MAXIMUM ENTROPY UNDER CARRÉ DU CHAMP CONSTRAINTS

Tanya Veeravalli
Centre for Frontier AI Research (CFAR)
Institute of High Performance Computing, Agency for Science, Technology and Research, 138632 Singapore
veeravalli_tanya@a-star.edu.sg

Atsushi Nitanda
Centre for Frontier AI Research (CFAR)
Institute of High Performance Computing, Agency for Science, Technology and Research, 138632 Singapore
Nanyang Technological University, 639798 Singapore
atsushi_nitanda@a-star.edu.sg

ABSTRACT

Recent works have shown generative models' striking capabilities to memorize training data, and encode the underlying lower-dimensional data manifold M. However, while this phenomenon has been studied from different perspectives, current training objectives do not explicitly *control* the geometric complexity of memorization. In this work, we propose a constrained maximum-entropy (MaxEnt) principle for learning a generative model density q that (i) fits the data, (ii) discourages tangential "memorization" (training-point attractors along the manifold) through a data-driven Carré-du-Champ (CDC) tangential Fisher constraint defined on the projected marginal of q onto M, and (iii) enforces bounded normal thickness so the ambient entropy objective is well-posed. Leveraging this, we prove KL-to-uniform control and explicit anti-concentration bounds near training points, and lift them to ambient space via thickness control. We also provide scalable kNN/local-PCA estimators for the data-driven geometric terms and demonstrate numerical and simulation results. More generally, to our knowledge, this is the first work proposing a principled variational principle that jointly controls entropy and geometric complexity through a data-dependent metric.

1 INTRODUCTION

Data distributions encountered in modern deep generative modeling problems are often effectively low-dimensional: observations in $\mathbb{R}^d$ concentrate around an embedded lower-dimensional manifold M of dimension $m \ll d$ Narayanan & Mitter (2010); Fefferman et al. (2013). Models aim to sample from an unknown density $p(x)$ concentrated on M given only a set of finite data points $S_N = \{x^{(i)}\}_{i=1}^N \subset \mathbb{R}^d$. Some recent examples where the manifold hypothesis acted as a guiding principle include diffusion models, flow models, stochastic interpolants, and normalizing flows Ho et al. (2020); Sohl-Dickstein et al. (2015); Lipman et al. (2023); Rezende & Mohamed (2015); Albergo et al. (2025), among various others. In Stanczuk et al. (2024), the authors estimate M's dimension by computing the intrinsic dimensionality of the normal bundle, and demonstrate that near the data manifold, the score (gradient of the log-density of the data distribution) is orthogonal. Like so, there are other studies of emerging geometric behavior in diffusion/flow/autoregressive models that offer increased understanding into various phenomena. **In this work, we are interested in maximizing the entropy of a generative model under the manifold hypothesis, while mitigating memorization.**

Maximum Entropy Maximum entropy (MaxEnt) has had a long history Jaynes (1957); Kullback (1997) and is enjoying a resurgence in the generative modeling and reinforcement learning (RL) communities for its wide applicability in problems requiring diversity, robustness, and/or generalizability

Yang et al. (2025). Eysenbach & Levine (2022) show that MaxEnt RL maximizes a lower bound on a robust RL objective, and learns policies that are robust to disturbances in the dynamics and the reward function. Another concurrent work Smith et al. (2026) deals with posing the calibration/fine-tuning of generative models as a constrained maximum entropy optimization problem with distributional constraints. Chaudhari et al. (2017) also leverage local geometry but to construct a local-entropy-based objective function that generalizable solutions lying in large flat regions. Meanwhile, MaxEnt along with a data-fit constraint alone has been observed to drive the distribution towards the data points, so there needs to be a principled approach to mitigate memorization in this case as well.

Memorization There has been a great deal of recent interest in studying the memorization and generalization abilities of such models, e.g. Ye et al. (2025); Wang et al. (2025). While these models can match marginals and produce high-quality samples, they can still exhibit memorization-like behavior, manifesting as reduced diversity and sample trajectories that drift toward specific training samples Somepalli et al. (2023). In other words, as Bamberger et al. (2025) remarked, high quality samples alone cannot measure the performance of a generative model as it can be easily be achieved by reproducing training samples; more importantly, they also highlight that memorization occurs due to tangential flows of the velocity field associated with the model, creating attractors Achilli et al. (2024); Biroli et al. (2024).

Standard fitting/learning objectives such as score-matching, maximum likelihood estimation, or flow matching prioritize data fit but do not directly control tangential concentration along M. Conversely, maximizing ambient differential entropy alone is ill-posed in $\mathbb{R}^d$ as vanishing mass can be placed arbitrarily far from the data. Santi et al. (2025) consider MaxEnt as a fine-tuning problem where the authors were interested in exploration over an approximate data manifold implicitly-defined by a pre-trained diffusion model. Other recent work leverages local geometric covariance information in generative path constructions in the setting of flow matching Bamberger et al. (2025).

In this work, we propose a constrained maximum entropy principle for selecting a model density q on $\mathbb{R}^d$: maximize ambient entropy subject to data fit and geometric regularity constraints. Data-fit can be enforced by any standard constraint; here, we took it to be the negative log likelihood (NLL). To make the ambient entropy objective well-posed and prevent mass-at-infinity solutions, we impose a thickness constraint that keeps q concentrated on a tubular neighborhood of M. To suppress memorization, we introduce a tangential Carré du Champ (CDC)-type constraint on the projected marginal q_M on M, defined by a CDC quadratic form that's estimated from data.

Contributions

- We propose a general MaxEnt-CDC problem with constraints to prevent formation of narrow modes along the data manifold M, targeting memorization.
- Under a log-Sobolev inequality (LSI) on the data distribution p on M and a quantitative projector-accuracy on the data-driven CDC, we prove intrinsic Kullback-Leibler (KL) control and explicit anti-concentration bounds near training points; thickness lifts these guarantees to ambient neighborhoods in $\mathbb{R}^d$.
- We provide scalable estimators for CDC regularization and distance-to-manifold proxies, demonstrate increased diversity, and reduced reproduction of training points through experiments (in Appendix A).

2 Geometric setting and measures on a manifold

Assumption 2.1 (Compact C^2 manifold with reach)**.** Let $M \subset \mathbb{R}^d$ be a compact, connected, C^2 embedded manifold of dimension m without boundary. Assume M has positive reach: there exists $r_0 > 0$ such that every x in the tubular neighborhood

$$U_{r_0}(M) := \{x \in \mathbb{R}^d : \operatorname{dist}(x, M) < r_0\}$$

has a unique nearest point on M. Denote by $\pi : U_{r_0}(M) \to M$ the nearest-point projection. Incidentally, π is locally L_π-Lipshitz.

Let $\sigma := \frac{\mathrm{vol}_M}{\mathrm{vol}_M(M)}$ be the canonical reference measure on M. Let q be a probability density on $\mathbb{R}^d$ with corresponding probability measure (still denoted q) and assume $q(U_{r_0}(M)) = 1$ for

simplicity (or $q(U_{r_0}(M)) \approx 1$). The pushforward on M is $q_M := \pi \# q$, where $q_M(A) := q(\pi^{-1}(A))$ for a measurable $A \subset M$. Assume q_M is absolutely continuous with respect to σ: $q_M(A) = \int_A \rho(y)\, \sigma(dy)$.

3 THE MAXENT-CDC PROBLEM

In this section, we develop the theoretical foundations of the proposed maximum entropy under the carré du champ constraint (MaxEnt-CDC) formulation. We begin by motivating the principle through the memorization-quality trade-off in generative models, then introduce the carré du champ Fisher information and derive the Euler-Lagrange conditions. We further prove existence and qualitative properties of solutions and connect the formulation to Bakry-Émery curvature, log-Sobolev inequalities, and the geometry learned by diffusion models. Throughout, we leverage the data-dependent diffusion geometry estimated via a nearest-neighbors approach. To our knowledge, this is the first variational principle connecting geometry, entropy, memorization, and generalization in generative modeling problems.

3.1 SETUP

Let q denote a probability density on $\mathbb{R}^d$, and $\Gamma(x) \in \mathbb{R}^{d\times d}$ be a symmetric positive semidefinite matrix, and uniformly bounded; it will be estimated from data as we will see in Section A. Denote our (unknown) data distribution by p^{data} – we only have access to samples $S_N = \{x^{(i)}\}_{i=1}^N \sim p^{\text{data}} \subset \mathbb{R}^d$.

Definition 3.1 (Entropy)**.** The (ambient) entropy is defined as

$$H(q) := -\int_{\mathbb{R}^d} q(x) \log q(x)\, \mathrm{d}x \tag{1}$$

It is unbounded over all densities on $\mathbb{R}^d$: a model can increase $H(q)$ by allocating a small amount of probability mass far from the data (where likelihood constraints are not evaluated), while changing the data-fit objective negligibly. To make this well-posed, we impose a *thickness* control:

Definition 3.2 (Normal thickness control)**.** This constraint controls the thickness around the manifold M, and prevents the model from meeting the negative log-likelihood but wasting entropy by going far off-manifold:

$$\mathbb{E}_{X\sim q}[\mathrm{dist}(X, M)^2] \le \delta^2$$

We propose the following constrained MaxEnt problem:

How can we do maximum entropy learning while ensuring that memorization of the training data is minimized?

$$\begin{aligned} &\max_q H(q) \ \text{ s.t.} \\ &\mathcal{I}_\Gamma(\rho) \le C_{\text{tan}}, \ \int q(x)\, \mathrm{d}x = 1, \ \ q \ge 0 \\ &\mathcal{D}(q, p^{\text{data}}) \le C_{\text{nll}}, \ \mathbb{E}_{X\sim q}[\mathrm{dist}(X, M)^2] \le \delta^2 \end{aligned} \tag{2}$$

$\mathcal{D}(\cdot,\cdot)$ can be an arbitrary metric or divergence between data density p^{data} and q. In practice, we will take this to be the negative log-likelihood (NLL) on data: $-\mathbb{E}_{x\sim p^{\text{data}}}[\log q(x)]$ (which is equivalent to taking the Kullback-Leibler (KL)). We think of this as a "data-anchoring" term so that the entropy does not converge to the uniform distribution. Notably, since p^{data} is unknown, *we use the empirical NLL (obtained from S_N)*. $\mathcal{I}_\Gamma$ is the CDC quadratic form containing the data-driven $\Gamma(x)$ (which measures local tangent projection). We will dive into the CDC regularization further in Section 4. This formalizes our goal to maximize diversity of the generative model subject to bounded geometric curvature in data-adaptive coordinates, while respecting the data manifold.

