LEARNING WITH EMPIRICAL MEASURES: PRACTICAL TRANSFORMERS WITH UNIVERSALITY GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose the *Quantile Transformer*, a transformer-based architecture for learning with empirical measures. This architecture alleviates the computational burden of existing transformer layers for this task, which either require all-to-all particle interactions or rely on expensive optimal transport subprocedures. For measures on $\mathbb{R}$, we encode each empirical measure as a finite-dimensional token derived from order statistics and establish universal approximation guarantees for continuous maps on compact classes. For measures on $\mathbb{R}^d$, we extend both the construction and the universality result to compact classes of elliptical distributions using finitely many one-dimensional projections. Preliminary synthetic experiments validate our approach in terms of both learning performance and computational efficiency.

1 INTRODUCTION

This paper proposes a transformer-based architecture for learning sequence-to-sequence maps over $\mathcal{P}_2(\mathbb{R}^d)$ from finite samples, and establishes its ability to approximate continuous functions on compact measure sets. Specifically, for a sequence of T probability measures $\{\mu_t\}_{t=1}^T \subset \mathcal{P}_2(\mathbb{R}^d)$ observed through their M-atom empirical counterparts $\{\widehat{\mu}_{t,M}\}_{t=1}^T \subseteq \mathcal{E}_M(\mathbb{R}^d)$, we target architectures $\mathcal{T}_{M,\theta} : \mathcal{E}_M(\mathbb{R}^d)^T \to \mathcal{E}_M(\mathbb{R}^d)^T,\ \theta \in \mathbb{R}^p$ capable of ensuring arbitrary closeness to some continuous $\varphi : \mathcal{K}^T \to \mathcal{K}^T$[1], as measured by

$$\max_{s\in\{1,\dots T\}} \mathrm{W}_2\left(\ \mathcal{T}_{M,\theta}\left(\{\widehat{\mu}_{t,M}\}_{t=1}^T\right)_s,\ \varphi(\{\mu_t\}_{t=1}^T)_s\ \right), \tag{1}$$

where $\mathcal{K} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ is W_2-compact and W_2 is the Wasserstein-2 distance. We balance practical concerns with theoretical guarantees by: i) controlling the forward pass complexity relative to M; ii) avoiding optimal transport subproblems during training or prediction; iii) ensuring consistency when $M \to \infty$, and iv) ensuring high-probability approximation universality for several measure classes.

Learning with distributional inputs is important when the objects of study are accessible only through samples. For example, in cellular dynamics and socio-economic systems, technological constraints (Macosko et al., 2015) and privacy concerns (Dwork & Roth, 2014) preclude longitudinal tracking, and dynamics must be inferred from sequences of distribution-valued cross-sectional measurements. More generally, distributions can be used to summarize large collections of observations without renouncing knowledge of within-group variation (Billard & Diday, 2003; Beranger et al., 2023). A prominent such example is AlphaFold (Jumper et al., 2021; Abramson et al., 2024), where distributions are used to model inter-residue distances for guiding protein structure prediction.

Related work. Existing measure-based approaches typically equate tokens with empirical measure atoms, placing expressive modeling at odds with practical sequence processing: this is the case for early measure-to-measure networks (De Bie et al., 2019), mean-field and measure-valued formulations of self-attention (Geshkovski et al., 2024; Biswal et al., 2024; Castin et al., 2025; Bach et al., 2025), and recent practical transformers (Vandergrift et al., 2026). At finite sample size, this particle-level representation entails quadratic interactions in M, and joint exposure to attention of T measures incurs an impractical $\mathcal{O}\left(T^2M^2d\right)$ cost. In particular for self attention, the interacting-particle viewpoint was introduced by Lu et al. (2019); Vuckovic et al. (2020), with the latter formulating self-attention as a nonlinear operator on probability measures and studying its Wasserstein regularity and stability.

[1]For notational simplicity, we take identical input and output classes. The universality arguments extend to continuous maps $\varphi : \mathcal{K}_{\mathrm{in}}^T \to \mathcal{K}_{\mathrm{out}}^T$ between distinct compact subclasses of the measure class under consideration.

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.

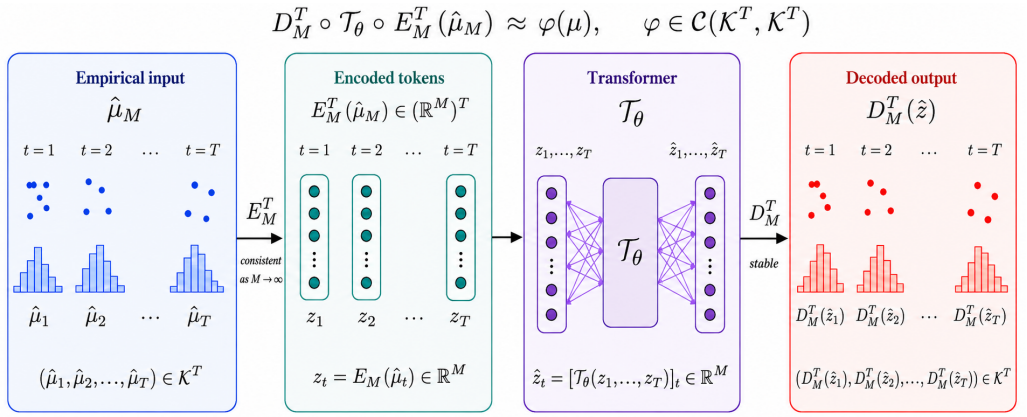

Figure 1: Conceptual pipeline for mapping sequences of empirical measures to sequences of empirical measures.

A separate line of work studies the continuous-depth regime of interacting particle transformers: Geshkovski et al. (2024); Rigollet (2026) model tokens as particles whose empirical measure evolves via a nonlinear continuity equation in the forward pass. Their simplified architecture yields a permutation-equivariant measure flow on $\mathbb{S}^{d-1}$, studied primarily for its long-time (large depth) clustering behavior. Subsequently, Geshkovski et al. (2026) show that a single such transformer augmented with ReLU feed-forward layers can interpolate finitely many measure-to-measure pairs on $\mathbb{S}^{d-1}$. Their inputs, however, evolve as separate initial conditions under a shared map, rather than being jointly coupled by attention.

Finally, Vandergrift et al. (2026) tackle measure-to-measure regression on independent input–output pairs, using transformers to learn nonlinear measure-dependent pushforwards. Their architecture similarly acts on the particles of a single measure, handling temporality by regressing on pairs of consecutive measures rather than joint modeling. Together, these works expose a vacant middle ground of theoretically backed yet still practical sequence-to-sequence architectures, which our paper addresses. Further related work is deferred to Appendix A and section-level discussions.

Contributions. We propose finite-dimensional measure representations that are compatible with W_2-geometry and consistent in the population limit. These representations form the basis of the *Quantile Transformer*, our architecture that casts learning over sequences of measures as Euclidean token prediction. A standard transformer can then be trained with Euclidean losses in this token space. Technically, we exploit the isometric representation of $(\mathcal{P}_2(\mathbb{R}), W_2)$ by quantile functions together with finite-dimensional projections onto step-function spaces, to represent each empirical measure by its normalized order statistics. We then extend this approach to elliptical measures in $\mathcal{P}(\mathbb{R}^d)$, where finitely many one-dimensional projections were shown to be measure-identifying in a recent Cramér–Wold-type result (Fraiman et al., 2023).

Consequently, attention operates over measures rather than particles, yielding a forward-pass complexity of $\mathcal{O}\left(T^2 M d^2\right)$ after a one-time offline preprocessing cost of $\mathcal{O}\left(TM \log(M) d^2\right)$. We highlight the modularity of our construction: geometry-preserving encoders and stable decoders transfer approximation results from Euclidean space transformers into probability measure spaces, while remaining compatible with standard training pipelines. Figure 1 summarizes the architecture, and our theoretical contributions (supported by preliminary experiments) are given below.

- **C1.** In Section 2.1, we formalize general encoder–decoder conditions for representing measures as finite-dimensional tokens and importing approximation universality from Euclidean sequence models. The framework is modular and allows alternative computable encoder–decoder pairs.
- **C2.** In Theorem 3.1, we establish high-probability approximation universality for continuous maps over compact subsets of $\mathcal{P}_2(\mathbb{R})^T$. Corollary 3.1 strengthens this to universality in expectation via clipping before decoding, while Proposition 3.2 provides explicit particle-count thresholds for the sampling and finite-resolution projection errors under standard regularity assumptions.
- **C3.** In Theorem 4.1, we extend the high probability universality result to the class of continuous functions over compact subsets of elliptical measures in $\mathcal{P}(\mathbb{R}^d)^T$.

2 PRELIMINARIES AND PROBLEM SETTING

Notation. We denote vectors and matrices by bold lowercase and uppercase letters, respectively, and scalars by regular lowercase letters. We write $\boldsymbol{e}_i$ for the ith canonical basis vector, and $\boldsymbol{I}_d$ for the identity matrix. Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices. We use $\langle\cdot,\cdot\rangle$ for the inner product indexed by the corresponding space. The cone of symmetric PSD matrices in $\mathbb{R}^{d\times d}$ is denoted by $\mathbb{S}^d_+$. We let $[M]$ be the set $\{1,\dots,M\}$, and $\mathfrak{S}_M \coloneqq \{\pi \mid \pi : [M] \to [M]\}$ be the permutation group over $[M]$. The notation 1_A gives the indicator function of a set A. We write W_p for the Wasserstein-p distance, $W_p(\mu,\nu) \coloneqq \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \, d\gamma(x,y)\right)^{1/p}$, where $\Gamma(\mu,\nu)$ gives all couplings between μ and ν. We use $\mathcal{P}_p(\mathbb{R}^d)$ for the set $\mathcal{P}_p(\mathbb{R}^d) = \left\{\mu \mid \int_{\mathbb{R}^d} \|x\|^p \, d\mu(x) < \infty\right\}$. $\mathcal{C}_{W_p}(\mathcal{X},\mathcal{Y})$ defines continuous maps $\mathcal{X} \mapsto \mathcal{Y}$, with product Wasserstein spaces carrying the product W_p topology.

Problem setting. We focus on learning functions $\varphi \in \mathcal{C}_{W_2}(\mathcal{K}^T, \mathcal{K}^T)$, where $\mathcal{K}$ satisfies the following assumption which is equivalent to W_2-compactness in $\mathcal{P}_2(\mathbb{R}^d)$.

Assumption 2.1. *The set $\mathcal{K} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ is W_2-closed and* $\lim_{R\to\infty} \sup_{\mu\in\mathcal{K}} \int_{\{\|x\|>R\}} \|x\|^2 \, d\mu(x) = 0$.

In our scenario where measures are accessed via samples, consider $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and a sequence $X_1, X_2, \dots \overset{\text{i.i.d.}}{\sim} \mu$. We define the equally weighted M-atom empirical measure as

$$\widehat{\mu}_M \coloneqq \frac{1}{M}\sum_{i=1}^{M} \delta_{X_i} \in \mathcal{E}_M(\mathbb{R}^d), \tag{2}$$

where $\mathcal{E}_M(\mathbb{R}^d) \coloneqq \{M^{-1}\sum_{i=1}^M \delta_{x_i} \mid x_i \in \mathbb{R}^d\}$. Letting $\boldsymbol{\mu} \coloneqq [\mu_1, \dots \mu_T]$ and $\widehat{\boldsymbol{\mu}}_M \coloneqq [\widehat{\mu}_{1,M}, \dots \widehat{\mu}_{T,M}]$, our work seeks approximation guarantees in terms of $W_2(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M), \varphi(\boldsymbol{\mu}))$. We focus on W_2 since our encoder–decoder schemes map the metric geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ to the Euclidean ℓ_2 geometry of standard transformers. However, the construction and qualitative universality results extend to W_p and ℓ_p spaces for $1 \le p < \infty$ with minor modifications.

Standard transformer architecture. Let $\mathcal{T}_\theta$ denote a standard unmasked softmax transformer (Vaswani et al., 2017), which was shown to be universal for continuous permutation-equivariant maps on compact product domains (Alberti et al., 2023, Theorem 4.3). For our purposes, we extend the result of Alberti et al. (2023, Theorem 4.3) with positional labels to obtain universality for ordered sequence maps — the result is stated below, with proof deferred to Appendix C.2.

Corollary 2.1. *Fix $T, q \in \mathbb{N}_+$, let $\mathcal{U} \subseteq \mathbb{R}^q$ be a nonempty compact set, and let $G : \mathcal{U}^T \longrightarrow (\mathbb{R}^q)^T$ be continuous. Then, for every $\eta > 0$, there exists a position-augmented unmasked softmax transformer $\mathcal{T}_\theta : \mathcal{U}^T \longrightarrow (\mathbb{R}^q)^T$ such that* $\sup_{\boldsymbol{z}\in\mathcal{U}^T} \max_{s\in[T]} \| \mathcal{T}_\theta(\boldsymbol{z})_s - G(\boldsymbol{z})_s \|_2 < \eta$.

2.1 TRANSFORMERS FOR EMPIRICAL MEASURES — AN ABSTRACT FRAMEWORK

We aim to reduce learning with empirical measures to standard finite-dimensional sequence learning while preserving population-level guarantees as $M \to \infty$. The construction requires three ingredients.

Step 1: an identifiable population representation. Fix a compact set $\mathcal{K} \subseteq \mathcal{P}_2(\mathbb{R}^d)$, and choose a separable Hilbert representation space $\mathcal{Z}$ and a continuous map $\Phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{Z}$ whose restriction $\Phi|_{\mathcal{K}}$ is injective[2]. Compactness then guarantees that $\Phi|_{\mathcal{K}}$ is a homeomorphism onto $\mathcal{Q} \coloneqq \Phi(\mathcal{K})$. Thus Φ preserves both underlying measure identifiability and W_2-topology, allowing any $\varphi \in C_{W_2}(\mathcal{K}^T, \mathcal{K}^T)$ to be represented as the continuous map $\widetilde{\varphi} : \mathcal{Q}^T \to \mathcal{Q}^T$, with $\widetilde{\varphi} \coloneqq \Phi^T \circ \varphi \circ (\Phi^{-1})^T$.

Step 2: consistent finite-dimensional encoding and stable decoding. First, choose a deterministic collection of finite-dimensional subspaces $\{\mathcal{Z}_r \subseteq \mathcal{Z}\}_{r\ge1}$ indexed by their dimension, $\dim(\mathcal{Z}_r) = r$, with orthogonal projections $\Pi_r : \mathcal{Z} \to \mathcal{Z}_r$[3] satisfying

$$\lim_{r\to\infty} \| z - \Pi_r z \|_{\mathcal{Z}} = 0, \qquad \forall z \in \mathcal{Z}. \tag{3}$$

[2]The pair $(\mathcal{Z}, \Phi)$ is guaranteed to exist. Indeed, by (Villani et al., 2009, Theorem 6.18) $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is Polish and hence separable and metrizable, while every separable metrizable space is homeomorphic to a subspace of a separable Hilbert space (Nagata, 1985, Theorem VI.10).

[3]Since $\mathcal{Z}_r$ is finite-dimensional, it is closed in $\mathcal{Z}$ and by the Hilbert projection theorem (Rudin, 1991, Theorem 12.4), the orthogonal projection Π_r exists uniquely.

Such a collection exists whenever $\mathcal{Z}$ is separable. For M-atom empirical measures, we let r depend on M, such that $\lim_{M\to\infty} r_M = \infty$. Then, to identify $\mathcal{Z}_r$ with $\mathbb{R}^r$, fix an orthonormal basis $(e_{r,j})_{j=1}^r$ of $\mathcal{Z}_r$ and define $C_r : \mathcal{Z}_r \to \mathbb{R}^r$ as $C_r(z) \coloneqq \big(\langle z, e_{r,1}\rangle_{\mathcal{Z}}, \ldots, \langle z, e_{r,r}\rangle_{\mathcal{Z}}\big)$. By Parseval's identity (Rudin, 1987, Theorem 4.18), C_r is a linear isometric isomorphism.

With these choices, we define the finite-resolution population encoder $\overline{E}_{r_M} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^{r_M}$ as $\overline{E}_{r_M} \coloneqq C_{r_M} \circ \Pi_{r_M} \circ \Phi$ and denote its restriction to empirical measures as $E_{r_M} \coloneqq \overline{E}_{r_M}|_{\mathcal{E}_M(\mathbb{R}^d)}$. Remark that since E_{r_M} is a function of the empirical measure, it is permutation-invariant in the sample order. The following encoder consistency result ensues (proof deferred to Appendix C.1).

Proposition 2.1. *Let $\lim_{M\to\infty} r_M = \infty$ and fix an arbitrary $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.*

(a) If $\mathcal{K} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ is compact, then $\lim_{M\to\infty} \sup_{\mu\in\mathcal{K}} \big\|C_{r_M}^{-1}\overline{E}_{r_M}(\mu) - \Phi(\mu)\big\|_{\mathcal{Z}} = 0$.

(b) Sample $\widehat{\mu}_M \sim \mu$ as in (2). Then, almost surely, $\lim_{M\to\infty} \big\|C_{r_M}^{-1}E_{r_M}(\widehat{\mu}_M) - \Phi(\mu)\big\|_{\mathcal{Z}} = 0$.

In words, the discretized population representation uniformly recovers the true representation, and the empirical representation is consistent. Note that finite-M injectivity is not required of Φ.

Second, choose a decoder defined on the entire space, $D_{r_M} : \mathbb{R}^{r_M} \to \mathcal{E}_M(\mathbb{R}^d)$, and which is asymptotically stable around finite-resolution population representations. Formally, we require

$$\lim_{\eta\downarrow 0} \limsup_{M\to\infty} \sup_{\substack{\mu\in\mathcal{K},\ z\in\mathbb{R}^{r_M} \\ \|z-\overline{E}_{r_M}(\mu)\|_2\le\eta}} W_2(D_{r_M}(z), \mu) \;=\; 0. \tag{4}$$

Uniformly Lipschitz decoders satisfying $\lim_{M\to\infty} \sup_{\mu\in\mathcal{K}} W_2\big(D_{r_M}(\overline{E}_{r_M}(\mu)), \mu\big) = 0$ are a sufficient special case. Intuitively, decoders must ensure that closeness in $\mathbb{R}^{r_M}$ transfers to $\mathcal{P}_2(\mathbb{R}^d)$.

Step 3: target function approximation and universality. Define the measure-based transformer $\mathcal{T}_{M,\theta} : \mathcal{E}_M(\mathbb{R}^d)^T \to \mathcal{E}_M(\mathbb{R}^d)^T,\ \theta \in \mathbb{R}^p$ as the composition

$$\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M) \;\coloneqq\; (D_{r_M} \circ \mathcal{T}_\theta \circ E_{r_M})(\widehat{\boldsymbol{\mu}}_M), \tag{5}$$

where $\mathcal{T}_\theta$ is a standard Euclidean transformer. An ideally-trained $\mathcal{T}_\theta$ predicts $\widehat{\boldsymbol{z}} \in \overline{E}_{r_M}(\mathcal{K})$ for which an $f \in \mathcal{Q}$ exists such that $\widehat{\boldsymbol{z}} = (C_{r_M} \circ \Pi_{r_M} \circ \widetilde{\varphi})\,(f)$. However, this need not be the case in practice (especially during early training), and any bound on (1) must quantify the distance to $\overline{E}^T_{r_M}(\mathcal{K})$. To this end, we proceed in three steps: first, we define a metric over the product space $\mathcal{Z}^T$,

$$\mathrm{dist}_{T,\mathcal{Z}}(\boldsymbol{q}, \boldsymbol{p}) \;\coloneqq\; \max_{s\in[T]} \mathrm{dist}_{\mathcal{Z}}(\boldsymbol{q}_s, \boldsymbol{p}_s), \quad \forall \boldsymbol{q}, \boldsymbol{p} \in \mathcal{Z}^T, \tag{6}$$

and build an α-controlled closed neighborhood $\mathcal{Q}_\alpha,\ \alpha > 0$, of $\mathcal{Q}$ relative to (6) on which a representation's departure from $\overline{E}^T_{r_M}(\mathcal{K})$ is quantifiable. We then define a continuous surrogate $\widetilde{\varphi}_\alpha : \mathcal{Q}^T_\alpha \to \mathcal{Z}^T$ that remains faithful to $\widetilde{\varphi}$ on $\mathcal{Q}_\alpha$. Specifically, whenever $\mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}) \le \alpha$ for $\boldsymbol{q} \in \mathcal{Q}^T,\ \boldsymbol{h} \in \mathcal{Q}^T_\alpha$, it holds that $\mathrm{dist}_{T,\mathcal{Z}}(\widetilde{\varphi}_\alpha(\boldsymbol{h}),\ \widetilde{\varphi}(\boldsymbol{q})) \le \omega(\alpha)$ for some $\omega : \mathbb{R} \to \mathbb{R}$ satisfying $\lim_{r\to 0} \omega(r) = 0$.

Second, we define an event granting our mathematical wish that $\Phi^T(\widehat{\boldsymbol{\mu}}_M) \in \mathcal{Q}^T_\alpha$, namely

$$\widetilde{\mathcal{A}}_{M,\alpha}(\boldsymbol{\mu}) \;\coloneqq\; \Big\{ \mathrm{dist}_{T,\mathcal{Z}}\Big(\big(C_{r_M}^{-1} \circ E_{r_M}\big)^T(\widehat{\boldsymbol{\mu}}_M),\ \Phi^T(\boldsymbol{\mu})\Big) \le \alpha \Big\}. \tag{7}$$

On this event, $E_{r_M}(\widehat{\mu}_M) \in C_{r_M}(\mathcal{Z}_{r_M} \cap \mathcal{Q}_\alpha)$ and we invoke Corollary 2.1 for the existence of a transformer $\mathcal{T}_\theta$ approximating the compactly-supported map $G_{r_M,\alpha} \coloneqq C^T_{r_M} \circ \Pi^T_{r_M} \circ \widetilde{\varphi}_\alpha \circ (C^{-1}_{r_M})^T : C^T_{r_M}\big(\mathcal{Z}^T_{r_M} \cap \mathcal{Q}^T_\alpha\big) \to (\mathbb{R}^{r_M})^T$ up to a desired accuracy $\eta > 0$.

Finally, having chosen $\mathcal{T}_\theta$ on the event (7), contractivity of Π_{r_M} and the consistency of $\widetilde{\varphi}_\alpha$ on $\mathcal{Q}_\alpha$ ensure that $\max_{s\in[T]} \Big\| \mathcal{T}_\theta(E^T_{r_M}(\widehat{\boldsymbol{\mu}}_M))_s - \overline{E}^T_{r_M}(\varphi(\boldsymbol{\mu}))_s \Big\|_2 \le \eta + \omega(\alpha)$. Condition (4) then converts this token-space approximation into W_2-closeness of the decoded measures. Choosing α and η appropriately small and M sufficiently large yields an arbitrarily small measure-space error, while encoder consistency drives the probability of (7) to one over $\mathcal{K}^T$.

Discussion. The main advantage of our framework is modularity. First, any universality result analogous to Corollary 2.1 for a sequence-to-sequence architecture transfers to measure space with high probability. This reduces the need for the simplified, measure-specific architectures of prior

works (Geshkovski et al., 2026; De Bie et al., 2019), and can leverage standard training pipelines. Second, our framework is flexible in its choice of the encoder-decoder pair, provided it satisfies the properties above. Finally, decoupling decoding from the trainable component $\mathcal{T}_\theta$ isolates the return to measure space in the inference step, accounting for the likely complexity of this procedure.

A parallel, modular approach to measure prediction is proposed by Yang et al. (2026), based on kernel mean embeddings encoding and Gaussian mixture decoding. However, their framework predicts parametric marginal densities rather than empirical measures, and does not have formal guarantees.

We likewise considered kernel mean embeddings for instantiating our abstract framework. Indeed, universal kernels such as RBF yield injective embeddings (Sriperumbudur et al., 2011), which can be approximated by finite-dimensional Random Fourier Features averaged over samples (Rahimi & Recht, 2007). Guarantee-backed efficient decoding, however, is challenging: distributional pre-image methods can involve non-convex optimization, while kernel herding requires iterative optimization over particle locations (Muandet et al., 2017, Section 3.8). Since we seek practicality in addition to formal guarantees, we instead adopt quantile function-based measure representations, which preclude such drawbacks. We now develop this approach for measures in $\mathcal{P}(\mathbb{R})$ and elliptical classes in $\mathcal{P}(\mathbb{R}^d)$.

3 The case of measures over $\mathbb{R}$

We fix a compact $\mathcal{K} \in \mathcal{P}_2(\mathbb{R})$ and, as per Section 2.1, choose $(\Phi, \mathcal{Z})$ defining the representation homeomorphism. We settle on $\mathcal{Z} = L^2(0,1) =: L^2$ mapped from $\mathcal{P}_2(\mathbb{R})$ by the quantile function $\Phi : \mathcal{P}_2(\mathbb{R}) \to L^2$ with $\Phi(\mu)(z) := \inf\{x \in \mathbb{R} \mid F_\mu(x) \geq z\}$[4]. It is well-known[5] that Φ is an isometry between $(\mathcal{P}_2(\mathbb{R}), W_2)$ and $(L^2, \|\cdot\|_{L^2})$ and therefore injective on $\Phi|_{\mathcal{K}}$ since

$$W_2(\mu,\nu) \;=\; \|\,\Phi(\mu) - \Phi(\nu)\,\|_{L^2}\,. \tag{8}$$

For the collection of finite-dimensional subspaces $\{\mathcal{Z}_r \subseteq \mathcal{Z}\}_{r\geq 1}$ we let $r = M$ and choose $\mathcal{S}_M$ to be the M-dimensional space of functions that are constant on each cell of the $(0,1)$ partition

$$I_{i,M} \;:=\; \Big(\frac{i-1}{M}, \frac{i}{M}\Big], \qquad i \in [M], \tag{9}$$

given by $\mathcal{S}_M := \{f \in L^2 \mid f = \sum_{i=1}^M a_i 1_{I_{i,M}}, a_i \in \mathbb{R}\}$. We define projections by the operator $\Pi_M : L^2 \to \mathcal{S}_M$ as

$$\Pi_M(g) \;:=\; \sum_{i=1}^{M} \Big(M \int_{I_{i,M}} g(u)\,du\Big)\mathbf{1}_{I_{i,M}}. \tag{10}$$

The next result (proven in Appendix B.2) confirms that the projection Π_M fulfils condition (3).

Proposition 3.1 (Properties of Π_M)**.** *The operator Π_M defined in (10) is an orthogonal projection onto $\mathcal{S}_M$ and $\lim_{M\to\infty} \|\,\Pi_M g - g\,\|_{L^2} = 0$, for every $g \in L^2$.*

Finally, we define the coordinate isometry $C_M : \mathcal{S}_M \to \mathbb{R}^M$ extracting the support of $f \in \mathcal{S}_M$ as $C_M(f) = \frac{1}{\sqrt{M}}[a_1, a_2, \ldots a_M]$, and characterize its properties in Proposition D.1.

Empirical token encoding and decoding. For an empirical measure $\widehat{\mu}_M$ realized on atoms $\{x_i\}_{i=1}^M$, consider the sorting permutation $\pi_\uparrow \in \mathfrak{S}_M$ for which $x_{\pi_\uparrow(1)} \leq \cdots \leq x_{\pi_\uparrow(M)}$. Further, let $\mathcal{S}_M^\uparrow := \{f \in \mathcal{S}_M : a_1 \leq a_2 \leq \cdots \leq a_M\}$ be the monotone cone of $\mathcal{S}_M$. Then, $\Phi(\widehat{\mu}_M) \in \mathcal{S}_M^\uparrow$ is given by $\Phi(\widehat{\mu}_M) = \sum_{i=1}^M x_{\pi_\uparrow(i)} 1_{I_{i,M}}$ and the empirical embedding amounts to sorting coordinates[6]

$$E_M(\widehat{\mu}_M) \;:=\; (C_M \circ \Phi)(\widehat{\mu}_M) \;=\; \frac{1}{\sqrt{M}}\big(x_{\pi_\uparrow(1)}, x_{\pi_\uparrow(2)}, \ldots, x_{\pi_\uparrow(M)}\big)^\top. \tag{11}$$

We define D_M starting from the inverse coordinate map $C_M^{-1} : \mathbb{R}^M \to \mathcal{S}_M$ with $C_M^{-1}(z) := \sqrt{M}\sum_{i=1}^M z_i 1_{I_{i,M}}$. We then return to measure space through $\Delta_M : \mathcal{S}_M \to \mathcal{P}_2(\mathbb{R})$,

[4] Note that $\Phi(\mu)$ is the inverse of the cumulative distribution function $F_\mu(x) = \mu((-\infty, x])$.

[5] See Villani (2003, Theorem 2.18).

[6] Projections are implicit in (11), since $\Phi(\widehat{\mu}_M)$ already belongs to $\mathcal{S}_M$.

for $\Delta_M(f) \coloneqq \frac{1}{M}\sum_{i=1}^{M}\delta_{a_i}$. The empirical token decoder therefore takes the form

$$D_M(z) \;\coloneqq\; (\Delta_M \circ C_M^{-1})(z) = \frac{1}{M}\sum_{i=1}^{M}\delta_{\sqrt{M}z_i}. \tag{12}$$

It straightforwardly follows from definitions that D_M is stable, since it is 1-Lipschitz,

$$W_2^2(D_M(z), D_M(w)) \le \frac{1}{M}\sum_{i=1}^{M}\left|\sqrt{M}z_i - \sqrt{M}w_i\right|^2 = \|z-w\|_2^2, \tag{13}$$

with equality holding over the monotone cone of $\mathbb{R}^M$ where $D_M(E_M(\widehat{\mu}_M)) = \frac{1}{M}\sum_{i=1}^{M}\delta_{x_{\pi_\uparrow(i)}}$.

Target function approximation and universality. Following Step 3 of Section 2.1, we define the product metric $\mathrm{dist}_T(\boldsymbol{p},\boldsymbol{q}) \coloneqq \max_{s\in[T]} \|\,\boldsymbol{p}_s - \boldsymbol{q}_s\,\|_{L^2}$ and build a continuous surrogate $\widetilde{\varphi}_\alpha : \mathcal{Q}_\alpha^T \to L^{2^T}$ using α-nets. We show that whenever $\mathrm{dist}_T(\boldsymbol{h},\boldsymbol{q}) \le \alpha$ for $\boldsymbol{q}\in\mathcal{Q}^T$ and $\boldsymbol{h}\in\mathcal{Q}_\alpha^T$, it holds that $\mathrm{dist}_T(\widetilde{\varphi}_\alpha(\boldsymbol{h}),\ \widetilde{\varphi}(\boldsymbol{q})) \le \omega_{\widetilde{\varphi}}(4\alpha)$, where $\omega_{\widetilde{\varphi}}$ is the continuity modulus of $\widetilde{\varphi}$. The construction and respective proofs are delegated to Appendix D.3.

For the remaining sections, we consider a stronger (and more natural) version of event (7),

$$\mathcal{A}_{M,\alpha}(\boldsymbol{\mu}) \;\coloneqq\; \left\{\max_{t\in[T]} W_2(\widehat{\mu}_{t,M},\mu_t) \le \alpha\right\}. \tag{14}$$

Indeed, Hallin et al. (2021) show the probability of (14) approaches one uniformly over $\mathcal{K}$ as $M\to\infty$. With the machinery in place, we obtain the following universality result, proven in Appendix D.6.

Theorem 3.1 (High-probability approximation universality in $\mathcal{P}(\mathbb{R})$)**.** *Let $\mathcal{K}\subseteq\mathcal{P}_2(\mathbb{R})$ be a non-empty compact and let $\varphi\in\mathcal{C}_{W_2}(\mathcal{K}^T,\mathcal{K}^T)$. For every $\varepsilon>0$ and every $\delta>0$, there exists $M_0\in\mathbb{N}$ such that, for every $M\ge M_0$, there exists θ and a transformer-based map of kind (5) ensuring*

$$\sup_{\boldsymbol{\mu}\in\mathcal{K}^T}\mathbb{P}_{\boldsymbol{\mu}}\left(\max_{s\in[T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) > \varepsilon\right) < \delta.$$

Interestingly, an additional clipping layer preceding D_M allows us to derive the stronger statement below: uniform universal approximation in expectation. The proof is deferred to Appendix D.7.

Corollary 3.1 (Expectation approximation universality in $\mathcal{P}(\mathbb{R})$)**.** *Let $\mathcal{K}\subseteq\mathcal{P}_2(\mathbb{R})$ be compact, $\varphi\in\mathcal{C}_{W_2}(\mathcal{K}^T,\mathcal{K}^T)$, and $B_{\mathcal{K}} \coloneqq \max_{\nu\in\mathcal{K}} W_2(\nu,\delta_0)$. For $B\ge 0$, let $P_{B,M}:\mathbb{R}^M\to\mathbb{R}^M$ denote the Euclidean projection onto $\{z\in\mathbb{R}^M : \|\,z\,\|_2\le B\}$. Then, for every $\varepsilon>0$, there exists $M_0\in\mathbb{N}$ such that, for every $M\ge M_0$, there exists a clipped transformer $\mathcal{T}^{\mathrm{clip}}_{M,\theta} \coloneqq D_M^T\circ P^T_{B_{\mathcal{K}},M}\circ\mathcal{T}_\theta\circ E_M^T$ satisfying $\sup_{\boldsymbol{\mu}\in\mathcal{K}^T}\mathbb{E}_{\boldsymbol{\mu}}\left[\max_{s\in[T]} W_2\left(\mathcal{T}^{\mathrm{clip}}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s,\varphi(\boldsymbol{\mu})_s\right)\right] < \varepsilon$.*

Sampling thresholds. The proof of Theorem 3.1 uses thresholds M_1 and M_2 to control, respectively, the sampling event $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C$ and the projection error, after choosing the neighborhood-defining α. Proposition 3.2 makes the sampling's dependence on $\alpha,\varepsilon,\delta$ explicit under three regularity assumptions (proof deferred to Appendix D.8).

Proposition 3.2. *Fix $\alpha,\varepsilon>0$ and $\delta\in(0,1)$. Under any one of the three alternative assumptions below, there exist $M_1,M_2\in\mathbb{N}_+$ such that $\sup_{\boldsymbol{\mu}\in\mathcal{K}^T}\mathbb{P}_{\boldsymbol{\mu}}\left(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C\right)\le\delta$, for every $M\ge M_1$, and $\sup_{\mu\in\mathcal{K}}\|\,\Pi_M\Phi(\mu)-\Phi(\mu)\,\|_{L^2}\le\frac{\varepsilon}{3}$, for every $M\ge M_2$.*

- *(a) Suppose that, for some $D>0$, every $\mu\in\mathcal{K}$ satisfies $\mathrm{diam}(\mathrm{supp}\,\mu)\le D$. Then $M_1 = O\left(D^4/\alpha^4\log\left(T/\delta\right)\right)$, and $M_2 = O\left(D^2/\varepsilon^2\right)$.*
- *(b) Suppose that, for some $\sigma>0$, $\mathcal{K}$ is uniformly σ-sub-Gaussian and that $0<\alpha\le\sigma$. Then $M_1 = O\left(\sigma^4/\alpha^4\log\left(T/\delta\right)\right)$, and $M_2 = O\left(\sigma^2/\varepsilon^2\log\left(\sigma/\varepsilon\right)\right)$.*
- *(c) Suppose that, for some $L>0$, the quantile functions are uniformly L-Lipschitz. Then $M_1 = O\left(L^2/\alpha^2\log\left(T/\delta\right)\right)$, and $M_2 = O\left(L/\varepsilon\right)$.*

Computational complexity. Quantile ordering maps point clouds to M-dimensional Euclidean vectors at a one-time $\mathcal{O}(M\log M)$ preprocessing cost per measure. Since we avoid all-to-all particle interactions, forward-pass complexity is reduced from a potential $\mathcal{O}\left(T^2M^2\right)$ to $\mathcal{O}\left(T^2M\right)$.

Furthermore, this encoding eschews optimal transport subproblems, as well as train-time decoding due to (13). Decoding costs $\mathcal{O}(M)$ per measure and is isolated to the inference step.

4 The case of elliptical measures over $\mathbb{R}^d$

To extend the quantile construction to measures on $\mathbb{R}^d$, we use the Sliced Wasserstein metric (Rabin et al., 2011; Bonnotte, 2013) which, for elliptical measures, reduces the problem to a finite set of one-dimensional projections which reuse the machinery of Section 3. Although Brenier maps (Brenier, 1991) are another higher-dimensional analogue of quantiles, their empirical computation for $d > 1$ requires solving semi-discrete optimal transport problems. Sliced representations avoid this cost by retaining simple one-dimensional quantile computations along projected directions.

4.1 The sliced Wasserstein distance and elliptical measure identifiability

Consider the unit sphere $\mathbb{S}^{d-1}$ and a vector $\boldsymbol{s} \in \mathbb{S}^{d-1}$, and define the one-dimensional projection operator $P_{\boldsymbol{s}} : \mathbb{R}^d \to \mathbb{R}$ as $P_{\boldsymbol{s}}(x) \coloneqq \langle \boldsymbol{s}, x \rangle$. For any measure $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ we define the induced one-dimensional push-forward measure by $\mu^{\boldsymbol{s}} \coloneqq (P_{\boldsymbol{s}})_{\#}\mu \in \mathcal{P}_p(\mathbb{R})$.

Definition 4.1. *Let $p \geq 1$ and σ_{d-1} be the uniform probability measure on $\mathbb{S}^{d-1}$. For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the Sliced Wasserstein-p distance is* $\mathrm{SW}_p^p(\mu, \nu) \coloneqq \int_{\mathbb{S}^{d-1}} W_p^p(\mu^{\boldsymbol{s}}, \nu^{\boldsymbol{s}})\, d\sigma_{d-1}(\boldsymbol{s})$.

The quantity SW_2 is a metric on $\mathcal{P}_2(\mathbb{R}^d)$ Bonnotte (2013, Proposition 5.1.2) and topologically equivalent to W_2 (Bayraktar & Guo, 2021, Theorem 2.3 (i)). For the elliptical distribution class introduced next, we show that a finite-slice version of SW_2 retains these properties.

Definition 4.2. *A Borel probability measure μ on $\mathbb{R}^d$ is called an* elliptical distribution *if its characteristic function has the form $\varphi_\mu(\boldsymbol{t}) = \exp(i\langle \boldsymbol{m}, \boldsymbol{t} \rangle)\, \psi\big(\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}\big)$, for $\boldsymbol{t}, \boldsymbol{m} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} \in \mathbb{S}_+^d$, and a continuous function $\psi : [0, \infty) \to \mathbb{C}$.*

This class includes Gaussian, multivariate Student, Cauchy, stable, logistic, and multivariate uniform-on-ellipsoid distributions. The recent result of Fraiman et al. (2023) states that for the set $\mathcal{V}_d \subseteq \mathbb{S}^{d-1}$

$$\mathcal{V}_d \;\coloneqq\; \{\boldsymbol{e}_i : i \in [d]\} \cup \left\{ \frac{1}{\sqrt{2}}(\boldsymbol{e}_j + \boldsymbol{e}_k) : 1 \leq j < k \leq d \right\}, \tag{15}$$

with cardinality $L_d \coloneqq |\,\mathcal{V}_d\,| = \frac{d(d+1)}{2}$, the following finite Cramér–Wold-type result holds.

Proposition 4.1 (Restricted version of Theorem 1 of Fraiman et al. (2023))**.** *Let μ and ν be elliptical probability measures on $\mathbb{R}^d$. If $\mu^{\boldsymbol{s}} = \nu^{\boldsymbol{s}}$ for every $\boldsymbol{s} \in \mathcal{V}_d$, then $\mu = \nu$.*

In what follows, we consider the compact $\mathcal{K}_{\mathrm{ell}} \subseteq \big\{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mu \text{ is elliptical}\big\}$ metrized by W_2, on which it can be shown that $\mathrm{SW}_{2,V_d}^2(\mu, \nu) \;\coloneqq\; L_d^{-1} \sum_{\ell=1}^{L_d} W_2^2(\mu^{s_\ell}, \nu^{s_\ell})$ is also a metric. We defer the verification of Steps 1 and 2 from Section 2.1 to Appendix E.1 since the operators $\Phi_{\mathcal{V}_d}$, $\Pi_M^{L_d}$ and $C_M^{L_d}$ are component-wise applications of their counterparts in Section 3 and properties readily extend.

Empirical token encoding and decoding. For an empirical measure $\widehat{\mu}_M$ realized on atoms $\{x_i\}_{i=1}^M \subset \mathbb{R}^d$ and every $s_\ell \in V_d$, choose $\pi_{\ell,\uparrow} \in \mathfrak{S}_M$ such that $\langle s_\ell, x_{\pi_{\ell,\uparrow}(1)} \rangle \leq \cdots \leq \langle s_\ell, x_{\pi_{\ell,\uparrow}(M)} \rangle$, and write $y_{\ell,(i)} \;\coloneqq\; \langle s_\ell, x_{\pi_{\ell,\uparrow}(i)} \rangle$. Then, the empirical token amounts to

$$E_M(\widehat{\mu}_M) \;\coloneqq\; \left(C_M^{L_d} \circ \Phi_{V_d}\right)(\widehat{\mu}_M) = \frac{1}{\sqrt{M L_d}} \left[y_{1,(1)}, \ldots, y_{1,(M)}, \ldots, y_{L_d,(1)}, \ldots, y_{L_d,(M)}\right]^\top.$$

Since all requirements of Section 2.1 are satisfied, the consistency result in Proposition 2.1 holds.

Decoding from $\mathbb{R}^d$, unlike in one dimension, suffers from the absence of a canonical inverse for the sliced token. This is due to independent sorting across slices, which destroys particle correspondence. However, for elliptical measures, decoding remains computationally tractable via a different route.

By the stochastic representation formula of Cambanis et al. (1981), an elliptical random vector can be written as $\boldsymbol{X} = \boldsymbol{m} + R\boldsymbol{C}^{1/2}\boldsymbol{u}$, where $\boldsymbol{m}$ and $\boldsymbol{C} \succ \boldsymbol{0}$ are respectively the mean and covariance matrix, $\boldsymbol{u}$ is a random vector with distribution $\mathrm{Unif}(S^{d-1})$ and radius $R \sim \nu_R$ is independent of $\boldsymbol{u}$.

The assumption $\boldsymbol{C} \succ \mathbf{0}$ is a simplifying one, however, the subsequent results have a straightforward extension to rank-deficient cases. We impose that $\mathbb{E}[R^2] = d$ to remove scale ambiguity.

Then, by $\boldsymbol{u}$'s rotational invariance and for every $\boldsymbol{s} \in \mathbb{S}^{d-1}$, $\langle \boldsymbol{s},\ \boldsymbol{X} \rangle \overset{d}{=} \langle \boldsymbol{s},\ \boldsymbol{m} \rangle + \sqrt{\boldsymbol{s}^\top \boldsymbol{C} \boldsymbol{s}}\, RU_1$, where U_1 is the first coordinate of $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Therefore, all nondegenerate one-dimensional projections have the same standardized marginal and, denoting the quantile function of RU_1 as ϕ_{RU_1},

$$\Phi(\mu^{\boldsymbol{s}}) = \langle \boldsymbol{s},\ \boldsymbol{m} \rangle \mathbf{1} + \sqrt{\boldsymbol{s}^\top \boldsymbol{C} \boldsymbol{s}}\ \phi_{RU_1}. \tag{16}$$

This suggests a three-step reconstruction of the elliptical law: i) recover $\boldsymbol{m}$ and $\boldsymbol{C}$; ii) recover the common standardized marginal ϕ_{RU_1} approximately, and thus estimate ν_R; and iii) generate a point cloud using the stochastic representation formula above. We construct such a decoder D_M with full detail in Appendix F, and give the following stability result (proven in Appendix F.6).

Proposition 4.2 (Uniform decoder stability)**.** *Let $d \geq 2$, and let $\mathcal{K}_{\mathrm{ell}} \subset \mathcal{P}_2(\mathbb{R}^d)$ be a nonempty W_2-compact family of elliptical laws with positive-definite covariances. Let D_M be the decoder constructed in Appendix F, with parameters chosen as specified there. Then, D_M satisfies* (4).

Target function approximation and universality. The surrogate construction from Section 3 is directly applicable here over $L^{2^{L_d T}}$ for $\mathrm{dist}_{T,\mathcal{V}_d}(\boldsymbol{p}, \boldsymbol{q}) := \max_{t \in [T]} \| \boldsymbol{p}_t - \boldsymbol{q}_t \|_{L^{2^{L_d}}}$. We defer details to Appendix E.3 and proceed with the universality result, proven in Appendix E.5.

Theorem 4.1 (High-probability approximation universality in $\mathcal{P}(\mathbb{R}^d)$)**.** *Let $d \geq 2$, let $\mathcal{K}_{\mathrm{ell}} \subset \mathcal{P}_2(\mathbb{R}^d)$ be a nonempty W_2-compact family of elliptical laws with positive-definite covariances and consider $\varphi \in \mathcal{C}_{W_2}\left((\mathcal{K}_{\mathrm{ell}})^T, (\mathcal{K}_{\mathrm{ell}})^T\right)$. For every $\varepsilon > 0$ and every $\delta > 0$, there exists $M_0 \in \mathbb{N}$ such that, for every $M \geq M_0$, there exists θ and a transformer-based map* (5) *ensuring*

$$\sup_{\boldsymbol{\mu} \in (\mathcal{K}_{\mathrm{ell}})^T} \mathbb{P}_{\boldsymbol{\mu}} \left(\max_{s \in [T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) > \varepsilon \right) < \delta.$$

Computational complexity. The L_d projections and their sorted order statistics map point clouds to ML_d-dimensional Euclidean vectors at a one-time $\mathcal{O}\left(Md^2 \log M\right)$ cost per measure. Attention thus acts over T tokens, with forward-pass complexity $\mathcal{O}\left(T^2 M d^2\right)$. As before, training uses ordinary Euclidean losses and eschews optimal transport subproblems. Our proposed three-step decoder has a complexity of $\mathcal{O}\left(M^{7/4} + Md^2 + d^3\right)$ (see Appendix F) accounting for optimizing a simply-constrained strongly convex objective to prescribed accuracy, moment computations across slices, and spectral manipulation of the estimated covariance $\widehat{\boldsymbol{C}}$, respectively.

5 Synthetic next-measure prediction experiments

We consider next-measure prediction for one- and five-dimensional Gaussian and Student-t distributions. We describe the 1D Gaussian experiment here and defer the rest to Appendix H. Its measures are generated from a linear dynamical system: if the latent state has distribution $\mathcal{N}(\boldsymbol{m}_t, \boldsymbol{P}_t)$, then observations follow the law μ_t where

$$\mu_t = \mathcal{N}\left(\boldsymbol{c}^\top \boldsymbol{m}_t, \boldsymbol{c}^\top \boldsymbol{P}_t \boldsymbol{c} + \sigma_v^2\right), \qquad \boldsymbol{m}_{t+1} = \boldsymbol{A}\boldsymbol{m}_t, \qquad \boldsymbol{P}_{t+1} = \boldsymbol{A}\boldsymbol{P}_t\boldsymbol{A}^\top + \boldsymbol{Q}. \tag{17}$$

The system parameters are fixed, while $\boldsymbol{m}_t \in \mathbb{R}^p$ and $\boldsymbol{P}_t \in \mathbb{R}^{p \times p}$, with $p = 10$, are resampled independently per sequence. We use $T = 15$.

Baselines. For the Gaussian and Student-t experiments, we use three baselines, two of which are parameterless (not learned): *Persistence*, which predicts $\widehat{\mu}_{t+1,M} = \widehat{\mu}_{t,M}$; the *Population oracle*, which converts the true parameters into a deterministic midpoint grid approximation of the quantile function; and the causal *Moment Transformer*, which performs next-token prediction using each point cloud's sample mean and log standard deviation (or covariance matrix, for $d \geq 2$). The latter two baselines are privileged since they assume the correct parameters and parametric family, respectively.

Evaluation. For each $(M, \boldsymbol{A}, s)$ in the Gaussian and Student-t experiments, we evaluate on $N_{\mathrm{test}} = 2048$ empirical sequences and compute $\mathcal{L}^{\mathrm{test}}_{W_2,\mathrm{emp}} := \frac{1}{T N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \sum_{t=1}^{T} W_2\left(\widehat{\mu}^{i,\mathrm{pred}}_{t,M}, \widehat{\mu}^{i}_{t,M}\right)$. The plotted values average this error over three seeds. The metric includes target sampling variability

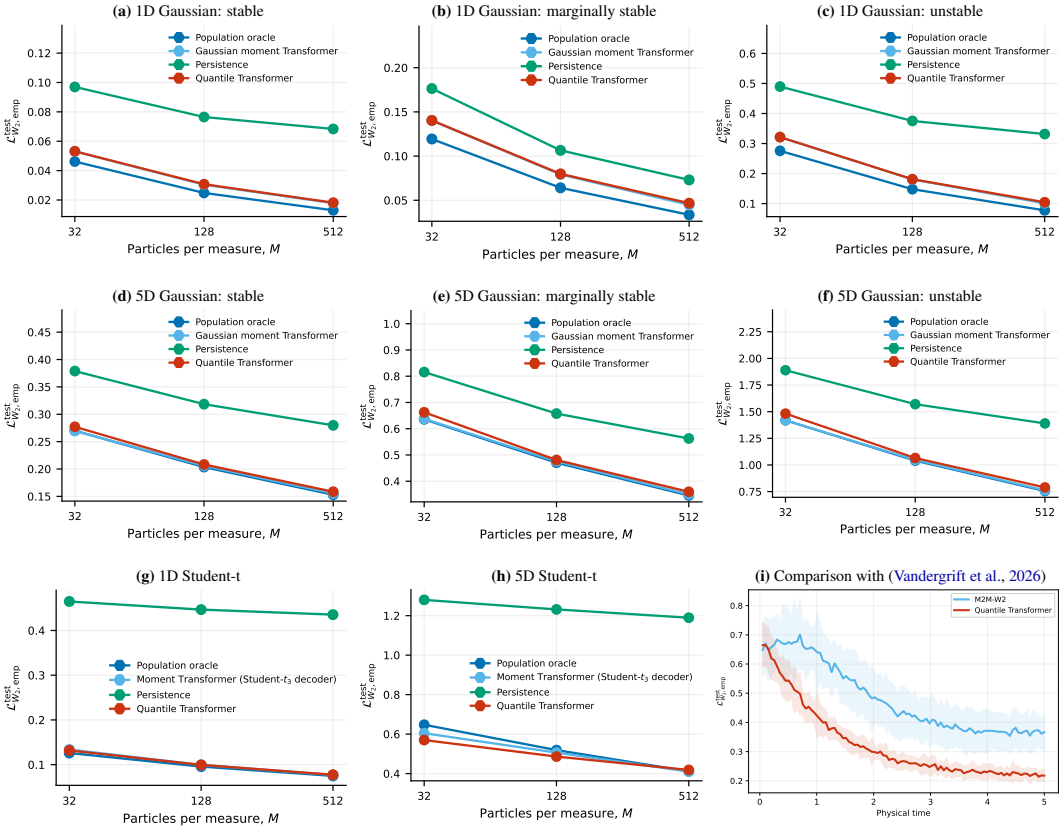

Figure 2: Panels (a)–(h) report three-seed mean empirical test W_2 error versus particles per measure $M \in \{32, 128, 512\}$. The 1D Gaussian panels (a)–(c) and 5D Gaussian panels (d)–(f) cover stable, marginally stable, and unstable regimes; panels (g) and (h) show the 1D and 5D Student-t cases. Panel (i) compares our method with M2M-W2 of Vandergrift et al. (2026) as the governing differential equation evolves in physical time.

via the population oracle. we use the same metric for comparing with M2M-W_2 of Vandergrift et al. (2026) for one training seed, with results averaged over $N_{\text{test}} = 10$ trajectories.

Results. In the 1D and 5D Gaussian panels of Figure 2, the Quantile Transformer remains close to the moment Transformer across particle counts and dynamical regimes. This observation transfers likewise to the 1D and 5D Student-t experiments. Both 5D experiments use the same decoder for the Quantile Transformer, which is agnostic of the observation family beyond ellipticity. Panel (i) presents the comparison with M2M-W2 (Vandergrift et al., 2026) for a gaussianity-preserving McKean–Vlasov system in dimension $d = 3$. Full experimental details, along with results on different metrics and new experimental scenarios, are given in Appendix H.

6 Conclusion, Limitations and Future Work

We introduced the *Quantile Transformer*, an architecture encoding empirical measures into W_2-compatible tokens in $\mathbb{R}^d$. We further proved its universal approximation ability relative to continuous sequence-to-sequence functions over arbitrary compacts in $\mathcal{P}_2(\mathbb{R})$ and compacts of elliptical measures in $\mathcal{P}_2(\mathbb{R}^d)$. Finally, we showedcased the architecture's competitive performance relative to measure class-aware baselines on synthetic experiments.

Together with its practical advantages, our construction has several limitations worth further investigation. First, measure-token parity is achieved at the cost of a linear-in-M token dimension. More compact, consistency-preserving representations ensuring decoder stability are thus desirable. Second, studying this framework for broader measure classes remains open. Finally, real-world experiments should be carried out to establish the Quantile Transformer's "in-the-wild" performance.

