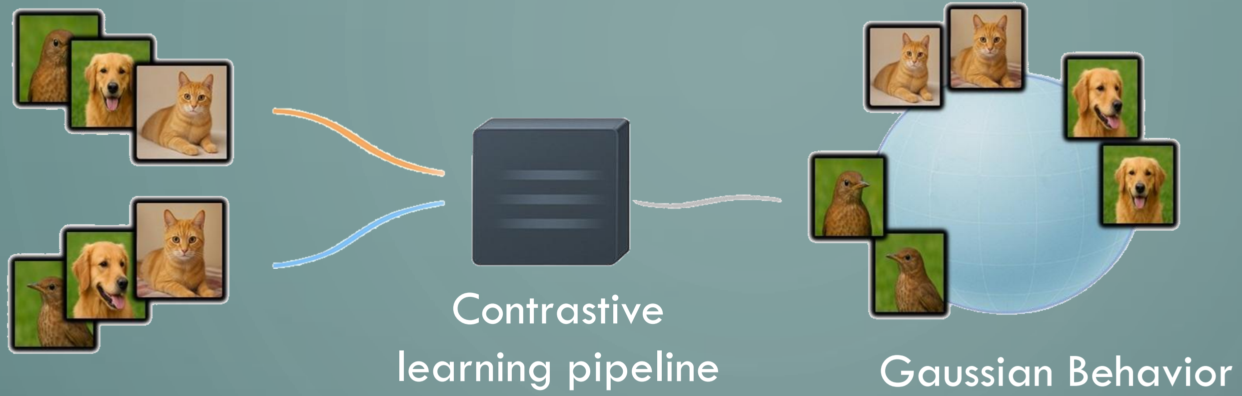


InfoNCE Induces Gaussian Distribution

ROY BETSER* · EYAL GOFER* · MEIR YOSSEF LEVI · GUY GILBOA

TECHNION – ISRAEL INSTITUTE OF TECHNOLOGY

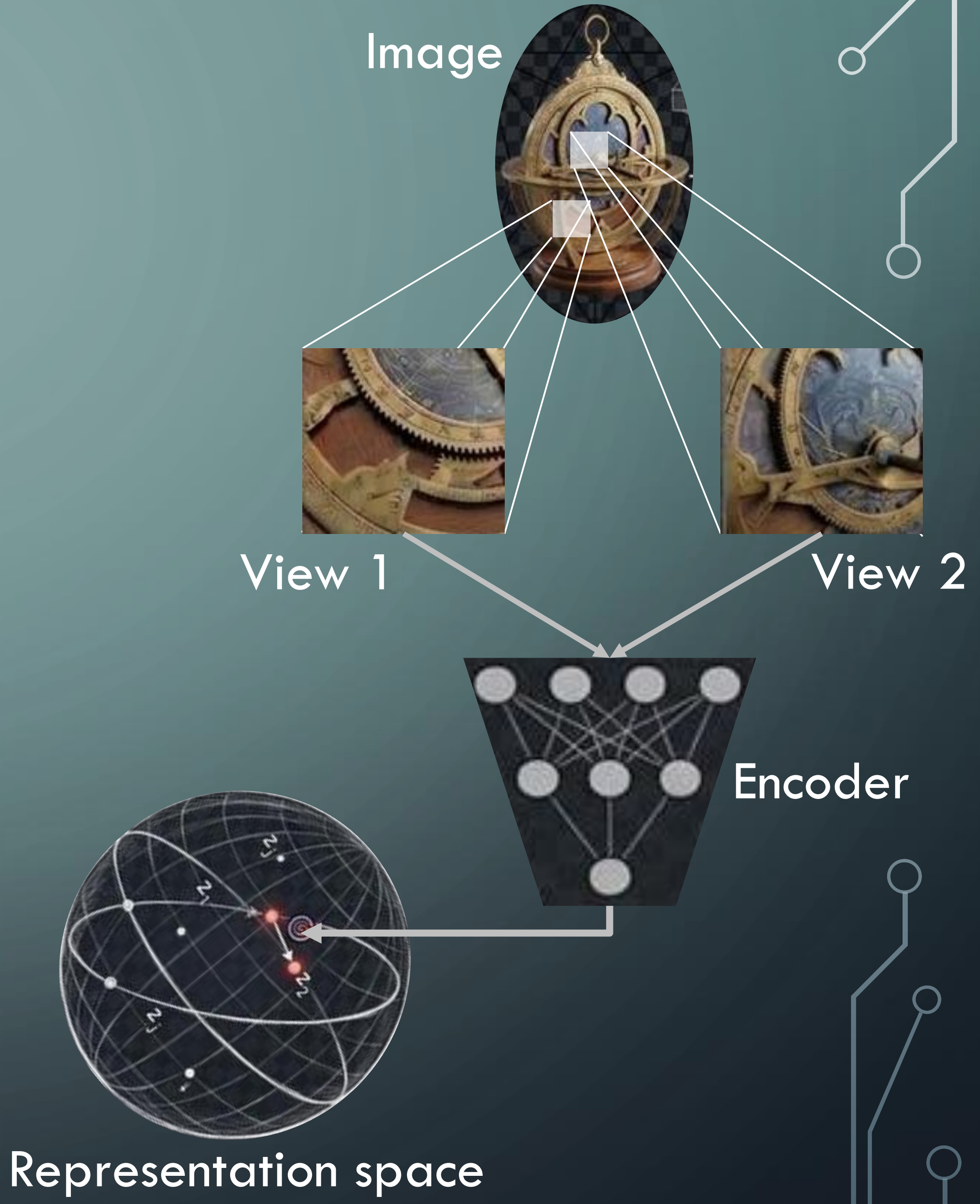
* EQUAL CONTRIBUTION

Paper page

ICLR 2026, ORAL PRESENTATION

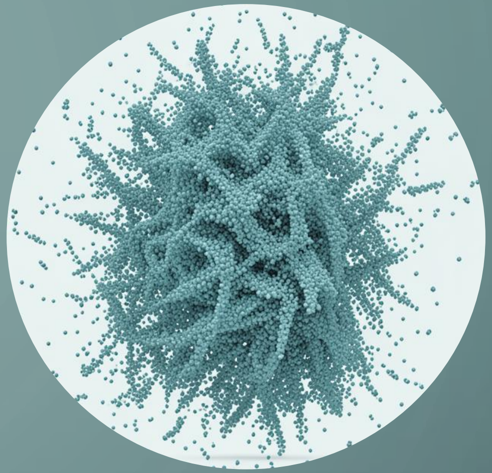
Contrastive learning

- A dominant paradigm for representation learning (e.g., SimCLR, CLIP, etc.).
- **Key idea:** Two augmented views of a sample form a positive pair.
- **Goal:** bringing positive pairs closer and separating other samples (negatives).



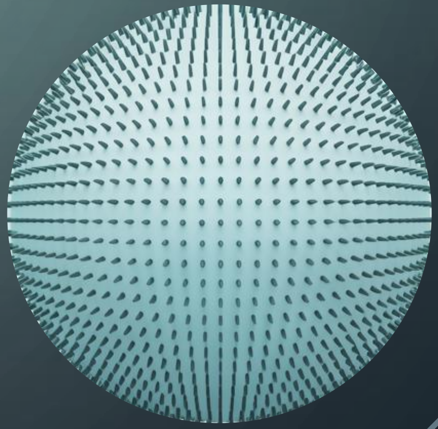
Representation space structure

- Encoding models produce high-dimensional embeddings.
- Many downstream applications reuse pretrained foundation models.



Random Structure

Uniform Hypersphere



If the embedding distribution has simple structure (e.g., Gaussian), analysis becomes tractable.

