VERTICAL SLICED-WASSERSTEIN DISTANCES FOR META-MEASURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Many modern learning problems require comparing probability distributions whose individual observations are themselves probability measures, naturally giving rise to *meta-measures* in $\mathcal{P}(\mathcal{P}(\Omega))$. However, optimal transport distances on this space can be prohibitively expensive to evaluate. We introduce the *vertical sliced Wasserstein* (VSW) distance, a scalable metric on $\mathcal{P}(\mathcal{P}(\Omega))$. Rather than slicing along the transport geometry of $\mathcal{P}(\Omega)$, VSW projects each inner measure using continuous test functions and compares the resulting one-dimensional distributions with one-dimensional Wasserstein distance. We prove that VSW is a valid metric on $\mathcal{P}(\mathcal{P}(\Omega))$, metrizes weak convergence, and admits dimension-independent statistical convergence guarantees. Empirically, VSW provides high accuracy and efficiency compared with existing distances for meta-measures. We further derive a practical particle gradient for the empirical VSW_2^2 objective and demonstrate its use for gradient-based transport of meta-measures. VSW provides a simple and scalable alternative for comparing and optimizing distributions of distributions.

1 INTRODUCTION AND RELATED WORK

Many learning problems naturally involve *distributions of distributions* (DoDs): a dataset can be viewed as a distribution over class-conditional distributions (Alvarez-Melis & Fusi, 2020), a collection of shapes as a distribution over point clouds where each point cloud is itself an empirical distribution (Haviv et al., 2025), and documents, cell populations, or mixture models can similarly be represented as collections of probability measures (Yurochkin et al., 2019; Nguyen & Mueller, 2026). Formally, these objects are *meta-measures* $\mathbb{P} \in \mathcal{P}(\mathcal{P}(\Omega))$, i.e., probability measures whose observations are themselves probability measures on Ω. Comparing such objects requires accounting for variation both *within* each inner distribution and *across* the population of distributions.

Wasserstein distances on meta-measures. A natural metric for this setting is Wasserstein-over-Wasserstein (WoW), which solves an outer optimal transport (OT) problem using Wasserstein distance between inner measures as its ground cost. Related constructions appear in hierarchical OT (Yurochkin et al., 2019), generative modeling (Dukler et al., 2019), and the optimal transport dataset distance (OTDD) (Alvarez-Melis & Fusi, 2020). While geometrically appealing, WoW is expensive: comparing empirical meta-measures requires many inner OT problems and inherits the statistical and computational difficulties of Wasserstein distance at both levels (Dukler et al., 2019; Bonet et al., 2025b). Additionally, Bonet et al. (2025b) develop WoW gradient flows and introduce tractable MMD objectives with Sliced-Wasserstein kernels.

Sliced and scalable meta-measure distances. Recent methods therefore seek scalable alternatives through slicing. Double-Sliced Wasserstein (DSW) (Piening & Beinert, 2026) uses one-dimensional Wasserstein projections and quantile representations, Busemann-based approaches slice along geodesic directions of Wasserstein space (Bonet et al., 2025a), and s-OTDD uses moment-based projections for dataset comparison (Nguyen et al., 2025). Although substantially cheaper than exact WoW, these approaches still construct their projections from the transport geometry of the inner measures or surrogates thereof.

We introduce *Vertical Sliced Wasserstein* (VSW), a complementary approach that slices a measure through its action on continuous test functions. For a test function $f \in C(\Omega)$, an inner measure μ is represented by the scalar $P_f(\mu) = \int_\Omega f \, \mathrm{d}\mu$. Applying this projection to every inner measure

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

converts a meta-measure into a one-dimensional distribution. Then, one can compare efficiently $\mathbb{P}$ and $\mathbb{Q}$ in $\mathcal{P}(\mathcal{P}(\Omega))$ using the one-dimensional p-Wasserstein distance W_p by $\mathrm{VSW}_p^p(\mathbb{P},\mathbb{Q}) := \mathbb{E}_{f\sim\sigma}[\mathrm{W}_p^p\big((P_f)_\#\mathbb{P},(P_f)_\#\mathbb{Q}\big)]$, where σ is the law of a Gaussian process with universal covariance kernel κ. We call this construction *vertical* because it acts through the dual representation of each measure, in contrast to methods that slice along its transport geometry.

This simple change has important practical consequences. Each slice compresses an entire inner distribution to one scalar, requiring only test-function evaluation and averaging before a single one-dimensional OT computation across the outer distributions. VSW therefore avoids inner OT, per-distribution quantile representations, and Gaussian surrogates. At the same time, the family of test-function projections is rich enough to distinguish meta-measures: we prove that VSW_p is a metric and metrizes weak convergence. For $p = 2$, VSW is additionally Hilbertian and closely connected to kernel mean embeddings, while retaining the outer Wasserstein comparison that distinguishes distributions beyond their mean measures.

Empirically, this combination yields an accurate and efficient meta-measure distance. Across dataset comparison, point-cloud and image-patch distributions, and shape classification, VSW closely tracks transport-based reference distances while consistently reducing runtime. We further derive an explicit gradient formulation for the empirical VSW_2^2 objective, supporting parametric generative modeling. Our main contributions are:

1. We introduce **Vertical Sliced Wasserstein (VSW)**, a new sliced distance for meta-measures based on test-function projections, and prove that it defines a metric that metrizes weak convergence.
2. We establish statistical and computational scalability, including dimension-independent two-stage convergence rates and efficient Monte Carlo estimation without inner OT.
3. We show that VSW_2 is Hilbertian, induces universal characteristic kernels on meta-measures, and admits, under discretization, an explicit gradient for optimization and generative modeling.
4. We provide broad empirical validation across dataset comparison, point clouds, image patches, shape classification, discrete particle gradient flows, and hierarchical generation, where VSW is fast while maintaining strong agreement with transport-based baselines.

The closest computational baseline is DSW (Piening & Beinert, 2026), which first projects multivariate inner measures to one dimension, represents the resulting measures through quantile functions in $L^2([0,1])$, and then applies functional slicing with Gaussian-process directions. VSW instead applies test functions directly to the original inner measures, reducing each to a single scalar per slice and avoiding both inner quantile construction and an additional Euclidean slicing stage. A related test-functional approach is the hierarchical IPM of Catalano & Lavenant (2024), which, unlike VSW, requires optimization over a class of test functions.

2 Background and standing assumptions

Spaces. Throughout, (Ω,ρ) is a compact metric space. When gradients on Ω are needed (Sections 3.2–3.4) we take $\Omega \subset \mathbb{R}^d$. $C(\Omega)$ denotes the continuous real functions with the sup norm $\|\cdot\|_\infty$. Since Ω is compact, $(C(\Omega),\|\cdot\|_\infty)$ is a separable Banach space. By the Riesz–Markov representation theorem its dual is the space of finite signed Radon measures $\mathcal{M}(\Omega)$ with the total variation norm, and $\mathcal{P}(\Omega)$ is a convex subset of the unit ball of $\mathcal{M}(\Omega)$. The weak (narrow) topology on $\mathcal{P}(\Omega)$ is the restriction of the weak* topology of $\mathcal{M}(\Omega) = C(\Omega)^*$.

By Prokhorov's theorem, $\mathcal{P}(\Omega)$ is weakly compact and it is metrizable because $C(\Omega)$ is separable. Its Borel σ-algebra, $\mathcal{B}(\mathcal{P}(\Omega))$, is generated by the maps $P_f : \mathcal{P}(\Omega) \to \mathbb{R}$

$$P_f : \mu \mapsto \langle f,\mu\rangle := \int_\Omega f \,\mathrm{d}\mu, \qquad f \in C(\Omega), \tag{1}$$

(Kallenberg, 2017, Lem. 4.5 and Thm. 4.2). Since $\mathcal{P}(\Omega)$ is itself a compact metric space, the same reasoning applies one level up: $\mathcal{P}(\mathcal{P}(\Omega))$, the space of Borel probability measures on $\mathcal{P}(\Omega)$, is compact and metrizable in its weak topology. For $\mathbb{P} \in \mathcal{P}(\mathcal{P}(\Omega))$ we write $\mu \sim \mathbb{P}$ for a random inner measure.

Note the two elementary but crucial algebraic properties of *vertical projections* defined by (1):

Lemma 2.1. *For all $f, g \in C(\Omega)$, $a, b \in \mathbb{R}$, $\mu, \mu_0, \mu_1 \in \mathcal{P}(\Omega)$, and $t \in [0,1]$,*

$$P_{af+bg}(\mu) = aP_f(\mu) + bP_g(\mu), \qquad P_f((1-t)\mu_0 + t\mu_1) = (1-t)P_f(\mu_0) + tP_f(\mu_1).$$

Thus, $P_f(\mu)$ is linear in f and affine in μ along mixtures.

Optimal transport between one-dimensional measures. For $p \geq 1$ and $\alpha, \beta \in \mathcal{P}(\mathbb{R})$, with finite p-moment, $1 \leq p < \infty$, let F_α, F_β denote their corresponding Cumulative Distribution Functions (CDFs), and $q_\alpha = F_\alpha^{-1}, q_\beta = F_\beta^{-1}$ be their quantile functions. The optimal transport cost is:

$$\mathrm{W}_p^p(\alpha, \beta) = \int_0^1 \left|q_\alpha(u) - q_\beta(u)\right|^p \mathrm{d}u. \tag{2}$$

We use $\mathrm{W}_p^p(\alpha,\beta) \leq (\operatorname{diam}\operatorname{supp})^{p-1}\mathrm{W}_1(\alpha,\beta)$ for compactly supported α, β, and the empirical rate: if α is supported in an interval of length $|I|$ and $\widehat{\alpha}_N$ is the empirical measure of N i.i.d. samples, then

$$\mathbb{E}\,\mathrm{W}_1(\widehat{\alpha}_N, \alpha) \;\leq\; \tfrac{1}{\sqrt{N}} \int_{\mathbb{R}} \sqrt{F_\alpha(1-F_\alpha)} \;\leq\; \tfrac{|I|}{2\sqrt{N}}, \tag{3}$$

where the first inequality follows from Bobkov & Ledoux (2019, Thm 3.2).

Gaussian slicing measure. We write σ for a Borel probability measure on $C(\Omega)$. A slicing direction is a realization $f \sim \sigma$, and L iid slices are $f_1, \ldots, f_L \sim \sigma$. This is akin to the sliced-Wasserstein distance with Gaussian projections introduced by Nadjahi et al. (2021), but for meta-measures.

Assumption 1. *We consider a continuous, symmetric, positive definite kernel and universal "slicing kernel" $\kappa : \Omega \times \Omega \to \mathbb{R}$, where universal means that the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\kappa$ associated to κ is dense in $(C(\Omega), \|\cdot\|_\infty)$, and such that the centered Gaussian process with covariance kernel κ, $\mathrm{GP}(0,\kappa)$, admits a version with almost surely continuous sample paths.*

Under the above assumption, we consider the model where $f \sim \sigma = \mathrm{GP}(0,\kappa)$. For every finite collection $x_1, \ldots, x_n \in \Omega$, the random vector $(f(x_1), \ldots, f(x_n))$ has law $\mathcal{N}(0, K)$, where $K_{ij} = \mathbb{E}_{f\sim\mathrm{GP}(0,\kappa)}[f(x_i)f(x_j)] = \kappa(x_i, x_j)$. See extra details in Appendix A.

Wasserstein-over-Wasserstein. For a Borel metric D on $\mathcal{P}(\Omega)$, $\mathbb{W}_{p,D}^p(\mathbb{P},\mathbb{Q}) := \inf_{\Gamma\in\Pi(\mathbb{P},\mathbb{Q})} \int D(\mu,\nu)^p \,\mathrm{d}\Gamma(\mu,\nu)$, with $\Pi(\mathbb{P},\mathbb{Q})$ the set of couplings between $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{P}(\Omega))$. When $D = \mathrm{W}_p$, it recovers the standard WoW distance, which we simply denote by $\mathbb{W}_p$. Additional details of this section are available in Appendix A.

3 Our method: The Vertical Sliced Wasserstein distance

Definition 3.1. *For $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{P}(\Omega))$, $p \in [1,\infty)$, and $\sigma = \mathrm{GP}(0,\kappa)$ the Gaussian slicing measure under Assumption 1, we define the Vertical Sliced Wasserstein discrepancy by*

$$\mathrm{VSW}_p(\mathbb{P},\mathbb{Q}) \;:=\; \Big(\int_{C(\Omega)} \mathrm{W}_p^p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big)\,\mathrm{d}\sigma(f)\Big)^{1/p}, \tag{4}$$

where, for each fixed $f \in C(\Omega)$, the pushforward $(P_f)_\#\mathbb{P} \in \mathcal{P}(\mathbb{R})$ is the law of the scalar vertical slice $P_f(\mu)$ when $\mu \sim \mathbb{P}$ defined precisely by $(P_f)_\#\mathbb{P}(B) = \mathbb{P}(\{\mu \in \mathcal{P}(\Omega) : P_f(\mu) \in B\})$ for all $B \in \mathcal{B}(\mathbb{R})$, which is supported on $[-\|f\|_\infty, \|f\|_\infty]$ (similarly for $(P_f)_\#\mathbb{Q}$).

The *"vertical"* viewpoint reflects the affine structure of the space of measures: the slicers (1) are affine along mixtures $(1-t)\mu_0 + t\mu_1$, as stated in Lemma 2.1. In contrast, *"horizontal"* or *transport geometry* describes the movement of mass within Ω, e.g., through displacement (McCann) interpolation $\mu_t = ((1-t)\mathrm{id} + tT)_\#\mu_0$, for an optimal Monge map T for the quadratic-cost. A related distinction between spatial transport and local mass variation appears in unbalanced OT (Chizat et al., 2018; Liero et al., 2018); here "vertical" specifically refers to affinity along probability mixtures.

3.1 Well-posedness, metricity, and topology of VSW

We show regularity of the slicing procedure to ensure that VSW_p in (4) is well-defined (Lemma 3.2). Moreover, VSW_p defines a true metric on $\mathcal{P}(\mathcal{P}(\Omega))$ (Theorem 3.3) which metrizes weak convergence (Proposition 3.5), and $(\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{VSW}_2)$ embeds isometrically into a Hilbert space (Theorem 3.6).

Lemma 3.2. *(i) The map $P : C(\Omega) \times \mathcal{P}(\Omega) \to \mathbb{R}$, $P(f, \mu) := \langle f, \mu \rangle$, is jointly continuous. (ii) For fixed $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{P}(\Omega))$ the map $f \mapsto \mathrm{W}_p\big((P_f)_\# \mathbb{P}, (P_f)_\# \mathbb{Q}\big)$ is 2-Lipschitz from $(C(\Omega), \|\cdot\|_\infty)$ to $\mathbb{R}$; in particular it is continuous, and the integral in (4) is well defined and finite.*

Theorem 3.3. VSW_p *is a metric on $\mathcal{P}(\mathcal{P}(\Omega))$ for every $p \in [1, \infty)$.*

Remark 3.4 (Comparison with horizontal slicing)**.** *Busemann projections are not yet known to distinguish meta-measures in general (Bonet et al., 2025a), while moment projections need additional assumptions ensuring that moments determine the measures (Nguyen et al., 2025). In our setting, vertical slicing distinguishes meta-measures whenever the kernel is universal, with no extra assumptions on the meta-measures themselves.*

Proposition 3.5. *For every $p \in [1, \infty)$, the metric VSW_p induces the weak topology on $\mathcal{P}(\mathcal{P}(\Omega))$. In particular $(\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{VSW}_p)$ is a compact metric space.*

Theorem 3.6. *Let λ be Lebesgue measure on $(0, 1)$. The map*

$$T : \mathcal{P}(\mathcal{P}(\Omega)) \to L^2(C(\Omega) \times (0,1), \sigma \otimes \lambda), \qquad T(\mathbb{P})(f, u) := q_{(P_f)_\# \mathbb{P}}(u),$$

is well defined and satisfies

$$\|T(\mathbb{P}) - T(\mathbb{Q})\|_{L^2(\sigma \otimes \lambda)} = \mathrm{VSW}_2(\mathbb{P}, \mathbb{Q}).$$

Hence $(\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{VSW}_2)$ embeds isometrically into a separable Hilbert space. Moreover, VSW_2^β is conditionally negative definite for $\beta \in (0, 2]$, so $\exp(-\gamma \mathrm{VSW}_2^\beta)$ is positive definite for all $\gamma > 0$.

