Memorization-Dilation:
Modeling Neural Collapse Under Noise

Duc Anh Nguyen1∗ Ron Levi2∗ Julian Lienen3∗ Gitta Kutyniok1,4 Eyke Hüllermeier5
1 Department of Mathematics, LMU Munich, Germany
2 Faculty of Mathematics, Technion – Israel Institute of Technology, Israel
3 Department of Computer Science, Paderborn University, Germany
4 Department of Physics and Technology, University of Tromsø, Norway
5 Institute of Informatics, LMU Munich, Germany

Abstract
The notion of neural collapse refers to several emergent phenomena that have been empirically observed across various canonical classification problems. During the terminal phase of training a deep neural network, the feature embedding of all examples of the same class tend to collapse to a single representation, and the features of different classes tend to separate as much as possible. Neural collapse is often studied through a simplified model, called the unconstrained feature representation, in which the model is assumed to have "infinite expressivity" and can map each data point to any arbitrary representation. In this work, we propose a more realistic variant of the unconstrained feature representation that takes the limited expressivity of the network into account. Empirical evidence suggests that the memorization of noisy data points leads to a degradation (dilation) of the neural collapse. Using a model of the memorization-dilation (M-D) phenomenon, we show one mechanism by which different losses lead to different performances of the trained network on noisy data. Our proofs reveal why label smoothing, a modification of cross-entropy empirically observed to produce a regularization effect, leads to improved generalization in classification tasks.

1 Introduction
The empirical success of deep neural networks has accelerated the introduction of new learning algorithms and triggered new applications, with a pace that makes it hard to keep up with profound theoretical foundations and insightful explanations. As one of the few yet particularly appealing theoretical characterizations of overparameterized models trained for canonical classification tasks, Neural Collapse (NC) provides a mathematically elegant formalization of learned feature representations [Papyan et al. (2020)].

To explain NC, consider the following setting. Suppose we are given a balanced dataset \( \mathcal{D} = \{ (x_n^{(k)}, y_n) \}_{k \in [K], n \in [N]} \subset \mathcal{X} \times \mathcal{Y} \) in the instance space \( \mathcal{X} = \mathbb{R}^d \) and label space \( \mathcal{Y} = [N] := \{1, \ldots, N\} \), i.e. each class \( n \in [N] \) has exactly \( K \) samples \( x_n^{(1)}, \ldots, x_n^{(K)} \). We consider network architectures commonly used in classification tasks that are composed of a feature engineering part \( g: \mathcal{X} \to \mathbb{R}^M \) (which maps an input signal \( x \in \mathcal{X} \) to its feature representation \( g(x) \in \mathbb{R}^M \)) and a linear classifier \( W(\cdot) + b \) given by a weight matrix \( W \in \mathbb{R}^{N \times M} \) as well as a bias vector \( b \in \mathbb{R}^N \). Let \( w_n \) denote the row vector of \( W \) associated with class \( n \in [N] \). During training, both classifier components are simultaneously optimized by minimizing the cross-entropy loss.

∗These authors contributed equally to this work.

Preprint. Under review.
Denoting the feature representations $g(x_n^{(k)})$ of the sample $x_n^{(k)}$ by $h_n^{(k)}$, the class means and the global mean of the features by

$$h_n := \frac{1}{K} \sum_{k=1}^{K} h_n^{(k)}, \quad h := \frac{1}{N} \sum_{n=1}^{N} h_n,$$

Following [Papyan et al., 2020], NC consists of the following interconnected phenomena (where the limits take place as training progresses):

**(NC1) Variability collapse.** For each class $n \in [N]$, we have

$$\frac{1}{K} \sum_{k=1}^{K} \|h_n^{(k)} - h_n\|^2 \to 0.$$

**(NC2) Convergence to simplex ETF structure.** For any $m, n \in [N]$ with $m \neq n$, we have

$$\frac{1}{N - 1} \left( \frac{\|h_m - h\|_2}{\|h_n - h\|_2} \right)^2 \to \frac{1}{N - 1}.$$

**(NC3) Convergence to self-duality.** For any $n \in [N]$, it holds

$$\frac{h_n - h}{\|h_n - h\|_2} - \frac{w_n}{\|w_n\|_2} \to 0.$$

**(NC4) Simplification to nearest class center behavior.** For any feature representation $u \in \mathbb{R}^M$, it holds

$$\arg \max_{n \in [N]} \langle w_n, u \rangle + b_n \to \arg \min_{y \in Y} \|u - h_n\|_2.$$

In this paper, we consider a well known simplified model, in which the features $h_n^{(k)}$ are not parameterized by the feature engineering network $g$ but are rather free variables. This model is often referred to as layer-peeled model or unconstrained features model, see e.g. Lu and Steinerberger [2020], Fang et al. [2021], Zhu et al. [2021]. However, as opposed to those contributions, in which the features $h$ can take any value in $\mathbb{R}^M$, we consider here the case $h \geq 0$ (understood component-wise). This is motivated by the fact that features are typically the outcome of some non-negative activation function, like the Rectified Linear Unit (ReLU) or sigmoid. Moreover, by incorporating the limited expressivity of the network to the layer-peeled model, we propose a new model, called memorization-dilation (MD). Given such model assumptions, we formally prove advantageous effects of the so-called label smoothing (LS) technique Szegedy et al. [2015] (training with a modification of cross-entropy (CE) loss), in terms of generalization performance. This is further confirmed empirically.

### 2 Related Work

Studying the nature of neural network optimization is challenging. In the past, a plethora of theoretical models has been proposed to do so [Sun, 2020]. These range from analyzing simple linear [Kunin et al., 2019], [Zhu et al., 2020], [Laurent and von Brecht, 2018] to non-linear deep neural networks [Saxe et al., 2014], [Yun et al., 2018]. As one prominent framework among others, Neural Tangent Kernels [Jacot et al., 2018], [Roberts et al., 2021], where neural networks are considered as linear models on top of randomized features, have been broadly leveraged for studying deep neural networks and their learning properties.

Many of the theoretical properties of deep neural networks in the regime of overparameterization are still unexplained. Nevertheless, certain peculiarities have emerged recently. Among those, so-called “benign overfitting” [Bartlett et al., 2019], [Li et al., 2021], where deep models are capable of perfectly fitting potentially noisy data by retaining accurate predictions, has recently attracted attention. Memorization has been identified as one significant factor contributing to this effect [Arpit et al., 2017], [Sanyal et al., 2021], which also relates to our studies. Not less interesting, the learning
risk of highly-overparameterized models shows a double-descent behavior when varying the model complexity [Nakkiran et al. 2020] as yet another phenomenon. Lastly, the concept of NC [Papyan et al. 2020] has recently shed light on symmetries in learned representations of overparameterized models.

After laying the foundation of a rigorous mathematical characterization of the NC phenomenon by [Papyan et al. 2020], several follow-up works have broadened the picture. As the former proceeds from studying CE loss, the collapsing behavior has been investigated for alternative loss functions. For instance, squared losses have shown similar collapsing characteristics [Poggio and Liao 2020, 2021], and have paved the way for more opportunities in its mathematical analysis, e.g., by an NC-interpretable decomposition [Han et al. 2021]. More recently, [Kornblith et al. 2021] provide an exhaustive overview over several commonly used loss functions for training deep neural networks regarding their feature collapses.

Besides varying the loss function, different theoretical models have been proposed to analyze NC. Most prominently, unconstrained feature models have been considered, which characterize the penultimate layer activations as free optimization variables as mentioned before [Mixon et al. 2020, Lu and Steinberger 2020, E and Wojtowytsch 2021]. This stems from the assumption of highly overparameterized models that are assumed to allow for approximating any point in the feature space [Mixon et al. 2020]. While unconstrained feature models typically only look at the last feature encoder layer, layer-peeling allows for “white-boxing” further layers before the last one for a more comprehensive theoretical analysis [Fang et al. 2021]. Moreover, [Zhu et al. 2021] extend the unconstrained feature model analysis by highlighting the structure of the global landscape and the optimization therein. Beyond unconstrained feature models, [Ergen and Pilanci 2021] introduce a convex analytical framework to characterize the encoder layers for a more profound understanding of the NC phenomenon. Referring to its usefulness in transfer learning, [Galanti et al. 2021] highlight favorable properties of NC on previously unseen data.

3 Layer-peeled model with positive features

As a prerequisite to the MD model, in this section we introduce a slightly modified version of the layer-peeled (or unconstrained features) model (see e.g. [Zhu et al. 2021, Fang et al. 2021]), in which the features have to be positive. Accordingly, we will show that the global minimizers of the modified layer-peeled model correspond to an NC configuration, which differs from the global minimizers specified in other works and captures more closely the NC phenomenon in practice.

