SPHERE-JEPA: Spherical Prediction with Homogeneous Embeddings

Léo Nicollier
Université Paris-Saclay, CNRS,
ENS Paris-Saclay, Centre Borelli,
Advanced Track and Trace
leo.nicollier@gmail.com

Max Dunitz
Advanced Track and Trace

Marc Pic
Advanced Track and Trace

Pablo Musé
Université Paris-Saclay, CNRS,
ENS Paris-Saclay, Centre Borelli

Enric Meinhardt-Llopis
Université Paris-Saclay, CNRS,
ENS Paris-Saclay, Centre Borelli

Gabriele Facciolo
Université Paris-Saclay, CNRS,
ENS Paris-Saclay, Centre Borelli,
Institut Universitaire de France

Abstract

A fundamental open question in self-supervised learning (SSL) is the explicit characterization of the optimal geometry of the learned representations. Recently, LeJEPA identified isotropic Gaussian embeddings as optimal for minimizing downstream prediction risk in Euclidean spaces. However, the corresponding problem for distributions supported on lower-dimensional manifolds, such as the hypersphere, remains unexplored. In this work, we demonstrate that extending this minimax analysis to smooth distributions on Riemannian manifolds fundamentally changes the optimal solution. We show that, under a worst-case formulation, both k-nearest neighbors and kernel ridge regression induce hyperspherical uniformity. More precisely, we show that uniform distributions on manifolds are optimal for k-nearest neighbors, and that the uniform distribution on the sphere is optimal for kernel ridge regression with both the exponential dot-product kernel and the linear kernel. This theoretical insight reveals a fundamental limitation of Gaussian embeddings: their non-uniform density induces anisotropic k-NN neighborhoods, severely biasing the estimator. To correct this, we introduce **SPHERE-JEPA**, a theoretically grounded SSL framework. We adapt LeJEPA's Cramér–Wold projection mechanism to enforce hyperspherical uniformity rather than a Gaussian prior. Empirically, SPHERE-JEPA achieves competitive results on standard benchmarks, including a $+1.8$ percentage-point linear-probing gain on ImageNet-1K (ViT-B/14), and improves texture-retrieval mAP by over six percentage points in a controlled nonparametric evaluation (GitHub repo).

1 Introduction

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning visual representations without manual annotations [Chen et al., 2020]. By leveraging invariances across multiple augmented views of the same data, modern SSL methods such as SimCLR [Chen et al., 2020], BYOL [Grill

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

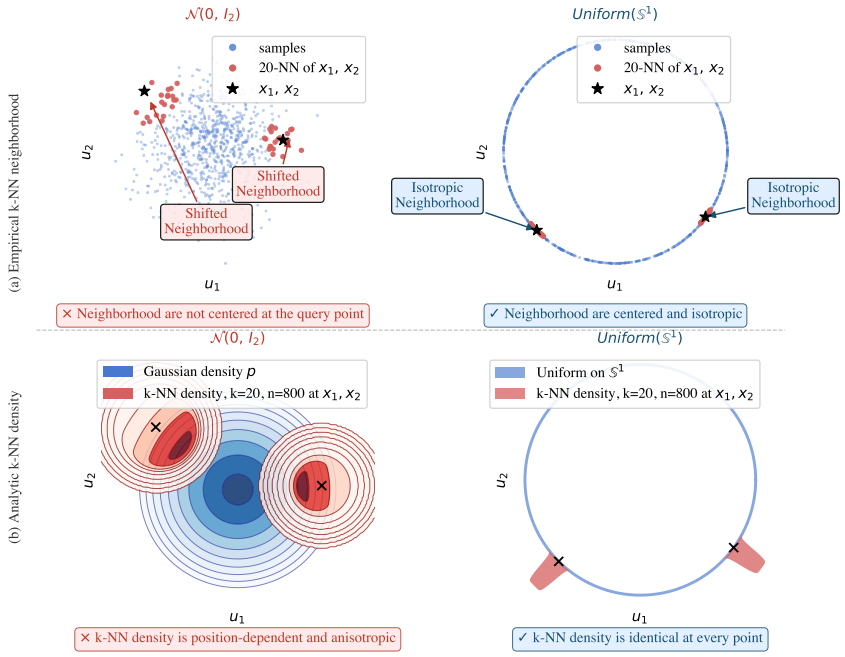


Figure 1: **k-NN neighborhoods are density-biased. (a)** Empirical k-NN neighborhoods ($k = 20$) at two query points x_1, x_2: under a Gaussian distribution on $\mathbb{R}^2$, neighborhoods are not centered at the query point but are skewed toward regions of higher density, resulting in anisotropic and directionally biased neighborhoods. In contrast, for a uniform distribution on $\mathbb{S}^1$ the neighborhoods are centered and isotropic everywhere. **(b)** Analytic k-NN density: under a Gaussian distribution, the induced densities are position-dependent and anisotropic, whereas the uniform distribution on $\mathcal{S}^1$ yields identical densities at every point, reflecting its intrinsic geometric symmetry.

et al., 2020], and DINO [Caron et al., 2021b] have achieved performance comparable to supervised learning across a wide range of tasks.

Until recently, most SSL methods relied primarily on heuristic mechanisms to prevent representation collapse and shape the geometry of learned embeddings, without explicitly characterizing what constitutes an optimal representation space. With LeJEPA, Balestriero and LeCun [2025] provided a first step towards such a characterization by showing that an isotropic Gaussian distribution is optimal for linear ridge regression and k-nearest neighbor regression in a worst-case formulation. These two evaluation protocols are commonly used to assess the quality of SSL representation, as they probe, respectively, the linear separability of features and the local geometric structure of the embedding space. However, this analysis is restricted to distributions with densities in $\mathbb{R}^d$, and therefore excludes a broad class of natural representations supported on lower-dimensional manifolds, such as the hypersphere.

In this work, we extend this analysis to smooth distributions supported on Riemannian manifolds. Under the same worst-case formulation as LeJEPA, we derive geometric optimality principles for both local and kernel-based learning methods.

First, we show that minimizing the worst-case bias of k-nearest neighbors implies that the representation distribution must be uniform on the underlying manifold. Second, we prove that minimizing the worst-case bias of kernel ridge regression with both the linear kernel and the exponential dot-product kernel $K(x, y) = e^{\kappa x^\top y}$ forces the representation distribution to be uniform on the hypersphere.

This result contrasts with Gaussian representations, whose non-uniform density induces anisotropic local neighborhoods, thereby distorting the local geometry and degrading the performance of non-parametric methods such as k-NN (see Figure 1). Building on this theoretical insight, we introduce **SPHERE-JEPA**, a self-supervised learning framework that enforces hyperspherical uniformity. Our approach follows the same design as LeJEPA and relies on the same projection-based regularization mechanism inspired by the Cramér–Wold theorem [Cuesta-Albertos et al., 2007], which enforces a

target distribution through random one-dimensional projections. We repurpose this projection-based mechanism to enforce a *uniform distribution on the hypersphere* rather than a Gaussian distribution in $\mathbb{R}^d$, leading to **Sketched Uniform Spherical Regularization (SUSReg)**.

We evaluate our method on ImageNet-100 [Deng et al., 2009] using a ResNet-18, and on ImageNet-1K using a ViT-B/14 [Oquab et al., 2023] backbone. Our results show broadly comparable performance on standard benchmarks, with a larger gain on a controlled texture-retrieval task designed to probe k-NN behavior in the learned representation space. Notably, on ImageNet-1K, it yields a $+1.8\%$ accuracy gain in linear probing and improves average transfer performance across diverse downstream datasets. Furthermore, in a controlled texture retrieval task designed to probe k-NN behavior, our approach yields significant gains, improving mean average precision by over 6% in the learned representation space, without incurring any additional computational cost.

Contributions. Our key contributions are as follows: *(i)* We extend the theoretical analysis of optimal SSL representations from Euclidean densities to distributions supported on smooth manifolds; *(ii)* We prove that the uniform distribution on the hypersphere is optimal under a worst-case formulation that combines radial kernel ridge regression and k-NN; *(iii)* We introduce SUSReg, a projection-based regularizer that promotes hyperspherical uniformity in learned representations; *(iv)* We report gains on ViT-B/14 linear probing and controlled nonparametric retrieval.

Organization. The remainder of the paper is structured as follows. Section 2 introduces the notation and setup, followed by related work in Section 3. Section 4 develops the theoretical analysis of optimal representation geometries for k-nearest neighbors and kernel ridge regression. In Section 5, we introduce SPHERE-JEPA and the SUSReg regularization. Experiments are reported in Section 6. We conclude in Section 7.

2 Notation and Setup

We adopt the same notation as in Balestriero and LeCun [2025]. We consider a dataset composed of N independent samples. Each sample is observed through V views, yielding data points $x_{n,v} \in \mathbb{R}^D$, $n = 1, \dots, N$, $v = 1, \dots, V$, where D denotes the input dimension (e.g., for an image of spatial resolution $H \times W$ with C channels, $D = C \times H \times W$). The views typically correspond to data augmentations of a given image (e.g., image crops or geometric transformations).

Following standard SSL practice [Caron et al., 2020, 2021b], we distinguish between *local* and *global* views [Caron et al., 2021a]. Both are obtained as crops from the input image: local views correspond to smaller crops that capture limited spatial context, while global views are larger crops that preserve most of the image content. We denote by V_g the number of global views, by V_l the number of local views, and by V_a the total number of views. We index all views $v = 1, \dots, V_a$, whith the first V_g indices corresponding to global views. We assume that the samples $\{x_n\}_{n=1}^N$ are independent and identically distributed.

Encoder and projection head. Let $g_\phi : \mathbb{R}^D \to \mathbb{R}^h$ denote the backbone encoder and $h_\psi : \mathbb{R}^h \to \mathbb{R}^d$ the MLP projection head. We write $f_\theta = h_\psi \circ g_\phi$, with $\theta = (\phi, \psi)$. The backbone architecture may be convolutional or based on self-attention. For each sample and view, the projected embedding is $z_{n,v} \coloneqq f_\theta(x_{n,v}) \in \mathbb{R}^d$. We denote its normalized version by

$$\tilde{z}_{n,v} \coloneqq \frac{z_{n,v}}{\|z_{n,v}\|} \in \mathbb{S}^{d-1}. \tag{1}$$

3 Related Work

3.1 Self-Supervised Learning

Self-supervised learning (SSL) [Chen et al., 2020] aims to learn transferable representations without manual annotations by leveraging the intrinsic structure and invariances of the data. This is typically achieved by enforcing consistency between representations of different augmented views of the same sample, while separating those of distinct samples. A fundamental challenge shared by most SSL objectives [Grill et al., 2020] is to avoid representation collapse, in which all inputs are mapped to a constant embedding. Existing methods address this issue through a variety of architectural and optimization strategies.

Contrastive methods, such as SimCLR [Chen et al., 2020] and MoCo [He et al., 2020], rely on negative samples to enforce discriminative representations. Momentum-based frameworks, such as MoCo or DINO [Caron et al., 2021b], further stabilize the training with an exponential moving average (EMA) teacher. Non-contrastive approaches, including BYOL [Grill et al., 2020] and SimSiam [Chen and He, 2021], eliminate the need for negative samples but introduce explicit asymmetries during training, such as stop-gradient operations or additional predictor networks. Decorrelation-based approaches, such as Barlow Twins [Zbontar et al., 2021] and VICReg [Bardes et al., 2021], prevent collapse by explicitly penalizing correlations across embedding dimensions. Beyond the choice of the objective, DINO shows that architectural design also plays a critical role in training stability. In particular, the use of register tokens in Vision Transformers helps maintain feature diversity and mitigates collapse by allocating dedicated capacity for global information [Oquab et al., 2023, Siméoni et al., 2025].

Although these strategies have demonstrated strong empirical performance, they rely primarily on heuristic mechanisms to prevent collapse. In most cases, the structure of the optimal representation is implicitly defined by training dynamics or data augmentations rather than explicitly characterized. As a result, many SSL objectives enforce invariances that are sufficient for good performance, but whose optimality properties remain only partially understood.

3.2 JEPA and LeJEPA

Joint Embedding Predictive Architectures (JEPAs), introduced by LeCun et al. [2022], proposes a representation learning framework based on predictability in the latent space. Rather than relying on pixel-level matching or contrastive objectives, a JEPA learns representations by predicting the latent representation of one view (or future state) from another, thereby emphasizing the structure and dynamics of the representation itself. In this sense, JEPAs can be viewed as a generalization of earlier non-contrastive methods such as BYOL [Grill et al., 2020] and SimSiam [Chen and He, 2021], which also rely on direct prediction in a latent space.

LeJEPA [Balestriero and LeCun, 2025] instantiates the JEPA framework by retaining its core latent prediction objective. Like BYOL or SimSiam, it learns representations by predicting the embedding of one view from another in the latent space. Overall, LeJEPA is composed of two complementary terms: *(i)* a predictive loss, similar to BYOL/SimSiam, implemented as a mean squared error between representations, which promotes alignment across views; *(ii)* a distributional regularization term that constrains the geometry of the representation space.

