A. Appendix

A.1. Extended Theory for Hyperspheres

A.1.1. Assumptions

Generative Process Let the generator $g : \mathcal{Z} \to \mathcal{X}$ be an injective function between the two spaces $\mathcal{Z} = \mathbb{S}^{N-1}$ and $\mathcal{X} \subseteq \mathbb{R}^K$ with $K \geq N$. We assume that the marginal distribution $p(\mathbf{z})$ over latent variables $\mathbf{z} \in \mathcal{Z}$ is uniform:

$$p(\mathbf{z}) = \frac{1}{|\mathcal{Z}|}. \tag{8}$$

Further, we assume that the conditional distribution over positive pairs $p(\tilde{\mathbf{z}}|\mathbf{z})$ is a von Mises-Fisher (vMF) distribution

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \tag{9}$$

$$\text{with} \quad C_p := \int e^{\kappa \boldsymbol{\eta}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{10}$$

where κ is a parameter controlling the width of the distribution and $\boldsymbol{\eta}$ is any vector on the hypersphere. Finally, we assume that during training one has access to observations $\mathbf{x}$, which are samples from these distributions transformed by the generator function g.

Model Let $f : \mathcal{X} \to \mathbb{S}_r^{N-1}$, where $\mathbb{S}_r^{N-1}$ denotes a hypersphere with radius r. The parameters of this model are optimized using contrastive learning. We associate a conditional distribution $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ with our model f through $h = f \circ g$ and

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}$$
$$\text{with} \quad C_q(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top \, h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{11}$$

where $C_q(\mathbf{z})$ is the partition function and $\tau > 0$ is a scale parameter.

A.1.2. Proofs for Sec. 3

We begin by recalling a result of Wang & Isola (2020), where the authors show an asymptotic relation between the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ and two loss functions, the *alignment* loss $\mathcal{L}_{\mathsf{align}}$ and the *uniformity* loss $\mathcal{L}_{\mathsf{uni}}$:

Proposition A (Asymptotics of $\mathcal{L}_{\mathsf{contr}}$, Wang & Isola, 2020)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M$$
$$= \mathcal{L}_{\mathsf{align}}(f;\tau) + \mathcal{L}_{\mathsf{uni}}(f;\tau), \tag{12}$$

where

$$\mathcal{L}_{\mathsf{align}}(f;\tau) := -\frac{1}{\tau} \mathop{\mathbb{E}}_{(\tilde{\mathbf{z}},\mathbf{z})\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[(f\circ g)(\mathbf{z})^\mathsf{T}(f\circ g)(\mathbf{z})\right]$$
$$\mathcal{L}_{\mathsf{uni}}(f;\tau) := \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{(f\circ g)(\tilde{\mathbf{z}})^\mathsf{T}(f\circ g)(\mathbf{z})/\tau}\right]\right]. \tag{13}$$

Proof. See Theorem 1 of Wang & Isola (2020). Note that they originally formulated the losses in terms of observations $\mathbf{x}$ and not in terms of the latent variables $\mathbf{z}$. However, this modified version simplifies notation in the following. □

Based on this result, we show that the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ asymptotically converges to the cross-entropy between the ground-truth conditional p and our assumed model conditional distribution q_h, up to a constant. This is notable, because given the correct model specification for q_h, it is well-known that the cross-entropy is minimized iff $q_h = p$, i.e., the ground-truth conditional distribution and the model distribution will match.

Theorem 1 ($\mathcal{L}_{\mathsf{contr}}$ converges to the cross-entropy between latent distributions)**.** *If the ground-truth marginal distribution p is uniform, then for fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| =$$
$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right] \tag{14}$$

where H is the cross-entropy between the ground-truth conditional distribution p over positive pairs and a conditional distribution q_{h} parameterized by the model f, and $C_h(\mathbf{z}) \in \mathbb{R}^+$ is the partition function of q_{h} (see Appendix A.1.1):

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\mathbf{z})^{-1} e^{h(\tilde{\mathbf{z}})^\mathsf{T} h(\mathbf{z})/\tau}$$
$$\textit{with} \quad C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\mathsf{T} h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}}. \tag{15}$$

Proof. The cross-entropy between the conditional distributions p and q_h is given by

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right] \tag{16}$$
$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})} \left[-\log q_h(\tilde{\mathbf{z}}|\mathbf{z})\right]\right] \tag{17}$$
$$= \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[-\frac{1}{\tau} h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) + \log C_h(\mathbf{z})\right] \tag{18}$$
$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z})\right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log C_h(\mathbf{z})\right]. \tag{19}$$

Using the definition of C_h in Eq. (15) we obtain

$$= -\frac{1}{\tau}\mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})}\left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z})\right] \tag{20}$$

$$+\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\int_{\mathcal{Z}} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}\,\mathrm{d}\tilde{\mathbf{z}}\right]. \tag{21}$$

By assumption the marginal distribution is uniform, i.e., $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$. We expand by $|\mathcal{Z}||\mathcal{Z}|^{-1}$ and estimate the integral by sampling from $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$, yielding

$$= -\frac{1}{\tau}\mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})}\left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z})\right] \tag{22}$$

$$+\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log|\mathcal{Z}|\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}\left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}\right]\right] \tag{23}$$

$$= -\frac{1}{\tau}\mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})}\left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z})\right] \tag{24}$$

$$+\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}\left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}\right]\right] + \log|\mathcal{Z}|. \tag{25}$$

