Unsupervised Learning of Semantic Mappings

Tomer Galanti¹ and Lior Wolf¹,²

¹The Blavatnik School of Computer Science, Tel Aviv University, Israel
²Facebook AI Research

Abstract

We discuss the feasibility of the following learning problem: given unmatched samples from two domains and nothing else, learn a mapping between the two, which preserves semantics. Due to the lack of paired samples and without any definition of the semantic information, the problem might seem ill-posed. Specifically, in typical cases, it seems possible to build infinitely many alternative mappings from every target mapping. This apparent ambiguity stands in sharp contrast to the recent empirical success in solving this problem. A theoretical framework for measuring the complexity of compositions of functions is developed in order to show that the target mapping is of lower complexity than all other mappings. The measured complexity is directly related to the depth of the neural networks being learned and the semantic mapping could be captured simply by learning using architectures that are not much bigger than the minimal architecture.

1. Introduction

Multiple recent reports (Xia et al., 2016; Kim et al., 2017; Zhu et al., 2017; Yi et al., 2017) convincingly demonstrated that one can learn to map between two domains that are each specified merely by a set of unlabeled examples. For example, given a set of unlabeled images of horses, and a set of unlabeled images of zebras, CycleGAN (Zhu et al., 2017) creates the analog zebra image for a new image of a horse and vice versa. These recent methods employ two types of constraints. First, when mapping from one domain to another, the output has to be indistinguishable from the samples of the new domain. This is enforced using GANs (Goodfellow et al., 2014) and is applied at the distribution level: the mapping of horse images to the zebra domain should create images that are indistinguishable from the training images of zebras and vice versa. The second type of constraints enforces that for every single sample, transforming it to the other domain and back (by a composition of the mappings in the two directions) results in the original sample. This is enforced for each training sample from either domain: every training image of a horse (zebra), which is mapped to a zebra (horse) image and then back to the source domain, should be as similar as possible to the original input image. In another example, taken from DiscoGAN (Kim et al., 2017), a function is learned to map a handbag to a shoe of a similar style. One may wonder why striped bags are not mapped, for example, to shoes with a checkerboard pattern. If every striped pattern in either domain is mapped to a checkerboard pattern in the other and vice-versa, then both the distribution constraints and the circularity constraints might hold. The former could hold since both striped and checkerboard patterned objects would be generated. Circularity could hold since, for example, a striped object would be mapped to a checkerboard object in the other domain and then back to the original striped object.

One may claim that the distribution of striped bags is similar to those of striped shoes and that the distribution of checkerboard patterns is also the same in both domains. In this case, the alignment follows from fitting the shapes of the distributions. This explanation is unlikely, since no effort is being made to create handbags and shoes that have the same distributions of these properties, as well as many other properties.
2. The Unsupervised Alignment Problem

The learning algorithm is provided with only two unlabeled datasets: one includes i.i.d samples from the first distribution and the second includes i.i.d samples from the other distribution.

\[
\begin{align*}
  x_i \in X_A & \text{ for } i = 1 \ldots m \text{ where } x_i \sim D_A \text{ and } X_A \text{ denotes the space of domain } A = (X_A, D_A) \\
  x_j \in X_B & \text{ for } j = 1 \ldots n \text{ where } x_j \sim D_B \text{ and } X_B \text{ denotes the space of domain } B = (X_B, D_B)
\end{align*}
\]  

(1)

(all notations are listed in the appendix, see Tab. 1). To semantically tie the two distributions together, our model is based on a generative approach, which is well aligned with the success of GAN-based image generation, e.g., (Radford et al., 2015), in mapping random input vectors into realistic-looking images.

Let \( z \in X \) be a random vector that is distributed according to the distribution \( D_Z \) and which we employ to denote the semantic essence of samples in \( X_A \) and \( X_B \). We denote \( D_A = y_A \circ D_Z \) and \( D_B = y_B \circ D_Z \), where the functions \( y_A : X \rightarrow X_A \) and \( y_B : X \rightarrow X_B \), and \( f \circ D \) denotes the distribution of \( f(x) \), where \( x \sim D \). It makes sense to assume that both \( y_A \) and \( y_B \) are invertible, since given training samples, one may be expected to be able to recover the underlying properties of the generated samples, even with very weak supervision (Chen et al., 2016).

We denote by \( y_{AB} = y_B \circ y_A^{-1} \), the function that maps the first domain to the second domain. It is semantic in the sense that it goes through the shared semantic space \( X \). The goal of the learner is to fit a function \( h \in H \), for some hypothesis class \( H \) that is closest to \( y_{AB} \),

\[
\inf_{h \in H} R_{D_A}[h, y_{AB}],
\]

(2)

where \( R_D[f_1, f_2] = E_{x \sim D} \ell(f_1(x), f_2(x)) \), for a loss function \( \ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \) and a distribution \( D \).

It is not clear that such fitting is possible without further information. Assume, for example, that there is a natural order on the samples in \( X_B \). A mapping that maps an input sample \( x \in X_A \) to the sample that is next in order to \( y_{AB}(x) \) could be just as feasible. More generally, one can permute the samples in \( X_A \) by some function \( \Pi \) that replaces each sample with another sample that has a similar likelihood (the formal definition is given in Sec. 4) and learn \( h \) that satisfies \( h = y_{AB} \circ \Pi \). We call this difficulty “the alignment problem” and our work is dedicated to understanding the plausibility of learning despite this problem.

In the cross domain transfer line of work (Taigman et al., 2017), the alignment problem is dealt with by incorporating a fixed pre-trained feature map \( f \) and requiring what is called \( f \)-constancy, namely that the following risk is small \( R_{D_A}[f, f \circ h] \), or informally, \( f(x) = f(h(x)) \).

In multiple recent contributions (Xia et al., 2016; Kim et al., 2017; Zhu et al., 2017; Yi et al., 2017) circularity is employed. Circularity requires the recovery of both \( y_{AB} \) and \( y_{BA} = y_A \circ y_B^{-1} \) simultaneously. Namely, functions \( h \) and \( h' \) are learned jointly by minimizing the risk:

\[
\inf_{h, h' \in H} \text{disc}_C(h \circ D_A, D_B) + \text{disc}_C(h' \circ D_B, D_A) + R_{D_A}[h' \circ h, Id_A] + R_{D_B}[h \circ h', Id_B]
\]

(3)

where \( \text{disc}_C(D_1, D_2) = \sup_{c_1, c_2 \in C} |R_{D_1}[c_1, c_2] - R_{D_2}[c_1, c_2]| \) denotes the discrepancy between distributions \( D_1 \) and \( D_2 \) that is implemented with a GAN (Ganin et al., 2016). The first term in Eq. 3 ensures that the samples generated by mapping domain \( A \) to domain \( B \) follow the distribution of samples in domain \( B \). The second term is the analog term for the mapping in the other direction. The last two terms ensure that mapping a sample from one domain to the second and back, results in the original sample.

While the circularity constraints, expressed as the last two terms in Eq. 3, are elegant and do not require additional supervision, for every invertible permutation \( \Pi \) of the samples in domain \( B \) (not to be confused with a permutation of the vector elements of the representation of samples in \( B \)) we have

\[
\begin{align*}
  (h' \circ \Pi^{-1}) \circ (\Pi \circ h) & = h \circ h' \approx Id_A, \text{ and} \\
  (\Pi \circ h) \circ (h' \circ \Pi^{-1}) & = \Pi \circ (h \circ h') \circ \Pi^{-1} \approx \Pi \circ Id_B \circ \Pi^{-1} = Id_B.
\end{align*}
\]

(4)
Therefore, every circularity preserving $h$ and $h'$ gives rise to many possible solutions of the form $\tilde{h} = h \circ \Pi$ and $\tilde{h}' = \Pi^{-1} \circ h'$. If $\Pi$ happens to satisfy $DB(x) \approx DB(\Pi(x))$, then the discrepancy terms in Eq. 3 also remain largely unchanged. Circularity by itself cannot, therefore, explain the recent success of unsupervised mapping.

2.1 An Illustrative Example

Despite the availability of a large number of alternative hypotheses $h'$ that satisfy the constraints of Eq. 3, the methods of (Xia et al., 2016; Kim et al., 2017; Zhu et al., 2017; Yi et al., 2017) enjoy empirical success, Why? In order to illustrate our main thesis, we present a very simple toy example, depicted in Fig. 1. Consider the domain $A$ of uniformly distributed points $(x_1, x_2)\top \in \mathbb{R}^2$, where $0 \leq x_1 < 1$ and $x_2 = 0.5$. Let $B$ be a similar domain, except $x_2 = 2$. We are interested in learning the mapping $y_{AB}^2((x_1, 0.5)\top) = (x_1, 2)\top$. We note that there are infinitely many mappings from domain $A$ to $B$ that satisfy the constraints of Eq. 3. However, when we learn the mapping using a neural network with one hidden layer of size 2 and Leaky ReLU activations, parameterized by the weights $W$ and the bias $b$.

While this is an extreme example, we are able to show that limiting the complexity of the admissible solutions

\[ \sigma_a(x) = \text{Ind}[x < 0]ax + \text{Ind}[x \geq 0]x, \text{ for the indicator function Ind}[q] which maps a true value to one, zero otherwise. \]
eliminates the solutions that are derived from $y_{AB}$ by permuting the samples in the space $X_A$, since such mixing requires added complexity.

2.2 Informal Statement of the Main Results

In this work, we show that what separates $h \approx y_{AB}$ and $\hat{h} \approx \Pi \circ y_{AB}$ is the complexity of these functions. Namely, that the complexity of $h$ is much lower and, therefore, it is learnable with much smaller networks than the alternatives. Therefore, in order to learn $h$ and not $\hat{h}$, all that one needs is to use a network that is not “too big”.

To show this, we develop a new framework for measuring the complexity of composition of functions. Our function complexity framework measures the complexity of a function as the depth of a neural network which implements it, or the shallowest network, if there are multiple such networks. In other words, we use the number of layers of a network as a proxy for the Kolmogorov complexity of functions, using layers in lieu of the primitives of the universal Turing machines, which is natural for studying functions that can be computed by feedforward neural networks.

To make it a useful tool, an elaborate theoretical system is developed for this complexity framework. The system is based on well-justified assumptions and is presented in the next section. In Sec. 4, we apply the new framework and study the properties of the low-complexity solutions when learning in an unsupervised manner. We provide a complexity-based definition to the illusive notion of semantics. A semantic mapping is a mapping with the lowest complexity among all permuted versions of it, i.e., it is the most straightforward one in this set. Using this notion, we are able to state our main results informally as: (1) There are only a handful of semantic mappings between two domains. (2) every semantic function between two domains $A$ and $B$ either passes through a shared space $Z$, or is very different than any composition of a semantic function form $A$ to $Z$ with a semantic function from $Z$ to $B$. Therefore, if there is a semantic function through $Z$, one obtains it by learning a minimal network between $A$ and $B$, or alternatively, but unlikely, learns a completely different function through another shared space $Z'$.

Based on our results, we are able to make concrete predictions. The first one, which is empirically validated in the appendix, states that in contrast to the current common wisdom, one can learn a semantic mapping between two spaces without any matching samples and even without circularity.

**Prediction 1** When learning with a small enough network in an unsupervised way a mapping between domains that share common characteristics, the GAN constraint in the target domain is sufficient to obtain a semantic mapping.

The strongest clue that helps identify the semantic mapping from the other mappings is the suitable complexity of the network that is learned. A network with a complexity that is too low cannot replicate the target distribution, when taking inputs in the source domain. A network that has a complexity that is too high, would not learn the semantic mapping.

We believe that the success of the recent methods arises from selecting the architecture used in an appropriate way. For example, DiscoGAN (Kim et al., 2017) employs either eight or ten layers, depending on the dataset. We make the following prediction, which is also validated empirically in the appendix:

**Prediction 2** When learning in an unsupervised way a mapping between domains, the size of the network needs to be carefully adjusted.

This prediction is also surprising, since in supervised learning, extra depth is not as detrimental, if at all. As far as we know, this is the first time that this clear distinction between supervised and unsupervised learning is made.

Although our first prediction states that circularity is not needed, there does seem to be an advantage to using it. We show that using circularity, ambiguity is reduced in a very specific way. Namely, the potential family of permutations $\Pi$ that extend the semantic solution $h$ to the alternatives $\Pi \circ h$ are constrained to be of lower complexity. Therefore, the argument made following Eq. 4 holds only for a very limited set of possible permutations $\Pi$. 
3. A Complexity Measure for Functions

In order to model the composition of neural networks, we define a complexity measurement that assigns a value based on the number of simple functions that make up a complex function.

**Definition 1 (Stratified complexity model (SCM))** A stratified complexity model $\mathcal{N} := \text{SCM}[\mathcal{C}]$ is a hypothesis class of functions $p : \mathbb{R}^M \rightarrow \mathbb{R}^M$ specified by a set of functions $\mathcal{C}$. Every function $p$ in $\mathcal{N}$ has an appropriate decomposition:

- There are $p_1, \ldots, p_n \in \mathcal{C}$ such that $p = p_n \circ p_{n-1} \circ \cdots \circ p_1$ (if $n = 0$ then $p = \text{Id}$ and if $n = 1$ then $p = p_1$).
- Every function in $\mathcal{C}$ is invertible.

Informally, a SCM partitions a set of invertible functions into disjoint complexity classes,

$$
\mathcal{C}_0 := \{\text{Id}\}
$$

$$
\mathcal{C}_n := \left\{ p = p_n \circ \cdots \circ p_1 \bigg| p_n \circ \cdots \circ p_1 \text{ is an appropriate decomposition of } p \right\} \setminus \bigcup_{i=0}^{n-1} \mathcal{C}_i
$$

When considering simple functions $p_i$ that are layers in a neural network, each complexity class contains the functions that are implemented by networks of $n$ layers. In addition, we denote the complexity of a function $p$:

$$
C(p) := \arg_{n} \{ p \in \mathcal{C}_n \}
$$

If the complexity of a function $p$ equals $n$, then any appropriate decomposition $p = p_n \circ \cdots \circ p_1$ will be called a minimal decomposition of $p$. If there is no such $n$, we denote $C(p) = \infty$.

According to this measurement, the complexity of a function $p$ is determined by the minimal number of primitive functions required in order to represent it. It is, therefore, necessary to understand the rules that dictate the complexity of a composition of functions, inverse functions, etc. We begin with the simplest relationship that occurs between two composed functions.

**Definition 2 (Unfused functions)** Let $\mathcal{N} = \text{SCM}[\mathcal{C}]$ and let $p, q \in \mathcal{N}$ be any two functions. We say that $p$ is unfused in $q$ if $C(p \circ q) = C(p) + C(q)$ and denote $p \not\bowtie q$. Otherwise, we say that $p$ is fused in $q$ and denote $p \bowtie q$. In addition, the function $p$ and $q$ will be called left and right partial functions of $p \circ q$ (resp.).

Informally, a function $p$ is unfused in another function $q$, if the first operations of $p$ do not invert the last processing steps of $q$. For instance, if we can represent $p = g_1 \circ g_2$ and $q = g_2^{-1} \circ g_3$ such that $C(g_1) < C(p)$ and $C(g_3) < C(q)$, then (by Lem. 4 below) the two functions are fused.

In addition to characterizing the result of composing two functions, we also define a measure for the complexity of transforming one function into the other. The conditional complexity between the functions $p$ and $q$ is the complexity of the function $g$ that satisfies: $p = g \circ q$.

**Definition 3 (Conditional complexity)** Let $\mathcal{N} = \text{SCM}[\mathcal{C}]$ and let $p, q \in \mathcal{N}$ be any two functions. The conditional complexity between $p$ and $q$ is denoted: $C(p||q) := C(p \circ q^{-1})$.

In this work, we focus our attention on SCMs that represent the architectures of fully connected neural networks with layers of a fixed size, i.e.,

**Definition 4 (NN-SCM)** A NN-SCM is a SCM $\mathcal{N} = \text{SCM}[\mathcal{C}]$ that satisfies the following conditions:

- $\mathcal{C} = \left\{ \sigma \circ W \bigg| W \in \mathbb{R}^{M \times M} \text{ and } W \text{ is invertible} \right\}$. Here, $W$ denotes both a linear transformation and the associated matrix form.

- $\sigma$ is a non-linear element-wise activation function.

For brevity, we denote $\mathcal{N} := \text{SCM}[\sigma]$ to refer to a NN-SCM with the activation function $\sigma$.

The NN-SCM with the Leaky ReLU activation function is of a particular interest, since (Kim et al., 2017; Zhu et al., 2017) employ it as the main activation function (plain ReLUs and $\tanh$ are also used).
3.1 Semantic mappings

The following definition of a semantic mapping is both intuitive and well defined in concrete complexity terms. We say that given two distributions \( D_A \) and \( D_B \), a semantic mapping \( f : X_A \to X_B \) between domains \( A \) and \( B \) is a mapping that has minimal complexity among the functions \( h : X_A \to X_B \) that satisfy \( h \circ D_A \approx D_B \). Consider, again, the example of a line segment in \( \mathbb{R}^M \) (Sec. 2.1) and the semantic space of the \([0, 1]\) \( \subset \mathbb{R} \) interval. The two linear mappings, which map either segment ends to 0 and the other to 1 are semantic, when using \( f \) that are ReLU based neural networks. Other mappings to this segment are possible, simply by permuting points on the segment in \( \mathbb{R}^M \). However, these are much more complicated. In order to measure the distance between \( h \circ D_A \) and \( D_B \) we will use the discrepancy distance, \( \text{disc}_D \). In this work, we will focus on classes of discriminators \( D \) of the form \( D_m := \{ u | C(u) \leq m \} \) for some \( m \in \mathbb{N} \). In addition, for simplicity, we will write \( \text{disc}_m := \text{disc}_{D_m} \).

**Definition 5 (Semantic mapping)** Let \( \mathcal{N} = \text{SCM}[^C] \). Let \( A = (X_A, D_A) \) and \( B = (X_B, D_B) \) be two domains. We define the \((m, \epsilon_0)\)-semantic complexity between \( A \) and \( B \) as:

\[
C_{A,B}^{m,\epsilon_0} := \min_{h \in \mathcal{N} \cup \{0\}} \left\{ \exists h \text{ s.t } C(h) = i \text{ and } \text{disc}_m(h \circ D_A, D_B) \leq \epsilon_0 \right\}
\]

(7)

The set of \((m, \epsilon_0)\)-semantic functions between \( A \) and \( B \) is:

\[
H_{\epsilon_0}(D_A, D_B; m) := H_{\epsilon_0}(D_A, D_B; m, C_{A,B}^{m,\epsilon_0}),
\]

where

\[
H_{\epsilon_0}(D_A, D_B; m, k) := \left\{ h \left| C(h) \leq k \text{ and } \text{disc}_m(h \circ D_A, D_B) \leq \epsilon_0 \right\}
\]

(8)

We note that for any fixed \( \epsilon_0 > 0 \), the sequence \( \{C_{A,B}^{m,\epsilon_0}\}_{m=0}^{\infty} \) is monotonically increasing as \( m \) tends to infinity. In addition, we assume that for every two distributions of interest, \( D_I \) and \( D_J \), and an error rate \( \epsilon_0 > 0 \), there is a function \( h \) of finite complexity such that \( \text{disc}_C(h \circ DI, DJ) \leq \epsilon_0 \). Therefore, the sequence \( \{C_{A,B}^{m,\epsilon_0}\}_{m=0}^{\infty} \) is upper bounded by \( C(h) \) for all \( m \in \mathbb{N} \cup \{0\} \). In particular, there is a minimal value \( m_0 > 0 \) such that \( C_{A,B}^{m,\epsilon_0} = C_{A,B}^{m_0,\epsilon_0} \) for all \( m \geq m_0 \). We denote: \( E_{A,B}^{\epsilon_0} := m_0 \) and \( C_{A,B}^{\epsilon_0} := C_{A,B}^{m_0,\epsilon_0} \).

3.2 Identifiability

Every neural network implementation gives rise to many alternative implementations by performing simple operations, such as permuting the units of any hidden layer, and then permuting back as part of the linear mapping in the next layer. Therefore, it is first required to identify and address the set of transformations that could be inconsequential to the function which the network computes.

**Definition 6 (Invariant set)** Let \( \mathcal{N} = \text{SCM}[\sigma] \) be a NN-SCM. The invariant set \( \text{Invariant}(\mathcal{N}) \) is the set of all \( \pi : \mathbb{R}^M \to \mathbb{R}^M \) that satisfy the following conditions:

- \( \pi : \mathbb{R}^M \to \mathbb{R}^M \) is an invertible linear transformation.
- \( \sigma \circ \pi = \pi \circ \sigma \).

Functions in \( \text{Invariant}(\mathcal{N}) \) are called invariants or invariant functions.

For example, for neural networks with the \( \tanh \) activation function, the set of invariant functions contains the linear transformations that take vectors, permute them and multiply each coordinate by \( \pm 1 \). Formally, each \( \pi = [\epsilon_1 \cdot e_{i(1)}, ..., \epsilon_M \cdot e_{i(M)}]^T \) where \( \epsilon_i \) is the \( i \)’th standard basis vector, \( t \) is a permutation over \([M]\) and \( \epsilon_i \in \{ \pm 1 \} \) (Fefferman and Markel, 1993).

Our analysis is made much simpler, if every function has one invariant representation up to a sequence of manipulations using invariant functions that do not change the essence of the processing at each layer.

**Definition 7 (Identifiability of minimal representation)** A NN-SCM \( \mathcal{N} = \text{SCM}[\sigma] \) obeys identifiability of minimal representation with respect to \( \text{Invariant}(\mathcal{N}) \), if for all \( n \in \mathbb{N} \cup \{0\} \) and \( p \in C_n \) such that there
are two appropriate decompositions \( p = p_n \circ \ldots \circ p_1 \) and \( p = q_n \circ \ldots \circ q_1 \), then there are invariants \( \pi_1, \ldots, \pi_n \in \text{Invariant}(N) \) such that:

\[
q_i = \sigma \circ (\pi_i \circ W_i) \quad \text{and} \quad p_i = \sigma \circ W_i \\
\forall 1 < i < n: q_i = \sigma \circ (\pi_i \circ W_i \circ \pi_i^{-1}) \quad \text{and} \quad p_i = \sigma \circ W_i \quad \text{(9)}
\]

We consider that since any \( \pi \in \text{Invariant}(N) \) commutes with \( \sigma \), then, an alternative writing could be:

\[
q_1 = \pi_1 \circ p_1, \forall i = 2, ..., n - 1; q_i = \pi_i \circ p_i \circ \pi_i^{-1} \quad \text{and} \quad q_n = p_n \circ \pi_n^{-1} \quad \text{(10)}
\]

A stronger identifiability condition requires that every non-minimal implementation of a function \( p \) goes through the same processing steps as dictated by the layers of the minimal representation, where each of these steps can be mapped to multiple layers in the longer implementation.

**Definition 8 (Identifiability)** Let \( N = \text{SCM}[\sigma] \) be a NN-SCM obeying identifiability of minimal representation. We say that \( N \) obeys identifiability if for every function \( p \in N \), such that \( p = p_n \circ \ldots \circ p_1 \) is a minimal decomposition, if \( p = q_n \circ \ldots \circ q_1 \), then:

\[
\exists j_1 < \ldots < j_{n+1} = m + 1, \pi_1, \ldots, \pi_n \in \text{Invariant}(N) : q_{j_1:j_i} = \pi_1 \circ p_1, \forall i = 2, ..., n - 1; q_{j_i:j_{i+1}} = \pi_i \circ p_i \circ \pi_i^{-1} \quad \text{and} \quad q_{j_n:j_{n+1}} = p_n \circ \pi_n^{-1} \quad \text{(11)}
\]

**Here, if** \( u = u_n \circ \ldots \circ u_1 \) **then** \( u_{i+1:j} = u_i \circ \ldots \circ u_j \) **if** \( j \leq i \) **and** \( u_{i:i} = \text{Id} \).

In the context of neural networks, the general question of uniqueness up to invariants, also known as identifiability, is an open question. Nevertheless, several authors have made progress in this area for different neural network architectures. The most notable work has been done by Fefferman and Markel (1993) that proves identifiability for \( \sigma = \text{tanh} \). Furthermore, the representation is unique up to the invariant functions. Other works (Williamson and Helmke, 1995; F. Albertini and Maillot, 1993; Kurková and Kainen, 2014; Sussmann, 1992) prove such uniqueness for neural networks with only one hidden layer and various activation functions. As far as we know, there are no recent results continuing this line of work for activation functions such as Leaky ReLU. Uniqueness, which is stronger than identifiability, since it means that even multiple representations with different number of layers do not exist, does not hold for these activation functions. To see this, note that for every \( M \times M \) mapping \( W \), the following holds:

\[
\sigma \circ W = (\sigma \circ W) \circ (\sigma \circ -\text{Id}) \circ (\sigma \circ -\text{Id}/a) \quad \text{(12)}
\]

where \( \sigma \) is the Leaky ReLU activation function with parameter \( a \).

We conjecture that for networks with Leaky ReLU activations identifiability holds, or at least for networks with a fixed number of neurons per layer.

### 3.3 Properties of Inverses and Compositions

It is necessary to study the effect of inversion on the complexity of functions, since, for example, we care about both \( h' = \Pi \circ h \) and \( h = \Pi^{-1} \circ h' \).

**Definition 9** Let \( N = \text{SCM}[\sigma] \) be an SCM. We say that \( N \) is d-inverse-complexity-preserving (d-ICP for short) if: \( \forall p \in C : C(p^{-1}) \leq d \cdot C(p) \). Sometimes, we will omit writing \( I(N) \) and write \( I \) instead, when \( N \) is obvious from the context.

An immediate consequence of the following theorem is that neural networks with Leaky ReLU activations are 3-ICP, see also Lem. 11 in the appendix, which is part of the theorem’s proof (all proofs can be found in the appendix).
Theorem 1 Let \( \mathcal{N} = \text{SCM}[\sigma] \) be a NN-SCM with \( \sigma \) that is the Leaky ReLU with parameter \( a > 0 \). Then, for any \( u \in \mathcal{N} \), \( |C(u^{-1}) \circ \ldots \circ D,D) \leq \epsilon_0 \) \( (14) \)

We denote \((m, \epsilon_0)\)-DPMs of complexity \( k \) by \( \text{DPM}_{\epsilon_0}(D; m, k) := \left\{ \Pi \mid \text{disc}_m(\Pi \circ D, D) \leq \epsilon_0 \text{ and } C(\Pi) = k \right\} \).

As we discussed earlier, two functions \( f \) and \( g \) can be either fused or unfused. In the following Theorem, it is shown that a generic decomposition \( u_{n+1:1} \) is unfused, i.e, \( C(u_{n+1:1}) = n \). In particular, for generic functions \( f \) and \( g \) are unfused.

Theorem 2 Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). The set of sequences of invertible matrices \((W_1, \ldots, W_n) \in \mathbb{R}^{M \times M \times n}\) such that \( C(u_{n+1:1}^{-1}) = C(u_{n+1:1}) + 2 \) and \( u_i = \sigma \circ W_i \) (for \( i \in \{n\} \)) is open and dense in \( \mathbb{R}^{M \times M \times n} \).

As we discussed earlier, two functions \( f \) and \( g \) can be either fused or unfused. In the following Theorem, it is shown that a generic decomposition \( u_{n+1:1} \) is unfused, i.e, \( C(u_{n+1:1}) = n \). In particular, for generic functions \( f \) and \( g \) are unfused.

Theorem 3 Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). The set of sequences of invertible matrices \((W_1, \ldots, W_n) \in \mathbb{R}^{M \times M \times n}\) such that \( C(u_{n+1:1}) = n \) where \( u_i = \sigma \circ W_i \) (for \( i \in \{n\} \)) is open and dense in \( \mathbb{R}^{M \times M \times n} \).

It is still important to cover both fused and unfused functions, since our main results are “for every” and not “for general” due to the fact that we study specific functions, and specifically we study relations of the form \( h' = \Pi \circ h \), and look for the \( h' \) with the minimal complexity, for which, \( C(h') < C(h) \).

If two functions are known to be unfused then, from definition, the complexity of their composition is exactly the sum of their complexities. It is natural to assume that the same holds for the composition of three unfused functions. However, this is not necessarily the situation when the middle function in this composition has a complexity that is lower than the ICP factor \( I \) (see Lem. 13 in the appendix). However, when the function in the middle of the composition has a complexity \( \geq I \), the complexity of the composition is the sum of the complexities.

Theorem 4 Let \( \mathcal{N} = \text{SCM}[\sigma] \) be a NN-SCM obeying identifiability. Then, for all \( f, g, h \in \mathcal{N} \) such that \( f \not\sim g \), \( g \not\sim h \) and \( I \leq C(g) \), we have:

\[
C(f \circ g \circ h) = C(f) + C(g) + C(h)
\]

(13)

4. Learning Semantic Mappings

In this section, we present the theoretical foundations for unsupervised alignment algorithms. In the unsupervised alignment problem, the algorithms are provided with only two unmatched datasets of samples from the domains \( A \) and \( B \) and the task is to learn a semantic function between them. The goal of this section is to understand under which constraints one is able to learn the semantic mapping using unsupervised alignment.

