Analysis of feature representations in Representation Learning

Roberto Halpin Gregorio

1 Introduction

Feature representations are the key topic of representation learning. Much work has been performed on learning feature representations on large image datasets such as ImageNet [2] and generalizing these representations to downstream tasks. Many results in this area are quite mysterious due to a lack of understanding and highly non-convex training procedures. Self supervised and Few-shot learning can be classified as sub-fields of representation learning that look to use robust and expressive representations in low-label datasets or tasks.

A recent paper caught my attention under review at ICLR 2022, called *A Theoretically Grounded Characterization of Feature Representations* [1]. The paper identifies two key properties of feature representations *local alignment* and *local congregation*. Using these two properties they derive bounds that measure the performance of downstream classifiers.

Specifically, they lower bound the minimum expected loss, upper bound the probability a test point has high loss, and upper bound the excess risk. They do all this analysis using an upper bound for the zero-one loss and focus on binary and multi-class classification.

My goals for this project were to outline the key properties that the original paper introduces and then extend their results as much as possible, looking for interesting avenues to analyze. My novelties focus around an alternate, more general, upper bound to the zero-one loss, deriving all the original bounds for both binary and multi-class classification, and even introducing an additional multi-class bound. Additionally, I looked into general Lipschitz losses and more complex, expressive classifiers, as the original paper focuses on linear classifiers with bounded norm.

For several of my proofs, I heavily used proof techniques and setup from the original paper. This is especially true when dealing with the more general upper bound to the zero-one loss because the claims and theorems turned out to be very similar. Also, any quotes used in the report are directly from the original paper [1].

The first order of business is to outline the overall problem setup.

2 Problem Setup

In general, I will use the problem setup as in the original paper.

Say we have an input space $\mathcal{X}$ and an output space of targets $\mathcal{Y}$. Additionally, there is an underlying distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We have a feature representation $\phi : \mathcal{X} \to \mathbb{R}^f$. This feature representation is most likely pre-trained on another dataset or task, but we have no additional information other than the feature representation itself. We also assume that this feature representation is bounded: $\|\phi(x)\| \leq B$.

The initial task I focus on is binary classification, which I later will extend to multi-class classification. In the case of binary classification, we have the label space $\mathcal{Y} = \{-1, 1\}$. The classifier is as follows: "Our classifier will use a scoring function that operates on feature space, $h : \mathbb{R}^f \to \mathbb{R}$. The predicted label will then be sign$(h(\phi(x)))$, where if $h(\phi(x))$ is 0, we will arbitrarily assign a label of -1. The set of possible functions h defines the hypothesis class for the classifier; denote this by $\mathcal{H}$. For most of our analysis, we primarily care about the *smoothness* of the functions in $\mathcal{H}$. We will assume that

all functions in $\mathcal{H}$ are Lipschitz continuous with Lipschitz constant less than W." Thus we have:

$$h(\phi(x)) - h(\phi(x')) \leq W\|\phi(x) - \phi(x')\| \quad \forall h \in \mathcal{H}$$

For binary classification, we focus on the commonly used hypothesis class of *linear classifiers of bounded norm*: $\mathcal{W} = \{\boldsymbol{v} \mapsto \boldsymbol{w}^\top \boldsymbol{v}; \|\boldsymbol{w}\| \leq W\}$. This hypothesis class is commonly used when evaluating robust and expressive representations, because the goal is for the representations to do the work in linearly separating the data.

The original paper claims that since the zero one loss, $l^*(h(\phi(x)), y) = \mathbb{I}[\text{sign}(h(\phi(x))) \neq y]$ is difficult to analyze, they use a continuous-margin based upper bound which is standard in theoretical treatments [4] instead:

$$l(h(\phi(x)), y) = \min(1, \max(0, 1 - yh(\phi(x))) \geq l^*(h(\phi(x)), y)$$

The original paper uses this upper bound without much discussion on its relation to the zero-one loss, there is not much intuition on its use and the tightness of the bound is not completely clear. To improve upon this I will introduce a more general form of the loss with a scaling parameter β that can dictate how tight of an upper bound we want:

$$\ell_\beta(h(\phi(x)), y) = \min(1, \max(0, 1 - \beta yh(\phi(x))), \qquad \beta \geq 1$$

Clearly $\ell_\beta(h(\phi(x)), y) \geq l^*(h(\phi(x)), y)$. Additionally, when $\beta = 1$, we have the paper's original loss, and when β approaches infinity we have the zero-one loss.