We form a variational principle using the Euler-Lagrange equations to solve this problem. For more details, we refer to Appendix D. The following sections will improve our understanding of the role of each constraint in Problem 2.

4 CDC TANGENTIAL FISHER CONTROL AND INTRINSIC FISHER INFORMATION

Recall that our goal is to suppress memorization understood as tangential collapse of the model distribution along the data manifold. A natural way to quantify tangential accumulation is the intrinsic Fisher information:

$$\mathcal{I}(\rho) = \int_M \left\|\nabla_M \log \rho(y)\right\|^2 \rho(y) d\sigma(y) \tag{3}$$

where ρ is the intrinsic density on M defined by q_M. $\mathcal{I}(\rho)$ penalizes high-frequency variation of $\log \rho$ along M. Unsurprisingly, this is hard to directly evaluate as M and the intrinsic gradient ∇_M are unknown. However, since $M \hookrightarrow \mathbb{R}^d$, intrinsic gradients can be expressed using the ambient gradients via the tangent projector:

Definition 4.1 (Intrinsic gradient ∇_M via the tangent projector onto M)**.** Let the orthogonal projector onto the tangent space $T_y M \subset \mathbb{R}^d$ be $P_T(y)$. Then, for any smooth function f on M and any smooth extension $\tilde{f}$ off M,

$$\nabla_M f(y) = P_T(y) \nabla \tilde{f}$$

The above suggests replacing the unknown projector $P_T(y)$ by a data-driven surrogate $\Gamma(y)$. We then have a CDC Fisher proxy

Definition 4.2 (Carré du champ (CDC) regularizer)**.** Given a C^2 function $f : \mathbb{R}^d \to \mathbb{R}$, the carré du champ operator is

$$\boxed{\mathcal{I}_\Gamma(\rho) \coloneqq \int_M \langle \nabla \log \rho(y), \Gamma(y) \nabla \log \rho(y) \rangle \rho(y) \, \mathrm{d}\sigma(y).} \tag{4}$$

$\Gamma(x)$ is estimated using local neighborhood statistics inspired by Bamberger et al. (2025); the data-driven estimation procedure can be found in Appendix A. In a sufficiently small neighborhood around y, the data cloud is approximately supported on an m-dimensional affine subspace aligned with $T_y M$, so the top-m eigenspace of a kernel-weighted local covariance provides a consistent estimate of the tangent space. The covariance behaves like a local tangent projector. The quadratic form $\langle \nabla f, \Gamma \nabla f \rangle$ is the standard carré du champ energy associated with a local metric Γ (in fact, it is a Riemannian metric); here, we use it solely as a geometry-aware regularizer.

Next, we make an assumption about the projector accuracy $\Gamma \approx P_T$ (which, as we will see, is verified numerically):

Assumption 4.3 (CDC projector accuracy)**.** Recall $\Gamma : M \to \mathbb{R}^{d \times d}$ is symmetric and PSD by virtue of it being a local covariance estimate. Assume there exists $\epsilon \in [0, 1)$ such that for all $y \in M$, $(1 - \epsilon) P_T(y) \preceq \Gamma(y) \preceq (1 + \epsilon) P_T(y)$, in the sense of quadratic forms.

The following shows us the equivalence of $\mathcal{I}_\Gamma$ and the intrinsic Fisher quantity:

Lemma 4.4 (Equivalence of $\mathcal{I}_\Gamma$ and $\mathcal{I}$)**.** *Under Assumption 4.3, for any smooth $\rho > 0$ with $\int \rho \, \mathrm{d}\sigma = 1$,* $(1 - \epsilon)\mathcal{I}(\rho) \leq \mathcal{I}_\Gamma(\rho) \leq (1 + \epsilon)\mathcal{I}(\rho)$.

Proof. We refer to Appendix B. □

4.1 LOG-SOBOLEV INEQUALITY (LSI) AND INTRINSIC DIVERSITY

The CDC tangential Fisher constraint introduced in the previous section penalizes sharp variation of ρ and hence $\log \rho$ along the data manifold. While this aligns with the heuristic notion that memorization corresponds to concentration near training points, Fisher-type quantities are gradient energies and do not, by themselves, directly quantify distributional closeness or diversity. To turn tangential smoothness into a quantitative non-collapse principle, we make use of a standard functional-analytic bridge: the log-Sobolev inequality (LSI) on (M, σ). LSI relates KL divergence (w.r.t. the uniform measure σ) to the intrinsic Fisher information, yielding an explicit bound of the form $\mathrm{KL}(q_M || \sigma) \lesssim \mathcal{I}(\rho)$. Combining this with projector-accuracy $\Gamma \approx P_T$ converts the CDC constraint $\mathcal{I}(\rho) \leq C_{\tan}$ into an intrinsic KL control guarantee. This, in turn, implies anti-concentration: the probability assigned by q_M to unions of small neighborhoods around projected training points cannot

exceed the corresponding uniform baseline volume by more than a factor controlled by said KL guarantee. This allows us a clean mechanism to show CDC tangential control prevents memorization in the sense of tangential collapse.

Assumption 4.5 (LSI on (M, σ))**.** There exists $\lambda > 0$ such that for all smooth $f : M \to \mathbb{R}$ with $\int_M f^2 \, \mathrm{d}\sigma = 1$,

$$\mathrm{Ent}_\sigma(f^2) := \int_M f^2 \log f^2 \, \mathrm{d}\sigma \leq \frac{2}{\lambda} \int_M \|\nabla_M f\|^2 \, d\sigma.$$

Remark 4.6. LSI holds for many compact manifolds under mild curvature/regularity conditions; moreover, it allows dimension-free conversion and thus we use it as a sufficient condition.

We will need the following lemma to help us with formalizing anti-concentration bounds later:

Lemma 4.7 (LSI implies KL-Fisher bound)**.** *Under Assumption 4.5, for any smooth density $\rho > 0$ with $\int \rho \, \mathrm{d}\sigma = 1$,*

$$\mathrm{KL}(\rho||1) := \int_M \rho \log \rho \, \mathrm{d}\sigma \leq \frac{\mathcal{I}(\rho)}{2\lambda}$$

Proof. We refer to Appendix B. □

Theorem 4.8 (CDC Fisher bound implies intrinsic KL control)**.** *Assume Assumptions 4.3 and 4.5. If $\mathcal{I}_\Gamma(\rho) \leq C_{\tan}$, then*

$$\mathrm{KL}(q_M||\sigma) = \int_M \rho \log \rho \, \mathrm{d}\sigma \leq \frac{C_{\tan}}{2\lambda(1-\epsilon)}$$

Proof. By Lemma 4.4, $\mathcal{I}(\rho) \leq \frac{1}{1-\epsilon}\mathcal{I}_\Gamma(\rho) \leq \frac{C_{\tan}}{1-\epsilon}$. Applying Lemma 4.7, we arrive at the statement. In essence, this KL-to-uniform control is a *diversity guarantee*. □

4.2 Intrinsic anti-concentration near training points

The previous section shows that the CDC Fisher constraint yields a bound on $\mathrm{KL}(q_M||\sigma)$, meaning that the projected model distribution q_M cannot deviate too far from the uniform reference measure σ on M. To connect this global notion of diversity to memorization, we study how much probability mass q_M assigns to neighborhoods of the training samples S_N. If the model "memorizes", then a non-negligible fraction of the model's probability must concentrate in unions of small balls around the projected training points.

We work with the projected training points $y^{(i)} = \pi(x^{(i)})$ (from tubularity). To make things clearer, and to avoid geodesic vs Euclidean subtleties, we work with extrinsic balls on M:

$$B_M^{\mathrm{ext}}(y, r) := \{z \in M : \|z - y\| < r\}.$$

Next, we define intrinsic memorization mass:

$$\mathrm{Mem}_M^r(q_M, S_N) := q_M\left(\cup_{i=1}^N B_M^{\mathrm{ext}}(y^{(i)}, r)\right)$$

We make use of Pinsker's inequaity, which states that for probability measures $\mu \ll \nu$ on the same measurable space, $\|\mu - \nu\|_{\mathrm{TV}} \leq \sqrt{1/2\mathrm{KL}(\mu||\nu)}$.

Theorem 4.9 (Anti-concentration bound)**.** *Under the assumptions of Theorem 4.8, for all $r > 0$,*

$$\mathrm{Mem}_M^r(q_M, S_N) \leq \sigma\left(\cup_{i=1}^N B_M^{\mathrm{ext}}(y^{(i)}, r)\right) + \sqrt{\frac{C_{\tan}}{4\lambda(1-\epsilon)}}.$$

Proof. Define the set $A := \cup_{i=1}^N B_M^{\mathrm{ext}}(y^{(i)}, r)$. Then, $q_M(A) \leq \sigma(A) + \|q_M - \sigma\|_{\mathrm{TV}}$. From here, a simple application of Pinsker's inequality suffices. □

5 LIFTING INTRINSIC CONTROL TO AMBIENT SPACE CONTROL VIA "NORMAL THICKNESS"

Theorem 4.9 yielded anti-concentration around the projected training points $y^{(i)}$, and controlled the projected model distribution $q_M = \pi_{\#}q$ on M. However, the model q and the observed samples $x^{(i)}$ live in the ambient $\mathbb{R}^d$ and memorization is also often assessed in ambient coordinates (say, for example, neighborhoods in pixel space). Therefore, this necessitates the translation of intrinsic non-collapse on M into an ambient space statement about the mass q assigns. The "lifting" relies on two geometric ingredients: (i) *projection regularity*, which ensures that nearby points in $\mathbb{R}^d$ have projections that are also nearby on M, and (ii) *thickness control*, guaranteeing most of the model mass lies within a tubular neighborhood where the projection map is well-defined.