AI USE STATEMENT

GPT-5.5, GPT-5.6 Sol, and GPT-6 Astra were used across the development of this work to assist with polishing the writing, retrieving and discovering relevant literature, and brainstorming proof strategies and research directions for the theoretical results. The authors independently developed, completed, and verified all final proofs, critically evaluated all AI-assisted suggestions. The code was fully developed with Codex, which was relentlessly micromanaged by the authors (in this instance, contrary to the real world, micromanaging proved useful). We take full responsibility for the final content of this work, including all text, claims, proofs, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work is theoretical and aims to advance the understanding of machine learning. We identify no direct ethical or societal implications specific to the results, and any downstream impacts would depend on future applications.

REPRODUCIBILITY STATEMENT

All assumptions underlying the theoretical results are stated explicitly, with complete proofs provided in the appendices. Appendix B provides the synthetic data generation, model architecture, hyperparameter selection, training, evaluation, and computational details for the experiments. All code used to produce the experimental results are provided as supplementary material alongside the submission.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal approximation for efficient transformers. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, volume 221 of *Proceedings of Machine Learning Research*, pp. 72–86. PMLR, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Tal Amir and Nadav Dym. Fourier sliced-wasserstein embedding for multisets and measures. In *International Conference on Learning Representations*, volume 2025, pp. 24590–24629, 2025.

Miriam Ayer, H Daniel Brunk, George M Ewing, William T Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pp. 641–647, 1955.

Eviatar Bach, Ricardo Baptista, Edoardo Calvello, Bohan Chen, and Andrew Stuart. Learning enhanced ensemble filters. *Journal of Computational Physics*, pp. 114550, 2025.

Erhan Bayraktar and Gaoyue Guo. Strong equivalence between metrics of wasserstein type. 2021.

Oumaima Bencheikh and Benjamin Jourdain. Approximation rate in wasserstein distance of probability measures on the real line by deterministic empirical measures. *Journal of Approximation Theory*, 274:105684, 2022.

Boris Beranger, Huan Lin, and Scott Sisson. New models for symbolic data analysis: B. beranger et al. *Advances in Data Analysis and Classification*, 17(3):659–699, 2023.

Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, 1997. doi: 10.1007/978-1-4612-0653-8.

Lynne Billard and Edwin Diday. From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, 98(462):470–487, 2003.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, New York, 3 edition, 1995. ISBN 978-0-471-00710-4.

Patrick Billingsley. *Convergence of probability measures*, volume 2. wiley New York, 1999.

Shiba Biswal, Karthik Elamvazhuthi, and Rishi Sonthalia. Universal approximation of mean-field models via transformers. *arXiv preprint arXiv:2410.16295*, 2024.

Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie), 2013.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.

Gwendoline De Bie, Gabriel Peyré, and Marco Cuturi. Stochastic deep networks. In *International Conference on Machine Learning*, pp. 1556–1565. PMLR, 2019.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Ricardo Fraiman, Leonardo Moreno, and Thomas Ransford. A cramér–wold theorem for elliptical distributions. *Journal of Multivariate Analysis*, 196:105176, 2023.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers, 2024. URL https://arxiv.org/abs/2312.10794.

Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *Foundations of Computational Mathematics*, pp. 1–50, 2026.

Marc Hallin, Gilles Mordant, and Johan Segers. Multivariate goodness-of-fit tests based on wasserstein distance. 2021.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

Jerry Yao-Chieh Hu, Hude Liu, Hong-Yu Chen, Weimin Wu, and Han Liu. Universal approximation with softmax attention. *arXiv preprint arXiv:2504.15956*, 2025.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5(Jul):819–844, 2004.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *International Conference on Learning Representations*, volume 2024, pp. 35177–35205, 2024.

Rainer Kress. *Linear Integral Equations*. Springer New York, 3rd edition, 2014. doi: 10.1007/978-1-4614-9593-2.

Christoph H Lampert. Predicting the future behavior of a time-varying probability distribution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 942–950, 2015.

Martin Larsson, Jonghwa Park, and Johannes Wiesel. On concentration of the empirical measure for radial transport costs. *Stochastic Processes and their Applications*, 178:104466, 2024.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Nimrod Megiddo and Arie Tamir. Linear time algorithms for some separable quadratic programming problems. *Operations Research Letters*, 13(4):203–211, 1993.

Dimitri Meunier, Massimiliano Pontil, and Carlo Ciliberto. Distribution regression with sliced wasserstein kernels. In *International Conference on Machine Learning*, pp. 15501–15523. PMLR, 2022.

Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. *Advances in neural information processing systems*, 25, 2012.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1982. ISBN 9780471094425.

James R. Munkres. *Topology*. Prentice Hall, 2 edition, 2000. ISBN 978-0-13-181629-9.

Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-wasserstein embedding. *Advances in Neural Information Processing Systems*, 34:3389–3400, 2021.

J-I Nagata. *Modern general topology*, volume 33. Elsevier, 1985.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pp. 435–446. Springer, 2011.

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

Philippe Rigollet. The mean-field dynamics of transformers. In *International Congress of Mathematicians 2026*, pp. 389–404. SIAM, 2026.

Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, 3 edition, 1976.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 3 edition, 1987.

Walter Rudin. *Functional Analysis*. McGraw-Hill, New York, 2 edition, 1991.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pp. 13–31. Springer, 2007.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

Matthew Vandergrift, Martha White, Yury Polyanskiy, Philippe Rigollet, and Lazar Atanackovic. Measure-to-measure regression with transformers. *arXiv preprint arXiv:2605.28075*, 2026.

Veeravalli S Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Roman Vershynin. High-dimensional probability. *Cambridge Series in Statistical and Probabilistic Mathematics*, 47, 2019.

Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 0-8218-3312-X.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020.

Chuang Xu and Arno Berger. Best finite constrained approximations of one-dimensional probabilities. *Journal of approximation theory*, 244:1–36, 2019.

Tianlin Yang, Hailiang Du, and Louis Aslett. A distribution-to-distribution neural probabilistic forecasting framework for dynamical systems. *arXiv preprint arXiv:2603.25370*, 2026.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

A Additional related literature

Classical learning methods with distributional inputs. Early approaches largely adapt classical kernel methods to distribution-valued data by defining positive-definite kernels directly on probability measures. Examples include the Probability Product Kernel of Jebara et al. (2004) and kernels based on Kernel Mean Embeddings (KMEs) (Smola et al., 2007; Muandet et al., 2012), which map distributions into an RKHS and are injective for characteristic kernels. This enables standard discriminative methods to operate on distributional inputs — see Muandet et al. (2017) for a review.

Early next-distribution-prediction work. To our knowledge, Lampert (2015) provide the earliest treatment of next-distribution prediction. They embed time-varying distributions via kernel mean embeddings (KMEs) and learn their evolution through a regularized linear operator in the corresponding RKHS, with optional measure reconstruction through kernel herding. While their framework similarly relies on measure embeddings, the representation and restrictions in the sequence model differ substantially.

De Bie et al. (2019) introduce an early general-purpose network for learning nonlinear maps between probability measures. Their structural and functional unit is the *elementary block* implementing $\mathbf{x} \mapsto \mathbb{E}_{\mathbf{X}' \sim \mu}\big[f(\mathbf{x}, \mathbf{X}')\big]$, which for M-atom empirical measures, amounts to $\mathcal{O}\left(M^2\right)$ pairwise interactions. They show that three blocks with an intermediate stochastic excitation yield W_1 universality for continuous maps between compactly supported probability measures (note that our work does not require a compact support assumption). Note that the elementary block unit can be viewed as a learned, query-dependent analogue of the above KME aggregation.

Standard transformer expressivity. Transformer universality has been established under increasingly restrictive architectures. Yun et al. (2019) cover continuous permutation-equivariant maps, and general sequence maps with positional encodings. Alberti et al. (2023) obtain uniform approximation on compact product domains with a single attention layer. Kajitsuka & Sato (2024) further restrict to single-head, low-rank self-attention. Finally, Hu et al. (2025) show universality with softmax attention alone. Our analysis builds on a direct extension of the uniform result of Alberti et al. (2023).

Other related works. Bach et al. (2025) propose *measure neural mappings* (MNMs), an operator-learning formulation combining measure-valued cross-attention and pushforward operations, with the former reducing to standard particle attention for empirical measures.

Among sliced-Wasserstein methods, Amir & Dym (2025) construct injective finite-dimensional encodings for measures with a bounded number of atoms. Naderializadeh et al. (2021) learn fixed-size pooling representations using reference measures and interpolation. Meunier et al. (2022) construct universal sliced-Wasserstein kernels on probability measures and derive learning bounds for scalar-valued regression when the measures are supported on a common compact domain.

B AUXILIARY RESULTS USED THROUGHOUT THE PAPER

B.1 SUPPORTING RESULTS: MATHEMATICAL ANALYSIS

Proposition B.1 (Uniform continuity on compact domains)**.** *Let X and Y be metric spaces, with X compact. Every continuous map $f : X \to Y$ is uniformly continuous.*

Proof. See Rudin (1976, Thm. 4.19). □

Proposition B.2 (Continuous images of compact sets)**.** *Let X be compact, let Y be a topological space, and let $f : X \to Y$ be continuous. Then $f(X)$ is compact.*

Proof. See Munkres (2000, Thm. 26.5). □

Proposition B.3 (Compact embeddings are homeomorphisms onto their images)**.** *Let X be compact and let Y be Hausdorff. If $f : X \to Y$ is continuous and injective, then*

$$f : X \to f(X) \tag{18}$$

is a homeomorphism. In particular, $f^{-1} : f(X) \to X$ is continuous.

Proof. See Munkres (2000, Thm. 26.6). In particular for what we need, $\mathbb{R}^M$ is Hausdorff. □

Proposition B.4 (Finite products of compact spaces)**.** *Let $X_1, \ldots, X_T$ be compact topological spaces. Then*

$$X_1 \times \cdots \times X_T \tag{19}$$

is compact in the product topology. In particular, if X is compact, then X^T is compact.

Proof. See Munkres (2000, Thm. 26.7). □

Proposition B.5 (Compact sets admit α-nets)**.** *A compact subset $\mathcal{Q}$ of a metric space $(\mathcal{X}, d)$ admits an α-net for every $\alpha > 0$.*

Proof. We start from Definition 2.32 of compactness from Rudin (1976), stating that a set $\mathcal{Q}$ is compact if every open cover admits a finite subcover.

Consider the family of open covers of $\mathcal{Q}$ indexed by $\alpha > 0$, $\mathcal{C}_\alpha = \{B(x, \alpha) \mid x \in \mathcal{Q}\}$. It is straightforward to see that for any $\alpha > 0$, $\mathcal{Q} \subseteq \bigcup_{Y \in \mathcal{C}_\alpha} Y$. By compactness, there exists for every $\alpha > 0$ an integer $J_\alpha \in \mathbb{N}_+$ and a set of points $\{x_1, x_2, \ldots x_{J_\alpha}\}$ such that $\mathcal{Q} \in \bigcup_{i=1}^{J_\alpha} B(x_i, \alpha)$.

From here, and observing that the finite cover means that every $x \in \mathcal{Q}$ belongs to at least one ball $B(x_j, \alpha)$, the result immediately follows. □

B.2 Supporting results: functional analysis

Proposition B.1. *Let $\mathcal{J}$ be a finite partition of $(0,1]$ into positive length intervals, and let $\mathcal{S}(\mathcal{J})$ be the space of functions constant on each interval. Define the operator $\Pi_{\mathcal{J}} : L^2 \to \mathcal{S}(\mathcal{J})$ by*

$$\Pi_{\mathcal{J}}(g) \;:=\; \sum_{J\in\mathcal{J}} \left(\frac{1}{|\,J\,|} \int_J g(u)\,du \right) \mathbf{1}_J, \tag{20}$$

where $|\,J\,|$ denotes interval length. Then,

(a) $\Pi_{\mathcal{J}}$ is the orthogonal projection onto $\mathcal{S}(\mathcal{J})$;

(b) for a sequence of partitions $(\mathcal{J}_M)_{M\geq 1}$ satisfying $\lim_{M\to\infty} \max_{J\in\mathcal{J}_M} |\,J\,| = 0$, it holds that $\lim_{M\to\infty} \| \Pi_{\mathcal{J}_M} g - g \|_{L^2} = 0$ for every $g \in L^2$;

(c) for the same sequence of partitions $(\mathcal{J}_M)_{M\geq 1}$ as above and any compact set $\mathcal{Q} \in L^2$, it holds that $\lim_{M\to\infty} \sup_{g\in\mathcal{Q}} \| \Pi_{\mathcal{J}_M} g - g \|_{L^2} = 0$.

Proof. We prove point (a). For every $h \in \mathcal{S}(\mathcal{J}_M)$, we have

$$\langle g - \Pi_{\mathcal{J}_M} g,\; h\rangle_{L^2} \;=\; \sum_{J\in\mathcal{J}} h_J \int_J \big(g - \Pi_{\mathcal{J}_M} g\big) = 0,$$

where h_J is the value of h on J and the last equality simply comes from the definition of $\Pi_{\mathcal{J}}$. We thus have orthogonality and the ensuing non-expansiveness.

To prove point (b), define $\rho_M := \max_{J\in\mathcal{J}_M} |\,J\,|$ and let $g \in C([0,1])$. Pick any interval $J \in \mathcal{J}_M$ and a point $u \in J$. By definition, we have

$$\begin{aligned} |\,\Pi_{\mathcal{J}_M} g(u) - g(u)\,| &= \left| \frac{1}{|\,J\,|} \int_J \big(g(z) - g(u)\big)\,dz \right| \\ &\leq \frac{1}{|\,J\,|} \int_J \big|\big(g(z) - g(u)\big)\big|\,dz \\ &\leq \sup_{|\,z-u\,|\leq \delta_M} |\,g(z) - g(u)\,|\,. \end{aligned}$$

Since g is continuous on the compact $[0,1]$ it is uniformly continuous which means $\lim_{M\to\infty} \sup_{|\,z-u\,|\leq\delta_M} |\,g(z) - g(u)\,| = 0$. We now bring back this property into L^2 by the density of $C([0,1])$ in L^2 (Rudin, 1987, Theorem 3.14).

Pick an arbitrary $h \in L^2$. Densness of $C([0,1])$ in L^2, means that for every $\varepsilon > 0$ there exists $g \in C([0,1])$ such that

$$\| h - g \|_{L^2} < \varepsilon.$$

Using that $\Pi_{\mathcal{J}_M}$ is a contraction, adding and subtracting $\Pi_{\mathcal{J}_M} g$ and g, and using the triangle inequality

$$\begin{aligned} \| \Pi_{\mathcal{J}_M} h - h \|_{L^2} &\leq \| \Pi_{\mathcal{J}_M} (h - g) \|_{L^2} + \| \Pi_{\mathcal{J}_M} g - g \|_{L^2} + \| g - h \|_{L^2} \\ &\leq 2\varepsilon + \| \Pi_{\mathcal{J}_M} g - g \|_{L^2}\,. \end{aligned}$$

The last term tends to zero by the first part of the argument, and since $\varepsilon > 0$ is arbitrary, $\lim_{M\to\infty} \| \Pi_{\mathcal{J}_M} h - h \|_{L^2} = 0$ for every $h \in L^2$.

Since $\lim_{M\to\infty} \Pi_{\mathcal{J}_M} = I$ pointwise on L^2 from before, where I is the identity map, and L^2 is a Banach space, the convergence is uniform on compact subsets of L^2 by Kress (2014, Corollary 10.7). Hence, for every compact $\mathcal{K} \subset L^2$, $\lim_{M\to\infty} \sup_{h\in\mathcal{K}} \| \Pi_{\mathcal{J}_M} h - h \|_{L^2} = 0$. □

Proposition 3.1 (Properties of Π_M)**.** *The operator Π_M defined in (10) is an orthogonal projection onto $\mathcal{S}_M$ and $\lim_{M\to\infty} \| \Pi_M g - g \|_{L^2} = 0$, for every $g \in L^2$.*

Proof. Orthogonality follows directly from point (a) of Proposition B.1. Convergence of projections to the identity map follows by noticing that partition (9) has equal interval length of $1/M$ which goes to zero as $M \to \infty$, and point (b) of Proposition B.1 applies. □

B.3 Supporting results: measure theory, probability theory and optimal transport

Proposition B.6 (One-dimensional quantile formula)**.** *Let $p \geq 1$ and let $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$. Then*

$$W_p^p(\mu, \nu) = \int_0^1 \left| F_\mu^{-1}(u) - F_\nu^{-1}(u) \right|^p \, du. \tag{21}$$

In particular, for $p = 2$,

$$W_2(\mu, \nu) = \left\| F_\mu^{-1} - F_\nu^{-1} \right\|_{L^2(0,1)}. \tag{22}$$

Proof. See Villani (2003)[Thm. 2.18, Remarks 2.19]. □

Theorem B.1 (Empirical convergence in Wasserstein distance)**.** *Let $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and let $X_1, X_2, \ldots \overset{i.i.d.}{\sim} \mu$. Define the empirical measure*

$$\widehat{\mu}_M = \frac{1}{M} \sum_{i=1}^{M} \delta_{X_i}. \tag{23}$$

Then $\lim_{M \to \infty} W_p(\widehat{\mu}_M, \mu) = 0$ almost surely.

Proof. We first remark the almost sure weak convergence result for probability measures given by Varadarajan (1958).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathcal{Z}, \mathrm{d}_{\mathcal{Z}})$ be a metric space and $\mathcal{B}(\mathcal{Z})$ the class of Borel sets of $\mathcal{Z}$. Let $\implies$ denote the weak convergence of probability measures, meaning that for a measure μ and a sequence $\{\mu_n\}_{n \geq 1}$ defined on $\mathcal{Z}$,

$$\mu_n \implies \mu \qquad \Longleftrightarrow \qquad \int_{\mathcal{Z}} g \mathrm{d}\mu_n \overset{n \to \infty}{\longrightarrow} \int_{\mathcal{Z}} g \mathrm{d}\mu,$$

for each bounded continuous function g on $\mathcal{Z}$. The corollary to Theorem 3 of Varadarajan (1958) states

Theorem B.2. *Consider a sequence of* i.i.d. *random variables $\{X_i\}_{i \geq 1}$ with $X_i : \Omega \to \mathcal{Z}$, and let $(\mathcal{Z}, \mathrm{d}_{\mathcal{Z}})$ be a separable metric space. For each $w \in \Omega$ we denote μ_n^w as the measure with masses $\frac{1}{n}$ at each of the n points $X_1(w), X_2(w), \ldots X_n(w)$. Then $\mathbb{P}(\{w \in \Omega \mid \mu_n^w \implies \mu\}) = 1$.*

Since for us $\mathcal{Z} = \mathbb{R}^d$ and $\mathrm{d}_{\mathcal{Z}} \equiv \|\cdot\|$, the corrolary applies and we have that $\widehat{\mu}_M \implies \mu$ almost surely w.r.t. $\mathbb{P}$.

Next, since for us $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, by the strong law of large numbers we have,

$$\int_{\mathbb{R}^d} \|x\|^p \, d\widehat{\mu}_M(x) = \frac{1}{M} \sum_{i=1}^{M} \|X_i\|^p \longrightarrow \mathbb{E}[\|X_1\|^p] = \int_{\mathbb{R}^d} \|x\|^p \, d\mu(x) \qquad \text{almost surely.} \tag{24}$$

Finally, we refer to Thm 7.12 of Villani (2003). Specifically, we have just shown that condition iii) in the aforementioned theorem holds almost surely for our setting, hence condition i) is implied, which is this theorem's claim.

□

C Omitted results in Section 2

C.1 Finite resolution empirical consistency result in Section 2.1

Proposition 2.1. *Let* $\lim_{M\to\infty} r_M = \infty$ *and fix an arbitrary* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

(a) If $\mathcal{K} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ *is compact, then* $\lim_{M\to\infty} \sup_{\mu\in\mathcal{K}} \left\| C_{r_M}^{-1} \overline{E}_{r_M}(\mu) - \Phi(\mu) \right\|_{\mathcal{Z}} = 0.$

(b) Sample $\widehat{\mu}_M \sim \mu$ *as in (2). Then, almost surely,* $\lim_{M\to\infty} \left\| C_{r_M}^{-1} E_{r_M}(\widehat{\mu}_M) - \Phi(\mu) \right\|_{\mathcal{Z}} = 0.$

Proof. For point (a), since $\mathcal{K}$ is compact and Φ is continuous, $\Phi(\mathcal{K})$ is compact in $\mathcal{Z}$. Moreover, the orthogonal projections Π_r are contractions and converge strongly to the identity by (3). Strong convergence of a uniformly bounded family of linear operators is uniform on compact subsets. Therefore,

$$\lim_{M\to\infty} \sup_{\mu\in\mathcal{K}} \|\Pi_{r_M}\Phi(\mu) - \Phi(\mu)\|_{\mathcal{Z}} = 0.$$

For point (b), by the definition of E_M, $C_{r_M}^{-1} E_M(\widehat{\mu}_M) = \Pi_{r_M}\Phi(\widehat{\mu}_M)$. Hence, using the triangle inequality relative to $\Pi_{r_M}\Phi(\mu)$ and the contractivity of orthogonal projections, we have

$$\| \Pi_{r_M}\Phi(\widehat{\mu}_M) - \Phi(\mu) \|_{\mathcal{Z}} \le \| \Phi(\widehat{\mu}_M) - \Phi(\mu) \|_{\mathcal{Z}} + \| \Pi_{r_M}\Phi(\mu) - \Phi(\mu) \|_{\mathcal{Z}} .$$

The second term has limit zero by (3). For the first term, we know that $\lim_{M\to\infty} W_2(\widehat{\mu}_M, \mu) = 0$ almost surely by Theorem B.1. Since Φ is continuous, it preserves almost sure convergence and thus the first term goes to zero almost surely. □

C.2 Uniform approximation of ordered sequence maps (Corollary 2.1)

Corollary 2.1. *Fix* $T, q \in \mathbb{N}_+$, *let* $\mathcal{U} \subseteq \mathbb{R}^q$ *be a nonempty compact set, and let* $G : \mathcal{U}^T \longrightarrow (\mathbb{R}^q)^T$ *be continuous. Then, for every* $\eta > 0$, *there exists a position-augmented unmasked softmax transformer* $\mathcal{T}_\theta : \mathcal{U}^T \longrightarrow (\mathbb{R}^q)^T$ *such that* $\sup_{\boldsymbol{z}\in\mathcal{U}^T} \max_{s\in[T]} \| \mathcal{T}_\theta(\boldsymbol{z})_s - G(\boldsymbol{z})_s \|_2 < \eta$.

Proof. We reduce the ordered map G to a continuous permutation-equivariant map on an augmented compact product domain and then apply Theorem 4.3 of Alberti et al. (2023). Choose pairwise distinct positional labels $\tau_1, \dots, \tau_T \in (0,1)$ and set $\mathcal{L} := \{\tau_1, \dots, \tau_T\}$. Define the linear embedding $\iota : \mathbb{R}^q \to \mathbb{R}^{q+1}$ by $\iota(z) := (z, 0)$ and the fixed positional vectors $\mathbf{p}_s := (\mathbf{0}_q, \tau_s)$. The corresponding positional lift is

$$P_T : \mathcal{U}^T \longrightarrow (\mathbb{R}^{q+1})^T, \qquad P_T(\boldsymbol{z})_s := \iota(z_s) + \mathbf{p}_s = (z_s, \tau_s).$$

Thus, P_T consists of a shared tokenwise linear input map followed by fixed absolute positional encodings. The augmented token space $\mathcal{X} := \mathcal{U} \times \mathcal{L} \subseteq \mathbb{R}^{q+1}$ is compact, and $P_T(\mathcal{U}^T) \subseteq \mathcal{X}^T$. Let $\mathfrak{S}_T$ act on token sequences by

$$(\pi \cdot \mathbf{x})_s := x_{\pi^{-1}(s)}, \qquad \pi \in \mathfrak{S}_T, \quad s \in [T].$$

For each $\pi \in \mathfrak{S}_T$, define $\mathcal{A}_\pi := \pi \cdot P_T(\mathcal{U}^T) = \prod_{s=1}^T \left(\mathcal{U} \times \{\tau_{\pi^{-1}(s)}\}\right)$. The displayed product identity is where the assumption that the original domain is the full product $\mathcal{U}^T$ is used. Because $\mathcal{L}$ is finite, every singleton $\{\tau_j\}$ is both open and closed in $\mathcal{L}$. Hence each $\mathcal{A}_\pi$ is both open and closed in the relative topology of $\mathcal{X}^T$. Moreover, the sets $\{\mathcal{A}_\pi : \pi \in \mathfrak{S}_T\}$ are pairwise disjoint because their ordered positional-label patterns are distinct. Set $\mathcal{A} := \bigcup_{\pi\in\mathfrak{S}_T} \mathcal{A}_\pi$. Since the union is finite, both $\mathcal{A}$ and $\mathcal{X}^T \setminus \mathcal{A}$ are open and closed in $\mathcal{X}^T$. Every $\mathbf{x} \in \mathcal{A}$ has a unique representation

$$\mathbf{x} = \pi \cdot P_T(\boldsymbol{z}), \qquad \pi \in \mathfrak{S}_T, \quad \boldsymbol{z} \in \mathcal{U}^T.$$

Indeed, in this representation, $x_s = \left(z_{\pi^{-1}(s)}, \tau_{\pi^{-1}(s)}\right)$. Since the positional labels are distinct, their order uniquely determines π. Once π is known, z_r is recovered from the first q coordinates of $x_{\pi(r)}$. Let $\iota^T : (\mathbb{R}^q)^T \to (\mathbb{R}^{q+1})^T$ denote the componentwise application of ι. The extra zero coordinate introduced by ι^T only matches the augmented token dimension and will be removed at the end of the proof. Define $F : \mathcal{X}^T \to (\mathbb{R}^{q+1})^T$ by

$$F(\mathbf{x}) := \begin{cases} \pi \cdot \iota^T\big(G(\boldsymbol{z})\big), & \mathbf{x} = \pi \cdot P_T(\boldsymbol{z}) \in \mathcal{A}, \\ \mathbf{0}, & \mathbf{x} \notin \mathcal{A}. \end{cases}$$

The uniqueness above ensures that F is well defined. For each $\pi \in \mathfrak{S}_T$ and $\mathbf{x} \in \mathcal{A}_\pi$, we have

$$F(\mathbf{x}) = \pi \cdot \iota^T \left(G \left(P_T^{-1}(\pi^{-1} \cdot \mathbf{x}) \right) \right).$$

Every map in this expression is continuous. In particular, the inverse of P_T on its image simply discards the positional coordinates and is continuous. Therefore, F is continuous on each $\mathcal{A}_\pi$. It is also continuous on $\mathcal{X}^T \setminus \mathcal{A}$, where it is identically zero. The sets $\mathcal{A}_\pi$, together with $\mathcal{X}^T \setminus \mathcal{A}$, are closed, pairwise disjoint, and form a finite cover of $\mathcal{X}^T$. Thus, repeated application of the pasting lemma (Munkres, 2000, Theorem 18.3) shows that F is continuous on $\mathcal{X}^T$. The map F is also permutation equivariant. First, for every $\sigma, \pi \in \mathfrak{S}_T$, we have $\sigma \cdot \mathcal{A}_\pi = \mathcal{A}_{\sigma\pi}$. Thus, $\mathcal{A}$ and its complement are invariant under permutations. If $\mathbf{x} = \pi \cdot P_T(\boldsymbol{z}) \in \mathcal{A}$, then

$$F(\sigma \cdot \mathbf{x}) = F\big((\sigma\pi) \cdot P_T(\boldsymbol{z})\big) = (\sigma\pi) \cdot \iota^T(G(\boldsymbol{z})) = \sigma \cdot F(\mathbf{x}).$$

If $\mathbf{x} \notin \mathcal{A}$, then also $\sigma \cdot \mathbf{x} \notin \mathcal{A}$, and both sides of the same identity are zero. Therefore, F is a continuous permutation-equivariant map on the compact product domain $\mathcal{X}^T$. For $\mathbf{v} = (v_1, \ldots, v_T) \in (\mathbb{R}^{q+1})^T$, let

$$\| \mathbf{v} \|_\infty \coloneqq \max_{s \in [T]} \max_{j \in [q+1]} |(v_s)_j|.$$

Applying the continuous case of Alberti et al. (2023, Theorem 4.3) to F, with approximation accuracy $\eta/\sqrt{q}$, yields an unmasked softmax transformer $\widetilde{\mathcal{T}}_\theta$ such that

$$\sup_{\mathbf{x} \in \mathcal{X}^T} \left\| \widetilde{\mathcal{T}}_\theta(\mathbf{x}) - F(\mathbf{x}) \right\|_\infty < \frac{\eta}{\sqrt{q}}.$$

Here $\widetilde{\mathcal{T}}_\theta$ belongs to the residual softmax-transformer class of Alberti et al. (2023, Definition 2.4). Let $R : \mathbb{R}^{q+1} \to \mathbb{R}^q$ be the linear projection $R(y, t) \coloneqq y$, and let R^T denote its componentwise application. Define $\mathcal{T}_\theta \coloneqq R^T \circ \widetilde{\mathcal{T}}_\theta \circ P_T$. This is a position-augmented unmasked softmax transformer with a tokenwise linear output projection. Moreover, $F(P_T(\boldsymbol{z})) = \iota^T(G(\boldsymbol{z}))$ and $R^T F(P_T(\boldsymbol{z})) = G(\boldsymbol{z})$ for every $\boldsymbol{z} \in \mathcal{U}^T$. Therefore,

$$\begin{aligned} \sup_{\boldsymbol{z} \in \mathcal{U}^T} \max_{s \in [T]} \| \mathcal{T}_\theta(\boldsymbol{z})_s - G(\boldsymbol{z})_s \|_2 &\le \sqrt{q} \sup_{\boldsymbol{z} \in \mathcal{U}^T} \left\| \widetilde{\mathcal{T}}_\theta(P_T(\boldsymbol{z})) - F(P_T(\boldsymbol{z})) \right\|_\infty \\ &\le \sqrt{q} \sup_{\mathbf{x} \in \mathcal{X}^T} \left\| \widetilde{\mathcal{T}}_\theta(\mathbf{x}) - F(\mathbf{x}) \right\|_\infty < \eta, \end{aligned}$$

where we used $\| u \|_2 \le \sqrt{q} \, \| u \|_\infty$ for $u \in \mathbb{R}^q$. This proves the result. □

D Omitted results in Section 3

D.1 Normalized coordinate map isometry

Proposition D.1 (The normalized coordinate map is an isometry)**.** *Let $\mathcal{S}_M \subseteq L^2$ be the step-function space associated with the partition $\{I_{i,M}\}_{i=1}^M$ defined in (9), and let $C_M : \mathcal{S}_M \to \mathbb{R}^M$ be defined by*

$$C_M\left(\sum_{i=1}^M a_i \mathbf{1}_{I_{i,M}}\right) \coloneqq \frac{1}{\sqrt{M}} [a_1, \dots, a_M]^\top . \tag{25}$$

Then C_M is a linear isometric isomorphism from $(\mathcal{S}_M, \|\cdot\|_{L^2})$ onto $(\mathbb{R}^M, \|\cdot\|_2)$. In particular, for every $f, g \in \mathcal{S}_M$,

$$\| C_M(f) - C_M(g) \|_2 = \| f - g \|_{L^2} . \tag{26}$$

Consequently, C_M is bijective and its inverse $C_M^{-1} : \mathbb{R}^M \to \mathcal{S}_M$ is also an isometry.

Proof. Let

$$f = \sum_{i=1}^M a_i \mathbf{1}_{I_{i,M}}, \qquad g = \sum_{i=1}^M b_i \mathbf{1}_{I_{i,M}}$$

be arbitrary elements of $\mathcal{S}_M$. By definition of C_M,

$$C_M(f) - C_M(g) = \frac{1}{\sqrt{M}} [a_1 - b_1, \dots, a_M - b_M]^\top .$$

Hence

$$\| C_M(f) - C_M(g) \|_2^2 = \frac{1}{M} \sum_{i=1}^M |a_i - b_i|^2. \tag{27}$$

On the other hand, since the intervals $I_{1,M}, \dots, I_{M,M}$ are pairwise disjoint and each has Lebesgue measure $1/M$,

$$\begin{aligned} \| f - g \|_{L^2}^2 &= \int_0^1 \left| \sum_{i=1}^M (a_i - b_i) \mathbf{1}_{I_{i,M}}(u) \right|^2 du \\ &= \sum_{i=1}^M |a_i - b_i|^2 \int_{I_{i,M}} 1 \, du \\ &= \frac{1}{M} \sum_{i=1}^M |a_i - b_i|^2. \end{aligned} \tag{28}$$

Comparing (27) and (28) gives

$$\| C_M(f) - C_M(g) \|_2^2 = \| f - g \|_{L^2}^2 .$$

Taking square roots proves (26).

Linearity follows directly from the definition. Moreover, for every $z = (z_1, \dots, z_M)^\top \in \mathbb{R}^M$,

$$C_M^{-1}(z) = \sqrt{M} \sum_{i=1}^M z_i \mathbf{1}_{I_{i,M}},$$

so C_M is bijective. Since C_M preserves distances, its inverse preserves distances as well. Thus C_M is a linear isometric isomorphism. □

D.2 Quantile identity after cell averaging

Proposition D.2 (Quantile identity after cell averaging)**.** *For every $\nu \in \mathcal{K}$, it holds that*

$$(\Phi \circ \Delta_M \circ \Pi_M \circ \Phi)(\nu) = (\Pi_M \circ \Phi)(\nu). \tag{29}$$

Proof. We recall first the definitions of all the operators.

$$\Phi : \mathcal{P}_2(\mathbb{R}) \to L^2 \text{ with } \Phi(\mu)(z) \;:=\; \inf\{x \in \mathbb{R} \mid F_\mu(x) \geq z\}, \tag{30}$$

$$\Delta_M : \mathcal{S}_M \to \mathcal{E}_M(\mathbb{R}) \text{ with } \Delta_M(f) \;:=\; \frac{1}{M}\sum_{i=1}^{M} \delta_{a_i}, \tag{31}$$

$$\Pi_M : L^2 \to \mathcal{S}_M \text{ with } \Pi_M(g) \;:=\; \sum_{i=1}^{M} \left(M \int_{I_{i,M}} g(u)\, du \right) \mathbf{1}_{I_{i,M}}, \tag{32}$$

where $I_{i,M}$ are the partition intervals defined in (9).

Now, notice that map Φ is non-decreasing, meaning that

$$z \leq u \implies \Phi(\mu)(z) \leq \Phi(\mu)(u).$$

This in turn implies that $(\Pi_M \circ \Phi)(\mu)$ belongs to the monotone cone as $\mathcal{S}_M^{\uparrow} := \{f \in \mathcal{S}_M : a_1 \leq a_2 \leq \cdots \leq a_M\}$ of the finite-dimensional step-function space $\mathcal{S}_M$. Specifically, for every $i, j \in [M]$ such that $i < j$, every $u \in I_{i,M}$, and every $v \in I_{j,M}$, we have $u \leq v$. Define the coefficients

$$a_{i,M}(\mu) \;:=\; M \int_{I_{i,M}} \Phi(\mu)(u)\, du, \qquad i \in [M]. \tag{33}$$

Then

$$(\Pi_M \circ \Phi)(\mu) = \sum_{i=1}^{M} a_{i,M}(\mu) \mathbf{1}_{I_{i,M}}. \tag{34}$$

Since each interval $I_{i,M}$ has Lebesgue measure $1/M$, for every $i < j$ we have

$$\begin{aligned} a_{j,M}(\mu) - a_{i,M}(\mu) &= M \int_{I_{j,M}} \Phi(\mu)(v)\, dv - M \int_{I_{i,M}} \Phi(\mu)(u)\, du \\ &= M^2 \int_{I_{i,M}} \int_{I_{j,M}} (\Phi(\mu)(v) - \Phi(\mu)(u))\, dv\, du \\ &\geq 0, \end{aligned} \tag{35}$$

where the second line comes from $1 = M \int_{I_{k,M}} du$ for all $k \in [M]$. Therefore, $a_{1,M}(\mu) \leq a_{2,M}(\mu) \leq \cdots \leq a_{M,M}(\mu)$, and $(\Pi_M \circ \Phi)(\mu) \in \mathcal{Z}_M$.

In parallel, note that $(\Phi \circ \Delta_M)|_{\mathcal{Z}_M} = \mathrm{Id}$. Specifically, take an arbitrary $f = \sum_{i=1}^{M} a_i \mathbf{1}_{I_{i,M}} \in \mathcal{S}_M^{\uparrow}$, so that $a_1 \leq a_2 \leq \cdots \leq a_M$. Let $\Phi^{-1}(\Delta_M(f))$ denote the cumulative distribution function of $\Delta_M(f)$. Then, for every $x \in \mathbb{R}$,

$$\Phi^{-1}(\Delta_M(f))(x) = \frac{1}{M}\sum_{k=1}^{M} \mathbf{1}_{\{a_k \leq x\}}. \tag{36}$$

We now show that the quantile function of $\Delta_M(f)$ is precisely f. Fix $i \in [M]$ and let $u \in I_{i,M} = \left(\frac{i-1}{M}, \frac{i}{M}\right]$. Since $a_1, \ldots, a_i \leq a_i$, it follows from (36) that

$$\Phi^{-1}(\Delta_M(f))(a_i) \geq \frac{i}{M} \geq u.$$

Therefore, by the definition of the generalized inverse,

$$\Phi(\Delta_M(f))(u) = \inf\left\{x \in \mathbb{R} : \Phi^{-1}(\Delta_M(f))(x) \geq u\right\} \leq a_i. \tag{37}$$

Conversely, let $x < a_i$. Since $a_i \le a_{i+1} \le \cdots \le a_M$, none of the indices $k \ge i$ satisfies $a_k \le x$. Hence, at most $i-1$ atoms of $\Delta_M(f)$ lie in $(-\infty, x]$, and consequently $\Phi^{-1}(\Delta_M(f))(x) \le \frac{i-1}{M} < u$. Thus no $x < a_i$ belongs to the set $\{x \in \mathbb{R} : F_{\Delta_M(f)}(x) \ge u\}$, which implies that

$$\Phi(\Delta_M(f))(u) \ge a_i. \tag{38}$$

Combining (37) and (38), we obtain, for every $i \in [M]$ and every $u \in I_{i,M}$,

$$\Phi(\Delta_M(f))(u) = a_i,$$

which is exactly $f \in \mathcal{S}_M^{\uparrow}$. The conclusion immediately follows. □

D.3 Construction of the surrogate in (Section 3)

To build the surrogate, we need to introduce the notion of α-net, and a basic helper result.

Definition D.1 (α-net)**.** *Let $(\mathcal{X}, d)$ be a metric space, let $\mathcal{A} \subseteq \mathcal{X}$, and fix $\alpha > 0$. A finite set $\mathcal{N}_\alpha := \{x^1, \dots, x^{J_\alpha}\} \subseteq \mathcal{A}$ is called an α-net of $\mathcal{A}$ if*

$$\sup_{x \in \mathcal{A}} \min_{j \in [J_\alpha]} d(x, x^j) \le \alpha. \tag{39}$$

Equivalently, the closed balls of radius α centered at the points of $\mathcal{N}_\alpha$ cover $\mathcal{A}$.

It is a known result that compact sets in a given metric space admit finite α-nets (see Proposition B.5). We construct a neighborhood enlargement of $\mathcal{Q}$, whose size is quantified α associated with an admissible α-net over $\mathcal{Q}$, and a "close-enough" continuous modification of $\widetilde{\varphi}$ over this neighborhood.

Define the α-net $\mathcal{N}_\alpha$ with cardinality J_α over $\mathcal{Q}^T$ with respect to dist_T, for some $\alpha > 0$. Further, let the α-sized neighborhood enlargement of $\mathcal{Q}^T$ with respect to this same metric be

$$\mathcal{Q}_\alpha := \left\{ \boldsymbol{h} \in L^2 \;\middle|\; \inf_{\boldsymbol{q} \in \mathcal{Q}} \mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}) \le \alpha \right\}. \tag{40}$$

The set $\mathcal{Q}_\alpha^T$ is closed in L^{2^T}, although it need not be compact since L^{2^T} is infinite-dimensional. Compactness will only be required after restricting this neighborhood to the finite-dimensional space $\mathcal{S}_M^T$. We now define the surrogate function $\widetilde{\varphi}_\alpha : \mathcal{Q}_\alpha^T \to L^{2^T}$,

$$\widetilde{\varphi}_\alpha(\boldsymbol{h}) := \sum_{i=1}^{J_\alpha} c_j(\boldsymbol{h}) \widetilde{\varphi}(\boldsymbol{q}_i), \tag{41}$$

where for every $j \in [J_\alpha]$, $c_j : \mathcal{Q}_\alpha^T \to [0,1]$ is defined as

$$c_j(\boldsymbol{h}) := \frac{\beta_j(\boldsymbol{h})}{\sum_{k=1}^{J_\alpha} \beta_k(\boldsymbol{h})}, \qquad \beta_j(\boldsymbol{h}) := \max\{0,\, 3\alpha - \mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}_j)\}. \tag{42}$$

The following proposition establishes the validity of our construction (proof in Appendix D.4).

Proposition D.3 (Validity of surrogate function)**.** *The following hold for construction (41).*

(a) The quantity $\sum_{k=1}^{J_\alpha} \beta_k(\boldsymbol{h})$ is positive and hence c_j is well-defined $\forall j \in [J_\alpha]$;

(b) The function $\widetilde{\varphi}_\alpha$ is continous on $\mathcal{Q}_\alpha^T$ relative to dist_T.

We further note some properties of the surrogate function that allow us to transfer bounds back onto $\widetilde{\varphi}$. Note that $\tilde{\varphi}$ is continuous over $\mathcal{Q}^T$ with respect to dist_T and hence uniformly continuous. We define its modulus of continuity below, and later use it to quantify the notion of closeness to $\widetilde{\varphi}$.

$$\omega_{\widetilde{\varphi}}(r) := \sup\left\{ \mathrm{dist}_T(\widetilde{\varphi}(\boldsymbol{q}), \widetilde{\varphi}(\boldsymbol{p})) \;\middle|\; \boldsymbol{q}, \boldsymbol{p} \in \mathcal{Q}^T,\ \mathrm{dist}_T(\boldsymbol{q}, \boldsymbol{p}) \le r \right\}, \tag{43}$$

with $\lim_{r \to 0} \omega_{\widetilde{\varphi}}(r) = 0$.

The following proposition establishes some useful properties of the surrogate function (proof in Appendix D.5).

Proposition D.4 (Properties of the surrogate function)**.** *The following hold for construction* (41)*.*

(a) If $\boldsymbol{q} \in \mathcal{Q}^T$ *and* $\boldsymbol{h} \in \mathcal{Q}_\alpha^T$ *satisfy* $\operatorname{dist}_T(\boldsymbol{h}, \boldsymbol{q}) \le \alpha$*, then* $\operatorname{dist}_T(\widetilde{\varphi}_\alpha(\boldsymbol{h}),\ \widetilde{\varphi}(\boldsymbol{q})) \le \omega_{\widetilde{\varphi}}(4\alpha)$*.*

(b) For any $s \in [T]$*, the function* $\widetilde{\varphi}_\alpha(\boldsymbol{h})_s$ *is a non-decreasing element of* L^2 *and hence the quantile function of some measure in* $\mathcal{P}_2(\mathbb{R})$*.*

Having defined the surrogate of $\widetilde{\varphi}$, we proceed with the construction that maps into the finite-dimensional space $(\mathbb{R}^M)^T$. Define

$$G_{M,\alpha} \;:=\; C_M^T \circ \Pi_M^T \circ \widetilde{\varphi}_\alpha \circ (C_M^{-1})^T : \mathcal{D}_{M,\alpha} \longrightarrow (\mathbb{R}^M)^T. \tag{44}$$

The map $G_{M,\alpha}$ is continuous and $\mathcal{D}_{M,\alpha} \;:=\; C_M^T\big(\mathcal{S}_M^T \cap \mathcal{Q}_\alpha^T\big) \subseteq (\mathbb{R}^M)^T$ is compact. To verify that the empirical token belongs to this domain with high probability, fix $\boldsymbol{\mu} = (\mu_1, \dots, \mu_T) \in \mathcal{K}^T$ and set

$$\boldsymbol{q} \;:=\; \Phi^T(\boldsymbol{\mu}) \in \mathcal{Q}^T, \qquad \widehat{\boldsymbol{q}}_M \;:=\; \Phi^T(\widehat{\boldsymbol{\mu}}_M) \in (\mathcal{S}_M)^T. \tag{45}$$

On $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$, the quantile isometry yields

$$d_T(\widehat{\boldsymbol{q}}_M, \boldsymbol{q}) = \max_{t\in[T]} \| \Phi(\widehat{\mu}_{t,M}) - \Phi(\mu_t) \|_{L^2} = \max_{t\in[T]} W_2(\widehat{\mu}_{t,M}, \mu_t) \le \alpha. \tag{46}$$

Since $\boldsymbol{q} \in \mathcal{Q}^T$, this implies $\widehat{\boldsymbol{q}}_M \in \mathcal{Q}_\alpha^T$. Since also $\widehat{\boldsymbol{q}}_M \in \mathcal{S}_M^T$, then on the event $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$,

$$E_M^T(\widehat{\boldsymbol{\mu}}_M) = C_M^T(\widehat{\boldsymbol{q}}_M) \in \mathcal{D}_{M,\alpha}. \tag{47}$$

The following uniform empirical convergence fact will be used to control the probability that an empirical token lies outside $\mathcal{D}_{M,a}$ (proof in Appendix B.3).

Proposition D.5 (Uniform empirical W_p consistency on compact classes; adaptation of Corollary 1 of Hallin et al. (2021))**.** *Let* $\mathcal{K} \subseteq \mathcal{P}_p(\mathbb{R}^d)$ *be compact. For every* $\mu \in \mathcal{K}$*, let* $X_1, X_2, \dots \overset{\text{i.i.d.}}{\sim} \mu$ *and* $\widehat{\mu}_M = M^{-1}\sum_{i=1}^M \delta_{X_i}$*. Then, for every* $\varepsilon > 0$ *and every fixed* $T \in \mathbb{N}_+$*,*

$$\lim_{M\to\infty} \sup_{\boldsymbol{\mu}\in\mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}}\left(\max_{t\in[T]} W_p^p(\widehat{\mu}_{t,M}, \mu_t) > \varepsilon\right) \;=\; 0. \tag{48}$$

Proof. The proof follows immediately from our assumption of compactness of $\mathcal{K}$ in $\mathcal{P}_p(\mathbb{R}^d)$ and Corollary 1 of Hallin et al. (2021), which we state here in full adapted to our notation.

Corollary D.1 (Corollary 1 of Hallin et al. (2021))**.** *Let* $\mathcal{K} \subseteq \mathcal{P}_p(\mathbb{R}^d)$ *be such that*

$$\lim_{r\to\infty} \sup_{\mu\in\mathcal{K}} \int_{\{\|x\|>r\}} \|x\|^p \, d\mu(x) = 0.$$

Then, it holds that

$$\forall \varepsilon > 0, \qquad \lim_{M\to\infty} \sup_{\mu\in\mathcal{K}} \mathbb{P}\left(W_p^p(\widehat{\mu}_M, \mu) > \varepsilon\right) = 0,$$

i.e., $W_p^p(\widehat{\mu}_M, \mu) \to 0$ *in probability as* $M \to \infty$*, uniformly in* $\mu \in \mathcal{K}$*.*

Since we assume $\mathcal{K}$ is compact w.r.t. W_p, by Assumption 2.1 the condition of the above corollary is fulfilled and the result follows. □

Using Proposition D.5, we have the high probability statement of the event $\mathcal{A}_{M,\alpha}$,

$$\sup_{\boldsymbol{\mu}\in\mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}}\big(E_M^T(\widehat{\boldsymbol{\mu}}_M) \notin \mathcal{D}_{M,\alpha}\big) \le \sup_{\boldsymbol{\mu}\in\mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}}\left(\max_{t\in[T]} W_2(\widehat{\mu}_{t,M}, \mu_t) > \alpha\right) \longrightarrow 0 \tag{49}$$

as $M \to \infty$, by uniform empirical Wasserstein consistency on the compact class $\mathcal{K}$.

D.4 Validity of the surrogate construction (Proposition D.3)

Proposition D.3 (Validity of surrogate function)**.** *The following hold for construction* (41)*.*

(a) The quantity $\sum_{k=1}^{J_\alpha} \beta_k(\boldsymbol{h})$ *is positive and hence* c_j *is well-defined* $\forall j \in [J_\alpha]$*;*

(b) The function $\widetilde{\varphi}_\alpha$ *is continous on* $\mathcal{Q}_\alpha^T$ *relative to* dist_T*.*

Proof. Point (a). For every $\boldsymbol{h} \in \mathcal{Q}_\alpha^T$, compactness of $\mathcal{Q}^T$ implies that there exists $\boldsymbol{q} \in \mathcal{Q}^T$ satisfying $\mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}) \le a$. By property (39), there is $j \in [J_a]$ such that $\mathrm{dist}_T(\boldsymbol{q}, \boldsymbol{q}^j) \le a$. Hence

$$\mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}^j) \le \mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}) + \mathrm{dist}_T(\boldsymbol{q}, \boldsymbol{q}^j) \le 2a, \tag{50}$$

and therefore $\beta_j(\boldsymbol{h}) \ge a > 0$.

Point (b). For each $j \in [J_a]$, the map $\boldsymbol{h} \mapsto d_T(\boldsymbol{h}, \boldsymbol{q}^j)$ is continuous. Hence β_j is continuous. The denominator in (42) is strictly positive on $\mathcal{Q}_\alpha^T$, so every c_j is continuous. Equation (41) then shows that $\widetilde{\varphi}_a$ is continuous. □

D.5 Properties of the surrogate construction (Proposition D.4)

Proposition D.4 (Properties of the surrogate function)**.** *The following hold for construction* (41)*.*

(a) If $\boldsymbol{q} \in \mathcal{Q}^T$ *and* $\boldsymbol{h} \in \mathcal{Q}_\alpha^T$ *satisfy* $\mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}) \le \alpha$*, then* $\mathrm{dist}_T(\widetilde{\varphi}_\alpha(\boldsymbol{h}),\ \widetilde{\varphi}(\boldsymbol{q})) \le \omega_{\widetilde{\varphi}}(4\alpha)$*.*

(b) For any $s \in [T]$*, the function* $\widetilde{\varphi}_\alpha(\boldsymbol{h})_s$ *is a non-decreasing element of* L^2 *and hence the quantile function of some measure in* $\mathcal{P}_2(\mathbb{R})$*.*

Proof. Point (a). Let $\boldsymbol{q} \in \mathcal{Q}^T$ and suppose that $\mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}) \le \alpha$. If $c_j(\boldsymbol{h}) > 0$, then $\mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}_j) < 3\alpha$, and therefore

$$\mathrm{dist}_T(\boldsymbol{q}, \boldsymbol{q}_j) \le \mathrm{dist}_T(\boldsymbol{q}, \boldsymbol{h}) + \mathrm{dist}_T(\boldsymbol{h}, \boldsymbol{q}_j) < 4\alpha. \tag{51}$$

Using the definition of dist_T, the triangle inequality in L^2, and the fact that the weights are nonnegative and sum to one, we get

$$\begin{aligned}
\mathrm{dist}_T(\widetilde{\varphi}_\alpha(\boldsymbol{h}), \widetilde{\varphi}(\boldsymbol{q})) &= \max_{s\in[T]} \left\| \sum_{j=1}^{J_\alpha} c_j(\boldsymbol{h}) \left([\widetilde{\varphi}(\boldsymbol{q}_j)]_s - [\widetilde{\varphi}(\boldsymbol{q})]_s \right) \right\|_{L^2} \\
&\le \sum_{j=1}^{J_\alpha} c_j(\boldsymbol{h}) \max_{s\in[T]} \| [\widetilde{\varphi}(\boldsymbol{q}_j)]_s - [\widetilde{\varphi}(\boldsymbol{q})]_s \|_{L^2} \\
&= \sum_{j=1}^{J_\alpha} c_j(\boldsymbol{h}) \, \mathrm{dist}_T(\widetilde{\varphi}(\boldsymbol{q}_j), \widetilde{\varphi}(\boldsymbol{q})) \\
&\le \omega_{\widetilde{\varphi}}(4\alpha),
\end{aligned} \tag{52}$$

where the last inequality follows from (51).