By inserting the definition $h = f \circ g$,

$$= -\frac{1}{\tau}\mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})}\left[(f\circ g)(\tilde{\mathbf{z}})^\top (f\circ g)(\mathbf{z})\right] \tag{26}$$

$$+\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}\left[e^{(f\circ g)(\tilde{\mathbf{z}})^\top (f\circ g)(\mathbf{z})/\tau}\right]\right] \tag{27}$$

$$+\log|\mathcal{Z}|, \tag{28}$$

we can identify the losses introduced in Proposition A,

$$= \mathcal{L}_{\mathsf{align}}(f;\tau) + \mathcal{L}_{\mathsf{uni}}(f;\tau) + \log|\mathcal{Z}|, \tag{29}$$

which recovers the original alignment term and the uniformity term for maximimizing entropy by means of a von Mises-Fisher KDE up to the constant $\log|\mathcal{Z}|$. According to Proposition A this equals

$$= \lim_{M\to\infty}\mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M + \log|Z|, \tag{30}$$

which concludes the proof. □

Proposition 1 (Minimizers of the cross-entropy maintain the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, $\tau > 0$ and consider the ground-truth conditional distribution of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}\exp(\kappa\tilde{\mathbf{z}}^\top\mathbf{z})$. Let h map onto a hypersphere with radius $\sqrt{\tau\kappa}$.*[4] *Consider the conditional distribution q_h parameterized by the model, as defined above in Theorem 1, where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h is a minimizer of the cross-entropy $\mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_h(\tilde{\mathbf{z}}|\mathbf{z})]$, then $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_h(\tilde{\mathbf{z}}|\mathbf{z})$ and $\forall\mathbf{z},\tilde{\mathbf{z}} : \kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$.*

[4]Note that in practice this can be implemented as a learnable rescaling operation of the network f.

Proof. By assumption, $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h—in particular, for $h(\mathbf{z}) = \sqrt{\tau\kappa}\mathbf{z}$. The global minimum of the cross-entropy between two distributions is reached if they match by value and have the same support. Thus, this means

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_h(\tilde{\mathbf{z}}|\mathbf{z}). \tag{31}$$

This expression also holds true for $\tilde{\mathbf{z}} = \mathbf{z}$; additionally using that h maps from a unit hypersphere to one with radius $\sqrt{\tau\kappa}$ yields

$$p(\mathbf{z}|\mathbf{z}) = q_h(\mathbf{z}|\mathbf{z}) \tag{32}$$

$$\Leftrightarrow\quad C_p^{-1}e^{\kappa\mathbf{z}^\top\mathbf{z}} = C_h(\mathbf{z})^{-1}e^{h(\mathbf{z})^\top h(\mathbf{z})/\tau} \tag{33}$$

$$\Leftrightarrow\quad C_p^{-1}e^{\kappa} = C_h(\mathbf{z})^{-1}e^{\kappa} \tag{34}$$

$$\Leftrightarrow\quad C_p = C_h. \tag{35}$$

As the normalization constants are identical we get for all $\mathbf{z},\tilde{\mathbf{z}}\in\mathcal{Z}$

$$e^{\kappa\mathbf{z}^\top\tilde{\mathbf{z}}} = e^{h(\mathbf{z})^\top h(\tilde{\mathbf{z}})} \Leftrightarrow \kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}}). \tag{36}$$

□

Proposition 2 (Extension of the Mazur-Ulam theorem to hyperspheres and the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$. If $h : \mathcal{Z}\to\mathcal{Z}$ maintains the dot product up to a constant factor, i.e., $\forall\mathbf{z},\tilde{\mathbf{z}}\in\mathcal{Z} : \kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$, then h is an orthogonal linear transformation.*

Proof. As h maintains the dot product up to a factor, this also holds true if one rotates the coordinate system by an arbitrary rotation matrix $\mathbf{R}\in\mathrm{SO}(N)$. Thus, we get

$$\forall\mathbf{R}\in\mathrm{SO}(N),\forall\mathbf{z},\tilde{\mathbf{z}}\in\mathcal{Z} : \tag{37}$$

$$\kappa\mathbf{z}^\top\mathbf{R}^\top\mathbf{R}\tilde{\mathbf{z}} = h(\mathbf{R}\mathbf{z})^\top h(\mathbf{R}\tilde{\mathbf{z}}). \tag{38}$$

We consider the partial derivatives w.r.t. $\mathbf{z}$ and obtain:

$$\forall\mathbf{R}\in\mathrm{SO}(N)\ \forall\mathbf{z},\tilde{\mathbf{z}}\in\mathcal{Z} : \tag{39}$$

$$\kappa\tilde{\mathbf{z}} = \mathbf{R}\mathbf{J}_h^\top(\mathbf{R}\mathbf{z})h(\mathbf{R}\tilde{\mathbf{z}}). \tag{40}$$

We can recover the initial dot product by multiplying both sides of the equation with $\mathbf{z}^\top$ to obtain

$$\forall\mathbf{R}\in\mathrm{SO}(N)\ \forall\mathbf{z},\tilde{\mathbf{z}}\in\mathcal{Z} : \tag{41}$$

$$\kappa\mathbf{z}^\top\tilde{\mathbf{z}} = \mathbf{z}^\top\mathbf{R}\mathbf{J}_h^\top(\mathbf{R}\mathbf{z})h(\mathbf{R}\tilde{\mathbf{z}}) \tag{42}$$

$$= h(\mathbf{R}\tilde{\mathbf{z}})^\top\mathbf{J}_h(\mathbf{R}\mathbf{z})\mathbf{R}^\top\mathbf{z}. \tag{43}$$

From here, we take the partial derivative on both sides, this time w.r.t. $\tilde{\mathbf{z}}$, yielding

$$\forall\mathbf{R}\in\mathrm{SO}(N)\ \forall\mathbf{z},\tilde{\mathbf{z}}\in\mathcal{Z} : \tag{44}$$

$$\kappa\mathbf{z} = [\mathbf{R}\mathbf{J}_h(\mathbf{R}\tilde{\mathbf{z}})\mathbf{J}_h^\top(\mathbf{R}\mathbf{z})\mathbf{R}^\top]\mathbf{z}. \tag{45}$$