For conciseness, we denote by $H$ the matrix formed by the features $h^{(k)}_n$, $n \in [N]$, $k \in [K]$ as columns, and define $\|W\|$ and $\|H\|$ to be the Frobenius norm of the respective matrices, i.e. $\|W\|^2 = \sum_{n=1}^{N} \|w_n\|^2$ and $\|H\|^2 = \sum_{k=1}^{K} \sum_{n=1}^{N} \|h^{(k)}_n\|^2$. We consider the regularized version of the model (instead of the norm constraint one as in [Fang et al. 2021])

$$\min_{W,H} \quad L_\alpha(W,H) := L_\alpha(W;H) + \lambda_W \|W\|^2 + \frac{\lambda_H}{K} \|H\|^2 \quad (P_\alpha)$$

s.t. $H \geq 0$,

where $\lambda_W, \lambda_H > 0$ are the penalty parameters for the weight decays. By $L_\alpha$ we denote empirical risk with respect to the LS loss with parameter $\alpha \in [0,1]$, where $\alpha = 0$ corresponds to the conventional CE loss. More precisely, given a value $\alpha > 0$, the label corresponding to class $n \in [N]$ is presented as the following probability vector:

$$y^{(\alpha)}_n = (1 - \alpha)e_n + \frac{\alpha}{N} \mathbf{1}_N \in [0,1]^N,$$

where $e_n$ denotes the $n$-th standard basis vector in $\mathbb{R}^N$ and $\mathbf{1}_N \in \mathbb{R}^N$ denotes the vector consisting of only ones. Let $p : \mathbb{R}^M \rightarrow \mathbb{R}^N$ be the function that assigns to each feature representation $z \in \mathbb{R}^N$ the probability scores of the classes (as a probability vector in $\mathbb{R}^N$),

$$p_W(z) := \text{softmax}(Wz) := \left[ \frac{e^{\langle w_m, z \rangle}}{\sum_{i=1}^{N} e^{\langle w_i, z \rangle}} \right]_m^N \in [0,1]^N.$$

\footnote{Note that for simplicity we assume that the last layer does not have bias terms, i.e. $b = 0$. The result can be however easily extended to the more general case when the biases do not vanish.}
We will show that in common settings, the minimizers of \((ii-iii)\) are replaced by

\[
\hat{W} \in \mathbb{R}^{N \times N}, \quad \hat{h}_n = \frac{1}{\sqrt{N}} \sum_{k=1}^{K} h_n^{(k)} \quad \text{for every } n \in [N].
\]

Theorem 3.2. The configurations defined in Definition 3.1 above differ from the ones specified in other works, and both collections \(\{h_n\}_{n=1}^N\) and \(\{w_n\}_{n=1}^N\) form \(N\)-simplex ETFs. In particular, this requires that the global feature mean \(\hat{h} = \frac{1}{N} \sum_{n=1}^N h_n\) lies at the origin, which can only hold true when we accept negative features. Accordingly, the property (NC2) (in the limit as training progresses), which involves the centralized class means \(h_n - \hat{h}\), become in this setting a property of the un-centralized class means \(h_n\),

\[
\|h_n\| = \|h_n\|, \quad \left\langle \frac{h_m}{\|h_m\|}, \frac{h_n}{\|h_n\|} \right\rangle = -\frac{1}{N-1} \quad \text{for any } m \neq n.
\]

Similarly, the property (NC3) in this setting becomes the duality between the weight vectors \(w_n\) and the un-centralized class means \(h_n\),

\[
\frac{h_n}{\|h_n\|} = \frac{w_n}{\|w_n\|},
\]
instead of the centralized ones $h_n - h$ as we would expect.

In opposition, our NC configurations defined in Definition 3.1 above matches perfectly the limit of (NC2) and (NC3). Indeed, it is not difficult to see that the projection of $h_n$ onto the subspace $P_{h_n}$ is a multiple of $h_n - h$, and hence the duality between the weights $w_n$ and the centralized class means $h_n - h$ follows straightforwardly from the condition (iii). Furthermore it follows from the condition (ii) that the centralized class means satisfy

$$\|h_m - h\| = \|h_n - h\|, \quad \frac{h_m - h}{\|h_m - h\|} = \frac{h_n - h}{\|h_n - h\|} \quad \text{for any } m \neq n.$$

4 The Memorization-Dilation model

4.1 Experimental Motivation

Previous studies of the NC phenomenon mainly focus on the collapsing variability of training activations, and make rather cautious statements about its effects on generalization. For instance, Papyan et al. [2020] report slightly improved test accuracies for training beyond zero training error. Going a step further, Zhu et al. [2021] show that the NC phenomenon also happens for overparameterized models when labels are completely randomized. Here, the models seem to memorize by overfitting the data points, however, a rigorous study how label corruption affects generalization in the regime of NC is still lacking.

To fill the gap, we advocate to analyze the effects of label corruption in the training data on the (previously unseen) test instead of the training feature collapse. Eventually, tight test class clusters go hand in hand with easier separation of the instances and, thus, a smaller generalization error. Following Zhu et al. [2021], we measure the collapse of the penultimate layer activations by the NC metric. Given the within-class covariance matrix $\Sigma_W \in \mathbb{R}^{d \times d}$, which is defined as

$$\Sigma_W := \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} (h^k_n - h_n)(h^k_n - h_n)^\top,$$

and the inter-class covariance matrix $\Sigma_B \in \mathbb{R}^{d \times d}$ given by

$$\Sigma_B := \frac{1}{N} \sum_{n=1}^{N} (h_n - h)(h_n - h)^\top$$

for the penultimate layer dimensionality $d$, the metric $\mathcal{NC}_1$ is defined as

$$\mathcal{NC}_1 := \frac{1}{N} \text{trace}(\Sigma_W \Sigma_B^\dagger). \quad (2)$$

Here, $\Sigma_B^\dagger$ denotes the pseudo-inverse of $\Sigma_B$. Moreover, we distinguish $\mathcal{NC}_1^{\text{train}}$ and $\mathcal{NC}_1^{\text{test}}$ to be calculated on the training and test instances, respectively. We call $\mathcal{NC}_1^{\text{test}}$ dilation.

To understand the effects of label corruption in the data, let us clarify the notion of memorization. While there exists a variety of definitions, we tailor our definition to the NC setting. Formally, suppose that label noise is incorporated by (independently) corrupting the instance of each class label $n$ in the training data with probability $\eta \in (0, 1)$, where corruption means drawing a label uniformly at random from the label space $\mathcal{Y}$. We denote the set of corrupted instances by $[\tilde{K}]$. For a given dataset $D$ (with label noise $\eta$), we define memorization as

$$\text{mem} := \sum_{n=1}^{N} \sum_{k \in [\tilde{K}]} \|h^{(k)}_n - h^*_n\|_2, \quad (3)$$

where $h^*_n$ denotes the mean of (unseen) test instances belonging to class $n$.

We call the original ground truth label of a sampled its true label. We call the label after corruption, which may be the true label or not, the observed label. Since instances of the same true label tend to have similar input features in some sense, the network is biased to map them to similar feature representations. Instances are corrupted randomly, and hence, instances of the same true label but
different observed labels do not have predictable characteristics that allow the network to separate them in a way that can be generalized. When the network nevertheless succeeds in separating such instances, we say that the network memorized the feature representations of the corrupted instances in the training set. The metric mem in (3) thus measures memorization.

To quantify the interaction between mem and $N_{\text{CE}}$, we analyzed the learned representations $h$ in the penultimate layer feature space for different noise configurations. One may wonder whether one can see a systematic trend in the test collapse given the memorization, and how this evolves over different loss functions.

To this end, we trained simple probabilistic multi-layer neural networks for two classes ($N = 2$), which we subsampled from the image classification datasets MNIST [LeCun et al., 1998], FashionMNIST [Xiao et al., 2017], CIFAR-10 [Krizhevsky and Hinton, 2009] and SVHN [Netzer et al., 2011]. The labels are corrupted with noise degrees $\eta \in [0.025, 0.4]$. The network consists of 9 hidden layers with 2048 neurons each, thus, it represents a vastly overparameterized model. The number of penultimate layer features $d$ is set to the number of classes $N$. We trained these networks on two variations of the CE loss, namely its conventional form and LS with a smoothing factor $\alpha = 0.1$, as well as the mean-squared error (MSE). Moreover, we consider label relaxation (LR) [Lienen and Hüllermeier, 2021] as a generalization to LS with a relaxation degree $\alpha = 0.1$. The networks were trained for 200 epochs using SGD with an initial learning rate of 0.1 multiplied by 0.1 each 40 epochs and a small weight decay of 0.001. Moreover, we considered ReLU as activation function throughout the network, as well as batch normalization in each hidden layer. A linear softmax classifier is composed on the encoder. We conducted each experiment ten times with different seeds. We refer to the appendix for a more comprehensive description of the experimental settings and parameters.

