INFONCE INDUCES GAUSSIAN DISTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning has been at the bedrock of unsupervised learning in recent years, allowing training with massive unlabeled data for both task-specific and general (foundation) models. A prototypical loss in contrastive training is InfoNCE and its variants. In this paper we show that the embedding of the features which emerge from InfoNCE training can be well approximated by a multivariate Gaussian distribution. We justify this claim by taking two approaches. First, we show that under certain alignment and concentration assumptions, finite projections of a high dimensional representation approach multivariate Gaussian distribution, as the representation dimensions approach infinity. Next, under less strict assumptions, we show that adding a small regularization term (which vanishes asymptotically) that promotes low feature norm and high feature entropy, we reach similar asymptotic results. We demonstrate experimentally, in a synthetic setting, CIFAR-10 and on pretrained foundation models, that the features indeed follow almost precise Gaussian distribution. One can use the Gaussian model to easily derive analytic expressions in the representation space and to obtain very useful measures, such as likelihood, data entropy and mutual information. Hence, we expect such theoretical grounding to be very useful in various applications involving contrastive learning.

1 INTRODUCTION

Self-supervised learning with contrastive objectives has transformed modern representation learning, enabling scalable training of encoders without labels (Oord et al., 2018; Chen et al., 2020a; He et al., 2020; Radford et al., 2021). Among these objectives, the InfoNCE loss balances two pressures: positive pairs are aligned while the batch is repelled to encourage uniformity (Wang & Isola, 2020). This uniformity is often described geometrically as "spreading out" the data on the hypersphere (Chen & He, 2021), but a deeper probabilistic question remains: *What is the actual distribution of representations trained with InfoNCE?*

Answering this is not only of theoretical interest. A Gaussian characterization is directly motivated by recent empirical findings that "more Gaussian" representations can yield better downstream performance (Eftekhari & Papyan, 2025); it also provides a principled justification for practical methods that already model contrastive representations as Gaussians for classification, uncertainty estimation, prompt learning, and test-time adaptation (Baumann et al., 2024; Morales-Álvarez et al., 2024; Lu et al., 2022). Moreover, assuming Gaussian structure makes entropy, likelihoods, Mahalanobis distances, and KL divergences available in closed form, which underpins many OOD, calibration, and density-based diagnostics (Lee et al., 2018; Tosh et al., 2021). These benefits are already exploited in applied work, but largely without a unifying theoretical foundation (Betser et al., 2025; Fort et al., 2021). At the same time, numerous empirical studies report that contrastive features and their low-dimensional projections are close to Gaussians (Tian et al., 2020a; Chen et al., 2020b; Bardes et al., 2022), motivating a precise population-level explanation of why such Gaussian structure emerges in the first place.

Analyzing the *population* InfoNCE objective, we identify that under different assumptions the population law becomes isotropic and yields Gaussian low-dimensional projections. First, an alignment-plateau assumption reduces training to a constrained uniformity problem on $\mathbb{S}^{d-1}$. With a norm-concentration assumption both the normalized (to unit norm) and unnormalized representations have Gaussian projections (Sec. 4.1); Finally, under the weaker "attainable alignment at uniformity" assumption, the same asymptotic conclusion holds: adding a small convex regularizer selects

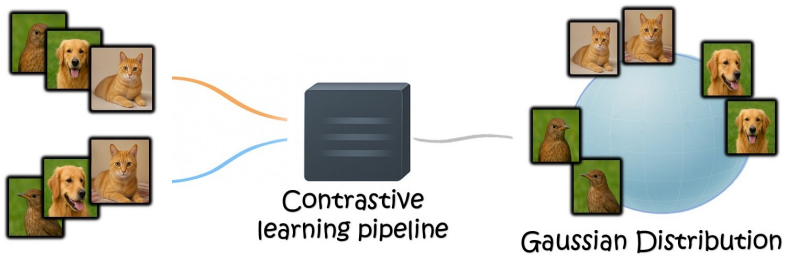

Figure 1: **Illustration.** Contrastive learning yields (approximately) Gaussian representations.

the uniform solution, and the required regularization vanishes as the dimension grows (Sec. 4.2). We validate our main conclusions on different data types (synthetic, CIFAR-10 (Krizhevsky et al., 2009) and MS-COCO (Lin et al., 2014)) and different models (linear encoder, MLP + activations, CLIP (Radford et al., 2021) and Dino Caron et al. (2021)).

Our main contributions are:

- **Bounded alignment.** In the infinite-negatives regime, alignment is upper-bounded by augmentation mildness, with additional dependence on the mean representation.
- **Uniformity on the sphere.** Under either of two alignment related assumptions, the law of the normalized representations asymptotically converges to the uniform law on the sphere.
- **Gaussian projections.** We show that fixed-k projections are asymptotically Gaussian for normalized representations. This includes any subset of k coordinates, considered separately or jointly. For the unnormalized representations the same holds.
- **Practical guidance.** We suggest that a small regularizer can steer representations toward isotropy and Gaussianity in practice.
- **Empirical evidence.** The main assumptions of our study are validated empirically. In addition, we demonstrate the emerging Gaussian statistics, as dimension grows, on synthetic and real data.

2 Related Work

Contrastive learning and InfoNCE. The InfoNCE loss (Oord et al., 2018) is the standard objective in self-supervised representation learning, underpinning frameworks such as SimCLR (Chen et al., 2020a), CLIP (Radford et al., 2021), and DINO (Caron et al., 2021). It balances two forces: alignment of positive pairs and batch-wise repulsion that encourages uniformity (Wang & Isola, 2020; Chen & He, 2021). Related concentration phenomena have also been documented (Caron et al., 2021; Draganov et al., 2025). Yet, despite these insights, the *probabilistic law* governing the representations remains unclear. In particular, there is little theoretical understanding of the distributional geometry of the raw, *unnormalized* representations, specifically their asymptotic laws. The direct encoder outputs are used in downstream tasks (Fort et al., 2021), such as image synthesis (Ramesh et al., 2022), conformity quantification (Levi & Gilboa, 2025), and representation regularization (Bardes et al., 2022). Understanding their distribution is essential for improving downstream applications. Some empirical studies observe that representations are approximately Gaussian (Tian et al., 2020a; Fort et al., 2021), but lack an explanatory theory. Our work fills this gap, providing a thorough characterization of the asymptotic laws of both normalized and unnormalized representations.

A complementary line of work studies *identifiability* of representations, including in the context of contrastive objectives. These works analyze which structural assumptions on the data-generating model ensure that its latent variables can be uniquely recovered (up to symmetries) by optimizing a contrastive loss (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Zimmermann et al., 2021; Roeder et al., 2021). In this setting, the focus is on when class-conditional or component-level structure remains identifiable in the learned representation. A second line of work analyzes contrastive learning from a *task-driven* perspective, showing when representations become linearly separable or cluster according to class labels (Saunshi et al., 2019; HaoChen et al., 2021); these results characterize the *class-conditional geometry* of the embeddings. By contrast, our work does not study recovery or class structure. Instead,

we analyze the *marginal* distribution induced by the population InfoNCE functional and show that its minimizers follow a strongly Gaussian law. This concerns the *overall* embedding distribution, aggregated over all samples, and is orthogonal to whether mixture components or semantic clusters remain distinguishable.

Regularization and design choices can promote isotropic, near-Gaussian representations, for instance via whitening-style objectives, neural collapse, or Gaussian-mixture structure (Ermolov et al., 2021; Papyan et al., 2020; Fort et al., 2021), with related NLP work (Zhuo et al., 2023). We introduce a light convex regularizer that biases training toward isotropy at finite d yet vanishes in the high-dimensional limit, providing a lightweight mechanism that complements InfoNCE's natural Gaussianization.

Mathematical tools. Independently of contrastive learning, a classical line of work studies high-dimensional uniform measures on the sphere and their connection to Gaussians, sometimes referred to as the "soap-bubble effect" (Vershynin, 2018; Wegner, 2021). Similar geometry is exploited in hyperspherical variational families and radial Bayesian priors (Davidson et al., 2018; Farquhar et al., 2020), which exploit the geometry of (approximately) uniform hyperspherical distributions. These results are not specific to InfoNCE, but they provide geometric intuition for why spherical uniformity and Gaussian structure are closely related.

We draw on classical tools from probability, spherical harmonic analysis, and information theory: (i) *Maximal correlation (Hirschfeld-Gebelein-Rényi, HGR)* and its data-processing inequality, which upper-bound the alignment achievable under augmentations (Hirschfeld, 1935; Gebelein, 1941; Rényi, 1959); (ii) a *polar (radial-angular) chain rule for KL divergence*, which separates angular regularization from radial structure (Dupuis & Ellis, 2011; Cover, 1999); and (iii) the *Maxwell-Poincaré spherical CLT*, yielding Gaussian fixed-k projections for the uniform law on $\mathbb{S}^{d-1}$ (Maxwell, 1860; Poincaré, 1912; Diaconis & Freedman, 1987). While uncommon in latent-space analysis, these tools are particularly useful for our setting.

3 Setup

Data domain. Let $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be a standard Borel space (a standard setting in probability) with a base probability p_{base}. We draw $X_0 \sim p_{\text{base}}$ as a single data item (e.g., an image).

Pairs via augmentation. Contrastive learning is built around pairs of related examples rather than individual samples. To form such pairs, we use an *augmentation channel* $\mathcal{A}$, which takes a base sample $X_0 \sim p_{\text{base}}$ and produces stochastic variations of it. Formally, given X_0, we draw two independent augmentations

$$X, Y \;\sim\; \mathcal{A}(\cdot \mid X_0). \tag{1}$$

Here X and Y are two different "views" of the same underlying example X_0 (e.g., two differently cropped or color-jittered images). We denote by p_X the marginal distribution of a single augmentation X, and by p_{XY} the joint distribution of a pair (X, Y). It is assumed that p_X is nonatomic (achievable in practice by infinitesimal dither if needed).

