Doppelgängers and Adversarial Vulnerability

Supplementary Material

A. Relations, Contrast, Coverings, Topology

Every reflexive binary relation (RBR) $\approx$ on a set $\mathbf{X}$ defines a covering $\{\mathfrak{g}(x) = \{y : x \approx y\}\}_{x\in\mathbf{X}}$, of $\mathbf{X}$ and vice versa one can define reflexive binary relationships through coverings of $\mathbf{X}$.

Lemma 1. (a) *A covering $\mathfrak{G} = \{\mathfrak{g}(x)\}_{x\in\mathbf{X}}$, s.t., $x \in \mathfrak{g}(x), \forall x \in \mathbf{X}$ defines a unique RBR,*

$$\approx_{\mathfrak{G}} = \{(x, y) : y \in \mathfrak{g}(x)\}_{x\in\mathbf{X}} \subset \mathbf{X} \times \mathbf{X};$$[17]

(b) *The RBR $\approx_{\mathfrak{G}}$ is symmetric if and only if $y \in \mathfrak{g}(x)$ whenever $x \in \mathfrak{g}(y)$;* **(c)** *The RBR $\approx_{\mathfrak{G}}$ is transitive if and only if $\mathfrak{g}(y) \subset \mathfrak{g}(x)$ for all $(x, y) \in \approx_{\mathfrak{G}}$.*

Proof of Lemma. Parts **(a)** and **(b)** are trivial. If $\approx_{\mathfrak{G}}$ is transitive, then let $x \approx_{\mathfrak{G}} y$. Since $\approx_{\mathfrak{G}}$ is transitive, every $z \in \mathfrak{g}(y)$, i.e., $y \approx_{\mathfrak{G}} z$, satisfies $x \approx_{\mathfrak{G}} z$, i.e., $z \in \mathfrak{g}(x)$.This proves the necessity in part **(c)**. On the other hand, suppose that $\mathfrak{g}(y) \subset \mathfrak{g}(x)$ for all $(x, y) \in \approx_{\mathfrak{G}}$, and so $(x, y) \in \approx_{\mathfrak{G}}$ and $(y, z) \in \approx_{\mathfrak{G}}$ imply $z \in \mathfrak{g}(z) \subset \mathfrak{g}(y) \subset \mathfrak{g}(x)$, which proves the sufficiency in part **(c)**. □

Arguing by symmetry proves the following

Observation 6. *A symmetric RBR is transitive if and only if the canonical covering $\{\mathfrak{g}(x) = \{y : y \approx x\}\}_{x\in\mathbf{X}}$ satisfies $\mathfrak{g}(x) = \mathfrak{g}(y)$ whenever $y \in \mathfrak{g}(x)$.*

Every RBR on a space $\mathbf{X}$ defines a canonical topology $\tau_{\approx}$ on $\mathbf{X}$:

Definition 10. Let $\approx$ be a RBR on a space $\mathbf{X}$, and let $\mathfrak{G} = \{\mathfrak{g}(x)\}_{x\in\mathbf{X}}$, s.t., $x \in \mathfrak{g}(x), \forall x \in \mathbf{X}$ be the covering generated by $\approx$, the topology generated by the sub-basis $\mathfrak{G}$ is called the **canonical topology** generated by the RBR $\approx$ and is denoted by $\tau_{\approx}$. If the RBR is symmetric, the canonical topology is called **tolerable topology**.

Tolerable/Tolerance topologies and some applications have been studied in [34, 59], and [57].

A.1. Example 2 Metric/Pseudometric

If the indiscriminability relation is transitive, then the perceptual topology may be optimal or not. In the former case

$$d_{\mathcal{K}}(x, y) = \begin{cases} 0, & x \overset{\alpha\delta}{\approx} y \\ 1, & x \overset{\alpha\delta}{\not\approx} y \end{cases} \tag{20}$$

is a metric generating the optimal perceptual topology, in the later case $d_{\mathcal{K}}$ is a non-separating pseudo-metric generating the perceptual topology.[18]

B. Features and Relations.

The attempts to understand the role of feature representations in the processes of discrimination and, in particular, establishing identity are ongoing. Much remains unknown, reduced to embracing items of faith, and is frequently and thoroughly revised. See for example, [4, 55]. Philso

Definition 11. We will say that the feature representation **satisfies the Law of Indiscriminability** if $\Phi_x = \Phi_y$ implies $x \overset{\alpha\delta}{\approx} y$.

Example 6: Suppose that the indiscriminability relation is transitive. Let $\Phi = \mathbf{X}$. The feature representation $\Phi_x = \mathfrak{d}(x)$ for every $x \in \mathbf{X}$, satisfies the Law of Indiscriminability. Indeed, the transitivity of $\overset{\alpha\delta}{\approx}$ implies that $\mathfrak{d}(x) = \mathfrak{d}(y)$ iff $x \overset{\alpha\delta}{\approx} y$, and hence $\Phi_x = \Phi_y$ iff $x \overset{\alpha\delta}{\approx} y$.

Observation 7. *If there exists a feature representation that satisfies the Law of Indiscriminability and such that $(x \overset{\alpha\delta}{\approx} y) \implies (\Phi_x = \Phi_y)$, then $\overset{\alpha\delta}{\approx}$ is transitive.*

Proof of Observation 7.
Let $x \overset{\alpha\delta}{\approx} y \overset{\alpha\delta}{\approx} z$, then under the assumption that $u \overset{\alpha\delta}{\approx} v$ implies $\Phi_u = \Phi_v$, we get $\Phi_x = \Phi_y = \Phi_z$. The feature representation satisfies the Law of Indiscriminability and so from $\Phi_x = \Phi_z$ we get $x \overset{\alpha\delta}{\approx} z$. □

For every discriminative feature representation, indiscernible inputs are indiscriminable, i.e., $(\Phi_x = \Phi_y) \implies (x \overset{\alpha\delta}{\approx} y)$. The feature representation in Example 6 is a discriminative feature representation and satisfies the Law of Indiscriminability. However, unless the corresponding perceptual topology is optimal, the Leibniz Law of Identity does not hold.

