Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections

SUPPLEMENTARY DOCUMENT

Kimia Nadjahi[1*], **Alain Durmus**[2], **Pierre E. Jacob**[3], **Roland Badeau**[1], **Umut Şimşekli**[4]

1: LTCI, Télécom Paris, Institut Polytechnique de Paris, France
2: Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190 Gif-sur-Yvette, France
3: Department of Information Systems, Decision Sciences and Statistics, ESSEC Business School, Cergy, France
4: INRIA - Département d'Informatique de l'École Normale Supérieure, PSL Research University, Paris, France

Abstract

This document provides details on our theoretical results and their proofs, and a complete description of the setup of our experiments.

S1 Conditional Central Limit Theorem for Gaussian Projections

We give the formal statement of the result presented in Section 2.2, corresponding to [1, Theorem 1] for the special case of one-dimensional projections.

Theorem S1 ([1, Theorem 1]). *There exists a constant C such that for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\int_{\mathbb{R}^d} \mathbf{W}_2^2\left(\theta_\sharp^\star \mu, \mathrm{N}\big(0, d^{-1}\mathtt{m}_2(\mu)\big)\right) \mathrm{d}\gamma_d(\theta) \leq C d^{-1}\{\alpha(\mu) + \big(\mathtt{m}_2(\mu)\beta_1(\mu)\big)^{1/2} + \mathtt{m}_2(\mu)^{1/5}\beta_2(\mu)^{4/5}\}\ , \tag{S1}$$

where

$$\mathtt{m}_2(\mu) = \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x)\ ,\ \ \alpha(\mu) = \int_{\mathbb{R}^d} \big|\ \|x\|^2 - \mathtt{m}_2(\mu)\big|\mathrm{d}\mu(x)\ , \tag{S2}$$

$$\beta_q(\mu) = \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} |\langle x, x'\rangle|^q \,\mathrm{d}(\mu\otimes\mu)(x,x')\right)^{\frac{1}{q}}\ , \tag{S3}$$

with $q \in \{1, 2\}$.

S2 Postponed proofs for Section 3

S2.1 Proof of Proposition 1

Proof of Proposition 1. Let $\theta \in \mathbb{R}^d$ and write $\theta = r\bar{\theta}$, $r \geq 0$ and $\bar{\theta} \in \mathbb{S}^{d-1}$. Then, we get

$$\mathbf{W}_p^p(\theta_\sharp^\star\mu, \theta_\sharp^\star\nu) = \mathbf{W}_p^p\big((r\bar{\theta})_\sharp^\star\mu, (r\bar{\theta})_\sharp^\star\nu\big) \tag{S4}$$

$$= \int_0^1 \big|F^{\leftarrow}_{(r\bar{\theta})_\sharp^\star\mu}(t) - F^{\leftarrow}_{(r\bar{\theta})_\sharp^\star\nu}(t)\big|^p \mathrm{d}t\ , \tag{S5}$$

*Corresponding author: kimia.nadjahi@telecom-paris.fr

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

where (S5) results from (3): $F_{\tilde{\mu}}$ and $F^{\leftarrow}_{\tilde{\mu}}$ denote the cumulative distribution and quantile function respectively, of a one-dimensional probability measure $\tilde{\mu}$, *i.e.* $F_{\tilde{\mu}}(s) = \tilde{\mu}((-\infty, s])$ and $F^{\leftarrow}_{\tilde{\mu}}(t) = \inf\{s' \in \mathbb{R} \,:\, F_{\tilde{\mu}}(s') \geq t\}$ for $s \in \mathbb{R}$ and $t \in [0,1]$. For any $r > 0$ and $\theta \in \mathbb{S}^{d-1}$, we get

$$F_{(r\bar{\theta})^\star_\sharp \mu}(s) = \big((r\bar{\theta})^\star_\sharp \mu\big)\{(-\infty, s]\} \tag{S6}$$

$$= \big(\bar{\theta}^\star_\sharp \mu\big)\{(-\infty, s/r]\} = F_{\bar{\theta}^\star_\sharp \mu}(s/r) \;, \tag{S7}$$

which easily implies that $F^{\leftarrow}_{(r\bar{\theta})^\star_\sharp \mu}(t) = rF^{\leftarrow}_{\bar{\theta}^\star_\sharp \mu}(t)$. Therefore, using this property in (S5), we obtain,

$$\mathbf{W}_p^p(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) = \int_0^1 \left| rF^{\leftarrow}_{\bar{\theta}^\star_\sharp \mu}(t) - rF^{\leftarrow}_{\bar{\theta}^\star_\sharp \nu}(t) \right|^p \mathrm{d}t \tag{S8}$$

$$= r^p \, \mathbf{W}_p^p(\bar{\theta}^\star_\sharp \mu, \bar{\theta}^\star_\sharp \nu) \;. \tag{S9}$$

By applying a d-spherical change of variables in the definition of $\widetilde{\mathbf{SW}}_p$ (9) and plugging (S9),

$$\widetilde{\mathbf{SW}}_p^p(\mu, \nu) = \int_{\mathbb{R}_+} \int_{\mathbb{S}^{d-1}} r^p \mathbf{W}_p^p\big(\bar{\theta}^\star_\sharp \mu, \bar{\theta}^\star_\sharp \nu\big) \, (2\pi)^{-\frac{d}{2}} d^{\frac{d}{2}} e^{-\frac{d}{2}\|r\bar{\theta}\|^2} r^{d-1} \mathrm{d}\bar{\theta}\mathrm{d}r \tag{S10}$$

$$= (2\pi)^{-\frac{d}{2}} d^{\frac{d}{2}} \int_{\mathbb{R}_+} r^{p+d-1} e^{-\frac{d}{2}r^2} \left(\int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p\big(\bar{\theta}^\star_\sharp \mu, \bar{\theta}^\star_\sharp \nu\big) \mathrm{d}\bar{\theta} \right) \mathrm{d}r \;. \tag{S11}$$

Since the surface area of $\mathbb{S}^{d-1}$ is equal to $2\pi^{\frac{d}{2}}\Gamma(d/2)^{-1}$ [2], and by definition of SW (4), $\int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p\big(\bar{\theta}^\star_\sharp \mu, \bar{\theta}^\star_\sharp \nu\big)\mathrm{d}\bar{\theta} = 2\pi^{\frac{d}{2}}\Gamma(d/2)^{-1}\mathbf{SW}_p^p(\mu, \nu)$.