InfoNCE loss. Let $f : \mathcal{X} \to \mathbb{R}^d$, $d \geq 2$, be a Borel-measurable encoder that maps inputs to representations. Training is performed using the InfoNCE loss, which operates on *normalized* representations:

$$\hat{f}(x) := \begin{cases} f(x)/\|f(x)\|, & \|f(x)\| > 0, \\ c_0, & \|f(x)\| = 0, \end{cases} \qquad c_0 \in \mathbb{S}^{d-1} \text{ fixed}, \tag{2}$$

where c_0 is arbitrary. Given a minibatch of N paired augmentations $\{(x_i, y_i)\}_{i=1}^N$ drawn i.i.d. from p_{XY}, let $u_i := \hat{f}(x_i)$ and $v_i := \hat{f}(y_i)$. The empirical InfoNCE loss is

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\frac{1}{\tau} \langle u_i,\, v_i \rangle\right)}{\sum_{j=1}^{N} \exp\left(\frac{1}{\tau} \langle u_i,\, v_j \rangle\right)}, \tag{3}$$

with a fixed temperature $\tau > 0$. Since u_i and v_j are unit-normalized, $\langle u_i, v_j \rangle$ equals cosine similarity. The numerator measures the similarity of the *positive* pair (u_i, v_i). The denominator compares each

anchor u_i to all candidates $\{v_j\}_{j=1}^N$, where $j \neq i$ serve as *negatives*. This softmax encourages u_i to rank its true partner highest while remaining distinct from negatives, preventing collapse.

Population InfoNCE. The empirical InfoNCE loss in (Eq. 3) depends on the batch size N. As $N \to \infty$, the empirical averages converge to expectations. Let

$$\mu := \hat{f}_* p_X, \qquad \pi := (\hat{f}, \hat{f})_* p_{XY}, \tag{4}$$

be the marginal distribution of representations and the joint distribution of positive pairs, respectively. Here $\hat{f}_* p_X$ denotes the *pushforward measure* of p_X by $\hat{f}$, which is the distribution of $\hat{f}(X)$. As shown by Wang & Isola (2020, Theorem 1, Eq. (2)), in the infinite-negatives limit $N \to \infty$ the empirical InfoNCE loss (up to the additive $\log N$ term) converges to the following population functional. With $\alpha = 1/\tau$ for fixed $\tau > 0$:

$$\mathcal{L}(\mu, \pi) \;=\; -\alpha\, \mathbb{E}_{(u,v)\sim\pi}[u \cdot v] \;+\; \Phi(\mu), \qquad \Phi(\mu) := \mathbb{E}_{u\sim\mu} \log \mathbb{E}_{v\sim\mu} \exp\big(\alpha\, u \cdot v\big). \tag{5}$$

The first term measures *alignment* of positive pairs, while the second is a *uniformity potential* depending only on μ.

3.1 Augmentation mildness.

We now introduce a new term which quantifies the degree of augmentation. The augmentation channel $\mathcal{A}$ limits how much positive-pair alignment can be induced. We quantify this with the *augmentation mildness* parameter

$$\eta_2 \;:=\; \sup_{\substack{g \in L^2(p_X) \\ \mathrm{Var}(g) > 0}} \frac{\mathrm{Var}\big(\mathbb{E}[g(X) \mid X_0]\big)}{\mathrm{Var}(g(X))} \;\in [0, 1], \tag{6}$$

which measures how predictable functions of the view X are from the base X_0. This quantity equals the squared Hirschfeld-Gebelein-Rényi (HGR) maximal correlation between X and X_0 (Appendix A.1). Intuitively, $\eta_2 = 0$ when X is (effectively) independent of X_0 (very strong/noisy augmentations), and $\eta_2 = 1$ when X is fully determined by X_0 (no augmentation noise).

Example. Consider the Gaussian channel $X = AX_0 + \sqrt{1 - A^2}\,\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ is independent of $X_0 \sim \mathcal{N}(0, 1)$. Since X and X_0 are jointly Gaussian with Pearson correlation A, the maximal correlation is $\rho_m(X, X_0) = |A|$, and hence $\eta_2 = A^2$ (more details in Appendix A.2).

Proposition 1 (Augmentation-controlled alignment bound)**.** *Let $X, Y \sim \mathcal{A}(\cdot \mid X_0)$ be conditionally independent given the base sample X_0, and let $u = \hat{f}(X)$, $v = \hat{f}(Y)$ be normalized representations in $\mathbb{S}^{d-1}$, i.e., $\|u\| = \|v\| = 1$. Then*

$$\mathbb{E}_{(u,v)\sim\pi}[u \cdot v] \;\leq\; \eta_2 \;+\; (1 - \eta_2)\, \|m(\mu)\|^2, \qquad m(\mu) := \mathbb{E}[u] = \mathbb{E}[v], \tag{7}$$

where $\eta_2 = \rho_m^2(X, X_0)$ is the squared HGR maximal correlation between the view and the base, and μ is the marginal law of u.

The proof appears in Appendix A.3. This upper bound links the alignment of positive pairs to the structure of the augmentation channel via maximal correlation. While HGR is well established in statistical dependence analysis (Huang & Xu, 2020; Zhang et al., 2024), it has not previously been used to control alignment in contrastive learning. Existing works (e.g., Tian et al. (2020b)) analyze augmentations empirically, but do not derive bounds of this form.

4 Gaussianity from InfoNCE

We study why minimizing the population InfoNCE objective (Eq. 5) yields (approximately) Gaussian low-dimensional projections of learned representations, for both *normalized* representations on the sphere and *unnormalized* representations in $\mathbb{R}^d$. Our analysis proceeds along two complementary routes, which differ in the strength of the assumptions they require.

Empirical idealization. We first analyze an idealized regime with infinite data, ambient dimension $d \to \infty$, and sufficient optimization. Guided by empirical observations, we assume an *alignment*

plateau and *thin-shell concentration*; these assumptions enable a simple derivation of Gaussian projections.

Regularized route. To weaken the assumptions, we then study a regularized variant of the population objective. Replacing exact plateau behavior with a milder alignment assumption and introducing a vanishing convex regularizer ensures a unique minimizer and again yields Gaussian low-dimensional projections. This route requires strictly weaker assumptions than the empirical idealization.

4.1 Gaussian projections at alignment plateau

We work in the population setting (Eq. 5) with positive pairs as defined earlier.

Assumption 1 (Alignment plateau)**.** *After sufficient training, the positive-pair alignment saturates at a ceiling; concretely,*

$$\mathbb{E}_{(u,v)\sim\pi}[u\cdot v] \;=\; \eta_2 \;+\; r_{\mathrm{plat}}, \tag{8}$$

where $r_{\mathrm{plat}} \le 0$ is a constant error term representing the difference between the alignment value at plateau and the maximal correlation defined by the augmentations (η_2).

Empirically, alignment saturation has been reported in some contrastive-learning settings (Wang & Isola, 2020; Fang et al., 2024), which motivates considering a plateau model as a plausible scenario rather than a universal requirement. In our experiments (Fig. 2, Appendix Figs. 7, 8), we frequently observe high alignment alongside improving uniformity with larger dimensions and batch sizes, suggesting that alignment may saturate before uniformity in at least some regimes. An extension that places the plateau at the alignment bound (Eq. 7) is discussed in Appendix D.

Corollary 1 (Gaussian k-projections at the plateau)**.** *Suppose the alignment plateau condition (Eq. 8) holds, and consider the population objective (Eq. 5). Let μ^* denote the global minimizer supported on $\mathbb{S}^{d-1}$. Then, in the limit $d \to \infty$, for every fixed $k \ge 1$ the k-dimensional marginal of $u \sim \mu^*$ satisfies*

$$\sqrt{d}\, u_k \;\Rightarrow\; \mathcal{N}(0, I_k), \tag{9}$$

where u_k denotes the projection of u onto a fixed k-dimensional coordinate subspace and I_k is the $k \times k$ identity matrix.

The proof is provided in Appendix C.1 and follows from two lemmas. The first establishes that $\Phi(\mu)$ attains a global minimum at the uniform law (Wang & Isola, 2020), while the second invokes the central limit theorem on the sphere (Diaconis & Freedman, 1987) to deduce Gaussian projections.

4.1.1 Gaussian projections for unnormalized representations.

So far we analyzed normalized representations on the sphere. We now extend the result to the original, unnormalized encoder outputs $z = f(X) \in \mathbb{R}^d$. Write $z = ru$, where $r = \|z\|$ is the representation radius and $u = z/\|z\| \in \mathbb{S}^{d-1}$ the normalized direction (Eq. (Eq. 2)).

Assumption 2 (Thin-shell concentration)**.** *We assume the representation radius concentrates in a thin shell:*

$$\frac{r}{r_0} \;\xrightarrow[d\to\infty]{}\; 1, \qquad r_0 \in (0,\infty). \tag{10}$$

Norm concentration is widely observed in contrastive learning: unnormalized representations cluster around a characteristic radius (Wang & Isola, 2020; HaoChen et al., 2021; Levi & Gilboa, 2025; Betser et al., 2025). This thin-shell effect is further promoted by weight decay, which penalizes norm growth and stabilizes a common scale. In particular, Draganov et al. (2025) show that appropriate weight decay suppresses norm inflation and tightens the dispersion of representation norms, lending empirical support to Assumption 2. Consistent with these reports, our experiments exhibit progressively sharper radius histograms as dimension and batch size increase (Figs. 3, 4, 10).

Proposition 2 (Gaussian projections for unnormalized representations)**.** *Let $z = f(x) \in \mathbb{R}^d$ be the unnormalized representation $u := z/\|z\|$. Assume $u \sim \sigma$ (the uniform distribution on $\mathbb{S}^{d-1}$) and that*

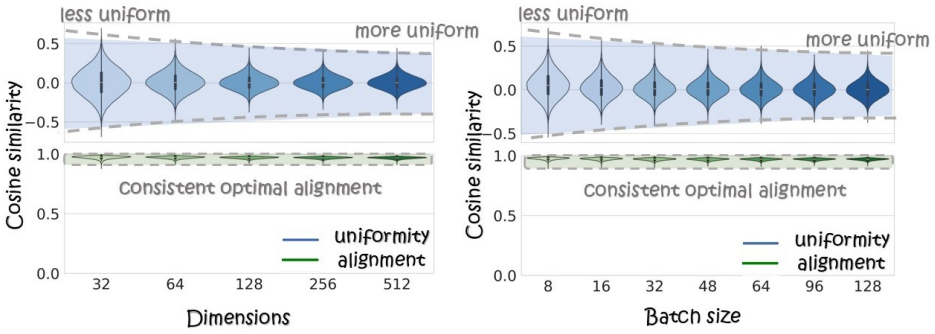

Figure 2: **Uniformity vs. alignment across settings.** A simple encoder trained on synthetic Laplace data exhibits (i) near-optimal alignment across all configurations and (ii) steadily improving uniformity as batch size and dimensionality grow.