Further details are provided in Appendices B and D, and all proofs of this section are in Appendix C.

3.2 Computation

Sampling slices: For translation-invariant κ on $\Omega \subset \mathbb{R}^d$ with spectral measure τ_κ (Bochner), an R-feature slice is

$$f_R(x) = \sqrt{\tfrac{2}{R}} \sum_{r=1}^{R} a_r \cos\big(\langle \omega_r, x \rangle + b_r\big), \qquad a_r \sim \mathcal{N}(0,1),\ \omega_r \sim \tau_\kappa,\ b_r \sim \mathrm{Unif}[0, 2\pi], \tag{5}$$

where we normalize $\kappa(x, x) = \kappa(0) = 1$ so that the spectral measure τ_κ is a probability distribution. Conditionally on (ω, b), f_R is a Gaussian process whose covariance converges uniformly to κ as $R \to \infty$ (Rahimi & Recht, 2007). Alternatively, a truncated Karhunen–Loève expansion may be used (see uniform approximation guarantees in Proposition D.2 and Lemma D.3). Gradients have the closed form $\nabla f_R(x) = -\sqrt{2/R} \sum_r a_r \sin(\langle \omega_r, x \rangle + b_r)\, \omega_r$.

Cost: The full VSW algorithm is shown in Algorithm 1. Let N denote the maximum number of inner measures in either empirical meta-measure, and n the maximum number of points per inner measure. Evaluating L independent R feature slices costs $O(LNnRd + LN \log N)$. In practice, we share the Fourier basis $(\omega_r, b_r)_{r=1}^R$ across slices and resample only the Gaussian coefficients (conditional on the basis, this estimates Gaussian-sliced Wasserstein between RFF mean embeddings in $\mathbb{R}^R$; see Appendix C for the scaling identity and pseudometric caveat). Computing and averaging the R basis features once for each inner measure, and reusing them across slices, reduces the cost to $O(NnRd + LNR + LN \log N)$. Unlike prior sliced WoW constructions, VSW requires no per inner-measure sorting or quantile representation. Its runtime advantages are verified empirically in Section 4.

3.3 Statistical properties

We consider the two-stage observation model of Szabó et al. (2016): inner measures $\mu_1, \dots, \mu_N \overset{\text{iid}}{\sim} \mathbb{P}$; for each i, points $x_{i1}, \dots, x_{in} \overset{\text{iid}}{\sim} \mu_i$ (conditionally independent given μ_i), empirical inner measures $\widehat{\mu}_i = \frac{1}{n} \sum_j \delta_{x_{ij}}$, and the plug-in meta-measures $\widehat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^N \delta_{\mu_i}$, $\widetilde{\mathbb{P}}_{N,n} = \frac{1}{N} \sum_{i=1}^N \delta_{\widehat{\mu}_i}$.

Theorem 3.7. *For every $p \in [1, \infty)$, there exists $C_p < \infty$, depending only on p (with $C_p = 1$ for $p \in [1, 2]$), such that*

$$\mathbb{E}\left[\mathrm{VSW}_p^p(\mathbb{P}, \widetilde{\mathbb{P}}_{N,n})\right] \le 2^{p-1}\Big(\tfrac{1}{2}\, 2^p \Lambda_p N^{-1/2} + C_p\, 2^p \Lambda_p n^{-p/2}\Big),$$

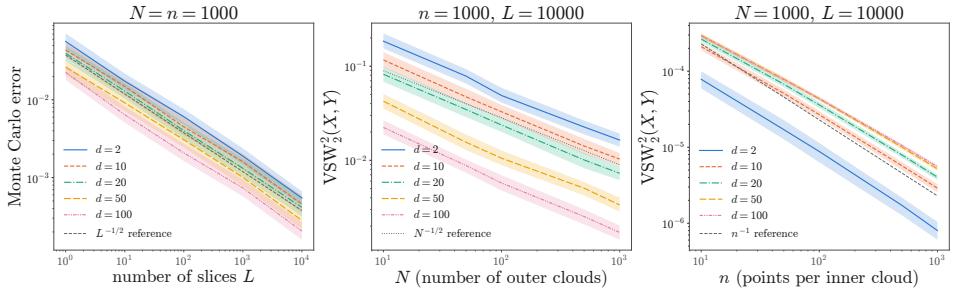


Figure 1: Statistical complexity analysis for empirical meta-measures X, Y. Setup in Appendix F. The outer law has three atoms, so the $N^{-1/2}$ reference for VSW_2^2 corresponds to the atomic-outer rate (Remark C.2). Left: distance error against a 10^5-slice reference. Right: X, Y share component labels to isolate inner sampling error.

Algorithm 1 $\widehat{\mathrm{VSW}}_p^p$ between empirical meta-measures

Require: clouds $\{X_i\}_{i\le N}$, $\{Y_k\}_{k\le M}$; kernel κ; slices L; features R; order p

1: **for** $l = 1, \dots, L$ **do**
2: sample f_l via (5) (or KL truncation)
3: $s_i \leftarrow \frac{1}{|X_i|}\sum_{x\in X_i} f_l(x)$ for all i; $t_k \leftarrow \frac{1}{|Y_k|}\sum_{y\in Y_k} f_l(y)$ for all k
4: $d_l \leftarrow \mathrm{W}_p^p\big(\frac{1}{N}\sum_i \delta_{s_i}, \frac{1}{M}\sum_k \delta_{t_k}\big)$ ▷ sort + quantile matching, closed form (2)
5: **end for**
6: **return** $\frac{1}{L}\sum_l d_l$

where $\Lambda_p = \mathbb{E}_\sigma \|f\|_\infty^p < \infty$. The rate exponents are dimension-free. Dependence on (Ω, κ) enters only through the sup-norm moment, which is finite but may grow with the size of Ω relative to the kernel bandwidth.

Remark 3.8. *Unlike $\mathbb{W}_{p,\mathrm{W}_p}$, which inherits the dimensional dependence of Wasserstein estimation at both levels (Fournier & Guillin, 2015; Catalano & Lavenant, 2024), Theorem 3.7, by Jensen, gives dimension-free rate exponents for $\mathbb{E}\,\mathrm{VSW}_p(\mathbb{P}, \widetilde{\mathbb{P}}_{N,n})$: $N^{-1/(2p)}$ at the outer level and $n^{-1/2}$ at the inner level, consistent with sliced-Wasserstein and two-stage sampling rates (Nadjahi et al., 2020; Szabó et al., 2016). For $p = 2$, the worst-case outer distance rate $N^{-1/4}$ is sharp and slower than HIPM's parametric $N^{-1/2}$ rate; it improves to $N^{-1/2}$ under the projected-density condition in Remark C.2.*

Finally, the Monte Carlo slice complexity reads as follows:

Proposition 3.9. *Let $f_1, \dots, f_L \overset{iid}{\sim} \sigma$ and $\widehat{\mathrm{VSW}}_L^p(\mathbb{P}, \mathbb{Q}) := \frac{1}{L}\sum_l \mathrm{W}_p^p\big((P_{f_l})_\#\mathbb{P}, (P_{f_l})_\#\mathbb{Q}\big)$. Then*

$$\mathbb{E}\Big|\widehat{\mathrm{VSW}}_L^p - \mathrm{VSW}_p^p\Big| \;\le\; \sqrt{\frac{\mathrm{Var}_\sigma\big[\mathrm{W}_p^p\big]}{L}} \;\le\; 2^p\sqrt{\Lambda_{2p}/L}.$$

Following Nadjahi et al. (2020), we examine empirical scaling on Gaussian-mixture meta-measures for $d \in \{2, 10, 20, 50, 100\}$ using finite Fourier features. Figure 1 shows an $L^{-1/2}$ trend for Monte Carlo *distance* error against a finite-slice reference, and VSW_2^2 trends of $N^{-1/2}$ and n^{-1} at the outer and inner levels, with nearly unchanged slopes across dimensions. This is an empirical scaling study: Theorem 3.7 assumes a compact domain, and Proposition 3.9 bounds the powered-distance estimator.

3.4 Gradient flows: Practical implementation

We optimize the empirical VSW_2^2 objective by moving the support points of its inner measures, thereby obtaining a finite-dimensional particle-level gradient flow. Throughout this calculation, the target and the slicing law are fixed. Fix a target $\mathbb{Q} \in \mathcal{P}(\mathcal{P}(\Omega))$ and define $\mathcal{F} : \mathcal{P}(\mathcal{P}(\Omega)) \to \mathbb{R}$ by $\mathcal{F}(\mathbb{P}) := \mathrm{VSW}_2^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{f\sim\sigma}[\mathrm{W}_2^2\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big)]$.

The discretization used in practice is as follows: Let $\mathbb{Q} = \frac{1}{N}\sum_{k=1}^{N} \delta_{\widehat{\nu}_k}$, $\widehat{\nu}_k = \frac{1}{m}\sum_l \delta_{y_{kl}}$, and $\mathbb{P}_X = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\mu}_i}$, $\widehat{\mu}_i = \frac{1}{n}\sum_j \delta_{x_{ij}}$, with $X = (x_{ij}) \in (\mathbb{R}^d)^{N\times n}$. For simplicity we take equal outer sizes. The quantile formulation also handles $N \neq M$. Define

$$\mathcal{F}(X) \;:=\; \mathrm{VSW}_2^2\big(\mathbb{P}_X, \mathbb{Q}\big) = \mathbb{E}_{f\sim\sigma}\Big[\frac{1}{N}\sum_{i=1}^{N}\big(s_{(i)}(f) - t_{(i)}(f)\big)^2\Big], \tag{6}$$

where $s_i(f) = \frac{1}{n}\sum_j f(x_{ij})$, $t_k(f) = \frac{1}{m}\sum_l f(y_{kl})$, and $(\cdot)_{(i)}$ denotes increasing rearrangement. Here the projected laws $\alpha_f = \frac{1}{N}\sum_i \delta_{s_i(f)}$ and $\beta_f = \frac{1}{N}\sum_k \delta_{t_k(f)}$ are atomic, so transport is determined by the sorting permutation rather than a continuous monotone map. Lemma C.5 ensures the required uniqueness.

Proposition 3.10. *Assume Assumptions 1–2, with the smoothness and moment conditions holding on a compact neighborhood of the particle configurations. If the source clouds and the target clouds are each pairwise distinct as measures, as in Lemma C.5, then $\mathcal{F}$ is differentiable at X, and*

$$\nabla_{x_{ij}}\mathcal{F}(X) \;=\; \frac{2}{Nn}\,\mathbb{E}_{f\sim\sigma}\Big[\big(s_i(f) - t_{\pi_f(i)}(f)\big)\,\nabla f(x_{ij})\Big]. \tag{7}$$

Consequently, the associated Euclidean particle gradient flow is $\dot{x}_{ij}(t) = -\nabla_{x_{ij}}\mathcal{F}(X(t))$, with the right-hand side given by (7). The proof and corresponding algorithm are given in Appendix C. In the ideal setting, replacing the expectation in Proposition 3.10 by an average over independent slices $f_1, \ldots, f_L \sim \sigma$ gives a Monte Carlo estimate of the particle gradient. In practice, we represent the slices using finitely many random Fourier features, yielding the corresponding finite-feature approximation of this gradient. When the particles are produced by a differentiable generator, the particle gradients can be backpropagated through the generator by the chain rule.

4 Experiments

We evaluate VSW across several tasks, demonstrating both performance and runtime advantages. For each experiment, we also optimize the baseline hyperparameters and implementations to ensure a fair comparison. Additional ablations and implementation details are provided in Appendices E and F.

Optimal Transport Dataset Distance. Dataset distances provide a model-agnostic way to quantify similarity between datasets, with applications such as transfer learning and dataset selection (Alvarez-Melis & Fusi, 2020; Liu et al., 2025). Following previous work (Nguyen et al., 2025; Piening & Beinert, 2026), we evaluate VSW as an efficient approximation to the optimal transport dataset distance (OTDD), comparing against s-OTDD (Nguyen et al., 2025) and DSW (Piening & Beinert, 2026), using OTDD (Alvarez-Melis & Fusi, 2020) as the reference. We construct random pairs of sub-datasets from MNIST (LeCun et al., 1998), FashionMNIST (Xiao et al., 2017), and CIFAR-10 (Krizhevsky & Hinton, 2009) and represent each dataset through its class-conditional empirical distributions. Following Piening & Beinert (2026), we use identical dataset pairs and feature representations across methods, remove the feature cost from OTDD so that it reduces to a Wasserstein-over-Wasserstein distance between class-conditional distributions, and use s-OTDD with its default implementation.

We report Pearson and Spearman correlations with OTDD while varying the number of slices. For DSW, the total projection budget is $S = PK$, where P and K denote the numbers of outer and inner projections, respectively. For each dataset, we use 30 pairs of disjoint sub-datasets drawn from the same training set. The two sub-datasets in each pair have equal size, with that size varying from $5{,}000$ to $10{,}000$ across pairs. Figure 2 shows that VSW achieves substantially stronger correlation in the low-slice regime across all three datasets and remains competitive at larger slice budgets under this protocol. Figure 3 compares runtime as the dataset size increases for matched budgets of $1\mathrm{k}$, $5\mathrm{k}$, and $10\mathrm{k}$ slices. VSW is consistently the fastest method across all three datasets and projection budgets under this resampling protocol.

Comparing Distributions of Point Clouds. Following the point-cloud distribution experiment of DSW (Piening & Beinert, 2026), we evaluate whether VSW captures changes in distributions of 3D point clouds from ModelNet-10 (Wu et al., 2015). Starting from a fixed reference distribution, we independently vary (i) the number of shapes in the target distribution, (ii) the amount of Gaussian

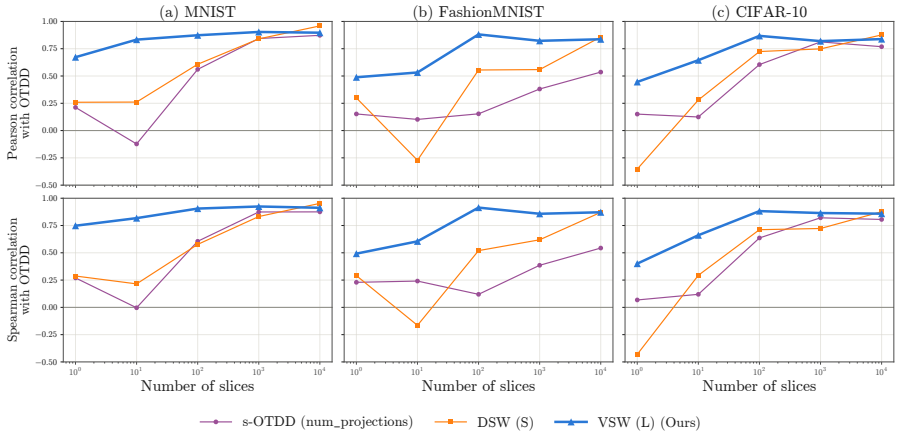


Figure 2: Pearson and Spearman correlations with exact OTDD versus the number of slices, computed over 30 pairs per dataset. Both sub-datasets in a pair are drawn from the same training set and have equal size (5,000–10,000 samples); sizes vary across pairs.