The Leibniz Law of Identity does hold in the following example.

Example 7: Let $\mathbf{X} = (0, +\infty)$. Suppose that Weber's law holds and let $k > 0$ be the Weber constant. Let $w = 1 + k$, and let

$$\mathfrak{d}(x) = (x/w, xw) \tag{21}$$

[17] We will use the notation $x \approx_{\mathfrak{G}} y$, whenever $(x, y) \in \approx_{\mathfrak{G}}$.

[18] $d_{\mathcal{K}}(x, y) = d(x, y)(\mathfrak{d}(x))$ where d is the continuity defined by Kopperman in [38].

The covering $\{\mathfrak{d}(x)\}_{x\in\mathbf{X}}$ defines a symmetric RBR on $\mathbf{X}$. Let Φ be the collection of all sub-intervals in $\mathbf{X}$. For every $x \in \mathbf{X}$, let $\Phi_+(x) \subset \Phi$ be the collection of all semi-closed intervals of the form $[b, bw)$, $x/w < b \leq x$ and let $\Phi_-(x) \subset \Phi$ be the collection of all semi-closed intervals of the form $(b/w, b]$, $x \leq b < xw$. Then, the assignment $\Phi_x = \Phi_-(x) \bigcup \Phi_+(x)$ defines a discriminative feature representation on $\mathbf{X}$ where the $\overset{\alpha\delta}{\approx}$ relationship is generated by the covering defined in Equation 21. Furthermore, the Leibniz Law of Identity holds and the feature representation satisfies the Law of Indiscriminability.

While indiscernibility has been discussed extensively, we believe that active discrimination of feature representations is equally important biologically and epistemologically.

Definition 12. We will say that x and y are **actively indiscernible**, in a given context, if the subject is not able to activate the relevant knowledge that the features attributed to x and the features attributed to y are distinct. We will use the notation $\Phi_x \overset{\alpha\delta}{\approx} \Phi_y$ to denote that x and y are actively indiscernible.

Indiscernibility does imply active indiscernabality, i.e, $(\Phi_x = \Phi_y) \implies (\Phi_x \overset{\alpha\delta}{\approx} \Phi_y)$. The former is a transitive relation on $\mathbf{X}$ but the later is a reflexive and symmetric relationship that may or may not be transitive. Many researchers have studied cases in which $(\Phi_x \overset{\alpha\delta}{\approx} \Phi_y) \implies (x \overset{\alpha\delta}{\approx} y)$, see for example [33, 56, 57]. Alternatively, it is possible that there exist biologically plausible contexts where $(x \overset{\alpha\delta}{\approx} y) \implies (\Phi_x \overset{\alpha\delta}{\approx} \Phi_y)$.

B.1. Discriminative feature representations.

Discriminative feature representations are very useful to model and study indiscrimination and categorization. However, the few explicit examples, including Example 6, Example 7, and the maximal cliques proposed in [66], may be too large to be biologically plausible. It is possible that some of the apparent bloat is due to the inclusion of hypothetical (non-attributable) and redundant features.

The existence of discriminative feature representations is a specific instance of a general property of reflexive and symmetric binary relations.

Definition 13. Let $\approx$ be a reflexive and symmetric binary relation on a set $\mathbf{X}$, and let Φ be a set (of features). A feature representation $\{\Phi_x \subset \Phi\}_{x\in\mathbf{X}}$ is called a $\approx$-**discriminative feature representation** if

$$\Phi_x \bigcap \Phi_y \neq \emptyset \Longleftrightarrow x \approx y. \tag{22}$$

Theorem 1. [19] *Every symmetric and reflexive binary representation $\approx$ admits a $\approx$-discriminative feature representation. Specifically, there exists a set of features Φ and a feature representation $\{\Phi_x \subset \Phi\}_{x\in\mathbf{X}}$ satisfying Condition (22).*

Proof of Theorem 1.
Let $\Gamma(\mathbf{X}, E_\approx)$ be the simple graph whose vertexes are the points in $\mathbf{X}$ and the set of edges is $E_\approx = \{\{x, y\}, x \approx y\}$. Denote by $\mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))$ the nonempty cliques in the graph $\Gamma(\mathbf{X}, E_\approx)$, and for every $x \in \mathbf{X}$ let $\mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))_x = \{\varkappa \in \mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx)) : x \in \varkappa\}$. We define the feature space Φ to be the nonempty subsets of $\mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))$ and the feature representation to be

$$\Phi_x = \mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))_x, \forall x \in \mathbf{X}.$$

For every pair $x \approx y$, $\Phi_x \cap \Phi_y \neq \emptyset$ since the clique $\{x, y\} \in \mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))_x \cap \mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))_y$. Vice versa, if $\Phi_x \cap \Phi_y \neq \emptyset$, then x, y belong to some clique $\kappa \in \mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))$ and so $x \overset{\alpha\delta}{\approx} y$. □

Discriminative feature representations are just $\overset{\alpha\delta}{\approx}$-discriminative feature representations.

Example 6 Continued: When indiscriminability is transitive, the feature representation $\{\Phi_x = \mathfrak{d}(x)\}_{x\in\mathbf{X}}$ is a discriminative feature representation, the Doppelgängers $y \overset{\alpha\delta}{\approx} x$ of every $x \in \mathbf{X}$ are the discriminative features of x and

$$\mathfrak{cl}(y) = \mathfrak{d}(x), \forall y \in \mathfrak{d}(x).$$

Example 6 and Example 7 show that building discriminative feature representations might be too resource demanding, because, the size of the feature representations Φ_x may be too large. On the other hand, even such large feature representations may be starting points in the search and identification of smaller representations through a refinement process outlined below. In particular, applying this process to the discriminative feature representation in Example 6, yields a new discriminative feature representation $\{\hat{\Phi}_x = \{\mathfrak{d}(x)\}\}_{x\in\mathbf{X}}$. The new discriminative feature representation appears simpler, the feature representation of each input is a single feature. However, the new features are much more complex than the original features.