Assumption 2 holds, i.e., $r \xrightarrow[d\to\infty]{} r_0 \in (0,\infty)$. *Then for any fixed* k*-dimensional subspace,*

$$\sqrt{d}\, z_k \;\Rightarrow\; \mathcal{N}\big(0,\, r_0^2 I_k\big) \qquad (d \to \infty), \tag{11}$$

where z_k *denotes the orthogonal projection of* z *onto that subspace and* I_k *is the* $k \times k$ *identity.*

See proof in Appendix C.2.

4.2 Gaussian projections using regularization

We now relax the two separate assumptions from the previous section and replace them with a single, weaker requirement regarding the achievable alignment with a uniform marginal (Assumption 3). We work in a regularized setting, where the regularization vanishes as $d \to \infty$. In a way that will be made precise, we show that the uniform distribution gets arbitrarily close to optimality or even reaches optimality. As before, this has direct implications to its low-dimensional projections, which are approximately Gaussian (Theorem 2). This result shows that Gaussianity can be obtained without relying on the stronger thin-shell or plateau conditions.

We constrain f to take values in $B \subseteq \mathbb{R}^d$, which is either some closed ball centered at 0 with positive radius or $\mathbb{R}^d$. We take the original loss and add two new losses: one to penalize large squared norms, and the other to encourage high entropy (We comment that both are commonly regarded as desirable goals, irrespective of our setup). Specifically, for fixed $\beta, \lambda > 0$,

$$J(f) = \Phi(\mu) - \alpha \mathbb{E}_{(u,v)\sim\pi}[u \cdot v] + \beta(-H(\rho) + \lambda \mathbb{E}_{Z\sim\rho}\|Z\|^2)\,, \tag{12}$$

where $\rho = f_* p_X$ is the unnormalized pushforward probability. Define the truncated Gaussian γ_λ^B,

$$\gamma_\lambda^B(dz) = c_{B,\lambda} e^{-\lambda\|z\|^2} \mathbf{1}_B(z) dz\,, \qquad c_{B,\lambda}^{-1} = \int_B e^{-\lambda\|z\|^2}\, dz\,. \tag{13}$$

If $\rho \ll \gamma_\lambda^B$, then

$$\mathrm{KL}(\rho\|\gamma_\lambda^B) = \int \log \frac{d\rho}{dz}\, d\rho - \int \log \frac{d\gamma_\lambda^B}{dz}\, d\rho = -H(\rho) \;+\; \lambda\, \mathbb{E}_\rho\|Z\|^2 \;+\; \log c_{B,\lambda}^{-1}\,, \tag{14}$$

that is, equality up to an additive constant. Since $\rho(B) = 1$, if $\rho \not\ll \gamma_\lambda^B$, then both $\mathrm{KL}(\rho\|\gamma_\lambda^B)$ and $-H(\rho)$ are $+\infty$. Thus, it is equivalent to minimize

$$J(f) = \Phi(\mu) - \alpha \mathbb{E}_{(u,v)\sim\pi}[u \cdot v] + \beta \mathrm{KL}(\rho\|\gamma_\lambda^B)\,, \tag{15}$$

and we thereby also implicitly restrict ρ to satisfy $\rho \ll \gamma_\lambda^B$ and in particular $\rho(B) = 1$.

Our goal is to prove that for $\beta \geq \beta_0$, taking the angular probability as σ approaches optimality and the optimal radial probability is that of γ_λ^B. If $B = \mathbb{R}^d$, this means that a Gaussian ρ approaches optimality. Furthermore, as $d \to \infty$, $\beta_0 \to 0$.

This will be done in several steps. First, ρ can be decomposed into a radial part and an angular part. We show that the radial part can be chosen optimally in a straightforward way.

Proposition 3. *Let $\rho(dz) = \mu(du)\kappa(dr \mid u)$ and $\gamma_\lambda^B(dz) = \sigma(du)\xi(dr \mid u)$ in polar coordinates $z = ru$. Then $\kappa = \xi$ is an optimal choice, yielding $\mathrm{KL}(\rho\|\gamma_\lambda^B) = \mathrm{KL}(\mu\|\sigma)$.*

The proof is given in Appendix B.1. The above proposition reduces the optimization problem for unnormalized embedding to normalized embeddings only. It also describes an optimal probability for embedding norms, in contrast to the original InfoNCE loss, which is completely oblivious to embedding norms.

It is important to note that because we are working with a standard Borel space with a nonatomic p_X, any probability $\rho \in \mathcal{P}(B)$ has $\rho = g_* p_X$ for some encoding g. In addition, any $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ has $\mu = h_* p_X$ for some encoding, and since B contains a ball around 0, there is an encoding f s.t $h = \hat{f}$. Thus we can legitimately speak about "choosing" ρ or μ, since suitable encodings exist that induce them. In addition, we may also define:

Definition 1. *For every $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$,*

$$\mathrm{Align}(\mu) = \sup_f \left\{ \mathbb{E}[\hat{f}(X) \cdot \hat{f}(Y)] : f \text{ measurable}, \ (\hat{f})_* p_X = \mu \right\} , \tag{16}$$

As was noted, the supremum is always taken on a nonempty set. We can write

$$\tilde{J}(\mu) = \Phi(\mu) - \alpha \mathrm{Align}(\mu) + \beta \mathrm{KL}(\mu\|\sigma) , \tag{17}$$

and it holds that $\inf_{\{\hat{f}: \hat{f}_* p_X = \mu\}} J(f) = \tilde{J}(\mu)$, and consequently $\inf_f J(f) = \inf_{\mu \in \mathcal{P}(\mathbb{S}^{d-1})} \tilde{J}(\mu)$.

The reason is that $\mathrm{Align}(\mu)$ can be approximated arbitrarily well by an encoding, and the KL divergence is optimized by taking the radial distribution given in Proposition 3. We can therefore focus on optimizing $\tilde{J}(\mu)$.

The assumption for which we will prove our result is the following:

Assumption 3. *It holds that $\alpha(\eta_2 - \mathrm{Align}(\sigma)) \xrightarrow{d\to\infty} 0$.*

We will require one more technical lemma before proceeding to prove the result.

Lemma 1. *If $d \geq 2$, then $\mathrm{KL}(\mu\|\sigma) \geq C(d-1)\|m(\mu)\|^2$, where $C > 0$ is a universal constant.*

Proof is provided in Appendix B.2. To understand the constant, see (Vershynin, 2018, Proposition 2.6.1).

Theorem 1. *Let $d \geq 2$. There is a universal constant $C > 0$ s.t. for $\beta \geq \beta_0 = \frac{\alpha(1-\eta_2)}{C(d-1)}$,*

- *Under Assumption 3, $\tilde{J}(\sigma) - \inf_\mu \tilde{J}(\mu) \xrightarrow{d\to\infty} 0$.*
- *Assuming further that $\mathrm{Align}(\sigma) = \eta_2$ yields that $\tilde{J}(\sigma) = \min_\mu \tilde{J}(\mu)$.*

Proof. Write $\delta(d) = \eta_2 - \mathrm{Align}(\sigma)$. For every μ, we have that $\Phi(\mu) - \Phi(\sigma) \geq 0$ (Wang & Isola, 2020, Theorem 1). In addition,

$$\mathrm{Align}(\mu) - \mathrm{Align}(\sigma) \leq \eta_2 + (1-\eta_2)\|m(\mu)\|^2 - (\eta_2 - \delta(d)) = (1-\eta_2)\|m(\mu)\|^2 + \delta(d) \tag{18}$$

by Proposition 1. Lastly,

$$\mathrm{KL}(\mu\|\sigma) - \mathrm{KL}(\sigma\|\sigma) = \mathrm{KL}(\mu\|\sigma) \geq C(d-1)\|m(\mu)\|^2 \tag{19}$$

by Lemma 1. Therefore,

$$\begin{aligned} \tilde{J}(\mu) - \tilde{J}(\sigma) &= (\Phi(\mu) - \Phi(\sigma)) - \alpha(\mathrm{Align}(\mu) - \mathrm{Align}(\sigma)) + \beta(\mathrm{KL}(\mu\|\sigma) - \mathrm{KL}(\sigma\|\sigma)) \\ &\geq -\alpha(1-\eta_2)\|m(\mu)\|^2 - \alpha\delta(d) + \beta C(d-1)\|m(\mu)\|^2 \\ &= (-\alpha(1-\eta_2) + \beta C(d-1))\,\|m(\mu)\|^2 - \alpha\delta(d) \geq -\alpha\delta(d) , \end{aligned} \tag{20}$$

where the last inequality is by the choice of β.

If we assume that $\alpha\delta(d) \xrightarrow{d\to\infty} 0$, then $\tilde{J}(\sigma) - \inf_\mu \tilde{J}(\mu) \leq \alpha\delta(d)$, so $\tilde{J}(\sigma) - \inf_\mu \tilde{J}(\mu) \xrightarrow{d\to\infty} 0$.