Multiplying with $\mathbf{R}^\top$ from the left and defining $\mathbf{z}' := \mathbf{R}^\top \mathbf{z}$ gives

$$\forall \mathbf{R} \in \mathrm{SO}(N)\ \forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}: \tag{46}$$

$$\kappa \mathbf{z}' = [\mathbf{J}_h(\mathbf{R}\tilde{\mathbf{z}})\mathbf{J}_h^\top(\mathbf{R}^2\mathbf{z}')]\mathbf{z}'. \tag{47}$$

We define a transform from $(\mathbf{R}, \mathbf{z}, \tilde{\mathbf{z}})$ to $(\mathbf{a}, \mathbf{b}, \mathbf{z}')$: First, we select $\mathbf{R}$ and $\mathbf{z}$ s.t. $\mathbf{z}' = \mathbf{R}^T\mathbf{z}$ and $\mathbf{b} = \mathbf{R}\mathbf{z} = \mathbf{R}^2\mathbf{z}'$. Then, we select $\tilde{\mathbf{z}}$ s.t. $\mathbf{a} = \mathbf{R}\tilde{\mathbf{z}}$. With this transform, we rewrite the aforementioned equation and obtain:

$$\forall \mathbf{a}, \mathbf{b}, \mathbf{z}' \in \mathcal{Z}: \kappa \mathbf{z}' = [\mathbf{J}_h(\mathbf{a})\mathbf{J}_h(\mathbf{b})^\top]\mathbf{z}', \tag{48}$$

which can only be satisfied iff

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z}: \mathbf{J}_h(\mathbf{a})\mathbf{J}_h(\mathbf{b})^\top = \kappa \mathbf{I}. \tag{49}$$

By evaluating this expression for $\mathbf{a} = \mathbf{b}$ we get

$$\forall \mathbf{b} \in \mathcal{Z}: \mathbf{J}_h(\mathbf{b})^\top = \kappa \mathbf{J}_h^{-1}(\mathbf{b}). \tag{50}$$

Inserting this property again in the previous expression yields

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z}: \mathbf{J}_h(\mathbf{a})\kappa\mathbf{J}_h(\mathbf{b})^{-1} = \kappa \mathbf{I}, \tag{51}$$

and finally:

$$\forall \mathbf{a}, \mathbf{b} \in \mathcal{Z}: \mathbf{J}_h(\mathbf{a}) = \mathbf{J}_h(\mathbf{b}) \tag{52}$$

$$\forall \mathbf{a} \in \mathcal{Z}: \kappa \mathbf{J}_h(\mathbf{a})^{-1} = \mathbf{J}_h^\top(\mathbf{a}). \tag{53}$$

□

Taking all of this together, we can now prove Theorem 2:

Theorem 2. *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, the ground-truth marginal be uniform, and the conditional a vMF distribution (cf. Eq. 2). Let the mixing function g be differentiable and injective. If the assumed form of q_h, as defined above, matches that of p, and if f is differentiable and minimizes the CL loss as defined in Eq. (1), then for fixed $\tau > 0$ and $M \to \infty$, $h = f \circ g$ is linear, i.e., f recovers the latent sources up to orthogonal linear transformations.*

Proof. As f minimzes the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ we can apply Theorem 1 to see that f also minimizes the cross-entropy between $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z})$ for any point $\mathbf{z}$ on $\mathcal{Z}$. This means, we can apply Proposition 1 to show that the concatenation $h = f \circ g$ is an isometry with respect to the dot product. Finally, according to Proposition 2, h must then be an orthogonal linear transformation on the hypersphere. Thus, f recovers the latent sources up to orthogonal linear transformations, concluding the proof. □

A.2. Extension of theory to subspaces of $\mathbb{R}^N$

Here, we show how one can generalize the theory above from $\mathcal{Z} = \mathbb{S}^{N-1}$ to $\mathcal{Z} \subseteq \mathbb{R}^N$. Under mild assumptions regarding the ground-truth conditional distribution p and the model distribution q_h, we prove that all minimizers of the cross-entropy between p and q_h are linear functions, if $\mathcal{Z}$ is a convex body. Note that the hyperrectangle $[a_1, b_1] \times \ldots \times [a_N, b_N]$ is an example of such a convex body.

A.2.1. Assumptions

First, we restate the core assumptions for this proof. The main difference to the assumptions for the hyperspherical case above is that we assume different conditional distributions: instead of rotation-invariant von Mises-Fisher distributions, we use translation-invariant distributions (up to restrictions determined by the finite size of the space) of the exponential family.

Generative process Let $g: \mathcal{Z} \to \mathcal{X}$ be an injective function between the two spaces $\mathcal{Z} \subseteq \mathbb{R}^N$ and $\mathcal{X} \subseteq \mathbb{R}^K$ with $K \geq N$ and where $\mathcal{Z}$ is a convex body (e.g., a hyperrectangle). Further, let the marginal distribution be uniform, i.e., $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$. We assume that the conditional distribution over positive pairs $p(\tilde{\mathbf{z}}|\mathbf{z})$ is an exponential distribution

$$\begin{aligned} p(\tilde{\mathbf{z}}|\mathbf{z}) &= C_p^{-1}(\mathbf{z}) e^{-\lambda \delta(\tilde{\mathbf{z}}, \mathbf{z})} \\ \text{with} \quad C_p(\mathbf{z}) &:= \int e^{-\lambda \delta(\mathbf{z}, \tilde{\mathbf{z}})}\, \mathrm{d}\tilde{\mathbf{z}}, \end{aligned} \tag{54}$$

where $\lambda > 0$ a parameter controlling the width of the distribution and δ is a (semi-)metric. If δ is a semi-metric, i.e., it does not fulfill the triangle inequality, there must exist a metric δ' such that δ can be written as the composition of a continuously invertible map $j: \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $j(0) = 0$ and the metric, i.e., $\delta = j \circ \delta'$. Finally, we assume that during training one has access to samples from both of these distributions.