Let Φ be a space of features, and let $\{\Phi_x \subset \Phi\}_{x\in\mathbf{X}}$ be a context dependent feature representation of the inputs $x \in \mathbf{X}$. The **semantic cluster** of inputs sharing a feature $\xi \in \Phi$ is the context-dependent cluster of inputs:

$$\mathfrak{cl}(\xi) = \{x : \xi \in \Phi_x.\} \tag{23}$$

The feature ξ is hypothetical in a given context if its semantic cluster is empty, $\mathfrak{cl}(\xi) = \emptyset$. The semantic clusters shared

[19] The theorem is attributed to Kalmar and Yakubovich, and is proven for reflexive and symmetric binary relations on finite sets in [66]. Here we provide simple proof of the general case, that does not rely on transfinite induction

by the attributable features define a new feature space $\mathfrak{cl}(\Phi)$ and a new feature representation. Specifically, for every $x \in \mathbf{X}$ define $\mathfrak{cl}(\Phi)_x = \{\mathfrak{cl}(\xi) : \xi \in \Phi_x\}$ and the feature representation $\{\mathfrak{cl}(\Phi)_x \subset \mathfrak{cl}(\Phi) = \bigcup_{x\in\mathbf{X}} \mathfrak{cl}(\Phi)_x\}_{x\in\mathbf{X}}$. There is a bijective mapping between the new feature space $\mathfrak{cl}(\Phi)$ and the quotient space $\Phi/\equiv$ where the equivalence relation $\equiv$ is defined by $\xi \equiv \eta \iff \mathfrak{cl}(\xi) = \mathfrak{cl}(\eta)$. In a sense the new feature space and representation are smaller (there is an on-to mapping $\Phi \to \mathfrak{cl}(\Phi)$). All features in $\mathfrak{cl}(\Phi)$ are attributed and there are no redundant, semantically synonymous features. Furthermore,

$$\mathfrak{cl}(\xi) = \mathfrak{cl}(\mathfrak{cl}(\xi)), \text{ for every attributed feature } \xi \in \Phi. \tag{24}$$

Indeed, if ξ is attributed feature, then $y \in \mathfrak{cl}(\xi) \iff \xi \in \Phi_y \iff \mathfrak{cl}(\xi) \in \mathfrak{cl}(\Phi)_y \iff y \in \mathfrak{cl}(\mathfrak{cl}(\xi))$.

Let $\approx$ be a reflexive and symmetric binary relation on a set $\mathbf{X}$, and let $\{\Phi_x \subset \Phi\}_{x\in\mathbf{X}}$ be a $\approx$- **discriminative feature representation**, i.e., the representation satisfies Condition 22. The clusters of inputs sharing attributed features have additional structure and define a new discriminative feature representation.

Observation 8. *Let $\approx$ be a reflexive and symmetric binary relation on a set $\mathbf{X}$, and let $\{\Phi_x \subset \Phi\}_{x\in\mathbf{X}}$ be a $\approx$-discriminative feature representation, i.e., the representation satisfies Condition 22.*

(i.) *For every attributed feature $\xi \in \Phi$, the cluster $\mathfrak{cl}(\xi)$ is a clique in $\Gamma(\mathbf{X}, E_\approx)$ and $\mathfrak{cl}(\xi) \in \mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))_x$, for every $x \in \mathfrak{cl}(\xi)$.*

(ii.) *$\{\mathfrak{cl}(\Phi)_x\}_{x\in\mathbf{X}}$ is a $\approx$-discriminative feature representation.*

Proof of Observation 8

(i.) Let $y, z \in \mathfrak{cl}(\xi)$, and so $\xi \in \Phi_y \cap \Phi_z$, but $\{\Phi_x \subset \Phi\}_{x\in\mathbf{X}}$ is a $\approx$-discriminative feature representation and so $y \approx z$. Thus $\mathfrak{cl}(\xi) \in \mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))_x \subset \mathfrak{Cl}(\Gamma(\mathbf{X}, E_\approx))$, for all $x \in \mathfrak{cl}(\xi)$.

(ii.) If $x \approx y$, then there exists $\xi \in \Phi_x \cap \Phi_y$ and so $\mathfrak{cl}(\xi) \in \mathfrak{cl}(\Phi)_x \cap \mathfrak{cl}(\Phi)_y$, and clearly $\mathfrak{cl}(\Phi)_x \cap \mathfrak{cl}(\Phi)_y \neq \emptyset$.

On the other hand, every feature $\xi \in \Phi$, such that $\mathfrak{cl}(\xi) \in \mathfrak{cl}(\Phi)_x \cap \mathfrak{cl}(\Phi)_y \neq \emptyset$, belongs to $\Phi_x \cap \Phi_y$, and so $\mathfrak{cl}(\Phi)_x \cap \mathfrak{cl}(\Phi)_y \neq \emptyset \implies \Phi_x \cap \Phi_y \neq \emptyset \implies x \approx y$.