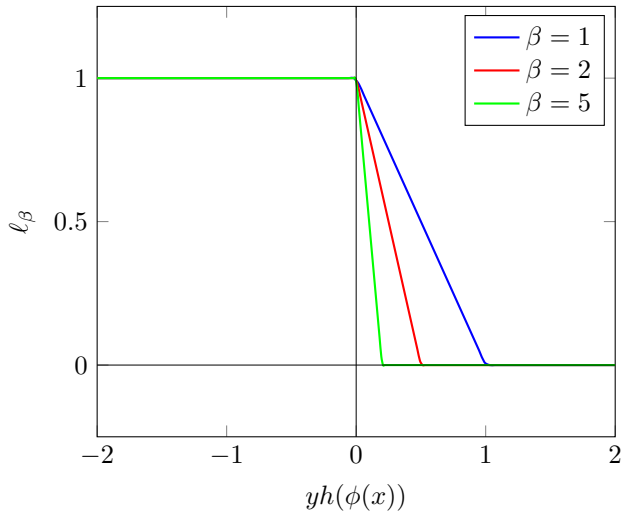

Figure 1: Comparing various values of β when $y = +1$ for ℓ_β.

Quantitatively we can measure how loose our upper bound is by calculating the area under the curve when the input is positive, as that is where the loss differs from the zero-one loss. The area under the curve can be easily calculated as $\frac{1}{2\beta}$, which is $O(\beta^{-1})$. Thus we can achieve a much tighter bound by increasing β, however, it will come at a cost, which will be discussed later.

2.1 Multi-class Classification

Although binary classification is nice for preliminary analysis, it lacks applicability in most modern problems. Representation learning as a whole, especially in computer vision, tends to deal with multiple classes when performing classification. The most popular datasets such as Imagenet [2] and CIFAR [3] are multi-class classification tasks. Thus, it is of great importance to analyze these representations in the multi-class classification framework.

As before we have the same input space $\mathcal{X}$, but now we have a target space $\mathcal{Y}$ consisting of K classes: $\mathcal{Y} = \{0, 1, \ldots, K-1\}$. Our underlying distribution $\mathcal{D}$ is over $\mathcal{X} \times \mathcal{Y}$. Our feature representation stays the same as $\phi : \mathcal{X} \to \mathbb{R}^f$ with $\|\phi(x)\| \leq B$.

The original paper explains the changes to the classifier: "For multi-class classification, our classifier will use a scoring function that scores how well a feature vector matches a class label, h: $\mathbb{R}^f \times \mathcal{Y} \to \mathbb{R}$. The predicted label will then be $\arg\max_y h(\phi(x), y)$, where ties are broken arbitrarily. The set of possible functions h defines the hypothesis class for the classifier; denote this by $\mathcal{H}$. For most of our analysis, we primarily care about the *smoothness* of the functions in $\mathcal{H}$. We will assume that all functions in $\mathcal{H}$ are Lipschitz continuous in the first argument with Lipschitz constant less than W." Thus we have:

$$h(\phi(x), y) - h(\phi(x'), y) \leq W\|\phi(x) - \phi(x')\| \quad \forall h \in \mathcal{H}, \forall y \in \mathcal{Y}$$

As before, I focus on the commonly used hypothesis class of *linear classifiers of bounded norm*: $\mathcal{W} = \{\boldsymbol{v} \mapsto \boldsymbol{w}_y^\top \boldsymbol{v}; \|\boldsymbol{w}_y\| \leq W\}$. For each class there is a weight vector w_y to score the feature representation. The predicted y is identified by finding the class that results in the greatest score.

In self supervised learning this is the predominantly used classifier since the focus is on building robust and expressive representations that should linearly separate the data. Thus, analysis using these classifiers will align well with modern techniques.

However, there is also benefit to shift to more complex classifiers such as N-layer neural networks due to their improved performance especially in generalization, using a pre-trained feature representation for different data or tasks. This is because although ideally we want our feature representations to do all the work and linearly separate the data for us, this is not always possible. Thus, it is beneficial to be able to train a more expressive classifier in downstream tasks when we have fixed feature representations. These classifiers will also be analyzed as they seem to be an interesting and useful extension.

The zero-one loss in the multi-class classification setting given a sample (x, y) is $l^*(h, \phi(x), y) = \mathbb{I}[\arg\max_{y'} h(\phi(x), y') \neq y]$. The multi-class loss the original paper uses follows their binary classification loss. They claim that the zero-one loss is difficult to analyze, so they use a continuous margin-based upper bound:

$$l(h, \phi(x), y) = \min(1, \max(0, \max_{y' \neq y}(1 + h(\phi(x), y')) - h(\phi(x), y))) \geq l^*(h, \phi(x), y)$$

As before, the paper uses this upper bound without much discussion on its relation to the multi-class zero-one loss. To improve I will extend the above loss with a scaling parameter β that can dictate how tight of an upper bound we want:

$$\ell_\beta(h, \phi(x), y) = \min(1, \max(0, 1 - \beta[h(\phi(x), y) - \max_{y' \neq y} h(\phi(x), y')])), \qquad \beta \geq 1$$

Note that when $\beta = 1$, we have the paper's original multi-class loss and when β approaches infinity we have the multi-class zero-one loss. Similar properties hold as the previous scaled loss, in terms of the effect of the β parameter.