- Likelihood, entropy, mutual information, etc.

Contrastive learning objective

InfoNCE loss:

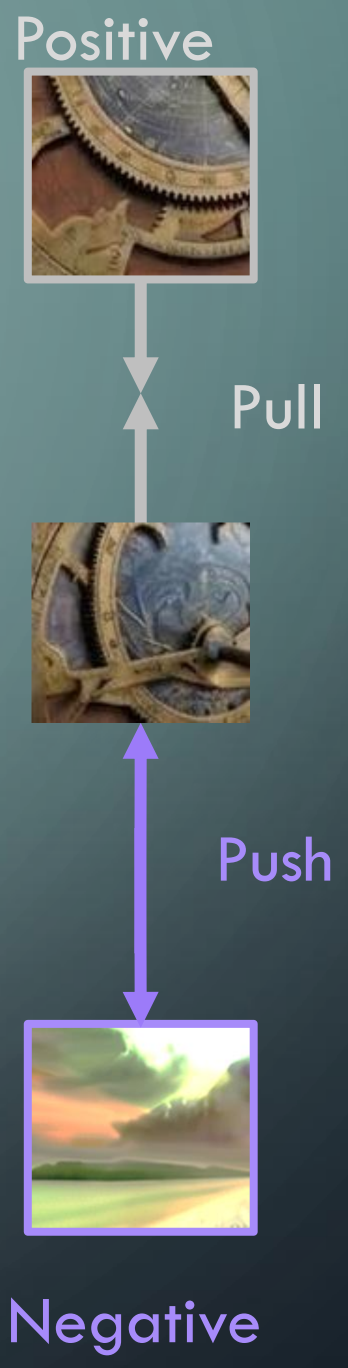
$$\mathcal{L}_{\mathrm{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{exp\left(\frac{1}{\tau} u_i \cdot v_i\right)}{\sum_{j=1}^{N} \exp\left(\frac{1}{\tau} u_i \cdot v_i\right)}$$

Alignment

Uniformity

Population objective:

$$\mathcal{L}(\mu, \pi) = \underbrace{-\alpha\, \mathbb{E}_{(u,v)\sim\pi}[u \cdot v]}_{\text{Alignment}} + \underbrace{\mathbb{E}_{u\sim\mu} \log \mathbb{E}_{v\sim\mu} \exp(\alpha\, u \cdot v)}_{\text{Uniformity } \Phi(\mu)}$$