We define ambient memorization at scale $r > 0$, to be

$$\mathrm{Mem}_r^{\mathbb{R}^d}(q, S_N) \coloneqq q\left(\cup_{i=1}^N B_{\mathbb{R}^d}(x^{(i)}, r)\right).$$

With this, we observe the relationship between intrinsic memorization and ambient memorization:

Lemma 5.1 (Projection inclusion)**.** *Under Assumption 2.1 and recalling $q(U_{r_0}(M)) = 1$ (for simplicity), for any $r < r_0$, $B_{\mathbb{R}^d}(x^{(i)}, r) \subset \pi^{-1}(B_M^{\mathrm{ext}}(y^{(i)}, L_\pi r))$ for each i. Consequently,*

$$\mathrm{Mem}_r^{\mathbb{R}^d}(q, S_N) \le q_M\left(\cup_{i=1}^N B_M^{\mathrm{ext}}(y^{(i)}, L_\pi r)\right) = \mathrm{Mem}_{L_\pi r}^M(q_M, S_N).$$

Proof. We refer to Appendix B. □

This lemma has a direct corollary, whereby we can formally derive an anti-concentration bound in ambient space

Corollary 5.2 (Ambient anti-concentration bound with thickness control)**.** *Using Theorem 4.9 and under the assumption of Lemma 5.1, for all $r > 0$ with $L_\pi r < r_0$,*

$$\mathrm{Mem}_r^{\mathbb{R}^d}(q, S_N) \le \sigma\left(\cup_{i=1}^N B_M^{\mathrm{ext}}(y^{(i)}, L_\pi r)\right) + \sqrt{\frac{C_{\tan}}{4\lambda(1-\epsilon)}}$$

Proof. The proof follows from Lemma 5.1 to reduce ambient mass to intrinsic mass at radius $L_\pi r$ and then applying Theorem 4.9. □

Remark 5.3. Note that we have worked under the assumption $q(U_{r_0}(M)) = 1$ (or ≈ 1). However, we have, in practice, the thickness constraint $\mathbb{E}_q[\mathrm{dist}(\mathrm{X}, \mathrm{M})] < \delta^2$. In this case, for any $t < r_0$,

$$q(\underbrace{U_t(M)^c}_{\text{the complement}}) = \mathbb{P}(\mathrm{dist}(X, M) \ge t) \le \frac{\delta^2}{t^2}$$

by Markov's inequality. Then, noticing that

$$q(\cup_i B(x^{(i)}, r)) \le q((\cup_i B(x^{(i)}, r) \cap U_t(M)) + q(U_t(M)^c),$$

we can get finer-grained estimates for Corollary 5.2 where we will incur an additive δ^2/t^2 tail term.

6 RATES OF APPROXIMATIONS OF THE DATA-DRIVEN ESTIMATORS

In this part, we analyze the gap between the estimators $\widehat{\Gamma}(x)$, $\widehat{\mathrm{dist}}(x, M)$ (estimated via local covariance procedures as in Appendix A) and the idealized assumptions $\Gamma \approx P_T$ and the Lipschitz projection. We state rates in a parameterized form.

Assumption 6.1 (Local sampling and noise model)**.** Let Z be a random point on M with density bounded above and below w.r.t. σ. Assume observations are of the form $x = Z + \xi$, where ξ is mean-zero noise, $\mathbb{E}[\xi\xi^\top] = \Sigma_\xi$ and $\|\xi\| \le \tau$ almost surely (or sub-Gaussian with parameter τ). Assume the second fundamental form of M is bounded so that locally M deviates from its tangent plane by $O(h^2)$ at scale h.

Remark 6.2. Assumption 6.1 is a convenient way to formalize "data concentrate near M" with thickness and curvature. The rates below separate (i) curvature bias (ii) statistical error from finite k (the number of neighbors). The following is a generic subspace perturbation lemma.

Lemma 6.3 (Projector perturbation bound)**.** *Let S be a symmetric PSD matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Let $U \in \mathbb{R}^{d\times m}$ span the top-m eigenspace, and $P := UU^\top$. Let $\widehat{S} = S + E$ with symmetric perturbation E, and let $\widehat{P}$ be the top-m projector of $\widehat{S}$. If the eigengap $\Delta := \lambda_m - \lambda_{m+1} > 0$, then*

$$\|\widehat{P} - P\|_{\mathrm{op}} \;\leq\; \frac{2\|E\|_{\mathrm{op}}}{\Delta}.$$

Proof. We refer to Appendix B. □

This justifies the following – if the local covariance has a clear eigengap between tangential and normal directions, then local PCA recovers the tangent space. The required projector-accuracy assumption $\left\|\widehat{\Gamma} - P_T\right\|_{\mathrm{op}} \leq \epsilon$ holds whenever the covariance estimation error $\left\|\widehat{S} - S\right\|_{\mathrm{op}}$ is $O(\epsilon\Delta)$.

6.1 Rate for $\widehat{\Gamma}$ as an approximation of P_T

We now interpret S as the *population* local covariance in a neighborhood of radius h around an anchor point, and $\widehat{S}$ as the empirical weighted covariance $\widehat{S}_i$ computed from k neighbors.

Proposition 6.4 (Parameterized local PCA tangent-space rate)**.** *Fix $y \in M$, and consider a neighborhood scale h (e.g. the typical k-NN radius). Let $P_h(y)$ be the population top-m projector of S at y, with $\Delta(h)$ the corresponding eigengap. Under Assumption 6.1, suppose S satisfies:*

1. *(Eigenspace separation) The top-m eigenspace equals T_yM and has eigengap $\Delta(h) > 0$ to the remaining $(d-m)$ directions;*
2. *(Curvature bias) Deviation from the true tangent projector obeys $\|P_h(y) - P_T(y)\|_{\mathrm{op}} \leq C_{\mathrm{curv}}h$ or $C_{\mathrm{curv}}h^2$.*

Let $\widehat{\Gamma}_i$ be the empirical top-m projector from k samples in the neighborhood. Then with probability at least $1 - \delta_{\mathrm{fail}}$,

$$\|\widehat{\Gamma}_i - P_T(y)\|_{\mathrm{op}} \;\leq\; \underbrace{\|P_h(y) - P_T(y)\|_{\mathrm{op}}}_{\textit{curvature bias}} + \underbrace{\frac{2\|\widehat{S} - S\|_{\mathrm{op}}}{\Delta(h)}}_{\textit{statistical \& noise term}},$$

and moreover, for sub-Gaussian noise and bounded moments one typically has a concentration bound of the form

$$\|\widehat{S} - S\|_{\mathrm{op}} \;\lesssim\; C_1\Big(\tau^2 + h^2\Big)\sqrt{\frac{\log(d/\delta_{\mathrm{fail}})}{k}} + C_2\Big(\tau^2 + h^2\Big)\frac{\log(d/\delta_{\mathrm{fail}})}{k},$$

where C_1, C_2 depend on the weighting scheme and moment constants.

Proof. We refer to Appendix B. □

Remark 6.5 (Regarding Assumption 4.3). If $\widehat{\Gamma}_i$ satisfies $\left\|\widehat{\Gamma}_i - P_T(y)\right\|_{\mathrm{op}} \leq \epsilon$, then $(1-\epsilon)P_T(y) \preceq \widehat{\Gamma}_i \preceq (1+\epsilon)P_T(y)$ holds as quadratic forms on $\mathbb{R}^d$ (a projector perturbation implies a sandwich bound on the tangent subspace), thus matching Assumption 4.3 with $\Gamma = \widehat{\Gamma}$.

6.2 Rate for $\widehat{\mathrm{dist}}(x, M)$ as a thickness proxy

Let $x \in \mathbb{R}^d$ be a point in a tube around M and let $y = \pi(x)$ (which only has a normal component). In reality, since we have data anchors, we will have a tangential error as well. If $\widehat{\Gamma}$ approximates $P_T(y)$ well, then the normal residual $(I - \widehat{\Gamma})(x - x^{(i)})$ approximates the normal component:

Proposition 6.6 (Parameterized distance proxy bound)**.** *Let i be a nearest neighbor anchor to $x \in \mathbb{R}^d$ and assume $x^{(i)} \in U_{r_0}(M)$ with $y^{(i)} = \pi(x^{(i)})$. Assume $||\widehat{\Gamma}_i - P_T(y^{(i)})||_{\mathrm{op}} \leq \epsilon$ and $||y^{(i)} - y|| \leq c_h$ (anchor proximity). Then*

$$|\widehat{\mathrm{dist}}(x, M) - \mathrm{dist}(x, M)| \leq C_{\mathrm{anchor}} c_h + \epsilon ||x - x^{(i)}|| + C_{\mathrm{curv}} \|x - y\|^2$$

Proof. See Appendix B. □

7 Numerical Experiments

In this part, we solve the MaxEnt-CDC problem equation 2 and see the effect each constraint has on the entropy. We train a Real NVP normalizing flow as our generative model, since the log density $\log q$ is readily available Dinh et al. (2017). We take $M = \mathbb{S}^2$, $k = 30$, $d = 3$, and $m = 2$. To solve Problem 2, we form a (relaxed) Lagrangian and perform appropriate penalization for the constraints. More details about leveraging a variational principle can be seen in Appendix D.

