Towards a Unified Framework of Contrastive Learning for Disentangled Representations (Supplementary Material)

Stefan Matthes, Zhiwei Han, Hao Shen
fortiss GmbH, Munich, Germany
{matthes,han,shen}@fortiss.org

A Proofs of the Theorems

In this section, we provide proofs of our theoretical claims. In many applications, the marginal distributions of the positive pairs $p_{\mathbf{x}}$ and $p_{\tilde{\mathbf{x}}}$ coincide, e.g., when $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are sampled as successive frames from a temporally stationary process. We consider here the case where $p_{\mathbf{x}}$ may also be different from $p_{\tilde{\mathbf{x}}}$ and the distribution of negative samples $p_{\mathbf{x}^-}$ is chosen to be nonzero on the support $\mathcal{X}$. In practice, given two batches of corresponding observations, it is often convenient to select the negative samples from the first, second, or both batches, resulting in $p_{\mathbf{x}^-} = p_{\mathbf{x}}$, $p_{\mathbf{x}^-} = p_{\tilde{\mathbf{x}}}$, or $p_{\mathbf{x}^-} = \frac{1}{2}(p_{\mathbf{x}} + p_{\tilde{\mathbf{x}}})$.

Lemma 1. *Let the data generating process follow Eq. 1 and g be differentiable and invertible. Further, assume that f and δ have universal approximation capability. Then for the optimal estimators of $\mathcal{L}_{\delta\text{-NCE}}(f,\delta)$, $\mathcal{L}_{\delta\text{-INCE}}(f,\delta;K)$, $\mathcal{L}_{\delta\text{-SCL}}(f,\delta)$ and $\mathcal{L}_{\delta\text{-NWJ}}(f,\delta)$, it holds that*

$$\delta(h(\mathbf{s}), h(\tilde{\mathbf{s}})) = -\log p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s}) + p_{\mathbf{s}^-}(\tilde{\mathbf{s}}) + \gamma(\mathbf{s}), \tag{11}$$

where $h = f \circ g$ and γ is some function that only depends on $\mathbf{s}$.

Proof. We will show below for each loss function separately that the optimal estimators satisfy the form

$$\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}})) = -\log p_{\tilde{\mathbf{x}}|\mathbf{x}}(\tilde{\mathbf{x}}|\mathbf{x}) + \log p_{\mathbf{x}^-}(\tilde{\mathbf{x}}) + \gamma_{\mathbf{x}}(\mathbf{x}), \tag{12}$$

where $\gamma_{\mathbf{x}}$ is a function that depends only on $\mathbf{x}$ and is zero for all but $\mathcal{L}_{\delta\text{-INCE}}(f,\delta;K)$.

We can then use the fact that the data generating function g is injective and differentiable. This allows us to apply the probability transformations $p_{\tilde{\mathbf{x}}|\mathbf{x}}(\tilde{\mathbf{x}}|\mathbf{x}) = p_{\tilde{\mathbf{s}}|\mathbf{s}}(g^{-1}(\tilde{\mathbf{x}})|g^{-1}(\mathbf{x}))\operatorname{vol} J_{g^{-1}}(\tilde{\mathbf{x}})$ and $p_{\mathbf{x}^-}(\tilde{\mathbf{x}}) = p_{\mathbf{s}^-}(g^{-1}(\tilde{\mathbf{x}}))\operatorname{vol} J_{g^{-1}}(\tilde{\mathbf{x}})$, where $\operatorname{vol} J_{g^{-1}}(\tilde{\mathbf{x}})$ is the product of the singular values of the Jacobian [1]. Thus, we obtain

$$\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}})) = -\log p_{\tilde{\mathbf{s}}|\mathbf{s}}(g^{-1}(\tilde{\mathbf{x}})|g^{-1}(\mathbf{x})) + \log p_{\mathbf{s}^-}(g^{-1}(\tilde{\mathbf{x}})) + \gamma_{\mathbf{x}}(\mathbf{x}),$$

where the Jacobians nicely cancel. Finally, we substitute $\mathbf{x} = g(\mathbf{s})$ and $\tilde{\mathbf{x}} = g(\tilde{\mathbf{s}})$ and get

$$\delta(h(\mathbf{s}), h(\tilde{\mathbf{s}})) = -\log p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s}) + \log p_{\mathbf{s}^-}(\tilde{\mathbf{s}}) + \gamma(\mathbf{s}),$$

where $\gamma(\mathbf{s}) = \gamma_{\mathbf{x}}(g(\mathbf{s}))$.

Part 1 (δ-NCE loss):

It is known that the NCE loss converges towards the log difference of the two data distributions of the positive and negative class [3, 6], that is

$$\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}})) = -\log \frac{p_{\text{pos}}(\mathbf{x}, \tilde{\mathbf{x}})}{p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{x}^-}(\tilde{\mathbf{x}})} = -\log p_{\tilde{\mathbf{x}}|\mathbf{x}}(\tilde{\mathbf{x}}|\mathbf{x}) + \log p_{\mathbf{x}^-}(\tilde{\mathbf{x}}).$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Part 2 (δ-INCE loss):

The stated result can be obtained using Theorem 3.2 from Ma and Collins [9]. In addition, we give an alternative proof in Section B. Theorem 3.2 in [9] assumes (Assumption 2.1 in their work) that there exist functions f and δ such that for all $(\mathbf{x}, \tilde{\mathbf{x}}) \in \mathcal{X} \times \mathcal{X}$

$$\frac{p_{\tilde{\mathbf{x}}|\mathbf{x}}(\tilde{\mathbf{x}}|\mathbf{x})}{p_{\mathbf{x}^-}(\tilde{\mathbf{x}})} = \frac{1}{Z(\mathbf{x})}\, e^{-\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))}, \tag{13}$$

where $Z(\mathbf{x}) = \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}} e^{-\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))}$. This assumption is satisfied since we optimize both f and δ using universal function approximators. Then the theorem states that in the limit of infinite data, this relation holds for the global optimum when we optimize the InfoNCE loss. To obtain their notation, simply replace $\mathbf{y} = \tilde{\mathbf{x}}$ and $s(\mathbf{x}, \mathbf{y}) = \log p_{\mathbf{x}^-}(\tilde{\mathbf{x}}) - \delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))$. Therefore, we get $\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}})) = -\log p_{\tilde{\mathbf{x}}|\mathbf{x}}(\tilde{\mathbf{x}}|\mathbf{x}) + \log p_{\mathbf{x}^-}(\tilde{\mathbf{x}}) + \gamma_{\mathbf{x}}(\mathbf{x})$ with $\gamma_{\mathbf{x}}(\mathbf{x}) = -\log Z(\mathbf{x})$.