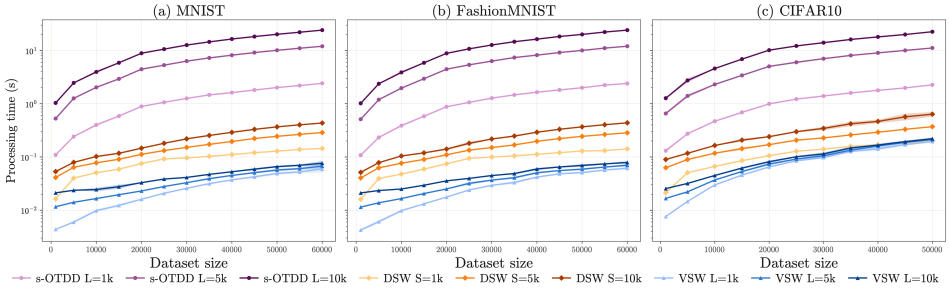


Figure 3: Runtime (s) vs the dataset size (each of 2 sets). VSW remains fastest among all methods.

Table 1: Wall-clock time of a single distance evaluation between the 662 ground-truth ShapeNet chairs and 662 generated samples (2048 points per cloud). † denotes a single run.

	OT-NNA	WoW	DSW	VSW (Ours)
runtime (s)	6077.65†	3094.22†	0.0709 ± 0.00017	$\mathbf{0.0125 \pm 0.00003}$

noise applied to the target points, and (iii) the number of points per target shape. We compare VSW with OT-NNA, exact WoW, and DSW. To facilitate direct comparison between methods with different scales, we report the normalized metrics together in Figure 4, with the reference configuration indicated by the dashed line.

VSW exhibits trends consistent with the OT-based baselines across all three perturbations. In particular, the distance decreases as the target contains more shapes, increases with stronger Gaussian perturbations, and remains comparatively stable once the point-cloud resolution is sufficiently high. These results indicate that VSW preserves the behavior of established distribution-comparison metrics while being substantially more efficient. We additionally report the wall-clock time in a realistic point cloud evaluation setting, comparing 662 ground-truth and 662 generated ShapeNet chairs (Chang et al., 2015) from PVD (Zhou et al., 2021), each containing 2048 points. The results are shown in Table 1. VSW requires only 0.0125 s per distance evaluation, compared with 0.0709 s for DSW, corresponding to approximately a $5.7\times$ reduction in runtime. OT-NNA and WoW are both substantially slower.

Comparing Image Distributions via Patch Distributions. Following DSW (Piening & Beinert, 2026), we evaluate VSW for comparing image distributions through their patch distributions. We use synthetic Perlin-noise textures and independently vary the *lacunarity* and *persistence* of the target distribution relative to a fixed reference. We compare VSW against DSW and exact Wasserstein-over-Wasserstein (WoW), and normalize each metric for visualization.

As shown in Figure 5, all three methods correctly attain their minimum near the reference parameters, while increasing as the target distribution deviates from the reference. VSW closely follows the trends

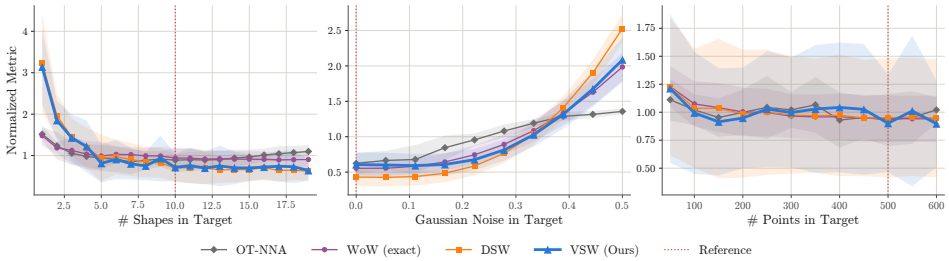


Figure 4: Point cloud comparison metric trends.

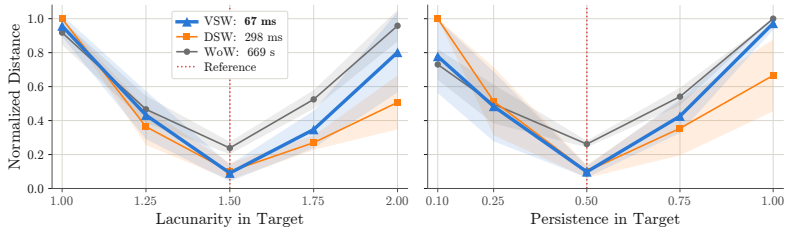


Figure 5: Comparing image distributions via patch distributions. VSW remains the fastest method, while achieving the same trend as baselines.

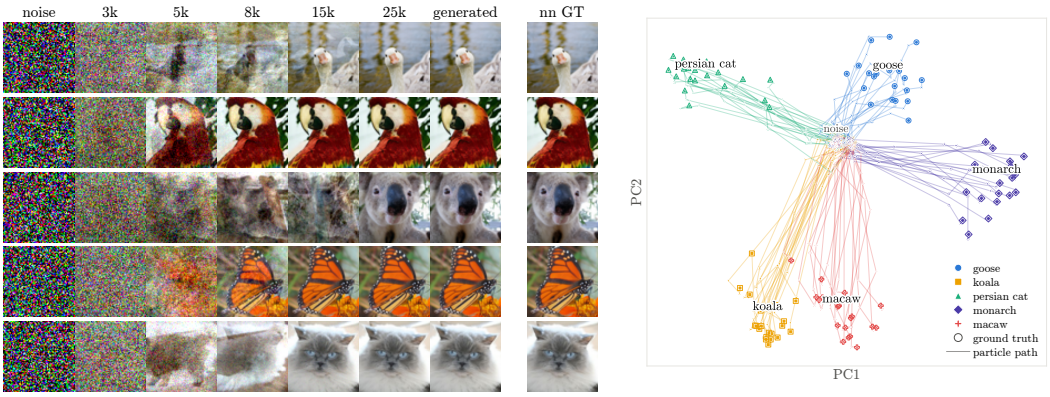


Figure 6: Particle optimization toward fixed ImageNet targets in pixel space **(Left)**, and in a 2D PCA projection of CLIP ViT-L/14 embeddings **(Right)**.

of DSW and WoW, indicating that it preserves meaningful sensitivity to changes in image-patch distributions. At the same time, VSW requires only 67 ms per comparison, compared with 298 ms for DSW and 669 s for exact WoW, making it approximately $4.4\times$ faster than DSW and nearly four orders of magnitude faster than exact WoW.

Gradient Flow on Natural Images. We use ImageNet (Deng et al., 2009) to evaluate the VSW particle gradient on a high-dimensional dataset. We optimize five clouds of twenty noise-initialized particles against a fixed target containing twenty training images from each of five classes. Figure 6 shows the particle trajectories in pixel space and in a two-dimensional PCA projection of CLIP ViT-L/14 embeddings (Radford et al., 2021), fitted to the target images. The particles progressively resemble target images, with projected trajectories moving toward the target clusters. Nearest-neighbor ground-truth images provide a qualitative reference for the final particles. These visualizations illustrate optimization toward the fixed empirical target.

Parametric Generative Modeling. We next evaluate VSW in a parametric generative modeling setting on two 2D hierarchical datasets: a checkerboard distribution and an eight-component Gaussian mixture. Each dataset is represented as a distribution of inner distributions, with an outer latent variable selecting the component and inner noise generating samples within it. We train a lightweight MLP using the corresponding objective. We compare VSW with DSW (Piening & Beinert, 2026) and MMDSW (Bonet et al., 2025b). As shown in Figure 7, VSW and MMDSW recover the main structure of both targets, and VSW closely matches the held-out distributions. VSW also reduces the Wasserstein error substantially faster in wall-clock time, reaching a low error earlier than DSW and MMDSW on both datasets.

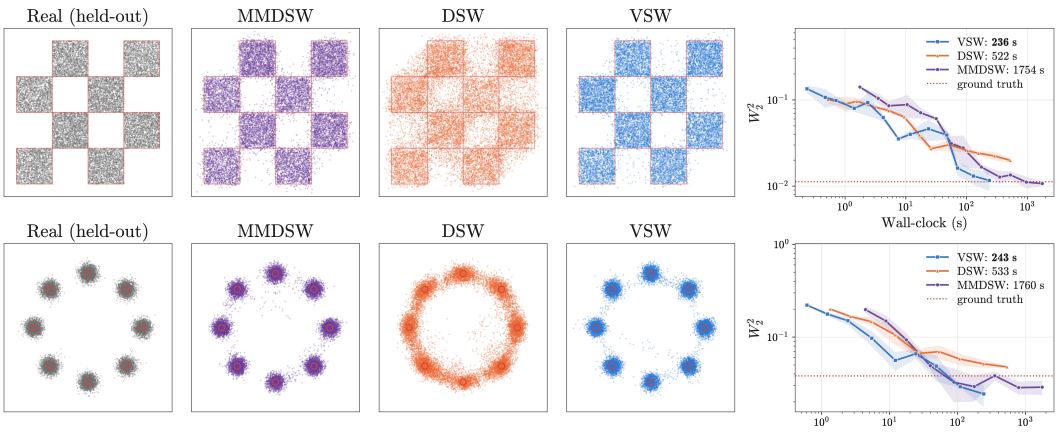


Figure 7: Parametric generation on Checkerboard and 8-Gaussians datasets.

Table 2: Shape classification with KNN (Acc., ↑) and runtime per distance computation (Time).

Distance	2D Shapes		Animals		FAUST-500		FAUST-1000		MNIST-2000	
	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)
VSW	97.7 ± 2.6	0.35	99.0 ± 1.3	0.34	39.0 ± 5.5	2.80	41.2 ± 6.0	3.76	96.5 ± 3.4	7.48
SQW	99.4 ± 1.2	0.93	99.3 ± 1.0	0.95	37.5 ± 5.5	6.51	40.8 ± 6.0	10.15	86.0 ± 4.3	18.46
TLB	100.0 ± 0.2	0.91	100.0 ± 0.0	0.97	36.7 ± 5.6	40.96	40.2 ± 6.0	163.06	95.5 ± 3.3	653.84
STLB	98.5 ± 2.0	0.99	99.0 ± 1.3	1.00	35.2 ± 5.5	10.01	38.4 ± 5.9	10.65	82.2 ± 3.8	18.36
AE	99.7 ± 0.8	0.13	97.9 ± 1.8	0.13	37.5 ± 5.7	2.63	42.0 ± 5.9	4.46	93.2 ± 4.0	19.79
GW	99.7 ± 0.6	3.1	100.0 ± 0.0	6.2	29.2 ± 4.4	346.3	33.0 ± 5.3	1824.7	— timeout —	

Shape Classification. Following the protocol of DSW (Piening & Beinert, 2026), we evaluate VSW on 2D Shapes (Beier et al., 2022), Animals (Sumner & Popović, 2004), FAUST-500/1000 (Bogo et al., 2014) (500 and 1000 vertices per shape), and MNIST-2000 (Piening & Beinert, 2026) using local distance distributions. For each distance, we precompute the pairwise distance matrix and perform 3-NN classification over 1000 random 25%/75% train/test splits, reporting mean accuracy and CPU time per distance. VSW, SQW, and STLB share one random projection basis across the distance matrix, while AE caches its per-shape energy self-terms. TLB and GW are deterministic and do not include such amortization.

Table 2 shows competitive mean accuracies across these datasets, with VSW consistently faster than the other sliced baselines. Compared with SQW, VSW is approximately 2.3–2.8× faster across all datasets. The advantage is especially clear for larger shapes, requiring only 3.76 ms on FAUST-1000 and 7.48 ms on MNIST-2000. VSW provides a favorable accuracy–runtime trade-off for this task.

5 Conclusion and Future Work

We introduced Vertical Sliced Wasserstein (VSW), a scalable distance for meta-measures based on slicing through continuous test-function projections rather than the transport geometry of the inner measures. We showed that VSW is a valid metric that metrizes weak convergence, and that VSW_2 admits a Hilbertian embedding, yielding characteristic kernels on meta-measures. We further established dimension-independent two-stage convergence rates and derived a practical particle gradient for optimization. Across dataset comparison, shape classification, point-cloud and image-distribution comparison, gradient flow and parametric generation, VSW provides competitive agreement with transport-based distances while substantially reducing computation.

Our analysis also leaves several limitations and open directions. Promising directions include extending parametric generative modeling beyond low-dimensional examples to higher-dimensional and structured domains, such as point cloud and image generation. For instance, a continuum WoW-flow interpretation (where, based on the differential geometry of Bonet et al. (2025b), VSW defines the objective, while $(\mathcal{P}(\mathcal{P}(\Omega)), \mathbb{W}_2)$ defines the geometry of the descent) together with a convergence of the discrete optimization schemes requires a separate theoretical analysis.

AI USE STATEMENT

Generative AI tools were used to assist with writing, editing, and polishing the manuscript, as well as with drafting and debugging code. All scientific and intellectual contributions, including the research ideas, proposed methodology, experimental design, analysis, and conclusions, were developed by the authors. The authors reviewed, verified, and, where necessary, revised all AI-assisted content and code, and take responsibility for the accuracy and integrity of the final work.

REPRODUCIBILITY STATEMENT

We provide detailed experimental and implementation settings in Appendix F, including dataset construction, hyperparameters, baseline configurations, evaluation protocols, and computational resources. The supplementary material includes the code used for our experiments. We plan to publicly release the complete codebase upon acceptance.

REFERENCES

Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer, 2007.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

Florian Beier, Robert Beinert, and Gabriele Steidl. On a linear Gromov–Wasserstein distance. *IEEE Transactions on Image Processing*, 31:7292–7305, 2022. doi: 10.1109/TIP.2022.3221286.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic analysis on semigroups: theory of positive definite and related functions*, volume 100. Springer, 1984.

Sergey G. Bobkov and Michel Ledoux. One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 2019. URL https://api.semanticscholar.org/CorpusID:209912842.

Vladimir I. Bogachev. *Gaussian Measures*. American Mathematical Society, 1998.

Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3794–3801, 2014. doi: 10.1109/CVPR.2014.491.

Clément Bonet, Elsa Cazelles, Lucas Drumetz, and Nicolas Courty. Busemann functions in the wasserstein space: existence, closed-forms, and applications to slicing. *arXiv preprint arXiv:2510.04579*, 2025a.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing datasets with wasserstein over wasserstein gradient flows. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=I1OHPb4zWo.

Marta Catalano and Hugo Lavenant. Hierarchical integral probability metrics: A distance on random probability measures with low sample complexity. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=cmy38XZlJu.

Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, L. Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *ArXiv*, abs/1512.03012, 2015. URL https://api.semanticscholar.org/CorpusID:2554264.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018. doi: 10.1007/s10208-016-9331-y. URL https://doi.org/10.1007/s10208-016-9331-y.

Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/4e0cb6fb5fb446d1c92ede2ed8780188-Paper.pdf.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Richard M Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Yonatan Dukler, Wuchen Li, Alex Lin, and Guido Montufar. Wasserstein of Wasserstein loss for learning generative models. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1716–1725. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/dukler19a.html.

Xavier Fernique. Intégrabilité des vecteurs gaussiens. *CR Acad. Sci. Paris Sér. AB*, 270:A1698–A1699, 1970.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Ruiyu Han. Sliced wasserstein distance between probability measures on hilbert spaces. *arXiv preprint arXiv:2307.05802*, 2023.

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions, 2025. URL https://openreview.net/forum?id=HB4lr0ykTi.

Olav Kallenberg. *Random Measures, Theory and Applications*. Springer, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211 (3):969–1117, 2018. doi: 10.1007/s00222-017-0759-8. URL https://doi.org/10.1007/s00222-017-0759-8.

Xinran Liu, Yikun Bai, Yuzhe Lu, Andrea Soltoggio, and Soheil Kolouri. Wasserstein task embedding for measuring task similarities. *Neural Networks*, 181:106796, 2025.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(95):2651–2667, 2006. URL http://jmlr.org/papers/v7/micchelli06a.html.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20802–20812. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/eefc9e10ebdc4a2333b42b2dbb8f27b6-Paper.pdf.

Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, and Umut Simsekli. Fast approximation of the sliced-Wasserstein distance using concentration of random projections. *Advances in Neural Information Processing Systems*, 34:12411–12424, 2021.

Masha Naslidnyk, Siu Lun Chau, Francois-Xavier Briol, and Krikamol Muandet. Kernel quantile embeddings and associated probability metrics. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45770–45800. PMLR, 2025. URL https://proceedings.mlr.press/v267/naslidnyk25a.html.

Khai Nguyen and Peter Mueller. Summarizing nonparametric bayesian mixture posteriors—sliced optimal transport metrics for gaussian mixtures. *Journal of Computational and Graphical Statistics*, 0(0):1–12, 2026. doi: 10.1080/10618600.2026.2634831.

Khai Nguyen, Hai Nguyen, Tuan Pham, and Nhat Ho. Lightspeed geometric dataset distance via sliced optimal transport. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Moritz Piening and Robert Beinert. Slicing Wasserstein over Wasserstein via functional optimal transport. In *International Conference on Learning Representations*, volume 2026, pp. 84708–84740, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/88549904b871acbd676fb7c81581b0c2-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Raif M. Rustamov and Subhabrata Majumdar. Intrinsic sliced Wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29388–29415. PMLR, 2023. URL https://proceedings.mlr.press/v202/rustamov23a.html.

Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL http://jmlr.org/papers/v12/sriperumbudur11a.html.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

Robert W. Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM Trans. Graph.*, 23(3):399–405, August 2004. ISSN 0730-0301. doi: 10.1145/1015706.1015736. URL https://doi.org/10.1145/1015706.1015736.

Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016. URL http://jmlr.org/papers/v17/14-510.html.

Aad W. van der Vaart and J. Harry van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pp. 200–222. Institute of Mathematical Statistics, 2008.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. PointFlow: 3D point cloud generation with continuous normalizing flows. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4540–4549, 2019. doi: 10.1109/ICCV.2019.00464.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin Solomon. *Hierarchical optimal transport for document representation*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3D shape generation and completion through point-voxel diffusion. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5806–5815. IEEE, 2021.

A Background

Proof of Lemma 2.1. Both identities follow from linearity of the integral, in the integrand and in the measure respectively. For the first, $af + bg \in C(\Omega)$ and

$$P_{af+bg}(\mu) = \int_\Omega (af + bg)\,\mathrm{d}\mu = a\int_\Omega f\,\mathrm{d}\mu + b\int_\Omega g\,\mathrm{d}\mu = a\,P_f(\mu) + b\,P_g(\mu).$$

For the second, $\mu_t := (1-t)\mu_0 + t\mu_1 \in \mathcal{P}(\Omega)$, and for every Borel $B \subseteq \Omega$ we have $\mu_t(B) = (1-t)\mu_0(B) + t\mu_1(B)$; the identity $\int_\Omega f\,\mathrm{d}\mu_t = (1-t)\int_\Omega f\,\mathrm{d}\mu_0 + t\int_\Omega f\,\mathrm{d}\mu_1$ therefore holds for simple f and extends to all $f \in C(\Omega)$ by uniform approximation and boundedness of f on the compact Ω. Hence $P_f(\mu_t) = (1-t)P_f(\mu_0) + tP_f(\mu_1)$. □

Details on the Slicing Kernel Assumption

We provide additional details for Assumption 1. Let $f \sim \sigma = \mathrm{GP}(0, \kappa)$, where κ is continuous and universal on the compact space Ω, i.e., its RKHS $\mathcal{H}_\kappa$ is dense in $C(\Omega)$. A standard example is the Gaussian radial basis function (RBF) kernel on compact $\Omega \subset \mathbb{R}^d$ (Steinwart, 2001; Micchelli et al., 2006).

The canonical metric induced by the Gaussian process is

$$d_\kappa(x, y) := \left(\mathbb{E}_{f\sim\sigma}\left[(f(x) - f(y))^2\right]\right)^{1/2} = \left(\kappa(x,x) - 2\kappa(x,y) + \kappa(y,y)\right)^{1/2}.$$

For a metric d, the covering number is defined as

$$N(\Omega, d, \varepsilon) := \min\left\{ n \in \mathbb{N} : \exists x_1, \ldots, x_n \in \Omega \text{ such that } \Omega = \bigcup_{i=1}^{n} B_d(x_i, \varepsilon) \right\}.$$

The sample-path continuity required by Assumption 1 is guaranteed by Dudley's entropy condition

$$\int_0^1 \sqrt{\log N(\Omega, d_\kappa, \varepsilon)}\,\mathrm{d}\varepsilon < \infty,$$

where N denotes the covering number. Under this condition, the Gaussian process admits a version with almost surely continuous sample paths (Dudley, 1967; Adler & Taylor, 2007). Universality remains a separate requirement on κ.

Lemma A.1. *Under Assumption 1, (i) there is $\varepsilon > 0$ with $\mathbb{E}_\sigma \exp(\varepsilon \|f\|_\infty^2) < \infty$; in particular $\Lambda_r := \mathbb{E}_\sigma \|f\|_\infty^r < \infty$ for all $r \geq 1$ (Fernique, 1970; Ledoux & Talagrand, 1991); (ii) $\operatorname{supp}(\sigma) = \overline{\mathcal{H}_\kappa}^{\|\cdot\|_\infty} = C(\Omega)$ (Bogachev, 1998, Thm. 3.6.1), see also (van der Vaart & van Zanten, 2008).*

Remark A.2. *More specifically, from (ii), we have that for any $f \in C(\Omega), \epsilon > 0$, there exists finitely many $x_1, \dots x_n \in \Omega, c_1, \dots c_n \in \mathbb{R}$, with $\|f - \sum_{i=1}^n c_i \kappa(\cdot, x_i)\|_\infty < \epsilon$.*

Assumption 2. *We assume that the underlying space $\Omega \subset \mathbb{R}^d$ is compact and that $\kappa \in C^2(\Omega \times \Omega)$. Then the derivative process $\nabla f = (\partial_1 f, \dots, \partial_d f)$ is a centered Gaussian field with covariance*

$$\mathbb{E}_\sigma\big[\partial_i f(x)\, \partial_j f(y)\big] = \frac{\partial^2 \kappa(x,y)}{\partial x_i\, \partial y_j}, \qquad i,j = 1, \dots, d,$$

and we assume that its canonical metric

$$d_\nabla(x,y) := \big(\mathbb{E}_\sigma \|\nabla f(x) - \nabla f(y)\|^2\big)^{1/2}$$

satisfies Dudley's entropy condition $\int_0^1 \sqrt{\log N(\Omega, d_\nabla, \varepsilon)}\, \mathrm{d}\varepsilon < \infty$. Under these hypotheses $f \in C^1(\Omega)$ for σ-a.e. f, i.e. $\sigma\big(C^1(\Omega)\big) = 1$, and the Borell–TIS inequality (Adler & Taylor, 2007) gives

$$\mathbb{E}_\sigma\Big[\sup_{x \in \Omega} \|\nabla f(x)\|^r\Big] < \infty \qquad \text{for all } r \geq 1.$$

For instance, the Gaussian RBF kernel is C^∞ and satisfies these conditions, with C^∞ sample paths.

Schoenberg theory. A metric d on a set X is *Hilbertian* if it is realized by an isometric feature map: there exist a Hilbert space $\mathcal{H}$ and a map $\phi : X \to \mathcal{H}$ with

$$d(x,y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}} \qquad \text{for all } x, y \in X.$$

A symmetric function $\psi : X \times X \to \mathbb{R}$ with $\psi(x,x) = 0$ is *conditionally negative definite* (CND) if

$$\sum_{i,j=1}^n c_i c_j\, \psi(x_i, x_j) \;\leq\; 0 \qquad \text{whenever } n \in \mathbb{N},\ x_1, \dots, x_n \in X,\ c \in \mathbb{R}^n,\ \sum_{i=1}^n c_i = 0,$$

that is, negative semi-definiteness restricted to the hyperplane of zero-sum coefficient vectors. Schoenberg's theorem states that d is Hilbertian if and only if d^2 is CND, and that a nonnegative CND function stays CND under fractional powers: ψ CND implies ψ^α CND for every $\alpha \in (0,1]$. Applying this with $\psi = d^2$ and $\alpha = \beta/2$ gives that d^β is CND for every $\beta \in (0,2]$, and by the CND–PD correspondence

$$d^\beta \text{ is CND} \iff \exp(-\gamma d^\beta) \text{ is positive definite for every } \gamma > 0,$$

we conclude that

$$K_{\gamma,\beta}(x,y) = \exp\big(-\gamma\, d(x,y)^\beta\big)$$

is a positive definite kernel on X for all $\gamma > 0$ and $\beta \in (0,2]$ (Schoenberg, 1938; Berg et al., 1984).

B Our method: Position between kernel and transport lifts

Writing $m_\mu := \int \kappa(\cdot, x)\, \mathrm{d}\mu(x) \in \mathcal{H}_\kappa$, the Gaussian field $\mu \mapsto P_f(\mu)$ has covariance $\mathbb{E}_\sigma[P_f(\mu) P_f(\nu)] = \langle m_\mu, m_\nu \rangle_{\mathcal{H}_\kappa}$. Thus VSW uses Gaussian linear probes of the embedded meta-measures, in the sense of an isonormal Gaussian process on $\mathcal{H}_\kappa$. This connects the construction to Hilbert-space slicing (Han, 2023) and kernel expected sliced Wasserstein (Naslidnyk et al., 2025). It does not identify their slicing laws: our GP is $C(\Omega)$-valued and need not be $\mathcal{H}_\kappa$-valued, whereas a Hilbert-valued Gaussian requires trace-class covariance. Intrinsic sliced Wasserstein (Rustamov & Majumdar, 2023) instead pushes each inner measure through Laplacian eigenfunctions to one-dimensional laws; VSW takes scalar expectations and then compares their outer distributions.

Define $\mathrm{MMD}_\kappa(\mu,\nu) = \big\|\int \kappa(\cdot, x)\, \mathrm{d}(\mu - \nu)(x)\big\|_{\mathcal{H}_\kappa}$, which satisfies

$$\mathrm{MMD}_\kappa^2(\mu,\nu) = \iint \kappa\ \mathrm{d}(\mu-\nu) \otimes \mathrm{d}(\mu-\nu) = \mathbb{E}_{f\sim\sigma}\big[\langle f, \mu - \nu\rangle^2\big]. \tag{8}$$

Write $c_p := \mathbb{E}|Z|^p = 2^{p/2}\Gamma\big(\frac{p+1}{2}\big)/\sqrt{\pi}$ for $Z \sim \mathcal{N}(0,1)$ (so $c_2 = 1$, $c_1 = \sqrt{2/\pi}$).

Proposition B.1 (Exact restriction to Diracs: VSW extends MMD)**.** *For all $\mu, \nu \in \mathcal{P}(\Omega)$ and $p \in [1, \infty)$,*

$$\mathrm{VSW}_p\big(\delta_\mu, \delta_\nu\big) \;=\; c_p^{1/p}\, \mathrm{MMD}_\kappa(\mu, \nu).$$

In particular the embedding $\mu \mapsto \delta_\mu$ is, up to the constant $c_p^{1/p}$, an isometry from $(\mathcal{P}(\Omega), \mathrm{MMD}_\kappa)$ into $(\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{VSW}_p)$.

Proof of Proposition B.1. Under δ_μ the projected law is the point mass at $\langle f, \mu\rangle$, so the slice value is $|\langle f, \mu - \nu\rangle|$. The scalar $\langle f, \mu - \nu\rangle$ is centered Gaussian with variance $\mathrm{MMD}_\kappa^2(\mu, \nu)$ by (8); taking p-th absolute moments and p-th roots gives the claim. □

Proposition B.2 (Upper bound: Wasserstein over MMD, tight on Diracs)**.** *For all $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{P}(\Omega))$ and $p \in [1, \infty)$,*

$$\mathrm{VSW}_p(\mathbb{P}, \mathbb{Q}) \;\le\; c_p^{1/p}\, \mathbb{W}_{p, \mathrm{MMD}_\kappa}(\mathbb{P}, \mathbb{Q}),$$

with equality when $\mathbb{P}, \mathbb{Q}$ are Dirac meta-measures. If in addition Assumption 2 holds and $L_p := \mathbb{E}_\sigma[\mathrm{Lip}(f)^p]^{1/p} < \infty$, then also $\mathrm{VSW}_p(\mathbb{P}, \mathbb{Q}) \le L_p\, \mathbb{W}_{p, \mathrm{W}_1}(\mathbb{P}, \mathbb{Q}) \le L_p\, \mathbb{W}_{p, \mathrm{W}_p}(\mathbb{P}, \mathbb{Q})$.

Proof of Proposition B.2. Let $\Gamma \in \Pi(\mathbb{P}, \mathbb{Q})$ be arbitrary. For each f, $(P_f \times P_f)_\# \Gamma$ couples the projected laws, so $\mathrm{W}_p^p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big) \le \int |\langle f, \mu - \nu\rangle|^p \,\mathrm{d}\Gamma(\mu, \nu)$. Integrate in f and use Fubini:

$$\mathrm{VSW}_p^p(\mathbb{P}, \mathbb{Q}) \le \int \Big(\mathbb{E}_\sigma |\langle f, \mu - \nu\rangle|^p\Big) \,\mathrm{d}\Gamma(\mu, \nu) = c_p \int \mathrm{MMD}_\kappa(\mu, \nu)^p \,\mathrm{d}\Gamma(\mu, \nu).$$

Taking the infimum over Γ proves the bound. □

Proposition B.3 (Lower bound: intensities in MMD)**.** *For all $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{P}(\Omega))$ and all $p \in [1, \infty)$,*

$$\mathrm{VSW}_p(\mathbb{P}, \mathbb{Q}) \;\ge\; c_p^{1/p}\, \mathrm{MMD}_\kappa\big(m_\mathbb{P},\, m_\mathbb{Q}\big),$$

where $m_\mathbb{P}, m_\mathbb{Q}$ are the intensity measures, in particular, $m_\mathbb{P}(B) = \mathbb{E}_{\mu\sim\mathbb{P}}[\mu(B)], \forall B \in \mathcal{B}(\Omega)$.

