Competitive Distance: A New Foundation for Classification Confidence

1 Introduction

1.1 What Are Measures?

Mathematics and science rely on measures to quantify and compare properties of objects, systems, and phenomena. A measure is a function that assigns a numerical value to an entity based on a well-defined rule, allowing for meaningful comparisons and computations. Measures appear across disciplines, from physics and engineering to probability theory and machine learning, forming the backbone of quantitative reasoning.

1.1.1 A Motivating Example

To illustrate the importance of measures, consider the problem of comparing the brightness of two light sources. While we might intuitively say that one is "twice as bright" as the other, this requires a clear, numerical definition of brightness. In physics, brightness is often measured using luminous intensity, defined as power per unit solid angle (candela). However, other contexts might define brightness differently, leading to inconsistencies. This highlights the necessity of precise, well-defined measures.

1.1.2 Formal Definition of a Measure

In mathematical analysis, a measure is formally defined within the framework of measure theory. Given a set X, a measure μ is a function that assigns a non-negative real number to subsets of X, satisfying certain axioms of additivity:

1. **Non-negativity:** For any measurable set $A \subseteq X$, the measure is always non-negative:
$$\mu(A) \geq 0.$$

2. **Null empty set:** The measure of the empty set is zero:
$$\mu(\emptyset) = 0.$$

3. **σ-additivity:** For any countable collection of disjoint measurable sets $A_1, A_2, \ldots$, the measure satisfies:

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

These properties ensure that a measure is well-defined and can be used to quantify size, probability, or other numerical attributes of sets. While this formalism applies primarily to measure theory, more general notions of measures appear throughout applied mathematics.

1.1.3 Examples of Common Measures

The concept of a measure extends far beyond abstract mathematics. Below are some fundamental examples:

- **Length (Metric Measure):** In Euclidean space, distance functions such as the Euclidean norm define a measure of length:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

- **Probability Measure:** A probability measure P assigns a value between 0 and 1 to events in a sample space Ω, such that:

$$P(\Omega) = 1.$$

- **Energy and Physical Quantities:** In physics, measures exist for energy, force, entropy, and more. For instance, kinetic energy is measured as:

$$E_k = \frac{1}{2}mv^2.$$

- **Similarity and Dissimilarity Measures:** In data science and machine learning, measures of similarity and dissimilarity include:

$$d_{\text{cosine}}(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}, \tag{1}$$

$$D_{\text{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{2}$$

These measures play essential roles in their respective domains, ensuring that comparisons and computations remain meaningful.

1.1.4 The Problem with Intensity Measures

Despite the ubiquity of measures, not all commonly used quantities are formally defined in a rigorous mathematical sense. One such concept is **intensity**. In various fields, intensity is often assumed to be a measure of "strength" or "amount" of a property, such as:

- The brightness of light in physics.

- The activation of a neuron in deep learning.

- The likelihood of a classification decision.

However, unlike distance or probability, intensity lacks a well-defined mathematical structure. It is often treated heuristically, with no universally agreed-upon formalism. This leads to inconsistencies in interpretation, particularly in fields like machine learning, where neural activations are commonly referred to as "intensities" without a rigorous definition.

1.1.5 Transition to the Next Section

The lack of a formal definition for intensity creates inconsistencies in classification confidence, neural activations, and probability scaling. In the next section, we will explore the fundamental differences between distance and intensity, highlighting why intensity must be understood as a function of contrastive distances rather than an independent measure.

1.2 Distance vs. Intensity: The Problem

In many fields, the concepts of *distance* and *intensity* are used interchangeably or without formal distinction. While distance is rigorously defined in mathematics, intensity is often treated as an intuitive, heuristic measure of strength or presence. This lack of formalism creates inconsistencies, particularly in classification systems where confidence and decision-making rely on well-defined measures.

To address these inconsistencies, we introduce the concept of **competitive distance** early in this section. Competitive distance provides a structured approach to defining classification confidence by comparing relative distances rather than relying on heuristic intensity values.

1.2.1 Distance: A Well-Defined Measure

Distance is one of the most fundamental concepts in mathematics. A properly defined distance function (or metric) satisfies key properties such as non-negativity, symmetry, and the triangle inequality. Some common distance measures include:

- **Euclidean Distance:** Measures the straight-line separation between two points:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

- **Mahalanobis Distance:** Adjusts for correlations and feature variances in multidimensional spaces:

$$D_M(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}.$$

- **Cosine Distance:** Measures angular separation between vectors rather than absolute magnitude:

$$d_{\text{cosine}}(x, y) = 1 - \frac{x \cdot y}{\|x\| \|y\|}.$$

These measures provide a **geometric basis for similarity and dissimilarity** in various mathematical spaces.

1.2.2 Competitive Distance: A More Rigorous Alternative to Intensity

Traditional intensity-based measures suffer from several shortcomings, particularly in classification and decision-making systems. Instead of using heuristic intensity measures, we define **competitive distance**, which quantifies classification confidence in terms of the relative separation between competing class distances.

Competitive distance is defined as:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)),$$

where:

- $D_c(x)$ is the distance from x to the target class prototype.

4

- $D_{\neg c}(x)$ is the distance from x to the nearest non-target class prototype.

- $S_c(x)$ represents a properly defined classification confidence measure based on relative distances.

This formulation ensures that classification confidence:

- Is grounded in geometric relationships between classes.

- Remains bounded and interpretable.

- Naturally aligns with decision boundaries and uncertainty estimation.

1.2.3 Why Intensity is a Concept Without Formal Definition

Unlike distance, intensity is often used informally to describe the strength of a feature, probability, or classification confidence. Examples include:

- **Brightness in Physics:** The intensity of light is often equated with amplitude squared.

- **Neural Network Activations:** The magnitude of an activation is assumed to correspond to the presence of a feature.

- **Classification Confidence:** Softmax outputs are interpreted as confidence scores, but their underlying mathematical meaning is ambiguous.

Unlike distance, intensity lacks a universal definition, leading to ambiguity in how it should be interpreted across different domains.

1.2.4 Why the Traditional View of Intensity is Problematic

Interpreting intensity as an independent measure leads to several problems:

1. **Lack of Scale Consistency:** Intensity is often defined in arbitrary units, making comparisons across different models or datasets difficult.

2. **Unbounded Growth:** If intensity were simply a negation of distance, then maximizing intensity would drive distances toward negative infinity, which is not meaningful.

3. **Overconfidence in Classification:** Traditional activation-based confidence measures can exaggerate certainty due to improper scaling.

4. **Lack of Decision Boundary Awareness:** Intensity-based confidence does not inherently reflect how close a sample is to a decision boundary.

1.2.5   The Need for a Competitive Distance Framework

Since intensity lacks a foundational definition, we propose redefining it in terms of *contrastive distance*. Instead of treating intensity as an absolute measure, we introduce a framework where classification confidence is derived from the **difference between distances to competing classes**:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)).$$

This formulation ensures that classification confidence is not an arbitrary activation magnitude but a function of how much closer an input is to one class compared to others. It naturally prevents unbounded confidence values, aligns with decision boundaries, and improves uncertainty estimation.

1.2.6   Conclusion and Transition

The lack of a formal definition for intensity has led to inconsistencies in classification confidence, neural activations, and probability scaling. By reframing intensity in terms of **competitive distance**, we provide a rigorous and interpretable measure that resolves these issues. The next section will outline the goals of this chapter and the key insights that will be developed in later sections.

1.3   What This Chapter Will Do

The previous section highlighted a fundamental issue in many classification and measurement systems: while **distance** is a rigorously defined mathematical concept, **intensity** lacks a formal foundation. This lack of definition has led to inconsistencies in fields ranging from machine learning to cognitive science, particularly in how classification confidence and feature representations are interpreted.

This chapter aims to resolve these issues by establishing a well-defined framework for intensity based on **competitive distance**. Specifically, we will:

1. **Define intensity as a function of contrastive distance.** We will show that intensity should not be treated as an absolute measure but rather as a transformation of relative distances between competing classes.

2. **Reinterpret classification confidence.** By analyzing how classification systems operate, we will demonstrate that confidence is best understood as a function of how much closer an input is to one class than to others.

3. **Provide a rigorous mathematical foundation.** We will formally establish how competitive distance satisfies necessary mathematical properties and aligns with well-defined concepts in metric spaces.

4. **Explain why softmax works.** By reinterpreting softmax as a competitive distance transformation, we will clarify why neural networks behave as they do and how confidence scores emerge.

5. **Propose improvements for machine learning models.** We will explore how explicitly incorporating competitive distance into neural networks can enhance robustness, calibration, and adversarial resistance.

6. **Bridge artificial intelligence with cognitive science.** We will show that the competitive distance framework aligns with human perception, decision-making, and category learning, suggesting that it may be a fundamental principle of intelligence.

