, we first prove that $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ implies $\mu_i \Rightarrow \mu$. Let $Z_i \sim \mu_i$, we have $E^\top Z_i \sim E^\star_\# \mu_i$. By the definition of the IPRW distance (cf. Definition 3) and using the fact that $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$, we have $(\|E^\top Z_i)\|^p)_{i \in \mathbb{N}}$ is uniformly integrable for all $E \in \mathbb{S}_{d,k}$. Since $\mathbb{S}_{d,k}$ is compact, there exists a finite set $\{E_1, E_2, \ldots, E_I\} \subseteq \mathbb{S}_{d,k}$ so that $\|x\| \leq \sum_{j=1}^I \|E_j^\top x\|$ for all $x \in \mathbb{R}^d$. Therefore, we have

$$\|Z_i\|^p \leq \left(\sum_{j=1}^I \|E_j^\top Z_i\| \right)^p \leq I^p \left(\max_{1 \leq j \leq I} \|E_j^\top Z_i\|^p \right) \leq I^p \left(\sum_{j=1}^I \|E_j^\top Z_i\|^p \right).$$

Therefore, we deduce that $(\|Z_i\|^p)_{i \in \mathbb{N}}$ is uniformly integrable which implies the tightness of $\{\mu_i\}_{i \in \mathbb{N}}$. Using the Prokhorov's theorem (cf. Theorem B.1), we obtain that every subsequence of $\{\mu_i\}_{i \in \mathbb{N}}$ has a weakly convergent subsequence.

The next step is to show that all the weakly convergent subsequences converge to the same probability measure μ. We fix an arbitrary subsequence and for simplicity abbreviate the subscripts and still denote it by $\{\mu_i\}_{i \in \mathbb{N}}$. Let $\tilde{\mu}_i$ be the limit of any given weakly convergent subsequence $(\mu_{i_j})_{j \in \mathbb{N}}$, we need to prove that $\tilde{\mu}_i = \mu$. In particular, we define the characteristic function for any probability measure ν as follows,

$$\Phi_\nu(z) := \int_{\mathbb{R}^d} e^{\mathrm{i}\langle z, x \rangle} \, d\nu(x) \quad \text{for all } z \in \mathbb{R}^d.$$

Since $\mu_{i_j} \Rightarrow \tilde{\mu}_i$, we have $\Phi_{\mu_{i_j}}(z) \to \Phi_{\tilde{\mu}_i}(z)$ for all $z \in \mathbb{R}^d$. Thus, we need to show that $\Phi_{\mu_{i_j}}(z) \to \Phi_\mu(z)$ for all $z \in \mathbb{R}^d$. This is trivial when $z = \mathbf{0}_d$ since $\Phi_{\mu_{i_j}}(\mathbf{0}_d) = \Phi_\mu(\mathbf{0}_d) = 1$ for all $j \in \mathbb{N}$. Otherwise, let $r := \|z\|$ and

$v := z/\|z\|$, we have

$$\lim_{j\to+\infty} \Phi_{\mu_{i_j}}(z) = \lim_{j\to+\infty} \int_{\mathbb{R}^d} e^{\mathrm{i}\langle z,x\rangle}\, d\mu_{i_j}(x) = \lim_{j\to+\infty} \int_{\mathbb{R}^d} e^{\mathrm{i}r\langle v,x\rangle}\, d\mu_{i_j}(x).$$

Since $\|v\| = 1$, we define $\bar{E} \in \mathbb{S}_{d,k}$ whose first column is v. Let $\bar{r}$ be a k-dimensional vector whose first coordinate is r and other coordinates are zero. Then we have $r\langle v, x\rangle = \langle \bar{r}, \bar{E}^\top x\rangle$. Putting these pieces together yields that

$$\lim_{j\to+\infty} \Phi_{\mu_{i_j}}(z) = \lim_{j\to+\infty} \int_{\mathbb{R}^k} e^{\mathrm{i}\langle \bar{r},y\rangle}\, d\bar{E}^\star_\# \mu_{i_j}(y).$$

For such fixed $\bar{E}$, we claim that $\mathcal{W}_p(\bar{E}^\star_\#\mu_{i_j}, \bar{E}^\star_\#\mu) \to 0$ holds true. More specifically, $\underline{\mathcal{PW}}_{p,k}(\mu_{i_j}, \mu) \to 0$ implies that $\int \mathcal{W}_p^p(E^\star_\#\mu_{i_j}, E^\star_\#\mu) d\sigma(E) \to 0$. Since $\mathcal{W}_p^p(E^\star_\#\mu_{i_j}, E^\star_\#\mu)$ is non-negative, it is easy to derive that $\mathcal{W}_p(\bar{E}^\star_\#\mu_{i_j}, \bar{E}^\star_\#\mu) \nrightarrow 0$ for almost every E. Nonetheless, by the continuity of $\mathcal{W}_p^p(E^\star_\#\mu_{i_j}, E^\star_\#\mu)$ with respect to E, we can obtain that $\mathcal{W}_p(\bar{E}^\star_\#\mu_{i_j}, \bar{E}^\star_\#\mu) \nrightarrow 0$ for all fixed E. Indeed, by the proof by contradiction, we assume that $\mathcal{W}_p(E^\star_\#\mu_{i_j}, E^\star_\#\mu) \nrightarrow 0$ for some fixed E. Then, there exists a neighborhood S of E (it is fixed) such that $\int_S \mathcal{W}_p^p(E^\star_\#\mu_{i_j}, E^\star_\#\mu) d\sigma(E) \nrightarrow 0$. This contradicts $\int \mathcal{W}_p^p(E^\star_\#\mu_{i_j}, E^\star_\#\mu) d\sigma(E) \to 0$ since the inside term is non-negative. Thus, we achieve the desired claim.

Using Theorem B.4, we have $\bar{E}^\star_\#\mu_{i_j} \Rightarrow \bar{E}^\star_\#\mu$. Since $r\langle v, x\rangle = \langle \bar{r}, \bar{E}^\top x\rangle$, we have

$$\lim_{j\to+\infty} \int_{\mathbb{R}^k} e^{\mathrm{i}\langle \bar{r},x\rangle}\, d\bar{E}^\star_\#\mu_{i_j}(x) = \int_{\mathbb{R}^k} e^{\mathrm{i}\langle \bar{r},x\rangle}\, d\bar{E}^\star_\#\mu(x) = \int_{\mathbb{R}^d} e^{\mathrm{i}r\langle v,x\rangle}\, d\mu(x) = \int_{\mathbb{R}^d} e^{\mathrm{i}\langle z,x\rangle}\, d\mu(x).$$

Putting these pieces together yields that $\Phi_{\mu_{i_j}}(z) \to \Phi_\mu(z)$ for all $z \in \mathbb{R}^d/\{\mathbf{0}_d\}$ and $\tilde{\mu}_i = \mu$ for all $i \in \mathbb{N}$. Using the Prokhorov's theorem again yields that the whole sequence $\{\mu_i\}_{i\in\mathbb{N}}$ has the limit μ in weak sense. Therefore, $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ implies $\mu_i \Rightarrow \mu$. Since the Wasserstein distances metrize the weak convergence (cf. Theorem B.4), we conclude that $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ implies $\mathcal{W}_p(\mu_i, \mu) \to 0$. This completes the proof.

B.3 Proof of Theorem 3.2

By Lemma 3.1, we have $\underline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ if and only if $\overline{\mathcal{PW}}_{p,k}(\mu_i, \mu) \to 0$ if and only if $\mathcal{W}_p(\mu_i, \mu) \to 0$. By Theorem B.4, we have $\mu_i \Rightarrow \mu$ if and only if $\mathcal{W}_p(\mu_i, \mu) \to 0$. Putting these pieces together yields the desired result.

B.4 Proof of Theorem 3.3

Fixing $E \in \mathbb{S}_{d,k}$, the mapping $x \mapsto E^\top x$ is continuous from $\mathbb{R}^d$ to $\mathbb{R}^k$. Since $\mu_i \Rightarrow \mu$ and $\nu_i \Rightarrow \nu$, the continuous mapping theorem implies that $E^\star_\#\mu_i \Rightarrow E^\star_\#\mu$ and $E^\star_\#\nu_i \Rightarrow E^\star_\#\nu$. The next step is the key ingredient in the proof and we hope to show that

$$\mathcal{W}_p^p(E^\star_\#\mu, E^\star_\#\nu) \le \liminf_{i\to+\infty} \mathcal{W}_p^p(E^\star_\#\mu_i, E^\star_\#\nu_i) \quad \text{for all } E \in \mathbb{S}_{d,k}. \tag{7}$$

From Theorem B.2, there exists a coupling $\pi_i \in \Pi(E^\star_\#\mu_i, E^\star_\#\nu_i)$ such that $\mathcal{W}_p^p(E^\star_\#\mu_i, E^\star_\#\nu_i) = \int_{\mathbb{R}^k\times\mathbb{R}^k} \|x - y\|^p\, d\pi_i(x, y)$. By the definition of lim inf, there exists a subsequence of $\{\pi_i\}_{i\in\mathbb{N}}$ such that $\int_{\mathbb{R}^k\times\mathbb{R}^k} \|x-y\|^p\, d\pi_i(x, y)$ converges to $\liminf_{i\to+\infty} \mathcal{W}_p^p(E^\star_\#\mu_i, E^\star_\#\nu_i)$. For the simplicity, we still denote it by $\{\pi_i\}_{i\in\mathbb{N}}$. By Lemma B.3 and Prokhorov's theorem (cf. Theorem B.1), $\{\pi_i\}_{i\in\mathbb{N}}$ is sequentially compact in weak sense. Thus, there exists a subsequence $\{\pi_{i_j}\}_{j\in\mathbb{N}}$ such that $\pi_{i_j} \Rightarrow \tilde{\pi} \in \mathscr{P}(\mathbb{R}^k \times \mathbb{R}^k)$. Putting these pieces together yields that

$$\liminf_{i\to+\infty} \mathcal{W}_p^p(E^\star_\#\mu_i, E^\star_\#\nu_i) = \int_{\mathbb{R}^k\times\mathbb{R}^k} \|x - y\|^p\, d\tilde{\pi}(x, y).$$

By the definition of the Wasserstein distance, it suffices to show that $\tilde{\pi} \in \Pi(E^\star_\#\mu, E^\star_\#\nu)$. Indeed, let $f : \mathbb{R}^k \to \mathbb{R}$ be a continuous and bounded function, we have

$$\int_{\mathbb{R}^k\times\mathbb{R}^k} f(x)\, d\tilde{\pi}(x, y) = \lim_{j\to+\infty} \int_{\mathbb{R}^k\times\mathbb{R}^k} f(x)\, d\pi_{i_j}(x, y).$$

Since $\pi_{i_j} \in \Pi(E^\star_\# \mu_{i_j}, E^\star_\# \nu_{i_j})$ and $E^\star_\# \mu_i \Rightarrow E^\star_\# \mu$, we have

$$\lim_{j \to +\infty} \int_{\mathbb{R}^k \times \mathbb{R}^k} f(x)\, d\pi_{i_j}(x,y) = \lim_{j \to +\infty} \int_{\mathbb{R}^k} f(x)\, dE^\star_\# \mu_{i_j}(x) = \int_{\mathbb{R}^k} f(x)\, dE^\star_\# \mu(x).$$

Since $E^\star_\# \nu_i \Rightarrow E^\star_\# \nu$, the same argument implies that $\int_{\mathbb{R}^k \times \mathbb{R}^k} f(y)\, d\tilde{\pi}(x,y) = \int_{\mathbb{R}^k} f(y)\, dE^\star_\# \nu(y)$. Putting these pieces together yields Eq. (7).

For the IPRW distance, we derive from Eq. (7) and the Fatou's lemma that

$$\underline{\mathcal{PW}}^p_{p,k}(\mu,\nu) = \int_{\mathbb{S}_{d,k}} \mathcal{W}^p_p(E^\star_\# \mu, E^\star_\# \nu)\, d\sigma(E) \leq \liminf_{i \to +\infty} \int_{\mathbb{S}_{d,k}} \mathcal{W}^p_p(E^\star_\# \mu_i, E^\star_\# \nu_i)\, d\sigma(E) = \liminf_{i \to +\infty} \underline{\mathcal{PW}}^p_{p,k}(\mu_i, \nu_i).$$

Since $\underline{\mathcal{PW}}_{p,k}(\mu,\nu)$ and $\underline{\mathcal{PW}}_{p,k}(\mu_i,\nu_i)$ are both nonnegative, we take the p-th root of both sides of the above inequality and have $\underline{\mathcal{PW}}_{p,k}(\mu,\nu) \leq \liminf_{i \to +\infty} \underline{\mathcal{PW}}_{p,k}(\mu_i,\nu_i)$.

For the PRW distance, we derive from Eq. (7) and the fact that the supremum of a sequence of lower semi-continuous mappings is lower semi-continuous that

$$\overline{\mathcal{PW}}^p_{p,k}(\mu,\nu) = \sup_{E \in \mathbb{S}_{d,k}} \mathcal{W}^p_p(E^\star_\# \mu, E^\star_\# \nu) \leq \liminf_{i \to +\infty} \overline{\mathcal{PW}}^p_{p,k}(\mu_i,\nu_i).$$

where the first equality holds true since the Wasserstein distance is nonnegative. Since $\overline{\mathcal{PW}}_{p,k}(\mu,\nu)$ and $\overline{\mathcal{PW}}_{p,k}(\mu_i,\nu_i)$ are both nonnegative, we have $\overline{\mathcal{PW}}_{p,k}(\mu,\nu) \leq \liminf_{i \to +\infty} \overline{\mathcal{PW}}_{p,k}(\mu_i,\nu_i)$.

C Postponed Proofs in Subsection 3.2

In this section, we provide the detailed proofs for Theorem 3.4-3.8.

C.1 Preliminary technical results

To facilitate reading, we collect several preliminary technical results which will be used in the postponed proofs in subsection 3.2.

Theorem C.1 (Tonelli's theorem) *if $(\mathcal{X}, A, \mu)$ and $(\mathcal{Y}, B, \nu)$ are σ-finite measure spaces, while $f : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ is non-negative measurable function, then*

$$\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} f(x,y)\, dy \right) dx = \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} f(x,y)\, dx \right) dy = \int_{\mathcal{X} \times \mathcal{Y}} f(x,y)\, d(x,y).$$

The following proposition provides the state-of-the-art general bound for the Wasserstein distance between the true measure and its empirical version in $\mathbb{R}^d$. Note that we do not assume any additional structures of the true measure. Similar results can be found in many classical works, e.g., Fournier and Guillin (2015, Theorem 1), Weed and Bach (2019, Theorem 1) and Lei (2020, Theorem 3.1). Since $p \geq 1$, we present the following results which directly follows the proof of Lei (2020, Theorem 3.1).

Proposition C.2 *Let $\mu_\star \in \mathscr{P}_q(\mathbb{R}^d)$ and $M_q := M_q(\mu_\star) < +\infty$. Then we have*

$$\mathbb{E}[\mathcal{W}_p(\widehat{\mu}_n, \mu_\star)] \leq (\mathbb{E}[\mathcal{W}^p_p(\widehat{\mu}_n, \mu_\star)])^{1/p} \lesssim_{p,q} n^{-[\frac{1}{(2p) \vee d} \wedge (\frac{1}{p} - \frac{1}{q})]} (\log(n))^{\frac{\zeta'_{p,q,d}}{p}}, \quad \text{for all } n \geq 1. \tag{8}$$

where $\lesssim_{p,q}$ refers to "less than" with a constant depending only on (p,q) and

$$\zeta'_{p,q,d} = \begin{cases} 2 & \text{if } d = q = 2p, \\ 1 & \text{if "}d \neq 2p \text{ and } q = \frac{dp}{d-p}\text{" or "}q > d = 2p\text{"}, \\ 0 & \text{otherwise.} \end{cases}$$

The following proposition provides a bound for the covering number of $\mathbb{S}_{d,k}$ in the operator norm of a matrix, denoted by $\|\cdot\|_{\mathrm{op}}$. This is a straightforward consequence of the classical results on the covering number of the

unit sphere in $\mathbb{R}^d$ in Euclidean norm. For the proof details, we refer the interested readers to Niles-Weed and Rigollet (2019, Lemma 4). For the background materials on the covering number, we refer the interested readers to Wainwright (2019, Chapter 5). For the ease of presentation, we provide a formal definition of covering number of $\mathbb{S}_{d,k}$ in $\|\cdot\|_{\mathrm{op}}$ as follows.

For any $\epsilon \in (0,1)$, the ϵ-covering number of $\mathbb{S}_{d,k}$ in $\|\cdot\|_{\mathrm{op}}$ is defined by

$$N(\mathbb{S}_{d,k}, \epsilon, \|\cdot\|_{\mathrm{op}}) = \inf\left\{N \in \mathbb{N} : \exists x_1, x_2, \ldots, x_N \in \mathbb{S}_{d,k}, \text{ s.t. } \mathbb{S}_{d,k} \subseteq \bigcup_{i=1}^{N} \mathbb{B}(x_i, \epsilon)\right\},$$

where $\mathbb{B}(x,r) = \{y \in \mathbb{S}_{d,k} : \|y - x\|_{\mathrm{op}} \leq r\}$ is the ball of radius $r > 0$ centered at $x \in \mathbb{S}_{d,k}$ in the operator norm of a matrix.

