APPENDIX

A. Related Work

Contrastive learning has received a surge of interest in the last few years. A large body of work investigates contrastive learning methods empirically [van den Oord et al., 2018, Bachman et al., 2019, Dwibedi et al., 2021, Chen et al., 2020]. Many recent works focus on understanding the impact of different approaches for sampling positive [Ho and Nvasconcelos, 2020, Tian et al., 2021, Zheng et al., 2021, Hayes et al., 2022, Wang et al., 2020] and negative instances [Chuang et al., 2020, Ma et al., 2021, Ash et al., 2020, 2022, Shah et al., 2022, Robinson et al., 2021], motivated by the intuitive idea that sampling instances that are more *informative* about the underlying structure could lead to more effective contrastive learning. Notably, Ho and Nvasconcelos [2020] propose to generate "challenging" positive pairs via adversarial perturbation (*adversarial augumentation*), which they demonstrate empirically to be promising. Zheng et al. [2021] empirically investigate an approach that selects positive instances via a graph-based weakly-supervised approach. To the best of our knowledge, sampling approaches for positive instances have not been studied systematically in a latent variable model. Yang et al. [2021] and Patrick et al. [2021] investigate approaches for optimizing the composition of transformations.

The theoretical analysis of contrastive learning approaches has recently received increasing attention [Saunshi et al., 2019, HaoChen et al., 2021, Graf et al., 2021, Wang and Isola, 2020, Zimmermann et al., 2021, Wang et al., 2022]. Notably, [Saunshi et al., 2019] proposed one of the first theoretical frameworks for contrastive learning, which evaluates the quality of the learned representations on a downstream classification task. However, they make no assumptions on the structure of the underlying latent space. Wang and Isola [2020] and Zimmermann et al. [2021] analyzed contrastive learning in a latent variable model, albeit without assuming additional structure, such as latent classes. Both works consider only classical sampling strategies, where positive instances are generated via random augmentations.

B. Reconstruction of Latent Space

We first introduce a framework for analyzing the quality of a representation function with respect to its ability to recover the latent class structure in $\mathcal{Z}$. Recall that we identify the latent space with the unit hypersphere, i.e., $\mathcal{Z} = \mathbb{S}^{k-1}$. The latent classes $\mathcal{C} = \{C_1, C_2, \dots\}$ form spherical caps in $\mathcal{Z}$ (see Fig. 1, main text). We denote class labels with lower case letters, i.e., $c_1, c_2, \dots$. In the following, we assume that the conditional distribution of positive pairs of latent variables (z, z^+) from which (x, x^+) are generated is von Mises-Fisher, i.e.,

$$p(z^+|z) = C_p^{-1} \exp(\tau^{-1} z^T z^+) \tag{A.1}$$

$$C_p = \int \exp(\tau^{-1} z^T z^+) dz^+ \,, \tag{A.2}$$

where $(z, z^+) \sim p(z^+|z)p_c(z)$ is a positive pair of latent variables sampled from class c and $\tau > 0$ a hyperparameter. The marginal distribution over the class c is assumed to be uniform, i.e., $p_c(z) = |\mathcal{C}|^{-1}$. Recall that observations $x \in \mathcal{X}$ are generated by an unknown map g, i.e., $x = g(z)$. We can define a posterior distribution $p(z|x)$ over the true latent variables that generated x. In Algorithms 1(i) and 2, positive instances are directly sampled from $p(z^+|z)$. In Algorithms 1(ii) and 1(iii), candidate positive instances are drawn from $p(z^+|z)$ and accepted according to the specified rules. Negative instances are sampled uniformly at random.

Recall that we want to learn a representation function $f : \mathcal{X} \to \mathbb{S}^{k-1}$, such that the composition $h : \mathbb{S}^{k-1} \to \mathbb{S}^{k-1}$, $h = f \circ g$ preserves the alignment between latent variables. A good representation function recovers the hidden latent variables, the underlying task is a demixing problem, where we learn to invert the generative process g (up to orthogonal linear transformations). This requires that h preserves the dot products between positive pairs (z, z^+) up to a constant, i.e., $\kappa z^T z^+ = h(z)^T h(z^+)$ (with $\kappa > 0$). This is equivalent to requiring that h locally reconstructs the latent space up to linear and orthogonal transformation.