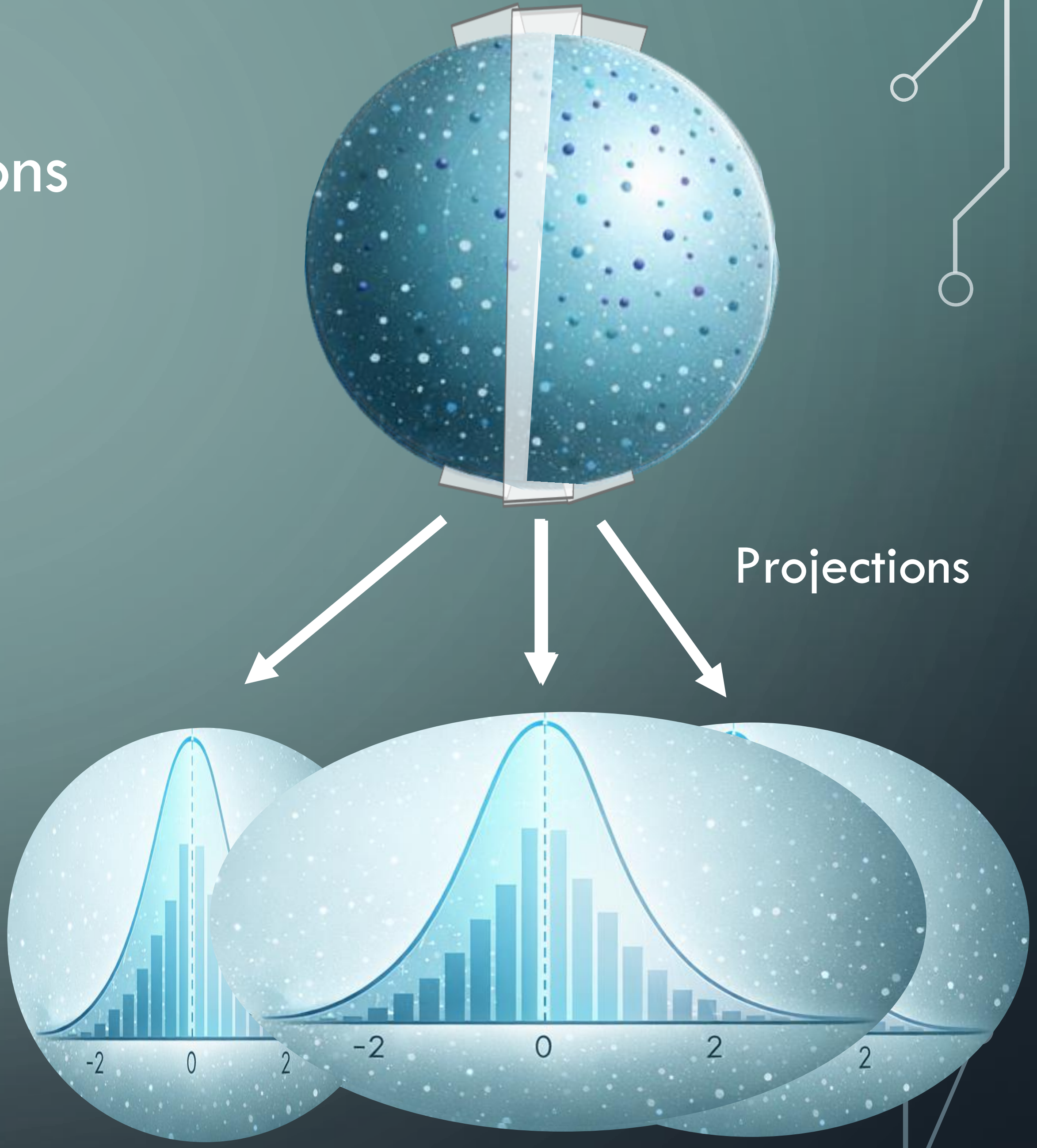


u_i, v_i - normalized representations

Uniformity and Gaussian projections

Maxwell–Poincaré Lemma[1]

- High dimensional uniform points on the sphere appear random.
- Projecting onto any fixed direction yields Gaussian behavior.