Proposition C.3 *There exists a universal constant $c > 0$ such that for all $\epsilon \in (0,1)$, the ϵ-covering number of $\mathbb{S}_{d,k}$ in $\|\cdot\|_{\mathrm{op}}$ satisfies that $N(\mathbb{S}_{d,k}, \epsilon, \|\cdot\|_{\mathrm{op}}) \leq (c\sqrt{k}\epsilon^{-1})^{dk}$.*

The following theorem (Lei, 2020) summarizes the concentration results assuming the Bernstein tail condition under product measure. Indeed, let $\{X_i\}_{i\in[n]}$ be independent samples from probability measure μ_i on spaces $\mathcal{X}_i$ and X_i' be independent copies of X_i for all $i \in [n]$. Denote $X = (X_1, \ldots, X_n)$ and $X'_{(i)} = (X_1, \ldots, X_i', \ldots, X_n)$ which is identical to X except for X_i'. Let $f : \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$ be a function such that $\mathbb{E}[|f(X)|] < +\infty$, and define $D_i = f(X) - f(X'_{(i)})$.

Theorem C.4 *Suppose that there exists some $\sigma_i, M > 0$ so that $\mathbb{E}[|D_i|^k \mid X_{-i}] \leq (1/2)\sigma_i^2 k! M^{k-2}$ for all $k \geq 2$. Then the following statement holds,*

$$\mathbb{P}(f(X) - \mathbb{E}(f(X)) > t) \leq \exp\left(-\frac{t^2}{2(\sum_{i=1}^n \sigma_i^2) + 2tM}\right).$$

The following theorem summarizes the concentration results assuming the Poincaré inequality under product measure. We denote by $\|\nabla_i f\|$ the length of the gradient with respect to the i^{th} coordinate.

Theorem C.5 (Corollary 4.6 in Ledoux (1999)) *Denote by μ^n the product of μ on $\otimes_{i=1}^n \mathbb{R}^d$ and $\mu \in \mathscr{P}(\mathbb{R}^d)$ satisfies the Poincaré inequality (cf. Definition 3.4). For every function f on $\otimes_{i=1}^n \mathbb{R}^d$ satisfying $\mathbb{E}(|f(X)|) < +\infty$, and $\sum_{i=1}^n \|\nabla_i f(X)\|^2 \leq \alpha^2$ and $\max_{1\leq i\leq n} \|\nabla_i f(X)\| \leq \beta$ almost surely. Then the following statement holds true for $X \sim \mu^n$ that,*

$$\mathbb{P}(f(X) - \mathbb{E}(f(X)) > t) \leq \exp\left(-\frac{1}{K}\min\left\{\frac{t}{\beta}, \frac{t^2}{\alpha^2}\right\}\right),$$

where $K > 0$ only depends on the constant M in the Poincaré inequality.

C.2 Proof of Theorem 3.4

Note that $\mu_\star \in \mathscr{P}_q(\mathbb{R}^d)$ and $M_q := M_q(\mu_\star) < +\infty$. Fixing $E \in \mathbb{S}_{d,k}$, we have $E^\star_\# \mu_\star \in \mathscr{P}_q(\mathbb{R}^k)$ and $M_q(E^\star_\# \mu_\star) \leq M_q < +\infty$. Then Proposition C.2 implies that

$$\left(\mathbb{E}[\mathcal{W}_p^p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star)]\right)^{1/p} \lesssim_{p,q} n^{-[\frac{1}{(2p)\vee k} \wedge (\frac{1}{p} - \frac{1}{q})]} (\log(n))^{\frac{\zeta'_{p,q,k}}{p}} \quad \text{for all } n \geq 1.$$

Since $\mathcal{W}_p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star) \geq 0$ for any $E \in \mathbb{S}_{d,k}$ and $\mu_\star \in \mathscr{P}_q(\mathbb{R}^d)$, Theorem C.1 implies that

$$\mathbb{E}[\underline{\mathcal{PW}}_{p,k}^p(\widehat{\mu}_n, \mu_\star)] = \mathbb{E}\left[\int_{\mathbb{S}_{d,k}} \mathcal{W}_p^p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star)\, d\sigma(E)\right] = \int_{\mathbb{S}_{d,k}} \mathbb{E}[\mathcal{W}_p^p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star)]\, d\sigma(E).$$

Note that $\zeta_{p,q,k} = \zeta'_{p,q,k}$ where $\zeta_{p,q,k}$ is defined in Theorem 3.4. Moreover, $p \geq 1$. By the Jensen's inequality, we have

$$\mathbb{E}[\underline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\star)] \leq (\mathbb{E}[\underline{\mathcal{PW}}_{p,k}^p(\widehat{\mu}_n, \mu_\star)])^{1/p}. \tag{9}$$

Putting these pieces together yields the desired result.

C.3 Proof of Theorem 3.5

By the definition of $\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\star)$, we have

$$\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\star)] \leq \sup_{E\in\mathbb{S}_{d,k}} \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)] + \mathbb{E}\left[\sup_{E\in\mathbb{S}_{d,k}} \left(\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)]\right)\right]. \quad (10)$$

Using the same arguments for proving Theorem 3.4, we have

$$\sup_{E\in\mathbb{S}_{d,k}} \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)] \lesssim_{p,q} n^{-[\frac{1}{(2p)\vee k}\wedge(\frac{1}{p}-\frac{1}{q})]}(\log(n))^{\frac{\zeta_{p,q,k}}{p}} \quad \text{for all } n \geq 1. \quad (11)$$

The remaining step is to bound the gap $\mathbb{E}[\sup_{E\in\mathbb{S}_{d,k}}(\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)])]$. We first claim that $\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)]$ is sub-exponential with parameters $(2\sigma n^{1/2-1/p}, 2Vn^{-1/p})$ for all $E \in \mathbb{S}_{d,k}$ if the true measure $\mu_\star$ satisfies the projection Bernstein-type tail condition (cf. Definition 3.1). Indeed, let $f(X) = \mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)$, we have

$$D_i = f(X) - f(X'_{(i)}) \leq \mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\widehat{\mu}'_n) \leq n^{-1/p}\left(\|E^\star_\#(X_i) - E^\star_\#(X'_i)\|\right).$$

By the triangle inequality and using the projection Bernstein-type tail condition, we have

$$\mathbb{E}[|D_i|^k \mid X_{-i}] \leq 2^k n^{-k/p}(\mathbb{E}_{X\sim E^\star_\#\mu}[|X|^k]) \leq 2^{k-1}n^{-k/p}\sigma^2 k! V^{k-2} = \frac{(2n^{-1/p}\sigma)^2 k!(2n^{-1/p}V)^{k-2}}{2}.$$

This implies that the condition in Theorem C.4 holds true with $\sigma_i = 2n^{-1/p}\sigma$ and $M = 2n^{-1/p}V$. Equipped with Theorem C.4 yields that

$$\mathbb{P}\left(\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)] \geq t\right) \leq \exp\left(-\frac{t^2}{8\sigma^2 n^{1-2/p} + 4tVn^{-1/p}}\right).$$

For the simplicity, let $Z_E = \mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)]$. Then we have $\mathbb{E}[Z_E] = 0$ and $\mathbb{P}(Z_E \geq t) \leq \exp(-t^2/(8\sigma^2 n^{1-2/p} + 4tVn^{-1/p}))$. This together with the definition of Z_E and Wainwright (2019, Theorem 2.2) yields the desired claim.

We then interpret $\{Z_E\}_{E\in\mathbb{S}_{d,k}}$ as an empirical process indexed by $E \in \mathbb{S}_{d,k}$ and claim that there exists a random variable L satisfying $\mathbb{E}[L] \leq 4M_q(\mu_\star)$ so that $|Z_U - Z_V| \leq L\|U - V\|_{\text{op}}$ for all $U, V \in \mathbb{S}_{d,k}$. More specifically, it follows from the definition that

$$Z_U - Z_V \;=\; \left(\mathcal{W}_p(U^\star_\#\widehat{\mu}_n, U^\star_\#\mu) - \mathcal{W}_p(V^\star_\#\widehat{\mu}_n, V^\star_\#\mu)\right) - \mathbb{E}\left[\mathcal{W}_p(U^\star_\#\widehat{\mu}_n, U^\star_\#\mu) - \mathcal{W}_p(V^\star_\#\widehat{\mu}_n, V^\star_\#\mu)\right].$$

Since the Wasserstein distance is nonnegative and satisfies the triangle inequality, we have

$$\begin{aligned}\mathcal{W}_p(U^\star_\#\widehat{\mu}_n, U^\star_\#\mu) - \mathcal{W}_p(V^\star_\#\widehat{\mu}_n, V^\star_\#\mu) \;&=\; \mathcal{W}_p(U^\star_\#\widehat{\mu}_n, U^\star_\#\mu) - \mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\mu) + \mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\mu) - \mathcal{W}_p(V^\star_\#\widehat{\mu}_n, V^\star_\#\mu)\\ &\leq\; \mathcal{W}_p(U^\star_\#\mu, V^\star_\#\mu) + \mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\widehat{\mu}_n)\end{aligned}$$

Putting these pieces together yields that

$$Z_U - Z_V \;\leq\; \mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\widehat{\mu}_n) + \mathcal{W}_p(U^\star_\#\mu_\star, V^\star_\#\mu_\star) + \mathbb{E}\left[\mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\widehat{\mu}_n) + \mathcal{W}_p(U^\star_\#\mu_\star, V^\star_\#\mu_\star)\right].$$

Since the Wasserstein distance is symmetrical, we have

$$Z_V - Z_U \;\leq\; \mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\widehat{\mu}_n) + \mathcal{W}_p(U^\star_\#\mu_\star, V^\star_\#\mu_\star) + \mathbb{E}\left[\mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\widehat{\mu}_n) + \mathcal{W}_p(U^\star_\#\mu_\star, V^\star_\#\mu_\star)\right].$$

Therefore, we conclude that

$$|Z_U - Z_V| \;\leq\; \mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\widehat{\mu}_n) + \mathcal{W}_p(U^\star_\#\mu_\star, V^\star_\#\mu_\star) + \mathbb{E}\left[\mathcal{W}_p(U^\star_\#\widehat{\mu}_n, V^\star_\#\widehat{\mu}_n) + \mathcal{W}_p(U^\star_\#\mu_\star, V^\star_\#\mu_\star)\right].$$

Let $X \sim \mu$, we have

$$\begin{aligned}|Z_U - Z_V| &\leq 2\left(\mathbb{E}(\|(U-V)X\|^p)\right)^{1/p} + \left(\frac{1}{n}\sum_{i=1}^n \|(U-V)X_i\|^p\right)^{1/p} + \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \|(U-V)X_i\|^p\right)^{1/p}\right] \\ &\leq \|U-V\|_{\rm op}\left(2(\mathbb{E}(\|X\|^p))^{1/p} + \left(\frac{1}{n}\sum_{i=1}^n \|X_i\|^p\right)^{1/p} + \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \|X_i\|^p\right)^{1/p}\right]\right) \\ &:= L\|U-V\|_{\rm op}\,.\end{aligned}$$

Note that $X_{1:n} = (X_1, \ldots, X_n)$ are independent and identically distributed samples according to $\mu_\star$. By the Jensen's inequality and using the fact that $q > p \geq 1$, we have

$$\mathbb{E}[L] \leq 4(\mathbb{E}(\|X\|^p))^{1/p} \leq 4(\mathbb{E}(\|X\|^q))^{1/q} = 4M_q(\mu_\star).$$

Thus, by a standard ϵ-net argument, we obtain that

$$\mathbb{E}[\sup_{E\in\mathbb{S}_{d,k}} Z_E] \leq \inf_{\epsilon>0}\left\{\epsilon\mathbb{E}[L] + 4\sigma n^{1/2-1/p}\sqrt{\log(N(\mathbb{S}_{d,k},\epsilon,\|\cdot\|_{\rm op}))} + 2Vn^{-1/p}\log(N(\mathbb{S}_{d,k},\epsilon,\|\cdot\|_{\rm op}))\right\}$$

Proposition C.3 shows that there exists a universal constant $c > 0$ such that

$$\log(N(\mathbb{S}_{d,k},\epsilon,\|\cdot\|_{\rm op})) \leq dk\log\left(\frac{c\sqrt{k}}{\epsilon}\right).$$

Putting these pieces together and choosing $\epsilon = \sqrt{k}n^{-1/p}$ (it is chosen to achieve the tight bound) yields that

$$\begin{aligned}\mathbb{E}\left[\sup_{E\in\mathbb{S}_{d,k}} Z_E\right] &\lesssim_{p,q} \inf_{\epsilon>0}\left\{\epsilon + n^{1/2-1/p}\sqrt{dk\log\left(\frac{\sqrt{k}}{\epsilon}\right)} + n^{-1/p}dk\log\left(\frac{\sqrt{k}}{\epsilon}\right)\right\} \\ &\lesssim_{p,q} n^{1/2-1/p}\sqrt{dk\log(n)} + n^{-1/p}dk\log(n).\end{aligned}$$

Therefore, we conclude that

$$\mathbb{E}\left[\sup_{E\in\mathbb{S}_{d,k}}\left(\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)]\right)\right] \lesssim_{p,q} n^{1/2-1/p}\sqrt{dk\log(n)} + n^{-1/p}dk\log(n).$$

This together with Eq. (10) and Eq. (11) yields the desired inequality.

C.4 Proof of Theorem 3.6

Using the same arguments in Theorem 3.5, we obtain Eq. (10) and Eq. (11). So it suffices to bound the gap $\mathbb{E}[\sup_{E\in\mathbb{S}_{d,k}}(\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)])]$ under different condition.

We first claim that $\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)]$ is sub-exponential with parameters $(\sqrt{K/2}n^{-1/(2\vee p)}, (K/2)n^{-1/p})$ for all $E \in \mathbb{S}_{d,k}$ if the true measure $\mu_\star$ satisfies the projection Poincaré inequality (cf. Definition 3.2). Indeed, we consider $X = (X_1, \ldots, X_n)$ and $X' = (X'_1, \ldots, X'_n)$ where X_i, X'_i are independent samples from $E^\star_\#\mu_\star$. Let $f(X) = \mathcal{W}_p(E^\star_\#\widehat{\mu}_n, E^\star_\#\mu_\star)$, we have $\mathbb{E}(|f(X)|) < +\infty$. By the triangle inequality, we have

$$|f(X) - f(X')| \leq n^{-1/p}\left(\sum_{i=1}^n \|X_i - X'_i\|^p\right)^{1/p} \leq n^{-\frac{1}{2\vee p}}\|X - X'\|.$$

This implies that the following statement holds almost surely,

$$\sum_{i=1}^n \|\nabla_i f(X)\|^2 \leq n^{-\frac{2}{2\vee p}} \quad \text{and} \quad \max_{1\leq i\leq n}\|\nabla_i f(X)\| \leq n^{-\frac{1}{p}}, \quad \text{almost surely.}$$

In addition, the probability measure $E^\star_\# \mu_\star \in \mathscr{P}(\mathbb{R}^k)$ is assumed to satisfy the Poincaré inequality. Equipped with Theorem C.5 yields that

$$\mathbb{P}\left(\mathcal{W}_p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star)] \geq t\right) \leq \exp\left(-\frac{1}{K}\min\left\{\frac{t}{n^{-1/p}}, \frac{t^2}{n^{-2/(2\vee p)}}\right\}\right),$$

For the simplicity, let $Z_E = \mathcal{W}_p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star)]$. Then we have $\mathbb{E}[Z_E] = 0$ and $\mathbb{P}(Z_E \geq t) \leq \exp(-K^{-1}\min\{n^{1/p}t, n^{2/(2\vee p)}t^2\})$. This together with the definition of Z_E and Wainwright (2019, Theorem 2.2) yields the desired claim.

Using the same argument in Theorem 3.5, we can interpret $\{Z_E\}_{E\in\mathbb{S}_{d,k}}$ as an empirical process indexed by $E \in \mathbb{S}_{d,k}$ and show that there exists a random variable L satisfying $\mathbb{E}[L] \leq 4M_q(\mu_\star)$ so that $|Z_U - Z_V| \leq L\|U - V\|_{\mathrm{op}}$ for all $U, V \in \mathbb{S}_{d,k}$. By a standard ϵ-net argument, we obtain that

$$\mathbb{E}[\sup_{E\in\mathbb{S}_{d,k}} Z_E] \leq \inf_{\epsilon>0}\left\{\epsilon\mathbb{E}[L] + \sqrt{2K}n^{-1/(2\vee p)}\sqrt{\log(N(\mathbb{S}_{d,k}, \epsilon, \|\cdot\|_{\mathrm{op}}))} + (K/2)n^{-1/p}\log(N(\mathbb{S}_{d,k}, \epsilon, \|\cdot\|_{\mathrm{op}}))\right\}.$$

Combining Proposition C.3 and choosing $\epsilon = \sqrt{k}n^{-1/p}$ (it is chosen to achieve the tight bound) yields that

$$\begin{aligned}\mathbb{E}\left[\sup_{E\in\mathbb{S}_{d,k}} Z_E\right] &\lesssim_{p,q} \inf_{\epsilon>0}\left\{\epsilon + n^{-1/(2\vee p)}\sqrt{dk\log\left(\frac{\sqrt{k}}{\epsilon}\right)} + n^{-1/p}dk\log\left(\frac{\sqrt{k}}{\epsilon}\right)\right\} \\ &\lesssim_{p,q} n^{-1/(2\vee p)}\sqrt{dk\log(n)} + n^{-1/p}dk\log(n).\end{aligned}$$

Therefore, we conclude that

$$\mathbb{E}\left[\sup_{E\in\mathbb{S}_{d,k}}\left(\mathcal{W}_p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star) - \mathbb{E}[\mathcal{W}_p(E^\star_\# \widehat{\mu}_n, E^\star_\# \mu_\star)]\right)\right] \lesssim_{p,q} n^{-1/(2\vee p)}\sqrt{dk\log(n)} + n^{-1/p}dk\log(n).$$

This together with Eq. (10) and Eq. (11) yields the desired inequality.