Besides, by applying the change of variables $t = (d/2)^{1/2} r$,

$$\int_{\mathbb{R}_+} r^{p+d-1} e^{-\frac{d}{2}r^2} \mathrm{d}r = 2^{(p+d)/2} d^{-(p+d)/2} \int_{\mathbb{R}_+} t^{p+d-1} e^{-t^2} \mathrm{d}t = 2^{(p+d)/2-1} d^{-(p+d)/2} \, \Gamma\big((d+p)/2\big)$$

We finally obtain,

$$\widetilde{\mathbf{SW}}_p^p(\mu, \nu) = (2/d)^{p/2} \, \frac{\Gamma\big(d/2 + p/2\big)}{\Gamma(d/2)} \, \mathbf{SW}_p^p(\mu, \nu) \;. \tag{S12}$$

□

S2.2 Proof of Theorem 1

Proof of Theorem 1. By the triangle inequality, for any $\theta \in \mathbb{R}^d$,

$$\left| \mathbf{W}_2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) - \mathbf{W}_2\{\mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \right| \tag{S13}$$

$$\leq \mathbf{W}_2\{\theta^\star_\sharp \mu_d, \mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d))\} + \mathbf{W}_2\{\theta^\star_\sharp \nu_d, \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \tag{S14}$$

Therefore, taking the integral with respect to $\boldsymbol{\gamma}_d$,

$$\int_{\mathbb{R}^d} \Big(\mathbf{W}_2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) - \mathbf{W}_2\{\mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \Big)^2 \mathrm{d}\boldsymbol{\gamma}_d(\theta) \tag{S15}$$

$$\leq \int_{\mathbb{R}^d} \Big(\mathbf{W}_2\{\theta^\star_\sharp \mu_d, \mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d))\} + \mathbf{W}_2\{\theta^\star_\sharp \nu_d, \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \Big)^2 \mathrm{d}\boldsymbol{\gamma}_d(\theta) \tag{S16}$$

$$\leq 2 \left\{ \int_{\mathbb{R}^d} \mathbf{W}_2^2\{\theta^\star_\sharp \mu_d, \mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d))\} \mathrm{d}\boldsymbol{\gamma}_d(\theta) + \int_{\mathbb{R}^d} \mathbf{W}_2^2\{\theta^\star_\sharp \nu_d, \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \mathrm{d}\boldsymbol{\gamma}_d(\theta) \right\} \;, \tag{S17}$$

where (S17) follows from $(a+b)^2 \leq 2(a^2+b^2)$. Then, we apply Theorem S1 to bound (S17), and we conclude there exists a universal constant $C > 0$ such that

$$\int_{\mathbb{R}^d} \Big(\mathbf{W}_2(\theta^\star_\sharp \mu_d, \theta^\star_\sharp \nu_d) - \mathbf{W}_2\{\mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\} \Big)^2 \mathrm{d}\boldsymbol{\gamma}_d(\theta) \tag{S18}$$

$$\leq C\big(\Xi_d(\mu_d) + \Xi_d(\nu_d)\big) \tag{S19}$$

Using $|\|a\| - \|b\|| \leq \|a - b\|$ in $\mathrm{L}^2(\boldsymbol{\gamma}_d)$ gives

$$\Big|\Big\{\int_{\mathbb{R}^d} \mathbf{W}_2^2(\theta_\sharp^\star \mu_d, \theta_\sharp^\star \nu_d)\mathrm{d}\boldsymbol{\gamma}_d(\theta)\Big\}^{1/2} - \Big\{\int_{\mathbb{R}^d} \mathbf{W}_2^2\{\mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\}\mathrm{d}\boldsymbol{\gamma}_d(\theta)\Big\}^{1/2}\Big| \tag{S20}$$

$$\leq \Big\{\int_{\mathbb{R}^d} \Big(\mathbf{W}_2(\theta_\sharp^\star \mu_d, \theta_\sharp^\star \nu_d) - \mathbf{W}_2\{\mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\}\Big)^2 \mathrm{d}\boldsymbol{\gamma}_d(\theta)\Big\}^{1/2} \tag{S21}$$

$$\leq C^{1/2}\big(\Xi_d(\mu_d) + \Xi_d(\nu_d)\big)^{1/2} \tag{S22}$$

By (9) and Proposition 1,

$$\int_{\mathbb{R}^d} \mathbf{W}_2^2(\theta_\sharp^\star \mu_d, \theta_\sharp^\star \nu_d)\mathrm{d}\boldsymbol{\gamma}_d(\theta) = \widetilde{\mathbf{SW}}_2^2(\mu_d, \nu_d) = \mathbf{SW}_2^2(\mu_d, \nu_d) \ .$$

We then obtain the final result by rewritting (S20) as $\big|\mathbf{SW}_2(\mu_d, \nu_d) - \mathbf{W}_2\{\mathrm{N}(0, d^{-1}\mathtt{m}_2(\mu_d)), \mathrm{N}(0, d^{-1}\mathtt{m}_2(\nu_d))\}\big|$.