□

B.2. Finite Discriminative Feature Representations

Observation 9. *Let $\approx$ be symmetric and reflexive binary relation on $\mathbf{X}$, Φ a finite set of attributed features. If there exists an $\approx$-discriminative feature representation $\{\Phi_x \subset \Phi\}_{x\in\mathbf{X}}$, then for every fully populated classifier $R = \{R_1, \ldots, R_m\}$ with more labels than the total number of attributed features $\#\Phi$, $m > \#\Phi$, there exist $x \approx y$ such that $\text{label}_R(x) \neq \text{label}_R(y)$.*

Proof of Observation 9

Let $\Phi_{(j)}$ be the collection of features attributed to inputs whose labels equal $j = 1, \ldots, m$. Specifically,

$$\Phi_{(j)} = \{\xi \in \Phi : \xi \in \Phi_x \text{ for some } x \in R_j \subset \mathbf{X}\}.$$

Denote by $\#_j$ the number of elements in $\Phi_{(j)}$, i.e., $\#_j = \#\Phi_{(j)}$. The classifier is fully populated and the feature representation is $\approx$-discriminative feature representation and so

$$\#_j \geq 1, \quad j = 1, \ldots, m. \tag{25}$$

The sets of attributed features $\Phi_{(j)}$ cannot be disjoint. Indeed, $\Phi_{(i)} \cap \Phi_{(j)} = \emptyset$, if $i \neq j$ together with Inequality 25would lead to the contradiction

$$m > \#\Phi = \sum_{j=1}^{m} \#_j \geq m.$$

Thus there exits a feature $\xi \in \Phi_{(i)} \cap \Phi_{(j)}$ for some labels $i \neq j$. And hence there exist two inputs $x \in R_i$ and $y \in R_j$, i.e., $\text{label}_R(x) = i \neq j = \text{label}_R(y)$, such that $\xi \in \Phi_x \cap \Phi_y$. The feature representation is $\approx$-discriminative and so $x \approx y$.

□

C. Sorites, Ill-posed Classification

Sorites chains and the related paradoxes are deeply related to indiscriminability. They have been studied and argued since at least the 4th century BCE.[20] Every pair of adversarial Doppelgängers $x \overset{\alpha\delta}{\approx} y$ such that $\text{label}_R(x) \neq \text{label}_R(y)$ is a sorites chain. On the other hand, every sorites chain $x_1 \overset{\alpha\delta}{\approx} x_2 \overset{\alpha\delta}{\approx} \cdots \overset{\alpha\delta}{\approx} x_n$ such that $\text{label}_R(x_1) \neq \text{label}_R(x_n)$ for some classifier R must contain a pair of adversarial Doppelgängers. Indeed,

Lemma 2. *Let R be classifier with labeling function $\text{label}_R : X \to \{1, 2, \ldots, m\}$. If there exists a chain of Doppelgängers $x_1 \overset{\alpha\delta}{\approx} x_2 \overset{\alpha\delta}{\approx} \cdots \overset{\alpha\delta}{\approx} x_n$ whose initial and final samples are assigned different labels $\text{label}_R(x_1) \neq \text{label}_R(x_n)$ by R then there exists a pair of adversarial Doppelgängers $x_i \overset{\alpha\delta}{\approx} x_{i+1}$, $\text{label}_R(x_i) \neq \text{label}_R(x_{i+1})$, where $i \in \{1, \ldots, n-1\}$. In fact i can be chosen so that $\text{label}_R(x_1) = \text{label}_R(x_i)$ and/or $\text{label}_R(x_{i+1}) = \text{label}_R(x_n)$.*

[20] At least since Eubulides of Miletus formulated the Heap Paradox.

Proof of Lemma. The short proof of the lemma is constructive. There are two types of constructions that might be used to produce the adversarial Doppelgängers, the "first encounter pair" and the "last encounter". The last encounter construction is based on identifying the last link in the chain that has the same label as x_1. Set

$$i = \max\{j \in \{1, 2, \dots, m\} \,:\, \text{label}_R(x_1) = \text{label}_R(x_j)\}.$$

Then $\text{label}_R(x_i) = \text{label}_R(x_1)$ and $\text{label}_R(x_i) = \text{label}_R(x_1) \neq \text{label}_R(x_{i+1})$. Thus $x_i \overset{\alpha\delta}{\approx} x_{i+1}$ is a pair of adversarial Doppelgängers. The first encounter construction is based on identifying the first link in the chain whose label is different from $\text{label}_R(x_1)$. Indeed, let

$$i = \min\{j \in \{1, 2, \dots, m\} \,:\, \text{label}_R(x_1) \neq \text{label}_R(x_j)\}.$$

Then $\text{label}_R(x_{i-1}) = \text{label}_R(x_1)$ and $\text{label}_R(x_{i-1}) = \text{label}_R(x_1) \neq \text{label}_R(x_i)$. Thus $x_{i-1} \overset{\alpha\delta}{\approx} x_i$, are a pair of adversarial Doppelgängers. To complete the proof of the lemma reverse the order of elements in the Doppelgänger chain.

□

Lemma 3. *Let* $\mathbf{X}$ *be the closed bounded interval* $[a, b] \subset (0, +\infty)$. *Suppose that Weber's law holds and let* k *be the Weber constant. Let* $w = 1 + k$, *and*

$$\mathfrak{d}(x) = \begin{cases} [a, xw)\,, & a \leq x < aw \\ (x/w, xw)\,, & aw \leq x \leq b/w \\ (x/w, b]\,, & b/w < x \leq b, \end{cases}$$

then $\mathbf{X}/\!\sim_\sigma$ *is a singleton.*

Proof of Lemma. Indeed, an argument by induction shows that every two inputs $x \in \mathbf{X}$ and $y \in \mathbf{X}$ can be "connected" by a finite chain of Doppelgängers $x \overset{\alpha\delta}{\approx} x_1 \overset{\alpha\delta}{\approx} x_2 \overset{\alpha\delta}{\approx} \cdots \overset{\alpha\delta}{\approx} x_m \overset{\alpha\delta}{\approx} y$. The induction will be on $\text{jump}(x; y) = \min\{l \in \mathbb{N} : \; y < w^l x\}$ **the number of (perceptual) jumps** one needs to get from x to y. Note that by the definition of the sub-basis $\text{jump}(x; y) = 1$ implies $x \overset{\alpha\delta}{\approx} y$.