Specifically, this regularization encourages the learned embeddings to match a prescribed isotropic Gaussian distribution. This constraint is implemented via random one-dimensional projections of the embeddings, leveraging the Cramér–Wold theorem [Cuesta-Albertos et al., 2007]. For each view $v \in \{1, \dots, V_a\}$, consider a mini-batch of embeddings $\{z_{n,v}\}_{n=1}^{N} \subseteq \mathbb{R}^d$. At each training step, LeJEPA draws a finite set of directions $\mathcal{A} = \{a_1, \dots, a_{|\mathcal{A}|}\}$ with $a \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and forms the sliced variables $t^a_{n,v} \coloneqq a^\top z_{n,v} \in \mathbb{R}$. It then enforces, for each view v and each direction $a \in \mathcal{A}$ (with $\mathcal{A}$ resampled at every mini-batch), that the empirical distribution of $\{t^a_{n,v}\}_{n=1}^{N}$ matches a standard normal $\mathcal{N}(0,1)$ via the Epps–Pulley test. By the Cramér–Wold theorem, matching the one-dimensional projections to $\mathcal{N}(0,1)$ across uniform directions enforces an isotropic Gaussian embedding distribution in $\mathbb{R}^d$.

3.3 Epps–Pulley test

LeJEPA uses the test by Epps and Pulley [1983] (EP) as a univariate statistical discrepancy based on characteristic functions. Given a scalar random variable X with samples $\{x_j\}_{j=1}^{N}$ and a reference distribution Y with characteristic function $\varphi_Y(t) = \mathbb{E}[e^{itY}]$, the empirical characteristic function of X is $\hat{\varphi}_X(t) = \frac{1}{N}\sum_{j=1}^{N} e^{itx_j}$. The EP discrepancy is defined as a weighted L^2 distance between characteristic functions:

$$\mathrm{EP}(X, Y) = N \int_{\mathbb{R}} \left|\hat{\varphi}_X(t) - \varphi_Y(t)\right|^2 w(t)\, dt, \tag{2}$$

where $w(t)$ is a Gaussian weight function. In LeJEPA, this test is applied to the projected scalars $x_j = t^a_{j,v} = a^\top z_{j,v}$, with $Y \sim \mathcal{N}(0,1)$, and averaged over a finite set of directions $a \in \mathcal{A}$.

3.4 Distributional Regularization of Latent Representations

In addition to LeJEPA, other recent self-supervised learning methods incorporate explicit regularization of the latent space to prevent collapse and shape the geometry of the learned representations [Sablayrolles et al., 2018, Bardes et al., 2021, Caron et al., 2021b]. These regularizers can be grouped according to the target distribution they enforce on the embeddings.

A first group of methods targets an isotropic Gaussian distribution in $\mathbb{R}^d$. VICReg [Bardes et al., 2021] enforces this only through low-order statistics: it maintains non-zero variance in each coordinate and decorrelates coordinates across the batch, promoting isotropy at the covariance level. LeJEPA enforces a stronger, distribution-level constraint by matching the embedding distribution to an isotropic Gaussian via random one-dimensional projections and the Epps–Pulley test.

Normalized spherical embeddings are also used in contrastive methods such as SimCLR [Chen et al., 2020] and MoCo [He et al., 2020]; spherical representations are therefore not specific to SPHERE-JEPA. Normalization alone, however, does not enforce a uniform distribution on the sphere. KoLeo [Sablayrolles et al., 2018], notably employed in DINO, encourages more uniform coverage through a mini-batch estimate of differential entropy. Its nearest-neighbor and logarithmic terms can be sensitive to nearly coincident embeddings, motivating numerical safeguards in implementations.

The choice of a regularizer poses a fundamental question regarding downstream performance: Do different distributional constraints favor specific estimators, such as linear models, kernel methods, or k-nearest neighbors?

4 Optimality Criteria for Self-Supervised Representations

A central challenge in self-supervised learning (SSL) is to learn representations that transfer efficiently to a broad range of downstream tasks with minimal supervision [Chen et al., 2020]. Since downstream tasks are unknown at training time, optimizing representations for a single task may lead to poor transferability. Following LeJEPA, we adopt a *minimax* perspective: we seek a representation that minimizes the worst-case prediction error over a class of downstream targets.

Representation quality is commonly assessed through two classical downstream protocols: linear ridge regression [Schölkopf and Smola, 2002] and k-nearest neighbors (k-NN) [Hastie et al., 2009].

Linear ridge regression. This criterion probes the *global structure* of the representation space. Given a dataset of representations and targets $\{(z_i, y_i)\}_{i=1}^N$, it estimates a linear predictor $g_w(z) = \langle w, z\rangle$ by minimizing the regularized empirical risk

$$\widehat{w}_\lambda = \arg\min_{w\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N}(y_i - \langle w, z_i\rangle)^2 + \lambda\|w\|^2, \qquad \lambda > 0.$$

k-nearest neighbors. Conversely, k-NN is sensitive to the *local geometry*, as its prediction bias depends heavily on the neighborhood shape and the local sampling density of the data. Together, they capture complementary aspects of representation quality.

Minimax formulation. Let p denote the marginal distribution of the representations $z = f_\theta(x) \in \mathbb{R}^d$. Given a downstream estimator $\widehat{g}$ trained on samples drawn from p, we measure its performance through the integrated squared bias (ISB)

$$\mathrm{ISB}(g, p) = \int \big(\mathbb{E}[\widehat{g}(z)] - g(z)\big)^2 \, p(z)\,\mathrm{dvol}(z).$$

Given a class $\mathcal{G}$ of downstream targets, we consider the minimax problem

$$p^\star = \arg\min_{p} \sup_{g\in\mathcal{G}} \mathrm{ISB}(g, p).$$

Different downstream protocols induce different geometric constraints on the optimal representation distribution.

Proposition 4.1 (Linear ridge regression, [Balestriero and LeCun, 2025])**.** *Consider the linear ridge regression with a downstream target class*

$$\mathcal{G} = \{g_w(x) = \langle w, x\rangle \ : \ \|w\| \leq 1\}\,.$$

Under the constraint $\mathbb{E}[\|X\|^2] = 1$, minimizing the worst-case ISB is equivalent to minimizing the top eigenvalue of the covariance matrix of p. Consequently, any optimal distribution is isotropic.

Proposition 4.2 (k-NN on compact and Euclidean supports)**.** *Let $c > 0$, $k \to \infty$, and $k/n \to 0$. Let $M_n \to \infty$ satisfy $M_n(k/n)^{1/m} \to 0$, and set $|f|_{C^3} = \max_{1\leq j\leq 3} \|\nabla^j f\|_\infty$. The ISB below refers to the population-radius estimator. If $\mathcal{M} \subset \mathbb{R}^d$ is a compact, connected, m-dimensional Riemannian submanifold without boundary, the uniform density on $\mathcal{M}$ is the unique pointwise asymptotic minimizer of the worst-case ISB over*

$$\mathcal{G}_n^{\mathcal{M}} = \left\{f \in C^3(\mathcal{M}) : \|\Delta_{\mathcal{M}} f\|_2 \leq c,\ |f|_{C^3} \leq M_n\right\}.$$

If $\mathcal{M} = \mathbb{R}^m$, then for every positive smooth probability density p, the worst-case leading-order ISB functional over the class obtained by setting $\mathcal{M} = \mathbb{R}^m$ in the definition above diverges as $n \to \infty$. In particular, this holds for Gaussian densities.

The previous two criteria are not sufficient to fully constrain the optimal geometry of the representation distribution. We therefore introduce a third minimax criterion based on kernel ridge regression [Schölkopf and Smola, 2002] (KRR) with the exponential dot-product kernel $K(x, y) = e^{\kappa x^\top y}$ with $\kappa > 0$.

Let T_p denote the associated population covariance operator (see Appendix F) and consider the source class

$$\mathcal{G}_p = \left\{g = T_p f \ : \ \|f\|_{L^2(p)} \leq 1\right\}.$$

Proposition 4.3 (Exponential KRR)**.** *For kernel ridge regression with kernel $K(x, y) = e^{\kappa x^\top y}$, the worst-case ISB over $\mathcal{G}_p$ is minimized by $p^\star = \mathrm{Unif}(\mathbb{S}^{d-1})$.*

Across the admissible distributions on Euclidean space and compact manifolds, the three minimax criteria single out the uniform distribution on the hypersphere as their unique common minimizer:

Theorem 4.4 (Optimal representation geometry)**.** *The uniform distribution on the hypersphere $\mathrm{Unif}(\mathbb{S}^{d-1})$ simultaneously minimizes the worst-case ISB for linear ridge regression, k-NN regression, and exponential-kernel ridge regression.*

Our analysis considers smooth densities on Euclidean space and on compact manifolds, including spheres and tori. For k-NN, the worst-case leading functional is unbounded for positive smooth densities on $\mathbb{R}^d$, while the uniform density minimizes it on each fixed compact manifold. The uniform distribution on the sphere also minimizes the linear-ridge and exponential-KRR criteria. It is therefore a common minimizer of the three criteria among the distributions studied here.

Our optimality guarantee concerns the three downstream estimators and target classes analyzed above. For representations on the sphere, the KRR argument also shows that the uniform distribution is a minimizer for any positive-definite rotationally invariant kernel; uniqueness need not hold for every kernel. This includes an MLP probe only when its training is described by regularized kernel regression with a fixed, rotationally invariant neural tangent kernel. The result does not cover finite-width MLP feature learning or end-to-end fine-tuning, where the representation geometry can change during training.

5 SPHERE-JEPA: Spherical Prediction with Homogeneous Embeddings

5.1 Sketched Uniform Spherical Regularization (SUSReg)

We introduce *Sketched Uniform Spherical Regularization* (SUSReg), a projection-based regularizer inspired by SIGReg [Balestriero and LeCun, 2025]. Let $\{\tilde{z}_{n,v}\}_{n=1}^N \subseteq \mathbb{S}^{d-1}$ be a normalized minibatch for view v, and let X_v (X when the context is clear) be uniform over this set. By the Cramér-Wold theorem, $X \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ if and only if every projection $a^\top X$, $a \in \mathbb{S}^{d-1}$, follows the density

$$\rho_d(t) = C_d\,(1 - t^2)^{\frac{d-3}{2}}, \qquad t \in (-1, 1), \tag{3}$$

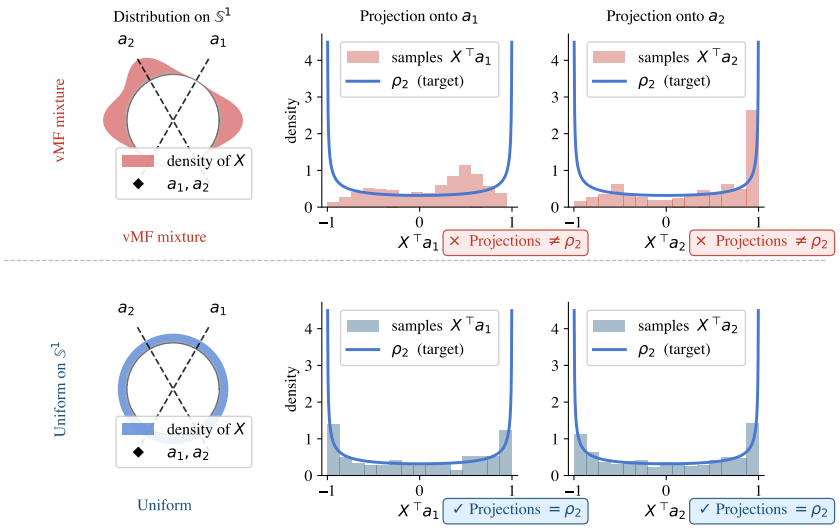


Figure 2: Illustration of the Cramér–Wold characterization underlying SUSReg on $\mathbb{S}^1$. **Top row:** a mixture of von Mises–Fisher (vMF) distributions—the spherical analogue of Gaussian distributions—induces a non-uniform distribution with mass concentrated around a few directions. The resulting projections $X^\top a_1$ and $X^\top a_2$ deviate from the target density ρ_2. **Bottom row:** a uniform distribution on $\mathbb{S}^1$ yields projections that match ρ_2 across directions. Histograms compare the distributions of $a^\top X$ to the target density ρ_2 (solid curve). Non-uniform representations (top) produce mismatched projections and are penalized by SUSReg, whereas uniform representations (bottom) satisfy the projection constraint and are therefore not penalized by SUSReg.

where C_d is a normalization constant that ensures that ρ_d integrates into one. For large d, ρ_d converges to the density of $\mathcal{N}(0, \frac{1}{d})$ by the central limit theorem [Billingsley, 1995]. We use this Gaussian approximation when $d > 256$. This characterization is illustrated in Figures 2–3, which provide an empirical visualization of the Cramér–Wold theorem: a distribution is uniform on the sphere if and only if its one-dimensional projections match ρ_d along all directions. Accordingly, only uniformly distributed representations yield projection distributions consistent with ρ_d, while non-uniform distributions lead to systematic, direction-dependent mismatches.