Proof of Proposition B.3. For laws α, β on $\mathbb{R}$, any coupling (X, Y) satisfies, by Jensen, $\mathbb{E}|X - Y|^p \ge |\mathbb{E}X - \mathbb{E}Y|^p$; hence $\mathrm{W}_p(\alpha, \beta) \ge |\,\mathrm{mean}(\alpha) - \mathrm{mean}(\beta)|$. The mean of $(P_f)_\#\mathbb{P}$ is $\mathbb{E}_{\mu\sim\mathbb{P}}\langle f, \mu\rangle = \langle f, m_\mathbb{P}\rangle$ by Fubini. Therefore $\mathrm{VSW}_p^p \ge \mathbb{E}_\sigma |\langle f, m_\mathbb{P} - m_\mathbb{Q}\rangle|^p$. The scalar $\langle f, m_\mathbb{P} - m_\mathbb{Q}\rangle$ is centered Gaussian with variance $\mathrm{MMD}_\kappa^2(m_\mathbb{P}, m_\mathbb{Q})$ by (8). □

Example B.4 (Strictness: identical intensities, positive VSW)**.** *Let $a \ne b \in \Omega$ and consider*

$$\mathbb{P} = \delta_{\frac{1}{2}(\delta_a + \delta_b)} \qquad \textit{versus} \qquad \mathbb{Q} = \tfrac{1}{2}\,\delta_{\delta_a} + \tfrac{1}{2}\,\delta_{\delta_b}.$$

Both have intensity $m_\mathbb{P} = m_\mathbb{Q} = \frac{1}{2}(\delta_a + \delta_b)$, so every comparison based on intensities—in particular the level-2 MMD of mean embeddings—vanishes, and Proposition B.3 is uninformative. Fix $f \in C(\Omega)$. Under $\mathbb{P}$ the projected law is the point mass at $s := \frac{1}{2}(f(a) + f(b))$; under $\mathbb{Q}$ it is $\frac{1}{2}\delta_{f(a)} + \frac{1}{2}\delta_{f(b)}$. The only coupling of a Dirac with another law is the product, and $f(a) - s = s - f(b) = \frac{1}{2}(f(a) - f(b))$, so

$$\mathrm{W}_p^p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big) = \tfrac{1}{2}|f(a) - s|^p + \tfrac{1}{2}|f(b) - s|^p = 2^{-p}\,|f(a) - f(b)|^p.$$

Since $f(a) - f(b) \sim \mathcal{N}\big(0, d_\kappa(a, b)^2\big)$, integrating in f gives

$$\mathrm{VSW}_p^p(\mathbb{P}, \mathbb{Q}) = 2^{-p}\,\mathbb{E}_\sigma|f(a) - f(b)|^p = 2^{-p} c_p\, d_\kappa(a, b)^p > 0,$$

the strict positivity because $d_\kappa(a, b) > 0$ whenever κ is strictly positive definite and $a \ne b$. Thus VSW *sees the outer (transport) structure that intensity comparisons cannot; every slice with $f(a) \ne f(b)$ separates the two meta-measures.*

Remark B.5 (Interpretation)**.** *Propositions B.1–B.3 identify VSW_p as* the sliced Wassersteinization of MMD at the meta level*: its inner geometry is exactly MMD_κ (tight on Diracs), while its outer geometry is transport—strictly finer than the mean-embedding (intensity) comparison, as Example B.4 shows, yet never exceeding the full outer OT cost over the MMD_κ ground metric.*

Corollary B.6 (Universal characteristic kernels on meta-measures)**.** *For every $\gamma > 0$ and $\beta \in (0, 2]$, the kernel*

$$K_{\gamma,\beta}(\mathbb{P}, \mathbb{Q}) = \exp\big(-\gamma\, \mathrm{VSW}_2(\mathbb{P}, \mathbb{Q})^\beta\big)$$

is positive definite on $\mathcal{P}(\mathcal{P}(\Omega))$. Moreover, it is continuous and universal *on the compact metric space $(\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{weak})$, hence characteristic. Consequently, the MMD induced by $K_{\gamma,\beta}$ is a metric on the space of Borel probability measures over $\mathcal{P}(\mathcal{P}(\Omega))$.*

Proof of Corollary B.6. Positive definiteness. By Theorem 3.6, there is an isometric embedding $T : (\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{VSW}_2) \to H$ into a Hilbert space H with $\mathrm{VSW}_2(\mathbb{P}, \mathbb{Q}) = \|T(\mathbb{P}) - T(\mathbb{Q})\|_H$. Hence VSW_2^2 is conditionally negative definite (CND). For every $\beta \in (0, 2]$, we have $\beta/2 \in (0, 1]$, so $r \mapsto r^{\beta/2}$ is a Bernstein function on $[0, \infty)$. Applying it to the nonnegative CND function VSW_2^2 shows that $(\mathrm{VSW}_2^2)^{\beta/2} = \mathrm{VSW}_2^\beta$ is CND. Schoenberg's theorem then yields that $K_{\gamma,\beta}(\mathbb{P}, \mathbb{Q}) = \exp\big(-\gamma\, \mathrm{VSW}_2(\mathbb{P}, \mathbb{Q})^\beta\big)$ is positive definite for all $\gamma > 0$ and $\beta \in (0, 2]$.

Continuity. By Proposition 3.5, VSW_2 induces the weak topology on $\mathcal{P}(\mathcal{P}(\Omega))$, hence $(\mathbb{P}, \mathbb{Q}) \mapsto \mathrm{VSW}_2(\mathbb{P}, \mathbb{Q})$ is continuous for the weak topology. Composing with $t \mapsto \exp(-\gamma t^\beta)$ gives continuity of $K_{\gamma,\beta}$.

Universality and characteristicness. The space $(\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{weak})$ is compact. For $\beta = 2$, $K_{\gamma,2}(\mathbb{P}, \mathbb{Q}) = \exp(-\gamma\|T(\mathbb{P}) - T(\mathbb{Q})\|_H^2)$ is the restriction of a Gaussian kernel on the Hilbert space H to the compact set $T(\mathcal{P}(\mathcal{P}(\Omega)))$. Such a Gaussian kernel is universal on compact subsets of a separable Hilbert space (e.g. Christmann & Steinwart, 2010); therefore $K_{\gamma,2}$ is universal on $\mathcal{P}(\mathcal{P}(\Omega))$.

For $\beta \in (0, 2)$, the function $\psi(t) := \exp(-\gamma t^{\beta/2})$ is completely monotone on $[0, \infty)$, so by Bernstein's theorem there exists a probability measure ν_β on $[0, \infty)$ such that $\psi(t) = \int_0^\infty e^{-st}\, \mathrm{d}\nu_\beta(s)$ and $\nu_\beta((0, \infty)) > 0$. Substituting $t = \mathrm{VSW}_2^2(\mathbb{P}, \mathbb{Q})$ gives

$$K_{\gamma,\beta}(\mathbb{P}, \mathbb{Q}) = \int_0^\infty \exp\big(-s\, \mathrm{VSW}_2^2(\mathbb{P}, \mathbb{Q})\big)\, \mathrm{d}\nu_\beta(s).$$

If a finite signed Borel measure η on $\mathcal{P}(\mathcal{P}(\Omega))$ satisfies $\iint K_{\gamma,\beta}\, \mathrm{d}\eta \otimes \mathrm{d}\eta = 0$, Fubini applies because the kernels are bounded and η has finite total variation. The nonnegative Gaussian-kernel quadratic forms therefore vanish for ν_β-a.e. s, hence for some $s > 0$. Universality of $\exp(-s\, \mathrm{VSW}_2^2)$ forces $\eta = 0$ by the equivalence between universality and integral strict positive definiteness on compact spaces (Sriperumbudur et al., 2011, Prop. 4). The same equivalence proves universality of $K_{\gamma,\beta}$; universal continuous kernels on compact metric spaces are characteristic. □

Remark B.7 (Uses)**.** *Corollary B.6 yields kernel two-sample tests between collections of datasets, distribution regression whose inputs are meta-measures (one level above Szabó et al., 2016), and Gaussian-process priors over functionals of meta-measures.*

C Proofs of the Statements in Section 3

Proof of Lemma 3.2. (i) Let $(f_n, \mu_n) \to (f, \mu)$ in $C(\Omega) \times \mathcal{P}(\Omega)$. In particular:

- $f_n \to f$ in $(C(\Omega), \|\cdot\|_\infty)$, i.e. $\|f_n - f\|_\infty = \sup_{x\in\Omega} |f_n(x) - f(x)| \to 0$, which is uniform convergence on Ω;
- $\mu_n \to \mu$ in $\mathcal{P}(\Omega)$, i.e. $\mu_n \rightharpoonup \mu$ weakly: $\int_\Omega g\, \mathrm{d}\mu_n \to \int_\Omega g\, \mathrm{d}\mu$ for every $g \in C(\Omega)$.

Then

$$|\langle f_n, \mu_n\rangle - \langle f, \mu\rangle| \le |\langle f_n - f, \mu_n\rangle| + |\langle f, \mu_n\rangle - \langle f, \mu\rangle| \le \|f_n - f\|_\infty + |\langle f, \mu_n\rangle - \langle f, \mu\rangle|\,.$$

The first term tends to zero by uniform convergence, while the second tends to zero by weak convergence, since the fixed function $f \in C(\Omega)$ is bounded and continuous. Hence P is jointly continuous.

(ii) Choose fixed $f, g \in C(\Omega)$, the deterministic map $\mu \mapsto (P_f(\mu), P_g(\mu))$ pushes $\mathbb{P}$ to a coupling between $(P_f)_\#\mathbb{P}$ and $(P_g)_\#\mathbb{P}$. We have:

$$\begin{aligned}\mathbb{E}_{\mu\sim\mathbb{P}}|P_f(\mu) - P_g(\mu)|^p &= \mathbb{E}_{\mu\sim\mathbb{P}}\Big|\int (f-g)d\mu\Big|^p \\ &\le \mathbb{E}_{\mu\sim\mathbb{P}}\int |f-g|^p d\mu \\ &\le \sup_{x\in\Omega}|f(x)-g(x)|^p \\ &= \|f-g\|_\infty^p ;\end{aligned}$$

hence $\mathrm{W}_p\big((P_f)_\#\mathbb{P}, (P_g)_\#\mathbb{P}\big) \le \|f-g\|_\infty$, and similarly for $\mathbb{Q}$. The triangle inequality for W_p on $\mathcal{P}_p(\mathbb{R})$ then gives $\big|\mathrm{W}_p((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}) - \mathrm{W}_p((P_g)_\#\mathbb{P}, (P_g)_\#\mathbb{Q})\big| \le 2\|f-g\|_\infty$. Thus $f \mapsto \mathrm{W}_p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big)$ is 2-Lipschitz.

Finiteness. Fix $f \in C(\Omega)$. As noted after Definition 3.1, both $(P_f)_\#\mathbb{P}$ and $(P_f)_\#\mathbb{Q}$ are supported in the interval $I_f := [-\|f\|_\infty, \|f\|_\infty]$, of length $|I_f| = 2\|f\|_\infty$. Consequently every coupling $\pi \in \Pi\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big)$ is supported in $I_f \times I_f$, on which $|x-y| \le 2\|f\|_\infty$, so

$$\mathrm{W}_p^p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big) = \inf_\pi \int_{\mathbb{R}\times\mathbb{R}} |x-y|^p \,\mathrm{d}\pi(x,y) \;\le\; \big(2\|f\|_\infty\big)^p.$$

Integrating this bound against σ and using Lemma A.1(i),

$$\mathrm{VSW}_p^p(\mathbb{P},\mathbb{Q}) = \int_{C(\Omega)} \mathrm{W}_p^p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big)\,\mathrm{d}\sigma(f) \;\le\; 2^p \int_{C(\Omega)} \|f\|_\infty^p \,\mathrm{d}\sigma(f) \;=\; 2^p\Lambda_p \;<\; \infty.$$

The integral in (4) is therefore well defined. □

Proof of Theorem 3.3. We first show that VSW_p is a pseudometric on $\mathcal{P}(\mathcal{P}(\Omega))$: it is finite, symmetric, vanishes on the diagonal, and satisfies the triangle inequality.

Finiteness was established in Lemma 3.2(ii): $\mathrm{VSW}_p^p(\mathbb{P},\mathbb{Q}) \le 2^p\Lambda_p < \infty$ for all $\mathbb{P},\mathbb{Q} \in \mathcal{P}(\mathcal{P}(\Omega))$.

Symmetry is inherited slicewise: W_p is symmetric on $\mathcal{P}(\mathbb{R})$.

Vanishing on the diagonal: if $\mathbb{P} = \mathbb{Q}$ then $(P_f)_\#\mathbb{P} = (P_f)_\#\mathbb{Q}$ for every $f \in C(\Omega)$.

Triangle inequality. Let $\mathbb{P},\mathbb{Q},\mathbb{R}' \in \mathcal{P}(\mathcal{P}(\Omega))$ and write $h_{\mathbb{P}\mathbb{Q}}(f) := \mathrm{W}_p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big)$, so that $\mathrm{VSW}_p(\mathbb{P},\mathbb{Q}) = \|h_{\mathbb{P}\mathbb{Q}}\|_{L^p(\sigma)}$ by (4). For each fixed f, the triangle inequality for W_p on $\mathcal{P}(\mathbb{R})$ gives the pointwise bound

$$h_{\mathbb{P}\mathbb{Q}}(f) \;\le\; h_{\mathbb{P}\mathbb{R}'}(f) + h_{\mathbb{R}'\mathbb{Q}}(f).$$

Taking $L^p(\sigma)$ norms and applying Minkowski's inequality gives the result.

The only property left to verify is identity of indiscernibles. Assume $\mathrm{VSW}_p(\mathbb{P},\mathbb{Q}) = 0$.

Define

$$A := \big\{f \in C(\Omega) :\ (P_f)_\#\mathbb{P} = (P_f)_\#\mathbb{Q}\big\} = \big\{f \in C(\Omega) :\ \mathrm{W}_p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big) = 0\big\}.$$

The two descriptions agree because both projected laws are compactly supported, hence lie in $\mathcal{P}_p(\mathbb{R})$, the subspace of $\mathcal{P}(\mathbb{R})$ of measures with finite p-moment, on which W_p is a metric.

Step 0 ($\sigma(A) = 1$). By assumption and (4),

$$0 = \mathrm{VSW}_p^p(\mathbb{P},\mathbb{Q}) = \int_{C(\Omega)} \mathrm{W}_p^p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big)\,\mathrm{d}\sigma(f).$$

The integrand is nonnegative and Borel (Lemma 3.2(ii)), so it vanishes σ-a.e., i.e. $\sigma(A) = 1$.