Point (b). For each $s \in [T]$,

$$[\widetilde{\varphi}_\alpha(\boldsymbol{h})]_s = \sum_{j=1}^{J_\alpha} c_j(\boldsymbol{h}) [\widetilde{\varphi}(\boldsymbol{q}_j)]_s.$$

By the definition of $\mathcal{Q}$ as $\Phi(\mathcal{K})$, each function $[\widetilde{\varphi}(\boldsymbol{q}_j)]_s$ is nondecreasing and belongs to L^2. A convex combination of such functions has the same properties. Hence $[\widetilde{\varphi}_\alpha(\boldsymbol{h})]_s$ is the quantile function, up to equality almost everywhere, of a measure in $\mathcal{P}_2(\mathbb{R})$. □

D.6 Uniform empirical quantile-token universality in probability on compact classes (Theorem 3.1)

Theorem 3.1 (High-probability approximation universality in $\mathcal{P}(\mathbb{R})$)**.** *Let $\mathcal{K} \subseteq \mathcal{P}_2(\mathbb{R})$ be a non-empty compact and let $\varphi \in \mathcal{C}_{W_2}(\mathcal{K}^T, \mathcal{K}^T)$. For every $\varepsilon > 0$ and every $\delta > 0$, there exists $M_0 \in \mathbb{N}$ such that, for every $M \geq M_0$, there exists θ and a transformer-based map of kind (5) ensuring*

$$\sup_{\boldsymbol{\mu} \in \mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}} \left(\max_{s \in [T]} W_2 \left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s \right) > \varepsilon \right) < \delta.$$

Proof. The route to proving this result goes through showing that the event

$$\mathcal{B}_{M,\theta,\alpha}(\boldsymbol{\mu})^C \;\coloneqq\; \left\{ \max_{s \in [T]} W_2(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s,\, \varphi(\boldsymbol{\mu})_s) > \varepsilon \right\} \;\subseteq\; \mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C, \tag{53}$$

where θ is the parameter that satisfies 14 and $A_{M,\alpha}(\boldsymbol{\mu})^C$ is the event defined in (14), repeated below for convenience

$$\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C \;\coloneqq\; \left\{ \max_{t \in [T]} W_2(\widehat{\mu}_{t,M}, \mu_t) > \alpha \right\}. \tag{54}$$

Once this result is established, we select an $M_1 = M_1(\delta, \alpha)$ such that for all $M \geq M_1$ it holds that $P(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C) < \delta$, it will immediately follow that

$$\mathbb{P}(\mathcal{B}_{M,\theta,\alpha}(\widehat{\boldsymbol{\mu}}_M, \boldsymbol{\mu})^C) \leq \mathbb{P}(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C) < \delta. \tag{55}$$

Moving forward, we study what happens on the event $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$. To begin, we use the triangle inequality with respect to $(D_M \circ G_{M,\alpha} \circ E_M^T)(\widehat{\boldsymbol{\mu}}_M)_s$ to get

$$\begin{aligned} \max_{s \in [T]} W_2(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s,\, \varphi(\boldsymbol{\mu})_s) \leq & \max_{s \in [T]} W_2(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s,\, (D_M \circ G_{M,\alpha} \circ E_M^T)(\widehat{\boldsymbol{\mu}}_M)_s) \\ & + \max_{s \in [T]} W_2((D_M \circ G_{M,\alpha} \circ E_M^T)(\widehat{\boldsymbol{\mu}}_M)_s,\, \varphi(\boldsymbol{\mu})_s) \end{aligned} \tag{56}$$

Note that the quantity $(D_M(G_{M,\alpha} \circ E_M^T)(\widehat{\boldsymbol{\mu}}_M)_s)$ is well defined on $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$, since, by definition

$$\begin{aligned} (G_{M,\alpha} \circ E_M^T)(\widehat{\boldsymbol{\mu}}_M) \;&\coloneqq\; C_M^T \circ \Pi_M^T \circ \widetilde{\varphi}_\alpha \circ (C_M^{-1})^T \circ C_M^T \circ \Phi^T(\widehat{\mu}_M) \\ &=\; C_M^T \circ \Pi_M^T \circ \widetilde{\varphi}_\alpha \circ \Phi^T(\widehat{\mu}_M), \end{aligned} \tag{57}$$

and by (46) we have that $\Phi^T(\widehat{\mu}_M) \in \mathcal{Q}_\alpha$, the domain of $\widetilde{\varphi}_\alpha$.

We now bound the terms of (56) separately, beginning with the first one. By construction,

$$\mathcal{D}_{M,\alpha} = C_M^T\big(\mathcal{S}_M^T \cap \mathcal{Q}_\alpha^T\big) = C_M^T\Big(\big(\mathcal{S}_M \cap \mathcal{Q}_\alpha\big)^T\Big) = \Big(C_M\left(\mathcal{S}_M \cap \mathcal{Q}_\alpha\right)\Big)^T$$

is a compact Cartesian product, and $G_{M,\alpha}$ is continuous on $\mathcal{D}_{M,\alpha}$. Therefore, by transformer universality Corollary 2.1, for any $\varepsilon > 0$ there exists θ such that

$$\sup_{\boldsymbol{z}_M \in \mathcal{D}_{M,\alpha}} \max_{s \in [T]} \| \mathcal{T}_\theta(\boldsymbol{z}_M)_s - G_{M,\alpha}(\boldsymbol{z}_M)_s \|_2 \leq \frac{\varepsilon}{3}. \tag{58}$$

Using the definition of $\mathcal{T}_{M,\theta}$, the decoder's Lipschitz property (13) and the notation $\boldsymbol{z}_M \coloneqq E_M^T(\widehat{\boldsymbol{\mu}}_M)$, the first term of (56) is processed as

$$\begin{aligned} & W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s,\, D_M\left(G_{M,\alpha} \circ E_M^T(\widehat{\boldsymbol{\mu}}_M)_s\right)\right) \\ & \qquad = W_2\left(D_M\left(\mathcal{T}_\theta \circ E_M^T(\widehat{\boldsymbol{\mu}}_M)_s\right),\, D_M\left(G_{M,\alpha} \circ E_M^T(\widehat{\boldsymbol{\mu}}_M)_s\right)\right) \\ & \qquad \leq \| \mathcal{T}_\theta(\boldsymbol{z}_M)_s - G_{M,\alpha}(\boldsymbol{z}_M)_s \|_2 . \end{aligned} \tag{59}$$

We have shown that for all realizations $\widehat{\boldsymbol{\mu}}_M$ of $\boldsymbol{\mu}$ satisfying event $\mathcal{A}_{M,\alpha}$, it holds that $\boldsymbol{z}_M \in D_{M,\alpha}$. Therefore, transformer universality applies and we have that for any $\epsilon > 0$, there exists θ such that for any realization of empirical measures within $\mathcal{A}_{M,\alpha}$

$$\max_{s \in [T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s,\, D_M\left(G_{M,\alpha} \circ E_M^T(\widehat{\boldsymbol{\mu}}_M)_s\right)\right) \leq \sup_{\boldsymbol{z}_M \in \mathcal{D}_{M,\alpha}} \max_{s \in [T]} \| \mathcal{T}_\theta(\boldsymbol{z}_M)_s - G_{M,\alpha}(\boldsymbol{z}_M)_s \|_2$$

$$\leq \frac{\varepsilon}{3} \tag{60}$$

We move to the second term of (56), which dictates the lower bounds on M. Note that, by definition, $\varphi = \Phi^{-1} \circ \widetilde{\varphi} \circ \Phi$. Using the quantile isometry and the expansion in (57), we get that

$$\begin{aligned} &W_2(D_M(G_{M,\alpha} \circ E_M^T(\widehat{\boldsymbol{\mu}}_M)_s),\, \varphi(\boldsymbol{\mu})_s) \\ &\qquad = \left\| \Phi \circ D_M(C_M^T \circ \Pi_M^T \circ \widetilde{\varphi}_\alpha \circ \Phi^T(\widehat{\mu}_M)_s) - (\widetilde{\varphi} \circ \Phi^T)(\boldsymbol{\mu})_s \right\|_{L^2} \\ &\qquad = \left\| \Phi \circ \Delta_M(\Pi_M^T \circ \widetilde{\varphi}_\alpha \circ \Phi^T(\widehat{\boldsymbol{\mu}}_M)_s) - (\widetilde{\varphi} \circ \Phi^T)(\boldsymbol{\mu})_s \right\|_{L^2}, \end{aligned} \tag{61}$$

where the second equality comes from the definition of D_M in (12). Next, let $\widehat{\boldsymbol{q}}_M := \Phi^T(\widehat{\boldsymbol{\mu}}_M) \in \mathcal{Q}_\alpha$ and $\boldsymbol{q} := \Phi^T(\boldsymbol{\mu}) \in \mathcal{Q}$. From Proposition D.4, we know that all components of $\widetilde{\varphi}_\alpha(\widehat{\boldsymbol{q}}_M)$ are non-decreasing elements of L^2 and hence the quantile functions of some measures in $\mathcal{P}_2(\mathbb{R})$. Therefore, letting $\boldsymbol{h}_M := \widetilde{\varphi}_\alpha(\widehat{\boldsymbol{q}}_M)$ and applying Proposition D.2 we continue (61) as

$$\begin{aligned} &\left\| \Phi \circ \Delta_M(\Pi_M^T \circ \widetilde{\varphi}_\alpha \circ \Phi^T(\widehat{\boldsymbol{\mu}}_M)_s) - (\widetilde{\varphi} \circ \Phi^T)(\boldsymbol{\mu})_s \right\|_{L^2} \\ &\qquad = \left\| \Pi_M^T(\boldsymbol{h}_M)_s - \widetilde{\varphi}(\boldsymbol{q})_s \right\|_{L^2} \\ &\qquad \leq \left\| \Pi_M^T(\boldsymbol{h}_M)_s - (\boldsymbol{h}_M)_s \right\|_{L^2} + \left\| \widetilde{\varphi}_\alpha(\widehat{\boldsymbol{q}}_M)_s - \widetilde{\varphi}(\boldsymbol{q})_s \right\|_{L^2} \end{aligned} \tag{62}$$

where the first inequality comes from the triangle inequality and the definition of $\boldsymbol{h}_M$. Since these developments hold for any $s \in [T]$, we apply the maximum and combine (61) with (62) to get

$$\begin{aligned} &\max_{s\in[T]} W_2(D_M(G_{M,\alpha} \circ E_M^T(\widehat{\boldsymbol{\mu}}_M)_s),\, \varphi(\boldsymbol{\mu})_s) \\ &\qquad \leq \max_{s\in[T]} \left\| \Pi_M^T(\boldsymbol{h}_M)_s - (\boldsymbol{h}_M)_s \right\|_{L^2} + \max_{s\in[T]} \left\| \widetilde{\varphi}_\alpha(\widehat{\boldsymbol{q}}_M)_s - \widetilde{\varphi}(\boldsymbol{q})_s \right\|_{L^2} \\ &\qquad = \max_{s\in[T]} \left\| \Pi_M^T(\boldsymbol{h}_M)_s - (\boldsymbol{h}_M)_s \right\|_{L^2} + \omega_{\widetilde{\varphi}}(4\alpha). \end{aligned} \tag{63}$$

We process (63) one last time,

$$\begin{aligned} \max_{s\in[T]} \left\| \Pi_M^T(\boldsymbol{h}_M)_s - (\boldsymbol{h}_M)_s \right\|_{L^2} &= \max_{s\in[T]} \left\| \Pi_M^T(\widetilde{\varphi}_\alpha(\widehat{\boldsymbol{q}}_M))_s - (\widetilde{\varphi}_\alpha(\widehat{\boldsymbol{q}}_M))_s \right\|_{L^2} \\ &= \max_{s\in[T]} \left\| \sum_{i=1}^{J_\alpha} c_j(\boldsymbol{h}_M) \left(\Pi_M(\widetilde{\varphi}(\boldsymbol{q}_i)_s) - \widetilde{\varphi}(\boldsymbol{q}_i)_s\right) \right\|_{L^2} \\ &\leq \max_{s\in[T]} \sum_{i=1}^{J_\alpha} c_j(\boldsymbol{h}_M) \left\| \Pi_M(\widetilde{\varphi}(\boldsymbol{q}_i)_s) - \widetilde{\varphi}(\boldsymbol{q}_i)_s \right\|_{L^2} \\ &\leq \max_{s\in[T]} \max_{i\in[J_\alpha]} \left\| \Pi_M(\widetilde{\varphi}(\boldsymbol{q}_i)_s) - \widetilde{\varphi}(\boldsymbol{q}_i)_s \right\|_{L^2}, \end{aligned} \tag{64}$$

where the first equality comes from the definition of $\boldsymbol{h}_M$; the second equality comes from the definition of $\widetilde{\varphi}_\alpha$; the first inequality comes from the triangle inequality; and the final inequality comes from the fact that for any $M \in \mathbb{N}_+$ and any α-net, $c_j(\boldsymbol{h}_M) \in [0, 1]$. Putting (64) and (63) together, we have that

$$\begin{aligned} &\max_{s\in[T]} W_2(D_M(G_{M,\alpha} \circ E_M^T(\widehat{\boldsymbol{\mu}}_M)_s),\, \varphi(\boldsymbol{\mu})_s) \\ &\qquad \leq \max_{s\in[T]} \max_{i\in[J_\alpha]} \left\| \Pi_M(\widetilde{\varphi}(\boldsymbol{q}_i)_s) - \widetilde{\varphi}(\boldsymbol{q}_i)_s \right\|_{L^2} + \omega_{\widetilde{\varphi}}(4\alpha) \end{aligned} \tag{65}$$

Since we know that $\lim_{M\to\infty} \left\| \Pi_M(\widetilde{\varphi}(\boldsymbol{q}_i)_s) - \widetilde{\varphi}(\boldsymbol{q}_i)_s \right\|_{L^2} = 0$ by Proposition 3.1 and $\lim_{\alpha\downarrow 0} \omega_{\widetilde{\varphi}}(4\alpha) = 0$ and $T, J_\alpha < \infty$, there exists $M_2 = M_2(\mathcal{K}) \in \mathbb{N}_+$ such that for any $M \geq M_2$

$$\max_{s\in[T]} \max_{i\in[J_\alpha]} \left\| \Pi_M(\widetilde{\varphi}(\boldsymbol{q}_i)_s) - \widetilde{\varphi}(\boldsymbol{q}_i)_s \right\|_{L^2} \leq \frac{\varepsilon}{3}, \tag{66}$$

and an α sufficiently small such that

$$\omega_{\widetilde{\varphi}}(4\alpha) \leq \frac{\varepsilon}{3}. \tag{67}$$

Finally, we have that for any $M \geq M_2$ and a small enough α, the second term of (56) is bounded as follows on the event $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$

$$W_2(D_M(G_{M,\alpha} \circ E_M^T(\widehat{\boldsymbol{\mu}}_M)_s),\, \varphi(\boldsymbol{\mu})_s) \leq \frac{2\varepsilon}{3}. \tag{68}$$

Putting (60) and (68) together, we have that for any fixed $\varepsilon, \delta > 0$, there exist $\alpha = \alpha(\varepsilon) > 0$ and $M_0 = M_0(\varepsilon, \delta, \alpha, \mathcal{K}) := \max\{M_2(\mathcal{K}), M_1(\alpha, \delta, \varepsilon)\}$ such that for any $M \geq M_0$ there exists θ which, for all $\boldsymbol{\mu} \in \mathcal{K}$ ensures that

$$\mathcal{B}_{M,\theta,\alpha}(\boldsymbol{\mu}) = \left\{ \max_{s\in[T]} W_2(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s) < \varepsilon \right\} \tag{69}$$

happens whenever $\mathcal{A}_{M,\alpha}$ happens. This means that $\mathcal{A}_{M,\alpha} \subseteq \mathcal{B}_{M,\theta,\alpha}$ and therefore $\mathcal{B}^C_{M,\theta,\alpha} \subseteq \mathcal{A}^C_{M,\alpha}$. The result immediately follows. □

D.7 Uniform empirical quantile-token universality in expectation on compact classes (Corollary 3.1)

Corollary 3.1 (Expectation approximation universality in $\mathcal{P}(\mathbb{R})$)**.** *Let $\mathcal{K} \subseteq \mathcal{P}_2(\mathbb{R})$ be compact, $\varphi \in \mathcal{C}_{W_2}(\mathcal{K}^T, \mathcal{K}^T)$, and $B_{\mathcal{K}} := \max_{\nu\in\mathcal{K}} W_2(\nu, \delta_0)$. For $B \geq 0$, let $P_{B,M} : \mathbb{R}^M \to \mathbb{R}^M$ denote the Euclidean projection onto $\{z \in \mathbb{R}^M : \| z \|_2 \leq B\}$. Then, for every $\varepsilon > 0$, there exists $M_0 \in \mathbb{N}$ such that, for every $M \geq M_0$, there exists a clipped transformer $\mathcal{T}^{\text{clip}}_{M,\theta} := D^T_M \circ P^T_{B_{\mathcal{K}},M} \circ \mathcal{T}_\theta \circ E^T_M$ satisfying $\sup_{\boldsymbol{\mu}\in\mathcal{K}^T} \mathbb{E}_{\boldsymbol{\mu}}\left[\max_{s\in[T]} W_2\left(\mathcal{T}^{\text{clip}}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right)\right] < \varepsilon$.*

Proof. Theorem 3.1 already provides uniform convergence in probability. This, however, does not by itself yield convergence in expectation, because the Wasserstein error could be arbitrarily large on the exceptional event where the high-probability approximation fails. The purpose of the clipping layer is to remove this possibility. More precisely, we first show that clipping cannot increase the distance to any possible target in $\mathcal{K}$, while at the same time placing every decoded output in a uniformly bounded W_2 ball. We then combine this deterministic bound with the high-probability conclusion of Theorem 3.1.

We begin by verifying that the radius is well-defined. By the reverse triangle inequality for W_2, for every $\nu, \nu' \in \mathcal{P}_2(\mathbb{R})$,

$$|W_2(\nu, \delta_0) - W_2(\nu', \delta_0)| \leq W_2(\nu, \nu').$$

Thus, the map $\nu \mapsto W_2(\nu, \delta_0)$ is continuous, and even 1-Lipschitz, on $\mathcal{P}_2(\mathbb{R})$. Since $\mathcal{K}$ is compact, this map attains a finite maximum on $\mathcal{K}$. We next establish the key geometric property of the clipping layer. Fix $M \in \mathbb{N}_+$, $z \in \mathbb{R}^M$, and $\nu \in \mathcal{K}$, and introduce the quantile functions

$$q_z := \Phi(D_M(z)), \qquad q_\nu := \Phi(\nu). \tag{70}$$

The normalization used in the decoder (12) implies that the Euclidean norm of the token is exactly the L^2 norm of the decoded quantile function. Indeed,

$$\| q_z \|^2_{L^2} = W_2^2(D_M(z), \delta_0) = \frac{1}{M} \sum_{i=1}^{M} \left|\sqrt{M} z_i\right|^2 = \| z \|_2^2 . \tag{71}$$

Thus, the radial size of a token in $\mathbb{R}^M$ agrees exactly with the second-moment radius of the measure obtained after decoding. Let

$$c_z := \begin{cases} 1, & \| z \|_2 \leq B_{\mathcal{K}}, \\ \dfrac{B_{\mathcal{K}}}{\| z \|_2}, & \| z \|_2 > B_{\mathcal{K}}. \end{cases} \tag{72}$$

Then

$$P_{B_{\mathcal{K}},M}(z) = c_z z, \qquad 0 \leq c_z \leq 1. \tag{73}$$

Multiplying z by the nonnegative scalar c_z multiplies every atom of $D_M(z)$ by the same scalar. Since a nonnegative scaling preserves the ordering of the atoms, the corresponding quantile function is scaled by the same factor. Consequently, up to equality almost everywhere,

$$\Phi\left(D_M(P_{B_{\mathcal{K}},M}(z))\right) = c_z q_z. \tag{74}$$

Consider the closed L^2 ball

$$\mathbb{B}_{\mathcal{K}} := \left\{ q \in L^2 : \| q \|_{L^2} \leq B_{\mathcal{K}} \right\}. \tag{75}$$

By (71), the right-hand side of (74) is precisely the metric projection of q_z onto $\mathbb{B}_{\mathcal{K}}$. In other words,

$$\Phi\left(D_M(P_{B_{\mathcal{K}},M}(z))\right) = \mathrm{proj}_{\mathbb{B}_{\mathcal{K}}}(q_z). \tag{76}$$

On the other hand, because $\nu \in \mathcal{K}$, the definition of $B_{\mathcal{K}}$ and the quantile isometry give

$$\| q_\nu \|_{L^2} = W_2(\nu, \delta_0) \le B_{\mathcal{K}}. \tag{77}$$

Hence $q_\nu \in \mathbb{B}_{\mathcal{K}}$ and therefore

$$\mathrm{proj}_{\mathbb{B}_{\mathcal{K}}}(q_\nu) = q_\nu. \tag{78}$$

The metric projection onto a closed convex subset of a Hilbert space is nonexpansive. Applying this fact to the ball $\mathbb{B}_{\mathcal{K}}$ and then using the quantile isometry (8), we obtain

$$\begin{aligned} W_2\left(D_M(P_{B_{\mathcal{K}},M}(z)), \nu\right) &= \left\| \mathrm{proj}_{\mathbb{B}_{\mathcal{K}}}(q_z) - q_\nu \right\|_{L^2} \\ &= \left\| \mathrm{proj}_{\mathbb{B}_{\mathcal{K}}}(q_z) - \mathrm{proj}_{\mathbb{B}_{\mathcal{K}}}(q_\nu) \right\|_{L^2} \\ &\le \| q_z - q_\nu \|_{L^2} \\ &= W_2(D_M(z), \nu). \end{aligned} \tag{79}$$

Therefore, clipping the output token cannot increase its decoded W_2 distance to any target measure in $\mathcal{K}$. The same quantile-space interpretation also gives the required uniform bound on the clipped output. Indeed,

$$\begin{aligned} W_2\left(D_M(P_{B_{\mathcal{K}},M}(z)), \delta_0\right) &= \| \Phi\left(D_M(P_{B_{\mathcal{K}},M}(z))\right) \|_{L^2} \\ &= \left\| \mathrm{proj}_{\mathbb{B}_{\mathcal{K}}}(q_z) \right\|_{L^2} \le B_{\mathcal{K}}. \end{aligned} \tag{80}$$

We now transfer these two deterministic properties to the full sequence-to-sequence architecture. For a fixed M and fixed transformer parameters θ, let

$$\mathcal{T}_{M,\theta} \;\coloneqq\; D_M^T \circ \mathcal{T}_\theta \circ E_M^T \tag{81}$$

denote the un-clipped architecture from Theorem 3.1. Define the un-clipped and clipped random losses, respectively, by

$$L_{M,\theta}(\boldsymbol{\mu}) \;\coloneqq\; \max_{s\in[T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right), \tag{82}$$

$$L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) \;\coloneqq\; \max_{s\in[T]} W_2\left(\mathcal{T}^{\mathrm{clip}}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right). \tag{83}$$

Both quantities are random through the empirical sequence $\widehat{\boldsymbol{\mu}}_M$. For each output index $s \in [T]$, the target $\varphi(\boldsymbol{\mu})_s$ belongs to $\mathcal{K}$. We may therefore apply (79) separately to every output token of the transformer. Taking the maximum over s gives the pointwise domination

$$L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) \le L_{M,\theta}(\boldsymbol{\mu}) \tag{84}$$

for every $\boldsymbol{\mu} \in \mathcal{K}^T$ and every realization of the empirical samples. Moreover, by (80), every clipped prediction has W_2 distance at most $B_{\mathcal{K}}$ from δ_0. Every target $\varphi(\boldsymbol{\mu})_s \in \mathcal{K}$ has the same property by the definition of $B_{\mathcal{K}}$. Consequently, the triangle inequality through δ_0 gives, for every $s \in [T]$,

$$W_2\left(\mathcal{T}^{\mathrm{clip}}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) \le W_2\left(\mathcal{T}^{\mathrm{clip}}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \delta_0\right) + W_2\left(\delta_0, \varphi(\boldsymbol{\mu})_s\right) \le 2B_{\mathcal{K}}. \tag{85}$$

Taking the maximum over the output indices yields the deterministic bound

$$0 \le L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) \le 2B_{\mathcal{K}} \tag{86}$$

for every $\boldsymbol{\mu} \in \mathcal{K}^T$ and every sample realization. This is the additional control that is not provided by convergence in probability alone. If $B_{\mathcal{K}} = 0$, then every $\nu \in \mathcal{K}$ satisfies $W_2(\nu, \delta_0) = 0$, and therefore $\mathcal{K} = \{\delta_0\}$. In this case, $P_{0,M}$ maps every output token to the origin and $D_M(0_M) = \delta_0$. Since the target also equals δ_0 in every component, the clipped loss is identically zero. We may therefore assume for the remainder of the proof that $B_{\mathcal{K}} > 0$. Fix $\varepsilon > 0$, and set

$$\eta \;\coloneqq\; \frac{\varepsilon}{2}, \qquad \delta \;\coloneqq\; \min\left\{1, \frac{\varepsilon}{8B_{\mathcal{K}}}\right\}. \tag{87}$$

Apply Theorem 3.1 with approximation threshold η and probability level δ. There exists $M_0 \in \mathbb{N}$ such that, for every $M \geq M_0$, one can choose transformer parameters θ satisfying

$$\sup_{\boldsymbol{\mu} \in \mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}} \left(L_{M,\theta}(\boldsymbol{\mu}) > \eta \right) < \delta. \tag{88}$$

We use exactly the same parameters θ in the clipped architecture. By (84),

$$\left\{ L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) > \eta \right\} \subseteq \left\{ L_{M,\theta}(\boldsymbol{\mu}) > \eta \right\}. \tag{89}$$

Thus, the clipped loss inherits the same high-probability guarantee. It remains to convert this probability bound into an expectation bound. Fix an arbitrary $\boldsymbol{\mu} \in \mathcal{K}^T$ and split the expectation according to whether the clipped loss is at most η or greater than η. Using (86), we obtain

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\mu}} \left[L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) \right] &= \mathbb{E}_{\boldsymbol{\mu}} \left[L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) \mathbf{1}_{\{L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) \leq \eta\}} \right] + \mathbb{E}_{\boldsymbol{\mu}} \left[L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) \mathbf{1}_{\{L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) > \eta\}} \right] \\ &\leq \eta + 2B_{\mathcal{K}} \mathbb{P}_{\boldsymbol{\mu}} \left(L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) > \eta \right) \\ &\leq \eta + 2B_{\mathcal{K}} \mathbb{P}_{\boldsymbol{\mu}} \left(L_{M,\theta}(\boldsymbol{\mu}) > \eta \right), \end{aligned} \tag{90}$$

where the last inequality follows from (89). Taking the supremum over $\boldsymbol{\mu} \in \mathcal{K}^T$ and using (88) gives

$$\sup_{\boldsymbol{\mu} \in \mathcal{K}^T} \mathbb{E}_{\boldsymbol{\mu}} \left[L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu}) \right] < \eta + 2B_{\mathcal{K}}\delta \leq \frac{\varepsilon}{2} + 2B_{\mathcal{K}} \frac{\varepsilon}{8B_{\mathcal{K}}} = \frac{3\varepsilon}{4} < \varepsilon. \tag{91}$$

By the definition of $L^{\mathrm{clip}}_{M,\theta}(\boldsymbol{\mu})$ in (83), this is precisely the final statement.

□

D.8 Quantitative sample-size thresholds (Proposition 3.2)

The proof of Theorem 3.1 first chooses $\alpha > 0$ so that the surrogate error $\omega_{\widetilde{\varphi}}(4\alpha)$ is at most $\varepsilon/3$ and then introduces two sample-size thresholds: M_1 controls the probability of the bad empirical event $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C$, whereas M_2 controls the cell-average projection error. Proposition 3.2 below makes the dependence of both thresholds on α, ε and δ explicit under three standard regularity assumptions.

In what follows, the empirical estimates use the Dvoretzky–Kiefer–Wolfowitz inequality with the sharp constant of Massart (1990) and the transportation-cost concentration bound of Larsson et al. (2024). The corresponding equal-weight quantization problem is studied in Xu & Berger (2019) and Bencheikh & Jourdain (2022).

Proposition 3.2. *Fix $\alpha, \varepsilon > 0$ and $\delta \in (0,1)$. Under any one of the three alternative assumptions below, there exist $M_1, M_2 \in \mathbb{N}_+$ such that $\sup_{\boldsymbol{\mu} \in \mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}} \left(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C \right) \leq \delta$, for every $M \geq M_1$, and $\sup_{\mu \in \mathcal{K}} \| \Pi_M \Phi(\mu) - \Phi(\mu) \|_{L^2} \leq \frac{\varepsilon}{3}$, for every $M \geq M_2$.*

- *(a) Suppose that, for some $D > 0$, every $\mu \in \mathcal{K}$ satisfies $\operatorname{diam}(\operatorname{supp} \mu) \leq D$. Then $M_1 = O\left(D^4/\alpha^4 \log\left(T/\delta\right) \right)$, and $M_2 = O\left(D^2/\varepsilon^2 \right)$.*
- *(b) Suppose that, for some $\sigma > 0$, $\mathcal{K}$ is uniformly σ-sub-Gaussian and that $0 < \alpha \leq \sigma$. Then $M_1 = O\left(\sigma^4/\alpha^4 \log\left(\mathrm{T}/\delta\right) \right)$, and $M_2 = O\left(\sigma^2/\varepsilon^2 \log\left(\sigma/\varepsilon\right) \right)$.*
- *(c) Suppose that, for some $L > 0$, the quantile functions are uniformly L-Lipschitz. Then $M_1 = O\left(L^2/\alpha^2 \log\left(T/\delta\right) \right)$, and $M_2 = O\left(L/\varepsilon \right)$.*

Proof. We first introduce the notation used in the proof. Set

$$\rho_M(\mathcal{K}) \coloneqq \sup_{\mu \in \mathcal{K}} \| \Pi_M \Phi(\mu) - \Phi(\mu) \|_{L^2}.$$

Fix $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_T) \in \mathcal{K}^T$. For each $t \in [T]$, let $X_{t,1}, \ldots, X_{t,M}$ be i.i.d. with law μ_t and define

$$\widehat{\mu}_{t,M} \coloneqq \frac{1}{M} \sum_{j=1}^{M} \delta_{X_{t,j}}, \qquad \mathcal{A}_{M,\alpha}(\boldsymbol{\mu}) \coloneqq \left\{ \max_{t \in [T]} W_2(\widehat{\mu}_{t,M}, \mu_t) \leq \alpha \right\}.$$

Let $\mathbb{P}_{\boldsymbol{\mu}}$ denote the joint sampling law of these random variables. We derive finite-M bounds in each of the three regimes and then select M_1 and M_2. The union bound gives

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C\right) \le \sum_{t=1}^{T} \mathbb{P}_{\mu_t}\left(W_2(\widehat{\mu}_{t,M}, \mu_t) > \alpha\right). \tag{92}$$

Thus, no independence across $t \in [T]$ is needed, and it suffices to control the empirical Wasserstein error for an arbitrary $\mu \in \mathcal{K}$. For any probability measure λ on $\mathbb{R}$, write $\widehat{\lambda}_M$ for the empirical measure of M i.i.d. samples with law λ, and write $\mathbb{P}_\lambda$ for the corresponding sampling law. For a probability measure ν on $\mathbb{R}$, write $F_\nu(x) := \nu((-\infty, x])$ and define

$$\kappa_M(\mu) := \sup_{x\in\mathbb{R}} |F_{\widehat{\mu}_M}(x) - F_\mu(x)|.$$

The Dvoretzky–Kiefer–Wolfowitz inequality gives

$$\mathbb{P}_\mu\left(\kappa_M(\mu) > \eta\right) \le 2\exp(-2M\eta^2), \qquad \eta > 0. \tag{93}$$

We begin with the bounded-support case. Fix $\mu \in \mathcal{K}$, write $q = \Phi(\mu)$ and $\widehat{q}_M = \Phi(\widehat{\mu}_M)$, and let I_μ be the smallest closed interval containing $\operatorname{supp}\mu$. The length of I_μ is at most D, and $\widehat{\mu}_M$ is supported on I_μ almost surely. Hence $|\widehat{q}_M(u) - q(u)| \le D$ for almost every $u \in (0,1)$. Using the one-dimensional representations of W_2 and W_1, we obtain

$$\begin{aligned} W_2^2(\widehat{\mu}_M, \mu) &= \int_0^1 |\widehat{q}_M(u) - q(u)|^2 \,\mathrm{d}u \\ &\le D\int_0^1 |\widehat{q}_M(u) - q(u)| \,\mathrm{d}u \\ &= D\, W_1(\widehat{\mu}_M, \mu) \\ &= D\int_{\mathbb{R}} |F_{\widehat{\mu}_M}(x) - F_\mu(x)| \,\mathrm{d}x \\ &\le D^2 \kappa_M(\mu). \end{aligned} \tag{94}$$

Consequently, $\{W_2(\widehat{\mu}_M, \mu) > \alpha\} \subseteq \{\kappa_M(\mu) > \alpha^2/D^2\}$. Combining this inclusion with (93), substituting the result into (92), and taking the supremum over $\boldsymbol{\mu} \in \mathcal{K}^T$ gives

$$\sup_{\boldsymbol{\mu}\in\mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}}\left(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C\right) \le 2T\exp\left(-\frac{2M\alpha^4}{D^4}\right). \tag{95}$$

It follows that one may take

$$M_1^{\mathrm{bd}}(\alpha,\delta) := \left\lceil \frac{D^4}{2\alpha^4}\log\left(\frac{2T}{\delta}\right)\right\rceil = O\left(1 + \frac{D^4}{\alpha^4}\log\left(\frac{2T}{\delta}\right)\right). \tag{96}$$

We next control the projection error under the same assumption. The bounded and nondecreasing function q has finite one-sided limits at 0 and 1, which we use to extend it to $[0,1]$. For every $i \in [M]$, let

$$\Delta_{i,M} := q\left(\frac{i}{M}\right) - q\left(\frac{i-1}{M}\right).$$

Since q is nondecreasing,

$$\Delta_{i,M} \ge 0, \qquad \sum_{i=1}^{M} \Delta_{i,M} = q(1) - q(0) \le D. \tag{97}$$

The values of q on $I_{i,M}$ lie, up to a null set, in an interval of length at most $\Delta_{i,M}$. The variance of a random variable supported on an interval of length Δ is at most $\Delta^2/4$. Because $(\Pi_M q)|_{I_{i,M}}$ is the average of q on $I_{i,M}$, we therefore have

$$\int_{I_{i,M}} |q(u) - (\Pi_M q)(u)|^2 \,\mathrm{d}u \le \frac{\Delta_{i,M}^2}{4M}.$$

Summing over the cells and using (97) yields

$$\| q - \Pi_M q \|_{L^2}^2 \le \frac{1}{4M} \sum_{i=1}^{M} \Delta_{i,M}^2 \le \frac{1}{4M} \left(\sum_{i=1}^{M} \Delta_{i,M} \right)^2 \le \frac{D^2}{4M}.$$

Taking the square root and then the supremum over $\mu \in \mathcal{K}$ gives

$$\rho_M(\mathcal{K}) \le \frac{D}{2\sqrt{M}}. \tag{98}$$

Hence it is enough to take

$$M_2^{\mathrm{bd}}(\varepsilon) := \left\lceil \frac{9D^2}{4\varepsilon^2} \right\rceil = O\left(1 + \frac{D^2}{\varepsilon^2}\right). \tag{99}$$

We next consider the sub-Gaussian case. By uniform σ-sub-Gaussianity, for every $\mu \in \mathcal{K}$ one can choose a center $a_\mu \in \mathbb{R}$ such that

$$\mu\big(|x - a_\mu| > t\big) \le 2 \exp\left(-\frac{t^2}{2\sigma^2}\right), \qquad t \ge 0. \tag{100}$$

Fix $\mu \in \mathcal{K}$, choose such a center, and define

$$S_\mu(x) := \frac{x - a_\mu}{\sigma}, \qquad \nu := (S_\mu)_{\#}\mu, \qquad \widehat{\nu}_M := (S_\mu)_{\#}\widehat{\mu}_M.$$

If $Y \sim \nu$, then $\mathbb{P}(|Y| > t) \le 2\exp(-t^2/2)$ for every $t \ge 0$. Applying Tonelli's theorem to $\exp(|Y|^2/4) = 1 + \int_0^{|Y|} (t/2)\exp(t^2/4)\,\mathrm{d}t$ gives

$$\begin{aligned} \int_{\mathbb{R}} \exp\left(\frac{|y|^2}{4}\right) \mathrm{d}\nu(y) &= 1 + \int_0^\infty \frac{t}{2} \exp\left(\frac{t^2}{4}\right) \nu(|y| > t)\,\mathrm{d}t \\ &\le 1 + \int_0^\infty t \exp\left(-\frac{t^2}{4}\right) \mathrm{d}t = 3. \end{aligned} \tag{101}$$

We apply Larsson et al. (2024, Example 3.4(a)) with dimension $d = 1$, transportation order $p = 2$, exponential-moment order $\beta = 2$, and exponential coefficient $1/4$. In the notation of that result, the transportation cost is the unrooted quantity W_2^2 and, since $p > d/2$, the small-deviation rate function is x^2. The parameters of the cited result are fixed, and (101) bounds its exponential-moment quantity by 3 uniformly over $\mu \in \mathcal{K}$. Consequently, there exist numerical constants $C \ge 1$ and $c > 0$ such that, for every $0 < x \le 1$,

$$\mathbb{P}_\nu\left(W_2^2(\widehat{\nu}_M, \nu) > x\right) \le C \exp(-cMx^2). \tag{102}$$

At $x = 1$, the two terms in the cited estimate are absorbed by enlarging C if necessary. Since S_μ is a similarity of ratio $1/\sigma$,

$$W_2(\widehat{\nu}_M, \nu) = \frac{1}{\sigma} W_2(\widehat{\mu}_M, \mu).$$

For $0 < \alpha \le \sigma$, substituting $x = (\alpha/\sigma)^2$ into (102), then using (92), gives

$$\sup_{\boldsymbol{\mu} \in \mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}}\left(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C\right) \le CT \exp\left(-cM\left(\frac{\alpha}{\sigma}\right)^4\right). \tag{103}$$

It is therefore enough to take

$$M_1^{\mathrm{sg}}(\alpha, \delta) := \left\lceil \frac{1}{c}\left(\frac{\sigma}{\alpha}\right)^4 \log\left(\frac{CT}{\delta}\right) \right\rceil = O\left(\left(\frac{\sigma}{\alpha}\right)^4 \log\left(\frac{\mathrm{e}T}{\delta}\right)\right), \tag{104}$$

where the last relation uses $0 < \alpha \le \sigma$ and the fact that C and c are numerical constants. We now turn to the projection error. Write $q = \Phi(\mu)$, let $a = a_\mu$, and, for $R > 0$, define the clipped quantile

$$q_R(u) := a + \max\{-R, \min\{q(u) - a, R\}\}.$$

The function q_R is nondecreasing, agrees with q whenever $q(u) \in [a-R, a+R]$, and has range diameter at most $2R$. Applying the bounded-range calculation leading to (98) directly to q_R, and using that $\mathrm{Id} - \Pi_M$ is an orthogonal projection and hence a contraction on L^2, yields

$$\begin{aligned} \| \Pi_M q - q \|_{L^2} &\le \| (\mathrm{Id} - \Pi_M) q_R \|_{L^2} + \| (\mathrm{Id} - \Pi_M)(q - q_R) \|_{L^2} \\ &\le \frac{R}{\sqrt{M}} + \| q - q_R \|_{L^2}. \end{aligned} \tag{105}$$

By the definition of q_R and the quantile transform,

$$\| q - q_R \|_{L^2}^2 = \int_{\mathbb{R}} (|x-a| - R)_+^2 \, \mathrm{d}\mu(x).$$

Moreover, Tonelli's theorem gives

$$\int_{\mathbb{R}} (|x-a| - R)_+^2 \, \mathrm{d}\mu(x) = 2 \int_R^\infty (t-R) \mu(|x-a| > t) \, \mathrm{d}t. \tag{106}$$

Combining (100) and (106), we obtain

$$\begin{aligned} \int_{\mathbb{R}} (|x-a_\mu| - R)_+^2 \, \mathrm{d}\mu(x) &\le 4 \int_R^\infty (t-R) \exp\left(-\frac{t^2}{2\sigma^2}\right) \mathrm{d}t \\ &\le 4 \int_R^\infty t \exp\left(-\frac{t^2}{2\sigma^2}\right) \mathrm{d}t \\ &= 4\sigma^2 \exp\left(-\frac{R^2}{2\sigma^2}\right). \end{aligned}$$

Substituting this estimate into (105) gives

$$\| \Pi_M \Phi(\mu) - \Phi(\mu) \|_{L^2} \le \frac{R}{\sqrt{M}} + 2\sigma \exp\left(-\frac{R^2}{4\sigma^2}\right).$$

Choosing $R = 2\sigma\sqrt{\log(\mathrm{e}M)}$ yields

$$\begin{aligned} \| \Pi_M \Phi(\mu) - \Phi(\mu) \|_{L^2} &\le 2\sigma \sqrt{\frac{\log(\mathrm{e}M)}{M}} + \frac{2\sigma}{\mathrm{e}M} \\ &\le 4\sigma \sqrt{\frac{\log(\mathrm{e}M)}{M}}. \end{aligned}$$

Taking the supremum over $\mu \in \mathcal{K}$ proves the finite-M bound

$$\rho_M(\mathcal{K}) \le 4\sigma \sqrt{\frac{\log(\mathrm{e}M)}{M}}. \tag{107}$$

Hence it is enough to take

$$M_2^{\mathrm{sg}}(\varepsilon) = O\left(1 + \frac{\sigma^2}{\varepsilon^2} \log\left(\mathrm{e} + \frac{\sigma}{\varepsilon}\right)\right). \tag{108}$$

Finally, suppose that the quantile functions are uniformly L-Lipschitz, that is,

$$|\Phi(\mu)(u) - \Phi(\mu)(v)| \le L|u-v|, \qquad u, v \in (0,1), \quad \mu \in \mathcal{K}. \tag{109}$$

Fix $\mu \in \mathcal{K}$, write $q = \Phi(\mu)$ and $\widehat{q}_M = \Phi(\widehat{\mu}_M)$, and extend q to $[0,1]$ by

$$q(0) \;:=\; \lim_{v \downarrow 0} q(v), \qquad q(1) \;:=\; \lim_{v \uparrow 1} q(v).$$

These limits are finite, the extension remains L-Lipschitz, and $\mu([q(0), q(1)]) = 1$. Hence every observation used to construct $\widehat{\mu}_M$ belongs to $[q(0), q(1)]$ almost surely. Let $F = F_\mu$, $\widehat{F} = F_{\widehat{\mu}_M}$, and $\kappa = \kappa_M(\mu)$. Fix $u \in (0,1)$ and set

$$u_- \;:=\; (u - \kappa) \vee 0, \qquad u_+ \;:=\; (u + \kappa) \wedge 1.$$

For every $v \in (0,1)$, the definition of the generalized inverse and the right-continuity of F imply

$$x < q(v) \quad \Longrightarrow \quad F(x) < v, \qquad F(q(v)) \geq v. \tag{110}$$

If $u_- > 0$, then, for every $x < q(u_-)$,

$$\widehat{F}(x) \leq F(x) + \kappa < u_- + \kappa = u.$$

It follows that $\widehat{q}_M(u) \geq q(u_-)$. If $u_- = 0$, the same conclusion follows because every observation is at least $q(0)$. Similarly, if $u_+ < 1$, then

$$\widehat{F}(q(u_+)) \geq F(q(u_+)) - \kappa \geq u_+ - \kappa = u,$$

and hence $\widehat{q}_M(u) \leq q(u_+)$. If $u_+ = 1$, the same conclusion follows because every observation is at most $q(1)$. Thus, almost surely, for every $u \in (0,1)$,

$$q\left((u - \kappa_M(\mu)) \vee 0\right) \leq \widehat{q}_M(u) \leq q\left((u + \kappa_M(\mu)) \wedge 1\right). \tag{111}$$

This argument does not require F to be continuous and therefore also covers measures with atoms. Since both clipped arguments in (111) differ from u by at most $\kappa_M(\mu)$, the Lipschitz property gives

$$W_2(\widehat{\mu}_M, \mu) = \| \widehat{q}_M - q \|_{L^2} \leq \| \widehat{q}_M - q \|_{L^\infty(0,1)} \leq L\kappa_M(\mu). \tag{112}$$

Consequently, $\{W_2(\widehat{\mu}_M, \mu) > \alpha\} \subseteq \{\kappa_M(\mu) > \alpha/L\}$. Combining this inclusion with (93), then using (92), gives

$$\sup_{\boldsymbol{\mu} \in \mathcal{K}^T} \mathbb{P}_{\boldsymbol{\mu}} \left(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^C \right) \leq 2T \exp\left(-\frac{2M\alpha^2}{L^2} \right). \tag{113}$$

It follows that one may take

$$M_1^{\mathrm{Lip}}(\alpha, \delta) \;:=\; \left\lceil \frac{L^2}{2\alpha^2} \log\left(\frac{2T}{\delta} \right) \right\rceil = O\left(1 + \frac{L^2}{\alpha^2} \log\left(\frac{2T}{\delta} \right) \right). \tag{114}$$

It remains to control the projection error. On each cell $I_{i,M}$, whose length is $1/M$, the cellwise variance identity gives

$$\begin{aligned} \int_{I_{i,M}} |q(u) - (\Pi_M q)(u)|^2 \, \mathrm{d}u &= \frac{M}{2} \int_{I_{i,M}} \int_{I_{i,M}} |q(u) - q(v)|^2 \, \mathrm{d}u \, \mathrm{d}v \\ &\leq \frac{L^2 M}{2} \int_{I_{i,M}} \int_{I_{i,M}} |u - v|^2 \, \mathrm{d}u \, \mathrm{d}v \\ &= \frac{L^2}{12M^3}, \end{aligned}$$

where

$$\int_{I_{i,M}} \int_{I_{i,M}} |u - v|^2 \, \mathrm{d}u \, \mathrm{d}v = \frac{1}{6M^4}.$$

Summing over $i \in [M]$ gives

$$\| q - \Pi_M q \|_{L^2}^2 \leq \frac{L^2}{12M^2}.$$

Taking the square root and then the supremum over $\mu \in \mathcal{K}$ yields

$$\rho_M(\mathcal{K}) \leq \frac{L}{\sqrt{12}\, M}. \tag{115}$$

Therefore, it is enough to take

$$M_2^{\mathrm{Lip}}(\varepsilon) \;:=\; \left\lceil \frac{\sqrt{3}\, L}{2\varepsilon} \right\rceil = O\left(1 + \frac{L}{\varepsilon} \right). \tag{116}$$

The choices in (96), (99), (104), (108), (114), and (116) prove all six claims. □

Table 1: Characteristic functions for representative elliptical distributions. We use the convention $\varphi_\mu(t) = \exp(im^\top t)\,\psi(t^\top \Sigma t)$, with $q := t^\top \Sigma t$. Here, Γ denotes the Gamma function, J_ν the Bessel function of the first kind of order ν, and K_ν the modified Bessel function of the second kind of order ν.

Distribution	Generator $\psi(q)$	Characteristic function $\varphi_\mu(t)$	In $\mathcal{P}_2(\mathbb{R}^d)$?
Gaussian	$e^{-q/2}$	$\exp\left(im^\top t - \frac{1}{2}t^\top \Sigma t\right)$	Yes
Student t_ν	$\frac{2^{1-\nu/2}}{\Gamma(\nu/2)}(\sqrt{\nu q})^{\nu/2} K_{\nu/2}(\sqrt{\nu q})$	$e^{im^\top t}\frac{2^{1-\nu/2}}{\Gamma(\nu/2)}\left(\sqrt{\nu\, t^\top \Sigma t}\right)^{\nu/2} K_{\nu/2}\left(\sqrt{\nu\, t^\top \Sigma t}\right)$	iff $\nu > 2$
Cauchy	$e^{-\sqrt{q}}$	$\exp\left(im^\top t - \sqrt{t^\top \Sigma t}\right)$	No
Symmetric α-stable	$e^{-q^{\alpha/2}}, \quad 0 < \alpha \le 2$	$\exp\left(im^\top t - (t^\top \Sigma t)^{\alpha/2}\right)$	Only $\alpha = 2$
Logistic	$\frac{\pi\sqrt{q}}{\sinh(\pi\sqrt{q})}$	$e^{im^\top t}\frac{\pi\sqrt{t^\top \Sigma t}}{\sinh\left(\pi\sqrt{t^\top \Sigma t}\right)}$	Yes
Uniform on solid ellipsoid	$2^{d/2}\Gamma\left(\frac{d}{2}+1\right)\frac{J_{d/2}(\sqrt{q})}{q^{d/4}}$	$e^{im^\top t} 2^{d/2}\Gamma\left(\frac{d}{2}+1\right)\frac{J_{d/2}\left(\sqrt{t^\top \Sigma t}\right)}{(t^\top \Sigma t)^{d/4}}$	Yes

E Omitted results in Section 4

E.1 Verification of Steps 1 and 2 from Section 2.1 in the construction of Section 4.1

Let $\Phi : \mathcal{P}_2(\mathbb{R}) \to L^2$ denote the one-dimensional quantile representation. We choose $\mathcal{Z} := L^{2^{L_d}}$, equipped with $\| h \|_{\mathcal{Z}}^2 := L_d^{-1} \sum_{\ell=1}^{L_d} \| h_\ell \|_{L^2}^2$, and define $\Phi_{V_d}(\mu) := [\Phi(\mu^{s_1}), \dots, \Phi(\mu^{s_{L_d}})]^\top$. Defining $\mathrm{SW}_{2,V_d}^2(\mu,\nu) := L_d^{-1} \sum_{\ell=1}^{L_d} W_2^2(\mu^{s_\ell}, \nu^{s_\ell})$, the one-dimensional quantile isometry gives

$$\| \Phi_{V_d}(\mu) - \Phi_{V_d}(\nu) \|_{\mathcal{Z}} = \mathrm{SW}_{2,V_d}(\mu,\nu) \le W_2(\mu,\nu). \tag{117}$$

Thus, Φ_{V_d} is continuous. Injectivity of $\Phi_{V_d}|_{\mathcal{K}_{\mathrm{ell}}}$ follows directly from injectivity of Φ in one dimension and Proposition 4.1. Therefore, $\Phi_{V_d}|_{\mathcal{K}_{\mathrm{ell}}}$ is a homeomorphism onto $\mathcal{Q}_{\mathrm{ell}} := \Phi_{V_d}(\mathcal{K}_{\mathrm{ell}})$ and satisfies the requirements of Step 1 in Section 2.1.

Next, we let $r_M := ML_d \to \infty$ and define the collection of finite-dimensional subspaces as carthesian products of $\mathcal{S}_M$ from Section 3, namely $\mathcal{Z}_{r_M} := \mathcal{S}_M^{L_d} \subseteq \mathcal{Z}$. Then, the projection operator $\Pi_M^{L_d} : L^{2^{L_d}} \to \mathcal{S}_M^{L_d}$ is a component-wise application of Π_M in (10) and satisfies condition (3) via Proposition 3.1. Likewise, $C_M^{L_d} : \mathcal{S}_M^{L_d} \to \mathbb{R}^{ML_d}$ is a component-wise application of C_M from Section 3 further normalized by $\sqrt{L_d}$. The population embedding is then $\overline{E}_M := C_M^{L_d} \circ \Pi_M^{L_d} \circ \Phi_{V_d}$.

Definition E.1 (Symmetric-matrix uniqueness set; Definition 2 of Fraiman et al. (2023))**.** *A set $\mathcal{S} \subseteq \mathbb{R}^d$ is called a* symmetric-matrix uniqueness set*, or* sm-uniqueness set*, if for every real symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$,*

$$\boldsymbol{s}^\top \boldsymbol{A}\boldsymbol{s} = 0,\ \forall \boldsymbol{s} \in \mathcal{S} \quad \Longrightarrow \quad \boldsymbol{A} = \boldsymbol{0}_{d\times d}.$$

Corollary E.1 (A minimal symmetric-matrix uniqueness set; Corollary 4 of Fraiman et al. (2023))**.** *Let $\mathcal{S}_d := \{\boldsymbol{e}_i : i \in [d]\} \cup \{\boldsymbol{e}_j + \boldsymbol{e}_k : 1 \le j < k \le d\}$. Then $\mathcal{S}_d$ is an sm-uniqueness set in $\mathbb{R}^d$ with cardinality*

$$|\,\mathcal{S}_d\,| = d + \binom{d}{2} = \frac{d(d+1)}{2}.$$

E.2 Empirical–population token bound (Proposition E.1)

The following useful properties of the embedding map immediately emerge (proof in Appendix E.2).

Proposition E.1. *The empirical quantile embedding E_M satisfies*

(a) For any $\widehat{\mu}_M, \widehat{\nu}_M \in \mathcal{E}_M(\mathbb{R}^d)$,

$$\| E_M(\widehat{\mu}_M) - E_M(\widehat{\nu}_M) \|_2^2 = \mathrm{SW}_{2,\mathcal{V}_d}^2(\widehat{\mu}_M, \widehat{\nu}_M). \tag{118}$$

(b) For every $\mu \in \mathcal{K}_{\mathrm{ell}}$ and every $\widehat{\mu}_M \in \mathcal{E}_M(\mathbb{R}^d)$,

$$\left\| E_M(\widehat{\mu}_M) - \overline{E}_M(\mu) \right\|_2 \le \mathrm{SW}_{2,\mathcal{V}_d}(\widehat{\mu}_M, \mu) \le W_2(\widehat{\mu}_M, \mu). \tag{119}$$

Proof. Point a) follows by applying the one-dimensional quantile identity componentwise across the slicing directions and then using the isometry of $C_M^{L_d}$. Indeed, for $\widehat{\mu}_M, \widehat{\nu}_M \in \mathcal{E}_M(\mathbb{R}^d)$,

$$\begin{aligned}
\| E_M(\widehat{\mu}_M) - E_M(\widehat{\nu}_M) \|_2^2 &= \| \Phi_{\mathcal{V}_d}(\widehat{\mu}_M) - \Phi_{\mathcal{V}_d}(\widehat{\nu}_M) \|_{L^{2^{L_d}}}^2 \\
&= \frac{1}{L_d} \sum_{\ell=1}^{L_d} \| \Phi(\widehat{\mu}_M^{\boldsymbol{s}_\ell}) - \Phi(\widehat{\nu}_M^{\boldsymbol{s}_\ell}) \|_{L^2}^2 \\
&= \frac{1}{L_d} \sum_{\ell=1}^{L_d} W_2^2(\widehat{\mu}_M^{\boldsymbol{s}_\ell}, \widehat{\nu}_M^{\boldsymbol{s}_\ell}) \\
&= \mathrm{SW}_{2,\mathcal{V}_d}^2(\widehat{\mu}_M, \widehat{\nu}_M),
\end{aligned} \tag{120}$$

where the first equality uses C_M's isometry from Proposition D.1, the second follows from the definition of the norm on $L^{2^{L_d}}$, and the third uses the one-dimensional quantile formula in Proposition B.6.