C.5 Proof of Theorem 3.7

Since the arguments in this proof hold true for both IPRW and PRW distances, we denote $W = \underline{\mathcal{PW}}_{p,k}$ or $W = \overline{\mathcal{PW}}_{p,k}$ for short. Let $f(X) = W(\widehat{\mu}_n, \mu_\star)$, we have

$$D_i = f(X) - f(X'_{(i)}) \leq W(\widehat{\mu}_n, \widehat{\mu}'_n) \leq n^{-1/p}\left(\sup_{E\in\mathbb{S}_{d,k}}\|E^\star_\#(X_i) - E^\star_\#(X'_i)\|\right).$$

By the triangle inequality, we have

$$\mathbb{E}\left[|D_i|^k \mid X_{-i}\right] \leq 2^k n^{-k/p}\left(\mathbb{E}\left[\sup_{E\in\mathbb{S}_{d,k}, X\sim E^\star_\#\mu}|X|^k\right]\right).$$

Since the true measure $\mu_\star$ satisfies the Bernstein-type tail condition (cf. Definition 3.3), we have

$$\mathbb{E}\left[|D_i|^k \mid X_{-i}\right] \leq 2^{k-1}n^{-k/p}\sigma^2 k! V^{k-2} = \frac{(2n^{-1/p}\sigma)^2 k!(2n^{-1/p}V)^{k-2}}{2}$$

This implies that the condition in Theorem C.4 holds true with $\sigma_i = 2n^{-1/p}\sigma$ and $M = 2n^{-1/p}V$. Equipped with Theorem C.4 yields the desired inequality.

C.6 Proof of Theorem 3.8

Since the arguments in this proof hold true for both IPRW and PRW distances, we denote $W = \underline{\mathcal{PW}}_{p,k}$ or $W = \overline{\mathcal{PW}}_{p,k}$ for short. We consider $X = (X_1, X_2, \ldots, X_n)$ and $X' = (X'_1, X'_2, \ldots, X'_n)$ where X_i, X'_i are

independent samples from $\mu_\star$. Let $f(X) = W(\widehat{\mu}_n, \mu_\star)$, we have $\mathbb{E}(|f(X)|) < +\infty$. By the triangle inequality, we have

$$|f(X) - f(X')| \leq n^{-1/p} \left(\sum_{i=1}^{n} \|X_i - X_i'\|^p \right)^{1/p} \leq n^{-\frac{1}{2\vee p}} \|X - X'\|.$$

This implies that the following statement holds almost surely,

$$\sum_{i=1}^{n} \|\nabla_i f(X)\|^2 \leq n^{-\frac{2}{2\vee p}} \quad \text{and} \quad \max_{1\leq i\leq n} \|\nabla_i f(X)\| \leq n^{-\frac{1}{p}}.$$

In addition, the true measure $\mu_\star$ satisfies the Poincaré inequality (cf. Definition 3.4). Equipped with Theorem C.5 yields the desired inequality.

D Postponed Proofs in Subsection 3.3

In this section, we provide the detailed proofs for Theorem 3.9-3.11 and Theorem A.1-A.2. Our results are derived analogously to the proof in Bernton et al. (2019) for the estimators based on Wasserstein distance and the proof in Nadjahi et al. (2019) for the estimators based on sliced-Wasserstein distance.

D.1 Preliminary technical results

To facilitate the reading, we collect several preliminary technical results which will be used in the postponed proofs in subsection 3.3.

Theorem D.1 (Theorem 2.43 in Aliprantis and Border (2006)) *A real-valued lower semi-continuous function on a compact space attains a minimum value, and the nonempty set of minimizers is compact. Similarly, an upper semicontinuous function on a compact set attains a maximum value, and the nonempty set of maximizers is compact.*

Definition D.1 (epiconvergence) *Let $\mathcal{X}$ be a metric space and $\{f_i\}_{i\in\mathbb{N}}$ be a sequence of real-valued function from $\mathcal{X}$ to $\mathbb{R}$. We say that the sequence $\{f_i\}_{i\in\mathbb{N}}$ epiconverges to a function $f : \mathcal{X} \to$ br if for each $x \in \mathcal{X}$, the following statement holds true,*

$$\liminf_{i\to+\infty} f_i(x_i) \geq f(x) \text{ for every sequence } \{x_i\}_{i\in\mathbb{N}} \text{ such that } x_i \to x,$$
$$\limsup_{i\to+\infty} f_i(x_i) \leq f(x) \text{ for some sequence } \{x_i\}_{i\in\mathbb{N}} \text{ such that } x_i \to x.$$

Proposition D.2 (Proposition 7.29 in Rockafellar and Wets (2009)) *Let $\mathcal{X}$ be a metric space and $\{f_i\}_{i\in\mathbb{N}}$ be a sequence of real-valued function from $\mathcal{X}$ to $\mathbb{R}$ with a lower semi-continuous function $f : \mathcal{X} \to \mathbb{R}$. Then the sequence $\{f_i\}_{i\in\mathbb{N}}$ epiconverges to f if and only if*

$$\liminf_{i\to+\infty}(\inf_{x\in K} f_i(x)) \geq \inf_{x\in K} f(x) \text{ for every compact set } K \subseteq \mathcal{X},$$
$$\limsup_{i\to+\infty}(\sup_{x\in O} f_i(x)) \leq \sup_{x\in O} f(x) \text{ for every open set } O \subseteq \mathcal{X}.$$

Recall that $\delta\text{-argmin}_{x\in\mathcal{X}} f = \{x \in \mathcal{X} : f(x) \leq \inf_{x\in\mathcal{X}} f + \delta\}$ for a generic function $f : \mathcal{X} \to \mathbb{R}$. The following theorem gives asymptotic properties for the infimum and δ-argmin of epiconvergent functions and thus a standard approach to prove the existence and consistency of the estimators.

Theorem D.3 (Theorem 7.31 in Rockafellar and Wets (2009)) *Let $\mathcal{X}$ be a metric space and $\{f_i\}_{i\in\mathbb{N}}$ be a sequence of function which epiconverges to a lower semi-continuous function f with $\inf_{x\in\mathcal{X}} f \in (-\infty, +\infty)$. Then we have the following statements,*

1. *$\inf_{x\in\mathcal{X}} f_i \to \inf_{x\in\mathcal{X}} f$ if and only if for every $\delta > 0$ there exists a compact set $B \subseteq \mathcal{X}$ and $N \in \mathbb{N}$ such that $\inf_{x\in B} f_i \leq \inf_{x\in\mathcal{X}} f_i + \delta$ for all $i \geq N$.*

2. $\limsup_{i\to+\infty}(\delta\text{-}\operatorname{argmin}_{x\in\mathcal{X}} f_i) \subseteq \delta\text{-}\operatorname{argmin}_{x\in\mathcal{X}} f$ *for any* $\delta \geq 0$ *and* $\limsup_{i\to+\infty}(\delta_i\text{-}\operatorname{argmin}_{x\in\mathcal{X}} f_i) \subseteq \operatorname{argmin}_{x\in\mathcal{X}} f$ *whenever* $\delta_i \downarrow 0$.

3. *Assume that* $\inf_{x\in\mathcal{X}} f_i \to \inf_{x\in\mathcal{X}} f$, *there exists a sequence* $\delta_i \downarrow 0$ *such that* $\delta_i\text{-}\operatorname{argmin}_{x\in\mathcal{X}} f_i \to \operatorname{argmin}_{x\in\mathcal{X}} f$. *Conversely, if* $\operatorname{argmin}_{x\in\mathcal{X}} f \neq \emptyset$ *and if such a sequence exists, then* $\inf_{x\in\mathcal{X}} f_i \to \inf_{x\in\mathcal{X}} f$.

The following theorem summarizes the well-known Skorokhod's representation theorem.

Theorem D.4 (Skorokhod's representation theorem) *Let* $\{\mu_n\}_{n\in\mathbb{N}}$ *be a sequence of probability measures on a metric space* $\mathcal{S}$ *such that* μ_n *converges weakly to some probability measure* μ_∞ *on* $\mathcal{S}$ *as* $n \to \infty$. *Suppose also that the support of* μ_∞ *is separable. Then there exist random variables* X_n *defined on a common probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that the law of* X_n *is* μ_n *for all* n *(including* $n = \infty$*) and such that* X_n *converges to* X_∞ *almost surely.*

The following theorem presents the classical results which lead to a standard approach for proving the measurability of the estimators. Note that the projection $\operatorname{proj}(D) = \{x \in \mathcal{X} : \exists y \in \mathcal{Y}, \text{s.t.}(x, y) \in D\}$ for each $D \subseteq \mathcal{X} \times \mathcal{Y}$ and the section $D_x = \{y \in \mathcal{Y} : (x, y) \in D\}$ for each $x \in \operatorname{proj}(D)$.

Theorem D.5 (Corollary 1 in Brown and Purves (1973)) *Let* $\mathcal{X}, \mathcal{Y}$ *be complete separable metric spaces and* f *be a real-valued Borel measurable function defined on a Borel subset* D *of* $\mathcal{X} \times \mathcal{Y}$. *Suppose that for each* $x \in \operatorname{proj}(D)$, *the section* D_x *is* σ*-compact and* $f(x, \cdot)$ *is lower semi-continuous with respect to the relative topology on* D_x. *Then*

1. *The sets* $G = \operatorname{proj}(D)$ *and* $I = \{x \in G : \exists y \in D_x \text{ s.t. } y = \operatorname{argmin}_{z\in\mathcal{Y}} f(x, z)\}$ *are Borel.*

2. *For each* $\epsilon > 0$, *there exists a Borel measure function* φ_ϵ *satisfying, for* $x \in G$ *that,*

$$f(x, \varphi_\epsilon(x)) \begin{cases} = \inf_{y\in G} f(x, y), & x \in I, \\ \leq \epsilon + \inf_{y\in G} f(x, y), & \text{if } x \notin I \text{ and } \inf_{y\in G} f(x, y) \neq -\infty, \\ \leq -\epsilon^{-1}, & x \notin I \text{ and } \inf_{y\in G} f(x, y) = -\infty. \end{cases}$$

To show that the MEPRW estimator is measurable, we establish the lower semi-continuity of the expectation of empirical PRW distance in the following lemma.

Lemma D.6 *The expected empirical PRW distance is lower semi-continuous in the usual weak topology. If the sequences* $\{\mu_i\}_{i\in\mathbb{N}}, \{\nu_i\}_{i\in\mathbb{N}} \subseteq \mathscr{P}(\mathbb{R}^d)$ *satisfying that* $\mu_i \Rightarrow \mu \in \mathscr{P}(\mathbb{R}^d)$ *and* $\nu_i \Rightarrow \nu \in \mathscr{P}(\mathbb{R}^d)$, *we have* $\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu, \widehat{\nu}_m)] \leq \liminf_{i\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_i, \widehat{\nu}_{i,m})]$, *where* $\widehat{\nu}_m = (1/m)\sum_{j=1}^m \delta_{Z_j}$ *for i.i.d. samples* $Z_{1:m}$ *according to* ν *and* $\{\widehat{\nu}_{i,m}\}_{i\in\mathbb{N}}$ *are defined similarly.*

D.2 Proof of Theorem 3.9

We first prove that $\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) \neq \emptyset$. Indeed, by Assumption 3.2 and Theorem 3.3, the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$ is lower semi-continuous. By Assumption 3.3, the set $\Theta_\star(\tau)$ is bounded for some $\tau > 0$. By the definition of inf, there exists $\theta' \in \Theta$ such that $\overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'}) = \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) + \tau/2$. This implies that $\theta' \in \Theta_\star(\tau)$ and $\Theta_\star(\tau)$ is nonempty. By the lower semi-continuity of the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$, the set $\Theta_\star(\tau)$ is closed. Putting these pieces together yields that $\Theta_\star(\tau)$ is compact. Therefore, we conclude the desired result from Theorem D.1.

Then we show that there exists a set $E \subseteq \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, the sequence of mappings $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta)$ epiconverges to the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$ as $n \to +\infty$. Indeed, we only need to prove that the conditions in Proposition D.2 hold true.

Fix $K \subseteq \Theta$ as a compact set. By the lower semi-continuity of the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta)$ (cf. Assumption 3.2 and Theorem 3.3), Theorem D.1 implies that

$$\inf_{\theta\in K} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) = \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_{\theta_n})$$

for some sequence $\theta_n = \theta_n(\omega) \in K$. Thus, we have

$$\liminf_{n\to+\infty} \inf_{\theta\in K} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) = \liminf_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_{\theta_n}).$$

By the definition of lim inf, there exists a subsequence of $\{\theta_n\}_{n\in\mathbb{N}}$ such that $\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_{\theta_n})$ converges to $\liminf_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_{\theta_n})$ along this subsequence. By the compactness of K, this subsequence must have a convergent subsubsequence. We denote this subsubsequence as $\{\theta_{n_j}\}_{j\in\mathbb{N}}$ and its limit as $\bar{\theta} \in K$. Then

$$\liminf_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_{\theta_n}) = \lim_{j\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_{n_j}(\omega), \mu_{\theta_{n_j}}).$$

Since $\omega \in E$ where $\mathbb{P}(E) = 1$, Assumption 3.1 and 3.2 imply $\widehat{\mu}_{n_j}(\omega) \Rightarrow \mu_\star$ and $\mu_{\theta_{n_j}} \Rightarrow \mu_{\bar{\theta}}$. These pieces together with the lower semi-continuity of the PRW distance (cf. Theorem 3.3) yields that $\lim_{j\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_{n_j}(\omega), \mu_{\theta_{n_j}}) \geq \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\bar{\theta}})$. Putting these pieces together yields that

$$\liminf_{n\to+\infty} \inf_{\theta\in K} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \geq \inf_{\theta\in K} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta).$$

Fix $O \subseteq \Theta$ as an arbitary open set. By the definition of inf, there exists a sequence $\theta'_n = \theta'_n(\omega) \in O$ such that $\overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'_n}) \to \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$. In addition, $\inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \leq \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_{\theta'_n})$. Thus, we have

$$\begin{aligned}\limsup_{n\to+\infty} \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) &\leq \limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_{\theta'_n}) \\ &\leq \limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\star) + \limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'_n}).\end{aligned}$$

Since $\omega \in E$ where $\mathbb{P}(E) = 1$, Assumption 3.1 implies $\limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\star) = 0$. By the definition of θ'_n, $\limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'_n}) = \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$. Putting these pieces together yields that $\limsup_{n\to+\infty} \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$.

Proposition D.2 guarantees that there exists a set $E \subseteq \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, the sequence of mappings $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta)$ epiconverges to the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$ as $n \to +\infty$. Then the second statement of Theorem D.3 implies that

$$\limsup_{n\to+\infty} \operatorname*{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \subseteq \operatorname*{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta). \tag{12}$$

The next step is to show that, for every $\delta > 0$, there exists a compact set $B \subseteq \Theta$ and $N \in \mathbb{N}$ such that $\inf_{\theta\in B} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) + \delta$. In what follows, we prove a stronger statement which states that the above inequality holds true with $\delta = 0$. Indeed, by the same reasoning for the open set case in the proof of epiconvergence, we have

$$\limsup_{n\to+\infty} \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta).$$

By Assumption 3.3 and using previous argument, $\Theta_\star(\tau)$ is nonempty and compact for some $\tau > 0$. The above inequality implies that there exists $n_1(\omega) > 0$ such that, for all $n \geq n_1(\omega)$, the set $\{\theta \in \Theta : \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta'\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'}) + \tau/2\}$ is nonempty. For any θ in this set and let $n \geq n_1(\omega)$, we have

$$\overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) \leq \overline{\mathcal{PW}}_{p,k}(\mu_\star, \widehat{\mu}_n(\omega)) + \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) + \frac{\tau}{2}.$$

By Assumption 3.1, there exists $n_2(\omega) > 0$ such that, for all $n \geq n_2(\omega)$, we have

$$\overline{\mathcal{PW}}_{p,k}(\mu_\star, \widehat{\mu}_n(\omega)) \leq \mathcal{W}_p(\mu_\star, \widehat{\mu}_n(\omega)) \leq \frac{\tau}{2}.$$

Putting these pieces together yields that, for all $n \geq \max\{n_1(\omega), n_2(\omega)\}$, we have $\overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) \leq \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) + \tau$. This implies that, for all $n \geq \max\{n_1(\omega), n_2(\omega)\}$ that,

$$\left\{\theta \in \Theta : \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \leq \inf_{\theta'\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'}) + \frac{\tau}{2}\right\} \subseteq \Theta_\star(\tau).$$

Therefore, we have $\inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta\in\Theta_\star(\tau)} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta)$. This together with the compactness of $\Theta_\star(\tau)$ yields the desired result.