4.1 Counting Semantic Mappings

Recall that \( \text{disc}_m \) is the discrepancy distance for discriminators of complexity up to \( m \). We have discussed the functions \( \Pi \) which replaces between members in the domain \( B \) that have similar probabilities. Formally, these are defined using the discrepancy distance.

Definition 10 (Density preserving mapping) Let \( \mathcal{N} = \text{SCM}[\mathcal{C}] \) and \( D \) a distribution. A \((m, \epsilon_0)\)-density preserving mapping over \( D \) (or an \((m, \epsilon_0)\)-DPM for short) is a function \( \Pi \) such that

\[
\text{disc}_m(\Pi \circ D, D) \leq \epsilon_0
\]

(14)

We denote \((m, \epsilon_0)\)-DPMs of complexity \( k \) by \( \text{DPM}_{\epsilon_0}(D; m, k) := \left\{ \Pi \mid \text{disc}_m(\Pi \circ D, D) \leq \epsilon_0 \text{ and } C(\Pi) = k \right\} \).
We would like to bound the number of shared semantic distributions by the number of DPMs. We consider that there are infinitely many DPMs and semantic mappings. For example, if we slightly perturb the weights of a minimal representation of DPM, II, we obtain a new DPM. Therefore, we define a relation between of functions that reflects whether the two are similar. In this way, we are able to bound the number of different (not-similar) semantic mappings by the number of different DPMs.

**Definition 11 (Similarity between pairs of distributions or functions)** Let \( \mathcal{N} = \text{SCM}[\sigma] \):

- **Distributions** \( D_1 \) and \( D_2 \) are \((m, \epsilon_0)\)-similar and we denote
  \[
  D_1 \sim_{m, \epsilon_0} D_2 \iff \text{disc}_m(D_1, D_2) \leq \epsilon_0
  \] (15)

- **Functions** \( f \) and \( g \) are \((D, m, \epsilon_0)\)-similar and we denote \( f \sim_{D, m, \epsilon_0} g \), if \( C(f) = C(g) =: n \) and there are minimal decompositions: \( f = f_{n+1,1} \) and \( g = g_{n+1,1} \) such that
  \[
  \forall i \in [n]: f_{i+1,1} \circ D \sim_{m, \epsilon_0} g_{i+1,1} \circ D
  \] (16)

The defined similarity is reflexive and symmetric, but not transitive. Therefore, there are many different ways to partition the space of functions into disjoint subsets such that in each subset, any two functions are similar. We count the number of functions up to the similarity as the minimal number of subsets required in order to cover the entire space. This idea is presented in the following Def. 12.

**Definition 12 (Covering numbers)** Let \( \langle \mathcal{U}, \sim_{\mathcal{U}} \rangle \) be a set and a reflexive and symmetric relation. A covering of \( \langle \mathcal{U}, \sim_{\mathcal{U}} \rangle \), is a tuple \( \langle \mathcal{U}, \equiv_{\mathcal{U}} \rangle \) such that: \( \equiv_{\mathcal{U}} \) is an equivalence relation and \( u_1 \equiv_{\mathcal{U}} u_2 \implies u_1 \sim_{\mathcal{U}} u_2 \). The covering number of \( \langle \mathcal{U}, \sim_{\mathcal{U}} \rangle \) is:

\[
\min \left| \mathcal{U} / \equiv_{\mathcal{U}} \right| \text{ s.t: the minimum is taken over } \langle \mathcal{U}, \equiv_{\mathcal{U}} \rangle \text{ that is a covering of } \langle \mathcal{U}, \sim_{\mathcal{U}} \rangle
\] (17)

We denote the covering number of \( \langle \mathcal{U}, \sim_{\mathcal{U}} \rangle \) by \( \text{Covering}(\langle \mathcal{U}, \sim_{\mathcal{U}} \rangle) \). Here, \( \mathcal{U} / \equiv_{\mathcal{U}} \) is the quotient set of \( \mathcal{U} \) by \( \equiv_{\mathcal{U}} \).

Informally, the following theorem states that the number of semantic mappings is upper bounded by the square root of the number of DPMs of size \( 2C_{A,B}^n \). This result is useful since DPMs are expected to be rare in real-world domains. When imagining mapping a space to itself, in a way that preserves the distribution, one first considers symmetries. Near-perfect symmetries are rare in natural domains, and when these occur, e.g., (Kim et al., 2017), they form well-understood ambiguities. Another option that can be considered is that of replacing specific samples in domain \( B \) with other samples of the same probability. However, these very local discontinuous mappings are of very high complexity, since this complexity is required in order to reduce the modeling error for discontinuous functions. One can also consider replacing larger sub-domains with other sub-domains such that the distribution is preserved. This could be possible, for example, if the distribution within the sub-domains is almost uniform (unlikely), or if it is estimated inaccurately due to the limitations of the training set.

Thm. 5 employs the following weak assumption. In Lem. 35 in the Appendix we prove that this assumption holds for the case of a continuous risk if the discriminators have bounded weights.

**Assumption 1** Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). For every \( m > 0 \) and \( n > 0 \), the function

\[
\text{disc}_m(f_{W_n,...,W_1} \circ D_1, D_2)
\] (18)

is continuous as a function of the weights of \( W_n, ..., W_1 \). Here, \( f_{W_n,...,W_1} = (\sigma \circ W_n) \circ ... \circ (\sigma \circ W_1) \).
Theorem 5 (Counting semantic mappings) Let $N = \text{SCM}[\sigma]$ be a NN-SCM with $\sigma$ that is a Leaky ReLU with parameter $a > 0$ and assume Assumption 1. Let $\epsilon_0, \epsilon_1$ and $\epsilon_2 < \epsilon_1 - 2\epsilon_0$ are three positive constants and $A = (X_A, D_A)$ and $B = (X_B, D_B)$ are two domains. Assume that $m \geq k + 2C_{A,B}^{(a_0)'} + 5$. Then,

$$\text{Covering} \left( H_{\epsilon_0}(D_A, D_B; m), \frac{D_A}{k, \epsilon_0} \right) \leq \lim_{\epsilon \to 0} \sqrt{\text{Covering} \left( \text{DPM}_2(\epsilon, \epsilon'; k, 2C_{A,B}^{(a_0)'} + 2), \frac{D_B}{m, \epsilon_2} \right)}$$

(19)

An interesting variant of this theorem follows from trying to count the number of mappings in $H_{\epsilon_0}^+(D_A, D_B; m) := H_{\epsilon_0}(D_A, D_B; m) \cap H_{\epsilon_0}(D_B, D_A; m))^{-1}$, i.e., the set of semantic mappings between $A$ and $B$ such that their inverse is semantic between $B$ and $A$. This is the case of the circularity constraint, which requires both sides to be semantic. In this case, we obtain the same bound with $\min\{2C_{A,B}^{(a_0)'} + 2, 2C_{B,A}^{(a_0)'} + 2\}$. Therefore, the number of two-sided semantic mappings is smaller.

To see an example where a function is semantic while its inverse is not, consider the network representation of a linear function $W = (\sigma \circ -\text{Id}) \circ (\sigma \circ -W/a)$. The inverse function $W^{-1}$ is of the same complexity (2). However, a small perturbation of the network representation, which is still semantic, has an inverse complexity of 4, while the original inverse function $W^{-1}$ is the semantic function and has a complexity of 2. Similar situations of higher complexities can be constructed, for example, by having linear functions as the top two layers of a neural network.

4.2 Shared Semantic Distributions

The recovery of an analog in domain $B$ for a sample in domain $A$, naturally takes place as a two step process: first recovering the properties of the source sample that can be transferred between the domains, and then generating a sample in $B$ that has these properties. Therefore, when discussing analogies, there are elementary compositions that play a major role. For example, drawing semantic analogies (i.e., mapping semantically) between domain $A$ and domain $B$ through domain $Z$ is naturally given by a composition of a mapping from $A$ to $Z$ and a mapping from $Z$ to $B$.

Definition 13 (Z-mappings and shared semantic distributions) Let $A = (X_A, D_A)$ and $B = (X_B, D_B)$ be two domains and $D_Z$ a distribution.

- The set of $Z$-mappings associated with $D_Z$ is denoted by:

$$Z(D_A, D_Z, D_B; m, \epsilon_0, \epsilon_1) := H_{\epsilon_0}(D_Z, D_B; m) \circ H_{\epsilon_0}(D_A, D_Z; m)$$

If $\epsilon_0 = \epsilon_1$, we write $Z(D_A, D_Z, D_B; m, \epsilon_0)$ for short.

- $D_Z$ is a $(m, \epsilon_0, \epsilon_1, \epsilon_2)$-shared semantic distribution between $A$ and $B$ if for all $y_B \in H_{\epsilon_0}(D_Z, D_B; m)$ and $y_A^1 \in H_{\epsilon_0}(D_A, D_Z; m)$ we have: $y_B \not\approx y_A^1$ and,

$$Z(D_A, D_Z, D_B; m, \epsilon_0, \epsilon_1) \cap H_{\epsilon_2}(D_A, D_B; m) \neq \emptyset$$

(21)

If $\epsilon_0 = \epsilon_1$ and $\epsilon_2 = 2\epsilon_0$ we write $(m, \epsilon_0)$ for short.

In general, there are many shared semantic distributions $D_Z$ between $A$ and $B$. For example, $D_A$ itself is a shared semantic distribution. In addition, under mild assumptions (see Thm. 40), for every semantic function $y \in H_{\epsilon_0}(D_A, D_B; m)$ and a minimal decomposition $y = y_n \circ \cdots \circ y_1$, for every $i \leq n$, $y_{i+1} \circ D_A$ is a shared semantic distribution between $A$ and $B$.

For simplicity, when considering a function $h = g \circ f \in Z(D_A, D_Z, D_B; m, \epsilon_0, \epsilon_1)$ such that $g \in H_{\epsilon_1}(D_Z, D_B; m)$ and $f \in H_{\epsilon_0}(D_A, D_Z; m)$ we will simply write $h = g \circ f \in Z(D_A, D_Z, D_B; m, \epsilon_0, \epsilon_1)$. A semantic mapping between $A$ and $B$ passes through a sequence of shared semantic distributions $D_{Z_1}, \ldots, D_{Z_n}$. One may expect that there exists a semantic mapping in the other direction, from $B$ to $A$, that passes through the same sequence. In the proof of the following theorem, it is shown that for the typical case of $C_{B,A}^{(a_0)'+\epsilon_1} = C_{A,B}^{(a_0)'+\epsilon_1} + 2$ (see Thm. 2), there is a semantic mapping that passes through $(-\sigma^{-1}) \circ D_{Z_n}, \ldots, (-\sigma^{-1}) \circ D_{Z_1}$.

The case $C_{B,A}^{(a_0)'+\epsilon_1} = C_{A,B}^{(a_0)'+\epsilon_1} - 2$ is similar and we do not cover it in this paper.
The limitations of unsupervised based learning that are due to symmetry, are also a part of our model. For example, the mapping of cars in one pose to cars in the mirrored pose that sometimes happens in (Kim et al., 2017), is similar in nature to the mapping of $x$ to $1 - x$ in the simple example given in Sec. 2.1. Such
symmetries occur when we can divide $y_{AB}$ into two functions $y_{AB} = y_2 \circ y_1$ such that a function $\pi$ in the invariant set is a DPM of $y_1 \circ D_A$ and, therefore, $D_B \approx y_2 \circ \pi \circ y_1$.

We base our work on the assumption of identifiability, which constitutes an open question for most activation functions. We hope that there would be a renewed interest in this question, which has been open for decades for networks with more than a single hidden layer and is unexplored for modern activation functions.

The stratified complexity model (SCM) is related to structural risk minimization by Vapnik and Chervonenkis (1971), which employs a hierarchy of nested subsets of hypothesis classes in order of increasing complexity. SCMs are specific to a hypothesis class of a certain recursive form, and the complexity classes are not nested.

While we focused on unsupervised learning, the emergence of semantics from learning with a restricted capacity is widely applicable, e.g., to autoencoders, transfer learning, semi-supervised learning and elsewhere.

As an extreme example, Sutskever et al. (2015) present empirical evidence that a semantic mapper can be learned, even from very few examples, if the network trained is kept small.

We point to a key difference between supervised learning and unsupervised learning. While in the former, deeper networks, which can learn even random labels, work well (Zhang et al., 2017), unsupervised learning requires a careful control of the network capacity.

6. Conclusion

The recent success in mapping between two domains in an unsupervised way and without any existing knowledge, other than network hyperparameters is nothing less than extraordinary and has far reaching consequences. As far as we know, nothing in the existing machine learning or cognitive science literature suggests that this would be possible.

We provide the necessary machinery for understanding such phenomena by presenting a framework for measuring the complexity of compositions of functions and by providing a concrete definition of semantics. Using the new machinery, we explain how, simply by training networks that are not too complex, the semantic mapping stands out from all other alternative mappings.

There are a few results that require further exploration. As one curious example, a surprising result is that all invertible functions can be divided into three classes, depending on the complexity of the inverse, which can be lower by 2, larger by 2, or the same. This can be an artifact of measuring complexity with Leaky ReLU networks, or something that has profound implications.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant ERC CoG 725974). The authors would like to thank Sagie Benaim for helping with the experiments and Moustapha Cisse, Léon Bottou and Ofir Yakovian for insightful discussions.

References

Xi Chen, Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In NIPS. 2016.

E.D. Sontag F. Albertini and V. Maillot. Uniqueness of weights for neural networks. In R. Mammone, editor, Artificial Neural Networks for Speech and Vision, 1993.

Charles Fefferman and Scott Markel. Recovering a feed-forward net from its output. In NIPS, 1993.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. J. Mach. Learn. Res., 17(1):2096–2030, 2016.
Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, pages 2672–2680, 2014.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. arXiv preprint arXiv:1703.05192, 2017.

Vera Kurková and Paul C. Kainen. Comparing fixed and variable-width gaussian networks. Neural Networks, 2014.

Andrew L. Maas, Awni Y Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

Héctor J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. Neural Networks, 1992.

Ilya Sutskever, Rafał Józefowicz, Karol Gregor, Danilo Jimenez Rezende, Timothy P. Lillicrap, and Oriol Vinyals. Towards principled unsupervised learning. arXiv preprint arXiv:1511.06440, 2015.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In International Conference on Learning Representations (ICLR), 2017.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probab. and its Applications, 16(2):264–280, 1971.

Robert C. Williamson and Uwe Helmke. Existence and uniqueness results for neural network approximations. IEEE Trans. Neural Networks, 6(1):2–13, 1995.

Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. arXiv preprint arXiv:1611.00179, 2016.

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. arXiv preprint arXiv:1704.02510, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In ICLR, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. arXiv preprint arXiv:1703.10593, 2017.
Appendix A. Summary of Notation

Tab. 1 lists the symbols used in our work.

| Symbol | Explanation |
|--------|-------------|
| $\mathcal{X}$ | A feature space |
| $A, B$ | Two domains; Specified by $(\mathcal{X}_A, D_A)$ and $(\mathcal{X}_B, D_B)$ (resp.) |
| $\mathcal{X}_A, \mathcal{X}_B$ | The sample spaces of $A$ and $B$ (resp.) |
| $D_A, D_B$ | Distributions over $\mathcal{X}_A$ and $\mathcal{X}_B$ (resp.) |
| $y_A, y_B$ | Functions from the feature space to the domains, $y_A : \mathcal{X} \rightarrow \mathcal{X}_A$ and $y_B : \mathcal{X} \rightarrow \mathcal{X}_B$ |
| $\mathcal{Z}$ | A distribution over a feature space $\mathcal{X}$ |
| $y_{AB}, y_{BA}$ | $y_{AB} = y_B \circ y_A^{-1}$ and $y_{BA} = y_A \circ y_B^{-1}$ |
| $\ell$ | Loss function $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ |
| $R_D[f_1, f_2]$ | The risk function $R_D[f_1, f_2] = \mathbb{E}_{x \sim D_\ell}\ell(f_1(x), f_2(x))$ where $\ell$ is a loss function and $D$ is a distribution |
| $\text{disc}_D(D_1, D_2)$ | The discrepancy between two distributions $D_1$ and $D_2$, i.e., $\text{disc}_D(D_1, D_2) = \sup_{c_1, c_2 \in D} |R_{D_1}[c_1, c_2] - R_{D_2}[c_1, c_2]|$ |
| $\sigma$ | A non-linear element-wise activation function |
| $\mathcal{C}$ | A class of functions; in most cases $\mathcal{C} = \{ \sigma \circ \mathcal{W} \mid \mathcal{W} \in \mathbb{R}^{M \times M}$ is an invertible linear transformation $\}$ |
| $\mathcal{N} = \text{SCM}[\mathcal{C}]$ | A SCM specified by a class of functions $\mathcal{C}$ (see Def. 1) |
| $\mathcal{N} = \text{SCM}[\sigma]$ | A NN-SCM specified by the activation function $\sigma$ (see Def. 4) |
| $\mathcal{C}(p)$ | The complexity of a function $p$ (see Eqs. 5, 6) |
| $\mathcal{C}(p)[q]$ | The conditional complexity between two functions $p$ and $q$ (see Def. 3) |
| $p \nless p, p \ngtr q$ | Fused and unfused functions, respectively (see Def. 2) |
| $\text{Invariant}(\mathcal{N})$ | The invariant set of $\mathcal{N}$ (see Def. 6) |
| $\pi$ | An invariant function (see Def. 6) |
| $I = I(\mathcal{N})$ | The ICP factor of $\mathcal{N}$ (see Def. 9) |
| $D_m$ | $D_m = \{ u | C(u) \leq m \}$ |
| $\text{disc}_m$ | $\text{disc}_m := \text{disc}_{D_m}$ |
| $C_{A,B}^m$ | The $(m, \epsilon_0)$-semantic complexity between $A$ and $B$ (see Def. 5) |
| $C_{A,B}$ | $C_{A,B} = \max_{m \geq 1} C_{A,B}^m$ |
| $E_{A,B}$ | The minimal integer $m$ such that $E_{A,B}^m = C_{A,B}^m$ |
| $H_{\epsilon_0}(D_A, D_B; m, k)$ | The set of functions that are $\epsilon_0$-close to semantic with complexity $\leq k$ under discriminators of complexity $\leq m$ (see Def. 5) |
| $H_{\epsilon_0}(D_A, D_B; m)$ | The set of $(m, \epsilon_0)$-semantic functions (see Def. 5) |
| $\mathcal{Z}(D_A, D_Z, D_B; m, \epsilon_0, \epsilon_1)$ | The set of $Z$-mappings that pass through $D_Z$ (see Def. 13) |
| $S_1 \circ S_2$ | A composition of sets, $S_1 \circ S_2 = \{ s_1 \circ s_2 | s_1 \in S_1 \text{ and } s_2 \in S_2 \}$ |
| $F$ | A function between decompositions (see Eq. 173) |
| $F^+$ | A function between decompositions (see Eq. 60) |
| $D_1 \sim_{D_2}$ | $D_1$ and $D_2$ are $(m, \epsilon)$-close (see Def. 11) |
| $f \nsim g$, $f \nsim g$ | $f$ and $g$ are $(D, m, \epsilon)$-close (see Def. 11) |
| $\text{Covering}(\mathcal{U}, \sim_{\mathcal{U}})$ | The covering number of $\mathcal{S}$ with respect to relation $\sim_{\mathcal{U}}$ on $\mathcal{U}$ (see Def. 12) |
| $x :\leftarrow x$ | $x$ is assigned to $X$ |
| $f_{W_n} \ldots W_1$ | $f_{W_n} \ldots W_1 = (\sigma \circ W_n) \circ \ldots \circ (\sigma \circ W_1)$ (see Assumption 1) |
Appendix B. Empirical Validation of Prediction 1

In the literature (Zhu et al., 2017; Kim et al., 2017), learning a mapping $h : X_A \to X_B$, based only on the GAN constraint on $B$, is presented as a failing baseline. In Yi et al., 2017), among many non-semantic mappings obtained by the GAN baseline, one can find images of GANs that are successful. However, this goes unnoticed.

In order to validate the prediction that a purely GAN based solution is viable, we conducted a series of experiments using the DiscoGAN architecture. We consider image domains $A$ and $B$, where $X_A = X_B = \mathbb{R}^{3\times64\times64}$.

In DiscoGAN, the generator is build of: (i) an encoder consisting of convolutional layers with $4 \times 4$ filters followed by Leaky ReLU activation units and (ii) a decoder consisting of deconvolutional layers with $4 \times 4$ filters followed by a ReLU activation units. Sigmoid is used for the output layer. Between 4 to 5 convolutional/deconvolutional layers are used, depending on the domains used in $A$ and $B$ (we match the published code architecture per dataset). The discriminator is similar to the encoder, but has an additional convolutional layer as the first layer and a sigmoid output unit.

The first set of experiments considers the CelebA face dataset. Transformations are learned between the subset of images labeled as “man” and those labeled as “woman”, as well as from blond to black hair and glasses to no eyewear. The results are shown in Fig. 3, 4, and 5, (resp.). It is evident that the output image is highly related to the input images.

In the case of mapping handbag to shoes, as seen in Fig. 6, the GAN does not provide a meaningful solution. However, in the case of edges to shoes and vice versa (Fig. 7), the GAN solution is successful.

Appendix C. Empirical Validation of Prediction 2

We predict that the selection of the right number of layers is crucial in unsupervised learning. Using fewer layers than needed will not support the modeling of the semantic transformation. In contrast, adding superfluous layers would mean that more and more alternative mappings obscure the semantic transformation.

In (Kim et al., 2017), 8 layers are sometimes employed while at other times 10 layers are used (counting both convolution and deconvolution). In our experiment we vary the number of layers and inspect the influence on the results.

These experiments were done on the CelebA gender conversion task, where eight layers are employed in the experiments of (Kim et al., 2017). Using the public implementation and adding and removing layers, we obtain the results in Fig. 8,9,10,11. Note that since the encoder and the decoder parts of the learned network are symmetrical, the number of layers is always even. As can be seen, changing the number of layers has a dramatic effect on the results and the best results are obtained at eight layers. The results degrade quickly as one deviates from the optimal value. Using fewer layers, the GAN fails to produce images of the desired class. Adding layers, the semantic alignment is lost, just as expected.
Appendix D. Assumptions

Assumption 1 Let $\mathcal{N} = \text{SCM}[\sigma]$ with $\sigma$ that is Leaky ReLU with parameter $a > 0$. For every $m > 0$ and $n > 0$, the function

$$\text{disc}_m(f_{W_n,\ldots,W_1} \circ D_1, D_2)$$

is continuous as a function of the weights of $W_n, \ldots, W_1$. Here, $f_{W_n,\ldots,W_1} = (\sigma \circ W_n) \circ \cdots \circ (\sigma \circ W_1)$.

Assumption 2 Let $\mathcal{N} = \text{SCM}[\sigma]$ with $\sigma$ that is Leaky ReLU with parameter $a > 0$. For all $m > 0$, the function

$$R_D[f_{V_m,\ldots,V_1}, f_{W_m,\ldots,W_1}]$$

is continuous as a function of $V_m, \ldots, V_1, W_m, \ldots, W_1$.

Appendix E. Lemmas

Lemma 1 Let $D_1$ and $D_2$ be two classes of functions and $D_1, D_2$ two distributions. Assume that $D_1 \circ \{p\} \subset D_2$ then,

$$\text{disc}_{D_1}(p \circ D_1, p \circ D_2) \leq \text{disc}_{D_2}(D_1, D_2)$$

In particular, if $m \geq k + C(p)$ then,

$$\text{disc}_k(p \circ D_1, p \circ D_2) \leq \text{disc}_m(D_1, D_2)$$

Proof By the definition of discrepancy:

$$\text{disc}_{D_1}(p \circ D_1, p \circ D_2) = \sup_{c_1, c_2 \in D_1} \left| R_{p \circ D_1}[c_1, c_2] - R_{p \circ D_2}[c_1, c_2] \right|$$

$$= \sup_{c_1, c_2 \in D_1} \left| R_{D_1}[c_1 \circ p, c_2 \circ p] - R_{D_2}[c_1 \circ p, c_2 \circ p] \right|$$

Since $D_1 \circ \{p\} \subset D_2$ we have:

$$\text{disc}_{D_1}(p \circ D_1, p \circ D_2) = \sup_{c_1, c_2 \in D_1} \left| R_{D_1}[c_1 \circ p, c_2 \circ p] - R_{D_2}[c_1 \circ p, c_2 \circ p] \right|$$

$$\leq \sup_{u_1, u_2 \in D_2} \left| R_{D_1}[u_1, u_2] - R_{D_2}[u_1, u_2] \right| = \text{disc}_{D_2}(D_1, D_2)$$

The second inequality is a special case for $D_1 = D_k$ and $D_2 = D_m$.

Lemma 2 Let $A = (\mathcal{X}_1, D_1)$ and $B = (\mathcal{X}_2, D_2)$ be two domains and $D_Z$ a distribution.

1. Assume that $m \geq k + C(p)$. Then,

$$\text{disc}_k(p \circ D_1, D_3) \leq \text{disc}_m(D_1, D_2) + \text{disc}_k(p \circ D_2, D_3)$$

2. Let $y_1, y_2$ and $y = y_2 \circ y_1^{-1}$ be three functions and $m \geq k + C(y_2)$. Then,

$$\text{disc}_k(y \circ D_1, D_2) \leq \text{disc}_m(D_Z, y_1^{-1} \circ D_1) + \text{disc}_k(y_2 \circ D_Z, D_2)$$

3. Let $h$ be any function and $m \geq k + C(h^{-1})$. Then,

$$\text{disc}_k(D_1, h^{-1} \circ D_2) \leq \text{disc}_m(h \circ D_1, D_2)$$
Proof
1. Follows from Lem. 1, since \( D_k \circ \{p\} \subset D_m \), we have:
\[
\text{disc}_k(p \circ D_1, p \circ D_2) \leq \text{disc}_m(D_1, D_2)
\]
(30)
Therefore, by the triangle inequality,
\[
\text{disc}_k(p \circ D_1, D_3) \leq \text{disc}_k(p \circ D_1, p \circ D_2) + \text{disc}_k(p \circ D_2, D_3)
\]
\[
\leq \text{disc}_m(D_1, D_2) + \text{disc}_m(p \circ D_2, D_3)
\]
(31)
2. We use Lem. 1 with \( p \leftarrow y_2, D_1 \leftarrow D_k, \) and \( D_2 \leftarrow D_m \) and \( D_k \circ \{y_2\} \subset D_2 \):
\[
\text{disc}_k(y_2 \circ D_2, y \circ D_1) = \text{disc}_k(y_2 \circ D_2, y \circ y_1^{-1} \circ D_1) \leq \text{disc}_m(D_2, y_1^{-1} \circ D_1)
\]
(32)
Therefore, by the triangle inequality,
\[
\text{disc}_k(y_2 \circ D_2, y \circ D_1) \leq \text{disc}_k(y_2 \circ D_2, D_3) + \text{disc}_k(D_2, y_2 \circ D_2)
\]
\[
\leq \text{disc}_k(y_2 \circ D_2, D_3) + \text{disc}_m(D_2, y_1^{-1} \circ D_1)
\]
(33)
3. Follows immediately from Lem. 1 for \( p \leftarrow h^{-1} \) and \( D_k \circ \{h^{-1}\} \subset D_m \).

Lemma 3 Let \( N = SCM[C] \). Assume that \( D_1 \sim_{m,\epsilon_1} D_2 \) and \( D_2 \sim_{m,\epsilon_2} D_3 \) then \( D_1 \sim_{m,\epsilon_1+\epsilon_2} D_3 \).

Proof We consider that
\[
D_1 \sim_{m,\epsilon_1} D_2 \implies \text{disc}_m(D_1, D_2) \leq \epsilon_1
\]
(34)
and,
\[
D_2 \sim_{m,\epsilon_2} D_3 \implies \text{disc}_m(D_2, D_3) \leq \epsilon_2
\]
(35)
Therefore, by the triangle inequality,
\[
\text{disc}_m(D_1, D_3) \leq \text{disc}_m(D_1, D_2) + \text{disc}_m(D_2, D_3) \leq \epsilon_1 + \epsilon_2
\]
(36)

Lemma 4 Let \( N = SCM[C] \). In addition, let \( u, v \) be any two functions. Then,
\[
\max\{C(u) - C(v^{-1}), C(v) - C(u^{-1})\} \leq C(u \circ v) \leq C(u) + C(v)
\]
(37)

Proof We begin by proving the upper bound. We denote \( C(u) = n \) and \( C(v) = m \). Let \( u = u_{n+1:1} \) and \( v = v_{m+1:1} \) be minimal decompositions of \( u \) and \( v \) (resp.). Therefore, we can represent, \( u \circ v = u_{n+1:1} \circ v_{m+1:1} \). In particular, \( C(u \circ v) \leq n + m = C(u) + C(v) \).

The lower bound follows immediately from the upper bound:
\[
C(u) = C(u \circ v \circ v^{-1}) \leq C(u \circ v) + C(v^{-1}) \implies C(u) - C(v^{-1}) \leq C(u \circ v)
\]
(38)

By similar considerations, \( C(v) - C(u^{-1}) \leq C(u \circ v) \).