Table 1: Results depicting the effects of the different constraints on the MaxEnt (M.E.) objective for a sphere. $\perp \equiv$ normal thickness; $\Gamma \equiv$ CDC constraint. Mem denotes memorization, and Cover $\equiv$ coverage of M

Obj.	$H(q)$	**NLL**	$\mathrm{Mem}_{0.1}^{\mathbb{R}^d}(q, S_N)$	**Cover**	$\perp$
NLL	13.36	-5.6108	0.6690	0.1880	0.0094
M.E.	20.17	-3.1619	0.9650	0.0680	0.0104
M.E. + Γ	14.37	-2.4852	0.6640	0.1700	0.0832
M.E. + $\perp$	20.00	-3.3725	0.4040	0.0680	0.0382
Pb. 2	13.54	-3.7177	**0.2390**	**0.2080**	0.0027

We set the distance for memorization of a point to count to be within 0.1 of a training data point. The results in Table 1 distinctly show that MaxEnt alone increases memorization. This confirms the paper's insight that maximizing entropy without geometric constraints leads to diffuse mass that does not respect manifold structure. The CDC + thickness constraints sacrifice some entropy but achieve meaningful diversity on the manifold. Both together achieve the best result – highest coverage and lowest memorization. **We defer, to Appendix C, other experiments showing sample diversity, memorization $\mathrm{Mem}_r(q, S_N)$ capacity, and verification of the algorithm in Section A.**

8 Limitations

The estimated projection $\widehat{\Gamma}$ is only geometrically meaningful when x is near a training anchor. Samples in "gap" regions receive gradient signals from distant anchors (though this is mitigated by the thickness constraint). We also need to compute ambient scores at each step, and in high-dimensional ambient spaces, this may become expensive unless we take into consideration known efficiencies when employing this principle for a specific generative model along with amortized score estimation (e.g. using a Score Network and score matching).

9 Conclusion

We introduced a geometry-aware maximum-entropy framework for modeling data in $\mathbb{R}^d$ that concentrate near an unknown low-dimensional manifold M. The key idea is to identify memorization-like tangential collapse along M and control it via a Carré du Champ-type Fisher constraint induced by a data-adaptive field Γ that approximates the local tangent projector. We obtain guarantees on the intrinsic diversity and lift them to the ambient space. Experimental results validate our MaxEnt problem and show that the constraints are necessary and complementary: diversity is increased significantly while suppressing memorization.

10 ACKNOWLEDGMENTS

This project is supported by the National Research Foundation, Singapore, under grant AIVP-2024-004 (Making Foundation Models More Reliable: From Fundamental Principles to Practical Applications).

REFERENCES

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion, 2024. URL https://arxiv.org/abs/2410.08727.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025. URL https://arxiv.org/abs/2303.08797.

Jacob Bamberger, Iolo Jones, Dennis Duncan, Michael M. Bronstein, Pierre Vandergheynst, and Adam Gosztolai. Carré du champ flow matching: better quality-generalisation tradeoff in generative models, 2025. URL https://arxiv.org/abs/2510.05930.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1), November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54281-3. URL http://dx.doi.org/10.1038/s41467-024-54281-3.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1YfAfcgl.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

Benjamin Eysenbach and Sergey Levine. Maximum entropy RL (provably) solves some robust RL problems. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PtSAD3caaA2.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis, 2013. URL https://arxiv.org/abs/1310.0425.

D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer Berlin Heidelberg, 2001. ISBN 9783540411604. URL https://books.google.com.sg/books?id=eoiGTf4cmhwC.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

A. D. Ioffe. On lower semicontinuity of integral functionals. i. *SIAM Journal on Control and Optimization*, 15(4):521–538, 1977. doi: 10.1137/0315035. URL https://doi.org/10.1137/0315035.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957. doi: 10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pp. 1786–1794, Red Hook, NY, USA, 2010. Curran Associates Inc.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

Kotaro Sakamoto, Ryosuke Sakamoto, Masato Tanabe, Masatomo Akagawa, Yusuke Hayashi, Manato Yaguchi, Masahiro Suzuki, and Yutaka Matsuo. The geometry of diffusion models: Tubular neighbourhoods and singularities. In Sharvaree Vadgama, Erik Bekkers, Alison Pouplin, Sekou-Oumar Kaba, Robin Walters, Hannah Lawrence, Tegan Emerson, Henry Kvinge, Jakub Tomczak, and Stephanie Jegelka (eds.), *Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM)*, volume 251 of *Proceedings of Machine Learning Research*, pp. 332–363. PMLR, 29 Jul 2024. URL https://proceedings.mlr.press/v251/sakamoto24a.html.

Riccardo De Santi, Marin Vlastelica, Ya-Ping Hsieh, Zebang Shen, Niao He, and Andreas Krause. Provable maximum entropy manifold exploration via diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DoaqUv7YQy.

Henry D. Smith, Nathaniel L. Diamant, and Brian L. Trippe. Calibrating generative models to distributional constraints, 2026. URL https://arxiv.org/abs/2510.10020.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 2256–2265. JMLR.org, 2015.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46412–46440. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/stanczuk24a.html.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, August 2011. ISSN 1615-3383. doi: 10.1007/s10208-011-9099-z. URL http://dx.doi.org/10.1007/s10208-011-9099-z.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IQxBDLmVpT.

Entao Yang, Xiaotian Zhang, Yue Shang, and Ge Zhang. High-entropy advantage in neural networks' generalizability. *CoRR*, abs/2503.13145, March 2025. URL https://doi.org/10.48550/arXiv.2503.13145.

Zeqi Ye, Qijie Zhu, Molei Tao, and Minshuo Chen. Provable separations between memorization and generalization in diffusion models, 2025. URL https://arxiv.org/abs/2511.03202.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians, 2014. URL https://arxiv.org/abs/1405.0680.

A EFFICIENT DATA-DRIVEN COMPUTATION OF $\Gamma(x)$ AND $\mathrm{dist}(X, M)$

So far, we have made use of $\Gamma(x)$ and $\mathrm{dist}(X, M)$ in formally showing anti-concentration, both intrinsically on M and in the ambient space where the model q lives. In this part, we describe an efficient way to estimate these objects; we will call the estimated objects $\widehat{\Gamma}(x)$ and $\widehat{\mathrm{dist}}(X, M)$.

A.1 LOCAL PCA

First, we fix k, the number of neighbors, and an intrinsic dimension estimate of m. The intrinsic dimension can be estimated from prior knowledge, the eigengap/spectral gap, or even through the method described in Stanczuk et al. (2024) as mentioned in Section 1. On the given dataset S_N, we build an approximate nearest neighbor (ANN) index (which is used to rapidly retrieve high-dimensional vectors close to a query point).

A.1.1 PRECOMPUTATION AT THE ANCHOR POINTS $x^{(i)}$

We precompute once, for each i:

1. Find $\mathcal{N}_k(i)$, the k-nearest neighbors of $x^{(i)}$ in S_N.
2. Choose weights $w_{ij} \geq 0$, $\sum_{j\in\mathcal{N}_k(i)} w_{ij} = 1$. For instance, we can take Gaussian weights Bamberger et al. (2025). Let $\mu_i \coloneqq \sum_{j\in\mathcal{N}_k(i)} x^{(j)}$.
3. Then, we form the weighted covariance
$$\widehat{S}_i \coloneqq \sum_{j\in\mathcal{N}_k(i)} w_{ij}(x^{(j)} - \mu_i)(x^{(j)} - \mu_i)^\top$$
4. Compute $\widehat{U}_i \in \mathbb{R}^{d\times m}$, containing the top-$m$ eigenvectors of $\widehat{S}_i$. Randomized SVD is efficient when $m \ll d$. Define the local tangent projector estimate
$$\widehat{\Gamma}_i \coloneqq \widehat{U}_i\widehat{U}_i^\top.$$

A.1.2 QUERY-TIME EVALUATION

Given a new $x \in \mathbb{R}^d$, we pick the nearest anchor $i = NN(x)$ and set $\widehat{\Gamma}(x) \coloneqq \widehat{\Gamma}_i$.

Remark A.1. If we interpolate projectors, this method becomes geometrically unsound. The average of projectors from different tangent spaces is generally not a projector onto a meaningful tangent space. Assumption 4.3 holds *locally*.

A.2 DISTANCE-TO-MANIFOLD PROXY VIA NORMAL RESIDUAL

Due to the tubular assumption of the manifold, we can decompose a point into its projection onto the manifold and curvature terms. To estimate r_0, known as the injectivity radius of the tubular neighborhood, we refer to Sakamoto et al. (2024).

Given the nearest anchor i and $\widehat{U}_i$, define

$$\widehat{\mathrm{dist}}(x, M) \coloneqq \left\|(I_m - \widehat{U}_i\widehat{U}_i^\top)(x - x^{(i)})\right\|.$$

This quantity is the distance from x to the estimated local tangent place at $x^{(i)}$, i.e. a proxy for the normal deviation from the manifold. To efficiently compute this, we compute without forming $\widehat{\Gamma}_i = \widehat{U}_i\widehat{U}_i^\top$ explicitly:

$$(I_m - \widehat{U}_i\widehat{U}_i^\top)(x - x^{(i)}) = (x - x^{(i)}) - \widehat{U}_i(\widehat{U}_i^\top(x - x^{(i)}))$$

which costs $O(dm)$.

B PROOFS

In this section, we state proofs of aforementioned lemmata/theorems.

Lemma B.1 (Equivalence of $\mathcal{I}_\Gamma$ and $\mathcal{I}$, Lemma 4.4)**.** *Under Assumption 4.3, for any smooth $\rho > 0$ with $\int \rho \, \mathrm{d}\sigma = 1$,*

$$(1-\epsilon)\mathcal{I}(\rho) \leq \mathcal{I}_\Gamma(\rho) \leq (1+\epsilon)\mathcal{I}(\rho)$$

Proof. Fix $y \in M$ and]let $\tilde{f}$ be a smooth extension off M of $f \coloneqq \log \rho$. Then $\left\|\nabla_M f(y)\right\|^2 = \langle \nabla \tilde{f}(y), P_T(y) \nabla \tilde{f}(y) \rangle$. Directly applying Assumption 4.3 implies

$$(1-\epsilon)\langle \nabla \tilde{f}(y), P_T(y)\nabla \tilde{f}(y)\rangle \leq \langle \nabla \tilde{f}(y), \Gamma(y) \nabla \tilde{f}(y)\rangle \leq (1+\epsilon)\langle \nabla \tilde{f}(y), P_T(y)\nabla \tilde{f}(y)\rangle$$

Thus $(1-\epsilon)\|\nabla_M \log \rho\|^2 \leq \langle \nabla \log \rho, \Gamma \nabla \log \rho \rangle \leq (1+\epsilon)\|\nabla_M \log \rho\|^2$. Multiplying by $\rho(y)$ and integrating over M w.r.t. σ gives the result. □