1.3.1 Chapter Structure

To achieve these goals, this chapter is structured as follows:

- **Section 2: The Foundation of Distance Measures**
  We establish the mathematical properties of distance measures and explain why distance is the fundamental quantity underlying similarity and classification.

- **Section 3: The Problem with Intensity**
  We demonstrate why intensity cannot be an independent measure, showing that absolute activation values and naive intensity interpretations lead to inconsistencies.

- **Section 4: Competitive Distance and Classification**
  We introduce a formal definition of intensity based on contrastive distance, derive decision boundaries, and explain classification confidence as a function of competitive separation.

- **Section 5: Implications and Applications**
  We explore how this framework improves neural network training, enhances robustness to adversarial attacks, and aligns with biological cognition.

- **Section 6: Conclusion and Future Research**
  We summarize key insights and propose future research directions in machine learning, neuroscience, and theoretical measurement.

1.3.2 A Paradigm Shift: From Absolute Activation to Competitive Distance

This chapter challenges the traditional view that classification systems operate based on absolute activation values or intensity measures. Instead, we propose a paradigm shift: **classification is a competitive process, and intensity should be defined as a function of relative distances between class representations.**

By the end of this chapter, we will have established a new theoretical foundation for classification confidence, one that is mathematically rigorous, biologically plausible, and practically useful for improving artificial intelligence and cognitive modeling.

2 The Foundation of Distance Measures

2.1 Defining Distance Formally

Distance is a fundamental measure in mathematics and applied sciences, quantifying the separation between two points in space. Unlike intensity,

which lacks a universal mathematical definition and is often used heuristically, distance is rigorously defined through metric space axioms. This section introduces the formalism of distance measures, explores different types of distances used in classification, and lays the groundwork for understanding competitive distance as a critical concept in classification confidence.

2.1.1 Metric Spaces and Distance Functions

A **metric space** is a set X equipped with a function $d : X \times X \to \mathbb{R}$ that satisfies specific properties ensuring a consistent measure of separation. This function $d(x, y)$, called a **metric**, must satisfy the following axioms for all $x, y, z \in X$:

1. **Non-negativity:** The distance between any two points is always non-negative:
$$d(x, y) \geq 0.$$

2. **Identity of Indiscernibles:** The distance between a point and itself is zero, and only identical points have zero distance:
$$d(x, y) = 0 \iff x = y.$$

3. **Symmetry:** The distance from x to y is the same as from y to x:
$$d(x, y) = d(y, x).$$

4. **Triangle Inequality:** The direct distance between two points is never greater than the sum of distances through an intermediate point:
$$d(x, z) \leq d(x, y) + d(y, z).$$

These properties ensure that distance behaves predictably and supports various mathematical operations, such as clustering, nearest-neighbor searches, and decision boundary calculations.

2.1.2 Common Distance Metrics

Several distance metrics are widely used in mathematics, data science, and machine learning. Each has unique properties that make it suitable for different applications:

Euclidean Distance The **Euclidean distance** is the most familiar metric, representing the straight-line distance between two points:

$$d_{\text{Euclidean}}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

It is commonly used in clustering, nearest-neighbor methods, and gradient-based optimization.

Manhattan Distance The **Manhattan distance** (or L_1-norm) measures distance along coordinate axes:

$$d_{\text{Manhattan}}(x, y) = \sum_{i=1}^{n}|x_i - y_i|.$$

This is useful in grid-based movement applications such as urban navigation and certain machine learning models where axis-aligned movement is relevant.

Mahalanobis Distance The **Mahalanobis distance** accounts for correlations between variables and adapts to different feature scales:

$$D_M(x, y) = \sqrt{(x - y)^T \Sigma^{-1}(x - y)},$$

where Σ is the covariance matrix of the dataset. This metric is widely used in anomaly detection and metric learning.

Cosine Distance The **cosine distance** measures the angular difference between two vectors:

$$d_{\text{cosine}}(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}.$$

It is commonly used in text analysis and information retrieval, where directional similarity matters more than magnitude.

Kullback-Leibler Divergence (KL-Divergence) Unlike the previous metrics, KL-divergence is a measure of how one probability distribution P diverges from another reference distribution Q:

$$D_{\mathrm{KL}}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

Since it does not satisfy the triangle inequality, KL-divergence is not a true metric but is essential in probabilistic modeling and information theory.

2.1.3 Distance as the Fundamental Measure

A key property of distance is its universality: almost all similarity measures can be derived from distance functions, either through negation or transformation. Intensity measures, often used in classification, are typically functions of distance rather than independent quantities.

Introducing Competitive Distance A crucial insight in classification tasks is that intensity (e.g., neural activations) should be understood not as an absolute quantity but as a measure of **relative separation** between competing classes. This leads to the concept of **competitive distance**, which will be formalized in later sections.

Instead of simply measuring distance to a single class prototype $D_c(x)$, competitive distance considers the separation between multiple competing classes:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)).$$

where: - $D_c(x)$ is the distance from x to the target class. - $D_{\neg c}(x)$ is the distance from x to the nearest competing class. - $S_c(x)$ represents intensity, now framed as a function of contrastive distances.

This framework provides a geometrically grounded approach to classification, improving interpretability and robustness in machine learning models. The next section will explore why intensity must be derived from contrastive distances rather than treated as an independent measure.

2.2 Distance as the Foundational Measure

Distance is one of the most rigorously defined and universally applicable measures in mathematics and applied sciences. Unlike intensity, which is often

treated as a heuristic quantity without a formal definition, distance satisfies well-defined mathematical properties and serves as the basis for measuring similarity, classification, and decision-making.

This section establishes why distance should be considered the primary measure underlying classification and why intensity must be derived from it rather than treated as an independent quantity.

2.2.1 Why Distance is a Fundamental Concept

A well-defined measure should be:

1. **Mathematically Rigorously Defined:** A proper measure must satisfy axioms that ensure consistent behavior across different contexts.

2. **Universally Applicable:** The measure should generalize across different spaces, domains, and problem types.

3. **Meaningful in Decision-Making:** The measure should be interpretable and provide actionable information.

Distance satisfies all three criteria. It is rigorously defined through the principles of metric spaces, appears in diverse fields from physics to machine learning, and provides meaningful comparisons between entities.

2.2.2 Distance and Similarity: Two Sides of the Same Concept

Distance functions are typically used to quantify dissimilarity, but they also inherently encode similarity. Given a distance function $d(x, y)$, we can define similarity as its inverse:

$$s(x, y) = \frac{1}{1 + d(x, y)}.$$

Thus, rather than defining similarity and intensity as independent concepts, we recognize that they are transformations of a single underlying measure—distance. This unification helps eliminate inconsistencies in classification models.

2.2.3 The Role of Distance in Classification

In classification tasks, distance is fundamental to decision boundaries and confidence estimation. Consider a classifier that assigns an input x to one of K possible classes, each represented by a prototype or decision region.

- The class assignment is determined by the smallest distance:

$$c^* = \arg\min_c D_c(x).$$

- Classification confidence is inherently a function of **contrastive distances**—how much closer x is to one class than to others.

This naturally leads to the concept of competitive distance.

Competitive Distance as a Classification Framework   Instead of treating classification confidence as an absolute function of activation magnitudes, we redefine it as a function of relative distances:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)),$$

where $D_c(x)$ is the distance to the target class and $D_{\neg c}(x)$ is the distance to the closest competing class. This ensures that classification is always based on relative separations rather than absolute activation values.

2.2.4 Distance and Decision Boundaries

Decision boundaries separate different class regions in feature space. These boundaries emerge naturally from distance-based classification:

- The decision boundary between two classes c_1 and c_2 occurs where their distances are equal:
$$D_{c_1}(x) = D_{c_2}(x).$$

- More generally, for multi-class problems, each class region consists of all points closest to a given prototype.

This formalism reinforces that classification is inherently a process of comparing distances, further proving that distance, not intensity, is the fundamental measure.

2.2.5 Distance-Based Measures in Machine Learning

Distance-based reasoning is widely used in machine learning:

- **Nearest Neighbor Classification:** Classifies an input based on the class of the closest training example.

- **Support Vector Machines (SVMs):** Maximize the margin (distance) between class decision boundaries to improve generalization and classification confidence.

- **Metric Learning:** Learns a distance function that best separates different categories based on the data, often optimizing for discriminative power.

These examples highlight how classification models operate by leveraging distance, often implicitly. Understanding this foundation allows us to redefine classification confidence as a competitive distance measure.

2.2.6 Conclusion

Distance is a rigorously defined, foundational measure that underlies classification, similarity, and decision boundaries. The next section will demonstrate why intensity cannot be treated as an independent quantity and must instead be formulated as a function of contrastive distances.