□

S2.3 Proof of Proposition 2

Proof of Proposition 2. This result follows from an analogous translation property of the Wasserstein distance: by [3, Remark 2.19], $\mathbf{W}_2$ (1) can factor out translations; in particular, for any $\xi, \xi' \in \mathcal{P}_2(\mathbb{R}^d)$ with respective means $\mathbf{m}_\xi, \mathbf{m}_{\xi'}$ and centered versions $\bar{\xi}, \bar{\xi}'$,

$$\mathbf{W}_2^2(\xi, \xi') = \mathbf{W}_2^2(\bar{\xi}, \bar{\xi}') + \|\mathbf{m}_\xi - \mathbf{m}_{\xi'}\|^2 \ . \tag{S23}$$

By using (S23) in the definition of SW of order 2 (4), we obtain for any $\mu_d, \nu_d \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathbf{SW}_2^2(\mu_d, \nu_d) = \int_{\mathbb{S}^{d-1}} \mathbf{W}_2^2(\theta_\sharp^\star \bar{\mu}_d, \theta_\sharp^\star \bar{\nu}_d)\mathrm{d}\boldsymbol{\sigma}(\theta) + \int_{\mathbb{S}^{d-1}} |\mathbf{m}_{\theta_\sharp^\star \mu_d} - \mathbf{m}_{\theta_\sharp^\star \nu_d}|^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{S24}$$

$$= \mathbf{SW}_2^2(\bar{\mu}_d, \bar{\nu}_d) + \int_{\mathbb{S}^{d-1}} |\mathbf{m}_{\theta_\sharp^\star \mu_d} - \mathbf{m}_{\theta_\sharp^\star \nu_d}|^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \ . \tag{S25}$$

By the properties of pushforward measures, $\mathbf{m}_{\theta_\sharp^\star \xi} = \langle \theta, \mathbf{m}_\xi \rangle$ for any $\theta \in \mathbb{S}^{d-1}$ and $\xi \in \mathcal{P}_2(\mathbb{R}^d)$. The second term of (S25) can thus be reformulated as

$$\int_{\mathbb{S}^{d-1}} |\mathbf{m}_{\theta_\sharp^\star \mu_d} - \mathbf{m}_{\theta_\sharp^\star \nu_d}|^2 \mathrm{d}\boldsymbol{\sigma}(\theta) = \int_{\mathbb{S}^{d-1}} |\langle \theta, \mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d} \rangle|^2 \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{S26}$$

$$= (\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d})^\top \left(\int_{\mathbb{S}^{d-1}} \theta\theta^\top \mathrm{d}\boldsymbol{\sigma}(\theta)\right)(\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d}) \tag{S27}$$

$$= (1/d)\ \|\mathbf{m}_{\mu_d} - \mathbf{m}_{\nu_d}\|^2 \ , \tag{S28}$$

where the last equation results from $\int_{\mathbb{S}^{d-1}} \theta\theta^\top \mathrm{d}\boldsymbol{\sigma}(\theta) = (1/d)\mathbf{I}_d$. The final result is obtained by incorporating (S28) in (S25).

□

S2.4 Error analysis under independence

This section gives a detailed analysis of the error bound under the first setting discussed in Section 3.3: we consider sequences of independent random variables which have zero means and finite fourth-order moments, and we derive an upper bound for Ξ_d in the next proposition.

Proposition S1. *Let $(X_j)_{j\in\mathbb{N}^*}$ be a sequence of independent random variables with zero means and $\mathbb{E}[X_j^4] < +\infty$ for $j \in \mathbb{N}^*$. Set for any $d \in \mathbb{N}^*$, $X_{1:d} = \{X_j\}_{j=1}^d$ and let μ_d be the distribution of $X_{1:d}$. Then, we have*

$$\Xi_d(\mu_d) \leq d^{-1/2}\big\{\max_{1\leq j\leq d} \mathrm{Var}[X_j^2]\big\}^{1/2} + \big\{d^{-1/4} + d^{-2/5}\big\} \max_{1\leq j\leq d} \mathrm{Var}[X_j] \ . \tag{S29}$$

Proof of Proposition S1. Given the definition of $\Xi_d(\mu_d)$ (7), the proof consists in bounding $\mathtt{m}_2(\mu_d)$, $\alpha(\mu_d)$ and $\beta_q(\mu_d)$ for $q \in \{1, 2\}$.

Since for any $j \in \{1, \dots, d\}$, $\mathbb{E}[X_j] = 0$, then $\mathrm{Var}\big[X_j\big] = \mathbb{E}[X_j^2]$ and

$$\mathtt{m}_2(\mu_d) = \sum_{j=1}^{d} \mathbb{E}[X_j^2] = \sum_{j=1}^{d} \mathrm{Var}[X_j] \leq d \max_{1 \leq j \leq d} \mathrm{Var}[X_j] \tag{S30}$$

To bound $\alpha(\mu_d)$, we first use the Cauchy–Schwarz inequality.

$$\alpha(\mu_d) \leq \left\{ \int_{\mathbb{R}^d} \big(\|x_{1:d}\|^2 - \mathtt{m}_2(\mu_d) \big)^2 \mathrm{d}\mu_d(x_{1:d}) \right\}^{1/2} \tag{S31}$$

Besides, $\int_{\mathbb{R}^d} \big(\|x_{1:d}\|^2 - \mathtt{m}_2(\mu_d) \big)^2 \mathrm{d}\mu_d(x_{1:d}) = \mathrm{Var}\big[\|X_{1:d}\|^2\big]$, and since the d components of $X_{1:d}$ are assumed to be pairwise independent, $\mathrm{Var}\big[\|X_{1:d}\|^2\big] = \sum_{j=1}^{d} \mathrm{Var}\big[X_j^2\big]$. We conclude that

$$\alpha(\mu_d) \leq \left(\sum_{j=1}^{d} \mathrm{Var}\big[X_j^2\big] \right)^{1/2} \leq \big(d \max_{1 \leq j \leq d} \mathrm{Var}[X_j^2] \big)^{1/2} . \tag{S32}$$

Finally, we bound $\beta_q(\mu_d)$ for $q \in \{1, 2\}$ by bounding $\beta_2(\mu_d)$ then using the fact that $\beta_1(\mu_d) \leq \beta_2(\mu_d)$ by the Cauchy–Schwarz inequality. Denote by $X'_{1:d}$ an independent copy of $X_{1:d}$.

$$\langle X_{1:d}, X'_{1:d} \rangle^2 = \Big(\sum_{j=1}^{d} X_j X'_j \Big)^2 = \sum_{j=1}^{d} {X_j}^2 {X'_j}^2 + 2 \sum_{i<j} X_i X'_i X_j X'_j \, . \tag{S33}$$

Since $X_{1:d}$ and $X'_{1:d}$ are independent on one hand, and they both are sequences of d independent random variables with zero means on the other hand, we have