In the absence of class structure, [Zimmermann et al., 2021, Prop.1] showed that if $\mathcal{F}$ is sufficiently rich, a suitable h minimizes the cross entropy of the ground-truth conditional distribution $p(z^+|z)$ and the conditional distribution of the recovered latent variables. In the presence of class structure, an analogous result can be shown for the distribution of positive pairs sampled from a class c:

Theorem A.1: Consider the minimizer

$$h^* = \operatorname{argmin} \mathop{\mathbb{E}}_{\substack{(z,z^+) \\ \sim p(z^+|z)p_c(z)}} \left[H(p(z^+|z,c), q_h(z^+|z,c)) \right] \,, \tag{A.3}$$

with

$$q_h(z^+|z,c) = C_h(z^+)^{-1} e^{h(z^+)^T h(z)/\tau} \tag{A.4}$$

$$C_h(z) := \int e^{h(z^+)^T h(z)/\tau} dz \,, \tag{A.5}$$

denoting the conditional distributions over reconstructed latent variables and $H(p, q)$ denoting the cross entropy between distributions p and q. Then h^* locally reconstructs latent space up to linear and orthogonal transformation.

We include a proof for completeness.

$$
\begin{aligned}
&\mathbb{E}_{c\sim p_c}\left[\mathbb{E}_{z\sim p_c(z)}\left[H\left(p(\cdot|z,c), q_h(\cdot|z,c)\right)\right]\right] \\
&\quad= \mathbb{E}_{c\sim p_c}\left[\mathbb{E}_{z\sim p_c(z)}\left[\mathbb{E}_{z^+\sim p(z^+|z,c)}\left(-\log q_h(z^+|z,c)\right)\right]\right] \\
&\quad\overset{(1)}{=} \mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[-\frac{1}{\tau}h(z^+)^T h(z) + \log C_h(z)\right] \\
&\quad= -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[h(z^+)^T h(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log C_h(z)\right] \\
&\quad\overset{(2)}{=} -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[h(z^+)^T h(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log \int_{z'} e^{h(z')^T h(z)/\tau} dz'\right] \\
&\quad\overset{(3)}{=} -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[h(z^+)^T h(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log\left(|\mathcal{C}|\cdot \mathbb{E}_{z'\sim p_c(z)}\left(e^{h(z')^T h(z)/\tau}\right)\right)\right] \\
&\quad= -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[h(z^+)^T h(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log \mathbb{E}_{z'\sim p_c(z)}\left(e^{h(z')^T h(z)/\tau}\right)\right] + \log|\mathcal{C}| \\
&\quad\overset{(4)}{=} -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[((f\circ g)(z^+))^T (f\circ g)(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log \mathbb{E}_{z'\sim p_c(z)}\left(e^{((f\circ g)(z'))^T (f\circ g)(z)/\tau}\right)\right] + \log|\mathcal{C}|\,,
\end{aligned}
$$

where in (1) we have inserted the definition of q_h and in (2) the definition of the partition function C_h. In (3) we have multiplied by 1 ($|\mathcal{C}|\,|\mathcal{C}|^{-1}$) and approximated the integral by sampling from $p_c(z) = |\mathcal{C}|^{-1}$. In (4), we have inserted $h = f \circ g$.

By expressing functions of latent variables with the corresponding expressions for observables, we get

$$
-\frac{1}{\tau}\mathbb{E}_{(x,x^+)\sim p_{pos}}\left[f(x^+)^T f(x)\right] + \mathbb{E}_{x\sim p_{data}}\left[\log \mathbb{E}_{x^-\sim p_{data}}\left(e^{f(x^-)^T f(x)/\tau}\right)\right] + \log|\mathcal{C}| = \mathcal{L}_{align}(f;\tau) + \mathcal{L}_{uni}(f;\tau) + \log|\mathcal{C}|\,.
$$

Geometrically, the cross-entropy encodes the concepts of *alignment* and *uniformity*, which are characterized by the following loss functions [Wang and Isola, 2020, Zimmermann et al., 2021]:

- *Alignment:* Positive pairs should be mapped to nearby feature representations. This is captured in the loss:

$$
\mathcal{L}_{align}(f;\tau) = -\frac{1}{\tau}\mathop{\mathbb{E}}_{(x,x^+)\sim p_{pos}}\left[f(x^+)^T f(x)\right]\,. \tag{A.6}
$$

- *Uniformity:* Feature vectors should be approximately uniformly distributed on $\mathbb{S}^{k-1}$ to encourage separability. This is encoded in the loss:

$$
\mathcal{L}_{uni}(f;\tau) = \mathop{\mathbb{E}}_{x\sim p_{data}}\left[\log \mathop{\mathbb{E}}_{x'\sim p_{data}}\left(e^{f(x')^T f(x)/\tau}\right)\right]\,. \tag{A.7}
$$

In particular, we can show the following relation:

Corollary A.2:

$$
\mathbb{E}_{z\sim p_c(z)}\left[H\left(p(\cdot|z), q_h(\cdot|z)\right)\right] = \mathcal{L}_{align}(f;\tau) + \mathcal{L}_{uni}(f;\tau) + \log|\mathcal{C}|\,.
$$

Thm. A.1 guarantees that representations learned via Algorithms 1 and 2 recover the latent space *locally*, i.e., recover the relationship between close-by points within the same class. What can we say about their ability to recover *global* structure, such as the relationship between classes?

To answer this question, we analyze which geometric assumptions are implicitly encoded in the contrastive loss $\mathcal{L}_{contr}$ via $\mathcal{L}_{align}$ and $\mathcal{L}_{uni}$. We find that $\mathcal{L}_{contr}$ encourages representations that recover a homogeneous reconstructed latent space, where the latent classes are well-concentrated and uniformly distributed in the latent space:

Thm. A.1 and Corr. A.2 suggest that minimizing $\mathcal{L}_{contr}$ implies a small uniformity loss ($\mathcal{L}_{uni}$) and alignment loss ($\mathcal{L}_{align}$). A more detailed analysis reveals that a small uniformity loss ensures that the angular separation between classes is not too uneven, favouring a distribution close to the uniform distribution in the true latent space. A small alignment loss implies that the angular sizes of the classes are not too large, i.e., that the classes are well concentrated. This can be seen with the following arguments:

1) A small uniformity loss $\mathcal{L}_{uni}(f)$ ensures that the angular separation between classes is not too uneven. In particular, note that

$$\begin{aligned}\mathcal{L}_{uni}(f) &= \mathbb{E}_{\substack{c\sim p_c \\ x\sim p_{data}(\cdot|c)}}\left[\log \mathbb{E}_{\substack{c'\sim p_c \\ x^-\sim p_{data}(\cdot|c')}}\left(e^{f(x)^T f(x^-)}\right)\right] \\ &= \rho\mathbb{E}_{\substack{c\sim p_c \\ (x,x^-)\sim p_{pos}(\cdot|c)}}\left[\log \mathbb{E}\left(e^{f(x)^T f(x^-)}\right)\right] + (1-\rho)\mathbb{E}_{\substack{c,c'\sim p_c \\ x\sim p_{data}(\cdot|c) \\ x^-\sim p_{data}(\cdot|c')}}\left[\log \mathbb{E}\left(e^{f(x)^T f(x^-)}\right)\right] \\ &= \rho\mathbb{E}_{(c;x,x^-)}\left[\log \mathbb{E}\left(e^{\tilde{z}^T \tilde{z}^-}\right)\right] + (1-\rho)\mathbb{E}_{(c,x),(c',x^-)}\left[\log \mathbb{E}\left(e^{\tilde{z}^T \tilde{z}^-}\right)\right] \\ &= \rho\mathbb{E}_{(c,x)}\left[\log e^{f(x)^T \mu_{\tilde{c}}}\right] + (1-\rho)\mathbb{E}_{(c,c')}\left[\log e^{\mu_{\tilde{c}}^T \mu_{\tilde{c}'}}\right] .\end{aligned}$$

Notably, a small uniformity loss ensures that the angular separation is not too large for any two classes, implying distribution close to the uniform distribution in the true latent space.