Following SIGReg, SUSReg enforces this condition by penalizing, via the Epps–Pulley test, the mismatch between $a^\top X$ and $Y \sim \rho_d$ over randomly sampled directions:

$$\mathcal{L}_{\text{SUSReg}} \;:=\; \frac{1}{V_a} \sum_{v=1}^{V_a} \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathrm{EP}\big(a^\top X,\; Y\big)\,. \tag{4}$$

Minimizing this loss enforces that the projected variables $a^\top X$ match the target distribution ρ_d across random directions, thereby enforcing uniformity on the hypersphere and aligning the representation geometry with the optimal structure characterized in Section 4.

5.2 Invariant Prediction Loss on the Hypersphere

While SUSReg enforces global uniformity on the hypersphere, it does not guarantee invariance to semantic-preserving transformations. To address this, we introduce an alignment objective. Unlike LeJEPA, which applies a mean squared error (MSE) to unnormalized embeddings in $\mathbb{R}^d$, we adapt this principle to normalized representations.

For a sample with multiple views, we define a prototype μ_n as the average of its global views and align all views to it via the invariance loss $\mathcal{L}_{\text{inv}}$:

$$\mu_n \;:=\; \frac{1}{V_g} \sum_{v=1}^{V_g} \tilde{z}_{n,v}, \qquad \mathcal{L}_{\text{inv}} \;:=\; \frac{1}{V_a} \sum_{v=1}^{V_a} \|\mu_n - \tilde{z}_{n,v}\|_2^2\,. \tag{5}$$

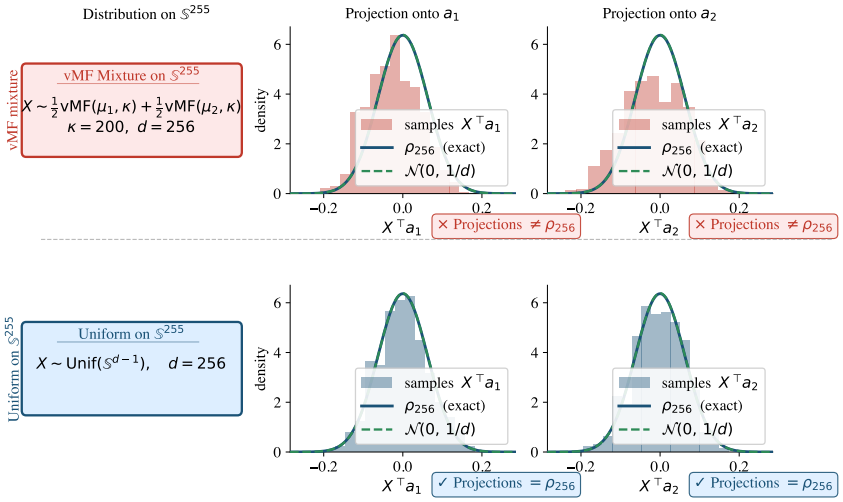


Figure 3: Illustration of the Cramér–Wold characterization underlying SUSReg on $\mathbb{S}^{255}$ ($d = 256$). **Top row:** a mixture of von Mises–Fisher (vMF) distributions induces a non-uniform distribution with mass concentrated around a few directions. The resulting projections $X^\top a_1$ and $X^\top a_2$ induce distributions that deviate from the target density ρ_{256}. **Bottom row:** a uniform distribution on $\mathbb{S}^{255}$ yields projections consistent with ρ_{256} across directions. Histograms compare the distributions of $a^\top X$ to ρ_{256} (solid curve). Non-uniform representations (top) produce mismatched projections and are penalized by SUSReg, whereas uniform representations (bottom) satisfy the projection constraint and are therefore not penalized by SUSReg. The Gaussian approximation $\mathcal{N}(0, 1/d)$ (dashed) closely matches ρ_{256}, justifying its use in high dimensions.

This MSE on normalized embeddings approximates the squared geodesic distance locally, offering more stable and bounded gradients.

5.3 SPHERE-JEPA Objective

Following LeJEPA, our final objective combines the invariance loss (for cross-view alignment) and SUSReg (for global hyperspherical geometry) via a balancing weight $\lambda \in [0, 1]$:

$$\mathcal{L}_{\text{SPHERE-JEPA}} \;=\; (1-\lambda)\,\mathcal{L}_{\text{inv}} \;+\; \lambda\,\mathcal{L}_{\text{SUSReg}}.$$

The choice of a spherical rather than Gaussian target raises a natural question: why not simply normalize LeJEPA's Gaussian embeddings? Indeed, if $Z \sim \mathcal{N}(0, I_d)$, then $Z/\|Z\| \sim \text{Unif}(\mathbb{S}^{d-1})$. First, both methods impose their distributional constraint at the projection head, while standard downstream benchmarks evaluate the backbone without that head. Normalizing LeJEPA's head output after training, therefore, does not imply equivalent backbone features. Second, even if the normalized head outputs were evaluated, the training objectives would differ. An exact Gaussian vector can be written as $Z = RU$, where U is uniform on the sphere and independent of the radius R. With finite batches and projections, Gaussian matching also penalizes radial mismatches, although R is discarded by normalization. SUSReg directly targets the retained directions.

6 Experiments

Architectures. To evaluate across different inductive biases architectures, we use both self-attention and convolutional backbones: a primary Vision Transformer (ViT-B/14) [Oquab et al., 2023], and standard ResNet-18/50 models [He et al., 2016]. Representations are extracted from the ViT class token and via global average pooling for the ResNets.

Embedding Head. The backbone output $g_\phi(x)$ is processed by a 3-layer MLP projection head h_ψ with GELU activations and hidden dimensions $[2048, 2048, 256]$. In both methods, distributional regularization acts on the projected representation $h_\psi(g_\phi(x))$, whereas downstream evaluation uses the backbone representation $g_\phi(x)$ without the projection head.

Training Data and Views. We evaluate on ImageNet-1K [Deng et al., 2009], ImageNet-100, and Galaxy10 [Leung and Bovy, 2018]. For ViT-B/14 on ImageNet-1K, we use a multi-crop strategy ($V_g = 2$ global, $V_l = 6$ local; $V_a = 8$ total views). ResNets use only two global views ($V_a = V_g = 2$). We apply standard augmentations: DINO [Caron et al., 2021b] and SimSiam [Chen and He, 2021] for ViT, and BYOL [Grill et al., 2020] for ResNets.

Optimization. Models are trained with AdamW and cosine annealing. On ImageNet-1K, ViT-B/14 is trained for 100 epochs (5-epoch warmup) with batch size 512, learning rate (LR) 5×10^{-4}, and weight decay (WD) 5×10^{-2}. ResNets are trained for 200 epochs (1-epoch warmup) with LR 5×10^{-2} and WD 5×10^{-4}. $|\mathcal{A}|$ for SUSReg and SIGReg is fixed to 1024 across all runs. We set the balancing weight to $\lambda = 0.05$ for both methods, following the LeJEPA configuration; we did not tune this value specifically for SPHERE-JEPA. Its sensitivity is reported in Appendix C.

EMA Teacher. When using an EMA teacher, the target prototype μ_n (Eq. 5) is generated by a momentum encoder and projection head. While the online network updates via gradient descent, its weights are used to update the EMA teacher with a momentum parameter $\tau = 0.996$. SUSReg constrains the marginal distribution of projected representations across samples, but does not control how the prediction target for an individual sample changes during training. The EMA teacher provides more slowly changing targets by averaging encoder weights over time. This offers a possible explanation for its optimization benefit, distinct from the geometric constraint imposed by SUSReg.

6.1 Main Results on ImageNet Benchmarks

ImageNet-1K. We evaluate ViT-B/14 representations after 100 epochs on ImageNet-1K (Table 3). Without an EMA teacher, SPHERE-JEPA matches LeJEPA's k-NN accuracy and improves linear-probing accuracy by 1.8 percentage points. Adding an EMA teacher yields substantial gains for both methods (Table 3). This shared effect is consistent with the teacher stabilizing prediction targets; the experiment does not isolate that mechanism. We therefore do not treat EMA as a requirement of SUSReg or of the geometric result.

ImageNet-100. We evaluate ResNet-18 backbones across four independent pretraining seeds (Table 4). LeJEPA obtains higher mean accuracy than SPHERE-JEPA on both linear probing and k-NN. Thus, the hyperspherical prior does not yield an improvement on this benchmark under the evaluated protocol.

Sensitivity to the regularizer weight. We vary the SUSReg balancing weight λ while keeping the architecture, training schedule, random seed, and evaluation protocol fixed. The single-seed ablation and its results are reported in Appendix C.

6.2 Downstream and Specialized Evaluations

Transfer Learning. We further assess transferability via linear probing on downstream tasks using the ImageNet-1K pretrained ViT-B/14. As shown in Table 1, SPHERE-JEPA consistently improves across most datasets, achieving a higher average accuracy (54.1%) than LeJEPA (52.4%).

Texture Retrieval. While standard benchmarks assess linear separability, our theory suggests hyperspherical representations excel with nonparametric estimators like k-NN. We test this directly on a controlled texture retrieval task (details in Appendix E). As reported in Table 2, SPHERE-JEPA significantly outperforms LeJEPA, improving mean Average Precision by $+6.36\%$ directly in the embedding space (0.9184 vs. 0.8548) and $+7.34\%$ after the projection head (0.9292 vs. 0.8558).

Galaxy10. We evaluate ResNet-50 backbones across four independent pretraining seeds on Galaxy10 (Table 5). LeJEPA obtains higher mean accuracy on both linear probing and k-NN. The gains observed in our other evaluations, therefore, do not extend to this setting.

Dense prediction. We pretrain ViT-Tiny backbones on ImageNet-100 with InfoNCE, VICReg, LeJEPA, and SPHERE-JEPA, then evaluate their frozen representations on ImageNet-100 and ADE20K. Across four pretraining seeds, SPHERE-JEPA obtains the highest mean on all four metrics. On ADE20K, it reaches $12.88 \pm 0.39\%$ mIoU, compared with $11.76 \pm 0.24\%$ for LeJEPA. The protocol and complete results are reported in Appendix D.

Table 1: Frozen-backbone linear probing accuracy (%) on downstream datasets using ViT-B/14 backbones pretrained for 100 epochs on ImageNet-1K. Avg denotes the mean across all datasets.

Method	DTD	Aircraft	CIFAR10	CIFAR100	Flowers	Food	Pets	Avg
LeJEPA	60.5	09.0	82.0	52.1	45.3	56.9	60.8	52.4
SPHERE-JEPA	60.0	10.0	82.5	53.1	51.1	60.3	62.0	**54.1**
LeJEPA (EMA)	58.5	10.4	81.9	49.2	59.0	68.3	79.6	58.1
SPHERE-JEPA (EMA)	60.4	10.5	82.2	51.8	52.8	68.6	81.8	**58.3**

Table 2: Nearest-neighbor retrieval performance averaged across four texture datasets.

Method	Recall@1	Recall@3	Recall@5	mAP	mAP (emb)
LeJEPA	78.2	91.6	95.1	85.6	85.5
SPHERE-JEPA	**89.0**	**96.3**	**97.9**	**92.9**	**91.8**

Table 3: ImageNet-1K (ViT-B/14, 100 ep.) evaluation.

Method	k-NN (%)	Linear (%)
Without EMA		
LeJEPA	**50.4**	59.2
SPHERE-JEPA	**50.4**	**61.0**
With EMA ($\tau = 0.996$)		
LeJEPA	61.7	68.2
SPHERE-JEPA	**62.6**	**69.3**

Table 4: ImageNet-100 (ResNet-18, 200 ep.). Mean $\pm$ sample standard deviation over four pretraining seeds.

Method	k-NN (%)	Linear (%)
LeJEPA	$\mathbf{67.61 \pm 0.30}$	$\mathbf{72.46 \pm 0.24}$
SPHERE-JEPA	65.67 ± 0.56	71.22 ± 0.58

Table 5: Galaxy10 (ResNet-50, 200 ep.). Mean $\pm$ sample standard deviation over four pretraining seeds.

Method	k-NN (%)	Linear (%)
LeJEPA	$\mathbf{70.56 \pm 0.51}$	$\mathbf{75.03 \pm 0.69}$
SPHERE-JEPA	68.67 ± 1.04	72.13 ± 1.39

7 Conclusion

We introduced a minimax perspective on optimal SSL representations, showing that linear ridge regression, k-NN regression, and exponential-kernel ridge regression collectively characterize the uniform distribution on the sphere as an optimal representation geometry. Motivated by this result, we introduced SPHERE-JEPA and its projection-based regularizer, SUSReg. Empirically, spherical uniformity improves linear separability on ImageNet1K and substantially boosts nonparametric retrieval performance. However, our empirical validation remains limited: our large-scale evaluation relies on a single ViT-B/14 training run without hyperparameter tuning. Validating robustness across multiple seeds, scaling to larger batch sizes and architectures remain open challenges.

Acknowledgements

This work was partially funded by ANRT with the CIFRE grant n° 2024/1221, Advanced Track and Trace, and Centre Borelli. This work was granted access to the HPC resources of IDRIS (Jean Zay supercomputer) under the allocation 2025-AD011017323 made by GENCI.