Let us assume that $\text{jump}(x; z) \leq n$ implies that x can be connected to z by a finite chain of Doppelgängers, and suppose that $\text{jump}(x; y) = n + 1$.

If $w^n x < y < w^{n+1} x$, choose any x_* such that $y/w < x_* < w^n x$ (see Figure 3).

Since $\text{jump}(x; x_*) = n$, the induction assumption implies that there exists a finite chain of Doppelgängers $x \overset{\alpha\delta}{\approx} x_1 \overset{\alpha\delta}{\approx} x_2 \overset{\alpha\delta}{\approx} \cdots \overset{\alpha\delta}{\approx} x_*$. By construction $x_* \overset{\alpha\delta}{\approx} y$, and so we can add this final link to obtain the finite chain $x \overset{\alpha\delta}{\approx} x_1 \overset{\alpha\delta}{\approx} x_2 \overset{\alpha\delta}{\approx} \cdots \overset{\alpha\delta}{\approx} x_* \overset{\alpha\delta}{\approx} y$. □

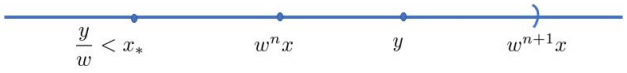

Figure 3. Linking $x \overset{\alpha\delta}{\approx} x_1 \overset{\alpha\delta}{\approx} x_2 \overset{\alpha\delta}{\approx} \cdots \overset{\alpha\delta}{\approx} x_*$ with $x_* \overset{\alpha\delta}{\approx} y$ to get the chain of Doppelgängers $x \overset{\alpha\delta}{\approx} x_1 \overset{\alpha\delta}{\approx} x_2 \overset{\alpha\delta}{\approx} \cdots \overset{\alpha\delta}{\approx} x_* \overset{\alpha\delta}{\approx} y$ from x to y.

Thus if R is any classifier defined on $[a, b]$ such that $\text{label}_R(x) \neq \text{label}_R(y)$ for some $a \leq x, y \leq b$, then $x \sim_\sigma y$ and Lemma 2 implies that R admits adversarial Doppelgängers and hence cannot be regular.

More generally, if the transitive closure $\sim_\sigma$ of $\overset{\alpha\delta}{\approx}$ is trivial then every fully populated classifier with two or more classes admits adversarial Doppelgängers. In particular, let R be a fully populated classifier with labeling function $\text{label}_R : X \twoheadrightarrow \{1, 2, \dots, m\}$, then for every label c there exist adversarial Doppelgängers $x(c) \overset{\alpha\delta}{\approx} x^*(c) \in \mathbf{X}$ such that $c = \text{label}_R(x(c)) \neq \text{label}_R(x^*(c))$.

D. Weber's Law and a Regular Classifier

Example 8: Let $w > 1$ and $\mathbf{X}$ be the closed bounded interval $[a, b'] \subset (0, +\infty)$ and let $a < b < b'/w$ and

$$\mathfrak{d}(x) = \begin{cases} [a, xw)\,, & a \leq x < aw \\ (x/w, xw)\,, & aw \leq x \leq b/w \\ (x/w, b]\,, & b/w < x \leq b \\ (b, xw)\,, & b < x \leq bw \\ (x/w, b']\,, & bw < x \leq b'. \end{cases} \tag{26}$$

There exists a unique regular fully populated classifier with two labels. Every other fully populated classifier with two or more labels must admit adversarial Doppelgängers.

E. Proof of Observation 4, Section 5.2.

Here we will prove that if $\inf_{x \in \mathbf{X}} \mu(\mathfrak{d}(x)) > 0$, then for every classifier $R = \{R_1, \dots, R_m\}$ whose recall rates are sufficiently high so that $\underline{\rho} > 1 - 1/\bar{k}(\Omega)$. i.e.,

$$\left(1 - \underline{\rho}\right) \bar{k}(\Omega) < 1 \tag{27}$$

Then every misclassified input x is an adversarial Doppelgänger.

Proof of Observation 4. Let x be misclassified by R, that is $x \in R_j \cap \Omega_{i(x)}$, where $j \in \{1, \dots, m\}$ and $j \neq i(x)$ and so

$$\mu\left(\Omega_{i(x)}\right) - \mu\left(\Omega_{i(x)} \cap R_{i(x)}\right) = \sum_{s \neq i(x)} \mu\left(\Omega_{i(x)} \cap R_s\right). \tag{28}$$

The input x is misclassified by R, i.e., $j = \text{label}_R(x) \neq i(x)$, and so $\sum_{s \neq i(x)} \mu\left(\Omega_{i(x)} \cap R_s\right) \geq \mu\left(\Omega_{i(x)} \cap R_j\right)$. Therefore, we get

$$\mu\left(\Omega_{i(x)}\right) - \mu\left(\Omega_{i(x)} \cap R_{i(x)}\right) \geq \mu\left(\Omega_{i(x)} \cap R_j\right) \tag{29}$$

and so

$$1 - \underline{\rho} \geq 1 - \frac{\mu\left(\Omega_{i(x)} \cap R_{i(x)}\right)}{\mu\left(\Omega_{i(x)}\right)} \geq \frac{\mu\left(\Omega_{i(x)} \cap R_j\right)}{\mu\left(\Omega_{i(x)}\right)}$$

The lower bound of the recall rates $\left(1 - \underline{\rho}\right) \bar{k}(\Omega) < 1$ and $\mathfrak{d}(x) \subset \Omega_{i(x)}$ yield the estimate

$$\begin{aligned} 1 &> \left(1 - \underline{\rho}\right) \bar{k}(\Omega) \\ &\geq \frac{\mu\left(\Omega_{i(x)}\right)}{\mu\left(\mathfrak{d}(x)\right)} \frac{\mu\left(\Omega_{i(x)} \cap R_j\right)}{\mu\left(\Omega_{i(x)}\right)} \\ &\geq \frac{\mu\left(\mathfrak{d}(x) \cap R_j\right)}{\mu\left(\mathfrak{d}(x)\right)}. \end{aligned} \tag{30}$$

Hence the set of adversarial Doppelgängers of x has positive measure, i.e., $\mu\left(\mathfrak{d}(x) \setminus R_j\right) > 0$.