For point b), let $\widehat{\mu}_M = \frac{1}{M}\sum_{i=1}^M \delta_{\boldsymbol{x}_i} \in \mathcal{E}_M(\mathbb{R}^d)$. For every $\ell \in [L_d]$, the projected empirical measure $\widehat{\mu}_M^{\boldsymbol{s}_\ell}$ is an equally weighted empirical measure on $\mathbb{R}$ with M atoms. Hence its quantile function is constant on each interval $I_{i,M}$ and therefore belongs to $\mathcal{S}_M$. Consequently,

$$\Pi_M^{L_d} \Phi_{\mathcal{V}_d}(\widehat{\mu}_M) = \Phi_{\mathcal{V}_d}(\widehat{\mu}_M). \tag{121}$$

Using the definitions of E_M and $\overline{E}_M$, (121), and the linearity of $\Pi_M^{L_d}$,

$$\begin{aligned}
E_M(\widehat{\mu}_M) - \overline{E}_M(\mu) &= C_M^{L_d} \Phi_{\mathcal{V}_d}(\widehat{\mu}_M) - C_M^{L_d} \Pi_M^{L_d} \Phi_{\mathcal{V}_d}(\mu) \\
&= C_M^{L_d} \Pi_M^{L_d} \left(\Phi_{\mathcal{V}_d}(\widehat{\mu}_M) - \Phi_{\mathcal{V}_d}(\mu) \right).
\end{aligned} \tag{122}$$

Since $C_M^{L_d}$ is an isometry and $\Pi_M^{L_d}$ is contractive,

$$\begin{aligned}
\left\| E_M(\widehat{\mu}_M) - \overline{E}_M(\mu) \right\|_2 &= \left\| \Pi_M^{L_d} \left(\Phi_{\mathcal{V}_d}(\widehat{\mu}_M) - \Phi_{\mathcal{V}_d}(\mu) \right) \right\|_{L^{2^{L_d}}} \\
&\le \| \Phi_{\mathcal{V}_d}(\widehat{\mu}_M) - \Phi_{\mathcal{V}_d}(\mu) \|_{L^{2^{L_d}}} \\
&= \mathrm{SW}_{2,\mathcal{V}_d}(\widehat{\mu}_M, \mu),
\end{aligned} \tag{123}$$

where the last equality follows from the elliptical finite sliced quantile isometry. Finally, each slicing projection map $P_{\boldsymbol{s}_\ell}$ is 1-Lipschitz, since

$$| P_{\boldsymbol{s}_\ell}(\boldsymbol{x}) - P_{\boldsymbol{s}_\ell}(\boldsymbol{y}) | = | \langle \boldsymbol{s}_\ell, \boldsymbol{x} - \boldsymbol{y} \rangle | \le \| \boldsymbol{s}_\ell \|_2 \| \boldsymbol{x} - \boldsymbol{y} \|_2 = \| \boldsymbol{x} - \boldsymbol{y} \|_2. \tag{124}$$

Therefore, we have

$$\begin{aligned}
W_2^2(\widehat{\mu}_M^{\boldsymbol{s}_\ell}, \mu^{\boldsymbol{s}_\ell}) &= \inf_{\gamma \in \Pi(\widehat{\mu}_M^{\boldsymbol{s}_\ell}, \mu^{\boldsymbol{s}_\ell})} \int_{\mathbb{R}\times\mathbb{R}} | x - y |^2 \, d\gamma(x,y) \\
&\le \inf_{\pi \in \Pi(\widehat{\mu}_M, \mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} | P_{\boldsymbol{s}_\ell}(\boldsymbol{x}) - P_{\boldsymbol{s}_\ell}(\boldsymbol{y}) |^2 \, d\pi(\boldsymbol{x}, \boldsymbol{y}) \\
&\le \inf_{\pi \in \Pi(\widehat{\mu}_M, \mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \| \boldsymbol{x} - \boldsymbol{y} \|_2^2 \, d\pi(\boldsymbol{x}, \boldsymbol{y}) \\
&= W_2^2(\widehat{\mu}_M, \mu),
\end{aligned} \tag{125}$$

Averaging over the directions gives

$$\mathrm{SW}_{2,\mathcal{V}_d}^2(\widehat{\mu}_M, \mu) = \frac{1}{L_d} \sum_{\ell=1}^{L_d} W_2^2(\widehat{\mu}_M^{\boldsymbol{s}_\ell}, \mu^{\boldsymbol{s}_\ell}) \le W_2^2(\widehat{\mu}_M, \mu).$$

Taking square roots proves the second inequality in (119). □

E.3 Construction of the surrogate in (Section 4)

We reuse the surrogate construction of Section 3. We first have to define the appropriate product metric, and equip $\left(L^{2^{L_d}}\right)^T$ with

$$\mathrm{dist}_{T,\mathcal{V}_d}(\boldsymbol{p},\boldsymbol{q}) \;:=\; \max_{t\in[T]} \|\, \boldsymbol{p}_t - \boldsymbol{q}_t \,\|_{L^{2^{L_d}}}\,. \tag{126}$$

We apply the α-net-based construction in (40)–(42) to the compact $\mathcal{Q}^T_{\mathrm{ell}}$, relative to (126). Denote the resulting map by $\widetilde{\varphi}_{\mathrm{ell},\alpha} : (\mathcal{Q}^T_{\mathrm{ell}})_\alpha \longrightarrow \left(L^{2^{L_d}}\right)^T$, and note that the validity statements of Proposition D.3 apply. We further obtain an impoverished analog of Proposition (D.4) (proof in Appendix E.4).

Proposition E.2 (Surrogate approximation on $\mathcal{K}_{\mathrm{ell}}$)**.** *Let $\boldsymbol{q} \in \mathcal{Q}^T_{\mathrm{ell}}$ and $\boldsymbol{h} \in (\mathcal{Q}^T_{\mathrm{ell}})_\alpha$ satisfy $\mathrm{dist}_{T,\mathcal{V}_d}(\boldsymbol{h},\boldsymbol{q}) \le \alpha$. Then,*

$$\mathrm{dist}_{T,\mathcal{V}_d}\left(\widetilde{\varphi}_{\mathrm{ell},\alpha}(\boldsymbol{h}), \widetilde{\varphi}_{\mathrm{ell}}(\boldsymbol{q})\right) \le \omega_{\widetilde{\varphi}_{\mathrm{ell}}}(4\alpha).$$

Define the compact finite-dimensional domain $\mathcal{D}^{\mathrm{ell}}_{M,\alpha} := (C_M^{L_d})^T \left((\mathcal{S}_M^{L_d})^T \cap (\mathcal{Q}^T_{\mathrm{ell}})_\alpha\right) \subseteq (\mathbb{R}^{ML_d})^T$ and the finite-dimensional, continuous surrogate $G^{\mathrm{ell}}_{M,\alpha} : \mathcal{D}^{\mathrm{ell}}_{M,\alpha} \to (\mathbb{R}^{ML_d})^T$ given by

$$G^{\mathrm{ell}}_{M,\alpha} \;:=\; (C_M^{L_d})^T \circ (\Pi_M^{L_d})^T \circ \widetilde{\varphi}_{\mathrm{ell},\alpha} \circ ((C_M^{L_d})^{-1})^T. \tag{127}$$

For $\boldsymbol{\mu} = (\mu_1,\dots,\mu_T) \in \mathcal{K}^T_{\mathrm{ell}}$, let $\widehat{\boldsymbol{\mu}}_M = (\widehat{\mu}_{1,M},\dots,\widehat{\mu}_{T,M})$ be the empirical sequence and consider the same event $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$ as in (14). On this event,

$$\mathrm{dist}_{T,\mathcal{V}_d}\left(\Phi^T_{\mathcal{V}_d}(\widehat{\boldsymbol{\mu}}_M), \Phi^T_{\mathcal{V}_d}(\boldsymbol{\mu})\right) = \max_{t\in[T]} \mathrm{SW}_{2,\mathcal{V}_d}(\widehat{\mu}_{t,M},\mu_t) \le \alpha,$$

and hence $E^T_M(\widehat{\boldsymbol{\mu}}_M) \in \mathcal{D}^{\mathrm{ell}}_{M,\alpha}$.

For fixed M and α, the map $G^{\mathrm{ell}}_{M,\alpha}$ is continuous on the compact finite-dimensional domain $\mathcal{D}^{\mathrm{ell}}_{M,\alpha}$. Applying the same finite-dimensional transformer universality result used in Section 3, for every $\eta > 0$ there exists θ such that

$$\sup_{\boldsymbol{z}\in\mathcal{D}^{\mathrm{ell}}_{M,\alpha}} \max_{s\in[T]} \left\| \mathcal{T}_\theta(\boldsymbol{z})_s - G^{\mathrm{ell}}_{M,\alpha}(\boldsymbol{z})_s \right\|_2 \le \eta. \tag{128}$$

E.4 Surrogate approximation on the compact elliptical class (Proposition E.2)

Proposition E.2 (Surrogate approximation on $\mathcal{K}_{\mathrm{ell}}$)**.** *Let $\boldsymbol{q} \in \mathcal{Q}^T_{\mathrm{ell}}$ and $\boldsymbol{h} \in (\mathcal{Q}^T_{\mathrm{ell}})_\alpha$ satisfy $\mathrm{dist}_{T,\mathcal{V}_d}(\boldsymbol{h},\boldsymbol{q}) \le \alpha$. Then,*

$$\mathrm{dist}_{T,\mathcal{V}_d}\left(\widetilde{\varphi}_{\mathrm{ell},\alpha}(\boldsymbol{h}), \widetilde{\varphi}_{\mathrm{ell}}(\boldsymbol{q})\right) \le \omega_{\widetilde{\varphi}_{\mathrm{ell}}}(4\alpha).$$

Proof. The surrogate $\widetilde{\varphi}_{\mathrm{ell},\alpha}$ is obtained from the same α-net construction as $\widetilde{\varphi}_\alpha$ in Proposition D.4, with $\mathcal{Q}^T$ replaced by $\mathcal{Q}^T_{\mathrm{ell}}$ and dist_T replaced by $\mathrm{dist}_{T,\mathcal{V}_d}$. Since $\mathcal{Q}^T_{\mathrm{ell}}$ is compact, the same construction applies.

Hence, if $\boldsymbol{q} \in \mathcal{Q}^T_{\mathrm{ell}}$ and $\boldsymbol{h} \in (\mathcal{Q}^T_{\mathrm{ell}})_\alpha$ satisfy $\mathrm{dist}_{T,\mathcal{V}_d}(\boldsymbol{h},\boldsymbol{q}) \le \alpha$, Point (a) of Proposition D.4 gives

$$\mathrm{dist}_{T,\mathcal{V}_d}\left(\widetilde{\varphi}_{\mathrm{ell},\alpha}(\boldsymbol{h}), \widetilde{\varphi}_{\mathrm{ell}}(\boldsymbol{q})\right) \le \omega_{\widetilde{\varphi}_{\mathrm{ell}}}(4\alpha).$$

□

Proposition E.3 (Uniform continuity of the finite sliced inverse map)**.** *Let $\mathcal{K}_{\mathrm{ell}} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ be a compact on elliptical measures. Then the identity map*

$$\mathrm{id} : (\mathcal{K}_{\mathrm{ell}}, \mathrm{SW}_{2,\mathcal{V}_d}) \longrightarrow (\mathcal{K}_{\mathrm{ell}}, W_2)$$

is uniformly continuous.

Proof. It holds that, for every $\mu, \nu \in \mathcal{K}_{\text{ell}}$,

$$\| \Phi_{\mathcal{V}_d}(\mu) - \Phi_{\mathcal{V}_d}(\nu) \|_{L^{2^{L_d}}} = \mathrm{SW}_{2,\mathcal{V}_d}(\mu, \nu). \tag{129}$$

Moreover, as established in Step 1, $\Phi_{\mathcal{V}_d} : \mathcal{K}_{\text{ell}} \longrightarrow \mathcal{Q}_{\text{ell}}$ is a homeomorphism onto its image. In particular, $\Phi_{\mathcal{V}_d}^{-1} : \mathcal{Q}_{\text{ell}} \to \mathcal{K}_{\text{ell}}$ is continuous when $\mathcal{Q}_{\text{ell}}$ is equipped with the $L^{2^{L_d}}$ norm and $\mathcal{K}_{\text{ell}}$ with W_2.

Now let $\{\mu_n\}_{n\geq 1} \subseteq \mathcal{K}_{\text{ell}}$ and $\mu \in \mathcal{K}_{\text{ell}}$ satisfy $\mathrm{SW}_{2,\mathcal{V}_d}(\mu_n, \mu) \longrightarrow 0$. By (129),

$$\| \Phi_{\mathcal{V}_d}(\mu_n) - \Phi_{\mathcal{V}_d}(\mu) \|_{L^{2^{L_d}}} \longrightarrow 0.$$

Continuity of $\Phi_{\mathcal{V}_d}^{-1}$ therefore gives

$$W_2(\mu_n, \mu) = W_2\big(\Phi_{\mathcal{V}_d}^{-1}\big(\Phi_{\mathcal{V}_d}(\mu_n)\big), \Phi_{\mathcal{V}_d}^{-1}\big(\Phi_{\mathcal{V}_d}(\mu)\big)\big) \longrightarrow 0.$$

Hence $\mathrm{id} : (\mathcal{K}_{\text{ell}}, \mathrm{SW}_{2,\mathcal{V}_d}) \longrightarrow (\mathcal{K}_{\text{ell}}, W_2)$ is continuous.

Finally, since $\mathcal{K}_{\text{ell}}$ is compact in W_2 and $\mathrm{SW}_{2,\mathcal{V}_d} \leq W_2$ by (129), the identity map

$$\mathrm{id} : (\mathcal{K}_{\text{ell}}, W_2) \longrightarrow (\mathcal{K}_{\text{ell}}, \mathrm{SW}_{2,\mathcal{V}_d})$$

is continuous. Uniform continuity now follows from Proposition B.1 applied to id. □

E.5 Uniform empirical quantile-token universality in probability on compact classes in $\mathbb{R}^d$ (Theorem 4.1)

Theorem 4.1 (High-probability approximation universality in $\mathcal{P}(\mathbb{R}^d)$)**.** *Let $d \geq 2$, let $\mathcal{K}_{\text{ell}} \subset \mathcal{P}_2(\mathbb{R}^d)$ be a nonempty W_2-compact family of elliptical laws with positive-definite covariances and consider $\varphi \in \mathcal{C}_{W_2}\left((\mathcal{K}_{\text{ell}})^T, (\mathcal{K}_{\text{ell}})^T\right)$. For every $\varepsilon > 0$ and every $\delta > 0$, there exists $M_0 \in \mathbb{N}$ such that, for every $M \geq M_0$, there exists θ and a transformer-based map (5) ensuring*

$$\sup_{\boldsymbol{\mu} \in (\mathcal{K}_{\text{ell}})^T} \mathbb{P}_{\boldsymbol{\mu}} \left(\max_{s\in[T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) > \varepsilon \right) < \delta.$$

Proof. The proof follows the same conceptual argument as that of Theorem 3.1. Fix $\varepsilon, \delta > 0$. The route to proving the result is to find $\alpha > 0$ and $M_0 \in \mathbb{N}$ such that, for every $M \geq M_0$, one can choose transformer parameters θ for which

$$\left\{ \max_{s\in[T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) > \varepsilon \right\} \subseteq \mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^c \tag{130}$$

for every $\boldsymbol{\mu} \in \mathcal{K}_{\text{ell}}^T$.

Indeed, once (130) is established, we have that

$$\sup_{\boldsymbol{\mu} \in \mathcal{K}_{\text{ell}}^T} \mathbb{P}_{\boldsymbol{\mu}} \left(\max_{s\in[T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) > \varepsilon \right) \leq \sup_{\boldsymbol{\mu} \in \mathcal{K}_{\text{ell}}^T} \mathbb{P}_{\boldsymbol{\mu}} \left(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^c \right), \tag{131}$$

and the right-hand side converges to zero as $M \to \infty$ via Proposition D.5.

We will now prove (130). Fix some $\alpha > 0$, $\eta > 0$, and $M \in \mathbb{N}$, and choose θ so that (128) holds. Fix also an arbitrary $\boldsymbol{\mu} \in \mathcal{K}_{\text{ell}}^T$ and let $\boldsymbol{\nu} := \varphi(\boldsymbol{\mu})$.

We first control the discrepancy between the transformer output token and the population token of $\boldsymbol{\nu}$ on the event $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$. Recall that

$$\alpha \geq \max_{t\in[T]} W_2(\widehat{\mu}_{t,M}, \mu_t) \geq \max_{t\in[T]} \mathrm{SW}_{2,\mathcal{V}_d}\left(\widehat{\mu}_{t,M}, \mu_t\right) = \mathrm{dist}_{T,\mathcal{V}_d}\left(\Phi_{\mathcal{V}_d}^T(\widehat{\boldsymbol{\mu}}_M), \Phi_{\mathcal{V}_d}^T(\boldsymbol{\mu})\right),$$

where the first inequality follows from the definition of $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$, the second from Proposition E.1 (b), and the equality from the definition of $\mathrm{SW}_{2,\mathcal{V}_d}$ and $\mathrm{dist}_{T,\mathcal{V}_d}$ in (126), together with the one-dimensional quantile identity. This implies that $E_M^T(\widehat{\boldsymbol{\mu}}_M) \in \mathcal{D}_{M,\alpha}^{\text{ell}}$, and hence the input of $\mathcal{T}_\theta$ is well defined.

Let $\boldsymbol{z}_M := E_M^T(\widehat{\boldsymbol{\mu}}_M)$. For every $s \in [T]$, the triangle inequality gives

$$\left\| \mathcal{T}_\theta(\boldsymbol{z}_M)_s - \overline{E}_M(\boldsymbol{\nu}_s) \right\|_2 \le \left\| \mathcal{T}_\theta(\boldsymbol{z}_M)_s - G^{\mathrm{ell}}_{M,\alpha}(\boldsymbol{z}_M)_s \right\|_2 + \left\| G^{\mathrm{ell}}_{M,\alpha}(\boldsymbol{z}_M)_s - \overline{E}_M(\boldsymbol{\nu}_s) \right\|_2 . \tag{132}$$

The first term is bounded by (128). For the second term, using the definitions of $G^{\mathrm{ell}}_{M,\alpha}$ and $\overline{E}_M$, we obtain

$$\begin{aligned}
&\left\| G^{\mathrm{ell}}_{M,\alpha}(\boldsymbol{z}_M)_s - \overline{E}_M(\boldsymbol{\nu}_s) \right\|_2 \\
&\quad = \left\| C_M^{L_d} \left(\left((\Pi_M^{L_d})^T \circ \widetilde{\varphi}_{\mathrm{ell},\alpha} \circ ((C_M^{L_d})^{-1})^T \right) (\boldsymbol{z}_M)_s \right) - (C_M^{L_d} \circ \Pi_M^{L_d} \circ \Phi_{\mathcal{V}_d})(\boldsymbol{\nu}_s) \right\|_2 \\
&\quad = \left\| \Pi_M^{L_d} \left(\left(\widetilde{\varphi}_{\mathrm{ell},\alpha} \circ ((C_M^{L_d})^{-1})^T \right) (\boldsymbol{z}_M)_s - \Phi_{\mathcal{V}_d}(\boldsymbol{\nu}_s) \right) \right\|_{L^{2^{L_d}}} \\
&\quad \le \left\| \left(\widetilde{\varphi}_{\mathrm{ell},\alpha} \circ ((C_M^{L_d})^{-1})^T \right) (\boldsymbol{z}_M)_s - \Phi_{\mathcal{V}_d}(\boldsymbol{\nu}_s) \right\|_{L^{2^{L_d}}} ,
\end{aligned} \tag{133}$$

where the second equality follows from the isometry of $C_M^{L_d}$ and the linearity of $\Pi_M^{L_d}$, while the final inequality follows from the contractiveness of $\Pi_M^{L_d}$.

Now recall that $\boldsymbol{z}_M = E_M^T(\widehat{\boldsymbol{\mu}}_M) = \left((C_M^{L_d})^T \circ \Phi_{\mathcal{V}_d}^T \right)(\widehat{\boldsymbol{\mu}}_M)$, and that $\Phi_{\mathcal{V}_d}(\boldsymbol{\nu}_s) = \widetilde{\varphi}_{\mathrm{ell}} \left(\Phi_{\mathcal{V}_d}^T(\boldsymbol{\mu}) \right)_s$. Therefore, (133) continues as

$$\begin{aligned}
\left\| G^{\mathrm{ell}}_{M,\alpha}(\boldsymbol{z}_M)_s - \overline{E}_M(\boldsymbol{\nu}_s) \right\|_2 &\le \left\| \widetilde{\varphi}_{\mathrm{ell},\alpha} \left(\Phi_{\mathcal{V}_d}^T(\widehat{\boldsymbol{\mu}}_M) \right)_s - \widetilde{\varphi}_{\mathrm{ell}} \left(\Phi_{\mathcal{V}_d}^T(\boldsymbol{\mu}) \right)_s \right\|_{L^{2^{L_d}}} \\
&\le \mathrm{dist}_{T,\mathcal{V}_d} \left(\widetilde{\varphi}_{\mathrm{ell},\alpha} \left(\Phi_{\mathcal{V}_d}^T(\widehat{\boldsymbol{\mu}}_M) \right), \widetilde{\varphi}_{\mathrm{ell}} \left(\Phi_{\mathcal{V}_d}^T(\boldsymbol{\mu}) \right) \right) \\
&\le \omega_{\widetilde{\varphi}_{\mathrm{ell}}}(4\alpha),
\end{aligned} \tag{134}$$

where the second inequality follows from the definition of $\mathrm{dist}_{T,\mathcal{V}_d}$ and the final inequality follows from Proposition E.2.

Combining (132), (128), and (134), we obtain

$$\left\| (\mathcal{T}_\theta \circ E_M^T)(\widehat{\boldsymbol{\mu}}_M)_s - \overline{E}_M(\boldsymbol{\nu}_s) \right\|_2 \le \eta + \omega_{\widetilde{\varphi}_{\mathrm{ell}}}(4\alpha) \tag{135}$$

for every $s \in [T]$ on the event $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$.

We now use the uniform stability of the decoder. For $r > 0$, define

$$\Delta_M(r) := \sup_{\substack{\nu \in \mathcal{K}_{\mathrm{ell}},\ \boldsymbol{z} \in \mathbb{R}^{ML_d} \\ \left\| \boldsymbol{z} - \overline{E}_M(\nu) \right\|_2 \le r}} W_2(D_M(\boldsymbol{z}), \nu). \tag{136}$$

By Proposition 4.2,

$$\lim_{r \downarrow 0} \limsup_{M \to \infty} \Delta_M(r) = 0. \tag{137}$$

Hence, there exists $r_\varepsilon > 0$ such that

$$\limsup_{M \to \infty} \Delta_M(r_\varepsilon) < \frac{\varepsilon}{2}.$$

By the definition of $\limsup$, there exists $M_1 \in \mathbb{N}$ such that, for every $M \ge M_1$,

$$\Delta_M(r_\varepsilon) < \varepsilon. \tag{138}$$

We now choose α and η so that the transformer output falls inside this decoder-stability neighborhood. Since $\lim_{r \downarrow 0} \omega_{\widetilde{\varphi}_{\mathrm{ell}}}(r) = 0$, choose $\alpha > 0$ such that

$$\omega_{\widetilde{\varphi}_{\mathrm{ell}}}(4\alpha) < \frac{r_\varepsilon}{2}. \tag{139}$$

Next, choose $\eta > 0$ such that

$$\eta < \frac{r_\varepsilon}{2}. \tag{140}$$

For every M, choose the transformer parameters θ so that (128) holds with this fixed η.

Then, on $\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})$, (135), (139), and (140) imply that, for every $s \in [T]$,

$$\left\| (\mathcal{T}_\theta \circ E_M^T)(\widehat{\boldsymbol{\mu}}_M)_s - \overline{E}_M(\boldsymbol{\nu}_s) \right\|_2 \le \eta + \omega_{\widetilde{\varphi}_{\mathrm{ell}}}(4\alpha) < r_\varepsilon. \tag{141}$$

Since $\boldsymbol{\nu} = \varphi(\boldsymbol{\mu}) \in \mathcal{K}_{\mathrm{ell}}^T$, we have $\boldsymbol{\nu}_s \in \mathcal{K}_{\mathrm{ell}}$ for every $s \in [T]$. Therefore, for every $M \ge M_1$, (138) and (141) yield

$$W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) = W_2\left(D_M\left((\mathcal{T}_\theta \circ E_M^T)(\widehat{\boldsymbol{\mu}}_M)_s\right), \boldsymbol{\nu}_s\right) \le \Delta_M(r_\varepsilon) < \varepsilon \tag{142}$$

for every $s \in [T]$. Consequently,

$$\mathcal{A}_{M,\alpha}(\boldsymbol{\mu}) \quad \Longrightarrow \quad \max_{s\in[T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) < \varepsilon.$$

Equivalently,

$$\left\{ \max_{s\in[T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) > \varepsilon \right\} \subseteq \mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^c \tag{143}$$

for every $M \ge M_1$. Finally, for the fixed α chosen above, Proposition D.5 implies that there exists $M_2 \in \mathbb{N}$ such that, for every $M \ge M_2$,

$$\sup_{\boldsymbol{\mu}\in\mathcal{K}_{\mathrm{ell}}^T} \mathbb{P}_{\boldsymbol{\mu}}\left(\mathcal{A}_{M,\alpha}(\boldsymbol{\mu})^c\right) < \delta. \tag{144}$$

Let $M_0 := \max\{M_1, M_2\}$. Then, for every $M \ge M_0$, choosing θ according to (128) with the fixed η above, we obtain from (131), (143), and (144) that

$$\sup_{\boldsymbol{\mu}\in\mathcal{K}_{\mathrm{ell}}^T} \mathbb{P}_{\boldsymbol{\mu}}\left(\max_{s\in[T]} W_2\left(\mathcal{T}_{M,\theta}(\widehat{\boldsymbol{\mu}}_M)_s, \varphi(\boldsymbol{\mu})_s\right) > \varepsilon \right) < \delta. \qquad \square$$

F Decoder reconstruction for elliptical measures (Section 4)

First, recall the stochastic representation of an elliptical measure given by Cambanis et al. (1981),

$$\boldsymbol{X} = \boldsymbol{m} + R\boldsymbol{C}^{1/2}\boldsymbol{u}. \tag{145}$$

Now, fix a token $\boldsymbol{z} = [\boldsymbol{z}_1^\top, \dots, \boldsymbol{z}_{L_d}^\top]^\top \in \mathbb{R}^{ML_d}$, $\boldsymbol{z}_\ell \in \mathbb{R}^M$, where the ℓth block corresponds to direction $\boldsymbol{s}_\ell \in V_d$. Let $\mathrm{proj}_M^\uparrow : \mathbb{R}^M \to \mathbb{R}_\uparrow^M$ denote the Euclidean projection onto the monotone cone. For every $\ell \in [L_d]$, the quantile function associated to $\boldsymbol{z}$ is

$$\phi_{\ell,\boldsymbol{z}} := C_M^{-1}\left(\mathrm{proj}_M^\uparrow\left(\sqrt{L_d}\,\boldsymbol{z}_\ell\right)\right) \in \mathcal{S}_M^\uparrow. \tag{146}$$

Whenever $\boldsymbol{z} = \overline{E}_M(\mu)$, it holds that $\phi_{\ell,\boldsymbol{z}} = \Pi_M \Phi(\mu^{\boldsymbol{s}_\ell})$. Since $\phi_{\ell,\boldsymbol{z}} \in \mathcal{S}_M$, it takes the form $\phi_{\ell,\boldsymbol{z}} = \sum_{i=1}^M a_{\ell,i}(\boldsymbol{z})\mathbf{1}_{I_{i,M}}$, which we use henceforth.

F.1 Recovery of the mean and covariance

The quantile function ϕ_μ of a one-dimensional distribution μ retains all its distributional information. This fact leads to an identity which we use repeatedly in our development: given a random variable $V \sim \mathrm{Unif}(0,1)$ and $Y \sim \mu$, it holds that

$$\phi_\mu(V) \overset{d}{=} Y. \tag{147}$$

For a canonical reference and proof, see Billingsley (1995, Thm. 14.1, second proof).

Since $\phi_{\ell,\boldsymbol{z}}$ in (146) is nondecreasing, it is the quantile function of the equal-weight measure $\frac{1}{M}\sum_{i=1}^M \delta_{a_{\ell,i}(\boldsymbol{z})}$. We thus use (147) to recover the directional mean and variance of this measure by

$$m_{\boldsymbol{s}_\ell}(\boldsymbol{z}) := \frac{1}{M}\sum_{i=1}^M a_{\ell,i}(\boldsymbol{z}), \qquad c_{\boldsymbol{s}_\ell}^2(\boldsymbol{z}) := \frac{1}{M}\sum_{i=1}^M \left(a_{\ell,i}(\boldsymbol{z}) - m_\ell(\boldsymbol{z})\right)^2. \tag{148}$$

Using the directions composing V_d in (15), we reconstruct the mean as

$$\widehat{\boldsymbol{m}}(\boldsymbol{z}) \;\coloneqq\; [m_{\boldsymbol{e}_1}(\boldsymbol{z}), m_{\boldsymbol{e}_2}(\boldsymbol{z}), \dots m_{\boldsymbol{e}_d}(\boldsymbol{z})]^\top, \tag{149}$$

and the individual covariance entries $(i,j) \in [d] \times [d]$ as

$$[\boldsymbol{C}^0(\boldsymbol{z})]_{ii} \;\coloneqq\; c^2_{\boldsymbol{e}_i}(\boldsymbol{z}), \quad [\boldsymbol{C}^0(\boldsymbol{z})]_{ij} = [\boldsymbol{C}^0(\boldsymbol{z})]_{ji} \;\coloneqq\; c^2_{(\boldsymbol{e}_i+\boldsymbol{e}_j)/\sqrt{2}}(\boldsymbol{z}) - \frac{1}{2}\left(c^2_{\boldsymbol{e}_i}(\boldsymbol{z}) + c^2_{\boldsymbol{e}_j}(\boldsymbol{z})\right). \tag{150}$$

Since for an arbitrary transformer output token $\boldsymbol{C}^0$ need not be positive-semidefinite, we further project onto $\mathbb{S}^d_+$ to get the final reconstruction

$$\widehat{\boldsymbol{C}}(\boldsymbol{z}) \;\coloneqq\; \operatorname*{arg\,min}_{\boldsymbol{C}\in\mathbb{S}^d_+} \left\| \boldsymbol{C} - \boldsymbol{C}^0(\boldsymbol{z}) \right\|_F^2. \tag{151}$$

F.2 Standardized marginal extraction

Recall that $\phi_{\ell,\boldsymbol{z}}$ approximates $\Phi(\mu^{\boldsymbol{s}_\ell})$ of the supposed measure μ encoded in $\boldsymbol{z}$. Since according to (16) every direction contains the same standardized quantile, we seek to recover it from our slice estimates $\phi_{\ell,\boldsymbol{z}}$.

For every $\ell \in [L_d]$, let $\widehat{c}_\ell(\boldsymbol{z}) \coloneqq \sqrt{\boldsymbol{s}_\ell^\top \widehat{\boldsymbol{C}}(\boldsymbol{z}) \boldsymbol{s}_\ell}$ as the recovered directional variance and define $\beta(\boldsymbol{z}) \coloneqq \sum_{\ell=1}^{L_d} \widehat{c}^2_\ell(\boldsymbol{z})$. Whenever $\beta(\boldsymbol{z}) > 0$, define $\phi^0_{\boldsymbol{z}} \in L^2$ as the solution of

$$\phi^0_{\boldsymbol{z}} \in \operatorname*{arg\,min}_{\varphi\in L^2} \sum_{\ell=1}^{L_d} \left\| \phi_{\ell,\boldsymbol{z}} - \langle \boldsymbol{s}_\ell,\, \widehat{\boldsymbol{m}}(\boldsymbol{z})\rangle \mathbf{1} - \widehat{c}_\ell(\boldsymbol{z})\varphi \right\|^2_{L^2}. \tag{152}$$

Note that since all compoenents of (152) are in $\mathcal{S}_M$, the minimizer necessarily belongs to $\mathcal{S}_M$. Therefore, the minimization is finite dimensional and amounts to recovering the coefficients of every component constant function on the partition (9). The minimizer is unique and is given by

$$\phi^0_{\boldsymbol{z}} = \frac{\displaystyle\sum_{\ell=1}^{L_d} \widehat{c}_\ell(\boldsymbol{z}) \left(\phi_{\ell,\boldsymbol{z}} - \langle \boldsymbol{s}_\ell,\, \widehat{\boldsymbol{m}}(\boldsymbol{z})\rangle \mathbf{1}\right)}{\beta(\boldsymbol{z})}. \tag{153}$$

On an exact elliptical law, this corresponds to (16), for which the common standardized marginal is Φ_{RU_1}. Since the recovery is only approximate, and we know that the law of RU_1 is symmetrical around zero, we enforce this structural property on our standardized quantile approximation by defining

$$\phi^{\mathrm{sym}}_{\boldsymbol{z}}(u) := \frac{\phi^0_{\boldsymbol{z}}(u) - \phi^0_{\boldsymbol{z}}(1-u)}{2}. \tag{154}$$

To control stability of the reconstruction in ill-defined cases, we choose a floor threshold κ_M with $\lim_{M\to\infty} \kappa_M = 0$. Then, the good scenario of $\beta(\boldsymbol{z}) > \kappa_M$ or $\| \phi^{\mathrm{sym}}_{\boldsymbol{z}} \|_{L^2} > \kappa_M$, gives a quantile approximation of

$$\widehat{\phi}_{RU_1,\boldsymbol{z}} \;\coloneqq\; \frac{\phi^{\mathrm{sym}}_{\boldsymbol{z}}}{\| \phi^{\mathrm{sym}}_{\boldsymbol{z}} \|_{L^2}}. \tag{155}$$

Otherwise, it means we are dealing either with a pathological output $\boldsymbol{z}$ for which $\beta(\boldsymbol{z}) \sim 0$ or $\| \phi^{\mathrm{sym}}_{\boldsymbol{z}} \|_{L^2} \sim 0$. Since D_M has to be defined for any $\boldsymbol{z} \in \mathbb{R}^{ML_d}$, we bypass the radial fitting step and directly set the radial law as

$$\widehat{\nu}_{R,M}(\boldsymbol{z}) \;\coloneqq\; \delta_{\sqrt{d}}, \tag{156}$$

so as to saisfy the second moment assumption on the random variable R, of $\mathbb{E}[R^2] = d$. This fallback branch thus produces a valid covariance-normalized elliptical law $\widehat{\boldsymbol{m}} + \widehat{\boldsymbol{C}}^{1/2}\sqrt{d}\,\boldsymbol{u}$. We will subsequently show that, asymptotically, this fallback branch is never accessed.

F.3 RADIAL LAW RECONSTRUCTION

If the transformer output $\boldsymbol{z}$ is ill-behaved, we set the radial law to (156). This is a valid covariance-normalized radial law since its second moment is d.

Otherwise, since U_1 is symmetric, the law of $|RU_1|$ determines the signed marginal law of RU_1. Because RU is spherically symmetric, this one-coordinate marginal determines the full law of RU, and hence the radial law $\mathrm{law}(R)$ through $R = \|RU\|$. Thus ν_R can be approximated from $\widehat{\phi}_{RU_1,\boldsymbol{z}}$.

For any fixed M, we choose a grid of J_M radii $0 = r_{1,M} < \cdots < r_{J_M,M}$ taken to contain $\sqrt{d}$ and assume that the radial law has the form $\nu_R \approx \sum_{j=1}^{J_M} w_j \delta_{r_{j,M}}$. Based on our normalization convention, the admissible weights lie in the set

$$\mathcal{W}_M := \left\{ \boldsymbol{w} \in \mathbb{R}^{J_M} : \boldsymbol{w} \geq \mathbf{0},\ \mathbf{1}^\top \boldsymbol{w} = 1,\ \sum_{j=1}^{J_M} w_j r_{j,M}^2 = d \right\}, \tag{157}$$

which is nonempty due to the inclusion of $\sqrt{d}$. We are then tasked with reconstructing the optimal vector of weights $\boldsymbol{w}^\star = [w_1, \dots w_{J_M}] \in \mathcal{W}_M$ from the distribution of the absolute marginal $|RU_1|$ which we recovered approximately (the absolute value is chosen for convenience, since it removes the redundant sign from the symmetric random variable U_1).

We select a set of K_M CDF threshold values against which to evaluate the distribution of $|\,RU_1\,|$, $\{x_{k,M}\}_{k=1}^{K_M} \subseteq \mathbb{R}_+$. Then, for ever $x_{k,M}$ we can evaluate the CDF of RU_1 at $x_{k,M}$ via the following relation

$$\mathbb{P}(|RU_1| \leq x_{k,M}) = \sum_{j=1}^{J_M} w_j\, \mathbb{P}(r_{j,M}|U_1| \leq x_{k,M}). \tag{158}$$

First, we compute the LHS of (158) using $\widehat{\phi}_{RU_1,\boldsymbol{z}}$ in (155). Again using (147), we obtain

$$\begin{aligned}
\mathbb{P}\left(|RU_1| \leq x_{k,M}\right) \approx\ & \mathbb{P}\left(|\widehat{\phi}_{RU_1,\boldsymbol{z}}(V)| \leq x_{k,M}\right) \\
=\ & \mathbb{P}\left(|\sum_{i=1}^{M} \widehat{a}_{i,\boldsymbol{z}} 1_{I_{i,M}}(V)| \leq x_{k,M}\right) \\
=\ & \mathbb{P}\left(\sum_{i=1}^{M} |\widehat{a}_{i,\boldsymbol{z}}| 1_{I_{i,M}}(V) \leq x_{k,M}\right) \\
=\ & \frac{1}{M} \sum_{i=1}^{M} 1_{\{|\,\widehat{a}_{i,\boldsymbol{z}}\,| \leq x_{k,M}\}},
\end{aligned} \tag{159}$$

where the second equality comes from $\widehat{\phi}_{RU_1,\boldsymbol{z}} \in \mathcal{S}_M$, while the rest follows from disjointness of $I_{i,M}$. We collect these probabilities in a vector $\widehat{\boldsymbol{p}}_M(\boldsymbol{z}) \in \mathbb{R}^{K_M}$ defined as

$$\widehat{\boldsymbol{p}}_M(\boldsymbol{z}) \ :=\ \left[\, \mathbb{P}\left(|RU_1| \leq x_{1,M}\right), \dots \mathbb{P}\left(|RU_1| \leq x_{K_M,M}\right) \,\right]^\top. \tag{160}$$

Second, we evaluate the probabilities on the RHS of (158). To this end, we use the result $U_1^2 \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$ (Muirhead, 1982, Section 1.5) which gives

$$P(r|U_1| \leq x) \ :=\ \begin{cases} I_{\min\{x^2/r^2,1\}}\left(\frac{1}{2}, \frac{d-1}{2}\right) & \text{if } r > 0 \\ 1, \text{ if } r = 0, \end{cases} \tag{161}$$

where $I_y(a,b)$ is the regularized incomplete beta function, and the minimum is taken to account for $|\,U_1\,| \in [0,1]$ sealing the probability to 1 once $x \geq r$. Then, for every pair $(j,k) \in [J_M] \times [K_M]$, we collect the probability $\mathbb{P}\left(r_{j,M}|U_1| \leq x_{k,M}\right)$ in a matrix $\boldsymbol{A}_M \in \mathbb{R}^{K_M \times J_M}$ as

$$[\boldsymbol{A}_M]_{kj} \ :=\ \mathbb{P}\left(r_{j,M}|U_1| \leq x_{k,M}\right). \tag{162}$$

Thus, recovering ν_R amounts to solving the linear system $\boldsymbol{A}_M \boldsymbol{w} = \widehat{\boldsymbol{p}}_M(\boldsymbol{z})$. Practically, we approximate a solution by optimizing the following ρ_M-regularized strongly convex constrained objective

$$\widehat{\boldsymbol{w}}_M(\boldsymbol{z}) \in \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathcal{W}_M} \left\{ \frac{1}{K_M} \| \boldsymbol{A}_M \boldsymbol{w} - \widehat{\boldsymbol{p}}_M(\boldsymbol{z}) \|_2^2 + \rho_M \| \boldsymbol{w} \|_2^2 \right\}. \tag{163}$$

The reconstructed radial law is then $\widehat{\nu}_{R,M}(\boldsymbol{z}) := \sum_{j=1}^{J_M} [\widehat{\boldsymbol{w}}_M(\boldsymbol{z})]_j \delta_{r_{j,M}}$, which, clearly satisfies $\mathbb{E}_{\widehat{\nu}_{R,M}(\boldsymbol{z})}[r^2] = d$. Optimization (163) is carried out up to an error of τ_M, which is theoretically determined in subsequent sections.

F.4 Deterministic reconstruction of the M-atom output empirical measure

We reconstruct the cloud using the stochastic representation in (145). For the fixed ambient dimension d, choose a sequence of directions $\{\boldsymbol{u}_i\}_{i \geq 1} \subseteq S^{d-1}$ satisfying

$$\lim_{k \to \infty} \mathrm{W}_2 \left(\frac{1}{k} \sum_{i=1}^{k} \delta_{\boldsymbol{u}_i}, \sigma_d \right) = 0, \tag{164}$$

where σ_d is the uniform probability measure on S^{d-1}. We use the same sequence for every input token and every output position.

Choosing and fixing the directions. A sequence satisfying (164) can be obtained by sampling independent uniform directions once. For example, draw independent vectors $\boldsymbol{g}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and set

$$\boldsymbol{u}_i = \frac{\boldsymbol{g}_i}{\| \boldsymbol{g}_i \|}, \qquad i \geq 1. \tag{165}$$

These directions are almost surely well defined and are independent with law σ_d (Muirhead, 1982, Theorem 1.5.6).

Lemma F.1 (Consistency of sampled directions)**.** *Independent directions with law σ_d satisfy* (164) *almost surely.*

Proof. By Varadarajan's theorem (Varadarajan, 1958), their empirical measures converge weakly to σ_d almost surely. All these measures have Euclidean second moment equal to 1. The characterization of Wasserstein convergence by weak convergence and convergence of second moments therefore gives (164) (Ambrosio et al., 2005, Remark 7.1.11). □

Fix any realization satisfying (164). The resulting decoder is deterministic as a function of $\boldsymbol{z}$. For each M, only a finite prefix of this sequence is needed; additional directions can be sampled and stored when a longer prefix is required. The directions are never resampled between tokens or output positions.

Radial allocation and synthesis. We construct M equally weighted particles using R_M representative radii, where $1 \leq R_M \leq M$. Divide the particles as evenly as possible into R_M groups. Thus the group sizes $k_{s,M}$ sum to M and belong to $\{\lfloor M/R_M \rfloor, \lceil M/R_M \rceil\}$, with the larger groups first.

Group s represents radial probability mass $p_{s,M} = k_{s,M}/M$. Partition $(0, 1]$ into consecutive cells with these lengths:

$$\ell_{0,M} = 0, \qquad \ell_{s,M} = \sum_{i=1}^{s} p_{i,M}, \qquad J_{s,M} = (\ell_{s-1,M}, \ell_{s,M}], \quad 1 \leq s \leq R_M. \tag{166}$$

For $\nu \in \mathcal{P}_2([0, \infty))$, define the representative radius of group s by

$$r_{s,M}(\nu) = \frac{1}{p_{s,M}} \int_{J_{s,M}} \Phi(\nu)(v)\, dv. \tag{167}$$

This average minimizes the squared error when replacing the radial quantile by a constant on $J_{s,M}$.

For the fitted radial law $\widehat{\nu}_R = \sum_{j=1}^{J_M} w_j \delta_{r_{j,M}}$, with dependence on $\boldsymbol{z}$ and M understood, the integral is a finite sum. Writing $W_j = \sum_{i=1}^{j} w_i$ and $W_0 = 0$, we obtain

$$r_{s,M}(\widehat{\nu}_R) = \frac{1}{p_{s,M}} \sum_{j=1}^{J_M} r_{j,M} \, | \, J_{s,M} \cap (W_{j-1}, W_j] \, | \, , \tag{168}$$

where $|\cdot|$ denotes interval length. For the fallback law $\delta_{\sqrt{d}}$, all representative radii equal $\sqrt{d}$.

At each representative radius, use the first $k_{s,M}$ directions of the fixed sequence. Applying the location and covariance transformation gives

$$\mathcal{R}_M(\boldsymbol{m}, \boldsymbol{C}, \nu) = \frac{1}{M} \sum_{s=1}^{R_M} \sum_{i=1}^{k_{s,M}} \delta_{\boldsymbol{m} + r_{s,M}(\nu) \boldsymbol{C}^{1/2} \boldsymbol{u}_i}, \quad \text{and} \quad D_M(\boldsymbol{z}) = \mathcal{R}_M\big(\widehat{\boldsymbol{m}}(\boldsymbol{z}), \widehat{\boldsymbol{C}}(\boldsymbol{z}), \widehat{\nu}_R\big). \tag{169}$$

We output exactly M equal-weight particles. Performing angular approximation at every representative radius approximates the independent radius–direction representation.

F.5 Computational complexity

Projecting onto the monotone cone of $\mathbb{R}$. The first step of projecting onto the monotone cone of R^M amounts to one-dimensional isotonic regression for which the PAV method (Ayer et al., 1955) requires $\mathcal{O}(M)$ time and $\mathcal{O}(M)$ memory. Hence projecting all $L_d = \mathcal{O}(d^2)$ blocks costs $\mathcal{O}(Md^2)$.

Recovering the mean and covariance. Computing directional means and variances costs $\mathcal{O}(M)$ and therefore totals to $\mathcal{O}(Md^2)$ for all directions. Assembling them into the mean vector and covariance matrix adds a further $\mathcal{O}(d^2)$. Finally, projecting $\boldsymbol{C}^0(\boldsymbol{z})$ onto $\mathbb{S}_+^d$ adds $\mathcal{O}(d^3)$. This step amounts to a total of order $\mathcal{O}(Md^2 + d^3)$.

Reconstructing the radial law $\widehat{\nu}_R$. We break the radial law reconstruction in Section F.3 into three blocks of significant cost. First, assembling $\widehat{\phi}_{RU_1,\boldsymbol{z}}$, requires a straightforward $\mathcal{O}(Md^2)$ operations. Second, computing $[\widehat{\boldsymbol{p}}_M]_k$ can be done in $\mathcal{O}(M + K_M)$ using the fact that quantile values and thresholds are ordered (a two-pointer linear scan and merge). Finally, consider solving (163) using accelerated projected gradient with $\rho_M = \tau_M = M^{-1}$, where τ_M is the target objective suboptimality. The objective is L-smooth with constant

$$L_M = \frac{2}{K_M} \|\boldsymbol{A}_M\|_{\mathrm{op}}^2 + \frac{2}{M} \le 2J_M + \frac{2}{M},$$

since the entries of $\boldsymbol{A}_M$ lie in $[0,1]$. As $\mathcal{W}_M$ is contained in the probability simplex, its Euclidean diameter is at most $\sqrt{2}$. The accelerated convergence bound therefore guarantees objective error at most M^{-1} after $\mathcal{O}(\sqrt{MJ_M})$ iterations.

Each iteration requires a gradient evaluation of cost $\mathcal{O}(K_M J_M)$, and a Euclidean projection

$$\Pi_{\mathcal{W}_M}(v) = \operatorname*{argmin}_{w \ge 0} \left\{ \frac{1}{2}\|w - v\|_2^2 : \sum_{j=1}^{J_M} w_j = 1, \quad \sum_{j=1}^{J_M} r_{j,M}^2 w_j = d \right\}.$$

This is a separable convex quadratic program with two linear equality constraints, admitting an exact $\mathcal{O}(J_M)$-time algorithm in the real-arithmetic model (Megiddo & Tamir, 1993). Including projections, the optimization cost is thus

$$\mathcal{O}\left((K_M J_M + J_M)\sqrt{MJ_M}\right).$$

For fixed d and our grids $J_M, K_M = \mathcal{O}(\sqrt{M})$, this becomes $\mathcal{O}(M^{7/4})$. This is an achievable theoretical bound; the experiments use CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018) with OSQP and a Clarabel fallback.

Generating the output empirical measure deterministically. For the practical reference cloud, we need only the first $k_M^{\max} = \lceil M/R_M \rceil$ directions. Generating these directions by (165) requires $dk_M^{\max}$ scalar standard normal draws and $\mathcal{O}(dk_M^{\max})$ arithmetic operations for normalization. Under a unit-cost model for normal sampling, this is an offline cost of $\mathcal{O}(dk_M^{\max})$ shared across reconstructions.

After the radial fit and computation of $\widehat{C}(\boldsymbol{z})^{1/2}$, we evaluate (168) by traversing the fitted-weight intervals $(W_{j-1}, W_j]$ and the cells $J_{s,M}$ in order. This requires $\mathcal{O}(J_M + R_M)$ operations, since each step advances at least one interval index.

Next, compute $\widehat{C}(\boldsymbol{z})^{1/2}\boldsymbol{u}_i$ once for each $i \le k_M^{\max}$ and reuse these vectors across radial groups. Including direction initialization, dense matrix–vector multiplication, and the final scaling and translation, the first reconstruction costs

$$O\big(dk_M^{\max} + J_M + R_M + d^2 k_M^{\max} + Md\big). \tag{170}$$

Subsequent reconstructions reuse the stored directions and omit their initialization cost. This count excludes moment extraction, covariance factorization, and radial optimization. For fixed d and $J_M = O(M)$, the synthesis cost is $O(M)$, matching the order required to write the output coordinates. This does not establish optimality of the complete decoder.

Approximate total cost. Under the explicit schedules,

$$J_M, K_M, R_M = \mathcal{O}\left(\sqrt{M}\right), \qquad k_M^{\max} = \left\lceil \frac{M}{R_M} \right\rceil = \mathcal{O}\left(\sqrt{M}\right), \qquad \rho_M = \tau_M = M^{-1}.$$

Combining the preceding costs, decoding one token into an M-atom empirical measure requires $\mathcal{O}\left(M^{7/4} + Md^2 + d^3\right)$ arithmetic operations, excluding the offline generation and storage of the fixed angular directions. For fixed ambient dimension d, the radial optimization dominates asymptotically, giving the bound $\mathcal{O}\left(M^{7/4}\right)$.

F.6 Resolution choices and decoder stability statement

We now specify the asymptotic resolutions used by the decoder. Recall the radial grid

$$0 = r_{1,M} < \cdots < r_{J_M,M},$$

introduced in Section F.3 and chosen to contain $\sqrt{d}$. Define its maximal radius and maximal spacing

$$B_M \;:=\; r_{J_M,M}, \qquad h_M \;:=\; \max_{1 \le j < J_M} (r_{j+1,M} - r_{j,M}). \tag{171}$$

Thus $B_M \to \infty$ allows the grid to capture arbitrarily large radii, while $h_M \to 0$ makes the grid asymptotically dense.

Recall also that K_M is the number of CDF evaluations in the radial fit (163), ρ_M is its regularization parameter, τ_M is the permitted objective suboptimality, κ_M is the fallback threshold used in the standardized-marginal reconstruction, and R_M is the number of representative radii used in the particle construction (166)–(169). We require

$$\begin{cases} B_M \to \infty, \quad h_M \to 0, \quad K_M \to \infty; \\ \rho_M > 0, \quad \rho_M \to 0, \quad \kappa_M > 0, \quad \kappa_M \to 0, \quad \tau_M \to 0; \\ R_M \to \infty, \quad \frac{R_M}{M} \to 0. \end{cases} \tag{172}$$

The first line ensures consistency of the radial-law discretization and of the finite CDF objective. The second removes the asymptotic effects of regularization, fallback thresholding, and numerical optimization. The final line ensures that the radial quantile partition becomes finer while the group sizes $k_{s,M}$ in (166) diverge, so that the angular approximation (164) is accurate per radial group.