Step 1 ($A = C(\Omega)$). The map $h(f) := \mathrm{W}_p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big)$ is continuous by Lemma 3.2(ii), hence $A = h^{-1}(\{0\})$ is closed. Since $\sigma(A) = 1$, closedness implies $\mathrm{supp}\sigma \subseteq A$. By Lemma A.1(ii) and universality (Assumption 1), $\mathrm{supp}\sigma = C(\Omega)$, hence $A = C(\Omega)$. Therefore

$$(P_f)_\#\mathbb{P} = (P_f)_\#\mathbb{Q} \qquad \text{for every } f \in C(\Omega). \tag{9}$$

Step 2 (finite-dimensional distributions agree). Fix $k \in \mathbb{N}$ and $f_1, \dots, f_k \in C(\Omega)$. Consider $X : \mathcal{P}(\Omega) \to \mathbb{R}^k$, $X(\mu) = (P_{f_1}(\mu), \dots, P_{f_k}(\mu))$. For any $t \in \mathbb{R}^k$, Lemma 2.1 gives

$$\langle t, X(\mu)\rangle = \sum_{j=1}^{k} t_j P_{f_j}(\mu) = P_{g_t}(\mu), \qquad g_t := \sum_{j=1}^{k} t_j f_j \in C(\Omega).$$

Applying (9) to g_t shows that $\langle t, X\rangle$ has the same law under $\mu \sim \mathbb{P}$ and under $\mu \sim \mathbb{Q}$. By the Cramér–Wold theorem, the laws of X under $\mathbb{P}$ and $\mathbb{Q}$ coincide, i.e.

$$\big(P_{f_1}, \dots, P_{f_k}\big)_{\#}\mathbb{P} = \big(P_{f_1}, \dots, P_{f_k}\big)_{\#}\mathbb{Q} \qquad \text{in } \mathcal{P}(\mathbb{R}^k). \tag{10}$$

Step 3 ($\mathbb{P} = \mathbb{Q}$). Let

$$\mathcal{C} = \Big\{\{\mu \in \mathcal{P}(\Omega) : (P_{f_1}(\mu), \dots, P_{f_k}(\mu)) \in B\} : \ k \in \mathbb{N},\ f_1, \dots, f_k \in C(\Omega),\ B \in \mathcal{B}(\mathbb{R}^k)\Big\}.$$

Then $\mathcal{C}$ is a π-system (closed under finite intersections). Moreover $\sigma(\mathcal{C}) = \mathcal{B}(\mathcal{P}(\Omega))$ because the weak Borel σ-algebra is generated by the maps P_f (Section 2).

Define $\mathcal{L} := \{C \in \mathcal{B}(\mathcal{P}(\Omega)) : \mathbb{P}(C) = \mathbb{Q}(C)\}$. This is a λ-system. By (10), we have $\mathcal{C} \subseteq \mathcal{L}$. Dynkin's π–λ theorem yields $\sigma(\mathcal{C}) \subseteq \mathcal{L}$, hence $\mathcal{B}(\mathcal{P}(\Omega)) \subseteq \mathcal{L}$ and therefore $\mathbb{P} = \mathbb{Q}$. □

Remark C.1. *Only Step 1 uses universality, and it does so through the support theorem: the closed agreement set A must contain* $\operatorname{supp}\sigma$. *If κ is not universal, Steps 0–3 still yield $(P_f)_{\#}\mathbb{P} = (P_f)_{\#}\mathbb{Q}$ for every $f \in \overline{\mathcal{H}_\kappa}$, hence $\mathbb{P} = \mathbb{Q}$ on $\mathcal{F}_\kappa := \sigma\big(P_f : f \in \overline{\mathcal{H}_\kappa}\big) \subseteq \mathcal{B}(\mathcal{P}(\Omega))$. Thus* VSW_p *remains a pseudometric on $\mathcal{P}(\mathcal{P}(\Omega))$ whose null relation is exactly agreement on $\mathcal{F}_\kappa$, and it is a metric on the corresponding quotient; when $\mathcal{F}_\kappa \subsetneq \mathcal{B}(\mathcal{P}(\Omega))$, distinct meta-measures can be at* VSW_p*-distance zero. Conversely, any slicing measure with full support in $C(\Omega)$ (not necessarily Gaussian) and finite Λ_p yields a metric by the same proof; the Gaussian choice is what later gives exact MMD compatibility (Propositions B.1–B.3) and Borell–TIS moment control.*

Proof of Proposition 3.5. Weak convergence implies $\mathrm{VSW}_p \to 0$. Let $\mathbb{P}_m \rightharpoonup \mathbb{P}$ in the weak topology on $\mathcal{P}(\mathcal{P}(\Omega))$, and fix $f \in C(\Omega)$. Since $P_f : \mathcal{P}(\Omega) \to \mathbb{R}$ is continuous (Lemma 3.2(i)), for every $g \in C_b(\mathbb{R})$ the composition $g \circ P_f$ belongs to $C(\mathcal{P}(\Omega))$; hence

$$\int_{\mathbb{R}} g \,\mathrm{d}\big((P_f)_{\#}\mathbb{P}_m\big) = \int_{\mathcal{P}(\Omega)} g \circ P_f \,\mathrm{d}\mathbb{P}_m \ \longrightarrow\ \int_{\mathcal{P}(\Omega)} g \circ P_f \,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}} g \,\mathrm{d}\big((P_f)_{\#}\mathbb{P}\big),$$

i.e. $(P_f)_{\#}\mathbb{P}_m \rightharpoonup (P_f)_{\#}\mathbb{P}$ in $\mathcal{P}(\mathbb{R})$. Moreover $|P_f| \le \|f\|_\infty$ on $\mathcal{P}(\Omega)$, so all these laws are supported in the fixed compact interval $I_f = [-\|f\|_\infty, \|f\|_\infty]$. On probability measures supported in a common compact interval, W_p metrizes weak convergence, so

$$\mathrm{W}_p^p\big((P_f)_{\#}\mathbb{P}_m, (P_f)_{\#}\mathbb{P}\big) \ \longrightarrow\ 0 \qquad \text{for each fixed } f \in C(\Omega).$$

The integrand is dominated by $(2\|f\|_\infty)^p$, which lies in $L^1(\sigma)$ by Lemma A.1(i). Dominated convergence therefore yields $\mathrm{VSW}_p(\mathbb{P}_m, \mathbb{P}) \to 0$.

$\mathrm{VSW}_p \to 0$ *implies weak convergence.* Consider the identity map

$$\iota : (\mathcal{P}(\mathcal{P}(\Omega)), \text{weak}) \to (\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{VSW}_p),$$

which is a bijection. The previous paragraph shows it is continuous. Since $(\mathcal{P}(\mathcal{P}(\Omega)), \text{weak})$ is compact and $(\mathcal{P}(\mathcal{P}(\Omega)), \mathrm{VSW}_p)$ is Hausdorff (being metric), a continuous bijection from a compact space onto a Hausdorff space is a homeomorphism. Hence VSW_p induces the weak topology. □

Proof of Theorem 3.6. Measurability. The evaluation map $(f, \mu) \mapsto P_f(\mu)$ is jointly continuous (Lemma 3.2(i)). Hence, for fixed $\mathbb{P} \in \mathcal{P}(\mathcal{P}(\Omega))$, the function

$$G(f, s) := \mathbb{P}\{\mu \in \mathcal{P}(\Omega) : P_f(\mu) \le s\}$$

is jointly Borel in (f, s): the set $\{(f, s, \mu) : P_f(\mu) - s \le 0\}$ is closed in $C(\Omega) \times \mathbb{R} \times \mathcal{P}(\Omega)$, so its indicator is jointly Borel, and Tonelli (integrating out μ against $\mathbb{P}$) yields joint measurability of G. Define the quantile

$$q(f, u) := \inf\{s \in \mathbb{Q}_{\mathrm{rat}} : G(f, s) \ge u\}.$$

Since the infimum runs over a countable set and each subgraph $\{(f,u) : u \le G(f,s)\}$ is Borel, q is jointly Borel in (f,u).

Integrability. Since $(P_f)_{\#}\mathbb{P}$ is supported in $[-\|f\|_\infty, \|f\|_\infty]$, we have $|q(f,u)| \le \|f\|_\infty$ for all (f,u). By Lemma A.1(i), $\mathbb{E}_\sigma \|f\|_\infty^2 < \infty$, so $T(\mathbb{P}) \in L^2(\sigma \otimes \lambda)$.

Isometry. For $p = 2$, the one-dimensional OT formula (2) gives, for each f,

$$\mathrm{W}_2^2\big((P_f)_{\#}\mathbb{P}, (P_f)_{\#}\mathbb{Q}\big) = \int_0^1 \big|q_{(P_f)_{\#}\mathbb{P}}(u) - q_{(P_f)_{\#}\mathbb{Q}}(u)\big|^2 \,\mathrm{d}u.$$

Integrating in f and applying Tonelli yields

$$\mathrm{VSW}_2^2(\mathbb{P},\mathbb{Q}) = \int_{C(\Omega)} \int_0^1 \big|T(\mathbb{P})(f,u) - T(\mathbb{Q})(f,u)\big|^2 \,\mathrm{d}u \,\mathrm{d}\sigma(f) = \|T(\mathbb{P}) - T(\mathbb{Q})\|_{L^2(\sigma\otimes\lambda)}^2 .$$

This proves the claimed isometric embedding. □

Proof of Theorem 3.7. By the triangle inequality (Theorem 3.3) and $(a+b)^p \le 2^{p-1}(a^p + b^p)$,

$$\mathbb{E}\,\mathrm{VSW}_p^p(\mathbb{P},\widetilde{\mathbb{P}}) \le 2^{p-1}\Big(\mathbb{E}\,\mathrm{VSW}_p^p(\mathbb{P},\widehat{\mathbb{P}}_N) + \mathbb{E}\,\mathrm{VSW}_p^p(\widehat{\mathbb{P}}_N,\widetilde{\mathbb{P}})\Big).$$

Outer term. Fix f and let $\nu_f := (P_f)_{\#}\mathbb{P} \in \mathcal{P}(I_f)$, $I_f = [-\|f\|_\infty, \|f\|_\infty]$. Then $(P_f)_{\#}\widehat{\mathbb{P}}_N$ is exactly the empirical measure of the N i.i.d. scalars $P_f(\mu_i) \sim \nu_f$. Using $\mathrm{W}_p^p \le |I_f|^{p-1}\mathrm{W}_1$ on I_f and (3),

$$\mathbb{E}\,\mathrm{W}_p^p\big((P_f)_{\#}\widehat{\mathbb{P}}_N, \nu_f\big) \le (2\|f\|_\infty)^{p-1} \cdot \frac{2\|f\|_\infty}{2\sqrt{N}} = \frac{(2\|f\|_\infty)^p}{2\sqrt{N}}.$$

Integrating in f (Tonelli; joint measurability as before) gives

$$\mathbb{E}\,\mathrm{VSW}_p^p(\mathbb{P},\widehat{\mathbb{P}}_N) \le \tfrac{1}{2}\cdot 2^p \Lambda_p N^{-1/2}.$$

Inner term. Pair μ_i with $\widehat{\mu}_i$: the coupling of $(P_f)_{\#}\widehat{\mathbb{P}}_N$ and $(P_f)_{\#}\widetilde{\mathbb{P}}$ that matches index i to index i yields

$$\mathrm{W}_p^p\big((P_f)_{\#}\widehat{\mathbb{P}}_N, (P_f)_{\#}\widetilde{\mathbb{P}}\big) \le \frac{1}{N}\sum_{i=1}^N \big|P_f(\mu_i) - P_f(\widehat{\mu}_i)\big|^p.$$

Conditionally on μ_i, $P_f(\widehat{\mu}_i) - P_f(\mu_i) = \frac{1}{n}\sum_j \big(f(x_{ij}) - \mathbb{E}f(x_{ij})\big)$ is a centered average of i.i.d. variables bounded by $2\|f\|_\infty$. For $p \in [1,2]$, Jensen and the variance bound give $\mathbb{E}\big|\cdot\big|^p \le \big(\frac{(2\|f\|_\infty)^2}{4n}\big)^{p/2} \le (2\|f\|_\infty)^p n^{-p/2}$; for $p > 2$, the Marcinkiewicz–Zygmund inequality gives $\mathbb{E}|\cdot|^p \le C_p(2\|f\|_\infty)^p n^{-p/2}$. Taking expectations, then the σ-integral (Tonelli), yields

$$\mathbb{E}\,\mathrm{VSW}_p^p(\widehat{\mathbb{P}}_N,\widetilde{\mathbb{P}}) \le C_p\, 2^p \Lambda_p\, n^{-p/2}.$$

Combining the outer and inner bounds proves the theorem. The distance rates in Remark 3.8 follow by Jensen ($\mathbb{E}\mathrm{VSW}_p \le (\mathbb{E}\mathrm{VSW}_p^p)^{1/p}$) and subadditivity of $x \mapsto x^{1/p}$. □

Remark C.2. *For $p = 2$, Theorem 3.7 gives $\mathbb{E}\,\mathrm{VSW}_2(\mathbb{P},\widetilde{\mathbb{P}}_{N,n}) = O(N^{-1/4} + n^{-1/2})$, whereas the hierarchical IPM of Catalano & Lavenant (2024) attains the parametric outer rate. The outer exponent is sharp in the worst case, as a two-atom meta-measure shows. For distinct $\mu_a, \mu_b \in \mathcal{P}(\Omega)$, set $\mathbb{P} = \frac{1}{2}\delta_{\mu_a} + \frac{1}{2}\delta_{\mu_b}$, and let $\widehat{w}_N$ be the proportion of μ_a among the N outer samples. Matching the common mass in each projected law gives*

$$\mathrm{VSW}_2^2(\mathbb{P},\widehat{\mathbb{P}}_N) = \big|\widehat{w}_N - \tfrac{1}{2}\big|\,\mathrm{MMD}_\kappa^2(\mu_a,\mu_b).$$

Since $N\widehat{w}_N \sim \mathrm{Binomial}(N,1/2)$, its centered absolute moments satisfy $\mathbb{E}|\widehat{w}_N - \frac{1}{2}| \asymp N^{-1/2}$ and $\mathbb{E}|\widehat{w}_N - \frac{1}{2}|^{1/2} \asymp N^{-1/4}$. Thus, for this fixed $\mathbb{P}$, the expected squared distance has order $N^{-1/2}$ and the expected distance has order $N^{-1/4}$. These are worst-case rates; faster outer convergence is

possible under additional regularity of the projected laws. Specifically, if for σ-a.e. f the law $(P_f)_\#\mathbb{P}$ admits a density ϱ_f with

$$J_2(f) := \int_{\mathbb{R}} \frac{F_{(P_f)_\#\mathbb{P}}(z)\big(1 - F_{(P_f)_\#\mathbb{P}}(z)\big)}{\varrho_f(z)}\,\mathrm{d}z, \qquad \mathbb{E}_\sigma\big[J_2(f)\big] < \infty,$$

then $\mathbb{E}\,\mathrm{W}_2^2\big((P_f)_\#\mathbb{P}, (P_f)_\#\widehat{\mathbb{P}}_N\big) \le 2J_2(f)/(N+1)$ (Bobkov & Ledoux, 2019, Thm. 5.1). Integrating in f yields $\mathbb{E}\,\mathrm{VSW}_2^2(\mathbb{P}, \widehat{\mathbb{P}}_N) \le 2\mathbb{E}_\sigma[J_2]/(N+1)$, hence $\mathbb{E}\,\mathrm{VSW}_2(\mathbb{P}, \widehat{\mathbb{P}}_N) = O(N^{-1/2})$. Combining this with the unchanged inner-sampling bound gives $\mathbb{E}\,\mathrm{VSW}_2(\mathbb{P}, \widetilde{\mathbb{P}}_{N,n}) = O(N^{-1/2} + n^{-1/2})$ under this additional regularity condition. In the definition of J_2, we use the conventions $0/0 = 0$ and $c/0 = +\infty$ for $c > 0$.