Fig. 1 shows the trends of $\sqrt{N_{\text{CE}}}$ per memorization for various configurations. As can be seen, the figure shows an approximately linear correspondence between $\sqrt{N_{\text{CE}}}$ and mem for the CE derivatives (CE and LS) on all datasets when mem is not large. Moreover, as CE and LS share the same slope, these results suggest that the degradation of the test collapse (aka dilation) is a function of memorization and the expressivity of the network, and not of the choice of the loss. The loss only affects how the noise translates to memorization, but not how memorization translates to dilation. Even though the same amount of noise is mapped to different memorization values in CE and LS, the memorization-dilation curve is nevertheless shared between CE and LS. Hence, since LS leads the network to memorize less, it results in improved generalization performance (cf. Fig 1). We can further see that MSE and LR show a different memorization-dilation correspondence, which means that these losses affect the inductive bias in a different way than CE and LS.
In the appendix, we provide additional results showing exemplary feature distributions and the behavior in the multi-class case with $N > 2$, which reveals a similar correspondence for higher feature and class dimensions. Furthermore, we consider different models for label noise. The results support our MD model, and show that the memorization-dilation curve is roughly independent of the noise model for low-to-mid noise levels.

4.2 The Memorization-Dilation Model

Motivated by the observations of the previous experiments, we propose the so-called memorization-dilation (MD) model, which extends the unconstrained feature model by incorporating the interaction between memorization and dilation as a model assumption. By this, we explicitly capture the limited expressivity of the network, thereby modeling the inductive bias of the underlying model.

This model shall provide a basis to mathematically characterize the difference in the learning behavior of CE and LS. More specifically, we would like to know why LS shows improved generalization performance over conventional CE, as was observed in past works [Müller et al., 2019]. The main idea can be explained as follows. We first note that dilation is directly linked to generalization, since the more concentrated the feature representations of each class are, the easier it is to separate the different classes with a linear classifier without having outliers crossing the decision boundary. The MD model asserts that dilation is a linear function of memorization. Hence, the only way that LS can lead to less dilation than CE is if LS memorizes less than CE. Hence, the goal in our analysis is to show that, under the MD model, LS indeed leads to less memorization than CE.

Note that this description is observed empirically in the experiments of Section 4.1.

Next we define the MD model in the two class setting.

**Definition 4.1.** We call the following minimization problem $\mathcal{MD}$. Minimize the MD risk

$$\mathcal{R}_{\lambda,\eta,\alpha}(U,r) := F_{\lambda,\alpha}(W,H,r) + \eta G_{\lambda,\alpha}(W,U,r),$$

with respect to the noisy feature embedding $U = [u_1, u_2] \in \mathbb{R}^{2 \times M}$ and the standard deviation $r \geq 0$, under the constraints

$$\eta \|h_1 - u_2\| \leq \frac{C_{MDR}}{\|h_1 - h_2\|} \quad (4)$$

$$\eta \|h_2 - u_1\| \leq \frac{C_{MDR}}{\|h_1 - h_2\|}. \quad (5)$$

Here,

- $H \in \mathbb{R}^{2 \times M}$ and $W \in \mathbb{R}^{M \times 2}$ form an NC configuration (see Definition 3.1).
- $C_{MD} > 0$ is called the memorization-dilation slope, $0 \leq \alpha < 1$ is called the LS parameter, $\eta > 0$ the noise level, and $\lambda > 0$ the regularization parameter.
- $F_{\lambda,\alpha}$ is the component in the (regularized) risk that is associated with the correctly labeled samples,

$$F_{\lambda,\alpha}(W,H,r) := \int \left( \ell_{\alpha}(W, h_1 + v, y_1^{(\alpha)}) + \lambda \|h_1 + v\|^2 \right) d\mu_1^1(v)$$

$$+ \int \left( \ell_{\alpha}(W, h_2 + v, y_2^{(\alpha)}) + \lambda \|h_2 + v\|^2 \right) d\mu_2^2(v),$$

where $\mu_1^1$ and $\mu_2^2$ are some distributions with mean 0 and standard deviation $r$, and $\ell_{\alpha}$ is the LS loss defined in [1].

- $G_{\lambda,\alpha}$ is the component in the (regularized) risk that is associated with the corrupted samples, defined as

$$G_{\lambda,\alpha}(W,U,r) = \ell_{\alpha}(W, u_1, y_1^{(\alpha)}) + \ell_{\alpha}(W, u_2, y_2^{(\alpha)}) + \lambda \|u_1\|^2 + \lambda \|u_2\|^2.$$

**Remark 1.** We skipped the normalization by $N = 2$ in the empirical risk to keep the definitions of $F_{\lambda,\alpha}$ and $G_{\lambda,\alpha}$ better readable. Certainly this has no significant effect on our results later on.
The $\mathcal{MD}$ problem can be interpreted as follows. The amount of memorization in the first class is defined to be $\eta \| h_2 - u_1 \|$, since the more noise $\eta$ there is, the more examples we need to memorize. The amount of memorization in the second class is defined the same. The dilation is defined to be $\frac{h_1 - h_2}{\| h_1 - h_2 \|^2}$, which models a similar quantity to $2$. The constraints of Assumptions 4.2. tell us that in order to map noisy samples $u_1$ away from $h_2$, we have to pay with dilation $r$. The larger $r$ is, the further away we can map $u_1$ from $h_2$. The correspondence between memorization and dilation is linear with slope $C_M$ by assumption. There are two main forces in the optimization problem—$u_1$ would like to be as close as possible to its optimal position $h_2$, and similarly $u_2$ likes to be close to $h_1$. In view of the constraints of Assumptions 4.2. to achieve this, $r$ has to be increased to $r_{\text{max}} := \frac{1}{\eta\| h_1 - h_2 \|^2}$. On the other hand, the optimal $r$ for the term $F_{\lambda, \alpha}$ is $r = 0$, namely, the layer-peeled NC configuration. An optimal solution hence balances between memorization and dilation. See Figure 2 for a visualization of the MD model.

Our goal is to compare the optimal dilation $r$ between the LS and CE losses, and for simplicity we choose to study the $\mathcal{MD}$ problem in the feature space of dimension 2, i.e. $M = 2$. Toward this goal, we will impose several model assumptions. These are stated in Assumption 4.2. below and are discussed with more details in the appendix.

Briefly speaking, we assume that $W$ and $H$ are given as an NC configuration of the noiseless model given in Theorem 3.2. Furthermore we distinguish between the two loss functions by setting $\alpha = 0$ for CE and $\alpha = \alpha_0 > 0$ for LS. This will give two different scales of the feature embeddings. More accurately, we denote the optimal $H$ corresponding to the CE loss by $H^{CE}$ and the optimal $G$ with respect to LS by $H^{LS}$. For a choice of LS parameters satisfying $\alpha_0 > 2\sqrt{\lambda_W \lambda_H}$, it can be proved that

$$
\gamma := \frac{\| H^{CE} \|}{\| H^{LS} \|} > 1. \tag{6}
$$

Since (6) is an important relation in our proofs, we restrict ourselves to the case $\alpha_0 > 2\sqrt{\lambda_W \lambda_H}$, which is a reasonable assumption.

Next, for the distributions $\mu^1_\alpha$ and $\mu^2_\alpha$, we will assume a certain symmetry about the means, as well as bounded supports with radius proportional to $r$. Finally, we assume that the noise level $\eta$ is small relatively to the LS parameter $\alpha_0$.

**Assumption 4.2.**

1. Let $\alpha_0 > 0$. We assume that $(W, H)$ is equal to the solution of

$$
\min_{W, H} \ell_{\alpha, W, h_1, y_1^{(\alpha)}} + \ell_{\alpha, W, h_2, y_2^{(\alpha)}} + \lambda_W \| W \|^2 + \lambda_H \| H \|^2
$$

s.t. $H \geq 0$.

where we denote $(W, H) = (W^{CE}, H^{CE})$ when $\alpha = 0$ and $(W, H) = (W^{LS}, H^{LS})$ when $\alpha = \alpha_0$.

2. Assume that the distributions $\mu^1_\alpha$ and $\mu^2_\alpha$ are centered, in the sense that

$$
\int (w_2 - w_1, v) d\mu^1_\alpha(v) = \int (w_1 - w_2, v) d\mu^2_\alpha(v) = 0,
$$

$$
\int (h_1, v) d\mu^1_\alpha(v) = \int (h_2, v) d\mu^2_\alpha(v) = 0.
$$

Furthermore, we assume that there exists a constant $A > 0$ such that $\|v\| \leq Ar$ for any vector $v$ that lies in the support of $\mu^1_\alpha$ or in the support of $\mu^2_\alpha$.
3. Assume that the noise level $\eta$ and the LS parameter $\alpha_0$ satisfy the following. We suppose $\alpha_0 > 4\sqrt{\lambda W\lambda_H}$, which guarantees $\gamma := \|H^CE\|/\|H^{LS}\| > 1$. We moreover suppose that $\eta$ is sufficiently small to guarantee $\eta^{1/2} \ll 1 - \frac{1}{\gamma}$.

Our main result in this section is that LS leads to lower dilation in comparison to CE. This is formally stated as below.