References

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of machine learning research*, 18(21):1–38, 2017.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 1995.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021a.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

John B Conway. *A course in functional analysis*. Springer, 2019.

J. A. Cuesta-Albertos, R. Fraiman, and T. Ransford. A sharp form of the Cramér–Wold theorem. *Journal of Theoretical Probability*, 2007.

Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, Francesca Odone, and Peter Bartlett. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(5), 2005.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. ISSN 00063444. URL `http://www.jstor.org/stable/2336512`.

Alfred Gray. The volume of a small geodesic ball of a riemannian manifold. *Michigan Mathematical Journal*, 20(4):329–344, 1974.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2 edition, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL `https://doi.org/10.1007/978-0-387-84858-7`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Guillermo Henry, Andrés Muñoz, and Daniela Rodriguez. k-nearest neighbor density estimation on riemannian manifolds, 2011. URL `https://arxiv.org/abs/1106.4763`.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Henry W Leung and Jo Bovy. Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astronomical Society*, November 2018. ISSN 1365-2966. doi: 10.1093/mnras/sty3217. URL http://dx.doi.org/10.1093/mnras/sty3217.

Charles A Micchelli and Grace Wahba. Design problems for optimal surface interpolation. Technical report, 1979.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Bruno Pelletier. Kernel density estimation on riemannian manifolds. *Statistics & probability letters*, 73(3):297–304, 2005.

P. Petersen. Riemannian geometry. *Graduate Texts in Mathematics*, 2006.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018.

B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. MIT Press, 2002. ISBN 9780262194754. URL https://books.google.ch/books?id=y8ORL3DWt4sC.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

A Computational Resources

This appendix provides the hardware specifications and computational requirements necessary to reproduce the results presented in this work.

Hardware Infrastructure. Our experiments were conducted using the following hardware setups:

- **ImageNet-1K Pretraining (ViT-B/14):** All runs (with and without EMA) were executed using 4 NVIDIA A100 (80GB) GPUs.
- **Other Models:** Pretraining for ResNet-18 on ImageNet-100 and ResNet-50 on Galaxy10, as well as all downstream linear probing evaluations, were performed on a single NVIDIA RTX-6000 GPU.

Training Duration and Efficiency. For the 100-epoch pretraining of a ViT-B/14 on ImageNet-1K, the execution time was approximately 2.5 days (60 hours) per run. It is important to note that **SPHERE-JEPA** does not incur any additional computational overhead compared to the LeJEPA baseline.

Software Environment. All models were implemented using PyTorch and trained using the AdamW optimizer. We will release the core implementation and training scripts to ensure full reproducibility.

B Existing Assets and Licenses

The datasets and software libraries used in this work are publicly available and strictly used for academic research purposes, in compliance with their respective licenses. PyTorch is released under the BSD-style license. ImageNet-1K and ImageNet-100 are subject to the ImageNet terms of access for non-commercial research. The Galaxy10 DECals dataset is derived from the DESI Legacy Imaging Surveys and Galaxy Zoo, distributed under the MIT license and CC BY 4.0, respectively.

C Sensitivity to the Regularizer Weight

We vary the balancing weight λ in

$$\mathcal{L}_{\text{SPHERE-JEPA}} = (1-\lambda)\mathcal{L}_{\text{inv}} + \lambda\mathcal{L}_{\text{SUSReg}}.$$

All runs, reported in Table 6, use the same ResNet-18 architecture, ImageNet-100 training schedule, random seed, and evaluation protocol. Thus, this experiment isolates the effect of λ for one pretraining seed; it is distinct from the four-seed comparison in Table 4.

Table 6: ImageNet-100 top-1 accuracy (%) for different values of the SUSReg balancing weight λ (one pretraining seed).

λ	Linear probe	k-NN
0.003	68.8	61.7
0.005	69.6	64.2
0.007	70.6	64.4
0.009	70.7	65.7
0.01	70.5	65.2
0.03	70.8	65.7
0.05	70.8	64.8
0.07	70.0	64.3
0.09	68.8	62.4

Performance varies with λ: both metrics decrease at the smallest and largest tested values. Between 0.007 and 0.05, linear-probe accuracy ranges from 70.5% to 70.8%, while k-NN accuracy ranges from 64.4% to 65.7%. The best linear-probe accuracy is obtained at 0.03 and 0.05; the best k-NN accuracy is obtained at 0.009 and 0.03.

D ViT-Tiny Evaluation and Dense Prediction

We pretrain ViT-Tiny backbones on ImageNet-100 using InfoNCE, VICReg, LeJEPA, and SPHERE-JEPA. For each method, we evaluate the pretrained representations with a linear probe and k-NN classification on ImageNet-100. To assess spatial transfer, we freeze the backbone and train a linear segmentation head on its patch features for ADE20K. Table 7 reports the mean and sample standard deviation over four independent pretraining seeds.

Table 7: Evaluation of ViT-Tiny representations pretrained on ImageNet-100. We report top-1 linear-probe and k-NN accuracy on ImageNet-100, and linear segmentation-probe performance with a frozen backbone on ADE20K. All values are percentages (mean $\pm$ sample standard deviation over four independent pretraining seeds). Bold indicates the highest mean in each column.

	ImageNet-100		ADE20K	
Method	Linear probe	k-NN	mIoU	Pixel acc.
InfoNCE	51.93 ± 4.01	56.93 ± 0.58	7.81 ± 0.45	55.26 ± 0.43
VICReg	57.74 ± 0.51	56.01 ± 0.48	9.20 ± 0.56	56.76 ± 0.59
LeJEPA	70.71 ± 0.38	67.39 ± 0.61	11.76 ± 0.24	58.94 ± 0.26
SPHERE-JEPA (SUSReg)	$\mathbf{71.91} \pm 0.90$	$\mathbf{69.03} \pm 0.43$	$\mathbf{12.88} \pm 0.39$	$\mathbf{59.92} \pm 0.32$

SPHERE-JEPA shows the highest observed mean on all four metrics. Relative to LeJEPA, its gains are 1.20 percentage points in linear-probe accuracy, 1.64 points in k-NN accuracy, 1.12 points in ADE20K mIoU, and 0.98 points in pixel accuracy.

E Texture Retrieval Details

E.1 Problem Setup

We consider a nonparametric retrieval task designed to evaluate the geometry of learned representations. Given a query image, the objective is to retrieve another view of the same texture instance among visually similar samples.

In this setting, each data point is defined as a set of transformed views generated from the same underlying texture. This formulation encourages representations of transformed views to remain close in the embedding space while preserving separation across different textures.

The resulting evaluation protocol depends on the relative distanceamong nearby samples in the representation space, making it particularly well-suited toor analyzing how representation regularization affects nearest-neighbor retrieval.

E.2 Datasets

We instantiate the retrieval protocol using four procedural texture datasets: *Disk*, *Cloud*, *Flake*, and *Wood* (see Figure 4). Each dataset is generated from a distinct stochastic process, defining a family of textures with shared statistical properties.

Disk and Flake textures are derived from heavy-tailed noise with additional smoothing or structured filtering, respectively. Cloud textures are generated from blurred Brownian motion, yielding smooth and self-similar patterns. Wood textures are obtained from Perlin noise using the `noise` library[1].

Within each dataset, samples are generated from the same underlying stochastic process with different random seeds, yielding images that share similar statistical structure yet remain visually distinct.

For each texture family, we generate $10{,}000$ samples for training, 500 for validation, and $10{,}000$ for testing, using independent random seeds to ensure no overlap between splits.

[1] `https://pypi.org/project/noise/`

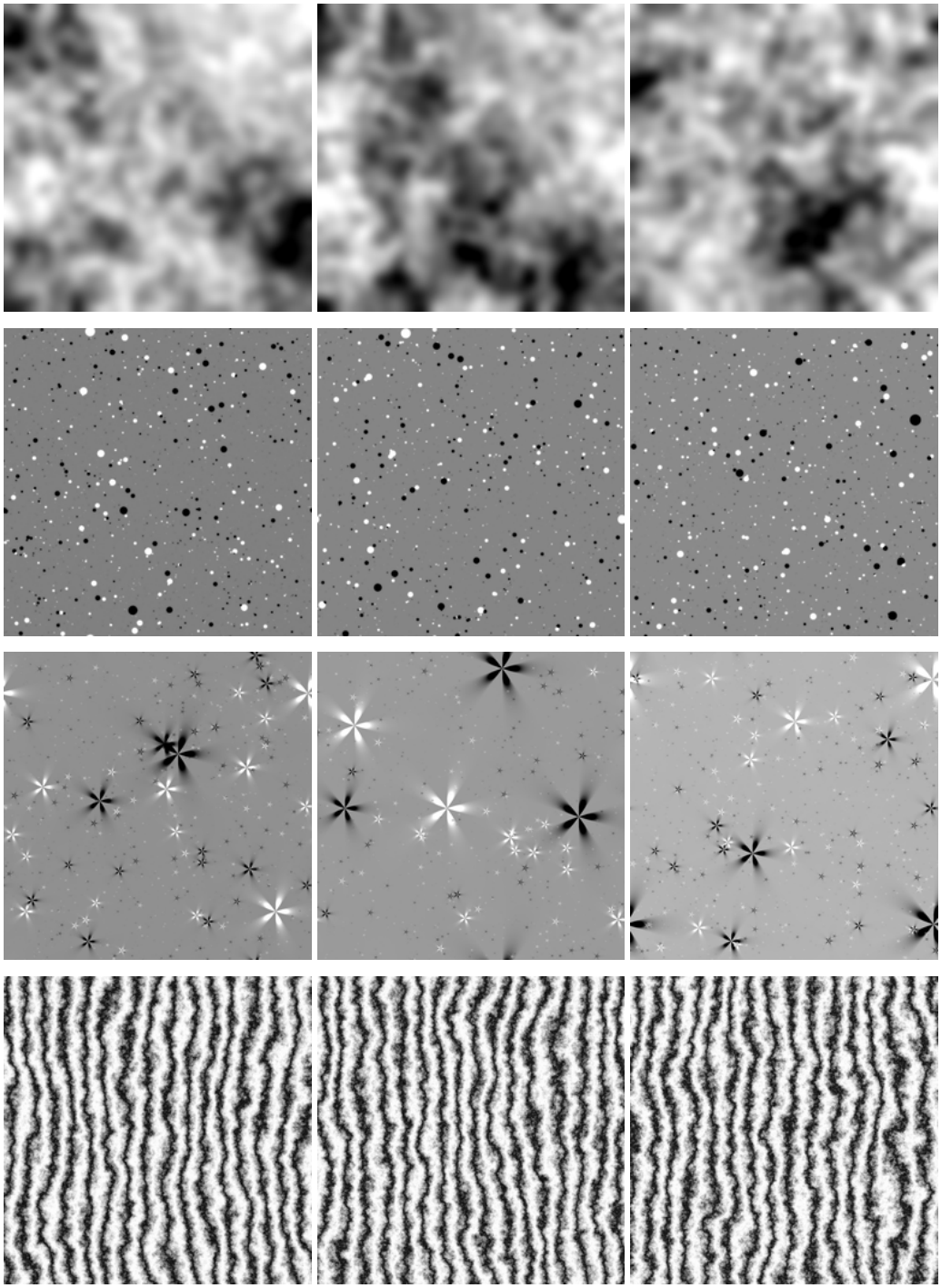


Figure 4: Examples of procedural textures used for retrieval evaluation. Each row corresponds to a texture family (top to bottom: cloud, disk, flake, wood). Images within a row are generated from the same stochastic process with different random seeds, resulting in strong statistical similarity but distinct visual realizations.

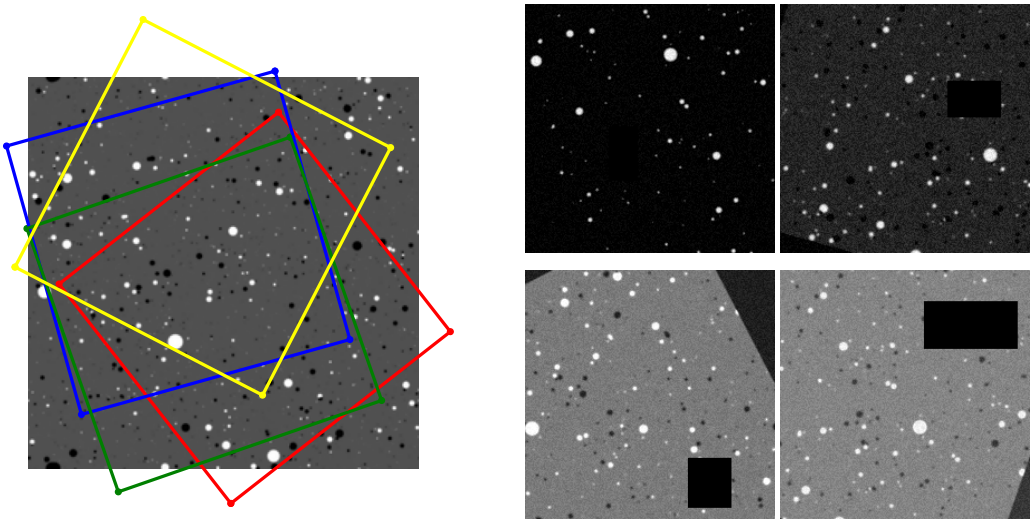


Figure 5: **Left:** effective source regions induced by independently sampled random affine transformations. **Right:** corresponding augmented views obtained after affine and photometric transformations.