Lemma 5 Let \( N = SCM[C] \). In addition, let \( u_1, u_2, u_3 \in N \) be three functions such that: \( C(u_1 \circ u_2 \circ u_3) = C(u_1) + C(u_2) + C(u_3) \). Then, \( u_1 \not\preceq u_2 \) and \( u_2 \not\preceq u_3 \).
Proof By Lem. 4, \( C(u_1 \circ u_2 \circ u_3) \leq C(u_1 \circ u_2) + C(u_3) \). In addition, by \( C(u_1 \circ u_2 \circ u_3) = C(u_1) + C(u_2) + C(u_3) \), we have: \( C(u_1) + C(u_2) \leq C(u_1 \circ u_2) \). Again by Lem. 4, \( C(u_1 \circ u_2) \leq C(u_1) + C(u_2) \) and conclude that \( C(u_1 \circ u_2) = C(u_1) + C(u_2) \). The second equation follows by similar arguments. ■

**Lemma 6** Invariant(\( \mathcal{N} \)) is closed under inverse and composition, i.e.,
\[
\pi \in \text{Invariant}(\mathcal{N}) \iff \pi^{-1} \in \text{Invariant}(\mathcal{N})
\]
(39)

And,
\[
\pi_1, \pi_2 \in \text{Invariant}(\mathcal{N}) \implies \pi_1 \cdot \pi_2 \in \text{Invariant}(\mathcal{N})
\]
(40)

**Proof Inverse:** Let \( \pi \in \text{Invariant}(\mathcal{N}) \). Then, by definition, \( \pi \) is an invertible linear mapping and \( \pi \circ \sigma = \sigma \circ \pi \). In particular, \( \pi^{-1} \) is also an invertible linear mapping and \( \pi^{-1} \circ \sigma = \sigma \circ \pi^{-1} \). Thus, \( \pi^{-1} \in \text{Invariant}(\mathcal{N}) \).

**Composition:** Let \( \pi_1, \pi_2 \in \text{Invariant}(\mathcal{N}) \). Then, \( \pi_1 \) is an invertible linear mapping and \( \pi_1 \circ \sigma = \sigma \circ \pi_1 \) for \( i = 1, 2 \). In particular, \( \pi_1 \circ \pi_2 \) is also an invertible linear mapping and \( \pi_1 \circ \pi_2 \circ \sigma = \pi_1 \circ \sigma \circ \pi_2 = \sigma \circ \pi_1 \circ \pi_2 \). Thus, \( \pi_1 \circ \pi_2 \in \text{Invariant}(\mathcal{N}) \).

Recall the notation introduced in Eq. 11.

**Lemma 7** Let \( \mathcal{N} = \text{SCM}[\sigma] \) obeying identifiability. If \( p = p_{n+1:1} = q_{n+1:1} \) are two minimal decompositions of \( p \) then:
\[
\forall i \in [n]: p_{i+1:1} \circ q_{i+1:1}^{-1} \in \text{Invariant}(\mathcal{N}) \text{ and } p_{n+1:1} \circ q_{n+1:1}^{-1} \in \text{Invariant}(\mathcal{N})
\]
(41)

**Proof** First, if \( i = n \) then \( p_{i+1:1} \circ q_{i+1:1}^{-1} = \text{Id} \in \text{Invariant}(\mathcal{N}) \).

Otherwise, by minimal identifiability,
\[
q_1 = \pi_1 \circ p_1, \forall i = 2, ..., n - 1: q_i = \pi_i \circ p_i \circ \pi_{i-1}^{-1} \text{ and } q_n = p_n \circ \pi_{n-1}^{-1}
\]
(42)

In addition,
\[
p_{i+1:1} = p_i \circ p_{i-1} \circ ... \circ p_1
\]
\[
q_{i+1:1} = (\pi_i \circ p_i \circ \pi_{i-1}^{-1}) \circ (\pi_{i-1} \circ p_{i-2} \circ \pi_{i-2}^{-1}) \circ ... \circ (\pi_1 \circ p_1)
\]
(43)

Therefore, \( q_{i+1:1} = \pi_i \circ p_{i+1:1} \) and \( p_{n+1:1} \circ q_{n+1:1}^{-1} = \pi_{n-1}^{-1} \in \text{Invariant}(\mathcal{N}) \). By similar considerations, \( p_{n+1:1} \circ q_{n+1:1}^{-1} \in \text{Invariant}(\mathcal{N}) \).

**Lemma 8** Let \( \mathcal{N} = \text{SCM}[\sigma] \) obeying identifiability and \( \sigma \) is a non-linear element-wise activation function. Let \( p_1, p_2 \) and \( p_3 \) be three functions such that \( C(p_i) = 1 \) for all \( i = 1, 2, 3 \). Then, \( p_2 \circ p_1 \neq p_3 \).

**Proof** There are invertible linear mappings, \( W_i \), such that \( p_i = \sigma \circ W_i \) for all \( i = 1, 2, 3 \). Therefore, if:
\[
p_2 \circ p_1 = p_3
\]
(44)

then,
\[
(\sigma \circ W_2) \circ (\sigma \circ W_1) = (\sigma \circ W_3)
\]
(45)

which is equivalent to:
\[
\sigma = W_2^{-1} \circ W_3 \circ W_1^{-1}
\]
(46)
in contradiction to the assumption that \( \sigma \) is a non-linear function.
Lemma 9  Let \( N = \text{SCM}[\sigma] \), \( f \neq \text{Id} \) is any function such that \( C(f) > 0 \) and \( \pi \in \text{Invariant}(N) \). Then, \( C(\pi \circ f) \leq C(f) \).

Proof  Let \( f = f_{n+1:1} \) be a minimal decomposition of \( f \) and \( f_i = \sigma \circ W_i \) for \( i \in [n] \) and \( W_i \) are invertible linear mappings. Since \( \pi \circ \sigma = \sigma \circ \pi \), we have:

\[
\pi \circ f = \sigma \circ (\pi \circ W_n) \circ \sigma \circ (\pi \circ W_{n-1}) \circ \ldots \circ \sigma \circ W_1
\]

This is a decomposition of length \( n \). Therefore, \( C(\pi \circ f) \leq n = C(f) \).

Lemma 10  Let \( N = \text{SCM}[\sigma] \), \( f \neq \text{Id} \) is any function and \( W \) is an invertible linear mapping. Then, \( C(f \circ W) \leq C(f) \).

Proof  Let \( f = f_{n+1:1} \) be a minimal decomposition of \( f \) and \( f_i = \sigma \circ W_i \) for \( i \in [n] \) and \( W_i \) are invertible linear mappings. We have:

\[
f \circ W = (\sigma \circ W_n) \circ \ldots \circ (\sigma \circ W_2) \circ (\sigma \circ W_1 \cdot W)
\]

This is a decomposition of length \( n \). Therefore, \( C(f \circ W) \leq n = C(f) \).
Appendix F. Proofs for the Thms. 1 and 4

F.1 Properties of inverses

Lemma 11 Let $\mathcal{N} = \text{SCM}[\sigma]$, where $\sigma$ is the Leaky ReLU activation function, with parameter $a > 0$. Then, $I(\mathcal{N}) \leq 3$.

**Proof** We show that the inverse of every function $u$ of complexity 1 can be represented as a decomposition $u_1 \circ u_2 \circ u_3$ for $u_1, u_2, u_3 \in C$. It follows from the following simple identities,

$$
\sigma^{-1} = -\text{Id} \circ \sigma \circ -\text{Id}/a
$$

(49)

Therefore,

$$
W^{-1} \circ \sigma^{-1} = -W^{-1} \circ \sigma \circ -\text{Id}/a
$$

(50)

And,

$$
\text{Id} = (\sigma \circ -\text{Id}) \circ (\sigma \circ -\text{Id}/a)
$$

(51)

We obtain,

$$
W^{-1} \circ \sigma^{-1} = (\sigma \circ -\text{Id}) \circ (\sigma \circ W^{-1}/a) \circ (\sigma \circ -\text{Id}/a)
$$

(52)

Therefore, $C(u^{-1}) \leq 3$. Finally, let $p \in \mathcal{N}$ be a function of complexity $n \geq 1$ with minimal decomposition $p = p_{n+1:1}$.

**Theorem 1** Let $\mathcal{N} = \text{SCM}[\sigma]$ be a NN-SCM with $\sigma$ that is the Leaky ReLU with parameter $a > 0$. Then, for any $u \in \mathcal{N}$, $|C(u^{-1}) - C(u)|$ is either 0 or 2.

**Proof**

**Part 1:** In this part, we prove by induction on $C(u) = n$ that:

$$
u^{-1} = (\sigma \circ -\text{Id}) \circ (\sigma \circ W^{-1} \circ W_{1}^{-1}) \circ \ldots \circ (\sigma \circ W^{-1} \circ W_{n}^{-1} \circ W_{1}) \circ (\sigma \circ -\text{Id}/a)
$$

(53)

Where $u = (\sigma \circ W_{n}) \circ \ldots \circ (\sigma \circ W_{1})$ is a minimal decomposition of $u$.

**Case $C(u) = 0$:** Then $u = \text{Id}$, $u^{-1} = (\sigma \circ \text{Id}) \circ (\sigma \circ -\text{Id}/a) = \text{Id}$ and $C(u^{-1}) = 0$.

**Case $C(u) = 1$:** Follows immediately from Lem. 11 and Eq. 52.

**Induction hypothesis:** Assume that:

$$
u^{-1} = (\sigma \circ -\text{Id}) \circ (\sigma \circ W_{1}^{-1} \circ W_{1}^{-1}) \circ \ldots \circ (\sigma \circ W_{n}^{-1} \circ W_{1}) \circ (\sigma \circ -\text{Id}/a)
$$

(54)

Where $u = (\sigma \circ W_{n}) \circ \ldots \circ (\sigma \circ W_{1})$ is a minimal decomposition of $u$.

**Case $C(u) = n + 1$:** Let $u = u_{n+1:2} = (\sigma \circ W_{n+1}) \circ \ldots \circ (\sigma \circ W_{1})$ be a minimal decomposition of $u$. We denote $v = u_{n+1:1}$. By the induction hypothesis,

$$v^{-1} = (\sigma \circ -\text{Id}) \circ (\sigma \circ W_{1}^{-1} \circ W_{1}^{-1}) \circ \ldots \circ (\sigma \circ W_{n}^{-1} \circ W_{1}) \circ (\sigma \circ -\text{Id}/a)
$$

(55)

By Eq. 50 we can represent: $u_{n+1}^{-1} = -W_{n+1}^{-1} \circ \sigma \circ -\text{Id}/a$. We consider that:

$$
u^{-1} = v^{-1} \circ u_{n+1}^{-1} = (\sigma \circ -\text{Id}) \circ (\sigma \circ W_{1}^{-1} \circ W_{1}^{-1}) \circ \ldots \circ (\sigma \circ W_{n}^{-1} \circ W_{1}) \circ (\sigma \circ -\text{Id}/a) \circ (W_{n+1}^{-1} \circ \sigma \circ -\text{Id}/a)
$$

$$= (\sigma \circ -\text{Id}) \circ (\sigma \circ W_{1}^{-1} \circ W_{1}^{-1}) \circ \ldots \circ (\sigma \circ W_{n}^{-1} \circ W_{1}) \circ (\sigma \circ W_{n+1}^{-1} \circ W_{1}) \circ (\sigma \circ -\text{Id}/a)
$$

(56)

In particular, $C(u^{-1}) \leq n + 3 = C(u) + 2$ and $C(u) = C((u^{-1})^{-1}) \leq C(u^{-1}) + 2$. Therefore,

$$|C(u^{-1}) - C(u)| \leq 2
$$

(57)
Part 2: In this part, we show that \[ |C(u^{-1}) - C(u)| \neq 1 \] (58)
Assume by contradiction that \( C(u^{-1}) = C(u) + 1 \). Therefore, there is a minimal decomposition \( u^{-1} = p_{n+1} \). By Part 1, there is a decomposition \( u^{-1} = q_{n+1} \). By identifiability,
\[
\exists j_1 < \ldots < j_{n+2} = n + 3, \pi_1, \ldots, \pi_n \in \text{Invariant}(N) : \qquad q_{j_2:j_1} = \pi_1 \circ p_1, \forall i = 2, \ldots, n : q_{j_{k+1}:j_k} = \pi_i \circ p_i \circ \pi_{i-1}^{-1} \text{ and } q_{j_{n+2}:j_{n+1}} = p_{n+1} \circ \pi_n^{-1} \quad (59)
\]
Since \( j_{n+2} = n + 3 \) and \( j_i < j_{i+1} \), there is \( k \in [n+1] \) such that \( j_{k+1} - j_k = 2 \) and for every \( i \in [n+1] \setminus \{k\} \) we have \( j_{i+1} - j_i = 1 \). In particular, \( C(q_{j_{k+2}:j_k}) = 1 \) in contradiction to Lem. 8. We conclude that \( C(u^{-1}) \neq C(u) + 1 \) and \( C(u) = C((u^{-1})^{-1}) \neq C(u^{-1}) + 1 \).

\textbf{Lemma 12} Let \( N = \text{SCM}[\sigma] \) where \( \sigma \) is the Leaky ReLU activation function, with parameter \( a > 0 \). We define a function from one decomposition to another as follows:
\[
\text{F}^+(u_{n+1:1}) = (\sigma \circ -\text{Id}) \circ (\sigma \circ W_1^{-1}/a) \circ \ldots \circ (\sigma \circ W_n^{-1}/a) \circ (\sigma \circ -\text{Id}/a) \quad (60)
\]
where \( u_i = \sigma \circ W_i \) for \( i \in [n] \). Then, for all \( j \in [n+1] \) we have:
\[
|\text{F}^+(u_{n+1:1})|_{n-j+3:1} \circ u \circ D = (-\sigma^{-1}) \circ u_{j-1} \circ D \quad (61)
\]

\textbf{Proof} In the proof of Thm. 1, we built recursively the composition:
\[
\text{F}^+(u_{n+1:1}) = (\sigma \circ -\text{Id}) \circ (\sigma \circ W_1^{-1}/a) \circ \ldots \circ (\sigma \circ W_n^{-1}/a) \circ (\sigma \circ -\text{Id}/a) \quad (62)
\]
to invert \( u_{n+1:1} \), i.e., \( \text{F}^+(u_{n+1:1}) = u_{n+1:1}^{-1} \). Similarly, for all \( j \in [n+1] \):
\[
(\sigma \circ -\text{Id}) \circ |\text{F}^+(u_{n+1:1})|_{n-j+3:1} = (\sigma \circ -\text{Id}) \circ (\sigma \circ W_j^{-1}/a) \circ \ldots \circ (\sigma \circ W_n^{-1}/a) \circ (\sigma \circ -\text{Id}/a) = u_{n+1:j}^{-1} \quad (63)
\]
In particular, for all \( j \in [n+1] \):
\[
|\text{F}^+(u_{n+1:1})|_{n-j+3:1} = (-\sigma^{-1}) \circ u_{n+1:j}^{-1} \quad (64)
\]
Therefore, for all \( j \in [n+1] \):
\[
|\text{F}^+(u_{n+1:1})|_{n-j+3:1} \circ u \circ D = (-\sigma^{-1}) \circ u_{n+1:j}^{-1} \circ u \circ D = (-\sigma^{-1}) \circ u_{j-1} \circ D \quad (65)
\]

\textbf{F.2 Properties of compositions}

\textbf{Lemma 13} Let \( N = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). Then, there are functions \( f, g, h \in N \) such that \( C(f \circ g \circ h) < C(f) + C(g) + C(h) \), \( f \not\preceq g \) and \( g \not\preceq h \).

\textbf{Proof} Let \( W \neq \text{Id} \) be any invertible \( M \times M \) real matrix. Consider the following functions:
\[
f = (\sigma \circ W), \quad g = (\sigma \circ -\text{Id}) \text{ and } h = (\sigma \circ -W^{-1}/a) \quad (66)
\]
Since both \( f \) and \( g \) are functions of complexity 1, the complexity of \( f \circ g \) is at most 2. The complexity of \( f \circ g \) is not 0, since \( f \circ g \neq \text{Id} \). In addition, by Lem. 8, \( C(f \circ g) \neq 1 \). Similarly, \( C(g \circ h) = 2 \). On the other hand, \( f \circ g \circ h = \sigma \) and \( C(\sigma) = 1 \).

21
Lemma 14 Let $\mathcal{N} = \text{SCM}[\sigma]$ be a NN-SCM obeying identifiability. Let $u \in \mathcal{N}$ be a function such that $C(u) = n > 0$ and $a \in \mathcal{N}$ be a function of complexity 1. Then, there is $0 \leq r \leq n$ such that:

$$C(u_{r+1} \circ a) = 1 \text{ and } C(u) = C(u \circ a) + r - 1$$

(67)

Proof If $u \not\triangleright a$ then we take $r = 0$. Otherwise, $u \triangleright a$. Let $u = u_{n+1:1}$ be a minimal decomposition of $u$ and let $u \circ a = v_{d+1:1}$ be a minimal decomposition of $u \circ a$. Therefore,

$$u \circ a = u_{n+1:1} \circ a = v_{d+1:1}$$

(68)

By identifiability, there is $0 \leq r \leq n$ such that:

$$u_{r+1:1} \circ a = \pi_1 \circ v_1$$

(69)

For some $\pi_1 \in \text{Invariant}(\mathcal{N})$. In addition,

$$\exists j_2 = r + 1 \leq \ldots \leq j_{d+1} = n + 1 \text{ and } \pi_2, \ldots, \pi_{d-1} \in \text{Invariant}(\mathcal{N}) :$$

$$\forall i = 2, \ldots, d-1 : u_{j_{i+1}:j_i} = \pi_i \circ v_i \circ \pi_i^{-1}$$

(70)

Therefore,

$$\forall i = 2, \ldots, d - 1 : C(u_{j_{i+1}:j_i}) = j_{i+1} - j_i$$

(71)

In particular,

$$\forall i = 2, \ldots, d : j_i = r + i - 1 \text{ and } u_i = \pi_i \circ v_i \circ \pi_i^{-1}$$

(72)

We also have, $j_{d+1} = n + 1$. Therefore,

$$C(u) = n = j_{d+1} - 1 = d + r - 1 = C(u \circ a) + r - 1$$

(73)

Lemma 15 Let $\mathcal{N} = \text{SCM}[\sigma]$ be a NN-SCM obeying identifiability. Let $u \in \mathcal{N}$ be a function such that $C(u) = n > 0$ and $a \in \mathcal{N}$ be a function of complexity 1. If $u \triangleright a$, $u \equiv b \circ d$ such that $b \not\triangleright d$ and $C(a^{-1}) = C(d)$, then we have $d \triangleright a$ or $d \circ a \in \text{Invariant}(\mathcal{N})$.

Proof Let $b = b_{n+1:1}$ and $d = d_{m+1:1}$ be minimal decompositions of $b$ and $d$ (resp.). We consider that $u = b \circ d = b_{n+1:1} \circ d_{m+1:1}$ is a minimal decomposition of $u$. We denote $u = u_{n+m+1:1}$ where $u_i = b_{i-m}$ for $m + 1 \leq i \leq n + m$ and $u_i = d_i$ for $i \leq m$. By Lem. 14 we have:

$$\exists 0 \leq r \leq n + m : C(u_{r+1:1} \circ a) = 1 \text{ and } C(u) = C(u \circ a) + r - 1$$

(74)

If $r \leq m$: then we have $C(d \circ a) = C(u_{m+1:r+1} \circ u_{r+1:1} \circ a) \leq m - r + C(u_{r+1:1} \circ a) = m - r + 1$. If $r \neq 0$, then $d \triangleright a$ as desired. On the other hand, if $r = 0$ then $C(u) = C(u \circ a) + 1$ and $u \not\triangleright a$ in contradiction. If $r \geq m$: then there is a function $v$ such that $C(v) = 1$ and:

$$v = b_{r-m+1:1} \circ d \circ a$$

(75)

Alternatively,

$$v \circ a^{-1} = b_{r-m+1:1} \circ d$$

(76)

Since $b \not\triangleright d$, by Lem. 5,

$$C(b_{r-m+1:1} \circ d) = C(b_{r-m+1:1}) + C(d) = C(d) + r - m$$

(77)
If $r - m > 1$: then $C(d) + r - m = C(v \circ a^{-1}) \leq C(a^{-1}) + 1 = C(d) + 1$ in contradiction to $r > 1$.
If $r - m = 0$: then $v = d \circ a$. Therefore, $d \circ a$ and $C(d \circ a) = 1$ (since $C(a) = 1$, $1 \leq C(d)$ and $C(d \circ a) = 1 < C(a) + C(d)$).
If $r - m = 1$: then, we have,
\[ C(v \circ a^{-1}) = C(b_1 \circ d) = C(b_1) + C(d) = 1 + C(d) = 1 + C(a^{-1}) = C(v) + C(a^{-1}) \] (78)
Alternatively, $v \not\subseteq a^{-1}$. By Lem. 7, for $p = v \circ a^{-1} = b_1 \circ d$ we have: $d \circ a \in \text{Invariant}(\mathcal{N})$. We conclude that $d \circ a$ or $d \circ a \in \text{Invariant}(\mathcal{N})$.

Lemma 16 (Composition reduction) Let $\mathcal{N} = \text{SCM}[\sigma]$ be a NN-SCM obeying identifiability. Let $u, v \in \mathcal{N}$. Then, we can represent $u = a \circ b$ and $v = b^{-1} \circ c$ such that: $a \not\subseteq c$, $b^{-1} \not\subseteq c$ and
\[ C(a) + C(b) - 1 \leq C(a \circ b) . \] (79)

Proof Let $u = u_{n+1:1}$ and $v = v_{m+1:1}$ be minimal decompositions of $u$ and $v$ (resp.). Let $u \circ v = z_{d+1:1}$ be a minimal decomposition of $u \circ v$. By identifiability we have:
\[ \exists j_i = 1 < ... < j_{d+1} = n + m + 1, \exists \pi_1, ..., \pi_{d-1} \in \text{Invariant}(\mathcal{N}) \]
\[ [u_n \circ ... \circ u_1 \circ v_m \circ ... \circ v_1]_{j_2:j_1} = \pi_1 \circ z_1 \]
\[ \forall i = 2, ..., d - 1: [u_n \circ ... \circ u_1 \circ v_m \circ ... \circ v_1]_{j_{i+1}:j_i} = \pi_i \circ z_i \circ \pi_{i-1}^{-1} \] (80)
\[ [u_n \circ ... \circ u_1 \circ v_m \circ ... \circ v_1]_{j_{d+1}:j_d} = z_d \circ \pi_{d-1}^{-1} \]
Here, $[f_k \circ ... \circ f_1]_{i:j} = f_{i-1} \circ ... \circ f_j$ for $0 \leq j < i$ and $u_{i:i} = \text{Id}$.
We consider two options. The first: there is no index $k$ such that $j_k \leq m$ and $m + 1 < j_{k+1}$. The second: there is such $k$.

Case 1: In this case, for every index $i = 2, ..., d - 1$:
\[ v_{j_{i+1}:j_i} = \pi_i \circ z_i \circ \pi_{i-1}^{-1} \text{ or } u_{j_{i+1} - m:j_i - m} = \pi_i \circ z_i \circ \pi_{i-1}^{-1} \] (81)
And for $i = 1$ and $i = d$:
\[ v_{j_2:j_1} = \pi_1 \circ z_1 \text{ and } v_{j_{d+1}:j_d} = z_d \circ \pi_{d-1}^{-1} \] (82)
If the first equation holds,
\[ 1 = C(\pi_1 \circ z_1 \circ \pi_{i-1}^{-1}) = C(v_{j_{i+1}:j_i}) = j_{i+1} - j_i \] (83)
Therefore, for any $i = 1, ..., d$: $j_{i+1} = j_i + 1$. We conclude that:
\[ C(u \circ v) = d = n + m = C(u) + C(v) \text{ and } u \not\subseteq v \] (84)
Finally, we choose $a = u$, $b = \text{Id}$ and $c = v$. This gives $a \not\subseteq b$, $b^{-1} \not\subseteq c$ and $a \not\subseteq c$.

Case 2: Let $1 \leq k \leq d$ be the index for which $j_k \leq m$ and $m + 1 < j_{k+1}$. With no loss of generality, we assume that $k \neq 1, d$. For any index $i \neq k$ such that $1 < i < d$, we have,
\[ v_{j_{i+1}:j_i} = \pi_i \circ z_i \circ \pi_{i-1}^{-1} \text{ or } u_{j_{i+1} - m:j_i - m} = \pi_i \circ z_i \circ \pi_{i-1}^{-1} \] (85)
And if $i = 1$ or $i = d$:
\[ v_{j_2:j_1} = \pi_1 \circ z_1 \text{ or } v_{j_{d+1}:j_d} = z_d \circ \pi_{d-1}^{-1} \] (86)
As in Case 1, we conclude that $j_{i+1} = j_i + 1$. Therefore, we have $j_k = k$ and we denote $j_{k+1} = m + r + 1$, i.e,
\[ [u_n \circ ... \circ u_1 \circ v_m \circ ... \circ v_1]_{j_{k+1}:j_k} = u_{r+1:1} \circ v_{m+1:k} = \pi_k \circ z_k \circ \pi_{k-1}^{-1} \] (87)

23
Finally, we choose:

\[ b = v_{m+1:k}^{-1}, \quad a = u \circ b^{-1} \quad \text{and} \quad c = v_{k:1} \]

(88)

It follows immediately that \( v = b^{-1} \circ c \) such that \( b^{-1} \not\triangleright c \). In addition, we have, \( u = a \circ b \) and \( u \circ v = a \circ c \).

We consider that:

\[ C(a \circ c) = C(z_{d+1:1}) = C(z_{d+1:k} \circ z_{k:1}) = C(z_{d+1:k}) + C(z_{k:1}) \]

(89)

By Lem. 5 we have \( C(c) = C(v_{k:1}) = k - 1 = C(z_{k:1}) \). In addition,

\[ a = u \circ b^{-1} = u_{n+1:r+1} \circ v_{m+1:k} \]

(90)

By identifiability, \( u_{r+1} \circ v_{m+1:k} = \pi_k \circ z_k \circ \pi_{k-1}^{-1} \) and \( u_{n+1:r+1} \circ \pi_k = z_{d+1:k+1} \). Therefore,

\[ a = u \circ b^{-1} = z_{d+1:k} \circ \pi_{k-1}^{-1} \]

(91)

Thus, by Lem. 10, \( C(a) \leq C(z_{d+1:k}) \) since \( z_{d+1:k} \not\in \text{Id} \) because \( k \leq d \). In particular,

\[ C(a \circ c) = C(z_{d+1:k}) + C(z_{k:1}) \geq C(a) + C(c) \]

(92)

Finally, by Lem. 4, we have: \( C(a \circ c) = C(a) + C(c) \) or \( a \not\triangleright c \).

Since \( u_{r+1} = \pi_k^{-1} \circ z_k \circ \pi_{k-1}^{-1} \circ b \) we have:

\[
C(u) = C(u_{n+1:r+1} \circ \pi_k \circ z_k \circ \pi_{k-1}^{-1} \circ b) \\
= C(u_{n+1:r+1}) + C(\pi_k \circ z_k \circ \pi_{k-1}^{-1} \circ b) \\
= C(a) - 1 + C(\pi_k \circ z_k \circ \pi_{k-1}^{-1} \circ b) \\
\geq C(a) - 1 + C(b) - C(\pi_k \circ z_k \circ \pi_{k-1}^{-1}) \\
\geq C(a) - 1 + C(b) - C(\pi_{k-1} \circ z_k \circ \pi_k^{-1}) \\
\geq C(a) - 1 + C(b) - I
\]

\[ \square \]

**Theorem 4** Let \( \mathcal{N} = \text{SCM}[\sigma] \) be an NN-SCM obeying identifiability. Then, for all \( f, g, h \in \mathcal{N} \) such that \( f \not\triangleright g \), \( g \not\triangleright h \) and \( I \leq C(g) \), we have:

\[ C(f \circ g \circ h) = C(f) + C(g) + C(h) \]

(13)

**Proof** Denote \( C(f) = i, C(g) = n, C(h) = m \). We prove the theorem by induction on \( C(f) = i \).

Case \( i = 0 \): In this case \( f = \text{Id} \) and \( C(f \circ g \circ h) = C(g \circ h) = C(g) + C(h) \) as desired.