Lemma B.2 (LSI implies KL-Fisher bound, lemma 4.7)**.** *Under Assumption 4.5, for any smooth density $\rho > 0$ with $\int \rho \, \mathrm{d}\sigma = 1$,*

$$\mathrm{KL}(\rho||1) \coloneqq \int_M \rho \log \rho \, \mathrm{d}\sigma \leq \frac{\mathcal{I}(\rho)}{2\lambda}$$

Proof. Let $f \coloneqq \sqrt{\rho}$, so $\int_M f^2 \, \mathrm{d}\sigma = 1$ and $\mathrm{Ent}_\sigma(f^2) = \mathrm{KL}(\rho||1)$. Using that $\nabla_M f = 1/2 \rho^{-1/2} \nabla_M \rho$,

$$\|\nabla_M f\|^2 = \frac{\|\nabla_M \rho\|^2}{4\rho} = \|\nabla_M \log \rho\|^2 / 4.$$

Integrating, $\int_M \|\nabla_M f\|^2 \, \mathrm{d}\sigma = \mathcal{I}(\rho)/4$ and Assumption 4.5 completes the proof. □

Lemma B.3 (Projection inclusion, Lemma 5.1)**.** *Under Assumption 2.1 and recalling $q(U_{r_0}(M)) = 1$ (for simplicity), for any $r < r_0$,*

$$B_{\mathbb{R}^d}(x^{(i)}, r) \subset \pi^{-1}(B_M^{\mathrm{ext}}(y^{(i)}, L_\pi r)) \text{ for each } i$$

Consequently,

$$\begin{aligned}\mathrm{Mem}_r^{\mathbb{R}^d}(q, S_N) &\leq q_M\left(\cup_{i=1}^N B_M^{\mathrm{ext}}(y^{(i)}, L_\pi r)\right) \\ &= \mathrm{Mem}_{L_\pi r}^M(q_M, S_N).\end{aligned}$$

Proof. Fix i and let $x \in B_{\mathbb{R}^d}(x^{(i)}, r)$. Since q is supported on $U_{r_0}(M)$, we may assume $x \in U_{r_0}(M)$. By Lipschitzness of π,

$$\|\pi(x) - y^{(i)}\| = \|\pi(x) - \pi(x^{(i)})\| \leq L_\pi \|x - x^{(i)}\| < L_\pi r,$$

so $\pi(x) \in B_M^{\mathrm{ext}}(y^{(i)})$ and hence $x \in \pi^{-1}(B_M^{\mathrm{ext}} y^{(i)}, L_\pi r)$. Taking unions over i, applying $q(\cdot)$, and using $q(\pi^{-1}(A)) = q_M(A)$, we get the desired statement. □

Lemma B.4 (Projector perturbation bound, Lemma 6.3)**.** *Let S be a symmetric PSD matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Let $U \in \mathbb{R}^{d \times m}$ span the top-m eigenspace, and $P \coloneqq UU^\top$. Let $\widehat{S} = S + E$ with symmetric perturbation E, and let $\widehat{P}$ be the top-m projector of $\widehat{S}$. If the eigengap $\Delta \coloneqq \lambda_m - \lambda_{m+1} > 0$, then*

$$\|\widehat{P} - P\|_{\mathrm{op}} \leq \frac{2\|E\|_{\mathrm{op}}}{\Delta}.$$

Proof. We utilize a standard Davis-Kahan-type argument. Let $U_\perp \in \mathbb{R}^{d \times (d-m)}$ span the orthogonal complement, so $P_\perp = U_\perp U_\perp^\top = I_d - P$. Similarly, let $\widehat{U}$ span the top-m eigenspace of $\widehat{S}$, so $\widehat{P} = \widehat{U}\widehat{U}^\top$.

A standard way to measure the distance between subspaces is via the matrix of principal angles Θ. One such identity is $\left\|\widehat{P} - P\right\|_{\mathrm{op}} = \|\sin \Theta\|_{\mathrm{op}}$. So, it suffices to bound $\|\sin \Theta\|_{\mathrm{op}}$.

Now we consider the off-diagonal block of $\widehat{U}$ in the basis $(U, U_\perp)$: $U_\perp^\top \widehat{U}$, quantifying how far the estimated subspace is from the true one. A standard Davis-Kahan inequality states

$$\left\|U_\perp^\top \widehat{U}\right\|_{\text{op}} \leq \frac{\left\|U_\perp^\top E \widehat{U}\right\|_{\text{op}}}{\text{gap}}$$

where gap is the separation between the top-m eigenvalues and the rest. The separation (eigengap), in this case, becomes $\Delta = \lambda_m - \lambda_{m+1}$. Taking operator norms gives

$$\left\|U_\perp^\top \widehat{U}\right\|_{\text{op}} \leq \frac{\|E\|_{\text{op}}}{\Delta}$$

It turns out $\left\|U_\perp^\top \widehat{U}\right\|_{\text{op}} = \|\sin\Theta\|_{\text{op}}$ (the sines of the principal angles between two subspaces), and therefore this is directly comparable to the projector difference $\left\|\widehat{P} - P\right\|_{\text{op}}$. A conservative bound is $\left\|\widehat{P} - P\right\|_{\text{op}} \leq 2\|\sin\Theta\|_{\text{op}} = \frac{2\|E\|_{\text{op}}}{\Delta}$. For more details, we refer the reader to Yu et al. (2014). □

Proposition B.5 (Parameterized local PCA tangent-space rate, Proposition 6.4)**.** *We work at a fixed anchor with projected location $y \in M$, and consider a neighborhood scale h (e.g. the typical k-NN radius). Let $P_h(y)$ be the population top-m projector of S at y, with $\Delta(h)$ the corresponding eigengap. Under Assumption 6.1, suppose the* population *local covariance satisfies:*

1. *(Eigenspace separation) The top-m eigenspace equals $T_y M$ and has eigengap $\Delta(h) > 0$ to the remaining $(d - m)$ directions;*
2. *(Curvature bias) Deviation from the true tangent projector obeys $\|P_h(y) - P_T(y)\|_{\text{op}} \leq C_{\text{curv}} h$ or $C_{\text{curv}} h^2$*

Let $\widehat{\Gamma}_i$ be the empirical top-m projector from k samples in the neighborhood. Then with probability at least $1 - \delta_{\text{fail}}$,

$$\|\widehat{\Gamma}_i - P_T(y)\|_{\text{op}} \;\leq\; \underbrace{\|P_h(y) - P_T(y)\|_{\text{op}}}_{\textit{curvature bias}} + \underbrace{\frac{2\|\widehat{S} - S\|_{\text{op}}}{\Delta(h)}}_{\textit{statistical \& noise term}},$$

and moreover, for sub-Gaussian noise and bounded moments one typically has a concentration bound of the form

$$\|\widehat{S} - S\|_{\text{op}} \;\lesssim\; C_1\left(\tau^2 + h^2\right)\sqrt{\frac{\log(d/\delta_{\text{fail}})}{k}} + \; C_2\left(\tau^2 + h^2\right)\frac{\log(d/\delta_{\text{fail}})}{k},$$

where C_1, C_2 depend on the weighting scheme and moment constants.

Proof. Recall $\widehat{S} = S + E$, and $P_h(y)$ is the tangent projector in the ideal noiseless case. Then,

$$\left\|\widehat{\Gamma}_i - P_T(y)\right\|_{\text{op}} \leq \left\|\widehat{\Gamma}_i - P_h(y)\right\|_{\text{op}} + \left\|P_h(y) - P_T(y)\right\|_{\text{op}}.$$

Applying Lemma 6.3 to the first term, we get $\left\|\widehat{\Gamma}_i - P_h(y)\right\|_{\text{op}} \leq 2\|E\|_{\text{op}}/\Delta(h)$ (well-defined if the population eigengap $\Delta(h) > 0$). The stated concentration form for $\|E\|_{\text{op}}$ comes from standard matrix concentration applied to the sample covariance and under sub-Gaussian noise τ Tropp (2011); Vershynin (2018). □

Proposition B.6 (Parameterized distance proxy bound, Proposition 6.6)**.** *Let i be a nearest neighbor anchor to $x \in \mathbb{R}^d$ and assume $x^{(i)} \in U_{r_0}(M)$ with $y^{(i)} = \pi(x^{(i)})$. Assume $\left\|\widehat{\Gamma}_i - P_T(y^{(i)}\right\|_{\text{op}} \leq \epsilon$ and $\left\|y^{(i)} - y\right\| \leq c_h$ (anchor proximity). Then*

$$|\widehat{\text{dist}}(x, M) - \text{dist}(x, M)| \leq C_{\text{anchor}} c_h + \epsilon\left\|x - x^{(i)}\right\| + C_{\text{curv}}\|x - y\|^2$$

Proof. Let $y = \pi(x) \in M$, so that $\operatorname{dist}(x, M) = \|x - y\|$. Define the proxy

$$\widehat{\operatorname{dist}}(x, M) := \big\|(I - \widehat{\Gamma}_i)(x - x^{(i)})\big\|.$$

Using the triangle inequality $|\|a\| - \|b\|| \le \|a - b\|$ with

$$a = (I - \widehat{\Gamma}_i)(x - x^{(i)}), \qquad b = x - y,$$

we obtain

$$\big|\widehat{\operatorname{dist}}(x, M) - \operatorname{dist}(x, M)\big| \le \big\|(I - \widehat{\Gamma}_i)(x - x^{(i)}) - (x - y)\big\|. \tag{5}$$

Add and subtract $(I - P_T(y^{(i)}))(x - x^{(i)})$ and apply the triangle inequality:

$$\begin{aligned}\big\|(I - \widehat{\Gamma}_i)(x - x^{(i)}) - (x - y)\big\| \le{}& \underbrace{\big\|(I - \widehat{\Gamma}_i)(x - x^{(i)}) - (I - P_T(y^{(i)}))(x - x^{(i)})\big\|}_{=:R_{\mathrm{proj}}} \\ &+ \underbrace{\big\|(I - P_T(y^{(i)}))(x - x^{(i)}) - (x - y)\big\|}_{=:R_{\mathrm{geom}}}.\end{aligned} \tag{6}$$