**Theorem 4.3.** Suppose that Assumption 4.2 holds true and $\lambda := \lambda_H$. Let $r^CE_*$ and $r^LS_*$ be the optimal dilations, i.e. the optimum $r$ in the MD problem, corresponding to the CE and LS loss (accordingly $\alpha = 0$ and $\alpha = \alpha_0$), respectively. Then it holds that

$$r^CE_* \|h^CE_1 - h^CE_2\| > r^LS_* \|h^{LS}_1 - h^{LS}_2\|.$$ 

Theorem 4.3 reveals a mechanism by which LS achieves better generalization than CE. It is proven that LS memorizes and dilates less than CE, which is associated with better generalization.

The memorization-dilation model combines a statistical term $F_{\lambda,\alpha}$, that describes the risk over the distribution of feature embeddings of samples with clean labels, and an empirical term $\eta G_{\lambda,\alpha}$ that describes the risk over training samples with noisy labels. One point of view that can motivate such a hybrid statistical-empirical definition is the assumption that the network only memorizes samples of noisy labels, but not samples of clean labels. Such a memorization degrades (dilates) both the collapse of the training and test samples, possibly with different memorization-dilation slopes. However, memorization is not limited to corrupted labels, but can also apply to samples of clean labels [Feldman and Zhang, 2020], by which the learner can partially negate the dilation of the training features (but not test features). The fact that our model does not take the memorization of clean samples into account is one of its limitations. We believe that future work should focus on modeling the total memorization of all examples. Nevertheless, we believe that our current MD model has merit, since 1) noisy labels are memorized more than clean labels, and especially in the low noise regime the assumption of observing memorization merely for corrupted labels appears reasonable, and 2) our approach and proof techniques can be the basis of more elaborate future MD models.

5 Conclusion

In this paper, we first characterized the global minimizers of the Layer-Peeled Model (or the Unconstrained Features Model) with the positivity condition on the feature representations. Our characterization captures better the empirically observed neural collapse (NC) behavior that has been studied in recent works. Besides the conventional cross-entropy (CE) loss, we studied the model in case of the label smoothing (LS) loss, showing that NC also occurs when applying this technique.

Then we extended the model to the so-called Memorization-Dilation (MD) Model by incorporating the limited expressivity of the network. Using the MD model, which is supported by our experimental observations, we show that when trained with the LS loss, the network memorizes less than when trained by the CE loss. This poses one explanation to the improved generalization performance of the LS technique over the conventional CE loss.

Our model has limitations, however. Firstly, our proof is currently restricted to a 2D feature space. Nevertheless, we believe that our proof techniques can be extended to the case of higher dimensional features. Secondly, and more importantly, our model is limited to the case of two classes. Motivated by promising results on the applicability of our model to the multi-class setting, we believe that future work should focus on extending the MD model in this respect. With such extensions, memorization-dilation analysis has the potential to underlie a systematic comparison of the generalization capabilities of different losses, such as CE, LS, and label relaxation, by analytically deriving formulas for the amount of memorization associated with each loss.

Acknowledgments and Disclosure of Funding

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center “On-The-Fly Computing” (CRC 901 project no. 160364472). Moreover,
the authors gratefully acknowledge the funding of this project by computing time provided by the Paderborn Center for Parallel Computing (PC$^2$).

References

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 233–242. PMLR, 2017.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. CoRR, abs/1906.11300, 2019.

Weinan E and Stephan Wojtowytsch. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers, 2021.

Tolga Ergen and Mert Pilanci. Revealing the structure of deep neural networks via convex duality. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 3004–3014. PMLR, 2021.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. Proceedings of the National Academy of Sciences of the United States of America, 118, 2021.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. CoRR, abs/2112.15121, 2021.

X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. CoRR, abs/2106.02073, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image classification. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778. IEEE Computer Society, 2016.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 8580–8589, 2018.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features?, 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Canada, 2009.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 3560–3569. PMLR, 09–15 Jun 2019.

Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 2908–2913. PMLR, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.
Liam Li, Kevin G. Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. In Proceedings of Machine Learning and Systems 2020, MLSys 2020. Austin, TX, USA, March 2-4, 2020. mlsys.org, 2020.

Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. CoRR, abs/2106.03212, 2021.

Julian Lienen and Eyke Hüllermeier. From label smoothing to label relaxation. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021, pages 8583–8591. AAAI Press, 2021.

Jianfeng Lu and Stefan Steinerberger. Neural collapse with cross-entropy loss. CoRR, abs/2012.08465, 2020.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. CoRR, abs/2011.11619, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 4696–4705, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NIPS, Granada, Spain, November 12-17, Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. CoRR, abs/2008.08186, 2020.

Tomaso A. Poggio and Qianli Liao. Generalization in deep network classifiers trained with the square loss. 2020.

Tomaso A. Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. CoRR, abs/2101.00072, 2021.

Daniel A. Roberts, Sho Yaida, and Boris Hanin. The principles of deep learning theory. CoRR, abs/2106.10165, 2021.

Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip H. S. Torr. How benign is benign overfitting? In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

Ruoyu Sun. Optimization for deep learning: An overview. Journal of the Operations Research Society of China, 8:249–294, 2020.
Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Re-thinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Global optimality conditions for deep neural networks. *In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Zhihui Zhu, Daniel Soudry, Yonina C. Eldar, and Michael B. Wakin. The global optimization geometry of shallow linear neural networks. *J. Math. Imaging Vis.*, 62(3):279–292, 2020.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *CoRR*, abs/2105.02375, 2021.
A Experimental details

A.1 Memorization experiments

A.1.1 Setting

To produce the results of Section 4.1, we trained simple multi-layer perceptron models with 9 hidden layers of width 2048. Each layer involves a batch normalization layer and uses the parameterized activation parameters (one of ReLU or sigmoid) throughout the network. To train the network, we employed SGD as optimizer with a learning rate of 0.1 that is multiplied by 0.1 each 40 epochs. We further employed a Nesterov momentum of 0.9. In total, we trained for 200 epochs, which was sufficient to observe the neural collapse phenomenon. We ensure that the parameterization works reasonably well for all losses for a fair and realistic comparison. We further use a weight decay regularization of 0.001. The batch size is set to 512 for all experiments. Each assessed parameter combination has been executed 5 times to gain statistically meaningful results.

The penultimate layer feature dimension was set to the number of classes $N$. On top of the encoding network architecture, a linear softmax classifier is attached. The entire model is optimized for four different losses: Conventional cross-entropy with degenerate target distributions, label smoothing with a default smoothing parameter of $\alpha = 0.1$, label relaxation with an imprecisiation degree of $\alpha = 0.1$ and mean squared error.

![Figure 3](image)

Figure 3: Exemplary penultimate layer activations (post training) of the clean and corrupted training data in the 2D feature space. Green 1 represent test instances of clean label 1 data, blue 2 represent clean test instances of label 2 data, red 1 represent instances of training samples that were originally labeled as 1 but were changed to label 2, red 2 represent instances of training samples that were originally labeled as 2 but were changed to label 1. (a) Collapse to a sub-optimal configuration, where one of the class centroids is at the origin. (b) The class centroids are along the axes, corresponding to the optimal NC configuration of Definition 3.1.

In the idealized experimental environment, we considered the datasets MNIST and CIFAR-10 as show cases. To reduce the problem complexity for the theoretical analysis, we subsampled the first $N$ classes of each dataset, all other instances were excluded. The binary case $N = 2$ allows for a convenient analysis of the learned feature representations of the penultimate layer with $d = N = 2$. In case of $N = 2$, cross-entropy and its derived losses did not always attain the optimal NC configuration through SGD, namely did not always align the class centroids along the axes. In some cases, the learned representation collapsed to one class centroid in the origin and the other one on a diagonal line in the positive quadrant in the 2D feature space. Figure 3 shows this case in (a) and a case the corresponds to the optimal NC configuration in (b). We filtered out the former examples, as these only infrequently occur in the $d = 2$ case.

A.1.2 Conventional label noise: Further results

In the first label noise setting, we considered conventional label corruption, which is described in the paper. Beyond the results shown in the main part, we provide further evidence of our findings here. To this end, we repeated the experiment with different numbers of classes, namely $N \in \{3, 5, 10\}$.
Figures 4, 5, and 6 show the results. Albeit not perfect, a similar dependence can be observed for higher dimensions.

Figure 4: Feature collapse of the test instances in terms of $\sqrt{NC_1^{\text{test}}}$ per memorization and the resulting test accuracies (averaged over ten seeds) for $N = 3$.

Figure 5: Feature collapse of the test instances in terms of $\sqrt{NC_1^{\text{test}}}$ per memorization and the resulting test accuracies (averaged over ten seeds) for $N = 5$.

A.1.3 Latent noise classes: Further results

While we considered “conventional” label noise in the first experiment, we extend our analysis to a different form of label noise: For each original class, we split an instance fraction $\eta \in [0.025, 0.2]$ of each class apart and introduce new latent (noise) classes. Thus, the learner has again to separate these instance from their original class as it is pretended to face different classes. We consider the same basic architectural framework, but with four classes instead of two. To preserve compatibility to the previous experiments, we keep $d = N = 2$. We repeated each run with 5 different random seeds.