2) A small alignment loss implies that the angular sizes of the classes are not too large, i.e., that the classes are well concentrated. For this, note that

$$\mathcal{L}_{align}(f) = \mathbb{E}_{\substack{c\sim p_c \\ (x,x^+)\sim p_{pos}(\cdot|c)}}\left[e^{f(x)^T f(x^+)}\right] = \mathbb{E}_{(c;x,x^+)}\left[e^{\tilde{z}^T \tilde{z}^+}\right] = \mathbb{E}_{(c;x)}\left[e^{\tilde{z}^T \mu_{\tilde{c}}}\right] .$$

This suggests that the classical contrastive loss $\mathcal{L}_{contr}$ may not capture heterogenity between classes or low-dimensional structure in latent space well. Such geometric information could be uncovered by sampling instances that are informative about the underlying structure. This observation motivates the design of active or adversarial sampling strategies that pick informative positive pairs, with the hope of incorporating more geometric information into the training process (Algorithms 1 and 2).

C. Recovering latent class structure

In this section we give theoretical evidence for the quality of the representation functions trained with Algorithm 1. Specifically, we analyze how well the representations recover the underlying latent class structure. We focus on the comparison of passive an active sampling strategies, i.e., *random augmentation* (Algorithm 1(i)) and *active selection* (Algorithm 1(ii)). Both approaches sample positive instances via random augmentation. However, while the random augmentation approach adds each of the sampled instances to the training data, the active selection approach rejects instances that are not close to the decision boundary and therefore less informative. We provide a theoretical argument in favour of such an approach.

Sampling from the Hypersphere On a k-dimensional hypersphere with radius r, caps are characterized by the polar angle θ, measured as the angle between rays from the center of the sphere to the pole and the base of the cap. The area of the spherical cap is given by [S, 2011] (assuming $\theta < \frac{\pi}{2}$)

$$A_k^{\text{cap}}(r,\theta) = \frac{1}{2}A_k(r)I_{\sin^2\theta}\left(\frac{k+1}{2},\frac{1}{2}\right) , \tag{A.8}$$

where

$$A_k(r) = \frac{2\pi^{k/2}}{\Gamma\left(\frac{k}{2}\right)}r^{k-1} . \tag{A.9}$$

denotes the area of the whole hypersphere, $\Gamma(y)$ the gamma function and $I_y(a,b)$ the incomplete beta function, both of which can be computed numerically. The factor $I_{\sin^2\theta}\left(\frac{n+1}{2},\frac{1}{2}\right)$ corresponds to the probability of receiving a point in the cap when sampling uniformly at random from the hypersphere.

Guarantees for passive sampling We now assume that we have trained a representation function $\hat{f}$ with Algorithm 1(i) and that we have trained a classifier $\hat{q}(x) = W\hat{f}(x)$ in the reconstructed latent space $\tilde{\mathcal{Z}}$. We want to derive error bounds for the representation function $\hat{f}$ in terms of its ability to recover the latent class structure. We assume that an $\hat{f}$ is an α-accurate minimizer[1] of the unsupervised training objective

$$\mathcal{L}_{un}(f) := \mathbb{E}_{(x,x^+,\{x_i^-\}_{i=1}^m)}\left[l\left(\{f(x)^T f(x^+) - \max_{1\le i\le m} f(x)^T f(x_i^-)\}_{i=1}^m\right)\right] , \tag{A.10}$$

which can be empirically estimated over a sample $\mathcal{D} = \{(x_j, x_j^+, x_{j1}^-, \ldots, x_{jm}^-)\}_{i=1}^n$ that contains n positive pairs and mn negative instances:

$$\hat{\mathcal{L}}_{un}(f) := \frac{1}{n}\sum_{j=1}^n l\left(\{f(x)^T f(x^+) - \max_{1\le i\le m} f(x)^T f(x_i^-)\}_{i=1}^m\right) . \tag{A.11}$$

[1] $z \in C$ implies $(g\circ f)(z) \in \tilde{C}$ with probability $1-\alpha$

We further define a supervised *margin loss* with respect to classifiers $q : \tilde{\mathcal{Z}} \to \mathbb{R}^{|\mathcal{C}|}$. For this, we first define a *margin function*

$$\gamma_q(f(x), c) := q_c(f(x)) - \max_{c' \neq c} q_{c'}(f(x)) \ . \tag{A.12}$$