Part 3 (δ-SCL loss):

To simplify the notation, we substitute $\psi(\mathbf{x}, \tilde{\mathbf{x}}) = -\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))$ in the δ-SCL loss and obtain

$$\tilde{\mathcal{L}}_{\delta\text{-SCL}}(\psi) = \mathop{\mathbb{E}}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \\ \sim p_{\text{pos}}}} -2\, e^{\psi(\mathbf{x}, \tilde{\mathbf{x}})} + \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim p_{\mathbf{x}} \\ \mathbf{x}^- \sim p_{\mathbf{x}^-}}} e^{2\psi(\mathbf{x}, \mathbf{x}^-)}\,. \tag{14}$$

First we derive the Taylor series of $\tilde{\mathcal{L}}_{\delta\text{-SCL}}$:

$$\begin{aligned}\tilde{\mathcal{L}}_{\delta\text{-SCL}}(\psi + \epsilon\eta) = \tilde{\mathcal{L}}_{\delta\text{-SCL}}(\psi) + 2\epsilon \left[\mathop{\mathbb{E}}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \\ \sim p_{\text{pos}}}} - e^{\psi(\mathbf{x}, \tilde{\mathbf{x}})}\, \eta(\mathbf{x}, \tilde{\mathbf{x}}) + \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim p_{\mathbf{x}} \\ \mathbf{x}^- \sim p_{\mathbf{x}^-}}} e^{2\psi(\mathbf{x}, \mathbf{x}^-)}\, \eta(\mathbf{x}, \mathbf{x}^-) \right] \\ + \epsilon^2 \left[\mathop{\mathbb{E}}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \\ \sim p_{\text{pos}}}} - e^{\psi(\mathbf{x}, \tilde{\mathbf{x}})}\, \eta(\mathbf{x}, \tilde{\mathbf{x}})^2 + 2 \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim p_{\mathbf{x}} \\ \mathbf{x}^- \sim p_{\mathbf{x}^-}}} e^{2\psi(\mathbf{x}, \mathbf{x}^-)}\, \eta(\mathbf{x}, \mathbf{x}^-)^2 \right] + \mathcal{O}(\epsilon^3). \end{aligned} \tag{15}$$

A necessary optimality condition is that in the expansion of $\tilde{\mathcal{L}}_{\delta\text{-SCL}}$, the term of order ϵ is zero for any perturbation η. We have

$$\begin{aligned} 0 &= \mathop{\mathbb{E}}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \\ \sim p_{\text{pos}}}} - e^{\psi(\mathbf{x}, \tilde{\mathbf{x}})}\, \eta(\mathbf{x}, \tilde{\mathbf{x}}) + \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim p_{\mathbf{x}} \\ \mathbf{x}^- \sim p_{\mathbf{x}^-}}} e^{2\psi(\mathbf{x}, \mathbf{x}^-)}\, \eta(\mathbf{x}, \mathbf{x}^-) \\ &= \int p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{x}^-}(\mathbf{x}^-)\, e^{2\psi(\mathbf{x}, \mathbf{x}^-)}\, \eta(\mathbf{x}, \mathbf{x}^-)\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{x}^- - \int p_{\text{pos}}(\mathbf{x}, \tilde{\mathbf{x}})\, e^{\psi(\mathbf{x}, \tilde{\mathbf{x}})}\, \eta(\mathbf{x}, \tilde{\mathbf{x}})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\tilde{\mathbf{x}} \\ &= \int p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{x}^-}(\tilde{\mathbf{x}})\, e^{2\psi(\mathbf{x}, \tilde{\mathbf{x}})}\, \eta(\mathbf{x}, \tilde{\mathbf{x}})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\tilde{\mathbf{x}} - \int p_{\text{pos}}(\mathbf{x}, \tilde{\mathbf{x}})\, e^{\psi(\mathbf{x}, \tilde{\mathbf{x}})}\, \eta(\mathbf{x}, \tilde{\mathbf{x}})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\tilde{\mathbf{x}} \\ &= \int \left(p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{x}^-}(\tilde{\mathbf{x}})\, e^{2\psi(\mathbf{x}, \tilde{\mathbf{x}})} - p_{\text{pos}}(\mathbf{x}, \tilde{\mathbf{x}})\, e^{\psi(\mathbf{x}, \tilde{\mathbf{x}})} \right) \eta(\mathbf{x}, \tilde{\mathbf{x}})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\tilde{\mathbf{x}}. \end{aligned}$$

This term vanishes if and only if

$$e^{\psi(\mathbf{x}, \tilde{\mathbf{x}})} = \frac{p_{\text{pos}}(\mathbf{x}, \tilde{\mathbf{x}})}{p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{x}^-}(\tilde{\mathbf{x}})},$$

that is $\psi^*(\mathbf{x}, \tilde{\mathbf{x}}) = \log p_{\tilde{\mathbf{x}}|\mathbf{x}}(\tilde{\mathbf{x}}|\mathbf{x}) - \log p_{\mathbf{x}^-}(\tilde{\mathbf{x}})$.