Model Let $\mathcal{Z}'$ be a subset of $\mathbb{R}^N$ that is a convex body and let $f: \mathcal{X} \to \mathcal{Z}'$ be the model whose parameters are optimized. We associate a conditional distribution $q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z})$ with our model f through

$$\begin{aligned} q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z}) &= C_q^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \\ \text{with} \quad C_q(\mathbf{z}) &:= \int e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau}\, \mathrm{d}\tilde{\mathbf{z}}, \end{aligned} \tag{55}$$

where $C_q(\mathbf{z})$ is the partition function and δ is defined above.

A.2.2. Minimizing the cross-entropy

In a first step, we show the analogue of Proposition A for $\mathcal{Z}$ being a convex body:

Proposition 3. *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the $\mathcal{L}_{\delta\text{-contr}}$ loss converges to*

$$\lim_{M\to\infty} \mathcal{L}_{\delta\text{-contr}}(f;\tau,M) - \log M = \mathcal{L}_{\delta\text{-align}}(f;\tau) + \mathcal{L}_{\delta\text{-uni}}(f;\tau), \tag{56}$$

where

$$\mathcal{L}_{\delta\text{-align}}(f;\tau) := \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} [\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))]$$
$$\mathcal{L}_{\delta\text{-uni}}(f;\tau) := \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau}\right]\right)\right], \tag{57}$$

and $\mathcal{L}_{\delta\text{-contr}}(f;\tau,M)$ is as defined in Eq. (6).

Proof. This proof is adapted from Wang & Isola (2020). By the Continuous Mapping Theorem and the law of large numbers, for any $\mathbf{x}, \tilde{\mathbf{x}}$ and $\{\mathbf{x}_i^-\}_{i=1}^M$ it follows almost surely

$$\begin{aligned}
&\lim_{M\to\infty} \log\left(\frac{1}{M} e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))/\tau} + \frac{1}{M}\sum_{i=1}^{M} e^{-\delta(f(\mathbf{x}),f(\mathbf{x}_i^-))/\tau}\right) \\
&= \log\left(\mathop{\mathbb{E}}_{\mathbf{x}^-\sim p_{\text{data}}} \left[e^{-\delta(f(\mathbf{x}),f(\mathbf{x}^-))/\tau}\right]\right) \\
&= \log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{-\delta(h(\mathbf{z}),h(\tilde{\mathbf{z}}))/\tau}\right]\right),
\end{aligned} \tag{58}$$

where in the last step we expressed the sample $\mathbf{x}$ and negative examples $\mathbf{x}^-$ in terms of their latent factors.

We can now express the limit of the entire loss function as

$$\begin{aligned}
&\lim_{M\to\infty} \mathcal{L}_{\delta\text{-contr}}(f;\tau,M) - \log M \\
&= \frac{1}{\tau} \mathop{\mathbb{E}}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\text{pos}}} [\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))] \\
&\quad + \lim_{M\to\infty} \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\text{pos}}\\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[\log\left(\frac{1}{M} e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))/\tau} + \frac{1}{M}\sum_{i=1}^{M} e^{-\delta(f(\mathbf{x}),f(\mathbf{x}^-))/\tau}\right)\right] \\
&= \frac{1}{\tau} \mathop{\mathbb{E}}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\text{pos}}} [\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))] \\
&\quad + \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\text{pos}}\\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[\lim_{M\to\infty} \log\left(\frac{1}{M} e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))/\tau} + \frac{1}{M}\sum_{i=1}^{M} e^{-\delta(f(\mathbf{x}),f(\mathbf{x}_i^-))/\tau}\right)\right].
\end{aligned} \tag{59}$$

Note that as δ is a (semi-)metric, the expression $e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))}$ is upper-bounded by 1. Hence, according to the Dominated Convergence Theorem one can switch the limit with the expectation value in the second step. Inserting the previous results yields

$$\begin{aligned}
&= \frac{1}{\tau} \mathop{\mathbb{E}}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\text{pos}}} [\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))] \\
&\quad + \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\text{data}}} \left[\log\left(\mathop{\mathbb{E}}_{\mathbf{x}^-\sim p_{\text{data}}} \left[e^{-\delta(f(\mathbf{x}),f(\mathbf{x}^-))/\tau}\right]\right)\right] \\
&= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z})\\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} [\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))] \\
&\quad + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{-\delta(h(\mathbf{z}),h(\tilde{\mathbf{z}}))/\tau}\right]\right)\right] \\
&= \mathcal{L}_{\delta\text{-align}}(f;\tau) + \mathcal{L}_{\delta\text{-uni}}(f;\tau).
\end{aligned} \tag{60}$$

□

Next, we derive a property similar to Theorem 1, which suggests a practical method to find minimizers of the cross-entropy between the ground-truth p and model conditional q_{h}. This property is based on our previously introduced objective function in Eq. (6), which is a modified version of the InfoNCE objective in Eq. (1).

Theorem 3. *Let δ be a semi-metric and $\tau, \lambda > 0$ and let the ground-truth marginal distribution p be uniform. Consider a ground-truth conditional distribution $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) \exp(-\lambda\delta(\tilde{\mathbf{z}}, \mathbf{z}))$ and the model conditional distribution*

$$q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau}$$
$$\text{with} \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} \mathrm{d}\tilde{\mathbf{z}}. \tag{61}$$

Then the cross-entropy between p and q_{h} is given by

$$\lim_{M\to\infty} \mathcal{L}_{\delta\text{-contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| = \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_{\text{h}}(\cdot|\mathbf{z})\right], \tag{62}$$

which can be implemented by sampling data from the accessible distributions.