With respect to a fixed margin $\gamma > 0$, we can define a loss function

$$\Phi_\gamma(v) := \min\left(1, \max\left(0, 1 - \frac{v}{\gamma}\right)\right) = \begin{cases} 1, & v \leq 0 \\ 1 - \frac{v}{\gamma}, & 0 \leq v \leq \gamma \\ 0, & \gamma \leq v \end{cases} , \tag{A.13}$$

and a margin loss

$$L_{class}(\mathcal{C}, q) := \mathbb{E}_{\substack{c \sim p_c \\ x \sim p_{data}(\cdot|c)}} \left[\Phi_\gamma\left(\gamma_q(f(x), c)\right)\right] \ . \tag{A.14}$$

We can empirically estimate L_{class} over $\mathcal{D}$ as

$$\hat{L}_{class}(\mathcal{C}, q, \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \Phi_\gamma(\gamma_q(f(x_i), c_i)) \ . \tag{A.15}$$

We can derive the following error bound:

Theorem A.3: Let $\gamma > 0$ (fixed) denote the margin in the true latent space and $\tilde{\gamma} \leq \gamma$ the margin in the reconstructed latent space. For any $\delta > 0$ we have with probability at least $1 - \delta$ that

$$L_{class}(f) \leq \frac{1}{1 - m\rho}\left(\hat{\mathcal{L}}_{un}(f) - m\rho\right) + \frac{1}{1 - m\rho}\left(4L_\alpha \mathcal{R}_{\mathcal{D}}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2|\mathcal{D}|}}\right)$$

for all $f \in \mathcal{F}$. Here, $\mathcal{R}_{\mathcal{D}}(\mathcal{F})$ denotes the Rademacher complexity of the function class $\mathcal{F}$ and $L_\alpha \leq \frac{2}{\tilde{\gamma}}(1 - 2\alpha)$ and ρ the probability of sampling a false negative instance, i.e., the probability of sampling twice from the same class.
To compute the Rademacher complexity, we restrict f to the sample set $\mathcal{D}$ (with $|\mathcal{D}| =: n$)

$$f\big|_{\mathcal{D}} = \{\big(f(x_j), f(x_j^+), f(x_{1j}^-), \ldots, f(x_{mj}^-)\big)\}_{j=1}^n \subseteq \mathbb{R}^{3dmn} \ .$$

The Rademacher complexity is then given as

$$\mathcal{R}_{\mathcal{D}}(\mathcal{F}) = \mathbb{E}_{\sigma \sim \{\pm 1\}^{3dmn}}\left[\sup_{f \in \mathcal{F}} \langle \sigma,\ f\big|_{\mathcal{D}} \rangle\right] \ . \tag{A.16}$$

Remark A.4: The proof of Theorem A.3 is similar to [Saunshi et al., 2019, Theorem 4.1]. We give a bound with respect to an α-accurate representation function f, which approximates an optimal representation function f^* that recovers g^{-1} up to orthogonal linear transformation.
The proof of Theorem A.3 relies on two auxiliary lemmas, which we state first. Note that Φ_γ is $\frac{1}{\gamma}$-Lipschitz. This ensures the validity of the following standard bound for learning with noisy labels [Natarajan et al., 2013]) for the unsupervised contrastive training loss $\mathcal{L}_{un}$:

Lemma A.5 ([Natarajan et al., 2013]): For any fixed margin $\tilde{\gamma} > 0$ and a $\delta > 0$ we have with probability at least $1 - \delta$ over a ground truth set $\mathcal{D}$ for all $f \in \mathcal{F}$

$$\mathcal{L}_{un}(f) \leq \widehat{\mathcal{L}}_{un}(\mathcal{D}, f) + 4L_\alpha \mathcal{R}_{\mathcal{D}}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2|\mathcal{D}|}} \ ,$$

where $\mathcal{R}_{\mathcal{D}}(\mathcal{F})$ denotes the Rademacher complexity of the function class $\mathcal{F}$ and $L_\alpha \leq \frac{2}{\tilde{\gamma}}(1 - 2\alpha)$.
We further need the following result, which relates the unsupervised contrastive loss $\mathcal{L}_{un}$ to the supervised loss L_{class}:

Lemma A.6: For any $f \in \mathcal{F}$ we have

$$L_{class}(\mathcal{C}, f) \leq \frac{1}{1 - m\rho}\left(\mathcal{L}_{un}(f) - m\rho\right) \ .$$

The lemma is a slight generalization of [Saunshi et al., 2019, Lemma 4.3].

$$\begin{aligned}
\mathcal{L}_{un}(f) &= \mathbb{E}_{c,c'\sim p_c,\ \substack{(x,x^+)\sim p_{pos}(\cdot|c)\\ x^-\sim p_{data}(\cdot|c')}} \left[\Phi_\gamma\left(f(x)^T f(x^+) - \max_{1\le i\le m} f(x)^T f(x^-)\right)\right] \\
&\ge \mathbb{E}_{\substack{c,c'\sim p_c\\ x\sim p_{data}(\cdot|c')}} \left[\Phi_\gamma\left(f(x)^T \mu_c - \max_{1\le i\le m} f(x)^T \mu_{c'}\right)\right] \\
&= (1-m\rho)\mathbb{E}_{\substack{c\neq c'\\ x\sim p_{data}(\cdot|c')}} \left[\Phi_\gamma\left(f(x)^T \mu_c - \max_{1\le i\le m} f(x)^T \mu_{c'}\right)\right] + m\rho \\
&= (1-m\rho)L_{class}(\mathcal{C}, f) + m\rho \ .
\end{aligned}$$

Then Theorem A.3 follows from combining Lemmas A.5 and A.6.

Comparison of passive and active sampling For our analysis of the active and passive sampling strategies, we pick two classes and focus on the problem of learning a separator between them. This reduces the problem to a binary classification task. Formally, let $|\mathcal{C}| = 2$ and

$$\begin{aligned}
C_+ &:= \{z|(z,1) \in \mathbb{S}^{k-1} \times \mathcal{C}\} \subseteq \mathcal{Z} \\
C_- &:= \{z|(z,-1) \in \mathbb{S}^{k-1} \times \mathcal{C}\} \subseteq \mathcal{Z} \ .
\end{aligned}$$

We assume that C_+, C_- are linearly separable with margin γ. Our goal is to to learn a classifier in $\tilde{\mathcal{Z}}$ that recovers the latent structure defined by $\{C_+, C_-\} \subseteq \mathcal{Z}$, i.e., we want to learn a classifier that separates $\{\tilde{C}_+, \tilde{C}_-\} \subseteq \tilde{\mathcal{Z}}$.

We want to compare the *passive* and *active* sampling approaches for positive instances. Note that the active sampling approach resembles classical active learning techniques for binary classification [Balcan et al., 2007], which allows us to utilize theoretical results from this literature. We show that the active selection approach reduces the number of samples that we need to add in each round (m_t), in comparison with the amount of samples needed, if positive instances are sampled passively. This indicates that representation functions can be trained more efficiently via active selection. Formally, we show the following result:

Theorem A.7: For any $\delta, \epsilon > 0$, we can recover the class structure up to error ϵ with probability $1-\delta$ with (1) sample complexity $m = O(\frac{d}{\epsilon})$ for random augmentations (Algorithm 1(i)) and (2) sample complexity $m = O\left(d^{3/2}\log(\frac{1}{\epsilon})\right)$ for active selection (Algorithm 1(ii), with rejection threshold $a_t = \frac{\pi}{2^{t-1}}$).

The proof follows results on active learning for binary classification [Balcan et al., 2007]. We outline the proof below. We will make use of the following standard result (see, e.g., [Anthony et al., 1999]):

Theorem A.8: Let H denote a set of functions from $\tilde{\mathcal{Z}}$ to ± 1 with finite VC dimension $V \ge 1$. Let D be a an arbitrary fixed distribution on $\tilde{\mathcal{Z}} \times \{\pm 1\}$. Then there exists a universal constant C, such that for any $\epsilon, \delta > 0$, if we draw a sample of size $N(\epsilon,\delta) = \frac{1}{\epsilon}(4V\log(\frac{1}{\epsilon}) + 2\log(\frac{2}{\delta}))$ from D, all hypotheses with error $\ge \epsilon$ are inconsistent with the data with probability $1-\delta$. (Thm. A.7)