E.3 Data Augmentation and View Generation

To generate multiple views from each texture image, we apply stochastic spatial and photometric augmentations. These transformations introduce significant variability while preserving the underlying texture statistics across views.

Each view is generated using a random affine transformation implemented with Kornia. The transformation parameters include a rotation angle uniformly sampled in $[0, 360]$ degrees, a translation of up to 20% of the image size along each spatial dimension, an isotropic scaling factor, and a random shear transformation.

The transformed image is obtained by zero-padding outside the image boundaries and is resized to a fixed resolution of 224×224 pixels. Figure 5 illustrates both the effective source regions induced by independently sampled transformations and the corresponding augmented views.

Each view is further modified through photometric augmentations, including brightness and contrast adjustments, additive noise, and random erasing. All images are converted to grayscale, clamped to valid intensity ranges, and normalized before being fed to the network.

Despite strong spatial and photometric variations, the resulting views preserve consistent texture statistics while exhibiting substantial local variability, as illustrated in Figure 5.

Unless otherwise specified, we use $V_g = 2$ views per image. The same augmentation pipeline is used at test time to ensure consistency between training and evaluation.

E.4 Retrieval Results

We evaluate retrieval performance using a nearest-neighbor protocol. For each query image, the goal is to retrieve another view of the same texture instance among a set of candidate samples.

Evaluation is performed within mini-batches of size $B = 100$. For each query, similarity is computed against the other samples in the batch, and retrieval performance is measured based on the ranking of these candidates.

We report nearest-neighbor retrieval performance using Recall@K and mean Average Precision (mAP), evaluated both on the projection head outputs and directly on backbone embeddings.

Tables 8, 9, 10, 11 report results for each dataset, while Table 2 summarizes the average performance across all four datasets.

These results highlight the importance of representation geometry in nonparametric settings, with hyperspherical regularization leading to improved performance when evaluated directly in the embedding space.

Table 8: Retrieval performance on the Disk texture dataset.

Method	Recall@1	Recall@3	Recall@5	mAP	mAP (emb)
byol	95.7	98.7	99.1	**97.3**	61.2
spherejepa	89.1	95.9	97.5	92.8	**90.6**
lejepa	82.3	92.4	95.2	87.9	88.0
vicreg	67.0	82.1	86.9	75.9	75.7

Table 9: Retrieval performance on the Flake texture dataset.

Method	Recall@1	Recall@3	Recall@5	mAP	mAP (emb)
byol	88.4	96.0	97.9	**92.5**	**89.3**
spherejepa	83.0	94.6	97.2	89.2	89.0
lejepa	79.5	93.2	96.3	86.9	87.5
vicreg	69.7	86.6	92.0	79.3	80.3

Table 10: Retrieval performance on the Cloud texture dataset.

Method	Recall@1	Recall@3	Recall@5	mAP	mAP (emb)
byol	93.6	97.3	98.2	**95.6**	80.4
spherejepa	90.1	95.7	97.3	93.3	**91.8**
lejepa	87.4	94.8	96.7	91.5	90.6
vicreg	81.8	90.5	93.3	86.9	87.4

F Worst-case integrated squared bias for kernel ridge regression

We briefly recall the population and empirical kernel ridge regression (KRR) formulations.

Let

$$K(x, y) = \exp(\kappa x^\top y), \qquad \kappa > 0,$$

let $\mathcal{H}_K$ denote the associated RKHS (see, e.g., Schölkopf and Smola [2002]), and let p be a probability distribution on $\mathcal{M} \subseteq \mathbb{R}^d$.

For $g_\star \in \mathcal{H}_K$, the empirical KRR estimator based on

$$x_1, \ldots, x_B \overset{i.i.d.}{\sim} p$$

is defined by

$$\widehat{g}_{B,\lambda} = \arg\min_{\hat{g} \in \mathcal{H}_K} \frac{1}{B} \sum_{i=1}^{B} (\hat{g}(x_i) - g_\star(x_i))^2 + \lambda \|\hat{g}\|^2_{\mathcal{H}_K}, \qquad \lambda > 0.$$

Equivalently, in operator form (see, e.g., Schölkopf and Smola [2002]),

$$\widehat{g}_{B,\lambda} = (T_B + \lambda I)^{-1} T_B g_\star,$$

where

$$T_B g = \frac{1}{B} \sum_{i=1}^{B} K(x_i, \cdot)\, g(x_i)$$

denotes the empirical covariance operator.

Table 11: Retrieval performance on the Wood texture dataset.

Method	Recall@1	Recall@3	Recall@5	mAP	mAP (emb)
byol	98.9	99.8	99.9	**99.4**	93.0
spherejepa	93.8	99.0	99.7	96.4	**95.9**
lejepa	63.8	86.0	92.3	76.0	75.7
vicreg	16.3	39.7	56.2	34.6	35.0

The corresponding population covariance operator is

$$T_p g = \int_{\mathcal{M}} K(x, \cdot)\, g(x)\, p(x)\, \mathrm{dvol}(x).$$

Assume that

$$\int_{\mathcal{M}} K(x, x)\, p(x)\, \mathrm{dvol}(x) < \infty.$$

Under this condition, the empirical covariance operator converges almost surely to the population covariance operator as $B \to \infty$ [De Vito et al., 2005]. Consequently, in the large-sample regime, the integrated squared bias of empirical KRR is asymptotically governed by its population counterpart. Therefore, throughout the remainder of the analysis, we study the population bias functional directly.

The population integrated squared bias associated with $g_\star$ is

$$\mathrm{ISB}(g_\star, p) = \int_{\mathcal{M}} \left[\lambda (T_p + \lambda I)^{-1} g_\star(x)\right]^2 p(x)\, \mathrm{dvol}(x).$$

We consider the source class

$$\mathcal{G}_p = \left\{ g_\star = T_p f \ : \ \|f\|_{L^2(p)} \le 1 \right\}.$$

We define the worst-case population integrated squared bias by

$$\mathcal{E}(p) = \sup_{g_\star \in \mathcal{G}_p} \mathrm{ISB}_p(g_\star).$$

Since K is positive definite,

$$|K(x, y)|^2 \le K(x, x) K(y, y).$$

Hence the assumption

$$\int_{\mathcal{M}} K(x, x)\, p(x)\, \mathrm{dvol}(x) < \infty$$

implies

$$\int_{\mathcal{M}} \int_{\mathcal{M}} |K(x, y)|^2 p(x) p(y)\, \mathrm{dvol}(x) \mathrm{dvol}(y) < \infty.$$

Therefore, the population covariance operator T_p is Hilbert–Schmidt, hence compact, on $L^2(p)$. Moreover, since K is symmetric and positive definite, T_p is self-adjoint and positive.

Consequently, there exist eigenvalues

$$\mu_1 \ge \mu_2 \ge \cdots \ge 0, \qquad \mu_j \to 0,$$

and an orthonormal basis

$$(e_j)_{j \ge 1}$$

of $L^2(p)$ such that

$$T_p e_j = \mu_j e_j, \qquad j \ge 1.$$

Lemma F.1 (Spectral form of the population worst-case ISB)**.** *Let $(\mu_j, e_j)_{j \ge 1}$ be the eigendecomposition of T_p. Then*

$$\mathcal{E}(p) = \max_{j \ge 1} \left(\frac{\mu_j \lambda}{\mu_j + \lambda} \right)^2.$$

Proposition F.1 (Uniform spherical distributions minimize the population worst-case KRR bias)**.** *Assume that p is a density on $\mathcal{M}$ satisfying*

$$\int_{\mathcal{M}} \|x\|^2 p(x)\,\mathrm{dvol}(x) = 1.$$

Then minimizing the worst-case population bias $\mathcal{E}(p)$ is equivalent to minimizing the top eigenvalue μ_1 of the population covariance operator T_p.

Moreover, for the exponential kernel

$$K(x, y) = e^{\kappa x^\top y},$$

the minimizer is

$$p^\star = \mathrm{Unif}(\mathbb{S}^{d-1}).$$

G Proofs: Kernel Ridge Regression

Lemma F.1 (Spectral form of the population worst-case ISB)**.** *Let $(\mu_j, e_j)_{j\geq 1}$ be the eigendecomposition of T_p. Then*

$$\mathcal{E}(p) = \max_{j\geq 1} \left(\frac{\mu_j \lambda}{\mu_j + \lambda}\right)^2.$$

Proof. Expand

$$f = \sum_{j\geq 1} a_j e_j, \qquad \sum_{j\geq 1} a_j^2 \leq 1.$$

Since $g_\star = T_p f$, we have

$$g_\star = \sum_{j\geq 1} \mu_j a_j e_j.$$

Moreover,

$$\lambda(T_p + \lambda I)^{-1} g_\star = \sum_{j\geq 1} \frac{\lambda \mu_j}{\mu_j + \lambda} a_j e_j.$$

Therefore,

$$\left\|\lambda(T_p + \lambda I)^{-1} g_\star\right\|^2_{L^2(p)} = \sum_{j\geq 1} \left(\frac{\lambda \mu_j}{\mu_j + \lambda}\right)^2 a_j^2.$$

Taking the supremum over $\sum_{j\geq 1} a_j^2 \leq 1$, the optimum is attained by concentrating all mass on the eigendirection maximizing $\left(\frac{\lambda \mu_j}{\mu_j + \lambda}\right)^2$.

Hence

$$\mathcal{E}(p) = \max_{j\geq 1} \left(\frac{\lambda \mu_j}{\mu_j + \lambda}\right)^2.$$

□

Proposition F.1 (Uniform spherical distributions minimize the population worst-case KRR bias)**.** *Assume that p is a density on $\mathcal{M}$ satisfying*

$$\int_{\mathcal{M}} \|x\|^2 p(x)\,\mathrm{dvol}(x) = 1.$$

Then minimizing the worst-case population bias $\mathcal{E}(p)$ is equivalent to minimizing the top eigenvalue μ_1 of the population covariance operator T_p.

Moreover, for the exponential kernel

$$K(x, y) = e^{\kappa x^\top y},$$

the minimizer is

$$p^\star = \mathrm{Unif}(\mathbb{S}^{d-1}).$$

Proof. Define

$$h_\lambda(\mu) = \left(\frac{\mu\lambda}{\mu+\lambda}\right)^2.$$

Since h_λ increases strictly on $[0, +\infty)$, Lemma F.1 implies

$$\mathcal{E}(p) = h_\lambda(\mu_1).$$

Therefore minimizing $\mathcal{E}(p)$ is equivalent to minimizing μ_1.

Using the expansion

$$e^{\kappa x^\top y} = \sum_{m=0}^{\infty} \frac{\kappa\langle x, y\rangle^m}{m!},$$

the kernel admits a feature representation

$$K(x, y) = \langle \Phi(x), \Phi(y)\rangle.$$

By the spectral characterization of compact positive self-adjoint operators (see, e.g., Conway [2019]),

$$\mu_1 = \sup_{\|a\|=1} \langle T_p a, a\rangle.$$

Using the feature representation

$$K(x, y) = \langle \Phi(x), \Phi(y)\rangle,$$

Rewriting the integral that defines T_p as an expectation and thanks to the feature representation, the covariance operator can be written as:

$$T_p a = \mathbb{E}_{X\sim p}\big[\langle a, \Phi(X)\rangle\, \Phi(X)\big].$$

Therefore,

$$\langle T_p a, a\rangle = \mathbb{E}_{X\sim p}\big[\langle a, \Phi(X)\rangle^2\big].$$

and hence

$$\mu_1 = \sup_{\|a\|=1} \mathbb{E}_{X\sim p}\big[\langle a, \Phi(X)\rangle^2\big]. \tag{6}$$

Thus μ_1 corresponds to the largest variance direction in feature space.

Next, the kernel is rotationally invariant:

$$K(Qx, Qy) = K(x, y), \qquad Q \in O(d).$$

Let dQ denote the normalized Haar probability measure on the orthogonal group $O(d)$. Given any admissible distribution p, define its rotational symmetrization by

$$\bar{p}(A) = \int_{O(d)} Q^* p(A)\, dQ,$$

with Q^* the pullback of p by Q, for every measurable set $A \subseteq \mathcal{M}$.

By construction, $\bar{p}$ is rotationally invariant and remains a probability distribution. Since orthogonal transformations preserve Euclidean norms,

$$\int_{\mathcal{M}} \|x\|^2 \bar{p}(x)\, \mathrm{dvol}(x) = 1.$$

Moreover, by (6) and the fact that the supremum of linear functionals is convex, $p \mapsto \mu_1(T_p)$ is convex. Therefore,

$$\mu_1(T_{\bar{p}}) \le \int_{O(d)} \mu_1(T_{Q^*p})\, dQ,$$

Since the kernel is rotationally invariant,

$$\mu_1(T_{Q^*p}) = \mu_1(T_p),$$

and hence

$$\mu_1(T_{\bar{p}}) \le \mu_1(T_p).$$

Therefore, without loss of generality, we may restrict attention to rotationally invariant distributions p.