If we assume further that $\mathrm{Align}(\sigma) = \eta_2$, then $\delta(d) = 0$, and since $\tilde{J}(\sigma) \leq \tilde{J}(\mu)$ for every μ, $\tilde{J}(\sigma) = \min_\mu \tilde{J}(\mu)$, completing the proof. □

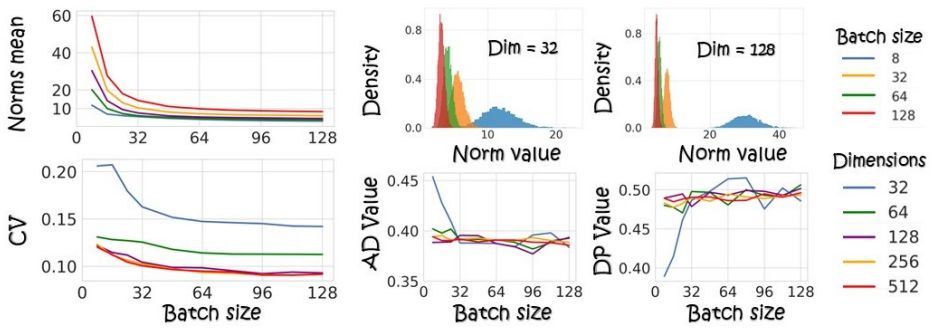

Figure 3: **Synthetic data experiments.** Left: radius statistics vs. batch size (curves: representation dimension) showing thin-shell concentration strengthening with d and N. Top middle/right: norm histograms across batch sizes illustrating radius tightening. Bottom: normality diagnostics - AD (lower is better; normality if < 0.752) and DP (higher p is better; normality if $p > 0.05$) - with averages comfortably in the normal range and 100% per-coordinate compliance.

Since the optimal radial component of the distribution is known, we can draw conclusions w.r.t. ρ as well. For example, we can directly obtain the following corollary.

Corollary 2. *Let $B = \mathbb{R}^d$ ($d \geq 2$) and $\beta \geq \beta_0$. If $\mathrm{Align}(\sigma) = \eta_2$, where σ is the uniform distribution on $\mathbb{S}^{d-1}$ and η_2 is the augmentation mildness, then $\mathcal{N}(0, (2\lambda)^{-1} I_d)$ is an optimal choice for ρ.*

5 Experiments

We conduct experiments under three different regimes: (i) synthetic data with a simple linear encoder, (ii) the CIFAR-10 dataset with both an MLP + activation encoder and a SimCLR-style contrastive encoder, and (iii) pretrained models, including several foundation-scale encoders. In all settings, we evaluate both *normalized* and *unnormalized* representations, mirroring our theoretical analysis. The experiments are designed to validate the assumptions underlying our analysis and to illustrate the emergence of Gaussian behavior in both regimes. We observe stable trends across runs, and all figures are shown for a representative seed. Full implementation details appear in Appendix E.1.

Metrics. We monitor norm concentration via the coefficient of variation (CV) of the representation norms:

$$\mathrm{CV} = \frac{\mathrm{std}(\|z_i\|)}{\mathrm{mean}(\|z_i\|)}. \tag{21}$$

A low CV indicates a tightened norm distribution and is consistent with thin-shell behavior. To assess Gaussianity of individual coordinates, we apply two standard one-dimensional normality tests: (i) the Anderson–Darling (AD) test (Anderson & Darling, 1954), where AD < 0.752 indicates failure to reject the null hypothesis of Gaussianity, and (ii) the D'Agostino–Pearson (DP) test (D'agostino & Pearson, 1973), where $p > 0.05$ indicates failure to reject the null. In both cases, the null hypothesis is that *each coordinate is Gaussian*; the alternative is that it is non-Gaussian. Taken together, these coordinate-wise tests and the global CV measure play complementary roles: AD/DP probe marginal normality of individual coordinates, while CV probes the high-dimensional radial law through norm concentration. This combination provides a strong finite-sample indicator of approximate Gaussianity and effectively rules out natural alternatives such as Student-t, Laplace, or Gaussian mixture distributions, which would typically fail at least one of these diagnostics.

Synthetic datasets. To validate our diagnostics in controlled settings, we evaluate two synthetic families: (i) a Laplace(0,1) distribution in $\mathbb{R}^{1024}$, and (ii) a Gaussian mixture with 25 equally weighted components (each with random means), also in $\mathbb{R}^{1024}$. Each dataset contains $10k$ samples, and we train linear encoders with varying representational dimensions and batch sizes. Figure 3 (left) shows that for the Laplace case the representation norms tighten as batch size (x-axis) and dimensionality

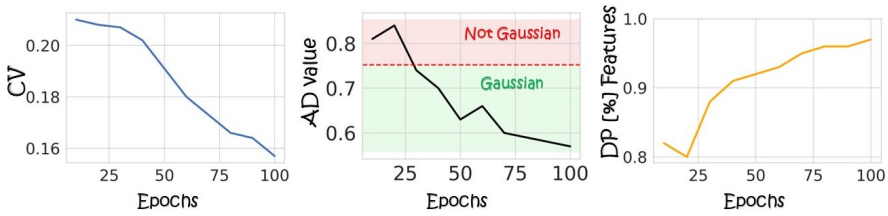

Figure 4: **CIFAR-10 training dynamics.** A two-layer MLP trained with InfoNCE on CIFAR-10 exhibits increasing Gaussianity over training. Left: representation norms concentrate as indicated by declining CV, (Eq. (Eq. 21)). Middle: the AD statistic decreases from non-Gaussian levels into the normal range. Right: the fraction of coordinates passing the DP normality test rises steadily.

(colored curves) increase: the *mean* norm decreases and the norms concentrate, as quantified by the decreasing CV. This monotonic decrease in CV is consistent with the thin-shell behavior predicted by our asymptotic analysis. Histograms of $\|z_i\|$ for different batch sizes further illustrate the emergence of norm concentration. We repeat the experiment on the synthetic *25-component Gaussian mixture*. Despite the clear non-Gaussian structure in the input space, the learned *marginal* representations again display strong Gaussian signatures. As reported in Table 1, all coordinates pass the AD and DP normality thresholds, and the CV values are low. Thus, even when the underlying data-generating process is a mixture, the resulting representations remain approximately Gaussian at the marginal level, consistent with our theoretical characterization of the population InfoNCE minimizers.

CIFAR-10. We train a two-layer MLP with a single ReLU nonlinearity using the InfoNCE objective on CIFAR-10. At regular intervals we evaluate on the test set, tracking radius concentration via the coefficient of variation (CV) and normality via AD and DP. Fig. 4 reports: (left) norms concentrate over training, CV declines; (middle) the AD statistic drops from a non-Gaussian level into the normal range; and (right) the fraction of coordinates passing normality (DP $p > 0.05$) increases steadily. This experiment shows how norm concentration and Gaussianity of representations increase as training proceeds. Overall, both thin-shell concentration and Gaussianity strengthen as training progresses.

Table 1: **Gaussianity diagnostics across datasets and training regimes.** Rows report five metrics: norm concentration via the coefficient of variation (**CV**, Eq. 21), and two normality test: Anderson-Darling (**AD**) and D'Agostino-Pearson (**DP**), summarized by the average test statistic (**Avg.**) and the percentage of coordinates whose statistics fall in the Gaussian-acceptance region (**Norm. Feat.**). Columns correspond to different data sources and training configurations: synthetic Laplace and Gaussian-mixture inputs (linear encoder), multiple CIFAR-10 regimes (supervised vs. contrastive, low/high augmentation, no/high weight decay, all with ResNet-18 encoder), and two ImageNet-R variants (Sketch, Painting, encoded with CLIP). Results are of the unnormalized embeddings.

Metric	**CIFAR-10**						**Synthetic**		**ImageNet-R**	
	Sup.	Contrastive					Contrastive		Contrastive	
	Sup.	Contr.	Low Aug.	High Aug.	No WD	High WD	Laplace	Mix	Sketch	Painting
CV	0.50	0.09	0.12	0.13	0.09	0.10	0.08	0.08	0.14	0.14
AD Avg.	3.3	0.43	0.39	0.49	0.41	0.42	0.38	0.39	0.41	0.40
AD Norm. Feat.	6.2%	96.1%	93.7%	92.1%	94.5%	93.7%	100%	100%	94.8%	95.3%
DP Avg.	0.041	0.39	0.46	0.32	0.46	0.45	0.49	0.46	0.44	0.43
DP Norm. Feat.	3.9%	94.5%	93.7%	91.5%	92.1%	91.5%	100%	100%	93.3%	94.2%

InfoNCE vs. Supervised training. We use the CIFAR-10 dataset and ResNet-18 (He et al., 2016) for a controlled comparison between supervised and contrastive learning. We use an initialized ResNet-18 model in both cases, with a 2 layer MLP (following SimCLR (Chen et al., 2020a) setting). In Table 1, we show that supervised training does not induce any Gaussianity while contrastive learning does. We also add ablations on augmentation strength (low/high, regular experiment is standard augmentations) and weight decay strength (none/strong ($1e-3$), regular experiment is standard - $1e-4$). Results show that in all settings the representations are approximately Gaussian, while stronger augmentations lead to lower alignment values and no weight decay leads to high norm values.

Table 2: **Normality test scores for pretrained models.** Each cell shows *Unnormalized / Normalized*. The *Avg.* column contains the average score for all features, and *Norm. Feat.* represents the percentage of features passing the normal distribution test. Thresholds are indicated in brackets, with the sign showing whether higher or lower results imply normality.