(1) follows from Thm. A.8. For (2), we first note that the error of a classifier q can be measured with respect to w as

$$\mathrm{err}(w) = \frac{\arccos(w \cdot w^*)}{\pi} \ ,$$

where w^* denotes an optimal seperator for the data. With this, $\mathrm{err}(w) \le \epsilon$ implies $\|w - w^*\|_2 \le \epsilon\pi$. We want to show via induction that m_t samples are sufficient to obtain a classifier with $\mathrm{err}(w_t) \le 2^{-t}$ with probability $1 - \delta(1 - 1/(t+1))$. The case $t = 1$ follows again from Thm. A.8, i.e., with $m_1 = O(k + \log(1/\delta))$ we have $\mathrm{err}(w_1) \le \frac{1}{2}$ with probability $1 - \frac{\delta}{2}$. We assume that the claim is true for some t (induction hypothesis) and want to prove the claim for $t+1$. For an anchor point $x \sim p_{data}(\cdot|c)$ we can define the following two sets:

$$\begin{aligned}
S_1^t(x) &:= \{f(x^+) \in \tilde{\mathcal{Z}} : |w_t \cdot f(x^+)| \le a_t\} \\
S_2^t(x) &:= \{f(x^+) \in \tilde{\mathcal{Z}} : |w_t \cdot f(x^+)| > a_t\} \ .
\end{aligned}$$

In round t, we can write the error of the classifier q_t as

$$\mathrm{err}(w_t) = \mathrm{err}(w_t|S_1^t)P(S_1^t) + \mathrm{err}(w_t|S_2^t)P(S_2^t) \ ,$$

where $P(S)$ denotes the probability of sampling from S and

$$\mathrm{err}(w|S) := \mathrm{Prob}((w \cdot f(x))(w^* f(x)) < 0 | x \in S) \ .$$

Consider a classifier $\hat{w}$ that is consistent with $\mathcal{D}_t$. By the induction hypothesis both w_t and $\hat{w}$ have error at most 2^{-t}, i.e., $\mathrm{err}(\hat{w}) \leq 2^{-t}$ and $\mathrm{err}(w_t) \leq 2^{-t}$ with probability $1 - \delta(1 - 1/(t+1))$. This implies

$$\|w_t - w^*\|_2 \leq 2^{-t}\pi$$
$$\|\hat{w} - w^*\|_2 \leq 2^{-t}\pi \ .$$

Now let $\tilde{x} \in S_2$. Then

$$(w_t \cdot \tilde{x})(\hat{w} \cdot \tilde{x}) > 0$$
$$(w_t \cdot \tilde{x})(w^* \cdot \tilde{x}) > 0 \ ,$$

which implies $\mathrm{err}(\hat{w}|S_2) = 0$.

We can compute the probability $\mathrm{Prob}(\mathrm{S}_1^{\mathrm{t}})$ of sampling from the region S_1^t (close to the decision boundary) with respect to the acceptance threshold as

$$\mathrm{Prob}(S_1) \leq \frac{a_t\sqrt{k}}{2\pi} \ .$$

The proof follows from a geometric calculation and can be found in [Balcan et al., 2007, Lemma 4]. Inserting this above, we have

$$\mathrm{err}(\hat{w}) \leq 2^{-(t-1)}\sqrt{4\pi k} \cdot \mathrm{err}(\hat{w}|S_1) \ ,$$

which holds for all $\hat{w}$ consistent with $\mathcal{D}_t$. By construction, we add m_t samples (from S_1) to $\mathcal{D}_t$ in iteration t. By Thm. A.8 there exists a constant, such that with probability $1 - \frac{\delta}{t^2+t}$ we have

$$\mathrm{err}(\hat{w}|S_1) \leq \frac{1}{4\sqrt{4\pi k}}$$

for all $\hat{w}$ consistent with $\mathcal{D}_{t+1}$. This implies $\mathrm{err}(\hat{w}) \leq 2^{-(t+1)}$ for all $\hat{w}$ consistent with $\mathcal{D}_{t+1}$ and therefore the claim as $\mathrm{err}(w_{t+1}) \leq 2^{-(t+1)}$.