To verify that the critical point is indeed a minimizer, we insert the solution back into Eq. 15 and check if the order ϵ^2 term is strictly positive for any direction η. This leads to

$$\begin{aligned}\tilde{\mathcal{L}}_{\delta\text{-SCL}}(\psi^* + \epsilon\eta) = \tilde{\mathcal{L}}_{\delta\text{-SCL}}(\psi^*) \\ + \epsilon^2 \left[\mathop{\mathbb{E}}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \\ \sim p_{\text{pos}}}} - \frac{p_{\text{pos}}(\mathbf{x}, \tilde{\mathbf{x}})}{p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{x}^-}(\tilde{\mathbf{x}})} \eta(\mathbf{x}, \tilde{\mathbf{x}})^2 + 2 \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim p_{\mathbf{x}} \\ \mathbf{x}^- \sim p_{\mathbf{x}^-}}} \left(\frac{p_{\text{pos}}(\mathbf{x}, \mathbf{x}^-)}{p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{x}^-}(\mathbf{x}^-)} \right)^2 \eta(\mathbf{x}, \mathbf{x}^-)^2 \right] + \mathcal{O}(\epsilon^3). \end{aligned} \tag{16}$$

The order ϵ^2 term can be simplified to

$$\int \frac{p_{\text{pos}}(\mathbf{x}, \tilde{\mathbf{x}})^2}{p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{x}^-}(\tilde{\mathbf{x}})} \eta(\mathbf{x}, \tilde{\mathbf{x}})^2 \, \mathrm{d}\mathbf{x} \, \mathrm{d}\tilde{\mathbf{x}},$$

which is clearly positive for any direction η. Thus, $\tilde{\mathcal{L}}_{\delta\text{-SCL}}$ reaches indeed a minimum at ψ^*, or in terms of δ and f, we have $\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}})) = -\log p_{\tilde{\mathbf{x}}|\mathbf{x}}(\tilde{\mathbf{x}}|\mathbf{x}) + \log p_{\mathbf{x}^-}(\tilde{\mathbf{x}})$.

Part 4 (δ-NWJ loss):

The proof for $\mathcal{L}_{\delta\text{-NWJ}}$ is analogous to $\mathcal{L}_{\delta\text{-SCL}}$. We use the same substitution as in the last part, i.e., $\psi(\mathbf{x}, \tilde{\mathbf{x}}) = -\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))$, and get

$$\tilde{\mathcal{L}}_{\delta\text{-NWJ}}(\psi) = \mathop{\mathbb{E}}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \\ \sim p_{\text{pos}}}} -\psi(\mathbf{x}, \tilde{\mathbf{x}}) + \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim p_{\mathbf{x}} \\ \mathbf{x}^- \sim p_{\mathbf{x}^-}}} e^{\psi(\mathbf{x}, \mathbf{x}^-)} \,. \tag{17}$$

The Taylor expansion of $\tilde{\mathcal{L}}_{\delta\text{-NWJ}}$ is given by

$$\begin{aligned} \tilde{\mathcal{L}}_{\delta\text{-NWJ}}(\psi + \epsilon\eta) = \tilde{\mathcal{L}}_{\delta\text{-NWJ}}(\psi) + \epsilon \left[\mathop{\mathbb{E}}_{\substack{(\mathbf{x}, \tilde{\mathbf{x}}) \\ \sim p_{\text{pos}}}} -\eta(\mathbf{x}, \tilde{\mathbf{x}}) + \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim p_{\mathbf{x}} \\ \mathbf{x}^- \sim p_{\mathbf{x}^-}}} e^{\psi(\mathbf{x}, \mathbf{x}^-)} \, \eta(\mathbf{x}, \mathbf{x}^-) \right] \\ + \frac{1}{2}\epsilon^2 \mathop{\mathbb{E}}_{\substack{\mathbf{x} \sim p_{\mathbf{x}} \\ \mathbf{x}^- \sim p_{\mathbf{x}^-}}} e^{\psi(\mathbf{x}, \mathbf{x}^-)} \, \eta(\mathbf{x}, \mathbf{x}^-)^2 + \mathcal{O}(\epsilon^3). \end{aligned} \tag{18}$$

Again, the term of order ϵ is zero at $\psi^*(\mathbf{x}, \tilde{\mathbf{x}}) = \log p_{\tilde{\mathbf{x}}|\mathbf{x}}(\tilde{\mathbf{x}}|\mathbf{x}) - \log p_{\mathbf{x}^-}(\tilde{\mathbf{x}})$ and we directly see that the ϵ^2 term is strictly positive for any direction η.

□

Lemma 2. *Let the data generating process follow Eq. 1, where d is a semi-metric (i.e., the triangle inequality does not necessarily hold), and $\mathcal{S}$ is open. Let us further assume that δ has the form as in Eq. 2 and satisfies $\delta(h(\mathbf{s}), h(\tilde{\mathbf{s}})) = -\log p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s}) + \log p_{\mathbf{s}^-}(\tilde{\mathbf{s}}) + \gamma(\mathbf{s})$. Then, for the optimal estimators of the contrastive losses presented above, h is a homeomorphism between $\mathcal{S}$ and $h(\mathcal{S})$.*

Proof. After inserting $p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s})$ from Eq. 1 and δ from Eq. 2, we obtain

$$d(h(\mathbf{s}), h(\tilde{\mathbf{s}})) + \alpha(h(\mathbf{s})) + \tilde{\alpha}(h(\tilde{\mathbf{s}})) = d(\mathbf{s}, \tilde{\mathbf{s}}) - \log Q(\tilde{\mathbf{s}}) + \log Z(\mathbf{s}) + \log p_{\mathbf{s}^-}(\tilde{\mathbf{s}}) + \gamma(\mathbf{s}). \tag{19}$$

This must also hold for all $\mathbf{s}$, $\tilde{\mathbf{s}}$, including $\mathbf{s} = \tilde{\mathbf{s}}$, and since $d(\mathbf{s}, \mathbf{s}) = d(h(\mathbf{s}), h(\mathbf{s}))$, we have

$$\alpha(h(\mathbf{s})) + \tilde{\alpha}(h(\mathbf{s})) = -\log Q(\mathbf{s}) + \log Z(\mathbf{s}) + \log p_{\mathbf{s}^-}(\mathbf{s}) + \gamma(\mathbf{s}). \tag{20}$$