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} (\langle x_{1:d}, x'_{1:d} \rangle)^2 \mathrm{d}(\mu_d \otimes \mu_d)(x_{1:d}, x'_{1:d}) \tag{S34}$$

$$= \sum_{j=1}^{d} \mathbb{E}\big[{X_j}^2\big] \mathbb{E}\big[{X'_j}^2\big] = \sum_{j=1}^{d} \mathbb{E}\big[X_j^2\big]^2 = \sum_{j=1}^{d} \mathrm{Var}\big[X_j\big]^2 \, . \tag{S35}$$

Therefore, $\beta_2(\mu_d) \leq (\sum_{j=1}^{d} \mathrm{Var}[X_j]^2)^{1/2} \leq (d \max_{1 \leq j \leq d} \mathrm{Var}[X_j]^2)^{1/2}$. Since $X_{1:d}$ has finite second and fourth-order moments, $\max_{1 \leq j \leq d} \mathrm{Var}[X_j]$, $\max_{1 \leq j \leq d} \mathrm{Var}[X_j^2] < \infty$, and we get

$$\mathtt{m}_2(\mu_d) \leq d \max_{1 \leq j \leq d} \mathrm{Var}[X_j], \quad \alpha(\mu_d) \leq d^{1/2} (\max_{1 \leq j \leq d} \mathrm{Var}[X_j^2])^{1/2}, \tag{S36}$$

$$\beta_1(\mu_d), \, \beta_2(\mu_d) \leq d^{1/2} \max_{1 \leq j \leq d} \mathrm{Var}[X_j] \, . \tag{S37}$$

The final result is obtained by bounding $\Xi(\mu_d)$ using (S36).

□

Note that the setting considered in Proposition S1 was mentioned in [1] to illustrate the conditions of [1, Corollary 3]. We derived an explicit upper bound of Ξ_d under this setting for completeness, showing that $\Xi_d(\mu_d)$ goes to zero as $d \to \infty$, which we can then use to refine the convergence rate in Theorem 1, as we explained in Section 3.3.

S2.5 Error analysis under weak dependence

We now analyze the error under the weak dependence condition introduced in Definition 1. Specifically, the proposition below gives the formal statement of the result mentioned before Corollary 1: we consider a sequence of fourth-order weakly dependent random variables, and we prove that $\Xi(\mu_d)$ goes to zero as $d \to \infty$, with a convergence rate that depends on $\{\rho(n)\}_{n \in \mathbb{N}^*}$.

Proposition S2. *Let $(X_j)_{j\in\mathbb{N}^*}$ be a sequence of random variables which is fourth-order weakly dependent. Set for any $d \in \mathbb{N}^*$, $X_{1:d} = \{X_j\}_{j=1}^d$ and denote by μ_d the distribution of $X_{1:d}$. Then, there exists a universal constant $C > 0$ such that*

$$\Xi_d(\mu_d) \leq C\Big\{ d^{-1/2}\big(\rho(0) + 2\rho_\infty\big)^{1/2} + d^{-1/4}\rho(0)^{1/2}\big(\rho(0)^2 + 2\rho_\infty \max_{1\leq k\leq d-1}\rho(k)\big)^{1/4} \tag{S38}$$

$$+ d^{-2/5}\rho(0)^{1/5}\big(\rho(0)^2 + 2\rho_\infty \max_{1\leq k\leq d-1}\rho(k)\big)^{2/5}\Big\} . \tag{S39}$$

Proof of Proposition S2. We proceed as in the proof of Proposition S1, *i.e.* by bounding $\mathtt{m}_2(\mu_d)$, $\alpha(\mu_d)$ and $\beta_2(\mu_d)$.

Since $(X_j)_{j\in\mathbb{N}^*}$ is assumed to be fourth-order weakly dependent, then by Definition 1, there exist some constant $K \geq 0$ and a nonincreasing sequence of real coefficients $\{\rho(n)\}_{n\in\mathbb{N}}$ such that, for any $1 \leq i \leq j \leq d$,

$$|\mathrm{Cov}(X_i^2, X_j^2)| \leq K\rho(j-i), \quad |\mathrm{Cov}(X_i, X_j)| \leq K\rho(j-i) \tag{S40}$$

First, using the same arguments as in (S30), we have $\mathtt{m}_2(\mu_d) = \sum_{j=1}^d \mathrm{Var}[X_j]$. We then use the second inequality in (S40) to bound $\mathtt{m}_2(\mu_d)$ as follows.

$$\mathtt{m}_2(\mu_d) = \sum_{j=1}^d \mathrm{Cov}(X_j, X_j) \leq dK\rho(0) \tag{S41}$$

Regarding $\alpha(\mu_d)$, we use the Cauchy–Schwarz inequality again (S31) but in this setting, the right-hand side features non-zero covariance terms:

$$\int_{\mathbb{R}^d} \big(\|x_{1:d}\|^2 - \mathtt{m}_2(\mu_d)\big)^2 \mathrm{d}\mu_d(x_{1:d}) = \mathrm{Var}\big[\|X_{1:d}\|^2\big] \tag{S42}$$

$$= \sum_{j=1}^d \mathrm{Var}\big[X_j^2\big] + 2\sum_{i<j} \mathrm{Cov}\big(X_i^2, X_j^2\big) . \tag{S43}$$

By using the first inequality in (S40), we get for any $d \in \mathbb{N}^*$,

$$\sum_{j=1}^d \mathrm{Var}\big[X_j^2\big] = \sum_{j=1}^d \mathrm{Cov}(X_j^2, X_j^2) \leq Kd\rho(0) , \tag{S44}$$

$$\sum_{i<j} \mathrm{Cov}\big(X_i^2, X_j^2\big) \leq \sum_{i<j} |\mathrm{Cov}\big(X_i^2, X_j^2\big)| \leq K\sum_{i<j}\rho(j-i) \tag{S45}$$

$$\leq K\sum_{n=1}^{d-1}(d-n)\rho(n) \tag{S46}$$

$$\leq Kd\sum_{n=1}^{d-1}\rho(n) \leq Kd\rho_\infty \tag{S47}$$

where (S46) results from the change of variable $n = j - i$. Besides, by Definition 1, $\{\rho(n)\}_{n\in\mathbb{N}}$ is a nonincreasing sequence satisfying $\sum_{n=0}^{+\infty}\rho(n) \leq \rho_\infty < +\infty$, hence (S47). We conclude that for any $d \in \mathbb{N}^*$,

$$\alpha(\mu_d) \leq d^{1/2}K^{1/2}\big(\rho(0) + 2\rho_\infty\big)^{1/2} \tag{S48}$$