The first statement of Theorem D.3 implies that

$$\inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta), \quad \text{as } n \to +\infty. \tag{13}$$

By Assumption 3.2 and Theorem 3.3, the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta)$ is lower semi-continuous. Theorem D.1 implies $\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\theta)$ are nonempty for all $n \geq \max\{n_1(\omega), n_2(\omega)\}$. Together with Eq. (12) and (13) yields the desired results.

Finally, we remark that these results hold true for δ_n-$\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$ with $\delta_n \to 0$. For Eq. (12) and (13), the analogous results can be derived by using the second and third statements of Theorem D.3. To show that δ_n-$\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$ is nonempty, we notice it contains the nonempty set $\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$.

D.3 Proof of Theorem 3.10

Following up the same approach used for analyzing Theorem 3.9, it is straightforward to derive that $\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta) \neq \emptyset$. Then we show that there exists a set $E \subseteq \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, the sequences $\theta \mapsto \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)}) \mid X_{1:n}]$ epiconverges $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$ as $n \to +\infty$. Indeed, it suffices to verify the conditions in Proposition D.2.

Fix $K \subseteq \Theta$ as an arbitrary compact set. By Assumption 3.2 and Lemma D.6, the mapping $\theta \mapsto \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)}) \mid X_{1:n}]$ is lower semi-continuous. Then Theorem D.1 implies that

$$\inf_{\theta\in K} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)}) \mid X_{1:n}] = \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta_n,m(n)}) \mid X_{1:n}]$$

for some sequence $\theta_n = \theta_n(\omega) \in K$. Thus, we have

$$\liminf_{n\to+\infty} \inf_{\theta\in K} \mathbb{E}\left[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)}) \mid X_{1:n}\right] = \liminf_{n\to+\infty} \mathbb{E}\left[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta_n,m(n)}) \mid X_{1:n}\right].$$

Following up the same approach used in the proof of Theorem 3.9, there exists a subsequence of $\{\theta_n\}_{n\in\mathbb{N}}$, denoted by $\{\theta_{n_j}\}_{j\in\mathbb{N}}$ with the limit $\bar{\theta} \in K$, such that

$$\begin{aligned}
\liminf_{n\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta_n,m(n)}) \mid X_{1:n}] &= \lim_{j\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_{n_j}(\omega), \widehat{\mu}_{\theta_{n_j},m(n_j)}) \mid X_{1:n_j}] \\
&\geq \liminf_{j\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_{n_j}(\omega), \mu_{\theta_{n_j}})] - \limsup_{j\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_{\theta_{n_j}}, \widehat{\mu}_{\theta_{n_j},m(n_j)}) \mid X_{1:n_j}].
\end{aligned}$$

Since $\omega \in E$ where $\mathbb{P}(E) = 1$, Assumption 3.1 and 3.2 imply $\widehat{\mu}_{n_j}(\omega) \Rightarrow \mu_\star$ and $\mu_{\theta_{n_j}} \Rightarrow \mu_{\bar{\theta}}$. These pieces together with the lower semi-continuity of the PRW distance (cf. Theorem 3.3) yields that $\liminf_{j\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_{n_j}(\omega), \mu_{\theta_{n_j}}) \geq \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\bar{\theta}})$. By Assumption 3.4 and using $\theta_{n_j} \to \bar{\theta}$, we have $\limsup_{j\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_{\theta_{n_j}}, \widehat{\mu}_{\theta_{n_j},m(n_j)}) \mid X_{1:n_j}] \to 0$. Putting these pieces together yields that

$$\liminf_{n\to+\infty} \inf_{\theta\in K} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)}) \mid X_{1:n}] \geq \inf_{\theta\in K} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta).$$

Fix $O \subseteq \Theta$ as an arbitary open set. By the definition of inf, there exists a sequence $\theta'_n = \theta'_n(\omega) \in O$ such that $\overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'_n}) \to \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$. In addition, we have

$$\inf_{\theta\in O} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)}) \mid X_{1:n}] \leq \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta'_n,m(n)}) \mid X_{1:n}].$$

Thus, we have

$$\begin{aligned}
&\limsup_{n\to+\infty} \inf_{\theta\in O} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta,m(n)}) \mid X_{1:n}] \leq \limsup_{n\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \widehat{\mu}_{\theta'_n,m(n)}) \mid X_{1:n}] \\
\leq\ &\limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\star) + \limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'_n}) + \limsup_{n\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_{\theta'_n}, \widehat{\mu}_{\theta'_n,m(n)}) \mid X_{1:n}].
\end{aligned}$$

Since $\omega \in E$ where $\mathbb{P}(E) = 1$, Assumption 3.1 implies $\limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega), \mu_\star) = 0$. By the definition of θ'_n, we have $\limsup_{n\to+\infty} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_{\theta'_n}) = \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\mu_\star, \mu_\theta)$. Using Assumption 3.4 and

$\lim_{j\to+\infty}\theta_{m_j}=\bar{\theta}$, we have $\limsup_{n\to+\infty}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_{\theta'_n},\widehat{\mu}_{\theta'_n,m(n)})\mid X_{1:n}]=0$. Putting these pieces together yields that $\limsup_{n\to+\infty}\inf_{\theta\in O}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]\le\inf_{\theta\in O}\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta)$.

Proposition D.2 guarantees that there exists a set $E\subseteq\Omega$ with $\mathbb{P}(E)=1$ such that, for all $\omega\in E$, the sequence of mappings $\theta\mapsto\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]$ epiconverges to the mapping $\theta\mapsto\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta)$ as $n\to+\infty$. Then the second statement of Theorem D.3 implies that

$$\limsup_{n\to+\infty}\operatorname*{argmin}_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]\subseteq\operatorname*{argmin}_{\theta\in\Theta}\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta).\tag{14}$$

The next step is to show that, for every $\delta>0$, there exists a compact set $B\subseteq\Theta$ and $N\in\mathbb{N}$ such that $\inf_{\theta\in B}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]\le\inf_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]+\delta$. In what follows, we prove a stronger statement which states that the above inequality holds true with $\delta=0$. Indeed, by the same reasoning for the open set case in the proof of epiconvergence, we have

$$\limsup_{n\to+\infty}\inf_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]\le\inf_{\theta\in\Theta}\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta).$$

By Assumption 3.3 and using previous argument, $\Theta_\star(\tau)$ is nonempty and compact for some $\tau>0$. The above inequality implies that there exists $n_1(\omega)>0$ such that, for all $n\ge n_1(\omega)$, the set $\{\theta\in\Theta:\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]\le\inf_{\theta'\in\Theta}\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_{\theta'})+\tau/3\}$ is nonempty. For any θ in this set and let $n\ge n_1(\omega)$, we have

$$\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta)\le\overline{\mathcal{PW}}_{p,k}(\mu_\star,\widehat{\mu}_n(\omega))+\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_\theta,\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]+\inf_{\theta\in\Theta}\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta)+\frac{\tau}{3}.$$

By Assumption 3.1, there exists $n_2(\omega)>0$ such that, for all $n\ge n_2(\omega)$, we have

$$\overline{\mathcal{PW}}_{p,k}(\mu_\star,\widehat{\mu}_n(\omega))\le\mathcal{W}_p(\mu_\star,\widehat{\mu}_n(\omega))\le\frac{\tau}{3}.$$

By Assumption 3.4, there exists $n_3(\omega)>0$ such that, for all $n\ge n_3(\omega)$, we have

$$\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_{\theta,m(n)},\mu_\theta)\mid X_{1:n}]\le\mathbb{E}[\mathcal{W}_p(\widehat{\mu}_{\theta,m(n)},\mu_\theta)\mid X_{1:n}]\le\frac{\tau}{3}.$$

Putting these pieces together yields that, for all $n\ge\max\{n_1(\omega),n_2(\omega),n_3(\omega)\}$ that,

$$\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta)\le\inf_{\theta\in\Theta}\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta)+\tau.$$

This implies that, for all $n\ge\max\{n_1(\omega),n_2(\omega),n_3(\omega)\}$ that,

$$\left\{\theta\in\Theta:\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]\le\inf_{\theta'\in\Theta}\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_{\theta'})+\frac{\tau}{3}\right\}\subseteq\Theta_\star(\tau).$$

Therefore, we have $\inf_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})|X_{1:n}]=\inf_{\theta\in\Theta_\star(\tau)}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})|X_{1:n}]$. This together with the compactness of $\Theta_\star(\tau)$ yields the desired result.

The first statement of Theorem D.3 implies that

$$\inf_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]\to\inf_{\theta\in\Theta}\overline{\mathcal{PW}}_{p,k}(\mu_\star,\mu_\theta),\quad\text{as } n\to+\infty.\tag{15}$$

By Assumption 3.2 and Lemma D.6, the mapping $\theta\mapsto\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]$ is lower semicontinuous. Theorem D.1 implies $\operatorname{argmin}_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]$ are nonempty for all $n\ge\max\{n_1(\omega),n_2(\omega),n_3(\omega)\}$. Together with Eq. (14) and (15) yields the desired results.

Finally, we remark that these results hold true for δ_n-$\operatorname{argmin}_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]$ with $\delta_n\to0$. For Eq. (14) and (15), the analogous results can be derived by using the second and third statements of Theorem D.3. To show that δ_n-$\operatorname{argmin}_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]$ is nonempty, we notice it contains the nonempty set $\operatorname{argmin}_{\theta\in\Theta}\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(\omega),\widehat{\mu}_{\theta,m(n)})\mid X_{1:n}]$.

D.4 Proof of Theorem 3.11

We first prove that $\operatorname{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta) \neq \emptyset$. Indeed, by Assumption 3.2 and Theorem 3.3, the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$ is lower semi-continuous. By Assumption 3.5, the set $\Theta_n(\tau)$ is bounded for some $\tau_n > 0$. By the definition of inf, there exists $\theta'_n \in \Theta$ such that $\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\theta'_n}) = \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta) + \tau_n/2$. This implies that $\theta'_n \in \Theta_n(\tau)$ and $\Theta_n(\tau)$ is nonempty. By the lower semi-continuity of the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$, the set $\Theta_n(\tau)$ is closed. Putting these pieces together yields that $\Theta_n(\tau)$ is compact. Therefore, we conclude the desired result from Theorem D.1.

Then we show that the sequences $\theta \mapsto \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$ epiconverges to $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$ as $m \to +\infty$. Indeed, it suffices to verify the conditions in Proposition D.2.

Fix $K \subseteq \Theta$ as an arbitrary compact set. By Assumption 3.2 and Lemma D.6, the mapping $\theta \mapsto \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$ is lower semi-continuous. Then Theorem D.1 implies that

$$\inf_{\theta\in K} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] = \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta_m,m}) \mid X_{1:n}]$$

for some sequence $\theta_m \in K$. Thus, we have

$$\liminf_{m\to+\infty} \inf_{\theta\in K} \mathbb{E}\left[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}\right] = \liminf_{m\to+\infty} \mathbb{E}\left[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta_m,m}) \mid X_{1:n}\right].$$

Following up the same approach used in the proof of Theorem 3.9, there exists a subsequence of $\{\theta_m\}_{m\in\mathbb{N}}$, denoted by $\{\theta_{m_j}\}_{j\in\mathbb{N}}$ with the limit $\bar{\theta} \in K$, such that

$$\begin{aligned}
\liminf_{m\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta_m,m}) \mid X_{1:n}] &= \lim_{j\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta_{m_j},m_j}) \mid X_{1:n}] \\
&\geq \liminf_{j\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\theta_{m_j}})] - \limsup_{j\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_{\theta_{m_j}}, \widehat{\mu}_{\theta_{m_j},m_j}) \mid X_{1:n}].
\end{aligned}$$

Assumption 3.1 and 3.2 imply $\widehat{\mu}_{m_j} \Rightarrow \mu_\star$ and $\mu_{\theta_{m_j}} \Rightarrow \mu_{\bar{\theta}}$. Together with the lower semi-continuity of the PRW distance yields that $\liminf_{j\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\theta_{m_j}}) \geq \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\bar{\theta}})$. By Assumption 3.4 and using $\theta_{m_j} \to \bar{\theta}$, we have $\limsup_{j\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_{\theta_{m_j}}, \widehat{\mu}_{\theta_{m_j},m_j}) \mid X_{1:n}] = 0$. Thus, we conclude that $\liminf_{m\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta_m,m}) \mid X_{1:n}] \geq \inf_{\theta\in K} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$.

Fix $O \subseteq \Theta$ as an arbitary open set. By the definition of inf, there exists a sequence $\theta'_m \in O$ such that $\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\theta'_m}) \to \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$. In addition, we have

$$\inf_{\theta\in O} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] \leq \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta'_m,m}) \mid X_{1:n}].$$

Thus, we have

$$\begin{aligned}
\limsup_{m\to+\infty} \inf_{\theta\in O} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] &\leq \limsup_{m\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta'_m,m}) \mid X_{1:n}] \\
&\leq \limsup_{m\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\theta'_m}) + \limsup_{n\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_{\theta'_n}, \widehat{\mu}_{\theta'_m,m}) \mid X_{1:n}].
\end{aligned}$$

By the definition of θ'_m, we have $\limsup_{m\to+\infty} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\theta'_m}) = \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$. Using Assumption 3.4 and $\lim_{j\to+\infty} \theta_{m_j} = \bar{\theta}$, we have $\limsup_{m\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_{\theta'_m}, \widehat{\mu}_{\theta'_m,m}) \mid X_{1:n}] = 0$. Putting these pieces together yields that $\limsup_{m\to+\infty} \inf_{\theta\in O} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] \leq \inf_{\theta\in O} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$.

Proposition D.2 guarantees that the sequence of mappings $\theta \mapsto \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$ epiconverges to the mapping $\theta \mapsto \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta)$ as $m \to +\infty$. Then the second statement of Theorem D.3 implies that

$$\limsup_{m\to+\infty} \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] \subseteq \operatorname*{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta). \tag{16}$$

The next step is to show that, for every $\delta > 0$, there exists a compact set $B \subseteq \Theta$ and $N \in \mathbb{N}$ such that $\inf_{\theta\in B} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] \leq \inf_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] + \delta$. In what follows, we prove a stronger statement which states that the above inequality holds true with $\delta = 0$. Indeed, by the same reasoning for the open set case in the proof of epiconvergence, we have

$$\limsup_{n\to+\infty} \inf_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] \leq \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta).$$

By Assumption 3.5 and using previous argument, $\Theta_n(\tau)$ is nonempty and compact for some $\tau > 0$. The above inequality implies that there exists $m_1 > 0$ such that, for all $m \geq m_1$, the set $\{\theta \in \Theta : \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] \leq \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta) + \tau/2\}$ is nonempty. For any θ in this set and let $m \geq m_1$, we have

$$\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta) \leq \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_{\theta,m}, \mu_\theta) \mid X_{1:n}] + \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta) + \frac{\tau}{2}.$$

By Assumption 3.4, there exists $m_2 > 0$ such that, for all $m \geq m_2$, we have

$$\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_{\theta,m}, \mu_\theta) \mid X_{1:n}] \leq \mathbb{E}[\mathcal{W}_p(\widehat{\mu}_{\theta,m}, \mu_\theta) \mid X_{1:n}] \leq \frac{\tau}{2}.$$

Putting these pieces together yields that $\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta) \leq \inf_{\theta'\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\theta'}) + \tau$ for all $m \geq \max\{m_1, m_2\}$. This implies that, for all $m \geq \max\{m_1, m_2\}$ that,

$$\left\{\theta \in \Theta : \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] \leq \inf_{\theta'\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_{\theta'}) + \frac{\tau}{2}\right\} \subseteq \Theta_n(\tau).$$

Therefore, we have $\inf_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m})|X_{1:n}] = \inf_{\theta\in\Theta_n(\tau)} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m})|X_{1:n}]$. This together with the compactness of $\Theta_n(\tau)$ yields the desired result.

The first statement of Theorem D.3 implies that

$$\inf_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}] \to \inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \mu_\theta), \quad \text{as } m \to +\infty. \tag{17}$$

By Assumption 3.2 and Lemma D.6, the mapping $\theta \mapsto \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$ is lower semi-continuous. Theorem D.1 implies $\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$ are nonempty for all $m \geq \max\{m_1, m_2\}$. Together with Eq. (16) and Eq. (17) yields the desired results.

Finally, we remark that these results hold true for δ_n-$\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$ with $\delta_n \to 0$. For Eq. (16) and (17), the analogous results can be derived by using the second and third statements of Theorem D.3. To show that δ_n-$\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$ is nonempty, we notice it contains the nonempty set $\operatorname{argmin}_{\theta\in\Theta} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$.