		Anderson-Darling (< 0.752)		**D'Agostino-Pearson** (> 0.05)	
		Avg.	**Norm. Feat.**	**Avg.**	**Norm. Feat.**
Self-supervised	**CLIP Img**	0.4749 / 0.4917	96.8% / 96.0%	0.4163 / 0.3988	99.6% / 99.4%
	CLIP Txt	0.5345 / 0.5368	94.0% / 93.6%	0.3775 / 0.3773	99.4% / 99.7%
	Dino	0.4415 / 0.4400	97.0% / 97.1%	0.4533 / 0.4544	99.2% / 99.3%
Supervised	**ResNet**	10.01 / 9.638	0.0% / 0.0%	2.2×10^{-6} / 3.2×10^{-6}	0.0% / 0.0%
	DenseNet	2.982 / 2.8538	42.2% / 41.6%	0.1550 / 0.1442	49.3% / 49.0%

Pretrained models. We generalize our evaluations to supervised and self-supervised vision backbones to assess whether Gaussian structure appears across common representation-learning paradigms, not only in the unimodal InfoNCE settings. Our supervised baselines are ResNet34 He et al. (2016) and DenseNet Huang et al. (2017), pretrained on ImageNet-1k Deng et al. (2009). Our self-supervised models include CLIP (Radford et al., 2021) (ViT-L/14) and DINO (Caron et al., 2021) (ViT-B/16). Although CLIP and DINO are not pure instances of the unimodal population InfoNCE objective they remain dominant SSL approaches and provide a natural testbed for examining whether our theoretical predictions manifest in practice. For CLIP, we analyze image and text encoders separately due to the known *modality gap* (Liang et al., 2022). Normality diagnostics (AD and DP) on the MS-COCO validation set (Lin et al., 2014) are reported in Table 2. We find that modern self-supervised models exhibit near-Gaussian low-dimensional projections, whereas standard supervised models deviate substantially. Additionally, we observe thin-shell concentration across *all* models (Fig. 6, Appendix E). We add experiments on images from ImageNet-R (Hendrycks et al., 2020), sketch and painting domains, to verify this phenomenon is not limited to natural images. Results are in Table 1, showing Gaussian behavior in these settings as well. These empirical regularities provide motivation for extending the population InfoNCE analysis to multimodal and self-distillation-based objectives.

6 Discussion and Conclusion

We showed that InfoNCE trained representations admit an asymptotic Gaussian law, via two routes: an alignment-plateau analysis with thin-shell concentration, and a regularized surrogate with milder assumptions. Experiments on synthetic data, CIFAR-10, and pretrained models (MS-COCO and ImageNer-R) are consistent with these assumptions and the Gaussian hypothesis, revealing norm concentration, alignment saturation, and near-Gaussian projections, and indicating that the Gaussian approximation remains accurate and informative well before the infinite-dimensional limit. This Gaussian view justifies common modeling choices (e.g., likelihood scoring, OOD detection) and suggests that explicit isotropy promoting regularizers may act as principled surrogates for InfoNCE's implicit bias. However, *limitations* remain: our results are asymptotic, relying on high-dimensional limits and idealized assumptions that may not capture all practical regimes. We therefore view our asymptotic framework as a principled starting point rather than a complete description of all practical regimes. For finite dimension d and batch size N, projections are close to Gaussian, with deviations vanishing as $d, N \to \infty$. Quantitative bounds follow from classical Berry-Esseen (Vershynin, 2018) rates in high dimension and uniform laws of large numbers for empirical objectives (Wellner et al., 2013). In particular, the minimizer of the empirical InfoNCE loss deviates from the population minimizer by $O(N^{-1/2})$ according to Wang & Isola (2020, Thm. 1), and the distribution of fixed-k projections deviates from Gaussian by $O(d^{-1})$ according to Diaconis & Freedman (1987) (see Theorem 2 in Appendix C.1). Thus, for large but finite d, N, the Gaussian limit provides a representative and empirically useful approximation. In addition, we do not analyze optimization dynamics or prove that training attains these minimizers in practice; our results are asymptotic and characterize the population optima under the stated assumptions. Overall, we provide a principled asymptotic explanation for Gaussianity in contrastive representations, grounding empirical observations and opening new directions for analysis and practical design.

Ethics Statement

This work is theoretical and empirical in nature, focused on understanding the statistical behavior of representations trained with contrastive learning. We do not foresee direct negative societal impacts. Potential downstream applications of Gaussian modeling (e.g., density estimation, OOD detection) could influence decisions in safety-critical domains, and care must be taken to ensure robustness and fairness.

Reproducibility

We provide detailed descriptions of theoretical assumptions, proofs, and experimental protocols. Datasets (Laplace synthetic data, CIFAR-10 (Krizhevsky et al., 2009), and MS-COCO (Lin et al., 2014)) are publicly available. Architectures, hyperparameters, and training settings are fully specified (Appendix E.1), and code for experiments will be released to ensure reproducibility.

References

Venkat Anantharam, Amin Gohari, Sudeep Kamath, and Chandra Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover. *arXiv preprint arXiv:1304.6133*, 2013.

Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769, 1954.

Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *Advances in Neural Information Processing Systems*, 35:8799–8810, 2022.

Anton Baumann, Rui Li, Marcus Klasson, Santeri Mentu, Shyamgopal Karthik, Zeynep Akata, Arno Solin, and Martin Trapp. Post-hoc probabilistic vision-language models. *arXiv preprint arXiv:2412.06014*, 2024.

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened clip as a likelihood surrogate of images and captions. In *42nd International conference on machine learning*, 2025.

Wlodzimierz Bryc and Amir Dembo. On the maximum correlation coefficient. *Theory of Probability & Its Applications*, 49(1):132–138, 2005.

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*, pp. 9650–9660, 2021.

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

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

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

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

RALPH D'agostino and Egon S Pearson. Tests for departure from normality. empirical results for the distributions of b2 and sqrt(b). *Biometrika*, 60(3):613–622, 1973.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The annals of statistics*, pp. 793–815, 1984.

Persi Diaconis and David Freedman. A dozen de finetti-style results in search of a theory. In *Annales de l'IHP Probabilités et statistiques*, volume 23, pp. 397–423, 1987.

Andrew Draganov, Sharvaree Vadgama, Sebastian Damrich, Jan Niklas Böhm, Lucas Maes, Dmitry Kobak, and Erik Bekkers. On the importance of embedding norms in self-supervised learning. *arXiv preprint arXiv:2502.09252*, 2025.

Paul Dupuis and Richard S Ellis. *A weak convergence approach to the theory of large deviations*. John Wiley & Sons, 2011.

Daniel Eftekhari and Vardan Papyan. On the importance of gaussianizing representations. *arXiv preprint arXiv:2505.00685*, 2025.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International conference on machine learning*, pp. 3015–3024. PMLR, 2021.

Xianghong Fang, Jian Li, Qiang Sun, and Benyou Wang. Rethinking the uniformity metric in self-supervised learning. *arXiv preprint arXiv:2403.00642*, 2024.

Sebastian Farquhar, Michael A Osborne, and Yarin Gal. Radial bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1352–1362. PMLR, 2020.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems*, 34:7068–7081, 2021.

Hans Gebelein. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

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*, pp. 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*, pp. 9729–9738, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, volume 2, 2020.

Hermann O Hirschfeld. A connection between correlation and contingency. In *Mathematical proceedings of the cambridge philosophical society*, volume 31, pp. 520–524. Cambridge University Press, 1935.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Shao-Lun Huang and Xiangxiang Xu. On the sample complexity of hgr maximal correlation functions for large datasets. *IEEE Transactions on Information Theory*, 67(3):1951–1980, 2020.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. PMLR, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of clip. In *42nd International conference on machine learning*, 2025.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5206–5215, 2022.

Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

James Clerk Maxwell. Ii. illustrations of the dynamical theory of gases. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 20(130):21–37, 1860.

Pablo Morales-Álvarez, Stergios Christodoulidis, Maria Vakalopoulou, Pablo Piantanida, and Jose Dolz. Bayesadapter: enhanced uncertainty estimation in clip few-shot adaptation. *arXiv preprint arXiv:2412.09718*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Henri Poincaré. *Calcul des probabilités*. Gauthier-Villars, 1912.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. *URL https://arxiv. org/abs/2204.06125*, 7(4), 2022.

Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pp. 9030–9039. PMLR, 2021.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pp. 5628–5637. PMLR, 2019.

Gabor Szeg. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020b.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pp. 1179–1206. PMLR, 2021.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Sven-Ake Wegner. Lecture notes on high-dimensional data. *arXiv preprint arXiv:2101.05841*, 2021.

Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Hongkang Zhang, Shao-Lun Huang, and Ercan Engin Kuruoglu. Hgr correlation pooling fusion framework for recognition and classification in multimodal remote sensing data. *Remote Sensing*, 16(10):1708, 2024.

Wenjie Zhuo, Yifan Sun, Xiaohan Wang, Linchao Zhu, and Yi Yang. Whitenedcse: Whitening-based contrastive learning of sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12135–12148, 2023.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International conference on machine learning*, pp. 12979–12990. PMLR, 2021.

LLM USAGE

Portions of this manuscript, including text editing, reference search, ideation, mathematical derivations, and summarization, were assisted by a large language model. The model was used interactively to refine exposition, suggest formulations, and check consistency of notation, but all results, proofs, and experiments were implemented and validated by the authors. All mathematical claims, experimental details, and citations were independently verified. No content was included without author review and approval.

OVERVIEW

This appendix provides complete proofs for all propositions, corollaries, lemmas, and theorems, along with additional derivations that did not fit in the main text. We also include supplementary experiments and implementation details. The appendices are organized as follows:

A. Proof and details of the alignment bound.

B. Proofs of some regularization surrogate-related claims.

C. Proof of the alignment-plateau approach. These include general claims, some are used in the regularization surrogate proof as well.

D. Discussion about exact alignment bound at plateau.

E. Experiment details.

A HGR MAXIMAL CORRELATION AND THE ALIGNMENT BOUND

A.1 HGR DEFINITION AND BASIC PROPERTIES

The Hirschfeld-Gebelein-Rényi (HGR) maximal correlation (Hirschfeld, 1935; Gebelein, 1941; Rényi, 1959) between random variables A and B is

$$\rho_m(A,B) \;:=\; \sup_{\substack{\mathbb{E}[\phi(A)]=\mathbb{E}[\psi(B)]=0\\ \mathrm{Var}(\phi)=\mathrm{Var}(\psi)=1}} \mathbb{E}[\phi(A)\psi(B)] \;\in [0,1]. \tag{22}$$

An equivalent "explained-variance" characterization (Gebelein, 1941; Rényi, 1959) is

$$\rho_m^2(A,B) = \sup_{\substack{g\in L^2(p_A)\\ \mathrm{Var}(g(A))>0}} \frac{\mathrm{Var}\big(\mathbb{E}[g(A)\mid B]\big)}{\mathrm{Var}(g(A))}. \tag{23}$$

Here p_A is the marginal law of A, and $L^2(p_A)$ denotes the square-integrable (measurable) functions of A under p_A. The numerator is the variance explained by the optimal L^2 predictor $\mathbb{E}[g(A)\mid B]$ and the denominator is its total variance. Hence, the ratio is a (generalized) coefficient of determination, i.e., the fraction of variance of $g(A)$ predictable from B, in $[0,1]$.