Let $X \sim p$. Since p is rotationally invariant, a standard spherical decomposition (see, e.g., Vershynin [2018]) yields $X = RU$, where $U \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, with $R \ge 0$, and R independent of U, and satisfying $\mathbb{E}[R^2] = 1$.

Let $Y = SV$ be an independent copy of X, where $V \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ is independent of S.

Then

$$X^\top Y = RS\, U^\top V,$$

and

$$\mathbb{E}[e^{\kappa X^\top Y}] = \mathbb{E}[g_d(RS)],$$

where

$$g_d(t) = \mathbb{E}_{U,V}[e^{t\kappa U^\top V}].$$

Since

$$e^{t\kappa U^\top V} = \sum_{m=0}^{\infty} \frac{t^m \kappa^m (U^\top V)^m}{m!},$$

symmetry implies

$$g_d(t) = \sum_{m=0}^{\infty} c_m t^{2m}, \qquad c_m \ge 0.$$

Therefore,

$$\mathbb{E}[g_d(RS)] = \sum_{m=0}^{\infty} c_m (\mathbb{E}[R^{2m}])^2.$$

By Jensen's inequality,

$$\mathbb{E}[R^{2m}] \ge (\mathbb{E}[R^2])^m = 1.$$

Hence

$$\mathbb{E}[g_d(RS)] \ge \sum_{m=0}^{\infty} c_m = g_d(1).$$

Equality holds only if

$$\mathbb{E}[R^{2m}] = 1, \qquad m \ge 1.$$

By Lemma G.1, this implies

$$R = 1 \qquad \text{a.s.}$$

Since we already reduced the problem to rotationally invariant distributions, any admissible optimizer must satisfy

$$X = RU, \qquad U \sim \mathrm{Unif}(\mathbb{S}^{d-1}).$$

Hence the condition $R = 1$ almost surely implies

$$X \sim \mathrm{Unif}(\mathbb{S}^{d-1}).$$

Therefore,

$$p^\star = \mathrm{Unif}(\mathbb{S}^{d-1})$$

minimizes μ_1, and consequently minimizes $\mathcal{E}(p)$. □

Lemma G.1. *Let $R \ge 0$ be a random variable such that*

$$\mathbb{E}[R^{2m}] = 1, \qquad m \ge 1.$$

Then

$$R = 1 \qquad \textit{almost surely}.$$

Proof. Taking $m = 1$ and $m = 2$ gives

$$\mathbb{E}[R^2] = 1, \qquad \mathbb{E}[R^4] = 1.$$

Hence

$$\mathrm{Var}(R^2) = \mathbb{E}[R^4] - \mathbb{E}[R^2]^2 = 0,$$

so $R^2 = 1$ almost surely. Since $R \geq 0$, we conclude that $R = 1$ almost surely. □

Remark G.2. The normalization

$$\int_{\mathcal{M}} \|x\|^2 p(x) \, \mathrm{dvol}(x) = 1$$

rules out the collapsed solution $p = \delta_0$, for which $K(x, y) \equiv \phi(0)$. The argument extends to dot-product kernels $K(x, y) = \phi(x^\top y)$ with

$$\phi(t) = \sum_{\ell \geq 0} a_\ell t^\ell, \qquad a_\ell \geq 0.$$

It is sufficient that $a_{2j} > 0$ for some $j \geq 2$. Indeed, equality then requires $\mathbb{E}[R^{2j}] = (\mathbb{E}[R^2])^j = 1$, and strict convexity implies $R = 1$ almost surely.

H Uniform Distributions Minimize k-NN Regression Bias Modified Fisher Information Term

The work of Balestriero and LeCun Balestriero and LeCun [2025] makes a theoretical case for isotropic Gaussian embeddings by demonstrating their optimal performance in both linear probing (ordinary least-squares regression) and nonlinear probing (k-NN and Nadaraya-Watson kernel regression). Specifically, they show that isotropic embeddings (not necessarily Gaussian) minimize the Tikhonov-regularized loss of ridge regression models fit to an arbitrary target on the embeddings using the linear kernel. Moreover, they show that isotropic Gaussian embeddings minimize, for a fixed number of training examples, the integrated square bias term of the bias-variance decomposition of the mean-squared error loss of k-NN regression and Nadaraya-Watson kernel regression under certain isotropy constraints (the prior introduced on p. 29 of that work). However, their proofs involve assumptions that rule out common choices of latent representations, such as spherical embeddings. For example, they linearize the integrated square bias term with a Euclidean Taylor expansion that imposes smoothness constraints on the embedding distribution and regression target function over $\mathbb{R}^d$ that fail for functions defined on the sphere (ibid., Lemmas 4 and 5). Moreover, their optimality results (ibid., Theorems 1 and 7-9) depend on the fact that the leading term of this expansion, the Euclidean Fisher information integral

$$\int_{\mathbb{R}^d} ||\nabla \log p(x)||_2^2 \, p(x) \, \mathrm{d}x, \tag{7}$$

is minimized (for fixed covariance matrix) by the Gaussian distribution (ibid., Lemma 6)–and, with additional scalar constraints on the distribution (ibid., Theorem 9), the isotropic Gaussian. But this integral is not defined when p is, for instance, a uniform probability measure on the sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$. The uniform measure on the sphere is singular with respect to the Lebesgue measure on the ambient Euclidean space $\mathbb{R}^d$ and does not have a probability density function with respect to the Lebesgue measure on $\mathbb{R}^d$. Moreover, the leading term of the bias of k-NN regression is not proportional to (7) when one accounts for the spatially varying k-NN radius.

The integrated squared bias of k-NN regression and kernel density estimation models on more general Riemannian manifolds can be performed using an expansion in covariant Taylor series, which does not impose these smoothness requirements in the ambient space. Indeed, this approach was used to estimate kernel density estimation and k-NN density estimation–where the regression target is the embedding distribution itself–in [Pelletier, 2005] and [Henry et al., 2011], respectively.

This more general setting tells a different story.

Let $\mathcal{M} \subseteq \mathbb{R}^d$ be an m-dimensional Riemannian manifold with positive injectivity radius $0 < r_{\min} \leq \mathrm{inj}_{\mathcal{M}}(x)$ at all points $x \in \mathcal{M}$ and volume form $\mathrm{dvol}(x)$. Let p be a $\mathcal{C}^3$-smooth probability density on $\mathcal{M}$. For each point $x \in \mathcal{M}$, let $B_r(x) \subseteq \mathcal{M}$ be the geodesic ball of radius r.

Given a target function $f \in C^3(\mathcal{M})$, define the population-radius local average

$$\widehat{f}_r(x) = \mathbb{E}_{Y\sim p}[f(Y) \mid Y \in B_r(x)] = \frac{\int_{B_r(x)} f(y)p(y)\,\mathrm{dvol}(y)}{\int_{B_r(x)} p(y)\,\mathrm{dvol}(y)}.$$

We evaluate it at the deterministic radius $r = r_k(x)$ defined by

$$\int_{B_{r_k(x)}(x)} p(y)\,\mathrm{dvol}(y) = \frac{k}{n}.$$

Throughout, $k \to \infty$ and $k/n \to 0$.

The pointwise bias of our estimate is simply

$$\mathrm{bias}(x; r_k(x)) = \mathbb{E}_{y\sim p}\big[f(y) \mid y \in B_r(x)\big] - f(x) = \widehat{f}_r(x) - f(x);$$

the integrated squared bias,

$$\mathrm{ISB}(f, p) = \int_{\mathcal{M}} \mathrm{bias}(x; r_k(x))^2\, p(x)\,\mathrm{dvol}(x). \tag{8}$$

Lemma H.1 (Riemannian analog of Balestriero and LeCun [2025], Lemma 4)**.** *For $f, p \in C^3(\mathcal{M})$ with $p(x) > 0$, the pointwise bias satisfies*

$$\mathrm{bias}(x; r_k(x)) = \frac{r_k(x)^2}{2(m+2)}\left(\Delta_{\mathcal{M}} f(x) + 2\langle \nabla f(x), \nabla \log p(x)\rangle_g\right) + O(r_k(x)^3).$$

We shall refer to the term $\Delta_{\mathcal{M}} f(x)$ as the wiggliness bias and $\langle \nabla f(x), \nabla p(x)\rangle_g / p(x)$ as the design bias.

Proposition H.1 (Riemannian analog of Balestriero and LeCun [2025], Theorem 7)**.** *Let $\mathcal{M}$ be a compact, smooth, m-dimensional Riemannian manifold without boundary. Let $f, p \in C^3(\mathcal{M})$ be fixed, with $p > 0$ and $\int_{\mathcal{M}} p\,\mathrm{dvol} = 1$. Assume $k \to \infty$ and $k/n \to 0$. Here $r_k(x)$ is the deterministic population radius satisfying $P(B_{r_k(x)}(x)) = k/n$. The leading term of the integrated squared bias of k-NN regression* (8) *is*

$$\mathrm{ISB}(f, p; k, n) = A(m, k, n) \int_{\mathcal{M}} \left(\Delta_{\mathcal{M}} f(x) + 2\frac{\langle \nabla f(x), \nabla p(x)\rangle_g}{p(x)}\right)^2 p(x)^{1-4/m}\,\mathrm{dvol}(x) + O\left(\left(\frac{k}{n}\right)^{5/m}\right), \tag{9}$$

where $A(m, k, n) \coloneqq \frac{1}{4(m+2)^2}\left(\frac{k\Gamma(1+m/2)}{n\pi^{m/2}}\right)^{4/m}$.

Remark H.1. The term $p(x)^{-4/m}$ does not appear in the result of Balestriero and LeCun [2025]; it arises due to the spatially varying bandwidth of a k-NN regression.

Remark H.2. Under the Laplacian constraint used below, the leading worst-case functional is finite for each fixed positive density on a compact manifold. Its asymptotic value is strictly larger for nonuniform densities than for the uniform density.

Using a minimax approach [Micchelli and Wahba, 1979, Bach, 2017], we find that the optimal design for minimizing k-NN regression is a uniform distribution.

Proposition H.2 (Asymptotic minimax optimality of the uniform design for k-NN)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact, connected, m-dimensional Riemannian submanifold without boundary. Let $c > 0$, $k = k(n) \to \infty$, and $k/n \to 0$. Suppose that $M_n \to \infty$ and*

$$M_n \left(\frac{k}{n}\right)^{1/m} \to 0.$$

Define

$$|f|_{C^3} = \max_{1\le j\le 3} \|\nabla^j f\|_\infty, \qquad \mathcal{G}_n = \left\{f \in C^3(\mathcal{M}) : \|\Delta_{\mathcal{M}} f\|_2 \le c,\ |f|_{C^3} \le M_n\right\}.$$

Set $q = 1 - 4/m$, $p_0 = \mathrm{vol}(\mathcal{M})^{-1}$, and

$$A(m, k, n) = \frac{1}{4(m+2)^2}\left(\frac{k\,\Gamma(1+m/2)}{n\pi^{m/2}}\right)^{4/m}.$$

For every fixed positive $p \in C^3(\mathcal{M})$ with $\int_{\mathcal{M}} p \, \mathrm{dvol} = 1$,

$$\liminf_{n\to\infty} \frac{\sup_{f\in\mathcal{G}_n} \mathrm{ISB}(f, p; k, n)}{A(m, k, n)} \geq c^2 \|p^q\|_\infty.$$

For $p = p_0$, the normalized supremum converges to $c^2 p_0^q$. Moreover, every nonconstant p has a strictly larger asymptotic worst-case value. Thus p_0 is the unique pointwise asymptotic minimizer among these fixed densities.

Proposition H.3 (Unbounded Euclidean case)**.** *Let $\mathcal{M} = \mathbb{R}^m$, let p be a positive nonconstant smooth density, and retain q and $\mathcal{G}_n$ as above, with $\mathcal{M} = \mathbb{R}^m$. Then*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{G}_n} \int_{\mathbb{R}^m} \left(\Delta f + 2\langle \nabla f, \nabla \log p\rangle\right)^2 p^q \, dx = \infty.$$

In particular, this holds for every nondegenerate Gaussian density.

I Proofs: k-Nearest Neighbors

Lemma H.1 (Riemannian analog of Balestriero and LeCun [2025], Lemma 4)**.** *For $f, p \in C^3(\mathcal{M})$ with $p(x) > 0$, the pointwise bias satisfies*

$$\mathrm{bias}(x; r_k(x)) = \frac{r_k(x)^2}{2(m+2)} \left(\Delta_{\mathcal{M}} f(x) + 2\langle \nabla f(x), \nabla \log p(x)\rangle_g\right) + O(r_k(x)^3).$$

Proof. For each $x \in \mathcal{M}$, we use the exponential map $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ to transform the integral over the geodesic ball $B_r(x)$ to the tangent ball $B_r^m(0) \subseteq \mathbb{R}^m$. In Riemannian normal coordinates at x, the metric g satisfies the following standard[2] approximation:

$$\sqrt{\det g} = 1 - \frac{1}{6}\mathrm{Ric}_{ij}(x) y^i y^j + O(\|y\|^3),$$

where Ric is the Ricci tensor and y^i the ith coordinate of a point y in an orthonormal basis of $T_x\mathcal{M}$. Thus,

$$\mathrm{dvol}(y) \approx 1 - \frac{1}{6}\mathrm{Ric}_{ij}(x) y^i y^j \mathrm{d}y + O(||y||^3)\mathrm{d}y$$

where we use the notation $\mathrm{d}y = \mathrm{d}y^1 \wedge \ldots \wedge \mathrm{d}y^n$.