Now we subtract the last two equations and get

$$d(h(\mathbf{s}), h(\tilde{\mathbf{s}})) + \tilde{\alpha}(h(\tilde{\mathbf{s}})) - \tilde{\alpha}(h(\mathbf{s})) = d(\mathbf{s}, \tilde{\mathbf{s}}) - \log Q(\tilde{\mathbf{s}}) + \log Q(\mathbf{s}) + \log p_{\mathbf{s}^-}(\tilde{\mathbf{s}}) - \log p_{\mathbf{s}^-}(\mathbf{s}). \tag{21}$$

Because of the symmetry of d we can conclude

$$d(h(\mathbf{s}), h(\tilde{\mathbf{s}})) = d(\mathbf{s}, \tilde{\mathbf{s}}). \tag{22}$$

If h were not injective, there would be some $\mathbf{s} \neq \tilde{\mathbf{s}}$ with $h(\mathbf{s}) = h(\tilde{\mathbf{s}})$. This would imply that $0 = d(h(\mathbf{s}), h(\tilde{\mathbf{s}})) = d(\mathbf{s}, \tilde{\mathbf{s}}) > 0$, which is a contradiction.

Since h is an injective continuous mapping and $\mathcal{S}$ is open, h is a homeomorphism due to the domain invariance theorem (see [2]). □

From Eq. (19), we also see that $\tilde{\alpha} \circ h = \log p_{\mathbf{s}^-} - \log Q$ and, except for $\mathcal{L}_{\delta\text{-INCE}}(f, \delta; K)$, $\alpha \circ h = \log Z$.

Before we continue, let us recall an important theorem by Mankiewicz [10] and the definition of 2-extremal points [11].

Theorem A. *(Mankiewicz, 1972) Suppose $\mathcal{X}$ and $\mathcal{Y}$ are real normed vector spaces, $\mathcal{U}$ a nonempty subset of $\mathcal{X}$, and let $f : \mathcal{U} \to f(\mathcal{U})$ be a surjective isometry, where $f(\mathcal{U})$ is a subset of $\mathcal{Y}$. If either both $\mathcal{U}$ and $f(\mathcal{U})$ are convex bodies or open and connected, then f can be uniquely extended to an affine isometry from $\mathcal{X}$ to $\mathcal{Y}$.*

Proof. See [10]. □

Definition 1. *As a 2-extremal point of a set $\mathcal{U}$ of a vector space we denote each element $\mathbf{x} \in \mathcal{U}$ such that from $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$ and $\mathbf{u} + \mathbf{u}' = 2\mathbf{x}$ follows $\mathbf{u} = \mathbf{u}' = \mathbf{x}$.*

Theorem 1 (Weak identifiability)**.** *Let $\mathcal{S} \subseteq \mathbb{R}^n$ be open and connected, $\mathcal{X} \subseteq \mathbb{R}^m$, and $g \colon \mathcal{S} \to \mathcal{X}$ invertible and differentiable. Let us further assume that the observed data satisfy the generative model given in Eq.* (1)*. If $d = \hat{d}$ has one of the following properties:*

(i) there exists a function $\xi \colon \mathbb{R}^+ \to \mathbb{R}^+$, such that $\xi \circ d$ is a norm-induced metric,

(ii) $d(\mathbf{s}, \tilde{\mathbf{s}}) = \sum_i d_i(|s_i - \tilde{s}_i|)$, where each d_i is continuous and strictly increasing,

then the optimal estimator of any of the contrastive losses presented above identifies the true latent factors up to affine transformations, i.e., $h = f \circ g$ is an affine mapping.

Proof. By Lemma 1 and Lemma 2, h is injective. Furthermore, since h is continuous and $\mathcal{S}$ is open, $h(\mathcal{S})$ is also open. Additionally, $h(\mathcal{S})$ is connected because $\mathcal{S}$ is connected.

Assume that condition *(i)* holds. Then, h is an isometry and we can apply Mankiewicz's theorem, which tells us that h is an affine mapping.

Let us now consider condition *(ii)*. We show that h is affine using central ideas of Lemma 2 in [11]. Let $\mathbf{s}_0$ be a 2-extremal point of $\mathcal{B}_r = \{\mathbf{s} \in \mathbb{R}^n : d(0, \mathbf{s}) \leq r\}$. Then $-\mathbf{s}_0$ is also a 2-extremal point of $\mathcal{B}_r$, since d is symmetric. Now define $\phi(\mathbf{s}) = h(\mathbf{s}) - h(0)$ and $\mathcal{A} = (\mathcal{B}_r + \mathbf{s} + \mathbf{s}_0) \cap (\mathcal{B}_r + \mathbf{s} - \mathbf{s}_0)$ for $\mathbf{s} \in \mathcal{S}$. It is straightforward to verify that $\mathbf{s}$ is the only element contained in $\mathcal{A}$. Since h is injective and d is translation invariant, we have

$$\phi(\mathcal{A}) = (\phi(\mathcal{B}_r) + \phi(\mathbf{s} + \mathbf{s}_0)) \cap (\phi(\mathcal{B}_r) + \phi(\mathbf{s} - \mathbf{s}_0)),$$

and because d is symmetric we can conclude

$$\phi(\mathbf{s}) = \frac{1}{2} \left(\phi(\mathbf{s} + \mathbf{s}_0) + \phi(\mathbf{s} - \mathbf{s}_0)\right). \tag{23}$$

It is obvious that for the standard basis vectors $\mathbf{e}_i \in \mathbb{R}^n$ each $a_i \mathbf{e}_i$ with $d(0, a_i \mathbf{e}_i) = r$ is a 2-extremal point of $\mathcal{B}_r$ since each d_i is strictly monotonically increasing. For each pair of different $\mathbf{s}, \mathbf{s}' \in \mathcal{B}_r$ with $\mathbf{s} + \mathbf{s}' = 2a_i \mathbf{e}_i$ we would have either $s_i = s'_i = a_i$ and for at least one coordinate $j \neq i$ $-s_j = s'_j \neq 0$ or one of $\mathbf{s}$, $\mathbf{s}'$ would have a higher value than a_i at the i-th coordinate. So in both cases $\mathbf{s}$ or $\mathbf{s}'$ would lie outside $\mathcal{B}_r$, which is a contradiction. It follows from Eq. (23) that straight lines in the direction of one of the standard basis vectors are mapped to straight lines, i.e., $\phi(\mathbf{s} + a_i \mathbf{e}_i) = \phi(\mathbf{s}) + a_i \left(\phi(\mathbf{s} + \mathbf{e}_i) - \phi(\mathbf{s})\right)$.