D.5 Proof of Lemma D.6

Since $\nu_i \Rightarrow \nu \in \mathscr{P}(\mathbb{R}^d)$ and $\mathbb{R}^d$ is separable, the Skorokhod's representation theorem (cf. Theorem D.4) implies that there exists m sequences of random variables $\{\{Z_i^k\}_{i\in\mathbb{N}}, k \in [m]\}$ and m random variables $\{Z^k, k \in [m]\}$ such that the distribution of Z_i^k is ν_i, the distribution of Z^k is ν and $\{Z_i^k\}_{i\in\mathbb{N}}$ converges to Z^k almost surely for all $k \in [m]$.

Suppose that $\widehat{\nu}_{i,m} = (1/m)(\sum_{k=1}^m \delta_{Z_i^k})$ and $\widehat{\nu}_m = (1/m)(\sum_{k=1}^m Z^k)$, we proceed to the key part of the proof and show that $\{\widehat{\nu}_{i,m}\}_{i\in\mathbb{N}}$ weakly converges to $\widehat{\nu}_m$. Indeed, it suffices to consider the deterministic case where $\widehat{\nu}_{i,m} = (1/m)(\sum_{k=1}^m \delta_{z_i^k})$ and $\widehat{\nu}_m = (1/m)(\sum_{k=1}^m z^k)$ where $\{\{z_i^k\}_{i\in\mathbb{N}}, k \in [m]\}$ and $\{z^k, k \in [m]\}$ are all deterministic such that $\lim_{i\to+\infty} \left(\max_{k\in[m]} \|z_i^k - z^k\|\right) = 0$. Since the Wasserstein distance metrizes the weak convergence (cf. Theorem B.4), we only need to show that $\lim_{i\to+\infty} \mathcal{W}_2(\widehat{\nu}_{i,m}, \widehat{\nu}_m) = 0$. By the definition of the Wasserstein distance, $\{\widehat{\nu}_{i,m}\}_{i\in\mathbb{N}}$ and $\widehat{\nu}_m$, we have $\mathcal{W}_2^2(\widehat{\nu}_{i,m}, \widehat{\nu}_m) \leq \max_{k\in[m]} \|z_i^k - z^k\|^2$. Putting these pieces together yields that $\{\widehat{\nu}_{i,m}\}_{i\in\mathbb{N}}$ weakly converges to $\widehat{\nu}_m$ almost surely.

Finally, we conclude from the lower semi-continuity of the PRW distance (cf. Theorem 3.3) and the Fatou's lemma that

$$\mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu, \widehat{\nu}_m)] \leq \mathbb{E}\left[\liminf_{i\to+\infty} \overline{\mathcal{PW}}_{p,k}(\mu_i, \widehat{\nu}_{i,m})\right] \leq \liminf_{i\to+\infty} \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\mu_i, \widehat{\nu}_{i,m})].$$

This completes the proof.

D.6 Proof of Theorem A.1

Using Assumption 3.2 and Theorem 3.3, the mapping $(\mu, \theta) \mapsto \overline{\mathcal{PW}}_{p,k}(\mu, \mu_\theta)$ is lower semi-continuous in $\mathcal{P}(\mathbb{R}^d) \times \Theta$. It remains to verify that the conditions in Theorem D.5 are satisfied.

We notice that the empirical measure $\widehat{\mu}_n(\omega)$ depends on $\omega \in \Omega$ only through $X_{1:n} \in \otimes_{i=1}^n \mathbb{R}^d$. Thus, we can write $\widehat{\mu}_n(\omega) = \widehat{\mu}_n(x)$ as a function in $\otimes_{i=1}^n \mathbb{R}^d$. Let $D = (\otimes_{i=1}^n \mathbb{R}^d) \times \Theta$, it is a Borel subset of $(\otimes_{i=1}^n \mathbb{R}^d) \times \mathbb{R}$. Since $\mathbb{R}^d$ is a Polish space, $\mathbb{R}^d \times \ldots \times \mathbb{R}^d$ endowed with the product topology is a Polish space. D_x is σ-compact for any $x \in \mathrm{proj}(D)$ since $D_x \subseteq \Theta$ and Θ is σ-compact.

Define $f(x, \theta) = \overline{\mathcal{PW}}_{p,k}(\widehat{\mu}_n(x), \mu_\theta)$, we claim that f is measurable on D and $f(x, \cdot)$ is lower semi-continuous on D_x. Indeed, we have shown that the mapping $(\mu, \theta) \mapsto \overline{\mathcal{PW}}_{p,k}(\mu, \mu_\theta)$ is lower semi-continuous and thus measurable in $\mathcal{P}(\mathbb{R}^d) \times \Theta$. The mapping $x \mapsto \widehat{\mu}_n(x)$ is measurable in $\otimes_{i=1}^n \mathbb{R}^d$. Since the composition of measurable functions is measure, f is measurable on D. Moreover, for any $x \in \otimes_{i=1}^n \mathbb{R}^d$, $f(x, \cdot)$ is lower semi-continuous on D_x since the mapping $(\mu, \theta) \mapsto \overline{\mathcal{PW}}_{p,k}(\mu, \mu_\theta)$ is lower semi-continuous on D. Putting these pieces together yields the desired results.

D.7 Proof of Theorem A.2

Using Assumption 3.2 and Lemma D.6, the mapping $(\nu, \theta) \mapsto \mathbb{E}[\overline{\mathcal{PW}}_{p,k}(\nu, \widehat{\mu}_{\theta,m}) \mid X_{1:n}]$ is lower semi-continuous in $\mathcal{P}(\mathbb{R}^d) \times \Theta$. Then the proof can be done similarly to the proof of Theorem A.1 using this result and Theorem D.5.

E Postponed Proofs in Subsection 3.4

In this section, we provide the detailed proofs for Theorem 3.12 and Theorem A.3. Our derivation is the refinement of the analysis in Bernton et al. (2019) for minimal Wasserstein estimators.

E.1 Preliminary technical results

To facilitate reading, we collect several preliminary technical results which will be used in the postponed proofs in subsection 3.4.

Let $(\mathcal{X}, \|\cdot\|_X)$ be a normed linear space and $\theta \mapsto f_\theta$ be a map from a subset Θ of $\mathbb{R}^d$ into $\mathcal{X}$. The statistical information comes from a sequence $\{f_n\}_{n \in \mathbb{N}}$ of random elements of $\mathcal{X}$, each of which is assumed to be measurable with respect to the σ-algebra generated by the balls in $\mathcal{X}$. In some sense f_n should converge to $f_{\theta_\star}$ where $\theta_\star$ is some fixed (but unknown) point in the interior of Θ. To avoid the abuse of notation, we use $K_1(x, \beta)$ here.

Theorem E.1 (Theorem 4.2 in Pollard (1980)) *Suppose the following assumptions hold:*

1. $\inf_{\theta \notin N} \|f_\theta - f_{\theta_\star}\|_X > 0$ *for every neighborhood* N *of* $\theta_\star$.
2. $\theta \mapsto f_\theta$ *is norm differentiable with non-singular derivative* $D_{\theta_\star}$ *at* $\theta_\star$.
3. *There exists a random element* $G_\star \in \mathcal{X}$ *for which* $G_n := \sqrt{n}(f_n - f_{\theta_\star}) \Rightarrow G_\star$ *in the sense for the metric induced by the norm* $\|\cdot\|_X$.

Then the limiting distribution of the goodness-of-fit statistic is given by

$$\sqrt{n} \inf_{\theta \in \Theta} \|f_n - f_\theta\|_X \Rightarrow \inf_{\theta \in \Theta} \|G_\star - \langle \theta, D_{\theta_\star} \rangle\|_X.$$

Let $K_1(x, \beta) = \{\theta : \|x - \langle \theta, D_{\theta_\star} \rangle\|_X \leq \inf_{\theta' \in \Theta} \|x - \langle \theta', D_{\theta_\star} \rangle\|_X + \beta\}$ and M_n is defined by

$$M_n = \left\{ \theta \in \Theta : \|f_n - f_\theta\|_X \leq \inf_{\theta' \in \Theta} \|f_n - f_{\theta'}\|_X + \eta_n / \sqrt{n} \right\},$$

where $\eta_n > 0$ is any sequence such that $\mathbb{P}(\eta_n \to 0) = 1$ and M_n is nonempty.

Theorem E.2 (Theorem 7.2 in Pollard (1980)) *Under the conditions of Theorem E.1, there exists a sequence of real number* $\beta_n \downarrow 0$ *satisfying*

$$\mathbb{P}_\star(M_n \subseteq \theta_\star + n^{-1/2} K_1(G_n, \beta_n)) \to 1, \quad \text{as } n \to +\infty.$$

Moreover, for any $\epsilon > 0$, *we have* $\mathbb{P}(d_H(K_1(G_n^\star, 0), K_1(G_n, \beta_n)) < \epsilon) \to 1$ *as* $n \to +\infty$.

E.2 Proof of Theorem 3.12

First, we show that $M_n \subseteq \mathcal{N}_1$ with (inner) probability approaching 1 as $n \to +\infty$. Indeed, with inner probability approaching 1, we have

$$\operatorname*{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_\theta) \subseteq \operatorname*{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{1,1}(\mu_\star, \mu_\theta).$$

By the definition of $\overline{\mathcal{PW}}_{1,1}$, we conclude that any minimizer of $\|\widehat{F}_n - F_\theta\|_L$ will be included in the set of minimizers of $\|F_\star - F_\theta\|_L$ with inner probability approaching 1. By Assumption 3.8, the minimizer of $\|F_\star - F_\theta\|_L$ is unique and $\mathcal{N}_1$ is the neighborhood of this minimizer. Putting these pieces together yields that the set $\inf_{\theta\in\Theta} \overline{\mathcal{PW}}_{1,1}(\widehat{\mu_n}, \mu_\theta)$ is contained in the set $\mathcal{N}_1$ with (inner) probability approaching 1 as $n \to +\infty$. By the definition of M_n, we achieve the desired result.

Then we make three key claims. First, we claim that $M_n \subseteq \Theta_n$ with (inner) probability approaching 1 as $n \to +\infty$, where Θ_n is defined by

$$\Theta_n = \left\{\theta \in \Theta : \|\theta - \theta_\star\|_\Theta \leq \frac{4\sqrt{n}\|\widehat{F}_n - F_\star\|_L + 2\eta_n}{c_\star\sqrt{n}}\right\}.$$

Indeed, for any $\theta \in \mathcal{N}_1$, we derive from the triangle inequality that

$$\|\widehat{F}_n - F_\theta\|_L - \|\widehat{F}_n - F_{\theta_\star}\|_L \geq \|F_\theta - F_\star\|_L - \|F_{\theta_\star} - F_\star\|_L - 2\|\widehat{F}_n - F_\star\|_L.$$

Using the definition of $\overline{\mathcal{PW}}_{1,1}$ together with Assumption 3.8, we have

$$\|\widehat{F}_n - F_\theta\|_L - \|\widehat{F}_n - F_{\theta_\star}\|_L \geq c_\star\|\theta - \theta_\star\|_\Theta - 2\|\widehat{F}_n - F_\star\|_L. \tag{18}$$

Since $M_n \subseteq \mathcal{N}_1$ with (inner) probability approaching one, Eq. (18) holds true for any $\theta \in M_n$ with (inner) probability approaching one. Moreover, by the definition of M_n, we have $\theta \in M_n$ satisfies

$$\|\widehat{F}_n - F_\theta\|_L \leq \inf_{\theta'\in\Theta} \overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_{\theta'}) + \frac{\eta_n}{\sqrt{n}} \leq \|\widehat{F}_n - F_{\theta_\star}\|_L + \frac{\eta_n}{\sqrt{n}} \tag{19}$$

Combining Eq. (18), Eq. (19) and the definition of Θ_n, we conclude that $\theta \in \Theta_n$ if $\theta \in M_n$ with (inner) probability approaching 1. This completes the proof the first claim.

Second, we claim that $\operatorname{argmin}_{\theta'\in\mathcal{N}_1} \|G_n - \langle\sqrt{n}(\theta' - \theta_\star), D_{\theta_\star}\rangle\|_L \subseteq \mathcal{N}_1 \cap \Theta_n$ with (inner) probability approaching 1 as $n \to +\infty$. Indeed, by the definition of G_n, we have

$$\|G_n - \langle\sqrt{n}(\theta' - \theta_\star), D_{\theta_\star}\rangle\|_L = \sqrt{n}\|\widehat{F}_n - F_{\theta_\star} - \langle\theta - \theta_\star, D_{\theta_\star}\rangle\|_L.$$

For the simplicity of notation, we let $R_\theta = F_\theta - F_{\theta_\star} - \langle\theta - \theta_\star, D_{\theta_\star}\rangle$. By Assumption 3.6, we have $\|R_\theta\|_L = o(\|\theta - \theta_\star\|_\Theta)$. By the definition of $\mathcal{N}_1$, we have $\|R_\theta\|_L \leq (1/2)c_\star\|\theta - \theta_\star\|_\Theta$. Therefore, for any $\theta \in \mathcal{N}_1$, we have

$$\begin{aligned}\|\widehat{F}_n - F_{\theta_\star} - \langle\theta - \theta_\star, D_{\theta_\star}\rangle\|_L \quad &\geq \quad \|\widehat{F}_n - F_\theta\|_L - \|R_\theta\|_L \\ &\overset{\text{Eq. (18)}}{\geq} \quad \|\widehat{F}_n - F_{\theta_\star}\|_L + (1/2)c_\star\|\theta - \theta_\star\|_\Theta - 2\|\widehat{F}_n - F_\star\|_L.\end{aligned}$$

This implies that, for any $\theta \in \mathcal{N}_1 \setminus \Theta_n$, we have

$$\|\widehat{F}_n - F_{\theta_\star} - \langle\theta - \theta_\star, D_{\theta_\star}\rangle\|_L \geq \|\widehat{F}_n - F_{\theta_\star}\|_L \geq \inf_{\theta'\in\mathcal{N}_1\cap\Theta_n} \|\widehat{F}_n - F_{\theta_\star} - \langle\theta' - \theta_\star, D_{\theta_\star}\rangle\|_L.$$

This completes the proof of the second claim.

Thirdly, we claim that there is an uniform control over the difference between $\theta \mapsto \sqrt{n}\|\widehat{F}_n - F_\theta\|_L$ and the convex map $\theta \mapsto \|G_n - \sqrt{n}\langle\theta - \theta_\star, D_{\theta_\star}\rangle\|_L$ over the set Ω_n with (inner) probability approaching 1 as $n \to +\infty$. Indeed, we define

$$\Gamma_n = \sup_{\theta\in\Omega_n} |\sqrt{n}\|\widehat{F}_n - F_\theta\|_L - \|G_n - \sqrt{n}\langle\theta - \theta_\star, D_{\theta_\star}\rangle\|_L|.$$

By the definition of G_n, we have

$$\begin{aligned}\Gamma_n &= \sup_{\theta\in\Omega_n} |\sqrt{n}\|\widehat{F}_n - F_{\theta_\star} - \langle\theta-\theta_\star, D_{\theta_\star}\rangle - R_\theta\|_L - \sqrt{n}\|\widehat{F}_n - F_{\theta_\star} - \langle\theta-\theta_\star, D_{\theta_\star}\rangle\|_L| \\ &= o\left(\sup_{\theta\in\Omega_n}\sqrt{n}\|\theta-\theta_\star\|_\Theta\right) = o(\sqrt{n}\|\widehat{F}_n - F_\star\|_L)\end{aligned}$$

By Assumption 3.7, we have $\Gamma_n \to 0$ as $\|\theta-\theta_\star\|_\Theta \to 0$ with (inner) probability approaching 1 as $n \to +\infty$. This completes the proof of the third claim.

By the definition of G_n and $G_n^\star$, we have $\|G_n - G_n^\star\|_L = \|\sqrt{n}(\widehat{F}_n - F_\star) - G_\star\|_L$. By Assumption 3.7, there exists a sequence $\tau_n^1 \to 0$ such that $\mathbb{P}(\|G_n - G_n^\star\|_L > \tau_n^1) \to 0$. By the definition of Γ_n and η_n, there exists two sequences $\tau_n^2 \to 0$ and $\tau_n^3 \to 0$ such that $\mathbb{P}(\Gamma_n > \tau_n^2) \to 0$ and $\mathbb{P}(\eta_n > \tau_n^3) \to 0$.