□

F. Adversarial Training May or May Not Work

Example 9: Let $\mathbf{X} = \mathbb{R}$ be the real line and let the probability measure μ be the Gaussian with mean 0 and variance $\sigma^2 = 1/2$. Suppose that Weber's law holds and let $k > 0$ be the Weber constant. Let $w = 1 + k$, and let

$$\mathfrak{d}(x) = \begin{cases} (wx, x/w), & x < 0 \\ \{0\}, & x = 0 \\ (x/w, xw), & x > 0 \end{cases} \tag{31}$$

Then $\mathbf{X}/\sim_\sigma = \{(-\infty, 0), \{0\}, (0, +\infty)\}$ and

$$\mu(\mathfrak{d}(x)) = \begin{cases} \dfrac{|\text{erf}(wx) - \text{erf}(x/w)|}{2}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{32}$$

Consider the regular classifier Ω two classes $\Omega_1 = (-\infty, 0)$ and $\Omega_2 = [0, +\infty)$. Let $\epsilon > 0$ and let $R(\varepsilon)$ be a two label classifier such that

$$\text{label}_{R(\varepsilon)}(x) = \begin{cases} 1, & x < \epsilon \\ 2, & x \geq \epsilon \end{cases} \tag{33}$$

Then the misclassified inputs are $[0, \epsilon)$, $\mathfrak{d}(\epsilon)$ is the collection of all adversarial Doppelgängers and the set of misclassified adversarial Doppelgängers is $(\epsilon/w, \epsilon)$.

Using the adversarial training [52], will move the decision boundary towards zero thus improving the accuracy of the classifier (at best the new decision boundary moves to $x = \epsilon/w$). The good news is that the "robustified" classifiers will converge to the perfect accuracy classifier Ω. See Figure 5(a)

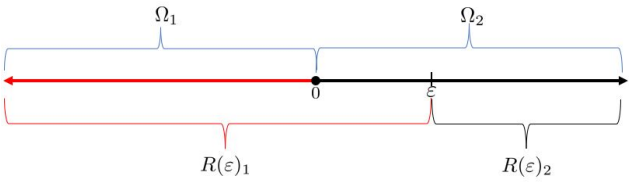

Figure 4. $X = \mathbb{R}$, $\mathfrak{d}(x)$ defined in (31), $\Omega_1 = (-\infty, 0)$ and $\Omega_2 = [0, +\infty)$, the classifier $R(\varepsilon)$ defined in (33).

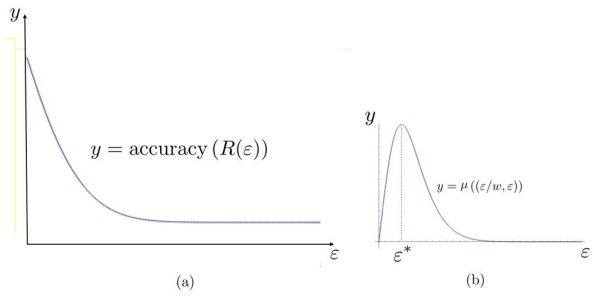

Figure 5. (a) The accuracies of the linear classifiers $R(\varepsilon)$ decrease as ε increases. (b) The sizes of the sets of adversarial Doppelgängers of the linear classifiers $R(\varepsilon)$ have a unique maximal value achieved at ε^*. The graph of the sizes of the sets of adversarial Doppelgängers is shown in this panel. To improve the readability the graph is scaled up to match the scales of the accuracy graph.

However, the function $\mu((\varepsilon/w, \varepsilon))$ has a unique global maximum on $[0, +\infty)$ achieved at $\varepsilon^* > 0$, see Figure 5(b) . Thus if we apply robust training starting with a classifier s.t. $\varepsilon > \varepsilon^*$ then we will improve accuracy but gain adversarial Doppelgängers (seemingly, we will be trading off adversarial Doppelgänger robustness to gain accuracy). However, if $\varepsilon < \varepsilon^*$, robust training will improve both accuracy and adversarial Doppelgänger robustness (i.e., there will be no trade-off just gain across the front) as we move towards the regular classifier.

Furthermore, the known methods to achieve certifiable robustness cannot prevent significant AD vulnerability. This is due to the misalignment between the l_p metric topology and human perception (the perceptual topology). In many cases the measure of the overlap of the phenomenal neighborhood $\mathfrak{d}(x)$ of an input x and the metric ball $B_{r(x)}(x)$ could have very small measure compared to the measure of the phenomenal neighborhood (here $r(x)$ is the robustness radius at x). For example, this happens when the input space is input spaces, specifically those that satisfy

Weber's and Weber-Fechner's laws.[21]

Thus robust training requires a way to locate and capture Doppelgängers.

We are not aware of any state of the art adversarial training including "salience-based adversarial training", which does that.

Perceptual discrimination is often a pre-attentive process, while the current salience-based training has been developed for higher level tasks that involve complex feature salience driven attention. at present, it is not known whether and when salience-based training can benefit the detection of Doppelgängers.

F.1. Where are the Doppelgängers?