Let us now bound $\beta_2(\mu_d)$. First, for any $d \in \mathbb{N}^*$,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} (\langle x_{1:d}, x'_{1:d} \rangle)^2 \mathrm{d}(\mu_d \otimes \mu_d)(x_{1:d}, x'_{1:d}) \tag{S49}$$

$$= \sum_{j=1}^{d} \mathbb{E}\big[{X_j}^2\big] \mathbb{E}\big[{X'_j}^2\big] + 2 \sum_{i<j} \mathbb{E}\big[X_i X_j\big] \mathbb{E}\big[X'_i X'_j\big] \tag{S50}$$

$$= \sum_{j=1}^{d} \mathbb{E}\big[{X_j}^2\big]^2 + 2 \sum_{i<j} \mathbb{E}\big[X_i X_j\big]^2 \tag{S51}$$

$$= \sum_{j=1}^{d} \mathrm{Var}\big[X_j\big]^2 + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j)^2 \,, \tag{S52}$$

where we used $\mathbb{E}[X_i] = 0$ for any $i \geq 1$. To bound (S52), we apply the second inequality in (S40), and adapt the arguments used to prove (S44) and (S46), .

$$\sum_{j=1}^{d} \mathrm{Var}\big[X_j\big]^2 \leq K^2 d \rho(0)^2 \tag{S53}$$

$$\sum_{i<j} \mathrm{Cov}(X_i, X_j)^2 \leq K^2 d \sum_{n=1}^{d-1} \rho(n)^2 \leq K^2 d \rho_\infty \max_{1 \leq n \leq d-1} \rho(n) \tag{S54}$$

Since $\sum_{n=0}^{+\infty} \rho(n) \leq \rho_\infty < \infty$, $\{\rho(n)\}_{n \in \mathbb{N}}$ converges to 0 and is thus bounded, so $\max_{1 \leq n \leq d-1} \rho(n) < \infty$. We then use (S53) and (S54) in the definition of $\beta_2(\mu_d)$, and $\beta_1(\mu_d) \leq \beta_2(\mu_d)$, to derive the upper-bound below for any $d \in \mathbb{N}^*$.

$$\beta_1(\mu_d), \beta_2(\mu_d) \leq d^{1/2} K \big\{ \rho(0)^2 + 2\rho_\infty \max_{1 \leq n \leq d-1} \rho(n) \big\}^{1/2} \tag{S55}$$

□

S3 Setup for synthetic experiments

We explain in more details the setup for the synthetic experiments discussed in Section 4, specifically the procedure to generate data.

For $d \in \mathbb{N}^*$, we generate $n = 10^4$ i.i.d. realizations of two random variables in $\mathbb{R}^d$, denoted by $X_{1:d} = \{X_j\}_{j=1}^d$ and $Y_{1:d} = \{Y_j\}_{j=1}^d$ and respectively distributed from μ_d, $\nu_d \in \mathcal{P}_2(\mathbb{R}^d)$. The n generated samples of $X_{1:d}$ and $Y_{1:d}$ are respectively denoted by $\{x^{(j)}\}_{j=1}^n$, $\{y^{(j)}\}_{j=1}^n$, with $x^{(j)}, y^{(j)} \in \mathbb{R}^d$ for $j \in \{1, \dots, n\}$. We approximate SW of order 2 between the empirical distributions of $\{x^{(j)}\}_{j=1}^n$ and $\{y^{(j)}\}_{j=1}^n$, given by $\hat{\mu}_{d,n} = n^{-1} \sum_{j=1}^n \delta_{x^{(j)}}$ and $\hat{\nu}_{d,n} = n^{-1} \sum_{j=1}^n \delta_{y^{(j)}}$ respectively. Note that in the main text (Section 4), these two distributions were denoted by μ_d, ν_d instead of $\hat{\mu}_{d,n}$, $\hat{\nu}_{d,n}$, to simplify the notation.

S3.1 Independent random variables

We first consider the setting described in Section S2.4, where $\mu_d = \mu^{(1)} \otimes \cdots \otimes \mu^{(d)}$ and $\nu_d = \nu^{(1)} \otimes \cdots \otimes \nu^{(d)}$ with $\mu^{(j)}, \nu^{(j)} \in \mathcal{P}_4(\mathbb{R})$ for $j \in \{1, \dots, d\}$. This means that $\{X_j\}_{j=1}^d$ and $\{Y_j\}_{j=1}^d$ are two sequences of d independent random variables. For each $j \in \{1, \dots, d\}$, $\mu^{(j)}$ (or $\nu^{(j)}$) refers to a Gaussian or a Gamma distribution, centered or not, as we explain hereafter.

Gaussian distributions (Figure 1(a)). For $j \in \{1, \dots, d\}$, $\mu^{(j)} = \mathrm{N}(m_1^{(j)}, \sigma_1^2)$ and $\nu^{(j)} = \mathrm{N}(m_2^{(j)}, \sigma_2^2)$, where $m_1^{(j)}$, $m_2^{(j)}$ are two i.i.d. samples from $\mathrm{N}(1, 1)$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 10$. Therefore, $\mu_d = \mathrm{N}(\mathbf{m}_1, \mathbf{I}_d)$ and $\nu_d = \mathrm{N}(\mathbf{m}_2, 10\,\mathbf{I}_d)$, where $\mathbf{I}_d$ denotes the identity matrix of size d, and $\mathbf{m}_1 = \{m_1^{(j)}\}_{j=1}^d$, $\mathbf{m}_2 = \{m_2^{(j)}\}_{j=1}^d \in \mathbb{R}^d$.