Proof. We use the definition of the cross-entropy to write

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} [H(p(\cdot|\mathbf{z}), q_{\text{h}}(\cdot|\mathbf{z})] \tag{63}$$
$$= -\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})} [\log(q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z}))]\right]. \tag{64}$$

We insert the definition of q_{h} and get

$$= - \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})} \left[\log(C_h^{-1}(\mathbf{z})) - \frac{1}{\tau}\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})) \right] \right] \tag{65}$$

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})} \left[\log(C_h(\mathbf{z})) + \frac{1}{\tau}\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})) \right] \right]. \tag{66}$$

As $C_h(\mathbf{z})$ does not depend on $\tilde{\mathbf{z}}$ it can be moved out of the inner expectation value, yielding

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})} \left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})) \right] + \log(C_h(\mathbf{z})) \right], \tag{67}$$

which can be written as

$$= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} \left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})) \right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log(C_h(\mathbf{z})) \right]. \tag{68}$$

Inserting the definition of C_h gives

$$= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} \left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})) \right] \tag{69}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \left(\int e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} d\tilde{\mathbf{z}} \right) \right]. \tag{70}$$

Next, the second term can be expanded by $1 = |\mathcal{Z}||\mathcal{Z}|^{-1}$, yielding

$$= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} \left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})) \right] \tag{71}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \left(\int \frac{|\mathcal{Z}|}{|\mathcal{Z}|} e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} d\tilde{\mathbf{z}} \right) \right]. \tag{72}$$

Finally, by using that the marginal is uniform, i.e., $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$, this can be simplified as

$$= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} \left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})) \right] \tag{73}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \right] \right) \right] \tag{74}$$

$$+ \log |\mathcal{Z}| \tag{75}$$

$$= \lim_{M\to\infty} \mathcal{L}_{\delta\text{-contr}}(f; \tau, M) - \log M + \log p|\mathcal{Z}|. \tag{76}$$

□

A.2.3. Cross-entropy minimizers are isometries

Now we show a version of Proposition 1, that is generalized from hyperspherical spaces to (subsets of) $\mathbb{R}^N$.

Proposition 4 (Minimizers of the cross-entropy are isometries)**.** *Let δ be a semi-metric. Consider the conditional distributions of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) \exp(-\delta(\tilde{\mathbf{z}}, \mathbf{z})/\lambda)$ and*

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau}$$
$$\text{with} \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \mathrm{d}\tilde{\mathbf{z}}, \tag{77}$$

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ can match for any point $\mathbf{z}$. If h is a minimizer of the cross-entropy $\mathcal{L}_{\mathsf{CE}} = \mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})]$, then h is an isometry, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \lambda\tau\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))$.

Proof. Note that $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h, e.g. the identity. The global minimum of cross-entropy between two distributions is reached if they match by value and have the same support. Hence, if p is a regular density, q_{h} will be a regular density, i.e., q_{h} is continuous and has only finite values $0 \leq q_{\mathrm{h}} < \infty$. As the two distributions match, this means

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}). \tag{78}$$

This expression also holds true for $\tilde{\mathbf{z}} = \mathbf{z}$; additionally using the property $\delta(\mathbf{z}, \mathbf{z}) = 0$ yields

$$p(\mathbf{z}|\mathbf{z}) = q_{\mathrm{h}}(\mathbf{z}|\mathbf{z}) \tag{79}$$

$$\Leftrightarrow \quad C_p^{-1}(\mathbf{z}) e^{-\delta(\mathbf{z},\mathbf{z})/\lambda} = C_h^{-1}(\mathbf{z}) e^{-\delta(h(\mathbf{z}), h(\mathbf{z}))/\tau} \tag{80}$$

$$\Leftrightarrow \quad C_p(\mathbf{z}) = C_h(\mathbf{z}). \tag{81}$$

As the normalization constants are identical, we obtain for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$

$$e^{-\delta(\tilde{\mathbf{z}},\mathbf{z})/\lambda} = e^{-\delta(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z}))/\tau} \tag{82}$$

$$\Leftrightarrow \quad \delta(\tilde{\mathbf{z}}, \mathbf{z}) = \frac{\lambda}{\tau}\delta(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z})). \tag{83}$$

By introducing a new semi-metric $\delta' := \lambda\tau^{-1}\delta$, we can write this as $\delta(\tilde{\mathbf{z}}, \mathbf{z}) = \delta'(h(\tilde{\mathbf{z}}), h(\mathbf{z}))$, which shows that h is an isometry. If there is no model mismatch, i.e., $\lambda = \tau$, this means $\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))$. □

Note, that this result does not depend on the choice of $\mathcal{Z}$ but just on the class of conditional distributions allowed.

A.2.4. Cross-entropy minimization identifies the ground-truth factors

Before we continue, let us recall a Theorem by Mankiewicz (1972):

Theorem C (Mankiewicz, 1972)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed linear spaces and let $\mathcal{V}$ be a convex body in $\mathcal{X}$ and $\mathcal{W}$ a convex body in $\mathcal{Y}$. Then every surjective isometry between $\mathcal{V}$ and $\mathcal{W}$ can be uniquely extended to an affine isometry between $\mathcal{X}$ and $\mathcal{Y}$.*

Proof. See Mankiewicz (1972). □

In addition, it is known that isometries on closed spaces are bijective:

Lemma A. *Assume h is an isometry of the closed space $\mathcal{Z}$ into itself, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} : \delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))$. Then h is bijective.*

Proof. See Lemma (2.6) in Całka (1982) for surjectivity. We show the injectivity by contradiction. Assume h is not injective. Then we can find a point $\tilde{\mathbf{z}} \neq \mathbf{z}$ where $h(\mathbf{z}) = h(\tilde{\mathbf{z}})$. But then $\delta(\mathbf{z}, \tilde{\mathbf{z}}) > \delta(\mathbf{z}, \mathbf{z})$ and $\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})) = \delta(h(\mathbf{z}), h(\mathbf{z})) = 0$ by the properties of δ. Hence, h is injective. □