The perceptual distances can be used to stratify an input space $\mathbf{X}$ into, at most countably many, disjoint spheres

$$S_\rho^w(x) = \{y \in X, \text{s.t. } d_w(x,y) = \rho \in [0,1]\}.$$

The open ball of radius one

$$\mathring{B}_1^w(x) = \{y \in X, \text{s.t. } d_w(x,y) < 1\}$$

is just the equivalence class $[x]_{\sim_\sigma}$ with respect to the equivalence relation $\sim_\sigma$. Some of the strata may be empty sets. For example, if $\sim_\sigma$ is trivial, then $S_1^w(x) = \emptyset$ and $\mathbf{X} = [x]_{\sim_\sigma} = \mathring{B}_1^w(x) = S_0^w(x) \bigcup S_{1/2}^w(x)$ for every $x \in \mathbf{X}$. On the other hand, if $\overset{\alpha\delta}{\approx}$ is transitive ($\overset{\alpha\delta}{\approx}$ equals $\sim_\sigma$), then $[x]_{\sim_\sigma} = \mathring{B}_1^w(x) = B_{1/2}^w(x)$ and $\mathbf{X} = S_0^w(x) \bigcup S_{1/2}^w(x) \bigcup S_1^w(x)$. The former case holds whenever the graph distance (d_∞) between any two inputs is finite (three, six or whatever). An example of the later case is provided in [6].

More interestingly the Doppelgängers of a point are precisely the nearest points to x if the distance is measured by d_w or d_∞.

G. Perceptually Harmonic Functions.

The labeling functions of regular classifiers are step functions and belong to the class of perceptually regular functions, i.e., functions that respect the regularity on the space of inputs imposed by the humans' inability to discriminate different "raw" signals.

Definition 14. A function $f : \mathbf{X} \to \mathbb{R}$ is called **perceptually regular** if $f = \text{const}$ on every equivalence class $\zeta \in \mathbf{X}/\sim_\sigma$.[22] Let $L^1_{\text{pr}}(\mathbf{X}, \mu, \overset{\alpha\delta}{\approx})$ denote the vector space of perceptually regular integrable functions.

[21] See Example 1 in the paper and the related examples in the Supplementary material, including the example discussed in Lemma 3, Example 8, Example 9, and Example 11.

[22] Thus f is perceptually regular if and only if $f : (\mathbf{X}, \overset{\alpha\delta}{\approx}) \to (\mathbb{R}, =)$ is a morphism of tolerance spaces, [69], mapping the (perceptual) tolerance space $(\mathbf{X}, \overset{\alpha\delta}{\approx})$ into the optimal tolerance space $(\mathbb{R}, =)$.

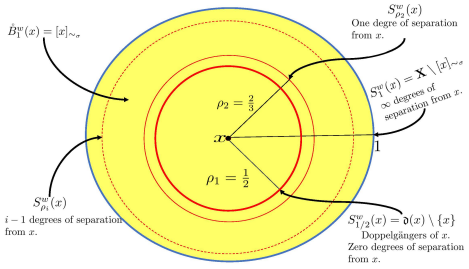

Figure 6. From the view point of an input/stimulus $x \in \mathbf{X}$, the space $\mathbf{X}$ is stratified into concentric spheres, the nearest neighbors of x are precisely its Doppelgängers some or all of which may be adversarial.

Assuming that the degree of a vertex x in the graph $\Gamma(\mathbf{X}, E_{\alpha\delta})$ defined by

$$d_{\alpha\delta}(x) = \mu(\mathfrak{d}(x)) \tag{34}$$

is integrable and $\inf_{x \in \mathbf{X}}(d_{\alpha\delta}(x)) > 0$, then we define the **discrimination Laplace operator** c.f. [7]:

$$\triangle_{\alpha\delta}(f)(x) = f(x) - \frac{1}{\sqrt{d_{\alpha\delta}(x)}} \int_{\mathfrak{d}(x)} \frac{f}{\sqrt{d_{\alpha\delta}}} \tag{35}$$

The kernel of $\triangle_{\alpha\delta}$ is nontrivial since $\triangle_{\alpha\delta}(\sqrt{d_{\alpha\delta}}) = 0$ and so 0 is an eigenvalue of $\triangle_{\alpha\delta}$. On the other hand, globally constant functions $f(x) \equiv c \in \mathbb{R} \setminus \{0\}$ are harmonic if and only if

$$\frac{1}{\sqrt{d_{\alpha\delta}(x)}} \int_{\mathfrak{d}(x)} \frac{1}{\sqrt{d_{\alpha\delta}}} = 1, \forall x \in \mathbf{X}. \tag{36}$$

It is easy to show that (36) fails in Example 1 and Example 8, where the probability measure μ is the uniform measure and so all nontrivial globally constant functions are not $\triangle_{\alpha\delta}$ harmonic.

Example 10: If $\overset{\alpha\delta}{\approx}$ is transitive (as in [6] for example) and hence the graph $\Gamma(\mathbf{X}, E_{\alpha\delta})$ is regular, then $\ker \triangle_{\alpha\delta} = L^1_{\text{pr}}(\mathbf{X}, \mu, \overset{\alpha\delta}{\approx})$ and the spectrum of the Laplace operator is the set $\{0, 1\}$.

The operator $\triangle_{\alpha\delta}$ has at least three aspects that make it hard to use. First, it may not be well defined in many cases. Second, piecewise constant functions are not necessarily $\triangle_{\alpha\delta}$-harmonic. Third, it is hard to describe the full spectrum of $\triangle_{\alpha\delta}$ for most perceptual topologies τ_δ. We will define another operator that exists in many situations when $\triangle_{\alpha\delta}$ is not well defined. Furthermore, all perceptually regular functions are harmonic with respect to it. Finally, its

spectrum is easy to compute. Assuming that the function $d_\sigma : \mathbf{X} \to [0, +\infty)$ defined by

$$d_\sigma(x) = \mu([x]_{\sim_\sigma}) \tag{37}$$

is positive, then we define the **Doppelgängers chain Laplace operator** by:

$$\triangle_\sigma(f)(x) = f(x) - \frac{1}{d_\sigma(x)} \int_{[x]_{\sim_\sigma}} f \tag{38}$$