Let $\beta_n = \max\{2\tau_n^1, 2\tau_n^2 + \tau_n^3\}$, we have $\beta_n \to 0$ with (inner) probability approaching 1 as $n \to +\infty$. It remains to show that $M_n \subseteq K(G_n, \beta_n)$ with (inner) probability approaching 1 as $n \to +\infty$. Indeed, we have

$$\inf_{\theta'\in\mathcal{N}_1}\|G_n - \langle\sqrt{n}(\theta'-\theta_\star), D_{\theta_\star}\rangle\|_L \geq \inf_{\theta'\in\mathcal{N}_1}\sqrt{n}\|\widehat{F}_n - F_{\theta'}\|_L - \tau_n^2.$$

By the definition of M_n, let $\theta \in M_n$, the above inequality implies

$$\inf_{\theta'\in\mathcal{N}_1}\|G_n - \langle\sqrt{n}(\theta'-\theta_\star), D_{\theta_\star}\rangle\|_L \geq \sqrt{n}\|\widehat{F}_n - F_\theta\|_L - \tau_n^2 - \tau_n^3.$$

Since $M_n \subseteq \Theta_n$ with (inner) probability approaching 1 as $n \to +\infty$, we have

$$\sqrt{n}\|\widehat{F}_n - F_\theta\|_L \geq \|G_n - \langle\sqrt{n}(\theta-\theta_\star), D_{\theta_\star}\rangle\|_L - \tau_n^2.$$

Putting these pieces together with $\beta_n \geq 2\tau_n^2 + \tau_n^3$ yields that $\theta \in K(G_n, \beta_n)$.

Finally, let $\epsilon > 0$, we prove that $\mathbb{P}(d_H(K(G_n^\star, 0), K(G_n, \beta_n)) < \epsilon) \to 1$ as $n \to +\infty$. Indeed, by the triangle inequality, $\theta \in K(G_n^\star, 0)$ implies $\theta \in K(G_n, 2\|G_n - G_n^\star\|_L)$. Therefore, we conclude that $K(G_n^\star, 0) \subseteq K(G_n, \beta_n)$ with (inner) probability approaching one as $n \to +\infty$. On the other hand, $\theta \in K(G_n, \beta_n)$ implies $\theta \in K(G_n^\star, \beta_n + 2\|G_n - G_n^\star\|_L)$. By the definition of β_n, G_n and $G_n^\star$, we obtain that $\beta_n + 2\|G_n - G_n^\star\|_L \to 0$ with (inner) probability approaching one as $n \to +\infty$. By the definition of the Hausdorff metric, we conclude the desired result.

E.3 Proof of Theorem A.3

Different from Theorem 3.12, the proof of Theorem A.3 is relatively straightforward and based on Theorem E.1 and E.2. It is mostly because there exists $\theta_\star$ in the interior of Θ such that $F_\star = F_{\theta_\star}$.

More specifically, we consider $f_\theta = F_\theta$ and $f_n = \widehat{F}_n$ such that

$$F_\theta(u,t) = \int_{\mathbb{R}^d} \mathbf{1}_{(-\infty,t]}(\langle u, x\rangle)\, d\mu_\theta(x), \quad \widehat{F}_n(u,t) = (1/n)|\{i \in [n] : \langle u, X_i\rangle \leq t\}|.$$

Let $\mathcal{X} = L(\mathbb{S}^{d-1}\times\mathbb{R})$ and $\|\cdot\|_X = \|\cdot\|_L$, we can check that $(\mathcal{X}, \|\cdot\|_X)$ is a normed linear space. By the definition of $\overline{\mathcal{PW}}_{1,1}$, we have $\overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_\theta) = \|\widehat{F}_n - F_\theta\|_X$. By Assumption 3.1, $\widehat{F}_n$ converges to $F_\star$. Moreover, in well-specified setting, $F_\star = F_{\theta_\star}$ where $\theta_\star$ is some fixed (but unknown) point in the interior of Θ. Now we are ready to check the conditions of Theorem E.1.

First, Assumption A.1 and $\overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_\theta) = \|\widehat{F}_n - F_\theta\|_X$ imply C1. Furthermore, by the definition of norm differentiable, Assumption 3.6 and Assumption A.2 imply C2. Finally, Assumption 3.7 and $F_\star = F_{\theta_\star}$ imply C3. Therefore, we conclude from Theorem E.1 that

$$\sqrt{n}\inf_{\theta\in\Theta}\overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_\theta) = \sqrt{n}\inf_{\theta\in\Theta}\|\widehat{F}_n - F_\theta\|_L \Rightarrow \inf_{t\in\Theta}\|G_\star - \langle t, D_{\theta_\star}\rangle\|_L.$$

in the sense for the metric induced by the norm $\|\cdot\|_L$. This together with the definition of the norm $\|\cdot\|_L$ implies the desired result for the goodness-of-fit statistics.

On the other hand, Theorem E.2 can be applied with specific choice of η_n. More specifically, we notice that the estimator $\widehat{\theta}_n$ is well defined by

$$\widehat{\theta}_n := \operatorname*{argmin}_{\theta\in\Theta} \overline{\mathcal{PW}}_{1,1}(\widehat{\mu}_n, \mu_\theta) = \operatorname*{argmin}_{\theta\in\Theta} \|\widehat{F}_n - F_\theta\|_L.$$

Let $\eta_n = 0$, the set $M_n = \{\widehat{\theta}_n\}$ is a singleton set. This implies that $\sqrt{n}(\widehat{\theta}_n - \theta_\star) \Rightarrow K_1(G_\star, 0)$ as $n \to +\infty$ under its Hausdorff metric topology. Since the random map $\theta \to \max_{u\in\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t)\rangle|\, dt$ has a unique infimum almost surely, we have $K_1(G_\star, 0)$ is a singleton set defined by

$$K_1(G_\star, 0) = \operatorname*{argmin}_{\theta\in\Theta} \max_{u\in\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t)\rangle|\, dt.$$

In this case, the Hausdorff metric is simply induced by the norm $\|\cdot\|_L$. Putting these pieces together yields the desired result for the MPRW estimator of order 1.

E.4 Minor Technical Issues

We use the notations of Bernton et al. (2019, Theorem B.8) throughout this subsection. Indeed, in page 38-39 of the recent arvix version of Bernton et al. (2019), the authors prove that $m(H_n) = \inf_{u\in L_n} f(H_n, u)$, implicitly assuming that the minimizer of the map $\theta \mapsto \sqrt{n}\|F_n - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle\|_{L_1}$ is contained in the set $\mathcal{N}_1 = \{\theta \in \mathcal{N} : \|\theta - \theta_\star\|_{\mathcal{H}} \le c_\star/2\}$. However, this result is not obvious. Indeed, it seems difficult to derive such results from the existing fact that the minimizer of $\theta \mapsto \sqrt{n}\|F_n - F_\theta\|_{L_1}$ is contained in $\mathcal{N}$. We **only** have the uniform control over the difference between $\theta \mapsto \sqrt{n}\|F_n - F_\theta\|_{L_1}$ and $\theta \mapsto \sqrt{n}\|F_n - F_{\theta_\star} - \langle \theta - \theta_\star, D_{\theta_\star}\rangle\|_{L_1}$ over the set S_n instead of the whole set. So there is few relationship between the minimizers of these two mappings. Moreover, the techniques from the proof of Pollard (1980, Theorem 7.2) can not be applicable to fix this issue here since the proof depends on the assumption that $\mu_\star = \mu_{\theta_\star}$ which does not hold under model misspecification yet.

F Computational Aspects

The computation of the PRW distance is in general computationally intractable when the projection dimension is $k \ge 2$ since this amounts to solving a nonconvex max-min optimization model. Despite several pessimistic results (Paty and Cuturi, 2019; Niles-Weed and Rigollet, 2019), we adopt the Riemannian optimization toolbox (Absil et al., 2009) to develop a Riemannian supergradient algorithm and empirically show that our algorithm can approximate $\overline{\mathcal{PW}}_{2,k}(\widehat{\mu}_n, \widehat{\nu}_n)$ when the projection dimension is $k \ge 2$. Part of results can be found in the appendix of concurrent work (Lin et al., 2020) and we provide the details for the sake of completeness.

Approximation of $\overline{\mathcal{PW}}_{2,k}$. We consider the computation of $\overline{\mathcal{PW}}_{2,k}$ between empirical measures. Indeed, let $\{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ and $\{y_1, y_2, \ldots, y_n\} \subseteq \mathbb{R}^d$ denote sets of n atoms, and let $(r_1, r_2, \ldots, r_n) \in \Delta^n$ and $(c_1, c_2, \ldots, c_n) \in \Delta^n$ denote weight vectors, we define discrete measures $\widehat{\mu}_n := \sum_{i=1}^n r_i \delta_{x_i}$ and $\widehat{\nu}_n := \sum_{j=1}^n c_j \delta_{y_j}$. The computation of $\overline{\mathcal{PW}}_{2,k}(\widehat{\mu}_n, \widehat{\nu}_n)$ is equivalent to solving a structured max-min optimization model where the maximization and minimization are performed over the Stiefel manifold $\mathrm{St}(d,k) := \{U \in \mathbb{R}^{d\times k} \mid U^\top U = I_k\}$ and the transportation polytope $\Pi(\mu,\nu) := \{\pi \in \mathbb{R}_+^{n\times n} \mid r(\pi) = r,\ c(\pi) = c\}$ respectively. Formally, we have

$$\max_{U\in\mathbb{R}^{d\times k}} \min_{\pi\in\mathbb{R}_+^{n\times n}} \sum_{i=1}^n \sum_{j=1}^n \pi_{i,j} \|U^\top x_i - U^\top y_j\|^2 \quad \text{s.t. } U^\top U = I_k,\ r(\pi) = r,\ c(\pi) = c. \tag{20}$$

Eq. (20) is equivalent to the non-convex nonsmooth optimization model as follows,

$$\max_{U\in\mathrm{St}(d,k)} \left\{ f(U) := \min_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^n \sum_{j=1}^n \pi_{i,j} \|U^\top x_i - U^\top y_j\|^2 \right\}. \tag{21}$$

Fixing $U \in \mathrm{St}(d,k)$, Eq. (21) becomes a classical OT problem which can be either solved by the Sinkhorn iteration (Cuturi, 2013) or the variant of network simplex method in the POT package (Flamary and Courty, 2017). The key challenge is the maximization over the Stiefel manifold $\mathrm{St}(d,k) := \{U \in \mathbb{R}^{d\times k} \mid U^\top U = I_k\}$.

Algorithm 1 Riemannian SuperGradient Ascent with Network Simplex Iteration (RSGAN)

Input: measures $\{(x_i, r_i)\}_{i\in[n]}$ and $\{(y_j, c_j)\}_{j\in[n]}$, dimension k and tolerance ϵ.
Initialize: $U_0 \in \mathrm{St}(d,k)$ and $\gamma_0 > 0$.
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Compute $\pi_{t+1} \leftarrow \mathrm{OT}(\{(x_i, r_i)\}_{i\in[n]}, \{(y_j, c_j)\}_{j\in[n]}, U_t)$.
  Compute $\xi_{t+1} \leftarrow P_{\mathrm{T}_{U_t}\mathrm{St}}(2V_{\pi_{t+1}}U_t)$.
  Compute $\gamma_{t+1} \leftarrow \gamma_0/\sqrt{t+1}$.
  Compute $U_{t+1} \leftarrow \mathrm{Retr}_{U_t}(\gamma_{t+1}\xi_{t+1})$.
end for

Eq. (21) is a special instance of the Stiefel manifold optimization problem. The dimension of $\mathrm{St}(d,k)$ is equal to $dk - k(k+1)/2$ and the tangent space at the point $Z \in \mathrm{St}(d,k)$ is defined by $\mathrm{T}_Z\mathrm{St} := \{\xi \in \mathbb{R}^{d\times k} : \xi^\top Z + Z^\top \xi = 0\}$. We endow $\mathrm{St}(d,k)$ with Riemannian metric inherited from the Euclidean inner product $\langle X, Y\rangle$ for any $X, Y \in \mathrm{T}_Z\mathrm{St}$ and $Z \in \mathrm{St}(d,k)$. Then the projection of $G \in \mathbb{R}^{d\times k}$ onto $\mathrm{T}_Z\mathrm{St}$ is given by Absil et al. (2009, Example 3.6.2): $P_{\mathrm{T}_Z\mathrm{St}}(G) = G - Z(G^\top Z + Z^\top G)/2$. We make use of the notion of a *retraction*, which is the first-order approximation of an exponential mapping on the manifold and which is amenable to computation (Absil et al., 2009, Definition 4.1.1). For the Stiefel manifold, we have the following definition:

Definition F.1 *A retraction on* $\mathrm{St} \equiv \mathrm{St}(d,k)$ *is a smooth mapping* $\mathrm{Retr} : \mathrm{TSt} \to \mathrm{St}$ *from the tangent bundle* TSt *onto* St *such that the restriction of* Retr *onto* $\mathrm{T}_Z\mathrm{St}$*, denoted by* Retr_Z*, satisfies that (i)* $\mathrm{Retr}_Z(0) = Z$ *for all* $Z \in \mathrm{St}$ *where* 0 *denotes the zero element of* TSt*, and (ii) for any* $Z \in \mathrm{St}$*, it holds that* $\lim_{\xi\in\mathrm{T}_Z\mathrm{St},\xi\to 0} \|\mathrm{Retr}_Z(\xi) - (Z+\xi)\|_F/\|\xi\|_F = 0$.

Our algorithm uses the retraction based on the **QR decomposition** as suggested by Liu et al. (2019). More specifically, $\mathrm{Retr}_Z^{\mathrm{qr}}(\xi) = \mathrm{qr}(Z+\xi)$ where $\mathrm{qr}(A)$ is the Q factor of the QR factorization of A.

We start with a brief overview of the Riemannian supergradient ascent algorithm for nonsmooth Stiefel optimization, denoted by $\max_{U\in\mathrm{St}(d,k)} F(U)$. A generic Riemannian supergradient ascent algorithm for solving this problem is given by

$$U_{t+1} \;\leftarrow\; \mathrm{Retr}_{U_t}(\gamma_{t+1}\xi_{t+1}) \quad \text{for any } \xi_{t+1} \in \mathrm{subdiff}\, F(U_t),$$

where $\mathrm{subdiff}\, F(U_t)$ is Riemannian subdifferential of F at U_t and Retr is any retraction on $\mathrm{St}(d,k)$. The step size is set as $\gamma_{t+1} = \gamma_0/\sqrt{t+1}$ as suggested by (Li et al., 2019). By the definition of Riemannian subdifferential, ξ_t can be obtained by taking $\xi \in \partial F(U)$ and by setting $\xi_t = P_{\mathrm{T}_U\mathrm{St}}(\xi)$. Thus, it is necessary for us to specify the subdifferential of f in Eq. (21). We define $V_\pi = \sum_{i=1}^n \sum_{j=1}^n \pi_{i,j}(x_i - y_j)(x_i - y_j)^\top \in \mathbb{R}^{d\times d}$ which is symmetry and derive that

$$\partial f(U) \;=\; \mathrm{Conv}\{2V_{\pi^\star}U \mid \pi^\star \in \operatorname*{argmin}_{\pi\in\Pi(\mu,\nu)} \langle UU^\top, V_\pi\rangle\}, \quad \text{for any } U \in \mathbb{R}^{d\times k},$$

It remains to solve an OT problem with a given U at each inner loop of the maximization and use the output $\pi(U)$ to obtain a supergradient of f. The network simplex method can exactly solve this LP. To this end, we summarize the pseudocode of the RSGAN algorithm in Algorithm 1.

Approximation of $\underline{\mathcal{PW}}_{2,k}$. We recall the definition of the IPRW distance of order 2 as follows,

$$\underline{\mathcal{PW}}_{2,k}^2(\mu,\nu) = \int_{\mathbb{S}_{d,k}} \mathcal{W}_2^2(E^\star_\#\mu, E^\star_\#\nu) d\sigma(E),$$

where σ is the uniform distribution on $\mathbb{S}_{d,k}$ and $E^\star$ is the linear transformation associated with E for any $x \in \mathbb{R}^d$ by $E^\star(x) = E^\top x$. For any measurable function f and $\mu \in \mathscr{P}(\mathbb{R}^d)$, we denote $f_\#\mu$ as the push-forward of μ by f, so that $f_\#\mu(A) = \mu(f^{-1}(A))$ where $f^{-1}(A) = \{x \in \mathbb{R}^d : f(x) \in A\}$ for any Borel set A. We approximate the integral by selecting a finite set of projections $\mathcal{S} \subseteq \mathbb{S}_{d,k}$ and computing the empirical average:

$$\underline{\mathcal{PW}}_{2,k}^2(\mu,\nu) \approx \frac{1}{\mathrm{card}(\mathcal{S})} \sum_{E\in\mathcal{S}} \mathcal{W}_2^2(E^\star_\#\mu, E^\star_\#\nu).$$

The quality of this approximation depends on the sampling of $\mathbb{S}_{d,k}$. In this paper, we use random projections picked uniformly on $\mathbb{S}_{d,k}$, which is analogues to the approach proposed by Bonneel et al. (2015) for the case of $k = 1$; see **Sampling schemes** for the details.