Let us now obtain the leading terms of the numerator and denominator using Taylor expansions in normal coordinates. Starting with the denominator, we multiply our expansion of $\mathrm{dvol}(y)$ with a Taylor expansion of $p(y)$

$$p(y) = p(0) + \nabla_k p(0) y^k + \frac{1}{2}\nabla_k \nabla_l p(0) y^k y^l + O(||y||^3)$$

and integrate. We note that by symmetry and linearity, odd terms $\nabla_k p y^k$ integrate to zero over the ball, and distinct coordinates $y^k y^l$ integrate to zero as well, so $\int_{B_r^m(0)} y^k y^l \mathrm{d}y = \delta_{kl} V_m r^{m+2}/(m+2)$,

[2] See [Gray, 1974, Corollary 2.9] or [Petersen, 2006, Exercise 5.9.43, p. 229]. Use the Jacobi identity and the fact that these are normal coordinates (so first derivatives of the metric vanish at x) to expand the metric in terms of the Riemannian curvature tensor R:

$$g_{kl}(y) = \delta_{kl} - \frac{1}{3} R_{kilj}(x) y^i y^j + O(||y||^3).$$

Then apply the approximation $\det(I + \epsilon A) \approx I + \epsilon \mathrm{tr}(A)$ to obtain, with $h_{kl} = -\frac{1}{3} R_{kilj}(x) y^i y^j$

$$\det g \asymp 1 + \mathrm{tr}(h) = 1 - \frac{1}{3}\delta^{kl} R_{kilj} y^i y^j = 1 - \frac{1}{3} R^l_{ilj} y^i y^j = 1 - \frac{1}{3}\mathrm{Ric}_{ij} y^i y^j,$$

since with the sign convention $R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$, the Ricci curvature is a contraction of the Riemannian curvature over the first and third indices, ensuring positive curvature for a sphere. The Taylor approximation $(1-\epsilon)^{1/2} \asymp 1 - \frac{1}{2}\epsilon$ yields

$$\sqrt{\det g} \asymp 1 - \frac{1}{6}\mathrm{Ric}_{ij} y^i y^j.$$

where $V_m \coloneqq \pi^{m/2}/\Gamma\left(\frac{m}{2}+1\right)$ is the volume of the unit ball in $\mathbb{R}^m$.

$$\begin{aligned}
D &= \int_{B_r^m(0)} p(y)\,\mathrm{dvol}(y) \\
&= \int_{B_r^m(0)} \Big[p(0) + \nabla_k p(0) y^k + \frac{1}{2}\nabla_k\nabla_l p(0) y^k y^l + O(\|y\|^3)\Big] \\
&\qquad \cdot \Big[1 - \frac{1}{6}\,\mathrm{Ric}_{ij}(0) y^i y^j + O(\|y\|^3)\Big]\,\mathrm{d}y \\
&= V_m r_k(x)^m \left[p(x) + \frac{r_k(x)^2}{2(m+2)}\Delta_{\mathcal{M}} p(x) - \frac{r_k(x)^2}{6(m+2)} p(x)\,\mathrm{Scal}(x) + O(r_k(x)^3)\right].
\end{aligned}$$

Since the symmetry collapses the sum of second-order derivatives into a trace:

$$\nabla_k\nabla_l p(0)\delta^{kl} = \Delta_{\mathcal{M}} p(x) \qquad \text{and} \qquad \mathrm{Ric}_{ij}(0)\delta^{ij} = \mathrm{Scal}(x)$$

where $\Delta_{\mathcal{M}}$ is the Laplace-Beltrami operator of p at x, and $\mathrm{Scal}(x)$ is the scalar curvature of the manifold at x.

We obtain the numerator N,

$$V_m r^m \left[f(x)p(x) + \frac{r_k(x)^2}{2(m+2)}\Delta_{\mathcal{M}}(fp)(x) - \frac{r_k(x)^2}{6(m+2)} f(x)p(x)\mathrm{Scal}(x)\right] + O(r^{m+3}),$$

by identical reasoning using our Taylor expansion of the product between the target and embedding density

$$f(y)p(y) \asymp \left(f(0) + \nabla_k f(0) y^k + \frac{1}{2}\nabla_k\nabla_l f(0) y^k y^l\right) \cdot \left(p(0) + \nabla_k p(0) y^k + \frac{1}{2}\nabla_k\nabla_l p(0) y^k y^l\right).$$

We write

$$\Delta_{\mathcal{M}}(fp)(x) = f(x)\Delta_{\mathcal{M}} p(x) + 2\langle \nabla f(x), \nabla p(x)\rangle_g + p(x)\Delta_{\mathcal{M}} f(x)$$

and compute the ratio between the numerator and the denominator:

$$\begin{aligned}
\widehat{f}_r(x) &= \frac{N}{D} \\
&= \frac{f(x)p(x) + \frac{r_k(x)^2\left[f(x)\Delta_{\mathcal{M}} p(x) + 2\langle \nabla f(x), \nabla p(x)\rangle_g + p(x)\Delta_{\mathcal{M}} f(x) - \frac{1}{3} f(x)p(x)\,\mathrm{Scal}(x)\right]}{2(m+2)}}{p(x) + \frac{r_k(x)^2}{2(m+2)}\left[\Delta_{\mathcal{M}} p(x) - \frac{1}{3} p(x)\,\mathrm{Scal}(x)\right]} + O(r_k(x)^3) \\
&= f(x) + \frac{r_k(x)^2}{2(m+2)}\left(\Delta_{\mathcal{M}} f(x) + 2\frac{\langle \nabla f(x), \nabla p(x)\rangle_g}{p(x)}\right) + O(r_k(x)^3),
\end{aligned}$$

where the last step used the approximation $\frac{A+\epsilon B}{C+\epsilon D} = \frac{A+\epsilon B}{C}(1+\epsilon\frac{D}{C})^{-1} \approx \frac{A+\epsilon B}{C}(1-\epsilon\frac{D}{C}) = \frac{A}{C} + \epsilon\frac{BC-AD}{C^2}$, which follows from the Taylor approximation $(1+\epsilon\frac{D}{C})^{-1} \approx 1 - \epsilon\frac{D}{C} + O(\epsilon^2)$.

Thus, the bias at x has the expansion $\frac{r_k(x)^2}{2(m+2)}\left(\Delta_{\mathcal{M}} f(x) + 2\frac{\langle \nabla f(x), \nabla p(x)\rangle_g}{p(x)}\right)$:

$$\begin{aligned}
\mathrm{bias}(x; r_k(x)) &= \widehat{f}_r(x) - f(x) \\
&= \frac{r_k(x)^2}{2(m+2)}\left(\Delta_{\mathcal{M}} f(x) + 2\frac{\langle \nabla f(x), \nabla p(x)\rangle_g}{p(x)}\right) + O(r_k(x)^3). \qquad (10)
\end{aligned}$$

□

Proposition H.1 (Riemannian analog of Balestriero and LeCun [2025], Theorem 7)**.** *Let $\mathcal{M}$ be a compact, smooth, m-dimensional Riemannian manifold without boundary. Let $f, p \in C^3(\mathcal{M})$ be fixed, with $p > 0$ and $\int_{\mathcal{M}} p\,\mathrm{dvol} = 1$. Assume $k \to \infty$ and $k/n \to 0$. Here $r_k(x)$ is the deterministic population radius satisfying $P(B_{r_k(x)}(x)) = k/n$. The leading term of the integrated squared bias of k-NN regression* (8) *is*

$$\mathrm{ISB}(f, p; k, n) = A(m,k,n)\int_{\mathcal{M}}\left(\Delta_{\mathcal{M}} f(x) + 2\frac{\langle \nabla f(x), \nabla p(x)\rangle_g}{p(x)}\right)^2 p(x)^{1-4/m}\,\mathrm{dvol}(x) + O\left(\left(\frac{k}{n}\right)^{5/m}\right), \qquad (9)$$

where $A(m,k,n) \coloneqq \frac{1}{4(m+2)^2}\left(\frac{k\Gamma(1+m/2)}{n\pi^{m/2}}\right)^{4/m}$.

Proof. Write

$$V_m = \frac{\pi^{m/2}}{\Gamma(1+m/2)}, \qquad \beta_n = \frac{k}{n}.$$

Throughout this proof, $r_k(x)$ denotes the deterministic population radius defined by

$$P(B_{r_k(x)}(x)) = \beta_n.$$

We work with fixed $f, p \in C^3(\mathcal{M})$, with $p > 0$, on a compact smooth manifold $\mathcal{M}$. The remainders below are then uniform in $x \in \mathcal{M}$. In geodesic normal coordinates y at x, the volume density is

$$\mathrm{dvol}(y) = \left(1 - \frac{1}{6}\,\mathrm{Ric}_{ij}(x)y^i y^j + O(\|y\|^3)\right) \mathrm{d}y.$$

For any $h \in C^3(\mathcal{M})$, Taylor expansion and symmetry of the Euclidean ball give

$$\begin{aligned}\int_{B_r(x)} h \,\mathrm{dvol} &= \int_{B_r^m(0)} \left(h(x) + \nabla_i h(x) y^i + \frac{1}{2}\nabla_i\nabla_j h(x) y^i y^j + O(\|y\|^3)\right) \\ &\quad \cdot \left(1 - \frac{1}{6}\,\mathrm{Ric}_{ij}(x)y^i y^j + O(\|y\|^3)\right) \mathrm{d}y \\ &= V_m r^m \left[h(x) + \frac{r^2}{2(m+2)}\Delta_{\mathcal{M}} h(x) - \frac{r^2}{6(m+2)} h(x)\,\mathrm{Scal}(x) + O(r^3)\right].\end{aligned}$$

In particular, taking $h = 1$ yields

$$\mathrm{vol}(B_r(x)) = V_m r^m \left[1 - \frac{\mathrm{Scal}(x)}{6(m+2)} r^2 + O(r^3)\right].$$

With additional smoothness, symmetry improves the last volume remainder to $O(r^4)$; this improvement is not needed here. Applying the expansion with $h = p$ gives

$$P(B_{r_k(x)}(x)) = V_m r_k(x)^m \left[p(x) + \frac{r_k(x)^2}{2(m+2)}\left(\Delta_{\mathcal{M}} p(x) - \frac{p(x)}{3}\,\mathrm{Scal}(x)\right) + O(r_k(x)^3)\right]. \tag{11}$$

Equivalently, the remainder outside the brackets is $O(r_k(x)^{m+3})$. Define the leading radius

$$r_0(x) = \left(\frac{\beta_n}{V_m p(x)}\right)^{1/m} = \left(\frac{k\Gamma(1+m/2)}{n\pi^{m/2} p(x)}\right)^{1/m}.$$

Since p is positive and bounded away from zero on $\mathcal{M}$, inverting (11) gives

$$r_k(x) = r_0(x) - \frac{r_0(x)^3}{2m(m+2)}\left(\frac{\Delta_{\mathcal{M}} p(x)}{p(x)} - \frac{1}{3}\,\mathrm{Scal}(x)\right) + O(r_0(x)^4). \tag{12}$$

Thus $r_k(x)^2 = r_0(x)^2 + O(r_0(x)^4)$. By Lemma H.1, the pointwise bias of the population-radius estimator is

$$\mathrm{bias}(x; r_k(x)) = \frac{r_k(x)^2}{2(m+2)}\left(\Delta_{\mathcal{M}} f(x) + 2\langle \nabla f(x), \nabla \log p(x)\rangle_g\right) + O(r_k(x)^3).$$

Because $r_0(x) = O(\beta_n^{1/m})$ uniformly in x, substituting (12) changes only the higher-order terms. Setting

$$A_{k,n} = \frac{1}{2(m+2)}\left(\frac{k\Gamma(1+m/2)}{n\pi^{m/2}}\right)^{2/m},$$

we obtain

$$\mathrm{bias}(x; r_k(x)) = A_{k,n} p(x)^{-2/m}\left(\Delta_{\mathcal{M}} f(x) + 2\langle \nabla f(x), \nabla \log p(x)\rangle_g\right) + O(\beta_n^{3/m}).$$

The leading term is $O(\beta_n^{2/m})$, so its product with the remainder is $O(\beta_n^{5/m})$. Consequently,

$$\mathrm{bias}(x; r_k(x))^2 = A_{k,n}^2 p(x)^{-4/m}\left(\Delta_{\mathcal{M}} f(x) + 2\langle \nabla f(x), \nabla \log p(x)\rangle_g\right)^2 + O(\beta_n^{5/m}). \tag{13}$$

Integrating against $p(x)\,\mathrm{dvol}(x)$, and using compactness, gives

$$\mathrm{ISB}(f, p; k, n) = A_{k,n}^2 \int_{\mathcal{M}} \left(\Delta_{\mathcal{M}} f(x) + 2\langle \nabla f(x), \nabla \log p(x)\rangle_g\right)^2 p(x)^{1-4/m}\,\mathrm{dvol}(x) + O\left(\left(\frac{k}{n}\right)^{5/m}\right).$$

□

Remark I.1. If f and p are C^4 and the metric is smooth, the symmetry of the geodesic ball cancels the third-order terms. The remainder in (9) then improves to $O((k/n)^{6/m})$.