Case \( i = 1 \): Let \( g = g_{n+1:1} \) and \( h = h_{m+1:1} \)

(94)

be minimal decompositions of \( g \) and \( h \) (resp.). In addition, we denote: \( b = g_{n+1:n-q+1} \) where \( q = C(f^{-1}) \) and \( c = g_{n-q+1:1} \circ h_{m+1:1} \). Assume by contradiction that \( f \triangleright g \circ h \). Then, we apply Lem. 15 and conclude that \( f \triangleright b \) or \( f \circ b \in \text{Invariant}(\mathcal{N}) \). In addition, since \( C(b) = C(f^{-1}) \leq I \leq C(g) \), we decompose \( c \) as follows: \( c = d_1 \circ d_2, d_1 = g_{n-q+1:1} \) and \( d_2 = h_{m+1:1} \). By Lem. 5, \( b \not\triangleright d_1 \) and \( d_1 \not\triangleright d_2 \). In addition,

\[ C(f \circ b \circ d_1) = C(f \circ g) = C(f) + C(g) = C(f) + C(b \circ d_1) = C(f) + C(b) + C(d_1) \]

(95)

Therefore, by Lem. 5, \( f \not\triangleright b \) and \( C(f \circ b \circ d_1) > C(d_1) \). This immediately eliminates the option \( f \triangleright b \). On the other hand, if \( f \circ b \in \text{Invariant}(\mathcal{N}) \) then by Lem. 9,

\[ C(f \circ b \circ d_1) \leq C(d_1) \]

(96)

24
in contradiction. Therefore, \( f \not\sqsubset g \circ h \) and
\[
C(f \circ g \circ h) = C(f) + C(g \circ h) = C(f) + C(g) + C(h)
\] (97)

**Induction hypothesis:** Assume that the statement holds for all \( f \) such that \( C(f) = i \geq 1 \), \( g \) and \( h \) that satisfy, i.e.,
\[
I \leq C(g), f \not\sqsubset g \text{ and } g \not\sqsubset h \implies C(f \circ g \circ h) = C(f) + C(g) + C(h)
\] (98)

**Induction step:** Let \( f \) be a function such that \( C(f) = i + 1 \) and \( f \not\sqsubset g \). We denote \( f' = f' \circ f_1 \) such that \( f' \not\sqsubset f_1, C(f_1) = 1 \) and \( C(f') = i \). Then, by Lem. 5, \( f_1 \not\sqsubset g \). Therefore, by case \( i = 1 \) we have:
\[
C(f_1 \circ g \circ h) = C(f_1) + C(g) + C(h) = C(f_1 \circ g) + C(h).
\]
Alternatively, \( f_1 \circ g \not\sqsubset h \).
In addition, since \( f \not\sqsubset g \) then \( C(f' \circ (f_1 \circ g)) = C(f) + C(g) = C(f') + 1 + C(g) = C(f') + C(f_1 \circ g) \).
Alternatively, \( f' \not\sqsubset f_1 \circ g \). Therefore, because \( I \leq C(g) < C(f_1 \circ g) \) by the induction hypothesis we have:
\[
I \leq C(f_1 \circ g), f' \not\sqsubset f_1 \circ g \text{ and } f_1 \circ g \not\sqsubset h
\]
\[
\implies C(f \circ g \circ h) = C(f' \circ (f_1 \circ g) \circ h) = C(f') + C(f_1 \circ g) + C(h)
\]
\[
= C(f') + 1 + C(g) + C(h) = C(f) + C(g) + C(h)
\] (99)

\[\blacksquare\]
Appendix G. Proofs for Thms. 2 and 3

Lemma 17 Let \( N = \text{SCM}[C] \). Let \( A = (X_A, D_A) \) and \( B = (X_B, D_B) \) be two domains and \( D_Z \) a distribution. Then, if \( m \geq k + C_{Z,B} \), we have:

\[
C^{k, \epsilon_0 + \epsilon_1}_{A,B} \leq C^{m, \epsilon_0}_{A,Z} + C^{m, \epsilon_1}_{Z,B}
\]

(100)

In particular,

\[
C^{\epsilon_0 + \epsilon_1}_{A,B} \leq C^{\epsilon_0}_{A,Z} + C^{\epsilon_1}_{Z,B}
\]

(101)

Proof Let \( f \in H_{\epsilon_0}(D_A, D_Z; m) \) and \( g \in H_{\epsilon_1}(D_Z, D_B; m) \). Then, by the first item of Lem. 2, for \( D_1 \leftarrow f \circ D_A, D_2 \leftarrow D_Z, D_3 \leftarrow D_B, p \leftarrow g \) and \( m \geq k + C_{Z,B} \), we have:

\[
\text{disc}_k(g \circ f \circ D_A, D_B) \leq \text{disc}_m(f \circ D_A, D_Z) + \text{disc}_m(g \circ D_Z, D_B) \leq \epsilon_0 + \epsilon_1
\]

Therefore,

\[
C^{k, \epsilon_0 + \epsilon_1}_{A,B} \leq C(f) + C(g) = C^{m, \epsilon_0}_{A,Z} + C^{m, \epsilon_1}_{Z,B}
\]

(103)

In order to prove the second inequality, we choose \( k \geq \max \{ E^{\epsilon_0 + \epsilon_1}_{A,B}, E^{\epsilon_1}_{A,Z}, E^{\epsilon_0}_{Z,B} \} \) and obtain:

\[
C^{\epsilon_0 + \epsilon_1}_{A,B} = C^{k, \epsilon_0 + \epsilon_1}_{A,B} \leq C^{m, \epsilon_0}_{A,Z} + C^{m, \epsilon_1}_{Z,B} = C^{\epsilon_0}_{A,Z} + C^{\epsilon_1}_{Z,B}
\]

(104)

\[ \blacksquare \]

Lemma 18 Let \( N = \text{SCM}[C] \) and \( A = (X_A, D_A) \) and \( B = (X_B, D_B) \) are two domains. If \( m \geq k \geq E^{\epsilon_0}_{A,B} \) then \( H_{\epsilon_0}(D_A, D_B; m) \subset H_{\epsilon_0}(D_A, D_B; k) \).

Proof Let \( y \in H_{\epsilon_0}(D_A, D_B; m) \). Since \( m \geq k \geq E^{\epsilon_0}_{A,B} \), we have: \( C(y) \leq C^{m, \epsilon_0}_{A,B} \leq C^{k, \epsilon_0}_{A,B} \). In addition, since \( D_B \subset D_m \), \( \text{disc}_k(g \circ D_A, D_B) \leq \text{disc}_m(g \circ D_A, D_B) \leq \epsilon_0 \). In particular, \( y \in H_{\epsilon_0}(D_A, D_B; k) \).

Lemma 19 Let \( N = \text{SCM}[C] \). Let \( A = (X_A, D_A) \) and \( B = (X_B, D_B) \) be two domains and \( D_Z \) a distribution.

If \( m \geq \max \{ E^{\epsilon_0 + \epsilon_1}_{A,B}, E^{\epsilon_1}_{A,Z}, E^{\epsilon_0}_{Z,B} \} \), then, \( D_Z \) is a \((m, \epsilon_0, \epsilon_1, \epsilon_0 + \epsilon_1)\)-shared semantic distribution between \( A \) and \( B \) iff:

\[
C^{\epsilon_0 + \epsilon_1}_{A,B} = C^{\epsilon_1}_{Z,B} + C^{\epsilon_0}_{A,Z}
\]

(105)

Proof

Part 1: Assume that \( D_Z \) is a \((m, \epsilon_0, \epsilon_1, \epsilon_0 + \epsilon_1)\)-shared semantic distribution between \( A \) and \( B \). Then, there is a function \( y_{AB} \in Z(D_A, D_Z; m; m, \epsilon_0, \epsilon_1) \) such that \( y_{AB} \in H_{\epsilon_0 + \epsilon_1}(D_A, D_B; m) \) and \( y^{-1}_A \in H_{\epsilon_0}(D_A, D_Z; m) \) and \( y_B \neq y^{-1}_A \). Therefore,

\[
C^{\epsilon_0 + \epsilon_1}_{A,B} = C^{m, \epsilon_0 + \epsilon_1}_{A,B} = C(y_{AB}) = C(y_B) + C(y^{-1}_A) = C^{\epsilon_0}_{A,Z} + C^{\epsilon_1}_{Z,B}
\]

(106)

Part 2: Assume that \( C^{\epsilon_0 + \epsilon_1}_{A,B} = C^{\epsilon_1}_{Z,B} + C^{\epsilon_0}_{A,Z} \). Let \( t = m + C^{\epsilon_1}_{Z,B} \), \( y_B \in H_{\epsilon_0}(D_Z, D_B; t) \) and \( y^{-1}_A \in H_{\epsilon_0}(D_A, D_Z; t) \) and denote \( y_{AB} = y_B \circ y^{-1}_A \). By the second item of Lem. 2, for \( D_1 \leftarrow D_A, D_2 \leftarrow D_B, D_3 \leftarrow D_Z, p \leftarrow y_B \) and \( t \geq m + C(y_B) \) we have:

\[
\text{disc}_m(y_{AB} \circ D_A, D_B) \leq \text{disc}(y^{-1}_A \circ D_A, D_Z) + \text{disc}(y_B \circ D_Z, D_B) \leq \epsilon_0 + \epsilon_1
\]

(107)

Therefore,

\[
C^{\epsilon_0 + \epsilon_1}_{A,B} \leq C(y_{AB}) \leq C(y_B) + C(y^{-1}_A) = C^{\epsilon_1}_{Z,B} + C^{\epsilon_0}_{A,Z} = C^{\epsilon_0 + \epsilon_1}_{A,B}
\]

(108)
In particular, \( y_{AB} \in \mathcal{Z}(D_A, D_Z, D_B; m, \epsilon_0, \epsilon_1) \cap H_{\epsilon_0 + \epsilon_1}(D_A, D_B; m) \) and \( y_B \not\in y_A^{-1} \). Alternatively, \( D_Z \) is a \((m, \epsilon_0, \epsilon_1, \epsilon_0 + \epsilon_1)\)-shared semantic distribution.

**Note:** we assumed that \( H_{\epsilon_1}(D_Z, D_B; t) \neq \emptyset \) and \( H_{\epsilon_0}(D_A, D_Z; t) \neq \emptyset \). It follows from the assumption that for every two distributions of interest in the paper, \( D_I \) and \( D_J \), and an error rate \( \epsilon > 0 \), there is a function \( h \) of finite complexity such that \( \text{disc}_\infty(h \circ D_I, D_J) \leq \epsilon \).
G.1 Topological properties of decompositions

Let $u_{n+1:1}$ be a decomposition such that $u_i = σ ◦ W_i$ for $i ∈ [n]$. We denote:

$$S_ε(W) = \{ \bar{W} \mid ||W - \bar{W}||_2 < ε \}$$

and,

$$S_ε(W_1, ..., W_n) = S_ε(W_1) × ... × S_ε(W_n)$$

**Definition 1 (Perturbation of a function)** Let $\mathcal{N} = SCM[σ]$ be a NN-SCM and a decomposition $u_{n+1:1}$.

The set of $ε$-perturbations of $u_{n+1:1}$ is:

$$S_ε(u_{n+1:1}) = \{ \bar{u}_{n+1:1} \mid \bar{u}_i = σ ◦ \bar{W}_i \text{ for } i ∈ [n] \text{ and } (\bar{W}_1, ..., \bar{W}_n) ∈ S_ε(W_1, ..., W_n) \}$$

**Lemma 20** Let $\mathcal{N} = SCM[σ]$ be a NN-SCM with $σ$ that is Leaky ReLU with parameter $a > 0$. Let $u_{n+1:1}$ be a decomposition such that $u_i = σ ◦ W_i$ for $i ∈ [n]$. Therefore, we have:

$$∀ x : ||u_{n+1:1}(x)||_2 ≤ \max\{a, 1\}^n ◦ \prod_{i=1}^n ||W_i||_2 ◦ ||x||_2$$

and,

$$∀ x : ||σ^{-1} ◦ u_{n+1:1}(x)||_2 ≤ \max\{a, 1\}^{n-1} ◦ \prod_{i=1}^n ||W_i||_2 ◦ ||x||_2$$

**Proof** We prove the second inequality by induction on $n ≥ 1$. The first inequality follows immediately from the second and $||σ(x)||_2 ≤ ||a \cdot x||_2 ≤ a \cdot ||x||_2$.

**Case $n = 1$** We have:

$$||σ^{-1} ◦ u_{n+1:1}(x)||_2 = ||W_1(x)||_2 ≤ ||W_1||_2 ◦ ||x||_2$$

**Induction hypothesis:** We assume that:

$$∀ x : ||σ^{-1} ◦ u_{n+1:1}(x)||_2 ≤ \max\{a, 1\}^{n-1} ◦ \prod_{i=1}^n ||W_i||_2 ◦ ||x||_2$$

As mentioned above, an immediate consequence is the following inequality,

$$∀ x : ||u_{n+1:1}(x)||_2 ≤ \max\{a, 1\}^n ◦ \prod_{i=1}^n ||W_i||_2 ◦ ||x||_2$$

**Case $n + 1$** We have:

$$||σ^{-1} ◦ u_{n+2:1}(x)||_2 = ||W_n ◦ u_{n+1:1}(x)||_2 ≤ ||W_n ◦ u_{n+1:1}(x)||_2$$

$$≤ ||W_n||_2 ◦ ||u_{n+1:1}(x)||_2$$

And by the induction hypothesis we have:

$$||σ^{-1} ◦ u_{n+2:1}(x)||_2 ≤ ||W_n||_2 ◦ \max\{a, 1\}^n ◦ \prod_{i=1}^n ||W_i||_2 ◦ ||x||_2$$

$$= \max\{a, 1\}^n ◦ \prod_{i=1}^{n+1} ||W_i||_2 ◦ ||x||_2$$

$$= \max\{a, 1\}^n ◦ \prod_{i=1}^{n+1} ||W_i||_2 ◦ ||x||_2$$
Lemma 21 Let $\mathcal{N} = \text{SCM}[\sigma]$ with $\sigma$ that is Leaky ReLU with parameter $a > 0$. In addition, let $u_{n+1}$ be a decomposition such that $C(u_{n+1}) = n$. Then, there is $\epsilon > 0$ such that if $\bar{u}_{n+1} \in S_{\epsilon}(u_{n+1})$ then: $C(\bar{u}_{n+1}) = n$.

Proof We denote $u_{n+1} = (\sigma \circ W_n) \circ \ldots \circ (\sigma \circ W_1)$. Assume by contradiction that for all $\epsilon > 0$, there is $\bar{u}_{n+1} = (\sigma \circ W_n) \circ \ldots \circ (\sigma \circ W_1)$ such that

$$C(\bar{u}_{n+1}) < n \text{ and } \forall i \in [n]: ||W_i - \bar{W}_i||_2 < \epsilon$$

Then, there is a sequence $\{u^k_{n+1}\}_{k=1}^\infty$ such that $\forall i \in [n]: ||W_i - W^k_i||_2 < \epsilon/k$ we have $C(u^k_{n+1}) < n$. Here, $u^k_{n+1} = u^k_n \circ \ldots \circ u^1_1 = (\sigma \circ W^k_n) \circ \ldots \circ (\sigma \circ W^k_1)$. By identifiability, for every $k$, there are indexes $p_k, q_k \leq n$ such that $p_k + 2 \leq q_k$ and

$$C(u^k_{p_k+1:q_k}) = 1$$

Since the sequence $\{u^k_{n+1}\}_{k=1}^\infty$ is infinitely long, by the pigeonhole principle, there is a tuple $(p, q)$ such that $p + 2 \leq q$ and there are infinitely many indexes $k$ that satisfy $C(u^k_{p+1:q}) = 1$. Therefore, with no loss of generality, we can assume that for all $k$ we have $C(u^k_{p+1:q}) = 1$ for a fixed tuple of indexes $p, q \leq n$ that satisfy $p + 2 \leq q$ (or replace the original sequence with such a sequence).

We denote $U^k$ an invertible linear mapping such that:

$$u^k_{q+1:p} = (\sigma \circ U^k)$$

In particular,

$$\sigma^{-1} \circ u^k_{q+1:p} = U^k$$

Therefore,

$$\forall x: ||\sigma^{-1} \circ u^k_{q+1:p}(x)||_2 = ||U^k(x)||_2$$

By Lem. 20,

$$||\sigma^{-1} \circ u^k_{q+1:p}(x)||_2 \leq \max\{a, 1\}^{q-p-1} \prod_{i=p}^q ||u^k_i|| \cdot ||x||_2$$

Therefore,

$$||U^k||_2 \leq \max\{a, 1\}^{q-p-1} \prod_{i=p}^q (||W_i|| + \epsilon/k)$$

And also, if the input dimension is $M$ and $|| \cdot ||_F$ is the Frobenius norm, then,

$$||U^k||_F \leq M \cdot \max\{a, 1\}^{q-p-1} \prod_{i=p}^q (||W_i|| + \epsilon/k)$$

In particular, the sequence $U^k$ is bounded. Thus, by the Bolzano-Weierstrass theorem, there is a subsequence $U^{k_i}$ that converges to a matrix $U$ (w.r.t the $\ell_2$ norm). With no loss of generality, we can replace $U^k$ with the sequence $U^{k_i}$. Alternatively, we can assume that $U^{k_i}$ converges to the matrix $U$ (w.r.t the $\ell_2$ norm). In addition, we have $W^k_i \rightarrow W_i$ (w.r.t the $\ell_2$ norm). Therefore, since for any $x$, $f_{E_n,...,E_1}(x)$ (see Tab. 1) is continuous as a function of a the matrices $E_n, \ldots, E_1$, we have:

$$\forall x: \lim_{k \rightarrow \infty} u^k_{q+1:p}(x) = u_{q+1:p}(x)$$

On the other hand, since for any $x$, the function $f_E(x)$ is continuous as a function of the matrix $E$, we have:

$$\lim_{k \rightarrow \infty} u^k_{q+1:p}(x) = \lim_{k \rightarrow \infty} (\sigma \circ U^k)(x) = (\sigma \circ U)(x)$$
Finally,
\[ \forall x : u_{q+1:p}(x) = (\sigma \circ U)(x) \] (129)

Alternatively,
\[ u_{q+1:p} = (\sigma \circ U) \] (130)

Therefore, \( C(u_{q+1:p}) \leq 1 \). On the other hand, since \( C(u_{n+1}) = n \), by Lem. 5 we have \( C(u_{q+1:p}) = q - p > 1 \) in contradiction.

**Lemma 22** Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). In addition, let \( u_{n+1} \) be a decomposition such that \( C(u_{n+1}) = 1 \). Then, for every \( \epsilon > 0 \) there is a \( \bar{u}_{n-1} \in S_c(u_{n-1}) \) such that \( C(u_n \circ \bar{u}_{n-1} \circ u_{n-2}) > 1 \).

**Proof** We assume that \( C(u_{n+1}) = 1 \) and show that for every \( \epsilon > 0 \) there is \( \bar{u}_{n-1} \in S_c(u_{n-1}) \) such that \( C(u_n \circ \bar{u}_{n-1} \circ u_{n-2}) > 1 \). Assume by contradiction that there is \( \epsilon > 0 \) such that for all \( \bar{u}_{n-1} \in S_c(u_{n-1}) \) we have \( C(u_n \circ \bar{u}_{n-1} \circ u_{n-2}) = 1 \). By Lem. 8, there is no possibility that \( n = 2 \). Therefore, \( n \geq 3 \). We denote \( u_i = \sigma \circ W_i \) for \( i \in [n] \). Let
\[ \forall k = 1, 2 : \bar{u}_{n+1}^k = u_n \circ u_{n-1}^k \circ u_{n-2} \] (131)
where \( u_{n-1}^k = u_{n-1} \circ V_k \) such that:
\[ \forall k = 1, 2 : ||W_{n-1} \cdot V_{k} - W_{n-1}||_2 \leq ||W_{n-1}||_2 \cdot ||V_k - \text{Id}||_2 < \epsilon \] (132)
and
\[ \sigma^{-1} \circ (W_{n-1} \circ V_1 \cdot V_2^{-1} \circ W_{n-1}^{-1}) \circ \sigma \text{ is not linear} \] (133)

Thus, \( \forall k = 1, 2 : \bar{u}_{n-1}^k \in S_c(u_{n-1}) \). Therefore,
\[ u_n \circ (u_{n-1} \circ V_k) \circ u_{n-2} = \sigma \circ U_k \] (134)
for some linear mappings \( U_k \) for \( k = 1, 2 \). In particular,
\[ \sigma \circ W_n \circ \sigma \circ W_{n-1} \cdot V_1 \cdot V_2^{-1} \cdot W_{n-1}^{-1} \circ \sigma^{-1} \circ W_{n-1} \circ \sigma^{-1} = u_n \circ u_{n-1} \circ V_1 \cdot V_2^{-1} \circ u_{n-1}^{-1} \circ u_n \]
\[ = (u_n \circ (u_{n-1} \circ V_1) \circ u_{n-2}^{-1}) \circ (u_n \circ (u_{n-1} \circ V_2) \circ u_{n-2}^{-1})^{-1} \] (135)
\[ = (\sigma \circ U_1 \circ (\sigma \circ U_2)^{-1}) = \sigma \circ U_1 \cdot U_2^{-1} \circ \sigma^{-1} \]

Alternatively,
\[ \sigma \circ W_n \cdot V_1 \cdot V_2^{-1} \cdot W_{n-1}^{-1} \cdot \sigma^{-1} = W_n^{-1} \cdot U_1 \cdot U_2^{-1} \cdot W_n \] (136)
in contradiction to the assumption that \( \sigma \circ W_n \cdot V_1 \cdot V_2^{-1} \cdot W_{n-1}^{-1} \cdot \sigma^{-1} \) is not linear.

**Lemma 23** Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). In addition, let \( u_{n+1} \) be a decomposition such that \( C(u_{n+1}) = 1 \). Then, for every \( \epsilon > 0 \) there is an \( \bar{u}_{n+1} \in S_c(u_{n+1}) \) such that \( C(u_n \circ \bar{u}_{n+1} \circ u_1) = n \).

**Proof** We prove this claim by induction on \( n \).

**Case** \( n = 2 \): By Lem. 8, there is no possibility that \( C(u_{3:1}) = 1 \). Therefore, the claim follows immediately.

**Case** \( n = 3 \): By Lem. 22, there is \( \bar{u}_{4:1} \in S_c(u_{4:1}) \) such that \( C(\bar{u}_{4:1}) = k > 1 \). If \( k = 2 \), then, by identifiability, \( C(\bar{u}_{4:2}) = 1 \) or \( C(\bar{u}_{3:1}) = 1 \), in contradiction to Lem. 8. Therefore, \( k = 3 = n \).
Case $n = 4$: By Lem. 22, there is $u_{3:1} \in S_{4:1}(u_{5:1})$ such that $C(u_{3:1}) = k > 1$. If $k = 3$, then, by identifiability, there is an index $i \leq n - 1$ such that $C(u_{i+2:1}) = 1$ in contradiction to Lem. 8. If $k = 2$, then, by identifiability, there are three options:

- $C(u_{3:1}) = 1$ and $C(u_{5:1}) = 1$.
- $C(u_{3:1}) = 1$.
- $C(u_{4:1}) = 1$.

The first option is not a possibility, since it contradicts Lem. 8. The second and third options are analogous - we prove them in parallel. By Case $n = 3$, there is $u_{4:2} \in S_{3:2}(u_{4:2})$ such that $C(u_{4:2}) = 3$. By Lem. 21, there is $\epsilon_0 > 0$ such that for every $u_{4:2} \in S_{4:1}(u_{4:2})$ we have: $C(u_{4:2}) = 3$. We denote $\epsilon' = \min\{\epsilon/4, \epsilon_0\}$.

Again, by Case $n = 3$, there is $u_{5:2} \in S_{5:2}(u_{5:2})$ such that $C(u_{5:2}) = 3$. Therefore, we have:

$$C(u_{4:2} \circ u_{5:2}) = 3 \quad \text{and} \quad C(u_{4:2} \circ u_{5:2}) = 3$$

By the above, $C(u_{4:2} \circ u_{5:2} \circ u_{1}) \neq 2, 3$. In addition, by Lem. 21 and Lem. 22 there is $\epsilon'' \leq \epsilon/4$ such that there is $u_{4:2} \in S_{4:2}(u_{4:2})$ that satisfies:

$$C(u_{4:2} \circ u_{5:2} \circ u_{1}) > 1, \quad C(u_{4:2} \circ u_{5:2}) = 3 \quad \text{and} \quad C(u_{4:2} \circ u_{5:2}) = 3$$

Therefore, $C(u_{4:2} \circ u_{5:2} \circ u_{1}) \neq 2, 3$. We conclude that $C(u_{4:2} \circ u_{5:2} \circ u_{1}) = 4$. We consider that $u_{4:2} \in S_{4:2}(u_{4:2}) \subset S_{3:2}(u_{4:2}) \subset S_{2:1}(u_{4:2}) \subset S_{1:1}(u_{4:2})$. Alternatively, we found $u_{4:2} \in S_{4:2}(u_{4:2})$ such that $C(u_{4:2} \circ u_{5:2} \circ u_{1}) = 4$.

**Induction hypothesis:** We assume that for every $k \leq n$ and decomposition $u_{k+1:1}$ such that $C(u_{k+1:1}) = 1$, for every $\epsilon > 0$ there is an $u_{k:2} \in S_{k}(u_{k:2})$ such that $C(u_{k:2} \circ u_{k:2} \circ u_{1}) = k$.

Case $n + 1$: By the induction hypothesis,

$$\exists u_{n:2} \in S_{n:2}(u_{n:2}) \quad \text{such that} \quad C(u_{n:2} \circ u_{n:2} \circ u_{1}) = n$$

In addition, by Lem. 21, there is $\epsilon_0 > 0$ such that

$$\forall u_{n:2} \in S_{n:2}(u_{n:2}) \quad \Rightarrow \quad C(u_{n:2} \circ u_{n:2} \circ u_{1}) = n$$

In particular,

$$\forall u_{n:2} \in S_{n:2}(u_{n:2}) \quad \Rightarrow \quad u_{n:2} \circ u_{1} \in S_{n:2}(u_{n:2} \circ u_{1}) \quad \Rightarrow \quad C(u_{n:2} \circ u_{1}) = n$$

We denote $\epsilon' = \min\{\epsilon/2, \epsilon_0\}$ and obtain: $\forall u_{n:2} \in S_{n:2}(u_{n:2}) \quad \Rightarrow \quad C(u_{n:2} \circ u_{1}) = n$. In addition, by Case $n = 4$, there is $u_{6:2} \in S_{6:2}(u_{6:2})$ such that $C(u_{6:2} \circ u_{6:2}) = 3$. Thus, we have:

$$C(u_{n:2} \circ u_{1}) = n \quad \text{and} \quad C(u_{n:1} \circ u_{n:1:n-2} = 4)$$

Therefore, by Thm. 4, since $u_{n:1} \neq u_{n:1:n-2} \neq u_{n:1:n-2}$ and $C(u_{n:1:n-2}) = 3 \quad \Rightarrow \quad C(u_{n:1:n-2} \circ u_{1}) = n + 1$

In addition,

$$u_{n:2} \in S_{n:2}(u_{n:2})$$

since $u_{n:2} \in S_{n:2}(u_{n:2})$ and $u_{n:2} \in S_{n:2}(u_{n:2})$ and $\epsilon' \leq \epsilon/2$. 

31
Lemma 24  Let $\mathcal{N} = \text{SCM}[\sigma]$ with $\sigma$ that is Leaky ReLU with parameter $a > 0$. In addition, let $u_{n+1:1}$ be a decomposition such that $C(u_{n+1:1}) = n$. Then, if for every $\epsilon > 0$ there is $\bar{u}_{n+1:1} \in S_k(u_{n+1:1})$ such that $C(\bar{u}_{q+1:p}) \leq 1$ then $C(u_{q+1:p}) \leq 1$.

Proof Assume that for every $\epsilon > 0$, there is $\bar{u}_{n+1:1} \in S_k(u_{n+1:1})$ such that $C(\bar{u}_{q+1:p}) = 1$. Let $\{u_{n+1:1}^k\}_{k=1}^\infty$ be a sequence such that $u_{n+1:1}^k \in S_{1/k}(u_{n+1:1})$ and $C(u_{q+1:p}) = 1$. We denote, $u_{q+1:p}^k = (\sigma \circ W_{q+1:p}^k)$ for $i \in [n]$ and $k \in \mathbb{N}$. We denote $U^k$ a invertible linear mapping such that:

$$u_{q+1:p}^k = (\sigma \circ U^k) \quad (145)$$

In particular,

$$\sigma^{-1} \circ u_{q+1:p}^k = U^k \quad (146)$$

Therefore, 

$$\forall x : ||\sigma^{-1} \circ u_{q+1:p}^k(x)||_2 = ||U^k(x)||_2 \quad (147)$$

By Lem. 20, 

$$||\sigma^{-1} \circ u_{q+1:p}^k(x)||_2 \leq \max\{a, 1\}^{q-p-1} \prod_{i=p}^q ||W_i^k||_2 \cdot ||x||_2 \quad (148)$$

$$\leq \max\{a, 1\}^{q-p-1} \prod_{i=p}^q (||W_i||_2 + \epsilon/k) \cdot ||x||_2$$

Therefore, 

$$||U^k||_2 \leq \max\{a, 1\}^{q-p+1} \prod_{i=p}^q (||W_i||_2 + \epsilon/k) \quad (149)$$

And also, if the input dimension is $M$ and $|| \cdot ||_F$ is the Frobenious norm, then,

$$||U^k||_F \leq M \cdot \max\{a, 1\}^{q-p-1} \prod_{i=p}^q (||W_i||_2 + \epsilon/k) \quad (150)$$

In particular, the sequence $U^k$ is bounded. Thus, by the Bolzano-Weierstrass theorem, there is a subsequence $U^{k_\epsilon}$ that converges to a matrix $U$ (w.r.t the $\ell_2$ norm). With no loss of generality, we can replace $U^k$ with the sequence $U^{k_\epsilon}$. Alternatively, we can assume that $U^k$ converges to the matrix $U$ (w.r.t the $\ell_2$ norm). In addition, we have $W_{q+1:p}^k \to W_q$ (w.r.t the $\ell_2$ norm). Therefore, since for any $x$, $f_{E_n, \ldots, E_1}(x)$ is continuous as a function of the matrices $E_n, \ldots, E_1$, we have:

$$\lim_{k \to \infty} u_{q+1:p}^k(x) = u_{q+1:p}(x) \quad (151)$$

On the other hand, since for any $x$, the function $f_E(x)$ is continuous as a function of a matrix $E$, we have:

$$\lim_{k \to \infty} u_{q+1:p}^k(x) = \lim_{k \to \infty} (\sigma \circ U^k)(x) = (\sigma \circ U)(x) \quad (152)$$

Finally, 

$$\forall x : u_{q+1:p}(x) = (\sigma \circ U)(x) \quad (153)$$

Alternatively, 

$$u_{q+1:p} = (\sigma \circ U) \quad (154)$$

Therefore, $C(u_{q+1:p}) \leq 1$. 

\[ \square \]
Lemma 25 Let $\mathcal{N} = SCM[\sigma]$ with $\sigma$ that is Leaky ReLU with parameter $\alpha > 0$. In addition, let $u_{n+1:1}$ and $v_{k+1:1}$ be two decompositions such that $C(u_{n+1:1}) = n$ and $C(v_{k+1:1}) = k$. Then, for every $\epsilon > 0$, there is $\bar{u}_{n:1} \circ \bar{v}_{k+1:2} \in S_{\epsilon}(u_{n:1} \circ v_{k+1:2})$ such that: $C(u_{n} \circ \bar{u}_{n:1} \circ \bar{v}_{k+1:2} \circ v_{1:1}) = n + k$.