3 The Problem with Intensity: Why It Must Be Defined Relative to Distance

3.1 The Illusion of Intensity

In many scientific and engineering disciplines, the term *intensity* is used informally to describe the strength or presence of a property. In physics, intensity often refers to measurable quantities such as the amplitude of a wave or the power of a signal. In deep learning, activation magnitudes are frequently interpreted as indicating the "strength" of a feature. However, despite its widespread use, intensity lacks a rigorous mathematical definition. This section explores why intensity is often an illusion—an artifact of how distance-based measures are interpreted rather than a fundamental property.

3.1.1 Distance vs. Intensity: A Fundamental Distinction

To clarify why intensity is problematic as an independent measure, it is essential to contrast it with distance:

- **Distance:** A well-defined mathematical measure satisfying metric space axioms (e.g., Euclidean, Mahalanobis distance).

- **Intensity:** A heuristic measure that lacks universal definition and consistency.

This distinction is crucial: while distances are rigorously defined and measurable, intensity is often arbitrarily assigned and scale-dependent.

3.1.2 Intensity as a Heuristic Concept

Unlike well-defined measures such as distance, probability, or energy, intensity is often assumed rather than derived. Consider common uses of intensity:

- **Physics:** The brightness of light is proportional to the square of the wave amplitude.

- **Neural Networks:** The magnitude of an activation is often assumed to represent feature strength.

- **Probability and Classification:** A higher softmax output is interpreted as greater confidence in a class.

Each of these cases suggests that intensity is a function of some underlying measure, yet no universal formulation exists. In particular, **the assumption that intensity corresponds to feature strength is misleading**, especially in machine learning.

3.1.3 Neural Activations Do Not Represent Absolute Feature Strength

In deep learning, neurons process information by transforming input data through activation functions. The magnitude of a neuron's activation is often interpreted as reflecting the presence or strength of a learned feature. However, this interpretation is problematic for several reasons:

1. **Relative, Not Absolute:** Neural networks make decisions based on *relative* activations rather than absolute values. A large activation in isolation does not indicate feature presence unless it is significantly larger than activations of competing neurons.

2. **Softmax Normalization:** In classification models, final activations are transformed through a softmax function:

$$p_c = \frac{e^{z_c}}{\sum_{c'} e^{z_{c'}}}$$

where z_c are the logits (pre-softmax activations). This transformation ensures that class scores are interpreted in a *relative* manner—an individual logit has no inherent meaning without comparison to others.

3. **Empirical Evidence from Perturbation Studies:** Studies show that **small adversarial perturbations can drastically change classification output**, even when they cause minimal change in activation magnitudes. This suggests that classification is not driven by feature "intensity" but rather by **how distances between class representations are modified**.

These observations suggest that the perception of intensity as an independent measure is an illusion. Instead, activation values should be interpreted through the lens of distance and competition between classes.

3.1.4   A Thought Experiment: Bigfoot and the Intensity Fallacy

To illustrate the illusion of intensity, consider the classic analogy of searching for Bigfoot. Suppose a group of researchers sets out to find evidence of Bigfoot in the forest. They define a set of features that might indicate Bigfoot's presence, such as large footprints, unusual hair samples, and distorted vocalizations.

A naive intensity-based approach suggests that finding a very large footprint is strong evidence of Bigfoot. However, this reasoning is flawed:

1. **Footprint Size Alone Is Not Evidence:** Large footprints could be caused by bears, weather effects, or hoaxes.

2. **The Strength of Evidence Comes from Competitive Comparison:** The key question is not "how large is this footprint?" but rather "how much more likely is this footprint to belong to Bigfoot than to any other known animal?"

3. **Classification is Relative, Not Absolute:** Without considering competing explanations, any measure of "Bigfoot-ness" is meaningless.

This parallels neural network classification: an activation's magnitude only matters in comparison to alternative activations. Just as a single footprint does not prove Bigfoot's existence, a high activation does not prove that a feature is truly present—it only suggests that the input is closer to one learned representation than another.

3.1.5 Transitioning from Intensity to Competitive Distance

Since intensity lacks an independent mathematical definition, we propose a shift in perspective: rather than viewing intensity as feature presence, it should be understood as a function of **competitive distances**. That is, instead of defining intensity as a direct activation value, it should be derived from the difference between distances to competing class representations:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x))$$

where:

- $D_c(x)$ is the distance of input x to the target class representation.

- $D_{\neg c}(x) = \min_{c' \neq c} D_{c'}(x)$ is the distance to the nearest non-target class.

- $S_c(x)$ is the intensity measure, which should be formulated as a function f of these relative distances.

This redefinition eliminates the illusion of intensity and provides a rigorous foundation for understanding classification confidence. In the next section, we will formalize this framework and show why intensity must be defined in terms of contrastive distance rather than absolute activations.

3.2 Why Intensity Cannot Be a Direct Negation of Distance

Given that intensity lacks an independent mathematical definition, one might propose a simple solution: define intensity as the negation of distance. That is, for a given class c, we could define an intensity function $S_c(x)$ as:

$$S_c(x) = -D_c(x)$$

where $D_c(x)$ is the distance from the input x to the prototype or learned representation of class c.

While this approach is intuitive, it leads to pathological behavior. In this section, we explore why intensity cannot be directly defined as the negation of distance and why a contrastive formulation is required.

3.2.1 The Unbounded Growth Problem

If intensity were simply the negation of distance, then maximizing $S_c(x)$ would require minimizing $D_c(x)$. However, in most practical settings, distances are non-negative and unbounded. This leads to the following issue:

- **As distance decreases, intensity increases, but with no upper bound.**

- **As distance grows, intensity becomes arbitrarily negative.**

- **A model optimizing for maximum intensity would push distances toward negative infinity, which is nonsensical.**

This is in direct contrast to how classification systems work in practice. In classification problems, confidence (which is often interpreted as intensity) does not increase indefinitely but rather saturates when an input is clearly closer to one class than another.

3.2.2 Competitive Distance is More Informative Than Absolute Distance

A deeper issue with defining intensity as $S_c(x) = -D_c(x)$ is that it ignores the *relative* structure of class distances. In a multi-class setting, classification is not about absolute closeness to a single prototype but about how much closer an input is to one class than to competing classes.

Consider the case of three class prototypes, c_1, c_2, and c_3, where an input x has distances:

$$D_{c_1}(x) = 5, \quad D_{c_2}(x) = 10, \quad D_{c_3}(x) = 20.$$

If we define intensity as $S_c(x) = -D_c(x)$, then the intensities for these classes would be:

$$S_{c_1}(x) = -5, \quad S_{c_2}(x) = -10, \quad S_{c_3}(x) = -20.$$

This suggests that x most strongly belongs to c_1, which is reasonable. However, consider a new input where all distances are doubled:

$$D_{c_1}(x') = 10, \quad D_{c_2}(x') = 20, \quad D_{c_3}(x') = 40.$$

Now, the corresponding intensities are:

$$S_{c_1}(x') = -10, \quad S_{c_2}(x') = -20, \quad S_{c_3}(x') = -40.$$

Even though the relationships between the classes remain the same, the absolute intensity values have changed significantly. This suggests that absolute distance is not the key factor in classification—**what matters is the relative difference between distances**.

3.2.3 A More Meaningful Formulation: Contrastive Distance

Instead of defining intensity as a simple negation of distance, we propose defining it as a function of *contrastive distance*, which captures the competitive relationship between different classes.

$$S_c(x) = f(D_{\neg c}(x) - D_c(x))$$

where:

- $D_c(x)$ is the distance from x to the target class prototype.

- $D_{\neg c}(x) = \min_{c' \neq c} D_{c'}(x)$ is the distance to the nearest non-target class.

- f is a transformation function ensuring appropriate scaling and boundedness.

This formulation guarantees that intensity:

- Is **bounded**—it does not grow arbitrarily large.

- Is **relative**—it depends on both the target and competing classes.

- Preserves the decision boundary—classification occurs when $S_c(x)$ is maximized, meaning x is closest to class c.

3.2.4 Ensuring Proper Scaling: Choosing f

The function f should be chosen to ensure stability and interpretability. Several common transformations achieve this:

- **Linear Difference:**

$$S_c(x) = D_{\neg c}(x) - D_c(x).$$

This preserves relative relationships while ensuring a contrastive interpretation.

- **Exponential Scaling (Softmax-like Behavior):**

$$S_c(x) = e^{-(D_c(x) - D_{\neg c}(x))}.$$

This ensures that the intensity remains positive and smoothly varies with distance differences.

- **Inverse Transformation (Similarity Scaling):**

$$S_c(x) = \frac{1}{1 + D_c(x) - D_{\neg c}(x)}.$$

This formulation keeps intensity in the range $(0, 1]$ and provides a natural probability-like interpretation.