D. Experiments: Informative positives for latent classes

In the main text (sec. 4.2, Tab. 1), we present results for three sampling techniques.

Experimental setup. Throughout the experiments, we defined class-conditioned distributions with Gaussian noise of size 0.3. We sampled positive instances (*baseline*) and candidate positive instances (*active* and *double-active*) with perturbations of size 0.2. The experimental setup was a standard MLP, which was trained with learning rate 0.0001. The hyperparameters in the reported experimental results are listed in the main text. We investigate the following four scenarios, which employ different passive and active sampling strategies:

1) **Baseline:** Anchor points and positive instances are sampled from the baseline prior (resembling Alg. 1(i)).
2) **Active:** Anchor points are sampled from the baseline. Candidate positive instances are sampled via perturbation in the observation space until one is found whose image lies in the acceptance region, i.e., within some ϵ of the decision boundary of q_{t-1} in the reconstructed latent space. This resembles the *active selection* strategy (Alg. 1(ii)).
3) **Double-active:** Anchor points sampled with a bias towards the region near the decision boundary of q_{t-1} in the reconstructed latent space. Candidate positive instances are sampled via perturbation in the observation space until one is found whose image lies in the acceptance region. With the additional preference for sampling near the decision boundary, this can be seen as closer to the *adversarial augmentation* idea (Alg. 1(iii)).

E. Experiments: Targeted augmentation in latent space

Experimental setup. Tab. 2 in sec. 5.2 gives experimental results for the second experimental setting. Again, our experimental setup was a standard MLP, which was trained with learning rate 0.0001. We investigate the following sampling strategies:

1) **Baseline:** Equal-sized random perturbations of all latents.
2) **Info-active:** Small, targeted perturbations of informative latents and independent random sampling of all others.
3) **Active:** Small, targeted perturbations of informative latents and larger, targeted perturbations of all others.
4) **Class-preserving:** Assume access to a class-preserving transformation, which is used to augment the anchor to generate a positive sample from the same class. This is an idealized setting, which can be thought of as a supervised approach. It is included for comparison; unsupervised approaches are not expected to match its accuracy.

An important hyperparameter in the experiments is the *perturbation scale*, i.e., the size of the perturbations (Gaussian noise) in the "informative" and "noisy" (i.e., uninformative) dimensions. In the *baseline* approach, the perturbation scale is 0.3 across all latents. In the *Info-active* approach, we apply perturbations of 0.3 to the informative latent and 9.0 to the noisy latents. In

the *active* approach, we apply again perturbations of 0.3 to the informative latents, but only perturbations of 0.9 to the noisy latents.

Hyperparameter choice. We investigate the impact of the choice of the perturbation scale for the noisy latents on the knn accuracy. Fig. 3 shows results for perturbation scales of $0.3 - 9.0$ for different hyperparameters. We notice that even small differences in the perturbation scales between informative and noisy latents (i.e., small targeted perturbations of the informative latents) improve over the baseline.

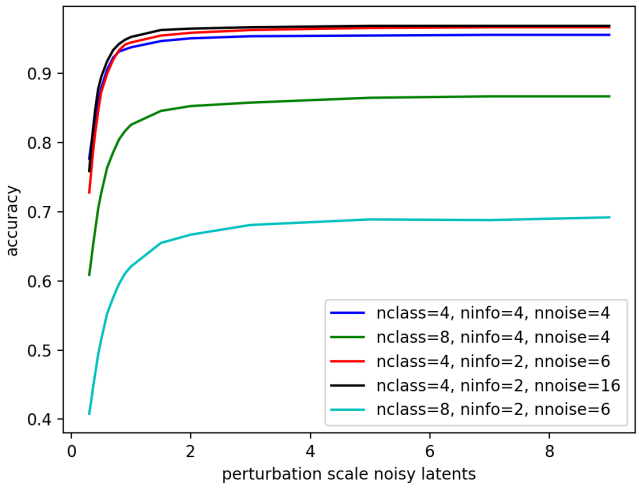

Fig. 3: Perturbation scale for "uninformative" or noisy latents.