Proof of Proposition 3.9. The summands are i.i.d. with mean VSW_p^p and second moment at most $\mathbb{E}_\sigma(2\,\|f\|_\infty)^{2p} = 2^{2p}\Lambda_{2p}$; apply $\mathbb{E}|X - \mathbb{E}X| \le \sqrt{\mathrm{Var}\,X}$ to the empirical mean. (Cf. the analogous bound for sliced divergences in Nadjahi et al., 2020.) □

Remark C.3 (Finite truncations are pseudometrics)**.** *For finite R the support of σ_R is the finite-dimensional subspace* $\mathrm{span}\{\varphi_1, \ldots, \varphi_R\} \subseteq C(\Omega)$. *When this subspace is proper, Step 1 of the proof of Theorem 3.3 does not apply;* $\mathrm{VSW}_p^{(\sigma_R)}$ *is in general only a pseudometric, with null relation given by agreement of the projected laws along that subspace (cf. Remark C.1). Lemma D.3 therefore describes a sequence of pseudometrics converging uniformly to the metric* $\mathrm{VSW}_p^{(\sigma)}$*; separation is guaranteed in the limit $R \to \infty$, and may already hold at finite R when the feature map separates inner measures.*

Remark C.4 (Random features are not covered by the coupling bound)**.** *Proposition D.2 requires a coupling of the two slicing measures. An independent random Fourier feature draw* (5) *is* not *naturally coupled to the limiting process $f \sim \sigma$, so Lemma D.3 does not apply to it; the covariance of* (5) *converges uniformly to κ (Rahimi & Recht, 2007), but a quantitative transfer from covariance convergence to convergence of the sliced distances is left open, and we treat the induced bias empirically.*

The shared finite-feature implementation. Condition on the sampled Fourier basis, let σ_R be the conditional law of f_R, and define $\phi_R(x) = \sqrt{2/R}\big(\cos(\langle\omega_r, x\rangle + b_r)\big)_{r=1}^R$ and $z_R(\mu) = \int \phi_R(x)\,\mathrm{d}\mu(x)$. Then $P_{f_R}(\mu) = a^\top z_R(\mu)$ for $a \sim \mathcal{N}(0, I_R)$. Writing SW_p for classical slicing with uniform unit directions, Gaussian radial symmetry gives the exact identity

$$\mathrm{VSW}_p^{(\sigma_R)}(\mathbb{P}, \mathbb{Q})^p = \mathbb{E}\,\|a\|^p\ \mathrm{SW}_p^p\big((z_R)_\#\mathbb{P}, (z_R)_\#\mathbb{Q}\big).$$

The distance itself has factor $(\mathbb{E}\,\|a\|^p)^{1/p}$, equal to $\sqrt{R}$ for $p = 2$. The finite-feature map can send distinct measures to the same vector, so this is generally a pseudometric on the original meta-measures. Conditional on the basis, Proposition 3.9's variance argument applies with σ_R and its sup-norm moments in place of σ. Increasing L controls this conditional Monte Carlo error, while no quantitative feature-error bound is provided here (Remark C.4). For a direct finite-feature reference check, Proposition B.1 gives $\mathrm{VSW}_2(\delta_\mu, \delta_\nu) = \mathrm{MMD}_\kappa(\mu, \nu)$, whereas the conditional finite-feature value is $\|z_R(\mu) - z_R(\nu)\|_2$.

Gradient Flow

Lemma C.5 (No ties, almost surely)**.** *Suppose κ is characteristic and the clouds $\widehat{\mu}_1, \ldots, \widehat{\mu}_N$ are pairwise distinct as measures (and similarly for the targets). Then for σ-a.e. f, the projected scalars $s_1(f), \ldots, s_N(f)$ are pairwise distinct (and similarly $t_1, \ldots, t_N$), so the sorting permutation π_f matching $s_{(i)}$ to $t_{(i)}$ is locally constant in X around any configuration with distinct clouds.*

Proof. For $i \ne i'$, $s_i(f) - s_{i'}(f) = \langle f, \widehat{\mu}_i - \widehat{\mu}_{i'}\rangle$ is a centered Gaussian with variance $\mathrm{MMD}_\kappa^2(\widehat{\mu}_i, \widehat{\mu}_{i'})$, which is strictly positive because κ is characteristic and the clouds are distinct; a nondegenerate Gaussian hits 0 with probability zero. A union over the finitely many pairs concludes. □

Algorithm 2 VSW flow for populations of point clouds (one step)

Require: particles $X = (x_{ij})$, targets $\{y_{kl}\}$, kernel κ, slices L, features R, step η

```
sample fresh slices f_1, ..., f_L via (5)
for l = 1, ..., L do
    compute s_i^(l), t_k^(l); sort both; obtain matching π_l
end for
g_ij ← (2/(NnL)) Σ_l (s_i^(l) − t_{π_l(i)}^(l)) ∇f_l(x_ij)
x_ij ← x_ij − η g_ij                    ▷ or pass g to Adam / reparametrized generator
```

Remark C.6 (Parametric generators.)**.** *For a differentiable generator $x_{ij} = g_\theta(z_i, \zeta_{ij})$, the particle gradients can be backpropagated through the generator by the chain rule, wherever the empirical objective is differentiable. For a fixed kernel and empirical minibatch, independent Gaussian-process slices give unbiased Monte Carlo estimates of the empirical loss and, when differentiation and expectation can be interchanged, its gradient. With finite random Fourier features, the corresponding statement holds for the finite-feature objective conditional on the sampled basis. Under suitable differentiation and integrability assumptions, averaging over random minibatches targets the gradient of the expected empirical minibatch objective, which generally differs from the population VSW_2^2 objective. Theorem 3.7 controls distance estimation error; it does not establish convergence of the corresponding gradients.*

Finally, we prove Proposition 3.10 which holds under Assumptions 1–2 and the hypotheses of Lemma C.5.

Proof of Proposition 3.10. For $f \in C^1$ write

$$F_f(X) := \min_{\pi \in S_N} \frac{1}{N} \sum_{i=1}^{N} \big(s_i(f) - t_{\pi(i)}(f)\big)^2,$$

so that $\mathcal{F} = \mathbb{E}_\sigma[F_f]$ by (2) (the increasing rearrangement realizes the minimum). Fix a coordinate x_{ij} and let $\Omega' \supset \Omega$ be a fixed compact neighborhood containing all configurations under consideration. In this proof, $\|f\|_\infty$ denotes $\sup_{x \in \Omega'} |f(x)|$. We assume that the regularity and moment conditions of Assumption 2 hold on Ω'; then the field ∇f is defined and Borell–TIS applies on Ω' (for the Gaussian RBF, f extends to all of $\mathbb{R}^d$). The interchange of one partial derivative with $\mathbb{E}_\sigma$ proceeds in three steps.

Step 1 (pointwise differentiability for σ-a.e. f). Let A be the full-measure set of Lemma C.5. For $f \in A$ the projected scalars $s_1(f), \dots, s_N(f)$ (resp. $t_1(f), \dots, t_N(f)$) are pairwise distinct, so the minimizing permutation π_f is unique and remains optimal on a ball $B(X, r_f)$ whose radius $r_f > 0$ *depends on* f, by continuity of $X \mapsto s_i(f)$. On this ball F_f coincides with the smooth function $\frac{1}{N}\sum_i (s_i - t_{\pi_f(i)})^2$, whence

$$\partial_{x_{ij}} F_f(X) = \tfrac{2}{Nn}\big(s_i(f) - t_{\pi_f(i)}(f)\big)\nabla f(x_{ij}), \qquad \partial s_i / \partial x_{ij} = \tfrac{1}{n}\nabla f(x_{ij}).$$

Step 2 (uniform integrable envelope for the difference quotients). For every $f \in C^1$ (not only $f \in A$) and every fixed $\pi \in S_N$, the map $X \mapsto \frac{1}{N}\sum_i (s_i - t_{\pi(i)})^2$ is differentiable with partial gradient in x_{ij} bounded in norm by

$$E(f) := \tfrac{4}{Nn}\, \|f\|_\infty \sup_{\Omega'} \|\nabla f\|,$$

since $|s_i - t_{\pi(i)}| \le 2\|f\|_\infty$. A pointwise minimum of finitely many functions that are M-Lipschitz along a segment is itself M-Lipschitz along that segment; hence every difference quotient of F_f in the coordinate x_{ij} is bounded by $E(f)$, *uniformly in the step size*. Moreover $E \in L^1(\sigma)$: by Cauchy–Schwarz,

$$\mathbb{E}_\sigma\big[\, \|f\|_\infty \sup_{\Omega'} \|\nabla f\| \,\big] \le \big(\mathbb{E}_\sigma \|f\|_\infty^2\big)^{1/2} \big(\mathbb{E}_\sigma \sup_{\Omega'} \|\nabla f\|^2\big)^{1/2} < \infty$$

by Lemma A.1(i) and Assumption 2.

Step 3 (dominated convergence). As $h \to 0$, the difference quotients $h^{-1}\big(F_f(X+he) - F_f(X)\big)$ converge for σ-a.e. f (Step 1) and are dominated by $E \in L^1(\sigma)$ (Step 2); dominated convergence yields the partial derivative of $\mathcal{F}$ and formula (7). Finally, the right-hand side of (7) is continuous at each configuration X with pairwise distinct clouds. Indeed, if $X_r \to X$, then for the full-measure set of slices having no ties at X, continuity of the projected scalars implies $\pi_f(X_r) = \pi_f(X)$ for all sufficiently large r. Hence the integrand in (7) converges pointwise, and the same integrable envelope E allows another application of dominated convergence. Thus, all partial derivatives of $\mathcal{F}$ exist and are continuous near X, and $\mathcal{F}$ is differentiable at X. □

D Statistical Properties

Corollary D.1 (Error of the fully empirical estimator)**.** *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{P}(\Omega))$ be observed through the two-stage scheme with parameters (N, n) and (M, m) respectively, and let $\widehat{\mathrm{VSW}}_{p,L}$ denote the L-slice Monte Carlo estimator of Proposition 3.9. Then*

$$\mathbb{E}\Big|\widehat{\mathrm{VSW}}_{p,L}\big(\widetilde{\mathbb{P}}_{N,n}, \widetilde{\mathbb{Q}}_{M,m}\big) - \mathrm{VSW}_p(\mathbb{P}, \mathbb{Q})\Big| \\ \leq C\Big(N^{-1/(2p)} + n^{-1/2} + M^{-1/(2p)} + m^{-1/2} + L^{-1/(2p)}\Big),$$

with $C = C\big(p, \Lambda_{2p}\big)$.

Proof. By the triangle inequality for VSW_p (Theorem 3.3), $\big|\mathrm{VSW}_p(\widetilde{\mathbb{P}}_{N,n}, \widetilde{\mathbb{Q}}_{M,m}) - \mathrm{VSW}_p(\mathbb{P}, \mathbb{Q})\big| \leq \mathrm{VSW}_p(\mathbb{P}, \widetilde{\mathbb{P}}_{N,n}) + \mathrm{VSW}_p(\mathbb{Q}, \widetilde{\mathbb{Q}}_{M,m})$, and Theorem 3.7 bounds the expectation of each term after taking p-th roots and applying Jensen. For the slicing error, Proposition 3.9 gives $\mathbb{E}|\widehat{\mathrm{VSW}}_{p,L}^p - \mathrm{VSW}_p^p| \leq 2^p(\Lambda_{2p}/L)^{1/2}$ at the empirical meta-measures, and $|a^{1/p} - b^{1/p}| \leq |a-b|^{1/p}$ for $a, b \geq 0$ converts this into a bound on the estimator itself. Combining the three contributions gives the claim. □

Proposition D.2 (Stability in the slicing measure)**.** *Let σ, σ' be two slicing measures on $C(\Omega)$ with finite sup-norm moments of order $2(p-1) \vee 2$, and let $\mathrm{VSW}_p^{(\sigma)}, \mathrm{VSW}_p^{(\sigma')}$ denote the corresponding distances. Then for any coupling Ξ of (σ, σ'),*

$$\big|\mathrm{VSW}_p^{(\sigma)}(\mathbb{P}, \mathbb{Q})^p - \mathrm{VSW}_p^{(\sigma')}(\mathbb{P}, \mathbb{Q})^p\big| \;\leq\; 2p\, \mathbb{E}_{(f,g)\sim\Xi}\Big[\big(2\max(\|f\|_\infty, \|g\|_\infty)\big)^{p-1}\; \|f-g\|_\infty\Big].$$

In particular, if a sequence of finite-rank approximations $f_R \sim \sigma_R$ can be coupled to $f \sim \sigma$ so that $\mathbb{E}\,\|f_R - f\|_\infty^2 \to 0$ and $\sup_R \mathbb{E}\,\|f_R\|_\infty^{2(p-1)} < \infty$ (both automatic from L^2 convergence when $p \leq 2$), then $\mathrm{VSW}_p^{(\sigma_R)} \to \mathrm{VSW}_p^{(\sigma)}$ uniformly on $\mathcal{P}(\mathcal{P}(\Omega)) \times \mathcal{P}(\mathcal{P}(\Omega))$. Lemma D.3 verifies this hypothesis for truncated Karhunen–Loève expansions.

Proof. Write $h(f) := \mathrm{W}_p\big((P_f)_\#\mathbb{P}, (P_f)_\#\mathbb{Q}\big) \in [0, 2\,\|f\|_\infty]$. By Lemma 3.2(ii), $|h(f) - h(g)| \leq 2\,\|f-g\|_\infty$, and $|h(f)^p - h(g)^p| \leq p\max(h(f), h(g))^{p-1}|h(f) - h(g)|$. Take expectations under Ξ and use $\max(h(f), h(g)) \leq 2\max(\|f\|_\infty, \|g\|_\infty)$; the marginals reproduce the two sliced distances. The uniformity claim follows by Cauchy–Schwarz on the right-hand side. □

Lemma D.3 (KL truncations satisfy the coupling hypothesis)**.** *Let κ satisfy Assumption 1, let $(\lambda_k, \varphi_k)_{k\geq 1}$ be the Mercer eigenpairs of κ on $L^2(\Omega, \iota)$ for a fully supported finite Borel measure ι on Ω, and set*

$$f_R \;=\; \sum_{k=1}^{R} \sqrt{\lambda_k}\,\xi_k\,\varphi_k, \qquad \xi_k \overset{iid}{\sim} \mathcal{N}(0,1).$$

Then $\{\sqrt{\lambda_k}\varphi_k\}_{k\geq 1}$ is an orthonormal basis of the Cameron–Martin space $\mathcal{H}_\kappa$ of σ, the series converges almost surely in $(C(\Omega), \|\cdot\|_\infty)$ with limit f distributed as σ, and $\mathbb{E}\,\|f - f_R\|_\infty^2 \to 0$. Consequently $\mathrm{VSW}_p^{(\sigma_R)} \to \mathrm{VSW}_p^{(\sigma)}$ uniformly on $\mathcal{P}(\mathcal{P}(\Omega)) \times \mathcal{P}(\mathcal{P}(\Omega))$ by Proposition D.2.