HGR satisfies a (multiplicative) data-processing inequality (DPI): if $A - B - C$ is a Markov chain, then

$$\rho_m(A,C) \;\le\; \rho_m(A,B)\,\rho_m(B,C) \quad \text{(Rényi, 1959; Anantharam et al., 2013).} \tag{24}$$

We work on a standard Borel space; conditional expectations exist in L^2. Our representations are normalized ($u, v \in \mathbb{S}^{d-1}$), hence bounded and in L^2.

A.2 GAUSSIAN EXAMPLE

If two random variables A and B are jointly Gaussian, then the HGR maximal correlation between them equals the absolute value of their Pearson correlation coefficient:

$$\rho_m(A,B) = |A|, \qquad A := \frac{\mathrm{Cov}(A,B)}{\sqrt{\mathrm{Var}(A)\mathrm{Var}(B)}}. \tag{25}$$

This is a special case where the supremum in the HGR definition is achieved by simple linear functions. More precisely, the optimal transformations are just standardized versions of A and B themselves. In other words, nonlinear functions cannot increase correlation beyond the linear one when the joint distribution is Gaussian. This result is well established; see, for example, Bryc & Dembo (2005).

A.3 PROOF OF THE ALIGNMENT BOUND

We prove the inequality

$$\mathbb{E}[u\cdot v] \;\le\; \eta_2 \;+\; (1-\eta_2)\,\|m(\mu)\|^2, \tag{26}$$

for normalized representations $u = \hat{f}(X)$ and $v = \hat{f}(Y)$ on $\mathbb{S}^{d-1}$, where $m(\mu) := \mathbb{E}[u] = \mathbb{E}[v]$ is their common mean.

Step 1: mean-residual decomposition. Since u and v share the same marginal μ, their means coincide:

$$m(\mu) \;:=\; \mathbb{E}[u] \;=\; \mathbb{E}[v]. \tag{27}$$

Define residuals

$$\tilde{u} \;:=\; u - m(\mu), \qquad \tilde{v} \;:=\; v - m(\mu), \tag{28}$$

so that $\mathbb{E}[\tilde{u}] = \mathbb{E}[\tilde{v}] = 0$. Expanding the inner product yields

$$\mathbb{E}[u \cdot v] \;=\; \mathbb{E}\big[(m(\mu) + \tilde{u}) \cdot (m(\mu) + \tilde{v})\big] \;=\; \|m(\mu)\|^2 \;+\; \mathbb{E}[\tilde{u} \cdot \tilde{v}]. \tag{29}$$

The cross terms vanish because $\mathbb{E}[\tilde{u}] = \mathbb{E}[\tilde{v}] = 0$, so

$$\mathbb{E}[m(\mu) \cdot \tilde{v}] = m(\mu) \cdot \mathbb{E}[\tilde{v}] = 0, \tag{30}$$

and

$$\mathbb{E}[\tilde{u} \cdot m(\mu)] = \mathbb{E}[\tilde{u}] \cdot m(\mu) = 0. \tag{31}$$

Step 2: bound the residual correlation via HGR. Fix a coordinate $k \in \{1, \dots, d\}$ and set

$$g_k(X) \;:=\; \tilde{u}_k, \qquad h_k(Y) \;:=\; \tilde{v}_k. \tag{32}$$

Then $\mathbb{E}[g_k(X)] = \mathbb{E}[h_k(Y)] = 0$ and, by the Markov structure $X - X_0 - Y$ the DPI for HGR maximal correlation gives

$$\rho_m(X, Y) \;\le\; \rho_m(X, X_0)\, \rho_m(X_0, Y) \;=\; \sqrt{\eta_2}\, \sqrt{\eta_2} \;=\; \eta_2, \tag{33}$$

as in Anantharam et al. (2013).

For any real-valued, square-integrable functions $g(X)$, $h(Y)$ with zero mean, we can apply the definition of HGR maximal correlation (Eq. (Eq. 22)) together with the Cauchy-Schwarz inequality to obtain:

$$\big|\mathbb{E}\big[g(X)\, h(Y)\big]\big| \;\le\; \rho_m(X, Y)\, \sqrt{\mathrm{Var}(g)\, \mathrm{Var}(h)}. \tag{34}$$

This inequality holds even when g and h are not normalized, since any such functions can be rescaled to have unit variance. In our case, the random variables X and Y are conditionally independent given X_0, and identically drawn from the same augmentation channel $\mathcal{A}(\cdot \mid X_0)$. Therefore, the Markov chain $X \leftarrow X_0 \rightarrow Y$ holds, and the multiplicative data-processing inequality (Eq. (Eq. 33)) gives:

$$\rho_m(X, Y) \;\le\; \rho_m(X, X_0)\, \rho_m(Y, X_0) \;=\; \eta_2. \tag{35}$$

Substituting (Eq. 35) into (Eq. 34) yields:

$$\big|\mathbb{E}\big[g(X)\, h(Y)\big]\big| \;\le\; \eta_2\, \sqrt{\mathrm{Var}(g)\, \mathrm{Var}(h)}. \tag{36}$$

Applying (Eq. 36) to (g_k, h_k) and summing over coordinates,

$$\mathbb{E}[\tilde{u} \cdot \tilde{v}] \;=\; \sum_{k=1}^{d} \mathbb{E}[\tilde{u}_k \tilde{v}_k] \;\le\; \eta_2 \sum_{k=1}^{d} \sqrt{\mathrm{Var}(\tilde{u}_k)\mathrm{Var}(\tilde{v}_k)} \;\le\; \eta_2 \sqrt{\sum_{k=1}^{d} \mathrm{Var}(\tilde{u}_k)}\, \sqrt{\sum_{k=1}^{d} \mathrm{Var}(\tilde{v}_k)}, \tag{37}$$

where the last step is Cauchy-Schwarz for sequences.

Step 3: compute the marginal variances. Because $\|u\| = \|v\| = 1$ and $m(\mu) = \mathbb{E}[u] = \mathbb{E}[v]$,

$$\sum_{k=1}^{d} \mathrm{Var}(\tilde{u}_k) \;=\; \mathbb{E}\big[\|\tilde{u}\|^2\big] \;=\; \mathbb{E}\big[\|u - m(\mu)\|^2\big] \;=\; \mathbb{E}\big[\|u\|^2\big] - \|m(\mu)\|^2 \;=\; 1 - \|m(\mu)\|^2, \tag{38}$$

and identically

$$\sum_{k=1}^{d} \mathrm{Var}(\tilde{v}_k) \;=\; 1 - \|m(\mu)\|^2. \tag{39}$$

Step 4: conclude. Combine (Eq. 37)-(Eq. 39) to get

$$\mathbb{E}[\tilde{u} \cdot \tilde{v}] \;\le\; \eta_2 \big(1 - \|m(\mu)\|^2\big). \tag{40}$$

B Regularized Surrogate proofs

B.1 Proof of Proposition 3

Proof. For any encoder f with angular law μ the KL term satisfies (by the KL chain rule, see e.g. Dupuis & Ellis, 2011, Theorem B.2.1)

$$\mathrm{KL}(\rho\|\gamma_\lambda^B) \;=\; \mathrm{KL}(\mu\|\sigma) \;+\; \int \mathrm{KL}\big(\kappa(\cdot \mid u) \,\|\, \xi(\cdot \mid u)\big)\, \mu(du), \tag{41}$$

where $\rho(dz) = \mu(du)\kappa(dr \mid u)$ and $\gamma_\lambda^B(dz) = \sigma(du)\xi(dr \mid u)$ in polar coordinates $z = ru$. Thus, at fixed μ, the KL term is minimized by choosing $\kappa(\cdot \mid u) = \xi(\cdot \mid u)$ μ-a.s., and then $\mathrm{KL}(\rho\|\gamma_\lambda^B) = \mathrm{KL}(\mu\|\sigma)$. □

B.2 Proof of Lemma 1

Proof. We can assume $\mu \ll \sigma$, otherwise $\mathrm{KL}(\mu\|\sigma) = +\infty$ and the claim is trivial. The claim is also trivially true if $m(\mu) = 0$, so assume $m(\mu) \neq 0$. By the Donsker-Varadhan variational formula (Dupuis & Ellis, 2011, Lemma 1.4.3)

$$\mathrm{KL}(\mu\|\sigma) = \sup_{\varphi} \Big\{ \mathbb{E}_{u\sim\mu}[\varphi(u)] - \log \mathbb{E}_{u\sim\sigma}\big[e^{\varphi(u)}\big] \Big\}, \tag{42}$$

where the supremum is taken over bounded measurable functions $\varphi : \mathbb{S}^{d-1} \to \mathbb{R}$. Taking $\varphi(u) = tw \cdot u$ for some unit vector $w \in \mathbb{R}^d$ and $t \in \mathbb{R}$, we have

$$\mathrm{KL}(\mu\|\sigma) \geq \mathbb{E}_{u\sim\mu}[tw \cdot u] - \log \mathbb{E}_{u\sim\sigma}\big[e^{tw\cdot u}\big] = tw \cdot m(\mu) - \log \mathbb{E}_{u\sim\sigma}\big[e^{tw\cdot u}\big]\,. \tag{43}$$

Suppose we showed that

$$\log \mathbb{E}_{u\sim\sigma}\big[e^{tw\cdot u}\big] \leq t^2/a \tag{44}$$

for some $a > 0$ for every choice of t and w. Then picking $t = \frac{a}{2}\|m(\mu)\|$ and $w = m(\mu)/\|m(\mu)\|$, we have

$$\mathrm{KL}(\mu\|\sigma) \geq tw \cdot m(\mu) - t^2/a = \frac{a}{2}\|m(\mu)\|^2 - \frac{a}{4}\|m(\mu)\|^2 = \frac{a}{4}\|m(\mu)\|^2\,. \tag{45}$$

It is left to show (Eq. 44) with $a = 4C(d-1)$. Now, since $g(u) = w \cdot u$ is 1-Lipschitz on the sphere, then by a corollary of Lévy's isoperimetric inequality, for all $s \geq 0$,

$$\sigma\left(|g| \geq s\right) \leq 2e^{-\frac{1}{2}(d-1)s^2}\,, \tag{46}$$

where we used the fact that the median of g is 0. Since $\mathbb{E}g = 0$, this implies that for some universal $C' > 0$,

$$\log \mathbb{E}e^{tg} \leq \frac{2C'^2t^2}{d-1} \tag{47}$$

(Vershynin, 2018, Proposition 2.6.1). This satisfies (Eq. 44) with $a = \frac{d-1}{2C'^2}$, and taking $C = 1/(8C'^2)$, we are done. □

C Alignment-plateau proofs

C.1 normalized representations

Lemma 2 (At the plateau the loss reduces to uniformity)**.** *Under Assumption 1, the population InfoNCE objective (Eq. 5) takes the form*

$$\mathcal{J}(\mu) \;=\; \Phi(\mu) \;-\; \alpha\,\mathbb{E}[u\cdot v] \;=\; \Phi(\mu) \;-\; \alpha\big(\eta_2 + r_{\mathrm{plat}}\big). \tag{48}$$

hence minimizing $\mathcal{J}$ over probability laws μ on $\mathbb{S}^{d-1}$ is equivalent to minimizing $\Phi(\mu)$. Moreover, $\Phi(\mu)$ is uniquely minimized by the uniform law σ on $\mathbb{S}^{d-1}$.