For this different type of label noise, the results shown in Figure 7 match the observations made before. Although the correspondence is not as clear as in the standard noise model, CE and LS are
close to sharing the same curve for both datasets. Similarly, one can see a linear trend in the test collapse per memorization, which is now defined between the instances of the latent class to the test centroid of the original class. Also, we see similar trends regarding the generalization performance.

A.2 Large-scale experiments

A.2.1 Setting

Beyond the experiments in the previous section, we analyzed the neural collapse properties when training commonly used architectures, such as ResNet [He et al. 2016] and VGG [Simonyan and Zisserman 2015] models. To this end, we trained the variants ResNet18 and VGG13 on the four benchmarks MNIST, FashionMNIST, CIFAR-10 and SVHN. Here, we consider conventional label noise degrees $\eta \in \{0, 0.1, 0.2, 0.3\}$. To ensure a fair comparison, we optimized hyperparameters, such as the learning rate schedule and the smoothing and relaxation parameters $\alpha$ for LS and LR, in a Bayesian optimization using Hyperband [Li et al. 2020]. We tuned these parameters based on a 20% separated validation split in the no-noise case $\eta = 0$, and applied the best parameters in the noise settings with $\eta > 0$. We refer to the appendix for a more comprehensive overview over the experimental setting and further result.

Just as in the previous experiments, we used SGD as optimizer with Nesterov momentum of 0.9, trained for 200 epochs with a batch size of 512. However, as opposed to the setting before, we performed a Bayesian hyperparameter optimization employing a Hyperband scheduler [Li et al. 2020] on a separated 20% validation split. To this end, we used the sklearn implementation and optimized for 30 iterations. Table 1 shows the considered hyperparameter space. The final model used within the evaluation was eventually trained on the complete training set (i.e., including the validation set).
Table 2 shows the resulting generalization performances as an average over 3 seeds. As can be seen, label smoothing consistently improves over cross-entropy, confirming both the empirical and theoretical observations as presented before. These results suggest that label smoothing is particularly appealing in case of label noise.

### A.3 Technical infrastructure

To realize the experiments, we proceeded from the official code base of Zhu et al. [2021] and augmented it by further baselines, models and our evaluation metrics. This implementation lever-
To execute the runs, we used Nvidia GPU accelerators (1080/2080 Ti, Titan RTX) in a modern cluster environment. Our code is publicly available at https://github.com/julilien/MemorizationDilation.

B Proof of Theorem 3.2

In this appendix we introduce the proof of our theorem on the layer-peeled model, i.e. Theorem 3.2

B.1 Reformulation of the LS empirical risk

Given a smoothing parameter $\alpha \in [0, 1)$, we will write the LS empirical risk introduced in Section 3 in more details,

$$L_\alpha(W, H) = \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \ell_\alpha \left( W \cdot h_n^{(k)}, y_n^{(\alpha)} \right)$$

$$= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m=1}^{N} -y_{nm}^{(\alpha)} \log \left( p_W(h_n^{(k)})_m \right)$$

$$= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \left[ (1 - \frac{N-1}{N} \alpha) \log \left( \sum_{i=1}^{N} e^{(w_i - w_n, h_n^{(k)})} \right) \right.$$  

$$\left. + \sum_{m=1}^{N} \frac{\alpha}{N} \log \left( \sum_{i=1}^{N} e^{(w_i - w_m, h_n^{(k)})} \right) \right]$$

$$= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \left[ (1 - \frac{N-1}{N} \alpha) \log \left( \sum_{i=1}^{N} e^{(w_i - w_n, h_n^{(k)})} \right) \right.$$  

$$\left. + \sum_{m=1}^{N} \frac{\alpha}{N} \log \left( \sum_{i=1}^{N} e^{(w_i - w_m, h_n^{(k)})} \right) \right]$$

$$= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \left[ (1 - \frac{N-1}{N} \alpha) \log \left( \sum_{i=1}^{N} e^{(w_i - w_n, h_n^{(k)})} \right) \right.$$  

$$\left. + \sum_{m=1}^{N} \frac{\alpha}{N} \sum_{m \neq n} \log \left( \sum_{i=1}^{N} e^{(w_i - w_m, h_n^{(k)})} \right) + \sum_{m=1}^{N} \frac{\alpha}{N} \left< w_n - w_m, h_n^{(k)} \right> \right]$$

$$= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \left[ (1 - \frac{N-1}{N} \alpha) \log \left( \sum_{i=1}^{N} e^{(w_i - w_n, h_n^{(k)})} \right) \right.$$  

$$\left. + \sum_{m=1}^{N} \frac{\alpha}{N} \sum_{m \neq n} \log \left( \sum_{i=1}^{N} e^{(w_i - w_m, h_n^{(k)})} \right) + \sum_{m=1}^{N} \frac{\alpha}{N} \left< w_n - w_m, h_n^{(k)} \right> \right]$$

$$= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \left[ (1 + \sum_{m=1}^{N} \frac{1}{m \neq n} \log \left( \sum_{i=1}^{N} e^{(w_i - w_m, h_n^{(k)})} \right) \right.$$  

$$\left. - \sum_{m=1}^{N} \frac{\alpha}{N} \left< w_m - w_n, h_n^{(k)} \right> \right].$$

https://pytorch.org/ BSD license
https://pytorch.org/vision/ BSD license
Shortly speaking, this differs from the conventional CE loss just by an additional bilinear term
\[ \frac{1}{N^2 K N^2} \sum_{k=1}^{K} \sum_{m=1}^{N} \sum_{m \neq n} \left\langle w_m - w_n, h_n^{(k)} \right\rangle. \]

B.2 Technical lemmata

Lemma B.1. We define
\[ P(W, H) := \frac{1}{KN(N-1)} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{m \neq n} \left\langle w_m - w_n, h_n^{(k)} - h_m^{(k)} \right\rangle. \]

Then under the condition \( H \geq 0 \) it holds
\[ P(W, H) \geq -\frac{1}{\sqrt{KN(N-1)}} \| W \| \| H \|. \tag{7} \]

The inequality (7) becomes an equality if and only if the following conditions hold simultaneously
\[ \sum_{n=1}^{N} w_n = 0 \tag{8} \]
\[ \left\langle h_n^{(k)}, h_m^{(k)} \right\rangle = 0 \quad \text{for all } m, n \in [N], k \in [K], m \neq n, \tag{9} \]
\[ \| h_n^{(k)} \| \quad \text{is independent of } n, k, \tag{10} \]
\[ w_m - w_n = c'(h_m^{(k)} - h_n^{(k)}) \quad \text{for some } c' > 0 \text{ not depending on } m, n, k. \tag{11} \]

Proof. Using the Cauchy-Schwarz inequality we get
\[ P(W, H) := \frac{1}{KN(N-1)} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{m \neq n} \left\langle w_m - w_n, h_n^{(k)} - h_m^{(k)} \right\rangle \]
\[ = \frac{1}{KN(N-1)} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{m \neq n} \left\langle w_m - w_n, h_n^{(k)} \right\rangle - \left\langle w_m - w_n, h_m^{(k)} \right\rangle \]
\[ \geq - \frac{1}{KN(N-1)} \sum_{k=1}^{K} \left( \sum_{n=1}^{N} \sum_{m=n+1}^{N} \| w_n - w_m \|^2 \right) \left( \sum_{n=1}^{N} \sum_{m=n+1}^{N} \| h_n^{(k)} - h_m^{(k)} \|^2 \right) \]
\[ = - \frac{1}{KN(N-1)} \left\{ \sum_{n=1}^{N} \sum_{m=n+1}^{N} \| w_n - w_m \|^2 \sum_{k=1}^{K} \left( \sum_{n=1}^{N} \sum_{m=n+1}^{N} \| h_n^{(k)} - h_m^{(k)} \|^2 \right) \right\} \]
\[ =: P_1 \]
\[ =: P_2 \]

Further application of Cauchy-Schwarz inequality yields
\[ P_1 = \sqrt{\sum_{n=1}^{N} \sum_{m=n+1}^{N} \| w_n - w_m \|^2} \]
\[ = \sqrt{N \sum_{n=1}^{N} \| w_n \|^2 - \sum_{n=1}^{N} \| w_n \|^2} \leq \sqrt{N} \| W \| \]

18
and

\[ P_2 = \sum_{k=1}^{K} \left( \sum_{n=1}^{N} \sum_{m=n+1}^{N} \left\| h_n^{(k)} - h_m^{(k)} \right\|^2 \right) \]

\[ = \sum_{k=1}^{K} \sqrt{\left( N - 1 \right) \sum_{n=1}^{N} \left\| h_n^{(k)} \right\|^2 - \sum_{n=1}^{N} \sum_{m=n+1}^{N} \left\langle h_n^{(k)}, h_m^{(k)} \right\rangle} \]

\[ \leq \sqrt{N - 1} \sum_{k=1}^{K} \sum_{n=1}^{N} \left\| h_n^{(k)} \right\| \]

\[ \leq \sqrt{KN(N - 1)} \sqrt{\sum_{k=1}^{K} \sum_{n=1}^{N} \left\| h_n^{(k)} \right\|^2} = \sqrt{KN(N - 1)} \| H \| \]