[1] Diaconis & Freedman, 1987

Bounded Alignment

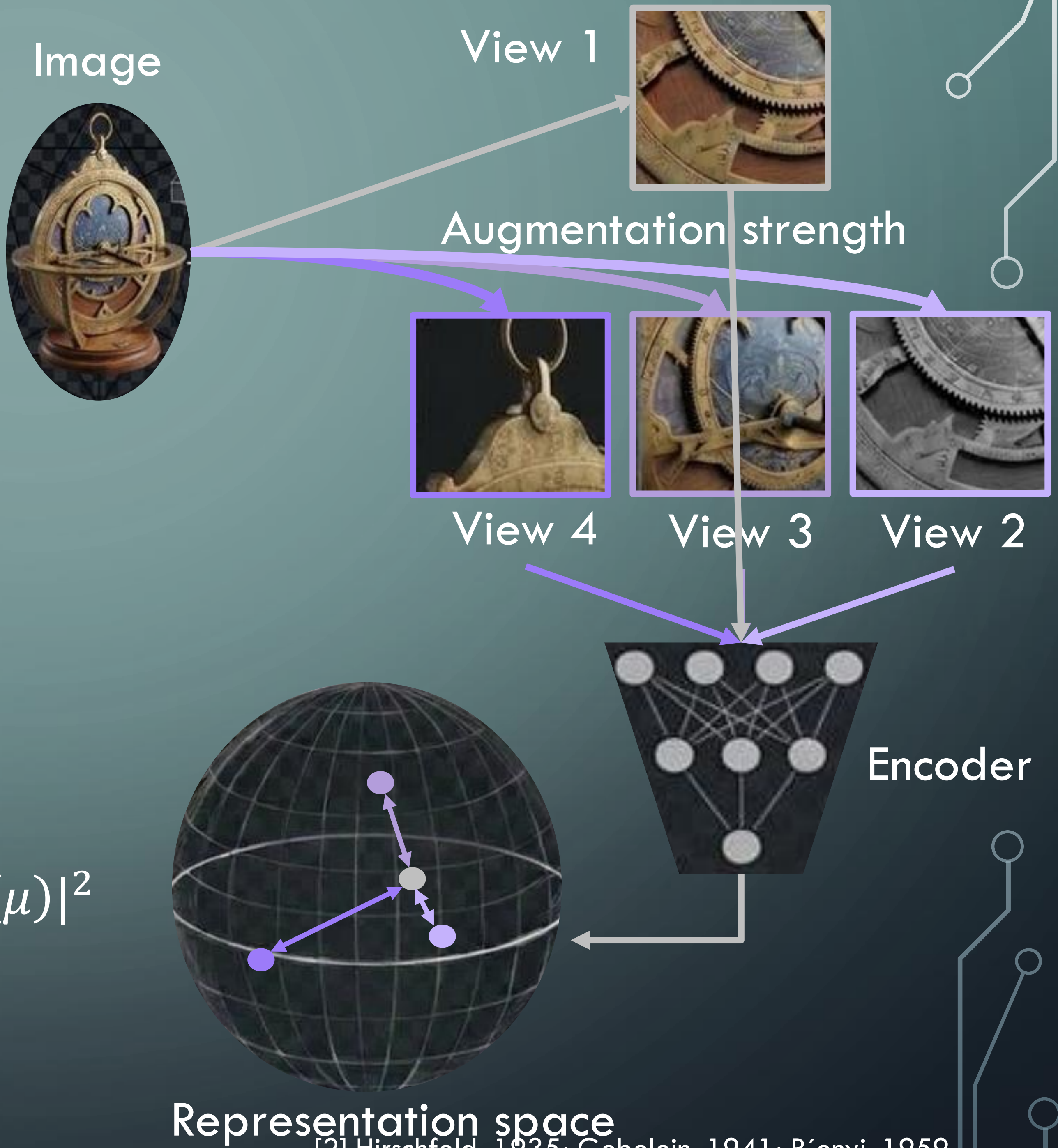
- Augmentation strength controls recoverability of original image.
- HGR maximal correlation[2]:

$$\eta_2 = \sup_{\substack{g \in L^2(p_X) \\ Var(g)>0}} \frac{Var(\mathbb{E}[g(X)|X_0])}{Var(g(X))} \in [0,1]$$

- Alignment bound:

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

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



[2] Hirschfeld, 1935; Gebelein, 1941; R´enyi, 1959

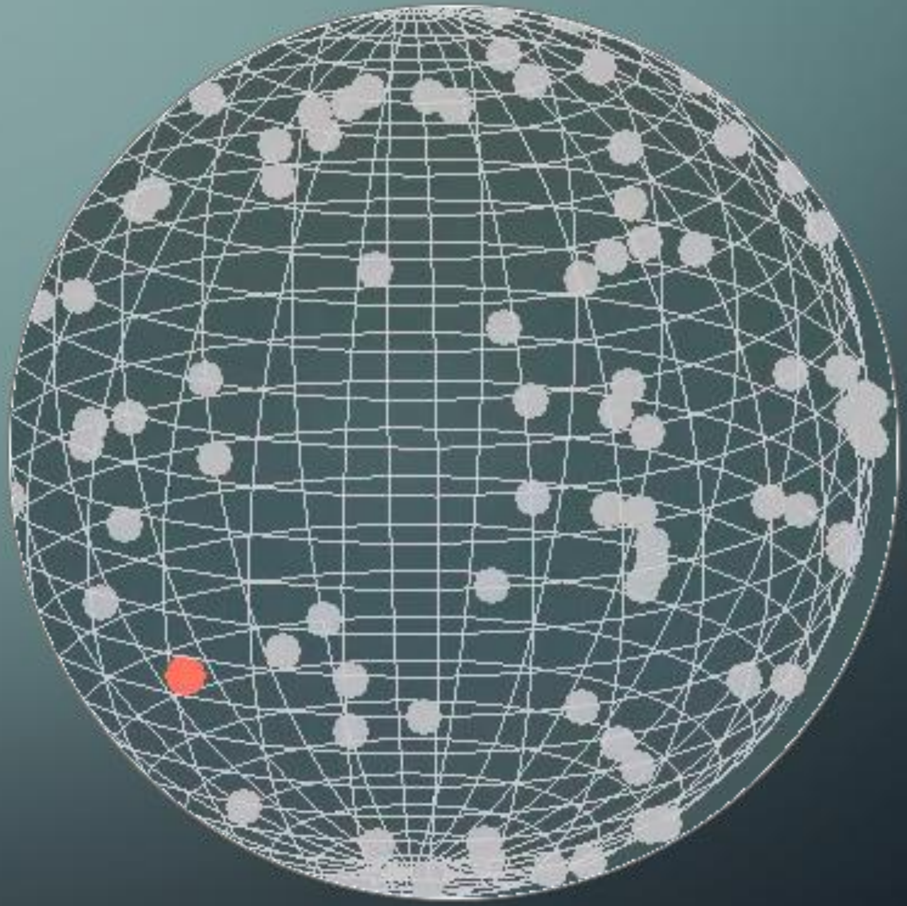
Alignment saturation case

When alignment saturates and training continues:

- Uniformity is the remaining objective[3].
- Uniform on sphere leads to Gaussian projections.

Phase 1: Both Alignment and uniformity improve until alignment saturation.

Phase 2: Alignment saturated, uniformity continues to improve.



[3] Wang & Isola, 2020

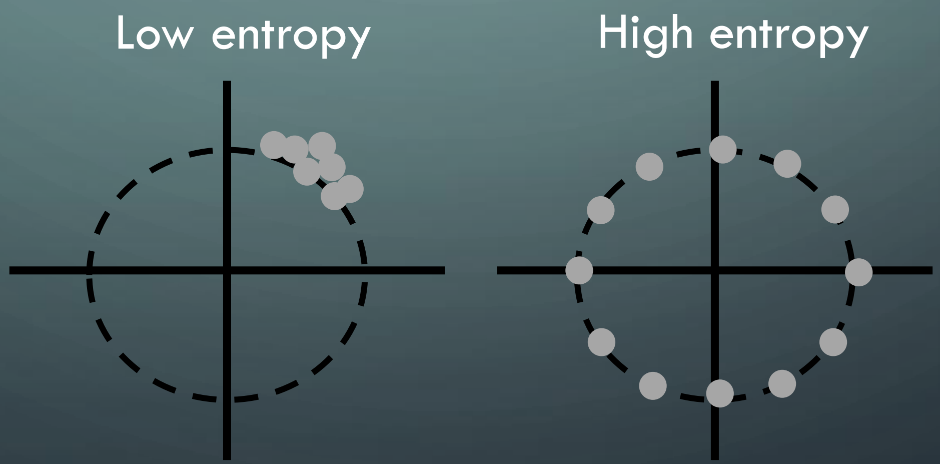
Regularized approach

- Add a regularizer with strength β.
  - Conclusions hold for $\beta \to 0$.