For the projector term,

$$R_{\mathrm{proj}} = \big\|(\widehat{\Gamma}_i - P_T(y^{(i)}))(x - x^{(i)})\big\| \le \|\widehat{\Gamma}_i - P_T(y^{(i)})\|_{\mathrm{op}}\, \|x - x^{(i)}\| \le \varepsilon\, \|x - x^{(i)}\|.$$

It remains to bound R_{geom}. Write

$$x - x^{(i)} = (x - y) + (y - y^{(i)}) + (y^{(i)} - x^{(i)}),$$

so that

$$\begin{aligned}(I - P_T(y^{(i)}))(x - x^{(i)}) - (x - y) &= (I - P_T(y^{(i)}))(x - y) - (x - y) \\ &\quad + (I - P_T(y^{(i)}))(y - y^{(i)}) + (I - P_T(y^{(i)}))(y^{(i)} - x^{(i)}) \\ &= -P_T(y^{(i)})(x - y) + (I - P_T(y^{(i)}))(y - y^{(i)}) \\ &\quad + (I - P_T(y^{(i)}))(y^{(i)} - x^{(i)}).\end{aligned}$$

Hence, using $\|I - P_T(y^{(i)})\|_{\mathrm{op}} = 1$,

$$R_{\mathrm{geom}} \le \underbrace{\|P_T(y^{(i)})(x - y)\|}_{(A)} + \underbrace{\|y - y^{(i)}\|}_{(B)} + \underbrace{\|y^{(i)} - x^{(i)}\|}_{(C)}. \tag{7}$$

We bound $(B) \le c_h$ by assumption. For (C), we absorb it into the anchor term by assuming (as in the standard tube/noise model) that $\|x^{(i)} - y^{(i)}\| \le C_{\mathrm{anchor}}\, c_h$; then $(B) + (C) \le C_{\mathrm{anchor}}\, c_h$ after adjusting constants.

To control (A), note that $x - y \in N_y M$, so $P_T(y)(x - y) = 0$ and

$$P_T(y^{(i)})(x - y) = (P_T(y^{(i)}) - P_T(y))(x - y).$$

By projection regularity / bounded curvature in a tubular neighborhood, the tangent projector is Lipschitz:

$$\|P_T(y^{(i)}) - P_T(y)\|_{\mathrm{op}} \le C_{\mathrm{curv}} \|y^{(i)} - y\|.$$

Therefore,

$$(A) \le \|P_T(y^{(i)}) - P_T(y)\|_{\mathrm{op}}\, \|x - y\| \le C_{\mathrm{curv}} \|y^{(i)} - y\|\, \|x - y\|.$$

Finally, using the anchor proximity $\|y^{(i)} - y\| \le c_h$ and (in the tube regime) the standard second-order approximation of the manifold by its tangent plane, we may upper bound this term by $C_{\mathrm{curv}} \|x - y\|^2$ after adjusting constants.[1]

Combining equation 5–equation 7 yields

$$\big|\widehat{\operatorname{dist}}(x, M) - \operatorname{dist}(x, M)\big| \le C_{\mathrm{anchor}}\, c_h + \varepsilon\, \|x - x^{(i)}\| + C_{\mathrm{curv}} \|x - y\|^2,$$

as claimed. □

Remark B.7 (Estimating the LSI constant). We note that for the sphere $\mathbb{S}^2$ with uniform measure σ, the log-Sobolev constant $\lambda = 1$. This follows from the Bakry-Émery criterion since $\mathbb{S}^2$ has constant positive Ricci curvature ≥ 1.

[1] Equivalently, one can keep the slightly more general bound $C_{\mathrm{curv}} \|y^{(i)} - y\|\, \|x - y\|$; the stated $\|x - y\|^2$ form follows whenever $\|y^{(i)} - y\| \lesssim \|x - y\|$ in the tubular neighborhood.

C Further Experimental Results

Here, we continue the analysis of our experimental setting: we solve the MaxEnt-CDC problem equation 2 and see the effect each constraint has on the entropy and memorization. We train a Real NVP normalizing flow as our generative model, since the log density $\log q$ is readily available Dinh et al. (2017). In practice, we can use diffusion/flow models as well and learn a score network which will allow for more efficient computations. We were interested in a simple architecture to more clearly understand the effect(s) of the constraints and objective function. We emphasize that this work establishes a principled variational framework and proves its theoretical properties. The following experiments on the manifold $\mathbb{S}^2$ serve as controlled validation where ground truth is available. Scaling to image-space diffusion models requires amortized score estimation (e.g., score networks) and approximate nearest-neighbor structures, which we leave for future work.

To that end, we repeat our settings here: We take $M = \mathbb{S}^2$, $k = 30$, the ambient dimension $d = 3$, and $m = 2$. To solve Problem 2, we form a (relaxed) Lagrangian and perform appropriate penalization for the constraints. We take the regularizer in front of the NLL term to be 1, the regularizer for the CDC term to be 0.01 (since the CDC Fisher $\mathcal{I}_\Gamma$ has larger magnitude than entropy/NLL terms, we scale it down to balance gradients), and the regularizer for the thickness control term to be 1. Note that one can also cap or rescale eigenvalues as in anisotropic CDC constructions. The following is a summary of the simulation settings:

- Architecture: RealNVP with 8 coupling layers, hidden dimension 64, 2 hidden layers per coupling.
- Optimizer: AdamW with learning rate 1e-3.
- Training: 800 epochs, batch size 64, gradient clipping at 5.
- CDC estimation $k = 30$ neighbors, Gaussian-weighted local covariance.

The following are the definitions of the metrics used:

- NLL: negative log-likelihood, reported as $-\mathbb{E}_{x \in S_N}[\log q(x)]$.
- Coverage (denoted **Cover** in Table 1): Fraction of uniformly sampled reference points within adaptive radius $1.5(\mathrm{Vol}(M)/(nc_m))^{1/m}$.
- Memorization (denoted by **Mem**): Fraction of generated samples within Euclidean distance 0.1 of any training point.

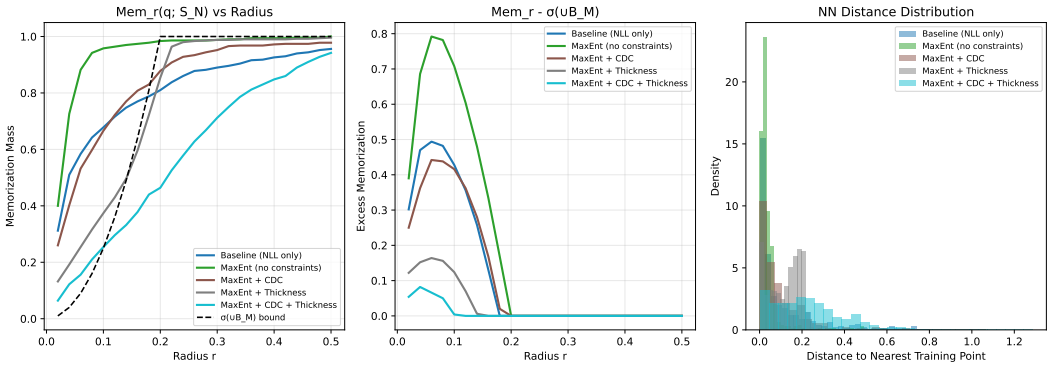

Figure 1: Comparison of Memorization across different components/constraints of the MaxEnt-CDC problem

In the first figure on the left, we see that MaxEnt + CDC + Thickness (i.e. the problem equation 2) suppresses memorization by comparing $\mathrm{Mem}_r(q, S_N)$ values we get. The third figure on the right shows the distance to the nearest training point is more distributed with this objective.

It is observed that we get increased sample diversity for MaxEnt-CDC problem 2 while respecting data manifold constraints, in Figure 2. Across the sphere, we get better more diverse coverage and do not just force learned samples to training points.

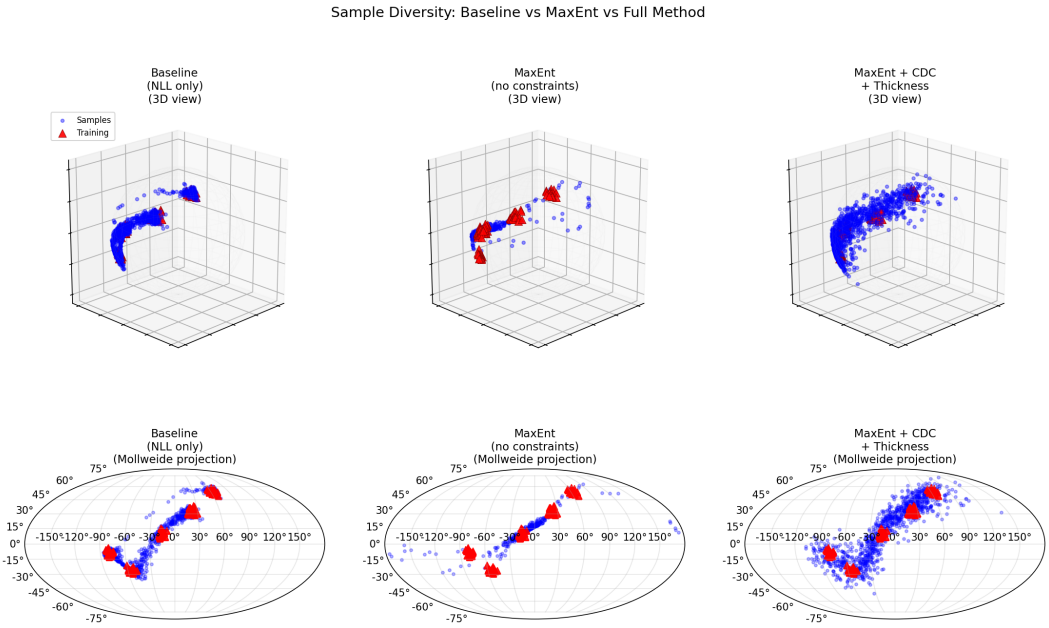

Figure 2: Comparing the samples obtained from different combinations of the objective and constraints. Red points are true data samples, and blue points are sampled from q

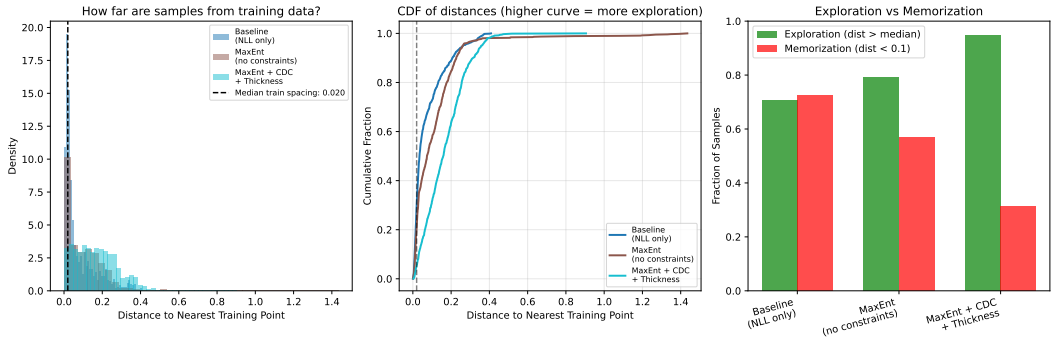

Figure 3: Sample Diversity Distances

In Figure 3, we observe the tradeoff between exploration and memorization. Additionally, we see the how sample distances are distributed.