3.2.5 Conclusion

Attempting to define intensity as $S_c(x) = -D_c(x)$ leads to unbounded values and does not capture the competitive nature of classification. Instead, intensity should be understood as a *contrastive measure*, computed from the relative distances between class prototypes.

This insight provides a rigorous foundation for understanding classification confidence. The next section will fully formalize this competitive distance framework and demonstrate its implications for machine learning and cognitive science.

3.3 The True Definition: Intensity as Competitive Distance

Having established that intensity cannot be a direct negation of distance, we now develop a formal definition based on contrastive distance. Instead of viewing intensity as an intrinsic property, we define it as a measure derived from the relative distances between an input and competing class representations. This framework ensures that intensity is well-defined, bounded, and meaningful in classification settings.

3.3.1 Reframing Intensity: A Function of Contrastive Distance

In a classification system, an input x is mapped to a class c based on its relationship to learned class representations. Instead of defining intensity as $S_c(x) = -D_c(x)$, we propose that intensity should be a function of the difference between the distance to the target class and the distance to the closest competing class:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)).$$

where:

- $D_c(x)$ is the distance from x to the target class representation.

- $D_{\neg c}(x) = \min_{c' \neq c} D_{c'}(x)$ is the distance to the closest non-target class.

- f is a transformation function that ensures appropriate scaling and interpretability.

This definition guarantees that intensity:

- **Captures class separability:** A higher intensity indicates that x is significantly closer to class c than any alternative.

- **Is bounded and stable:** Unlike raw distances, which can grow arbitrarily large, intensity is naturally constrained by the contrast between competing distances.

- **Naturally aligns with classification confidence:** In probabilistic classification, confidence scores are often functions of relative distances, making this formulation a direct theoretical justification for softmax-like behavior.

3.3.2 Choosing an Appropriate Transformation Function f

The function f should ensure that intensity remains meaningful and interpretable. Several choices exist:

- **Linear Contrast:**

$$S_c(x) = D_{\neg c}(x) - D_c(x).$$

  This preserves the raw difference in distances, maintaining contrast while ensuring a simple interpretation.

- **Softmax Scaling:**

$$S_c(x) = e^{-(D_c(x) - D_{\neg c}(x))}.$$

  This ensures that small distance differences produce smoothly varying intensities and prevents extreme variations.

- **Inverse Distance Scaling:**

$$S_c(x) = \frac{1}{1 + D_c(x) - D_{\neg c}(x)}.$$

  This formulation constrains intensity to $(0, 1]$, offering a probability-like interpretation.

- **Max-Normalization:**

$$S_c(x) = \frac{D_{\neg c}(x) - D_c(x)}{\max(D_{\neg c}(x), D_c(x))}.$$

  This normalizes intensity relative to the largest relevant distance, preventing extreme scaling effects.

Each of these functions ensures that intensity remains a measure of class separability rather than absolute feature strength.

3.3.3   Interpretation: Intensity as Decision Boundary Margin

Under this framework, classification confidence is directly tied to how well-separated an input is from competing classes. The decision boundary occurs when:

$$D_c(x) = D_{\neg c}(x).$$

At this point, intensity reaches a neutral state, meaning that the input is equidistant between two classes. This provides a direct geometric explanation for decision regions in neural networks:

- When $D_c(x) < D_{\neg c}(x)$, intensity is positive, indicating confident classification.

- When $D_c(x) > D_{\neg c}(x)$, intensity is negative, suggesting that another class is closer.

- When $D_c(x) \approx D_{\neg c}(x)$, intensity is near zero, indicating uncertainty.

3.3.4   Connections to Existing Classification Systems

Softmax Logits as Competitive Distance   Many classification systems, including neural networks, use softmax to produce class probabilities:

$$p_c = \frac{e^{z_c}}{\sum_{c'} e^{z_{c'}}}.$$

where z_c is the logit for class c. In our framework, logits can be interpreted as contrastive distances:

$$z_c = -D_c(x) + D_{\neg c}(x).$$

This reinterpretation suggests that neural networks implicitly learn a distance-based representation, where softmax is simply a nonlinear transformation of contrastive distance.

Margin-Based Classification   Support vector machines (SVMs) use margin-based classification, where a decision boundary is determined by maximizing separation. Our framework aligns with this principle, as intensity naturally encodes the margin between competing distances.

Probabilistic Models In Gaussian mixture models (GMMs), classification is based on likelihood ratios, which can be reinterpreted as contrastive distance comparisons. This further supports our claim that classification confidence is inherently a function of relative distances.

3.3.5 Implications for Neural Network Training

Redefining intensity as competitive distance suggests potential improvements in training and architecture design:

- **Alternative loss functions:** Instead of cross-entropy, loss functions could explicitly model contrastive distances.

- **Robust classification:** Systems that track $D_{\neg c}(x)$ explicitly could be more robust to adversarial attacks.

- **Improved interpretability:** Feature importance could be analyzed in terms of competitive distances rather than absolute activations.

3.3.6 Conclusion

Intensity is not a direct measure of feature strength, nor can it be naively defined as the negation of distance. Instead, we define it as a function of contrastive distance, ensuring that it is bounded, stable, and meaningful. This framework provides a rigorous foundation for understanding classification confidence, bridging the gap between metric learning, probabilistic classification, and deep learning.

The next section will explore how this framework naturally explains classification confidence, decision boundaries, and network activations.

4 Competitive Distance and Classification

4.1 The Real Meaning of Classification Confidence

Traditional interpretations of classification confidence suggest that larger activations in neural networks correspond to stronger feature presence. However, our competitive distance framework provides a more rigorous explanation: classification confidence is not a measure of absolute activation magnitude but rather a function of relative distances between an input and competing class prototypes.

In this section, we redefine classification confidence as a function of contrastive distance and show how this naturally explains decision boundaries, uncertainty, and neural activations.

4.1.1 Confidence as a Function of Competitive Distance

In a multi-class classification setting, a model assigns a confidence score $S_c(x)$ to each class c. Instead of viewing this as an absolute measure of feature strength, we define it in terms of the contrast between the distance to the target class and the distance to the closest competing class:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)).$$

where:

- $D_c(x)$ is the distance from input x to the prototype or learned representation of class c.

- $D_{\neg c}(x) = \min_{c' \neq c} D_{c'}(x)$ is the distance to the closest non-target class.

- f is a transformation function that ensures appropriate scaling.

This formulation provides a direct interpretation of confidence:

- **High confidence:** $S_c(x)$ is maximized when $D_c(x)$ is much smaller than $D_{\neg c}(x)$, meaning that x is clearly closer to class c than to any alternative.

- **Low confidence:** $S_c(x)$ is near zero when $D_c(x) \approx D_{\neg c}(x)$, meaning that x is near a decision boundary and classification is uncertain.

- **Misclassification:** If $D_c(x) > D_{\neg c}(x)$, then $S_c(x)$ is negative, indicating that x is closer to another class.

4.1.2 Decision Boundaries and Uncertainty

A key property of this framework is that classification confidence naturally reflects decision boundaries. The boundary between two classes c_1 and c_2 occurs where:

$$D_{c_1}(x) = D_{c_2}(x).$$

At this point:

$$S_{c_1}(x) = S_{c_2}(x) = 0.$$

This aligns with standard decision theory: when two competing classes are equidistant from x, the classification decision is maximally uncertain.

Beyond binary classification, the framework extends to multi-class problems, where decision boundaries form Voronoi-like regions in feature space. The class assigned to an input x is the one for which $S_c(x)$ is maximized, ensuring that each decision is based on the contrastive distance to competing classes.

4.1.3   Softmax Reinterpreted as Competitive Distance Scaling

Softmax is widely used to convert logits into probability-like confidence scores:

$$p_c = \frac{e^{z_c}}{\sum_{c'} e^{z_{c'}}}.$$

where z_c represents the logit for class c. Under our framework, logits can be directly interpreted as contrastive distances:

$$z_c = -D_c(x) + D_{\neg c}(x).$$

This shows that softmax is not computing confidence based on raw activations but rather on the difference between distances to competing classes. The exponentiation in softmax serves to ensure smooth transitions between probabilities and prevents negative values from dominating the classification decision.

This reinterpretation also clarifies why softmax confidence can be overconfident in high-dimensional spaces: when distances between class prototypes are large, the difference $D_{\neg c}(x) - D_c(x)$ becomes exaggerated, leading to probability distributions that favor one class disproportionately.

4.1.4   Geometric Interpretation of Activation Space

When visualizing activations in feature space, our framework provides a clearer understanding of neural network behavior:

- **Class Prototypes Form Distance-Based Regions:** Each learned class representation serves as a prototype, and classification occurs based on proximity to these points.

- **High-Confidence Regions Are Separated by Large Distance Differences:** Inputs far from decision boundaries have large contrastive distance differences, resulting in high classification confidence.