With the same goal as the original paper, our focus is to analyze how certain properties of the feature representation ϕ affects the lowest average loss we can achieve and the excess risk when generalizing to other datasets.

Let us first look at the two main properties the original paper identifies.

3 Properties of Feature Representations

The goal that the original paper had in mind is to find properties of feature representations that are easy to measure, easy to reason about, and predictive of downstream accuracy.

The original paper identifies that it is important to "look at pairs of examples that embed close to each other in feature space and ask if they are indeed similar in terms of their ground truth labels." They label this property as local alignment, specifically:

Definition 1. *"Suppose $\alpha > 0$. The **local alignment** of the feature space ϕ, denoted by $p_a^\phi(\alpha)$ is the probability that two data points $(x, y), (x', y') \sim \mathcal{D}$ share a label given that they embed within a distance of α:*

$$p_a^\phi(\alpha) \triangleq P(y = y' \mid \|\phi(x) - \phi(x')\| \leq \alpha, (x, y), (x', y') \sim \mathcal{D}),"$$

α is a hyperparameter that governs the the closeness of the feature representations.

They also remark that local alignment by itself can be "meaningless if data points do not generally embed close to each other." Thus the natural property that follows is a measure of data point congregation:

Definition 2. *"Suppose $\alpha > 0$. The **degree of congregation** of the feature space ϕ, denoted by p_c^ϕ, is the probability that two points x, x' sampled from $\mathcal{D}$ embed within a distance of α:*

$$p_c^\phi(\alpha) \triangleq P(\|\phi(x) - \phi(x')\| \leq \alpha \mid x, x' \sim \mathcal{D}),"$$

α is a hyperparameter here as well.

Figure 2: "An illustration of the properties of local alignment and congregation. Feature representations on the left are more congregated than those on the right. The first and third feature representations on the top are more aligned locally than the other two. Classical error bounds depend only on the norm of the feature vectors and so cannot distinguish between these." Illustration from [1]

Both these definitions are introduced in the original paper and using both *local alignment* and *congregation* we can analyze our minimum expected loss and generalization.

4 Lower bounding the minimum expected loss

Given that we have a feature space ϕ, we want to analyze how good can ϕ perform using a scoring function from our class $\mathcal{H}$. Since all the scoring functions in $\mathcal{H}$ are W-Lipschitz, we know that nearby data points, which are those with close feature representations, will be scored similarly. Thus, when a pair of these data points have different labels we expect to observe at least one classification. Formalizing this statement gives us the following:

Claim 1. *Consider two data points (x, y) and (x', y'). If $\|\phi(x) - \phi(x')\| < \frac{1}{\beta W}$ and $y \neq y'$, then:*

$$\ell(h(\phi(x)), y) + \ell(h(\phi(x'))), y') \geq 1$$

Proof. Due to the Lipschitz continuity of h we have:

$$|h(\phi(x)) - h(\phi(x'))| \leq W||\phi(x) - \phi(x')|| \leq \frac{1}{\beta}$$
$$\Rightarrow h(\phi(x')) > h(\phi(x)) - \frac{1}{\beta}$$

Without loss of generality we can assume $y = 1$. For convenience let us denote $\ell(h(\phi(x)), y)$ as a and $\ell(h(\phi(x')), y')$ as b. We can prove the claim by contradiction: suppose $a + b < 1$. We know that the loss is bounded between 0 and 1, so $a < 1$ and $b < 1$. Thus,

$$a = \max(0, 1 - \beta y h(\phi(x))$$
$$\geq 1 - \beta y h(\phi(x)) = 1 - \beta h(\phi(x))$$
$$\Rightarrow \beta h(\phi(x)) \geq 1 - a$$

Looking at the other loss,

$$\begin{aligned} b &= \max(0, 1 - \beta y' h(\phi(x')) \\ &\geq 1 - \beta y' h(\phi(x')) = 1 + \beta h(\phi(x')) \\ &\geq 1 + \beta(h(\phi(x)) - \frac{1}{\beta}) = \beta h(\phi(x)) \quad \text{(due to Lipschitz analysis done above)} \\ &\geq 1 - a \\ \Rightarrow a + b &\geq 1 \end{aligned}$$

which is a contradiction. □

We know that if a pair of data points have close feature representations and different labels, at least one is misclassified. Our goal is now to bound the best expected loss by using this fact:

Theorem 1. *Let ℓ_β be the loss function defined above, and $\mathcal{H}$ be a hypothesis class of W-Lipschitz functions. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, and p_a^ϕ and p_c^ϕ defined as above. Then*

$$\inf_{h \in \mathcal{H}} \mathbb{E}_{x,y \sim \mathcal{D}}[\ell(h(\phi(x)), y)] > \frac{1}{2}\left(1 - p_a^\phi\left(\frac{1}{\beta W}\right)\right) p_c^\phi\left(\frac{1}{\beta W}\right)$$

Proof. Given $(x, y), (x', y') \sim \mathcal{D}$, we have that $\|\phi(x) - \phi(x')\| < \frac{1}{\beta W}$ and $y \neq y'$ occurs with probability $\tilde{p} = (1 - p_a^\phi(1/\beta W))p_c^\phi(1/\beta W)$. We have for any $h \in \mathcal{H}$:

$$\begin{aligned} &\mathbb{E}_{x,y \sim \mathcal{D}}[\ell(h(\phi(x)), y)] = \frac{1}{2}\mathbb{E}_{(x,y),(x',y') \sim \mathcal{D}}[\ell(h(\phi(x)), y) + \ell(h(\phi(x')), y')] \\ &\geq \frac{\tilde{p}}{2}\mathbb{E}[\ell(h(\phi(x)), y) + \ell(h(\phi(x')), y') \mid \|\phi(x) - \phi(x')\| < \frac{1}{\beta W} \text{ and } y \neq y'] \\ &\geq \frac{\tilde{p}}{2} \qquad \text{(due to Claim 1)} \end{aligned}$$

Since this is valid for all $h \in \mathcal{H}$ we can take the h that minimizes the expected loss, resulting in:

$$\inf_{h \in \mathcal{H}} \mathbb{E}_{x,y \sim \mathcal{D}}[\ell(h(\phi(x)), y)] \geq \frac{\tilde{p}}{2}$$

□

The original paper has this proof for their margin loss, but we extend the results to a variable margin loss. As we can see if we choose a β that causes ℓ_β to be a tighter upper-bound to the zero-one loss, it affects our lower bound on the minimum expected loss. In this case, lower congregation and high local alignment for our feature representation, reduces the lower bound on the minimum expected loss.

4.1 Multi-class Classification

These results can be extended to multi-class classification by adjusting Claim 1.

Claim 2. *Consider two data points (x, y) and (x', y'). If $\|\phi(x) - \phi(x')\| < \frac{2}{\beta W}$ and $y \neq y'$, then:*

$$\ell(h(\phi(x)), y) + \ell(h(\phi(x'))), y') \geq 1$$

Proof. Due to the Lipschitz continuity of h we have:

$$\begin{aligned} |h(\phi(x), y) - h(\phi(x'), y)| \leq W||\phi(x) - \phi(x')|| \leq \frac{1}{2\beta} \quad \forall y \\ \Rightarrow h(\phi(x'), y) \in \left[h(\phi(x), y) - \frac{1}{2\beta}, h(\phi(x), y) + \frac{1}{2\beta}\right] \end{aligned}$$

Thus, for any pair of labels y_1, y_2,

$$h(\phi(x'), y_1) - h(\phi(x'), y_2) + \frac{1}{\beta} \geq h(\phi(x), y_1) - \frac{1}{2\beta} - \left(h(\phi(x), y_2) + \frac{1}{2\beta}\right) + \frac{1}{\beta} \geq h(\phi(x), y_1) - h(\phi(x), y_2)$$

For convenience let us denote $\ell(h(\phi(x), y))$ as a and $\ell(h(\phi(x'), y'))$ as b. We can prove the claim by contradiction: suppose $a + b < 1$. We know that the loss is bounded between 0 and 1, so $a < 1$ and $b < 1$. Thus,

$$\begin{aligned} a &= \max(0, 1 - \beta[h(\phi(x), y) - \max_{y'' \neq y} h(\phi(x), y'')]) \\ &\geq 1 - \beta[h(\phi(x), y) - \max_{y'' \neq y} h(\phi(x), y')] \\ &\geq 1 - \beta[h(\phi(x), y) - h(\phi(x), y')] \\ \Rightarrow \beta[h(\phi(x), y) - h(\phi(x), y')]) &\geq 1 - a \end{aligned}$$

Looking at the other loss,

$$\begin{aligned} b &= \max(0, 1 - \beta[h(\phi(x'), y') - \max_{y'' \neq y} h(\phi(x'), y'')]) \\ &\geq 1 - \beta[h(\phi(x'), y') - \max_{y'' \neq y} h(\phi(x'), y'')] \\ &\geq 1 - \beta[h(\phi(x'), y') - h(\phi(x'), y)] \\ &\geq 1 - \beta[h(\phi(x), y') - h(\phi(x), y) + \frac{1}{\beta}] = \beta[h(\phi(x), y) - h(\phi(x), y')] \quad \text{(due to Lipschitz analysis done above)} \\ &\geq 1 - a \\ \Rightarrow a + b &\geq 1 \end{aligned}$$

which is a contradiction. □

Using this claim and our proof for Theorem 1, we have the following lower bound for the multi-class case:

Theorem 2. *Let ℓ_β be the loss function defined above, and $\mathcal{H}$ be a hypothesis class of W-Lipschitz functions. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$, and p_a^ϕ and p_c^ϕ defined as above. Then*

$$\inf_{h \in \mathcal{H}} \mathbb{E}_{x,y \sim \mathcal{D}}[\ell(h(\phi(x)), y)] > \frac{1}{2}\left(1 - p_a^\phi\left(\frac{1}{2\beta W}\right)\right) p_c^\phi\left(\frac{1}{2\beta W}\right)$$

Note that the only change when extending the theorem to the multi-class case is that our dependence on βW became $2\beta W$.

5 Generalization Analysis

Even though we have classifiers that can produce low loss based on feature representation properties, this has little implication on the ability to generalize. Next we will look at how classifiers trained on small datasets generalize.

Most generalization bounds are note heavily impacted by the properties of the feature representations. Normally they are limited to the maximum norm of feature vectors. We will extend some generalization bounds presented in the original paper, which are based on our outlined properties of feature representations.

5.1 Bounding the probability of high test loss samples

When we have few labels, which is common in self-supervised learning or few-shot learning, classic generalization bound tend to perform poorly due to their reliance on sample size. Thus we present a bound based on the Lipschitz continuity of our classifier, properties of our feature representation, and our proposed loss.

First we will show that

Claim 3. *For any y, $\ell_\beta(h(\phi(x)), y) - \ell_\beta(h(\phi(x')), y) \leq \beta |h(\phi(x)) - h(\phi(x'))|$*

Proof. Let $\ell_\beta(h(\phi(x)), y) = f(1 - \beta y h(\phi(x)))$ where f is defined as in Theorem 7 in the Appendix. Thus, we have

$$\begin{aligned}\ell_\beta(h(\phi(x)), y) - \ell_\beta(h(\phi(x')), y) &\leq |1 - \beta y h(\phi(x)) - (1 - \beta y h(\phi(x')))| \\ &\leq |\beta h(\phi(x)) - \beta h(\phi(x'))| \\ &= \beta |h(\phi(x)) - h(\phi(x'))|\end{aligned}$$

□

Theorem 3. *Suppose S is a sampled training set of m points. For any $h \in \mathcal{H}$, let $\ell_\beta^{max}(h, S) = \max_{(x,y)\in S} \ell_\beta(h(\phi(x)), y)$ be the maximum loss h incurs on S. Then, for all $\epsilon > 0$ and $(x, y) \sim \mathcal{D}$:*

$$P(\ell_\beta(h(\phi(x)), y) > \ell_\beta^{max}(z, S) + \epsilon) \leq \left(1 - p_a^\phi\left(\frac{\epsilon}{\beta W}\right) p_c^\phi\left(\frac{\epsilon}{\beta W}\right)\right)^m.$$

Proof. For any y, $\ell_\beta(h(\phi(x)), y) - \ell_\beta(h(\phi(x')), y)) \leq \beta |h(\phi(x)) - h(\phi(x'))|$ (shown in Claim 3). In addition, since h is W-Lipschitz, we have for all x, x', y, h:

$$\ell_\beta(h(\phi(x)), y) - \ell_\beta(h(\phi(x')), y)) \leq \beta |h(\phi(x)) - h(\phi(x'))| \leq \beta W |\phi(x) - \phi(x')|$$

Thus, for any pair of samples $(x, y), (x', y')$:

$$\left(\|\phi(x) - \phi(x')\| < \frac{\epsilon}{W} \text{ and } y = y'\right) \Rightarrow \ell_\beta(h(\phi(x)), y)) \leq \ell_\beta(h(\phi(x')), y')) + \epsilon$$

Sampling this (x', y') has a probability of $p_a^\phi(\epsilon/\beta W) p_c^\phi(\epsilon/\beta W)$. Additionally, we know that if $(x', y') \in S$, then $\ell_\beta(h(\phi(x')), y') \leq \ell_\beta^{max}(h, S)$.

