SINKHORN BASED ASSOCIATIVE MEMORY RETRIEVAL USING SPHERICAL HELLINGER KANTOROVICH DYNAMICS

Aratrika Mustafi*
Department of Statistics,
Pennsylvania State University
abm6733@psu.edu

Soumya Mukherjee*
Department of Statistics,
Pennsylvania State University
szm6510@psu.edu

ABSTRACT

We propose a dense associative memory for empirical measures (weighted point clouds). Stored patterns and queries are finitely supported probability measures, and retrieval is defined by minimizing a Hopfield-style log-sum-exp energy built from the debiased Sinkhorn divergence. We derive retrieval dynamics as a spherical Hellinger Kantorovich (SHK) gradient flow, which updates both support locations and weights. Discretizing the flow yields a deterministic algorithm that uses Sinkhorn potentials to compute barycentric transport steps and a multiplicative simplex reweighting. Under local separation and PL-type conditions we prove basin invariance, geometric convergence to a local minimizer, and a bound showing the minimizer remains close to the corresponding stored pattern. Under a random pattern model, we further show that these Sinkhorn basins are disjoint with high probability, implying exponential capacity in the ambient dimension. Experiments on synthetic Gaussian point-cloud memories demonstrate robust recovery from perturbed queries versus a Euclidean Hopfield-type baseline.

1 INTRODUCTION

Associative memories are a class of models designed to store and retrieve information through patterns of association instead of recalling data by specifying its location. These models frame retrieval as a dynamical process : given a partial or corrupted query, the system evolves toward an attractor that represents a stored pattern. The query evolves through a dynamical process that decreases an energy whose local minima encode the memories. This viewpoint goes back to Hopfield's classical construction, where one designs recurrent interactions that induce a Lyapunov (energy) function whose local minima correspond to memories, yielding content-addressable recall from partial or corrupted cues (Hopfield (1982)). More recently, some modern Hopfield style models (Krotov & Hopfield (2016),Ramsauer et al. (2020)) use the soft-min/ log-sum-exp (LSE) energy to create high-capacity retrieval dynamics and improve efficiency. In the continuous state setting, these retrieval updates can be related to the attention mechanism used in transformer architectures, suggesting a broad role for energy-based retrieval as a reusable computational primitive inside learned systems.

A key limitation of much of the associative-memory literature is that patterns are typically vectors. A natural next step is to move beyond vector-valued patterns and treat probability measures as the objects being stored and retrieved. This shift is motivated by the growing role of distributional representations in modern learning, where uncertainty, multimodality, and population-level structure are often more naturally encoded by measures than by single points. In this direction, Tankala & Balasubramanian (2026) develop a distributional dense associative memory for Gaussian distributions in the 2-Wasserstein (Bures-Wasserstein) geometry, defining a log-sum-exp energy over stored distributions and deriving a retrieval dynamics that aggregates optimal transport maps in a Gibbs-weighted manner; their stationary points correspond to self-consistent Wasserstein barycenters. Many practical settings, however, are not well captured by a single Gaussian, instead distributions are often represented nonparametrically through weighted samples - for example, weighted point clouds, histograms, particle approximations of posteriors, or sets of learned feature vectors, suggesting the need for an associative-memory retrieval principle that operates directly on discrete measures.

In this paper we pursue this analogous extension of dense-associative-memory retrieval when both stored patterns and *queries are empirical measures (weighted point clouds)*, using a log-sum-exp energy built from the debiased Sinkhorn divergence. Sinkhorn divergences provide a computationally tractable way to compare discrete measures and yield deterministic barycentric projections from entropic optimal-transport couplings. A key issue is that transport only (Wasserstein) dynamics moves support locations while keeping the query weights fixed - an intrinsic limitation when the weights carry information such as saliency, mixture proportions, or coarse discretization effects. To evolve both support points and weights in a principled way while staying within probability measures, we therefore use a *spherical Hellinger Kantorovich* (SHK) dynamics [see Appendix D, Liero et al. (2018)], which couples Wasserstein (aka Kantorovich) transport with a spherical Hellinger-type reaction component. In the empirical setting, this leads to a fully deterministic retrieval operator that updates support points via Sinkhorn barycentric maps and updates weights via a multiplicative reweighting on the simplex. This combination is attractive because it aligns the retrieval dynamics with the underlying geometry of measures while allowing both *where mass* is located and *how much mass* is assigned to each particle, to adapt during recall. With this objective in mind, our goal is to construct a dense associative memory (DAM) capable of storing discrete measures and accurately recovering the corresponding distribution when provided with perturbed/noisy queries.

Given a collection of finitely supported discrete distributions (stored patterns) $\{X_i\}_{i=1}^N$ and a query finitely supported discrete distribution ξ, we define the LSE functional using the Sinkhorn divergence S_ε (defined in 2),

$$E(\xi) \;=\; -\frac{1}{\beta}\log\left(\sum_{i=1}^{N}\exp\big(-\beta\, S_\varepsilon(\xi, X_i)\big)\right), \tag{1}$$

with inverse temperature $\beta > 0$. This is the functional we aim to minimize along the SHK gradient flow. To this end, we develop a practically implementable algorithm (SinkhornAlgo) to carry out the optimization. In parallel, we establish theoretical guarantees ensuring convergence and near-accurate retrieval, and we present empirical evidence to validate these findings.

- **Local convergence and stability guarantees in SHK geometry -** Under standard local assumptions (margin separation, SHK smoothness, bounded gradients, and a local PL inequality), we prove descent and contraction properties for the softmin energy, quantify basin interference through separation margins, and establish geometric convergence of SHK gradient descent iterates to the unique local minimizer within a basin. We also bound the deviation between the basin minimizer and the corresponding stored pattern and provide a local stability bound showing stored patterns are approximate fixed points of one-step retrieval.
- **Separation results under a random pattern model.** We introduce a sampling mechanism for generating general M-atom measures (random supports and weights) and prove that, with high probability, the induced Sinkhorn neighborhoods around stored patterns are pairwise disjoint. This yields an exponential-in-dimension scaling of the number of storable patterns under the stated conditions.
- **Empirical evidence against a Euclidean Hopfield-type baseline-** On synthetic point-cloud patterns sampled from Gaussians, we compare our SinkhornSHK retrieval against a vectorized Euclidean Hopfield-style baseline and observe robust recovery from noisy queries, particularly in regimes where Euclidean similarity is ambiguous but distributional (OT/Sinkhorn) geometry remains discriminative.

2 Setting and notations

Let $\Omega \subset \mathbb{R}^d$ be open, bounded and convex with diameter $D := \sup_{x,y\in\Omega}\|x-y\| < \infty$ and boundary $\partial\Omega$, and we equip it with the usual Euclidean geometry induced by the ℓ_2 norm. For any $\omega \in \Omega$ and any set $S \subset \mathbb{R}^d$, define $\operatorname{dist}(\omega, S) = \min_{s\in S}\|\omega - s\|$. For any $x = (x_1, \ldots, x_M) \in \Omega^M$, let $d_\partial(x) := \min_{1\le m\le M}\operatorname{dist}(x_m, \partial\Omega)$ denote the distance of x from the boundary of Ω and $\operatorname{sep}(x) := \min_{i\ne j}\|x_i - x_j\|$ denote the minimum pairwise separation between any two components of x. Let

$\mathbb{B}(x,r) \coloneqq \{y \in \Omega : \|y-x\| < r\}$ and its closure $\bar{\mathbb{B}}(x,r) \coloneqq \{y \in \Omega : \|y-x\| \leq r\}$ be the open and closed ball of radius r around $x \in \Omega$. We denote by $\|\cdot\|_1$ and $\|\cdot\|_\infty$ the ℓ^1 and ℓ^∞ norms, respectively and $\|\cdot\|$ will be interpreted in the natural context.

Fix any $M \in \mathbb{N}$, $\frac{1}{M} > a_{\min} > 0$ and $\Delta_{\min} > 0$. Stored patterns are finitely supported discrete measures $X_i = \sum_{m=1}^M b_{i,m}\,\delta_{y_{i,m}}, b_{i,m} > a_{\min} > 0, \sum_{m=1}^M b_{i,m} = 1$, with all the locations $y_{i,m}$'s being separated from each other with minimum pairwise distance $\min_{m\neq k} \|y_{i,m} - y_{i,k}\| > \Delta_{\min}$, and the query is also a finitely supported discrete measures $\xi = \sum_{m=1}^M a_m\,\delta_{x_m}, a_m > a_{\min} > 0, \sum_{m=1}^M a_m = 1$. with all the locations x_m's being pairwise separated with margin $\Delta_{\min}$.

For any $M \in \mathbb{N}$ $\frac{1}{M} > a_{\min} > 0$ and $\Delta_{\min} > 0$, define the space of bounded weights $\Delta^\circ_{M,a_{\min}} \coloneqq \left\{a \in \mathbb{R}^M : a_m > a_{\min}, \sum_{m=1}^M a_m = 1\right\}$ and the space of M pairwise separated locations/atoms $\mathrm{Loc}_{M,\Delta_{\min}}(\Omega) \coloneqq \left\{x = (x_1,\ldots,x_M) \in \Omega^M : \min_{m\neq n}\|x_m - x_n\| > \Delta_{\min}\right\} \in \Delta^\circ_{M,a_{\min}} \times \mathrm{Loc}_{M,\Delta_{\min}}(\Omega)$. We will often consider the unordered particle parameterization of a finitely supported discrete measure with M atoms $(a,x) = \left((a_m)_{m=1}^M, (x_m)_{m=1}^M\right)$. Let $\mathcal{P}(\Omega)$ be the collection of probability measures on Ω and $\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega) = \left\{\xi \in \mathcal{P}(\Omega) : \xi = \sum_{l=1}^M a_l \delta_{x_l} \text{ for } (a_1,\ldots,a_M) \in \Delta^\circ_{M,a_{\min}} \text{ and } (x_1,\ldots,x_M) \in \mathrm{Loc}_{M,\Delta_{\min}}(\Omega)\right\}$.

Throughout the paper, we use the quadratic cost $c(x,y) = \frac{1}{2}\|x-y\|^2$ for defining the Wasserstein, spherical Hellinger Kantorovich distances and the sinkhorn divergence.

For $\varepsilon > 0$, define the entropic OT cost $\mathrm{OT}_\varepsilon(\mu,\nu) \coloneqq \min_{\pi\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} c(x,y)\,d\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi \,||\, \mu\otimes\nu)$, where $\Pi(\mu,\nu)$ denotes couplings with marginals μ,ν.

The debiased *Sinkhorn divergence* is defined as -

$$S_\varepsilon(\mu,\nu) \coloneqq \mathrm{OT}_\varepsilon(\mu,\nu) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(\mu,\mu) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(\nu,\nu). \tag{2}$$

A standard dual formulation of OT_ε uses Schrödinger (entropic OT) potentials $(f_{\mu,\nu}, g_{\mu,\nu})$, defined up to an additive constant. We do not require a specific normalization; only gradients ∇f and differences of potentials will matter. One can define the entropic soft c-transform operator A_ε via an expression of the form

$$A_\varepsilon(g,\nu)(x) \coloneqq -\varepsilon \log \int_\Omega \exp\Big(\frac{g_{\mu,\nu}(y) - c(x,y)}{\varepsilon}\Big)\,d\nu(y) \quad \text{(defined up to an additive constant).}$$

Then the optimal potentials $(f_{\mu,\nu}, g_{\mu,\nu})$ (Schrödinger potentials) satisfy the Schrödinger system

$$f_{\mu,\nu} = A_\varepsilon(g_{\mu,\nu},\nu) \quad \mu\text{-a.e.,} \qquad g_{\mu,\nu} = A_\varepsilon(f_{\mu,\nu},\mu) \quad \nu\text{-a.e..}$$

These potentials are unique up to adding a constant to $f_{\mu,\nu}$ and subtracting the same constant from $g_{\mu,\nu}$ (gauge invariance). This does not affect any gradient $\nabla f_{\mu,\nu}(x)$ or $\nabla g_{\mu,\nu}(x)$, which is what we ultimately use. The optimal entropic coupling has Gibbs form

$$\frac{d\pi^\varepsilon_{\mu,\nu}}{d(\mu\otimes\nu)}(x,y) = \exp\left(\frac{f_{\mu,\nu}(x) + g_{\mu,\nu}(y) - c(x,y)}{\varepsilon}\right).$$

For more details on the Sinkhorn divergence, we refer our readers to Feydy et al. (2019); Hardion & Lavenant (2025)

3 Theoretical Guarantees

In this section, we state all our theoretical results. The detailed proofs are available in Section G. Retrieval guarantees largely rely on ensuring sufficient separation in terms of Sinkhorn divergence among the stored patterns. We propose a sampling mechanism that allows the number of patterns N to be exponentially large in the dimension d and Theorem 1 establishes a high-probability separation guarantee for these N randomly generated stored patterns, ensuring that the associated Sinkhorn neighborhoods (basins) are pairwise disjoint.

Theorem 1 (Exponential storage capacity and high probability separation of patterns)**.** *Fix $d \geq 1, M \geq 2$, $a_{\min} > 0$, $\Delta_{\min} > 0$ and let $\Omega \subset \mathbb{R}^d$ be open, bounded and convex. Assume there exists $c \in \Omega$ and $R > 0$ such that the closed ball $\bar{\mathbb{B}}(c, R) \subset \Omega$. Fix any $\sigma \in (0, R/4)$, such that $\mathcal{Z}_{\sigma,\Delta_{\min}} := \left\{(z_1, \ldots, z_M) \in \mathbb{B}(0,\sigma)^M : \min_{n\neq m} \|z_n - z_m\| > \Delta_{\min}\right\}$ is non-empty, and set $R_0 := R - 2\sigma$. Let $\gamma, p \in (0,1)$ and choose*

$$N := \left\lfloor \sqrt{2p} \exp\left(\frac{\gamma^2}{4} d\right) \right\rfloor.$$

Let $X_1, \ldots, X_N$ be generated by the sampling mechanism described as SampAlgo (see Sec F). Assume $\varepsilon > 0$ is chosen small enough so that

$$\varepsilon \log M < \frac{1-\gamma}{16} R_0^2,$$

and define

$$d_{\min} := \sqrt{2(1-\gamma)} R_0, \quad r := \frac{d_{\min}^2}{32} - \varepsilon \log M, \quad \Delta := \frac{d_{\min}^2}{4}.$$

Then, with probability atleast $1 - p$, the following pairwise separation of Sinkhorn neighbourhoods/basins holds true (i.e. Assumption (A1)):

For each i, for every $\xi \in B_i(r) = \{\nu \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega) : S_\varepsilon(\nu, X_i) \leq r\}$, and every $j \neq i$,

$$S_\varepsilon(\xi, X_j) - S_\varepsilon(\xi, X_i) \geq \Delta.$$

In particular, $B_i(r) \cap B_i(j) = \emptyset, \quad \forall i \neq j$ i.e. the Sinkhorn neighborhoods of X_i's are pairwise disjoint.

The Sinkhorn margin separation property provides the foundation required for reliable retrieval, which ensures that if a query is close to a stored pattern, it is separated enough from other patterns to ensure accurate retrieval. In fact, we can prove retrieval guarantees in much greater generality than afforded by the particular sampling algorithm we propose. We can establish these results under some regularity assumptions regarding the energy functional E and the local Sinkhorn energies $F_i(\cdot) := S_\varepsilon(\cdot, X_i)$, which are presented in Section E.

Given the energy functional E along with a query $\xi \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ and a step-size η, we can define a gradient based evolution of E by equipping the space $\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ of finitely supported M-atom discrete probability measures with a choice of geometry. We choose the spherical Hellinger-Kantorovich (SHK) geometry and the gradient descent based one-step retrieval operator can be defined as

$$\Phi_\eta(\xi) = \mathrm{Ret}_\xi(-\eta \, \mathrm{grad}_{\mathrm{SHK}} \, E(\xi)) \tag{3}$$

where Ret_ξ is a retraction map that evolves the query ξ along the negative SHK gradient of E, given by $\mathrm{grad}_{\mathrm{SHK}} \, E(\xi)$, for a small step controlled by η and ensures that the resulting object is still a probability distribution (in fact, an element of $\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$). We relegate all details to the Appendix.

Theorem 2 (Geometric convergence in Sinkhorn divergence to the local minimizer and local basin invariance of gradient descent iterates)**.** *Let Assumptions (A2), (A3) and (A4) hold. Define the stored pattern margins $w_i := \min_{1\leq m\leq M}(b_{i,m} - a_{\min}) > 0, d_i^\partial := \min_{1\leq m\leq M} \mathrm{dist}(y_{i,m}, \partial\Omega) > 0$ and $s_i := \mathrm{sep}(y_i) = \min_{m\neq n} \|y_{i,m} - y_{i,n}\|_2 > \Delta_{\min}$. Let $\delta_i, \tau_i > 0$ such that $0 < \delta_i < \bar{\delta}_i := \min\left\{d_i^\partial, \frac{s_i - \Delta_{\min}}{2}\right\} > 0$ and $0 < \tau_i < w_i$. Let $r > 0$ be such that $0 < r < r_i^{\mathrm{loc}}(\delta_i, \tau_i) = \min\left\{\frac{a_{\min}\delta_i^2}{2} - \varepsilon \log M, \frac{\tau_i(s_i - \delta_i)^2}{4} - \varepsilon \log M\right\}$. Let $E_i^*(r)$ be the minimum value of $E(\xi)$ in the local basin $B_i(r)$ and $X_i^*(r)$ be a minimizer of E in $B_i(r)$ i.e. $E(X_i^*(r)) = E_i^*(r) := \inf_{\xi\in B_i(r)} E(\xi)$. Finally, define $\eta_{\mathrm{ret},i} := \min\left\{\frac{\lambda^2}{2D^2} \log \frac{\min_m b_{i,m} - \tau_i}{a_{\min}}, \frac{1}{2D} \min\left\{d_i^\partial - \delta_i, s_i - 2\delta_i - \Delta_{\min}\right\}\right\}$.*

Then for any step size $0 < \eta < \min\{\frac{1}{L}, \frac{1}{\mu}, \eta_{\mathrm{ret},i}\}$, the following hold true:

1. *If $\xi^{(k)} \in B_i(r)$ for all $k \in \mathbb{N} \cup \{0\}$, then the sequence of SHK gradient iterates $\left(\xi^{(k)}\right)_{k\geq 0}$ satisfy the explicit geometric bound*

$$S_\varepsilon\left(\xi^{(k)}, X_i^*(r)\right) \leq \frac{GS_\eta \sqrt{2\eta\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}}{1 - (1-\mu\eta)^{\frac{1}{2}}} \cdot (1-\eta\mu)^{\frac{k}{2}},$$

where $S_\eta \leq \min\left\{e^{\frac{\eta D^2}{\lambda^2}}, \frac{1}{\sqrt{a_{\min}}}\right\}$ *and* $G = D\sqrt{1+\frac{D^2}{\lambda^2}}$ *with* $D = \sup_{x,y\in\Omega}\|x-y\|$ *and* λ *being the relative strength scale of the spherical Hellinger component.*

2. *If* $\xi^{(k)} \in B_i(r)$ *for all* $k \in \mathbb{N}\cup\{0\}$*, the sequence of SHK gradient iterates* $\left(\xi^{(k)}\right)_{k\geq 0}$ *converges weakly to* $X_i^*(r)$ *in* $\mathcal{P}(\Omega)$*.*

3. *If* $\xi^{(k)} \in B_i(r)$ *for all* $k \in \mathbb{N}\cup\{0\}$*, then for any* $\delta > 0$*, the error bound* $S_\varepsilon\left(\xi^{(k)}, X_i^*(r)\right) \leq \delta$ *is guaranteed to be achieved once the number of iterations* k *is greater than or equal to* $\frac{\min\left\{\frac{2\eta D^2}{\lambda^2}, -\log a_{\min}\right\}+2\log G+\log(2\eta\left(E\left(\xi^{(0)}\right)-E_i^*(r)\right))+2\log(\frac{1}{\delta(1-\sqrt{1-\mu\eta})})}{-\log(1-\mu\eta)}$*. A simpler sufficient condition on the no. of iterations to achieve the same error bound is* $k \geq \min\left\{\frac{2D^2}{\mu\lambda^2}, -\frac{\log a_{\min}}{\mu\eta}\right\} + \frac{1}{\mu\eta}\log(2\eta\left(E\left(\xi^{(0)}\right)-E_i^*(r)\right)) + \frac{2}{\mu\eta}\log(\frac{2G}{\delta\mu\eta})$*.*

In addition, for $\rho_i(\eta, r, \xi^{(0)}) := G\frac{\sqrt{2\eta\left(E\left(\xi^{(0)}\right)-E_i^*(r)\right)}}{\sqrt{a_{\min}}(1-\sqrt{1-\mu\eta})}$ *and* $\alpha(r,\xi^{(0)}) := r^2 \times \frac{\mu a_{\min}}{2G^2\left(E\left(\xi^{(0)}\right)-E_i^*(r)\right)}$*, if the step-size* η *and the initial iterate* $\xi^{(0)}$ *satisfies the conditions* $\alpha(r,\xi^{(0)}) > 1$*,* $\frac{4\alpha(r,\xi^{(0)})}{\mu(\alpha(r,\xi^{(0)})+1)^2} \leq \eta < \min\{1/L, 1/\mu, \eta_{\mathrm{ret},i}\}$ *and* $S_\varepsilon(\xi^{(0)}, X_i) \leq r - \rho_i(\eta, r, \xi^{(0)})$ *i.e.* $\xi^{(0)} \in B_i(r-\rho_i)$*, then the sequence of SHK gradient descent iterates* $(\xi^{(k)})_{k\geq 0}$ *all belong to the local basin* $B_i(r)$*. Consequently, the above 3 properties hold true without the apriori basin invariance assumption* $\xi^{(k)} \in B_i(r)$ *for all* $k \in \mathbb{N}\cup\{0\}$*.*

Theorem 2 provides the core algorithmic guarantee, showing that the SHK gradient descent dynamics of E converges geometrically (in terms of no. of iterations) to the unique local minimizer within that basin and remain invariant inside it. Together, these results connect statistical separation with dynamical stability, yielding rigorous theoretical guarantees for accurate associative-memory retrieval.