Furthermore we have

$$d(0, a_i \mathbf{e}_i) = d(\mathbf{s}, \mathbf{s} + a_i \mathbf{e}_i) = d(\phi(\mathbf{s}), \phi(\mathbf{s} + a_i \mathbf{e}_i)) = d(0, a_i(\phi(\mathbf{s} + \mathbf{e}_i) - \phi(\mathbf{s}))),$$

and since the distance does not depend on $\mathbf{s}$, we obtain $\phi(\mathbf{s} + a_i \mathbf{e}_i) = \phi(\mathbf{s}) + a_i \phi(\mathbf{e}_i)$. After iterating over all $\mathbf{e}_i$, we can conclude $\phi(\mathbf{s} + \sum_i a_i \mathbf{e}_i) = \phi(\mathbf{s}) + \sum_i a_i \phi(\mathbf{e}_i)$.

□

Theorem 2 (Strong identifiability)**.** *Assume that all conditions in Theorem 1 are satisfied. Let the function d in Eq.* (1) *be defined by*

$$d(\mathbf{s}, \tilde{\mathbf{s}}) = \sum_i (|s_i - \tilde{s}_i| / \sigma_i)^\beta, \tag{24}$$

with $\beta \in (0, 2) \cup (2, \infty)$ and $\sigma_i > 0$ for all i, then $h = f \circ g$ is a generalized permutation matrix, i.e., a composition of a permutation and element-wise scaling and sign flips.

Proof. From Theorem 1 we know that $h(\mathbf{s}) = \mathbf{A}\mathbf{s} + \mathbf{b}$. For $p \geq 1$ we can define $\xi(x) = x^{1/p}$. Thus, h preserves the p-norm. In this case it has already been proved that $\mathbf{A}$ is a generalized permutation matrix (see, for example, [8]).

Let us now look at the case $0 < p < 1$. We denote the element in the i-th row and j-th column of $\mathbf{A}$ with a_{ij}. For any standard basis vector $\mathbf{e}_i$, it holds

$$1 = d(0, \mathbf{e}_i) = d(0, \mathbf{A}\mathbf{e}_i) = \sum_j |a_{ji}|^p.$$

Similarly for two different basis vectors $\mathbf{e}_i$ and $\mathbf{e}_j$, we have

$$2 = d(0, \mathbf{e}_i + \mathbf{e}_j) = d(0, \mathbf{A}(\mathbf{e}_i + \mathbf{e}_j)) = \sum_k |a_{ki} + a_{kj}|^p \leq \sum_k |a_{ki}|^p + \sum_k |a_{kj}|^p.$$

The last part is due to the triangle inequality. This implies that a_{ki} and a_{kj} must have the same sign or at least one of them must be zero.

On the other hand we have

$$2 = d(0, \mathbf{e}_i - \mathbf{e}_j) = d(0, \mathbf{A}(\mathbf{e}_i - \mathbf{e}_j)) = \sum_k |a_{ki} - a_{kj}|^p \leq \sum_k |a_{ki}|^p + \sum_k |a_{kj}|^p.$$

Therefore, a_{ki} and a_{kj} must have different signs or at least one of them must be zero. Taken together, this means that each row of $\mathbf{A}$ can have at most one nonzero entry. And since $\mathbf{A}$ is invertible, it is a generalized permutation matrix. □

B Alternative Convergence Proof for the δ-INCE Loss

Proof. We first make the substitution $\psi(\mathbf{x}, \tilde{\mathbf{x}}) = -\delta(f(\mathbf{x}), f(\tilde{\mathbf{x}}))$ and replace $\tilde{\mathbf{x}}$ with $\tilde{\mathbf{x}}_0$ and $\mathbf{x}_i^-$ with $\tilde{\mathbf{x}}_i$ for $i \neq 0$. The δ-INCE loss can then be formulated as

$$\tilde{\mathcal{L}}_{\delta\text{-INCE}}(\psi) = \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}}_0)\sim p_{\text{pos}} \\ \{\tilde{\mathbf{x}}_i\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} p_{\mathbf{x}^-}}} -\log \frac{e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_0)}}{\sum_{i=0}^K e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}}. \tag{25}$$

The logarithmic term can be also written as

$$r(\psi) = -\log \frac{e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_0)}}{\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}} = -\psi(\mathbf{x},\tilde{\mathbf{x}}_0) + \log \sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}$$

and has the following Gateaux derivatives:

$$\left.\frac{\mathrm{d}}{\mathrm{d}\epsilon} r(\psi + \epsilon\eta)\right|_{\epsilon=0} = \frac{\sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)}(\eta(\mathbf{x},\tilde{\mathbf{x}}_j) - \eta(\mathbf{x},\tilde{\mathbf{x}}_0))}{\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}}$$

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}\epsilon^2} r(\psi + \epsilon\eta)\right|_{\epsilon=0} = \frac{\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)} \sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)} \eta(\mathbf{x},\tilde{\mathbf{x}}_j)^2 - \left(\sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)} \eta(\mathbf{x},\tilde{\mathbf{x}}_j)\right)^2}{\left(\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}\right)^2}.$$