We prove that the SW of order 2 between such Gaussian distributions admits a closed-form expression: for any $\mathbf{m}_1, \mathbf{m}_2 \in \mathbb{R}^d$ and $\sigma_1^2, \sigma_2^2 > 0$,

$$\mathbf{SW}_2^2\{\mathrm{N}(\mathbf{m}_1, \sigma_1^2\, \mathbf{I}_d), \mathrm{N}(\mathbf{m}_2, \sigma_2^2\, \mathbf{I}_d)\} = \frac{1}{d}\|\mathbf{m}_1 - \mathbf{m}_2\|^2 + (\sigma_1 - \sigma_2)^2 \tag{S56}$$

Proof. First, given the properties of affine transformations of Gaussian random variables, we know that for any $\theta \in \mathbb{S}^{d-1}$, $\mathbf{m} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ symmetric positive-definite, $\theta^\star_\sharp \mathrm{N}(\mathbf{m}, \Sigma)$ is the univariate Gaussian distribution $\mathrm{N}(\langle \theta, \mathbf{m}\rangle, \theta^\intercal \Sigma \theta)$.

Using this property in the definition of SW (4) and the fact that $\|\theta\| = 1$ for $\theta \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
&\mathbf{SW}_2^2\{\mathrm{N}(\mathbf{m}_1, \sigma_1^2\, \mathbf{I}_d), \mathrm{N}(\mathbf{m}_2, \sigma_2^2\, \mathbf{I}_d)\} \\
&= \int_{\mathbb{S}^{d-1}} \mathbf{W}_2^2\{\mathrm{N}(\langle \theta, \mathbf{m}_1\rangle, \sigma_1^2), \mathrm{N}(\langle \theta, \mathbf{m}_2\rangle, \sigma_2^2)\}\mathrm{d}\boldsymbol{\sigma}(\theta) &\text{(S57)}\\
&= \int_{\mathbb{S}^{d-1}} \big\{ \langle \theta, \mathbf{m}_1 - \mathbf{m}_2\rangle^2 + (\sigma_1 - \sigma_2)^2\big\}\mathrm{d}\boldsymbol{\sigma}(\theta)\,, &\text{(S58)}
\end{aligned}$$

where (S58) results from the closed-form solution of the Wasserstein distance of order 2 between Gaussian distributions (2). Besides, by definition of the Euclidean inner-product, for any $\theta \in \mathbb{S}^{d-1}$,

$$\langle \theta, \mathbf{m}_1 - \mathbf{m}_2\rangle^2 = \big(\theta^\intercal(\mathbf{m}_1 - \mathbf{m}_2)\big)^2 = (\mathbf{m}_1 - \mathbf{m}_2)^\intercal \theta\theta^\intercal(\mathbf{m}_1 - \mathbf{m}_2)\,. \tag{S59}$$

We can thus rewrite (S58) to obtain

$$\begin{aligned}
&\mathbf{SW}_2^2\{\mathrm{N}(\mathbf{m}_1, \sigma_1^2\, \mathbf{I}_d), \mathrm{N}(\mathbf{m}_2, \sigma_2^2\, \mathbf{I}_d)\} \\
&= (\mathbf{m}_1 - \mathbf{m}_2)^\intercal \Big\{ \int_{\mathbb{S}^{d-1}} \theta\theta^\intercal \mathrm{d}\boldsymbol{\sigma}(\theta)\Big\}(\mathbf{m}_1 - \mathbf{m}_2) + (\sigma_1 - \sigma_2)^2\,. &\text{(S60)}
\end{aligned}$$

We conclude by using the fact that $\int_{\mathbb{S}^{d-1}} \theta\theta^\intercal \mathrm{d}\boldsymbol{\sigma}(\theta) = (1/d)\mathbf{I}_d$.

□

Gamma distributions (Figure 1(a)). Denote by $\Gamma(k, s)$ the Gamma distribution with shape parameter $k > 0$ and scale $s > 0$. For $j \in \{1, \ldots, d\}$, $\mu^{(j)} = \Gamma(k_1^{(j)}, s_1)$ and $\nu^{(j)} = \Gamma(k_2^{(j)}, s_2)$, where $k_1^{(j)}$ (respectively, $k_2^{(j)}$) is drawn from the uniform distribution over $[1, 5)$ (respectively, over $[5, 10)$), $s_1 = 2$ and $s_2 = 3$.

Centered (Gaussian or Gamma) distributions (Figures 1(b) and 2). We first generate $\{x^{(j)}\}_{j=1}^n$, $\{y^{(j)}\}_{j=1}^n$ using the Gaussian (or Gamma) distributions described in the two paragraphs above. Then, we center the data: for $j \in \{1, \ldots, n\}$, $\bar{x}^{(j)} = x^{(j)} - n^{-1}\sum_{i=1}^n x^{(i)}$ and $\bar{y}^{(j)} = y^{(j)} - n^{-1}\sum_{i=1}^n y^{(i)}$. The two distributions that we compare with SW, referred to as $\bar{\mu}_d, \bar{\nu}_d$ in Section 4, correspond to the empirical distributions of the centered datasets $\{\bar{x}^{(j)}\}_{j=1}^n$, $\{\bar{y}^{(j)}\}_{j=1}^n$, which can be denoted by $\bar{\mu}_{d,n}$ and $\bar{\nu}_{d,n}$.

We prove in the next proposition that our theoretical bounds derived in Section S2.4 can be improved for centered Gaussian distributions: in this setting, the expected approximation error is upper-bounded by a term in $d^{-1/2}$, which is consistent with the slope observed in Figure 1(b).