Theorem 3 (Sinkhorn distance between minimizer in local basin and stored pattern and stability of stored pattern)**.** *Let Assumptions (A1) and (A3) hold true. Then, we have that*

1. $S_\varepsilon(X_i^*(r), X_i) \leq \frac{1}{\beta}\log\left(1+(N-1)e^{-\beta\Delta}\right) \leq \frac{N-1}{\beta}e^{-\beta\Delta}$.

2. *for* $\eta \leq \eta_{\mathrm{ret},i}$ *as defined in Theorem 2,* $S_\varepsilon(X_i, \Phi_\eta(X_i)) \leq \frac{\min\left\{e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}}\right\}\eta G^2(N-1)e^{-\beta\Delta}}{1+(N-1)e^{-\beta\Delta}}$.

While Theorem 2 establishes geometric convergence of the SHK gradient descent iterates to the unique local minimizer within a Sinkhorn basin, it does not yet quantify how well this minimizer approximates the original stored pattern. Theorem 3 closes this gap by showing that the basin minimizer remains exponentially close (in terms of the inverse temperature β and margin separation Δ) to the corresponding stored pattern in Sinkhorn divergence, and that each stored pattern is an approximate fixed point of the retrieval operator. Thus, beyond dynamical convergence, Theorem 3 provides a fidelity guarantee: the attractor reached by the algorithm is not merely stable, but provably close to the intended memory, ensuring accurate and stable associative recall.

4 Conclusion

We developed a dense associative memory for empirical measures based on a Sinkhorn log-sum-exp energy and spherical Hellinger Kantorovich gradient dynamics, yielding deterministic transport-reaction retrieval with provable local convergence and separation guarantees. The full retrieval algorithm (pseudo-code and implementation details) along with the numerical experiments is provided in the Appendix (see Sections H and I). Future directions include faster retrieval implementations, adaptive regularization and applying our algorithm on real point clouds.

REFERENCES

Lenaic Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1):487–532, 2022.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2681–2690. PMLR, 2019.

Mathis Hardion and Hugo Lavenant. Gradient flows of potential energies in the geometry of sinkhorn divergences. *arXiv preprint arXiv:2511.14278*, 2025.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf`.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211 (3):969–1117, 2018.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Chandan Tankala and Krishnakumar Balasubramanian. Dense associative memory for gaussian distributions, 2026. URL `https://arxiv.org/abs/2509.23162`.

A First variation of the Sinkhorn divergence

Let $F_\nu(\mu) := S_\varepsilon(\mu,\nu)$ with ν fixed. A first variation $\delta F_\nu/\delta\mu$ is defined (up to an additive constant) by

$$\frac{d}{dt}F_\nu(\mu+t\chi)\Big|_{t=0} = \int_\Omega \left(\frac{\delta F_\nu}{\delta\mu}(\mu)(x)\right) d\chi(x), \qquad \int d\chi = 0.$$

It is standard to see that (derivation available in Hardion & Lavenant (2025))

$$\frac{\delta S_\varepsilon(\mu,\nu)}{\delta\mu} = f_{\mu,\nu} - \tfrac{1}{2}\big(f_{\mu,\mu} + g_{\mu,\mu}\big) \qquad \text{in } C(\Omega)/\mathbb{R}. \tag{4}$$

Intuitively, $\frac{\delta\,\mathrm{OT}_\varepsilon(\mu,\nu)}{\delta\mu} = f_{\mu,\nu}$ and $\frac{\delta\,\mathrm{OT}_\varepsilon(\mu,\nu)}{\delta\nu} = g_{\mu,\nu}$. Since μ appears in *both* marginals of the self term $\mathrm{OT}_\varepsilon(\mu,\mu)$, the first variation of $\frac{1}{2}\,\mathrm{OT}_\varepsilon(\mu,\mu)$ w.r.t. μ is the average potential

$$f^{\mathrm{sym}}_{\mu,\mu} := \tfrac{1}{2}\big(f_{\mu,\mu} + g_{\mu,\mu}\big),$$

which is invariant under the Sinkhorn gauge transformation $(f,g) \mapsto (f+c, g-c)$. When one chooses a symmetric gauge for the self problem (i.e. $f_{\mu,\mu} = g_{\mu,\mu}$), which is possible for symmetric costs), equation 4 reduces to the commonly stated formula $f_{\mu,\nu} - f_{\mu,\mu}$ in $C(\Omega)/\mathbb{R}$.

B First variation of the log-sum-exp energy

Let $Z(\xi) = \sum_{i=1}^N \exp(-\beta S_\varepsilon(\xi, X_i))$ and define Gibbs weights

$$w_i(\xi) := \frac{\exp(-\beta S_\varepsilon(\xi, X_i))}{\sum_{j=1}^N \exp(-\beta S_\varepsilon(\xi, X_j))}. \tag{5}$$

Differentiating equation 1 and using equation 4 gives

$$\frac{\delta E}{\delta \xi}(\xi) = \sum_{i=1}^N w_i(\xi)\Big(f_{\xi,X_i} - \tfrac{1}{2}\big(f_{\xi,\xi}+g_{\xi,\xi}\big)\Big) = \left(\sum_{i=1}^N w_i(\xi) f_{\xi,X_i}\right) - \tfrac{1}{2}\big(f_{\xi,\xi}+g_{\xi,\xi}\big), \qquad \text{in } C(\Omega)/\mathbb{R}. \tag{6}$$

We will denote the centered first variation of the energy functional $E(\cdot)$ as

$$u_\xi := \frac{\delta E}{\delta\xi}(\xi) - \Big\langle \frac{\delta E}{\delta \xi}(\xi), \xi \Big\rangle = \frac{\delta E}{\delta\xi}(\xi) - \int \frac{\delta E}{\delta\xi}(\xi) d\xi$$

where $\langle\cdot,\cdot\rangle$ is the duality pairing.

C From entropic potentials to deterministic barycentric maps

For quadratic cost $c(x,y) = \frac{1}{2}\|x-y\|^2$, the gradient of a Schrödinger potential has the barycentric form

$$\nabla f_{\mu,\nu}(x) = x - T^\varepsilon_{\mu\to\nu}(x), \qquad T^\varepsilon_{\mu\to\nu}(x) := \int y\, \pi^\varepsilon_{\mu,\nu}(dy \mid x), \tag{7}$$

where $\pi^\varepsilon_{\mu,\nu}(dy \mid x)$ is the conditional distribution under the optimal entropic coupling. The full derivation is provided in Section J.3. Thus, using equation 6, the transport velocity becomes

$$v(x) = \sum_{i=1}^N w_i(\xi)\, T^\varepsilon_{\xi\to X_i}(x) - T^\varepsilon_{\xi\to\xi}(x). \tag{8}$$

Even though equation 6 involves the symmetric self potential $\frac{1}{2}(f_{\xi,\xi}+g_{\xi,\xi})$, the *transport* velocity depends only on gradients. For symmetric costs (such as $\frac{1}{2}\|x-y\|^2$) and a self-coupling, the optimal entropic plan is symmetric and one has $\nabla f_{\xi,\xi} = \nabla g_{\xi,\xi}$, so

$$\nabla \tfrac{1}{2}\big(f_{\xi,\xi} + g_{\xi,\xi}\big) = \nabla f_{\xi,\xi},$$

and the self-correction in equation 8 remains the usual barycentric map $T^\varepsilon_{\xi\to\xi}$. This is a deterministic vector field on the query support computed from barycentric projections of Sinkhorn plans.

D SPHERICAL HELLINGER-KANTOROVICH GRADIENT FLOW (CONTINUOUS-TIME)

D.1 TRANSPORT+REACTION BASED CONTINUITY EQUATION

We consider a transport-reaction dynamics of the form

$$\partial_t \xi_t + \nabla \cdot (\xi_t v_t) = \xi_t \, r_t, \tag{9}$$

where v_t is a velocity field (transport) and r_t is a scalar reaction rate (mass reweighting). A SHK gradient flow of E sets

$$v_t(x) = -\nabla \left(\frac{\delta E}{\delta \xi}(\xi_t)(x) \right) = -\nabla u_{\xi_t}(x), \quad r_t(x) = -\frac{1}{\lambda^2} \left(\frac{\delta E}{\delta \xi}(\xi_t)(x) - \left\langle \frac{\delta E}{\delta \xi}(\xi_t), \xi_t \right\rangle \right) = -\frac{1}{\lambda^2} u_{\xi_t}(x), \tag{10}$$

with a scale parameter $\lambda > 0$ controlling the relative strength of the spherical Hellinger component. The subtraction of the mean ensures $\frac{d}{dt} \int d\xi_t = 0$, so probability mass is preserved.

D.2 RIEMMANIAN STRUCTURE INDUCED BY SPHERICAL HELLINGER-KANTOROVICH GEOMETRY ON THE SPACE OF PROBABILITY MEASURES

Let $\xi \in \mathcal{P}(\Omega)$ where $\mathcal{P}(\Omega)$ is the set of all probability measures defined on Ω equipped with the Spherical Hellinger-Kantorovich geometry. A (sufficiently regular) tangent vector at ξ can be represented by a pair (r, v) consisting of a vector field $v : \Omega \to \mathbb{R}^d$ (transport velocity) and a scalar field $r : \Omega \to \mathbb{R}$ (reaction rate), subject to the mass constraint $\int_\Omega r(x) d\xi(x) = 0$

Given such (v, r), the induced infinitesimal change of measure is the distribution $\dot{\xi}$ defined by the weak form

$$\frac{d}{dt} \int_\Omega \varphi d\xi_t = \int_\Omega \nabla \varphi(x) \cdot v(x) d\xi_t(x) + \int_\Omega \varphi(x) r(x) d\xi_t(x) \tag{11}$$

which corresponds to the PDE

$$\partial_t \xi_t + \nabla \cdot (\xi_t v) = \xi_t r. \tag{12}$$

Fix $\lambda > 0$. Define the inner product on the tangent space at ξ by

$$\langle (r, v), (r', v') \rangle_{\mathrm{SHK}, \xi} := \int_\Omega \left(v \cdot v' + \lambda^2 r r' \right) d\xi$$

for pairs satisfying $\int r d\xi = \int r' d\xi = 0$ and this induces the metric tensor at ξ to be

$$g_\xi^{\mathrm{SHK}} \left((r, v), (r', v') \right) = \langle (r, v), (r', v') \rangle_{\mathrm{SHK}, \xi} \, .$$

The induced norm on the tangent space at ξ is

$$\|(r, v)\|_{\mathrm{SHK}, \xi}^2 = \int_\Omega \left(\|v\|_2^2 + \lambda^2 r^2 \right) d\xi.$$

D.3 THE MANIFOLD OF FINITELY SUPPORTED DISCRETE MEASURES EQUIPPED WITH SHK GEOMETRY

For any $M \in \mathbb{N}$, $\frac{1}{M} > a_{\min} > 0$ and $\Delta_{\min} > 0$, define $\mathrm{Loc}_{M, \Delta_{\min}}(\Omega) := \left\{ x = (x_1, \dots, x_M) \in \Omega^M : \min_{m \neq n} \|x_m - x_n\| > \Delta_{\min} \right\} \in \Delta^\circ_{M, a_{\min}} \times \mathrm{Loc}_{M, \Delta_{\min}}(\Omega)$ as the space of M pairwise separated locations/atoms corresponding to the space of finitely supported discrete probability distributions with exactly M atoms and let $\Delta^\circ_{M, a_{\min}} := \left\{ a \in \mathbb{R}^M : a_m > a_{\min}, \sum_{m=1}^M a_m = 1 \right\}$ be the space of bounded probability weights associated with the M atoms. Further, define the associated parameter space (ordered particles) to be

$$\mathcal{M}_M := \Delta^\circ_{M, a_{\min}} \times \mathrm{Loc}_{M, \Delta_{\min}}(\Omega).$$

For any $(a, x) \in \mathcal{M}_M$, we associate to it the discrete measure $\sum_{m=1}^{M} a_m \delta_{x_m}$ through the parametrization mapping $\Xi : \mathcal{M}_M \to \mathcal{P}(\Omega) \subset \left(C^1(\Omega)\right)^*$, defined as

$$\Xi(a, x) \coloneqq \sum_{m=1}^{M} a_m \delta_{x_m} \in \mathcal{P}(\Omega).$$

Here $C^k(\Omega)$ is the class of k-times continuously differentiable functions on the domain Ω, and $\left(C^k(\Omega)\right)^*$ represents the space of all continuous linear functionals on $C^k(\Omega)$ i.e. the dual space of $C^k(\Omega)$. Similarly, $C_c^\infty(\Omega)$ represents the the class of compactly supported infinitely differentiable functions on the domain Ω, with $C_c^\infty(\Omega)$, with its corresponding continuous dual being $(C_c^\infty(\Omega))^*$. Since all of our theory is focused on the discrete measures $\Xi(a, x)$ and functionals defined using such measures, and all such objects are invariant under permutations of the labels associated with the weight-location pairs, we will consider the quotient space $\widetilde{\mathcal{M}}_M = \mathcal{M}_M / \mathfrak{S}_M$ where $\mathfrak{S}_M$ is the symmetric group that acts freely on $\mathcal{M}_M$ by relabeling:

$$\sigma \cdot (a_1, \ldots, a_M, x_1, \ldots, x_M) = \left(a_{\sigma^{-1}(1)}, \ldots, a_{\sigma^{-1}(M)}, x_{\sigma^{-1}(1)}, \ldots, x_{\sigma^{-1}(M)}\right).$$

Since the action is free and $\mathfrak{S}_M$ is finite, the quotient $\widetilde{\mathcal{M}}_M$ is a smooth finite-dimensional manifold, and it identifies canonically with the set of probability measures on Ω having exactly M pairwise separated support points and positive weights.

For computations and developing the theory without introducing extra notational overhead, it is simplest to work on the ordered cover $\mathcal{M}_M$. The mathematical objects of interest are permutation invariant, so the theory developed on $\mathcal{M}_M$ descends to the quotient space $\widetilde{\mathcal{M}}_M$. Hence, from here on we identify $\widetilde{\mathcal{M}}_M$ with $\mathcal{M}_M$ itself, which is equivalent to treating $(a, x) \in \mathcal{M}_M$ as unordered tuples. We also identify (a, x) with $\xi = \Xi(a, x)$ when convenient.

A tangent vector at (a, x) is a pair $(\delta a, \delta x) \in \mathbb{R}^M \times \left(\mathbb{R}^d\right)^M$ with the simplex constraint $\sum_{m=1}^{M} \delta a_m = 0$. Thus

$$T_{(a,x)}\mathcal{M}_M = \left\{(\delta a, \delta x) : \sum_m \delta a_m = 0\right\}$$

For a given $(a, x) \in \mathcal{M}_M$ and $(\delta a, \delta x) \in T_{(a,x)}\mathcal{M}_M$, we now proceed to compute the differential of Ξ at (a, x) in the direction $(\delta a, \delta x)$. Let $(a(t), x(t))$ be any C^1 curve in $\mathcal{M}_M$ such that

$$(a(0), x(0)) = (a, x) \quad \text{and} \quad \frac{d}{dt}(a(t), x(t))\Big|_{t=0} = (\dot{a}(0), \dot{x}(0)) = (\delta a, \delta x).$$

Let us define the induced curve of measures

$$\xi_t \coloneqq \Xi(a(t), x(t)) = \sum_{m=1}^{M} a_m(t) \delta_{x_{m(t)}}.$$

For every test function $\varphi \in C^1(\Omega)$, we have that

$$\int_\Omega \varphi d\xi_t = \sum_{m=1}^{M} a_m(t) \varphi\left(x_m(t)\right).$$

Differentiating at $t = 0$, we obtain

$$\begin{aligned} \frac{d}{dt}\Big|_{t=0} \int_\Omega \varphi d\xi_t &= \sum_{m=1}^{M} \dot{a}_m(0) \varphi\left(x_m\right) + \sum_{m=1}^{M} a_m \nabla\varphi\left(x_m\right) \cdot \dot{x}_m(0) \\ &= \sum_{m=1}^{M} \delta a_m \varphi\left(x_m\right) + \sum_{m=1}^{M} a_m \nabla\varphi\left(x_m\right) \cdot \delta x_m. \end{aligned}$$

Hence

$$d\Xi_{(a,x)}(\delta a, \delta x)$$

is the distribution characterized by

$$\left\langle d\Xi_{(a,x)}(\delta a, \delta x), \varphi \right\rangle = \sum_{m=1}^{M} \delta a_m \varphi\left(x_m\right) + \sum_{m=1}^{M} a_m \nabla \varphi\left(x_m\right) \cdot \delta x_m. \tag{13}$$

Equivalently,

$$d\Xi_{(a,x)}(\delta a, \delta x) = \sum_{m=1}^{M} \delta a_m \delta_{x_m} - \nabla \cdot \left(\sum_{m=1}^{M} a_m \delta x_m \delta_{x_m} \right)$$

in the sense of distributions.

Comparing Equation 13 with the weak form of the SHK evolution PDE in Equation 11, we can uniquely represent the tangent vector (distribution) $d\Xi_{(a,x)}(\delta a, \delta x)$ using any pair (r, v) such that $r(x_m) = \frac{\delta a_m}{a_m}$ and $v(x_m) = \delta x_m$. The mass constraint is automatically satisfied since $\int_\Omega r d\xi = \sum_{m=1}^{M} a_m \frac{\delta a_m}{a_m} = \sum_{m=1}^{M} \delta a_m = 0$. The uniqueness on the support points $x_1, \ldots, x_M$ can be established using Lemma 17.

Now, we define the pullback metric tensor on $\mathcal{M}_M$ at (a, x) induced by $g^{\mathrm{SHK}}_{\Xi(a,x)}$ as

$$g_{(a,x)} \coloneqq \Xi^* g^{\mathrm{SHK}}_{\Xi(a,x)},$$

which means that, for $\zeta = (\delta a, \delta x)$ and $\eta = (\delta a', \delta x')$ belonging to $T_{(a,x)}\mathcal{M}_M$,

$$g_{(a,x)}(\zeta, \eta) = g^{\mathrm{SHK}}_{\xi} \left(d\Xi_{(a,x)}\zeta, d\Xi_{(a,x)}\eta \right), \quad \xi = \Xi(a, x).$$

Using the identification above, we have that

$$r\left(x_m\right) = \frac{\delta a_m}{a_m}, \quad v\left(x_m\right) = \delta x_m, \quad r'\left(x_m\right) = \frac{\delta a'_m}{a_m}, \quad v'\left(x_m\right) = \delta x'_m.$$

Therefore,

$$\begin{aligned} g_{(a,x)}\left((\delta a, \delta x), (\delta a', \delta x')\right) &= \sum_{m=1}^{M} a_m \left(\delta x_m \cdot \delta x'_m + \lambda^2 \frac{\delta a_m}{a_m} \frac{\delta a'_m}{a_m} \right) \\ &= \sum_{m=1}^{M} \left(a_m \delta x_m \cdot \delta x'_m + \frac{\lambda^2}{a_m} \delta a_m \delta a'_m \right). \end{aligned}$$

Consequently, the SHK-induced Riemannian structure on $\mathcal{M}_M$ can be described as follows. Define for $(\delta a, \delta x), (\delta a', \delta x') \in T_{(a,x)}\mathcal{M}_M$:

$$\langle (\delta a, \delta x), (\delta a', \delta x') \rangle_{(a,x)} \coloneqq g_{(a,x)}\left((\delta a, \delta x), (\delta a', \delta x')\right) = \sum_{m=1}^{M} \left(\frac{\lambda^2}{a_m} \delta a_m \delta a'_m + a_m \delta x_m \cdot \delta x'_m \right).$$

The corresponding norm is

$$\|(\delta a, \delta x)\|^2_{(a,x)} = \sum_{m=1}^{M} \left(\frac{\lambda^2}{a_m} \left(\delta a_m\right)^2 + a_m \left\|\delta x_m\right\|_2^2 \right).$$

D.4 Defining an appropriate retraction map for $\mathcal{M}_M$

A local retraction is the standard way to define a “first-order accurate exponential map” used in Riemannian gradient descent. Let $T\mathcal{M}_M$ denote the tangent bundle corresponding to the manifold $\mathcal{M}_M$, given by $T\mathcal{M}_M \coloneqq \sqcup_{(a,x)\in\mathcal{M}_M} T_{(a,x)}\mathcal{M}_M = \left\{ ((a, x), (\delta a, \delta x)) : (a, x) \in \mathcal{M}_M, (\delta a, \delta x) \in T_{(a,x)}\mathcal{M}_M \right\}$, where $T_{(a,x)}\mathcal{M}_M$ is the tangent space at the parameter $(a, x) \in \mathcal{M}_M$. Following the general definition (and consistent with retractions are used in particle evolutions for gradient descent, for e.g. see Chizat (2022)) we use the following definition of a retraction map:

A smooth map Ret : $T\mathcal{M}_M \to \mathcal{M}_M$ is a retraction if for each base point $z = (a, x) \in \mathcal{M}_M$, its restriction $\mathrm{Ret}_z : T_z\mathcal{M}_M \to \mathcal{M}_M$ satisfies (i) $\mathrm{Ret}_z(0) = z$, (ii) $\mathrm{DRet}_z(0) = \mathrm{Id}$ on $T_z\mathcal{M}_M$. It need not be well-defined everywhere but there exists an open set $U_{(a,x)} \subset T_{(a,x)}\mathcal{M}_M$ containing the origin on which it is well-defined.