Approximation of $\overline{\mathcal{PW}}_{p,1}$. We recall the definition of the PRW distance of order p with the projection dimension $k = 1$ as follows,

$$\overline{\mathcal{PW}}^p_{p,1}(\mu,\nu) := \sup_{u\in\mathbb{S}_{d,1}} \mathcal{W}^p_p(u^\star_\#\mu, u^\star_\#\nu) = \sup_{u\in\mathbb{S}_{d,1}} \int_0^1 |F^{-1}_{u^\star_\#\mu}(t) - F^{-1}_{u^\star_\#\nu}(t)|^p dt.$$

where $u \in \mathbb{S}_{d,1}$ is an unit d-dimensional vector, $u^\star$ is the linear transformation associated with u for any $x \in \mathbb{R}^d$ by $u^\star(x) = u^\top x$, and F^{-1}_ξ is the quantile function of ξ. This integral can be estimated using a Monte Carlo estimate and a linear interpolation of the quantile function. Following up Nadjahi et al. (2019, Appendix 4), we consider two approximations of this quantity. The first one is given by,

$$\overline{\mathcal{PW}}^p_{p,1}(\mu,\nu) = \sup_{u\in\mathbb{S}_{d,1}} \frac{1}{K}\sum_{k=1}^K |\tilde{F}^{-1}_{u^\star_\#\mu}(t_k) - \tilde{F}^{-1}_{u^\star_\#\nu}(t_k)|^p, \tag{22}$$

where $\{t_k\}_{k=1}^K$ are uniform and independent samples from $[0,1]$ and $\tilde{F}^{-1}_\xi$ is a linear interpolation of F^{-1}_ξ which denotes either the exact quantile function of a discrete measure ξ, or an approximation by a Monte Carlo procedure. The second one is given by

$$\overline{\mathcal{PW}}^p_{p,1}(\mu,\nu) = \sup_{u\in\mathbb{S}_{d,1}} \frac{1}{K}\sum_{k=1}^K |s_k - \tilde{F}^{-1}_{u^\star_\#\nu}(\tilde{F}_{u^\star_\#\mu}(s_k))|^p, \tag{23}$$

where $\{s_k\}_{k=1}^K$ are uniform and independent samples from $u^\star_\#\mu$ and $\tilde{F}_\xi$ (resp. $\tilde{F}^{-1}_\xi$) is a linear interpolation of F_ξ (resp. F^{-1}_ξ) which denotes either the exact cumulative distribution function (resp. quantile function) of a discrete measure ξ, or an approximation by a Monte Carlo procedure.

Sampling schemes. We explain the methods that we use to generate the i.i.d. samples from the uniform distribution on the set of $d \times k$ orthogonal matrices, i.e., $\mathbb{S}_{d,k} = \{E \in \mathbb{R}^{d\times k} : E^\top E = I_k\}$ and the i.i.d. samples from multivariate elliptically contoured stable distributions.

To sample from $\mathbb{S}_{d,k}$, we first construct the $(d \times k)$-dimensional matrix Z by drawing each of its components from the standard normal distribution $\mathcal{N}(0,1)$ and then perform the QR decomposition of it: $E = \mathrm{qr}(Z)$. By the definition, $E \in \mathbb{S}_{d,k}$ is an uniform sample.

To sample from multivariate elliptically contoured stable distributions, we follows the approach presented in Nadjahi et al. (2019, Appendix 4). Indeed, we recall that if $Y \in \mathbb{R}^d$ is α-stable and elliptically contoured, i.e., $Y \in \mathcal{E}\alpha\mathcal{S}_c(\Sigma, \mathbf{m})$, then its joint characteristic function is defined as, for any $t \in \mathbb{R}^d$ that,

$$\mathbb{E}\left[\exp(it^\top Y)\right] \;=\; \exp\left(-(t^\top\Sigma t)^{\alpha/2} + it^\top\mathbf{m}\right), \tag{24}$$

where Σ is a positive definite matrix (akin to a correlation matrix), $\mathbf{m} \in \mathbb{R}^d$ is a location vector (equal to the mean if it exists) and $\alpha \in (0,2)$ controls the thickness of the tail. Elliptically contoured stable distributions are scale mixtures of multivariate Gaussian distributions (Samoradnitsky, 2017, Proposition 2.5.2) with computationally intractable densities. Fortunately, it was shown by Nolan (2013) that sampling from multivariate elliptically contoured stable distributions is possible: let $A \sim \mathcal{S}_{\alpha/2}(\beta,\gamma,\delta)$ be a one-dimensional positive $(\alpha/2)$-stable random variable with $\beta = 1$, $\gamma = 2\cos(\pi\alpha/4)^{2/\alpha}$ and $\delta = 0$, and $G \sim \mathcal{N}(0,\Sigma)$. By the definition, $Y = \sqrt{A}G + \mathbf{m}$ satisfies Eq. (24) and $Y \sim \mathcal{E}\alpha\mathcal{S}_c(\Sigma,\mathbf{m})$.

Optimization methods. Computing the MPRW and MEPRW estimators are intractable in general. This is mainly because the PRW distance requires a maximization over infinitely many projections. Formally, we hope to solve the following minimax optimization model,

$$\min_{\theta\in\Theta} \overline{\mathcal{PW}}^p_{p,1}(\mu_\theta,\mu_\star) \;=\; \min_{\theta\in\Theta}\max_{u\in\mathbb{S}_{d,1}} \int_0^1 |F^{-1}_{u^\star_\#\mu_\theta}(t) - F^{-1}_{u^\star_\#\mu_\star}(t)|^p dt,$$

where $\{\mu_\theta : \theta \in \Theta\}$ is the model and $\mu_\star$ is the data-generating process. Following up the approach presented in Nadjahi et al. (2019) together with the approximation of $\overline{\mathcal{PW}}_{p,1}$, we consider using the ADAM optimization method to minimize the (expected) PRW distance over the set of parameters while applying multiple projected supergradient ascent to find an approximate projection u which maximizes over $\mathbb{S}_{d,1}$ at each inner loop. The ADAM optimization method is associated with the default parameter setting as suggested by Kingma and Ba (2015). At each inner loop, we run 5 projected supergradient ascent with the learning rate 10^{-3}.

Gaussian models. For the MPRW estimator, we consider the approximate $\overline{\mathcal{PW}}^2_{2,1}$ distance based on Eq. (23). Indeed, let μ denote $\mathcal{N}(\mathbf{m}, \sigma^2\mathbf{I})$ and $\widehat{\nu}$ denote the empirical probability measures of n samples drawn from the data-generating process, we define the function $f_1(\mathbf{m}, \sigma^2, u)$ as

$$f_1(\mathbf{m}, \sigma^2, u) \;=\; \frac{1}{\text{card}(\mathcal{S})}\sum_{s\in\mathcal{S}} |s - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(\tilde{F}_{u^\star_\#\mu}(s))|^2 \mathcal{N}(s; u^\top\mathbf{m}, \sigma^2\mathbf{I}),$$

where $\mathcal{S} \subseteq \mathbb{R}$ and $\mathcal{N}(s; u^\top\mathbf{m}, \sigma^2\mathbf{I})$ refers to the density function of Gaussian of parameters $(u^\top\mathbf{m}, \sigma^2\mathbf{I})$ evaluated at $s \in \mathcal{S}$. We compute the explicit gradient expression of $f_1(\mathbf{m}, \sigma^2, u)$ with respect to the mean $\mathbf{m}$, the variance σ^2 and the projection vector u as follows,

$$\begin{aligned}
\nabla_{\mathbf{m}} f_1(\mathbf{m}, \sigma^2, u) &= \frac{1}{\sigma^2\text{card}(\mathcal{S})}\sum_{s\in\mathcal{S}} \Big(|s - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(\tilde{F}_{u^\star_\#\mu}(s))|^2 \mathcal{N}(s; u^\top\mathbf{m}, \sigma^2\mathbf{I})(s - u^\top\mathbf{m})u\Big),\\
\nabla_{\sigma^2} f_1(\mathbf{m}, \sigma^2, u) &= \frac{1}{2\sigma^4\text{card}(\mathcal{S})}\sum_{s\in\mathcal{S}} \Big(|s - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(\tilde{F}_{u^\star_\#\mu}(s))|^2 \mathcal{N}(s; u^\top\mathbf{m}, \sigma^2\mathbf{I})((s - u^\top\mathbf{m})^2 - \sigma^2)\Big),\\
\nabla_{u} f_1(\mathbf{m}, \sigma^2, u) &= \frac{1}{\sigma^2\text{card}(\mathcal{S})}\sum_{s\in\mathcal{S}} \Big(|s - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(\tilde{F}_{u^\star_\#\mu}(s))|^2 \mathcal{N}(s; u^\top\mathbf{m}, \sigma^2\mathbf{I})(s - u^\top\mathbf{m})\mathbf{m}\Big).
\end{aligned}$$

For the MEPRW estimator, we consider the approximate $\overline{\mathcal{PW}}^2_{2,1}$ distance based on Eq. (22). Indeed, let $\widehat{\mu}$ and $\widehat{\nu}$ denote the empirical probability measures of m samples drawn from $\mathcal{N}(\mathbf{m}, \sigma^2\mathbf{I})$ and n samples drawn from the data-generating process, we define the function $f_2(\mathbf{m}, \sigma^2, u)$ as

$$f_2(\mathbf{m}, \sigma^2, u) \;=\; \frac{1}{K}\sum_{k=1}^K |\tilde{F}^{-1}_{u^\star_\#\widehat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(t_k)|^2,$$

where $\{t_k\}_{k=1}^K$ are uniform and independent samples from $[0, 1]$. We compute the explicit gradient expression of $f_2(\mathbf{m}, \sigma^2, u)$ with respect to the mean $\mathbf{m}$, the variance σ^2 and the projection vector u as follows,

$$\begin{aligned}
\nabla_{\mathbf{m}} f_2(\mathbf{m}, \sigma^2, u) &= -\frac{2}{K}\sum_{k=1}^K |\tilde{F}^{-1}_{u^\star_\#\widehat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(t_k)|u,\\
\nabla_{\sigma^2} f_2(\mathbf{m}, \sigma^2, u) &= -\frac{2}{K}\sum_{k=1}^K |\tilde{F}^{-1}_{u^\star_\#\widehat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(t_k)|\mathbf{m},\\
\nabla_{u} f_2(\mathbf{m}, \sigma^2, u) &= -\frac{1}{\sigma^2 K}\sum_{k=1}^K \Big(|\tilde{F}^{-1}_{u^\star_\#\widehat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(t_k)|(u^\top\mathbf{m} - \tilde{F}^{-1}_{u^\star_\#\widehat{\mu}}(t_k))\Big).
\end{aligned}$$

Elliptically contoured stable models. When comparing the MEPRW estimator with the MPRW estimator using elliptically contoured stable models, we also approximate these estimators using the ADAM optimization method with the default parameter setting.

We consider the approximate $\overline{\mathcal{PW}}^2_{2,1}$ distance based on Eq. (22). Indeed, let $\widehat{\mu}$ and $\widehat{\nu}$ denote the empirical probability measures of m samples drawn from $\mathcal{E}\alpha\mathcal{S}_c(\mathbf{I}, \mathbf{m})$ and n samples drawn from the data-generating process, we define the function $f_3(\mathbf{m}, u)$ as

$$f_3(\mathbf{m}, u) \;=\; \frac{1}{K}\sum_{k=1}^K |\tilde{F}^{-1}_{u^\star_\#\widehat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\#\widehat{\nu}}(t_k)|^2.$$

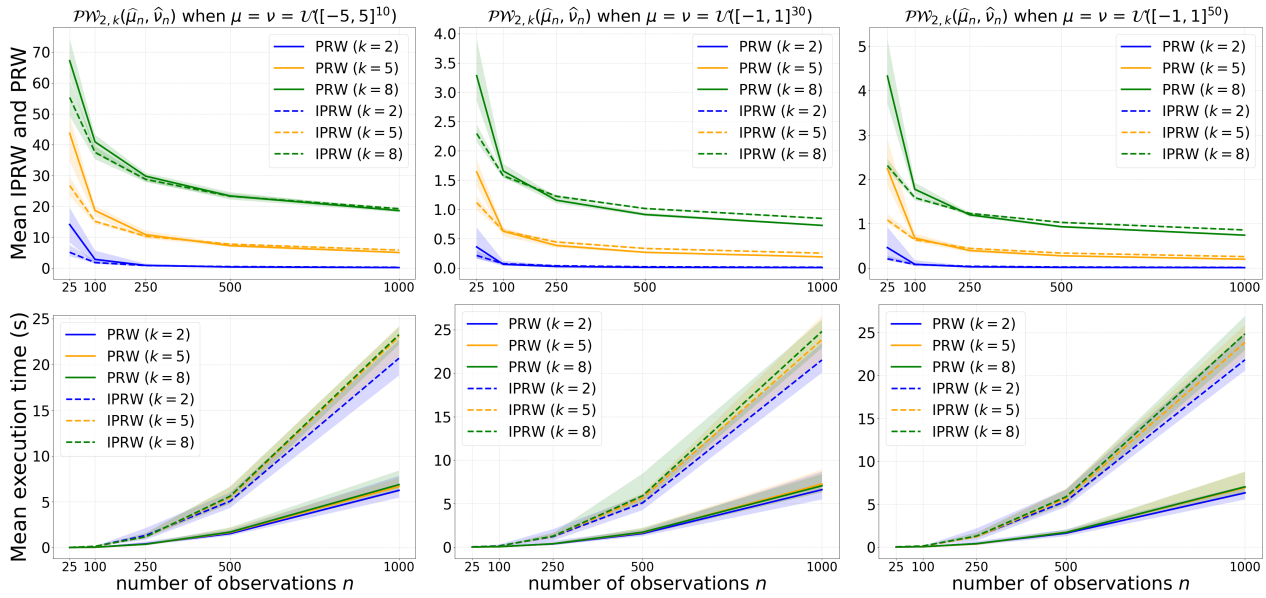


Figure 5: Mean values (Top) and mean computational time (Bottom) of the IPRW and PRW distances of order 2 between empirical measures $\widehat{\mu}_n$ and $\widehat{\nu}_n$ as the number of points n varies. Results are averaged over 100 runs.

where $\{t_k\}_{k=1}^K$ are uniform and independent samples from $[0,1]$. We compute the explicit gradient expression of $f(\mathbf{m}, u)$ with respect to the location parameter $\mathbf{m}$ and the projection vector u as follows,

$$\nabla_{\mathbf{m}} f_3(\mathbf{m}, u) = -\frac{2}{K}\sum_{k=1}^{K} |\tilde{F}^{-1}_{u^\star_\# \widehat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\# \widehat{\nu}}(t_k)| u,$$

$$\nabla_{u} f_3(\mathbf{m}, u) = -\frac{2}{K}\sum_{k=1}^{K} |\tilde{F}^{-1}_{u^\star_\# \widehat{\mu}}(t_k) - \tilde{F}^{-1}_{u^\star_\# \widehat{\nu}}(t_k)| \mathbf{m}.$$

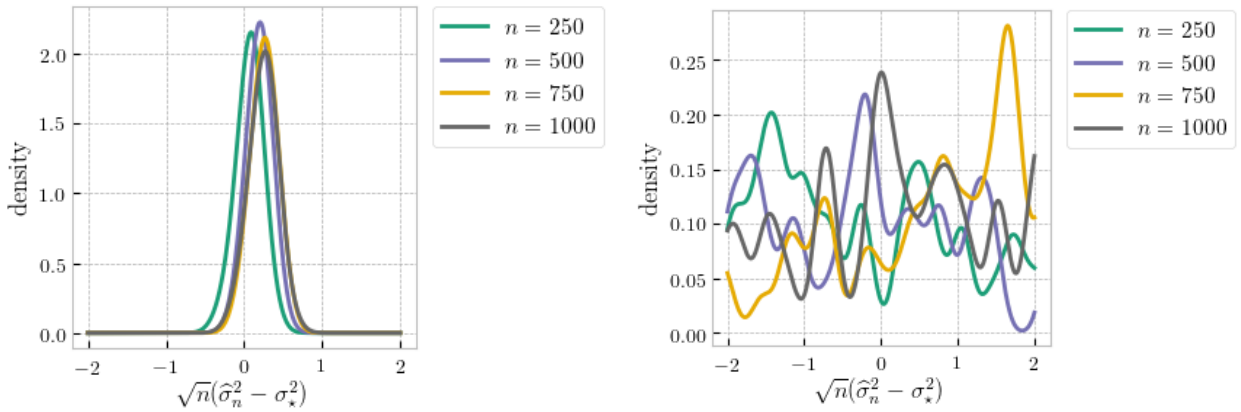
Generative modeling. We use the ADAM optimizer provided Pytorch GPU.

G Experimental Setup

Computing infrastructure. For the experiments on the uniform distribution over hypercube, we implement in Python 3.7 with Numpy 1.18 on a workstation with an Intel Core i5-9400F (6 cores and 6 threads) and 32GB memory, equipped with Ubuntu 18.04. For the experiments on MPRW and MEPRW estimators, we implement in Python 2.7 with Numpy 1.16 and IPython 5.8 on the same machine. These experiments were not conducted with GPU. For the experiments on neural networks, we implement on the same machine with 2 GPUs (GeForce GTX 1070 and GeForce GTX 2070).

Convergence and concentration. We conduct the experiment on the uniform distribution over different hypercubes which are also used in the experiment (Paty and Cuturi, 2019). In particular, we consider $\mu = \nu = \mathcal{U}([-v, v]^d)$ which is an uniform distribution over an hypercube and where d and v stand for the dimension and scale of the distribution respectively. $\widehat{\mu}_n$ and $\widehat{\nu}_n$ are empirical distributions corresponding to μ and ν with n samples. We evaluate the PRW and IPRW distance in terms of mean values and mean computational times over 100 runs for $(d, v) \in \{(10,1), (10,3), (30,1), (30,5), (50,1), (50,5)\}$. For the PRW distance, we run Algorithm 1 with EMD solver in the POT package (Flamary and Courty, 2017) and terminate the algorithm either when the maximum number of iterations $T = 30$ is reached or when $\|U_{t+1} - U_t\|_F \leq 10^{-6}$. For the IPRW distance, we draw 100 uniform and independent projections from $\mathbb{S}_{d,k}$ and compute each Wasserstein distances using EMD solver in the POT package again.