Therefore

\[ P(W, H) \geq -\frac{1}{KN(N - 1)} P_1 P_2 \geq -\frac{1}{\sqrt{K(N - 1)}} \| W \| \| H \|. \]

This becomes an equality if and only if

- The upper bound on \( P_1 \) becomes equality, i.e.
  \[ \sum_{n=1}^{N} w_n = 0 \]

- The upper bound on \( P_2 \) becomes equality, i.e.
  \[ \left\langle h_n^{(k)}, h_m^{(k)} \right\rangle = 0 \quad \text{for all } m, n \in [N], k \in [K], m \neq n, \]

  \[ \left\| h_n^{(k)} \right\| \text{ is independent of } n, k. \]

- The estimate \( P \geq -\frac{1}{KN(N - 1)} P_1 P_2 \) becomes an equality, i.e.
  \[ w_m - w_n = c (h_m^{(k)} - h_n^{(k)}) \quad \text{for some } c' > 0 \text{ not depending on } m, n, k \]

Lemma B.2. Assume that the inequality \( \[ \] \) shown in Lemma B.1 equalizes. Furthermore assume that there exist constants \( c_{n, k} \in \mathbb{R} \) (depending on \( n \in [N] \) and \( k \in [K] \)) and \( c \in \mathbb{R} \) such that

\[ \left\langle w_m, h_n^{(k)} \right\rangle = c_{n, k} \quad \text{for every } m \in [N] \setminus \{n\}, \]

\[ \sum_{m=1}^{N} \left\langle w_m - w_n, h_n^{(k)} \right\rangle = c \text{ (not depending on } n, k) \],

for all \( n \in [N] \) and \( k \in [K] \). Then, the pair \((W, H)\) must form a neural collapse configuration. Conversely, if \((W, H)\) is a neural collapse configuration, then \( \[ \] \) becomes an equality and the conditions \( \{12, 13\} \) both hold true.

Proof: The converse implication is straightforward. We prove here the forward implication. By \( \{8, 13\} \) we have for any \( n \in [N] \) and \( k \in [K] \) that

\[ 0 = \sum_{m=1}^{N} \left\langle w_m, h_n^{(k)} \right\rangle = \sum_{m=1}^{N} \left\langle w_m - w_n, h_n^{(k)} \right\rangle + N \left\langle w_n, h_n^{(k)} \right\rangle, \]

\[ \text{substituting } c \text{ for } c', \text{ we get } \]

\[ 0 = \sum_{m=1}^{N} \left\langle w_m, h_n^{(k)} \right\rangle = \sum_{m=1}^{N} \left\langle w_m - w_n, h_n^{(k)} \right\rangle + N \left\langle w_n, h_n^{(k)} \right\rangle, \]
Combining (14,15) with (9,11) gives
\[
\langle w_n, h_n^{(k)} \rangle = -\frac{c}{N}.
\] (14)

Combining this with (13,12) gives
\[
c = \sum_{m=1}^{N} \langle w_m - w_n, h_n^{(k)} \rangle = (N - 1)c_{n,k} - (N - 1)\frac{-c}{N},
\]
and hence
\[
\langle w_m, h_n^{(k)} \rangle = c_{n,k} = \frac{c}{N(N - 1)}.
\] (15)

Combining (14,15) with (9,11) gives
\[
-\frac{2c}{N - 1} = \langle w_n - w_m, h_n^{(k)} - h_m^{(k)} \rangle = c' \left( \|h_n^{(k)} - h_m^{(k)}\|^2 = c' \left( \|h_n^{(k)}\|^2 + \|h_m^{(k)}\|^2 \right) \right),
\] (16)

Combining (16) with (10) shows that for every \(n \in [N]\) and \(k \in [K]\), it holds
\[
\|h_n^{(k)}\|^2 = \frac{-c}{c'(N - 1)}.
\] (17)

On the other hand, it follows also from (14,15) that
\[
\|w_n\|^2 - \|w_m\|^2 = \langle w_n - w_m, w_n + w_m \rangle = c' \left( \langle h_n^{(k)} - h_m^{(k)}, w_n + w_m \rangle = 0, \right)
\] (18)

and hence the vectors \(w_n, n \in [N]\) have the same length, which can be computed via
\[
N^2 \|w_1\|^2 = N \sum_{n=1}^{N} \|w_n\|^2 = \sum_{n>m} \|w_n - w_m\|^2
\]
\[
= c' \sum_{n>m} \langle w_n - w_m, h_n^{(k)} - h_m^{(k)} \rangle
\]
\[
= c' \cdot \frac{N(N - 1)}{2} \cdot \frac{-2c}{N - 1}
\]
\[
= -cc'N.
\]

Hence, for each \(n \in [N]\), it holds
\[
\|w_n\|^2 = -\frac{cc'}{N}.
\] (19)

Now let \(h^{(k)} := \frac{1}{N} \sum_{m=1}^{N} h_m^{(k)}\) for each \(k \in [K]\). Observe that it holds
\[
\langle w_n, h_n^{(k)} - h^{(k)} \rangle = \frac{N - 1}{N} \langle w_n, h_n^{(k)} \rangle - \frac{1}{N} \sum_{m=1}^{N} \langle w_n, h_m^{(k)} \rangle
\]
\[
= -\frac{(N - 1)c}{N^2} - \frac{c}{N^2}
\]
\[
= \frac{c}{N}.
\] (20)

On the other hand, from (17) we have for each \(n \in [N]\) and \(k \in [K]\) that
\[
\|h_n^{(k)} - h^{(k)}\|^2 = \|h_n^{(k)}\|^2 - 2 \left( h_n^{(k)} \cdot \frac{1}{N} \sum_m h_m^{(k)} \right) + \frac{1}{N^2} \left( \sum_m h_m^{(k)} \right)^2
\]
\[
= \frac{N - 1}{N} \|h_n^{(k)}\|^2
\]
\[
= \frac{N - 1}{N} \cdot \frac{-c}{c'(N - 1)}
\]
\[
= -\frac{c}{Nc'}.
\] (21)
From (19) it follows that
\[ \langle w_n, h_n^{(k)} - h^{(k)} \rangle = \| w_n \| \| h_n^{(k)} - h^{(k)} \| , \]
which implies that \( w_n \) is parallel to \( h_n^{(k)} - h^{(k)} \) for every \( n \in [N] \) and \( k \in [K] \). More precisely, by combining this finding with the above calculation of \( \| w_n \| \) and \( \| h_n^{(k)} - h^{(k)} \| \) in (19) we obtain
\[ w_n = c \left( h_n^{(k)} - h^{(k)} \right) . \tag{22} \]

Finally it is left to show that \( h_n^{(k)} = h_n^{(k)} \) for any \( k, \ell \in [K] \). For this observe that \( h_n^{(k)} - h^{(k)} = h_n^{(\ell)} - h^{(\ell)} = w_n \) implies
\[ \| h_n^{(k)} \|^2 = \| h_n^{(\ell)} + h^{(k)} - h^{(\ell)} \|^2 = \| h_n^{(\ell)} \|^2 + 2 \langle h^{(k)} - h^{(\ell)}, h_n^{(k)} \rangle + \| h^{(k)} - h^{(\ell)} \|^2 , \]
and thus
\[ 2 \langle h^{(k)} - h^{(\ell)}, h_n^{(k)} \rangle + \| h^{(k)} - h^{(\ell)} \|^2 = 0 . \]

Similarly
\[ 2 \langle h^{(\ell)} - h^{(k)}, h_n^{(\ell)} \rangle + \| h^{(k)} - h^{(\ell)} \|^2 = 0 . \]

Combining the two equalities and taking the sum over \( n \) we obtain
\[ \| h^{(k)} - h^{(\ell)} \|^2 = 0 , \]
which means that \( h^{(k)} = h^{(\ell)} \) and therefore \( h_n^{(k)} = h_n^{(\ell)} \).

\[ \square \]

**B.3 Proof of Theorem 3.2**

**Proof.**

Step 1. First we introduce a lower bound on the (unregularized) loss. Using Jensen’s inequality for the convex function \( t \mapsto e^t \) we obtain that for each \( n \in [N] \) and \( k \in [K] \) it holds
\[ \sum_{m=1}^{N} e^{\langle w_m - w_n, h_n^{(k)} \rangle} \geq (N-1) e^{\frac{1}{N} \sum_{m \neq n}^{N} \langle w_m - w_n, h_n^{(k)} \rangle} , \]
with equality if and only if \( \langle w_m, h_n \rangle = c_n \) for every \( m \neq n \), independently of \( m \), for some constant \( c_n \). Inserting this into the formulation of \( L_{\alpha} \) in Section 5.1 we get
\[ L_{\alpha}(W, H) \geq \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \left[ \log \left( 1 + (N-1)e^{\frac{1}{N} \sum_{m \neq n}^{N} \langle w_m - w_n, h_n^{(k)} \rangle} \right) \right. \]
\[ \left. - \sum_{m=1}^{N} \frac{\alpha}{N} \left\langle w_m - w_n, h_n^{(k)} \right\rangle \right] . \]