Before continuing, we need to generalize the class of functions we consider as distance measures:

Lemma 1. *Let δ' be a the composition of a continuously invertible function $j : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $j(0) = 0$ and a metric δ, i.e., $\delta' := j \circ \delta$. Then, (i) δ' is a semi-metric and (ii) if a function $h : \mathbb{R}^n \to \mathbb{R}^n$ is an isometry of a space with the semi-metric δ', it is also an isometry of the space with the metric δ.*

Proof. (i) Let $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$. Per assumption j must be strictly monotonically increasing on $\mathbb{R}_{\geq 0}$. Since δ is a metric it follows $\delta(\mathbf{z}, \tilde{\mathbf{z}}) \geq 0 \Rightarrow \delta'(\mathbf{z}, \tilde{\mathbf{z}}) = j(\delta(\mathbf{z}, \tilde{\mathbf{z}})) \geq 0$, with equality iff $\mathbf{z} = \tilde{\mathbf{z}}$. Furthermore, since δ is a metric it is symmetric in its arguments and, hence, δ' is symmetric in its arguments. Thus, δ' is a semi-metric.

(ii) h is an isometry of a space with the semi-metric δ', allowing to derive that for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$,

$$\delta'(h(\mathbf{z}), h(\tilde{\mathbf{z}})) = \delta'(\mathbf{z}, \tilde{\mathbf{z}}) \tag{84}$$

$$j(\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))) = j(\delta(\mathbf{z}, \tilde{\mathbf{z}})) \tag{85}$$

and, applying the inverse j^{-1} which exists by assumption, yields

$$\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})) = \delta(\mathbf{z}, \tilde{\mathbf{z}}), \tag{86}$$

concluding the proof. □

By combining the properties derived before we can show that h is an affine function:

Theorem 4. *Let $\mathcal{Z} = \mathcal{Z}'$ be a convex body in $\mathbb{R}^N$. Let the mixing function g be differentiable and invertible. If the assumed form of q_{h} as defined in Eq. (55) matches that of p, and if f is differentiable and minimizes the cross-entropy between p and q_{h}, then we find that $h = f \circ g$ is affine, i.e., we recover the latent sources up to affine transformations.*

Proof. According to Proposition 4 h is an isometry and q_{h} is a regular probability density function. If the distance δ used in the conditional distributions p and q_{h} is a semi-metric as in Lemma 1, it follows that h is also an isometry for a proper metric. This also means that h is bijective according to Lemma A. Finally, Theorem C says that h is an affine transformation. □

We use the assumption that the marginal $p(\mathbf{z})$ is uniform, to show

Theorem 5. *Let $\mathcal{Z}$ be a convex body in $\mathbb{R}^N$, $h = f \circ g : \mathcal{Z} \to \mathcal{Z}$, and δ be a metric or a semi-metric as defined in Lemma 1. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as (5). Let the mixing function g be differentiable and injective. If the assumed form of q_{h} matches that of p, i.e.,*

$$\begin{aligned} q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) &= C_q^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \\ \text{with} \quad C_q(\mathbf{z}) &:= \int e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \, \mathrm{d}\tilde{\mathbf{z}}, \end{aligned} \tag{87}$$

and if f is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in (6) for $M \to \infty$, we find that $h = f \circ g$ is invertible and affine, i.e., we recover the latent sources up to affine transformations.

Proof. According to Theorem 3 h minimizes the cross-entropy between p and q_{h} as defined in Eq. (4). Then according to Theorem 4, h is an affine transformation. □

This result can be seen as a generalized version of Theorem 2, as it is valid for any convex body $\mathcal{Z} \subseteq \mathbb{R}^N$ and allows a larger variety of conditional distributions. A missing step is to extend this theory beyond uniform marginal distributions. This will be addressed in future work.

Under some assumptions we can further narrow down possible forms of h, thus, showing that h in fact solves the nonlinear ICA problem only up to permutations and elementwise transformations.

For this, let us first repeat a result from Li & So (1994), that shows an important property of isometric matrices:

Theorem D. *Suppose $1 \leq \alpha \leq \infty$ and $\alpha \neq 2$. An $n \times n$ matrix $\mathbf{A}$ is an isometry of L^α-norm if and only if $\mathbf{A}$ is a generalized permutation matrix, i.e., $\forall \mathbf{z} : (\mathbf{A}\mathbf{z})_\mathbf{i} = \alpha_\mathbf{i} \mathbf{z}_{\sigma(\mathbf{i})}$, with $\alpha_i = \pm 2$ and σ being a permutation.*

Proof. See Li & So (1994). Note that this can also be concluded from the Banach-Lamperti Theorem (Lamperti et al., 1958). □

Leveraging this insight, we can finally show:

Theorem 6. *Let $\mathcal{Z}$ be a convex body in $\mathbb{R}^N$, $h : \mathcal{Z} \to \mathcal{Z}$, and δ be an L^α metric for $\alpha \geq 1, \alpha \neq 2$ or the α-th power of such an L^α metric. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as in Eq. (5), and let the mixing function g be differentiable and invertible. If the assumed form of $q_h(\cdot|\mathbf{z})$ matches that of $p(\cdot|\mathbf{z})$, i.e., both use the same metric δ up to a constant scaling factor, and if f is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in Eq. (6) for $M \to \infty$ we find that $h = f \circ g$ is a composition of input independent permutations, sign flips and rescalings.*

Proof. First, we prove the case where both conditional distributions use exactly the same metric. By Theorem 5 h is an affine transformation. Moreover, according to Proposition 4 is an isometry. Thus, by Theorem D, h is a generalized permutation matrix, i.e., a composition of permutations and sign flips.