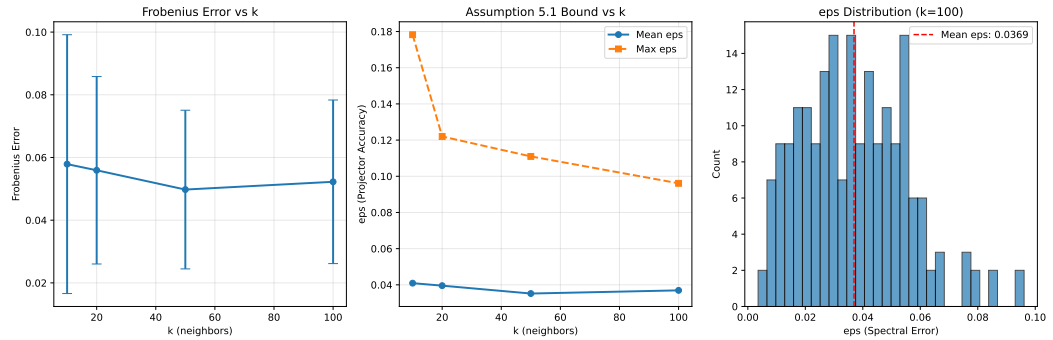

Figure 4: Accuracy and verification of Assumption 4.3

Figure 4 actually verifies Assumption 4.3! In particular, the middle plot shows how the projector accuracy error ϵ drops as the number of neighbors k increases, validating the assumption with small ϵ. That is, the data-driven local PCA algorithm outlined in Section A works sufficiently well.

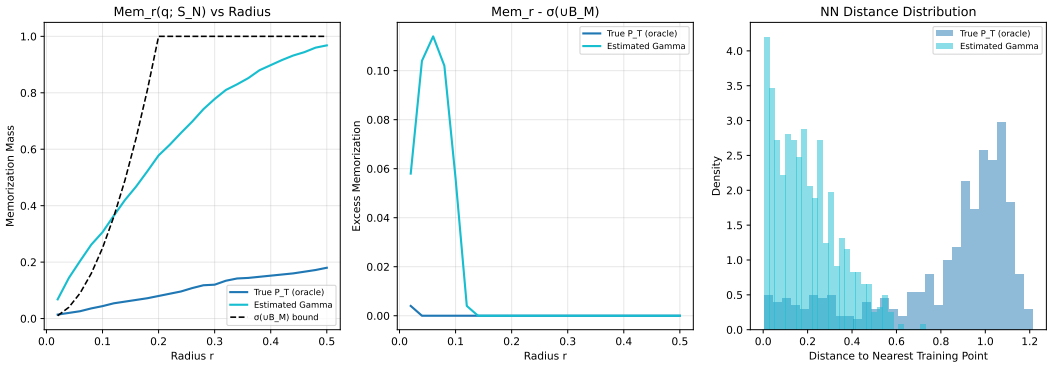

Figure 5: Memorization using true P_T on the sphere vs the estimated Γ.

Table 2: Results from solving equation 2 with true projector vs $\widehat{\Gamma}$

Proj.	$H(q)$	**NLL**	**Mem**	**Cover**	$\perp$
True P_T	12.52	-1.5507	0.4340	0.2280	0.0466
$\widehat{\Gamma}$	13.73	-1.2880	0.6220	0.1280	0.0047

Figure 5 and Table 2 shows that the true P_T outperforms the estimated tangent projector $\widehat{\Gamma}$ and shows the effect of clustered training data. The leftmost figure in Figure 5 shows that P_T has lower memorization than $\widehat{\Gamma}$, but both exceed the theoretical bound because training data is clustered. The middle shows the cost of estimation error, and the rightmost figure shows that true P_T can help spread samples farther from training points.

D Euler-Lagrange and variational formulation for MaxEnt-CDC equation 2

To analyze this problem, we derive the necessary condition in the form of an Euler-Lagrange equation (in its weak form). To this end, we introduce multipliers to form the (negative) Lagrangian functional. Let $\gamma \geq 0$ be the multiplier for the CDC constraint, $\alpha \geq 0$ be the multiplier for the NLL constraint, η be the multiplier for the normal thickness constraint, and μ be the multiplier for the normalization constraint.

That is, we convert problem 2 as follows:

$$\inf_{q \in \mathcal{P}(\Omega)} \mathcal{F}(q) \coloneqq \alpha \mathrm{NLL}(q) - H(q) + \gamma \mathcal{I}_\Gamma(q) + \eta \mathbb{E}_{X \sim q}[\underbrace{\mathrm{dist}(X, M)^2}_{=:d(x)^2}] \tag{8}$$

$$\text{s.t. } \alpha \geq 0, \lambda > 0 \tag{9}$$

$$\mathcal{P}(\Omega) \coloneqq \{q \in L^1(\Omega) : q \geq 0, \int q = 1\} \tag{10}$$

The Lagrangian, denoted by $\mathcal{F}(q)$, is given by

$$\int_{\mathbb{R}^d} q(x) \left[\log q(x) + \gamma \mathcal{I}_\Gamma(q)(x) - \mu - \eta d(x)^2\right] \mathrm{d}x - \alpha \int p^{\mathrm{data}}(x) \log q(x) \, \mathrm{d}x. \tag{11}$$

Note that we can rewrite

$$\mathcal{I}_\Gamma(q) = \int_{\mathbb{R}^d} \frac{(\nabla q(x))^\top \Gamma(x) (\nabla q(x))}{q(x)} \, \mathrm{d}x.$$

Writing the integrand in terms of q and ∇q, we have that $\mathcal{F}(q) = \int F(x, q, \nabla q)\,\mathrm{d}x$. We seek the functional derivative $\delta\mathcal{F}/\delta q = 0$, and obtain the Euler-Lagrange equation. First, we compute $\partial F/\partial q$:

$$\frac{\partial F}{\partial q} = -\alpha\frac{p^{\text{data}}}{q(x)} + \log q(x) + 1 - \gamma\frac{(\nabla q(x))^\top \Gamma(\nabla q(x))}{q(x)^2} - \mu - \eta d(x)^2$$

Next we compute $\partial F/\partial(\nabla q)$. Note that only the CDC term depends on ∇q. For a symmetric $\Gamma(x)$,

$$\frac{\partial F}{\partial(\nabla q)} = \gamma\frac{2\Gamma\nabla q(x)}{q(x)}.$$

The celebrated Euler-Lagrange (EL) equation $\frac{\partial F}{\partial q} - \nabla\cdot\left(\frac{\partial F}{\partial(\nabla q)}\right) = 0$ is

$$-\alpha\frac{p(x)}{q(x)} + \log q(x) + 1 - \gamma\mathcal{I}_\Gamma(\log q(x)) - \eta d(x)^2 - \mu - 2\gamma\nabla\cdot(\Gamma(x)\nabla\log q(x)) = 0. \quad (12)$$

Note that we have assumed that the integration domain is $\mathbb{R}^d$ and that $q(x)$ decays fast enough so that the boundary term in integration-by-parts is zero.

D.1 Reparameterization $u(x) = -\log q(x)$

Setting $q = e^{-u}$, $\log q = -u$, $\nabla\log q = -\nabla u$, we can arrive at a different form of Equation (12). This reparameterization is useful as positivity of q is automatic. The resulting EL equation (after absorbing constants) is the following nonlinear *elliptic* partial differential equation (PDE)

$$\underbrace{2\gamma\nabla\cdot(\Gamma\nabla u)}_{\text{(A)}} - \underbrace{\gamma(\nabla u)^\top\Gamma(\nabla u)}_{\text{(B)}} - \underbrace{\alpha p e^u}_{\text{(C)}} - u - \eta d^2 + \underbrace{1-\mu}_{=:D} = 0 \quad (13)$$

The EL equation reveals that the optimal density q arises from a balance between the different terms. (A) is an anisotropic diffusion operator induced by the CDC geometry that enforces geometric smoothing of the log-density. (B) is a nonlinear Hamilton-Jacobi-type term penalizing sharp score magnitudes, also weighted by geometry. (C) is a data-anchoring potential determined by the empirical data distribution and ensures fidelity to observed samples. Importantly, without this term, the PDE admits high-entropy solutions unrelated to the data manifold. The presence of u acts as a soft confining potential. Lastly, the constant D aids in enforcing the normalization constraint of q. The ηd^2 term is the contribution from the normal thickness constraint. Together, these terms enforce smoothness of the log-density in directions of low variance while allowing high variance along the data manifold.