Thus, for any $(x, y) \in \mathcal{D}$:

$$\begin{aligned}P(\ell_\beta(h(\phi(x)), y) \leq \ell_\beta^{max}(z, S) + \epsilon) &\geq P\left(\exists (x', y') \in S \mid \|\phi(x) - \phi(x')\| < \frac{\epsilon}{\beta W} \text{ and } y = y'\right) \\ &= 1 - P\left(\nexists (x', y') \in S \mid \|\phi(x) - \phi(x')\| < \frac{\epsilon}{\beta W} \text{ and } y = y'\right) \\ &= 1 - \left(1 - P\left((x', y') \in S \mid \|\phi(x) - \phi(x')\| < \frac{\epsilon}{\beta W} \text{ and } y = y'\right)\right)^m \\ &= 1 - \left(1 - p_a^\phi\left(\frac{\epsilon}{\beta W}\right) p_c^\phi\left(\frac{\epsilon}{\beta W}\right)\right)^m\end{aligned}$$

Taking complements of both sides completes the proof. □

5.1.1 Multi-class Classification

Again we can extend to multi-class classification by adjusting Claim 3.

Claim 4. *For any y, $\ell_\beta(h, \phi(x), y) - \ell_\beta(h, \phi(x'), y) \leq 2\beta W |\phi(x) - \phi(x')|$*

Proof. Let $\ell_\beta(h(\phi(x)), y) = f(1 - \beta[h(\phi(x), y) - \max_{y' \neq y} h(\phi(x), y')])$ where f is defined as in Theorem 7 in the Appendix. Thus, we have

$$\begin{aligned}&\ell_\beta(h, \phi(x), y) - \ell_\beta(h, \phi(x'), y) \\ &\leq \left|\left(1 - \beta[h(\phi(x), y) - \max_{y' \neq y} h(\phi(x), y')]\right) - \left(1 - \beta[h(\phi(x'), y) - \max_{y' \neq y} h(\phi(x'), y')]\right)\right| \\ &= \beta \left|(h(\phi(x'), y) - h(\phi(x), y)) + \left(\max_{y' \neq y} h(\phi(x), y') - \max_{y' \neq y} h(\phi(x'), y')\right)\right| \\ &\leq \beta |h(\phi(x'), y) - h(\phi(x), y)| + \beta \max_{y' \neq y} |h(\phi(x), y') - h(\phi(x'), y')| \\ &\leq \beta W |h(\phi(x), y) - h(\phi(x'), y)| + \beta W |h(\phi(x), y) - h(\phi(x'), y)| \\ &\leq 2\beta W |\phi(x) - \phi(x')|\end{aligned}$$

□

This claim yields the following theorem:

Theorem 4. *Suppose S is a sampled training set of m points. For any $h \in \mathcal{H}$, let $\ell_\beta^{max}(h, S) = \max_{(x,y)\in S} \ell_\beta(h, \phi(x), y)$ be the maximum loss h incurs on S. Then, for all $\epsilon > 0$ and $(x, y) \sim \mathcal{D}$:*

$$P(\ell_\beta(h, \phi(x), y) > \ell_\beta^{max}(h, S) + \epsilon) \leq \left(1 - p_a^\phi\left(\frac{\epsilon}{2\beta W}\right) p_c^\phi\left(\frac{\epsilon}{2\beta W}\right)\right)^m.$$

Proof. Follows directly from Claim 4 and Theorem 3. □

As before, utilizing this general, margin loss adds a flexible β hyperparameter that lets us adjust the tightness of our loss upper-bound, but also affecting our generalization bound. In this case, the more congregated and locally aligned our feature representation is, the less likely we observe high test loss samples. Interestingly, this is different than before where we wanted to observe a low congregation to reduce our lower bound on the minimum expected loss. Both bounds suggest for a high local alignment, but contrasting congregation.

5.1.2 Other common losses

Any loss that is L-Lipschitz with respect to $h(\phi(x))$, will result in the following:

$$\begin{aligned}
&\ell(h, \phi(x), y) - \ell(h, \phi(x'), y) \\
&\leq |\ell(h, \phi(x), y) - \ell(h, \phi(x'), y)| \\
&\leq L|h(\phi(x), y) - h(\phi(x'), y)| \\
&\leq LW|\phi(x) - \phi(x')|
\end{aligned}$$

This means that we have Theorem 4 with $2\beta = L$.

Now we may also have some loss functions that are L-Lipschitz with respect to $\phi(x)$. Following the above steps, we no longer have to rely on h being Lipschitz, which results in:

$$\begin{aligned}
&\ell(h, \phi(x), y) - \ell(h, \phi(x'), y) \\
&\leq L|\phi(x) - \phi(x')|
\end{aligned}$$

Thus we have Theorem 4 with $2\beta W = L$.

5.2 Bounding excess risk

Using a classic excess risk bound we can perform a traditional analysis. The original paper has the following bound, for any $\delta > 0$ with probability of at least $1 - \delta$, the following holds for all $h \in \mathcal{H}$:

$$R(h) - \hat{R}(h, S) \leq \Re_m(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

where we have $R(h)$ as the expected risk, $\hat{R}(h, S)$ as the empirical risk, and $\Re_m(\mathcal{H})$ as the Rademacher complexity of $\mathcal{H}$.