For the CDF evaluations in (158)–(163), for every $k \in [K_M]$ we use the tangent thresholds

$$x_{k,M} \;:=\; \tan\left(\frac{\pi}{2}\,\frac{k - \frac{1}{2}}{K_M}\right). \tag{173}$$

These are the images under $t \mapsto \tan(\pi t/2)$ of the midpoints of the uniform K_M-partition of $(0,1)$. One explicit choice satisfying (172) is

$$\begin{aligned} n_M &= \lceil M^{1/4} \rceil + \lceil \sqrt{d} \rceil, \qquad \{r_{j,M}\}_{j=1}^{J_M} = \{j/n_M : 0 \le j \le n_M^2\} \cup \{\sqrt{d}\}, \\ K_M &= \lceil \sqrt{M} \rceil, \qquad R_M = \lfloor \sqrt{M} \rfloor, \qquad \rho_M = \kappa_M = \tau_M = M^{-1}, \end{aligned} \tag{174}$$

where the radial grid is sorted and duplicate points are removed. For this choice,

$$B_M = n_M \to \infty, \qquad h_M \le n_M^{-1} \to 0,$$

and the inclusion of $\sqrt{d}$ guarantees that the feasible set $\mathcal{W}_M$ in (157) is nonempty. For fixed d,

$$J_M = \Theta(\sqrt{M}), \qquad K_M = \Theta(\sqrt{M}), \qquad R_M = \Theta(\sqrt{M}),$$

while $k_{s,M} \in \{\lfloor M/R_M \rfloor, \lceil M/R_M \rceil\} = \Theta(\sqrt{M})$. The particular exponents are not essential, since the stability argument uses only the limiting conditions in (172).

Proposition 4.2 (Uniform decoder stability)**.** *Let $d \ge 2$, and let $\mathcal{K}_{\mathrm{ell}} \subset \mathcal{P}_2(\mathbb{R}^d)$ be a nonempty W_2-compact family of elliptical laws with positive-definite covariances. Let D_M be the decoder constructed in Appendix F, with parameters chosen as specified there. Then, D_M satisfies (4).*

Proof. The proof follows one conceptual chain: how token $\boldsymbol{z}$ error propagates through slice error to $(\widehat{m}, \widehat{C}, \widehat{\phi})$ reconstructions, then through to the radial law approximation $\widehat{\nu}_R$, then finally through the decoded particles to μ.

So we fix μ and a token $\boldsymbol{z}$ with $\|\boldsymbol{z} - \overline{E}_M(\mu)\|_2 \le \eta$, and ask: how does η propagate through D_M as we have defined it in (169)?

For simplicity of notation, let $\nu_\mu := \nu_{R,\mu}$, $\widehat{\boldsymbol{m}} := \widehat{\boldsymbol{m}}(\boldsymbol{z})$, $\widehat{\boldsymbol{C}} := \widehat{\boldsymbol{C}}(\boldsymbol{z})$, $\widehat{\nu}_R := \widehat{\nu}_{R,M}(\boldsymbol{z})$. From (145) and (169), respectively, recall that

$$D_M(\boldsymbol{z}) = \mathcal{R}_M(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{C}}, \widehat{\nu}_R), \quad \text{and} \quad \mu = \mathrm{law}(\boldsymbol{m}_\mu + \boldsymbol{C}_\mu^{1/2} R_\mu \boldsymbol{u}). \tag{175}$$

Now, using two intermediate measures, $\mathcal{R}_M(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{C}}, \nu_\mu)$ and $\mathrm{law}(\widehat{\boldsymbol{m}} + \widehat{\boldsymbol{C}}^{1/2} R_\mu \boldsymbol{u})$, we apply the triangle inequality as follows.

$$\begin{aligned} \mathrm{W}_2(D_M(\boldsymbol{z}), \mu) \le{}& \underbrace{\mathrm{W}_2\Big(\mathcal{R}_M(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{C}}, \widehat{\nu}_R), \mathcal{R}_M(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{C}}, \nu_\mu)\Big)}_{T_{1,M}} \\ &+ \underbrace{\mathrm{W}_2\Big(\mathcal{R}_M(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{C}}, \nu_\mu), \mathrm{law}(\widehat{\boldsymbol{m}} + \widehat{\boldsymbol{C}}^{1/2} R_\mu \boldsymbol{u})\Big)}_{T_{2,M}} \\ &+ \underbrace{\mathrm{W}_2\Big(\mathrm{law}(\widehat{\boldsymbol{m}} + \widehat{\boldsymbol{C}}^{1/2} R_\mu \boldsymbol{u}), \mathrm{law}(\boldsymbol{m}_\mu + \boldsymbol{C}_\mu^{1/2} R_\mu \boldsymbol{u})\Big)}_{T_{3,M}}. \end{aligned} \tag{176}$$

We start with $T_{3,M}$ and notice that the two laws are based on the same true radial variable R_μ and the same spherical direction law $\boldsymbol{u} \sim \mathbb{S}^{d-1}$. We upper bound $T_{3,M}$ by constructing one admissible coupling of the two laws. To this end, we define the random vectors $\widehat{X} := \widehat{\boldsymbol{m}} + \widehat{\boldsymbol{C}}^{1/2} R_\mu \boldsymbol{u}$, and $X := \boldsymbol{m}_\mu + \boldsymbol{C}_\mu^{1/2} R_\mu \boldsymbol{u}$ which obey the two laws inside $T_{3,M}$, respectively. Therefore, $(\widehat{X}, X)$ is a valid coupling, and using the optimal property of the W_2 definition we have that

$$\begin{aligned} T_{3,M} &\le \Big(\mathbb{E}\|\widehat{X} - X\|^2\Big)^{1/2} \\ &= \left(\mathbb{E}\left\|(\widehat{\boldsymbol{m}} - \boldsymbol{m}_\mu) + (\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2}) R_\mu \boldsymbol{u}\right\|^2\right)^{1/2} \\ &\le \|\widehat{\boldsymbol{m}} - \boldsymbol{m}_\mu\| + \left(\mathbb{E}\left\|(\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2}) R_\mu \boldsymbol{u}\right\|^2\right)^{1/2} \end{aligned}$$

$$= \|\widehat{\boldsymbol{m}} - \boldsymbol{m}_\mu\| + \sqrt{d}\left(\mathbb{E}\left\|(\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})\boldsymbol{u}\right\|^2\right)^{1/2}, \tag{177}$$

where for the last inequality we applied Minkowski's inequality, and the equality comes from $\boldsymbol{u}$ and R's independence and the radius normalization assumption. We further process the second term on the RHS of (177).

$$\begin{aligned}
\left(\mathbb{E}\left\|(\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})\boldsymbol{u}\right\|^2\right)^{1/2} &= \left(\mathbb{E}\left[\boldsymbol{u}^\top(\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})^\top(\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})\boldsymbol{u}\right]\right)^{1/2} \\
&= \left(\mathbb{E}\left[\mathrm{Tr}\left((\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})^\top(\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})\boldsymbol{u}\boldsymbol{u}^\top\right)\right]\right)^{1/2} \\
&= \left(\mathrm{Tr}\left((\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})^\top(\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\top\right]\right)\right)^{1/2} \\
&= \frac{1}{\sqrt{d}}\left(\mathrm{Tr}\left((\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})^\top(\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2})\right)\right)^{1/2},
\end{aligned} \tag{178}$$

where the first equality comes from circularity of the trace, the second comes from linearity of expectation, and the third comes form $\mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\top\right] = \frac{1}{d}\boldsymbol{I}_d$ for uniform vectors on the sphere. Putting (177) and (178) together, and using the definition of the Frobenius norm, we get that

$$T_{3,M} \le \|\widehat{\boldsymbol{m}} - \boldsymbol{m}_\mu\| + \left\|\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2}\right\|_F \tag{179}$$

It remains to replace the square-root covariance error by the covariance error controlled in Lemma G.2. By point (c) of Lemma G.2, $\widehat{\boldsymbol{C}} \succeq \frac{\lambda_-}{2}\boldsymbol{I}_d$, while Lemma G.1 gives $\boldsymbol{C}_\mu \succeq \lambda_-\boldsymbol{I}_d$. Hence both matrices are bounded below by $(\lambda_-/2)\boldsymbol{I}_d$, and the matrix square-root estimate yields

$$\left\|\widehat{\boldsymbol{C}}^{1/2} - \boldsymbol{C}_\mu^{1/2}\right\|_F \le \frac{\left\|\widehat{\boldsymbol{C}} - \boldsymbol{C}_\mu\right\|_F}{2\sqrt{\lambda_-/2}},$$

where we used the fact that if $A, B \succeq aI_d$,, then $\|A^{1/2} - B^{1/2}\|_F \le \frac{\|A-B\|_F}{2\sqrt{a}}$ (Higham, 2008, Theorem 6.2). Therefore,

$$T_{3,M} \le \|\,\widehat{\boldsymbol{m}} - \boldsymbol{m}_\mu\,\| + \frac{\left\|\widehat{\boldsymbol{C}} - \boldsymbol{C}_\mu\right\|_F}{2\sqrt{\lambda_-/2}}.$$

Applying point (a) of Lemma G.2, we obtain

$$T_{3,M} \le C(\eta + b_M), \tag{180}$$

where the constant $C < \infty$ is independent of M, η, μ, and $\boldsymbol{z}$.

We move on to $T_{1,M}$ where we shoudl estimate the discrepancy between $\widehat{\nu}_R$ and ν_μ. We begin by using the Lipschitz property proven in Lemma G.7 (b) and point (c) in the covariance extraction lemma G.2 to bound

$$\begin{aligned}
T_{1,M} &:= \mathrm{W}_2\Big(\mathcal{R}_M(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{C}}, \widehat{\nu}_R), \mathcal{R}_M(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{C}}, \nu_{R,\mu})\Big) \\
&\le \left\|\widehat{\boldsymbol{C}}^{1/2}\right\|_{\mathrm{op}} \mathrm{W}_2(\widehat{\nu}_R, \nu_{R,\mu}) \\
&\le \sqrt{2\lambda_+}\mathrm{W}_2(\widehat{\nu}_R, \nu_{R,\mu})
\end{aligned} \tag{181}$$

To control the radial-law error $\mathrm{W}_2(\widehat{\nu}_R, \nu_{R,\mu})$, we use Proposition G.1. We first check the conditions of the proposition. Let sequences $\eta_n \downarrow 0$ and $M_n \to \infty$, and choose arbitrary $\mu_n \in \mathcal{K}_{\mathrm{ell}}$ and $\boldsymbol{z}_n \in \mathbb{R}^{M_n L_d}$ satisfying

$$\left\|\,\boldsymbol{z}_n - \overline{E}_{M_n}(\mu_n)\,\right\|_2 \le \eta_n.$$

Define $\nu_n := \nu_{R,\mu_n}$, and $\widehat{\nu}_n := \widehat{\nu}_{R,M_n}(\boldsymbol{z}_n)$. Since $\nu_n \in \mathcal{V}_R$, these form an admissible sequence of target radial laws for Proposition G.1.

Let $V \sim \mathrm{Unif}(0,1)$ and define $\delta_n := \mathrm{W}_2\Big(\mathrm{law}\Big(\Big|\widehat{\phi}_{RU_1,\boldsymbol{z}_n}(V)\Big|\Big), \mathcal{A}_d\nu_n\Big)$. By (147),

$$\mathcal{A}_d\nu_n = \mathrm{law}\big(\big|\phi_{R_{\mu_n}U_1}(V)\big|\big).$$

Hence, coupling the two laws with the same V and using $|\,|a| - |b|\,| \le |\,a-b\,|$,

$$\delta_n \le \left(\mathbb{E}\left[\left(\Big|\widehat{\phi}_{RU_1,\boldsymbol{z}_n}(V)\Big| - \big|\phi_{R_{\mu_n}U_1}(V)\big|\right)^2\right]\right)^{1/2} \le \Big\|\,\widehat{\phi}_{RU_1,\boldsymbol{z}_n} - \phi_{R_{\mu_n}U_1}\,\Big\|_{L^2} \le C(\eta_n + b_{M_n}), \tag{182}$$

where the last inequality follows from point (a) of Lemma G.2. Since $\eta_n \to 0$ and $b_M \to 0$, we obtain

$$\delta_n \longrightarrow 0.$$

Moreover, Lemma G.6 and the decoder schedules give $a_{M_n} \longrightarrow 0$, $\rho_{M_n} \longrightarrow 0$, $\tau_{M_n} \longrightarrow 0$, $K_{M_n} \longrightarrow \infty$. Therefore Proposition G.1, applied with target $\nu_n \in \mathcal{V}_R$ in conjunction with (181), yields

$$T_{1,M} = \sqrt{2\lambda_+}\,\mathrm{W}_2(\widehat{\nu}_n, \nu_n) \longrightarrow 0. \tag{183}$$

Finally, we bound the second term in (176). Since $\nu_\mu = \nu_{R,\mu} \in \mathcal{V}_R$, point (c) of Lemma G.7 yields

$$\begin{aligned} T_{2,M} &= \mathrm{W}_2\Big(\mathcal{R}_M(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{C}}, \nu_\mu), \mathrm{law}(\widehat{\boldsymbol{m}} + \widehat{\boldsymbol{C}}^{1/2} R_\mu \boldsymbol{u})\Big) \\ &\le \Big\|\,\widehat{\boldsymbol{C}}^{1/2}\,\Big\|_{\mathrm{op}} \Big(\vartheta_M + \sqrt{d}\,\gamma_M\Big) \\ &\le \sqrt{2\lambda_+}\,\Big(\vartheta_M + \sqrt{d}\,\gamma_M\Big), \end{aligned} \tag{184}$$

By point (a) of Lemma G.7, $\vartheta_M \longrightarrow 0$,$\gamma_M \longrightarrow 0$. Hence

$$T_{2,M} \longrightarrow 0.$$

Finally, since we introduced the sequence M_n to handle T_1, we evaluate all three terms along that same sequence. Therefore Proposition G.1, applied with target $\nu_n \in \mathcal{V}_R$, yields

$$\mathrm{W}_2(\widehat{\nu}_n, \nu_n) \longrightarrow 0.$$

Hence, by (181),

$$T_{1,M_n} \le \sqrt{2\lambda_+}\,\mathrm{W}_2(\widehat{\nu}_n, \nu_n) \longrightarrow 0.$$

For the second term, the bound (184) gives

$$T_{2,M_n} \le \sqrt{2\lambda_+}\,\Big(\vartheta_{M_n} + \sqrt{d}\,\gamma_{M_n}\Big) \longrightarrow 0,$$

since $\vartheta_M \to 0$ and $\gamma_M \to 0$ by Lemma G.7.

Likewise, from (180),

$$T_{3,M_n} \le C(\eta_n + b_{M_n}) \longrightarrow 0,$$

since $\eta_n \to 0$ and $b_M \to 0$.

Returning to (176), we obtain

$$\mathrm{W}_2(D_{M_n}(\boldsymbol{z}_n), \mu_n) \le T_{1,M_n} + T_{2,M_n} + T_{3,M_n} \longrightarrow 0.$$

Recall that the sequences $\eta_n \downarrow 0$ and $M_n \to \infty$, as well as the admissible choices $\mu_n \in \mathcal{K}_{\mathrm{ell}}$ and $\boldsymbol{z}_n \in \mathbb{R}^{M_n L_d}$ satisfying

$$\big\|\,\boldsymbol{z}_n - \overline{E}_{M_n}(\mu_n)\,\big\|_2 \le \eta_n,$$

were arbitrary. Therefore,

$$\lim_{\eta\downarrow 0} \limsup_{M\to\infty} \sup_{\substack{\mu\in\mathcal{K}_{\mathrm{ell}},\ \boldsymbol{z}\in\mathbb{R}^{ML_d} \\ \|\boldsymbol{z}-\overline{E}_M(\mu)\|_2\le\eta}} \mathrm{W}_2(D_M(\boldsymbol{z}), \mu) = 0. \qquad \square$$

G SUPPORTING RESULTS FOR THE PROOF OF DECODER STABILITY IN PROPOSITION 4.2

The proof proceeds in three steps: we first control the recovered geometry and standardized marginal, then the radial fit, and finally the particle reconstruction. This section recurrently uses the stochastic representation of elliptical laws in (145).

G.1 COMPACTNESS AND RECOVERY OF THE GEOMETRY AND MARGINAL

The purpose of this subsection is to show that the geometric and standardized-marginal components of the decoder are stable under small perturbations of the sliced token. More precisely, we show that closeness of $\boldsymbol{z}$ to the population token $\overline{E}_M(\mu)$, together with vanishing finite-resolution slicing error, implies uniform control of the recovered mean, covariance, and standardized marginal. Consequently, the recovered covariance remains uniformly nondegenerate and the fallback branch is eventually inactive. This prepares the quantities required for the radial-law reconstruction in the next subsection.

Lemma G.1 (Uniform bounds for mean and covariance)**.** *Let $d \geq 2$ and consider the W_2-compact set $\mathcal{K}_{\text{ell}} \subset \mathcal{P}_2(\mathbb{R}^d)$ defined in Section 4.1. For every $\mu \in \mathcal{K}_{\text{ell}}$, denote its mean, covariance and radial law respectively by $\boldsymbol{m}_\mu$, $\boldsymbol{C}_\mu$ and $\nu_{R,\mu}$. Then,*

(a) the mean is uniformly bounded over $\mathcal{K}_{\text{ell}}$, namely $\sup_{\mu\in\mathcal{K}_{\text{ell}}} \| \boldsymbol{m}_\mu \| < \infty$;

(b) if $\boldsymbol{C}_\mu \succ \boldsymbol{0}$ for all $\mu \in \mathcal{K}_{\text{ell}}$, there exist $0 < \lambda_- \leq \lambda_+ < \infty$ such that $\lambda_- \boldsymbol{I}_d \preceq \boldsymbol{C}_\mu \preceq \lambda_+ \boldsymbol{I}_d$, $\forall \mu \in \mathcal{K}_{\text{ell}}$;

(c) the set of $\mathcal{K}_{\text{ell}}$-associated radial laws $\mathcal{V}_R := \{\nu_{R,\mu} \,|\, \mu \in \mathcal{K}_{\text{ell}}\}$ is W_2-compact;

(d) the set of $\mathcal{K}_{\text{ell}}$-associated marginal laws $\mathcal{L}_{RU_1} := \{\mathrm{law}(RU_1) \,|\, R \sim \nu_{R,\mu},\ \mu \in \mathcal{K}_{\text{ell}},\ U_1 = \boldsymbol{e}_1^\top \boldsymbol{u}$ with $\boldsymbol{u} \sim S^{d-1}$ and $\boldsymbol{u} \perp R\}$ is W_2-compact.

Proof. The broad approach for proving (a) and (b)is to show that the maps $\mu \mapsto \boldsymbol{m}_\mu$, $\mu \mapsto \lambda_{\min}(\boldsymbol{C}_\mu)$ and $\mu \mapsto \lambda_{\max}(\boldsymbol{C}_\mu)$ are continuous in $\mathcal{K}_{\text{ell}}$. Then, since $\mathcal{K}_{\text{ell}}$ is nonempty and W_2-compact, these continuous functions attain their extrema.

We use Theorem 7.12 implication i) $\implies$ iv) from (Villani, 2003), which states that for a sequence of measures $\mu_n \in \mathcal{P}_2(\mathbb{R}^d)$ and a measure $\mu \in \mathcal{P}(\mathbb{R}^d)$,

$$\lim_{n\to\infty} \mathrm{W}_2(\mu_n, \mu) = 0 \implies \lim_{n\to\infty} \int g d\mu_n = \int g d\mu \tag{185}$$

for any continuous function $g : \mathbb{R}^d \to \mathbb{R}$ such that $| g(\boldsymbol{x}) | \leq C \left[1 + \| \boldsymbol{x} - \boldsymbol{x}_0 \|^2\right]$ for some $\boldsymbol{x}_0 \in \mathbb{R}^d$ and some $C \in \mathbb{R}_+$. For our purposes, we define functions $g_i(\boldsymbol{x}) = x_i$ and $g_{i,j}(\boldsymbol{x}) = x_i x_j$ for $i, j \in [d]$ which satisfy the quadratic growth condition. Then from (185) and g_i we get that $\mu \mapsto \boldsymbol{m}_\mu$ is continuous w.r.t. W_2 and this settles point (a).

Similarly, letting $\boldsymbol{C}_\mu := \int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^\top d\mu(\boldsymbol{x}) - \boldsymbol{m}_\mu \boldsymbol{m}_\mu^\top$ and considering (185) together with $g_{i,j}$ we get that $\mu \mapsto \boldsymbol{C}_\mu$ is continuous w.r.t W_2. Furthermore, we recall the eigenvalue bound (Bhatia, 1997, Corollary III.2.6) for any two symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$

$$\max_{1\leq j\leq n} |\lambda_j(\boldsymbol{A}) - \lambda_j(\boldsymbol{B})| \leq \|\boldsymbol{A} - \boldsymbol{B}\|, \tag{186}$$

which implies that $\lambda_{\min}$ and $\lambda_{\max}$ are 1-Lipschitz with respect to the operator norm. Since $\mu \mapsto \boldsymbol{C}_\mu$ is continuous and $\mathcal{K}_{\text{ell}}$ is compact, the quantities

$$\lambda_- := \min_{\mu\in\mathcal{K}_{\text{ell}}} \lambda_{\min}(\boldsymbol{C}_\mu), \qquad \lambda_+ := \max_{\mu\in\mathcal{K}_{\text{ell}}} \lambda_{\max}(\boldsymbol{C}_\mu)$$

are attained and finite. Positive definiteness of every $\boldsymbol{C}_\mu$ implies $\lambda_- > 0$, proving point (b).

For proving point (c) we we show that the map $\mu \mapsto \nu_{R,\mu}$ is continuous, which implies that it will map the compact $\mathcal{K}_{\text{ell}}$ to a compact image, namely $\mathcal{V}_R$. Let $\mu_n, \mu \in \mathcal{P}_2(\mathbb{R}^d)$. By Proposition 2.1 of Villani (2003), there exists an optimal coupling $\gamma_n^\star(\mu_n, \mu)$ for every n.

Recall from (145) that $\boldsymbol{C}_{\mu_n}^{-1/2}(X-\boldsymbol{m})$ defines the law of $R\boldsymbol{u}$, where $R \perp \boldsymbol{u}$, and $\boldsymbol{u} \sim S^{d-1}$. We define the push-forward maps

$$\mathcal{M}_n(x) \;\coloneqq\; \boldsymbol{C}_{\mu_n}^{-1/2}(x-\boldsymbol{m}_{\mu_n}), \qquad \mathcal{M}(x) \;\coloneqq\; \boldsymbol{C}_{\mu}^{-1/2}(x-\boldsymbol{m}_{\mu}),$$

and apply them to μ_n and μ, respectively, to get the push-forward measures

$$\bar{\mu}_n \;\coloneqq\; (\mathcal{M}_n)_{\#}\mu_n, \qquad \bar{\mu} \;\coloneqq\; \mathcal{M}_{\#}\mu. \tag{187}$$

By the definition of the wasserstein-2 distance, we have that

$$W_2^2(\bar{\mu}_n,\bar{\mu}) \le \int \|\mathcal{M}_n(x)-\mathcal{M}(y)\|^2 \, d\pi_n(x,y), \tag{188}$$

since $\bar{\gamma}_n \;\coloneqq\; (\mathcal{M}_n,\mathcal{M})_{\#}\gamma_n^\star$ is also a coupling. We process the norm on the RHS of (188).

$$\begin{aligned}
\|\,\mathcal{M}_n(x)-\mathcal{M}(y)\,\|^2 &= \Big\| \boldsymbol{C}_{\mu_n}^{-1/2}\big[(x-y)-(\boldsymbol{m}_{\mu_n}-\boldsymbol{m}_{\mu})\big] + \big(\boldsymbol{C}_{\mu_n}^{-1/2}-\boldsymbol{C}_{\mu}^{-1/2}\big)(y-\boldsymbol{m}_{\mu})\Big\|^2 \\
&\le 4\|\boldsymbol{C}_{\mu_n}^{-1/2}\|_{\mathrm{op}}^2\left(\|x-y\|^2+\|\boldsymbol{m}_{\mu_n}-\boldsymbol{m}_{\mu}\|^2\right) \\
&\qquad + 2\|\boldsymbol{C}_{\mu_n}^{-1/2}-\boldsymbol{C}_{\mu}^{-1/2}\|_{\mathrm{op}}^2\|y-\boldsymbol{m}_{\mu}\|^2,
\end{aligned}$$

where we used $\|a+b\|^2 \le 2\|a\|^2+2\|b\|^2$ and $\|Av\| \le \|A\|_{\mathrm{op}}\|v\|$. We then continue (188) as

$$\begin{aligned}
W_2^2(\bar{\mu}_n,\bar{\mu}) &\le \int 4\|\boldsymbol{C}_{\mu_n}^{-1/2}\|_{\mathrm{op}}^2\left(\|x-y\|^2+\|\boldsymbol{m}_{\mu_n}-\boldsymbol{m}_{\mu}\|^2\right) \\
&\qquad + 2\|\boldsymbol{C}_{\mu_n}^{-1/2}-\boldsymbol{C}_{\mu}^{-1/2}\|_{\mathrm{op}}^2\|y-\boldsymbol{m}_{\mu}\|^2 \, d\gamma_n^\star(x,y), \\
&= 4\|\boldsymbol{C}_{\mu_n}^{-1/2}\|_{\mathrm{op}}^2\left(\mathrm{W}_2(\mu_n,\mu)+\|\boldsymbol{m}_{\mu_n}-\boldsymbol{m}_{\mu}\|^2\right) \\
&\qquad + 2\|\boldsymbol{C}_{\mu_n}^{-1/2}-\boldsymbol{C}_{\mu}^{-1/2}\|_{\mathrm{op}}^2\int \|y-\boldsymbol{m}_{\mu}\|^2 \, d\gamma_n^\star(x,y)
\end{aligned} \tag{189}$$

Now, whenever $\lim_{n\to\infty}\mathrm{W}_2(\mu_n,\mu)=0$ points a) and b) ensure the RHS goes to zero and hence, $\lim_{n\to\infty}\mathrm{W}_2(\bar{\mu}_n,\bar{\mu})=0$.

Now, by our definition (187), $\mu_n = \mathrm{law}(R_{\mu_n}\boldsymbol{u})$. Notice that since $\|\,\boldsymbol{u}\,\|=1$ almost surely, we obtain $\nu_{R,\mu_n} = (\|\cdot\|)_{\#}\bar{\mu}$. Since the $\|\cdot\|$ is a 1-Lipschitz map, it contracts the W_2 distance giving

$$\mathrm{W}_2(\nu_{R,\mu_n},\nu_{R,\mu}) \le \mathrm{W}_2(\bar{\mu}_n,\bar{\mu}), \tag{190}$$

and we have our result.

Proving point (d) follows a similar route as (c), noting that the first coordinate projection of $\boldsymbol{u}$ is a 1-Lipschitz map. □

Lemma G.2 (Stability of moment and marginal extraction)**.** *There exist constants $\eta_0 \in (0,1)$, $C<\infty$, and $M_0 \in \mathbb{N}$ such that, for every $\eta \in (0,\eta_0]$, every $M \ge M_0$, every $\mu \in \mathcal{K}_{\mathrm{ell}}$, and every $\boldsymbol{z} \in \mathbb{R}^{ML_d}$ satisfying $\left\|\,\boldsymbol{z}-\overline{E}_M(\mu)\,\right\|_2 \le \eta$,*

(a) the moment, covariance, and radial law quantile reconstruction are bounded by

$$\|\,\widehat{\boldsymbol{m}}(\boldsymbol{z})-\boldsymbol{m}_{\mu}\,\| + \left\|\,\widehat{\boldsymbol{C}}(\boldsymbol{z})-\boldsymbol{C}_{\mu}\,\right\|_F + \left\|\,\widehat{\phi}_{RU_1,\boldsymbol{z}}-\phi_{R_\mu U_1}\,\right\|_{L^2} \le C(\eta+b_M),$$

where $\lim_{M\to\infty} b_M = 0$, and C depends on d, λ_+ and λ_- from Lemma (G.1);

(b) the decoder fallback branch is inactive;

(c) and the reconstructed covariance has uniform spectral bounds, $\frac{\lambda_-}{2}\boldsymbol{I}_d \preceq \widehat{\boldsymbol{C}}(\boldsymbol{z}) \preceq 2\lambda_+\boldsymbol{I}_d$.

Proof. Bounding each term will be done as a function of the total slice error, which we define as $\varepsilon_\ell \coloneqq \|\,\phi_{\ell,\boldsymbol{z}}-\Phi(\mu^{\boldsymbol{s}_\ell})\,\|_{L^2}$ for $\ell \in [L_d]$. Let an exact population token be $\overline{\boldsymbol{z}} \;\coloneqq\; \overline{E}_M(\mu)$, where $\overline{\boldsymbol{z}} = [\overline{\boldsymbol{z}}_1^\top,\dots,\overline{\boldsymbol{z}}_{L_d}^\top]^\top$. By (146) and the exact-token identity following it,

$$\Pi_M\Phi(\mu^{\boldsymbol{s}_\ell}) = C_M^{-1}\left(\mathrm{proj}_M^{\uparrow}\left(\sqrt{L_d}\,\overline{\boldsymbol{z}}_\ell\right)\right).$$

Hence, using that C_M^{-1} is an isometry and the nonexpansiveness of Euclideanprojections onto closed convex sets, and the token $\boldsymbol{z}$ along with the quantile repair procedure,

$$\| \phi_{\ell,\boldsymbol{z}} - \Pi_M \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2} = \left\| \mathrm{proj}_M^{\uparrow}(\sqrt{L_d}\, \boldsymbol{z}_\ell) - \mathrm{proj}_M^{\uparrow}(\sqrt{L_d}\, \overline{\boldsymbol{z}}_\ell) \right\|_2 \leq \sqrt{L_d}\, \| \boldsymbol{z}_\ell - \overline{\boldsymbol{z}}_\ell \|_2 . \tag{191}$$

By the triangle inequality, for every slicing direction $\ell \in [L_d]$,

$$\begin{aligned} \varepsilon_\ell &\leq \| \phi_{\ell,\boldsymbol{z}} - \Pi_M \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2} + \| \Pi_M \Phi(\mu^{\boldsymbol{s}_\ell}) - \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2} \\ &\leq \sqrt{L_d}\, \| \boldsymbol{z}_\ell - \overline{\boldsymbol{z}}_\ell \|_2 + \| \Pi_M \Phi(\mu^{\boldsymbol{s}_\ell}) - \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2} , \end{aligned} \tag{192}$$

where the second inequality comes from (191). Define the worst case total slicing error as

$$b_M \coloneqq \sup_{\mu \in \mathcal{K}_{\mathrm{ell}}} \left(\frac{1}{L_d} \sum_{\ell=1}^{L_d} \| \Pi_M \Phi(\mu^{\boldsymbol{s}_\ell}) - \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2}^2 \right)^{1/2} \tag{193}$$

. Squaring both sides of (192), averaging both sides over L_d and taking square root, we get

$$\begin{aligned} \left(\frac{1}{L_d} \sum_{\ell=1}^{L_d} \varepsilon_\ell^2 \right)^{1/2} &\leq \left(\frac{1}{L_d} \sum_{\ell} \left(\sqrt{L_d} \| \boldsymbol{z}_\ell - \overline{\boldsymbol{z}}_\ell \|_2 + \| \Pi_M \Phi(\mu^{\boldsymbol{s}_\ell}) - \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2} \right)^2 \right)^{1/2} \\ &\leq \left(\sum_{\ell=1}^{L_d} \| \boldsymbol{z}_\ell - \overline{\boldsymbol{z}}_\ell \|_2^2 \right)^{1/2} + \left(\frac{1}{L_d} \sum_{\ell=1}^{L_d} \| \Pi_M \Phi(\mu^{\boldsymbol{s}_\ell}) - \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2}^2 \right)^{1/2} \\ &\leq \left\| \boldsymbol{z} - \overline{E}_M(\mu) \right\|_2 + b_M \\ &\leq \eta + b_M, \end{aligned} \tag{194}$$

where the second inequality comes from applying Minkowski's inequality in $\mathbb{R}^{L_d}$ and the third inequality comes from the definitions of $\boldsymbol{z}$, $\overline{\boldsymbol{z}}$ and the ℓ_2-norm, and (193). Define the RHS of (194) as $e_M \coloneqq \eta + b_M$.

Next we show that L^2 closeness of the quantiles implies closeness of the means. We first bound the error in the recovered directional means. By (148) and the step-function representation of $\phi_{\ell,\boldsymbol{z}}$,

$$m_{\boldsymbol{s}_\ell}(\boldsymbol{z}) = \frac{1}{M} \sum_{i=1}^{M} a_{\ell,i}(\boldsymbol{z}) = \int_0^1 \phi_{\ell,\boldsymbol{z}}(u)\, du.$$

On the other hand, applying (147) to the projected law $\mu^{\boldsymbol{s}_\ell}$ gives

$$\int_0^1 \Phi(\mu^{\boldsymbol{s}_\ell})(u)\, du = \mathbb{E}_{\mu^{\boldsymbol{s}_\ell}}[X] = \langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu \rangle.$$

Therefore, using Cauchy–Schwarz and $\| \mathbf{1} \|_{L^2} = 1$, we have

$$| m_{\boldsymbol{s}_\ell}(\boldsymbol{z}) - \langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu \rangle | = \left| \int_0^1 (\phi_{\ell,\boldsymbol{z}}(u) - \Phi(\mu^{\boldsymbol{s}_\ell})(u))\, du \right| \leq \| \phi_{\ell,\boldsymbol{z}} - \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2} = \varepsilon_\ell. \tag{195}$$

We use these directional moment bounds to control the reconstructed mean. By the definition of $\widehat{\boldsymbol{m}}(\boldsymbol{z})$ and (195),

$$\| \widehat{\boldsymbol{m}}(\boldsymbol{z}) - \boldsymbol{m}_\mu \|^2 = \sum_{i=1}^{d} | m_{\boldsymbol{e}_i}(\boldsymbol{z}) - \langle \boldsymbol{e}_i,\, \boldsymbol{m}_\mu \rangle |^2 \leq \sum_{\ell=1}^{L_d} \varepsilon_\ell^2 \leq L_d e_M^2, \tag{196}$$

where the last inequality follows from the definition of e_M following (194). Let constant $C_1 = \sqrt{L_d}$.

Turning to the covariance, we first control the recovered directional variances. By (148) and the step-function representation of $\phi_{\ell,\boldsymbol{z}}$, and by (147), respectively, we have

$$c_{\boldsymbol{s}_\ell}(\boldsymbol{z}) = \| \phi_{\ell,\boldsymbol{z}} - m_{\boldsymbol{s}_\ell}(\boldsymbol{z}) \mathbf{1} \|_{L^2} , \qquad \sqrt{\boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell} = \| \Phi(\mu^{\boldsymbol{s}_\ell}) - \langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu \rangle \mathbf{1} \|_{L^2} .$$

Hence, by the reverse triangle inequality,

$$\begin{aligned}\left| c_{\boldsymbol{s}_\ell}(\boldsymbol{z}) - \sqrt{\boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell} \right| &\le \| \phi_{\ell,\boldsymbol{z}} - m_{\boldsymbol{s}_\ell}(\boldsymbol{z})\mathbf{1} - \Phi(\mu^{\boldsymbol{s}_\ell}) + \langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu \rangle \mathbf{1} \|_{L^2} \\ &\le \| \, \phi_{\ell,\boldsymbol{z}} - \Phi(\mu^{\boldsymbol{s}_\ell}) \, \|_{L^2} + | \, m_{\boldsymbol{s}_\ell}(\boldsymbol{z}) - \langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu \rangle \, | \\ &\le 2\varepsilon_\ell, \end{aligned} \tag{197}$$

where the last inequality follows from (195). We go from standard deviations to variances using (197),

$$\begin{aligned} | \, c^2_{\boldsymbol{s}_\ell}(\boldsymbol{z}) - \boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell \, | &= \left| c_{\boldsymbol{s}_\ell}(\boldsymbol{z}) - \sqrt{\boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell} \right| \left(c_{\boldsymbol{s}_\ell}(\boldsymbol{z}) + \sqrt{\boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell} \right) \\ &\le 2\varepsilon_\ell \left(c_{\boldsymbol{s}_\ell}(\boldsymbol{z}) + \sqrt{\boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell} \right). \end{aligned} \tag{198}$$

By Lemma G.1, $\boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell \le \lambda_+$ over $\mathcal{K}_{\text{ell}}$. Moreover, (197) implies $c_{\boldsymbol{s}_\ell}(\boldsymbol{z}) \le \sqrt{\boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell} + 2\varepsilon_\ell \le \sqrt{\lambda_+} + 2\varepsilon_\ell$, and therefore,

$$| \, c^2_{\boldsymbol{s}_\ell}(\boldsymbol{z}) - \boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell \, | \le 4\varepsilon_\ell \left(\sqrt{\lambda_+} + \varepsilon_\ell \right). \tag{199}$$

We now use the directional moment bounds to control the reconstructed covariance. Define for each $\ell \in [L_d]$ the quantity $\delta_\ell \coloneqq c^2_{\boldsymbol{s}_\ell}(\boldsymbol{z}) - \boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell$. By (199), $\delta_\ell^2 \le 32\lambda_+ \varepsilon_\ell^2 + 32\varepsilon_\ell^4$ and summing over ℓ gives

$$\sum_{\ell=1}^{L_d} \delta_\ell^2 \le 32\lambda_+ \sum_{\ell=1}^{L_d} \varepsilon_\ell^2 + 32 \sum_{\ell=1}^{L_d} \varepsilon_\ell^4 \le 32\lambda_+ L_d e_M^2 + 32 L_d^2 e_M^4, \tag{200}$$

where we used $\sum_{\ell=1}^{L_d} \varepsilon_\ell^2 \le L_d e_M^2$ from (194), together with $\sum_{i=1}^n a_i^4 \le \left(\sum_{i=1}^n a_i^2 \right)^2$.

We now choose constants. Fix $\eta_0 < 1/2$. From Lemma B.1 we know there exists an M_1 such that $\forall M \ge M_1$, $b_M \le \eta_0$. Then, for every $\eta \le \eta_0$ and therefore $e_M = \eta + b_M \le 2\eta_0 < 1$. Then, the RHS of (200) is in $\mathcal{O}\left(e_M^2\right)$ and there exists a constant $C_2 = C_2(d, \lambda_+) < \infty$, such that

$$\left(\sum_{\ell=1}^{L_d} \delta_\ell^2 \right)^{1/2} \le C_2 e_M. \tag{201}$$

Note that C_2 is independent of M, μ, and $\boldsymbol{z}$.

We next transfer this bound to the reconstructed covariance matrix $\boldsymbol{C}^0(\boldsymbol{z})$ defined in (150). For the diagonal entries,

$$[\boldsymbol{C}^0(\boldsymbol{z}) - \boldsymbol{C}_\mu]_{ii} = \delta_{\boldsymbol{e}_i}.$$

For $i < j$, letting $\boldsymbol{s}_{ij} = (\boldsymbol{e}_i + \boldsymbol{e}_j)/\sqrt{2}$, the covariance reconstruction formula and the identity

$$\boldsymbol{s}_{ij}^\top \boldsymbol{C}_\mu \boldsymbol{s}_{ij} = \frac{1}{2} \left([\boldsymbol{C}_\mu]_{ii} + [\boldsymbol{C}_\mu]_{jj} + 2[\boldsymbol{C}_\mu]_{ij} \right)$$

give $[\boldsymbol{C}^0(\boldsymbol{z}) - \boldsymbol{C}_\mu]_{ij} = \delta_{\boldsymbol{s}_{ij}} - \frac{1}{2}\left(\delta_{\boldsymbol{e}_i} + \delta_{\boldsymbol{e}_j}\right)$. Therefore we have $\left| \, [\boldsymbol{C}^0(\boldsymbol{z}) - \boldsymbol{C}_\mu]_{ij} \, \right|^2 \le 3\left(\delta^2_{\boldsymbol{s}_{ij}} + \delta^2_{\boldsymbol{e}_i} + \delta^2_{\boldsymbol{e}_j} \right)$. Summing over all entries finally yields

$$\| \, \boldsymbol{C}^0(\boldsymbol{z}) - \boldsymbol{C}_\mu \, \|_F \le C_3(d) \left(\sum_{\ell=1}^{L_d} \delta_\ell^2 \right)^{1/2} \le C_3(d) C_2 e_M, \tag{202}$$

where $C_3(d) < \infty$ is a dimension dependent factor that comes out of summing over all matrix entries. To pass the bound back onto $\widehat{\boldsymbol{C}}(\boldsymbol{z})$ recall that it is the Euclidean projection of $\boldsymbol{C}^0(\boldsymbol{z})$ onto $\mathbb{S}^d_+$ by (151), that $\boldsymbol{C}_\mu \in \mathbb{S}^d_+$, and that the Euclidean projection is nonexpansive. We get

$$\left\| \, \widehat{\boldsymbol{C}}(\boldsymbol{z}) - \boldsymbol{C}_\mu \, \right\|_F \le \| \, \boldsymbol{C}^0(\boldsymbol{z}) - \boldsymbol{C}_\mu \, \|_F \le C_3(d) C_2 e_M. \tag{203}$$

Therefore, we have covered the first two terms of our target inequality.

At this point, we interrupt the proof of point (a) to prove point (c), since it is convenient. Using the operator norm inequality and (203), we have

$$\left\| \widehat{\boldsymbol{C}}(\boldsymbol{z}) - \boldsymbol{C}_\mu \right\|_{\mathrm{op}} \le \left\| \widehat{\boldsymbol{C}}(\boldsymbol{z}) - \boldsymbol{C}_\mu \right\|_F \le C_3(d) C_2 e_M .$$

Since $e_M \le 2\eta_0$ for every $\eta \in (0, \eta_0]$ and every $M \ge M_0$, we choose $\eta_0 > 0$ sufficiently small so that $2C_3(d)C_2\eta_0 \le \frac{\lambda_-}{2}$. It then follows that $\left\| \widehat{\boldsymbol{C}}(\boldsymbol{z}) - \boldsymbol{C}_\mu \right\|_{\mathrm{op}} \le \frac{\lambda_-}{2}$ and therefore

$$\boldsymbol{C}_\mu - \frac{\lambda_-}{2}\boldsymbol{I}_d \preceq \widehat{\boldsymbol{C}}(\boldsymbol{z}) \preceq \boldsymbol{C}_\mu + \frac{\lambda_-}{2}\boldsymbol{I}_d .$$

By Lemma G.1, we have that $\lambda_- \boldsymbol{I}_d \preceq \boldsymbol{C}_\mu \preceq \lambda_+ \boldsymbol{I}_d$, and therefore

$$\frac{\lambda_-}{2}\boldsymbol{I}_d \preceq \widehat{\boldsymbol{C}}(\boldsymbol{z}) \preceq \left(\lambda_+ + \frac{\lambda_-}{2}\right)\boldsymbol{I}_d \preceq 2\lambda_+ \boldsymbol{I}_d . \tag{204}$$

With this, we have proven pont (c). Note that at this stage, the choice of η_0 has been updated to $\eta_0 \le \min\left\{\frac{1}{2}, \frac{\lambda_-}{4C_3(d)C_2}\right\}$.

We continue with bounding the third term in the LHS of point (a). We prepare a few details beforehand: first recall that computing $\phi^0_{\boldsymbol{z}}$ in (153) we divide by $\beta(\boldsymbol{z}) = \sum_{\ell=1}^{L_d} \widehat{c}^{\,2}_\ell(\boldsymbol{z})$. Under the conditions of this lemma, $\beta(\boldsymbol{z})$ is positive by using $\widehat{c}_\ell(\boldsymbol{z}) = \sqrt{\boldsymbol{s}_\ell^\top \widehat{\boldsymbol{C}}(\boldsymbol{z}) \boldsymbol{s}_\ell}$, and $\| \boldsymbol{s}_\ell \| = 1$ in the lower bound of (204)

$$\widehat{c}^{\,2}_\ell(\boldsymbol{z}) = \boldsymbol{s}_\ell^\top \widehat{\boldsymbol{C}}(\boldsymbol{z}) \boldsymbol{s}_\ell \ge \frac{\lambda_-}{2}. \tag{205}$$

Therefore, $\beta(\boldsymbol{z}) \ge \frac{L_d \lambda_-}{2}$. Second, by (204), $\widehat{c}_\ell(\boldsymbol{z}) \le \sqrt{2\lambda_+}$. Third, define the true directional scale by $c_{\ell,\mu} \coloneqq \sqrt{\boldsymbol{s}_\ell^\top \boldsymbol{C}_\mu \boldsymbol{s}_\ell}$. Then, we have the following bound on the standard deviation error of the predicted $\widehat{c}_\ell(\boldsymbol{z})$,

$$\begin{aligned}
|\, \widehat{c}_\ell(\boldsymbol{z}) - c_{\ell,\mu} \,| &= \frac{\left| \widehat{c}^{\,2}_\ell(\boldsymbol{z}) - c^2_{\ell,\mu} \right|}{\widehat{c}_\ell(\boldsymbol{z}) + c_{\ell,\mu}} \\
&= \frac{\left| \boldsymbol{s}_\ell^\top \left(\widehat{\boldsymbol{C}}(\boldsymbol{z}) - \boldsymbol{C}_\mu \right) \boldsymbol{s}_\ell \right|}{\widehat{c}_\ell(\boldsymbol{z}) + c_{\ell,\mu}} \\
&\le \frac{\left\| \widehat{\boldsymbol{C}}(\boldsymbol{z}) - \boldsymbol{C}_\mu \right\|_{\mathrm{op}}}{\sqrt{\lambda_-}} \\
&\le \frac{C_3(d)C_2}{\sqrt{\lambda_-}} e_M ,
\end{aligned} \tag{206}$$

where we used via Lemma G.1 that $\widehat{c}_\ell(\boldsymbol{z}) + c_{\ell,\mu} \ge \sqrt{\lambda_-} + \frac{\sqrt{\lambda_-}}{2} \ge \sqrt{\lambda_-}$. Finally, for simplicity, write $\phi_\mu \coloneqq \phi_{R_\mu U_1}$. Recall from (16) that $\Phi(\mu^{\boldsymbol{s}_\ell}) = \langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu \rangle \mathbf{1} + c_{\ell,\mu}\phi_\mu$. Furthermore, the covariance normalization $\mathbb{E}[R_\mu^2] = d$ and $\mathbb{E}[U_1^2] = 1/d$ imply

$$\| \phi_\mu \|^2_{L^2} = \mathbb{E}[(R_\mu U_1)^2] = 1. \tag{207}$$

We're now ready to estimate the discrepancy between ϕ_μ and $\phi^0_{\boldsymbol{z}}$ defined in (153), which gives

$$\begin{aligned}
&\left\| \phi^0_{\boldsymbol{z}} - \phi_\mu \right\|_{L^2} \\
&= \left\| \frac{1}{\beta(\boldsymbol{z})} \sum_{\ell=1}^{L_d} \widehat{c}_\ell(\boldsymbol{z}) \Big[\phi_{\ell,\boldsymbol{z}} - \Phi(\mu^{\boldsymbol{s}_\ell}) + \langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu - \widehat{\boldsymbol{m}}(\boldsymbol{z}) \rangle \mathbf{1} + \big(c_{\ell,\mu} - \widehat{c}_\ell(\boldsymbol{z}) \big) \phi_\mu \Big] \right\|_{L^2}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{\beta(\boldsymbol{z})} \sum_{\ell=1}^{L_d} \widehat{c}_\ell(\boldsymbol{z}) \Bigg(\| \phi_{\ell,\boldsymbol{z}} - \Phi(\mu^{\boldsymbol{s}_\ell}) \|_{L^2} \\
&\qquad\qquad + |\langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu - \widehat{\boldsymbol{m}}(\boldsymbol{z}) \rangle| \, \| \mathbf{1} \|_{L^2} + |c_{\ell,\mu} - \widehat{c}_\ell(\boldsymbol{z})| \, \| \phi_\mu \|_{L^2} \Bigg) \\
&= \frac{1}{\beta(\boldsymbol{z})} \sum_{\ell=1}^{L_d} \widehat{c}_\ell(\boldsymbol{z}) \Big(\varepsilon_\ell + |\langle \boldsymbol{s}_\ell,\, \boldsymbol{m}_\mu - \widehat{\boldsymbol{m}}(\boldsymbol{z}) \rangle| + |c_{\ell,\mu} - \widehat{c}_\ell(\boldsymbol{z})| \Big) \\
&\leq \frac{1}{\beta(\boldsymbol{z})} \sum_{\ell=1}^{L_d} \widehat{c}_\ell(\boldsymbol{z}) \Big(\varepsilon_\ell + \|\widehat{\boldsymbol{m}} - \boldsymbol{m}_\mu\| + |c_{\ell,\mu} - \widehat{c}_\ell(\boldsymbol{z})| \Big) \\
&\leq \frac{2\sqrt{2\lambda_+}}{L_d \lambda_-} \left[\sum_{\ell=1}^{L_d} \varepsilon_\ell + L_d \| \widehat{\boldsymbol{m}} - \boldsymbol{m}_\mu \| + \sum_{\ell=1}^{L_d} |c_{\ell,\mu} - \widehat{c}_\ell| \right]
\end{aligned} \tag{208}
$$

where we applied the triangle inequality, and the final equality follows from definitions and the aforeproven properties and bounds. We upper bound the remaining quantities.

We know from before that $\left(\frac{1}{L_d}\sum_{\ell=1}^{L_d} \varepsilon_\ell^2\right)^{1/2} \leq e_M$. Cauchy–Schwarz then gives

$$
\sum_{\ell=1}^{L_d} \varepsilon_\ell \leq \sqrt{L_d} \left(\sum_{\ell=1}^{L_d} \varepsilon_\ell^2 \right)^{1/2} \leq L_d e_M.
$$

For the second term in the sum we use the bound (196).Finally, for the third term, we use (206) to get

$$
\sum_{\ell=1}^{L_d} |\, \widehat{c}_\ell - c_{\ell,\mu}\,| \leq L_d \frac{C_3(d) C_2}{\sqrt{\lambda_-}} e_M.
$$

Substituting everything gives

$$
\begin{aligned}
\| \phi^0_{\boldsymbol{z}} - \phi_\mu \|_{L^2} &\leq \frac{2\sqrt{2\lambda_+}}{L_d \lambda_-} \left[L_d e_M + L_d \sqrt{L_d}\, e_M + L_d \frac{C_3(d) C_2}{\sqrt{\lambda_-}} e_M \right] \\
&= \frac{2\sqrt{2\lambda_+}}{\lambda_-} \left(1 + \sqrt{L_d} + \frac{C_3(d) C_2}{\sqrt{\lambda_-}} \right) e_M,
\end{aligned}
$$

and for simplicity we define $C_4 \coloneqq \frac{2\sqrt{2\lambda_+}}{\lambda_-} \left(1 + \sqrt{L_d} + \frac{C_3(d) C_2}{\sqrt{\lambda_-}} \right)$.

Since $U_1 \overset{d}{=} -U_1$, the law of $R_\mu U_1$ is symmetric about zero, and therefore $\phi_\mu(u) = -\phi_\mu(1-u)$ for almost every $u \in (0,1)$. Thus the symmetrization in (154) leaves ϕ_μ unchanged. Using the triangle inequality and the fact that $u \mapsto 1-u$ preserves the L^2 norm,

$$
\| \phi^{\mathrm{sym}}_{\boldsymbol{z}} - \phi_\mu \|_{L^2} \leq \| \phi^0_{\boldsymbol{z}} - \phi_\mu \|_{L^2} \leq C_4 e_M. \tag{209}
$$

We now prove point (b), namely that the fallback branch is inactive. Since $e_M \leq 2\eta_0$, we may choose $\eta_0 > 0$ sufficiently small so that $2C_4\eta_0 \leq \frac{1}{2}$. Hence, using $\| \phi_\mu \|_{L^2} = 1$ from (207), the reverse triangle inequality and (209),

$$
\| \phi^{\mathrm{sym}}_{\boldsymbol{z}} \|_{L^2} \geq \| \phi_\mu \|_{L^2} - \| \phi^{\mathrm{sym}}_{\boldsymbol{z}} - \phi_\mu \|_{L^2} \geq 1 - C_4 e_M \geq \frac{1}{2}.
$$

Moreover, by (205), $\beta(\boldsymbol{z}) \geq \frac{L_d \lambda_-}{2}$. Since $\kappa_M \to 0$ by (172), we may increase M_0, if necessary, so that for every $M \geq M_0$,

$$
\kappa_M < \min\left\{ \frac{L_d \lambda_-}{2}, \frac{1}{2} \right\}.
$$

Therefore, for every $\eta \in (0, \eta_0]$ and every $M \geq M_0$, $\beta(\boldsymbol{z}) > \kappa_M$, $\| \phi^{\mathrm{sym}}_{\boldsymbol{z}} \|_{L^2} > \kappa_M$, and both conditions required for the non-fallback branch are satisfied.