Proof Assume by contradiction that there is $\epsilon > 0$ such that:

$$\forall \bar{u}_{n:1} \circ \bar{v}_{k+1:2} \in S_{\epsilon}(u_{n:1} \circ v_{k+1:2}) \implies C(u_{n} \circ \bar{u}_{n:1} \circ \bar{v}_{k+1:2} \circ v_{1:1}) < n + k$$

(155)

If $u_{n+1:1} \not\in v_{k+1:1}$ then we take $\bar{u}_{n:1} = u_{n:1}$ and $\bar{v}_{k+1:2} = v_{k+1:2}$. Otherwise, by identifiability, there are indexes $p \leq k$ and $q \leq n$ such that

$$C(u_{q+1:1} \circ v_{k+1:p}) = 1$$

and

$$C(u_{n+1:q+1} \circ v_{q+1:1} \circ v_{k+1:p} \circ v_{1:1}) = (n - q) + 1 + (p - 1) = n - q + p$$

(156)

By Lem. 21, there is $\epsilon_0 > 0$ such that $\bar{u}_{n:1} \in S_{\epsilon_0}(u_{n:1})$ and $\bar{v}_{k+1:2} \in S_{\epsilon_0}(v_{k+1:2})$ then $C(u_{n} \circ \bar{u}_{n:1}) = n$ and $C(\bar{v}_{k+1:2} \circ v_{1:1}) = k$. We denote $\epsilon_1 = \min(\epsilon_0, \epsilon)$. We assumed that if $u_{n:1} \circ \bar{v}_{k+1:2} \in S_{\epsilon}(u_{n:1} \circ v_{k+1:2})$ then $C(u_{n} \circ \bar{u}_{n:1} \circ \bar{v}_{k+1:2} \circ v_{1:1}) < n + k$. Thus, by identifiability, for every $\bar{u}_{n:1} \circ \bar{v}_{k+1:2} \in S_{\epsilon_1}(u_{n:1} \circ v_{k+1:2})$ there are $r < s$ such that:

$$C(\bar{u}_{r+1:1} \circ \bar{v}_{s+1:1}) = 1$$

and

$$C(u_{n} \circ \bar{u}_{r+1:1} \circ \bar{v}_{s+1:1} \circ v_{1:1}) = (n - r) + 1 + (s - 1) = n - r + s$$

(157)

Thus, by Lem. 22, for every sequence $\{u_{n+1:1} \circ v_{k+1:2}\}_{i=1}^{\infty}$ such that

$$u_{n+1:1} \circ v_{k+1:2} \in S_{\epsilon_1}(u_{n:1} \circ v_{k+1:2})$$

$$C(u_{n+1:1} \circ v_{k+1:2}) = 1$$

where $s \leq k$ and $r \leq n$

$$C(u_{n} \circ u_{1:1} \circ v_{k+1:2} \circ v_{1:1}) = n - r + s$$

(158)

$$\implies r = q$$

In particular, there is an $0 < \epsilon_2 \leq \epsilon_1$ such that

$$\bar{u}_{n:1} \circ \bar{v}_{k+1:2} \in S_{\epsilon_2}(u_{n:1} \circ v_{k+1:2})$$

$$C(u_{n} \circ \bar{u}_{n:1} \circ \bar{v}_{k+1:2} \circ v_{1:1}) = n - r + s$$

(159)

Therefore,

$$\bar{u}_{n:1} \circ \bar{v}_{k+1:2} \in S_{\epsilon_2}(u_{n:1} \circ v_{k+1:2}) \implies C(\bar{u}_{q+1:1} \circ \bar{v}_{k+1:1}) = 1$$

(160)

In particular,

$$\bar{u}_{q+1:1} \circ \bar{v}_{k+1:1} \in S_{\epsilon_2}(u_{q+1:1} \circ v_{k+1:1}) \implies C(\bar{u}_{q+1:1} \circ \bar{v}_{k+1:1}) = 1$$

(161)

in contradiction to Lem. 23.

Lemma 26 Let $\mathcal{N} = SCM[\sigma]$ with $\sigma$ that is Leaky ReLU with parameter $\alpha > 0$. In addition, let $u_{n+1:1}$ be a decomposition. Then, for every $\epsilon > 0$ there is $\bar{u}_{n:2} \in S_{\epsilon}(u_{n:2})$ such that: $C(u_{n} \circ \bar{u}_{n:2} \circ u_{1:1}) = n$.

Proof If $C(u_{n+1:1}) = n$ then we take $\bar{u}_{n+1:1} = u_{n+1:1}$. Otherwise, we denote $C(u_{n+1:1}) = k$ and by identifiability,

$$\exists j_1 = 1 < \ldots < j_{k+1} = n + 1 : C(u_{j_{k+1}:1}) = 1$$

(162)

By Lem. 8, for every $i$ such that $j_i + 1 \not\in j_{i+1}$, we have: $j_{i+1} + 3 \leq j_{i+1}$. By Lem. 23, for each $i$ such that $j_i + 1 \not\in j_{i+1}$

$$\exists u_{j_i+1:1-1:j_i+1} \in S_{\epsilon/k}(u_{j_i+1-1:j_i+1})$$

such that $C(u_{j_i+1} \circ \bar{u}_{j_i+1-1:j_i+1} \circ u_{j_i:1}) = j_{i+1} - j_i$

(163)

We denote:

$$\forall i \in [k] : v_{j_i:1} = u_{j_i}$$

and

$$\forall t \in [n] \setminus \{j_1, \ldots, j_k\} : v_{t:1} = \bar{u}_{t}$$

(164)
We denote:
\[ \forall i \in [k+1] : j^1_i = j_i \]
(165)

We have:
\[ \forall i \in [k] : C(v^1_{j^1_{i+1}:j^1_i}) = j^1_{i+1} - j^1_i \]
(166)

\[ v^1_{n:2} \in S_{\epsilon/k}(u_{n:2}) \]
\[ v^1_1 = u_1 \]

By Lem. 25, there is a decomposition
\[ v^1_{j^1_{i}:j^1_{i+1}} \in S_{\epsilon/k}(v^1_{j^1_{i}:j^1_{i+1}}) \]
(167)
such that \( C(v^1_{j^1_{i}:j^1_{i+1}} \circ v^1_{j^1_{i}:j^1_{i+1}} \circ v^1_{j^1_{i}:j^1_{i+1}}) = j^1_{i+1} - j^1_i \). Next, we replace \( v^1_{n+1:1} \) with \( v^2_{n+1:1} \) defined as follows:
\[ \forall i \geq j^1_i : v^2_i = v^1_i, \forall 1 < i < j^1_i : v^2_i = v^1_i \text{ and } v^2_{n+1} = v^1_n = u_1 \]
(168)

and we denote:
\[ j^2_i = 1 \text{ and } \forall i \in [k] \setminus \{1\} : j^2_i = j^1_{i+1} \]
(169)

We have:
\[ \forall i \in [k-1] : C(v^2_{j^2_{i+1}:j^2_i}) = j^2_{i+1} - j^2_i \]
(170)
\[ v^2_{n:2} \in S_{\epsilon/k}(u_{n:2}) \]
\[ v^2_1 = v^1_1 = u_1 \]

We continue the process of replacing each \( v^k_{n+1:1} \) with \( v^{k+1}_{n+1:1} \) for \( k - 1 \) times. For each iteration, we have a decomposition that satisfies:
\[ \forall i \in [k-t+1] : C(v^t_{j^t_{i+1}:j^t_i}) = j^t_{i+1} - j^t_i, j^t_1 = 1, j^t_{n+2} = n + 1, v^t_1 = u_1, \]
(171)
\[ v^t_n = u_n \text{ and } v^t_{n+2} \in S_{\epsilon/k}(u_{n+2}) \]

and obtain that \( \bar{u}_{n+1:1} = v^k_{n+1:1} \) satisfies:
\[ C(\bar{u}_{n+1:1}) = n, \bar{u}_1 = u_1, \bar{u}_n = u_n \text{ and } \bar{u}_{n+2} \in S_{\epsilon}(u_{n+2}) \]
(172)

Lemma 27 Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). In addition, let \( u_{n+1:1} \) such that \( C(u_{n+1:1}) = n \). Then, for all \( \epsilon > 0 \) there is \( u_{n+1:1} \in S_{\epsilon}(u_{n+1:1}) \) such that \( C(u_{n+1:1}) = n + 2 \).

Proof We define a one-to-one function between compositions:
\[ F((\sigma \circ U_n) \circ ... \circ (\sigma \circ U_1)) = (\sigma \circ U_n^{-1}/a) \circ ... \circ (\sigma \circ U_1^{-1}/a) \]
(173)

We consider that for any decomposition \( d_{n+1:1} \), we have:
\[ d_{n+1:1} = (\sigma \circ -\text{Id}) \circ F(d_{n+1:1}) \circ (\sigma \circ -\text{Id}/a) \]
(174)

Let \( u = u_{n+1:1} \) and \( u_{n+1:1} \) be two decompositions such that \( u_i = \sigma \circ W_i \) and \( u_i = \sigma \circ U_i \), where \( W_i \) and \( U_i \) are invertible linear mapping. Since \( W^{-1} \) is continuous as a function of \( W \), there is \( \epsilon' > 0 \) such that:
\[ \forall i \in [n] : ||W^{-1}_{i}/a - U^{-1}_{i}/a|| < \epsilon' \implies \forall i \in [n] : ||W_i - U_i|| < \epsilon \]
(175)
In particular,
\[ S_ϵ′(F(u_{n+1:1})) \subset \{ \bar{v}_{n+1:1} \mid F^{-1}(\bar{v}_{n+1:1}) \in S_ϵ(u_{n+1:1}) \} \] (176)

By Lem. 26, there is \( \bar{v}_{n+1:1} \in S_ϵ′(F(u_{n+1:1})) \) such that
\[ C((\sigma \circ -\text{Id}) \circ \bar{v}_{n+1:1} \circ (\sigma \circ -\text{Id}/a)) = n + 2 \] (177)

Therefore, there is a decomposition \( u_{n+1:1} = F^{-1}(\bar{v}_{n+1:1}) \in S_ϵ(u_{n+1:1}) \) such that:
\[ \tilde{u}_{n+1:1}^{-1} = (\sigma \circ -\text{Id}) \circ \bar{v}_{n+1:1} \circ (\sigma \circ -\text{Id}/a) \] (178)
Thus, \( C(\tilde{u}_{n+1:1}^{-1}) = n + 2 \).

**Theorem 2** Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). The set of sequences of invertible matrices \( (W_1, ..., W_n) \in \mathbb{R}^{M \times M \times n} \) such that \( C(u_{n+1:1}^{-1}) = C(u_{n+1:1}^{-1}) + 2 \) and \( u_i = \sigma \circ W_i \) (for \( i \in [n] \)) is open and dense in \( \mathbb{R}^{M \times M \times n} \).

**Proof** We denote by \( F \) the function from Eq. 173,
\[ Z = \{ (W_1, ..., W_n) \in \mathbb{R}^{M \times M \times n} \mid \forall i \in [n] : u_i = \sigma \circ W_i, W_i \text{ is invertible and } C(u_{n+1:1}^{-1}) = C(u_{n+1:1}) + 2 \} \] (179)
and,
\[ Z' = \{ (W_1, ..., W_n) \in \mathbb{R}^{M \times M \times n} \mid \forall i \in [n] : W_i \text{ is invertible} \} \] (180)

**Openness:** Let \( u_{n+1:1} \) be a decomposition such that \( C(u_{n+1:1}^{-1}) = n + 2 \) and \( u_i = \sigma \circ W_i \). We have:
\[ u_{n+1:1}^{-1} = (\sigma \circ -\text{Id}) \circ F(u_{n+1:1}) \circ (\sigma \circ -\text{Id}/a) \] (181)
is a minimal decomposition. By Lem. 21, there is an \( \epsilon > 0 \) such that:
\[ \epsilon_{n+1:1} \in S_ϵ(F(u_{n+1:1})) \implies C((\sigma \circ -\text{Id}) \circ \epsilon_{n+1:1} \circ (\sigma \circ -\text{Id}/a)) = n + 2 \] (182)
Since \( W^{-1} \) is continuous as a function of \( W \), there is \( \epsilon' > 0 \) such that
\[ \forall i \in [n] : ||W_i - U_i|| < \epsilon' \implies i \in [n] : ||W_i^{-1}/a - U_i^{-1}/a|| < \epsilon \] (183)
In particular,
\[ S_ϵ(u_{n+1:1}) \subset \{ \bar{u}_{n+1:1} \mid F(\bar{u}_{n+1:1}) \in S_ϵ(F(u_{n+1:1})) \} \] (184)
Therefore,
\[ \bar{u}_{n+1:1} \in S_ϵ(u_{n+1:1}) \implies C(\bar{u}_{n+1:1}^{-1}) = C((\sigma \circ -\text{Id}) \circ F(\bar{u}_{n+1:1}) \circ (\sigma \circ -\text{Id}/a)) = n + 2 \] (185)

**Density:** By Lem. 27, the set \( Z \) is dense in \( Z' \). In addition, the set \( Z' \) is dense in \( \mathbb{R}^{M \times M \times n} \). Since density is a transitive relation, \( Z \) is also dense in \( \mathbb{R}^{M \times M \times n} \).

**Theorem 3** Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). The set of sequences of invertible matrices \( (W_1, ..., W_n) \in \mathbb{R}^{M \times M \times n} \) such that \( C(u_{n+1:1}) = n \) where \( u_i = \sigma \circ W_i \) (for \( i \in [n] \)) is open and dense in \( \mathbb{R}^{M \times M \times n} \).

**Proof** We denote:
\[ Z = \{ (W_1, ..., W_n) \in \mathbb{R}^{M \times M \times n} \mid \forall i \in [n] : u_i = \sigma \circ W_i, W_i \text{ is invertible and } C(u_{n+1:1}) = n \} \] (186)
and,
\[ Z' = \{ (W_1, ..., W_n) \in \mathbb{R}^{M \times M \times n} \mid \forall i \in [n] : W_i \text{ is invertible} \} \] (187)
By Lem. 21 and Lem. 27, the set \( Z \) is open and dense in \( Z' \). In addition, \( Z' \) is open and dense in \( \mathbb{R}^{M \times M \times n} \). Openness and density are both transitive relations. Therefore, \( Z \) is also open and dense in \( \mathbb{R}^{M \times M \times n} \).
Appendix H. Proof of Thm. 5

H.1 Covering numbers

Definition 14 (Set embedding) Let $(U, \sim_U)$ and $(V, \sim_V)$ be two tuples of sets and symmetric and reflexive relations on them (resp.). A function $F : U \to V$ is an embedding of $(U, \sim_U)$ in $(V, \sim_V)$ and we denote $(U, \sim_U) \preceq (V, \sim_V)$ if:

$$\forall u_1, u_2 \in U : F(u_1) \sim_V F(u_2) \implies u_1 \sim_U u_2$$  \hspace{1cm} (188)

Lemma 28 Let $(U, \sim_U)$ and $(V, \sim_V)$ be two tuples of sets and equivalence relations on them (resp.). Assume that $|U|/\sim_U$ and $|V|/\sim_V$ are finite. Then,

$$(U, \sim_U) \preceq (V, \sim_V) \iff |U|/\sim_U \leq |V|/\sim_V$$  \hspace{1cm} (189)

Proof Assume that $(U, \sim_U) \preceq (V, \sim_V)$. Then, there is a function $F : U \to V$ such if $F(u_1) \sim_V F(u_2)$ then $u_1 \sim_U u_2$. We denote $n := |U|/\sim_U$. Let $d_1, \ldots, d_n$ be $n$ representatives of the $n$ classes in $U/\sim_U$, i.e., $\forall i \neq j \leq n : d_i \not\sim_U d_j$. Then, $\forall i \neq j \leq n : F(d_i) \not\sim_V F(d_j)$. Therefore, we found $n$ representatives of $n$ different equivalence classes of $(V, \sim_V)$. In particular, $|U|/\sim_U \leq |V|/\sim_V$.

Assume that $|U|/\sim_U \leq |V|/\sim_V$. We denote $n := |U|/\sim_U$. Let $d_1, \ldots, d_n$ be $n$ representatives of the $n$ classes in $U/\sim_U$, i.e., $\forall i \neq j \leq n : d_i \not\sim_U d_j$. Similarly, $e_1, \ldots, e_m$ are $m$ representatives of the $m$ classes in $V/\sim_V$, i.e., $\forall i \neq j \leq m : e_i \not\sim_V e_j$. We define a mapping $F(u_1) = F(d_i) = e_i$ if $u \sim_U d_i$. Thus, for any $u_1, u_2 \in U$, if $u_1 \sim_U u_2$ then $F(u_1) = F(u_2)$. In addition, if $F(u_1) \not\sim_V F(u_2)$ then $F(u_1) = e_i$ and $F(u_2) = e_j$ for $i \neq j \leq n$. Thus, $u_1 \sim_U d_i$ and $u_2 \sim_U d_j$ such that $d_i \neq d_j$. But, by the way we defined $d_1, \ldots, d_n$ we have: $d_i \neq d_j$ implies $d_i \not\sim_U d_j$. Therefore, $u_1 \sim_U u_2$ iff $F(u_1) \sim_V F(u_2)$. Alternatively, $F$ is an embedding from $(U, \sim_U)$ to $(V, \sim_V)$.

Lemma 29 Let $(U, \sim_U)$ be a tuples of a set and an equivalence relation on it (resp.). Then,

$$\text{Covering}(U, \sim_U) = |U|/\sim_U$$  \hspace{1cm} (190)

Proof First, we consider that $(U, \sim_U)$ is a covering of itself and therefore, $\text{Covering}(U, \sim_U) \leq |U|/\sim_U$. Assume by contradiction that $\text{Covering}(U, \sim_U) < |U|/\sim_U$. Thus, there is a covering $(U, \equiv_U)$ of $(U, \sim_U)$ such that $|U|/\equiv_U < |U|/\sim_U$. But, by definition $u_1 \equiv_U u_2 \implies u_1 \sim_U u_2$. Thus, if $u_1, \ldots, u_n \in U$ are $n$ representatives of $n$ different equivalence classes in $(U, \sim_U)$ then $u_1, \ldots, u_n \in U$ are also $n$ representatives of $n$ different equivalence classes in $(U, \equiv_U)$. Therefore, $|U|/\sim_U < |U|/\equiv_U$ in contradiction to $|U|/\equiv_U < |U|/\sim_U$. Finally, we conclude that: $\text{Covering}(U, \sim_U) = |U|/\sim_U$.

Lemma 30 Let $(U, \sim_U)$ and $(V, \sim_V)$ be two tuples of sets and reflexive and symmetric relations on them (resp.). If $(U, \sim_U) \preceq (V, \sim_V)$ then $\text{Covering}(U, \sim_U) \preceq \text{Covering}(V, \sim_V)$.

Proof Assume that $(U, \sim_U) \preceq (V, \sim_V)$. Then, by definition, there is an embedding function $F : U \to V$ such that:

$$\forall u_1, u_2 \in U : F(u_1) \sim_V F(u_2) \implies u_1 \sim_U u_2$$  \hspace{1cm} (191)

Let $(V, \equiv_V)$ be a covering of $(V, \sim_V)$. We define a covering $(U, \equiv_U)$ of $(U, \sim_U)$ as follows:

$$u_1 \equiv_U u_2 \iff F(u_1) \equiv_V F(u_2)$$  \hspace{1cm} (192)
Part 1: We would like to prove that \((\mathcal{U}, \equiv_\mathcal{U})\) is a covering of \((\mathcal{U}, \sim_\mathcal{U})\). It is easy to see that \(\equiv_\mathcal{U}\) is an equivalence relation since \(\equiv_\mathcal{V}\) is an equivalence relation. Next, we would like to prove that \(u_1 \equiv_\mathcal{U} u_2 \implies u_1 \sim_\mathcal{U} u_2\). By the definition of \(\equiv_\mathcal{U}\):
\[
  u_1 \equiv_\mathcal{U} u_2 \implies F(u_1) \equiv_\mathcal{V} F(u_2)
\] (193)
In addition, since \((\mathcal{V}, \equiv_\mathcal{V})\) is a covering of \((\mathcal{V}, \sim_\mathcal{V})\):
\[
  F(u_1) \equiv_\mathcal{V} F(u_2) \implies F(u_1) \sim_\mathcal{V} F(u_2)
\] (194)
Finally, since \(F\) is an embedding:
\[
  F(u_1) \sim_\mathcal{V} F(u_2) \implies u_1 \sim_\mathcal{U} u_2
\] (195)
We conclude:
\[
  u_1 \equiv_\mathcal{U} u_2 \implies u_1 \sim_\mathcal{U} u_2
\] (196)
Therefore, \((\mathcal{U}, \equiv_\mathcal{U})\) is indeed a covering of \((\mathcal{U}, \sim_\mathcal{U})\).

Part 2: We would like to prove that \(|\mathcal{U}/ \equiv_\mathcal{U}| \leq |\mathcal{V}/ \equiv_\mathcal{V}|\). Let \(u_1, u_2 \in \mathcal{U}\) such that \(u_1 \not\equiv_\mathcal{U} u_2\). Then, by definition of \(\equiv_\mathcal{U}\) we have: \(F(u_1) \not\equiv_\mathcal{V} F(u_2)\). Therefore, if we take \(u_1, \ldots, u_n \in \mathcal{U}\) representations of \(n\) different equivalence classes in \((\mathcal{U}, \equiv_\mathcal{U})\) then, \(F(u_1), \ldots, F(u_n) \in \mathcal{V}\) are \(n\) representations of \(n\) different equivalence classes in \((\mathcal{V}, \equiv_\mathcal{V})\). In particular, \(|\mathcal{U}/ \equiv_\mathcal{U}| \leq |\mathcal{V}/ \equiv_\mathcal{V}|\). Therefore, the covering number of \((\mathcal{U}, \sim_\mathcal{U})\) is at most the covering number of \((\mathcal{V}, \sim_\mathcal{V})\).

Lemma 31 Let \((\mathcal{U}, \equiv_1)\) and \((\mathcal{U}, \equiv_2)\) be coverings of \((\mathcal{U}, \sim_\mathcal{U})\). Then, \((\mathcal{U}^2, \equiv_1 \times \equiv_2)\) is a covering of \((\mathcal{U}^2, \sim_\mathcal{U}^2)\). Where \(\mathcal{U}^2 = \mathcal{U} \times \mathcal{U}\) and the relation \(\equiv_\mathcal{U}^2\) is defined as follows:
\[
  (a, b) \sim_\mathcal{U}^2 (c, d) \iff a \sim_\mathcal{U} c \text{ and } b \sim_\mathcal{U} d
\] (197)
and \(\equiv_1 \times \equiv_2\) is defined as:
\[
  (a, b) \equiv_1 \times \equiv_2 (c, d) \iff a \equiv_1 c \text{ and } b \equiv_2 d
\] (198)

Proof We have to prove that \(\equiv_1 \times \equiv_2\) is an equivalence relation and that \((u_1, u_2) \equiv_1 \times \equiv_2 (v_1, v_2) \implies (u_1, u_2) \sim_\mathcal{U}^2 (v_1, v_2)\).

Reflexivity:
\[
  (u_1, u_2) \equiv_1 \times \equiv_2 (u_1, u_2) \iff u_1 \equiv_1 u_1 \text{ and } u_2 \equiv_1 u_2
\] (199)
The RHS is true since \(\equiv_1\) and \(\equiv_2\) are reflexive relations.

Symmetry:
\[
  (u_1, u_2) \equiv_1 \times \equiv_2 (v_1, v_2) \iff u_1 \equiv_1 v_1 \text{ and } u_2 \equiv_2 v_2
\] (200)
Since \(\equiv_1\) and \(\equiv_2\) are symmetric, we have:
\[
  u_1 \equiv_1 v_1 \text{ and } u_2 \equiv_2 v_2 \iff v_1 \equiv_1 u_1 \text{ and } v_2 \equiv_2 u_2
\] (201)
In addition,
\[
  (v_1, v_2) \equiv_1 \times \equiv_2 (u_1, u_2) \iff v_1 \equiv_1 u_1 \text{ and } v_2 \equiv_2 u_2
\] (202)
Therefore,\n\[
  (u_1, u_2) \equiv_1 \times \equiv_2 (v_1, v_2) \iff (v_1, v_2) \equiv_1 \times \equiv_2 (u_1, u_2)
\] (203)

Transitivity: follows from similar arguments.