Thus, the Tylor series of $\tilde{\mathcal{L}}_{\delta\text{-INCE}}$ is given by

$$\tilde{\mathcal{L}}_{\delta\text{-INCE}}(\psi + \epsilon\eta) = \tilde{\mathcal{L}}_{\delta\text{-INCE}}(\psi) + \epsilon \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}}_0)\sim p_{\text{pos}} \\ \{\tilde{\mathbf{x}}_i\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} p_{\mathbf{x}^-}}} \frac{\sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)}(\eta(\mathbf{x},\tilde{\mathbf{x}}_j) - \eta(\mathbf{x},\tilde{\mathbf{x}}_0))}{\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}}$$

$$+ \frac{1}{2}\epsilon^2 \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}}_0)\sim p_{\text{pos}} \\ \{\tilde{\mathbf{x}}_i\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} p_{\mathbf{x}^-}}} \frac{\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)} \sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)} \eta(\mathbf{x},\tilde{\mathbf{x}}_j)^2 - \left(\sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)} \eta(\mathbf{x},\tilde{\mathbf{x}}_j)\right)^2}{\left(\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}\right)^2} + \mathcal{O}(\epsilon^3). \tag{26}$$

A sufficient condition for $\tilde{\mathcal{L}}_{\delta\text{-INCE}}$ to reach an optimum is that the term of order ϵ vanishes and the term of order ϵ^2 is strictly positive for all directions η. Note that the choice of the variable $\tilde{\mathbf{x}}_0$ for the positive example was arbitrary. By iteratively renaming the positive example to $\tilde{\mathbf{x}}_k$ for $k = 0, \dots, K$ and calculating the mean, we can transform the term of order ϵ in the following way:

$$\mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}}_0)\sim p_{\text{pos}} \\ \{\tilde{\mathbf{x}}_i\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} p_{\mathbf{x}^-}}} \frac{\sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)}(\eta(\mathbf{x},\tilde{\mathbf{x}}_j) - \eta(\mathbf{x},\tilde{\mathbf{x}}_0))}{\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}}$$

$$= \frac{1}{K+1} \sum_{k=0}^{K} \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}}_k)\sim p_{\text{pos}} \\ \{\tilde{\mathbf{x}}_i\}_{i\neq k}^K \overset{\text{i.i.d.}}{\sim} p_{\mathbf{x}^-}}} \frac{\sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)}(\eta(\mathbf{x},\tilde{\mathbf{x}}_j) - \eta(\mathbf{x},\tilde{\mathbf{x}}_k))}{\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}}. \quad (27)$$

This expression vanishes for all η if and only if

$$\prod_l p_{\mathbf{x}^-}(\tilde{\mathbf{x}}_l) \sum_k p(\mathbf{x}|\tilde{\mathbf{x}}_k) \frac{\sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)}(\eta(\mathbf{x},\tilde{\mathbf{x}}_j) - \eta(\mathbf{x},\tilde{\mathbf{x}}_k))}{\sum_i e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}} = 0,$$

which can be simplified to

$$\sum_k p(\mathbf{x}|\tilde{\mathbf{x}}_k) \sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)} \eta(\mathbf{x},\tilde{\mathbf{x}}_j) = \sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)} \sum_k p(\mathbf{x}|\tilde{\mathbf{x}}_k)\eta(\mathbf{x},\tilde{\mathbf{x}}_k).$$

Since this must hold for all $\tilde{\mathbf{x}}_i$ and arbitrary η, we obtain

$$\frac{e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_i)}}{\sum_j e^{\psi(\mathbf{x},\tilde{\mathbf{x}}_j)}} = \frac{p(\mathbf{x}|\tilde{\mathbf{x}}_i)}{\sum_j p(\mathbf{x}|\tilde{\mathbf{x}}_j)}.$$

Obviously, for any solution ψ^*, the function $\psi^*(\mathbf{x},\tilde{\mathbf{x}}) + c(\mathbf{x})$ is also a solution, where $c(\mathbf{x})$ is a function that only depends on $\mathbf{x}$. Thus, we finally obtain $\psi^*(\mathbf{x},\tilde{\mathbf{x}}) = \log p(\mathbf{x}|\tilde{\mathbf{x}}) + c(\mathbf{x})$ or equivalently $\psi^*(\mathbf{x},\tilde{\mathbf{x}}) = \log p(\tilde{\mathbf{x}}|\mathbf{x}) - \log p_{\mathbf{x}^-}(\tilde{\mathbf{x}}) + \gamma_{\mathbf{x}}(\mathbf{x})$ using Bayes' theorem, where $\gamma_{\mathbf{x}}(\mathbf{x}) = c(\mathbf{x}) + \log p_{\mathbf{x}}(\mathbf{x})$.

Plugging this back into $\tilde{\mathcal{L}}_{\delta\text{-INCE}}$ we get

$$\tilde{\mathcal{L}}_{\delta\text{-INCE}}(\psi^* + \epsilon\eta) = \tilde{\mathcal{L}}_{\delta\text{-INCE}}(\psi^*)$$

$$+ \frac{1}{2}\epsilon^2 \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}}_0)\sim p_{\text{pos}} \\ \{\tilde{\mathbf{x}}_i\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} p_{\mathbf{x}^-}}} \frac{\sum_i p(\mathbf{x}|\tilde{\mathbf{x}}_i) \sum_j p(\mathbf{x}|\tilde{\mathbf{x}}_j)\eta(\mathbf{x},\tilde{\mathbf{x}}_j)^2 - \left(\sum_j p(\mathbf{x}|\tilde{\mathbf{x}}_j)\eta(\mathbf{x},\tilde{\mathbf{x}}_j)\right)^2}{\left(\sum_i p(\mathbf{x}|\tilde{\mathbf{x}}_i)\right)^2} + \mathcal{O}(\epsilon^3). \quad (28)$$

The term of order ϵ^2 can be rearranged into

$$\frac{1}{4}\epsilon^2 \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}}_0)\sim p_{\text{pos}} \\ \{\tilde{\mathbf{x}}_i\}_{i=1}^K \overset{\text{i.i.d.}}{\sim} p_{\mathbf{x}^-}}} \frac{\sum_i \sum_j p(\mathbf{x}|\tilde{\mathbf{x}}_i)p(\mathbf{x}|\tilde{\mathbf{x}}_j) \left(\eta(\mathbf{x},\tilde{\mathbf{x}}_i) - \eta(\mathbf{x},\tilde{\mathbf{x}}_j)\right)^2}{\left(\sum_i p(\mathbf{x}|\tilde{\mathbf{x}}_i)\right)^2}, \quad (29)$$

which is strictly positive in any direction.