- **Low-Confidence Regions Exist Near Decision Boundaries:** Inputs near a boundary have small contrastive distance differences, making them more sensitive to perturbations.

- **Misclassified Inputs Have Negative Contrastive Intensity:** Inputs incorrectly classified are those where the model has a stronger distance match with a non-target class.

This interpretation naturally aligns with findings in adversarial robustness: small perturbations that move an input across a decision boundary shift the contrastive distance balance, leading to misclassification.

4.1.5 Implications for Model Robustness and Calibration

Understanding classification confidence as a function of contrastive distance provides new insights into:

- **Calibration:** Many models are overconfident because softmax exaggerates small differences in contrastive distance. Adjusting $f(D_{\neg c}(x) - D_c(x))$ could improve calibration.

- **Adversarial Defense:** Tracking both $D_c(x)$ and $D_{\neg c}(x)$ explicitly could improve robustness against attacks that push inputs across decision boundaries.

- **Uncertainty Estimation:** Instead of relying on entropy of softmax probabilities, uncertainty can be directly measured using the raw difference in class distances.

4.1.6 Conclusion

We have redefined classification confidence as a function of competitive distance, providing a more rigorous explanation for decision boundaries, uncertainty, and neural activations. This framework naturally explains why softmax probabilities behave as they do and suggests new approaches for improving model calibration and robustness.

The next section will explore why this formulation prevents pathological behavior and ensures stability in classification decisions.

4.2 Why This Definition Prevents Pathological Behavior

A well-defined measure of classification confidence must be both stable and meaningful across a range of inputs. Traditional interpretations of intensity, which rely on absolute activations, lead to pathological behaviors such as unbounded growth, overconfidence, and susceptibility to adversarial perturbations. By redefining intensity as a function of contrastive distance, we eliminate these issues and ensure a principled and robust classification framework.

4.2.1 The Pathologies of Absolute Activation-Based Confidence

If classification confidence were purely a function of absolute activation values, we would encounter several well-documented issues:

1. **Unbounded Growth:** If confidence were directly defined as $S_c(x) = -D_c(x)$, optimizing for confidence would drive distances toward negative infinity, which is meaningless in practical scenarios.

2. **Overconfidence in High-Dimensional Spaces:** Softmax-based classification often leads to extreme confidence scores due to the exponential transformation of logits. When feature spaces have high dimensionality, even small absolute activation differences translate into near-certain probabilities, making models excessively confident.

3. **Sensitivity to Feature Scaling:** If activations encode absolute intensity, then uniform rescaling of all activations should not affect classification. However, empirical evidence suggests that such scaling does

alter confidence scores, indicating that absolute activations alone are insufficient for meaningful classification.

4. **Vulnerability to Adversarial Attacks:** Small, imperceptible changes in input space can drastically shift activations, leading to confident misclassification. This suggests that models are not truly measuring intrinsic feature strength, but rather responding to small shifts in contrastive distances.

These issues arise because absolute activations do not inherently encode class separability. A more robust approach must account for the competitive nature of classification.

4.2.2 Bounding Confidence with Contrastive Distance

By defining classification confidence as a function of contrastive distance:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)),$$

we introduce a bounded and stable formulation that avoids pathological behavior.

4.2.3 Preventing Unbounded Growth

Since confidence is now computed as a difference between distances, it remains constrained within meaningful limits:

$$-\infty < D_{\neg c}(x) - D_c(x) < \infty.$$

Unlike absolute activations, which can grow indefinitely, the contrastive distance difference is naturally bounded by the separability of class prototypes in the feature space.

4.2.4 Mitigating Overconfidence in High-Dimensional Spaces

In high-dimensional settings, absolute activations tend to spread apart, causing naive softmax probabilities to exaggerate confidence. However, under the contrastive distance formulation, confidence depends only on *relative* class distances, not their absolute magnitudes. This prevents extreme confidence shifts due to high-dimensional scaling effects.

4.2.5 Ensuring Scale-Invariance

A fundamental requirement of a well-behaved classification confidence measure is that it should remain unchanged under uniform scaling transformations. If all activations were multiplied by a constant factor α, then under the naive intensity interpretation:

$$S'_c(x) = \alpha S_c(x),$$

which arbitrarily alters classification confidence.

In contrast, contrastive distance formulations remain invariant under uniform scaling, since both $D_c(x)$ and $D_{\neg c}(x)$ scale together:

$$S'_c(x) = f(D_{\neg c}(x) - D_c(x)) = S_c(x).$$

This ensures that confidence is derived purely from the competitive relationships between classes, rather than from extrinsic scaling factors.

4.2.6 Robustness Against Adversarial Perturbations

One of the most striking failures of traditional classification systems is their susceptibility to adversarial attacks—small, carefully crafted perturbations that shift activations enough to cause misclassification. Many such attacks exploit the fact that classification confidence is overly dependent on small absolute activation changes rather than true class separability.

Under the contrastive distance framework:

- **Perturbations must significantly alter relative distances to affect confidence.** Small, local perturbations that do not meaningfully affect contrastive distances will not induce overconfidence or misclassification.

- **Confidence degradation is a function of decision boundary movement.** Rather than treating activations as independent measures of class presence, the framework explicitly tracks how distances to competing classes change. This makes adversarial manipulations easier to detect.

- **Classification is more stable to input-space transformations.** If an adversarial attack distorts feature space without substantially shifting relative distances, confidence remains unchanged.

This suggests that neural networks implementing explicit competitive distance tracking may be inherently more robust to adversarial perturbations than those relying on traditional activation-based confidence measures.

4.2.7 A Well-Behaved Classification Confidence Measure

To summarize, defining classification confidence in terms of contrastive distance ensures that:

- **Confidence remains bounded,** preventing infinite growth.

- **Overconfidence in high dimensions is mitigated,** making models more calibrated.

- **Scale-invariance is preserved,** ensuring robustness to activation transformations.

- **Adversarial perturbations must alter relative distances,** making confidence shifts more interpretable.

This formulation provides a principled approach to defining confidence that aligns with both theoretical and empirical requirements for well-behaved classification systems.

4.2.8 Conclusion

By replacing absolute activation-based confidence with contrastive distance, we eliminate key failure modes of existing classification frameworks. The next section will explore how this formulation translates into practical implementations, including probability scaling and decision-making in neural networks.

4.3 Transforming Competitive Distance into a Probability Measure

In practical classification systems, confidence values are often transformed into probability-like measures to facilitate decision-making and interpretability. Many existing approaches, such as softmax, implicitly assume that classification confidence is derived from absolute activations. However, our framework, which defines confidence in terms of contrastive distance, provides a more principled way to compute probability scores.

This section explores how contrastive distance can be transformed into a probability measure while maintaining its theoretical advantages.

4.3.1 Probability as a Function of Contrastive Distance

Given an input x, we define classification confidence using contrastive distance:

$$S_c(x) = D_{\neg c}(x) - D_c(x).$$

where:

- $D_c(x)$ is the distance from x to the prototype of class c.

- $D_{\neg c}(x) = \min_{c' \neq c} D_{c'}(x)$ is the distance to the closest non-target class.

To convert this into a probability-like measure, we apply a transformation function f that ensures proper scaling and interpretability.

4.3.2 Common Transformations for Probability Scaling

Several functions can be used to map contrastive distance into a probability space:

4.3.3 Softmax Transformation

The most commonly used probability transformation in neural networks is the softmax function:

$$p_c = \frac{e^{z_c}}{\sum_{c'} e^{z_{c'}}}.$$

Under our framework, we redefine logits as contrastive distances:

$$z_c = -D_c(x) + D_{\neg c}(x).$$

Substituting into softmax:

$$p_c = \frac{e^{-(D_c(x) - D_{\neg c}(x))}}{\sum_{c'} e^{-(D_{c'}(x) - D_{\neg c'}(x))}}.$$

This transformation ensures that probability is a smooth function of relative class separability.

4.3.4 Logistic Sigmoid Scaling

For binary classification, we can use a logistic sigmoid transformation:

$$p_c = \frac{1}{1 + e^{D_c(x) - D_{\neg c}(x)}}.$$

This function maps contrastive distance into the range $(0, 1)$, ensuring a probability-like interpretation without requiring explicit normalization.

4.3.5 Inverse Contrastive Distance

An alternative approach is to define probability as an inverse function of distance:

$$p_c = \frac{1}{1 + D_c(x) - D_{\neg c}(x)}.$$

This formulation is particularly useful for nearest-neighbor-based classification, where probability should decay smoothly with increasing distance.

4.3.6 Decision Boundaries in Probability Space

Regardless of the transformation function used, the decision boundary occurs when:

$$D_c(x) = D_{\neg c}(x).$$

At this point:

$$p_c = \frac{1}{2}, \quad \text{or} \quad p_c = \frac{1}{K} \text{ for multi-class settings.}$$

Thus, probability scores in this framework naturally reflect class separability without requiring heuristic adjustments.