Model misspecification. We conduct the experiments on three type of data: the mixture of 8, 12 and 25 Gaussian distributions with Gaussian models $\mathcal{M}_1 = \{\mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{I}) : \mathbf{m} \in \mathbb{R}^2, \sigma^2 > 0\}$ and elliptically contoured



(a) Mixture of 12 Gaussian distributions

(b) Mixture of 25 Gaussian distributions

Figure 6: Probability density of estimation of centered and rescaled $\widehat{\sigma}_n$ on the Gaussian model for different n.

stable models $\mathcal{M}_2 = \{\mathcal{E}\alpha\mathcal{S}_c(\mathbf{I}, \mathbf{m}) : \mathbf{m} \in \mathbb{R}^2\}$. For data-generating process, we fix k centers $\{(a_i, b_i)\}_{1 \le i \le k}$. For each sample, we first randomly select $\mathbf{m}$ from the centers at uniform and then draw the sample from $\mathcal{N}(2\mathbf{m}, 0.01)$. For the mixture of 8 and 12 Gaussian distributions, the fixed set of centers are evenly distributed around a unit circle. For the mixture of 25 Gaussian distributions, the fixed set of centers are 25 grid points in $[-2, 2]^2$.

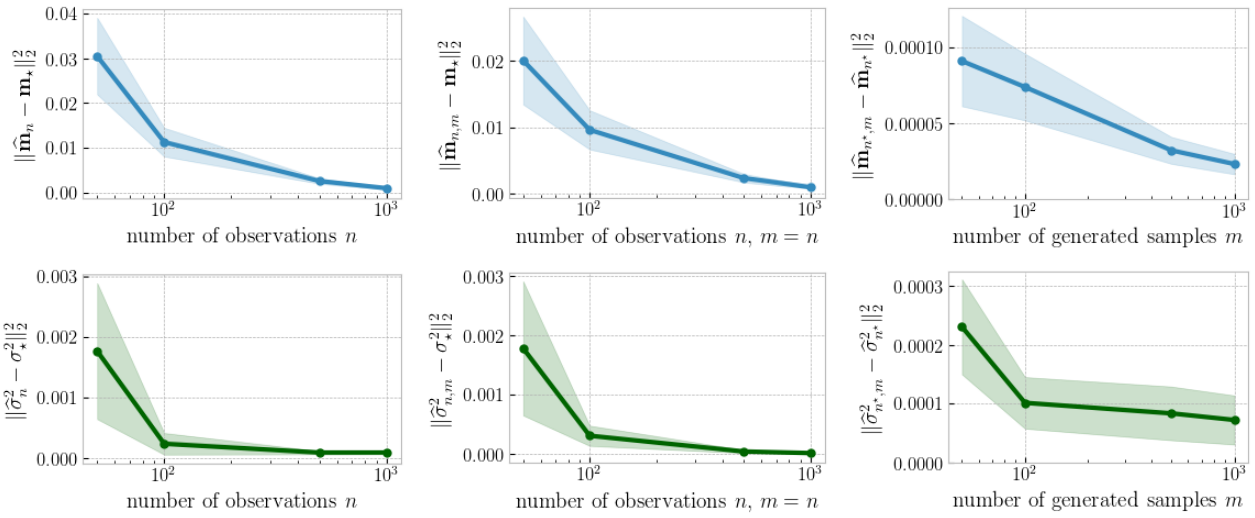
We use the ADAM optimization method with the default parameter setting to compute the MPRW and MEPRW estimators. At each inner loop, we run 5 projected supergradient ascent with the learning rate 10^{-3}. For the Gaussian models, we estimate the densities of $\widehat{\sigma}_n^2$ with a kernel density estimator by computing 100 times MPRW estimator of order 1. The maximum number of ADAM iterations is set as 20000. To illustrate the consistency of MPRW and MEPRW estimators, we compute 100 times MPRW and MEPRW estimators of order 2, where the maximum number of ADAM iterations are set as 20000 and 10000 respectively. We also verify the convergence of MEPRW to MPRW by computing 100 times these estimators on a fixed set of $n = 2000$ observations for different m generated samples from the model. The maximum number of ADAM iterations for MPRW and MEPRW estimators are set as 20000 and 10000. For the elliptically contoured stable models, we verify the consistency property of MEPRW and the convergence of MEPRW to MPRW. For the former one, we compute 100 times MEPRW estimator of order 2 and set the maximum number of ADAM iterations as 10000. For the latter one, we compute 100 times MPRW and MEPRW estimators of order 2 on a fixed set of $n = 100$ observations for different m generated samples from the model. The maximum number of ADAM iterations are set as 20000 and 10000. All of these settings are consistently used on the mixture of 8, 12 and 25 Gaussian distributions.

Generative modeling. The procedure of the max-SW generator is summarized as follows: we first sample a random variable Z from a fixed distribution on the base space $\mathcal{Z}$, and then transforms Z through a neural network parametrized by θ. This provides a parametric function $T_\theta : \mathcal{Z} \to \mathbb{R}^d$ which allows us to generate images from a distribution μ_θ. Our goal is to optimize the neural network parameters θ by minimizing the max-SW distance (Deshpande et al., 2019) between μ_θ and data-generating distribution. We use a neural network with the fully-connected configuration from Deshpande et al. (2018, Appendix D) and train our model with CIFAR10[7] and ImageNet200[8]. The former one consists of 60000 and 10000 images of size $3 \times 32 \times 32$ for training and testing while the latter one consists of 100000 and 10000 images for training and testing. We use the minimal expected max-SW estimator of order 2 approximated with 50 projected gradient ascent steps and 10^{-4} learning rate. We train for 1000 iterations with the ADAM optimizer (Kingma and Ba, 2015) and 10^{-4} learning rate.

H Additional Experimental Results

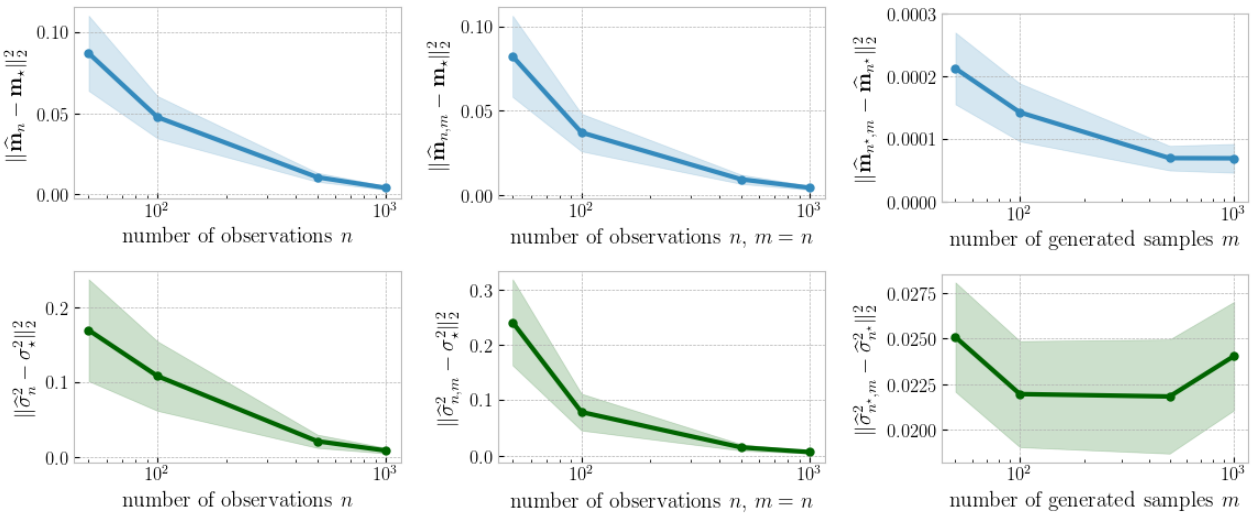
Convergence and concentration. Figure 5 presents average distances and computational times for $(d, v) \in \{(10, 5), (30, 1), (50, 1)\}$, where the shaded areas show the max-min values over 100 runs. We also observe that the IPRW distance is smaller than the PRW distance for small n, especially so when d and v are large. The two

[7]Available in https://www.cs.toronto.edu/ kriz/cifar.html
[8]Available in https://tiny-imagenet.herokuapp.com/



(a) MPRW vs. n (b) MEPRW vs. $n = m$ (c) MEPRW with $n = 2000$ vs. m

Figure 7: Minimal PRW and expected PRW estimations using Gaussian models and n samples from the mixture of 12 Gaussian distributions. Results are averaged over 100 runs and shaded areas represent standard deviation.



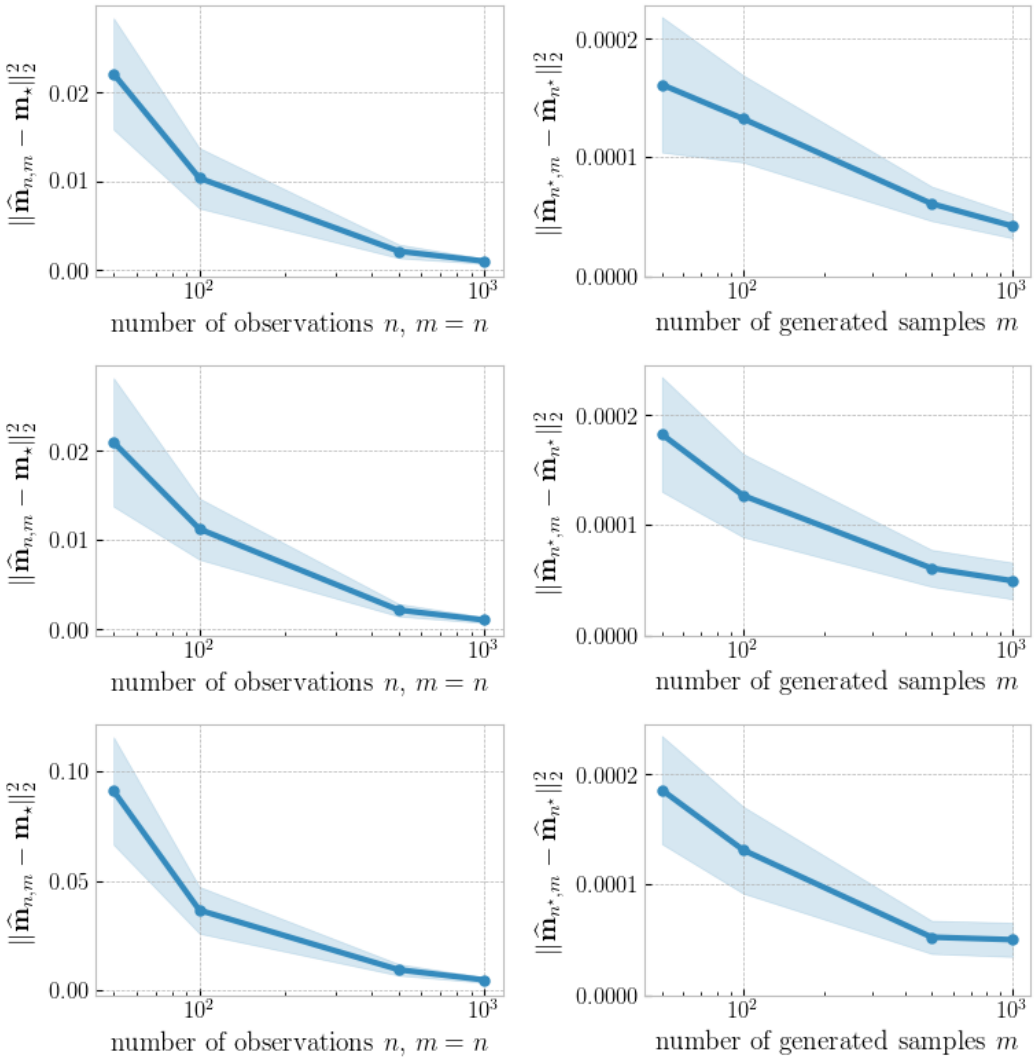
(a) MPRW vs. n (b) MEPRW vs. $n = m$ (c) MEPRW with $n = 2000$ vs. m

Figure 8: Minimal PRW and expected PRW estimations using Gaussian models and n samples from the mixture of 25 Gaussian distributions. Results are averaged over 100 runs and shaded areas represent standard deviation.

distances are close when n is large, supporting the theoretical results given by Theorem 3.4 and Theorem 3.6 in practice. The computation of the PRW distance is relatively faster than that of the IPRW distance.

Model misspecification: Gaussian models. Figure 6 shows the distributions centered and rescaled by $\sqrt{n}$ for a range of moderately large n, based on the two underlying models including the mixture of 12 Gaussian distributions and the mixture of 25 Gaussian distributions. The left figure supports the convergence rate and the limiting distribution of the estimator as derived in Theorem 3.12 on the mixture of 12 Gaussian distributions. The right figure suggests that the limiting distribution is not normal when the underlying model is given by the mixture of 25 Gaussian distributions. For the latter case, the result is not as anticipated by Theorem 3.12. This is possibly because we only conduct 5 projected supergradient ascent at each inner loop, which may not be enough to achieve a good approximate projection $u \in \mathbb{S}_{d,1}$.

Figure 7 and 8 demonstrate the large-sample consistency behavior of MPRW and MEPRW estimators on the



(a) MEPRW (b) MEPRW, $n^\star = 100$

Figure 9: Minimal expected PRW estimations using elliptically contoured stable models and n samples from the mixture of 8 Gaussian distributions (top), 12 Gaussian distributions (middle) and 25 Gaussian distributions (bottom), and m samples generated from the model. Results are averaged over 100 runs and shaded areas represent standard deviation.

mixture of 12 and 25 Gaussian distributions, which are expected since Assumption 3.1-3.3 are mild. The MEPRW estimator also converges to the MPRW estimator on the mixture of 12 Gaussian distributions, confirming Theorem 3.11. One exception in these experiments is the failure of convergence of MEPRW to MPRW on the mixture of 25 Gaussian distributions. Apparently, the results from Theorem 3.11 do not hold in this experiment setting. This is likely due to the violation of Assumption 3.5 that is necessary for Theorem 3.11 to hold.

Model misspecification: Elliptically contoured stable models. Figure 9 (a) illustrates the consistency of the MEPRW estimator $\widehat{\mathbf{m}}_{n,m}$, approximated with 5 projected supergradient ascent, the same way as for the Gaussian models. Figure 9 (b) confirms the convergence of $\widehat{\mathbf{m}}_{n,m}$ to the MPRW estimator $\widehat{\mathbf{m}}_n$, where we fix $n = 100$ observations and compute the mean squared error between these two estimators (using 5 projected supergradient ascent) for different values of m. Note that the MPRW estimator is approximated with the MEPRW obtained for a large enough value of m: $\widehat{\mathbf{m}}_n = \widehat{\mathbf{m}}_{n,10^4}$. To this end, our results on elliptically contoured stable models confirm Theorem 3.9, Theorem 3.10 and Theorem 3.11 in practice.

Generative modeling. Figure 10 presents the mean test loss on CIFAR10 over 10 runs, where the shaded areas show the max-min values over the runs. Here the minimal expected max-SW estimator of order 2 is

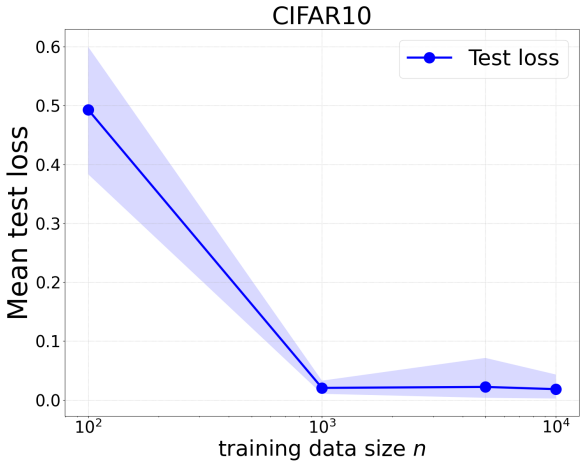


Figure 10: Mean test loss for different value of (n, m) on CIFAR10.

approximated with 20 projected gradient ascent steps and 10^{-4} learning rate. We trained for 1000 iterations with the ADAM optimizer (Kingma and Ba, 2015) and 10^{-4} learning rate. We also train the NNs with $(n, m) \in \{(100, 20), (1000, 40), (5000, 60), (10000, 100)\}$ where n is the number of training samples and m is the number of generated samples and compute the testing losses using the trained models on the testing dataset ($n = 10000$) with $m = 250$ generated samples. We compare these testing losses to that of a NN trained using $n = 60000$ (i.e., the training dataset) and $m = 200$ in Figure 10. Again, our results confirm Theorem 3.10 in practice.