Finally, using (155), (207), and the reverse triangle inequality,

$$\begin{aligned}\left\| \widehat{\phi}_{RU_1,\boldsymbol{z}} - \phi_\mu \right\|_{L^2} &\le \left\| \frac{\phi_{\boldsymbol{z}}^{\mathrm{sym}}}{\| \phi_{\boldsymbol{z}}^{\mathrm{sym}} \|_{L^2}} - \phi_{\boldsymbol{z}}^{\mathrm{sym}} \right\|_{L^2} + \| \phi_{\boldsymbol{z}}^{\mathrm{sym}} - \phi_\mu \|_{L^2} \\ &= | 1 - \| \phi_{\boldsymbol{z}}^{\mathrm{sym}} \|_{L^2} | + \| \phi_{\boldsymbol{z}}^{\mathrm{sym}} - \phi_\mu \|_{L^2} \\ &\le 2 \| \phi_{\boldsymbol{z}}^{\mathrm{sym}} - \phi_\mu \|_{L^2} \\ &\le C_5 e_M. \end{aligned} \tag{210}$$

Combining (196), (203), and (210), and enlarging the constant if necessary, gives

$$\| \widehat{\boldsymbol{m}}(\boldsymbol{z}) - \boldsymbol{m}_\mu \| + \left\| \widehat{\boldsymbol{C}}(\boldsymbol{z}) - \boldsymbol{C}_\mu \right\|_F + \left\| \widehat{\phi}_{RU_1,\boldsymbol{z}} - \phi_{R_\mu U_1} \right\|_{L^2} \le C e_M,$$

which proves (191). □

G.2 From marginal comparison to radial-law stability

The radial optimizer (163) is not minimizing W_2, but only differences of CDF values at finitely many thresholds. Consequently, for fixed M, two very different probability laws can still agree at those finitely many points. We therefore need a continuous-law distance sitting between the finite objective and the convergence of distributions. To this end, we introduce the quantity $\Delta : \mathcal{P}([0,\infty)) \times \mathcal{P}([0,\infty)) \to [0,\infty)$ given by

$$\Delta(\mu,\nu) \;:=\; \int_0^\infty \big(F_\mu(x) - F_\nu(x)\big)^2 \frac{2\,dx}{\pi(1+x^2)}, \tag{211}$$

where F_μ is the CDF of a measure μ. We next show this is a metric for measures supported on $[0,\infty)$.

Lemma G.3 (CDF distance and thresholds)**.** *The function Δ in (211) is a metric $\mathcal{P}([0,\infty))$ and induces weak convergence. Furthermore, for $\mu,\nu \in \mathcal{P}_2([0,\infty))$,*

$$\Delta(\mu,\nu)^2 \le \frac{2}{\pi} \mathrm{W}_2(\mu,\nu). \tag{212}$$

Finally, the error between the metric Δ and its discrete approximation evaluated at thresholds (173) is uniformly bounded for any fixed M,

$$\sup_{\mu,\nu\in\mathcal{P}([0,\infty))} \left| \Delta(\mu,\nu)^2 - \frac{1}{K_M} \sum_{k=1}^{K_M} \big(F_\mu(x_{k,M}) - F_\nu(x_{k,M})\big)^2 \right| \le \frac{4}{K_M}. \tag{213}$$

Proof. First, we show that Δ is a metric that metrizes weak convergence on $\mathcal{P}([0,\infty))$. From definition (211) of Δ, nonnegativity and symmetry are immediate. Moreover, since Δ is the weighted L^2 distance between the two CDFs, the L^2 triangle inequality gives, for every $\mu,\nu,\lambda \in \mathcal{P}([0,\infty))$,

$$\Delta(\mu,\lambda) \le \Delta(\mu,\nu) + \Delta(\nu,\lambda).$$

Finally, we need to verify that $\Delta(\mu,\nu) = 0$ implies $\mu = \nu$. Since the weight $x \longmapsto \frac{2}{\pi(1+x^2)} > 0$ on $[0,\infty)$, the equality $\Delta(\mu,\nu) = 0$ implies $F_\mu(x) = F_\nu(x)$ for Lebesgue-almost every $x \ge 0$. Since both CDFs are nondecreasing and right-continuous, equality almost everywhere implies equality everywhere. Hence $F_\mu = F_\nu$, and therefore $\mu = \nu$. Thus Δ is a metric.

We next show that this metric induces weak convergence. Suppose first that $\mu_n \xrightarrow{d} \mu$. By the CDF characterization of weak convergence, $F_{\mu_n}(x) \longrightarrow F_\mu(x)$ at every continuity point x of F_μ. A CDF has at most countably many discontinuities, so this convergence holds for Lebesgue-almost every $x \ge 0$. Moreover, since $F_\mu, F_\nu \in [-1,1]$,

$$0 \le \big(F_{\mu_n}(x) - F_\mu(x)\big)^2 \frac{2}{\pi(1+x^2)} \le \frac{2}{\pi(1+x^2)},$$

Since $\frac{2}{\pi(1+x^2)}$ is integrable, the dominated convergence theorem gives $\lim_{n\to\infty} \Delta(\mu_n,\mu) = 0$.

In reverse, suppose that $\lim_{n\to\infty} \Delta(\mu_n, \mu) = 0$. We show that $\lim_{n\to\infty} F_{\mu_n}(x) = F_\mu(x)$ at every continuity point $x > 0$ of F_μ. By contradiction, suppose the latter doesm't hold. Then, there exist an x, an $\varepsilon > 0$, and a subsequence, still denoted by μ_n, such that $| F_{\mu_n}(x) - F_\mu(x) | \geq \varepsilon$ for every n.

Consider first the case $F_{\mu_n}(x) \geq F_\mu(x) + \varepsilon$. Since F_μ is continuous at x, there exists $h > 0$ such that $F_\mu(x+h) - F_\mu(x) < \frac{\varepsilon}{2}$. For every $y \in [x, x+h]$, monotonicity gives

$$F_{\mu_n}(y) \geq F_{\mu_n}(x) \geq F_\mu(x) + \varepsilon,$$

while

$$F_\mu(y) \leq F_\mu(x+h) < F_\mu(x) + \frac{\varepsilon}{2}.$$

Subtracting, $F_{\mu_n}(y) - F_\mu(y) \geq \frac{\varepsilon}{2}$ for $y \in [x, x+h]$. It follows that, uniformly in n,

$$\Delta(\mu_n, \mu)^2 \geq \frac{\varepsilon^2}{4} \int_x^{x+h} \frac{2\,dy}{\pi(1+y^2)} > 0,$$

which contradicts $\Delta(\mu_n, \mu) \to 0$.

The case $F_{\mu_n}(x) \leq F_\mu(x) - \varepsilon$ is analogous. Therefore $F_{\mu_n}(x) \longrightarrow F_\mu(x)$ at every positive continuity point of F_μ.

If 0 is also a continuity point of F_μ, then $F_\mu(0) = 0$, since the measures are supported on $[0, \infty)$. For every positive continuity point x,

$$0 \leq F_{\mu_n}(0) \leq F_{\mu_n}(x).$$

Taking $n \to \infty$ and then letting positive continuity points $x \downarrow 0$ gives

$$F_{\mu_n}(0) \longrightarrow 0 = F_\mu(0).$$

Thus the CDFs converge at every continuity point of F_μ, and the CDF characterization of weak convergence yields $\mu_n \xrightarrow{d} \mu$ (Billingsley, 1999, Section 2).

We now prove the W_2 bound. Let $\mu, \nu \in \mathcal{P}_2([0,\infty))$, and let (X, Y) be any coupling of μ and ν. We want to bout the CDF difference by a function of the coupling, so that we can transition to a W_2 distance-based bound. Since $| F_\mu(x) - F_\nu(x) | \leq 1$, we have

$$\big(F_\mu(x) - F_\nu(x)\big)^2 \leq | F_\mu(x) - F_\nu(x) | = \big| \mathbb{E}\big[\mathbf{1}_{\{X \leq x\}} - \mathbf{1}_{\{Y \leq x\}}\big] \big| \leq \mathbb{E}\big[\big| \mathbf{1}_{\{X \leq x\}} - \mathbf{1}_{\{Y \leq x\}} \big|\big],$$

Furthermore, $\frac{1}{1+x^2} \leq 1$ on $[0, \infty)$, and using Tonelli's theorem to swap expectation and integration,

$$\begin{aligned}
\Delta(\mu, \nu)^2 &= \int_0^\infty \big(F_\mu(x) - F_\nu(x)\big)^2 \frac{2\,dx}{\pi(1+x^2)} \\
&\leq \frac{2}{\pi} \int_0^\infty | F_\mu(x) - F_\nu(x) |\, dx \\
&\leq \frac{2}{\pi} \int_0^\infty \mathbb{E}\big[\big| \mathbf{1}_{\{X \leq x\}} - \mathbf{1}_{\{Y \leq x\}} \big|\big]\, dx. \\
&\leq \frac{2}{\pi} \mathbb{E} \int_0^\infty \big| \mathbf{1}_{\{X \leq x\}} - \mathbf{1}_{\{Y \leq x\}} \big|\, dx.
\end{aligned}$$

For each realization of (X, Y), the integral argument is 1 only when x lies between the respective realized values and is 0 otherwise. Therefore

$$\int_0^\infty \big| \mathbf{1}_{\{X \leq x\}} - \mathbf{1}_{\{Y \leq x\}} \big|\, dx = | X - Y |,$$

and hence

$$\Delta(\mu, \nu)^2 \leq \frac{2}{\pi} \mathbb{E} | X - Y | \leq \frac{2}{\pi} \Big(\mathbb{E} | X - Y |^2 \Big)^{1/2},$$

where the final inequality is an application of Cauchy–Schwarz. Since the coupling (X, Y) was arbitrary, taking the infimum over all couplings yields

$$\Delta(\mu, \nu)^2 \leq \frac{2}{\pi} \mathrm{W}_2(\mu, \nu),$$

which proves our second point.

Finally, we prove the uniform quadrature bound. Our goal is to compare the continuous discrepancy $\Delta(\mu,\nu)^2$ with the finite CDF objective evaluated at the thresholds (173). We use a tangent change of variables to rewrite $\Delta(\mu,\nu)^2$ as an integral over $(0,1)$. Under this transformation, the thresholds $x_{k,M}$ correspond exactly to the midpoints of a uniform partition of $(0,1)$, like the one we use in (9). We can therefore interpret the finite CDF objective as a midpoint approximation of this integral and bound its error using the monotonicity of the two CDFs.

In the definition of Δ (211) we make the change of variables $x = \tan\left(\frac{\pi t}{2}\right)$, $t \in (0,1)$ and get

$$\Delta(\mu,\nu)^2 = \int_0^1 \left[F_\mu\left(\tan\frac{\pi t}{2}\right) - F_\nu\left(\tan\frac{\pi t}{2}\right)\right]^2 dt. \tag{214}$$

We partition the integration interval $[0,1]$ into the intervals $I_{k,M} = \left[\frac{k-1}{K_M}, \frac{k}{K_M}\right]$, $k \in [K_M]$, and denote the midpoint of $I_{k,M}$ by $t_{k,M} = \frac{k-\frac{1}{2}}{K_M}$. By the definition of the tangent thresholds in (173),

$$x_{k,M} = \tan\left(\frac{\pi t_{k,M}}{2}\right),$$

It is then easy to see that the quantity $\frac{1}{K_M}\sum_{k=1}^{K_M}\left(F_\mu(x_{k,M}) - F_\nu(x_{k,M})\right)^2$ is precisely the midpoint approximation of the integral defining $\Delta(\mu,\nu)^2$.

We now bound the midpoint approximation error. First, for $t \in [0,1)$ define

$$g_{\mu,\nu}(t) \;\coloneqq\; \left(F_\mu\left(\tan\frac{\pi t}{2}\right) - F_\nu\left(\tan\frac{\pi t}{2}\right)\right)^2, \tag{215}$$

Fix $k \in [K_M]$ and write $a_{k,M} \coloneqq \frac{k-1}{K_M}$, $b_{k,M} \coloneqq \frac{k}{K_M}$. For every $t \in I_{k,M}$

$$\begin{aligned}
|\, g_{\mu,\nu}(t) - g_{\mu,\nu}(t_{k,M})\,| &= \left|\left(F_\mu\left(\tan\frac{\pi t}{2}\right) - F_\nu\left(\tan\frac{\pi t}{2}\right)\right)^2 - \left(F_\mu\left(\tan\frac{\pi t_{k,M}}{2}\right) - F_\nu\left(\tan\frac{\pi t_{k,M}}{2}\right)\right)^2\right| \\
&\le 2\left|F_\mu\left(\tan\frac{\pi t}{2}\right) - F_\nu\left(\tan\frac{\pi t}{2}\right) - \left(F_\mu\left(\tan\frac{\pi t_{k,M}}{2}\right) - F_\nu\left(\tan\frac{\pi t_{k,M}}{2}\right)\right)\right| \\
&\le 2\left|F_\mu\left(\tan\frac{\pi t}{2}\right) - F_\mu\left(\tan\frac{\pi t_{k,M}}{2}\right)\right| + 2\left|F_\nu\left(\tan\frac{\pi t}{2}\right) - F_\nu\left(\tan\frac{\pi t_{k,M}}{2}\right)\right|,
\end{aligned}$$

where we used $\left|\,a^2 - b^2\,\right| = |\,a-b\,|\,|\,a+b\,| \le 2\,|\,a-b\,|$ for $a,b \in [-1,1]$. Since F_μ, F_ν are nondecreasing and the tangent function increases on $(0,\pi/2)$, and $t, t_{k,M} \in I_{k,M}$, we have that

$$\begin{cases}
\left|F_\mu\left(\tan\frac{\pi t}{2}\right) - F_\mu\left(\tan\frac{\pi t_{k,M}}{2}\right)\right| \le F_\mu\left(\tan\frac{\pi b_{k,M}}{2}\right) - F_\mu\left(\tan\frac{\pi a_{k,M}}{2}\right), \\
\left|F_\nu\left(\tan\frac{\pi t}{2}\right) - F_\nu\left(\tan\frac{\pi t_{k,M}}{2}\right)\right| \le F_\nu\left(\tan\frac{\pi b_{k,M}}{2}\right) - F_\nu\left(\tan\frac{\pi a_{k,M}}{2}\right).
\end{cases}$$

Consequently, in our target quantity for interval $I_{k,M}$,

$$\begin{aligned}
\left|\int_{I_{k,M}} g_{\mu,\nu}(t)\,dt - \frac{1}{K_M} g_{\mu,\nu}(t_{k,M})\right| &= \left|\int_{I_{k,M}} \left(g_{\mu,\nu}(t) - g_{\mu,\nu}(t_{k,M})\right) dt\right| \\
&\le \frac{2}{K_M}\left[F_\mu\left(\tan\frac{\pi b_{k,M}}{2}\right) - F_\mu\left(\tan\frac{\pi a_{k,M}}{2}\right)\right. \\
&\qquad \left. + F_\nu\left(\tan\frac{\pi b_{k,M}}{2}\right) - F_\nu\left(\tan\frac{\pi a_{k,M}}{2}\right)\right].
\end{aligned}$$

We now sum this estimate over all $k \in [K_M]$. The increments of each nondecreasing function telescope to $F_\mu\left(\tan\frac{\pi}{2}\right) - F_\mu(\tan(0)) \le 1$ and we get, via a final application of the triangle inequality

$$\left|\Delta(\mu,\nu)^2 - \frac{1}{K_M}\sum_{k=1}^{K_M}\left(F_\mu(x_{k,M}) - F_\nu(x_{k,M})\right)^2\right| \le \frac{2}{K_M}(1+1) = \frac{4}{K_M}.$$

Since the bound does not depend on μ, ν, it holds uniformly over $\mathcal{P}([0,\infty)])$. □

Lemma G.4 (Identification and stability of the radius)**.** *Let ν be a radial law such that, for $R \sim \nu$ and $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ with $\boldsymbol{u} \perp R$, we define the map $\mathcal{A}_d : \mathcal{P}([0,\infty)) \to \mathcal{P}([0,\infty))$ as*

$$\mathcal{A}_d\nu \;\coloneqq\; \mathrm{law}(R|U_1|),$$

where U_1 is the first coordinate of $\boldsymbol{u}$. The map $\mathcal{A}_d$ is weakly continuous and injective. Moreover, for $\nu, \nu' \in \mathcal{P}_2([0,\infty))$,

$$\mathrm{W}_2(\mathcal{A}_d\nu, \mathcal{A}_d\nu') \le d^{-1/2}\mathrm{W}_2(\nu,\nu'). \tag{216}$$

Proof. We start with the weak continuity of $\mathcal{A}_d$. Let $\nu_n, \nu \in \mathcal{P}([0,\infty))$ satisfy be such that $\nu_n \xrightarrow{d} \nu$. We want to show that $\mathcal{A}_d\nu_n \xrightarrow{d} \mathcal{A}_d\nu$, for which we use the the bounded-continuous-function characterization of weak convergence,

$$\mu_n \xrightarrow{d} \mu \quad\Longleftrightarrow\quad \int h\, d\mu_n \to \int h\, d\mu$$

for every $h \in C_b([0,\infty))$.

Consider $f : [0,\infty) \to \mathbb{R}$ with $f \in C_b([0,\infty))$, and define

$$g(r) \;\coloneqq\; \mathbb{E}[f(r|U_1|)]\,.$$

Since f is bounded, g is bounded. Moreover, if $r_n \to r$, then $r_n|U_1| \to r|U_1|$ almost surely, and thus $f(r_n|U_1|) \longrightarrow f(r|U_1|)$ almost surely. Since f is bounded, dominated convergence gives

$$g(r_n) \longrightarrow g(r).$$

Thus $g \in C_b([0,\infty))$.

Now, if $R_n \sim \nu_n$ and $R \sim \nu$, both independent of U_1, it holds that

$$\int f\, d(\mathcal{A}_d\nu_n) = \mathbb{E}[f(R_n|U_1|)] = \mathbb{E}[\mathbb{E}[f(R_n|U_1|) \mid R_n]] = \mathbb{E}[g(R_n)] = \int g(r)\, d\nu_n(r).$$

Since $\nu_n \xrightarrow{d} \nu$ and g is bounded and continuous,

$$\int g(r)\, d\nu_n(r) \longrightarrow \int g(r)\, d\nu(r) = \mathbb{E}[g(R)] = \mathbb{E}[\mathbb{E}[f(R|U_1|) \mid R]] = \mathbb{E}[f(R|U_1|)] = \int f\, d(\mathcal{A}_d\nu),$$

and we have our result.

We now show injectivity of $\mathcal{A}_d$, namely that $\mathcal{A}_d\nu = \mathcal{A}_d\nu' \implies \nu = \nu'$. Let $R \sim \nu$, $R' \sim \nu'$, independently of $\boldsymbol{u} \sim \mathrm{Unif}\mathbb{S}^{d-1}$.

By assumption, $R|U_1| \overset{d}{=} R'|U_1|$. This implies that $RU_1 \overset{d}{=} R'U_1$ since, for an independent Rademacher random variable ξ and a symmetric random variable X we know that $\xi \,|\, X \,| \overset{d}{=} X$ Vershynin (2019, Lemma 6.3.1) and therefore

$$RU_1 \overset{d}{=} \varepsilon R|U_1| \overset{d}{=} \varepsilon R'|U_1| \overset{d}{=} R'U_1.$$

Now notice that for a random variable $Y \coloneqq R\boldsymbol{u}$, with $\boldsymbol{u} \perp R$, $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ we have that $\| Y \| = R$ almost surely. We leverage this fact to arrive at $R \overset{d}{=} R'$.

First, construct $\boldsymbol{Y} \coloneqq R\boldsymbol{u}$, $\boldsymbol{Y'} \coloneqq R'\boldsymbol{u}$. Both laws are spherically symmetric, meaning that multiplication with an orthogonal matrix $\boldsymbol{Q}$ does not change their distribution. Therefore, for a vector $\boldsymbol{t} \in \mathbb{R}^d$ constructed as $\boldsymbol{t} = \| \boldsymbol{t} \| \boldsymbol{v}$ with some $\boldsymbol{v} \in \mathbb{S}^{d-1}$, and the orthogonal $\boldsymbol{Q}$ for which $\boldsymbol{Q}\boldsymbol{v} = \boldsymbol{e}_1$,

$$\langle \boldsymbol{t},\, \boldsymbol{Y} \rangle = \| \boldsymbol{t} \| \langle \boldsymbol{v},\, \boldsymbol{Y} \rangle = \| \boldsymbol{t} \| \langle \boldsymbol{Q}e_1,\, \boldsymbol{Y} \rangle = \| \boldsymbol{t} \| \langle e_1,\, \boldsymbol{Q}^\top \boldsymbol{Y} \rangle \overset{d}{=} \| \boldsymbol{t} \| \langle e_1,\, \boldsymbol{Y} \rangle = \| \boldsymbol{t} \| Y_1$$

Similarly, $\langle \boldsymbol{t},\, \boldsymbol{Y'} \rangle \overset{d}{=} \| \boldsymbol{t} \| \; Y_1'$. Since $Y_1 = RU_1$, $Y_1' = R'U_1$, and these two random variables have the same law from before, we have that $Y_1 \overset{d}{=} Y_1'$. Immediately, we have that for any $\boldsymbol{t} \in R^d$,

$$\langle \boldsymbol{t},\, \boldsymbol{Y} \rangle \overset{d}{=} \langle \boldsymbol{t},\, \boldsymbol{Y'} \rangle,$$

and by the Cramér–Wold theorem (Billingsley, 1995, Theorem 29.4) this implies that $\boldsymbol{Y} \overset{d}{=} \boldsymbol{Y}'$. Therefore, $\|\boldsymbol{Y}\| \overset{d}{=} \|\boldsymbol{Y}'\|$ and we have obtained that $R \overset{d}{=} R'$ and thus $\nu = \nu'$.

We now prove the forward Lipschitz estimate. Let $\nu, \nu' \in \mathcal{P}_2([0,\infty))$. We want to prove

$$\mathrm{W}_2(\mathcal{A}_d\nu, \mathcal{A}_d\nu') \le d^{-1/2}\mathrm{W}_2(\nu, \nu').$$

Take any coupling (R, R') of ν and ν', and let U_1 be independent of this coupling. Then $(R|U_1|, R'|U_1|)$ is a coupling of $\mathcal{A}_d\nu$ and $\mathcal{A}_d\nu'$. Therefore, by the definition of W_2,

$$\begin{aligned}\mathrm{W}_2^2(\mathcal{A}_d\nu, \mathcal{A}_d\nu') &\le \mathbb{E}\Big[\big(R|U_1| - R'|U_1|\big)^2\Big] \\ &= \mathbb{E}\big[(R-R')^2U_1^2\big]\,. \end{aligned} \tag{217}$$

Since U_1 is independent of (R, R'), and $\mathbb{E}[U_1^2] = \frac{1}{d}$, we have

$$\mathbb{E}\big[(R-R')^2U_1^2\big] = \mathbb{E}[(R-R')^2]\mathbb{E}[U_1^2] = \frac{1}{d}\mathbb{E}[(R-R')^2].$$

and therefore, $\mathrm{W}_2^2(\mathcal{A}_d\nu, \mathcal{A}_d\nu') \le \frac{1}{d}\mathbb{E}[(R-R')^2]$.

Since the coupling (R, R') was arbitrary, taking the infimum over all couplings on the RHS of (217) followed by square root yields

$$\mathrm{W}_2(\mathcal{A}_d\nu, \mathcal{A}_d\nu') \le d^{-1/2}\mathrm{W}_2(\nu, \nu'). \qquad \square$$

Lemma G.5 (Uniform inverse stability of the radial law)**.** *Let $\mathcal{A}_d$ be the absolute-marginal map introduced in Lemma G.4 and Δ be the metric in (211). Then, for every $\varepsilon > 0$, there exists $\delta > 0$ such that, for every $\nu \in \mathcal{V}_R$ and every $\nu' \in \mathcal{P}_2([0,\infty))$ satisfying*

$$\int r^2\, d\nu'(r) = d \qquad \text{and} \qquad \Delta(\mathcal{A}_d\nu', \mathcal{A}_d\nu) \le \delta,$$

we have $\mathrm{W}_2(\nu', \nu) \le \varepsilon$.

Proof. We argue by contradiction. The statement says that, uniformly over the target family $\mathcal{V}_R$, two radial laws must be close in W_2 whenever their absolute marginals are sufficiently close in Δ and both radial laws have the prescribed second moment.

Suppose that the conclusion fails. Then there exist $\varepsilon_0 > 0$ and sequences $\nu_n \in \mathcal{V}_R$, $\nu_n' \in \mathcal{P}_2([0,\infty))$ such that

$$\int r^2\, d\nu_n'(r) = d, \qquad \Delta(\mathcal{A}_d\nu_n', \mathcal{A}_d\nu_n) \longrightarrow 0, \tag{218}$$

while $\mathrm{W}_2(\nu_n', \nu_n) \ge \varepsilon_0$ for every n.

Since $\mathcal{V}_R$ is compact in W_2 by Lemma G.1, and $\nu_n \in \mathcal{V}_R$ for every n, there exist a subsequence $(\nu_{n_k})_k$ and some $\nu^\star \in \mathcal{V}_R$ such that

$$\mathrm{W}_2(\nu_{n_k}, \nu^\star) \longrightarrow 0.$$

By the Lipschitz property in Lemma G.4 and the W_2 bound on Δ in Lemma G.3, we have that also

$$\mathrm{W}_2(\mathcal{A}_d\nu_{n_k}, \mathcal{A}_d\nu^\star) \longrightarrow 0, \qquad \text{and} \qquad \Delta(\mathcal{A}_d\nu_{n_k}, \mathcal{A}_d\nu^\star) \longrightarrow 0,$$

respectively. Furthermore, by (218) we have that, along the same subsequence $\Delta(\mathcal{A}_d\nu_{n_k}', \mathcal{A}_d\nu_{n_k}) \longrightarrow 0$. Then, by the triangle inequality for Δ we have that

$$\Delta(\mathcal{A}_d\nu_{n_k}', \mathcal{A}_d\nu^\star) \le \Delta(\mathcal{A}_d\nu_{n_k}', \mathcal{A}_d\nu_{n_k}) + \Delta(\mathcal{A}_d\nu_{n_k}, \mathcal{A}_d\nu^\star) \longrightarrow 0. \tag{219}$$

Since Δ metrizes weak convergence by Lemma G.3, we obtain

$$\mathcal{A}_d\nu_{n_k}' \xrightarrow{d} \mathcal{A}_d\nu^\star. \tag{220}$$

We now show that this leads to $\mathrm{W}_2(\nu'_{n_k}, \nu^\star) \longrightarrow 0$. By the moment condition we assumed, Markov's inequality gives, for every $B > 0$,

$$\nu'_{n_k}((B,\infty)) \le \frac{1}{B^2}\int r^2\, d\nu'_{n_k}(r) = \frac{d}{B^2}.$$

Hence, for every $\varepsilon > 0$, choosing B large enough that $d/B^2 < \varepsilon$ yields $\nu'_{n_k}([0,B]) \ge 1-\varepsilon$ for every k. Therefore $(\nu'_{n_k})_k$ is tight, and by Prokhorov's theorem, (Billingsley, 1999, Theorem 5.1), it is relatively compact for weak convergence.

Let $\widetilde{\nu}$ be any weak subsequential limit of $(\nu'_{n_k})_k$. Let $\widetilde{\nu}$ be any weak subsequential limit of $(\nu'_{n_k})_k$, so that for some further subsequence $(\nu'_{n_{k_j}})_j$, $\nu'_{n_{k_j}} \xrightarrow{d} \widetilde{\nu}$. By the weak continuity of $\mathcal{A}_d$ from Lemma G.4, $\mathcal{A}_d\nu'_{n_{k_j}} \xrightarrow{d} \mathcal{A}_d\widetilde{\nu}$. On the other hand, (220) implies along the same further subsequence that $\mathcal{A}_d\nu'_{n_{k_j}} \xrightarrow{d} \mathcal{A}_d\nu^\star$. By uniqueness of weak limits,

$$\mathcal{A}_d\widetilde{\nu} = \mathcal{A}_d\nu^\star,$$

and, since $\mathcal{A}_d$ is injective by Lemma G.4, we obtain $\widetilde{\nu} = \nu^\star$. Thus $\nu^\star$ is the unique weak subsequential limit, and therefore

$$\nu'_{n_k} \xrightarrow{d} \nu^\star.$$

Finally, $\int r^2\, d\nu'_{n_k}(r) = \int r^2\, d\nu^\star(r) = d$, so weak convergence together with convergence of the second moments implies

$$\mathrm{W}_2(\nu'_{n_k}, \nu^\star) \longrightarrow 0$$

by the standard characterization of W_2 convergence (Ambrosio et al., 2005, Remark 7.1.11).

Since also $\mathrm{W}_2(\nu_{n_k}, \nu^\star) \longrightarrow 0$, the triangle inequality yields

$$\mathrm{W}_2(\nu'_{n_k}, \nu_{n_k}) \le \mathrm{W}_2(\nu'_{n_k}, \nu^\star) + \mathrm{W}_2(\nu^\star, \nu_{n_k}) \longrightarrow 0,$$

contradicting $\mathrm{W}_2(\nu'_n, \nu_n) \ge \varepsilon_0$. □

G.3 Approximation by radial grids and stability of the fit

The radial reconstruction step optimizes only over laws supported on the finite grid $\{r_{j,M}\}_{j=1}^{J_M}$ and satisfying the exact second-moment constraint. Before analyzing the fitted law, we therefore need to verify that this feasible class is rich enough to approximate every target radial law in $\mathcal{V}_R$.

Lemma G.6 (Approximation with the prescribed second moment)**.** *Let $\boldsymbol{w} \in \mathcal{W}_M$ of (157) and define $\nu_{R,\boldsymbol{w}} := \sum_{j=1}^{J_M} w_j \delta_{r_{j,M}}$. Further, define the uniform radial-grid approximation error*

$$a_M := \sup_{\nu\in\mathcal{V}_R}\inf_{\boldsymbol{w}\in\mathcal{W}_M} \mathrm{W}_2(\nu, \nu_{R,\boldsymbol{w}}). \tag{221}$$

Recall the radial grid $0 = r_{1,M} < \cdots < r_{J_M,M}$, and define $B_M := r_{J_M,M}$ and let its maximal spacing by $h_M := \max_{1\le j<J_M}(r_{j+1,M} - r_{j,M})$. If $\lim_{M\to\infty} B_M = \infty$ and $\lim_{M\to\infty} h_M = 0$, then $\lim_{M\to\infty} a_M = 0$.

Proof. Fix $\nu \in \mathcal{V}_R$ and let $R \sim \nu$. We construct a feasible grid-supported law that approximates ν uniformly over $\mathcal{V}_R$.

Define $\underline{R}_M$ be obtained by truncating R at B_M and rounding downward to the nearest grid point,

$$\underline{R}_M := \sum_{j=1}^{J_M-1} r_{j,M}\,\mathbf{1}_{\{r_{j,M}\le R<r_{j+1,M}\}} + B_M\,\mathbf{1}_{\{R\ge B_M\}}.$$

Then $\underline{R}_M$ is grid-valued, $0 \le \underline{R}_M \le R$ almost surely, and

$$(R-\underline{R}_M)^2 \le h_M^2\mathbf{1}_{\{R\le B_M\}} + R^2\mathbf{1}_{\{R>B_M\}} \quad \text{almost surely.}$$

Hence

$$\mathbb{E}(R-\underline{R}_M)^2 \le h_M^2 + \int_{(B_M,\infty)} r^2\, d\nu(r). \tag{222}$$

Since $\mathcal{V}_R$ is compact in W_2, its squared radii are uniformly integrable (Ambrosio et al., 2005, Proposition 7.1.5). Therefore, using $h_M \to 0$ and $B_M \to \infty$,

$$\sup_{\nu \in \mathcal{V}_R} \mathbb{E}(R - \underline{R}_M)^2 \longrightarrow 0. \tag{223}$$

In particular, the coupling $(R, \underline{R}_M)$ together with (223) gives

$$\sup_{\nu \in \mathcal{V}_R} \mathrm{W}_2(\nu, \mathrm{law}(\underline{R}_M)) \longrightarrow 0. \tag{224}$$

The law $\mathrm{law}(\underline{R}_M)$ is supported on the radial grid, but it need not satisfy the prescribed second moment exactly. To obtain a feasible law in $\mathcal{W}_M$, we must therefore restore the constraint $\mathbb{E}[\underline{R}_M^2] = d$. Write

$$m_{M,\nu} \;\coloneqq\; \mathbb{E}[\underline{R}_M^2], \qquad \epsilon_{M,\nu} \;\coloneqq\; d - m_{M,\nu}.$$

Since $\underline{R}_M \le R$ almost surely and $\mathbb{E}[R^2] = d$, we have $\epsilon_{M,\nu} \ge 0$ almost surely. Moreover,

$$\epsilon_{M,\nu} = \mathbb{E}[(R - \underline{R}_M)(R + \underline{R}_M)] \le \sqrt{\mathbb{E}(R - \underline{R}_M)^2}\sqrt{\mathbb{E}(R + \underline{R}_M)^2}' \le 2\sqrt{d}\,\sqrt{\mathbb{E}(R - \underline{R}_M)^2}, \tag{225}$$

where the first inequality is by Cauchy–Schwarz, and the second comes from $R + \underline{R}_M \le 2R$ almost surely along with R's moment constraint. Combining (225) with (223) we get

$$\sup_{\nu \in \mathcal{V}_R} \epsilon_{M,\nu} \longrightarrow 0.$$

However, the constraint set $\mathcal{W}_M$ imposes the moment constraint exaclty. To ensure this, me must make a correction to $\mathrm{law}(\underline{R}_M)$ in the form of a new law $\nu_M^\star$. For all sufficiently large M, it holds that $B_M^2 > d$. Define

$$\alpha_{M,\nu} \;\coloneqq\; \frac{\epsilon_{M,\nu}}{B_M^2 - m_{M,\nu}}, \qquad \nu_M^\star \;\coloneqq\; (1 - \alpha_{M,\nu})\,\mathrm{law}(\underline{R}_M) + \alpha_{M,\nu}\delta_{B_M}.$$

Since $m_{M,\nu} \le d < B_M^2$, we have $0 \le \alpha_{M,\nu} \le 1$, and

$$\int r^2 \, d\nu_M^\star(r) = m_{M,\nu} + \alpha_{M,\nu}(B_M^2 - m_{M,\nu}) = d.$$

Thus $\nu_M^\star = \nu_{R,\boldsymbol{w}^\star}$ for some $\boldsymbol{w}^\star \in \mathcal{W}_M$. Coupling $\mathrm{law}(\underline{R}_M)$ with $\nu_M^\star$ by moving an $\alpha_{M,\nu}$-fraction of the mass to B_M yields

$$\mathrm{W}_2^2(\mathrm{law}(\underline{R}_M), \nu_M^\star) \le \alpha_{M,\nu} B_M^2 \le \epsilon_{M,\nu} \frac{B_M^2}{B_M^2 - d}. \tag{226}$$

Since $B_M \to \infty$ and $\sup_{\nu \in \mathcal{V}_R} \epsilon_{M,\nu} \to 0$, the right-hand side vanishes uniformly over $\mathcal{V}_R$. Therefore, by (224), (226) and the triangle inequality,

$$\sup_{\nu \in \mathcal{V}_R} \mathrm{W}_2(\nu, \nu_M^\star) \longrightarrow 0.$$

Since $\nu_M^\star$ is supported on the radial grid and has second moment d, there exists a weight vector $\boldsymbol{w}_M^\star \in \mathcal{W}_M$ such that $\nu_M^\star = \nu_{R,\boldsymbol{w}_M^\star}$. Therefore, for every $\nu \in \mathcal{V}_R$,

$$\inf_{\boldsymbol{w} \in \mathcal{W}_M} \mathrm{W}_2(\nu, \nu_{R,\boldsymbol{w}}) \le \mathrm{W}_2(\nu, \nu_{R,\boldsymbol{w}_M^\star}) = \mathrm{W}_2(\nu, \nu_M^\star).$$

Taking the supremum over $\nu \in \mathcal{V}_R$ yields

$$a_M \le \sup_{\nu \in \mathcal{V}_R} \mathrm{W}_2(\nu, \nu_M^\star).$$

Since the RHS converges to zero, we conclude that $a_M \longrightarrow 0$. □

Proposition G.1 (Stability of the radial fit). *For each M, let $\nu_M \in \mathcal{V}_R$ be a target radial law and let $V \sim \mathrm{Unif}(0,1)$. Define*

$$\delta_M \;:=\; \mathrm{W}_2\Big(\mathrm{law}\Big(\Big|\widehat{\phi}_{RU_1,\boldsymbol{z}}(V)\Big|\Big)\,,\mathcal{A}_d\nu_M\Big)\,.$$

Let $\tau_M \geq 0$ be the objective approximation error for solving problem (163) with regularization ρ_M, and recall a_M from Lemma G.6 and the number of CDF threshold points K_M. If

$$a_M \longrightarrow 0, \qquad \delta_M \longrightarrow 0, \qquad \rho_M \longrightarrow 0, \qquad \tau_M \longrightarrow 0, \qquad K_M \longrightarrow \infty,$$

then $\mathrm{W}_2(\widehat{\nu}_R, \nu_M) \longrightarrow 0$.

Proof. We show that the absolute marginal generated by the fitted radial law converges in Δ to the true absolute marginal $\mathcal{A}_d\nu_M$. The conclusion will then follow directly from the uniform inverse-stability result of Lemma G.5.

For brevity, set $\psi_M \;:=\; \mathrm{law}\Big(\Big|\widehat{\phi}_{RU_1,\boldsymbol{z}}(V)\Big|\Big)$ where $V \sim \mathrm{Unif}(0,1)$, so that $\delta_M = \mathrm{W}_2(\psi_M, \mathcal{A}_d\nu_M)$. Also define, for probability measures $\mu, \nu \in \mathcal{P}([0,\infty))$,

$$L_M(\mu,\nu) \;:=\; \frac{1}{K_M}\sum_{k=1}^{K_M}\left(F_\mu(x_{k,M}) - F_\nu(x_{k,M})\right)^2.$$

By the definitions of $\boldsymbol{A}_M$ and the recovered CDF vector in problem (163), its objective can be written

$$J_{M,\boldsymbol{z}}(\boldsymbol{w}) \;:=\; L_M(\mathcal{A}_d\nu_{R,\boldsymbol{w}}, \psi_M) + \rho_M \,\|\,\boldsymbol{w}\,\|_2^2\,.$$

For each M, consider the target $\nu_M \in \mathcal{V}_R$. By the definition of a_M and the compactness of $\mathcal{W}_M$, there exists $\boldsymbol{w}_M^\star \in \mathcal{W}_M$ such that

$$\mathrm{W}_2(\nu_M, \nu_{R,\boldsymbol{w}_M^\star}) \leq a_M. \tag{227}$$

Since τ_M bounds the optimization error, it holds that $J_{M,\boldsymbol{z}}(\widehat{\boldsymbol{w}}_M) \leq J_{M,\boldsymbol{z}}(\boldsymbol{w}_M^\star) + \tau_M$. Therefore,

$$L_M(\mathcal{A}_d\widehat{\nu}_R, \psi_M) + \rho_M \,\|\,\widehat{\boldsymbol{w}}_M\,\|_2^2 \leq L_M(\mathcal{A}_d\nu_M^\star, \psi_M) + \rho_M \,\|\,\boldsymbol{w}_M^\star\,\|_2^2 + \tau_M. \tag{228}$$

Since $\boldsymbol{w}_M^\star$ is a simplex vector, $\|\,\boldsymbol{w}_M^\star\,\|_2^2 \leq 1$. Dropping the nonnegative term involving $\widehat{\boldsymbol{w}}_M$ gives

$$L_M(\mathcal{A}_d\widehat{\nu}_R, \psi_M) \leq L_M(\mathcal{A}_d\nu_M^\star, \psi_M) + \rho_M + \tau_M. \tag{229}$$

We now use the quadrature estimate of Lemma G.3 to replace the finite CDF discrepancy by Δ. Applying Lemma G.3 on both sides of (229) yields

$$\begin{aligned}
\Delta(\mathcal{A}_d\widehat{\nu}_R, \psi_M)^2 &\leq L_M(\mathcal{A}_d\widehat{\nu}_R, \psi_M) + \frac{4}{K_M} \\
&\leq L_M(\mathcal{A}_d\nu_M^\star, \psi_M) + \rho_M + \tau_M + \frac{4}{K_M} \\
&\leq \Delta(\mathcal{A}_d\nu_M^\star, \psi_M)^2 + \rho_M + \tau_M + \frac{8}{K_M}.
\end{aligned} \tag{230}$$

We're left to bound the first term on the RHS of (230). By the triangle inequality for W_2, the Lipschitz property of $\mathcal{A}_d$ from Lemma G.4, and (227),

$$\mathrm{W}_2(\mathcal{A}_d\nu_M^\star, \psi_M) \leq \mathrm{W}_2(\mathcal{A}_d\nu_M^\star, \mathcal{A}_d\nu_M) + \mathrm{W}_2(\mathcal{A}_d\nu_M, \psi_M) \leq \frac{1}{\sqrt{d}}\mathrm{W}_2(\nu_M^\star, \nu_M) + \delta_M \leq \frac{a_M}{\sqrt{d}} + \delta_M.$$

Hence, by the W_2 bound from Lemma G.3,

$$\Delta(\mathcal{A}_d\nu_M^\star, \psi_M)^2 \leq \frac{2}{\pi}\left(\frac{a_M}{\sqrt{d}} + \delta_M\right).$$

Substituting this into (230) gives

$$\Delta(\mathcal{A}_d\widehat{\nu}_R, \psi_M)^2 \leq \frac{2}{\pi}\left(\frac{a_M}{\sqrt{d}} + \delta_M\right) + \rho_M + \tau_M + \frac{8}{K_M}. \tag{231}$$

By our assumptions of the proposition, the RHS goes to zero, and therefore $\Delta(\mathcal{A}_d\widehat{\nu}_R, \psi_M) \longrightarrow 0$. Moreover, again by by the W_2 bound from Lemma G.3, $\Delta(\psi_M, \mathcal{A}_d\nu_M)^2 \leq \frac{2}{\pi}\delta_M \longrightarrow 0$. The triangle inequality for Δ thus gives

$$\Delta(\mathcal{A}_d\widehat{\nu}_R, \mathcal{A}_d\nu_M) \leq \Delta(\mathcal{A}_d\widehat{\nu}_R, \psi_M) + \Delta(\psi_M, \mathcal{A}_d\nu_M) \longrightarrow 0.$$

Finally, $\widehat{\nu}_R$ has second moment exactly d because $\widehat{\boldsymbol{w}}_M \in \mathcal{W}_M$, while $\nu_M \in \mathcal{V}_R$ also has second moment d. Hence Lemma G.5 applies and yields $\mathrm{W}_2(\widehat{\nu}_R, \nu_M) \longrightarrow 0$. □

G.4 Particle approximation

Lemma G.7 (Stability and consistency of particle reconstruction)**.** *Let $\mathcal{J}_M = \{J_{s,M}\}_{s=1}^{R_M}$ be the radial quantile partition from (166), and let $\Pi_{\mathcal{J}_M}$ denote the corresponding cell-average projection. Define the radial law quantile approximation error, and two angular approximation errors, as follows.*

$$\begin{cases} \vartheta_M \;:=\; \sup_{\nu\in\mathcal{V}_R} \| \Phi(\nu) - \Pi_{\mathcal{J}_M}\Phi(\nu) \|_{L^2}\,, \\ \varepsilon_k \;:=\; \mathrm{W}_2\Big(\frac{1}{k}\sum_{i=1}^{k}\delta_{\boldsymbol{u}_i}, \sigma_d\Big)\,, \\ \gamma_M \;:=\; \max_{1\le s\le R_M} \varepsilon_{k_{s,M}}\,, \end{cases}$$

where $\sigma_d \equiv \mathrm{Unif}(\mathbb{S}^{d-1})$. Then,

(a) $\lim_{M\to\infty}\vartheta_M = 0$ and $\lim_{M\to\infty}\gamma_M = 0$;

(b) for every $\boldsymbol{m}\in\mathbb{R}^d$, every $\boldsymbol{C}\succeq 0$, and every $\nu,\nu'\in\mathcal{P}_2([0,\infty))$,

$$\mathrm{W}_2(\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu),\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu')) \le \big\|\boldsymbol{C}^{1/2}\big\|_{\mathrm{op}}\,\mathrm{W}_2(\nu,\nu');$$

(c) for every $\nu\in\mathcal{V}_R$, if $R\sim\nu$ and $\boldsymbol{u}\sim\mathbb{S}^{d-1}$ are independent, then

$$\mathrm{W}_2\Big(\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu), \mathrm{law}(\boldsymbol{m}+\boldsymbol{C}^{1/2}R\boldsymbol{u})\Big) \le \big\|\boldsymbol{C}^{1/2}\big\|_{\mathrm{op}}\Big(\vartheta_M + \sqrt{d}\,\gamma_M\Big)\,.$$

Proof. First, point (a). By the balanced allocation in (166),

$$\max_{1\le s\le R_M} p_{s,M} \le \frac{1}{R_M} + \frac{1}{M} \longrightarrow 0,$$

where we used $R_M\to\infty$. Thus the mesh size of the partition $\mathcal{J}_M$ tends to zero, satisfying the hypothesis of Lemma B.1. Define the family of radial quantile functions $\mathcal{Q}_R := \{\Phi(\nu) : \nu\in\mathcal{V}_R\}\subset L^2$. Via Lemma G.1 point (c), $\mathcal{Q}_R$ is compact in L^2, since $\mathcal{V}_R$ is W_2-compact and, on the real line, the quantile isometry holds (8). Therefore, via Lemma B.1 we have

$$\vartheta_M = \sup_{\nu\in\mathcal{V}_R} \| \Phi(\nu) - \Pi_{\mathcal{J}_M}\Phi(\nu) \|_{L^2} \longrightarrow 0.$$

For the angular error γ_M, the balanced allocation also gives $\min_{1\le s\le R_M} k_{s,M} = \big\lfloor \frac{M}{R_M} \big\rfloor \longrightarrow \infty$, because $R_M/M\to 0$. Since $\varepsilon_k\to 0$ by (164),

$$\gamma_M = \max_{1\le s\le R_M} \varepsilon_{k_{s,M}} \le \sup_{k\ge\lfloor M/R_M\rfloor} \varepsilon_k \longrightarrow 0.$$

We now prove point (b). Consider two radial laws $\nu,\nu'\in\mathcal{P}_2([0,\infty))$. By (169), their reconstructions have particles

$$\boldsymbol{m} + r_{s,M}(\nu)\boldsymbol{C}^{1/2}\boldsymbol{u}_i \qquad\text{and}\qquad \boldsymbol{m} + r_{s,M}(\nu')\boldsymbol{C}^{1/2}\boldsymbol{u}_i,$$

respectively, for the same pairs (s,i). We construct a coupling by matching particles with the same radial-group index s and the same direction index i. Since every particle has mass $1/M$, the squared cost of this coupling is

$$\begin{aligned} \frac{1}{M}\sum_{s=1}^{R_M}\sum_{i=1}^{k_{s,M}} \Big\| \big(r_{s,M}(\nu) - r_{s,M}(\nu')\big)\boldsymbol{C}^{1/2}\boldsymbol{u}_i \Big\|^2 &\le \big\|\boldsymbol{C}^{1/2}\big\|_{\mathrm{op}}^2 \sum_{s=1}^{R_M}\frac{k_{s,M}}{M}\big(r_{s,M}(\nu) - r_{s,M}(\nu')\big)^2 \\ &= \big\|\boldsymbol{C}^{1/2}\big\|_{\mathrm{op}}^2 \sum_{s=1}^{R_M} p_{s,M}\big(r_{s,M}(\nu) - r_{s,M}(\nu')\big)^2, \end{aligned} \tag{232}$$

where the last equality follows from (166) where $p_{s,M} = k_{s,M}/M$. By definition (167), $r_{s,M}(\nu)$ and $r_{s,M}(\nu')$ are the cell averages of $\Phi(\nu)$ and $\Phi(\nu')$, respectively. Hence

$$\begin{aligned}\sum_{s=1}^{R_M} p_{s,M}\big(r_{s,M}(\nu) - r_{s,M}(\nu')\big)^2 &= \| \Pi_{\mathcal{J}_M}\Phi(\nu) - \Pi_{\mathcal{J}_M}\Phi(\nu') \|_{L^2}^2 \\ &\le \| \Phi(\nu) - \Phi(\nu') \|_{L^2} \\ &= \mathrm{W}_2(\nu,\nu') \end{aligned} \tag{233}$$

where the last lines comes from the nonexpansiveness of orthogonal projections $\Pi_{\mathcal{J}_M}$ and the quantile isometry (8). Piecing everything together, the following proves point (b) after taking square roots.

$$\mathrm{W}_2^2(\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu),\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu')) \le \left\| \boldsymbol{C}^{1/2} \right\|_{\mathrm{op}}^2 \mathrm{W}_2^2(\nu,\nu'),$$

Finally, we prove point (c), following the same deconstruction step using an intermediate law, in which only the radius has been discretized. Specifically, we decompose th edistance between $\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu)$ and $\mathrm{law}(\boldsymbol{m} + \boldsymbol{C}^{1/2}R\boldsymbol{u})$ via $\mathrm{law}(\boldsymbol{m} + \boldsymbol{C}^{1/2}\overline{R}_M\boldsymbol{u})$, where the first comparison controls the error from discretizing the spherical direction, while the second controls the error from replacing the true radius by its cell-average approximation.

Using the mainstay triangle inequality, we have

$$\begin{aligned}&\mathrm{W}_2\Big(\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu), \mathrm{law}(\boldsymbol{m} + \boldsymbol{C}^{1/2}R\boldsymbol{u})\Big) \\ &\qquad \le \mathrm{W}_2\Big(\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu), \mathrm{law}(\boldsymbol{m} + \boldsymbol{C}^{1/2}\overline{R}_M\boldsymbol{u})\Big) \\ &\qquad\quad + \mathrm{W}_2\Big(\mathrm{law}(\boldsymbol{m} + \boldsymbol{C}^{1/2}\overline{R}_M\boldsymbol{u}), \mathrm{law}(\boldsymbol{m} + \boldsymbol{C}^{1/2}R\boldsymbol{u})\Big). \end{aligned} \tag{234}$$

We bound these two terms separately.

Let $V \sim \mathrm{Unif}(0,1)$ and set $R = \Phi(\nu)(V)$ (as per (147)). Recall from (166) that the intervals $J_{s,M}$ form a partition of $(0,1]$ with $|\,J_{s,M}\,| = p_{s,M}$, and from (167) that $r_{s,M}(\nu)$ is the average of $\Phi(\nu)$ over $J_{s,M}$. Accordingly, define the cell-averaged radius

$$\overline{R}_M \;\coloneqq\; \Pi_{\mathcal{J}_M}\Phi(\nu)(V).$$

Equivalently, on the event $V \in J_{s,M}$, $\overline{R}_M = r_{s,M}(\nu)$. Thus $\overline{R}_M$ takes the representative radius $r_{s,M}(\nu)$ with probability

$$\mathbb{P}(V \in J_{s,M}) = |\,J_{s,M}\,| = p_{s,M}. \tag{235}$$

In particular, its radial masses agree exactly with the group masses used in the renderer (169).

We first bound the second term on the RHS of (234) using coupling of the shared V and the same independent $\boldsymbol{u} \sim \sigma_d$. This gives

$$\begin{aligned}\mathrm{W}_2^2\Big(\mathrm{law}(\boldsymbol{m} + \boldsymbol{C}^{1/2}\overline{R}_M\boldsymbol{u}), \mathrm{law}(\boldsymbol{m} + \boldsymbol{C}^{1/2}R\boldsymbol{u})\Big) &\le \mathbb{E}\left\| \boldsymbol{C}^{1/2}(\overline{R}_M - R)\boldsymbol{u} \right\|^2 \\ &\le \left\| \boldsymbol{C}^{1/2} \right\|_{\mathrm{op}}^2 \mathbb{E}\,|\,\overline{R}_M - R\,|^2 \\ &= \left\| \boldsymbol{C}^{1/2} \right\|_{\mathrm{op}}^2 \| \Pi_{\mathcal{J}_M}\Phi(\nu) - \Phi(\nu) \|_{L^2}^2 \\ &\le \left\| \boldsymbol{C}^{1/2} \right\|_{\mathrm{op}}^2 \vartheta_M^2, \end{aligned} \tag{236}$$

where the second inequality comes from independence of V and $\boldsymbol{u}$ and the fact that $\| \boldsymbol{u} \| = 1$ almost surely. The rest come from definitions.