For our purpose, we will build Ret as a product of a retraction on $\Delta^{\circ}_{M,a_{\min}}$ and one on $\mathrm{Loc}_{M,\Delta_{\min}}(\Omega)$. For the retraction map on $\mathrm{Loc}_{M,\Delta_{\min}}(\Omega)$, define:

$$\mathrm{Ret}^{\mathrm{pos}}_x(\delta x) \coloneqq x + \delta x, \quad \text{i.e.} \quad \mathrm{Ret}^{\mathrm{pos}}_{(x_1,\dots,x_M)}(\delta x_1,\dots,\delta x_M) = (x_1+\delta x_1,\dots,x_M+\delta x_M)\,.$$

and $\mathrm{Ret}^{\mathrm{pos}}_x(\cdot)$ is well-defined on $U_x = \left\{\delta x \in T_x\,\mathrm{Loc}_M(\Omega) : \|\delta x\|_\infty < \frac{1}{2}\min\left(d_\partial(x), \mathrm{sep}(x) - \Delta_{\min}\right)\right\} \subset T_x\,\mathrm{Loc}_M$. Clearly $\mathrm{Ret}^{\mathrm{pos}}_x(0) = x$ and $\mathrm{DRet}^{\mathrm{pos}}_x(0) = \mathrm{Id}$. Therefore, this is a retraction on $\mathrm{Loc}_{M,\Delta_{\min}}(\Omega)$.

Let $a \in \Delta^{\circ}_{M,a_{\min}}$. Its tangent space is

$$T_a\Delta^{\circ}_{M,a_{\min}} = \left\{\delta a \in \mathbb{R}^M : \sum_m \delta a_m = 0\right\}.$$

Define, for $\delta a \in T_a\Delta^{\circ}_{M,a_{\min}}$, the retraction map

$$\mathrm{Ret}^{\mathrm{w}}_a(\delta a) \coloneqq \frac{a \odot \exp(\delta a/a)}{\sum_{j=1}^M a_j \exp\left(\delta a_j/a_j\right)} \tag{14}$$

where $(\delta a/a)_m \coloneqq \delta a_m/a_m$, exp acts coordinatewise and $\odot$ is coordinatewise product. This is well-defined for all $\delta a \in U_a$ where $U_a = \left\{\delta a \in T_a\Delta^{\circ}_{M,a_{\min}} : \left\|\frac{\delta a}{a}\right\|_\infty < \frac{1}{2}\min_m \log\frac{a_m}{a_{\min}}\right\}$.. Further, it can be shown that $\mathrm{Ret}^{\mathrm{w}}_a(\delta a)$ is indeed a retraction map for $\Delta^{\circ}_{M,a_{\min}}$.

Finally, the product retraction on $\mathcal{M}_M$ is defined as

$$\mathrm{Ret}_{(a,x)}(\delta a, \delta x) \equiv \mathrm{Ret}_{\Xi(a,x)}(\delta a, \delta x) \coloneqq \left(\mathrm{Ret}^{\mathrm{w}}_a(\delta a), \mathrm{Ret}^{\mathrm{pos}}_x(\delta x)\right).$$

which is a retraction on $U_{(a,x)} = \left\{(\delta a, \delta x) \in T_{(a,x)}\mathcal{M}_M : \delta a \in U_a \text{ and } \delta x \in U_x\right\} \subset T_{(a,x)}\mathcal{M}_M$.

D.5 Riemannian gradient of the energy functional E on $\mathcal{M}_M$

Let $\xi = \Xi(a,x)$. For discrete ξ, write $u_m \coloneqq u_\xi\left(x_m\right)$.

Differential of E in particle coordinates : Consider a tangent perturbation $(\delta a, \delta x)$. Assume that the functional E defined on $\mathcal{P}(\Omega)$ admits a $C^1(\Omega)$ first variation $\frac{\delta E}{\delta \xi}(\xi)$ for distributions of interest $\xi \in \mathcal{P}(\Omega)$, whose centered version is $u_\xi \coloneqq \frac{\delta E}{\delta \xi}(\xi) - \langle \frac{\delta E}{\delta \xi}(\xi), \xi\rangle = \frac{\delta E}{\delta \xi}(\xi) - \int_\Omega \frac{\delta E}{\delta \xi}(\xi) d\xi$. Now, for a given $(a,x) \in \mathcal{M}_M$ and $(\delta a, \delta x) \in T_{(a,x)}\mathcal{M}_M$, $(a(t), x(t))$ be any C^1 curve in $\mathcal{M}_M$ such that

$$(a(0), x(0)) = (a, x) \quad \text{and} \quad \frac{d}{dt}(a(t), x(t))\Big|_{t=0} = (\dot a(0), \dot x(0)) = (\delta a, \delta x).$$

and let the induced curve of measures be

$$\xi_t \coloneqq \Xi(a(t), x(t)) = \sum_{m=1}^M a_m(t)\delta_{x_{m(t)}}.$$

Along the curve of measures ξ_t, by the chain rule, we have that the differential of $E \circ \Xi$ is given by

$$\begin{aligned} d\left(E \circ \Xi\right)_{(a,x)}\left[\delta a, \delta x\right] =& \frac{d}{dt}\left(E \circ \Xi\right)\left(a(t), x(t)\right)\Big|_{t=0} \\ =& \frac{d}{dt} E\left(\xi_t\right)\Big|_{t=0} \\ =& \left\langle \frac{\delta E}{\delta \xi}(\xi_0), \dot{\xi}_0 \right\rangle \\ =& \left\langle \frac{\delta E}{\delta \xi}(\xi_0), d\Xi(a,x)(\delta a, \delta x) \right\rangle \\ =& \left\langle \frac{\delta E}{\delta \xi}(\xi_0), \sum_{m=1}^{M} \delta a_m \delta_{x_m} - \nabla \cdot \left(\sum_{m=1}^{M} a_m \delta x_m \delta_{x_m} \right) \right\rangle \\ =& \sum_{m=1}^{M} \frac{\delta E}{\delta \xi}(\xi_0)\left(x_m\right) \delta a_m + \sum_{m=1}^{M} a_m \nabla \frac{\delta E}{\delta \xi}(\xi_0)\left(x_m\right) \cdot \delta x_m .. \end{aligned} \tag{15}$$

We note that, since $\sum_{m=1}^{M} \delta_m = 0$, one can replace $\frac{\delta E}{\delta \xi}(\xi_0)$ by $\frac{\delta E}{\delta \xi}(\xi_0) + c$ in Equation 15 for any constant c without changing the result, which is often termed as gauge-invariance. In particular,

$$d\left(E \circ \Xi\right)_{(a,x)}\left[\delta a, \delta x\right] = \sum_{m=1}^{M} u_\xi\left(x_m\right) \delta a_m + \sum_{m=1}^{M} a_m \nabla u_\xi\left(x_m\right) \cdot \delta x_m. \tag{16}$$

We will often abuse notation by using the shorthand $\xi \equiv \Xi[a, x]$ and treating E as a functional directly over $\mathcal{M}_M$, in which case we will represent the differential of E as

$$dE(\xi)[\delta a, \delta x] = \sum_{m=1}^{M} u_\xi\left(x_m\right) \delta a_m + \sum_{m=1}^{M} a_m \nabla u_\xi\left(x_m\right) \cdot \delta x_m$$

Riemannian gradient of E : The Riemannian gradient $\mathrm{grad}_{\mathrm{SHK}}\, E(a,x) \in T_{(a,x)}\mathcal{M}_M$ is the unique tangent vector (β, ζ) such that for all $(\delta a, \delta x)$,

$$g_{(a,x)}\left(\left(\beta, \zeta\right), \left(\delta a, \delta x\right)\right) = \left\langle\left(\beta, \zeta\right), \left(\delta a, \delta x\right)\right\rangle_{(a,x)} = dE(\xi)[\delta a, \delta x].$$

With $\xi \equiv \Xi(a, x)$, define $t_{\xi,m} = \frac{\delta E}{\delta \xi}(\xi)(x_m) + c$ for any constant $c \in \mathbb{R}$. Then, we have that

$$\sum_m \left(\frac{\lambda^2}{a_m} \beta_m \delta a_m + a_m \zeta_m \cdot \delta x_m \right) = \sum_m \left(t_{\xi,m} \delta a_m + a_m \nabla t_{\xi,m} \cdot \delta x_m \right).$$

Matching the δx_m terms gives

$$\zeta_m = \nabla t_{\xi,m} = \nabla u_\xi\left(x_m\right).$$

For weights, note δa is constrained by $\sum_m \delta a_m = 0$. The identity

$$\sum_m \left(\frac{\lambda^2}{a_m} \beta_m - t_{\xi,m} \right) \delta a_m = 0 \quad \forall \delta a : \sum_m \delta a_m = 0$$

holds if and only if the coefficients

$$k_m \coloneqq \frac{\lambda^2}{a_m} \beta_m - t_{\xi,m}$$

are all equal to the same constant k, which can be verified by taking $\delta a = e_i - e_j$ for all $i \neq j$. Hence

$$\frac{\lambda^2}{a_m} \beta_m = t_{\xi,m} + k.$$

Since, we must have $\sum_{m=1}^{M} \beta_m = 0$, we must have that

$$0 = \sum_{m=1}^{M} \beta_m = \frac{1}{\lambda^2} \sum_{m=1}^{M} a_m \left(t_{\xi,m} + k\right) = \frac{1}{\lambda^2}(\bar{t} + k),$$

where

$$\bar{t} \coloneqq \sum_{m=1}^{M} a_m t_{\xi,m}.$$

Therefore $k = -\bar{t}$, so

$$\beta_m = \frac{a_m}{\lambda^2} \left(t_{\xi,m} - \bar{t}\right) = \frac{a_m}{\lambda^2} u_m = \frac{a_m}{\lambda^2} u_\xi(x_m).$$

So the Riemannian gradient of E is

$$\operatorname{grad}_{\mathrm{SHK}} E(a, x) = \left(\left(\frac{a_m}{\lambda^2} u_m \right)_{m=1}^{M}, (\nabla u_m)_{m=1}^{M} \right).$$

D.6 SHK GRADIENT DESCENT UPDATES IN TERMS OF RETRACTION MAPS

Fix a step size $\eta > 0$. The retraction-based SHK gradient descent update is as follows. Given $\xi^{(k)} = \Xi\left(a^{(k)}, x^{(k)}\right)$, define

$$\left(a^{(k+1)}, x^{(k+1)}\right) \coloneqq \operatorname{Ret}_{\left(a^{(k)}, x^{(k)}\right)} \left(-\eta \operatorname{grad}_{\mathrm{SHK}} E\left(a^{(k)}, x^{(k)}\right)\right) = \operatorname{Ret}_{\left(a^{(k)}, x^{(k)}\right)} \left(-\eta \operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right),$$

and set

$$\xi^{(k+1)} \coloneqq \Xi\left(a^{(k+1)}, x^{(k+1)}\right)$$

We can express the SHK gradient descent update using the operator $\Phi_\eta\left(\xi^{(k)}\right) = \xi^{(k+1)}$.

For a measurable map $T : \Omega \to \Omega$, the pushforward $T_\# \xi$ is defined by

$$\left(T_\# \xi\right)(A) = \xi\left(T^{-1}(A)\right), \quad \text{equivalently} \quad \int \varphi d\left(T_\# \xi\right) = \int \varphi \circ T d\xi.$$

For a measurable function $\rho : \Omega \to (0, \infty)$, the reweighted measure $\rho\xi$ is defined by

$$(\rho\xi)(A) \coloneqq \int_A \rho(x) d\xi(x).$$

For a given ξ, define the pushforward map

$$T_\eta^\xi(x) \coloneqq x - \eta \nabla u_\xi(x).$$

and the reweighting map

$$\rho_\eta^\xi(x) \coloneqq \frac{\exp\left(-\frac{\eta}{\lambda^2} u_\xi(x)\right)}{\int_\Omega \exp\left(-\frac{\eta}{\lambda^2} u_\xi(z)\right) d\xi(z)}.$$

Then the SHK gradient descent update is

$$\Phi_\eta(\xi) = \left(T_\eta^\xi\right)_\# \left(\rho_\eta^\xi \xi\right).$$

D.7 Particle (empirical-measure) dynamics

Let $\xi_t = \sum_{m=1}^{M} a_m(t)\delta_{x_m(t)}$. Plugging this ansatz into equation 9 yields the Lagrangian system

$$\dot{x}_m(t) = v_t(x_m(t)), \qquad \dot{a}_m(t) = a_m(t)\, r_t(x_m(t)), \qquad m = 1, \dots, M. \tag{17}$$

Define the particle-wise first-variation values

$$z_m \;:=\; \left(\frac{\delta E}{\delta \xi}(\xi)\right)(x_m) = \sum_{i=1}^{N} w_i(\xi)\Big(f_{\xi,X_i}(x_m) - \tfrac{1}{2}\big(f_{\xi,\xi}(x_m) + g_{\xi,\xi}(x_m)\big)\Big), \tag{18}$$

and their ξ-average $\bar{z} := \sum_{m=1}^{M} a_m z_m$. Then equation 10 gives the weight dynamics

$$\dot{a}_m(t) = -\frac{1}{\lambda^2}\, a_m(t)\,\big(z_m - \bar{z}\big), \qquad \sum_m a_m(t) = 1. \tag{19}$$

Equation equation 19 is the (mass-preserving) *replicator* equation and can be viewed as a spherical Hellinger / natural-gradient flow on the simplex.

E Assumptions for analysis of the Sinkhorn and SHK gradient flow based retrieval

In this section, we list down some useful assumptions are useful to prove certain properties of the Spherical Hellinger-Kantorovich gradient descent. Define $F_i(\xi) = S_\varepsilon(\xi, X_i)$ for $= 1, \dots, N$.

Assumption (A1) (Margin separation)**.** *The stored patterns $X_1, \dots, X_N$ and ε are such that there exists a radius $r > 0$ and a corresponding $\Delta > 0$ such that for any $\xi \in B_i(r) := \{\mu \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega) : S_\varepsilon(\mu, X_i) \le r\}$, we have that, for all $j \neq i$,*

$$F_j(\xi) - F_i(\xi) \ge \Delta. \tag{20}$$

Assumption (A2) (Local Retraction L-smoothness of energy functional E in SHK geometry)**.** *For the same choice of r as in Assumption (A1), there exists $L > 0$ such that for every $\xi \in B_i(r)$, every tangent vector $(w, v) \in T_\xi(w, v)$ and every $\eta > 0$ small enough so that $\mathrm{Ret}_\xi(\eta(w, v))$ is well-defined, we have*

$$E\left(\mathrm{Ret}_\xi(\eta(r, v))\right) \le E(\xi) + \eta \langle \mathrm{grad}_{\mathrm{SHK}}\, E(\xi), (r, v)\rangle_{\mathrm{SHK},\xi} + \frac{L\eta^2}{2}\|(r, v)\|^2_{\mathrm{SHK},\xi} \tag{21}$$

where $\langle \cdot, \cdot \rangle_{\mathrm{SHK},\xi}$ and $\|\cdot\|_{\mathrm{SHK},\xi}$ is the SHK inner product and norm respectively .

Let $E_i^*(r) := \inf_{\xi \in\, B_i(r)} E(\xi)$. We will often denote $E_i^*(r)$ using E_i^* for convenience.

Assumption (A3) (Existence and uniqueness of minimizer of E in local basin)**.** *For the same choice of r as in Assumption (A1), there exists a minimizer $X_i^*(r) \in B_i(r)$ of E i.e. $E(X_i^*(r)) = E_i^*(r) := \inf_{\xi \in B_i(r)} E(\xi)$. Further the minimizer $X_i^*(r)$ is unique.*

Assumption (A4) (PL inequality in local basin)**.** *For the same choice of r as in Assumption (A1) and conditional on Assumption (A3) being true, there exists $\mu > 0$ such that for every $\xi \in B_i(r)$,*

$$\frac{1}{2}\|\,\mathrm{grad}_{\mathrm{SHK}}\, E(\xi)\|^2_{\mathrm{SHK},\xi} \ge \mu\left(E(\xi) - E_i^*(r)\right) \tag{22}$$

F Sampling Algorithm to ensure high probability separation of measures in $\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$

We construct a sampling model that (i) produces fully general M-atom measures (random weights, random supports) that belong to $\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$, and (ii) ensures pattern separation with high probability for number of patterns N exponentially large in d. For simplicity and concreteness, we assume that the domain Ω is such that there exists some point $c \in \Omega$ and some $R > 0$ such that the Euclidean ball

$$\bar{\mathbb{B}}(c, R) := \{x \in \Omega : \|x - c\| \le R\} \subset \Omega.$$

Sampling algorithm (SampAlgo) : Consider the domain Ω such that $\mathbb{B}(c, R) \coloneqq \{x : \|x - c\| \leq R\} \subseteq \Omega \subset \mathbb{R}^d$. Consider the shape radius $\sigma \in (0, \frac{R}{4})$, margin parameter $\gamma \in (0, 1)$ and inradius parameter $R_0 = R - 2\sigma$. Fix any $\varepsilon > 0$ such that $\varepsilon \log M < \frac{1-\gamma}{16} R_0^2$. For each pattern $i = 1, \ldots, N$,

1. **Generate random mean using Rademacher random variables :** Sample $s_i \in \{\pm 1\}^d$ with i.i.d. coordinates, $\mathbb{P}(s_{i,k} = +1) = \mathbb{P}(s_{i,k} = -1) = 1/2$. Define $\mu_i \coloneqq c + \frac{R_0}{\sqrt{d}} s_i$.
2. **Generate random weights :** Sample $b_i = (b_{i,1}, \ldots, b_{i,M})$ in i.i.d manner using any probability distribution supported on $\Delta^\circ_{M, a_{\min}}$ for some fixed $a_{\min} > 0$.
3. **Sample random point cloud around means :** Sample $z_i = (z_{i,1}, \ldots, z_{i,M}) \in (\mathbb{R}^d)^M$ from any probability distribution supported in $\mathcal{Z}_{\sigma, \Delta_{\min}} \coloneqq \big\{(z_1, \ldots, z_M) \in \mathbb{B}(0, \sigma)^M : \min_{n \neq m} \|z_n - z_m\| > \Delta_{\min}\big\}$.
4. **Set means using mean correction and define support points:** Compute the weighted shape mean $\bar{z}_i \coloneqq \sum_{m=1}^M b_{i,m} z_{i,m}$ and define support points $y_{i,m} \coloneqq \mu_i + (z_{i,m} - \bar{z}_i), \quad m = 1, \ldots, M$.
5. **Define the discrete measure :** Set the pattern as the discrete $X_i \coloneqq \sum_{m=1}^M b_{i,m} \delta_{y_{i,m}} \in \mathcal{P}_{M, a_{\min}, \Delta_{\min}}(\Omega)$

G Proof of main theoretical results

G.1 Proof of Theorem 1

Proof. We first verify that the probability distributions $X_1, \ldots, X_N$ generated by the sampling algorithm SampAlgo (see Sec F) belong to $\mathcal{P}_{M, a_{\min}, \Delta_{\min}}(\Omega)$. First, by construction of Step 2 in SampAlgo (see Sec F), we have that, for $i = 1, \ldots, N$, $b_i \in \Delta^\circ_{M, a_{\min}}$. Next, we prove strict pairwise separation of the support points $y_{i,m}$. For $m \neq n$,

$$y_{i,n} - y_{i,m} = (\mu_i + z_{i,n} - \bar{z}_i) - (\mu_i + z_{i,m} - \bar{z}_i) = z_{i,n} - z_{i,m}.$$

and hence $\|y_{i,n} - y_{i,m}\| = \|z_{i,n} - z_{i,m}\|$ for $i = 1, \ldots, N$. Since $z_i \in \mathcal{Z}_{\sigma, \Delta_{\min}}$, we have that $\min_{m \neq n} \|z_{i,n} - z_{i,m}\| > \Delta_{\min}$. Therefore, we have that

$$\min_{m \neq n} \|y_{i,n} - y_{i,m}\| > \Delta_{\min}.$$

Finally, we prove that every support point $y_{i,m}$ lies in Ω. Since each $z_{i,m} \in \bar{\mathbb{B}}(0, \sigma)$, we have that $\|z_{i,m}\| \leq \sigma$. Further, since $\bar{z}_i = \sum_{m=1}^M b_{i,m} z_{i,m}$ is a convex combination of the $z_{i,m}$, we have that

$$\|\bar{z}_i\| \leq \sum_{m=1}^M b_{i,m} \|z_{i,m}\| \leq \sum_{m=1}^M b_{i,m} \sigma = \sigma.$$

Further,

$$\|\mu_i - c\| = \left\| \frac{R_0}{\sqrt{d}} s_i \right\| = \frac{R_0}{\sqrt{d}} \|s_i\| = \frac{R_0}{\sqrt{d}} \sqrt{d} = R_0$$

Hence, for each m, we have that

$$\|y_{i,m} - c\| \leq \|\mu_i - c\| + \|z_{i,m}\| + \|\bar{z}_i\| \leq R_0 + \sigma + \sigma = R.$$

Therefore, we have that

$$y_{i,m} \in \bar{\mathbb{B}}(c, R) \subset \Omega.$$

Therefore, for $i = 1, \ldots, N$, the generated pattern X_i indeed belongs to $\mathcal{P}_{M, a_{\min}, \Delta_{\min}}(\Omega)$. Now, we compute its mean:

$$\begin{aligned} m(X_i) = \sum_{m=1}^M b_{i,m} y_{i,m} &= \sum_{m=1}^M b_{i,m} (\mu_i + z_{i,m} - \bar{z}_i) \\ &= \mu_i + \sum_{m=1}^M b_{i,m} z_{i,m} - \sum_{m=1}^M b_{i,m} \bar{z}_i \\ &= \mu_i + \bar{z}_i - \bar{z}_i = \mu_i. \end{aligned}$$

Now, define the event

$$A := \left\{ \forall 1 \le i < j \le N, \|\mu_i - \mu_j\| \ge d_{\min} = \sqrt{2(1-\gamma)} R_0 \right\}.$$

choosing $N := \left\lfloor \sqrt{2p} \exp\left(\frac{\gamma^2}{4} d\right) \right\rfloor$, we have that $\binom{N}{2} \exp\left(-\frac{\gamma^2 d}{2}\right) = \frac{N(N-1)}{2} \exp\left(-\frac{\gamma^2 d}{2}\right) \le p$. Therefore, by Lemma 16, we have that

$$\mathbb{P}(A) \ge 1 - p.$$

Now , under the event A, the means μ_i are pairwise $d_{\min}$-separated, Lemma 11 applies. Thus, for every i, every $\xi \in \mathcal{P}_{M, a_{\min}, \Delta_{\min}}(\Omega)$ with $S_\varepsilon(\xi, X_i) \le r$, and every $j \ne i$,

$$S_\varepsilon(\xi, X_j) - S_\varepsilon(\xi, X_i) \ge \Delta = \frac{d_{\min}^2}{4}.$$

This proves the margin-separation statement. It remains to prove pairwise disjointness of the basins. Suppose for contradiction that for some $i \ne j$ there exists

$$\xi \in B_i(r) \cap B_j(r).$$

Since $\xi \in B_i(r)$, the margin-separation statement with indices (i, j) gives

$$S_\varepsilon(\xi, X_j) - S_\varepsilon(\xi, X_i) \ge \Delta.$$

Since $\xi \in B_j(r)$, the same statement with indices (j, i) gives

$$S_\varepsilon(\xi, X_i) - S_\varepsilon(\xi, X_j) \ge \Delta.$$

Adding these two inequalities yields

$$0 \ge 2\Delta,$$

which is impossible since $\Delta = \frac{d_{\min}^2}{4} > 0$. Hence

$$B_i(r) \cap B_j(r) = \emptyset \quad \text{for all } i \ne j.$$

This completes the proof.