□

C Experimental Details

In all our experiments, we largely follow the experimental setup and evaluation protocol of [12], with some differences due to our novel approach and hardware limitations, which we describe below. Both α and $\tilde{\alpha}$ are parameterized by separate three-layer neural networks with dimensions 20, 20, and 1. The second layer has an additional skip connection and we use GELUs [5] as hidden activation. We normalize both networks to a mean of zero over a batch and use an additional learnable bias c.

Table 5: Identifiability on synthetic data. R^2 [%] mean $\pm$ standard deviation over 2 random seeds

Scenario	β	δ-NCE	δ-INCE	δ-SCL	δ-NWJ
Box (simple)	1/2	99.48 ± 0.03	99.26 ± 0.35	86.38 ± 2.55	99.71 ± 0.03
Box (simple)	1	99.89 ± 0.01	99.88 ± 0.01	92.80 ± 1.03	99.83 ± 0.04
Box (simple)	3	99.68 ± 0.17	94.49 ± 4.90	97.21 ± 0.14	99.65 ± 0.11
Box (simple)	5	99.74 ± 0.06	94.31 ± 4.66	97.85 ± 0.59	99.73 ± 0.02
Box (complex)	1	98.72 ± 0.54	99.80 ± 0.09	91.51 ± 1.25	97.17 ± 2.75
Box (complex)	3	99.87 ± 0.01	99.76 ± 0.01	97.37 ± 0.18	99.79 ± 0.04
Hollow ball	1	99.54 ± 0.21	99.51 ± 0.13	87.88 ± 6.63	99.56 ± 0.14
Hollow ball	3	95.39 ± 4.34	96.78 ± 0.03	96.05 ± 0.09	98.41 ± 0.15
Hollow ball	5	97.61 ± 0.33	94.08 ± 0.87	95.98 ± 0.16	97.90 ± 0.29
Cube grid	1	99.82 ± 0.03	99.54 ± 0.03	95.67 ± 2.37	99.72 ± 0.01
Cube grid	5	96.80 ± 0.09	88.74 ± 1.03	94.65 ± 0.06	97.83 ± 0.01

Table 6: R^2 [%] scores on synthetic data for $\alpha = \tilde{\alpha} = c$

Scenario	β	δ-NCE	δ-INCE	δ-SCL	δ-NWJ
Box (simple)	1/2	99.21	99.69	86.83	88.30
Box (simple)	1	99.77	99.91	92.85	99.70
Box (simple)	3	99.61	99.31	99.65	99.59
Box (simple)	5	99.84	99.13	94.15	99.83
Box (complex)	1	99.54	99.71	85.08	99.45
Box (complex)	3	99.46	99.70	91.79	99.10
Hollow ball	1	97.13	98.91	79.50	96.06
Hollow ball	3	97.55	95.64	94.21	97.69
Hollow ball	5	97.19	93.77	97.11	97.18
Cube grid	1	99.75	99.14	97.20	99.62
Cube grid	5	95.88	83.73	97.20	97.41

Our model is optimized using Adam [7] with a base learning rate of 10^{-4} for the encoder and c. For α and $\tilde{\alpha}$ we use a learning rate of 10^{-2}.

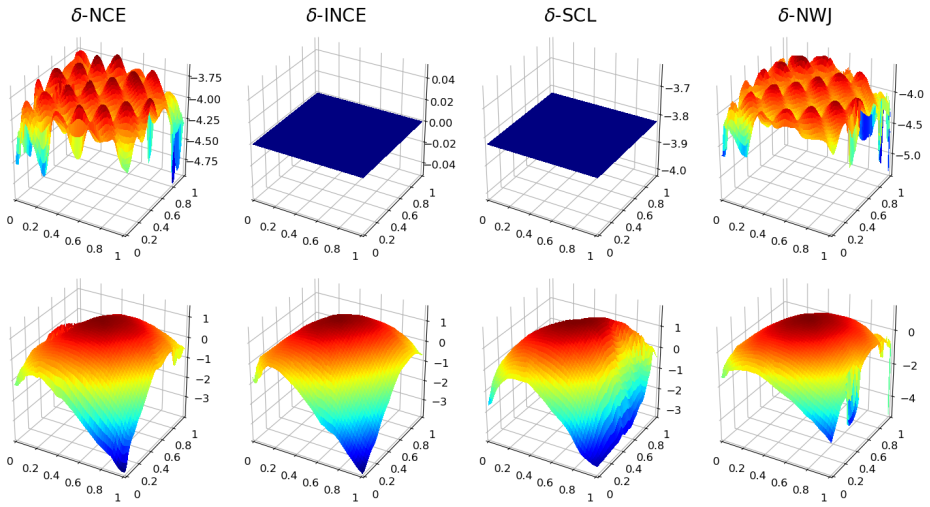
We train all models for 3×10^5 iterations, except for 3DIdent, where we train for 1×10^5 iterations, and the experiments in Section 4.5 when $n > 10$, where we set the number of iterations to 8×10^5.

We use the same encoder architecture and batch size as in [12] on KITTI Masks (64) and 3DIdent (512). However, for the experiments in sections 4.1, 4.2 and 4.5, we use a smaller neural network with residual connections and a smaller batch size of 5120 when $n \leq 10$ and 4096 when $n > 10$. The residual network has 2 hidden layers with $n \cdot 10$ and $n \cdot 20$ dimensions followed by 3 residual blocks and the output layer. Each residual block has 2 layers with $n \cdot 20$ dimensions each. In all hidden layers we use leaky ReLU activations and batch normalization.