We now bound the first term of (234). Fix $s \in [R_M]$ and define $\mu_{s,M}^{\mathrm{ang}} \coloneqq \frac{1}{k_{s,M}} \sum_{i=1}^{k_{s,M}} \delta_{\boldsymbol{u}_i}$. By the definition of $\varepsilon_{k_{s,M}}$ in the lemma body,

$$\varepsilon_{k_{s,M}} = \mathrm{W}_2\Big(\mu_{s,M}^{\mathrm{ang}}, \sigma_d\Big).$$

Since both measures are supported on the compact set $\mathbb{S}^{d-1}$ and the cost $(\boldsymbol{u},\boldsymbol{v}) \mapsto \| \boldsymbol{u}-\boldsymbol{v} \|^2$ is continuous, there exists an optimal coupling $\pi_{s,M} \in \Pi\Big(\sigma_d, \mu^{\text{ang}}_{s,M}\Big)$ such that

$$\int_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \| \boldsymbol{u}-\boldsymbol{v} \|^2 \, d\pi_{s,M}(\boldsymbol{u},\boldsymbol{v}) = \varepsilon^2_{k_{s,M}}.$$

For this fixed group s, define a map $\mathcal{M}_{s,M}(\boldsymbol{u}) \coloneqq \boldsymbol{m}+r_{s,M}(\nu)\boldsymbol{C}^{1/2}\boldsymbol{u}$. Then $(\mathcal{M}_{s,M},\, \mathcal{M}_{s,M})_{\#}\pi_{s,M}$ is a coupling of

$$\text{law}\Big(\boldsymbol{m}+r_{s,M}(\nu)\boldsymbol{C}^{1/2}\boldsymbol{u}\Big), \boldsymbol{u}\sim\sigma_d, \qquad \text{and} \qquad \frac{1}{k_{s,M}}\sum_{i=1}^{k_{s,M}} \delta_{\boldsymbol{m}+r_{s,M}(\nu)\boldsymbol{C}^{1/2}\boldsymbol{u}_i}. \tag{237}$$

Its squared transport cost is

$$\begin{aligned}
\int \| \mathcal{M}_{s,M}(\boldsymbol{u}) - \mathcal{M}_{s,M}(\boldsymbol{v}) \|^2 \, d\pi_{s,M}(\boldsymbol{u},\boldsymbol{v}) &= r_{s,M}(\nu)^2 \int \Big\| \boldsymbol{C}^{1/2}(\boldsymbol{u}-\boldsymbol{v}) \Big\|^2 \, d\pi_{s,M}(\boldsymbol{u},\boldsymbol{v}) \\
&\le r_{s,M}(\nu)^2 \Big\| \boldsymbol{C}^{1/2} \Big\|^2_{\text{op}} \int \| \boldsymbol{u}-\boldsymbol{v} \|^2 \, d\pi_{s,M}(\boldsymbol{u},\boldsymbol{v}) \\
&= r_{s,M}(\nu)^2 \Big\| \boldsymbol{C}^{1/2} \Big\|^2_{\text{op}} \varepsilon^2_{k_{s,M}}.
\end{aligned} \tag{238}$$

Recall from (235) that $\mathbb{P}(\overline{R}_M = r_{s,M}(\nu)) = p_{s,M}$. Hence the intermediate law can be written as the mixture

$$\text{law}(\boldsymbol{m}+\boldsymbol{C}^{1/2}\overline{R}_M\boldsymbol{u}) = \sum_{s=1}^{R_M} p_{s,M}\, \text{law}\Big(\boldsymbol{m}+r_{s,M}(\nu)\boldsymbol{C}^{1/2}\boldsymbol{u}\Big).$$

Moreover, by (169) and $p_{s,M} = k_{s,M}/M$,

$$\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu) = \frac{1}{M}\sum_{s=1}^{R_M}\sum_{i=1}^{k_{s,M}} \delta_{\boldsymbol{m}+r_{s,M}(\nu)\boldsymbol{C}^{1/2}\boldsymbol{u}_i} = \sum_{s=1}^{R_M} p_{s,M} \left(\frac{1}{k_{s,M}}\sum_{i=1}^{k_{s,M}} \delta_{\boldsymbol{m}+r_{s,M}(\nu)\boldsymbol{C}^{1/2}\boldsymbol{u}_i} \right). \tag{239}$$

Therefore, mixing the groupwise couplings $(\mathcal{M}_{s,M},\mathcal{M}_{s,M})_{\#}\pi_{s,M}$ with weights $p_{s,M}$ gives an admissible coupling between $\text{law}(\boldsymbol{m}+\boldsymbol{C}^{1/2}\overline{R}_M\boldsymbol{u})$ and $\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu)$. Using (238),

$$\begin{aligned}
\text{W}_2^2\Big(\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu), \text{law}(\boldsymbol{m}+\boldsymbol{C}^{1/2}\overline{R}_M\boldsymbol{u})\Big) &\le \Big\| \boldsymbol{C}^{1/2} \Big\|^2_{\text{op}} \sum_{s=1}^{R_M} p_{s,M} r_{s,M}(\nu)^2 \varepsilon^2_{k_{s,M}} \\
&\le \Big\| \boldsymbol{C}^{1/2} \Big\|^2_{\text{op}} \gamma_M^2 \sum_{s=1}^{R_M} p_{s,M} r_{s,M}(\nu)^2.
\end{aligned} \tag{240}$$

Since $r_{s,M}(\nu)$ is the average of $\Phi(\nu)$ on $J_{s,M}$, Jensen's inequality gives

$$\sum_{s=1}^{R_M} p_{s,M} r_{s,M}(\nu)^2 \le \int_0^1 \Phi(\nu)(v)^2 \, dv = \int_0^\infty r^2 \, d\nu(r) = d.$$

Consequently,

$$\text{W}_2\Big(\mathcal{R}_M(\boldsymbol{m},\boldsymbol{C},\nu), \text{law}(\boldsymbol{m}+\boldsymbol{C}^{1/2}\overline{R}_M\boldsymbol{u})\Big) \le \sqrt{d}\, \Big\| \boldsymbol{C}^{1/2} \Big\|_{\text{op}} \gamma_M. \tag{241}$$

Substituting (241) and (236) into (234) gives point (c). □

Table 2: Transformer architecture used in the experiments.

Component	Choice
Attention heads	4
Feed-forward dimension	512
Normalization	Pre-LayerNorm with final LayerNorm
Activation	GELU
Positional encoding	Fixed sinusoidal
Attention mask	Causal
Dropout	0
Input and output maps	Learned linear projections

H Additional experiments and experimental details

H.1 Common protocols and settings across experiments

Computational infrastructure. Most experiments were implemented in Python 3.12.10 and executed on a MacBook Pro equipped with an Apple M4 Pro SoC, a 12-core CPU, a 16-core GPU, and 24 GB of unified memory, using the PyTorch MPS backend. In the multivariate case, hyperparameter tuning and model training were carried out on a single NVIDIA L4 GPU with 24 GB of GPU memory, using PyTorch 2.13.0 and CUDA 13.0 in FP32 precision with TF32 disabled.

Architecture of standard transformer module. We use a causal pre-normalized transformer operating on the sequence of measure tokens, with the encoder and decoder varying with the experiment. Given hidden states $\boldsymbol{h}^{(\ell-1)}$, each block computes

$$\widetilde{\boldsymbol{h}}^{(\ell)} = \boldsymbol{h}^{(\ell-1)} + \mathrm{MHA}\Big(\mathrm{LN}(\boldsymbol{h}^{(\ell-1)})\Big),$$
$$\boldsymbol{h}^{(\ell)} = \widetilde{\boldsymbol{h}}^{(\ell)} + \mathrm{MLP}\Big(\mathrm{LN}(\widetilde{\boldsymbol{h}}^{(\ell)})\Big),$$

where self-attention is causally masked. A linear map embeds each input token, fixed sinusoidal positional encodings are added, and a final layer normalization and linear map produce the next-token prediction. The fixed architecture choices are summarized in Table 2 and hyperparameter tuning is described for each experiment separately.

Training and checkpoint selection. All transformers are trained to minimize $\mathcal{L}(\theta) = \frac{1}{T}\sum_{t=0}^{T-1} \| \widehat{\boldsymbol{z}}_{t+1} - \boldsymbol{z}_{t+1} \|^2$ for their respective token types, with target $\boldsymbol{z}_{t+1}$ and prediction $\widehat{\boldsymbol{z}}_{t+1}$.

During training, checkpoints are selected by the model's native token loss on a fixed validation set: quantile-token loss for the quantile Transformer and moment-token loss for the Gaussian moment Transformer. The best checkpoint is reloaded for all reported test metrics and figures. Neither decoded Wasserstein error nor test performance is used for selection; encoding is applied before each forward pass, while decoding is reserved for evaluation and visualization. Thus, neither decoding nor optimal transport is required during optimization or validation. Optimization hyperparameters not subject to tuning are reported in Table 3.

Hyperparameter tuning protocol. For each experiment, hyperparameters are tuned separately and per task for every learned model on a fixed pilot configuration (i.e., fixed particle count, seed, data generation regime, and train/validation data streams). All models and all hyperparameter candidates are trained on exactly the same sequence of sampled empirical measures. Training data are generated online, but every candidate uses an identically seeded, model-independent data-generation stream. All candidates are evaluated on the same fixed validation set, generated once from a separate random-number stream.

We evaluate the full search-space Cartesian product on the fixed pilot configuration and select by best validation MSE in the model's native representation space (e.g., quantile-token, moment-token error). The selected configuration is then fixed across all particle counts, seeds, and dynamical regimes in that experiment. Deterministic baselines require no tuning, and scaling-only models are not tuned unless stated otherwise. Each experiment specifies its pilot configuration, tuned models, search space, and tuning results.

Table 3: Optimization parameters used for the learned models.

Parameter	Value
Optimizer	AdamW
Peak learning rate	10^{-3}
Adam coefficients	$(\beta_1, \beta_2) = (0.9, 0.999)$
Adam ϵ	10^{-8}
Weight decay	10^{-4}
Gradient-norm threshold	1
Learning-rate schedule	500-step linear warmup, then cosine decay
Minimum learning rate	10^{-4}
Optimization steps	5000
Batch size	1024
Numerical precision	FP32
Dropout	0
Validation frequency	Every 500 steps, including step 0
Validation checkpoint criterion	Token-space loss

Table 4: Hyperparameter search space and selected values for each learned prediction model.

Hyperparameter	Candidate values	Best for Quantile Transformer	Best for Moment Transformer
Peak learning rate	$\{10^{-3}, 3 \times 10^{-4}\}$	10^{-3}	10^{-3}
Hidden dimension	$\{64, 128\}$	128	128
Transformer blocks	$\{2, 3\}$	3	3

Validation and test sets and protocol. For each experiment e, condition $c \in \mathcal{C}_e$, particle count $M \in \mathcal{M}_e$, and seed $s \in \mathcal{S}_e$, we generate $4,096$ evaluation trajectories using a random-number stream that is independent of the online training-data stream. The first $2,048$ trajectories form the validation set, and the remaining $2,048$ form the test set. The validation and test sets are therefore disjoint. Each set is processed internally as two batches of $1,024$ trajectories. Therefore, for fixed (e, c, M, s), all learned models, baselines, and oracles use the same validation and test trajectories.

The test set is unused for training, hyperparameter tuning, or checkpoint selection and is evaluated once after checkpoint selection to compute the reported empirical-measure and population-level metrics. Each metric is first averaged over test trajectories and scored prediction times within a run, then aggregated across seeds by reporting the mean and sample standard deviation.

H.2 Additional details and experiments for the LDS setting in Section 5

Dynamical system parameters. The latent state space is p-dimensional, with $p = 10$, and the observations are scalar. We consider the sequence of Gaussian marginals induced by the linear dynamical system

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_{t+1}, \qquad y_t = \boldsymbol{c}^\top \boldsymbol{x}_t + v_t, \tag{242}$$

where $\boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{Q})$, $v_t \sim \mathcal{N}(0, R)$, $\boldsymbol{Q} = 0.03^2 \boldsymbol{I}_p$, $R = 0.10^2$, and $\boldsymbol{c} = \frac{1}{\sqrt{p}} \boldsymbol{1}_p$ (every latent coordinate contributes equally to the scalar observation). Starting from $\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{P}_0)$, we propagate the population parameters analytically:

$$\boldsymbol{m}_{t+1} = \boldsymbol{A}\boldsymbol{m}_t, \qquad \boldsymbol{P}_{t+1} = \boldsymbol{A}\boldsymbol{P}_t\boldsymbol{A}^\top + \boldsymbol{Q}. \tag{243}$$

The corresponding scalar observation law is $\mu_t = \mathcal{N}\left(\boldsymbol{c}^\top \boldsymbol{m}_t,\ \boldsymbol{c}^\top \boldsymbol{P}_t \boldsymbol{c} + R\right)$. We do not simulate a realized latent-state trajectory, but rather draw a fresh empirical point cloud at every time step $y_{t,1}, \ldots, y_{t,M} \overset{\text{iid}}{\sim} \mu_t$, independently across time. Thus, particles have no persistent identities; temporal dependence occurs at the level of the evolving population distributions.

We consider three hidden state transition regimes. Let $\boldsymbol{A}_{\text{st}} \in \mathbb{R}^{p \times p}$ be the tridiagonal matrix satisfying $[\boldsymbol{A}_{\text{st}}]_{i,i} = 0.80$, $[\boldsymbol{A}_{\text{st}}]_{i,i+1} = 0.12$, $[\boldsymbol{A}_{\text{st}}]_{i+1,i} = -0.08$, with all remaining entries equal

to zero. Since the spectral radius $\rho(\boldsymbol{A}_{\mathrm{st}}) \approx 0.8218$, we scale $\boldsymbol{A}_{\mathrm{st}}$ to obtain transition matrices for the marginally stable and unstable regimes, respectively

$$\boldsymbol{A}_{\mathrm{ms}} \coloneqq \frac{\boldsymbol{A}_{\mathrm{st}}}{\rho(\boldsymbol{A}_{\mathrm{st}})}, \qquad \boldsymbol{A}_{\mathrm{un}} \coloneqq \frac{1.1\,\boldsymbol{A}_{\mathrm{st}}}{\rho(\boldsymbol{A}_{\mathrm{st}})}.$$

The resulting spectral radii are $\rho(\boldsymbol{A}_{\mathrm{st}}) \approx 0.8218,\ \rho(\boldsymbol{A}_{\mathrm{ms}}) = 1,\ \rho(\boldsymbol{A}_{\mathrm{un}}) = 1.1.$

For every trajectory, the initial population parameters are sampled independently as $\boldsymbol{m}_0 \sim \mathrm{Unif}([-2,2]^p),\ \boldsymbol{p}_0 \sim \mathrm{Unif}([0.1^2, 0.6^2]^p)$, with $\boldsymbol{P}_0 = \mathrm{Diag}(\boldsymbol{p}_0)$. The population parameters then evolve analytically according to (17).

Each sequence contains $T+1$ population measures $(\mu_0, \ldots, \mu_T)$, yielding T causal next-measure predictions. At every time step, M particles are sampled independently from μ_t. We use $M \in \{32, 128, 512\}$ and seeds $s \in \{0,1,2\}$. Apart from $\boldsymbol{A}$, all data-generation, model, and optimization parameters are held fixed across the three regimes, and every learned model is trained separately in each regime.

Data generation. The experiment tests the generalization to new trajectories of a single dynamical system. Each trajectory consists of point clouds sampled from μ_t independently across time yielding $\widehat{\mu}_{t,M}$. We consider three different seeds $s \in \{0,1,2\}$, three levels of sampling $M \in \{32, 128, 512\}$ and three different stability regimes depending on the spectral radius of $\boldsymbol{A}$: stable ($\boldsymbol{A}_{\mathrm{st}}$), marginally stable ($\boldsymbol{A}_{\mathrm{ms}}$), and unstable ($\boldsymbol{A}_{\mathrm{un}}$) (see Appendix H.2 for details). Validation and test sets are fixed for every $(M, \boldsymbol{A}, s)$ tuple, comprising 2048 independent empirical measure sequences each.

Validation and test sets. Here, $\mathcal{C}_e = \{\boldsymbol{A}_{\mathrm{st}}, \boldsymbol{A}_{\mathrm{ms}}, \boldsymbol{A}_{\mathrm{un}}\}$, $\mathcal{M}_e = \{32, 128, 512\}$, $\mathcal{S}_e = \{0,1,2\}$. Each trajectory contains the empirical measure sequence $(\widehat{\mu}_{0,M}, \ldots, \widehat{\mu}_{T,M})$ with $T = 15$ and therefore provides T causal input–target pairs.

Hyperparameter tuning. The fixed pilot configuration on which hyperparameters are tuned is $(s = 0, M = 128, \boldsymbol{A} = \boldsymbol{A}_{\mathrm{st}})$. We search for the best combination of maximal learning rate, transformer hidden dimension and number of transformer blocks. The search space combinations and final selections are listed in Table 4. Tuning results are reported in Figure 3 and the two final columns of Tbale 4.

Evaluation. Recall the equal-mass cells $I_{i,M}$ from (9) and the empirical decoder D_M from (12). The Gaussian moment baseline predicts a mean $\widehat{m}$ and a log standard deviation $\widehat{\ell}$. The following proposition describes how this output is mapped to the same empirical-measure space as the quantile Transformer.

Proposition H.1 (Decoding a Gaussian quantile token)**.** *Let $m \in \mathbb{R}$, $\sigma > 0$, and let $\phi_0 \coloneqq \Phi(\mathcal{N}(0,1))$ denote the standard-normal quantile function. Let $u_{i,M} \coloneqq (i - \frac{1}{2})/M$ be the midpoint of $I_{i,M}$ for $i \in [M]$. Define the Gaussian midpoint quantile token $\boldsymbol{q}_M(m,\sigma) \in \mathbb{R}^M$ by*

$$[\boldsymbol{q}_M(m,\sigma)]_i \coloneqq \frac{1}{\sqrt{M}} \Phi\big(\mathcal{N}(m,\sigma^2)\big)\,(u_{i,M}) = \frac{1}{\sqrt{M}}\,(m + \sigma\phi_0(u_{i,M}))\,. \tag{244}$$

Then $D_M(\boldsymbol{q}_M(m,\sigma)) = \frac{1}{M}\sum_{i=1}^{M} \delta_{\Phi(\mathcal{N}(m,\sigma^2))(u_{i,M})}$. Moreover, $\boldsymbol{q}_M(m,\sigma) \in \mathbb{R}^M_{\uparrow}$, so the decoder acts isometrically on such tokens by (13).

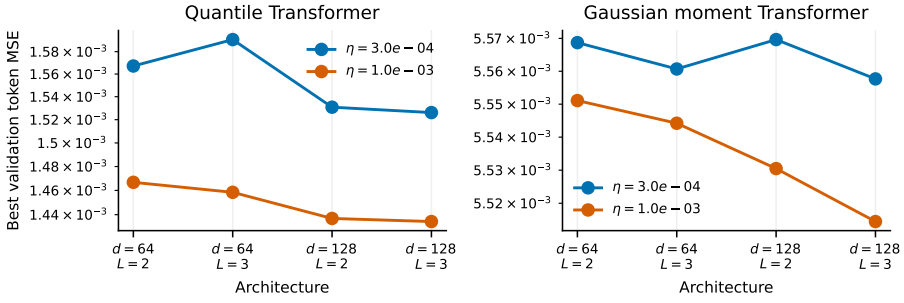

Figure 3: Hyperparameter tuning for Gaussian LDS experiments for the quantile and moment transformers.

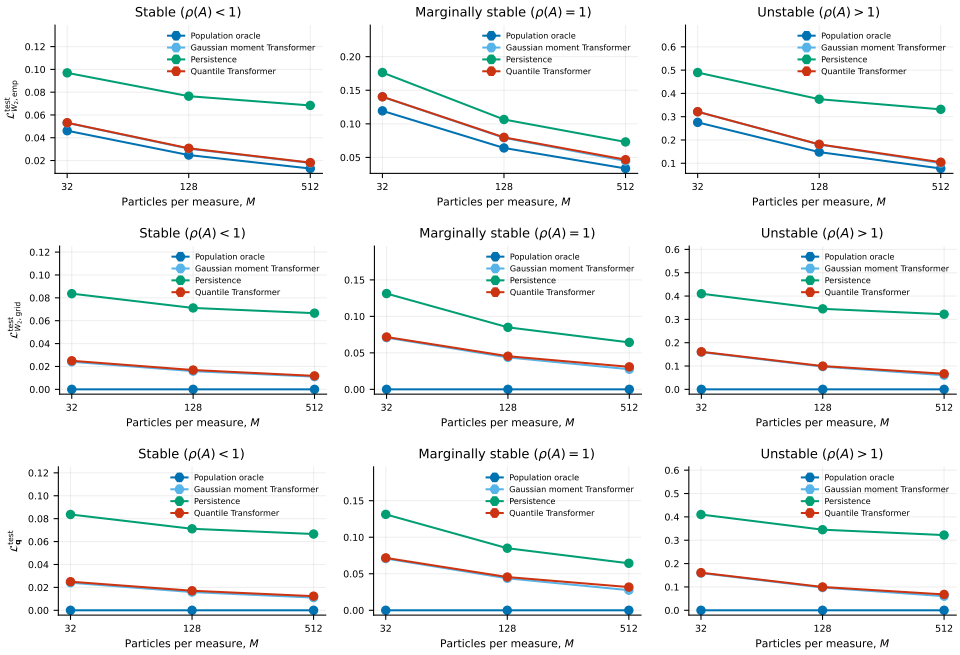

Figure 4: Three-seed average of test set accuracy at three levels of sampling $M \in \{32, 128, 512\}$.

Proof. The first identity follows from the Gaussian location–scale formula $\Phi(\mathcal{N}(m, \sigma^2)) = m + \sigma\phi_0$. Applying the definition of D_M in (12) gives the decoder form. Since ϕ_0 is nondecreasing and the $u_{i,M}$ are increasing, $\boldsymbol{q}_M(m, \sigma) \in \mathbb{R}^M_\uparrow$; the final claim follows from (13). □

For each $(M, \boldsymbol{A}, s)$, we average over $N_{\text{test}} = 2048$ empirical sequences. Define $\boldsymbol{q}^{i,\star}_{t,M} := \boldsymbol{q}_M(\boldsymbol{c}^\top \boldsymbol{m}^i_t, \boldsymbol{c}^\top \boldsymbol{P}^i_t \boldsymbol{c} + \sigma_v^2)$, $\widehat{\mu}^{i,\star}_{t,M} := D_M(\boldsymbol{q}^{i,\star}_{t,M})$. Performance is measured as the seed average of $\mathcal{L}^{\text{test}}_{W_2,\text{emp}} := \frac{1}{TN_{\text{test}}} \sum_{i,t} W_2(\widehat{\mu}^{i,\text{pred}}_{t,M}, \widehat{\mu}^i_{t,M})$, $\mathcal{L}^{\text{test}}_{W_2,\text{grid}} := \frac{1}{TN_{\text{test}}} \sum_{i,t} W_2(\widehat{\mu}^{i,\text{pred}}_{t,M}, \widehat{\mu}^{i,\star}_{t,M})$, $\mathcal{L}^{\text{test}}_{\boldsymbol{q}} := \frac{1}{TN_{\text{test}}} \sum_{i,t} \left\| \boldsymbol{q}^{i,\text{pred}}_{t,M} - \boldsymbol{q}^{i,\star}_{t,M} \right\|_2$, where $i \in [N_{\text{test}}]$ and $t \in [T]$, and

$$(\widehat{\mu}^{i,\text{pred}}_{t,M}, \boldsymbol{q}^{i,\text{pred}}_{t,M}) = \begin{cases} (D_M(\widehat{\boldsymbol{z}}^i_{t,M}), \widehat{\boldsymbol{z}}^i_{t,M}), & \text{Quantile transformer,} \\ (D_M(\boldsymbol{q}_M(\widehat{m}^i_{t,M}, \widehat{\sigma}^i_{t,M})), \boldsymbol{q}_M(\widehat{m}^i_{t,M}, \widehat{\sigma}^i_{t,M})), & \text{Gaussian moment transformer,} \\ (\widehat{\mu}^{i,\star}_{t,M}, \boldsymbol{q}^{i,\star}_{t,M}), & \text{Analytic oracle,} \\ (\widehat{\mu}^i_{t-1,M}, \boldsymbol{z}^i_{t-1,M}), & \text{Persistence.} \end{cases}$$

Thus, the baseline does not sample from its predicted Gaussian. It returns an equally weighted, M-atom midpoint approximation of that law. The midpoint levels are induced by the same $1/M$ partition used by the empirical quantile representation and avoid the endpoints of $(0, 1)$, where Gaussian quantiles diverge. Figure 4 shows the model-agnostic Quantile transformer performing almost indistinguishably from the model-aware Gaussian moment transformer across all metrics.

Training convergence. Training is done over 5000 steps. Training plots are given in Figures 6 and 7. Both learned models converge rapidly and remain stable, with validation loss closely tracking training loss and showing no evidence of overfitting. Performance improves with M, since larger particle sets provide less noisy representations. Higher residual errors can be seen in the unstable regime due to amplification by the expanding dynamics. The quantile Transformer benefits more visibly from continued optimization, whereas the Gaussian moment Transformer reaches its plateau earlier.

Geometry diagnostic. For equally weighted one-dimensional empirical measures, the monotone coupling is W_2-optimal, so the Euclidean distance between normalized quantile tokens equals the W_2-distance between their decoded measures. Figure 5 checks this identity, thereby verifying the encoder–decoder implementation and confirming that token-space errors have the intended Wasserstein interpretation.

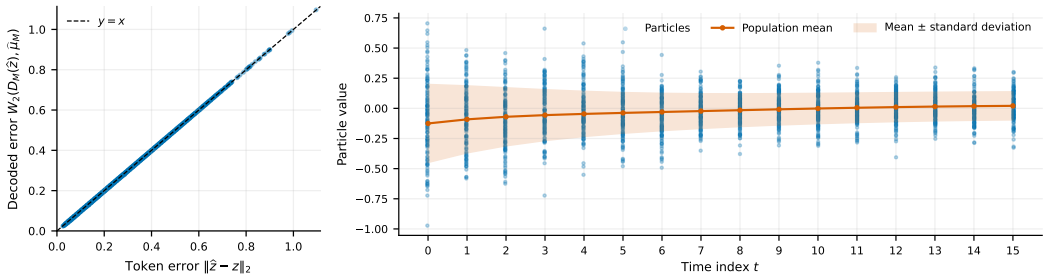

Figure 5: Left: ratio between W_2 distance of the decoded predicted quantile vs. the empirical measure, and the ℓ_2 distance between the predicted quantile token and the target quantile token. Right: point cloud dynamics.

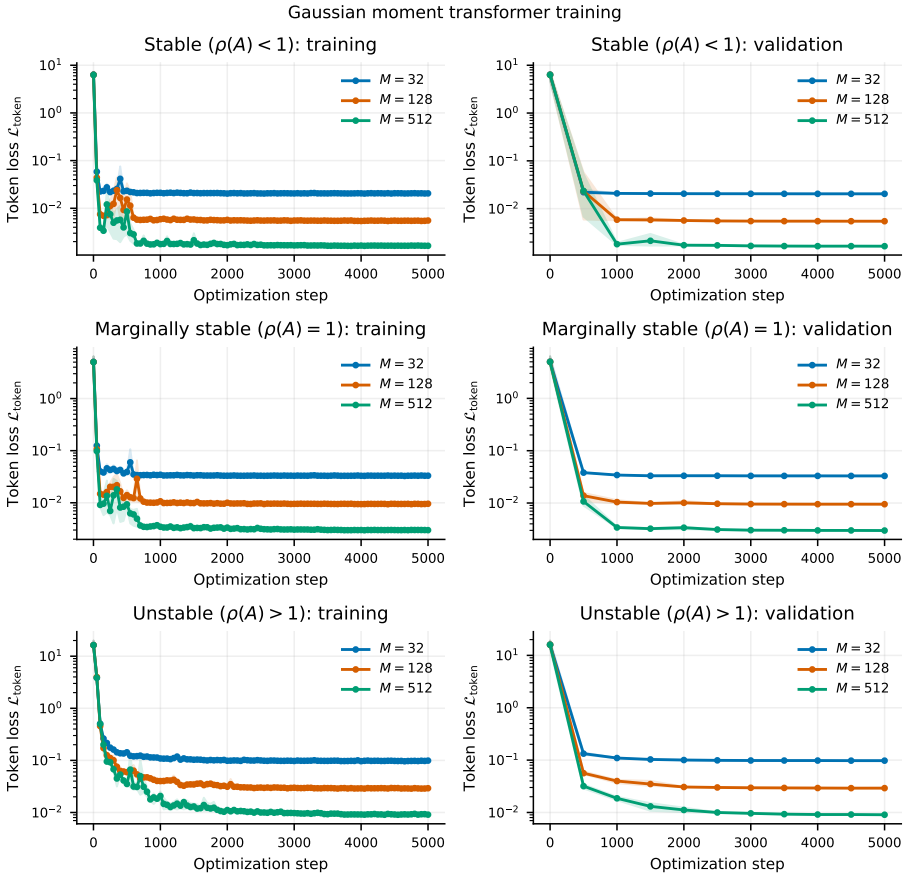

Figure 6: Training and validation losses for the Gaussian moment transformer on every setting.

Computational scaling. We compare the quantile Transformer with an all-particle-attention baseline as M increases as powers of two between 2^4 and 2^12, fixing $T = 15$, batch size 16, and the architecture set to that of Table 2. We report median model-only (i.e., without token sorting and quantile encoding) and end-to-end inference latency over 30 repetitions after 10 warm-up iterations. The results are depicted in Figure 8. Configurations exceeding the available memory are recorded as infeasible and skipped. In contrast, the quantile Transformer aggregates each empirical measure into a single token and attends over a sequence of length T, os the attention matrix dimensions are independent of M.

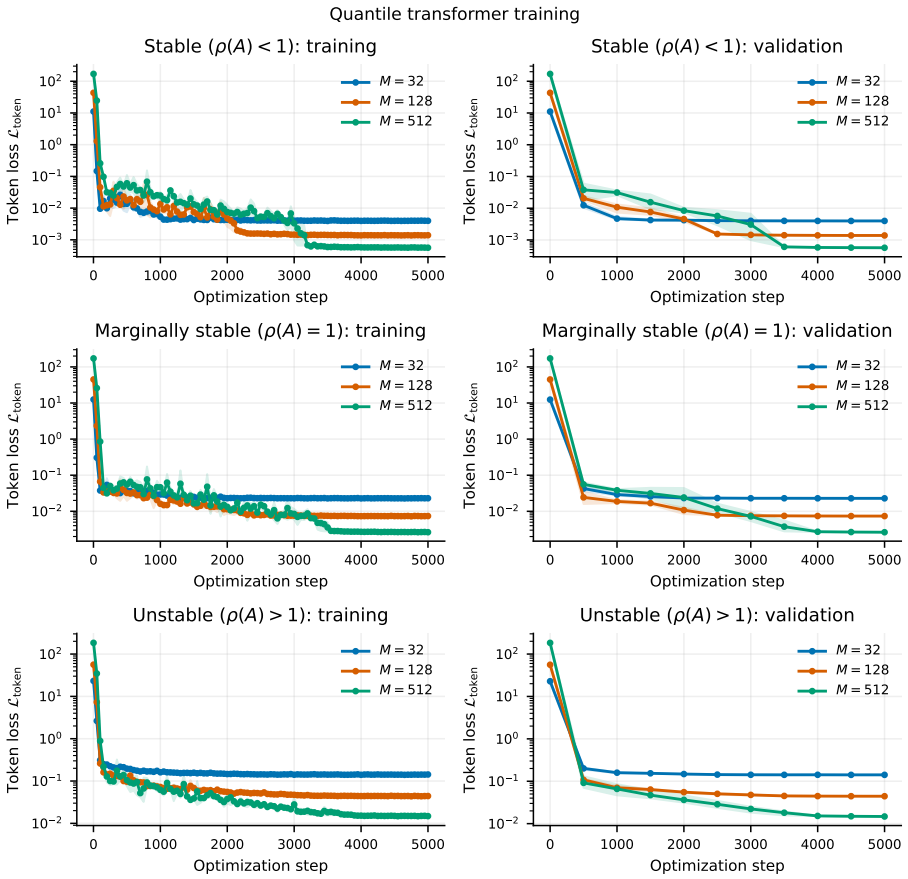

Figure 7: Training and validation losses for the Quantile transformer on every setting.

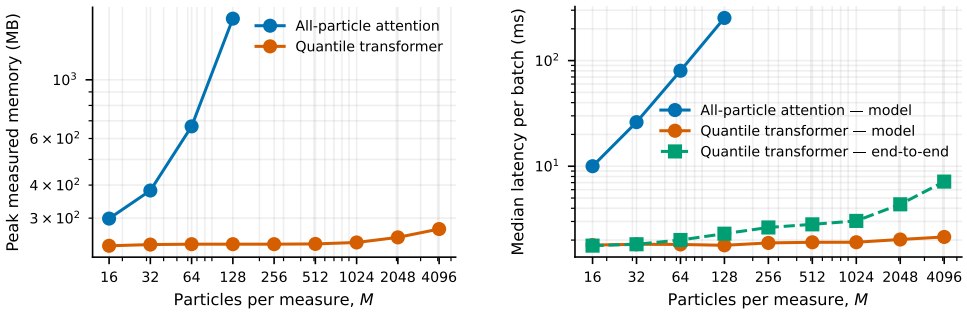

Figure 8: Forward pass scaling of a hypothetical all-to-all particle and sequence-to-sequence transformer versus the Quantile transformer.

Table 5: Hyperparameter search space and selected values for the nonlinear Gaussian experiment.

Hyperparameter	Candidate values	Quantile Transformer	Gaussian moment Transformer
Learning rate	$\{10^{-3}, 3 \times 10^{-4}\}$	10^{-3}	10^{-3}
Hidden dimension	$\{64, 128\}$	128	64
Transformer blocks	$\{2, 3\}$	3	3

H.3 Additional experiments: non-linear Gaussian dynamics in Section 5

Nonlinear Gaussian dynamics. We additionally consider a scalar Gaussian measure sequence $\mu_t = \mathcal{N}(m_t, \sigma_t^2)$ whose parameters follow the autonomous second-order dynamics

$$m_{t+1} = 2r\cos(\omega)m_t - r^2 m_{t-1} - \alpha m_t^3, \qquad \log \sigma_{t+1} = \rho \log \sigma_t + \beta m_t^2 + \eta.$$

We use $r = 0.98$, $\omega = 0.55$, $\alpha = 0.03$, $\rho = 0.75$, $\beta = 0.15$, and $\eta = -0.40$. For each trajectory, the initial parameters are sampled independently as

$$m_0, m_1 \sim \mathrm{Unif}([-1, 1]), \qquad \sigma_0, \sigma_1 \sim \mathrm{Unif}([0.12, 0.35]).$$

The cubic term produces amplitude-dependent oscillations, while the mean-dependent log-standard-deviation dynamics make the dispersion evolve nonlinearly with the location. A visualization of point clouds generated from this dynamics is given in Figure 11 (right). Since the transition depends on both m_{t-1} and m_t, the first scored prediction is μ_2 after observing (μ_0, μ_1). The mean recurrence depends on two consecutive marginals, providing a setting in which the current measure alone does not generally determine the next measure. All remaining data-generation and evaluation settings are identical to those of Section H.2.

Data generation. We assess generalization to new initial conditions of this fixed dynamical system. As in Appendix H.2, each empirical measure consists of M independent samples from μ_t, with fresh sampling at every time step and no persistent particle identities. We use $M \in \{32, 128, 512\}$ and seeds $s \in \{0, 1, 2\}$, training each learned model separately for every (M, s) pair.

Validation and test sets. The common validation/test construction applies, with one dynamical condition, $\mathcal{M}_e = \{32, 128, 512\}$, and $\mathcal{S}_e = \{0, 1, 2\}$. Each trajectory contains $T + 1 = 16$ empirical measures. Because μ_1 is initialized independently of μ_0, the prediction of μ_1 is excluded from both training and evaluation. The scored targets are therefore $\mu_2, \ldots, \mu_{15}$, giving 14 predictions per trajectory. For final training, validation and test each contain $2,048$ trajectories per (M, s) pair.

Hyperparameter tuning. The fixed pilot configuration on which hyperparameters are tuned is $(s = 0, M = 128, T = 15)$. We search over the hyperparameter combinations listed in Table 5. Tuning results are reported in Figure 9 and the final selections in Table 5.

Baselines. We use the same baselines as Sections 5 and H.2.

Evaluation. Performance evaluation uses the same metrics reported for Sections 5 and H.2.

Training. Training is done over 5000 steps. Training convergence plots for both transformers are given in Figure 10. Both learned models converge rapidly and remain stable, with validation loss

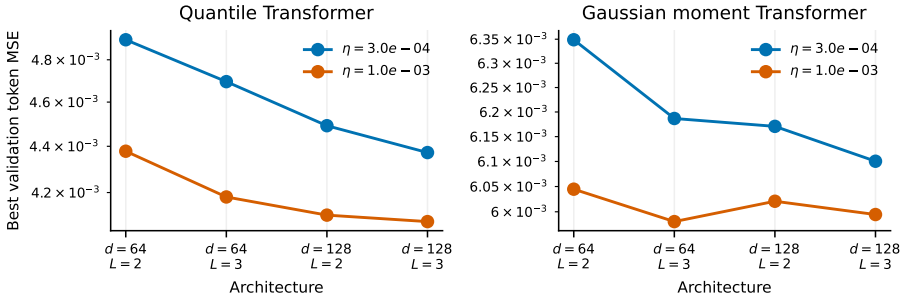

Figure 9: Hyperparameter tuning for the quantile and moment transformers.

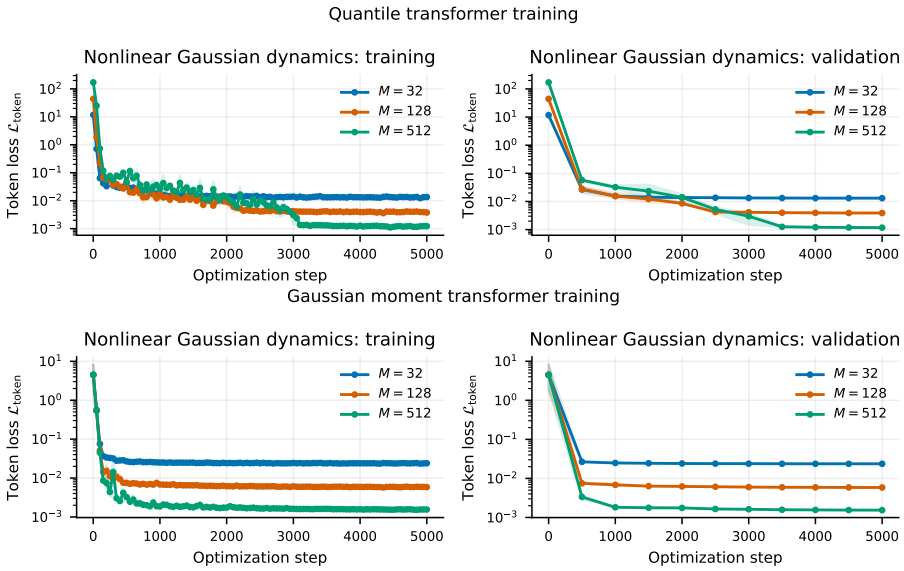

Figure 10: Training and validation losses for every considered model.

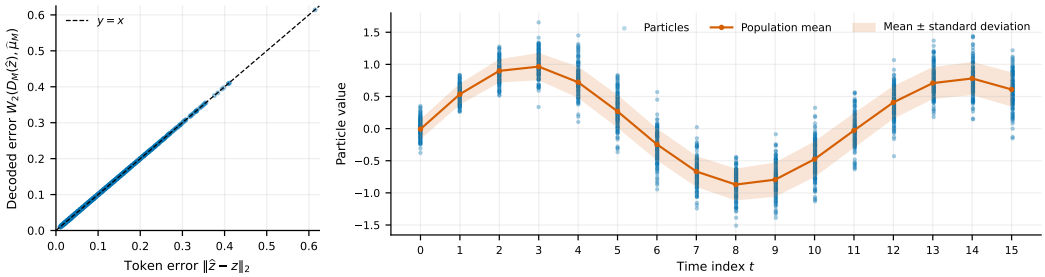

Figure 11: Left: the ratio between W_2 distance of the decoded predicted quantile vs. the empirical measure, and the ℓ_2 distance between the predicted quantile token and the target quantile token. Right: a depiction of the point cloud dynamics.

closely tracking training loss and showing no evidence of overfitting. Performance improves with M, since larger particle sets provide less noisy representations. Again, the quantile Transformer benefits more visibly from continued optimization, whereas the Gaussian moment Transformer reaches its plateau earlier.

Geometry diagnostic. As in Section H.2, Figure 11 (left) verifies the encoder–decoder implementation and confirms that token-space errors have the intended Wasserstein interpretation.

Model performance. The prediction accuracy comparison against the baselines is given in Figure 12. We note that both learned models substantially outperform Persistence across all particle counts and approach the analytic population oracle as M increases. The Gaussian moment Transformer performs slightly better than the Quantile Transformer, consistent with the former's privileged knowledge of the Gaussian family. Persistence improves little with M, indicating that its error is dominated by the changing nonlinear dynamics rather than sampling noise. The oracle has zero grid and quantile discrepancy by construction, while its empirical W_2 error decreases with M because of finite-sample variability. These experiments re-confirm the competitive performance of the Quantile Transformer observed in the linear case.

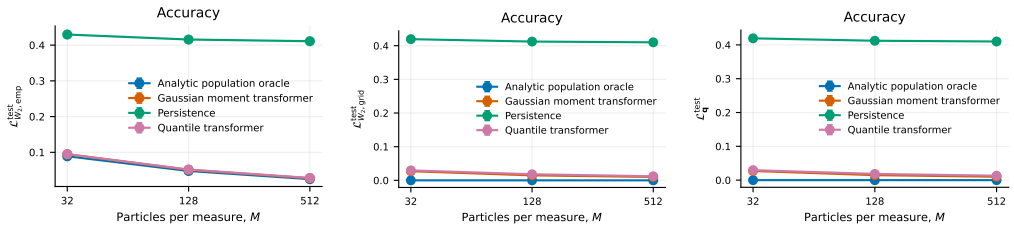

Figure 12: Test set accuracy at three levels of sampling $M \in \{32, 128, 512\}$.

Table 6: Hyperparameter search space and selected values for the evolving-bimodality experiment.

Hyperparameter	Candidate values	Quantile transf.	Mean–var. transf.	Loc–scale–4th-mom transf.	Mixture-parameter transf.
Peak learning rate	$\{10^{-3}, 3 \times 10^{-4}\}$	10^{-3}	3×10^{-4}	10^{-3}	10^{-3}
Hidden dimension	$\{64, 128\}$	128	128	128	128
Transformer blocks	$\{2, 3\}$	3	2	3	3

H.4 Additional experiments: time-changing bimodal distribution

Bimodal dynamics. We consider symmetric Gaussian mixtures

$$\mu_t = \frac{1}{2}\mathcal{N}(-a_t, 1 - a_t^2) + \frac{1}{2}\mathcal{N}(a_t, 1 - a_t^2), \qquad a_t = 0.55 + 0.40 \cos\left(\varphi + \frac{4\pi t}{15}\right), \tag{245}$$

where $\varphi \sim \mathrm{Unif}([0, 2\pi))$ is sampled once per trajectory. Particles are sampled independently from μ_t, with fresh sampling noise across time. Since $a_t \in [0.15, 0.95]$, the component variances remain positive. The population mean and variance are constant at zero and one, respectively, while the distribution transitions between unimodal and bimodal shapes. This isolates forecasting of shape changes beyond location and scale. We chose this example such that the moment transformer fails, since mean and variance are constant through the sequence. We give an example in Figure 15 (right).

Validation and test sets. There is only one dynamical condition, given by (245), with $\mathcal{M}_e = \{32, 128, 512\}$ and $\mathcal{S}_e = \{0, 1, 2\}$. Each trajectory contains the empirical measure sequence $(\widehat{\mu}_{0,M}, \ldots, \widehat{\mu}_{T,M})$ with $T = 15$ and therefore provides T causal input–target pairs. All remaining settings follow the common evaluation protocol.

Hyperparameter tuning. We separately tune the Quantile Transformer and three summary Transformers on $(s = 0, M = 128, T = 15)$. The latter encode empirical mean and log standard deviation (*mean–variance*), these coordinates augmented with standardized kurtosis (*location–scale–fourth-moment*), or an inferred mixture separation (*mixture-parameter*). We search the Cartesian product of peak learning rates $\{10^{-3}, 3 \times 10^{-4}\}$, hidden dimensions $\{64, 128\}$, and block counts $\{2, 3\}$. Selected configurations and tuning results appear in Table 6 and Figure 13. The minimum learning rate is one tenth of the selected peak; the mean–variance model therefore uses different learning rates from those in Table 3.

Baselines. We compare against Persistence and three learned summary Transformers. The *mean–variance Transformer* predicts empirical mean and log standard deviation and decodes them as

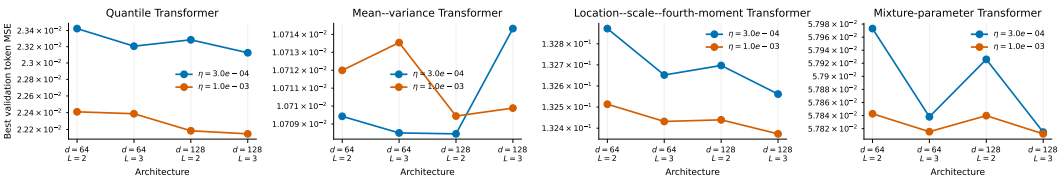

Figure 13: Hyperparameter tuning for the four learned models. Selection uses each model's native validation token loss.

Gaussian midpoint quantiles (i.e., uses the wrong measure class and only works in the steps where the two distributions merge —- See Figure 14, steps 0–2, 6–9, 13–15). The *location–scale–fourth-moment Transformer* assumes the family of measures $\mu_{m,\sigma,a} = \frac{1}{2}\mathcal{N}(m-\sigma a, \sigma^2(1-a^2)) + \frac{1}{2}\mathcal{N}(m+\sigma a, \sigma^2(1-a^2))$. It predicts $(\widehat{m}, \widehat{\ell}, \widehat{\kappa})$ from empirical mean, log standard deviation, and standardized kurtosis, then decodes as $\mu_{\widehat{m},e^{\widehat{\ell}},\widehat{a}}$, where $\widehat{a} = \left(\frac{3-\mathrm{clip}(\widehat{\kappa},1,3)}{2}\right)^{1/4}$. Thus, the mixture family and its kurtosis–separation relation are known, while location, scale, and separation are predicted. The *mixture-parameter Transformer* assumes even more information embodied as the family $\mu_a = \frac{1}{2}\mathcal{N}(-a, 1-a^2) + \frac{1}{2}\mathcal{N}(a, 1-a^2)$ (i.e., it knows that the mean is zero and the standard deviation is one). It predicts only the separation parameter.

Both mixture baselines use deterministic numerical inverse-CDF decoding on the midpoint quantile grid. All learned baselines receive only empirical summaries, whereas the *numerical population oracle* uses the true separation a_t to construct the corresponding midpoint empirical measure.

Evaluation. For the standardized mixture $\mu_a = \frac{1}{2}\mathcal{N}(-a, 1-a^2) + \frac{1}{2}\mathcal{N}(a, 1-a^2)$, the mean and variance are zero and one, respectively, so its standardized kurtosis is $\kappa(a) = \mathbb{E}[Y_a^4] = 3 - 2a^4$. Hence a (i.e., the separation) can be recovered from kurtosis through

$$g(\kappa) = \left(\frac{3 - \mathrm{clip}(\kappa, 1, 3)}{2}\right)^{1/4}.$$

Empirical standardized kurtosis is computed from centered particles using the empirical variance with denominator M, floored at 10^{-6}. The location–scale–fourth-moment model applies g to its predicted kurtosis, while the mixture-parameter model predicts the corresponding separation directly. For the latter, the the unconstrained scalar predicted by the transformer, b, is mapped to the final prediction $\widehat{a} = \min\{|b|, 1 - 10^{-5}\}$ to enforce a valid parameter range.

To decode the mixture baselines, let $F_a = \frac{1}{2}F_0\left(\frac{x+a}{\sqrt{1-a^2}}\right) + \frac{1}{2}F_0\left(\frac{x-a}{\sqrt{1-a^2}}\right)$ denote the CDF of the standardized mixture μ_a, where F_0 is the standard normal CDF. A mixture with location m, scale σ, and separation a is the law of $X = m + \sigma Y_a$, where $Y_a \sim \mu_a$. Its quantile function therefore satisfies $F^{-1}_{m,\sigma,a}(u) = m + \sigma F_a^{-1}(u)$. Evaluating this quantile function on the midpoint grid $u_i = (i - \frac{1}{2})/M$

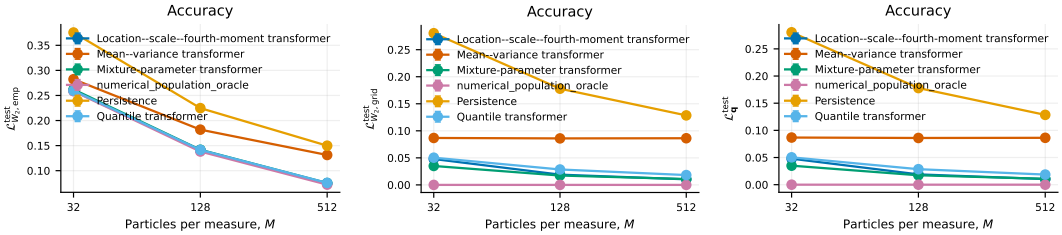

Figure 14: Test accuracy for $M \in \{32, 128, 512\}$: empirical W_2 error (left), population-grid W_2 discrepancy (middle), and population-token discrepancy (right). Results show the mean and sample standard deviation over three seeds.

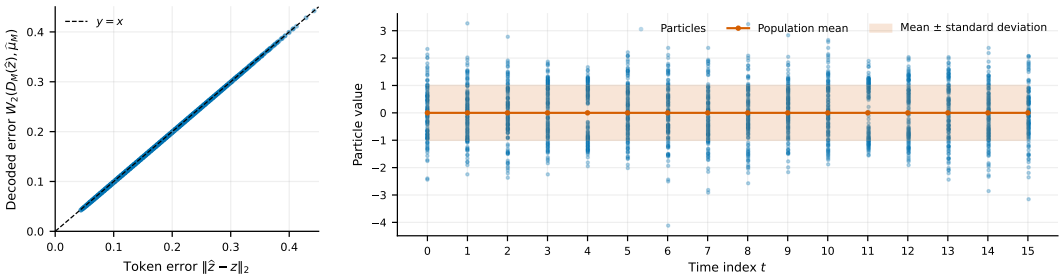

Figure 15: Left: diagnostic comparing decoded empirical W_2 error with quantile-token error. Right: an example of evolving empirical measures with constant population mean and variance.

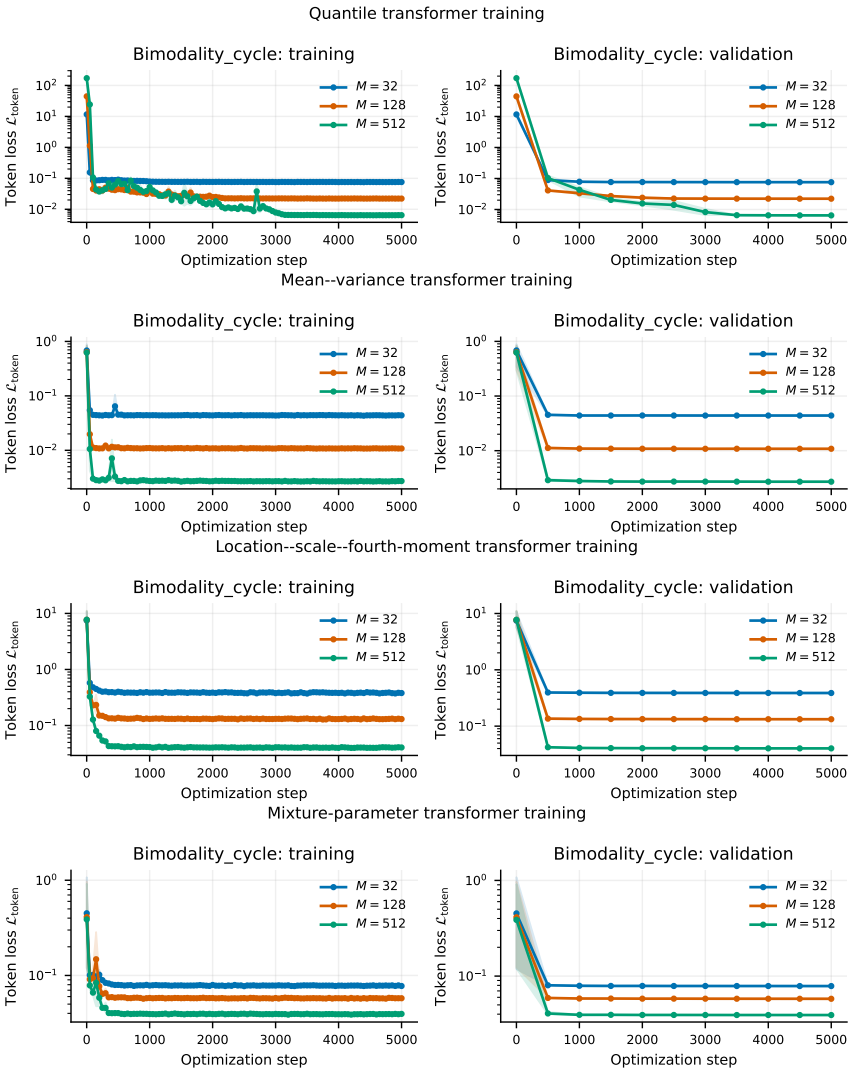

Figure 16: Training and validation losses for the Quantile, mean–variance, location–scale–fourth-moment, and mixture-parameter Transformers, respectively, in reading order.

gives the normalized population token

$$[\boldsymbol{q}_M^{\text{mix}}(m, \sigma, a)]_i = \frac{m + \sigma F_a^{-1}((i - \frac{1}{2})/M)}{\sqrt{M}}.$$

Since F_a^{-1} has no closed-form expression, each midpoint quantile $F_a^{-1}((i - \frac{1}{2})/M)$ is computed numerically by solving $F_a(x) = (i - \frac{1}{2})/M$ with 48 bisection iterations over $x \in [-10, 10]$. The empirical, population-grid, and population-token discrepancies are defined as in Sections 5 and H.2, using $\boldsymbol{q}_M^{\text{mix}}(0, 1, a_t)$ as the population reference and averaging over $t \in [T]$. The grid metric compares finite empirical measures obtained from the midpoint quantiles, rather than computing the exact Wasserstein distance to the continuous mixture. The oracle has zero grid and population-token discrepancy by construction.