□

G.2 Proof of Theorem 2

Proof. Assume first that $\xi^{(k)} \in B_i(r)$ for all $k \ge 0$. By Lemma 1, after relabeling the atoms of each iterate, there is a unique ordered representative

$$z^{(k)} = \left(a^{(k)}, x^{(k)}\right) \in K_i(\delta_i, \tau_i)$$

such that

$$\xi^{(k)} = \Xi\left(z^{(k)}\right).$$

Since $\eta < \eta_{\mathrm{ret},i}$, Lemma 2 implies that the local retraction is well-defined at each iterate.

Let $\xi^{(k+1)} = \Phi_\eta\left(\xi^{(k)}\right)$. For each k, define the one-step retraction curve

$$\gamma_k(t) := \mathrm{Ret}_{\xi^{(k)}}\left(-t\eta\, \mathrm{grad}_{\mathrm{SHK}}\, E\left(\xi^{(k)}\right)\right), \quad t \in [0,1].$$

Then $\gamma_k(0) = \xi^{(k)}$ and $\gamma_k(1) = \xi^{(k+1)}$. Applying Lemma 7 with $(\delta a, \delta x) = -\eta\, \mathrm{grad}_{\mathrm{SHK}}\, E\left(\xi^{(k)}\right)$ in particle coordinates yields

$$\mathrm{Length}(\gamma_k) \le e^{\|\delta a / a\|_\infty} \eta \left\| \mathrm{grad}_{\mathrm{SHK}}\, E\left(\xi^{(k)}\right) \right\|_{\mathrm{SHK}, \xi^{(k)}}.$$

Now we proceed to bound $\|\delta a / a\|_\infty$. For the SHK gradient descent direction, the weight update follows

$$\delta a_m = -\eta \frac{a_m}{\lambda^2} u_{\xi^{(k)}}(x_m) \quad \Longrightarrow \quad \frac{\delta a_m}{a_m} = -\frac{\eta}{\lambda^2} u_{\xi^{(k)}}(x_m)$$

where

$$u_\xi := \frac{\delta E}{\delta \xi}(\xi) - \left\langle \frac{\delta E}{\delta \xi}(\xi), \xi \right\rangle = \frac{\delta E}{\delta \xi}(\xi) - \int \frac{\delta E}{\delta \xi}(\xi) d\xi.$$

Thus

$$\left\| \frac{\delta a}{a} \right\|_\infty \leq \frac{\eta}{\lambda^2} \sup_{x \in \Omega} \left| u_{\xi^{(k)}}(x) \right|.$$

Now, $\frac{\delta E}{\delta \xi}(\xi) = \sum_{j=1}^N w_j(\xi) \frac{\delta S_\varepsilon(\xi, X_j)}{\delta \xi}$ and hence $u_\xi(x) = \sum_{j=1}^N w_j(\xi) \left(\phi_i(x) - \int \phi_i(y) d\xi(y)\right)$ where

$$\phi_i(x) := \left(\frac{\delta S_\varepsilon(\xi, X_i)}{\delta \xi} \right)(x)$$

From Lemma 2 Part (ii), we have that $\sup_{x \in \Omega} |u_\xi(x)| \leq D^2$. Consequently, we have that

$$\left\| \frac{\delta a}{a} \right\|_\infty \leq \frac{\eta D^2}{\lambda^2}.$$

Therefore, we have that

$$\text{Length}(\gamma_k) \leq e^{\eta D^2/\lambda^2} \eta \left\| \text{grad}_{\text{SHK}} E\left(\xi^{(k)}\right) \right\|_{\text{SHK}, \xi^{(k)}}.$$

Now, using Lemmas 7 and 4, we have that

$$\begin{aligned} d_{\text{SHK}}\left(\xi^{(k+1)}, \xi^{(k)}\right) &\leq \text{Length}(\gamma_k) \leq \min\left\{ e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}} \right\} \eta \left\| \text{grad}_{\text{SHK}} E\left(\xi^{(k)}\right) \right\|_{\text{SHK}, \xi^{(k)}} \\ &\leq \min\left\{ e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}} \right\} \eta \cdot \sqrt{\frac{2\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}{\eta}} (1 - \eta\mu)^{\frac{k}{2}} = C q^{\frac{k}{2}} \end{aligned}$$

where $C := \min\left\{ e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}} \right\} \cdot \sqrt{2\eta\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}$ and $q := 1 - \mu\eta \in (0, 1)$.

Now, for integers $m > k$,

$$d_{\text{SHK}}\left(\xi^{(m)}, \xi^{(k)}\right) \leq \sum_{t=k}^{m-1} d_{\text{SHK}}\left(\xi^{(t+1)}, \xi^{(t)}\right) \leq \sum_{t=k}^{m-1} C q^{t/2} \leq C \sum_{t=k}^{\infty} q^{t/2} = C \frac{q^{k/2}}{1 - \sqrt{q}}.$$

As $k \to \infty$, the RHS converges to zero. Therefore, $\left(\xi^{(k)}\right)$ is Cauchy in d_{SHK}.

The set $K_i(\delta_i, \tau_i)$ is compact by Lemma 1. Therefore the sequence $z^{(k)}$ has a Euclidean-convergent subsequence:

$$z^{(k_j)} \to z^\infty = (a^\infty, x^\infty) \in K_i(\delta_i, \tau_i).$$

Define

$$\xi^\infty := \Xi(z^\infty) \in P_M(\Omega)$$

Because $K_i(\delta_i, \tau_i) \subset \Delta^\circ_{M, a_{\min}} \times \text{Loc}_{M, \Delta_{\min}}(\Omega)$, the limit belongs to the parameter space $\Delta^\circ_{M, a_{\min}} \times \text{Loc}_{M, \Delta_{\min}}(\Omega)$.

Using Lemma 1, we have that

$$d_{\text{SHK}}\left(\xi^{(k_j)}, \xi^\infty\right) \leq L_i \left\| z^{(k_j)} - z^\infty \right\|_E \to 0.$$

Hence, the subsequence converges to ξ^∞ in d_{SHK}.

Let $\delta > 0$. Since $\left(\xi^{(k)}\right)$ is d_{SHK}-Cauchy, there exists N such that $d_{\text{SHK}}\left(\xi^{(m)}, \xi^{(n)}\right) < \delta/2 \quad \forall m, n \geq N$. Choose j so large that $k_j \geq N$ and $d_{\text{SHK}}\left(\xi^{(k_j)}, \xi^\infty\right) < \delta/2$. Then, for every $k \geq N$,

$$d_{\text{SHK}}\left(\xi^{(k)}, \xi^\infty\right) \leq d_{\text{SHK}}\left(\xi^{(k)}, \xi^{(k_j)}\right) + d_{\text{SHK}}\left(\xi^{(k_j)}, \xi^\infty\right) < \delta.$$

Therefore, $\xi^{(k)}$ converges to ξ^∞ in d_{SHK}.

Using Lemma 10, each $F_i = S_\varepsilon(\cdot, X_i)$ is G-Lipschitz in d_{SHK}, and hence,

$$\left|F_i\left(\xi^{(k)}\right) - F_i\left(\xi^{\infty}\right)\right| \le G d_{\text{SHK}}\left(\xi^{(k)}, \xi^{\infty}\right) \to 0.$$

Since $F_i\left(\xi^{(k)}\right) \le r$ for all k, $F_i\left(\xi^{\infty}\right) \le r$. Hence, we have that $\xi^{\infty} \in B_i(r)$.

Again, using Lemma 10, we have that

$$\left|E\left(\xi^{(k)}\right) - E\left(\xi^{\infty}\right)\right| \le G d_{\text{SHK}}\left(\xi^{(k)}, \xi^{\infty}\right) \to 0.$$

On the other hand, Lemma 4 gives $E\left(\xi^{(k)}\right) \to E_i^*(r)$. Therefore, we have that $E\left(\xi^{\infty}\right) = E_i^*(r)$. Therefore, ξ^{∞} is a minimizer of E on $B_i(r)$. Under Assumption (A3), by uniqueness, we have that

$$\xi^{\infty} = X_i^*(r)$$

This proves the convergence of the SHK gradient descent iterates in d_{SHK} to $X_i * (r)$.

Now, since $d_{\text{SHK}}\left(\xi^{(m)}, \xi^{(k)}\right) \le C\frac{q^{k/2}}{1-\sqrt{q}}$ and d_{SHK} is continuous, we have that

$$d_{\text{SHK}}\left(X_i^*(r), \xi^{(k)}\right) = d_{\text{SHK}}\left(\xi^{(\infty)}, \xi^{(k)}\right) = \lim_{m\to\infty} d_{\text{SHK}}\left(\xi^{(m)}, \xi^{(k)}\right) \le C\frac{q^{k/2}}{1-\sqrt{q}}.$$

Finally, using Lemma 8 and the fact that $S_\varepsilon\left(X_i^*(r), X_i^*(r)\right) = 0$, we have that

$$S_\varepsilon\left(\xi^{(k)}, X_i^*(r)\right) \le G d_{\text{SHK}}\left(X_i^*(r), \xi^{(k)}\right) \le GC\frac{q^{k/2}}{1-\sqrt{q}}.$$

Consequently, $\lim_{k\to\infty} S_\varepsilon(\xi^{(k)}, X_i^*(r)) = 0$ and since the Sinkhorn divergence metrizes weak convergence (Theorem 1 of Feydy et al. (2019)), we have that $\xi^{(k)}$ converges weakly to $X_i^*(r)$.

Now, $S_\varepsilon\left(\xi^{(k)}, X_i^*(r)\right) \le GC\frac{q^{k/2}}{1-\sqrt{q}} \le \delta$ if

$$\begin{aligned} k \ge & \frac{2\log\left(\frac{\min\left\{e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}}\right\}\cdot G\sqrt{2\eta\left(E\left(\xi^{(0)}\right)-E_i^*(r)\right)}}{\delta(1-\sqrt{1-\mu\eta})}\right)}{-\log(1-\mu\eta)} \\ = & \frac{\min\left\{\frac{2\eta D^2}{\lambda^2}, -\log a_{\min}\right\} + 2\log G + \log(2\eta\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)) + 2\log(\frac{1}{\delta(1-\sqrt{1-\mu\eta})})}{-\log(1-\mu\eta)}. \end{aligned}$$

Using $1 + \log x \le x$ for $0 < x < 1$, we have that $\frac{1}{-\log(1-\mu\eta)} \le \frac{1}{\mu\eta}$. Further, we have that $1 - \sqrt{1-\mu\eta} = \frac{\mu\eta}{1+\sqrt{1-\mu\eta}} \ge \frac{\mu\eta}{2} \implies \frac{1}{1-\sqrt{1-\mu\eta}} \le \frac{2}{\mu\eta}$. Consequently, we have the sufficient condition $k \ge \min\left\{\frac{2D^2}{\mu\lambda^2}, -\frac{\log a_{\min}}{\mu\eta}\right\} + \frac{1}{\mu\eta}\log(2\eta\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)) + \frac{2}{\mu\eta}\log(\frac{2G}{\delta\mu\eta})$.

We will now establish that if

$$F_i\left(\xi^{(0)}\right) \le r - \rho_i(\eta, r, \xi^{(0)}),$$

then $\xi^{(k)} \in B_i(r)$ for every $k \in \mathbb{N}$. We will prove by strong induction that $F_i\left(\xi^{(k)}\right) \le r$ for all k.

Note that, by assumption $F_i\left(\xi^{(0)}\right) \le r - \rho_i(\eta, r, \xi^{(0)}) < r$.

Now, assume that $F_i\left(\xi^{(t)}\right) \le r$ for all $t = 0, 1, \ldots, k$. Then $\xi^{(t)} \in B_i(r)$ for those t, so Assumptions (A2) and (A4) apply on each of these iterates, and consequently Lemma 4 applies up to time k, yielding

$$E\left(\xi^{(t)}\right) - E_i^\star(r) \le (1-\mu\eta)^t\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right), \quad t = 0, \ldots, k$$

We now bound the increment of F_i from $\xi^{(t)}$ to $\xi^{(t+1)}$. To do so, we construct an explicit smooth curve in $\mathcal{M}_M$ connecting them and bound its SHK length.

Along any absolutely continuous curve $\zeta(s)$ connecting $\zeta(0) = \xi^{(t)}$ to $\zeta(1) = \xi^{(t+1)}$ restricted to $\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$,

$$\frac{d}{ds}F_i(\zeta(s)) = \left\langle \mathrm{grad}_{\mathrm{SHK}}\, F_i(\zeta(s)), \frac{d}{ds}\zeta(s) \right\rangle_{\mathrm{SHK},\zeta(s)} \leq \|\mathrm{grad}_{\mathrm{SHK}}\, F_i(\zeta(s))\|_{\mathrm{SHK},\zeta(s)} \cdot \left\| \frac{d}{ds}\zeta(s) \right\|_{\mathrm{SHK},\zeta(s)}.$$

Integrating between 0 and 1 and using the global bound $\|\mathrm{grad}_{\mathrm{SHK}}\, F_i(\xi)\|_{\mathrm{SHK},\xi} \leq G$ from Lemma 8, we obtain

$$F_i\left(\xi^{(t+1)}\right) - F_i\left(\xi^{(t)}\right) \leq G \cdot \mathrm{Length}(\zeta).$$

Using Lemmas 4 and 7, we have that

$$\begin{aligned} F_i\left(\xi^{(t+1)}\right) - F_i\left(\xi^{(t)}\right) \leq & G \cdot \frac{\eta}{\sqrt{a_{\min}}} \cdot \sqrt{\frac{2\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}{\eta}}(1-\mu\eta)^{\frac{t}{2}} \\ = & G \cdot \sqrt{\frac{2\eta}{a_{\min}}} \sqrt{\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}(1-\mu\eta)^{\frac{t}{2}}. \end{aligned}$$

Now summing from $t = 0$ to $t = k$:

$$F_i\left(\xi^{(k+1)}\right) \leq F_i\left(\xi^{(0)}\right) + G\sqrt{\frac{2\eta}{a_{\min}}}\sqrt{\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)} \sum_{t=0}^{k}(1-\mu\eta)^{\frac{t}{2}}.$$

Since $\sum_{t=0}^{k}(1-\mu\eta)^{\frac{t}{2}} \leq \sum_{t=0}^{\infty}(1-\mu\eta)^{\frac{t}{2}} = \frac{1}{1-\sqrt{1-\mu\eta}}$,

$$F_i\left(\xi^{(k+1)}\right) \leq F_i\left(\xi^{(0)}\right) + G\frac{\sqrt{2\eta\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}}{\sqrt{a_{\min}}(1-\sqrt{1-\mu\eta})} = F_i\left(\xi^{(0)}\right) + \rho_i(\eta, r, \xi^{(0)}).$$

By the assumed initialization condition $F_i\left(\xi^{(0)}\right) \leq r - \rho_i(\eta, r, \xi^{(0)})$, we get $F_i\left(\xi^{(k+1)}\right) \leq r$. This completes the induction. Thus all iterates remain in $B_i(r)$.

The lower bound on η can be derived based on the natural constraint that $\rho_i(\eta, r, \xi^{(0)}) \leq r$ must be satisfied. The, we must have that

$$G\frac{\sqrt{2\eta\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}}{\sqrt{a_{\min}}(1-\sqrt{1-\mu\eta})} \leq r \iff \frac{2G^2\eta\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}{a_{\min}(1-\sqrt{1-\mu\eta})^2} \leq r^2.$$

Define $u := \sqrt{1-\mu\eta}$. Since $0 < \eta < \frac{1}{\mu}$, we have that $0 < u < 1$. Further, $\eta = \frac{1-u^2}{\mu} = \frac{(1-u)(1+u)}{\mu}$. Therefore, the condition reduces to

$$\frac{1+u}{1-u} \leq r^2 \times \frac{\mu a_{\min}}{2G^2\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)} =: \alpha(r, \xi^{(0)}).$$

Because $0 < u < 1$, we have that $\frac{1+u}{1-u} > 1$. Therefore, a necessary condition is $\alpha(r, \xi^{(0)}) > 1$. Moreover, if $\alpha(r, \xi^{(0)}) > 1$, then the condition is equivalent to

$$1 + u \leq \alpha(r, \xi^{(0)})(1-u) \iff u \leq \frac{\alpha(r, \xi^{(0)}) - 1}{\alpha(r, \xi^{(0)}) + 1} \iff 1 - \mu\eta \leq \left(\frac{\alpha(r, \xi^{(0)}) - 1}{\alpha(r, \xi^{(0)}) + 1}\right)^2.$$

Consequently, the condition reduces to $\alpha(r, \xi^{(0)}) > 1$ and $\eta \geq \frac{1}{\mu} \times \left[1 - \left(\frac{\alpha(r,\xi^{(0)})-1}{\alpha(r,\xi^{(0)})+1}\right)^2\right] = \frac{4\alpha(r,\xi^{(0)})}{\mu(\alpha(r,\xi^{(0)})+1)^2}$.

This completes the proof.

□

G.2.1 AUXILIARY RESULTS FOR PROVING THEOREM 2

For any $i = 1, \ldots, N$ consider the stored pattern $X_i = \sum_{m=1}^{M} b_{i,m}\delta_{y_{i,m}} \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ whose weight and locations parameters are denoted as $b_i := (b_{i,1}, \ldots, b_{i,M})$ and $y_i := (y_{i,1}, \ldots, y_{i,M})$, respectively. Let us define the stored pattern margins

$$w_i := \min_{1\le m\le M} (b_{i,m} - a_{\min}) > 0,$$

$$d_i^{\partial} := \min_{1\le m\le M} \operatorname{dist}(y_{i,m}, \partial\Omega) > 0,$$

$$s_i := \operatorname{sep}(y_i) = \min_{m\neq n} \|y_{i,m} - y_{i,n}\|_2 > \Delta_{\min}.$$

Further, define $\bar{\delta}_i := \min\left\{d_i^{\partial}, \frac{s_i-\Delta_{\min}}{2}\right\} > 0$. Then, for any $0 < \delta < \bar{\delta}_i$ and $0 < \tau < w_i$, let us define

$$r_i^{\text{loc}}(\delta,\tau) := \min\left\{\frac{a_{\min}\delta^2}{2} - \varepsilon\log M, \frac{\tau\,(s_i-\delta)^2}{4} - \varepsilon\log M\right\}. \tag{23}$$

Lemma 1 (local basin compactness inside parameter space)**.** *Fix $i \in \{1, \ldots, N\}$, and choose numbers δ_i, τ_i such that $0 < \delta_i < \bar{\delta}_i$, $\quad 0 < \tau_i < w_i$. Assume $0 < r < r_i^{\text{loc}}(\delta_i, \tau_i)$. Then every $\xi = \sum_{m=1}^{M} a_m\delta_{x_m} \in B_i(r)$ admits, after relabeling of its atoms, a unique ordered representative (a, x) satisfying*

$$\|a - b_i\|_1 \le \tau_i, \quad x_m \in \bar{\mathbb{B}}(y_{i,m}, \delta_i) \quad \textit{for every } m = 1, \ldots, M.$$

Consequently,

$$B_i(r) \subset \Xi\left(K_i(\delta_i, \tau_i)\right),$$

where

$$K_i(\delta_i, \tau_i) := \left\{(a,x) : \sum_{m=1}^{M} a_m = 1, \|a - b_i\|_1 \le \tau_i, x_m \in \bar{\mathbb{B}}(y_{i,m}, \delta_i)\,\forall m\right\},$$

and $K_i(\delta_i, \tau_i)$ is a compact subset of the parameter space $\mathcal{M}_M = \Delta^{\circ}_{M,a_{\min}} \times \mathrm{Loc}_{M,\Delta_{\min}}(\Omega)$.