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

High entropy → $-H(\rho)$

Low norms → $\lambda \mathbb{E}_{Z\sim\rho} |Z|^2$



Regularized approach

- Add a regularizer with strength β.
  - Conclusions hold for $\beta \to 0$.

$$J(f) = \underbrace{\Phi(\mu)}_{\text{Uniformity}} - \underbrace{\alpha \mathbb{E}_{(u,v)\sim\pi}[u \cdot v]}_{\text{Alignment}} + \underbrace{\beta\Big(\overset{\text{High entropy}}{-H(\rho)} + \overset{\text{Low norms}}{\lambda \mathbb{E}_{Z\sim\rho}\, |Z|^2}\Big)}_{\text{Regularizer}}$$

Low norms

High norms

Regularized approach

f – encoder
$f(\mathrm{X})$ – representation
$\hat{f}(\mathrm{X})$ – representation
p_X - data distribution.
$(\hat{f})_* p_X = \mu$ - push forward of the data distribution to the representation distribution.

- Define the best attainable alignment:

$$Align(\mu) = \sup_{f} \left\{ \mathbb{E}\left[\hat{f}(X) \cdot \hat{f}(Y)\right] : \left(\hat{f}\right)_* p_X = \mu \right\}$$

- Recall the alignment bound:

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

Minimal for the uniform distribution!

$\mathrm{m}(\mu) := \mathbb{E}[\mathrm{u}] = \mathbb{E}[\mathrm{v}]$
u_i, v_i - normalized representations

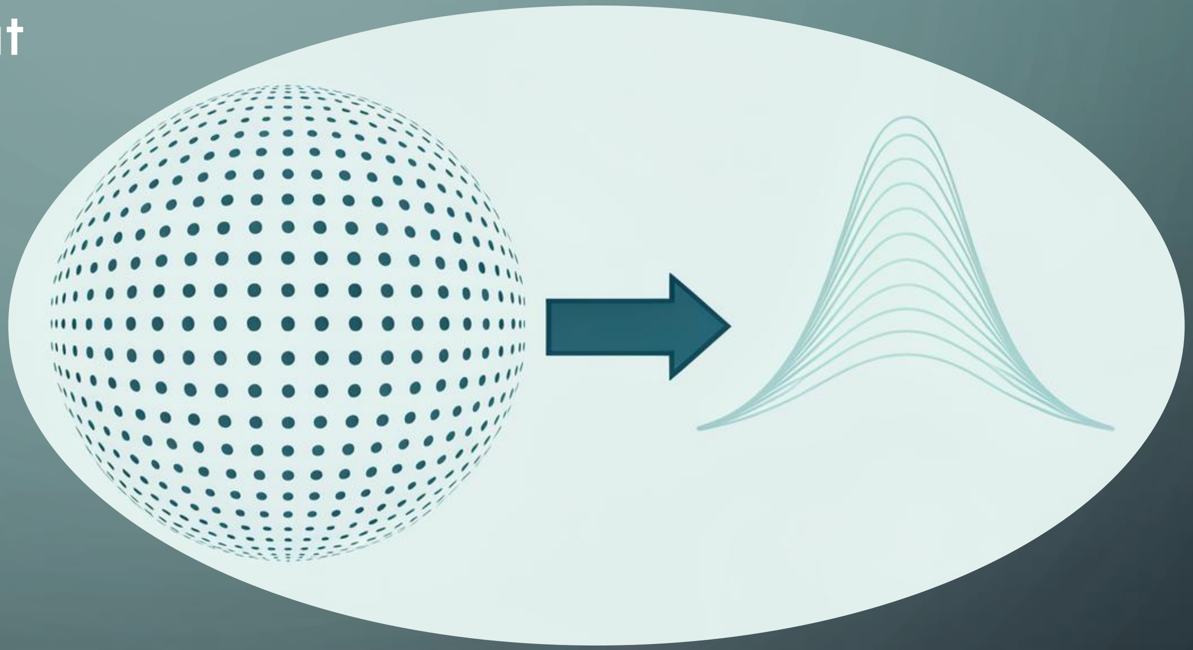
Regularized approach

If alignment bound is attainable at the uniform distribution:

- It is the global minimizer!
- Regularization strength:

$$\beta \geq \beta_0 = \frac{\alpha(1-\eta_2)}{C(d-1)} \xrightarrow[\mathrm{d}\to\infty]{} 0$$

Reminder: $J(f) = \mathcal{L}_{\mathrm{InfoNCE}} + \beta(\mathrm{regularizer})$



In the asymptotic case, with infinite dimensions, regularization vanishes.

Empirical results

Gaussianity tests in finite data and dimensions:

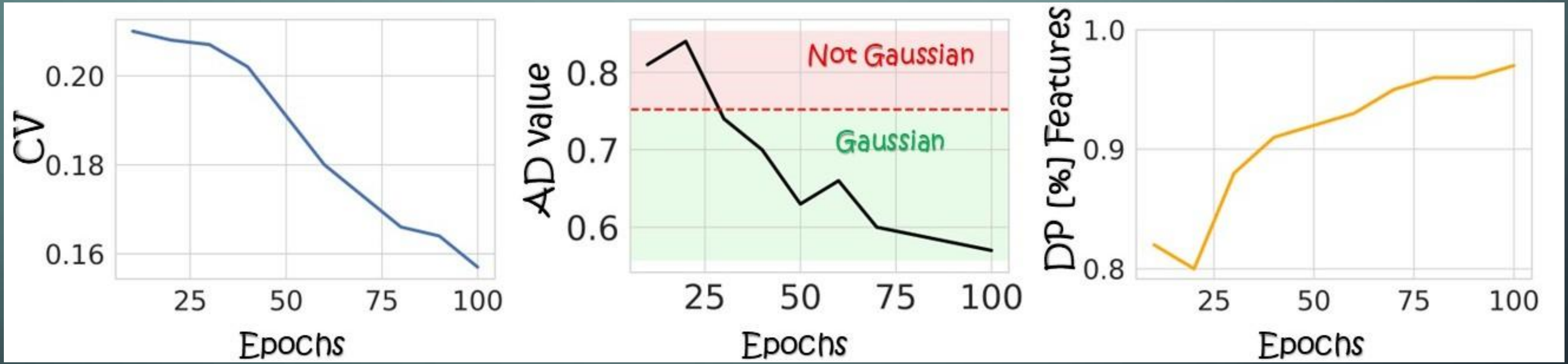
Metric	Synthetic (Linear)					CIFAR-10 (ResNet-18)	
	Laplace	GMM	Binary E0	E50	E100	Supervised	Contrastive
CV	0.08	0.08	0.36	0.12	0.09	0.5	0.09
AD Avg. (< 0.752)	0.38	0.39	1.64	0.41	0.42	3.3	0.43
AD Norm. Feat.	100%	100%	30%	93%	97%	6.2%	96.1%
DP Avg. (> 0.05)	0.49	0.46	0.02	0.44	0.46	0.041	0.39
DP Norm. Feat.	100%	100%	15%	89%	98%	3.9%	94.5%
Gaussian?	✓	✓	✗	✓	✓	✗	✓

- AD – Anderson-Darling test[4]
- DP – D'Agostino-Pearson test[5]

[4] Anderson & Darling, 1954, [5] D'Agostino & Pearson 1973

Empirical results

Gaussianity tests along training:



Main takeaways

- Contrastive learning naturally induces Gaussian structure into learned representations.
- Augmentations controls maximal alignment between positive pairs.
- Asymptotically, uniformity on the sphere is the global minimizer under different assumptions.
- Classical mathematical tools can help us analyze representation spaces.

THANK YOU

QUESTIONS?

Project page