4.3.7 Advantages Over Traditional Probability Scaling

Using contrastive distance for probability scaling provides several advantages:

- **Avoids Arbitrary Scaling:** Unlike raw logits, which can be arbitrarily shifted, contrastive distance ensures probabilities are based on meaningful geometric relationships.

- **Improves Calibration:** Many deep learning models suffer from over-confidence. Since contrastive distance explicitly models separability, probability scores better reflect classification uncertainty.

- **Provides a Geometric Interpretation:** Unlike heuristic activation scaling, probability scores directly correspond to distances in feature space.

4.3.8 Conclusion

Transforming contrastive distance into a probability measure preserves the theoretical rigor of our framework while ensuring interpretability in classification systems. By replacing naive activation-based logits with distance-based confidence, we obtain a more meaningful probabilistic model that improves calibration and decision-making.

The next section explores how this framework aligns with softmax classification and provides insights into improving deep learning architectures.

5 Broader Implications of This Framework

5.1 Why Softmax Works: A Distance Perspective

Softmax is the most widely used function for converting neural network outputs into probability distributions for classification tasks. Traditionally, softmax is interpreted as a way to normalize raw activations (logits) into a probability distribution, where larger activations indicate stronger feature presence. However, our contrastive distance framework provides a more fundamental explanation: softmax works because it approximates a transformation of relative class distances.

In this section, we reinterpret softmax as a competitive distance function, demonstrating how it aligns with our theoretical framework.

5.1.1 Traditional View of Softmax

In standard classification models, the softmax function is applied to logits z_c to compute class probabilities:

$$p_c = \frac{e^{z_c}}{\sum_{c'} e^{z_{c'}}}.$$

where z_c is typically an unbounded activation value produced by the last layer of a neural network.

This formulation ensures that:

- p_c is always positive and sums to 1 across all classes.

- The largest z_c receives the highest probability, reinforcing the idea that larger activations correspond to stronger class confidence.

- The exponential function magnifies differences between logits, exaggerating small activation differences.

However, this interpretation relies on the assumption that logits represent an intrinsic measure of class presence, which, as we have shown, is misleading. Instead, logits can be understood as *contrastive distance functions*.

5.1.2 Reinterpreting Softmax as a Competitive Distance Function

Instead of treating logits as absolute indicators of class strength, we define them in terms of contrastive distance:

$$z_c = -D_c(x) + D_{\neg c}(x).$$

where:

- $D_c(x)$ is the distance to the prototype of class c.

- $D_{\neg c}(x) = \min_{c' \neq c} D_{c'}(x)$ is the distance to the closest competing class.

Substituting this into the softmax equation:

$$p_c = \frac{e^{-(D_c(x) - D_{\neg c}(x))}}{\sum_{c'} e^{-(D_{c'}(x) - D_{\neg c'}(x))}}.$$

This formulation provides a geometric justification for softmax:

- Softmax probability is maximized when $D_c(x)$ is significantly smaller than $D_{\neg c}(x)$, meaning the input is clearly closer to class c than any alternative.

- The exponential function ensures smooth probability transitions, preventing abrupt confidence shifts near decision boundaries.

- The denominator normalizes probabilities across all classes, preserving relative ranking among competing distances.

5.1.3   Decision Boundaries in Softmax

A key property of softmax is that classification is based on which p_c is largest, which occurs when:

$$D_c(x) < D_{\neg c}(x).$$

This aligns with our contrastive distance framework: an input is assigned to the class whose prototype it is closest to. The probability difference between classes reflects the separation between competing distances.

5.1.4   Softmax Overconfidence and High-Dimensional Scaling

One common issue with softmax is that it tends to produce overconfident predictions, particularly in high-dimensional feature spaces. This is because the exponential function exaggerates small differences in logits, making even slightly preferred classes appear overwhelmingly likely.

Under our framework, this can be reinterpreted as a consequence of high-dimensional distance scaling:

- In high-dimensional spaces, feature distances tend to grow larger due to the *curse of dimensionality.*

- Since softmax probabilities are computed using exponentials of these distances, even small relative differences can lead to extreme confidence scores.

- This explains why deep learning models often require calibration techniques to avoid overconfidence.

A potential solution is to modify the softmax transformation by incorporating explicit distance normalization:

$$p_c = \frac{e^{-\lambda(D_c(x) - D_{\neg c}(x))}}{\sum_{c'} e^{-\lambda(D_{c'}(x) - D_{\neg c'}(x))}}.$$

where λ is a scaling factor that adjusts the sensitivity of probability estimates to contrastive distance differences.

5.1.5   Implications for Model Design

Understanding softmax as a function of competitive distance suggests several improvements to model architecture and training:

- **Distance-Based Training Objectives:** Instead of training models to maximize raw activation differences, loss functions could be designed to maximize contrastive class separation.

- **Improved Calibration:** Introducing explicit contrastive distance constraints could help prevent overconfidence in probabilistic predictions.

- **Robust Classification:** By explicitly tracking relative distances between competing classes, models may be more resistant to adversarial perturbations that attempt to manipulate activation magnitudes.

5.1.6   Conclusion

Softmax works not because it maps raw activations into probabilities, but because it acts as a transformation of contrastive class distances. This reinterpretation aligns with our framework and provides a theoretical foundation for understanding classification confidence.

The next section explores how this competitive distance perspective can improve neural network architectures and classification robustness.

5.2   Improving Neural Networks with Explicit Competitive Distance

Our reinterpretation of classification confidence as a function of contrastive distance suggests potential improvements to neural network architectures and training strategies. By explicitly modeling competitive distance, networks can become more robust, better calibrated, and less prone to adversarial perturbations. This section explores modifications to neural network design that leverage explicit distance-based representations.

5.2.1   Limitations of Current Neural Network Architectures

Traditional neural networks, particularly deep classifiers, operate under the assumption that larger activations indicate stronger feature presence. How-

ever, as we have shown, classification decisions are better understood as a function of competitive distances. This misalignment leads to several issues:

- **Overconfidence in Predictions:** Networks trained using standard softmax and cross-entropy loss tend to produce high-confidence outputs even when predictions are uncertain.

- **Sensitivity to Activation Scaling:** Scaling all activations by a constant should not affect classification decisions, yet empirical evidence shows that softmax-based models exhibit instability under such transformations.

- **Adversarial Vulnerability:** Since neural networks do not explicitly model contrastive distance, adversarial perturbations can effectively shift activation values in a way that changes classification outcomes without significantly altering perceptible features.

These issues suggest that explicitly incorporating competitive distance into network design may lead to improvements in classification robustness and interpretability.

5.2.2 Incorporating Competitive Distance into Neural Representations

We propose modifying neural network architectures to explicitly compute and track competitive distances during training and inference.

5.2.3 Distance-Based Logits

Instead of computing class logits as arbitrary learned feature activations, we redefine them in terms of distance to class prototypes. That is, rather than having the final layer output arbitrary scores z_c, we compute logits as:

$$z_c = -D_c(x) + D_{\neg c}(x),$$

where $D_c(x)$ is the distance to the class prototype and $D_{\neg c}(x)$ is the distance to the closest competing class.

This ensures that:

- Logits directly encode class separability.

- Classification confidence reflects true geometric relationships in feature space.

- Scaling feature activations does not distort classification confidence.

5.2.4 Contrastive Distance Loss Functions

Standard neural networks are trained using the cross-entropy loss:

$$\mathcal{L} = -\sum_c y_c \log p_c.$$

where p_c is computed via softmax. Instead, we propose a distance-aware loss function that explicitly optimizes for class separability:

$$\mathcal{L}_{\text{contrastive}} = \sum_c y_c (D_c(x) - D_{\neg c}(x)).$$

This loss function encourages networks to minimize the distance to the correct class while maximizing the distance to the nearest incorrect class. This has several advantages:

- Encourages larger decision margins, improving robustness.

- Reduces the risk of adversarial misclassification by ensuring inputs are farther from incorrect classes.

- Prevents overconfidence by naturally constraining the scale of logits.

5.2.5 Distance-Normalized Activations

One challenge in competitive distance modeling is ensuring that distances remain on a consistent scale across different inputs. One approach to address this is by normalizing distance calculations using batch statistics. Specifically, we modify the network to output *normalized distances*:

$$\tilde{D}_c(x) = \frac{D_c(x) - \mu_{\text{batch}}}{\sigma_{\text{batch}}}.$$

This transformation ensures that:

- Distances are comparable across different input distributions.

- Class confidence is less sensitive to feature magnitude variations.

- Decision boundaries remain stable across training epochs.