Proposition H.2 (Asymptotic minimax optimality of the uniform design for k-NN)**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact, connected, m-dimensional Riemannian submanifold without boundary. Let $c > 0$, $k = k(n) \to \infty$, and $k/n \to 0$. Suppose that $M_n \to \infty$ and*

$$M_n \left(\frac{k}{n}\right)^{1/m} \to 0.$$

Define

$$|f|_{C^3} = \max_{1 \le j \le 3} \|\nabla^j f\|_\infty, \qquad \mathcal{G}_n = \left\{ f \in C^3(\mathcal{M}) : \|\Delta_{\mathcal{M}} f\|_2 \le c,\ |f|_{C^3} \le M_n \right\}.$$

Set $q = 1 - 4/m$, $p_0 = \mathrm{vol}(\mathcal{M})^{-1}$, and

$$A(m,k,n) = \frac{1}{4(m+2)^2} \left(\frac{k\,\Gamma(1+m/2)}{n \pi^{m/2}} \right)^{4/m}.$$

For every fixed positive $p \in C^3(\mathcal{M})$ with $\int_{\mathcal{M}} p \,\mathrm{dvol} = 1$,

$$\liminf_{n \to \infty} \frac{\sup_{f \in \mathcal{G}_n} \mathrm{ISB}(f, p; k, n)}{A(m,k,n)} \ge c^2 \|p^q\|_\infty.$$

For $p = p_0$, the normalized supremum converges to $c^2 p_0^q$. Moreover, every nonconstant p has a strictly larger asymptotic worst-case value. Thus p_0 is the unique pointwise asymptotic minimizer among these fixed densities.

Proof. Set $\rho_n = (k/n)^{1/m}$. Adding a constant to f changes neither its bias nor the constraints defining $\mathcal{G}_n$. We may therefore assume $\int_{\mathcal{M}} f \,\mathrm{dvol} = 0$. By compactness, $\|f\|_\infty \le C_{\mathcal{M}} \|\nabla f\|_\infty \le C_{\mathcal{M}} M_n$.

In the Taylor expansion used to prove Proposition H.1, the pointwise bias remainder is bounded by $C_p M_n \rho_n^3$, uniformly in $x \in \mathcal{M}$ and $f \in \mathcal{G}_n$. Moreover, the elliptic estimate $\|\nabla f\|_2 \le C_{\mathcal{M}} \|\Delta f\|_2$ bounds the L^2 norm of the leading bias uniformly on $\mathcal{G}_n$ by $C_p \rho_n^2$. Squaring and integrating thus gives, uniformly for $f \in \mathcal{G}_n$,

$$\mathrm{ISB}(f, p; k, n) = A(m,k,n) \int_{\mathcal{M}} \big(\Delta f + 2 \langle \nabla f, \nabla \log p \rangle_g \big)^2 p^q \,\mathrm{dvol} + O\big(M_n \rho_n^5 + M_n^2 \rho_n^6 \big).$$

Since $A(m,k,n) \asymp \rho_n^4$ and $M_n \rho_n \to 0$, the remainder is $o(A(m,k,n))$ uniformly on $\mathcal{G}_n$.

For p_0, $\nabla \log p_0 = 0$, so

$$\frac{1}{A(m,k,n)} \sup_{f \in \mathcal{G}_n} \mathrm{ISB}(f, p_0; k, n) \longrightarrow c^2 p_0^q.$$

The upper bound follows from $\|\Delta f\|_2 \le c$; equality in the limit is obtained by taking any fixed smooth nonconstant function and scaling it so that $\|\Delta f\|_2 = c$.

Fix p and $\varepsilon > 0$. Choose a coordinate ball U where $p^q \ge \|p^q\|_\infty - \varepsilon$, a nonzero $\chi \in C_c^\infty(U)$, and set

$$u_\ell(x) = \chi(x) \cos(\ell x^1), \qquad f_\ell = \frac{c}{\|\Delta u_\ell\|_2} \left(u_\ell - \mathrm{vol}(\mathcal{M})^{-1} \int_{\mathcal{M}} u_\ell \,\mathrm{dvol} \right).$$

Then $\|\Delta f_\ell\|_2 = c$ and $\|\nabla f_\ell\|_2 = O(\ell^{-1})$. The reverse triangle inequality therefore yields

$$\left[\int_{\mathcal{M}} \big(\Delta f_\ell + 2 \langle \nabla f_\ell, \nabla \log p \rangle_g \big)^2 p^q \,\mathrm{dvol} \right]^{1/2} \ge c \sqrt{\|p^q\|_\infty - \varepsilon} - \|p^q\|_\infty^{1/2} \, \|2 \nabla \log p\|_\infty \, \|\nabla f_\ell\|_2.$$

Each fixed f_ℓ belongs to $\mathcal{G}_n$ for all sufficiently large n. Therefore, for every fixed ℓ, the asymptotic expansion gives

$$\liminf_{n \to \infty} \frac{1}{A(m,k,n)} \sup_{f \in \mathcal{G}_n} \mathrm{ISB}(f, p; k, n) \ge \int_{\mathcal{M}} \big(\Delta f_\ell + 2 \langle \nabla f_\ell, \nabla \log p \rangle_g \big)^2 p^q \,\mathrm{dvol}.$$

Letting $\ell \to \infty$ in the preceding lower bound and then $\varepsilon \to 0$ yields

$$\liminf_{n\to\infty} \frac{1}{A(m,k,n)} \sup_{f\in\mathcal{G}_n} \mathrm{ISB}(f,p;k,n) \geq c^2 \|p^q\|_\infty.$$

Finally, since the average of p is p_0,

$$\|p^q\|_\infty \geq p_0^q.$$

The inequality is strict for every nonconstant p when $m \neq 4$.

When $m = 4$, both sides of the preceding inequality equal one, so the previous argument establishes optimality but does not distinguish the uniform density from nonuniform ones. Uniqueness can nevertheless be recovered through a separate argument exploiting the embedding $\mathcal{M} \subset \mathbb{R}^d$. Let p be nonuniform. We show that its leading worst-case ISB is strictly larger than that of the uniform density. Let $x_1, \ldots, x_d$ be the coordinate functions of $\mathbb{R}^d$ restricted to $\mathcal{M}$, and let $H = (\Delta x_1, \ldots, \Delta x_d)$ denote the mean-curvature vector of $\mathcal{M}$. Since $m = 4$, we have $p^q = 1$, and the leading ISB functional is proportional to

$$Q(f) \coloneqq \int_{\mathcal{M}} (\Delta f + 2\langle \nabla f, \nabla \log p \rangle_g)^2 \,\mathrm{dvol}.$$

We show that

$$\sup_{||\Delta f||_2 \leq c} Q(f) \geq c^2 \left(1 + \frac{4\int_{\mathcal{M}} ||\nabla \log p||_g^2 \,\mathrm{dvol}}{\int_{\mathcal{M}} ||H||^2 \,\mathrm{dvol}}\right) > c^2$$

by summing Q over the coordinate functions:

$$\sum_{j=1}^d Q(x_j) = \sum_{j=1}^d ||\Delta x_j||_2^2 + 4\int_{\mathcal{M}} \left\langle \textstyle\sum_{j=1}^d (\Delta x_j)\nabla x_j,\, \nabla \log p \right\rangle_g \,\mathrm{dvol}$$
$$+ 4\sum_{j=1}^d \int_{\mathcal{M}} \langle \nabla x_j, \nabla \log p\rangle_g^2 \,\mathrm{dvol}.$$

The second of the three terms vanishes identically: H is in the normal bundle of $\mathcal{M}$ by the Gauss formula, so its projection onto the tangent space $\sum_{j=1}^d (\Delta x_j)\nabla x_j$ is zero. The third term also simplifies: ∇x_j is the tangential projection of the jth standard basis vector, so $\sum_{j=1}^d \langle \nabla x_j, X\rangle_g^2 = ||X||_g^2$ pointwise for every tangent field X. If we take $X = \nabla \log p$, then our equality becomes

$$\sum_{j=1}^d \left(Q(x_j) - ||\Delta x_j||_2^2\right) = 4\int_{\mathcal{M}} ||\nabla \log p||_g^2 \,\mathrm{dvol} > 0.$$

Indices j for which $\Delta x_j \equiv 0$ contribute 0 to both sums; in this case, x_j is then harmonic, and hence constant on the closed connected $\mathcal{M}$, so $\nabla x_j = 0$. We can discard them; not all will be discarded, since $\mathcal{M}$ is not a point. Letting i be a non-discarded index, we apply the mediant inequality and the identity $\sum_{j=1}^d ||\Delta x_j||_2^2 = \int_{\mathcal{M}} ||H||_{\mathbb{R}^d}^2 \,\mathrm{dvol}$ and rearrange to obtain

$$\frac{Q(x_i)}{\|\Delta x_i\|_2^2} \geq \frac{\sum_{j=1}^d Q(x_j)}{\sum_{j=1}^d \|\Delta x_j\|_2^2} = 1 + \frac{4\int_{\mathcal{M}} \|\nabla \log p\|_g^2 \,\mathrm{dvol}}{\int_{\mathcal{M}} \|H\|_{\mathbb{R}^d}^2 \,\mathrm{dvol}}.$$

Choose $f = c\, x_i / \|\Delta x_i\|_2$. Then $\|\Delta f\|_2 = c$, and this fixed smooth function belongs to $\mathcal{G}_n$ for all sufficiently large n. The uniform asymptotic expansion, therefore gives

$$\liminf_{n\to\infty} \frac{\sup_{f\in\mathcal{G}_n} \mathrm{ISB}(f,p;k,n)}{A(m,k,n)} \geq c^2 \left(1 + \frac{4\int_{\mathcal{M}} \|\nabla \log p\|_g^2 \,\mathrm{dvol}}{\int_{\mathcal{M}} \|H\|_{\mathbb{R}^d}^2 \,\mathrm{dvol}}\right) > c^2.$$

□

Proposition H.3 (Unbounded Euclidean case)**.** *Let $\mathcal{M} = \mathbb{R}^m$, let p be a positive nonconstant smooth density, and retain q and $\mathcal{G}_n$ as above, with $\mathcal{M} = \mathbb{R}^m$. Then*

$$\lim_{n\to\infty} \sup_{f\in\mathcal{G}_n} \int_{\mathbb{R}^m} \left(\Delta f + 2\langle \nabla f, \nabla \log p\rangle\right)^2 p^q \, dx = \infty.$$

In particular, this holds for every nondegenerate Gaussian density.

Proof. Since p is nonconstant, there exists $x_0 \in \mathbb{R}^m$ such that $\nabla \log p(x_0) \neq 0$. Set

$$v = \nabla \log p(x_0).$$

By continuity and positivity of p,

$$\int_{\mathbb{R}^m} \langle v, \nabla \log p(x) \rangle^2 p(x)^q \, dx > 0.$$

For $R > 0$, set $f_R(x) = R\, v^\top x$. Then

$$\nabla f_R = Rv, \qquad \Delta f_R = 0, \qquad \|\Delta f_R\|_2 = 0, \qquad |f_R|_{C^3} = R\|v\|.$$

Hence, for every fixed R, $f_R \in \mathcal{G}_n$ for all sufficiently large n, since $M_n \to \infty$. Moreover,

$$\int_{\mathbb{R}^m} \big(\Delta f_R + 2\langle \nabla f_R, \nabla \log p \rangle\big)^2 p^q \, dx = 4R^2 \int_{\mathbb{R}^m} \langle v, \nabla \log p(x) \rangle^2 p(x)^q \, dx.$$

Therefore,

$$\liminf_{n\to\infty} \sup_{f \in \mathcal{G}_n} \int_{\mathbb{R}^m} \big(\Delta f + 2\langle \nabla f, \nabla \log p \rangle\big)^2 p^q \, dx \geq 4R^2 \int_{\mathbb{R}^m} \langle v, \nabla \log p(x) \rangle^2 p(x)^q \, dx$$

for every $R > 0$. Letting $R \to \infty$ proves the claim.

For completeness, if $p = \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$, then

$$\nabla \log p(x) = -\Sigma^{-1}(x - \mu).$$

Thus, for every $v \neq 0$, $\langle v, \nabla \log p \rangle$ is not identically zero, and the integral in the assumption is positive (possibly $+\infty$ when $q \leq 0$). Hence, every nondegenerate Gaussian satisfies the assumption. □