Finally, for the case that δ matches the similarity measure in the ground-truth conditional distribution defined in Eq. (5) (denoted as δ^*) only up to a constant rescaling factor r, we know

$$\begin{aligned} &\forall \mathbf{z}, \tilde{\mathbf{z}} : \delta^*(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})) \\ \Leftrightarrow\ &\delta^*(\mathbf{z}, \tilde{\mathbf{z}}) = \delta^*\left(\frac{1}{r}h(\mathbf{z}), \frac{1}{r}h(\tilde{\mathbf{z}})\right). \end{aligned} \tag{88}$$

Thus, $\frac{1}{r}h$ is a δ^* isometry and the same argument as above holds, concluding the proof. □

A.3. Experimental details

For the experiments presented in Sec. 4.1 we train our feature encoder for 300 000 iterations with a batch size of 6144 utilizing Adam (Kingma & Ba, 2015) with a learning rate of 10^{-4}. Like Hyvärinen & Morioka (2016; 2017), for the mixing network, we i) use 0.2 for the angle of the negative slope[5], ii) use L^2 normalized weight matrices with minimum condition number of 25 000 uniformly distributed samples. For the encoder, we i) use the default (0.01) negative slope ii) use 6 hidden layers with dimensionality [$N \cdot 10$, $N \cdot 50$, $N \cdot 50$, $N \cdot 50$, $N \cdot 50$, $N \cdot 10$] and iii) initialize the normalization magnitude as 1. We sample 4096 latents from the marginal for evaluation. For MCC (Hyvärinen & Morioka, 2016; 2017) we use the Pearson correlation coefficient[6]; we found there to be no difference with Spearman[7].

For the experiments presented in Sec. 4.2.1, we use the same architecture as the encoder in (Klindt et al., 2021). As in (Klindt et al., 2021), we train for 300 000 iterations with a batch size of 64 utilizing Adam (Kingma & Ba, 2015) with a learning rate of 10^{-4}. For evaluation, as in (Klindt et al., 2021), we use 10 000 samples and the Spearman correlation coefficient.

For the experiments presented in Sec. 4.2.2, we train the feature encoder for 200 000 iterations using Adam with a learning rate of 10^{-4}. For the encoder we use a ResNet18 (He et al., 2016) architecture followed by a single hidden layer with dimensionality $N \cdot 10$ and LeakyReLU activation function using the default (0.01) negative slope. The scores on the training set are evaluated on 10% of the whole training set, 25 000 random samples. The test set consists of 25 000 samples not included in the training set. For the last row of Tab. 4 and Tab. 5 we used the best-working combination of image augmentations found by Chen et al. (2020a) to sample positive pairs. To be precise, we used a random crop and resize operation followed by a color distortion augmentation. The random crops had a uniformly distributed size (between 8% and 100% of the original image area) and a random aspect ration (between 3/4 and 4/3); subsequently, they were resized to the original image dimension (224×224) again. The color distortion operation itself combined color jittering (i.e., random changes of the brightness, contrast, saturation and hue) with color dropping (i.e., random grayscale conversations). We used the same parameters for these augmentations as recommended by Chen et al. (2020a).

The experiments in Sec. 4.1 took on the order of 5-10 hours on a GeForce RTX 2080 Ti GPU, the experiments on KITTI Masks took 1.5 hours on a GeForce RTX 2080 Ti GPU and those on 3DIdent took 28 hours on four GeForce RTX 2080 Ti GPUs. The creation of the 3DIdent dataset additionally required approximately 150 hours of compute time on a GeForce RTX 2080 Ti.

A.4. Details on 3DIdent

We build on the rendering pipeline of Johnson et al. (2017b) and use the Blender engine (Blender Online Community, 2021), as of version 2.91.0, for image rendering. The scenes depicted in the dataset show a rotated and translated object onto which a spotlight is directed. The spotlight is located on a half-circle above the scene and shines down. The scenes can be described by 10 parameters: the position of the object along the X-, Y- and Z-axis, the rotation of the object described by Euler angles (3), the position of the spotlight described by a polar angle, and the hue of the object, the ground and the spotlight. The value range is $[-3, 3]$ for all position parameters, and is $[-\pi/2, \pi/2]$ for the remaining parameters. The parameters are sampled from a 10-dimensional unit hyperrectangle, then rescaled to their corresponding value range. This ensures that the variance

[5] See e.g. https://pytorch.org/docs/stable/generated/torch.nn.LeakyReLU.html

[6] See e.g. https://numpy.org/doc/stable/reference/generated/numpy.corrcoef.html

[7] See e.g. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

Table 5. Identifiability up to affine transformations on the training set of 3DIdent. Mean ± standard deviation over 3 random seeds. As earlier, only the first row corresponds to a setting that matches the theoretical assumptions for linear identifiability; the others show distinct violations. Supervised training with unbounded space achieves scores of $R^2 = (99.98 \pm 0.01)\%$ and MCC $= (99.99 \pm 0.01)\%$. The last row refers to using the SimCLR (Chen et al., 2020a) augmentations to generate positive pairs. The last row refers to using the image augmentations suggested by Chen et al. (2020a) to generate positive image pairs; for details see Sec. A.3. In contrast to Table 4, the scores here are reported on the same data the models were trained on.

Dataset	Model f			Identity [%]	Unsupervised [%]	
$p(\cdot\|\cdot)$	Space	$q_{\mathrm{h}}(\cdot\|\cdot)$	M.	R^2	R^2	MCC
Normal	Box	Normal	✓	5.35 ± 0.72	97.83 ± 0.13	98.85 ± 0.07
Normal	Unbounded	Normal	✗	—"—	97.72 ± 0.02	55.90 ± 2.22
Laplace	Box	Normal	✗	—"—	97.95 ± 0.05	98.94 ± 0.03
Normal	Sphere	vMF	✗	—"—	66.73 ± 0.03	42.72 ± 3.20
Augm.	Sphere	vMF	✗	—"—	45.94 ± 1.80	47.6 ± 1.45

of the latent factors is the same for all latent dimensions.