Proof. By Mercer's theorem on the compact space Ω (with κ continuous and ι fully supported), the eigenfunctions φ_k are continuous and $\{\sqrt{\lambda_k}\varphi_k\}$ is an orthonormal basis of $\mathcal{H}_\kappa$, which coincides with the Cameron–Martin space of the Gaussian measure σ on $C(\Omega)$ (Bogachev, 1998, Ch. 2–3). For any orthonormal basis $\{e_k\}$ of the Cameron–Martin space of a centered Radon Gaussian measure on a separable Banach space, the series $\sum_k \xi_k e_k$ with i.i.d. standard Gaussian coefficients converges almost surely in the Banach norm, with limit distributed as the measure (Bogachev, 1998, Thm. 3.5.1); this is an Itô–Nisio-type theorem. Finally, almost surely convergent Gaussian series in a separable Banach space converge in L^r for every $r < \infty$ (Ledoux & Talagrand, 1991, Ch. 3), so $\mathbb{E}\,\|f - f_R\|_\infty^2 \to 0$; the coupling $\Xi = \mathrm{law}(f_R, f)$ in Proposition D.2 together with Cauchy–Schwarz gives the uniform convergence of the sliced distances. □

E Ablation

Ablation on R. The number of random Fourier features R is specific to VSW. We therefore sweep $R \in \{8, 16, \ldots, 1024\}$ in the dataset-distance setting of Figure 2 on FashionMNIST, at the two largest matched budgets $L = S = 10^3$ and $L = S = 10^4$ (Table 3), with the baseline configurations of Appendix F. Each entry is the mean $\pm$ standard deviation over 5 independent draws of the features and slices, for VSW and both baselines. Runtimes follow the protocol of Figure 3 and are given on the 30 pairs at their own sizes, which is the cost of the correlation beside them, and at fixed sizes 10^4 and 6×10^4. Larger R generally improves correlation at increased runtime, with limited gains beyond $R = 512$. For slicing budget $L = S = 1000$, all tested $R \geq 16$ give higher mean Pearson and Spearman correlations than DSW and s-OTDD; $R = 8$ is below DSW. At $L = S = 10^4$, $R = 128$ is competitive with DSW (Pearson 0.8468 ± 0.0080 versus 0.8396 ± 0.0125), with overlapping variability across random draws. We use $R = 128$ for its balance of correlation and runtime.

Table 3: Ablation on the number of random Fourier features R on FashionMNIST, at the two matched slicing budgets $L = S = 10^3$ (top) and $L = S = 10^4$ (bottom). Each entry is the mean $\pm$ standard deviation over 5 independent draws of the features and slices. Times are per distance evaluation: on the 30 pairs at their own sizes, at a fixed size of 10^4, and on the full training set of 6×10^4.

Method	Correlation with OTDD (↑)		Time per evaluation (ms, ↓)		
	Pearson	Spearman	5–10k	10^4	6×10^4
Matched slicing budget $L = S = 10^3$					
VSW, $R = 8$	0.6030 ± 0.0548	0.6015 ± 0.0695	3.4	3.2	12.8
VSW, $R = 16$	0.7931 ± 0.0362	0.7750 ± 0.0519	3.6	3.8	17.7
VSW, $R = 32$	0.7698 ± 0.0393	0.7623 ± 0.0653	4.6	4.8	25.3
VSW, $R = 64$	0.8044 ± 0.0164	0.8174 ± 0.0309	5.6	6.5	36.4
VSW, $R = 128$	0.8413 ± 0.0115	0.8575 ± 0.0151	8.2	10.4	61.4
VSW, $R = 256$	0.8525 ± 0.0049	0.8767 ± 0.0194	12.7	15.0	130.2
VSW, $R = 512$	0.8730 ± 0.0116	0.8878 ± 0.0051	22.1	25.9	224.3
VSW, $R = 1024$	0.8803 ± 0.0049	0.8897 ± 0.0117	54.9	75.5	401.0
DSW	0.7215 ± 0.0373	0.7356 ± 0.0359	46.9	47.7	141.9
s-OTDD	0.2813 ± 0.1932	0.2948 ± 0.1429	291.7	388.6	2403.4
Matched slicing budget $L = S = 10^4$					
VSW, $R = 8$	0.5971 ± 0.0676	0.6032 ± 0.0817	4.5	4.9	13.6
VSW, $R = 16$	0.7891 ± 0.0374	0.7733 ± 0.0577	5.8	6.3	21.0
VSW, $R = 32$	0.7723 ± 0.0421	0.7691 ± 0.0734	8.6	9.4	29.5
VSW, $R = 64$	0.8056 ± 0.0173	0.8145 ± 0.0322	14.2	15.2	44.5
VSW, $R = 128$	0.8468 ± 0.0080	0.8629 ± 0.0095	25.8	27.2	79.0
VSW, $R = 256$	0.8610 ± 0.0104	0.8842 ± 0.0203	45.3	43.8	158.8
VSW, $R = 512$	0.8768 ± 0.0070	0.8893 ± 0.0116	84.3	82.3	289.0
VSW, $R = 1024$	0.8789 ± 0.0031	0.8938 ± 0.0124	158.9	175.2	527.5
DSW	0.8396 ± 0.0125	0.8385 ± 0.0151	90.6	104.4	438.2
s-OTDD	0.4897 ± 0.0670	0.5305 ± 0.0608	2955.9	3889.8	24193.5

F Implementation Details

Below, we provide experimental details for all of the reported experiments in the paper. All experiments are conducted on an AMD EPYC 7702P CPU, or NVIDIA RTX A5000 GPUs. Unless stated otherwise, VSW uses the Gaussian RBF kernel $\kappa_\ell(x, y) = \exp\big(-\|x-y\|^2/(2\ell^2)\big)$, whose spectral measure in (5) is $\tau_{\kappa_\ell} = \mathcal{N}(0, \ell^{-2} I_d)$. The bandwidth $\ell > 0$ is set by the median heuristic (Gretton et al., 2012). We pool the points of all inner measures of the two meta-measures being compared, draw 2000 of them uniformly at random, and set ℓ to the median of their pairwise Euclidean distances. The bandwidth is computed once, before the slices are drawn.

Statistical Properties. Each meta-measure is a mixture of 3 Gaussian inner measures with uniform weights. The meta-measure is $\mathbb{P} = \frac{1}{3}\sum_{k=1}^{3} \delta_{\mathcal{N}(\mu_k, I_d)}$, a uniform distribution over three Gaussian inner measures whose means μ_k are equally spaced on a circle of radius 4 in the first two coordinates. To sample an empirical meta-measure, we draw N labels $k_1, \ldots, k_N$ independently and uniformly from $\{1, 2, 3\}$, and observe each inner measure $\mathcal{N}(\mu_{k_i}, I_d)$ through n i.i.d. points. Each empirical meta-measure has N inner measures observed through n points each. VSW uses $R = 256$, a median-heuristic bandwidth, and every curve in Figure 1 is the mean over 200 runs. For the Monte Carlo rate, we compare two independent samples with $N = n = 1000$ and report $|\widehat{\mathrm{VSW}}_{2,L} - \widehat{\mathrm{VSW}}_{2,L_{\max}}|$ for $L \le 10^4$, where the reference $L_{\max} = 10^5$ uses the same slices. For the outer and inner rates, both meta-measures are drawn from the same distribution, and we use $L = 10^4$. The outer rate varies N with $n = 1000$. The inner rate varies n with $N = 1000$, and both meta-measures share the same N component labels, so that only the inner sampling error remains.

Optimal Transport Dataset Distance. Following Nguyen et al. (2025), we draw 30 random pairs of disjoint sub-datasets from the training sets of MNIST, FashionMNIST, and CIFAR-10. For each pair, we sample a size uniformly from $\{5000, 6000, \ldots, 10000\}$ and take the first and the last that many indices of a random permutation of the training set. Each sub-dataset is a meta-measure whose inner measures are its class-conditional distributions of flattened images. As a reference, we compute the exact OTDD (Alvarez-Melis & Fusi, 2020) with its default implementation. Following Piening & Beinert (2026), we set the weight of its feature cost to zero, so that OTDD reduces to a Wasserstein-over-Wasserstein distance between the class-conditional distributions. s-OTDD (Nguyen et al., 2025) and DSW (Piening & Beinert, 2026) are used with their default hyper-parameters. We use a vectorized implementation of DSW that returns the same values as the released code under identical random draws, while being optimized so that the runtime comparison remains fair. VSW uses $R = 128$, with a bandwidth set once per pair of sub-datasets by the median heuristic.

The correlations in Figure 2 pool all 30 pairs across the sampled sizes. They are not conditioned on sample size, so shared sample-size dependence can contribute to the reported agreement. This protocol introduces no controlled class-proportion or domain shifts.

For Figure 3, we time one distance evaluation between two sub-datasets whose size ranges from 1000 to the full training set in steps of 5000, for $L = S \in \{10^3, 5 \times 10^3, 10^4\}$. All methods run on CPU. We report the mean and standard deviation over 10 random draws of the sub-datasets.

Comparing Distributions of Point Clouds. We follow the protocol and released code of Piening & Beinert (2026, Sec. 5.3). The reference meta-measure contains $N = 10$ training shapes of one class, and the target contains M test shapes of the same class. We vary one factor at a time: (i) the number of target shapes $M \in \{1, \ldots, 19\}$ (class *bed*, $n = m = 50$); (ii) the standard deviation $\sigma_{\mathrm{noise}} \in [0, 0.5]$ of Gaussian noise added to the target points (class *sofa*, $M = 10$, $n = m = 50$); and (iii) the target resolution $m \in \{50, 100, \ldots, 600\}$ (class *chair*, $M = 10$, $n = 500$). Each setting is repeated over 5 random draws of shapes. VSW uses $L = 2000$ slices with $R = 128$, and bandwidth is set by the median heuristic. For OT-NNA (Yang et al., 2019), WoW, and DSW, we use the implementations and hyperparameters of Piening & Beinert (2026). Our only change is a vectorized DSW implementation, which ensures optimal implementation of DSW and fair runtime comparison. Under identical random draws it returns the same values as the released code, so it reduces the runtime of DSW without changing its values. In Figure 4, each curve is divided by its mean over the sweep, and shaded regions show ± 1 standard deviation over the repetitions.

For Table 1, every method compares the full sets of 662 ShapeNet test chairs (Chang et al., 2015) and 662 PVD samples (Zhou et al., 2021), with 2048 points per cloud. All methods use the settings above

and run on one NVIDIA RTX A5000 GPU. For VSW and DSW, we report the mean $\pm$ standard deviation over 6 timed batches, while OT-NNA and WoW are timed once.

Comparing Image Distributions via Patch Distributions. We follow the protocol and released code of Piening & Beinert (2026, Sec. 5.4). Each meta-measure is a batch of 32 grayscale 64×64 Perlin-noise images (scale 100, pixel values rescaled to $[0, 1]$), and each image is represented by the uniform empirical measure of its 3249 overlapping 8×8 patches in $\mathbb{R}^{64}$. In the persistence sweep, the reference batch has persistence 0.5 (lacunarity 2.5, 5 octaves) and the target persistence varies over $\{0.1, 0.25, 0.5, 0.75, 1\}$. In the lacunarity sweep, the reference has lacunarity 1.5 (persistence 1, 6 octaves) and the target lacunarity varies over $\{1, 1.25, 1.5, 1.75, 2\}$. Reference and target batches are generated from different random seeds, each sweep is repeated 5 times, and all methods are evaluated on the same pairs of batches. VSW uses $L = 2000$ slices with $R = 128$ random Fourier features of a Gaussian RBF kernel, whose bandwidth is set by the median heuristic on the pooled patches. WoW and DSW are implemented as in the point-cloud experiment above. Following the released code, DSW uses 100×100 slices, a 50-point quantile grid, and length scale 0.1, and we again use a vectorized implementation that returns the same values as the released code. In Figure 5, each curve is divided by its maximum within a run before averaging, which preserves the depth of its minimum at the reference parameter. Shaded regions show ± 1 standard deviation over the 5 runs. The runtimes in Figure 5 are for one distance evaluation between two batches. For VSW and DSW, each evaluation is repeated 5 times, and the mean is averaged over all 50 pairs of batches. All runtimes are measured on the CPU.

Gradient Flow on Natural Images. The target meta-measure $\mathbb{Q}$ consists of $N = 5$ ImageNet classes (Deng et al., 2009) (goose, koala, Persian cat, monarch, and macaw), each represented by the uniform empirical measure of its first 20 training images. Each image is resized and centre-cropped to 64×64, scaled to $[0, 1]$, and flattened, so that $\Omega \subset \mathbb{R}^{12288}$. The source meta-measure has the same shape, with $N = 5$ clouds of $n = 20$ particles drawn i.i.d. from $\mathcal{N}(0, I)$. We run the particle scheme of Section 3.4 for 110,000 steps. At every step, we draw fresh slices ($L = 8000$ slices, $R = 4096$ random Fourier features of a Gaussian RBF kernel) and a Monte Carlo estimate of the gradient in Eq. (7) by automatic differentiation. Since each particle carries mass $1/(Nn)$, we multiply the gradient by Nn so that the step size does not depend on the number of particles. The particles are updated by gradient descent with heavy-ball momentum 0.9, and the momentum is reset when the bandwidth changes. The bandwidth follows a two-phase schedule. The large bandwidth first moves each cloud towards a target cloud, and the smaller one then resolves individual images. The full run takes about 6 minutes on one GPU.

In Figure 6 (left), each row follows one particle of one class, and the last column shows its nearest ground-truth image of the matched class in Euclidean distance. For the right panel, we clip all snapshots to $[0, 1]$, embed them with CLIP ViT-L/14 (Radford et al., 2021) using its standard preprocessing, and ℓ_2-normalize the embeddings. We then project them with a 2D PCA fitted on the 100 ground-truth embeddings only, so that all snapshots share the same axes. Each line traces one particle through the snapshots at steps 0, 1k, 2k, 3k, 5k, 8k, 15k, 25k, and 110k.

Parametric Generative Modeling. DSW (Piening & Beinert, 2026) uses 100×100 slices and a 50-point quantile grid, with its GP length scale tuned per dataset (0.03 and 0.003). MMDSW (Bonet et al., 2025b) is the MMD with the sliced-Wasserstein Riesz kernel $-\mathrm{SW}_2^r$, with 128 projections and $r = 1$ and $r = 0.5$, respectively. For DSW (Piening & Beinert, 2026), we again use a vectorized implementation that returns the same values as the reference one. We report the point-level W_2^2 between 4096 generated points, one from each of 4096 generated clouds, and 4096 points from the target, computed with POT (Flamary et al., 2021) and averaged over 2 draws. The ground-truth level is W_2^2 between two independent target samples of the same size. Curves show the mean and standard deviation over 3 seeds, and the scatter plots in Figure 7 show seed 1. The wall-clock is measured on GPU.

Shape Classification. We follow the protocol of Piening & Beinert (2026, Sec. 5.1) and use their released implementations of all baselines. Each shape is a metric-measure space given by its pairwise distance matrix, which we map to a meta-measure on $\Omega \subset \mathbb{R}$: the outer measure is uniform over the points, and the inner measure of point i is the empirical distribution of its distances to all points, i.e., row i of the distance matrix. For every method, we compute the full matrix of distances between shapes and report the mean and standard deviation of the 3-NN accuracy over 1000 random stratified $25\%/75\%$ train/test splits. FAUST-500 and FAUST-1000 use the 100 MPI-FAUST training

registrations (Bogo et al., 2014). For 2D Shapes (Beier et al., 2022) and Animals (Sumner & Popović, 2004), we use the preprocessed data released by Beier et al. (2022). For MNIST-2000, we follow the construction of Piening & Beinert (2026): for each digit 0–4, we build 20 shapes by sampling 2000 points from the mean training image proportionally to pixel intensity, adding Gaussian noise with standard deviation 3 pixels, and applying a random rotation, with Euclidean distances.

VSW uses $L = 100$ and $R = 16$. SQW and STLB use 100 projections and a 10-point quantile grid (length scale 0.01 for SQW), and TLB, AE, and GW are used as in the released code, without a feature term. To compare the methods fairly, VSW, SQW, and STLB each draw a single random basis for the whole distance matrix, so that every shape is projected once rather than once per pair. AE caches its per-shape energy self-terms, which changes its distances by less than 0.2%. TLB and GW are deterministic and admit no such amortization. Runtimes in Table 2 are the mean time per distance on CPU.