5.2.6 Improving Adversarial Robustness

One of the key weaknesses of traditional neural networks is their vulnerability to adversarial perturbations—small, carefully crafted modifications to input data that cause misclassification. Explicitly incorporating competitive distance into neural networks can improve adversarial robustness by:

1. **Tracking Distance to Decision Boundaries:** By explicitly modeling $D_{\neg c}(x) - D_c(x)$, networks can detect when inputs are near classification boundaries, making them more resistant to adversarial examples.

2. **Detecting Confidence Shifts:** Since adversarial attacks often reduce class separability, monitoring contrastive distance provides a way to flag potentially manipulated inputs.

3. **Enhancing Margin Maximization:** Contrastive loss functions naturally encourage larger class margins, making it harder for adversarial perturbations to flip classification outcomes.

5.2.7 Implications for Network Calibration and Uncertainty Estimation

Traditional deep learning models tend to be poorly calibrated, meaning that softmax probabilities do not accurately reflect classification confidence. Explicitly modeling competitive distance can improve calibration by:

- Replacing arbitrary activation-based logits with well-defined distance measures.

- Using contrastive distance to define uncertainty, rather than relying on heuristic entropy measures.

- Avoiding overconfidence in ambiguous cases by ensuring that probability scores reflect true class separability.

This suggests that distance-aware classification models may provide more reliable confidence estimates, improving real-world applicability in high-stakes decision-making scenarios.

5.2.8 Conclusion

By explicitly incorporating competitive distance into neural network architectures, we can address key weaknesses in traditional classification models. This framework offers a principled way to improve robustness, calibration, and adversarial resistance. The next section explores how this approach extends beyond neural networks to broader cognitive and perceptual systems.

5.3 Cognitive Science and Perception Models

The competitive distance framework we have developed for neural networks aligns with principles found in human cognition and perception. Rather than relying on absolute intensity measures, the human brain appears to make decisions based on relative contrasts between competing categories. This section explores how competitive distance-based classification aligns with cognitive models, perception theories, and neuroscience.

5.3.1 Perception as a Contrastive Process

Human perception is inherently contrastive. Psychophysical studies suggest that perception is not based on absolute intensity, but rather on the relative difference between stimuli. Several key findings support this:

- **Weber's Law:** The perceived change in a stimulus is proportional to the background intensity, meaning perception depends on relative differences rather than absolute values.

- **Lateral Inhibition in Vision:** The human visual system enhances contrast between neighboring regions to emphasize boundaries and edges, rather than responding directly to absolute brightness values.

- **Categorization in Language and Thought:** Cognitive categorization relies on discriminating between similar concepts rather than detecting individual features in isolation.

These findings suggest that the brain, like our competitive distance framework, does not rely on an independent measure of intensity but rather makes decisions based on contrastive relationships.

5.3.2 Similarity Judgments as Distance Comparisons

In cognitive psychology, similarity judgments are central to perception and classification. The **Tversky contrast model** [?] suggests that humans judge similarity not by direct feature matching, but by weighing shared and unique features between compared objects:

$$S(A, B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A),$$

where $S(A, B)$ is the perceived similarity, and f is a function that assigns weight to shared or unique features. This mirrors our contrastive distance formulation:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)).$$

Both models emphasize that classification and recognition are driven by relative comparisons rather than absolute measurements.

5.3.3 Competitive Distance in Neural Representations

Neuroscientific evidence suggests that the brain represents categories using **population coding** and **distributed representations**:

- **Prototype-based representations:** Studies in category learning show that humans classify new stimuli based on similarity to prototype examples, akin to our distance-based class separation model.

- **Opponent Processing:** The human visual system processes color through opponent channels (e.g., red-green, blue-yellow), suggesting that perceptual intensity is encoded as a relative contrast between competing signals rather than as an absolute value.

- **Decision Thresholds in Neural Circuits:** Neural decision-making models, such as **drift-diffusion models (DDMs)**, describe classification as a process of accumulating evidence until a threshold is reached. This aligns with our framework, where classification is determined by the point at which $D_c(x)$ is sufficiently smaller than $D_{\neg c}(x)$.

These findings suggest that human cognition naturally follows contrastive distance principles, reinforcing our model's relevance beyond artificial neural networks.

5.3.4 Implications for Cognitive Modeling

Reinterpreting classification as a competitive distance process has implications for cognitive models:

1. **Explaining Perceptual Ambiguity:** Just as our framework predicts low confidence near decision boundaries, human perception exhibits uncertainty in ambiguous cases where contrastive evidence is weak (e.g., optical illusions, ambiguous figures).

2. **Improving Cognitive Models of Categorization:** Traditional models assume feature-based classification, but competitive distance may provide a more biologically plausible mechanism.

3. **Connecting Perception and Decision Theory:** The framework aligns with Bayesian models of decision-making, where classification confidence depends on relative likelihoods rather than independent feature strengths.

5.3.5 Future Directions: Unifying Cognitive and Machine Learning Models

If human cognition operates through contrastive distance, machine learning models may benefit from explicitly incorporating these principles. Potential areas of research include:

- Designing machine learning models that better mimic human perception by integrating distance-based representations.

- Using contrastive distance principles to improve explainability and interpretability in AI systems.

- Applying cognitive insights to improve adversarial robustness, inspired by how humans handle noisy or ambiguous inputs.

5.3.6 Conclusion

Our competitive distance framework not only improves neural network classification but also aligns with cognitive science findings. By shifting from intensity-based to contrastive models, we bridge gaps between artificial intelligence, neuroscience, and human perception. The next section will summarize our key findings and discuss future research directions.

5.4 How BatchNorm Reinforces Competitive Distance Learning

Batch Normalization (BatchNorm) is a widely used technique in deep learning that normalizes activations across a batch of inputs to improve training stability, generalization, and convergence speed. While BatchNorm is often interpreted as a method for reducing internal covariate shift, we argue that it also plays a crucial role in **enforcing a form of competitive distance normalization**.

In this section, we examine how BatchNorm implicitly aligns with the competitive distance framework and propose modifications that could explicitly incorporate contrastive distance in neural networks.

5.4.1 BatchNorm as an Approximate Mahalanobis Transform

BatchNorm operates by normalizing activations x based on the mean and variance computed over a batch:

$$\hat{x} = \frac{x - \mu_{\text{batch}}}{\sigma_{\text{batch}}}$$

$$y = \gamma \hat{x} + \beta$$

where μ_{batch} and σ_{batch} are the batch mean and standard deviation, and γ, β are learnable scale and shift parameters.

This transformation closely resembles the **Mahalanobis distance normalization**:

$$D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

where Σ is the covariance matrix. BatchNorm, however, performs only a diagonal normalization (using variance rather than full covariance), making it an approximation of Mahalanobis normalization.

Since Mahalanobis distance is crucial for measuring feature separability in metric learning, this suggests that BatchNorm is implicitly encouraging networks to learn representations where distances between class prototypes are more meaningful.

5.4.2 BatchNorm as a Competitive Distance Normalizer

From the competitive distance perspective, classification decisions should not be based on absolute activations but on relative distances between classes. BatchNorm reinforces this by:

- **Standardizing Distance Measures:** By normalizing activations across a batch, BatchNorm ensures that learned representations remain comparable in terms of their relative distances.

- **Preventing Overdominant Features:** Without normalization, features with large magnitude differences may distort classification confidence. BatchNorm ensures that no single feature dominates purely due to scale.

- **Enhancing Separability in Feature Space:** Normalizing activations makes it easier for networks to learn stable decision boundaries, aligning with our framework that views classification as a function of contrastive distances.

By implicitly enforcing a form of competitive distance scaling, BatchNorm enhances a network's ability to learn well-separated class prototypes.

5.4.3 Reinterpreting BatchNorm in Competitive Distance Learning

Our competitive distance framework suggests that classification confidence should be computed as:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x)).$$

If we consider a network where BatchNorm is applied, we can reinterpret its function as ensuring that $D_c(x)$ and $D_{\neg c}(x)$ are computed over normalized feature representations, improving their consistency and preventing scale-induced distortions.

5.4.4 Can BatchNorm Be Improved for Competitive Distance Learning?

While BatchNorm provides a form of competitive distance normalization, it does so indirectly. We propose modifications that could explicitly incorporate contrastive distance principles:

- **Distance-Aware Normalization:** Instead of normalizing activations independently, normalize them relative to class prototypes, ensuring that the contrastive distance between class representations remains stable across batches.

- **Class-Wise BatchNorm:** Instead of computing a single mean and variance for the entire batch, compute separate statistics per class and normalize activations relative to their respective class distributions.

- **Gradient-Based Competitive Distance Regularization:** Introduce a regularization term that explicitly penalizes small differences between $D_c(x)$ and $D_{\neg c}(x)$, encouraging well-separated feature spaces.

These modifications could enhance the role of BatchNorm in maintaining stable, interpretable contrastive distances throughout network training.