Training. All four learned models follow the common training and checkpoint selection protocol with the selected hyperparameters above. Their training and validation histories appear in Figure 16. Native losses use different representation spaces and are therefore not directly comparable across models.

Geometry diagnostic. Figure 15 checks the decoder bound $W_2(D_M(\widehat{\boldsymbol{z}}), \widehat{\mu}) \leq \|\widehat{\boldsymbol{z}} - E_M(\widehat{\mu})\|_2$. No violations beyond numerical tolerance are recorded in the paper-training quantile runs. Equality is not required for unrestricted, potentially unordered predictions.

Model performance. Figure 14 reports prediction errors. As M increases from 32 to 512, the Quantile Transformer's empirical W_2 and its population grid errors decrease. The mean–variance model's grid discrepancy instead remains relatively constant, reflecting its inability to represent the changing shape.

At $M = 512$, the location–scale–fourth-moment and mixture-parameter models achieve relatively smaller grid discrepancies, but their empirical errors are close to the Quantile Transformer's 0.07566 and the population oracle's 0.07216. Thus, our quantile tokens capture evolving distributional shape without a prescribed mixture decoder, while family-aware summaries retain an advantage in population-grid accuracy.

Table 7: Hyperparameter search space and selected values for the nonlinear Student-t experiment.

Hyperparameter	Candidate values	Quantile Transformer	Shared moment Transformer
Peak learning rate	$\{10^{-3}, 3 \times 10^{-4}\}$	10^{-3}	10^{-3}
Hidden dimension	$\{64, 128\}$	128	128
Transformer blocks	$\{2, 3\}$	3	3

H.5 Additional experiments: nonlinear Student-t dynamics

Nonlinear Student-t dynamics. We retain the location and scale dynamics and their initialization from Section H.3, replacing only the Gaussian observation law by

$$Y_t = m_t + \sigma_t Z_\nu, \qquad Z_\nu = \sqrt{\frac{\nu-2}{\nu}}\, U_\nu, \qquad U_\nu \sim t_\nu, \qquad \nu = 3. \tag{246}$$

The standardized variable Z_ν has zero mean and unit variance, so m_t and σ_t retain their interpretation as population mean and standard deviation. Fresh independent observation noise is sampled for every particle and time step. This experiment tests distributional forecasting under heavy-tailed observations while preserving the underlying parameter dynamics. An example sequence appears in Figure 19.

Validation and test sets. There is one nonlinear dynamical condition with $\nu = 3$, $\mathcal{M}_e = \{32, 128, 512\}$, and $\mathcal{S}_e = \{0, 1, 2\}$. Each trajectory contains $T + 1 = 16$ empirical measures. As in the nonlinear Gaussian experiment, predictions are scored at $t \in [T]$, giving 14 scored times after observing at least two measures. All remaining settings follow the common evaluation protocol.

Hyperparameter tuning. We independently tune the Quantile Transformer and a single moment Transformer on $(s = 0, M = 128, T = 15, \nu = 3)$ using the common protocol. The moment model encodes empirical mean and log standard deviation. Its Gaussian and Student-t decoder variants share the same trained network and are not tuned separately. The search space and selected configurations are given in Table 7; tuning results appear in Figure 17.

Baselines. We use *Persistence*, as before, together with a learned *moment Transformer* that predicts $(\widehat{m}, \widehat{\ell})$ from empirical mean and log standard deviation. The trained moment Transformer is evaluated assuming the following output family:

$$\widehat{Y}_t = \widehat{m} + e^{\widehat{\ell}} \sqrt{\frac{\nu-2}{\nu}}\, U_\nu, \qquad U_\nu \sim t_\nu, \qquad \nu = 3.$$

Since $\mathrm{Var}(U_\nu) = \nu/(\nu-2)$ for $\nu > 2$, the factor $\sqrt{(\nu-2)/\nu}$ standardizes the Student-t variable to unit variance. Hence, the decoder's $\widehat{m}$ and $e^{\widehat{\ell}}$ represent the predicted population mean and standard deviation, respectively.

The Student-t decoder is therefore given the correct observation family and degrees of freedom, and is thus privileged, though it does not receive the true population parameters. The privileged

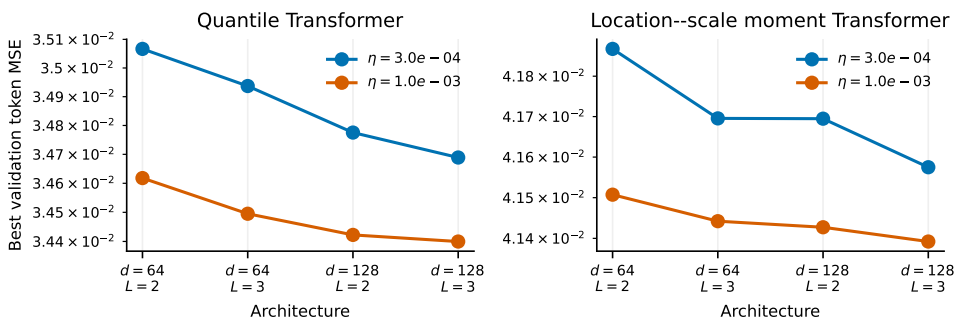

Figure 17: Hyperparameter tuning for the Quantile Transformer and the shared moment Transformer under Student-t observations. Selection uses each model's native validation token loss.

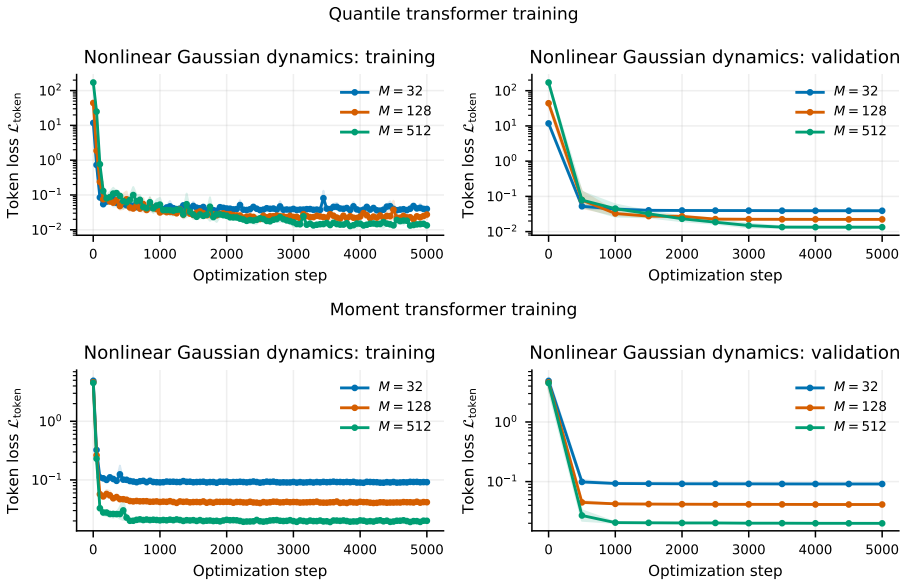

Figure 18: Training and validation losses for the Quantile Transformer (top) and shared moment Transformer (bottom) in the nonlinear Student-t experiment.

population oracle instead uses the true (m_t, σ_t) together with the known Student-t observation law $Y_t = m_t + \sigma_t \sqrt{\frac{\nu-2}{\nu}}\, U_\nu,\ \nu = 3$, and returns the corresponding midpoint-quantile empirical measure.

Evaluation. The moment-model variant is decoded by evaluating the quantile function of their assumed continuous output law on the midpoint grid $u_i = \frac{i-\frac{1}{2}}{M},\ i \in [M]$.

For the Student-t decoder, let F_ν denote the CDF of the standard Student-t distribution with ν degrees of freedom. If $U_\nu \sim t_\nu$, then $F_\nu^{-1}(u)$ is its quantile function. Since $Z_\nu = \sqrt{\frac{\nu-2}{\nu}}\, U_\nu$ has unit variance, and $Y = m + \sigma Z_\nu$, the quantile function of the corresponding location–scale Student-t law is $F^{-1}_{t_\nu;m,\sigma}(u) = m + \sigma\sqrt{\frac{\nu-2}{\nu}}\, F_\nu^{-1}(u)$. Evaluating this quantile function on the midpoint grid and applying the $1/\sqrt{M}$ token normalization gives

$$[\boldsymbol{q}_M^{t_\nu}(m,\sigma)]_i = \frac{1}{\sqrt{M}}\left[m + \sigma\sqrt{\frac{\nu-2}{\nu}}\, F_\nu^{-1}\left(\frac{i-\frac{1}{2}}{M}\right)\right], \qquad i \in [M]. \tag{247}$$

The Student-t quantiles $F_3^{-1}(u_i)$ are evaluated numerically. Applying the empirical decoder to a normalized token returns the equally weighted empirical measure supported on the corresponding midpoint quantiles. In particular, $D_M\left(\boldsymbol{q}_M^{t_3}(\widehat{m}, e^{\widehat{\ell}})\right) = \frac{1}{M}\sum_{i=1}^M \delta_{\widehat{m}+\frac{e^{\widehat{\ell}}}{\sqrt{3}}F_3^{-1}((i-\frac{1}{2})/M)}$.

The empirical, population-grid, and population-token losses are defined as in Sections 5 and H.2. For the population-based metrics, the reference token is the correctly specified Student-t token $\boldsymbol{q}_M^{t_3}(m_t, \sigma_t)$ constructed from the true population parameters.

Training. Both networks follow the common training and checkpoint selection protocol. Exactly one moment network is trained for each (M, s), and its selected checkpoint is reused for both decoders. Training and validation histories are shown in Figure 18. Their native loss magnitudes are not directly comparable across representations.

Geometry diagnostic. Figure 19 checks the decoder bound $W_2(D_M(\widehat{\boldsymbol{z}}), \widehat{\mu}) \le \|\widehat{\boldsymbol{z}} - E_M(\widehat{\mu})\|_2$. No violations beyond numerical tolerance are recorded in the paper-training quantile runs. The bound applies independently of the generating distribution; equality is not required for potentially unordered predicted tokens.

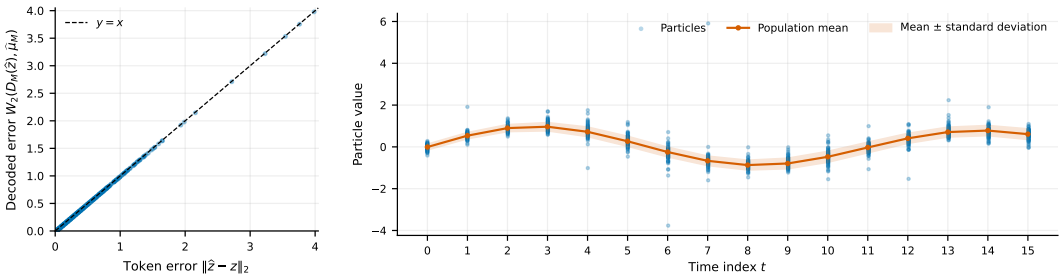

Figure 19: Left: diagnostic comparing decoded empirical W_2 error with quantile-token error. Right: an example empirical measure sequence under nonlinear Student-t dynamics.

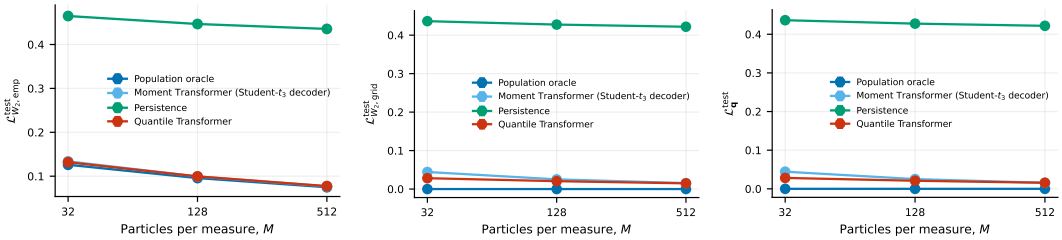

Figure 20: Test accuracy for $M \in \{32, 128, 512\}$: empirical W_2 error (left), population-grid W_2 discrepancy (middle), and population-token discrepancy (right). Results show the mean and sample standard deviation over three seeds. The two moment-model curves use the same checkpoints and differ only in their decoders.

Model performance. Figure 20 reports prediction errors. As M increases from 32 to 512, the Quantile Transformer's empirical W_2 and population grid errors decrease. At $M = 512$, the correctly specified moment decoder achieves empirical error close to the Quantile Transformer and the population oracle. Its grid discrepancy stays slightly above the Quantile Transformer's value.

Persistence remains substantially less accurate, as expected. These results show that quantile prediction remains competitive with a correctly specified parametric decoder even under heavy-tailed observations.

Table 8: Hyperparameter search space and selected values for the multivariate Gaussian experiment.

Hyperparameter	Candidate values	Sliced-quantile Transformer	Gaussian moment Transformer
Peak learning rate	$\{3 \times 10^{-4}, 10^{-4}\}$	3×10^{-4}	3×10^{-4}
Hidden dimension	$\{256, 512\}$	512	512
Transformer blocks	$\{3, 4\}$	4	4

H.6 Additional experiments: multivariate Gaussian dynamics

Dynamical system parameters. We extend the Gaussian linear dynamical system of Appendix H.2 to observations in $\mathbb{R}^5$, retaining latent dimension $p = 10$. The population parameters evolve as

$$\boldsymbol{m}_{t+1} = \boldsymbol{A}\boldsymbol{m}_t, \qquad \boldsymbol{P}_{t+1} = \boldsymbol{A}\boldsymbol{P}_t\boldsymbol{A}^\top + \boldsymbol{Q},$$

and the observation law is

$$\mu_t = \mathcal{N}(\boldsymbol{b}_t, \boldsymbol{C}_t), \qquad \boldsymbol{b}_t = \boldsymbol{H}\boldsymbol{m}_t, \qquad \boldsymbol{C}_t = \boldsymbol{H}\boldsymbol{P}_t\boldsymbol{H}^\top + \boldsymbol{R}.$$

Define the equicorrelation matrix $\boldsymbol{E}_r(a) = (1-a)\boldsymbol{I}_r + a\mathbf{1}_r\mathbf{1}_r^\top$. We use

$$\boldsymbol{Q} = 0.03^2\boldsymbol{E}_{10}(0.25), \qquad \boldsymbol{R} = 0.10^2\boldsymbol{E}_5(0.30).$$

Writing $\boldsymbol{e}_j$ for the coordinate vectors in $\mathbb{R}^{10}$, row i of the observation matrix is

$$\boldsymbol{H}_{i,:} = \frac{0.8e_i^\top + 0.45\boldsymbol{e}_{i+5}^\top + 0.2e_{1+(i \bmod 5)}^\top}{\sqrt{0.8^2 + 0.45^2 + 0.2^2}}, \qquad i = 1, \ldots, 5.$$

For each trajectory, independently sample

$$\boldsymbol{m}_0 \sim \mathrm{Unif}([-2, 2]^{10}), \qquad a_{0,j} \overset{\text{iid}}{\sim} \mathrm{Unif}([0.10, 0.60]),$$

and set $\boldsymbol{P}_0 = \mathrm{Diag}(\boldsymbol{a}_0)\boldsymbol{E}_{10}(0.35)\,\mathrm{Diag}(\boldsymbol{a}_0)$.

We use the same three transition regimes as in the scalar Gaussian experiment. The stable matrix is tridiagonal:

$$[\boldsymbol{A}_{\mathrm{st}}]_{ii} = 0.80, \qquad [\boldsymbol{A}_{\mathrm{st}}]_{i,i+1} = 0.12, \qquad [\boldsymbol{A}_{\mathrm{st}}]_{i+1,i} = -0.08,$$

with all other entries zero. Its spectral radius is $\rho(\boldsymbol{A}_{\mathrm{st}}) \approx 0.821798$. The marginally stable and unstable matrices are

$$\boldsymbol{A}_{\mathrm{ms}} = \frac{A_{\mathrm{st}}}{\rho(A_{\mathrm{st}})}, \qquad \boldsymbol{A}_{\mathrm{un}} = \frac{1.1\boldsymbol{A}_{\mathrm{st}}}{\rho(\boldsymbol{A}_{\mathrm{st}})},$$

with spectral radii 1 and 1.1, respectively. All remaining data-generation parameters are shared.

Data generation. As in the scalar Gaussian experiment, we propagate population parameters rather than realized latent-state trajectories. At every time, we independently draw M particles from μ_t; particles have no persistent identities across time. We use $M \in \{32, 128, 512\}$ and seeds

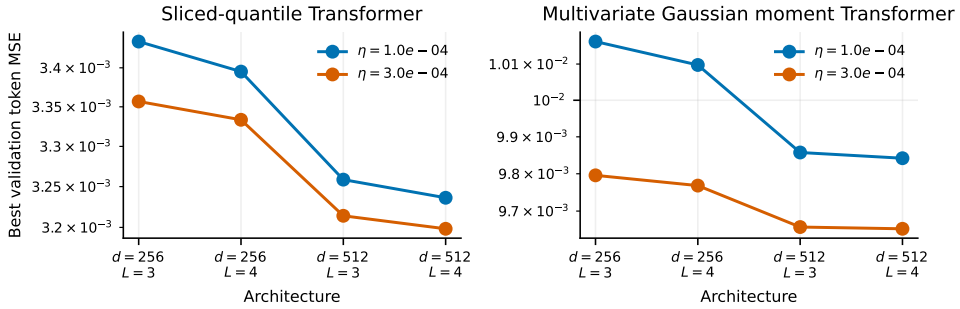

Figure 21: Best validation token loss for each hyperparameter candidate, with independent selection for each model.

$s \in \{0, 1, 2\}$. Each learned model is trained separately for every $(\boldsymbol{A}, M, s)$ combination, with hyperparameters shared across these settings.

Validation and test sets. Here, $\mathcal{C}_e = \{A_{\mathrm{st}}, A_{\mathrm{ms}}, A_{\mathrm{un}}\}$, $\mathcal{M}_e = \{32, 128, 512\}$, $\mathcal{S}_e = \{0, 1, 2\}$. Each trajectory contains $T + 1 = 16$ measures and supplies $T = 15$ scored next-measure predictions. The common disjoint validation/test construction applies, with $2,048$ trajectories per split, processed in eight batches of 256. For the decoder comparison, the selected CUDA training checkpoints (described shortly) are reevaluated locally on a shared test set regenerated using the original split construction and evaluation seeds.

Hyperparameter tuning. The fixed pilot configuration is $(A, M, s) = (A_{\mathrm{ms}}, 512, 0)$. For each learned model, we independently search

$$\eta_{\max} \in \{10^{-4}, 3 \times 10^{-4}\}, \qquad d_{\mathrm{model}} \in \{256, 512\}, \qquad L \in \{3, 4\},$$

giving eight candidates per model. Both models select $(\eta_{\max}, d_{\mathrm{model}}, L) = (3 \times 10^{-4}, 512, 4)$ by their best native validation token loss. The selected configurations are reused across all particle counts, seeds, and transition regimes.

Relative to the common architecture and optimization tables, this experiment uses eight attention heads, feed-forward width $4d_{\mathrm{model}}$, batch size 256, and $20,000$ updates. Validation occurs every $2,000$ updates, including initialization. The learning rate increases linearly for $2,000$ updates and then follows cosine decay to $0.1\eta_{\max}$. The selected models therefore have feed-forward width $2,048$ and minimum learning rate 3×10^{-5}. Transformer blocks are independently initialized, and biases and LayerNorm parameters are excluded from weight decay. Table 8 reports the eight candidates per model and the selected values. Tuning uses FP32 on an NVIDIA L4 GPU. Results appear in Figure 21.

Baselines. We compare the Quantile Transformer with Persistence, a Gaussian moment Transformer, and a population oracle. The moment Transformer uses a 20-dimensional token containing the empirical mean and the lower-triangular entries of $\mathrm{chol}(\widehat{\Sigma} + 10^{-5} I_5)$, where $\widehat{\Sigma}$ uses divisor M. Its predicted mean and triangular factor specify $\mathcal{N}(\widehat{b}, LL^\top)$, after replacing diagonal entries of L by $\max\{|L_{ii}|, 10^{-3}\}$.

The moment baseline and population oracle are materialized as equally weighted M-particle clouds using the same fixed standard Gaussian reference cloud. This reference is obtained by applying the componentwise standard-normal inverse CDF to a scrambled five-dimensional Sobol design with seed zero, clipping its coordinates to $[10^{-6}, 1 - 10^{-6}]$. It is transformed by the predicted or true location and symmetric covariance square root, respectively. The population oracle is therefore a numerical discretization of the known law. Persistence returns the preceding observed empirical measure.

The Quantile Transformer uses the identifying directions

$$V_5 = \{e_i : 1 \le i \le 5\} \cup \left\{ \frac{e_i + e_j}{\sqrt{2}} : 1 \le i < j \le 5 \right\}, \qquad |V_5| = 15.$$

Projected particles are sorted separately along each direction, concatenated, and normalized by $1/\sqrt{15M}$. The input to the learned linear projection is multiplied by $\sqrt{15M}$, and the output projection is divided by the same factor. Training targets and losses retain their normalized units.

Predicted tokens are decoded using the variable-generator elliptical construction developed in Appendix F. Directional predictions are projected onto the cone of nondecreasing sequences by equal-weight isotonic regression. Their moments determine a reconstructed location $\widehat{b}$ and covariance $\widehat{C}$, with negative covariance eigenvalues replaced by zero. The standardized, symmetrized marginal constructed from these repaired slices is then used to estimate the radial law.

The numerical schedules are

$$n_M = \lceil M^{1/4} \rceil + \lceil \sqrt{5} \rceil, \qquad K_M = \lceil \sqrt{M} \rceil, \qquad \rho_M = \kappa_M = \tau_M = M^{-1}.$$

The radial support is

$$\mathcal{R}_M = \{j / n_M : j = 0, \ldots, n_M^2\} \cup \{\sqrt{5}\},$$

and the absolute-marginal CDF is evaluated at $x_k = \tan\big(\frac{\pi}{2}(k-\frac{1}{2})/K_M\big)$, $k = 1, \dots, K_M$. Writing $\widehat{p}_k$ for these CDF values and $A_{kj} = \mathbb{P}(|r_j U_1| \le x_k)$ for U uniform on the unit sphere in $\mathbb{R}^5$, the radial weights solve

$$\min_{w \ge 0} \frac{1}{K_M}\|Aw - \widehat{p}\|_2^2 + \rho_M \|w\|_2^2 \quad \text{subject to} \quad \sum_j w_j = 1, \qquad \sum_j w_j r_j^2 = 5.$$

Here ρ_M is the regularization coefficient, κ_M the degeneracy threshold, and τ_M the allowed numerical objective gap.

The problem is implemented in CPU float64 using CVXPY 1.7.5 and OSQP 1.1.3, with absolute and relative tolerances 10^{-8}, polishing enabled, and at most $20,000$ iterations. An unsuccessful numerical acceptance check triggers an OSQP retry at tolerance 10^{-10} with $100,000$ iterations, followed, if necessary, by Clarabel 0.11.1 at tolerance 10^{-10} with 500 iterations. Warm starts are disabled. Accepted weights satisfy post-repair feasibility tolerance 10^{-10} and a computed primal–dual objective-gap bound of at most τ_M. If either degeneracy condition from the appendix holds, $\beta \le \kappa_M$ or $\|\phi^{\mathrm{sym}}\|_{L^2} \le \kappa_M$, the radial law is replaced by $\delta_{\sqrt{5}}$, retaining the reconstructed location and covariance as parameters of the underlying elliptical law.

For particle synthesis, the number of radial groups is

$$R_M = \lfloor \sqrt{M} \rfloor,$$

without dimension adjustment, giving $R_M = 5, 11, 22$ for $M = 32, 128, 512$, respectively. Write $M = qR_M + r$, with $0 \le r < R_M$, and assign

$$k_g = q + \mathbf{1}_{\{g<r\}}, \qquad g = 0, \dots, R_M - 1.$$

Let $p_g = k_g/M$, $\ell_0 = 0$, and $\ell_{g+1} = \ell_g + p_g$. For the fitted radial law $\widehat{\nu}_M$, group g uses the radial quantile average

$$\bar{r}_g = \frac{1}{p_g}\int_{\ell_g}^{\ell_{g+1}} F^{-1}_{\widehat{\nu}_M}(u)\, du.$$

We use randomly rotated antipodal orthogonal blocks to reduce finite-sample angular imbalance, since every complete block has zero mean and isotropic second moment. Formally, for each group g and block b, draw $G_{g,b} \in \mathbb{R}^{5\times 5}$ with independent standard-normal entries and compute its QR decomposition, adjusting signs so that the diagonal of the triangular factor is nonnegative. If $u_{g,b,1}, \dots, u_{g,b,5}$ are the columns of the resulting orthogonal matrix, form

$$\mathcal{B}_{g,b} = (u_{g,b,1}, -u_{g,b,1}, \dots, u_{g,b,5}, -u_{g,b,5}).$$

Each complete block satisfies

$$\frac{1}{10}\sum_{u \in \mathcal{B}_{g,b}} u = 0, \qquad \frac{1}{10}\sum_{u \in \mathcal{B}_{g,b}} uu^\top = \frac{1}{5} I_5.$$

Concatenate independently generated blocks within each group and retain the first k_g directions, denoted $u_{g,a}$. A truncated block need not retain exact isotropy. The decoded measure is

$$\widehat{\mu}_M^{\mathrm{pred}} = \frac{1}{M}\sum_{g=0}^{R_M-1}\sum_{a=1}^{k_g} \delta_{\widehat{b}+\bar{r}_g \widehat{C}^{1/2} u_{g,a}},$$

where $\widehat{C}^{1/2}$ is the symmetric positive-semidefinite square root. Radial averaging and angular truncation mean that this finite cloud need not have covariance exactly $\widehat{C}$.

For construction seed $c \in \{0, \dots, 9\}$ and zero-indexed group g, NumPy PCG64 is initialized with `SeedSequence([c,5,`R_M`,g,314159])`. For each (M, c), the resulting direction arrays are fixed and reused across all prediction times, trajectories, training seeds, and transition regimes. Different groups use separately randomized block sequences.

Evaluation. We report $\mathcal{L}^{\mathrm{test}}_{W_2,\mathrm{emp}}$ and $\mathcal{L}^{\mathrm{test}}_{\boldsymbol{q}}$. The former compares each predicted M-particle measure with the observed target cloud. For equally weighted clouds $X = (x_a)_{a=1}^M$ and $Y = (y_a)_{a=1}^M$ in $\mathbb{R}^5$, we compute

$$W_2(X,Y) = \left(\min_{\pi \in \mathfrak{S}_M} \frac{1}{M}\sum_{a=1}^M \|x_a - y_{\pi(a)}\|_2^2\right)^{1/2}.$$

The assignment problem is solved in CPU float64 using `scipy.optimize.linear_sum_assignment`, without slicing or entropic regularization. For the Quantile Transformer, distances are evaluated separately for the ten fixed angular constructions and then averaged; their clouds are never pooled.

The population reference for the token metric uses Gaussian quantile cell averages:

$$[\boldsymbol{q}^{\star}_{t,M}]_{\ell,i} = \frac{v_\ell^\top \boldsymbol{b}_t + \sqrt{v_\ell^\top \boldsymbol{C}_t v_\ell}\,\gamma_{i,M}}{\sqrt{15M}}, \qquad \gamma_{i,M} = M \int_{(i-1)/M}^{i/M} \Phi^{-1}_{\mathcal{N}(0,1)}(u)\,du.$$

The token loss compares this reference with the raw predicted token, before isotonic repair or decoding. Moment predictions are converted to population tokens using their predicted Gaussian parameters; Persistence uses the preceding empirical token, and the oracle uses $\boldsymbol{q}^{\star}_{t,M}$ itself. Thus, the oracle has zero token error but generally nonzero empirical Wasserstein error.

Both losses are averaged over the $2,048$ test trajectories and 15 scored prediction times, then summarized by the mean and sample standard deviation across the three training seeds. For the Quantile Transformer's Wasserstein loss, averaging over angular constructions occurs within each run before aggregation across training seeds. Figure 22 reports these two metrics.

Training. Tuning and final training use an NVIDIA L4 GPU in FP32, with Python 3.12.3 and PyTorch 2.13.0. All final runs complete $20,000$ updates; their selected checkpoints occur at update $18,000$ or $20,000$. The common native-token checkpoint criterion applies, so the subsequent choice of decoder does not alter training or checkpoint selection. The reported decoder evaluation uses these frozen checkpoints with MPS inference and CPU float64 decoding and transport. Figures 24 and 27 shows a rapid initial loss reduction followed by slower improvement. Training and validation losses exhibit similar trends, with lower final losses at larger M and higher residual losses in the unstable regime.

Model performance. Both learned models substantially outperform Persistence across all three regimes, with errors decreasing as M increases in Figure 22. The Gaussian moment Transformer consistently achieves lower errors, but the Quantile Transformer with balanced elliptical decoding remains close to it and to the population oracle in empirical W_2, despite having less class knowledge. Absolute errors are largest in the unstable regime, as expected.

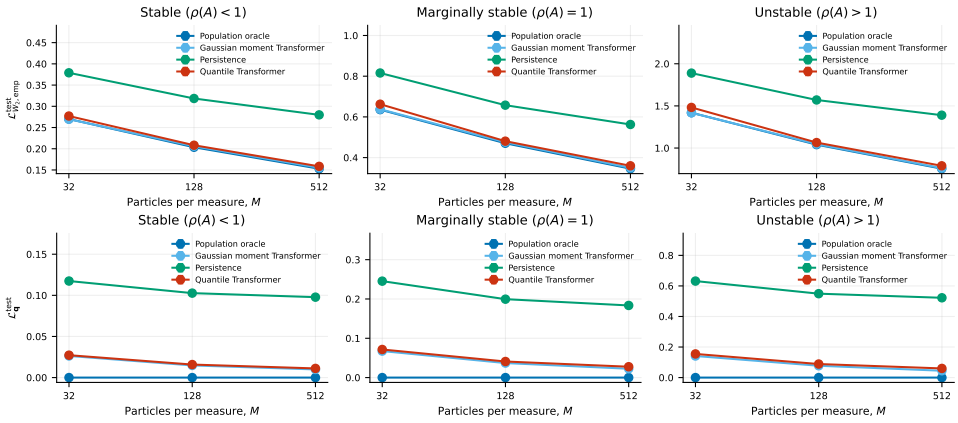

Figure 22: Test errors versus particle count: empirical W_2 distance (top) and population-token Euclidean error (bottom). Results show means and sample standard deviations across three seeds.

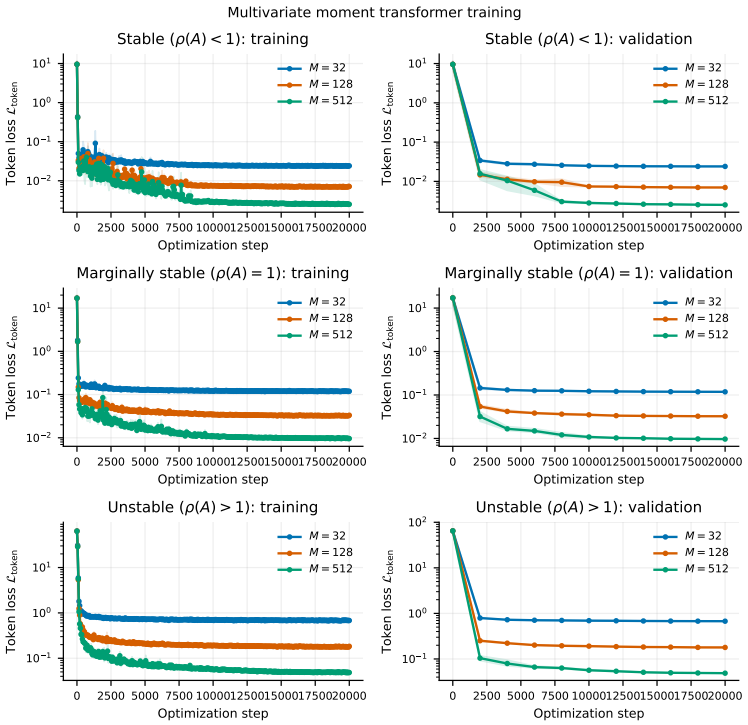

Figure 23: Training convergence for the multivariate Gaussian moment transformer.

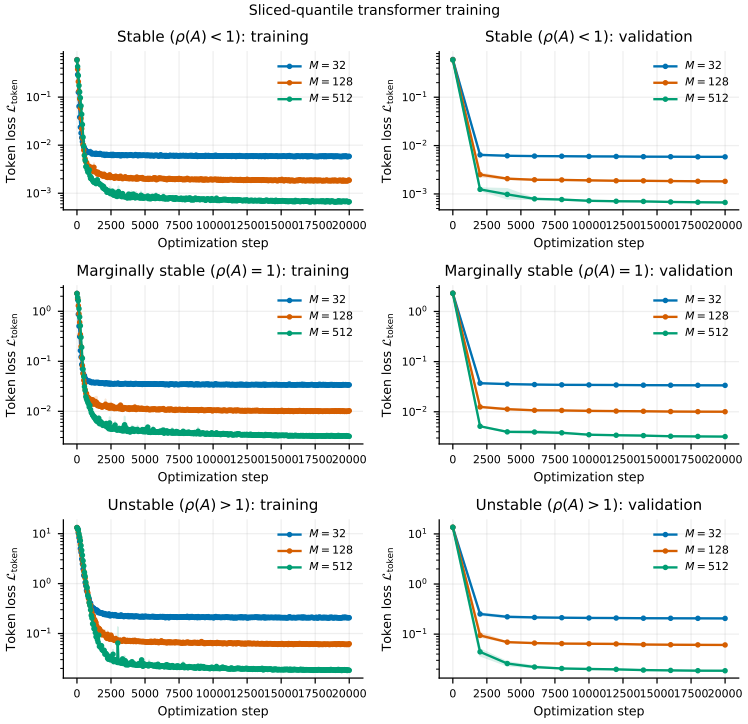

Figure 24: Training convergence for the Quantile Transformer.

Table 9: Hyperparameter search space and selected values for the multivariate Student-t experiment.

Hyperparameter	Candidate values	Sliced-quantile Transformer	Student-t moment Transformer
Peak learning rate	$\{10^{-4}, 3 \times 10^{-4}\}$	3×10^{-4}	3×10^{-4}
Hidden dimension	$\{256, 512\}$	512	512
Transformer blocks	$\{3, 4\}$	4	4

H.7 Additional experiments: multivariate Student-t dynamics

Dynamical system parameters. We consider correlated Student-t measures in dimension $d = 5$ with $\nu = 3$ degrees of freedom. For each coordinate $j = 1, \ldots, 5$, the location and marginal log standard deviation evolve according to

$$m_{t+1,j} = 2(0.98)\cos(0.55)m_{t,j} - (0.98)^2 m_{t-1,j} - 0.03m_{t,j}^3,$$
$$\ell_{t+1,j} = 0.75\ell_{t,j} + 0.15m_{t,j}^2 - 0.40, \qquad t \geq 1,$$

where $\sigma_{t,j} = \exp(\ell_{t,j})$. For each trajectory, all initial coordinates are sampled independently:

$$m_{0,j}, m_{1,j} \overset{\text{iid}}{\sim} \text{Unif}([-1, 1]), \qquad \sigma_{0,j}, \sigma_{1,j} \overset{\text{iid}}{\sim} \text{Unif}([0.12, 0.35]).$$

Set $\boldsymbol{b}_t = (m_{t,1}, \ldots, m_{t,5})^\top$, $D_t = \text{Diag}(\sigma_{t,1}, \ldots, \sigma_{t,5})$, $\boldsymbol{C}_t = D_t E_5(0.35) D_t$, using the equicorrelation notation from Section H.6. The population law has location $\boldsymbol{b}_t$ and covariance $\boldsymbol{C}_t$; its conventional Student-t scale matrix is $(\nu - 2)\boldsymbol{C}_t/\nu = \boldsymbol{C}_t/3$. There is one fixed nonlinear dynamical condition.

Data generation. Given $L_t L_t^\top = \boldsymbol{C}_t$, with L_t the Cholesky factor, each particle is generated as

$$Y_{t,a} = \boldsymbol{b}_t + \sqrt{\frac{\nu - 2}{U_{t,a}}}\, L_t Z_{t,a}, \qquad Z_{t,a} \sim \mathcal{N}(0, I_5), \qquad U_{t,a} \sim \chi^2_\nu.$$

The Gaussian vector and chi-square variable are independent, and fresh draws are used across particles, times, and trajectories. A single $U_{t,a}$ scales all five coordinates of a particle, yielding a multivariate Student-t law. The particle counts, training seeds, online sampling protocol, and separate training for each (M, s) follow Section H.6.

Validation and test sets. We retain the split sizes, shared evaluation trajectories, and observed-history prediction protocol of Section H.6. Each sequence contains $T + 1 = 16$ measures. Because the parameters of μ_1 are initialized independently of those of μ_0, the prediction of μ_1 is excluded from both training and evaluation. The scored targets are $\mu_2, \ldots, \mu_{15}$, giving 14 predictions per trajectory. The decoder comparison uses the saved CUDA checkpoints and a shared test set regenerated with the original split construction and evaluation seeds.

Hyperparameter tuning. Both learned models are tuned independently on Student-t sequences at the fixed pilot configuration $(M, s) = (512, 0)$. We use the same eight-candidate search grid, training

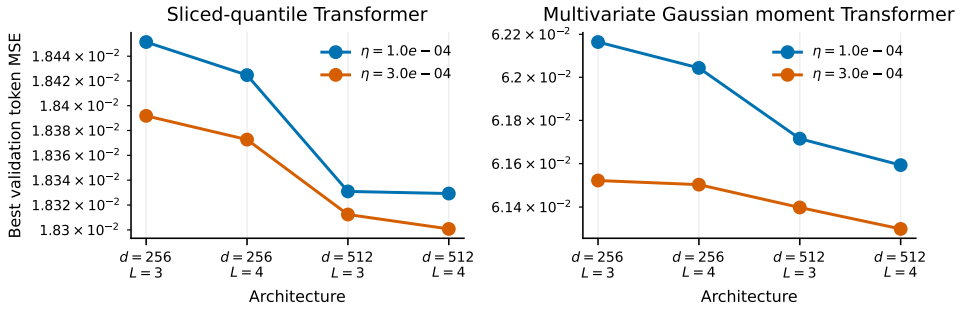

Figure 25: Best validation token loss for each hyperparameter candidate, with independent selection for each model.

budget, validation protocol, and native-token selection criterion as in Section H.6. Both models select

$$(\eta_{\max}, d_{\text{model}}, L) = (3 \times 10^{-4}, 512, 4).$$

The selected configurations are fixed across particle counts and training seeds. The remaining architecture and optimization settings are unchanged from the multivariate Gaussian experiment. Table 9 and Figure 25 summarize the hyperparameter tuning results.

Baselines. We retain Persistence and the 20-dimensional empirical mean/Cholesky moment representation from Section H.6, including its covariance regularization and predicted-diagonal correction. The moment baseline now decodes its predicted location and covariance using the known Student-t_3 family. The population oracle uses the same construction with the true population parameters.

For these two methods, the fixed reference cloud is constructed from a scrambled eight-dimensional Sobol design with seed zero. After clipping coordinates to $[10^{-12}, 1 - 10^{-12}]$ and applying the standard-normal inverse CDF componentwise, let $z_a \in \mathbb{R}^5$ contain the first five coordinates and $h_a \in \mathbb{R}^3$ the remaining three. The reference particles are

$$\xi_a = \frac{z_a}{\|h_a\|_2}, \qquad a = 1, \dots, M.$$

They are mapped to $\widehat{b} + \widehat{C}^{1/2}\xi_a$ for the moment baseline and $\boldsymbol{b}_t + \boldsymbol{C}_t^{1/2}\xi_a$ for the oracle. This is a deterministic finite approximation of the unit-covariance Student-t_3 reference law; its empirical covariance need not equal I_5.

The Quantile Transformer retains the identifying directions, normalized sliced-quantile representation, and input/output projection scaling of Section H.6. Its predictions are decoded using exactly the same variable-generator elliptical decoder, numerical schedules, solver tolerances, and fallback rules described there. In particular, synthesis uses $R_M = \lfloor \sqrt{M} \rfloor$ radial groups and the same balanced antipodal orthogonal-block construction, with ten fixed construction seeds. Direction arrays are reused across reconstructions as in the Gaussian experiment. Neither the Student-t family nor its degrees of freedom is supplied to this general decoder.

Evaluation. We report the empirical full-dimensional Wasserstein loss $\mathcal{L}^{\text{test}}_{W_2,\text{emp}}$ and population-token loss $\mathcal{L}^{\text{test}}_{\boldsymbol{q}}$. Their computation follows Section H.6, including exact particle assignment for empirical W_2, separate evaluation and averaging of the ten balanced constructions, and aggregation across training seeds. Time averages now run over $t = 2, \dots, 15$, with denominator $14N_{\text{test}}$.

The population token uses the standardized Student-t_3 quantile basis in place of the Gaussian basis. Writing F_3 for the standard Student-t_3 CDF, define

$$\gamma^{(3)}_{i,M} = \frac{M}{\sqrt{3}} \int_{(i-1)/M}^{i/M} F_3^{-1}(u)\, du.$$

Then

$$[\boldsymbol{q}^\star_{t,M}]_{\ell,i} = \frac{v_\ell^\top \boldsymbol{b}_t + \sqrt{v_\ell^\top \boldsymbol{C}_t v_\ell}\, \gamma^{(3)}_{i,M}}{\sqrt{15M}}.$$

The moment baseline is mapped to this token space using its predicted location and covariance. The Quantile Transformer is evaluated using its raw predicted token before repair or decoding; Persistence uses the preceding empirical token. The oracle uses the population token itself and therefore has zero token error, while its empirical Wasserstein error is generally nonzero. Only these two metrics are included in the paper comparison. Results are reported in Figure 26.

Training. Hyperparameter tuning is performed on MPS, while final training uses CUDA on the NVIDIA L4. The decoder evaluation uses frozen checkpoints with MPS inference and CPU float64 decoding and transport, as in Section H.6. All final runs complete $20,000$ updates, and every selected checkpoint occurs at the final update.

Figure 27 shows a rapid initial reduction in loss followed by comparatively stable validation curves. Online training losses exhibit fluctuations and occasional spikes. Both models attain lower final native losses as the number of particles increases.

Model performance. In Figure 26, both learned models substantially outperform Persistence, with errors decreasing as M increases. The Quantile Transformer achieves lower empirical W_2 and population-token errors than the Student-t_3 moment baseline at $M = 32$ and 128, while the moment baseline performs slightly better at $M = 512$.

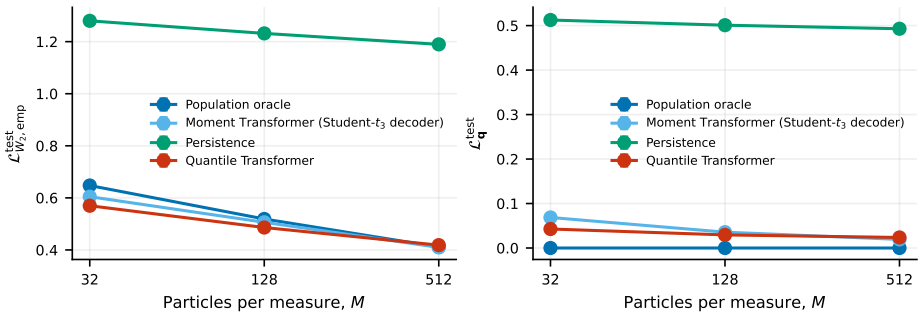

Figure 26: Test errors versus particle count: empirical W_2 distance (top) and population-token Euclidean error (bottom). Results show means and sample standard deviations across three seeds.

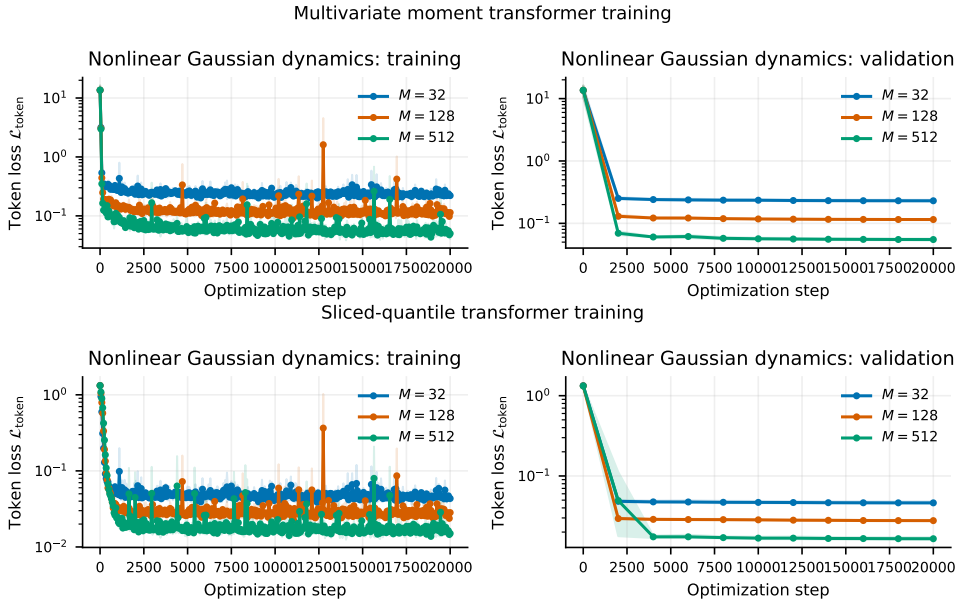

Figure 27: Training convergence for the multivariate Gaussian moment transformer (top) and Quantile Transformer (bottom) .

Table 10: Hyperparameter search space and selected values for the comparison with Vandergrift et al. (2026).

Hyperparameter	Candidate values	Quantile Transformer	M2M-W_2 Transformer
Peak learning rate	3×10^{-5}, 10^{-4}, 3×10^{-4}, 5×10^{-4}, 10^{-3}	3×10^{-5}	5×10^{-4}
Hidden dimension	Fixed	512	512
Transformer blocks	Fixed per model	4	5

H.8 Additional experiments: comparison with the M2M models of Vandergrift et al. (2026)

Dynamical system parameters. We compare the models on a Gaussian-preserving McKean–Vlasov system in dimension $d = 3$. Its marginals $\mu_t = \mathcal{N}(m_t, \Sigma_t)$ satisfy

$$\begin{cases} \dot{m}_t &= -0.4 m_t + 0.35 \tanh(B m_t) + 0.15 \tanh(\mathrm{diag}(\Sigma_t) - \mathbf{1}), \\ \dot{\Sigma}_t &= A_t \Sigma_t + \Sigma_t A_t^\top + 0.09 I_3, \\ A_t &= -0.6 I_3 + 0.2 \tanh\left(\frac{\mathrm{tr}\,\Sigma_t}{3} - 1\right) H + 0.1 \tanh\left(\frac{\|m_t\|^2}{3}\right) K. \end{cases} \tag{248}$$

Here $\tanh$ acts componentwise where appropriate. The matrices B, H, K are generated once from independent standard Gaussian matrices using seed 100, symmetrizing H, K and normalizing each matrix to unit operator norm. They are shared across all trajectories and data splits.

Data generation. We retain the training-trajectory count, number of observed marginals and stored-particle count of Vandergrift et al. (2026, Section 5.2 and Appendix A.2.2), using the Gaussian dynamics above. Initial means are uniform on $[-1.5, 1.5]^3$; initial covariance eigenvalues are independently uniform on $[0.5, 1.5]$, with independent orientations obtained by sign-corrected QR decomposition of standard Gaussian matrices. We integrate the moment equations in double precision by RK4, using 20 substeps between observation times $t_j = 5j/99$. Particles are sampled independently from the resulting Gaussian marginals, including independently across observation times.

Validation and test sets. We use ten validation and ten test trajectories, each containing 100 marginals with 500 stored particles per marginal. Checkpoint validation uses a fixed bank of $10 \times 99 \times 4 = 3{,}960$ adjacent-measure pairs, with 128 particles per measure: four subsamples for each validation transition. Validation and test rollouts start from 128-particle initial clouds and are evaluated against independently sampled 128-particle reference clouds at each forecast time. Initial and reference clouds are fixed and shared across methods.

Hyperparameter tuning. We tune only the learning rate, independently for each model, over $\{3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$. M2M-$W_2$ candidates receive 15,000 updates, whereas

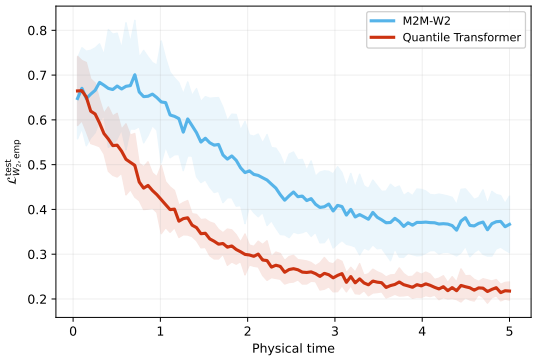

Figure 28: Performance of the Quantile Transformer and the M2M-W2 model of Vandergrift et al. (2026).

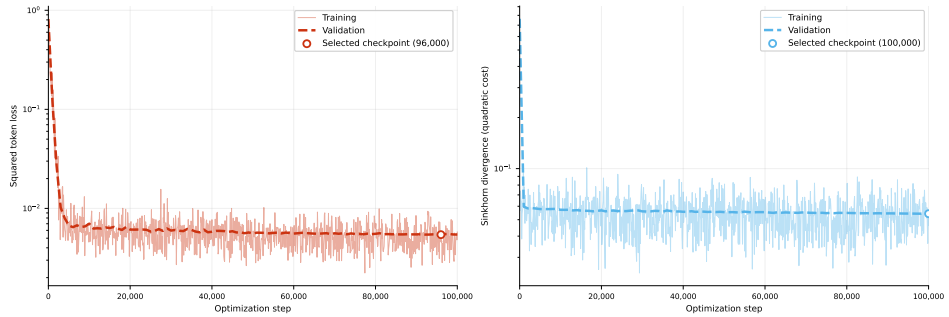

Figure 29: Training convergence for the quantile transformer (left) and M2M-W2 transformer (bottom) .

quantile candidates receive 100,000 updates. Every 1,000 updates, each candidate is assessed using its own training objective on the fixed validation bank. Its best checkpoint is then ranked by exact empirical validation-rollout W_2. The selected rates are 5×10^{-4} for M2M-W_2 and 3×10^{-5} for the quantile Transformer. Selected hyperparameters are reported in Table 10.

Baselines. We reconstruct the static M2M-W_2 method of Vandergrift et al. (2026, Section 3.3), applying their Table 8 static architecture to this objective and our three-dimensional data. It maps the current particle cloud directly to the next cloud. For the GeomLoss training objective, we use the debiased $p = 2$ Sinkhorn divergence with ground cost $\|x - y\|^2/2$, blur 0.05, scaling 0.9, and the tensorized backend, without taking a square root. Our reconstruction uses learned constant modulation, zero-initialized modulation and output layers, and sampling with replacement from exact minibatch OT plans.

Our quantile Transformer has width 512, four blocks, eight heads, feedforward multiplier 4, no dropout, and a maximum context of 15 measures. Its six identifying directions give 768-dimensional tokens normalized by $1/\sqrt{768}$. We retain projection scaling and the balanced general elliptical decoder from Section 4, with $R_M = \lfloor\sqrt{128}\rfloor = 11$ radial groups. Both methods recursively forecast from the same initial 128-particle cloud. M2M feeds back predicted particles; the quantile Transformer feeds back predicted tokens and decodes each forecast. This token-feedback policy was adopted following an earlier test inspection and a validation diagnostic; learning-rate selection originally used decoded-cloud feedback.

Evaluation. We report $\mathcal{L}^{\text{test}}_{W_2,\text{emp}}$, computed by exact minimum-cost assignment between predicted and independent reference clouds of 128 equally weighted particles, followed by a square root. Curves average over the ten test trajectories; aggregate scores additionally average over the 99 forecast marginals, excluding initialization. For the quantile model, distances are first averaged over ten fixed balanced angular constructions, without pooling particles. Shading denotes one standard deviation across test trajectories after this construction averaging.

Training. M2M-W_2 retains the optimizer, batching, full training budget and linear learning-rate schedule of Vandergrift et al. (2026, Appendices A.2.2, A.6 and A.8); after screening, it is retrained from initialization at the selected rate. The quantile model uses equally sized batches of observed contexts, scoring only the next-token prediction at each sampled endpoint. It minimizes squared token error using AdamW with weight decay 10^{-4}, excluding biases and normalization parameters, gradient clipping at 1, 10,000 warm-up updates and cosine decay to 10% of the peak rate. Its winning tuning checkpoint is reused without retraining. Both models use initialization seed 0 and FP32 training on an NVIDIA L4; test inference is performed on MPS. The selected checkpoints occur at updates 100,000 and 96,000 for M2M-W_2 and the quantile model, respectively.

Model performance. The quantile Transformer achieves mean empirical test $W_2 = 0.3226$, compared with 0.4807 for static M2M-W_2, a reduction of 32.9%. These results concern one training seed per model; the shaded bands quantify variation across test trajectories. Results are reported in Figure 28, for rolling out system (248) over five physical time steps.