Note that $\triangle_\sigma$ is defined whenever $\triangle_{\alpha\delta}$ is defined and sometimes $\triangle_{\alpha\delta} = \triangle_\sigma$.[23] The spectrum of $\triangle_\sigma$ is the set $\{0, 1\}$,

$$\ker \triangle_\sigma = L^1_{\mathrm{pr}}(\mathbf{X}, \mu, \overset{\alpha\delta}{\approx}) \tag{39}$$

and

$$\ker(\triangle_\sigma - \mathrm{Id}) = \left\{ f : \int_{[x]_{\sim_\sigma}} f = 0, \forall x \in \mathbf{X} \right\} \neq \{0\}. \tag{40}$$

H. Harmonic salience functions

By definition, a salience scalef_Φ is perceptually regular iff $f_\Phi(\Phi_x) = f_\Phi(\Phi_y)$ whenever $x \sim_\sigma y$, i.e., whenever the corresponding salience function $f : \mathbf{X} \to \mathbb{R}$ is perceptually regular and hence harmonic with respect to $\triangle_\sigma$ ($\triangle_\sigma f = 0$).

I. Class Invariants.

We will call a set $D \subset \mathbf{X}$ **perceptually regular** if $[x]_{\sim_\sigma} \subset D$ for every $x \in D$. The classes in $R_c = \{x : \mathrm{label}_R(x) = c\} \subset \mathbf{X}$, where R is a perceptually regular classifier are perceptually regular subsets. Thus from now on we will use we will use the shorter name **regular class** instead of a perceptually regular subset.

The **structural entropy** of the perceptually regular set D is defined as:

$$H_{\sim_\sigma}(D) = -\frac{1}{\mu(D)} \int_{y \in D} \log\left(\frac{\mu([y]_{\sim_\sigma})}{\mu(D)}\right) \text{ and} \tag{41}$$

the **index of coincidence** is

$$\mathrm{IC}_{\sim_\sigma}(D) = \frac{1}{\mu(D)^2} \int_{y \in D} \mu([y]_{\sim_\sigma}). \tag{42}$$

The total/cumulative importance (salience) of the inputs in a regular class $D \subset \mathbf{X}$ and the expected affinity/similarity between pairs of inputs in D can be defined and exploited if we have access to a well behaved (integrable) importance scale $f_\Phi : \Upsilon(\Phi) \to [0, +\infty)$ and similarity scale $s : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$, we define **the importance of** D, $I_\Phi(D)$, by

$$I_\Phi(D) = \int_D f_\Phi(\Phi_y),$$

and the **the expected affinity** of D by

$$\mathcal{R}_\Phi(D) = \frac{1}{\mu(D)} \int_D \int_D s(x, y) = \frac{1}{\mu(D)} \int_D P(x, D).$$

The expected affinity defined above involves an iterated integral and is essentially the resemblance attribute defined in [79].

J. Analyzing Doppelgänger Vulnerability.

Example 11: Let $\mu(A) = \frac{2\sqrt{\pi}}{\pi} \int_A e^{-t^2}\, dt$ be the probability measure on $\mathbf{X} = (0, +\infty)$ and let the indiscriminability relation on $\mathbf{X}$ be defined by the covering $\mathfrak{D}_{\alpha\delta} = \{\mathfrak{d}(x) = (x/w, xw)\}_{x \in \mathbf{X}}$, where $w > 1$ is a fixed constant. Let $\epsilon > 0$ and let $R(\epsilon)$ be the linear classifier defined by

$$\mathrm{label}_{R(\varepsilon)}(x) = \begin{cases} 1, & 0 < x < \epsilon \\ 2, & x \geq \epsilon. \end{cases} \tag{43}$$

The conceptual entropy of $H_{R(\epsilon)}(x)$ is positive if and only if $x \in (\epsilon/w, w\epsilon)$. In particular, the points outside the region of conceptual ambiguity $\mathbf{X} \setminus [\epsilon/w, w\epsilon]$ are not vulnerable to adversarial Doppelgängers attacks. The conceptual entropy achieves its global maximum $H_{R(\epsilon)}(x_*) = 1/2$ at a $x_* = x_*(\epsilon, w)$. It is equal to zero on $(0, \epsilon/w] \cup [\epsilon w, +\infty)$, and increases monotonically on $[\epsilon/w, x_*]$, and then decreases monotonically on $[x_*, \epsilon w]$. The vulnerability to an adversarial Doppelgänger attack is maximized at the point x_*. The measure of the region of ambiguity $\mathrm{A}(R(\epsilon))$ is $\mu(\mathrm{A}(R(\epsilon)) = \mathrm{erf}(w\epsilon) - \mathrm{erf}(\epsilon/w) < 1$. The $R(\epsilon)$-fooling rate $F_{R(\epsilon)}(\hat{a}) \leq \mathrm{erf}(w\epsilon) - \mathrm{erf}(\epsilon/w) < 1$ of an adversarial Doppelgänger attack $\hat{a}$ is safely bounded away from 1.

K. ANN Discrimination

The indiscriminability of inputs by VGG-19, ResNet, and Inception-V3 has been studied by Feather et al., [18]. Two inputs x and y are indiscriminable by these ANN models, $x \overset{\text{ANN}}{\approx} y$, iff they "produce the same activations in a model layer". The relation $\overset{\text{ANN}}{\approx}$ is transitive. Indeed, let x and y produce the same activations at some level, then they produce the same activations in all subsequent levels. Thus if $x \overset{\text{ANN}}{\approx} y$ and $y \overset{\text{ANN}}{\approx} z$, then x and z produce the same activations at all sufficiently high levels, and therefore, $x \overset{\text{ANN}}{\approx} z$. The same argument does imply that $\overset{\text{ANN}}{\approx}$ is transitive for all free-forward models and for all recurrent neural network models.

[23] If $\overset{\alpha\delta}{\approx}$ is transitive, then $\mathfrak{d}(x) = [x]_{\sim_\sigma}, \forall x \in \mathbf{X}$, and so $\triangle_{\alpha\delta} = \triangle_\sigma$.