5.4.5 Conclusion

BatchNorm already reinforces elements of competitive distance learning by ensuring that feature activations are normalized in a way that prevents scale-based distortions. However, by modifying BatchNorm to explicitly incorporate competitive distance principles, we may further improve classification robustness, calibration, and interpretability.

The next section will summarize the key insights of this work and outline future directions for research in competitive distance-based neural network architectures.

6 Conclusion and Future Work

6.1 What We Have Established

In this work, we have redefined classification confidence and intensity as functions of competitive distance rather than absolute activation strength. Our framework provides a principled, mathematically rigorous, and biologically plausible explanation for decision-making in both artificial and natural systems. This section summarizes the key insights and contributions of our work.

6.1.1 Distance as the Fundamental Measure

We began by analyzing different types of measures and showed that **distance is the primary and rigorously defined measure**, while intensity lacks a formal definition. We established that:

- Distance functions satisfy well-defined mathematical properties (e.g., metric space axioms).

- Most similarity and probability measures are derived from transformed or negated distances.

- Traditional interpretations of intensity as feature presence are heuristically defined and fail to account for classification dynamics.

6.1.2 The Illusion of Intensity and the Competitive Distance Framework

We demonstrated that intensity, as commonly understood in neural networks, is an illusion. Instead of representing absolute feature strength, intensity should be understood as a function of contrastive distance. Specifically:

- Absolute activations do not determine classification confidence—relative distances do.

- Decision-making is inherently a *competitive process* where an input is classified based on *how much closer it is to one class than to others.*

- This led us to define classification intensity as:

$$S_c(x) = f(D_{\neg c}(x) - D_c(x))$$

  where $D_c(x)$ is the distance to the target class and $D_{\neg c}(x)$ is the distance to the closest competing class.

This formulation resolved the issue of unbounded intensity growth, ensured proper scaling, and aligned with how classification systems actually operate.

6.1.3 A Distance-Based Interpretation of Softmax and Probability

By redefining logits in neural networks as contrastive distances:

$$z_c = -D_c(x) + D_{\neg c}(x),$$

we showed that softmax does not simply map activations to probabilities but rather performs a transformation of competitive distances. This new perspective explains:

- Why softmax confidence behaves as it does.

- Why overconfidence emerges in high-dimensional spaces.

- How alternative transformations (e.g., logistic sigmoid, inverse contrastive distance) could provide more calibrated probability estimates.

6.1.4 Implications for Neural Network Design and Adversarial Robustness

With our new framework, we proposed several modifications to neural networks:

- **Distance-Based Logits:** Rather than arbitrary activations, class logits should explicitly encode contrastive distances.

- **Contrastive Loss Functions:** Training objectives should optimize for class separability rather than raw activation differences.

- **Improved Robustness:** Tracking $D_{\neg c}(x)$ provides stronger defenses against adversarial attacks by preventing confidence from being easily shifted.

- **Calibration Improvements:** Our approach offers a natural way to correct overconfidence in deep learning models.

6.1.5 Connections to Cognitive Science and Perception

Finally, we showed that this framework aligns with human cognition and perception. Evidence from psychology and neuroscience suggests that:

- Humans classify objects based on relative similarity rather than absolute feature presence.

- Neural representations in the brain are structured around contrastive encoding (e.g., opponent processing, population coding).

- Perceptual decision-making aligns with our framework's predictions of confidence and uncertainty.

This suggests that artificial intelligence systems designed with competitive distance principles may better mimic human-like reasoning and robustness.

6.1.6 Final Summary

To summarize, we have established that:

1. **Distance is the primary measure**—intensity should be derived from contrastive distance.

2. **Classification is a competitive process**—confidence depends on relative distances, not absolute activations.

3. **Softmax transforms competitive distances**—this explains its behavior and limitations.

4. **Neural networks can be improved**—explicitly modeling competitive distance can enhance robustness, calibration, and interpretability.

5. **Cognition supports this framework**—human perception operates contrastively, aligning with our theoretical model.

The next section will discuss future directions, including how this framework can be further developed and tested in machine learning, neuroscience, and cognitive modeling.

6.2 Future Research Directions

The competitive distance framework we have developed provides a new perspective on classification confidence, neural network design, and cognitive modeling. This shift in understanding opens up multiple avenues for future research, spanning machine learning, neuroscience, and theoretical foundations of measurement. In this section, we outline key directions for further exploration.

6.2.1 Refining Distance-Based Neural Architectures

One of the most immediate applications of our framework is the development of neural architectures that explicitly incorporate competitive distance. Future research should explore:

- **Distance-Based Logits:** Instead of arbitrary activation-based logits, models should explicitly compute class scores as functions of relative distances.

- **Contrastive Distance Training:** Training loss functions should be reformulated to directly optimize class separability rather than relying on indirect activation maximization.

- **Alternative Probability Transformations:** While softmax is widely used, alternative functions (e.g., logistic contrastive transformations, inverse distance functions) should be explored for improved calibration and robustness.

- **Integration with Metric Learning:** Metric learning techniques, such as triplet loss and contrastive loss, align naturally with our framework and could be leveraged to build better distance-based classifiers.

Developing models that explicitly track contrastive distances could improve classification performance, reduce overconfidence, and enhance interpretability.

6.2.2 Improving Calibration and Uncertainty Estimation

Deep learning models often exhibit poor calibration, meaning that their confidence estimates do not accurately reflect real-world uncertainty. Our frame-

work provides a natural way to improve this by reformulating confidence as a function of class separability. Future research should investigate:

- **Measuring Model Calibration:** Evaluate how contrastive distance-based probabilities compare to standard softmax probabilities in terms of confidence reliability.

- **Redefining Uncertainty Metrics:** Instead of entropy-based uncertainty measures, models could track the raw contrastive distance $D_{\neg c}(x) - D_c(x)$ as a direct measure of classification uncertainty.

- **Calibration in Real-World Applications:** Apply competitive distance-based confidence measures to critical domains such as medical diagnosis and autonomous decision-making.

By replacing heuristically scaled activations with grounded distance-based measures, deep learning models could achieve better reliability in high-stakes applications.

6.2.3 Enhancing Adversarial Robustness

Adversarial examples pose a major challenge in deep learning, with small, imperceptible input perturbations causing confident misclassification. Our framework suggests that many adversarial attacks exploit the lack of explicit contrastive distance modeling. Future research should explore:

- **Adversarial Training with Competitive Distance:** Develop training procedures that explicitly increase the gap between $D_c(x)$ and $D_{\neg c}(x)$ to make decision boundaries more robust.

- **Detecting Adversarial Inputs:** Monitor changes in competitive distance metrics to flag inputs that are close to decision boundaries.

- **Interpreting Adversarial Attacks:** Analyze how existing adversarial perturbations manipulate contrastive distances and propose defenses that counteract these changes.

By ensuring that classification confidence is grounded in well-defined distance relationships, neural networks may become more resistant to adversarial manipulation.

6.2.4 Bridging Machine Learning and Cognitive Science

Our framework suggests that competitive distance is not only relevant for artificial systems but also aligns with principles observed in human cognition. Future work should investigate:

- **Empirical Testing in Human Perception:** Conduct psychophysical experiments to measure how human classification confidence corresponds to contrastive distance effects.

- **Computational Cognitive Models:** Adapt our framework to explain categorization, perceptual decision-making, and uncertainty estimation in the human brain.

- **Neuroscientific Validation:** Investigate whether neural representations in biological brains encode decision confidence using contrastive distance rather than absolute feature strength.

Understanding how the brain leverages competitive distance could inform the design of more human-like artificial intelligence systems.

6.2.5 Generalizing the Competitive Distance Framework

Beyond classification tasks, competitive distance may provide insights into a broader range of problems in mathematics, physics, and theoretical measurement. Future research could explore:

- **Extensions to Regression and Clustering:** Investigate whether contrastive distance can improve models that involve continuous predictions and clustering algorithms.

- **Application to Information Theory:** Explore whether competitive distance provides a meaningful reformulation of concepts like entropy, mutual information, and divergence measures.

- **Connections to Geometry and Topology:** Examine whether competitive distance can be formalized within existing mathematical frameworks, such as Riemannian geometry and manifold learning.

These investigations could reveal deeper connections between competitive distance, fundamental measurement theory, and applied machine learning.

6.2.6 Conclusion

Our framework redefines classification confidence as a function of contrastive distance, providing a more principled, interpretable, and biologically plausible foundation for decision-making. Future research should explore how this paradigm shift can be applied to improve deep learning models, enhance robustness, and unify cognitive science with artificial intelligence.

By moving beyond heuristic intensity-based interpretations and embracing competitive distance as the fundamental measure of classification confidence, we open new pathways for theoretical and applied advancements in AI, neuroscience, and mathematics.