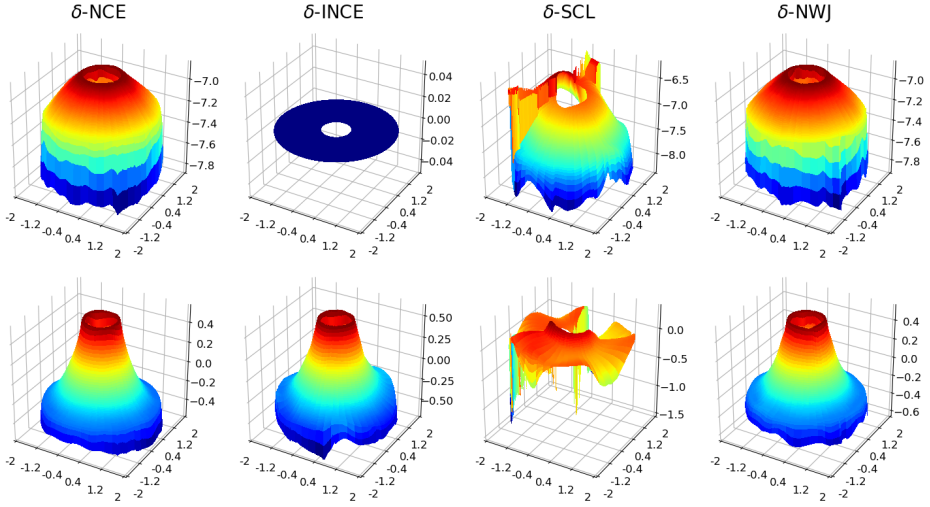
The experiments in Section 4.1 and Section 4.2 ran on a GeForce GTX 1080 Ti GPU for between 7 and 30 hours, depending on the space configuration, shape parameter, and loss function. The experiments on KITTI Masks took on average 2 hours on a GeForce GTX 1080 Ti GPU and the experiments on 3DIdent took on average 24 hours on four GeForce GTX 1080 Ti GPUs.

D Additional Experiments

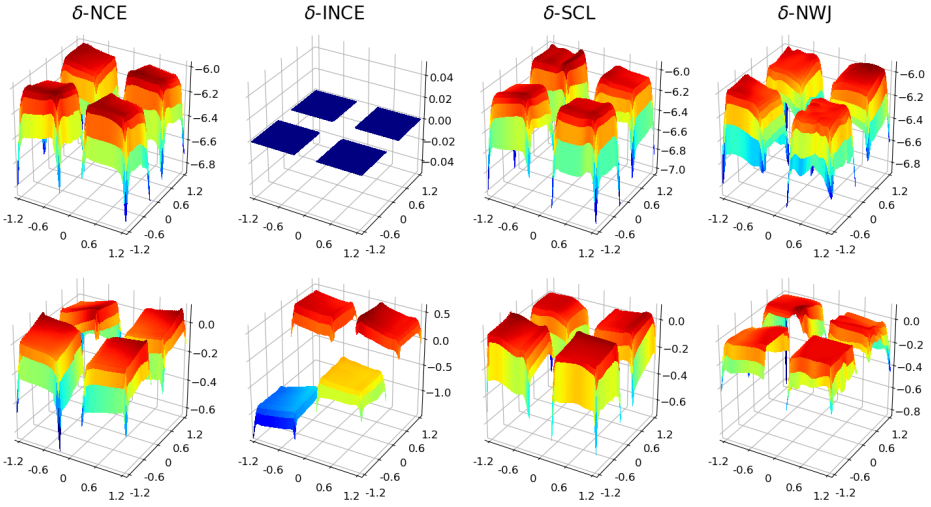
Table 5 and Table 6 contain the corresponding R^2 values for the experiments in Section 4.1 and Section 4.2, respectively. Figure 4 shows the learned functions α and $\tilde{\alpha}$ for the three configurations *box (complex)*, *hollow ball* and *cube grid*. Furthermore, Table 7 shows R^2 and MCC values for the original SCL [4] loss.



(a) Box (complex)



(b) Hollow ball



(c) Cube grid

Figure 4: Learned functions $\alpha \circ h$ (top rows) and $\tilde{\alpha} \circ h$ (bottom rows) for configurations (a) box (complex), (b) hollow ball and (c) cube grid. In the case of δ-INCE, α is set to zero.

Table 7: Identifiability on synthetic data for the original SCL [4]. R^2 and MCC [%] mean $\pm$ standard deviation over 2 random seeds

Scenario	β	R^2	MCC
Box (simple)	1/2	94.46 ± 2.94	55.52 ± 4.63
Box (simple)	1	89.06 ± 0.09	57.38 ± 0.89
Box (simple)	3	88.80 ± 0.08	57.05 ± 1.62
Box (simple)	5	89.84 ± 0.71	55.51 ± 2.19
Box (complex)	1	91.12 ± 0.07	56.02 ± 0.83
Box (complex)	3	91.02 ± 0.06	51.45 ± 3.14
Hollow ball	1	84.69 ± 1.81	52.97 ± 2.08
Hollow ball	3	22.59 ± 1.11	22.16 ± 1.24
Hollow ball	5	19.97 ± 1.31	23.46 ± 1.22
Cube grid	1	43.88 ± 4.91	32.45 ± 4.24
Cube grid	5	20.09 ± 1.64	23.53 ± 2.19

References

[1] A. Ben-Israel. The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999.

[2] L. E. Brouwer. Beweis der invarianz des n-dimensionalen gebiets. *Mathematische Annalen*, 71:305–313, 1911.

[3] M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2), 2012.

[4] J. Z. HaoChen, C. Wei, A. Gaidon, and T. Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

[5] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[6] A. Hyvärinen and H. Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[7] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] C.-K. Li and W. So. Isometries of ℓ_p-norm. *The American mathematical monthly*, 101(5):452–453, 1994.

[9] Z. Ma and M. Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. *arXiv preprint arXiv:1809.01812*, 2018.

[10] P. Mankiewicz. On extension of isometries in normed linear spaces. *Bull. Acad. Pol. Sci., Sér. Sci. Math. Astron. Phys*, 20:367–371, 1972.

[11] R. Wobst. Isometrien in metrischen vektorräumen. *Studia Mathematica*, 1(54):41–54, 1975.

[12] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.