Observe that the function \( t \mapsto \log \left( 1 + (N-1)e^{\frac{1}{N} t} \right) \) is also convex, hence applying again Jensen’s inequality we can lower bound the right-hand side in the estimate above, and obtain
\[ L_{\alpha}(W, H) \geq \log \left( 1 + (N-1)e^{\frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m \neq n}^{N} \langle w_m - w_n, h_n^{(k)} \rangle} \right) \]
\[ - \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m \neq n}^{N} \frac{\alpha}{N} \left\langle w_m - w_n, h_n^{(k)} \right\rangle . \tag{23} \]

Equality in (23) occurs if and only if the conditions (12) (see Lemma B.2) hold simultaneously.
Step 2. Recall that with the notation $P = P(W, H)$ from Lemma 3.1, the inequality (23) becomes
\[ \mathcal{L}_\alpha(W, H) \geq \log \left( 1 + (N-1)e^P \right) - \beta P + \lambda_W \|W\|^2 + \frac{\lambda_H}{K} \|H\|^2 =: \hat{L}(W, H), \]
with $\beta := \frac{N-1}{N} \alpha > 0$. Consider the function $g : \mathbb{R} \to \mathbb{R}$,
\[ g(t) := \log \left( 1 + (N-1)e^t \right) - \beta t. \]
Since $g$ is convex (as it differs from a convex function only by an additional linear function), it has a unique minimum specified as the root of the derivative
\[ g'(t) = \frac{(N-1)e^t}{1 + (N-1)e^t} - \beta. \]
We now aim to find a constant lower bound on the right-hand side $\hat{L}(W, H)$ of (24). We consider the following three cases, corresponding to three different regions of the feasible set of $(W, H)$:

(a) Case $t_0 > P(W, H)$: We will show that the minimizers of $\mathcal{L}_\alpha$ cannot be in this region. Toward a contradiction, assume that there is a minimizer $(W_0, H_0)$ s.t. $P(W_0, H_0) < t_0$. We construct $(W_1, H_1)$ to be a NC configuration (according to Definition 3.1) satisfying $\|W_0\| = \|W_1\|$ and $\|H_0\| = \|H_1\|$. Then we have
\[ P(W_1, H_1) = -\frac{1}{\sqrt{K(N-1)}} \|W_1\| \|H_1\| \leq \frac{1}{\sqrt{K(N-1)}} \|W_0\| \|H_0\| = P(W_0, H_0) < t_0 < 0. \]
By rescaling $W_1, H_1$ (with a constant smaller than 1) we obtain a pair $(W, H)$ with $P(W, H) = t_0$ and $\|W\| < \|W_0\|$, $\|H\| < \|H_0\|$. Thus it holds
\[ \mathcal{L}_\alpha(W_0, H_0) \geq \log \left( 1 + (N-1)e^{P(W_0, H_0)} \right) - \beta P(W_0, H_0) + \lambda_W \|W_0\|^2 + \lambda_H \|H_0\|^2 \]
\[ > \log \left( 1 + (N-1)e^{t_0} \right) - \beta t_0 + \lambda_W \|W_0\|^2 + \frac{\lambda_H}{K} \|H_0\|^2 = \mathcal{L}_\alpha(W, H), \]
which means that $(W_0, H_0)$ cannot be a minimizer of $\mathcal{L}_\alpha$. Note that the last equality holds because the inequality (23) equalizes when $(W, H)$ is a NC configuration (see Lemma 3.2).

(b) Case $P(W, H) \geq t_0$ 
\[ \geq -\frac{1}{\sqrt{K(N-1)}} \|W\| \|H\| : \]
We will show that at the minimizers in this region, $P$ must be $t_0$. Assume that $(W_0, H_0)$ is a minimizer of $\hat{L}$ in this region with $P(W_0, H_0) \neq t_0$. Then we consider all pairs $(W, H)$ with $\|W\| \leq \|W_0\|$ and $\|H\| \leq \|H_0\|$. By continuity we have that $P(W, H)$ can take all values in the interval
\[ \left[ -\frac{1}{\sqrt{K(N-1)}} \|W_0\| \|H_0\|, \frac{1}{\sqrt{K(N-1)}} \|W_0\| \|H_0\| \right]. \]

*Note that here the root $t_0$ exists as long as $\beta > 0$, for $\beta = 0$ we may, for convenience, define $t_0 := -\infty$ (this will correspond to Case (c) below).
which also includes \( t_0 \). It follows that \( \tilde{L}(W, H) < \tilde{L}(W_0, H_0) \), so \((W_0, H_0)\)
cannot be a minimizer of \( \tilde{L} \), meaning that a minimizer \((W, H)\) of \( \tilde{L} \) must satisfy
\[ P(W, H) = t_0. \]
The minimization of \( \tilde{L} \) then reduces to
\[
\min_{W, H} \lambda_W \|W\|^2 + \lambda_H \|H\|^2 \quad \text{s.t.} \quad - \frac{1}{\sqrt{K(N - 1)}} \|W\| \|H\| = t_0.
\]
Observe that
\[
\lambda_W \|W\|^2 + \frac{\lambda_H}{K} \|H\|^2 \geq 2 \sqrt{\frac{\lambda_W \lambda_H}{K}} \|W\| \|H\|
\]
\[
\geq -2t_0 \sqrt{(N - 1)\lambda_W \lambda_H}.
\]
Therefore we have \( \tilde{L}(W, H) \geq g(t_0) - 2t_0 \sqrt{(N - 1)\lambda_W \lambda_H} \) and this equals if and
only if the following conditions hold:
\begin{itemize}
  \item \( P(W, H) = t_0 \)
  \item \( \lambda_W \|W\|^2 = \lambda_H \|H\|^2 \) and \( \|W\| \|H\| = -\sqrt{K(N - 1)t_0} \).
\end{itemize}
(c) Case \( P(W, H) \geq - \frac{1}{\sqrt{K(N - 1)}} \|W\| \|H\| \geq t_0 \):
In this region, it holds
\[
g(P(W, H)) \geq g\left(- \frac{1}{\sqrt{K(N - 1)}} \|W\| \|H\|\right),
\]
so
\[
\tilde{L}(W, H) \geq f(\|W\|, \|H\|),
\]
with \( f : \mathbb{R}^2 \to \mathbb{R} \),
\[
f(w, h) := \log \left(1 + (N - 1)e^{-Ch}\right) + \beta Ch + \lambda_W w^2 + \frac{\lambda_H}{K} h^2
\]
where we set \( C := \frac{1}{\sqrt{K(N - 1)}} \) to shorten notation. Observe that even though \( w \) and \( h \),
as representatives for \( \|W\| \) and \( \|H\| \) respectively, must be positive, we can consider
them as real number (without positivity). This can be explained as follows. On the one
hand, we are interested in the global minimum of \( f \), at which \( w \) and \( h \) should have the
same sign. On the other hand, since \( f(w, h) = f(-w, -h) \), if \((w, h)\) is a minimum point then certainly \((-w, -h)\) is a minimum point of \( f \).
This observation allows us to set the derivatives of \( f \) to be 0 at the minimum, i.e.
\[
0 = \nabla_w f(w, h) = -\frac{(N - 1)e^{-Cwh}}{1 + (N - 1)e^{-Cwh}} Cb + 2\beta Ch + 2\lambda_W w,
\]
\[
0 = \nabla_h f(w, h) = -\frac{(N - 1)e^{-Cwh}}{1 + (N - 1)e^{-Cwh}} Ca + 2\beta Cw + 2\frac{\lambda_H}{K} h.
\]
Multiplying the first equality with \( w \) and the second with \( h \), we obtain in particular that
\( \lambda_W w^2 = \frac{\lambda_W}{K} h^2 \), and hence \( h = \sqrt{\frac{K\lambda_W}{\lambda_H} w} \). Inserting this into the first inequality while
denoting \( C' := C\sqrt{\frac{K\lambda_W}{\lambda_H}} \) yields
\[
-\frac{(N - 1)e^{-C' a^2}}{1 + (N - 1)e^{-C' a^2}} C' w + 2\beta C' w + 2\lambda_W w = 0.
\]
Excluding the trivial solution \((w, h) = (0, 0)\), so that we can multiply both sides with
\( 1/w \), we get
\[
w^2 = \frac{1}{C} \sqrt{\frac{\lambda_H}{K\lambda_W}} \log \left( (N - 1) \frac{1 - \beta - 2\sqrt{(N - 1)\lambda_W \lambda_H}}{\beta + 2\sqrt{(N - 1)\lambda_W \lambda_H}} \right)
\]
and
\[
h^2 = \frac{1}{C} \sqrt{\frac{K\lambda_W}{\lambda_H}} \log \left( (N - 1) \frac{1 - \beta - 2\sqrt{(N - 1)\lambda_W \lambda_H}}{\beta + 2\sqrt{(N - 1)\lambda_W \lambda_H}} \right)
\]
Finally, it is easy to check that
\[-CWh = \log \left( \frac{1}{N-1} \beta + 2\sqrt{(N-1)\lambda W \lambda H} \right) \geq \log \left( \frac{1}{N-1} \cdot \beta \right) = t_0,\]
i.e. the solution found above belongs indeed to the current region of the feasible set. In summary, we have shown in this case that \( L(W, H) \geq f(w_0, h_0) \) with \((w_0, h_0)\) specified as in \((25, 26)\), and this becomes equality if and only if \( P(W, H) = \frac{1}{\sqrt{R(N-1)}} \|W\| \|H\| \) and \( \|w\| = w_0, \|H\| = h_0 \).