Covering:
\[
  (u_1, u_2) \equiv_1 \times \equiv_2 (v_1, v_2) \iff u_1 \equiv_1 v_1 \text{ and } u_2 \equiv_2 v_2
\] (204)
Since \((\mathcal{U}, \equiv_i)\) is a covering of \((\mathcal{U}, \sim_U)\), for \(i = 1, 2\), we have:
\[
u_1 \equiv_1 v_1 \text{ and } u_2 \equiv_2 v_2 \implies u_1 \sim_U v_1 \text{ and } u_2 \sim_U v_2
\]
(205)
By the definition of \(\sim_U^2\) we have:
\[
u_1 \sim_U v_1 \text{ and } u_2 \sim_U v_2 \iff (u_1, u_2) \sim_U (v_1, v_2)
\]
(206)
Therefore,
\[(u_1, u_2) \equiv_1 \times \equiv_2 (v_1, v_2) \implies (u_1, u_2) \sim_U^2 (v_1, v_2)
\]
(207)

\[
\begin{align*}
\text{Lemma 32} & \quad \text{Let } (\mathcal{U}, \sim_U) \text{ be a tuple of a set and a reflexive and symmetric relation on it (resp.). Then,} \\
\text{Covering}(\mathcal{U}^2, \sim_U^2) & = \text{Covering}(\mathcal{U}, \sim_U)^2
\end{align*}
\]
\[
(208)
\]

\textbf{Proof}
(\(\leq\)):
Let \(\equiv_U\) be an equivalence relation such that \((\mathcal{U}, \equiv_U)\) is a covering of \((\mathcal{U}, \sim_U)\). By Lem. 31, \((\mathcal{U}^2, \equiv_U^2)\) is a covering of \((\mathcal{U}^2, \sim_U^2)\). In addition,
\[
|\mathcal{U}^2/ \equiv_U^2| = |\mathcal{U}/ \equiv_U|^2
\]
(209)
Thus, for every covering \((\mathcal{U}, \equiv_U)\) of \((\mathcal{U}, \sim_U)\) we can construct a covering of \((\mathcal{U}^2, \sim_U^2)\) of size \(|\mathcal{U}/ \equiv_U|^2\). In particular,
\[
\text{Covering}(\mathcal{U}^2, \sim_U^2) \leq \text{Covering}(\mathcal{U}, \sim_U)^2
\]
(210)
(\(\geq\)):
Let \((\mathcal{U}^2, \equiv)\) be a covering of \((\mathcal{U}^2, \sim_U^2)\). Let \(\equiv\) be the following equivalence relation on \(\mathcal{U}\) (proving that \(\equiv_1, \equiv_2\) are equivalence relations is left for the reader),
\[
u_1 \equiv_1 v_1 \iff \exists u_1, u_2 \in \mathcal{U} \text{ s.t. } (u_1, u_2) \equiv (v_1, v_2)
\]
(211)
we similarly define \(\equiv_2\) (w.r.t the second coordinate). We also note that \((\mathcal{U}, \equiv_1)\) and \((\mathcal{U}, \equiv_2)\) are coverings of \((\mathcal{U}, \sim_U)\). That is because,
\[
u_1 \equiv_1 v_1 \iff \exists u_1, v_2 \text{ s.t. } (u_1, v_1) \equiv (u_2, v_2)
\]
\[
\implies u_1 \sim_U u_2
\]
(212)
Let \(u_1, \ldots, u_n\) are representations of the \(n\) different equivalence classes of \(\equiv_1\) and \(v_1, \ldots, v_k\) are representations of the \(k\) different equivalence classes of \(\equiv_2\) then: \((u_i, v_j) \equiv (u_s, v_t)\) iff \(u_i = u_s\) and \(v_j = v_t\). Otherwise, there are \(u_i, u_s\) and \(v_j, v_t\) such that (with no loss of generality), \(u_i \neq u_s\) and \((u_i, v_j) \equiv (u_s, v_t)\). Since \(u_i \neq u_s\) then by the way we defined \(u_1, \ldots, u_n\) we have \(u_i \neq v_1\). Thus, by definition, there are no \(v, v' \in \mathcal{U}\) such that \((u_i, v) \equiv (u_s, v')\) in contradiction to \((u_i, v_j) \equiv (u_s, v_t)\). Thus, \(\{(u_i, v_j)\}_{i,j \in [n] \times [k]}\) are in different equivalence classes by \(\equiv\). In particular,
\[
|\mathcal{U}^2/ \equiv_1 \times \equiv_2| \leq |\mathcal{U}/ \equiv_1| \cdot |\mathcal{U}/ \equiv_2| \leq |\mathcal{U}^2/ \equiv_U^2|
\]
(213)
And by Lem. 31,
\[
\text{Covering}(\mathcal{U}, \sim_U)^2 \leq |\mathcal{U}^2/ \equiv_1 \times \equiv_2|
\]
(214)
By combining Eq. 213 and 214,
\[
\text{Covering}(\mathcal{U}, \sim_U)^2 \leq |\mathcal{U}^2/ \equiv_U^2|
\]
(215)
Therefore,
\[
\text{Covering}(\mathcal{U}, \sim_U)^2 \leq \text{Covering}(\mathcal{U}^2, \sim_U^2)
\]
(216)
By taking minimum over the RHS of Eq. 215 over the different possibilities of \(\equiv_U^2\).
**Lemma 33** Let \((U, \sim_U)\) and \((V, \sim_V)\) be two tuples of sets and reflexive and symmetric relations on them (resp.). Assume that \(U \subset V\) and \(\sim_U = (\sim_V)|_U\), i.e.,
\[
\forall u, v \in U : u \sim_U v \iff u \sim_V v
\]
(217)

Then,
\[
\text{Covering}(U, \sim_U) \leq \text{Covering}(V, \sim_V)
\]
(218)

**Proof** Let \((V, \equiv_V)\) be a covering of \((V, \sim_V)\). Then, it is easy to see that \((U, \equiv_U)\) is a covering of \((U, \sim_U)\), where \(\equiv_U = (\equiv_V)|_U\). In addition, we have: \(|U/ \equiv_U| \leq |V/ \equiv_V|\). Thus, for every covering of \((V, \sim_V)\), we can find a smaller covering for \((U, \sim_U)\). In particular, \(\text{Covering}(U, \sim_U) \leq \text{Covering}(V, \sim_V)\). 
\(\blacksquare\)

**Lemma 34** Let \((U_1, \sim_{U_1})\) and \((U_2, \sim_{U_2})\) be two tuples of sets and reflexive and symmetric relations on them (resp.). Assume that \(U_1 \cap U_2 = \emptyset\). We define \((U, \sim_U)\) as follows:
\[
U = U_1 \cup U_2
\]
\[
u \sim_U v \iff \exists i = 1, 2 : u, v \in U_i and u \sim_{U_i} v
\]
(219)

Then,
\[
\text{Covering}(U, \sim_U) \leq \text{Covering}(U_1, \sim_{U_1}) + \text{Covering}(U_2, \sim_{U_2})
\]
(220)

**Proof** Let \((U_1, \equiv_{U_1})\) and \((U_2, \equiv_{U_2})\) be minimal coverings of \((U_1, \sim_{U_1})\) and \((U_2, \sim_{U_2})\) (resp.). We define a covering \((U, \equiv_U)\) of \((U, \sim_U)\) as follows:
\[
u \equiv_U v \iff \exists i = 1, 2 : u, v \in U_i and u \equiv_{U_i} v
\]
(221)

Since both \(\equiv_{U_1}\) and \(\equiv_{U_2}\) are equivalence relations, \(\equiv_U\) is also an equivalence relation. In addition,
\[
\begin{align*}
u \equiv_U v & \Rightarrow \exists i = 1, 2 : u, v \in U_i and u \equiv_{U_i} v \\
& \Rightarrow \exists i = 1, 2 : u, v \in U_i and u \sim_{U_i} v \\
& \Rightarrow u \sim_U v
\end{align*}
\]
(222)

We also consider that the members \(U_1\) and the members of \(U_2\) do not share equivalence classes by \(\equiv_U\). In addition, the members of \(U_1\) are partitioned into equivalence classes by \(\equiv_{U_1}\). Similarly, \(U_2\) w.r.t \(\equiv_{U_2}\). Therefore,
\[
\text{Covering}(U, \sim_U) \leq |U/ \equiv_U| = |U_1/ \equiv_{U_1}| + |U_2/ \equiv_{U_2}|
\]
\[
= \text{Covering}(U_1, \sim_{U_1}) + \text{Covering}(U_2, \sim_{U_2})
\]
(223)
\(\blacksquare\)
H.2 Perturbations and discrepancy

Assumption 1 Let \( N = SCM[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). For every \( m > 0 \) and \( n > 0 \), the function

\[
\text{disc}_m( f_{W_{n}, \ldots, W_1} \circ D_1, D_2 )
\]

is continuous as a function of the weights of \( W_{n}, \ldots, W_1 \). Here, \( f_{W_{n}, \ldots, W_1} = (\sigma \circ W_n) \circ \ldots \circ (\sigma \circ W_1) \).

Assumption 2 Let \( N = SCM[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). For all \( m > 0 \), the function

\[
R_D[ f_{V_{m}, \ldots, V_1}, f_{W_{m}, \ldots, W_1} ]
\]

is continuous as a function of \( V_{m}, \ldots, V_1, W_{m}, \ldots, W_1 \).

Lemma 35 Let \( N = SCM[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \) and assume Assumption 2 with \( D : \leftarrow D_1 \). Let \( \text{disc}_{m,E} = \text{disc}_{c_m,E} \) for \( c_m,E = \{ u_{m+1,1} \mid vi \in [m] : u_i = \sigma \circ W_i \text{ s.t. } ||W_i|| \leq E \} \). For all \( m > 0, n > 0 \) and \( E > 0 \), the function

\[
\text{disc}_{m,E}( f_{W_{n}, \ldots, W_1} \circ D_1, D_2 )
\]

is continuous as a function of \( W_{n}, \ldots, W_1 \).

Proof Let \( W_{n}, \ldots, W_1 \) and \( W_{n}^k, \ldots, W_1^k \) be any invertible matrices in \( \mathbb{R}^{M \times M} \) such that for all \( i \in [n] \), \( W_i^k \rightarrow W_i \). We denote \( G_E = \{ W \in \mathbb{R}^{M \times M} \mid ||W|| \leq E \} \). By the triangle inequality,

\[
\text{disc}_{m,E}( D_1, D_2 ) \leq \text{disc}_{m,E}( D_1, D_3 ) + \text{disc}_{m,E}( D_3, D_2 ) \]

Similarly,

\[
\text{disc}_{m,E}( D_3, D_2 ) \leq \text{disc}_{m,E}( D_1, D_3 ) + \text{disc}_{m,E}( D_3, D_2 )
\]

therefore,

\[
|\text{disc}_{m,E}( D_3, D_2 ) - \text{disc}_{m,E}( D_1, D_2 )| \leq \text{disc}_{m,E}( D_1, D_3 )
\]

In particular,

\[
|\text{disc}_{m,E}( f_{W_{n}, \ldots, W_1} \circ D_1, D_2 ) - \text{disc}_{m,E}( f_{W_{n}^k, \ldots, W_1^k} \circ D_1, D_2 )| \leq \text{disc}_{m,E}( f_{W_{n}, \ldots, W_1} \circ D_1, f_{W_{n}^k, \ldots, W_1^k} \circ D_1 )
\]

Assume by contradiction that the last expression does not converge to 0. Therefore, there is a sequence \( (V_{m}^k, \ldots, V_1^k, U_{m}^k, \ldots, U_1^k) \) such that \( V_{m}^k, \ldots, V_1^k, U_{m}^k, \ldots, U_1^k \in G_E \) and

\[
|R_D[ f_{V_{m}^k, \ldots, V_1^k}, f_{U_{m}^k, \ldots, U_1^k} ]| \neq 0
\]
Since \((V^{k}_1, ..., V^{k}_m, U^{k}_1, ..., U^{k}_m) \in G^{2m}_E\) and \(G^{2m}_E\) is compact in \(\mathbb{R}^{M \times M \times 2m}\), by the Bolzano-Weierstrass theorem, there is a converging subsequence. Alternatively, there is an increasing sequence \(\{k_j\}_{j=1}^{\infty} \subset \mathbb{N}\) such that
\[
(V^{k_j}_1, ..., V^{k_j}_m, U^{k_j}_1, ..., U^{k_j}_m) \to (V_1, ..., V_m, U_1, ..., U_m) \in G^{2m}_E
\]  
(230)

In particular,
\[
(V^{k_j}_m, ..., V^{k_j}_1, W^{k_j}_m, ..., W^{k_j}_1) \to (V_m, ..., V_1, W_m, ..., W_1)
\]
(231)
\[
(U^{k_j}_m, ..., U^{k_j}_1, W^{k_j}_m, ..., W^{k_j}_1) \to (U_m, ..., U_1, W_m, ..., W_1)
\]

By Assumption 2, the function \(R_{D_i} [f_{V_{m}, ..., V_{1}, W_{m}, ..., W_{1}}, f_{U_{m}, ..., U_{1}, W_{m}, ..., W_{1}}]\) is a continuous. Therefore,
\[
\left| R_{D_1} [f_{V^{k_j}_m, ..., V^{k_j}_1, W^{k_j}_m, ..., W^{k_j}_1}, f_{U^{k_j}_m, ..., U^{k_j}_1, W^{k_j}_m, ..., W^{k_j}_1}] - R_{D_1} [f_{V_{m}, ..., V_{1}, W_{m}, ..., W_{1}}, f_{U_{m}, ..., U_{1}, W_{m}, ..., W_{1}}] \right| \to 0
\]
(232)
and,
\[
\left| R_{D_1} [f_{V^{k_j}_m, ..., V^{k_j}_1, W^{k_j}_m, ..., W^{k_j}_1}, f_{U^{k_j}_m, ..., U^{k_j}_1, W^{k_j}_m, ..., W^{k_j}_1}] - R_{D_1} [f_{V_{m}, ..., V_{1}, W_{m}, ..., W_{1}}, f_{U_{m}, ..., U_{1}, W_{m}, ..., W_{1}}] \right| \to 0
\]
(233)

and therefore, by the triangle inequality,
\[
\left| R_{D_1} [f_{V^{k_j}_m, ..., V^{k_j}_1, W^{k_j}_m, ..., W^{k_j}_1}, f_{U^{k_j}_m, ..., U^{k_j}_1, W^{k_j}_m, ..., W^{k_j}_1}] - R_{D_1} [f_{V_{m}, ..., V_{1}, W_{m}, ..., W_{1}}, f_{U_{m}, ..., U_{1}, W_{m}, ..., W_{1}}] \right| \to 0
\]
(234)
in contradiction. Thus, we conclude that:
\[
\lim_{k \to \infty} \left| \text{disc}_{m,E}(f_{W_{m}, ..., W_{1}} \circ D_1, D_2) - \text{disc}_{m,E}(f_{W_{m}, ..., W_{1}} \circ D_1, D_2) \right| = 0
\]
(235)

\[\blacksquare\]

**Lemma 36** Let \(\mathcal{N} = \text{SCM}[\sigma]\) with \(\sigma\) that is Leaky ReLU with parameter \(\alpha > 0\). In addition, let \(f = f_{n+1:1}\) and \(g = g_{n+1:1}\) be two decompositions. Assume that \(\text{disc}_{m,E}(g \circ D_A, D_B) \leq \epsilon_0\) and \(m \geq t + n + 2\). Then, there are \(\bar{f} = f_{n+1:1}, \bar{g} = g_{n+1:1}\) and \(\bar{q} = q_{n+3:1}\) such that:

1. \(\bar{q}_{n+2:1} = F_+ (\bar{g}_{n+1:1}) \circ (\sigma \circ \text{Id}/\alpha)\).
2. \(C(\bar{f}_{n+1:1} \circ \bar{g}_{n+3:1}) = 2(n + 2)\).
3. \(\forall j \in [n - 1]: \bar{q}_{n-j+3:1} \circ \bar{g} \circ D_A = (-\sigma^{-1}) \circ \bar{g}_{j+1:1} \circ D_A\).
4. \(\text{disc}_{m}(\bar{g} \circ D_B, D_A) \leq \epsilon_0 + \epsilon\).
5. \(\forall j \in [n]: \text{disc}_{m}(f_{j+1:1} \circ D_A, f_{j+1:1} \circ D_A) \leq \epsilon\).
6. \(\forall j \in [n]: \text{disc}_{m}(\bar{g}_{j+1:1} \circ D_A, g_{j+1:1} \circ D_A) \leq \epsilon\).

**Proof** We denote \(F^+\) and \(F\) are the functions from Eq 60 and Eq. 173 (resp.). Let:
\[
q_{n+3:1} = F^+ (g_{n+1:1})
\]
(236)
Let $q = \tilde{q}_{n+3:1} = \tilde{q}_{n+3:2} \circ (\sigma - \text{Id}/a)$ and $\bar{g} = F^{-1}(\tilde{q}_{n+2:2})$. By the first item of Lem. 2, for $D_1 : \leftarrow D_B$, $D_2 : \leftarrow \bar{g} \circ D_A$, $D_3 : \leftarrow D_A$, $p : \leftarrow q$, $m \geq t + n + 2 \geq t + C(\bar{q})$, we have:

$$\text{disc}(\bar{q} \circ D_B, D_A) \leq \text{disc}_m(\bar{q} \circ \bar{g} \circ D_A, D_A) + \text{disc}_m(\bar{g} \circ D_A, D_B)$$

$$= \text{disc}_m(\tilde{q}_{n+2} \circ q^{-1}_{n+2} \circ D_A, D_A) + \text{disc}_m(\bar{g} \circ D_A, D_B)$$

$$\leq \text{disc}_m(\tilde{q}_{n+2} \circ q^{-1}_{n+2} \circ D_A, D_A) + \text{disc}_m(g \circ D_A, D_B) + \text{disc}_m(\bar{g} \circ D_A, g \circ D_A)$$

$$\leq \text{disc}_m(\tilde{q}_{n+2} \circ q^{-1}_{n+2} \circ D_A, D_A) + \epsilon_0 + \text{disc}_m(\bar{g} \circ D_A, g \circ D_A)$$

(237)

The weights of $\tilde{q}_{n+1:1}$ are continuous functions of the weights of $\tilde{q}_{n+2:2}$. In addition, for all $j \in [n]$, the functions $\text{disc}_m(\tilde{f}_{j+1:1} \circ D_A, \tilde{f}_{j+1:1} \circ D_A)$, $\text{disc}_m(\tilde{g}_{j+1:1} \circ D_A, \tilde{g}_{j+1:1} \circ D_A)$ and $\text{disc}_m(\tilde{q}_{n+2} \circ q^{-1}_{n+2} \circ D_A, D_A)$ are continuous as a function of the weights of $\tilde{f}_{n+1:1}$ and $\tilde{g}_{n+1:1}$ (resp.). In particular,

$$\left(\text{disc}_m(\tilde{f}_{j+1:1} \circ D_A, \tilde{f}_{j+1:1} \circ D_A)\right)^n_{j=1}$$

$$\left|\text{disc}_m(\tilde{g}_{j+1:1} \circ D_A, \tilde{g}_{j+1:1} \circ D_A)\right|^n_{j=1}$$

$$\left|\text{disc}_m(\tilde{q}_{n+2} \circ q^{-1}_{n+2} \circ D_A, D_A)\right|$$

is a continuous function of the weights of $\tilde{f}_{n+1:1} \circ \tilde{q}_{n+3:1}$. Here, $||$ is the concatenations operator between tuples, i.e. $(x_1, \ldots, x_i)||(y_1, \ldots, y_m) = (x_1, \ldots, x_i, y_1, \ldots, y_m)$.

Therefore, for every $\epsilon > 0$ there is $\epsilon' > 0$ such that for all $f_{n:1} \circ \tilde{q}_{n+3:2} \in \mathcal{C}(f_{n:1} \circ q_{n+3:2})$ we have:

$$\forall j \in [n] : \text{disc}_m(\tilde{f}_{j+1:1} \circ D_A, \tilde{f}_{j+1:1} \circ D_A) \leq \epsilon/2$$

$$\text{disc}_m(\tilde{g}_{j+1:1} \circ D_A, \tilde{g}_{j+1:1} \circ D_A) \leq \epsilon/2$$

$$\text{disc}_m(\tilde{q}_{n+2} \circ q^{-1}_{n+2} \circ D_A, D_A) \leq \epsilon/2$$

(239)

Where $\bar{q} = \tilde{q}_{n+3:1} = \tilde{q}_{n+3:2} \circ (\sigma - \text{Id}/a)$ and $\bar{g} = F^{-1}(\tilde{q}_{n+2:2})$.

By Lem. 26, there is $f_{n+1:1} \circ \tilde{q}_{n+3:1}$ such that:

- $f_{n+1} \circ \tilde{q}_{n+3:2} \in \mathcal{C}(f_{n+1} \circ q_{n+3:2})$.
- $C(f_{n+1:1} \circ \tilde{q}_{n+3:1}) = 2n + 2$.
- $\tilde{q}_1 = q_1 = (\sigma - \text{Id}/a)$.
- $f_{n+1} = \tilde{f}_n$.

Finally, by Lem. 12,

$$\forall j \in [n+1] : \tilde{q}_{j-3:1} \circ \tilde{g} \circ D_A = (-\sigma^{-1}) \circ \tilde{g}_{j:1} \circ D_A$$

(240)

Therefore, we found $\tilde{f}_{n+1:1} \circ \tilde{q}_{n+3:1}$ with all the desired properties.

**Lemma 37** Let $\mathcal{N} = \text{SCM}[\sigma]$ with $\sigma$ that is Leaky ReLU with parameter $a > 0$. Let $f \overset{D}{\sim}_{m,\epsilon} g$. Then, for every minimal decomposition $f = f'_{n+1:1}$ there is a minimal decomposition $g = g'_{n+1:1}$ such that:

$$\forall i \in [n] : f'_{i+1:1} \circ D \sim_{m,\epsilon} g'_{i+1:1} \circ D$$

(241)

**Proof** Since $f \overset{D}{\sim}_{m,\epsilon} g$ there are minimal decompositions $f = f_{n+1:1}$ and $g = g_{n+1:1}$ such that:

$$\forall i \in [n] : f_{i+1:1} \circ D \sim_{m,\epsilon} g_{i+1:1} \circ D$$

(242)

42
By minimal identifiability, $f'_i = \pi_1 \circ f_i$, for all $i = 2, \ldots, n - 1$: $f'_{i+1} = \pi_1 \circ f_{i+1} \circ \pi_{i+1}^{-1}$ and $f'_n = f_n \circ \pi_n^{-1}$. Therefore, we define a minimal decomposition for $g$ as follows: $g = g_{n+1:1} \circ \pi_{n+1}^{-1}$ such that $g'_i = \pi_i \circ g_i$, for all $i = 2, \ldots, n - 1: g'_i = \pi_i \circ g_i \circ \pi_i^{-1}$ and $g'_n = g_n \circ \pi_n^{-1}$. This is a minimal decomposition of $g$, since each invariant function is an invertible linear mapping and commutes with $\sigma$. By Lem. 7 we have:

$$f'_{i+1} = \pi_1 \circ f_{i+1} \circ \pi_{i+1}^{-1} = \pi_1 \circ g_{i+1} \quad (243)$$

Therefore,

$$\text{disc}_m(f'_{i+1} \circ D, g'_{i+1} \circ D) = \text{disc}_m(f_{i+1} \circ D, g_{i+1} \circ D) \leq \epsilon \quad (244)$$

and,

$$\forall i \in [n - 1]: \text{disc}_m(f'_{i+1} \circ D, g'_{i+1} \circ D) = \text{disc}_m(f_i \circ D, g_i \circ D) \leq \text{disc}_m(f_{i+1} \circ D, g_{i+1} \circ D) \leq \epsilon \quad (245)$$

By Lem. 1, since $C(\pi_i) = 2$ we have:

$$\forall i \in [n - 1]: \text{disc}_m(f'_{i+1} \circ D, g'_{i+1} \circ D) = \text{disc}_m(\pi_i \circ f_{i+1} \circ D, \pi_i \circ g_{i+1} \circ D) \leq \text{disc}_m(f_{i+1} \circ D, g_{i+1} \circ D) \leq \epsilon \quad (246)$$

Alternatively,

$$\forall i \in [n]: f'_{i+1} \circ D \sim m, \epsilon g'_{i+1} \circ D \quad (247)$$

\[\square\]

**Lemma 38** Let $N = \text{SCM}[\sigma]$ with $\sigma$ that is Leaky ReLU with parameter $a > 0$. Let $h = h_{n+1:1}$ be any function such that $C(h) = n$ and $q = q_{n+3:1}$ such that $q_{n+2:1} = F(h_{n+1:1}) \circ (\sigma \circ \text{Id}/a)$ ($F$ is the function in Eq. 175) and $C(q) = n+2$. Then, for any minimal decomposition $q = q_{n+3:1}$ there is a minimal decomposition $h = h_{n+1:1}$ such that:

$$\forall j \in [n \setminus \{1\}]: q_{n-j+3:1} \circ h \circ D_A = (-\sigma)^{-1} \circ h_{j+1} \circ D_A \quad (248)$$

**Proof** Let $q = q'_{n+3:1}$ and $h = h_{n+1:1}$ be minimal decompositions of $q$ and $h$ (resp.). By Lem. 12,

$$\forall j \in [n + 1]: q_{n-j+3:1} \circ h \circ D_A = (-\sigma)^{-1} \circ h_{j+1} \circ D_A \quad (249)$$

By minimal identifiability, for decompositions $q = q'_{n+3:1}$ we have:

$$\exists \pi_1, \ldots, \pi_n \in \text{Invariant}(N) : \quad q'_i = \pi_1 \circ q_1, \forall j = 2, \ldots, n : q'_j = \pi_j \circ q_j \circ \pi_j^{-1} \quad (250)$$

By Lem. 7,

$$\forall j \in [n + 1]: q'_{n-j+3:1} = \pi_{n-j+2} \circ q_{n-j+3:1} \quad (251)$$

Therefore, for all $j \in [n + 1]$ we have:

$$(\sigma \circ \text{Id}) \circ q'_{n-j+3:1} \circ h \circ D_A = (\sigma \circ \text{Id}) \circ \pi_{n-j+2} \circ q_{n-j+3:1} \circ h \circ D_A \quad (252)$$

We define a decomposition $h = h_{n+1:1}$ as follows:

$$h'_{1} = \pi_n \circ h_1, \forall j = 2, \ldots, n - 1: h'_j = \pi_{n-j+1} \circ h_j \circ \pi_{n-j+2}^{-1} \quad (253)$$

In particular,

$$\forall j \in [n \setminus \{1\}]: h'_{j+1} = \pi_{n-j+2} \circ h_{j+1} \quad (254)$$
But, \( \pi_{n-j+2} \) commutes with both \( \sigma \) and \(-\text{Id}\) and therefore, for all \( j \in [n] \) such that \( j \neq 1 \) we have:

\[
(\sigma \circ -\text{Id}) \circ q'_{n-j+3,1} = (\sigma \circ -\text{Id}) \circ \pi_{n-j+2} \circ q_{n-j+3,1} \circ h \circ D_A \\
= \pi_{n-j+2} \circ (\sigma \circ -\text{Id}) \circ q_{n-j+3,1} \circ h \circ D_A \\
= \pi_{n-j+2} \circ h_{j,1} \circ D_A \\
= h'_{j,1} \circ D_A
\]

(255)

\[\square\]

**Lemma 39** Let \( \mathcal{N} = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). Assume that:

\[
\bar{f} \bigotimes_{k \leq 1} f' \bigotimes_{k+2, \epsilon} D_A, \bar{f}' \bigotimes_{k+2, \epsilon} f'
\]

Then, \( \bar{f} \bigotimes_{k \leq 1-2\epsilon} f' \).

**Proof** Assume by contradiction that \( \bar{f} \bigotimes_{k \leq 1-2\epsilon} f' \). Then, there are decompositions \( f = \bar{f}_{n+1,1} \) and \( f' = \bar{f'}_{n+1,1} \) such that:

\[
\forall j \in [n] : \text{disc}_k(\bar{f}_{j+1,1} \circ D_A, \bar{f'}_{j+1,1} \circ D_A) \leq \epsilon_1 - 2\epsilon
\]

(257)

By Lem. 37, since \( \bar{f} \bigotimes_{k+2, \epsilon} f' \bigotimes_{k+2, \epsilon} f' \), there are minimal decompositions \( f = f_{n+1,1} \) and \( f' = f'_{n+1,1} \) such that:

\[
\forall j \in [n] : \text{disc}_k(f_{j+1,1} \circ D_A, f'_{j+1,1} \circ D_A), \text{disc}_k(f'_{j+1,1} \circ D_A, f_{j+1,1} \circ D_A) \leq \epsilon
\]

(258)

Since \( \bar{f} \bigotimes_{k \leq 1} f' \), there is an index \( i \in [n] \) such that:

\[
\text{disc}_k(f_{i+1,1} \circ D_A, f'_{i+1,1} \circ D_A) > \epsilon_1
\]

(259)

Therefore, by the triangle inequality,

\[
\text{disc}_k(f_{i+1,1} \circ D_A, f'_{i+1,1} \circ D_A) \leq \text{disc}_k(f_{i+1,1} \circ D_A, f'_{i+1,1} \circ D_A) + \text{disc}_k(f_{i+1,1} \circ D_A, f_{i+1,1} \circ D_A) \\
\leq \text{disc}_k(f'_{i+1,1} \circ D_A, f_{i+1,1} \circ D_A) + \text{disc}_k(f_{i+1,1} \circ D_A, f_{i+1,1} \circ D_A)
\]

(260)

\[
\leq \epsilon_1 - 2\epsilon + \epsilon + \epsilon = \epsilon_1
\]

in contradiction.  

\[\square\]
H.3 Proof of Thm. 5

Theorem 5 (Counting semantic mappings) Let $N = SCM[\sigma]$ be a NN-SCM with $\sigma$ that is a Leaky ReLU with parameter $a > 0$ and assume Assumption 1. Let $\epsilon_0$, $\epsilon_1$ and $\epsilon_2 < \epsilon_1 - 2\epsilon_0$ are three positive constants and $A = (X_A, D_A)$ and $B = (X_B, D_B)$ are two domains. Assume that $m \geq k + 2C^{\epsilon_0}_{A,B} + 5$. Then,

$$\text{Covering} \left( H_{\epsilon_0}(D_A, D_B; m_i) \right) \leq \lim_{\epsilon \to 0} \sqrt{\text{Covering} \left( DPM_2(\epsilon_0 + \epsilon) \right)}$$

(19)

Proof We denote by $\epsilon$ any positive constant such that: $\epsilon < (\epsilon_1 - 2\epsilon_0 - \epsilon_2)/4$ and $t := k + C^{\epsilon_0}_{A,B} + 3$.

We would like to find an embedding mapping:

$$G : (H_{\epsilon_0}(D_A, D_B; m_i))^2 \rightarrow DPM_{2(\epsilon_0 + \epsilon)}(k, 2C^{\epsilon_0}_{A,B} + 2)$$

(261)

Part 1: In this part, we show how to construct $G$. Let $(f, g) \in (H_{\epsilon_0}(D_A, D_B; m_i))^2$ and $f = f_{n+1,1}$ and $g = g_{n+1,1}$ are decompositions of $f$ and $g$ (resp.). Let $F$ be the function from Eq. 173 and $q_{n+1,1} = F^*(g_{n+1,1})$.

Then, by Lem. 36 there are $\bar{f} = f_{n+1,1}, \bar{g} = g_{n+1,1}$ and $\bar{q} = q_{n+1,1}$ such that:

- $\bar{q}_{n+2,1} = F(\bar{g}_{n+1,1}) \circ (\sigma \circ -\text{Id}/a)$.
- $C(\bar{f}_{n+1,1} \circ \bar{q}_{n+3,1}) = 2n + 2$.
- $\forall j \in [n+1] : \bar{q}_{n-j+1,1} \circ \bar{g} \circ D_A = (-\sigma^{-1}) \circ \bar{g}_{j,1} \circ D_A$.
- $\text{disc}_t(\bar{q} \circ D_B, D_A) \leq \epsilon_0 + \epsilon$.
- $\forall j \in [n] : \text{disc}_m(\bar{f}_{j+1,1} \circ D_A, \bar{f}_{j+1,1} \circ D_A) \leq \epsilon$.
- $\forall j \in [n] : \text{disc}_m(\bar{g}_{j+1,1} \circ D_A, \bar{g}_{j+1,1} \circ D_A) \leq \epsilon$.