D.2 Existence of minimizers to the EL equation

We emphasize that different parameterizations of the density q are employed for different analytical purposes. For the derivation and interpretation of the EL Equation (12), we worked with $u = -\log q$, since $\nabla u = -\nabla\log q$ aligned with the score field and facilitated a clear geometric interpretation. However, for the analysis of existence of minimizers, compactness, etc.. it is technically advantageous to work with the quadratic parameterization $q = \phi^2$. Under this change of variables, the CDC energy becomes the Dirichlet form

$$\mathcal{I}_\Gamma(q) = 4\int_\Omega(\nabla\phi(x))^\top\Gamma(x)\nabla\phi(x)\,\mathrm{d}x$$

which is coercive and weakly lower semicontinuous in $H^1(\Omega)$; this allows the application of the direct methods of the calculus of variations.

Remark D.1. The two parameterizations are equivalent on the interior of the feasible set of the problem equation 2.

With this in mind, let us consider the problem of existence of the EL PDE equation 12. In Section 4, we had assumed $\Gamma(x)$ is a symmetric PSD measure of the tangent projector that is estimated from the data. In reality, it is indeed rank-deficient. However, for the following EL analysis, for ease of exposition and a demonstration of different parts of the EL equation, we assume that we analyze a full-rank variant (either by assumption or by adding a regularization). With this, we can therefore assume that $\Gamma(x)$ is uniformly elliptic in the following. That is, $0 < \beta_1 I \le \Gamma(x) \le \beta_2 I < \infty$ a.e.x. The following theorem states the existence of the maximizer of the MaxEnt-CDC problem:

Theorem D.2 (Existence of solution). *Let $\Omega \subset \mathbb{R}^d$ be compact without boundary (e.g. $\mathbb{T}^d$. Then the (negative) MaxEnt-CDC functional Equation* (2) *admits a minimizer q^* over the set*

$$\mathcal{Q} \coloneqq \left\{ q \in L^1(\Omega) : q > 0, \int q = 1, H(q) > -\infty, \mathcal{I}_\Gamma(q) < \infty \right\}. \tag{14}$$

Moreover, $q^* > 0$ a.e. on $\{p^{\text{data}} > 0\}$

Proof. The proof proceeds in a series of steps

We now show that minimizers of $\mathcal{F}$ exist under the assumptions of the theorem. We work in the Sobolev space $H^1(\Omega)$ with the usual norm

$$\|u\|_{H^1}^2 = \|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2. \tag{15}$$

1. Coercivity of the ϕ-transform Let $(q_n)_n \subset \mathcal{P}(\Omega)$ be a minimizing sequence for $\mathcal{F}(\cdot)$. That is, $\mathcal{F}(q_n) \to \inf \mathcal{F}(q)$. W.l.o.g. assume $\sup_n \mathcal{F}(q_n) < \infty$ (otherwise, there is nothing to prove). Define $\phi_n \coloneqq \sqrt{q_n}$, so $\|\phi_n\|_{L^2}^2 = 1$. Then, from definition we have

$$\mathcal{I}_\Gamma(q_n) = 4 \int \nabla \phi_n^\top \Gamma \nabla \phi_n \geq 4\beta_1 \|\nabla \phi_n\|_{L^2}^2 .$$

Since $\gamma \mathcal{I}_\Gamma(q_n)$ is bounded above along the minimizing sequence, $(\phi_n)_n$ is bounded in $H^1(\Omega)$.

2. Compactness of minimizing sequence There exists a sequence (Banach-Alaoglu) that's not relabeled and $\phi \in H^1(\Omega)$ such that $\phi_n \rightharpoonup \phi$ weakly in $H^1(\Omega)$. By reflexivity and Rellich–Kondrachov, there exists a subsequence (still denoted ϕ_n) and $\phi \in H^1(\Omega)$ such that $\phi_n \rightharpoonup \phi$ in H^1 and $\phi_n \to \phi$ in L^2 and a.e. Recall $q \coloneqq \phi^2$ and $q_n = \phi_n^2$. Then strong L^2 convergence implies strong L^1 convergence of squares:

$$\|q_n - q\|_{L^1} = \|\phi_n^2 - \phi^2\|_{L^1} \leq \|\phi_n - \phi\|_{L^2} \, \|\phi_n + \phi\|_{L^2} \to 0,$$

since (ϕ_n) is bounded in L^2. Also $\int q = \|\phi\|_{L^2}^2 = \lim_n \|\phi_n\|_{L^2}^2 = 1$, and $q \geq 0$ a.e., so $q \in \mathcal{P}(\Omega)$.

Lemma D.3 (Lower semicontinuity of the entropy term). *Let $(\phi_n) \subset \mathcal{A}$ and $\phi \in \mathcal{A}$ be such that $\phi_n \to \phi$ in $L^2(\Omega)$. Then*

$$\int_\Omega \phi^2 \log \phi^2 \, dx \;\leq\; \liminf_{n\to\infty} \int_\Omega \phi_n^2 \log \phi_n^2 \, dx. \tag{16}$$

Proof. Let $q_n = \phi_n^2$ and $q = \phi^2$. Since $\phi_n \to \phi$ in L^2, we claim that $q_n \to q$ in $L^1(\Omega)$. Indeed,

$$\int_\Omega |q_n - q| = \int_\Omega |\phi_n^2 - \phi^2| = \int_\Omega |\phi_n - \phi| \, |\phi_n + \phi| \tag{17}$$

$$\leq \|\phi_n - \phi\|_{L^2} \, \|\phi_n + \phi\|_{L^2}. \tag{18}$$

The second factor is bounded (since (ϕ_n) is bounded in L^2 and $\phi \in L^2$), and $\|\phi_n - \phi\|_{L^2} \to 0$, so $\|q_n - q\|_{L^1} \to 0$. □

3. Lower semicontinuity of the CDC term The map $\phi \mapsto \int \nabla \phi^\top \Gamma \nabla \phi$ is convex in $\nabla \phi$ and, since Γ is bounded and symmetric PSD, the functional

$$E(\phi) \coloneqq \int_\Omega \nabla \phi^\top \Gamma \nabla \phi$$

is weakly lower semicontinuous on $H^1(\Omega)$. This follows from Ioffe (1977). Thus,

$$\liminf_{n\to\infty} \mathcal{I}_A(q_n) = 4 \liminf_{n\to\infty} E(\phi_n) \geq 4E(\phi) = \mathcal{I}_A(q).$$

4. Lower semicontinuity of the NLL term We show

$$\liminf_{n\to\infty}\left(-\int p\log q_n\right)\geq-\int p\log q.$$

Fix $\varepsilon\in(0,1)$ and define the truncated function

$$\log_\varepsilon(t):=\log(t\vee\varepsilon),$$

which is bounded and continuous on $[0,\infty)$. Since $q_n\to q$ in L^1 and Ω has finite measure, after passing to a subsequence we may assume $q_n\to q$ a.e. Then $\log_\varepsilon(q_n)\to\log_\varepsilon(q)$ a.e., and since $\log_\varepsilon$ is bounded we have $\log_\varepsilon(q_n)\to\log_\varepsilon(q)$ in L^1 and also in $L^1(p^{\mathrm{data}}\,dx)$ because $p^{\mathrm{data}}\in L^\infty$:

$$\begin{aligned}&\int p^{\mathrm{data}}\,|\log_\varepsilon(q_n)-\log_\varepsilon(q)|\,dx\\&\leq\|p^{\mathrm{data}}\|_\infty\int|\log_\varepsilon(q_n)-\log_\varepsilon(q)|\,dx\to0.\end{aligned}$$

Therefore,

$$-\int p^{\mathrm{data}}\log_\varepsilon(q_n)\,dx\to-\int p^{\mathrm{data}}\log_\varepsilon(q)\,dx.$$

Finally, note $\log_\varepsilon(t)\downarrow\log t$ pointwise as $\varepsilon\downarrow0$, so by monotone convergence (applied to $-\log_\varepsilon$which increases to $-\log$ on $(0,\infty)$),

$$\begin{aligned}-\int p^{\mathrm{data}}\log(q)\,dx&=\lim_{\varepsilon\downarrow0}\Big(-\int p^{\mathrm{data}}\log_\varepsilon(q)\,dx\Big),\\-\int p^{\mathrm{data}}\log(q_n)\,dx&=\lim_{\varepsilon\downarrow0}\Big(-\int p\log_\varepsilon(q_n)\,dx\Big).\end{aligned}$$

Taking $\liminf_{n\to\infty}$ and using the convergence for fixed ε yields

$$\liminf_{n\to\infty}\Big(-\int p^{\mathrm{data}}\log(q_n)\,dx\Big)\geq-\int p^{\mathrm{data}}\log(q)\,dx,$$

as desired.

5. Lower semicontinuity of the normal thickness term Again, let $d(x):=\mathrm{dist}(x,M)$. d is continuous under the conditions of the manifold M. Consider

$$\Phi(q):=\mathbb{E}_q[d(X)^2]$$

Since M is compact, $d(x)^2$ is continuous and bounded, so $\int d^2\,\mathrm{d}q_n\to\int d^2\,\mathrm{d}q$, showing that Φ is continuous and thus lower semi-continuous. Likewise, it is easy to see $d(x)^2$ also satisfies coercivity.

Now we use the above results to conclude a minimizer exists.

The functional $\mathcal{F}$ admits a minimizer on $\mathcal{Q}$, and $q>0$ a.e. on $\{p^{\mathrm{data}}>0\}$

Combining Steps 3-5 gives $\mathcal{F}(q)\leq\liminf_{n\to\infty}\mathcal{F}(q_n)=\inf\mathcal{F}(q)$, so q^* exists. Additionally, if $q=0$ on some measurable set E with $\int_E p^{\mathrm{data}}>0$, then $\log q=-\infty$ on E and $-\int p\log q=\infty$, contradicting finiteness of $\mathcal{F}(q)$. Thus $q>0$ a.e. on $\{p^{\mathrm{data}}>0\}$, and we have shown the required.

With this, we conclude the proof of Theorem D.2. □

Under additional smoothness assumptions on $\Gamma(x)$ and standard boundary conditions, elliptic regularity implies that weak solutions of equation 12 are smooth in the interior of Ω; see, e.g., Gilbarg & Trudinger (2001) for details.

Finally, we note that it is possible to prove the existence of a minimizer if the domain $\Omega=\mathbb{R}^d$. If that is the case, then we need to assume that finite second moments of the density q and p exist.