Proof. At the plateau, $\mathbb{E}[u\cdot v]$ is the constant in (Eq. 8), so the alignment term is independent of μ, leaving the uniformity potential $\Phi(\mu)$ as the only objective. By Wang & Isola (2020, Appendix A), Φ is uniquely minimized at the uniform distribution on the sphere, i.e. $\mu = \sigma$. For consistency, the plateau value in (Eq. 8) must be feasible at $\mu = \sigma$. □

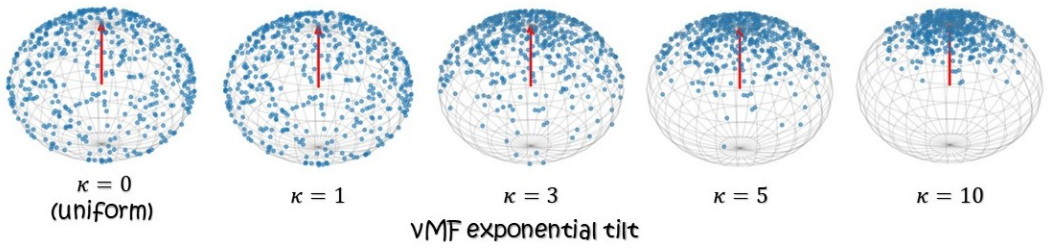

Figure 5: **vMF exponential tilt distribution for different concentration scales κ.**

Remark. In Eq. (Eq. 8), $r_{\text{plat}} \le 0$. By Eq. (Eq. 7) at $\mu = \sigma$ ($m(\mu) = 0$) the alignment ceiling is η_2; the plateau value is not guaranteed to be feasible at $\mu = \sigma$ and must be verified.

Lemma 3 (Maxwell-Poincaré (Diaconis & Freedman, 1984))**.** *Let U_d be uniform on $\mathbb{S}^{d-1}$ and fix $k \in \mathbb{N}$. Then*

$$\sqrt{d}(U_{d,1}, \dots, U_{d,k}) \;\Rightarrow\; \mathcal{N}(0, I_k) \qquad (d \to \infty). \tag{49}$$

A concrete rate of convergence was given by Diaconis & Freedman (1987).

Theorem 2. *(Diaconis & Freedman, 1987) If $1 \le k \le d-4$, then*

$$d_{\text{TV}}\big(\sqrt{d}(U_{d,1}, \dots, U_{d,k}), Z\big) \;\le\; \frac{2(k+3)}{d-k-3}\,, \tag{50}$$

where $Z \sim \mathcal{N}(0, I_k)$.

Clearly, Lemma 3 and Theorem 2 hold for any k indices, or for any orthonormal projection of U_d to k dimensions. Combining Lemmas 2 and 3, we get Corollary 1.

C.2 Unnormalized representations

We now prove Proposition 2 by reducing to the normalized case established above.

Proof. Let $z = f(X) \in \mathbb{R}^d$ denote the unnormalized representation and write its polar decomposition as $z = r\,u$ with $r = \|z\| > 0$ and $u := z/\|z\| \in \mathbb{S}^{d-1}$. By Lemma 2, at the alignment plateau the population objective reduces to minimizing $\Phi(\mu)$, whose unique minimizer is the uniform law σ on $\mathbb{S}^{d-1}$. Hence the angular component of any global minimizer satisfies $u \sim \sigma$ on $\mathbb{S}^{d-1}$. Assumption 2 further gives thin-shell concentration of the radius: $r \xrightarrow[d\to\infty]{\mathsf{P}} r_0 \in (0, \infty)$.

For any fixed $k \ge 1$ and any fixed k-dimensional subspace, let P_k be the corresponding orthogonal projector and set $u_k := P_k u$. By the Maxwell-Poincaré spherical CLT (Lemma 3),

$$\sqrt{d}\,u_k \;\Rightarrow\; \mathcal{N}(0, I_k) \qquad (d \to \infty). \tag{51}$$

Let $z_k := P_k z = r\,u_k$. Since $r \xrightarrow[d\to\infty]{\mathsf{P}} r_0$ and (Eq. 51) holds, Slutsky's theorem (Van der Vaart, 2000) yields

$$\sqrt{d}\,z_k \;=\; r\sqrt{d}\,u_k \;\Rightarrow\; \mathcal{N}\big(0,\, r_0^2 I_k\big) \qquad (d \to \infty). \tag{52}$$

This proves Proposition 2. □

D Exact alignment bound in plateau discussion

The following analysis begins from the alignment ceiling (Eq. 7): under a generalized plateau assumption (extending Assumption 1), the expected alignment is determined by the augmentation mildness η_2 and the squared mean norm $\|m(\mu)\|^2$, up to a negligible residual (noted as r_{plat} in Eq. (Eq. 8)). Substituting this relation into the population InfoNCE objective (Eq. 5) yields the surrogate

$$\mathcal{J}_q(\mu) \;=\; \Phi(\mu) \;-\; q\,\|m(\mu)\|^2, \qquad q \;=\; \alpha(1-\eta_2), \tag{53}$$

where $\Phi(\mu)$ is the uniformity potential of Wang & Isola (2020). Thus, at the plateau, the population loss reduces to a trade-off between uniformity and the mean vector length.

Stationary points. The surrogate involves the spherical convolution operator P with kernel $e^{\alpha\xi\cdot\eta}$, which diagonalizes in spherical harmonics by the Funk-Hecke theorem (Atkinson & Han, 2012). Analyzing the Euler-Lagrange condition shows that in high dimensions Ph must asymptotically take an exponential tilt form $Ph(\xi) \propto \exp(\beta w \cdot \xi)$. Inverting this relation via Gegenbauer expansions and their decay properties (Szeg, 1939) indicates that, under mild regularity, the stationary density h is well-approximated in its leading modes by either the uniform law or a von Mises–Fisher (vMF) tilt (Mardia & Jupp, 2009). This captures the dominant low-degree structure in high dimensions, though more complex stationary forms cannot be excluded.

Implications. Consequently, in high dimension the stationary points of the plateau surrogate are *well-approximated* by either the uniform distribution (when $m(\mu) = 0$) or a von Mises-Fisher (vMF) tilt aligned with an axis w (when $m(\mu) \neq 0$); see Fig. 5. The vMF concentration parameter κ quantifies the strength of angular concentration around w (larger $\kappa \Rightarrow$ narrower cone). This perspective helps explain why contrastive encoders often yield nearly uniform representations, with occasional vMF-like bias. For example, in CLIP, where a narrow-cone structure (a modality-dependent angular bias) has been observed (Liang et al., 2022).

E Experimental details

E.1 Implementation details

Code and reproducibility. Code will be released upon publication. All experiments were implemented in `PyTorch` with `torchvision`. Training was performed on a single 3090 NVIDIA RTX GPU with CUDA 11.8.

Synthetic Data Experiments

- **Dataset.** Laplace$(0, 1)$ vectors of dimensions - $d_{data} = 1024$. We use a set of 20k samples for training, and 5k samples for testing.
- **Representation dimensions.** The dimensions of representations vary: $d \in \{32, 64, 128, 256\}$.
- **Batch size.** Batch size in our experiments varies: $N \in \{8, 16, 32, 48, 64, 96, 128\}$.
- **Training objective.** InfoNCE loss with temperature $\tau \in \{0.1, 0.2\}$. We report results for $\tau = 0.1$, but note that results are similar.
- **Augmentations.** Each synthetic sample x is perturbed to form two correlated views

$$x_1 \;=\; A\,x + \sqrt{1 - A^2}\,\varepsilon_1, \qquad x_2 \;=\; A\,x + \sqrt{1 - A^2}\,\varepsilon_2, \tag{54}$$

  where $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, I)$ are independent. The parameter $A \in (0, 1)$ controls the correlation between views. After this linear Gaussian mixing, we apply light, independent jitter: additive Gaussian noise with std 0.2, feature dropout with probability 0.1, and random multiplicative scaling by $\exp(\mathcal{N}(0, 0.1^2))$. Unless otherwise stated, we use $A = 0.6$ (results for $A \in \{0.2, 0.5, 0.8\}$ appear in Fig. 11).
- **Optimization.** Optimizer: Adam. Learning rate $= 10^{-3}$. We ran 50-250 epochs depending on setup; unless stated otherwise, we report results at 150 epochs.
- **Evaluation metrics.** norm concentration (CV), mean norm values, Gaussianity diagnostics (AD/DP) tests and uniformity vs. alignment comparison (based on cosine similarity).