Proof. From Lemma 11, we have that $\mathrm{OT}_\varepsilon(\xi, \xi) \le \varepsilon\log M$ and $\mathrm{OT}_\varepsilon(X_i, X_i) \le \varepsilon\log M$.

Therefore,

$$S_\varepsilon(\xi, X_i) \ge \mathrm{OT}_\varepsilon(\xi, X_i) - \varepsilon\log M. \tag{24}$$

Now suppose $\xi \in B_i(r)$. Assume for the sake of contradiction that some query atom x_{m_0} does not belong to $\bigcup_{m=1}^{M} \bar{\mathbb{B}}(y_{i,m}, \delta_i)$. Then, for every target atom $y_{i,m}$,

$$\|x_{m_0} - y_{i,m}\|_2 > \delta_i$$

and hence every unit of mass transported out of row m_0 must pay least $\delta_i^2/2$ in transport cost. Since the row mass equals $a_{m_0} \ge a_{\min}$, we must have that

$$\mathrm{OT}_\varepsilon(\xi, X_i) \ge \frac{a_{\min}\delta_i^2}{2}.$$

Using equation 24,

$$S_\varepsilon(\xi, X_i) \ge \frac{a_{\min}\delta_i^2}{2} - \varepsilon\log M.$$

But $r < r_i^{\text{loc}}(\delta_i, \tau_i) \le \frac{a_{\min}\delta_i^2}{2} - \varepsilon\log M$, contradicting $S_\varepsilon(\xi, X_i) \le r$. Hence, each query atom lies in $\bigcup_m \bar{\mathbb{B}}(y_{i,m}, \delta_i)$.

Now, assume for the sake of contradiction that for some n_0, no query atom belongs to $\bar{\mathbb{B}}(y_{i,n_0}, \delta_i)$. Then every unit of mass transported into column n_0 pays at least $\delta_i^2/2$. Since the column mass equals $b_{i,n_0} > a_{\min}$,

$$\mathrm{OT}_\varepsilon(\xi, X_i) \ge \frac{b_{i,n_0}\delta_i^2}{2} \ge \frac{a_{\min}\delta_i^2}{2}.$$

Again equation 24 gives

$$S_\varepsilon(\xi, X_i) \geq \frac{a_{\min}\delta_i^2}{2} - \varepsilon \log M > r,$$

which is a contradiction. So every closed ball $\bar{\mathbb{B}}(y_{i,n}, \delta_i)$ contains at least one query atom.

Since $\delta_i < \bar{\delta}_i \leq \frac{s_i - \Delta_{\min}}{2} < \frac{s_i}{2}$, the balls $\bar{\mathbb{B}}(y_{i,n}, \delta_i)$ are pairwise disjoint. Based on our arguments above, all M query atoms lie in the union of these M disjoint balls, and each ball contains at least one query atom. Since there are exactly M query atoms, each ball contains exactly one. Therefore, after a unique relabeling, we may assume

$$x_m \in \bar{\mathbb{B}}(y_{i,m}, \delta_i), \quad m = 1, \ldots, M. \tag{25}$$

Let $P = (P_{mn})_{m=1,n=1}^{M,M}$ be any coupling matrix between ξ and X_i, i.e.

$$P \geq 0, \quad P\mathbf{1} = a, \quad P^\top \mathbf{1} = b_i.$$

Then, we have that $\sum_{m=1}^{M} P_{mm} \leq \sum_{m=1}^{M} \min(a_m, b_{i,m}) = 1 - \frac{1}{2}\|a - b_i\|_1$ and the off-diagonal mass satisfies

$$\sum_{m \neq n} P_{mn} = 1 - \sum_{m=1}^{M} P_{mm} \geq \frac{1}{2}\|a - b_i\|_1. \tag{26}$$

Now, fix some $m \neq n$. Then, using equation 25, we have that, $\|x_m - y_{i,n}\|_2 \geq \|y_{i,m} - y_{i,n}\|_2 - \|x_m - y_{i,m}\|_2 \geq s_i - \delta_i$. Therefore

$$c(x_m, y_{i,n}) = \frac{1}{2}\|x_m - y_{i,n}\|_2^2 \geq \frac{1}{2}(s_i - \delta_i)^2.$$

So every unit of off-diagonal mass contributes a transport cost of at least $\frac{1}{2}(s_i - \delta_i)^2$. Using equation 25, we have that

$$\mathrm{OT}_\varepsilon(\xi, X_i) \geq \frac{1}{2}(s_i - \delta_i)^2 \sum_{m \neq n} P_{mn} \geq \frac{(s_i - \delta_i)^2}{4}\|a - b_i\|_1.$$

Combining with equation 24, we have that

$$S_\varepsilon(\xi, X_i) \geq \frac{(s_i - \delta_i)^2}{4}\|a - b_i\|_1 - \varepsilon \log M.$$

If $\|a - b_i\|_1 > \tau_i$, then

$$S_\varepsilon(\xi, X_i) > \frac{(s_i - \delta_i)^2}{4}\tau_i - \varepsilon \log M \geq r_i^{\mathrm{loc}}(\delta_i, \tau_i) > r$$

which leads to a contradiction. Thus, we must have that

$$\|a - b_i\|_1 \leq \tau_i.$$

The defining constraints of $K_i(\delta_i, \tau_i)$ are closed and bounded in the finite-dimensional Euclidean space $\mathbb{R}^M \times (\mathbb{R}^d)^M$, so $K_i(\delta_i, \tau_i)$ is compact. We only need to verify that $K_i(\delta_i, \tau_i)$ is indeed a subset of the parameter space $\Delta^\circ_{M,a_{\min}} \times \mathrm{Loc}_{M,\Delta_{\min}}(\Omega)$. If $(a, x) \in K_i(\delta_i, \tau_i)$, then, for all $m = 1, \ldots, M$,

$$a_m \geq b_{i,m} - \tau_i > a_{\min},$$

since $\tau_i < w_i = \min_m(b_{i,m} - a_{\min})$. Also, since $\delta_i < d_i^\partial$, we have that, for every $m = 1, \ldots, M$

$$\bar{\mathbb{B}}(y_{i,m}, \delta_i) \subset \Omega \quad \text{for every } m.$$

Finally, for $m \neq n$, we have that

$$\|x_m - x_n\| \geq \|y_{i,m} - y_{i,n}\| - 2\delta_i \geq s_i - 2\delta_i > \Delta_{\min}$$

since $\delta_i < \frac{s_i - \Delta_{\min}}{2}$. Therefore, we finally have that

$$K_i(\delta_i, \tau_i) \subset \Delta^\circ_{M,a_{\min}} \times \mathrm{Loc}_{M,\Delta_{\min}}(\Omega).$$

This completes the proof. □

Lemma 2 (uniform retraction domain and local metric upper bound)**.** *Assume $0 < r < r_i^{\mathrm{loc}}(\delta_i, \tau_i)$, and let $K_i(\delta_i, \tau_i)$ be as in Lemma 1. Define*

$$a_i^- := \min_{1 \le m \le M} b_{i,m} - \tau_i > a_{\min}$$

and

$$\eta_{\mathrm{ret},i} := \min\left\{ \frac{\lambda^2}{2D^2} \log \frac{a_i^-}{a_{\min}}, \frac{1}{2D} \min\left\{ d_i^\partial - \delta_i, s_i - 2\delta_i - \Delta_{\min} \right\} \right\}.$$

Then, the following hold true

1. *For every $\xi = \Xi(a, x) \in \Xi(K_i(\delta_i, \tau_i))$,*

$$\sup_{z \in \Omega} |u_\xi(z)| \le D^2, \quad \sup_{z \in \Omega} \|\nabla u_\xi(z)\| \le D.$$

2. *If $0 < \eta < \eta_{\mathrm{ret},i}$, then for every $\xi = \Xi(a, x) \in \Xi(K_i(\delta_i, \tau_i))$, the retraction $\mathrm{Ret}_\xi(-\eta \, \mathrm{grad}_{\mathrm{SHK}} E(\xi))$ is well-defined.*

3. *If*

$$\|(\delta a, \delta x)\|_E^2 := \sum_{m=1}^M (\delta a_m)^2 + \sum_{m=1}^M \|\delta x_m\|^2,$$

then on $K_i(\delta_i, \tau_i)$,

$$\|(\delta a, \delta x)\|_{\mathrm{SHK},(a,x)}^2 \le L_i^2 \|(\delta a, \delta x)\|_E^2, \quad L_i := \sqrt{\max\left\{ \frac{\lambda^2}{a_i^-}, 1 \right\}}.$$

Consequently, for all $(a, x), (a', x') \in K_i(\delta_i, \tau_i)$,

$$d_{\mathrm{SHK}}(\Xi(a, x), \Xi(a', x')) \le L_i \|(a - a', x - x')\|_E.$$

Proof. For $j = 1, \ldots, N$, let

$$\phi_{j,\xi}(z) := \frac{\delta S_\varepsilon(\xi, X_j)}{\delta \xi}(z)$$

and let

$$\bar{\phi}_{j,\xi}(z) := \phi_{j,\xi}(z) - \int \phi_{j,\xi} d\xi.$$

Then

$$u_\xi(z) = \sum_{j=1}^N w_j(\xi) \bar{\phi}_{j,\xi}(z), \quad \sum_{j=1}^N w_j(\xi) = 1, \quad w_j(\xi) \ge 0.$$

Following the derivations in the proof of Lemma 8, specifically equation 31 and equation 32, for each $j = 1, \ldots, N$, we have that

$$\sup_{z \in \Omega} \|\nabla \phi_{j,\xi}(z)\| \le D, \quad \sup_{z \in \Omega} \phi_{j,\xi}(z) - \inf_{z \in \Omega} \phi_{j,\xi}(z) \le D^2.$$

Hence, we obtain

$$\sup_{z \in \Omega} \left| \bar{\phi}_{j,\xi}(z) \right| \le D^2 \text{ and } \sup_{z \in \Omega} \left\| \nabla \bar{\phi}_{j,\xi}(z) \right\| \le D.$$

Taking the convex combination with weights $w_j(\xi)$, we have that,

$$\sup_{z \in \Omega} |u_\xi(z)| \le D^2 \text{ and } \sup_{z \in \Omega} \|\nabla u_\xi(z)\| \le D.$$

Let

$$\mathrm{grad}_{\mathrm{SHK}} E(a, x) = \left(\left(\frac{a_m}{\lambda^2} u_\xi(x_m) \right)_{m=1}^M, (\nabla u_\xi(x_m))_{m=1}^M \right).$$

The actual retraction step is therefore

$$\delta a_m = -\eta \frac{a_m}{\lambda^2} u_\xi(x_m), \quad \delta x_m = -\eta \nabla u_\xi(x_m).$$

Using the above bound,

$$\left\|\frac{\delta a}{a}\right\|_{\infty} = \max_m \frac{|\delta a_m|}{a_m} \leq \eta \frac{D^2}{\lambda^2}, \quad \|\delta x\|_{\infty,2} := \max_m \|\delta x_m\| \leq \eta D. \tag{27}$$

By Lemma 1, every $(a, x) \in K_i(\delta_i, \tau_i)$ satisfies $a_m \geq a_i^-$ for all $m = 1, \dots, M$, $d_\partial(x) \geq d_i^\partial - \delta_i$ and $\mathrm{sep}(x) - \Delta_{\min} \geq s_i - 2\delta_i - \Delta_{\min}$.

Hence, under the local region of validity of the retraction map, as introduced in Section D.4,

$$\left\|\frac{\delta a}{a}\right\|_{\infty} < \frac{1}{2} \log \frac{a_i^-}{a_{\min}}$$

and

$$\|\delta x\|_{\infty,2} < \frac{1}{2} \min\left\{d_i^\partial - \delta_i, s_i - 2\delta_i - \Delta_{\min}\right\}$$

are sufficient to make the weight and position retractions well-defined. This is guaranteed by $0 < \eta < \eta_{\mathrm{ret},i}$.

On $K_i(\delta_i, \tau_i)$, we have $a_m \geq a_i^-$ and $a_m \leq 1$. Therefore,

$$\|(\delta a, \delta x)\|^2_{\mathrm{SHK},(a,x)} = \sum_{m=1}^{M} \frac{\lambda^2}{a_m} (\delta a_m)^2 + \sum_{m=1}^{M} a_m \|\delta x_m\|^2 \leq \frac{\lambda^2}{a_i^-} \sum_m (\delta a_m)^2 + \sum_m \|\delta x_m\|^2 \leq L_i^2 \|(\delta a, \delta x)\|^2_E.$$

Now $K_i(\delta_i, \tau_i)$ is convex, since the weight constraints define a convex set, and each position constraint $x_m \in \bar{B}(y_{i,m}, \delta_i)$ defines a convex set, and intersection of convex sets is a convex set again. Hence the straight segment joining two points of $K_i(\delta_i, \tau_i)$ stays in $K_i(\delta_i, \tau_i)$. Integrating the above pointwise metric upper bound along that straight segment yields

$$d_{\mathrm{SHK}}(\Xi(a, x), \Xi(a', x')) \leq L_i \|(a - a', x - x')\|_E .$$

This completes the proof. □

G.3 Proof of Theorem 3

Proof. Since $F_i(X_i) = S_\varepsilon(X_i, X_i) = 0 \leq r$, we have $X_i \in B_i(r)$. Assumption (A3) guarantees the existence and uniqueness of the minimizer $X_i^*(r) \in B_i(r)$. Now, we apply Lemma 5 to $\xi = X_i^*(r)$, giving

$$0 \leq F_i(X_i^*(r)) - E(X_i^*(r)) \leq \frac{1}{\beta} \log\left(1 + (N-1)e^{-\beta\Delta}\right)$$

Thus

$$F_i(X_i^*(r)) \leq E(X_i^*(r)) + \frac{1}{\beta} \log\left(1 + (N-1)e^{-\beta\Delta}\right).$$

But $X_i^*(r)$ minimizes E over $B_i(r)$, and $X_i \in B_i(r)$, hence

$$E(X_i^*(r)) \leq E(X_i).$$

Also $E(\xi) \leq F_i(\xi)$ for every $\xi \in B_i(r)$, so

$$E(X_i) \leq F_i(X_i) = 0.$$

Combining, we have that

$$\begin{aligned}
S_\varepsilon(X_i^*(r), X_i) =& F_i\left(X_i^*(r)\right) \\
\leq& E\left(X_i^*(r)\right) + \frac{1}{\beta}\log\left(1 + (N-1)e^{-\beta\Delta}\right) \\
\leq& E\left(X_i\right) + \frac{1}{\beta}\log\left(1 + (N-1)e^{-\beta\Delta}\right) \\
\leq& 0 + \frac{1}{\beta}\log\left(1 + (N-1)e^{-\beta\Delta}\right) \\
=& \frac{1}{\beta}\log\left(1 + (N-1)e^{-\beta\Delta}\right).
\end{aligned}$$

Now, consider the one-step retraction curve

$$\gamma_i(t) := \mathrm{Ret}_{X_i}\left(-t\eta\, \mathrm{grad}_{\mathrm{SHK}}\, E\left(X_i\right)\right), \quad t \in [0,1].$$

Then $\gamma_i(0) = X_i$ and $\gamma_i(1) = \Phi_\eta(X_i)$. Note that, as derived in Theorem 2, when $\eta \leq \eta_{\mathrm{ret},i}$, the retraction $\Phi_\eta(X_i)$ is well defined. Applying Lemma 7 with $(\delta a, \delta x) = -\eta\, \mathrm{grad}_{\mathrm{SHK}}\, E\left(X_i\right)$ in particle coordinates yields

$$\mathrm{Length}\left(\gamma_i\right) \leq \min\left\{e^{\|\frac{\delta a}{a}\|_\infty}, \frac{1}{\sqrt{a_{\min}}}\right\}\eta\left\|\mathrm{grad}_{\mathrm{SHK}}\, E\left(X_i\right)\right\|_{\mathrm{SHK},X_i}.$$

Now we proceed to bound $\|\frac{\delta a}{a}\|_\infty$. For the SHK gradient descent direction, the weight update follows

$$\delta a_m = -\eta\frac{a_m}{\lambda^2}u_{X_i}\left(x_m\right) \quad \Longrightarrow \quad \frac{\delta a_m}{a_m} = -\frac{\eta}{\lambda^2}u_{X_i}\left(x_m\right).$$

where

$$u_\xi := \frac{\delta E}{\delta \xi}(\xi) - \left\langle\frac{\delta E}{\delta \xi}(\xi), \xi\right\rangle = \frac{\delta E}{\delta \xi}(\xi) - \int\frac{\delta E}{\delta \xi}(\xi)d\xi.$$

Thus

$$\left\|\frac{\delta a}{a}\right\|_\infty \leq \frac{\eta}{\lambda^2}\sup_{x\in\Omega}\left|u_{X_i}(x)\right|.$$

Now, $\frac{\delta E}{\delta \xi}(\xi) = \sum_{j=1}^N w_j(\xi)\frac{\delta S_\varepsilon(\xi, X_j)}{\delta \xi}$ and hence $u_\xi(x) = \sum_{j=1}^N w_j(\xi)\left(\phi_i(x) - \int\phi_i(y)d\xi(y)\right)$ where

$$\phi_i(x) := \left(\frac{\delta S_\varepsilon(\xi, X_i)}{\delta \xi}\right)(x)$$

Following the argument presented in the proof of Lemma 8, specifically equation 32, we have that, for any $x, y \in \Omega$, $|\phi_i(x) - \phi_i(y)| \leq D^2$ and hence for any $\xi \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$, we have that $\sup_{x\in\Omega}|u_\xi(x)| \leq D^2$. Consequently, we have that

$$\left\|\frac{\delta a}{a}\right\|_\infty \leq \frac{\eta D^2}{\lambda^2}.$$

Using Lemma 8, the fact that $S_\varepsilon(X_i, X_i) = 0$ and following the same argument as in the proof of Theorem 2, we have that

$$\begin{aligned}
S_\varepsilon(X_i, \Phi_\eta(X_i)) \leq& G d_{\mathrm{SHK}}(X_i, \Phi_\eta(X_i)) \\
\leq& G\,\mathrm{Length}\left(\gamma_i\right) \\
\leq& \min\left\{e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}}\right\}\eta G\left\|\mathrm{grad}_{\mathrm{SHK}}\, E\left(X_i\right)\right\|_{\mathrm{SHK},X_i}.
\end{aligned}$$

Now, since $F_i(\xi) = S_\varepsilon(\xi, X_i)$ has a global minimum at $\xi = X_i$ and F_i is differentiable in the SHK sense, we must have that $\mathrm{grad}_{\mathrm{SHK}}\, F_i(X_i) = 0$. Now, using Lemma 6, we have that

$$\mathrm{grad}_{\mathrm{SHK}}\, E(X_i) = \sum_{j=1}^N w_j(X_i)\,\mathrm{grad}_{\mathrm{SHK}}\, F_j(X_i) = \sum_{j\neq i}^N w_j(X_i)\,\mathrm{grad}_{\mathrm{SHK}}\, F_j(X_i).$$

Consequently, using the gradient bound from Lemma 8, we have that

$$\begin{aligned}\|\mathrm{grad}_{\mathrm{SHK}}\, E(X_i)\|_{\mathrm{SHK},X_i} \leq& \sum_{j\neq i} w_j(X_i)\, \|\mathrm{grad}_{\mathrm{SHK}}\, F_j(X_i)\|_{\mathrm{SHK},X_i} \\ \leq& G\sum_{j\neq i} w_j(X_i) \leq \frac{G(N-1)e^{-\beta\Delta}}{1+(N-1)e^{-\beta\Delta}}\end{aligned}$$

Therefore, we have that

$$\begin{aligned}S_\varepsilon(X_i, \Phi_\eta(X_i)) &\leq \min\left\{e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}}\right\} \eta G\, \|\mathrm{grad}_{\mathrm{SHK}}\, E\,(X_i)\|_{\mathrm{SHK},X_i} \\ &\leq \frac{\min\left\{e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}}\right\} \eta G^2(N-1)e^{-\beta\Delta}}{1+(N-1)e^{-\beta\Delta}} \\ &\leq \min\left\{e^{\eta D^2/\lambda^2}, \frac{1}{\sqrt{a_{\min}}}\right\} \eta G^2(N-1)e^{-\beta\Delta}\end{aligned}$$

where we used $\frac{u}{1+u} \leq u$ for $u \geq 0$ in the last inequality.

This completes the proof. □

H Discrete retrieval algorithm for empirical measures

We now give a fully explicit deterministic retrieval scheme obtained by an explicit Euler discretization of equation 17. The resulting update alternates a Kantorovich (support) step and a spherical Hellinger (weight) step, both driven by the same Sinkhorn computations.

H.1 Discrete Sinkhorn objects

Let $\xi = \sum_{m=1}^{M} a_m \delta_{x_m}$ and $X_i = \sum_{n=1}^{M} b_{i,n}\delta_{y_{i,n}}$. An entropic coupling is a matrix $P_i \in \mathbb{R}_+^{M\times M}$ with row/column sums a and b_i. The barycentric projection map (as derived in Section C) on the query support is

$$T^{\varepsilon}_{\xi\to X_i}(x_m) = \sum_{n=1}^{M} \frac{P_i[m,n]}{a_m}\, y_{i,n}, \qquad T^{\varepsilon}_{\xi\to \xi}(x_m) = \sum_{\ell=1}^{M} \frac{P_0[m,\ell]}{a_m}\, x_\ell.$$

Moreover, the dual potentials returned by Sinkhorn (in the first argument) provide the discrete values $f_{\xi,X_i}(x_m)$ and $f_{\xi,\xi}(x_m)$ needed for equation 18. For the self-coupling $\mathrm{OT}_\varepsilon(\xi,\xi)$, ξ appears in *both* marginals, so the weight update requires the symmetric combination

$$f^{\mathrm{sym}}_{\xi,\xi}(x_m) \coloneqq \tfrac{1}{2}\big(f_{\xi,\xi}(x_m) + g_{\xi,\xi}(x_m)\big),$$

where $g_{\xi,\xi}$ is the Sinkhorn potential in the second argument. For the self-coupling, the correct (gauge-invariant) contribution to the *weight* gradient is the symmetric average $\frac{1}{2}(f_{\xi,\xi} + g_{\xi,\xi})$.