We define: $G(f, g) = \bar{f}_{n+1,1} \circ \bar{q}_{n+3,1}$.

Part 2: In this part, we show that:

$$(f, g) \in (H_{\epsilon_0}(D_A, D_B; m_i))^2 \implies G(f, g) \in DPM_{2(\epsilon_0 + \epsilon)}(k, 2C^{\epsilon_0}_{A,B} + 2)$$

(262)

By Part 1, $C(\bar{f}_{n+1,1} \circ \bar{q}_{n+3,1}) = 2n + 2 = 2C^{\epsilon_0}_{A,B} + 2$.

In addition, by the first item of Lem. 2, for $D_1 := \bar{q} \circ D_B, D_2 := D_A, D_3 := \bar{f} \circ D_B, p := \bar{f} \circ D_A, t \geq k + C^{\epsilon_0}_{A,B}$ we have:

$$\text{disc}_t(\bar{f} \circ D_A, D_B) \leq \text{disc}_t(\bar{f} \circ D_A, D_B) + \text{disc}_t(\bar{q} \circ D_B, D_A)$$

(263)

Since $f \in H_{\epsilon_0}(D_A, D_B; m)$:

$$\text{disc}_t(\bar{f} \circ D_A, D_B) \leq \text{disc}_m(f \circ D_A, D_B) + \text{disc}_m(\bar{f} \circ D_A, f \circ D_A) \leq \epsilon_0 + \epsilon$$

(264)

Finally:

$$\text{disc}_k(\bar{f} \circ \bar{q} \circ D_B, D_B) \leq 2(\epsilon_0 + \epsilon)$$

(265)

We conclude that

$$G(f, g) \in DPM_{2(\epsilon_0 + \epsilon)}(k, 2C^{\epsilon_0}_{A,B} + 2)$$

(266)
Part 3: In this part, we show that $G$ is an embedding. It requires showing that

$$G(f, g) \xrightarrow{\mathcal{D}_m, \epsilon} G(f', g')$$

Assume by contradiction that $G(f, g) \xrightarrow{\mathcal{D}_m, \epsilon} G(f', g')$ and that $(f, g) \xrightarrow{\mathcal{D}_A} (f', g')$. Then, we have

$$f \xrightarrow{\mathcal{D}_A} f'$$

We denote $G(f, g) = f \circ q$ and $G(f', g') = f' \circ q'$ (see Part 1).

Assume that $f \xrightarrow{\mathcal{D}_A} f'$: By Lem. 39, $f \xrightarrow{\mathcal{D}_A} f'$. In particular, for every $f = f_{n+1:1}$ and $f' = f_{n+1:1}$, there is an index $i \in [n]$ such that:

$$\text{disc}_k(f_{i+1:1} \circ D_A, f_{i+1:1} \circ D_A) > \epsilon_1 - 2\epsilon$$

As we showed in Part 2,

$$\text{disc}_i(q \circ D_B, D_A), \text{disc}_i(q' \circ D_B, D_A) \leq \epsilon_0 + \epsilon$$

By the first item of Lem. 2, for $D_1 := D_A, D_2 := q \circ D_B, D_3 := f_{i+1:1} \circ D_A, t \geq k + C_{A, B} \geq k + C(f_{i+1:1})$, we have:

$$\text{disc}_k(f_{i+1:1} \circ D_A, f_{i+1:1} \circ D_A) \leq \text{disc}_i(f_{i+1:1} \circ q \circ D_B, f_{i+1:1} \circ D_A) + \text{disc}_i(q \circ D_B, D_A)$$

$$\leq \text{disc}_i(f_{i+1:1} \circ q \circ D_B, f_{i+1:1} \circ D_A) + \epsilon_0 + \epsilon$$

Again, by the first item of Lem. 2, for $D_1 := D_A, D_2 := q' \circ D_B, D_3 := f_{i+1:1} \circ q \circ D_B, m \geq t + C_{A, B} \geq t + C(f'_{i+1:1})$, we have:

$$\text{disc}_i(f_{i+1:1} \circ q \circ D_B, f_{i+1:1} \circ D_A) \leq \text{disc}_m(f_{i+1:1} \circ q \circ D_B, f_{i+1:1} \circ q' \circ D_B)$$

$$\leq \text{disc}_m(f_{i+1:1} \circ q \circ D_B, f_{i+1:1} \circ q' \circ D_B) + \epsilon_0 + \epsilon$$

Therefore, we conclude that:

$$\epsilon_1 - 2\epsilon_0 - 4\epsilon < \text{disc}_m(f_{i+1:1} \circ q \circ D_B, f_{i+1:1} \circ q' \circ D_B)$$

Alternatively, for any minimal decompositions $f \circ \bar{q} = f_{n+1:1} \circ \bar{q}_{n+1:3:1}$ and $f' \circ \bar{q}' = f_{n+1:1} \circ \bar{q}'_{n+1:3:1}$ there are right partial functions $f_{i+1:1} \circ \bar{q}_{n+3:1}$ and $f'_{i+1:1} \circ \bar{q}'_{n+3:1}$ such that:

$$\epsilon_2 \leq \epsilon_1 - 2\epsilon_0 - 4\epsilon < \text{disc}_m(f_{i+1:1} \circ \bar{q}_{n+3:1} \circ D_B, f'_{i+1:1} \circ \bar{q}'_{n+3:1} \circ D_B)$$

in contradiction to $F(f, g) \xrightarrow{\mathcal{D}_m, \epsilon} F(f', g')$.

Assume that $g \xrightarrow{\mathcal{D}_A} g'$: Let $\bar{q} = \bar{q}_{n+3:1}$ and $\bar{q}' = \bar{q}'_{n+3:1}$ be any two minimal decompositions of $\bar{q}$ and $\bar{q}'$ (resp.). Then, by Lem. 38, there are minimal decompositions $\bar{g} = \bar{g}_{n+1:1}$ and $\bar{g}' = \bar{g}'_{n+1:1}$ such that:

$$\forall j \in [n] \setminus \{1\}: \bar{g}_{n-j+3:1} \circ \bar{g} \circ D_A = (-\sigma^{-1}) \circ \bar{g}_{j+1:1} \circ D_A$$

and $\bar{g}'_{n-j+3:1} \circ \bar{g}' \circ D_A = (-\sigma^{-1}) \circ \bar{g}'_{j+1:1} \circ D_A$
By Lem. 39, since $\bar{g} \sim_D m, \epsilon \bar{g}$ and $\bar{g}' \sim_D m, \epsilon \bar{g}'$, we have: $\bar{g} \sim_D m, \epsilon \bar{g}'$. In particular, there is an index $i \in [n + 1]$ such that:

$$\text{disc}_k (\bar{g}_{i+1} \circ D_A, \bar{g}'_{i+1} \circ D_A) > \epsilon_1 - 2\epsilon$$

(276)

and,

$$\text{disc}_k (\bar{g}_{i+1} \circ D_A, \bar{g}'_{i+1} \circ D_A) = 0 \leq \epsilon_1 - 2\epsilon$$

(277)

and,

$$\text{disc}_k (\bar{g}_{n+1+1} \circ D_A, \bar{g}'_{n+1+1} \circ D_A) = \text{disc}_k (\bar{g} \circ D_A, \bar{g}' \circ D_A)$$

$$\leq \text{disc}_k (\bar{g} \circ D_A, D_B) + \text{disc}_k (D_B, \bar{g}' \circ D_A)$$

$$\leq \text{disc}_k (\bar{g} \circ D_A, D_B) + \text{disc}_k (D_B, \bar{g}' \circ D_A) + \text{disc}_k (\bar{g} \circ D_A, g \circ D_A) + \text{disc}_k (g' \circ D_A, \bar{g}' \circ D_A)$$

$$\leq 2(\epsilon_0 + \epsilon) \leq \epsilon_1 - 2\epsilon$$

Therefore, $i \neq n + 1, 1$. Thus, $i \in [n] \setminus \{1\}$ and:

$$\text{disc}_{k+1} ((-\sigma^{-1}) \circ \bar{g}_{i+1} \circ D_A, (-\sigma^{-1}) \circ \bar{g}'_{i+1} \circ D_A)$$

$$= \text{disc}_{k+1} ((-\sigma^{-1}) \circ \bar{g}_{n-i+3:1} \circ D_A, (-\sigma^{-1}) \circ \bar{g}'_{n-i+3:1} \circ D_A)$$

(279)

By Lem. 1, for $p := (\sigma \circ \text{Id})$ of complexity 1 we have:

$$\epsilon_1 - 2\epsilon < \text{disc}_k (\bar{g}_{i+1} \circ D_A, \bar{g}'_{i+1} \circ D_A)$$

$$\leq \epsilon_1 - 2\epsilon$$

(280)

In addition, by Lem. 2, for $D_1 := \bar{g} \circ D_A, D_2 := D_B, D_3 := \bar{g}' \circ D_A, t \geq (k + 1) + (C^m_{A,B} + 2) \geq (k + 1) + C(\bar{q}_{n-i+3:1})$, we have:

$$\text{disc}_{k+1} (\bar{q}_{n-i+3:1} \circ D_B, \bar{q}'_{n-i+3:1} \circ D_A)$$

$$\leq \text{disc}_m (\bar{q}_{n-i+3:1} \circ D_B, \bar{q}'_{n-i+3:1} \circ D_A + \text{disc}_m (\bar{g} \circ D_A, D_B)$$

(281)

Again, by Lem. 2, for $D_1 := \bar{g}' \circ D_A, D_2 := D_B, D_3 := \bar{q}_{n-i+3:1} \circ D_B, m \geq t + (C^m_{A,B} + 2) \geq t + C(\bar{q}_{n-i+3:1})$, we have:

$$\text{disc}_m (\bar{q}_{n-i+3:1} \circ D_B, \bar{q}'_{n-i+3:1} \circ D_A + \text{disc}_m (\bar{g} \circ D_B, D_A)$$

(282)

Finally,

$$\epsilon_1 - 2\epsilon < \text{disc}_k (\bar{g}_{i+1} \circ D_A, \bar{g}'_{i+1} \circ D_A)$$

$$\leq \text{disc}_m (\bar{q}_{n-i+3:1} \circ D_B, \bar{q}'_{n-i+3:1} \circ D_B) + \epsilon_0 + \epsilon$$

(283)

In particular,

$$\epsilon_2 \leq \epsilon_1 - 2\epsilon_0 - 4\epsilon < \text{disc}_m (\bar{q}_{n-i+3:1} \circ D_B, \bar{q}'_{n-i+3:1} \circ D_B)$$

(284)

Alternatively, for any minimal decompositions $\bar{f} \circ \bar{q} = \bar{f}_{n+1+1} \circ \bar{q}_{n+3:1}$ and $\bar{f}' \circ \bar{q}' = \bar{f}'_{n+1+1} \circ \bar{q}'_{n+3:1}$ there are right partial functions $\bar{q}_{n-i+3:1}$ and $\bar{q}'_{n-i+3:1}$ such that:

$$\epsilon_2 \leq \epsilon_1 - 2\epsilon_0 - 4\epsilon < \text{disc}_m (\bar{q}_{n-i+3:1} \circ D_B, \bar{q}'_{n-i+3:1} \circ D_B)$$

(285)

in contradiction to $F(f, g) \sim_D m, \epsilon_2 F'(f', g')$. 47
Part 3: Finally, by Lem. 32 and Lem. 30,

\[
\text{Covering} \left( H_{\epsilon_0}(D_A, D_B; m), \frac{D_A}{k, \epsilon_1} \right)^2 = \text{Covering} \left( (H_{\epsilon_0}(D_A, D_B; m))^2, \frac{D_A}{k, \epsilon_1} \right)^2 \leq \text{Covering} \left( \text{DPM}_{2(\epsilon_0 + \epsilon)} \left( k, 2C_{A,B}^{\epsilon_0} + 2 \right), \frac{D_B}{m, \epsilon_2} \right) \quad (286)
\]

Alternatively, for all \( \epsilon_0, \epsilon_1, \epsilon_2 \) such that \( \epsilon < (\epsilon_1 - 2\epsilon_0 - \epsilon_2) / 4 \),

\[
\text{Covering} \left( H_{\epsilon_0}(D_A, D_B; m), \frac{D_A}{k, \epsilon_1} \right)^2 \leq \text{Covering} \left( \text{DPM}_{2(\epsilon_0 + \epsilon)} \left( k, 2C_{A,B}^{\epsilon_0} + 2 \right), \frac{D_B}{m, \epsilon_2} \right) \quad (287)
\]

By Lem. 33, the function \( \text{Covering} \left( \text{DPM}_{2(\epsilon_0 + \epsilon)} \left( k, 2C_{A,B}^{\epsilon_0} + 2 \right), \frac{D_B}{m, \epsilon_2} \right) \) is monotonically decreasing as \( \epsilon \) tends to 0 and is lower bounded by \( \text{Covering} \left( \text{DPM}_{2\epsilon_0} \left( k, 2C_{A,B}^{\epsilon_0} + 2 \right), \frac{D_B}{m, \epsilon_2} \right) \). Thus, the limit

\[
\lim_{\epsilon \to 0} \text{Covering} \left( \text{DPM}_{2(\epsilon_0 + \epsilon)} \left( k, 2C_{A,B}^{\epsilon_0} + 2 \right), \frac{D_B}{m, \epsilon_2} \right) \quad (288)
\]

exists and

\[
\text{Covering} \left( H_{\epsilon_0}(D_A, D_B; m), \frac{D_A}{k, \epsilon_1} \right) \leq \sqrt{\lim_{\epsilon \to 0} \text{Covering} \left( \text{DPM}_{2(\epsilon_0 + \epsilon)} \left( k, 2C_{A,B}^{\epsilon_0} + 2 \right), \frac{D_B}{m, \epsilon_2} \right)} \quad (289)
\]

\[
= \lim_{\epsilon \to 0} \sqrt{\text{Covering} \left( \text{DPM}_{2(\epsilon_0 + \epsilon)} \left( k, 2C_{A,B}^{\epsilon_0} + 2 \right), \frac{D_B}{m, \epsilon_2} \right)} \quad (289)
\]

\[
= \lim_{\epsilon \to 0} \sqrt{\text{Covering} \left( \text{DPM}_{2(\epsilon_0 + \epsilon)} \left( k, 2C_{A,B}^{\epsilon_0} + 2 \right), \frac{D_B}{m, \epsilon_2} \right)}
\]
Appendix I. Proof of Thms. 6 and 7

Lemma 40 Let $N = SCM[C]$, $\epsilon_0, \epsilon_1, \epsilon_2 > 0$ are three constants, $A = (X_A, DA)$ and $B = (X_B, DB)$ are two domains and $y \in H_\alpha(D_A, D_B; m)$ such that $y = y_{n+1}$ is a minimal decomposition of $y$. Assume that $k \geq \max \left\{ E_{A,B}^\alpha, C_{A,B}^\alpha \right\}$, $m \geq k + 2C_{A,B}^\alpha$ and that $C_{A,B}^\alpha = C_{A,B}^\alpha + 2k + 2\epsilon_1$. Let $D_1$ and $D_2$ be two distributions such that:

$$\text{disc}_m(y_{i+1:1} \circ D_A, D_I) \leq \epsilon_1 \quad \text{and} \quad \text{disc}_m(y_{j+1:1} \circ D_A, D_J) \leq \epsilon_2$$ \hfill (290)

Then, $y_{j+1:i+1} \in H_{\epsilon_1 + \epsilon_2}(D_1, D_J; k)$.

Proof Assume by contradiction that $y_{j+1:i+1} \notin H_{\epsilon_1 + \epsilon_2}(D_1, D_J; k)$. Then, exactly one of the following options hold:

- There is a function $u$ such that $C(u) < C(y_{j+1:i+1}) = j - i$ and $\text{disc}_k(u \circ D_I, D_J) \leq \epsilon_1 + \epsilon_2$.
- $\text{disc}_k(y_{j+1:i+1} \circ D_I, D_J) \geq \epsilon_1 + \epsilon_2$.

If the second option holds, then, by the triangle inequality,

$$\epsilon_1 + \epsilon_2 < \text{disc}_k(y_{j+1:i+1} \circ D_I, D_J) \leq \text{disc}_k(y_{j+1:i+1} \circ D_I, y_{j+1:i} \circ D_A) + \text{disc}_k(D_J, y_{j+1:i} \circ D_A) + \epsilon_2$$

In addition, by the first part of Lem. 2, for $D_1 :\leftarrow D_1, D_2 :\leftarrow y_{i+1:1} \circ D_A, D_3 :\leftarrow y_{j+1:1} \circ D_A, p :\leftarrow y_{j+1:i+1}$ and $m \geq k + j - i$,

$$\text{disc}_k(y_{j+1:i+1} \circ D_I, y_{j+1:i} \circ D_A) \leq \text{disc}_m(y_{j+1:i} \circ D_A, y_{j+1:i} \circ D_A) + \epsilon_1 \quad \text{and} \quad \text{disc}_m(y_{j+1:i} \circ D_A, D_I) \leq \epsilon_1 \quad \text{and} \quad \text{disc}_m(y_{j+1:i} \circ D_A, D_I) \leq \epsilon_1$$

Therefore, $\epsilon_1 + \epsilon_2 < \text{disc}_k(y_{j+1:i+1} \circ D_I, D_J) \leq \epsilon_1 + \epsilon_2$ in contradiction. Thus, we conclude that the first option must hold.

We denote $t = k + C_{A,B}^\alpha$. We note that since $m \geq t \geq k \geq E_{A,B}^\alpha$, by Lem. 18, we have:

$$y \in H_{\alpha}(D_A, D_B; m) \subset H_{\alpha}(D_A, D_B; t)$$ \hfill (293)

By the triangle inequality,

$$\text{disc}_t(u \circ y_{i+1:1} \circ D_A, y_{j+1:1} \circ D_A) \leq \text{disc}_m(y_{j+1:1} \circ D_A, D_J) + \text{disc}_m(u \circ y_{i+1:1} \circ D_A, D_J) \leq \text{disc}_m(u \circ y_{i+1:1} \circ D_A, D_J) + \epsilon_2 \quad \text{and} \quad \text{disc}_m(y_{i+1:1} \circ D_A, D_I) \leq \epsilon_1$$

By the first item of Lem. 2; for $D_1 :\leftarrow y_{i+1:1} \circ D_A, D_2 :\leftarrow D_1, D_3 :\leftarrow D_J, p :\leftarrow u$, and $m \geq t + C_{A,B}^\alpha \geq t + C(u)$, we have:

$$\text{disc}_t(u \circ y_{i+1:1} \circ D_A, D_J) \leq \text{disc}_m(u \circ D_I, D_J) + \text{disc}_m(y_{i+1:1} \circ D_A, D_I) \leq 2\epsilon_1 + \epsilon_2$$

Therefore,

$$\text{disc}_t(u \circ y_{i+1:1} \circ D_A, y_{j+1:1} \circ D_A) \leq 2(\epsilon_1 + \epsilon_2)$$

By the first item of Lem. 2; for $D_1 :\leftarrow u \circ y_{i+1:1} \circ D_A, D_2 :\leftarrow y_{j+1:1} \circ D_A, D_3 :\leftarrow D_B, p :\leftarrow y_{n+1:1} \circ D_A$, and $t = k + C_{A,B}^\alpha \geq k + C(y_{n+1:1})$, we have:

$$\text{disc}_k(y_{n+1:1} \circ u \circ y_{i+1:1} \circ D_A, D_B) \leq \text{disc}_k(y_{n+1:1} \circ y_{j+1:1} \circ D_A, D_B) + \text{disc}_k(u \circ y_{i+1:1} \circ D_A, D_J) \leq \text{disc}_t(y \circ D_A, D_B) + 2(\epsilon_1 + \epsilon_2) \leq \epsilon_0 + 2(\epsilon_1 + \epsilon_2)$$
On the other hand,
\[ C(y_{n+1,j+1} \circ u \circ y_{i+1:1}) \leq n - j + C(u) + i < n - j + (j - i) + i = n = C(y) \] (298)

Thus, we found a function \( g = y_{n+1,j+1} \circ u \circ y_{i+1:1} \) such that \( C(g) < C(y) \) and \( \text{disc}(g \circ D_A, D_B) \leq \epsilon_0 + 2(\epsilon_1 + \epsilon_2) \). Since \( k \geq E_{A,B}^{\epsilon_0+2\epsilon_1+2\epsilon_2} \), we have:

\[ C_{A,B}^{k,\epsilon_0+2\epsilon_1+2\epsilon_2} = C(g) < C(y) = C_{A,B}^{\epsilon_0} \] (299)

in contradiction to \( C_{A,B}^{\epsilon_0+2\epsilon_1+2\epsilon_2} = C_{A,B}^{\epsilon_0} \).

\[ \square \]

**Proof of Thm. 6**

**Theorem 6** Let \( N = \text{SCM}[\sigma] \) with \( \sigma \) that is Leaky ReLU with parameter \( a > 0 \). Let \( A = (X_A, D_A) \) and \( B = (X_B, D_B) \) be two domains and \( D_Z \neq D_A, D_B \) is a \((m, \epsilon_0, \epsilon_1, \epsilon_0 + \epsilon_1)\)-shared semantic distribution between \( A \) and \( B \). Let \( k = \max \{E_{\epsilon_0}^{A}, E_{\epsilon_0+4\epsilon_1}^{B,A}, E_{\epsilon_0}^{Z,A,Z} + 1\} \) and \( m \geq k + 3C_{A,B}^{\epsilon_0+\epsilon_1} + 4 \). Assume that \( C_{A,B}^{\epsilon_0+3\epsilon_1} = C_{A,B}^{\epsilon_0+\epsilon_1} + 2 \). Then, \( (-\sigma^{-1}) \circ D_Z \) is a \((k, \epsilon_1, \epsilon_0, \epsilon_0 + \epsilon_1)\)-shared semantic distribution between \( B \) and \( A \).

**Proof** Let \( y_{AB} \in Z(D_A, D_Z, D_B; m, \epsilon_0, \epsilon_1) \cap H_{\epsilon_0+\epsilon_1}(D_A, D_B; m) \). Assume that \( y_{AB} = p_{n+1:1} \) is a minimal decomposition of \( y_{AB} \), such that \( \text{disc}_m(p_{n+1:1} \circ D_A, D_Z) \leq \epsilon_0 \) and \( \text{disc}_m(D_B, p_{n+1:1} \circ D_Z) \leq \epsilon_1 \) for some \( i \in \{1, ..., n - 1\} \). We denote \( y_{BA} = y_{AB}^{-1}, t := m - C_{A,B}^{\epsilon_0+\epsilon_1} - 2 \).

**Part 1:** In this part, we show that

\[ y_{BA} \in H_{\epsilon_0+\epsilon_1}(D_B, D_A; t) \subset H_{\epsilon_0+\epsilon_1}(D_B, D_A; k) \] (300)

By the third item of Lem. 2, for: \( D_1 \leftarrow D_A, D_2 \leftarrow D_B, h \leftarrow y_{AB}, h^{-1} \leftarrow y_{BA} \) and \( m = t + C_{A,B}^{\epsilon_0+\epsilon_1} + 2 \), we have:

\[ \text{disc}_t(D_A, y_{BA} \circ D_B) \leq \text{disc}_m(y_{AB} \circ D_A, D_B) \leq \epsilon_0 + \epsilon_1 \] (301)

Since \( m \geq t \geq k \geq E_{B,A}^{\epsilon_0+\epsilon_1} \), we have:

\[ C_{A,B}^{\epsilon_0+\epsilon_1} = C_{B,A}^{\epsilon_0+\epsilon_1} = C_{A,B}^{\epsilon_0+\epsilon_1} + 2 = C_{A,B}^{\epsilon_0+\epsilon_1} + 2 = n + 2 \] (302)

Therefore, \( C(y_{BA}) + 2 = C_{A,B}^{\epsilon_0+\epsilon_1} \) and by Thm. 1, \( C(y_{BA}) = C(y_{AB}) + 2 \). In particular,

\[ y_{BA} \in H_{\epsilon_0+\epsilon_1}(D_B, D_A; t) \subset H_{\epsilon_0+\epsilon_1}(D_B, D_A; k) \] (303)

In addition, since \( C(y_{BA}) = C(y_{AB}) + 2 \), the following is a minimal decomposition of \( y_{BA} \):

\[ y_{BA} = F^+(p_{n+1:1}) \] (304)

Where, \( F^+ \) is the function defined in Eq. 173 (see the proof of Thm. 1).

**Part 2:** In this part, we show that:

\[ \text{disc}_t((-\sigma^{-1}) \circ D_Z, [F^+(p_{n+1:1})]_{n-i+2:1} \circ D_B) \leq \epsilon_1 \] (305)

and

\[ \text{disc}_t(D_A, [F^+(p_{n+1:1})]_{n+3:n-i+2} \circ (-\sigma^{-1}) \circ D_Z) \leq \epsilon_0 \] (306)

By Eq. 64, for \( i \in [n] \),

\[ [F^+(p_{n+1:1})]_{n-i+2:1} = (-\sigma^{-1}) \circ p_{n+1:i+1} \] (306)
In particular, for $i \in [n],$

$$[F^+(p_{n+1:1})]_{n+3:n-i+2} \circ (-\sigma^{-1}) = p_{i+1:1}^{-1}$$ (307)

By Lem. 12, for $i \in [n],$

$$[F^+(p_{n+1:1})]_{n-i+2:1} \circ y_{AB} \circ D_A = (-\sigma^{-1}) \circ p_{i+1:1} \circ D_A$$ (308)

By Lem. 1, for $p := (-\sigma^{-1})$ of complexity 1,

$$\text{disc}_t \left( (-\sigma^{-1}) \circ D_Z, [F^+(p_{n+1:1})]_{n-i+2:1} \circ D_B \right) = \text{disc}_t \left( (-\sigma^{-1}) \circ D_Z, (-\sigma^{-1}) \circ p_{n+1:1+i+1} \circ D_B \right) \leq \text{disc}_{t+1} \left( D_Z, p_{n+1:1+i+1}^{-1} \circ D_B \right) \tag{309}$$

By the third item of Lem. 1, for $h := p_{i+1:1}$ and $m = (t + 1) + C_{A,B}^{\epsilon_0+\epsilon_1} + 2 = (t + 1) + C(y_{AB}) + 2 = (t + 1) + C(p_{n+1:1+i+1})$ and by Eq. 307 we have:

$$\text{disc}_t \left( (-\sigma^{-1}) \circ D_Z, [F^+(p_{n+1:1})]_{n-i+2:1} \circ D_B \right) \leq \text{disc}_{t+1} \left( D_Z, p_{n+1:1+i+1}^{-1} \circ D_B \right) \leq \text{disc}_m \left( p_{n+1:1+i+1} \circ D_Z, D_B \right) \leq \epsilon_1$$ (310)

Again, by Eq. 307 and by the third item of Lem. 1, for $h := p_{i+1:1}$ and $m \geq t + C_{A,B}^{\epsilon_0+\epsilon_1} + 2 \geq t + C(y_{AB}) + 2 \geq t + C(p_{n+1:1})$ we have:

$$\text{disc}_t \left( [F^+(p_{n+1:1})]_{n+3:n-i+2} \circ (-\sigma^{-1}) \circ D_Z, D_A \right) = \text{disc}_t \left( (p_{i+1:1})^{-1} \circ D_Z, D_A \right) \leq \text{disc}_m \left( D_Z, p_{i+1:1} \circ D_B \right) \leq \epsilon_0$$ (311)

**Part 3:** In this part, we show that:

$$[F^+(p_{n+1:1})]_{n-i+2:1} \in H_{\epsilon_1}(D_B, (-\sigma^{-1}) \circ D_Z; k)$$

and

$$[F^+(p_{n+1:1})]_{n+3:n-i+2} \in H_{\epsilon_0}((-\sigma^{-1}) \circ D_Z, D_A; k)$$ (312)

The following conditions hold:

- $t \geq k + 2C_{A,B}^{\epsilon_0+\epsilon_1} + 4 = k + 2C_{B,A}^{\epsilon_0+\epsilon_1}$.
- $k \geq \max \left\{ E_{B,A}^{\epsilon_0+\epsilon_1}, E_{B,A}^{4\epsilon_0+4\epsilon_1} \right\}$.
- $C_{A,B}^{\epsilon_0+\epsilon_1} = C_{B,A}^{3\epsilon_0+\epsilon_1} = C_{B,A}^{\epsilon_0+3\epsilon_1}$.
- $\text{disc}_t ([F^+(p_{n+1:1})]_{n-i+2:1} \circ D_B, (-\sigma^{-1}) \circ D_Z) \leq \epsilon_1$.
- $\text{disc}_t ([F^+(p_{n+1:1})]_{n+3:n-i+2} \circ (-\sigma^{-1}) \circ D_Z, D_A) \leq \epsilon_0$.
- $y_{BA} = F^+(p_{n+1:1}) \in H_{\epsilon_0+\epsilon_1}(D_B, D_A; t)$.