Proposition S3. *For $d \in \mathbb{N}^*$, let $\mu_d = \mathrm{N}(\mathbf{m}_1, \sigma_1^2\mathbf{I}_d)$ and $\nu_d = \mathrm{N}(\mathbf{m}_2, \sigma_2^2\mathbf{I}_d)$, and denote by $\bar{\mu}_d$, $\bar{\nu}_d$ their centered versions, i.e. $\bar{\mu}_d = \mathrm{N}(\mathbf{0}, \sigma_1^2\mathbf{I}_d)$ and $\bar{\nu}_d = \mathrm{N}(\mathbf{0}, \sigma_2^2\mathbf{I}_d)$. Consider the empirical distributions $\bar{\mu}_{d,n}$, $\bar{\nu}_{d,n}$ given by*

$$\bar{\mu}_{d,n} = (1/n)\sum_{j=1}^n \delta_{(X_{1:d}^{(j)} - \bar{X}_{1:d})}\,, \quad \bar{\nu}_{d,n} = (1/n)\sum_{j=1}^n \delta_{(Y_{1:d}^{(j)} - \bar{Y}_{1:d})}\,, \tag{S61}$$

where $\{X_{1:d}^{(j)}\}_{j=1}^n$ (respectively, $\{Y_{1:d}^{(j)}\}_{j=1}^n$) is a sequence of n random variables i.i.d. from μ_d (respectively, from ν_d), $\bar{X}_{1:d} = n^{-1}\sum_{j=1}^n X_{1:d}^{(j)}$, and $\bar{Y}_{1:d} = n^{-1}\sum_{j=1}^n Y_{1:d}^{(j)}$. Then,

$$\mathbb{E}\big|\mathbf{SW}_2(\bar{\mu}_d, \bar{\nu}_d) - \mathbf{W}_2\{\mathrm{N}(0, d^{-1}\mathtt{m}_2(\bar{\mu}_{d,n})), \mathrm{N}(0, d^{-1}\mathtt{m}_2(\bar{\nu}_{d,n}))\}\big| \leq \frac{\sigma_1 + \sigma_2}{(2dn)^{1/2}} + \mathcal{O}\left(\frac{1}{dn}\right)\,,$$

where $\mathbb{E}$ *is the expectation with respect to* $\{X_{1:d}^{(j)}\}_{j=1}^n$ *and* $\{Y_{1:d}^{(j)}\}_{j=1}^n$*, and* $\mathtt{m}_2(\bar{\mu}_{d,n}), \mathtt{m}_2(\bar{\nu}_{d,n})$ *are defined in* (8)*, i.e.* $\mathtt{m}_2(\bar{\mu}_{d,n}) = n^{-1}\sum_{j=1}^n \|X_{1:d}^{(j)} - \bar{X}_{1:d}\|^2$ *and* $\mathtt{m}_2(\bar{\nu}_{d,n}) = n^{-1}\sum_{j=1}^n \|Y_{1:d}^{(j)} - \bar{Y}_{1:d}\|^2$.

Proof of Proposition S3. Given the closed-form expressions in (S56) and (2), we have

$$\begin{aligned}
&\mathbb{E}\big|\mathbf{SW}_2(\bar{\mu}_d, \bar{\nu}_d) - \mathbf{W}_2\{\mathrm{N}(0, d^{-1}\mathtt{m}_2(\bar{\mu}_{d,n})), \mathrm{N}(0, d^{-1}\mathtt{m}_2(\bar{\nu}_{d,n}))\}\big| \\
&= \mathbb{E}\big||\sigma_1 - \sigma_2| - |d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2} - d^{-1/2}\mathtt{m}_2(\bar{\nu}_{d,n})^{1/2}|\big| \\
&\leq \mathbb{E}\big|\sigma_1 - \sigma_2 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2} + d^{-1/2}\mathtt{m}_2(\bar{\nu}_{d,n})^{1/2}\big| \qquad \text{(S62)} \\
&\leq \mathbb{E}\big|\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2}\big| + \mathbb{E}\big|\sigma_2 - d^{-1/2}\mathtt{m}_2(\bar{\nu}_{d,n})^{1/2}\big| \,. \qquad \text{(S63)}
\end{aligned}$$

where (S62) results from applying the reverse triangle inequality, and (S63) follows from the triangle inequality and the linearity of the expectation.

The final result follows from bounding $\mathbb{E}\big|\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2}\big|$ and $\mathbb{E}\big|\sigma_2 - d^{-1/2}\mathtt{m}_2(\bar{\nu}_{d,n})^{1/2}\big|$ from above. First, by the Cauchy–Schwarz inequality,

$$\mathbb{E}\big|\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2}\big| \leq \Big\{\mathbb{E}\big[(\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2})^2\big]\Big\}^{1/2} , \qquad \text{(S64)}$$

with

$$\mathbb{E}\big[(\sigma_1 - d^{-1/2}\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2})^2\big] = \sigma_1^2 - 2\sigma_1 d^{-1/2}\mathbb{E}[\mathtt{m}_2(\bar{\mu}_{d,n})^{1/2}] + \mathbb{E}[d^{-1}\mathtt{m}_2(\bar{\mu}_{d,n})] \,. \qquad \text{(S65)}$$

Consider the random variable defined as $Z = \sqrt{\sum_{i=1}^{dn}\{(X_i - \bar{X})^2/\sigma_1^2\}}$, where $\{X_i\}_{i=1}^{dn}$ are i.i.d. from $\mathrm{N}(0, \sigma_1^2)$ and $\bar{X} = (dn)^{-1}\sum_{i=1}^{dn} X_i$. Then, by Cochran's theorem, Z is distributed from the chi distribution with $dn - 1$ degrees of freedom. This implies that,

$$\begin{aligned}
&\mathbb{E}[d^{-1}\mathtt{m}_2(\bar{\mu}_{d,n})] = \sigma_1^2\,\frac{dn-1}{dn} \,, \\
&\mathbb{E}[Z] = \sqrt{2}\,\frac{\Gamma(dn/2)}{\Gamma((dn-1)/2)} = \sqrt{dn-1}\left[1 - \frac{1}{4dn} + \mathcal{O}\left(\frac{1}{(dn)^2}\right)\right] .
\end{aligned}$$

Hence, (S65) boils down to

$$\mathbb{E}\big[(\sigma_1 - d^{-1/2}\mathtt{m}_2(\hat{\mu}_{d,n})^{1/2})^2\big] = \sigma_1^2\left[2 - \frac{1}{dn} - 2\left(1 - \frac{1}{dn}\right)^{1/2}\left\{1 - \frac{1}{4dn} + \mathcal{O}\left(\frac{1}{(dn)^2}\right)\right\}\right] . \qquad \text{(S66)}$$