The original paper is able to prove a bound on $\Re_m(\mathcal{H})$ on *linear classifiers with bounded norm*, $\mathcal{W}$, using the two properties discussed:

Theorem 5. *Letting p_c^ϕ be defined as above. The Rademacher complexity of $\mathcal{W}$ is bounded above by:*

$$\Re_m(\mathcal{W}) = \frac{W\left(\alpha\sqrt{p_c^\phi(\alpha)/2} + 2B\sqrt{(1 - p_c^\phi(\alpha)) + 2(1 - p_c^\phi(\alpha))^2}\right)}{2\sqrt{m}}$$

Proof. The proof is a fair bit involved and can be found in the Appendix of the original paper under Theorem 5. □

A slightly more general risk bound is the following (using our loss):

$$R(h) - \hat{R}(h, S) \leq \Re_m(\ell_\beta \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

We can extend this bound to our variable loss noting that our loss, ℓ_β, is a β-Lipschitz function and using the property that given a L-Lipschitz loss function ℓ, $\Re_m(\ell \circ \mathcal{H}) = L\Re_m(\mathcal{H})$. The excess risk bound becomes

$$R(h) - \hat{R}(h, S) \leq \beta\Re_m(\mathcal{W}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

The Rademacher bound is loss agnostic, so the initial bound can be directly applied to this risk bound. Overall, it is interesting how this risk bound is only affected by the congregation property, specifically it benefits from high congregation just like the other upper bound and unlike the lower bound. Also in this case the locally aligned property seems to have no effect. Additionally, we pay extra here when we want to use a larger β.

5.2.1 Multi-class Classification

The original paper does not have any mention of a multi-class classification variant of the excess risk bound. However, it is known that in the multi-class setting with k classes [4], that the following holds, for any $\delta > 0$ with probability of at least $1 - \delta$, the following holds for all $h \in \mathcal{H}$:

$$R(h) - \hat{R}(h, S) \leq 4k\Re_m(\ell \circ \Pi_1(\mathcal{H})) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

Using $\ell = \ell_\beta$ and recalling that in the multi-class case it is 2β-Lipschitz, it is clear that $4k\Re_m(\ell \circ \Pi_1(\mathcal{H})) \leq 8k\beta\Re_m(\Pi_1(\mathcal{H}))$.

Now we will show the following:

Claim 5. *Letting $\mathcal{H}$ be the linear classifier with bounded norm in the multi-label case and $\mathcal{W}$ be the linear classifier with bounded norm in the binary classification case, the following holds:*

$$\Re_m(\Pi_1(\mathcal{H})) = \Re_m(\mathcal{W})$$

Proof.

$$\begin{aligned}
\Re_m(\Pi_1(\mathcal{H})) &= \frac{1}{m}\mathbb{E}_{S,\sigma}\left[\sup_{y \in \mathcal{Y}, \|\boldsymbol{w}_y\| \leq W} \boldsymbol{w}_y^\top \sum_{i=1}^m \sigma_i \phi(x_i)\right] \text{ (from [4])} \\
&= \frac{1}{m}\mathbb{E}_{S,\sigma}\left[\sup_{h \in \mathcal{W}} \boldsymbol{w}_h^\top \sum_{i=1}^m \sigma_i \phi(x_i)\right] \\
&= \Re_m(\mathcal{W})
\end{aligned}$$

□

Thus, in the multi-class classification case, we have

$$R(h) - \hat{R}(h, S) \leq 8k\beta\Re_m(\mathcal{W}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

with $\Re_m(\mathcal{W})$ being bound by Theorem 5.

5.2.2 Other common losses

Any loss that is L-Lipschitz with respect to $h(\phi(x), y)$, will result in the following, for any $\delta > 0$ with probability of at least $1 - \delta$, the following holds for all $h \in \mathcal{W}$:

$$R(h) - \hat{R}(h, S) \leq 4kL\Re_m(\mathcal{W}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

Additionally, any loss that is L-Lipschitz with respect to $\phi(x)$, will result in the following bound due to, in essence, a tighter Rademacher complexity bound. For any $\delta > 0$ with probability of at least $1 - \delta$, the following holds for all $h \in \mathcal{W}$:

$$R(h) - \hat{R}(h, S) \leq 4k\frac{L}{W}\Re_m(\mathcal{W}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

5.3 Dealing with a more complex predictor

As mentioned earlier linear classifiers can run into some issues. When you have a fixed feature representation and want to fine-tune to a new task or dataset you are extremely limited with a linear classifier. You can only hope that the features are linearly separable or close to being separated by a hyperplane. However, if we allow the use of a non-linear classifier such as a $N-$layer neural network, these issues can be avoided and we can greatly improve generalization in certain tasks.