Therefore, by Lem. 40,

$$[F^+(p_{n+1:1})]_{n-i+2:1} \in H_{\epsilon_1}(D_B, (-\sigma^{-1}) \circ D_Z; k)$$ (313)

and

$$[F^+(p_{n+1:1})]_{n+3:n-i+2} \in H_{\epsilon_0}((-\sigma^{-1}) \circ D_Z, D_A; k)$$ (314)

We conclude that:

$$\mathcal{Z}(D_B, (-\sigma^{-1}) \circ D_Z, D_A; k; \epsilon_1, \epsilon_0) \cap H_{\epsilon_0+\epsilon_1}(D_B, D_A; k)$$ (315)
Finally, let $f \in \mathcal{H}_{\epsilon_1}(D_B, (-\sigma^{-1}) \circ D_Z; k)$ and $g \in \mathcal{H}_{\epsilon_0}((-\sigma^{-1}) \circ D_Z, D_B; k)$ and $s = k - C_{Z,B}^{\epsilon_0 + \epsilon_1} - 1$, then:
\[
\operatorname{disc}_s(g \circ f \circ D_B, D_A) \leq \operatorname{disc}_k(f \circ D_A, (-\sigma^{-1}) \circ D_Z) + \operatorname{disc}_k(g \circ (-\sigma^{-1}) \circ D_Z, D_A) \leq \epsilon_0 + \epsilon_1 \tag{316}
\]

Since $s \geq E_{B,A}^{\epsilon_0 + \epsilon_1}$, we have: $C(g \circ f) \geq C_{B,A}^{\epsilon_0 + \epsilon_1}$. Therefore,
\[
C_{B,A}^{\epsilon_0 + \epsilon_1} \leq C(g \circ f) \leq C(g) + C(f) \leq C([F^+(p_{n+1:1}^n)]_{n+3:n-i+2}) + C([F^+(p_{n+1:1}^n)]_{n-i+1})
\]
\[
= C(y_{BA}) \leq C_{B,A}^{k,\epsilon_0 + \epsilon_1} = C_{B,A}^{\epsilon_0 + \epsilon_1} \tag{317}
\]

In particular, $g \not\preceq f$. Alternatively, $(-\sigma^{-1}) \circ D_Z$ is a $(k, \epsilon_1, \epsilon_0, \epsilon_1)$-shared semantic distribution between $B$ and $A$.

\[\square\]

**Proof of Thm. 7**

**Definition 15** Let $A = (X_A, D_A)$ and $B = (X_B, D_B)$ be two domains and $D_Z$ a distribution. We say that $D_Z$ is a $(m, \epsilon_0, \epsilon_1)$-shared irreducible distribution between $A$ and $B$, if for all $y_B \in \mathcal{H}_{\epsilon_0}(D_Z, D_B; m)$ and $y_A^{-1} \in \mathcal{H}_{\epsilon_0}(D_A, D_Z; m)$ we have: $y_B \not\preceq y_A^{-1}$. If $\epsilon_0 = \epsilon_1$ we write $(m, \epsilon_0)$ for short.

**Lemma 41 (Alignment)** Let $\mathcal{N} = \text{SCM}[\sigma]$ be a NN-SCM with $\sigma$ that is Leaky ReLU with parameter $a > 0$ and $A = (X_A, D_A)$ and $B = (X_B, D_B)$ are two domains and $D_Z$ a distribution. Let $k \geq \max\left\{E_{A,Z}^0, E_{Z,B}^0\right\}$ and $m \geq k + C_{Z,B}^0$. Let $D_Z$ is a $(m, \epsilon_0)$-shared irreducible distribution between $A$ and $B$. Let $h$ be a function that satisfies:

- $C(h) \leq C_{A,Z}^{\epsilon_0} + C_{Z,B}^{\epsilon_0}$.
- $\operatorname{disc}_m(h \circ D_A, D_B) \leq \epsilon_1$.

Then, one of the following holds:

- $h \in \mathcal{H}_{\epsilon_0 + \epsilon_1}(D_Z, D_B; k, C_{Z,B}^{\epsilon_0}) \circ \mathcal{H}_{\epsilon_0}(D_A, D_Z; m)$.
- For every $y \in \mathcal{Z}(D_A, D_Z, D_B; m, \epsilon_0)$ we have: $C(h \| y) \geq 2C_{Z,B}^{\epsilon_0} - 9$.

**Proof** We denote $i = C_{A,Z}^{\epsilon_0} + C_{Z,B}^{\epsilon_0}$, $j = 2C_{Z,B}^{\epsilon_0} - 9$ and:
\[
H(D_A, D_B; i, j, \epsilon_0, \epsilon_1) := \left\{ h \mid \exists y \in \mathcal{Z}(D_A, D_Z, D_B; m, \epsilon_0) \quad \operatorname{disc}_m(h \circ D_A, D_B) \leq \epsilon_1, C(h) \leq i \text{ and } C(h \| y) < j \right\} \tag{318}
\]

We would like to show that if $C(h) \leq C_{A,Z}^{\epsilon_0} + C_{Z,B}^{\epsilon_0}$ and $\operatorname{disc}_m(h \circ D_A, D_B) \leq \epsilon_1$, then exactly one of the following holds:

- $h \in \mathcal{H}_{\epsilon_0 + \epsilon_1}(D_Z, D_B; k, C_{Z,B}^{\epsilon_0}) \circ \mathcal{H}_{\epsilon_0}(D_A, D_Z; m)$.
- $h \notin H(D_A, D_B; i, j, \epsilon_0, \epsilon_1)$.

If the second option holds; then, for every $y \in \mathcal{Z}(D_A, D_Z, D_B; m, \epsilon_0)$ we have $C(h \| y) = C(h \circ y^{-1}) \geq 2C_{Z,B}^{\epsilon_0} - 9$. Therefore, we assume that $h \in H(D_A, D_B; i, j, \epsilon_0, \epsilon_1)$ and prove that the first option holds. We denote by $y_{YAB} \in \mathcal{Z}(D_A, D_Z, D_B; m, \epsilon_0)$ the function, $y$, that corresponds to $h$ (see Eq. 318). In addition, we denote $y_{A}^{-1} \in \mathcal{H}_{\epsilon_0}(D_A, D_Z; m)$ and $y_{B} \in \mathcal{H}_{\epsilon_0}(D_Z, D_B; m)$ such that $y_{YAB} = y_B \circ y_A^{-1}$ and $\Pi := h \circ y_{AB}$. Since $D_Z$ is a $(m, \epsilon_0)$-shared irreducible distribution, $y_B \not\preceq y_A^{-1}$.
Part 1: We would like to prove that $C(\Pi \circ y_B) \leq C(y_B)$. By Thm. 16, there are decompositions $\Pi = a \circ b$ and $y_B = b^{-1} \circ c$ such that: $b^{-1} \not\preceq c$, $a \not\preceq c$ and $C(\Pi) \geq C(a) + C(b) - I - 1$.

Case 1: Assume that $C(c) \geq I$. By Thm. 4, for $f : \leftarrow a, g : \leftarrow c, h : \leftarrow y_A^{-1}$, we have,

$$C(\Pi \circ y_{AB}) = C(a) + C(c) + C(y_A^{-1})$$

$$= C(a \circ c) + C(y_A^{-1}) = C(\Pi \circ y_B) + C(y_A^{-1})$$ (319)

In particular,

$$C(\Pi \circ y_{AB}) - C(y_B) = C(\Pi \circ y_{AB}) - C(y_{AB}) \geq C(\Pi \circ y_B) - C(y_B)$$ (320)

But, $C(h) \leq C_{A,Z}^\alpha + C_{B,z}^\alpha = C(y_B)$ and therefore, $C(\Pi \circ y_B) \leq C(y_B)$.

Case 2: Assume that $C(c) < I$. If $C(y_B) < C(\Pi \circ y_B)$ then we have: $\Pi \circ y_B = a \circ c$ such that $C(a \circ c) = C(a) + C(c) > C(y_B)$ and since $C(c) < I$ we have:

$$C(a) \geq C(y_B) + 1 - C(c) \geq C(y_B) + 1 - (I - 1) = C(y_B) - I + 2$$ (321)

In addition,

$$C(\Pi) \geq C(a) + C(b) - I - 1 \geq C(y_B) + C(b) - 2I + 1$$ (322)

Since $C(y_B) = C(b^{-1}) + C(c)$, by Thm. 1, we have:

$$C(b) \geq C(b^{-1}) - 2 = C(y_B) - C(c) - 2 \geq C(y_B) - I - 1$$ (323)

We conclude that:

$$C(\Pi) \geq 2C(y_B) - 3I$$ (324)

Since $\sigma$ is a Leaky ReLU with parameter $a > 0$, we have, $I \leq 3$, and, therefore, we can replace every instance of $I$ in the proof with 3. In contradiction to $C(\Pi) < 2C(y_B) - 9$. In particular, $C(\Pi \circ y_B) \leq C(y_B)$.

Part 2: In this part, we show that:

$$\text{disc}_k(\Pi \circ y_B \circ D_Z, D_B) \leq \epsilon_0 + \epsilon_1$$ (325)

By the first item of Lem. 2; for $D_1 : \leftarrow D_Z, D_2 : \leftarrow y_A^{-1} \circ D_A, D_3 : \leftarrow D_B, p : \leftarrow \Pi \circ y_B$ and $m \geq k + C_{A,Z}^\alpha \geq k + C(y_B) \geq k + C(\Pi \circ y_B)$, we have:

$$\text{disc}_k(\Pi \circ y_B \circ D_Z, D_B) \leq \text{disc}_m(y_A^{-1} \circ D_Z, D_B) + \text{disc}_m(\Pi \circ y_B \circ y_A^{-1} \circ D_A, D_B)$$

$$= \text{disc}_m(y_A^{-1} \circ D_Z, D_B) + \text{disc}_m(h \circ D_A, D_B) \leq \epsilon_0 + \epsilon_1$$ (326)

Part 3: As we mentioned earlier, one can represent $h = \Pi \circ y_{AB} = (\Pi \circ y_B) \circ y_A^{-1}$. In Parts 1 and 2, we showed that

$$\Pi \circ y_B \in H_{\epsilon_0 + \epsilon_1}(D_Z, D_B; k, C_{A,Z}^\alpha)$$ (327)

since $C(\Pi \circ y_B) \leq C(y_B) = C_{A,Z}^\alpha$ and $\text{disc}_k(\Pi \circ y_B \circ D_Z, D_B) \leq \epsilon_0 + \epsilon_1$. In addition, by definition, $y_A^{-1} \in H_{\epsilon_0}(D_A, D_Z; m)$. Therefore, we conclude that:

$$h \in H_{\epsilon_0 + \epsilon_1}(D_Z, D_B; k, C_{A,Z}^\alpha) \circ H_{\epsilon_0}(D_A, D_Z; m)$$ (328)

Lemma 42 Let $\mathcal{N} = \text{SCM}[\sigma]$ be a NN-SCM with $\sigma$ that is Leaky ReLU with parameter $a > 0$, $A = (X_A, D_A)$ and $B = (X_B, D_B)$ are two domains and $D_Z$ is a distribution. $m \geq \max \{ F_{A,B}^{\alpha + \epsilon_1}, F_{A,Z}^{\alpha - \epsilon_1}, F_{B,Z}^{\alpha - \epsilon_1} \}$ and assume that $D_Z$ is a $(m, \epsilon_0, \epsilon_1, \epsilon_0 + \epsilon_1)$-shared semantic distribution between $A$ and $B$. Let $f \in H_{\epsilon_0}(D_A, D_Z; m)$ and $h = g \circ f$ is a function such that:
• \( \text{disc}_m(h \circ D_A, D_B) \leq \epsilon_0 + \epsilon_1 \).
• \( C(g) \leq C_{Z,B}^{\epsilon_1} \).

Then, \( g \not\succ f \).

**Proof** Since \( m \geq \max \{ E_{A,B}^{\epsilon_0 + \epsilon_1}, E_{A,Z}^{\epsilon_0 + \epsilon_1}, E_{Z,B}^{\epsilon_0 + \epsilon_1} \} \) and \( D_Z \) being a \((m, \epsilon_0, \epsilon_1, \epsilon_0 + \epsilon_1)\)-shared semantic distribution between \( A \) and \( B \), by Lem. 19, we have:

\[
C_{Z,B}^{\epsilon_1} + C_{A,Z}^{\epsilon_0} = C_{A,B}^{\epsilon_1} + C_{Z,B}^{\epsilon_0 + \epsilon_1} \tag{329}
\]

On the other hand, we assumed that \( C(g) \leq C_{Z,B}^{\epsilon_1} \) and that \( f \in H_{\epsilon_0}(D_A, D_Z; m) \). In particular,

\[
C(h) \leq C(g) + C(f) \leq C_{Z,B}^{\epsilon_1} + C_{A,Z}^{\epsilon_0} \tag{330}
\]

By the definition of \( C_{A,B}^{m,\epsilon_0 + \epsilon_1} \) as the minimal complexity of a function \( h' \) satisfying \( \text{disc}_{\epsilon_0}(h' \circ D_A, D_B) \leq \epsilon_0 + \epsilon_1 \), we have:

\[
C_{A,B}^{\epsilon_0 + \epsilon_1} = C_{A,B}^{m,\epsilon_0 + \epsilon_1} \leq C(h) \tag{331}
\]

Therefore, \( C(h) = C(g) + C(f) \) and \( g \not\succ f \).

---

**Theorem 7 (Alignment)** Let \( \mathcal{N} = \text{SCM}[\sigma] \) be a NN-SCM with \( \sigma \) that is Leaky ReLU with parameter \( \alpha > 0 \) and \( A = (X_A, D_A) \) and \( B = (X_B, D_B) \) are two domains. Let \( k \geq \max \{ E_{A,B}^{\epsilon_0}, E_{A,Z}^{\epsilon_0}, E_{Z,B}^{\epsilon_0} \} \) and \( m \geq k + C_{Z,B}^{\epsilon_0} \). Assume that \( D_Z \) is a \((m, \epsilon_0)\)-shared semantic distribution between \( A \) and \( B \) such that \( C_{Z,B}^{m,\epsilon_0} = C_{Z,B}^{\epsilon_0} \). Let \( h \in H_{2\epsilon_0}(D_A, D_B; m) \). Then, one of the following holds:

- \( h = g \circ f \in \mathcal{Z}(D_A, D_Z, D_B; k, \epsilon_0, 3\epsilon_0) \) such that \( g \not\succ f \).
- For every \( y \in \mathcal{Z}(D_A, D_Z, D_B; m, \epsilon_0) \) we have: \( C(h|y) \geq 2C_{Z,B}^{\epsilon_0} - 9 \).

**Proof** By Lem. 19, we have:

\[
C_{A,Z}^{\epsilon_0} + C_{Z,B}^{\epsilon_0} = C_{A,B}^{\epsilon_0} \tag{332}
\]

Therefore, since \( h \in H_{2\epsilon_0}(D_A, D_B; m) \) we obtain:

- \( C(h) \leq C_{A,Z}^{\epsilon_0} + C_{Z,B}^{\epsilon_0} \).
- \( \text{disc}_m(h \circ D_A, D_B) \leq \epsilon_0 \).

In addition, \( D_Z \) is a shared \((m, \epsilon_0)\)-semantic distribution between \( A \) and \( B \). Thus, \( D_Z \) is a shared \((m, \epsilon_0)\)-irreducible distribution between \( A \) and \( B \). By Lem. 41 for \( \epsilon_1 \leftarrow \epsilon_0 \), one of the following holds:

- \( h \in H_{3\epsilon_0}(D_Z, D_B; k, C_{Z,B}^{\epsilon_0}) \circ H_{\epsilon_0}(D_A, D_Z; m) \).
- For all \( y \in \mathcal{Z}(D_A, D_Z, D_B; m, \epsilon_0) \) we have: \( C(h|y) \geq 2C_{Z,B}^{\epsilon_0} - 9 \).

If the first option holds, then \( h = g \circ f \) such that \( f \in H_{\epsilon_0}(D_A, D_Z; m) \) and \( C(g) \leq C_{Z,B}^{\epsilon_0} \). Therefore, by Lem. 42 for \( \epsilon_1 \leftarrow \epsilon_0 \) we have: \( g \not\succ f \). In addition, by Lem. 18 we have:

\[
H_{\epsilon_0}(D_A, D_Z; m) \subset H_{\epsilon_0}(D_A, D_Z; k) \tag{333}
\]

and since \( C_{Z,B}^{m,\epsilon_0} = C_{Z,B}^{\epsilon_0} \) we can replace the first option by:

\[
h = g \circ f \in H_{3\epsilon_0}(D_Z, D_B; k, C_{Z,B}^{\epsilon_0}) \circ H_{\epsilon_0}(D_A, D_Z; k) = \mathcal{Z}(D_A, D_Z, D_B; k, \epsilon_0, 3\epsilon_0) \text{ s.t. } g \not\succ f \tag{334}
\]
Appendix J. Irreducible distributions

If \( y_{AB} = y_B \circ y_A^{-1} \) is the minimal mapping that passes through a distribution \( D_Z \) as \( D_A \xrightarrow{y_A^{-1}} D_Z \xrightarrow{y_B} D_B \), then one can replace \( D_Z \) with a shared irreducible distribution \( D_T \), such that \( y_{AB} = g_B \circ y_A^{-1} \), \( D_A \xrightarrow{y_A^{-1}} D_T \xrightarrow{g_B} D_B \) and there is a function \( b \) such that \( y_A = g_A \circ b \), \( y_B = g_B \circ b \). For simplicity we assume that the discriminators are \( D_\infty \). We will omit writing that the discriminators are of complexity \( \leq \infty \), i.e., \( H_i(D_1, D_2) := H_i(D_1, D_2; \infty) \), \( Z(D_1, D_2; \epsilon) := Z(D_1, D_2; \infty, \epsilon) \), etc.

**Lemma 43** Let \( \mathcal{N} = \text{SCM}[\sigma] \) be a NN-SCM with \( \sigma \) that is Leaky ReLU with parameter \( \alpha > 0 \). Let \( A = (X_A, D_A) \) and \( B = (X_B, D_B) \) be two domains and \( D_Z \) a distribution. Assume that \( C_{a,Z}^{\alpha/2} = C_{A,Z}^{\alpha/2} \) and \( C_{Z,B}^{\alpha} = C_{Z,B}^{\alpha/2} \). Let \( y_{AB} = y_B \circ y_A^{-1} \in Z(D_A, D_Z, D_B; \epsilon_0/2) \) such that:

\[ u \in H_{2\epsilon_0}(D_Z, D_B; C_{Z,B}^{\alpha} + 4) \circ H_{2\epsilon_0}(D_A, D_Z) \text{ we have: } C(y_{AB}) \leq C(u) \quad (335) \]

Then, there is a \( \epsilon_0 \)-shared irreducible distribution \( D_T \) such that:

\[ y_{AB} = g_B \circ y_A^{-1} \in H_{2\epsilon_0}(D_T, D_B; C_{T,B}^{\alpha} + 4) \circ H_{\epsilon_0}(D_A, D_T) \]

such that: \( g_B \not\approx g_A^{-1} \) and \( \exists b : y_A = g_A \circ b \), \( y_B = g_B \circ b \) \quad (336)

**Proof** We divide the proof into a few parts.

**Part 1:** In this part, we find a distribution \( D_T \) such that:

\[ y_{AB} \in H_{2\epsilon_0}(D_T, D_B; C_{T,B}^{\alpha} + 4) \circ H_{\epsilon_0}(D_A, D_T) \quad (337) \]

By Thm. 16, there are decompositions, \( y_B = g_B \circ b \) and \( y_A^{-1} = b^{-1} \circ g_A^{-1} \) such that \( g_B \not\approx g_A^{-1}, b^{-1} \not\approx g_A^{-1} \) and \( C(g_B) + C(b) - 4 \leq C(g_B \circ b) \). We define \( D_T := g_A^{-1} \circ D_A \). In particular,

\[ \text{disc}_\infty(g_A^{-1} \circ D_A, D_T) \leq \epsilon_0 \quad (338) \]

Since \( b^{-1} \not\approx g_A^{-1} \) and \( C_{A,Z}^{\alpha/2} = C_{A,Z}^{2\alpha/3} \), we have:

\[ C(b^{-1}) + C(g_A^{-1}) = C(b^{-1} \circ g_A^{-1}) = C(y_A^{-1}) \leq C_{A,Z}^{\alpha/2} = C_{A,Z}^{2\alpha/3} \quad (339) \]

In addition,

\[ \text{disc}_\infty(b^{-1} \circ D_T, D_Z) = \text{disc}_\infty(b^{-1} \circ g_A^{-1} \circ D_A, D_Z) = \text{disc}_\infty(y_A^{-1} \circ D_A, D_Z) \leq \epsilon_0 \quad (340) \]

Thus, \( C(b^{-1}) \geq C_{T,Z}^{\alpha/2} \) and by Lem. 17, we have:

\[ C(b^{-1}) + C(y_A^{-1}) \leq C_{A,Z}^{\alpha/2} + C_{T,Z}^{\alpha/2} \leq C_{A,T}^{\alpha/2} + C(b^{-1}) \quad (341) \]

In particular, \( C(g_A^{-1}) \leq C_{A,T}^{\alpha/2} \). Therefore, \( g_A^{-1} \in H_{\epsilon_0}(D_A, D_T) \). Thus, \( C(g_A^{-1}) = C_{A,T}^{\alpha/2} \).

Next, we would like to prove that \( g_B \in H_{\epsilon_0}(D_T, D_B; C_{T,B}^{\alpha} + 4) \). We consider that:

\[ \text{disc}_\infty(g_B \circ D_T, D_B) = \text{disc}_\infty(g_B \circ g_A^{-1} \circ D_A, D_B) = \text{disc}_\infty(y_{AB} \circ D_A, D_B) \leq \epsilon_0 \quad (342) \]

In particular, \( C(g_A^{-1}) \geq C_{T,B}^{\alpha/2} \). And by the third part of Lem. 2, for \( p \leftarrow b \),

\[ \text{disc}_\infty(b \circ D_T, D_T) \leq \text{disc}_\infty(D_Z, b^{-1} \circ D_T) \leq \epsilon_0 \quad (343) \]

In particular, \( C(b) \geq C_{T,Z}^{\alpha/2} \). By Lem. 17,

\[ C(g_B) + C(b) - 4 \leq C(g_B \circ b) = C(y_B) \leq C_{Z,B}^{\alpha/2} = C_{Z,B}^{2\alpha/3} \]

\[ \leq C_{Z,T}^{\alpha/2} + C_{Z,T}^{\alpha/2} \leq C_{T,B}^{\alpha/2} + C(b) \quad (344) \]

Therefore, \( C(g_B) \leq C_{T,B}^{\alpha/2} + 4 \). In particular,

\[ y_{AB} \in H_{2\epsilon_0}(D_T, D_B; C_{T,B}^{\alpha} + 4) \circ H_{\epsilon_0}(D_A, D_T) \quad (345) \]
Part 2: We would like to prove that:
\[ Z(D_A, D_T, D_B; \epsilon_0) \subset H_{2\epsilon_0}(D_Z, D_B; C_{Z,B}^{\epsilon_0} + 4) \circ H_{2\epsilon_0}(D_A, D_Z) \] (346)

Let \( u_{AB} = u_B \circ u_A^{-1} \in Z(D_A, D_T, D_B; \epsilon_0) \). Then, \( u_{AB} = u_B \circ u_A^{-1} \) such that \( u_A^{-1} \in H_{\epsilon_0}(D_A, D_T) \) and \( u_B \in H_{\epsilon_0}(D_T, D_B) \). By the first item of Lem. 2, for \( D_1 := u_A^{-1} \circ D_A, D_2 := D_T, D_3 := D_Z \) we have:
\[
disc_\infty(b^{-1} \circ u_A^{-1} \circ D_A, D_Z) \leq disc_\infty(b^{-1} \circ D_T, D_Z) + disc_\infty(D_T, u_A^{-1} \circ D_A) \\
= disc_\infty(y_A^{-1} \circ D_A, D_Z) + disc_\infty(D_T, u_A^{-1} \circ D_A) \leq 2\epsilon_0 \] (347)

In addition,
\[
C(b^{-1} \circ u_A^{-1}) \leq C(b^{-1}) + C(u_A^{-1}) = C(b^{-1} \circ g_A^{-1}) = C(y_A^{-1}) \] (348)

In particular,
\[
b^{-1} \circ u_A^{-1} \in H_{2\epsilon_0}(D_A, D_Z) \] (349)

On the other hand, by the first item of Lem. 2, for \( D_1 := b \circ D_Z, D_2 := D_T, D_3 := D_B, p := u_B \),
\[
disc_\infty(u_B \circ b \circ D_Z, D_B) \leq disc_\infty(u_B \circ D_T, D_B) + disc_\infty(b \circ D_Z, D_T) \\
= disc_\infty(u_B \circ D_T, D_B) + disc_\infty(b \circ D_Z, b \circ y_A^{-1} \circ D_A) \leq \epsilon_0 + disc_\infty(b \circ D_Z, b \circ y_A^{-1} \circ D_A) \] (350)

By Lem. 1, for \( p := b \), we have:
\[
disc_\infty(b \circ D_Z, b \circ y_A^{-1} \circ D_A) \leq disc_\infty(D_Z, y_A^{-1} \circ D_A) \leq \epsilon_0 \] (351)

Finally,
\[
disc_\infty(u_B \circ b \circ D_Z, D_B) \leq 2\epsilon_0 \] (352)

In addition, since \( disc_\infty(g_B \circ D_T, D_B) \leq \epsilon_0 \) we have \( C(g_B) \geq C_{T,B}^{\epsilon_0} = C(u_B) \). Therefore,
\[
C(u_B \circ b) \leq C(u_B) + C(b) \leq C(g_B) + C(b) \leq C(g_B \circ b) + 4 = C(y_B) + 4 = C_{Z,B}^{\epsilon_0} + 4 \] (353)

We conclude that:
\[
b^{-1} \circ u_A^{-1} \in H_{2\epsilon_0}(D_A, D_T) \text{ and } u_B \circ b \in H_{2\epsilon_0}(D_T, D_B; C_{Z,B}^{\epsilon_0} + 4) \] (354)

Therefore,
\[
u_{AB} \in H_{2\epsilon_0}(D_T, D_B; C_{Z,B}^{\epsilon_0} + 4) \circ H_{2\epsilon_0}(D_A, D_T) \] (355)

Part 3: Now, let \( u_{AB} = u_B \circ u_A^{-1} \in Z(D_A, D_T, D_B; \epsilon_0) \) such that \( u_A^{-1} \in H_{\epsilon_0}(D_A, D_T) \) and \( u_B \in H_{\epsilon_0}(D_T, D_B) \). Therefore,
\[
C(u_B) \leq C_{T,B}^{\epsilon_0} \leq C(g_B) \text{ and } C(u_A^{-1}) \leq C_{A,T}^{\epsilon_0} = C(g_A^{-1}) \] (356)

In addition, by part 2:
\[
u_{AB} \in H_{2\epsilon_0}(D_T, D_B; C_{Z,B}^{\epsilon_0} + 4) \circ H_{2\epsilon_0}(D_A, D_T) := F \] (357)

Since \( y_{AB} \) has a smaller complexity than all of the functions in \( F \), we have \( C(y_{AB}) \leq C(u_{AB}) \) and:
\[
C(y_{AB}) \leq C(u_B) \leq C(u_B) + C(u_A^{-1}) \] (358)

In addition, since \( g_B \not\in g_A^{-1} \) and the fact that \( C(u_B) \leq C(g_B) \) and \( C(u_A^{-1}) \leq C(g_A^{-1}) \), we have:
\[
C(u_B) + C(u_A^{-1}) \leq C(g_B) + C(g_A^{-1}) = C(y_{AB}) \leq C(u_B \circ u_A^{-1}) \leq C(u_B) + C(u_A^{-1}) \] (359)

In particular, \( u_B \not\in u_A^{-1} \).
Figure 3: Results for celebA Male to Female transfer (a) Input (b) The mapping obtained by the GAN loss without additional losses.

Figure 4: Same as Fig. 3 for black to blond hair conversion.

Figure 5: Same as Fig. 3 for eyeglasses to no eyeglasses conversion.

Figure 6: Same as Fig. 3 for handbag to shoes and shoes to handbag mapping.

Figure 7: Same as Fig. 3 for edges to shoes and shoes to edges conversion.
Figure 8: Results for celebA Male to Female transfer for networks with different number of layers.
Figure 9: Additional results for celebA Male to Female transfer for networks with different number of layers.
Figure 10: Results for celebA Female to Male transfer for networks with different number of layers.
Figure 11: Additional results for celebA Female to Male transfer for networks with different number of layers.
Figure 12: An illustration of Thm. 43. $D_Z$ is the original distribution and $D_T$ is the shared irreducible distribution. The function $y_{-1}^A$ maps between $D_A$ and $D_Z$ and the function $y_B$ maps $D_Z$ to $D_B$. In addition, $y_{-1}^A = b^{-1} \circ g_{A}^{-1}$ and $y_B = b \circ g_B$, such that $g_{A}^{-1}$ maps between $D_A$ and $D_T$ and $g_B$ maps between $D_T$ and $D_B$. 