H.2 Explicit Euler (Kantorovich transport) update

Using equation 8, a step of size $\eta > 0$ updates the support points by

$$x_m^{k+1} = x_m^k + \eta\left(\sum_{i=1}^{N} w_i^k\, T_i^k(x_m^k) - T_0^k(x_m^k)\right), \tag{28}$$

where T_i^k denotes the barycentric map computed from the Sinkhorn coupling between ξ^k and X_i, and T_0^k is the barycentric map from the self-coupling between ξ^k and itself.

H.3 Multiplicative (spherical Hellinger) weight update

A first-order discretization of equation 19 can be implemented as the multiplicative update

$$a_m^{k+1} \propto a_m^k \exp\Big(-\frac{\eta}{\lambda^2} z_m^k\Big), \qquad \sum_{m=1}^{M} a_m^{k+1} = 1, \tag{29}$$

where z_m^k is defined by equation 18 at (x^k, a^k). The normalization removes any additive-constant ambiguity in the potentials and guarantees $a_m^{k+1} > 0$ whenever $a_m^k > 0$.

H.4 Full deterministic retrieval algorithm

Algorithm 1: Entropic DDAM with spherical Hellinger-Kantorovich retrieval (empirical measures)

Inputs: stored empirical measures $\{X_i\}_{i=1}^N$; query $\xi^0 = \sum_m a_m^0 \delta_{x_m^0}$; parameters $\beta > 0$, $\varepsilon > 0$, step size $\eta > 0$, spherical Hellinger scale $\lambda > 0$; number of iterations K (or a stopping criterion).

For $k = 0, 1, \ldots, K-1$ **do:**

1. **Sinkhorn couplings and costs.** For each i, run Sinkhorn between ξ^k and X_i to obtain: (a) coupling matrix P_i^k, (b) source potential values $f_i^k[m] = f_{\xi^k, X_i}(x_m^k)$, and (c) the entropic OT cost $\mathrm{OT}_\varepsilon(\xi^k, X_i)$. Compute also the self-coupling between ξ^k and itself to obtain P_0^k, both potential vectors $f_0^k[m] = f_{\xi^k,\xi^k}(x_m^k)$ and $g_0^k[m] = g_{\xi^k,\xi^k}(x_m^k)$ and the entropic self-OT cost $\mathrm{OT}_\varepsilon(\xi^k, \xi^k)$. Finally, for each i, compute the OT cost $\mathrm{OT}_\varepsilon(X_i, X_i)$.
2. **Sinkhorn divergences and Gibbs weights.** For each i, compute $S_\varepsilon(\xi^k, X_i)$ using equation 2, and set
$$w_i^k = \frac{\exp(-\beta S_\varepsilon(\xi^k, X_i))}{\sum_{j=1}^N \exp(-\beta S_\varepsilon(\xi^k, X_j))}.$$
3. **Barycentric maps.** For each particle m and each pattern i, compute
$$T_i^k(x_m^k) = \sum_n \frac{P_i^k[m,n]}{a_m^k} y_{i,n}, \qquad T_0^k(x_m^k) = \sum_\ell \frac{P_0^k[m,\ell]}{a_m^k} x_\ell^k.$$
4. **Support update (transport).** Update x_m^{k+1} using equation 28.
5. **Weight update (reaction).** Compute
$$z_m^k = \sum_{i=1}^N w_i^k \Big(f_i^k[m] - \tfrac{1}{2}\big(f_0^k[m] + g_0^k[m]\big)\Big),$$
then update (a_m^{k+1}) by equation 29.

Output: retrieved empirical measure $\xi^K = \sum_m a_m^K \delta_{x_m^K}$.

All steps above are deterministic given the Sinkhorn solver (which itself is deterministic for fixed initialization and tolerance). The support update is a pushforward by a deterministic barycentric map, and the weight update is a deterministic multiplicative reweighting on the simplex. Thus the overall retrieval operator is deterministic.

I Numerical Experiments

In this section, we demonstrate the empirical performance of our proposed retrieval algorithm, referred to as SinkhornSHK Algo, for finitely supported discrete measures and compare it to a baseline Euclidean geometry based classical Hopfield-type algorithm, which we refer to as Euclidean Algo. The Euclidean algo vectorizes $(a, x) \in \mathcal{M}_M$ into $\xi_{\text{vec}} = [x_{11}, \ldots, x_{1d}, \ldots, x_{M1}, \ldots, x_{Md}, \log a_1, \ldots, \log a_m] \in \mathbb{R}^{(d+1)M}$ and applied the classical Hopfield fixed point algorithm for vector inputs based on Euclidean ℓ_2 inner product similarity using

Equation 3 of Ramsauer et al. (2020) with the same choice of β as for our proposed SinkhornSHK Algo.

We consider a toy experiment where the stored patterns $X_1, \dots, X_N$ are uniformly weighted and the support points are sampled from Gaussian distributions. We choose $N = 5$ and data dimension $d = 2$.

In Experiment 1, we choose the means of the Gaussian distributions to be $(-4.0, -1.0), (-2.0, 2.2), (1.0, -6.0), (4.0, -4.2)$ and $(4.2, -0.8)$, while the covariance matrices were chosen to be $\begin{bmatrix} 0.60 & 0.20 \\ 0.20 & 0.90 \end{bmatrix}$, $\begin{bmatrix} 0.80 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$, $\begin{bmatrix} 0.65 & 0.00 \\ 0.00 & 0.65 \end{bmatrix}$, $\begin{bmatrix} 0.55 & 0.10 \\ 0.10 & 1.00 \end{bmatrix}$ and $\begin{bmatrix} 0.95 & 0.00 \\ 0.00 & 0.50 \end{bmatrix}$. $M = 30$ support points were sampled in i.i.d manner from the 5 Gaussian distributions determined by each pair of mean and covariance parameters and the resulting uniformly weighted discrete distributions were set as the patterns to be stored.

In Experiment 2, we choose the means of the 5 Gaussian distributions to be all equal to $(0, 0)$ and the covariance matrices were randomly sampled using random orthogonal matrices coupled with uniformly sampled eigenvalues between 0.15 and 1.75. We sample $M = 25$ support points in i.i.d manner from each of these Gaussian distributions and the resulting uniformly weighted discrete distributions were set as the patterns to be stored.

In both the experiments, we first fix a pattern that we want to retrieve, then perturb the support points individually using i.i.d Gaussian noise (sd 0.5 in Experiment 1 and sd 0.2 in Experiment 2) to generate a query distribution that serves as the initial iterate $\xi^{(0)}$ for both algorithms. We chose $\beta = 50$, $\varepsilon = 0.05$ and step-size $\eta = 1.3$ for the SinkhornSHK Algo, and the same β for Euclidean Algo. We do not use the spherical Hellinger update step in these simple experiments since the all discrete measures involved are uniformly weighted. We use a maximum iteration threshold $k \leq 200$ for both algorithms, and the Sinkhorn algorithm for computing entropic OT transport plans were capped at 120 iterations.

In Experiment 1, we see that both algorithms are able to retrieve the correct discrete distributions when considering the support points that are returned by either algorithm. However, Experiment 2 clearly shows the superiority of SinkhornSHK Algo over Euclidean Algo, since the Sinkhorn Algo is able to converge to the correct pattern even when a noisy query is given. We believe that the ability of Sinkhorn ALgo to leverage the distributional perspective gives it the advantage over Euclidean Algo, since the latter relies on Euclidean inner products and is expected to fail in cases where Euclidean separation between support points is small, but separation in distributional metrics is still feasible.

J AUXILIARY RESULTS

J.1 DUAL AND OPTIMAL POTENTIALS

A standard dual form of OT_ε can be written (up to equivalent normalizations) in terms of potentials $f, g \in C(\Omega)$.One defines the entropic soft c-transform operator A_ε via an expression of the form

$$A_\varepsilon(g, \nu)(x) := -\varepsilon \log \int_\Omega \exp\Big(\frac{g_{\mu,\nu}(y) - c(x, y)}{\varepsilon}\Big)\, d\nu(y) \quad \text{(defined up to an additive constant).}$$

Then the optimal potentials $(f_{\mu,\nu}, g_{\mu,\nu})$ (Schrödinger potentials) satisfy the Schrödinger system

$$f_{\mu,\nu} = A_\varepsilon(g_{\mu,\nu}, \nu) \quad \mu\text{-a.e.}, \qquad g_{\mu,\nu} = A_\varepsilon(f_{\mu,\nu}, \mu) \quad \nu\text{-a.e..}$$

These potentials are unique up to adding a constant to $f_{\mu,\nu}$ and subtracting the same constant from $g_{\mu,\nu}$ (gauge invariance). This does not affect any gradient $\nabla f_{\mu,\nu}(x)$ or $\nabla g_{\mu,\nu}(x)$, which is what we ultimately use.

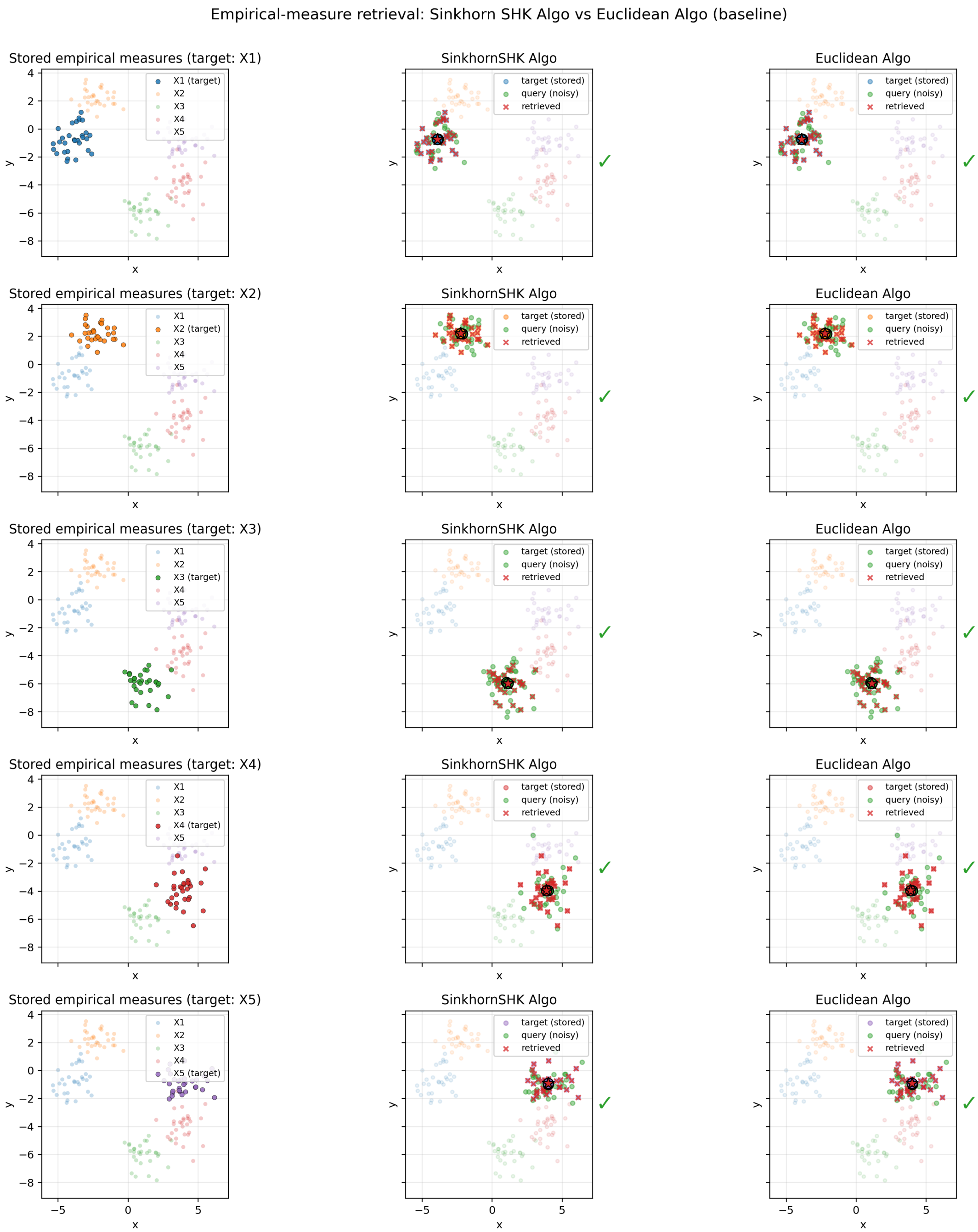

Figure 1: Experiment 1: Sinkhorn Algo and Euclidean Algo are both able to retrieve correct patterns from noisy queries

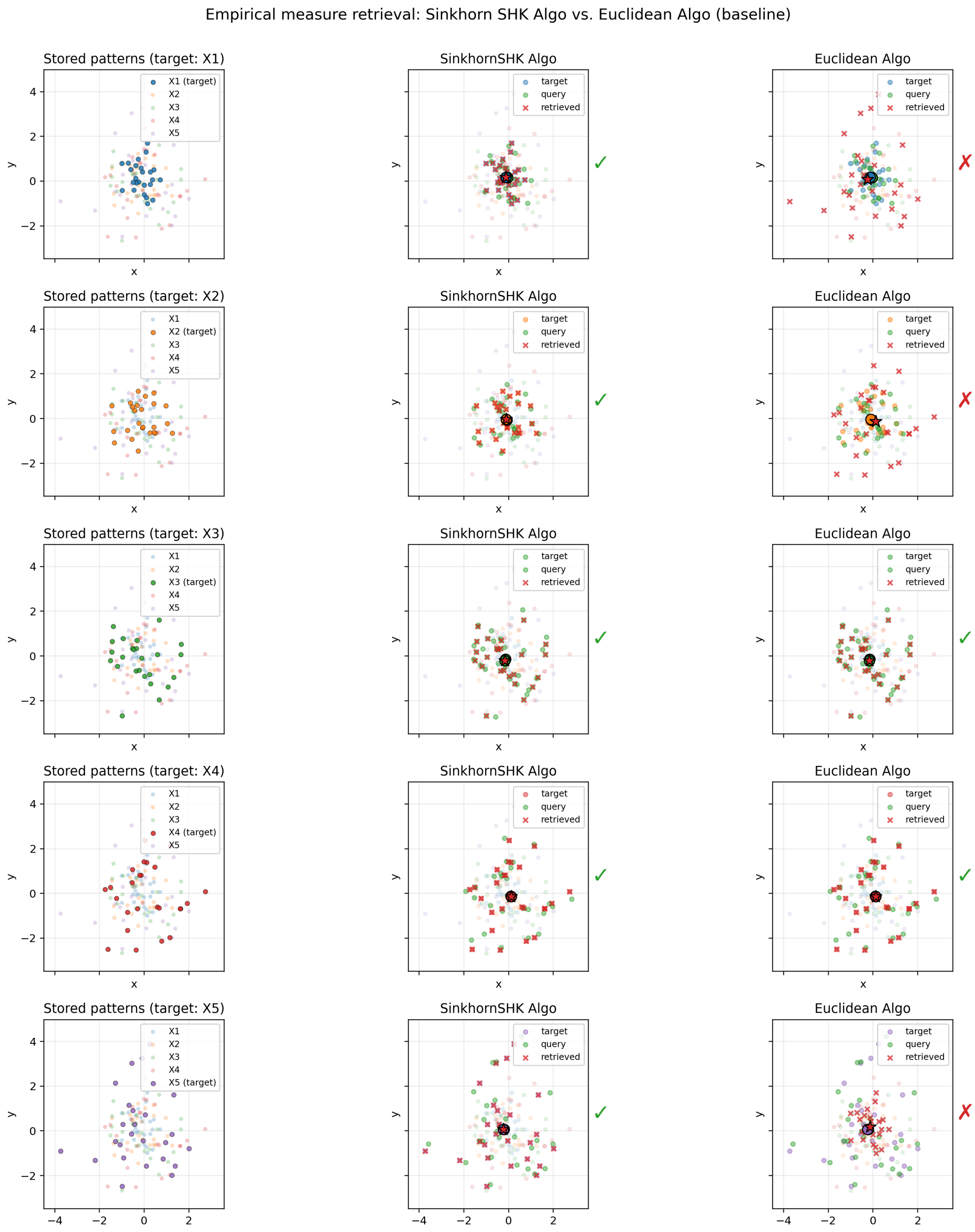

Figure 2: Experiment 2: Sinkhorn Algo succeeds in retrieving correct patterns from noisy queries in all instances, but Euclidean Algo fails in 3 cases.

J.2 Barycentric projection map

Let $\pi^{\varepsilon}_{\mu,\nu}$ be the optimal entropic coupling between μ and ν, which can be disintegrated into a marginal and conditional distribution as follows :

$$\pi^{\varepsilon}_{\mu,\nu}(dx, dy) = \mu(dx)\, \pi^{\varepsilon}_{\mu,\nu}(dy \mid x).$$

Define the barycentric projection (conditional mean) map:

$$T^{\varepsilon}_{\mu\to\nu}(x) \coloneqq \int_{\Omega} y\, \pi^{\varepsilon}_{\mu,\nu}(dy \mid x).$$

This is defined μ-a.e. and takes values in $\mathrm{conv}(\Omega) \subset \mathbb{R}^d$.

J.3 Computing gradient of Schrödinger potentials explicitly for quadratic costs

Recall (from the Schrödinger system) that

$$f_{\mu,\nu}(x) = -\varepsilon \log \int_{\Omega} \exp\Big(\frac{g_{\mu,\nu}(y) - c(x,y)}{\varepsilon}\Big)\, d\nu(y) \quad \text{(up to constant)}.$$

Differentiate with respect to x. Denote

$$Z(x) \coloneqq \int_{\Omega} \exp\Big(\frac{g_{\mu.\nu}(y) - c(x,y)}{\varepsilon}\Big)\, d\nu(y).$$

Then $f_{\mu,\nu}(x) = -\varepsilon \log Z(x)$, so

$$\nabla f_{\mu,\nu}(x) = -\varepsilon \frac{1}{Z(x)} \nabla Z(x).$$

Compute $\nabla Z(x)$:

$$\nabla Z(x) = \int_{\Omega} \exp\Big(\frac{g_{\mu,\nu}(y) - c(x,y)}{\varepsilon}\Big) \cdot \frac{1}{\varepsilon}\big(- \nabla c(x,y)\big)\, d\nu(y).$$

Therefore

$$\begin{aligned}\nabla f_{\mu,\nu}(x) &= -\,\varepsilon \frac{1}{Z(x)} \int \exp\Big(\frac{g_{\mu,\nu}(y) - c(x,y)}{\varepsilon}\Big) \cdot \frac{1}{\varepsilon}\big(- \nabla c(x,y)\big)\, d\nu(y) \\ &= \frac{1}{Z(x)} \int \exp\Big(\frac{g_{\mu,\nu}(y) - c(x,y)}{\varepsilon}\Big) \nabla c(x,y)\, d\nu(y).\end{aligned}$$

But the conditional distribution $\pi^{\varepsilon}(dy \mid x)$ has density proportional to $\exp((g_{\mu,\nu}(y) - c(x,y))/\varepsilon), d\nu(y)$. Hence

$$\nabla f_{\mu,\nu}(x) = \int \nabla c(x,y)\, \pi^{\varepsilon}_{\mu,\nu}(dy \mid x).$$

Now plug $c(x,y) = \frac{1}{2}\|x-y\|_2^2$, so $\nabla c(x,y) = x - y$. Then

$$\nabla f_{\mu,\nu}(x) = \int (x-y)\, \pi^{\varepsilon}_{\mu,\nu}(dy \mid x) = x - \int y\, \pi^{\varepsilon}_{\mu,\nu}(dy \mid x) = x - T^{\varepsilon}_{\mu\to\nu}(x).$$

Thus we have the fundamental identity:

$$\nabla f_{\mu,\nu}(x) = x - T^{\varepsilon}_{\mu\to\nu}(x) \quad \text{for } c(x,y) = \tfrac{1}{2}\|x-y\|_2^2.$$

This is precisely why barycentric projections give a transport-map-like representation of entropic OT gradients.

Similarly, we have that $\nabla g_{\mu,\nu}(y) = y - T^{\varepsilon}_{\nu\to\mu}(y)$. In particular, we have that

$$\frac{1}{2}\left(\nabla f_{\xi,\xi}(x) + \nabla g_{\xi,\xi}(x)\right) = x - T^{\varepsilon}_{\xi\to\xi}(x). \tag{30}$$

J.4 ADDITIONAL LEMMAS

Lemma 3 (One step descent of softmin energy)**.** *Let Assumptions (A1) and (A2) hold. Let* $\xi^+ = \Phi_\eta(\xi) = \mathrm{Ret}_\xi(-\eta \operatorname{grad}_{\mathrm{SHK}} E(\xi))$. *Then*

$$E\left(\xi^+\right) \leq E(\xi) - \eta\left(1 - \frac{L\eta}{2}\right) \| \operatorname{grad}_{\mathrm{SHK}} E(\xi)\|^2_{\mathrm{SHK},\xi}.$$

In particular, if $0 < \eta \leq \frac{1}{L}$, *then*

$$E\left(\xi^+\right) \leq E(\xi) - \frac{\eta}{2}\| \operatorname{grad}_{\mathrm{SHK}} E(\xi)\|^2_{\mathrm{SHK},\xi},$$

so E *strictly decreases unless* $\operatorname{grad}_{\mathrm{SHK}} E(\xi) = (0,0)$.