To ensure that the generative process is injective, we take two measures: First, we use a non-rotationally symmetric object (Utah tea pot, Newell, 1975), thus the rotation information is unambiguous. Second, we use different levels of color saturation for the object, the spotlight and the ground (1.0, 0.8 and 0.6, respectively), thus the object is always distinguishable from the ground.

A.4.1. Comparison to existing datasets

The proposed dataset contains high-resolution renderings of an object in a 3D scene. It features some aspects of natural scenes, e.g. complex 3D objects, different lighting conditions and continuous variables. Existing benchmarks (Klindt et al., 2021; Burgess & Kim, 2018; Gondal et al., 2019; Dittadi et al., 2021) for disentanglement in 3D scenes differ in important aspects to 3DIdent.

KITTI Masks (Klindt et al., 2021) only enables evaluating identification of the two-dimensional position and scale of the object instance. In addition, the observed segmentation masks are significantly lower resolution than examples in our dataset. 3D Shapes (Burgess & Kim, 2018) and MPI3D (Gondal et al., 2019) are rendered at the same resolution (64×64) as KITTI Masks. Whereas the dataset contributed by (Dittadi et al., 2021) is rendered at $2\times$ that resolution (128×128), our dataset is rendered at $3.5\times$ that resolution (224×224), the resolution at which natural image classification is typically evaluated (Deng et al., 2009). With that being said, we do note that KITTI Masks is unique in containing frames of natural video, and we thus consider it complementary to 3DIdent.

Burgess & Kim (2018), Dittadi et al. (2021), and Gondal et al. (2019) contribute datasets which contain variable object rotations around one, one, and two rotation axes, respectively, while 3DIdent contains variable object rotation around all three rotation axes as well as variable lighting conditions. Furthermore, each of these datasets were generated by sampling latent factors from an equidistant grid, thus only covering a limited number values along each axis of variation, effectively resulting in a highly coarse discretization of naturally continuous variables. As 3DIdent instead samples the latent factors uniformly in the latent space, this better reflects the continuous nature of the latent dimensions.

A.5. Effects of the Uniformity Loss

In previous work, Wang & Isola (2020) showed that a part of the contrastive (InfoNCE) loss — the uniformity loss — effectively ensures that the encoded features are uniformly distributed over a hypersphere. We now show that this part is crucial to ensure that the mapping is bijective. More precisely, we demonstrate that if the distribution of the encoded/reconstructed latents $h(\mathbf{z})$ has the same support as the distribution of $\mathbf{z}$, and both distributions are regular, i.e., their densities are non-zero and finite, then the transformation h is bijective.

First, we focus on the more general case of a map between manifolds:

Proposition 5. *Let $\mathcal{M}, \mathcal{N}$ be simply connected and oriented $\mathcal{C}^1$ manifolds without boundaries and $h : \mathcal{M} \to \mathcal{N}$ be a differentiable map. Further, let the random variable $\mathbf{z} \in \mathcal{M}$ be distributed according to $\mathbf{z} \sim p(\mathbf{z})$ for a regular density function p, i.e., $0 < p < \infty$. If the pushforward $p_{\#h}(\mathbf{z})$ of p through h is also a regular density, i.e., $0 < p_{\#h} < \infty$, then h is a bijection.*

Proof. We begin by showing by contradiction that the Jacobian determinant of h does not vanish, i.e., $|\det J_h| > 0$:

Suppose that the Jacobian determinant $|\det J_h|$ vanishes for some $\mathbf{z} \in \mathcal{M}$. Then the inverse of the Jacobian determinant goes to infinity at this point and so does the density of $h(\mathbf{z})$ according to the well-known transformation of probability densities. By assumption, both p and $p_{\#h}$ must be regular density functions and, thus, be finite. This contradicts the initial assumption and so the Jacobian determinant $|\det J_h|$

cannot vanish.

Next, we show that the mapping h is proper. Note that a map is called proper if pre-images of compact sets are compact (Ruzhansky & Sugimoto, 2015). Firstly, a continuous mapping between $\mathcal{M}$ and $\mathcal{N}$ is also closed, i.e., pre-images of closed subsets are also closed (Lee, 2013). In addition, it is well-known that continuous functions on compact sets are bounded. Lastly, according to the Heine–Borel theorem, compact subsets of $\mathbb{R}^D$ are closed and bounded. Taken together, this shows that h is proper.

Finally, according to Theorem 2.1 in (Ruzhansky & Sugimoto, 2015) a proper h with non-vanishing Jacobian determinant is bijective, concluding the proof. □

This theorem directly applies to the case of hyperspheres, which are simply connected and oriented manifolds without boundary. This yields:

Lemma 2. *Let $\mathcal{Z}$ be a hypersphere and $h : \mathcal{Z} \to \mathcal{Z}$ be a differentiable map. Further, let the marginal distribution $p(\mathbf{z})$ of the variable $\mathbf{z} \in \mathcal{Z}$ be a regular density function, i.e., $0 < p < \infty$. If the pushforward $p_{\#h}$ of p through h is also a regular density, i.e., $0 < p_{\#h} < \infty$, then h is a bijection.*

Therefore, we can conclude that a loss term ensuring that the encoded features are distributed according to a regular density function, such as the uniformity term, makes the map h bijective and prevents an information loss. Note that this does not assume that the marginal distribution of the ground-truth latents $p(\mathbf{z})$ is uniform but only that it is regular and non-vanishing.

Note that while the proposition shows that the uniformity loss is sufficient to ensure bijectivity, we can construct counterexamples if its assumptions (like differentiability) are violated even in just a single point. For instance, the requirement of h being fully differentiable is most likely violated in large unregularized neural networks with ReLU nonlinearities. Here, one might need the full contrastive loss to ensure bijectivity of h.