Besides, we know that

$$\left(1 - \frac{1}{dn}\right)^{1/2} = 1 - \frac{1}{2dn} + \mathcal{O}\left(\frac{1}{(dn)^2}\right) , \qquad \text{(S67)}$$

so we can write (S66) as

$$\mathbb{E}\big[(\sigma_1 - d^{-1/2}\mathtt{m}_2(\hat{\mu}_{d,n})^{1/2})^2\big] = \frac{\sigma_1^2}{2dn} + \mathcal{O}\left(\frac{1}{(dn)^2}\right) . \qquad \text{(S68)}$$

By plugging (S68) in (S64), we conclude that

$$\mathbb{E}\big|\sigma_1 - d^{-1/2}\mathtt{m}_2(\hat{\mu}_{d,n})^{1/2}\big| \leq \frac{\sigma_1}{(2dn)^{1/2}} + \mathcal{O}\left(\frac{1}{dn}\right) . \qquad \text{(S69)}$$

We can use the same reasoning to prove that

$$\mathbb{E}\big|\sigma_2 - d^{-1/2}\mathtt{m}_2(\hat{\nu}_{d,n})^{1/2}\big| \leq \frac{\sigma_2}{(2dn)^{1/2}} + \mathcal{O}\left(\frac{1}{dn}\right) , \qquad \text{(S70)}$$

and we use (S69) and (S70) to bound (S63), which concludes the proof.

□

S3.2 Autoregressive processes

Let $(X_j)_{j\in\mathbb{N}^*}$ be an autoregressive process of order 1 defined as $X_1 = \varepsilon_1$ and for $t \in \mathbb{N}^*$, $t > 1$, $X_t = \alpha X_{t-1} + \varepsilon_t$, where $\alpha \in [0, 1)$ and $(\varepsilon_j)_{j\in\mathbb{N}^*}$ is a sequence of i.i.d. real random variables such that $\mathbb{E}[\varepsilon_1] = 0$ and $\mathbb{E}[\varepsilon_1^2] < \infty$.

For $d \in \mathbb{N}^*$ and $B = 10^4$, we generate n realizations of $\{X_j\}_{j=B+1}^{B+d} \in \mathbb{R}^d$ using the aforementionned recursion. This gives us our first dataset $\{x^{(j)}\}_{j=1}^n$, where $x^{(j)} \in \mathbb{R}^d$ for $j \in \{1, \dots, n\}$. Note that the first B steps of the process are discarded in order to reach its stationary regime (which exists since $|\alpha| < 1$), and thus meet the weak dependence condition [4]. We repeat the same procedure to obtain the second dataset, $\{y^{(j)}\}_{j=1}^n$. Since the two datasets are generated using the same AR(1) model, μ_d and ν_d are the same distribution, so the exact value of SW is zero.

We conducted our experiments on two types of AR(1) processes, which differ from the distribution used to draw n i.i.d. samples of $\{\varepsilon_j\}_{j=1}^{B+d}$. The two settings are specified below.

Gaussian noise (Figure 1(c)). For $j \in \{1, \dots, B + d\}$, $\varepsilon_j \sim \mathrm{N}(0, 1)$.

Student's t noise (Figure 1(d)). Denote by $t(r)$ the Student's t distribution with $r > 0$ degrees of freedom. For $j \in \{1, \dots, B + d\}$, $\varepsilon_j \sim t(10)$.

S3.3 Computing infrastructure

The experiment comparing the computation time of our methodology against Monte Carlo estimation (Figure 2) was conducted on a daily-use laptop equipped with 8 $\times$ Intel Core i7-8650U CPU @ 1.90GHz, 16GB of RAM.

S4 Experimental details for image generation

Architecture. For each model (SWG, reg-SWG or reg-det-SWG), we used the architectures described in [5]: the "Conv & Deconv" generator and discriminator in [5, Section D] for MNIST, and DCGAN [6] with layernorm for both the generator and discriminator for CelebA.

Data preprocessing. For MNIST, we do not apply any specific preprocessing. For CelebA, each image is cropped at the center and resized to 140×140 (using the notation width $\times$ height, both in pixels), then resized to 64×64.

Optimization. For each model, we used the same optimization routine as in [5]: one training iteration consists in performing one update for the generator then one update for the discriminator, both with the default setting of Adam [7] (*i.e.* $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$). The values of other important hyperparameters are given in Table S1.

Dataset	Batch size	Learning rate	Total number of epochs
MNIST	512	5×10^{-4}	200
CelebA	64	1×10^{-4}	20

Table S1: Hyperparameters used when training each model.

Regularization parameters. For reg-SWG and reg-det-SWG, we tuned the regularization coefficients (λ_1, λ_2) via cross-validation: we trained the models for $\lambda_1 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and $\lambda_2 \in \{0, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and selected the tuple that minimizes the average FID over 5 runs.

Computing infrastructure. The FID and computation times on GPU reported in Table 1 (columns 'FID', 'T_{SW}, GPU' and 'T_{tot}, GPU') were obtained by training each model on a computer cluster equipped with 3 GPUs (NVIDIA Tesla V100-PCIE-32GB and 2 $\times$ NVIDIA Tesla V100-PCIE-16GB) for CelebA, and with 1 GPU (NVIDIA GP100GL, Tesla P100 PCIe 16GB) for MNIST. To obtain the computation times on CPU (Table 1, columns 'T_{SW}, CPU' and 'T_{tot}, CPU'), we used a workstation equipped with 24 $\times$ Intel Xeon CPU E5-2620 v3 @ 2.40GHz.

References

[1] Galen Reeves. Conditional central limit theorems for Gaussian projections. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3045–3049, 2017.

[2] Greg Huber. Gamma function derivation of n-sphere volumes. *The American Mathematical Monthly*, 89(5):301–302, 1982.

[3] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[4] Paul Doukhan and Michael H. Neumann. The notion of Ψ-weak dependence and its applications to bootstrapping time series. *Probability Surveys*, 5(none):146 – 168, 2008.

[5] Ishan Deshpande, Ziyu Zhang, and Alexander Schwing. Generative Modeling using the Sliced Wasserstein Distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

[6] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.