Proof. Using Equation 21, with $(r, v) = -\operatorname{grad}_{\mathrm{SHK}} E(\xi)$, we have that

$$\begin{aligned}
E\left(\xi^+\right) =& E\left(\mathrm{Ret}_\xi(-\operatorname{grad}_{\mathrm{SHK}} E(\xi)\right) \\
\leq& E(\xi) - \eta\langle \operatorname{grad}_{\mathrm{SHK}} E(\xi), \operatorname{grad}_{\mathrm{SHK}} E(\xi)\rangle_{\mathrm{SHK},\xi} + \frac{L\eta^2}{2}\| \operatorname{grad}_{\mathrm{SHK}} E(\xi)\|^2_{\mathrm{SHK},\xi} \\
=& E(\xi) - \eta\left(1 - \frac{L\eta}{2}\right) \| \operatorname{grad}_{\mathrm{SHK}} E(\xi)\|^2_{\mathrm{SHK},\xi}.
\end{aligned}$$

If $\eta \leq \frac{1}{L}$, then $1 - \frac{L\eta}{2} \geq \frac{1}{2}$, giving the stated bound. □

Lemma 4 (Energy gap contraction)**.** *Let Assumptions (A1), (A2) and (A4) hold and choose* $0 < \eta \leq \min\left\{\frac{1}{L}, \frac{1}{\mu}\right\}$. *Then for* $i = 1, \dots, N$, *if* $\xi^{(0)} \in B_i(r)$, $\xi^{(k+1)} = \Phi_\eta\left(\xi^{(k)}\right)$ *and* $\xi^{(k)} \in B_i(r)$ *for* $k \in \mathbb{N}$, *we have*

$$E\left(\xi^{(k+1)}\right) - E_i^*(r) \leq (1 - \mu\eta)\left(E\left(\xi^{(k)}\right) - E_i^*(r)\right).$$

Hence, for any $k \in \mathbb{N}$,

$$E\left(\xi^{(k)}\right) - E_i^*(r) \leq (1 - \mu\eta)^k\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)$$

and consequently, $\lim_{k\to\infty} E(\xi^{(k)}) = E_i^*(r)$.
Further, for each $k \in \mathbb{N}$,

$$\left\|\operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right\|^2_{\mathrm{SHK},\xi^{(k)}} \leq \frac{2}{\eta}\left(E\left(\xi^{(k)}\right) - E\left(\xi^{(k+1)}\right)\right).$$

Consequently, for each $k \in \mathbb{N}$,

$$\left\|\operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right\|_{\mathrm{SHK},\xi^{(k)}} \leq \sqrt{\frac{2}{\eta}\left(E\left(\xi^{(k)}\right) - E_i^*(r)\right)} \leq \sqrt{\frac{2\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)}{\eta}(1 - \eta\mu)^{\frac{k}{2}}}.$$

Proof. Using Lemma 3, we have that, provided $\xi^{(k)} \in B_i(r)$,

$$E\left(\xi^{(k+1)}\right) \leq E\left(\xi^{(k)}\right) - \frac{\eta}{2}\left\|\operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right\|^2_{\mathrm{SHK},\xi^{(k)}}.$$

Subtract $E_i^*(r)$ from both sides:

$$E\left(\xi^{(k+1)}\right) - E_i^*(r) \leq E\left(\xi^{(k)}\right) - E_i^*(r) - \frac{\eta}{2}\left\|\operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right\|^2_{\mathrm{SHK},\xi^{(k)}}.$$

Using Equation 22, we have that

$$\frac{1}{2}\left\|\operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right\|^2_{\mathrm{SHK},\xi^{(k)}} \geq \mu\left(E\left(\xi^{(k)}\right) - E_i^*(r)\right)$$
$$\implies \left\|\operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right\|^2_{\mathrm{SHK},\xi^{(k)}} \geq 2\mu\left(E\left(\xi^{(k)}\right) - E_i^*(r)\right).$$

Thus, provided $\xi^{(k)} \in B_i(r)$,

$$E\left(\xi^{(k+1)}\right) - E_i^*(r) \leq E\left(\xi^{(k)}\right) - E_i^*(r) - \eta\mu\left(E\left(\xi^{(k)}\right) - E_i^*(r)\right) = (1-\mu\eta)\left(E\left(\xi^{(k)}\right) - E_i^*(r)\right).$$

Iterating the inequality, we obtain

$$0 \leq E\left(\xi^{(k)}\right) - E_i^*(r) \leq (1-\mu\eta)^k\left(E\left(\xi^{(0)}\right) - E_i^*(r)\right)$$

and consequently $\lim_{k\to\infty} E(\xi^{(k)}) = E_i^*(r)$.

Using Lemma 3, since $0 < \eta \leq \frac{1}{L}$, we have that

$$E\left(\xi^{(k+1)}\right) \leq E\left(\xi^{(k)}\right) - \frac{\eta}{2}\left\|\operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right\|^2_{\mathrm{SHK},\xi^{(k)}}$$

which upon rearrangement gives

$$\left\|\operatorname{grad}_{\mathrm{SHK}} E\left(\xi^{(k)}\right)\right\|^2_{\mathrm{SHK},\xi^{(k)}} \leq \frac{2}{\eta}\left(E\left(\xi^{(k)}\right) - E\left(\xi^{(k+1)}\right)\right).$$

Again, since we assume that $\xi^{(k+1)} \in B_i(r)$, we have that $E_i^*(r) \leq E\left(\xi^{(k+1)}\right)$. Consequently,

$$E\left(\xi^{(k)}\right) - E\left(\xi^{(k+1)}\right) \leq E\left(\xi^{(k)}\right) - E_i^*(r).$$

This completes the proof.

□

Lemma 5 (Control on basin interference and softmin perturbation bounds of energy E inside local basin using separation margin)**.** *Let Assumption (A1) hold. Define the Gibbs weights $w_i(\xi)$ corresponding to any fixed $\xi \in B_i(r)$ as in Equation 5. Then, we have that*

1. *The weights satisfy*

$$w_i(\xi) \geq \frac{1}{1+(N-1)e^{-\beta\Delta}}, \quad \sum_{j\neq i} w_j(\xi) \leq \frac{(N-1)e^{-\beta\Delta}}{1+(N-1)e^{-\beta\Delta}}.$$

2. *The gap between $F_i(\xi)$ and $E(\xi)$ satisfies*

$$0 \leq F_i(\xi) - E(\xi) \leq \frac{1}{\beta}\log\left(1+(N-1)e^{-\beta\Delta}\right) \leq \frac{N-1}{\beta}e^{-\beta\Delta}.$$

Proof. Using Equation 20, we have that, for $\xi \in B_i(r)$,

$$\sum_{j\neq i} e^{-\beta F_j(\xi)} = e^{-\beta F_i(\xi)} \times \sum_{j\neq i} e^{-\beta(F_j(\xi)-F_i(\xi))} \leq e^{-\beta F_i(\xi)} \times \sum_{j\neq i} e^{-\beta\Delta_r} \leq e^{-\beta F_i(\xi)} \times (N-1)e^{-\beta\Delta}.$$

Therefore, we have that

$$w_i(\xi) = \frac{e^{-\beta F_i(\xi)}}{e^{-\beta F_i(\xi)} + \sum_{j\neq i} e^{-\beta F_j(\xi)}} \geq \frac{1}{1+(N-1)e^{-\beta\Delta}}$$

Therefore, we have that

$$\sum_{j\neq i} w_j(\xi) = 1 - w_i(\xi) \leq \frac{(N-1)e^{-\beta\Delta}}{1+(N-1)e^{-\beta\Delta}}.$$

Now, we have that

$$E(\xi) = -\frac{1}{\beta}\log\left(e^{-\beta F_i(\xi)}\left[1+\sum_{j\neq i} e^{-\beta(F_j(\xi)-F_i(\xi))}\right]\right) = F_i(\xi) - \frac{1}{\beta}\log\left(1+\sum_{j\neq i} e^{-\beta(F_j(\xi)-F_i(\xi))}\right).$$

Since each $F_j(\xi) - F_i(\xi) \geq \Delta$,

$$0 \leq \frac{1}{\beta}\log\left(1+\sum_{j\neq i} e^{-\beta(F_j-F_i)}\right) \leq \frac{1}{\beta}\log\left(1+(N-1)e^{-\beta\Delta}\right).$$

Finally, we use $\log(1+u) \leq u$ to get the last bound. ☐

Lemma 6 (Expression of SHK gradient oof energy functional E in terms of SHK gradients of Sinkhorn divergence)**.**

$$\mathrm{grad}_{\mathrm{SHK}}\, E(\xi) = \sum_{j=1}^{N} w_j(\xi)\, \mathrm{grad}_{\mathrm{SHK}}\, F_j(\xi)$$

Proof. Proof is obvious. ☐

Lemma 7 (Bounding SHK distance in terms of SHK gradient norm along retraction curve)**.** *Fix any* (a, x) *and any tangent increment* $(\delta a, \delta x) \in T_{(a,x)}\mathcal{M}_M$*. Define the retraction curve*

$$\gamma(t) := \mathrm{Ret}_{(a,x)}(t\delta a, t\delta x), \quad t \in [0,1].$$

Let d_{SHK} *denote the Riemannian distance induced by the SHK metric on* $\mathcal{M}_M = \Delta^{\circ}_{M,a_{\min}} \times \mathrm{Loc}_{M,\Delta_{\min}}$*. Then*

$$d_{\mathrm{SHK}}\left((a,x), \mathrm{Ret}_{(a,x)}(\delta a, \delta x)\right) \leq \mathrm{Length}(\gamma) \leq \min\left\{e^{\|\frac{\delta a}{a}\|_\infty}, \frac{1}{\sqrt{a_{\min}}}\right\}\|(\delta a, \delta x)\|_{\mathrm{SHK},(a,x)}.$$

where $\mathrm{Length}(\gamma) := \int_0^1 \|\frac{d}{dt}\gamma(t)\|_{\mathrm{SHK},\gamma(t)} dt$ *and* $(\frac{\delta a}{a})_m := \frac{\delta a_m}{a_m}$*.*

Proof. By definition of the Riemannian distance as the infimum of lengths over all curves connecting the points,

$$d_{\mathrm{SHK}}(\gamma(0), \gamma(1)) \leq \mathrm{Length}(\gamma).$$

So it suffices to bound the length of γ. We have that $\gamma(t) = (a(t), x(t))$ with

$$x(t) = \mathrm{Ret}^{\mathrm{pos}}_x(t\delta x) = x + t\delta x, \quad a(t) = \mathrm{Ret}^{w}_a(t\delta a).$$

Then, $x'(t) := \frac{d}{dt}x(t) = \delta x$. For the weights, introduce

$$s_m := \frac{\delta a_m}{a_m} \quad \left(\text{so } s \in \mathbb{R}^M \text{ and } \sum_m a_m s_m = 0 \text{ since } \sum_m \delta a_m = 0\right)$$

By the definition of Ret_a^w, we have that

$$a_m(t) = \frac{a_m e^{ts_m}}{Z(t)}, \quad Z(t) \coloneqq \sum_{\ell=1}^{M} a_\ell e^{ts_\ell}.$$

Differentiating, we have that

$$\frac{d}{dt} \log a_m(t) = s_m - \frac{Z'(t)}{Z(t)}, \quad \frac{Z'(t)}{Z(t)} = \sum_{\ell=1}^{M} a_\ell(t) s_\ell =: \bar{s}(t).$$

Thus

$$a_m'(t) \coloneqq \frac{d}{dt} a_m(t) = a_m(t) \left(s_m - \bar{s}(t)\right)$$

The SHK gradient norm at $\gamma(t) = (a(t), x(t))$ is given by

$$\|(u,v)\|^2_{\mathrm{SHK},(a(t),x(t))} = \sum_{m=1}^{M} \left(\frac{\lambda^2}{a_m(t)} u_m^2 + a_m(t) \left\|v_m\right\|^2 \right).$$

Applying it to $(u, v) = \frac{d}{dt}\gamma(t) \coloneqq \frac{d}{dt}\gamma(t) = (a'(t), x'(t))$, we have that

$$\begin{aligned}
\left\| \frac{d}{dt}\gamma(t) \right\|^2_{\mathrm{SHK},(a(t),x(t))} &= \sum_{m=1}^{M} \frac{\lambda^2}{a_m(t)} \left(a_m'(t)\right)^2 + \sum_{m=1}^{M} a_m(t) \left\|x_m'(t)\right\|^2 \\
&= \lambda^2 \sum_{m=1}^{M} a_m(t) \left(s_m - \bar{s}(t)\right)^2 + \sum_{m=1}^{M} a_m(t) \left\|\delta x_m\right\|^2 \\
&= \lambda^2 \left[\sum_{m=1}^{M} a_m(t) s_m^2 - 2\bar{s}(t) \sum_{m=1}^{M} a_m(t) s_m + \left(\bar{s}(t)\right)^2 \right] + \sum_{m=1}^{M} a_m(t) \left\|\delta x_m\right\|^2 \\
&= \lambda^2 \left[\sum_{m=1}^{M} a_m(t) s_m^2 - \left(\bar{s}(t)\right)^2 \right] + \sum_{m=1}^{M} a_m(t) \left\|\delta x_m\right\|^2 \\
&\leq \lambda^2 \sum_{m=1}^{M} a_m(t) s_m^2 + \sum_{m=1}^{M} a_m(t) \left\|\delta x_m\right\|^2.
\end{aligned}$$

Let $S \coloneqq \|s\|_\infty = \|\delta a / a\|_\infty$. Then for all $t \in [0, 1]$,

$$e^{-S} \leq e^{ts_m} \leq e^S, \quad e^{-S} \leq Z(t) = \sum_\ell a_\ell e^{ts_\ell} \leq e^S.$$

Hence

$$\frac{a_m e^{-S}}{e^S} \leq a_m(t) = \frac{a_m e^{ts_m}}{Z(t)} \leq \frac{a_m e^S}{e^{-S}},$$

i.e.

$$a_m e^{-2S} \leq a_m(t) \leq a_m e^{2S}.$$

Using $a_m(t) \leq a_m e^{2S}$,

$$\left\| \frac{d}{dt}\gamma(t) \right\|^2_{\mathrm{SHK},(a(t),x(t))} \leq e^{2S} \left(\lambda^2 \sum_{m=1}^{M} a_m s_m^2 + \sum_{m=1}^{M} a_m \left\|\delta x_m\right\|^2 \right) = e^{2S} \|(\delta a, \delta x)\|^2_{\mathrm{SHK},(a,x)}.$$

Taking square roots gives, for all $t \in [0, 1]$,

$$\left\|\frac{d}{dt}\gamma(t)\right\|_{(a(t),x(t))} \leq e^S \|(\delta a, \delta x)\|_{\mathrm{SHK},(a,x)}.$$

Therefore,

$$\mathrm{Length}(\gamma) = \int_0^1 \left\|\frac{d}{dt}\gamma(t)\right\|_{\mathrm{SHK},(a(t),x(t))} dt \leq \int_0^1 e^S \|(\delta a, \delta x)\|_{\mathrm{SHK},(a,x)} dt = e^S \|(\delta a, \delta x)\|_{\mathrm{SHK},(a,x)}.$$

Now, note that.

$$\sum_{m=1}^{M} a_m(t) \left\|\delta x_m\right\|^2 \leq \max_m \left\|\delta x_m\right\|^2 \leq \frac{1}{a_{\min}} \sum_{m=1}^{M} a_m \left\|\delta x_m\right\|^2$$

and

$$\begin{aligned}
\lambda^2 \sum_{m=1}^{M} a_m(t) \left(s_m - \bar{s}(t)\right)^2 \leq & \lambda^2 \sum_{m=1}^{M} a_m(t) \left(s_m - \bar{s}(0)\right)^2 \\
\leq & \lambda^2 \max_m \left(s_m - \bar{s}(0)\right)^2 \\
\leq & \frac{\lambda^2}{a_{\min}} \sum_{m=1}^{M} a_m \left(s_m - \bar{s}(0)\right)^2 = \frac{\lambda^2}{a_{\min}} \sum_{m=1}^{M} \frac{(\delta a_m)^2}{a_m}.
\end{aligned}$$

Combining, we have that,

$$\left\|\frac{d}{dt}\gamma(t)\right\|_{\mathrm{SHK},(a(t),x(t))} \leq \frac{1}{\sqrt{a_{\min}}} \|(\delta a, \delta x)\|_{\mathrm{SHK},(a,x)}$$

and hence,

$$\mathrm{Length}(\gamma) \leq \frac{1}{\sqrt{a_{\min}}} \|(\delta a, \delta x)\|_{\mathrm{SHK},(a,x)}.$$

□

Lemma 8 (Open bounded convex domain implies boundedness of SHK gradient of Sinkhorn divergence)**.** *Assume that the domain Ω is open, bounded and convex with diameter $D := \sup_{x,y\in\Omega} \|x - y\| < \infty$. Then, for $\xi, \nu \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$, we have that*

$$\| \operatorname{grad}_{\mathrm{SHK}} S_\varepsilon(\xi, \nu)\|_{\mathrm{SHK},\xi} \leq G.$$

where $G := D\sqrt{1 + \frac{D^2}{\lambda^2}}$. In particular, for $i = 1, \ldots, N$, we consequently obtain

$$\| \operatorname{grad}_{\mathrm{SHK}} F_i(\xi)\|_{\mathrm{SHK},\xi} \leq G.$$

Further, for any fixed $\xi, \xi', \nu \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$, the following holds

$$|S_\varepsilon(\xi, \nu) - S_\varepsilon(\xi', \nu)| \leq G d_{\mathrm{SHK}}\left(\xi, \xi'\right)$$

where d_{SHK} is the SHK metric restricted to $\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$. Therefore, $S_\epsilon(\cdot, \nu)$ is G-Lipschitz continuous in the SHK metric.

Proof. Note that, for any fixed $\nu \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ and $\xi \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$, we have that

$$\phi(x) := \left(\frac{\delta S_\varepsilon(\xi,\nu)}{\delta \xi}\right)(x) = f_{\xi,\nu}(x) - \tfrac{1}{2}\big(f_{\xi,\xi}(x) + g_{\xi,\xi}(x)\big)$$

where $w_i(\xi)$ are defined as in Equation 5.

Note that $\operatorname{grad}_{\mathrm{SHK}} S_\varepsilon(\xi, \nu)(x) = (r(x), v(x))$ where $v(x) = \nabla\phi(x)$ and $r(x) = \frac{1}{\lambda^2}\left(\phi(x) - \int \phi(y) d\xi(y)\right)$.

Then,

$$\|\text{grad}_{\text{SHK}}\, S_\varepsilon(\xi,\nu)\|^2_{\text{SHK},\xi} = \int_\Omega \|v(x)\|^2 d\xi(x) + \lambda^2 \int_\Omega [r(x)]^2\, d\xi(x).$$

Note that,

$$v(x) = \nabla\left(f_{\xi,\nu}(x) - \frac{1}{2}\left(f_{\xi,\xi}(x) + g_{\xi,\xi}(x)\right)\right) = -\left[T^\epsilon_{\xi\to\nu}(x) - T^\epsilon_{\xi\to\xi}(x)\right].$$

Each barycentric map $T^\epsilon_{\xi\to\nu}(x)$ lies in Ω, since it is a conditional expectation defined over the domain Ω, and similarly $T^\epsilon_{\xi\to\xi}(x)\in\Omega$ as well. Hence

$$\|v(x)\| = \|\nabla\phi(x)\| = \left\|T^\epsilon_{\xi\to\nu}(x) - T^\epsilon_{\xi\to\xi}(x)\right\| \le D. \tag{31}$$

Since Ω is convex and open, every line segment joining two points of Ω stays in Ω. By the mean value theorem and the bound $\|\nabla\phi(x)\|\le D$, for any $x, y\in\Omega$,

$$|\phi(x)-\phi(y)| \le \sup_{z\in[x,y]} \|\nabla\phi(z)\|_2\|x-y\|_2 \le D\|x-y\|_2 \le D^2. \tag{32}$$

Therefore, we have that,

$$|r(x)| = \frac{1}{\lambda^2}\left|\int [\phi(x)-\phi(y)]\, d\xi(y)\right| \le \frac{1}{\lambda^2}\int |\phi(x)-\phi(y)|\, d\xi(y) \le \frac{D^2}{\lambda^2}.$$

Therefore, we have that

$$\|\text{grad}_{\text{SHK}}\, S_\varepsilon(\xi,\nu)\|^2_{\text{SHK},\xi} \le D^2 + \lambda^2 \times \frac{D^4}{\lambda^4} \le D^2 + \frac{D^4}{\lambda^2}.$$

In particular, we can choose $\nu = X_i$.