Let us define an ℓ_1 neural network with N layers defined as follows

$$\mathcal{F}_i = \left\{ x \mapsto \sum_j w_j^i \sigma_i(f_j(x)) : \forall j, \quad f_j \in \mathcal{F}_{i-1}, \quad \left\|w^i\right\|_1 \leq B_i \right\}$$

where σ_i is a L_i-Lipschitz function. Additionally, let $\mathcal{F}_1$ be equal to $\ell \circ W$, where ℓ is L_1-Lipschitz. This final layer is very common in most neural networks used in representation learning.

Theorem 6. *Given our definition of $\mathcal{F}_i$ above,*

$$\Re_m(\mathcal{F}_i) \leq \left(\prod_{i=1}^{N} 2B_i L_i \right) \Re_m(W)$$

Proof.

$$\begin{aligned} \Re_m(\mathcal{F}_i) &\leq 2B_i \Re_m(\sigma_i \circ \mathcal{F}_{i-1}) \quad \text{(from lecture)} \\ &\leq 2B_i L_i \Re_m(\mathcal{F}_{i-1}) \quad \text{(Talagrand's lemma works with any } l\text{-Lipschitz fucntions from } \mathbb{R} \text{ to } \mathbb{R} \text{ [4])} \end{aligned}$$

Thus, we end up with

$$\begin{aligned} \Re_m(\mathcal{F}_i) &\leq \left(\prod_{i=1}^{N} 2B_i \right) \Re_m(\mathcal{F}_1) \\ &\leq \left(\prod_{i=2}^{N} 2B_i L_i \right) 2B_1 \Re_m(\ell \circ W) \\ &\leq \left(\prod_{i=1}^{N} 2B_i L_i \right) \Re_m(W) \end{aligned}$$

□

Now using our previous work, we can bound the excess risk in our multi-class classification task given an L_1-Lipschitz loss function and an ℓ_1 N-layer neural network as follows, for any $\delta > 0$ with probability of at least $1 - \delta$, the following holds for all $h \in \mathcal{H}$:

$$R(h) - \hat{R}(h, S) \leq 4k \left(\prod_{i=1}^{N} 2B_i L_i \right) \Re_m(W) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

where we still have the Rademacher upper bound for $\mathcal{W}$.

If our N-layer neural network has small magnitude weights and activations near 1-Lipschitz, our excess risk bound can improve compared to the case of using a linear classifier with bounded norm. Otherwise there is the danger of our bound exploding, but in most cases, a shallow neural net will be used as a classifier in representation learning, due to the very expressive feature representations.

6 Conclusion

In summary, the original paper identified some key properties of feature representations that are predictive of how well downstream classifiers perform. Specifically, a lower bound on the minimum expected loss, an upper bound on the probability of high test loss samples, and an upper bound on excess risk. These properties were the local alignment and degree of congregation of the feature space. This analysis was performed on classification with an arbitrary upper-bound to the zero-one loss.

My contributions on extending the original paper were to first to use alternate losses. I decided to derive all the derived bounds using a more general margin loss, both for binary and multi-class classification, that can be tuned to affect the tightness of the upper bound on the zero-one loss. Additionally, I provided an upper bound to the excess risk in the multi-class case for the general margin loss, which was missing entirely in the original paper.

In addition to looking at this generalized margin loss, I discussed the implications of using any Lipschitz loss in the general multi-class classification setup with regards to all the bounds. Lastly, I analyzed the effect of a more complex and expressive classifier, the N-layer ℓ_1 neural network when it came to bounding the excess risk. Both these extensions where quite important for my to analyze since most modern research in the representation learning scene tends to use alternate losses and potentially alternate classifiers.

More intuition, interesting results, and perhaps fat trimming could have been done with more time. There is a large overlap between the binary and multi-class classification proofs, but I felt that they were both important to show fully. The Rademacher upper bound proof from the original paper was omitted due to its length (Theorem 5 in the Appendix of the original paper), and proof techniques from the original paper where heavily used on certain proofs here.

References

[1] Anonymous. "A theoretically grounded characterization of feature representations". In: *Submitted to The Tenth International Conference on Learning Representations*. under review. 2022. URL: https://openreview.net/forum?id=7ADMMyZpeY.

[2] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[3] Alex Krizhevsky and Geoffrey Hinton. *Learning multiple layers of features from tiny images*. Tech. rep. University of Toronto, 2009. URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[4] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Cambridge, MA: The MIT Press, 2018.

A Appendix

Theorem 7. *"Let* $f(x) = \min(1, \max(0, x)$*. Then* $f(x) - f(x') \leq |x - x'|$*."*

Proof. Exact theorem is stated and proved in the original paper Appendix. To summarize the proof, a case-by-case analysis is performed for all values of x and x', resulting in the theorem. □