CIFAR-10 Experiments

- **Dataset.** CIFAR-10, training set size 50k, test set size 10k.
- **Augmentations.** We apply the standard SimCLR-style augmentation pipeline: a random resized crop to 32×32 pixels with scale uniformly sampled from $(0.2, 1.0)$, a random horizontal flip, color jitter with strengths $(0.8, 0.8, 0.8, 0.2)$, and random conversion to grayscale with probability 0.2r.

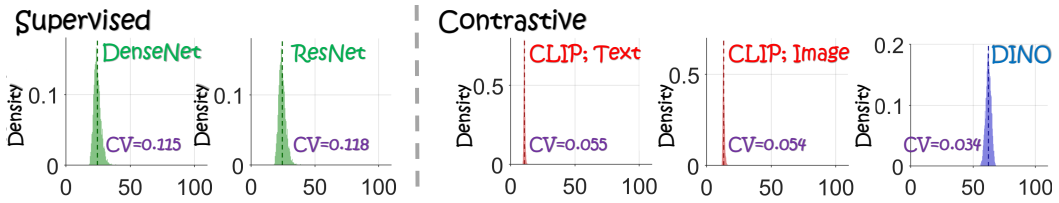

Figure 6: **Thin-shell concentration across pretrained models.** Radius distributions of representations from supervised models (DenseNet, ResNet) and contrastive models (CLIP, DINO). All models exhibit thin-shell concentration, with contrastive methods showing tighter clustering (lower CV, (Eq. 21)).

- **Architecture.** ResNet-18 encoder (pretrained on ImageNet (Deng et al., 2009)) with a two-layer MLP projection head (hidden dim = 512, output dim = 128).
- **Training objective.** InfoNCE with temperature $\tau = 0.1$.
- **Optimization.** Adam optimizer, learning rate $= 10^{-3}$, weight decay = 10^{-4}, batch size $= 256$, epochs = 100.
- **Evaluation metrics.** norm concentration (CV), Gaussianity diagnostics (AD/DP) tests.

Pretrained Model Diagnostics

- **Models.** CLIP (ViT-L/14, text and image modalities), DINO (ViT-B/32), ResNet-34 and DenseNet.
- **Datasets.** Full MS-COCO validation set (5k images).
- **Feature extraction.** Last-layer embeddings; whitening applied when noted.
- **Evaluation metrics.** norm concentration (CV), Gaussianity diagnostics (AD/DP) tests and uniformity before and after whitening.

E.2 Additional Experiments

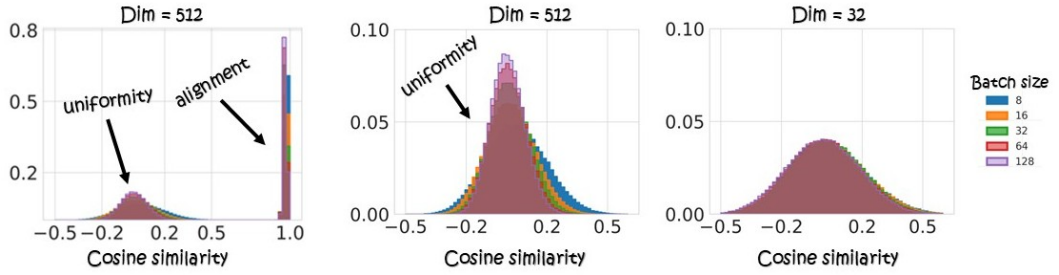

Figure 7: **Alignment and uniformity vs. batch size.** Histogram view of cosine similarities for positive pairs (alignment) and negatives (uniformity), corresponding to Fig. 2. As batch size increases, alignment remains high while uniformity improves, with negative-pair similarities concentrating near zero. The middle panel is a zoom of the left; the right panel shows that at very low dimensionality, increasing batch size yields little uniformity gain.

Figs. 7 and 8 provide alternative visualizations of Fig. 2, presenting the same experiments with a different display. Both figures plot the distributions of cosine similarities for positive pairs (alignment) and for negatives (uniformity). As batch size (Fig. 7) or dimensionality (Fig. 8) increases, uniformity improves (negative-pair similarities concentrate near zero) while alignment remains consistently high across settings. These complementary views reinforce the observation from the main body: uniformity continues to improve with larger batches and higher dimensions, whereas alignment appears to saturate early.

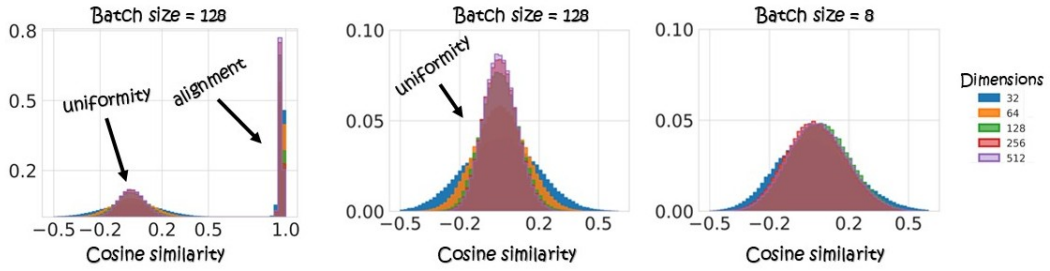

Figure 8: **Alignment and uniformity vs. dimensionality.** Histogram view of cosine similarities for positive pairs (alignment) and negatives (uniformity), corresponding to Fig. 2. As dimensionality increases, alignment stays high while uniformity improves, pushing negative-pair similarities toward zero. The middle panel is a zoom of the left; the right panel highlights that with very small batch sizes, increasing dimensionality offers limited uniformity improvement.

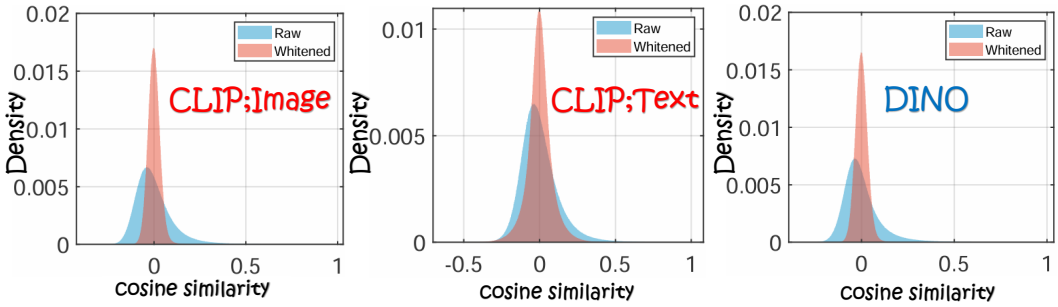

Figure 9: **Whitening and uniformity: unnormalized representations.** Cosine similarity histograms of negatives for CLIP (image, text) and DINO, before (raw) and after whitening. Unnormalized representations benefit from whitening, with distributions pushed closer to zero, reflecting enhanced uniformity.

Additionally, we assess uniformity in several pretrained models before and after whitening. Whitening consistently increases uniformity, indicating that these representations, which are already close to uniform (and approximately Gaussian; see Table 2), become more isotropic once decorrelated and rescaled. This effect holds consistently across pretrained models (CLIP image, CLIP text, and DINO), for both normalized and unnormalized representations, see Figs. 9, 10. Thus, a simple post hoc projection via whitening can further enhance uniformity in practice.

We examine the correlation between the data distribution and the representation distribution. Using Laplace data as input and observing Gaussian representations at the output, we can compute likelihoods for both input and output sets. Comparing these scores reveals strong correlation (Fig. 11), indicating that the distribution is indeed "pushed forward" through the encoder. This correlation remains stable across different augmentation strengths, showing that this "pushforward" behavior is insensitive to the level of augmentation.

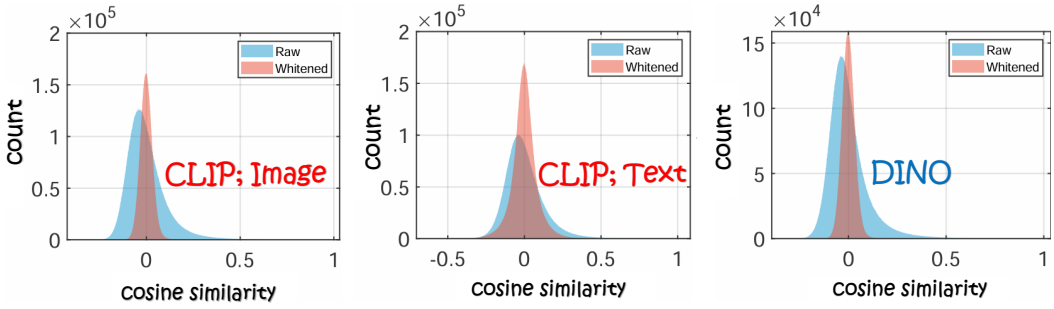

Figure 10: **Whitening and uniformity: normalized representations.** Cosine similarity histograms of negatives for CLIP (image, text) and DINO, before (normalized) and after whitening. Normalized representations are already close to uniform; whitening provides a modest but consistent improvement.

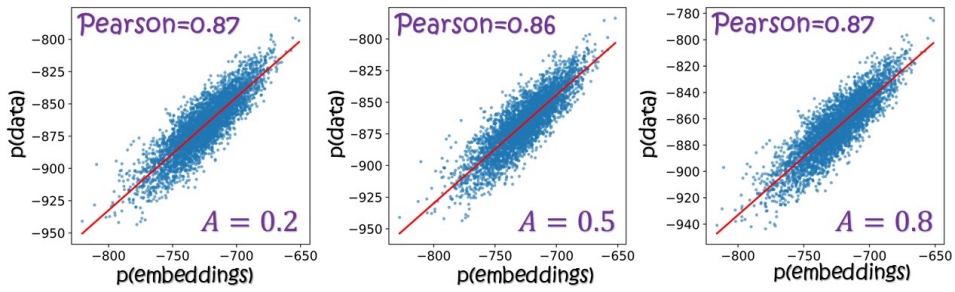

Figure 11: **Encoder "pushforward".** On synthetic data, the encoder maps Laplace-distributed inputs to approximately Gaussian representations. Because both source and target families admit tractable likelihoods, we can score entire sets and observe consistently high correlation across different augmentation strengths.