For the proof of the Lipschitz continuity of $F_\nu(\cdot) := S_\varepsilon(\cdot,\nu)$, consider $\gamma: [0,1]\to\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ to be an absolutely continuous path in $\mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ between ξ and ξ'. Then, by the chain rule,

$$\frac{d}{dt}F_\nu(\gamma(t)) = \left\langle \text{grad}\, F_\nu(\gamma(t)), \frac{d}{dt}\gamma(t)\right\rangle_{\text{SHK},\gamma(t)}$$

Apply Cauchy-Schwarz and the gradient bound:

$$\left|\frac{d}{dt}F_\nu(\gamma(t))\right| \le \|\text{grad}_{\text{SHK}}\, F_\nu(\gamma(t))\|_{\text{SHK},\gamma(t)} \left\|\frac{d}{dt}\gamma(t)\right\|_{\text{SHK},\gamma(t)} \le G\left\|\frac{d}{dt}\gamma(t)\right\|_{\text{SHK},\gamma(t)}$$

Integrating from 0 to 1, we have that,

$$|F_\nu(\xi') - F_\nu(\xi)| \le G\int_0^1 \left\|\frac{d}{dt}\gamma(t)\right\|_{\text{SHK},\gamma(t)} dt = G\ \text{Length}\ (\gamma)$$

Taking the infimum over all curves γ from ξ to ξ' gives

$$|F_\nu(\xi') - F_\nu(\xi)| \le G d_{\text{SHK}}(\xi,\xi')$$

□

Lemma 9 (Softmin function sensitivity)**.** *Define* $\text{softmin}_\beta(z_1,\ldots,z_N) := -\frac{1}{\beta}\log\left(\sum_{i=1}^N \exp(-\beta z_i)\right)$. *Then, we have that*

$$|\text{softmin}_\beta(z) - \text{softmin}_\beta(z')| \le \|z-z'\|_\infty \quad \forall z,z'\in\mathbb{R}^N.$$

Proof. Let us define $M_{\text{diff}} \coloneqq \max_{1 \le i \le N} (z_i - z'_i)$. Then, we have that $z_i \le z'_i + M_{\text{diff}}$ for $i = 1, \dots, N$. Note that that $\frac{\partial}{\partial z_i} \operatorname{softmin}_\beta(z) = \frac{\exp(-\beta z_i)}{\sum_{j=1}^N \exp(-\beta z_j)} > 0$. Since every partial derivative is positive, the softmin function is increasing in each variable z_i. Therefore, if two vectors $z, w \in \mathbb{R}^N$ satisfy $z_i \le w_i$ for $i = 1, \dots,$ then

$$\operatorname{softmin}_\beta(z) \le \operatorname{softmin}_\beta(w).$$

Further, for any vector $w \in \mathbb{R}^N$, by direct computation, we have that $\operatorname{softmin}_\beta(w + C\mathbf{1}) = \operatorname{softmin}_\beta(w_1 + C, \dots, w_N + C) = \operatorname{softmin}_\beta(w) + C$, which show the translation-equivariant nature of softmin.

Therefore, we have that

$$\operatorname{softmin}_\beta(z) \le \operatorname{softmin}_\beta(Z' + M_{\text{diff}}\mathbf{1}) = \operatorname{softmin}_\beta(z) + M_{\text{diff}}.$$

Consequently, we have that,

$$\|\operatorname{softmin}_\beta(z) - \operatorname{softmin}_\beta(z')\|_\infty \le M_{\text{diff}} \le \|z - z'\|_\infty.$$

□

Lemma 10 (Lipschitz continuity of LSE functional with respect to SHK metric)**.** *Assume that the domain Ω is open, bounded and convex with diameter $D \coloneqq \sup_{x,y \in \Omega} \|x - y\| < \infty$. Then, for stored patterns $X_1, \dots, X_N \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ and $F_i \coloneqq S_\varepsilon(\cdot, X_i)$, we have that*

$$|F_i(\xi) - F_i(\xi')| \le G d_{\text{SHK}}(\xi, \xi')$$

and

$$|E(\xi) - E(\xi')| \le G d_{\text{SHK}}(\xi, \xi') \quad \forall \xi, \xi' \in P_M(\Omega).$$

where $G \coloneqq D\sqrt{1 + \frac{D^2}{\lambda^2}}$. Therefore,$F_i$ and E are both Lipschitz continuous with respect to the SHK metric with Lipschitz constant G, and therefore are continuous as well.

Proof. With $F_i = S_\varepsilon(\cdot, X_i)$, applying Lemma 9 with $z_i = F_i(\xi)$ and $z'_i = F_i(\xi')$, we have that

$$|E(\xi) - E(\xi')| \le \max_{1 \le i \le N} |F_i(\xi) - F_i(\xi')|.$$

Again, by Lemma 8, we have, for $i = 1, \dots, N$,

$$|F_i(\xi) - F_i(\xi')| \le G d_{\text{SHK}}(\xi, \xi').$$

Taking maximum over i and combining with the previous inequality, we have that

$$|E(\xi) - E(\xi')| \le G d_{\text{SHK}}(\xi, \xi').$$

□

Lemma 11 (Mean separation of patterns implies margin separation of Sinkhorn divergences)**.** *Let $X_1, \dots, X_N \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ and define $\mu_i \coloneqq m(X_i) \in \mathbb{R}^d$. Assume the means are pairwise separated:*

$$\|\mu_i - \mu_j\| \ge d_{\min} \quad \textit{for all } i \ne j.$$

Fix any $\varepsilon > 0$ such that

$$d_{\min}^2 > 32\varepsilon \log M$$

and define

$$r \coloneqq \frac{d_{\min}^2}{32} - \varepsilon \log M, \quad \Delta \coloneqq \frac{d_{\min}^2}{4}$$

Then Assumption (A1) holds with the given choice of r and Δ; i.e. for every i, for every $\xi \in B_i(r)$, and every $j \ne i$,

$$F_j(\xi) - F_i(\xi) \ge \Delta.$$

Proof. Fix an index i and let $\xi \in B_i(r)$, so $S_\varepsilon(\xi, X_i) \le r$. Using Lemma 14 with $(\mu,\nu) = (\xi, X_i)$, we have that:

$$S_\varepsilon(\xi, X_i) \ge \frac{1}{2}\|m(\xi) - \mu_i\|^2 - \varepsilon \log M.$$

Combining with $S_\varepsilon(\xi, X_i) \le r$, we have that

$$r \ge \frac{1}{2}\|m(\xi) - \mu_i\|^2 - \varepsilon \log M \Longrightarrow \frac{1}{2}\|m(\xi) - \mu_i\|^2 \le r + \varepsilon \log M.$$

By the definition of r, we have that $r + \varepsilon \log M = \frac{d_{\min}^2}{32}$. Hence

$$\|m(\xi) - \mu_i\|^2 \le 2 \cdot \frac{d_{\min}^2}{32} = \frac{d_{\min}^2}{16} \Longrightarrow \|m(\xi) - \mu_i\| \le \frac{d_{\min}}{4}.$$

Fix $j \ne i$. Then, triangle inequality gives

$$\|m(\xi) - \mu_j\| \ge \|\mu_i - \mu_j\| - \|m(\xi) - \mu_i\| \ge d_{\min} - \frac{d_{\min}}{4} = \frac{3d_{\min}}{4}.$$

Using this along wih Lemma 14, for (ξ, X_j), we have that

$$S_\varepsilon(\xi, X_j) \ge \frac{1}{2}\|m(\xi) - \mu_j\|^2 - \varepsilon \log M \ge \frac{1}{2}\left(\frac{3d_{\min}}{4}\right)^2 - \varepsilon \log M = \frac{9d_{\min}^2}{32} - \varepsilon \log M.$$

Since $\xi \in B_i(r)$, $S_\varepsilon(\xi, X_i) \le r = \frac{d_{\min}^2}{32} - \varepsilon \log M$. Hence , we have that

$$S_\varepsilon(\xi, X_j) - S_\varepsilon(\xi, X_i) \ge \left(\frac{9d_{\min}^2}{32} - \varepsilon \log M\right) - \left(\frac{d_{\min}^2}{32} - \varepsilon \log M\right) = \frac{8d_{\min}^2}{32} = \frac{d_{\min}^2}{4} = \Delta.$$

Thus, Assumption (A1) holds under the given conditions. □

Lemma 12 (Entropic OT lower bound in terms of mean differences)**.** *For any $\mu, \nu \in \mathcal{P}(\Omega)$ and any $\varepsilon > 0$,*

$$\mathrm{OT}_\varepsilon(\mu,\nu) \ge \frac{1}{2}\|m(\mu) - m(\nu)\|^2$$

where $m(\mu) := \int_\Omega x d\mu(x) \in \mathrm{conv}(\Omega) \subset \mathbb{R}^d$ is the mean of μ.

Proof. Given any coupling $\pi \in \Pi(\mu,\nu)$ and random pair $(X,Y) \sim \pi$, define $Z := X - Y$. Then, we have that

$$\mathbb{E}_\pi[Z] = \mathbb{E}_\pi[X] - \mathbb{E}_\pi[Y] = m(\mu) - m(\nu).$$

By Jensen's inequality, we have that

$$\mathbb{E}_\pi\left[\|Z\|^2\right] \ge \|\mathbb{E}_\pi[Z]\|^2 = \|m(\mu) - m(\nu)\|^2.$$

Consequently, using the fact $\mathrm{KL}(\pi \,||\, \mu \otimes \nu) \ge 0$, we have that

$$\begin{aligned}\int_{\Omega\times\Omega} c(x,y)d\pi(x,y) + \varepsilon\,\mathrm{KL}(\pi \,||\, \mu\otimes\nu) \ge &\int_{\Omega\times\Omega} c(x,y)d\pi(x,y)\\ =&\frac{1}{2}\mathbb{E}_\pi\left[\|X-Y\|^2\right]\\ \ge&\frac{1}{2}\|m(\mu) - m(\nu)\|^2.\end{aligned}$$

Since this is true for any $\pi \in \Pi(\mu,\nu)$, taking the infimum over $\pi \in \Pi(\mu,\nu)$, we have that

$$\mathrm{OT}_\varepsilon(\mu,\nu) \ge \frac{1}{2}\|m(\mu) - m(\nu)\|^2.$$

□

Lemma 13 (Self Entropic OT distance upper bound)**.** *If $\mu = \sum_{m=1}^M a_m \delta_{x_m} \in \mathcal{P}_{M,a_{\min},\Delta_{\min}}(\Omega)$ with $a_m > 0$, then*

$$\mathrm{OT}_\varepsilon(\mu,\mu) \le \varepsilon \log M.$$

Proof. Let $\pi_{\text{diag}} \coloneqq \sum_{m=1}^{M} a_m \delta_{(x_m, x_m)} \in \Pi(\mu, \mu)$ denote the identity/diagonal coupling and its corresponding transport cost is given by

$$\int c(x,y) d\pi_{\text{diag}}(x,y) = \sum_{m=1}^{M} a_m \times \frac{1}{2} \|x_m - x_m\|^2 = 0.$$

On the discrete support $\{(x_m, x_\ell)\}_{m,\ell}$, we have that

$$(\mu \otimes \mu)(x_m, x_\ell) = a_m a_\ell \quad \text{and} \quad \pi_{\text{diag}}(x_m, x_\ell) = \begin{cases} a_m, & m = \ell, \\ 0, & m \neq \ell. \end{cases}$$

Consequently, we have that,

$$\mathrm{KL}(\pi_{\text{diag}} \mid\mid \mu \otimes \mu) = \sum_{m=1}^{M} a_m \log \frac{a_m}{a_m^2} = \sum_{m=1}^{M} a_m \log \frac{1}{a_m} = H(a)$$

where $H(a) \coloneqq -\sum_{m=1}^{M} a_m \log a_m$ is the Shannon entropy of the coupling π_{diag} corresponding to the probability vector $a = (a_1, \ldots, a_m)$ or equivalently, that of the discrete distribution $\sum_{m=1}^{M} a_m \delta_{x_m}$. Using the strict concavity of $x \mapsto \log x$ and Jensen's inequality (Polyanskiy & Wu, 2025, Theorem 1.4(b)), we have that

$$H(a) = \mathbb{E}_{X \sim \mu} \log \left[\frac{1}{\mu(X)} \right] \leq \log \mathbb{E} \left[\frac{1}{\mu(X)} \right] = \log M.$$

Therefore, we have that,

$$\mathrm{OT}_\varepsilon(\mu, \mu) \leq \int c(x,y) d\pi_{\text{diag}}(x,y) + \varepsilon \,\mathrm{KL}(\pi_{\text{diag}} \mid\mid \mu \otimes \mu) = 0 + \varepsilon H(a) \leq \varepsilon \log M.$$

□

Lemma 14 (Sinkhorn divergence lower bound in terms of mean differences)**.** *For any* $\mu, \nu \in \mathcal{P}_{M, a_{\min}, \Delta_{\min}}(\Omega)$,

$$S_\varepsilon(\mu, \nu) \geq \frac{1}{2} \|m(\mu) - m(\nu)\|^2 - \varepsilon \log M.$$

Proof. Using Lemma 12 for the first term and Lemma 13 for the last two terms, we have, from the definition of $S_\varepsilon(\mu, \nu)$:

$$\begin{aligned} &S_\varepsilon(\mu, \nu) \\ =& \mathrm{OT}_\varepsilon(\mu, \nu) - \frac{1}{2} \mathrm{OT}_\varepsilon(\mu, \mu) - \frac{1}{2} \mathrm{OT}_\varepsilon(\nu, \nu) \\ \geq& \frac{1}{2} \|m(\mu) - m(\nu)\|^2 - \frac{1}{2}(\varepsilon \log M) - \frac{1}{2}(\varepsilon \log M) \\ =& \frac{1}{2} \|m(\mu) - m(\nu)\|^2 - \varepsilon \log M. \end{aligned}$$

□

Lemma 15 (pairwise separation of random sign vectors)**.** *Let* $s_i, s_j \in \{\pm 1\}^d$ *be componentwise independent Rademacher random variables and define*

$$\mu_i = c + \frac{R_0}{\sqrt{d}} s_i, \quad \mu_j = c + \frac{R_0}{\sqrt{d}} s_j.$$

Fix $\gamma \in (0,1)$*. Then*

$$\mathbb{P}\left(\|\mu_i - \mu_j\|^2 < 2(1-\gamma) R_0^2 \right) \leq \exp\left(-\frac{\gamma^2}{2} d \right).$$

Proof. Define the random variable $H \coloneqq \sum_{k=1}^{d} \mathbf{1}\{s_{ik} \neq s_{jk}\}$. For any $k = 1, \dots, d$, since s_{ik} and s_{jk} are independent Rademacher random variables, $\mathbf{1}\{s_{ik} \neq s_{jk}\}$ is a Bernoulli random variable with probability parameter $p = \frac{1}{2}$. Using the independence across k, we have that H is the sum of d i.i.d Bernoulli random variables with common probability parameter $p = \frac{1}{2}$. Consequently, $H \sim \text{Binomial}(d, \frac{1}{2})$ and $\mathbb{E}(H) = \frac{d}{2}$.

Note that $|s_{ik} - s_{jk}| = 0$ if they are equal and 2 is different. Further, $\mu_i - \mu_j = \frac{R_0}{\sqrt{d}}(s_i - s_j)$. Therefore

$$\|\mu_i - \mu_j\|^2 = \frac{R_0^2}{d} \sum_{k=1}^{d} (s_{i,k} - s_{j,k})^2 = \frac{4R_0^2}{d} H$$

Thus, the event $\|\mu_i - \mu_j\|^2 < 2(1-\gamma)R_0^2$ is equivalent to

$$\frac{4R_0^2}{d} H < 2(1-\gamma)R_0^2 \quad \Longleftrightarrow \quad H < \frac{1-\gamma}{2} d = \frac{d}{2} - \frac{\gamma d}{2}.$$

Applying Hoeffding's inequality to H, we have that, for any $t > 0$,

$$\mathbb{P}(H - \mathbb{E}H \leq -t) \leq \exp\left(-\frac{2t^2}{d}\right)$$
$$\Longleftrightarrow \mathbb{P}(H - \frac{d}{2} \leq -t) \leq \exp\left(-\frac{2t^2}{d}\right).$$

Choosing $t = \frac{\gamma d}{2}$, we have that

$$\mathbb{P}(H - \mathbb{E}H \leq -\frac{\gamma d}{2}) \leq \exp\left(-\frac{\gamma^2}{2} d\right)$$
$$\Longleftrightarrow \mathbb{P}(H \leq \frac{d}{2} - \frac{\gamma d}{2}) \leq \exp\left(-\frac{\gamma^2}{2} d\right)$$
$$\Longleftrightarrow \mathbb{P}(H \leq \frac{1-\gamma}{2} d) \leq \exp\left(-\frac{\gamma^2}{2} d\right)$$
$$\Longleftrightarrow \mathbb{P}\left(\|\mu_i - \mu_j\|^2 < 2(1-\gamma)R_0^2\right) \leq \exp\left(-\frac{\gamma^2}{2} d\right).$$

□

Lemma 16 (uniform separation for all pairs of means of stored patterns)**.** *Let $\mu_1, \dots, \mu_N$ be defined as in Lemma 15 using the i.i.d. componentwise Rademacher random variables $s_1, \dots, s_N \in \{\pm 1\}^d$. Fix $p \in (0,1)$ and $\gamma \in (0,1)$. If*

$$N \coloneqq \left\lfloor \sqrt{2p} \exp\left(\frac{\gamma^2}{4} d\right) \right\rfloor.$$

then with probability at least $1 - p$,

$$\|\mu_i - \mu_j\| \geq d_{\min} \coloneqq \sqrt{2(1-\gamma)} R_0 \quad \textit{for all } i \neq j.$$

Proof. Let A_{ij} be the bad event $\left\{\|\mu_i - \mu_j\|^2 < 2(1-\gamma)R_0^2\right\}$. By Lemma 15,

$$\mathbb{P}(A_{ij}) \leq \exp\left(-\frac{\gamma^2}{2} d\right) \qquad \text{for each } i \neq j.$$

By the union bound over $\binom{N}{2} \leq \frac{N^2}{2}$ pairs,

$$\mathbb{P}(\exists\ i < j : A_{ij}) \leq \frac{N^2}{2} \exp\left(-\frac{\gamma^2}{2} d\right).$$

Under the stated condition on N, the RHS is $\leq p$. Therefore, with probability at least $1 - p$, no bad event occurs, i.e. all pairs satisfy

$$\|\mu_i - \mu_j\|^2 \geq 2(1-\gamma)R_0^2 \Longleftrightarrow \|\mu_i - \mu_j\| \geq \sqrt{2(1-\gamma)} R_0 = d_{\min}.$$

□

Lemma 17 (Linear independence of Dirac distribution and its distributional derivative)**.** *Let $(x_1, \ldots, x_M) \in \Omega \subset \mathbb{R}^d$ be pairwise distinct location parameters. Suppose*

$$\sum_{i=1}^{M} c_i \delta_{x_i} + \sum_{i=1}^{M} v_i \cdot \nabla \delta_{x_i} = 0 \quad \textit{in } \left(C^\infty(\mathbb{R}^d)\right)',$$

with $c_i \in \mathbb{R}$ and $v_i \in \mathbb{R}^d$. Then $c_i = 0$ and $v_i = 0$ for all $i = 1, \ldots, M$.

Proof. Fix $k \in \{1, \ldots, M\}$. Given any $x \in \Omega$ and $r \geq 0$, define the Euclidean ball $B(x, r) \coloneqq \{y \in \Omega : \|y - x\| \leq r\}$. Since the x_i's are distinct, there exists $r_k > 0$ such that $B(x_k, r_k) \cap \{x_j : j \neq k\} = \emptyset$. Let us choose $\psi \in C_c^\infty\left(\mathbb{R}^d\right) \subset C^\infty(\mathbb{R}^d)$ supported in $B(x_k, r_k)$ and equal to 1 in a neighborhood of x_k. Then, we have that, $\langle \delta_{x_i}, \psi \rangle = \begin{cases} \psi(x_i) = 1, & i = k \\ \psi(x_k) = 0, & i \neq k \end{cases}$ and $\langle \nabla \delta_{x_i}, \psi \rangle = -\nabla \psi(x_i) = 0$ for $i = 1, \ldots, M$, since ψ is constant near x_k and vanishes near x_i for $i \neq k$. Consequently, we have that

$$\begin{aligned}
&\sum_{i=1}^{M} \langle c_i \delta_{x_i}, \psi \rangle + \sum_{i=1}^{M} \langle v_i \cdot \nabla \delta_{x_i}, \psi \rangle = \langle 0, \psi \rangle \\
\iff &\sum_{i=1}^{M} c_i \psi(x_i) - \sum_{i=1}^{M} v_i \cdot \nabla \psi(x_i) = 0 \\
\iff &c_k = 0.
\end{aligned}$$

Now, let us fix any vector $u \in x \subset \mathbb{R}^d$. Then, let us choose $\psi_u \in C_c^\infty\left(\mathbb{R}^d\right) \subset C^\infty(\mathbb{R}^d)$ supported in $B(x_k, r_k)$ with $\psi_u(x_k) = 0$ and $\nabla \psi_u(x_k) = u$. Then, we have that $\langle \delta_{x_i}, \psi_u \rangle = 0 \quad$ for $i = 1, \ldots, M$ and $\langle \nabla \delta_{x_i}, \psi_u \rangle = -\nabla \psi_u(x_i) = \begin{cases} -\nabla \psi_u(x_k) = -u, & i = k, \\ -\nabla \psi_u(x_i) = 0, & i \neq k \end{cases}$. Consequently, we have that

$$\begin{aligned}
&\sum_{i=1}^{M} \langle c_i \delta_{x_i}, \psi_u \rangle + \sum_{i=1}^{M} \langle v_i \cdot \nabla \delta_{x_i}, \psi_u \rangle = \langle 0, \psi_u \rangle \\
\iff &\sum_{i=1}^{M} c_i \psi_u(x_i) - \sum_{i=1}^{M} v_i \cdot \nabla \psi_u(x_i) = 0 \\
\iff &- v_k \cdot u = 0.
\end{aligned}$$

Since this is true for any $u \in \Omega \in \mathbb{R}^d$, we must have that $v_k = 0$. Finally, since k was arbitrary, we must have that $c_i = 0$ and $v_i = 0$ for all $i = 1, \ldots, M$. $\square$