Step 3. We now come back to the actual loss \( L_\alpha \). In both cases (b) and (c) discussed above, we have shown that \( L_\alpha (W, H) \geq L(W, H) \geq \text{const} \) and this can equalize when the conditions in Lemma \( B.2 \) are satisfied. We deduce that \( L_\alpha \) achieves its minimum at either case (b) or (c), while both lead to a NC configuration by Lemma \( B.2 \).

\[\square\]

C Proof of Theorem 4.3

In this appendix we prove our theoretical result on the MD model, namely Theorem 4.3.

C.1 Preparation for the proof

The problem from Definition 4.1 is
\[\min_{U \geq 0, r \geq 0} R_{\lambda, \eta, \alpha} (U, r) := F_{\lambda, \alpha} (W, H, r) + \eta G_{\lambda, \alpha} (W, U, r)\]
under the constraints
\[\eta \|h_1 - u_2\| \leq \frac{C_{MDr}}{\|h_1 - h_2\|},\]
\[\eta \|h_2 - u_1\| \leq \frac{C_{MDr}}{\|h_1 - h_2\|} .\]

Observe that \( F_{\lambda, \alpha} \) does not depend on \( U \). Hence, for each \( r \geq 0 \) we can first solve the problem
\[\min_{U \geq 0} G_{\lambda, \alpha} (W, U, r)\]
under the same constraints to obtain the optimal configuration of \( U = U(r) \), and then solve
\[\min_{r \geq 0} R_{\lambda, \eta, \alpha} (U(r), r) .\]

The problem of optimizing \( G_{\lambda, \alpha} (W, U, r) \) over \( U \) can be separated into two subproblems over \( u_1 \) and \( u_2 \), which are independent and symmetric. We hence consider only the problem over \( u_1 \), namely
\[\min_{u_1 \in R^d_+} \log \left( 1 + e^{(w_2 - w_1, u_1)} \right) - \frac{\alpha}{2} \langle w_2 - w_1, u_1 \rangle + \lambda \|u_1\|^2 \quad (P_{u_1})\]
s.t. \( \eta \|h_2 - u_1\| \leq \frac{C_{MDr}}{\|h_1 - h_2\|} .\)

Remark 2. Without its constraint, the minimization of \( G_{\lambda, \alpha} (W, U, r) \) over \( U \) becomes a restriction of the problem
\[\min_{W, H} \ell_{\alpha} (W, p(h_1), y_1^{(n)}) + \ell_{\alpha} (W, p(h_2), y_2^{(n)}) + \lambda_W \|W\|^2 + \lambda_H \|H\|^2\]
where \( W \) is restricted to be in the optimal NC configuration (see Definition 3.1). Hence the problem \( (P_{u_1}) \) without its constraint has the minimizer at \( u_1 = h_1 \). Furthermore the problem \( (P_{u_1}) \) itself also has its minimizer at \( h_1 \) if \( h_1 \) is feasible under the side constraint. Namely, when
\[\eta \|h_2 - h_1\| \leq \frac{C_{MDr}}{\|h_1 - h_2\|} ,\]
or equivalently when
\[ r \geq \frac{\eta \|h_1 - h_2\|}{C_{MD}} =: r_{\text{max}}. \]

Thus we only need to study the problem \( (P_{u_1}) \) in case \( r < r_{\text{max}} \).

The rest of the proof can be summarized as follows: First, in Subsection C.2 we show that the solution \( u_1(r) \) to \( (P_{u_1}) \) must be on a small subset of the feasible set (see Lemma C.1 and C.2), which allows us to prove the (almost) linear dependence of \( u_1(r) - u_1(r_{\text{max}}) \) on the distance \( r_{\text{max}} - r \) (see Lemma C.3). Next, we study the behavior of \( r \mapsto G_{\lambda,\alpha}(W, U(r), r) \) locally around \( r_{\text{max}} \) and the behavior of \( r \mapsto F_{\lambda,\alpha}(W, H, r) \) around 0. This lets us show that the decay of the former function near \( r_{\text{max}} \) dominates the increasing of the latter one, and hence the optimal dilation \( r_* \) must be close to \( r_{\text{max}} \). The details of this argument are introduced in Subsection C.5. Finally in Subsection C.6 we apply this to each value \( \alpha \in \{0, \alpha_0\} \) and get the desired statement of Theorem 4.3.

C.2 Estimating the solution of \( (P_{u_1}) \)

**Lemma C.1.** Let \( r < r_{\text{max}} \), then the minimizer of \( (P_{u_1}) \) lies on the circle of radius \( \frac{C_{MD}}{\eta \|h_1 - h_2\|} \) around \( h_2 \), i.e. the inequality constraint must equalize at the minimizer.

![Figure 8: Illustration for Lemma C.1 and C.2.](image)

The feasible set of \( (P_{u_1}) \) is the intersection of the positive quadrant and the disc with boundary given by the red circle \( C \). We consider the case \( C \) has an intersection \( u_1^* \) with the segment \((0, h_1) \) and \( u_1^{**} \) with the segment \((h_1, h_2) \). The minimizer of \( (P_{u_1}) \) must lie on the red arc between \( u_1^* \) and \( u_1^{**} \).

**Proof.** We denote by \( F \subset \mathbb{R}^2 \) the set of all positions of \( u_1 \) that is feasible to \( (P_{u_1}) \), or equivalently the intersection of the positive quadrant and the disc of radius \( \frac{C_{MD}}{\eta \|h_1 - h_2\|} \) around \( h_2 \). Furthermore, we denote the circle on the boundary of this disc by \( C \) and assume without loss of generality that \( h_1 \) lies on the \( x \)-axis and \( h_2 \) on the \( y \)-axis (see Figure 8).

Consider an arbitrary feasible point \( u_1 \) in the interior of \( F \), we show that \( u_1 \) is not the minimum of the problem \( (P_{u_1}) \). Since \( u_1 \) is an interior point, there exists a disc \( B \) around \( u_1 \) which lies completely inside \( F \). Let the arc \( A \) be the intersection of \( B \) and the circle of radius \( \|u_1\| \) around the origin. Observe that moving \( u_1 \) along the arc \( A \) keeps its norm unchanged, but can increase or decrease the value of \( \langle w_2 - w_1, u_1 \rangle \). Hence the objective in \( (P_{u_1}) \) cannot reach its minimum at \( u_1 \), unless \( \langle w_2 - w_1, u_1 \rangle \) is equal to the minimizer \( t_0 \) of the function
\[
    t \mapsto \log(1 + e^t) - \frac{\alpha}{2} t
\]

However, in case \( \langle w_2 - w_1, u_1 \rangle = t_0 \), the point \( u_1 \) lies on a line that is orthogonal to \( w_2 - w_1 \), and one can choose another point \( u_1' \) on the intersection of this line and the disc \( B \) such that \( \|u_1'\| < \|u_1\| \). In particular, \( u_1' \) is a better feasible candidate in comparison to \( u_1 \).

Excluding all interior points, we now consider the boundary set \( \partial F \) of the feasible set, which consists of points on the circle \( C \) that lie in the positive quadrant (denoted by \( \partial F_1 \)), points on the \( x \)-axis between 0 and the intersection point \( u_1^* \) of \( C \) with the \( x \)-axis (denoted by \( \partial F_2 \)), as well as points on the \( y \)-axis between 0 and the intersection point of \( C \) with the \( y \)-axis (denoted by \( \partial F_3 \)). Note that the circle \( C \) may have no intersection with the \( x \)-axis, in that case we simply consider \( \partial F_2 \) as the empty set.

Now we find the possible optimal positions of \( u_1 \) on each of the sets \( \partial F_3 \) and \( \partial F_2 \) (in case \( \partial F_2 \neq \emptyset \)).

25
1. On $\partial F_3$: Observe that moving a point $u_1$ along $\partial F_3$ in the direction toward the origin will decrease both the scalar product $\langle w_2 - w_1, u_1 \rangle > 0$ (as the angle is kept unchanged while the length of $u_1$ is decreased) and the regularization term $\|u_1\|^2$. Since the function $t \mapsto \log(1 + e^{\alpha t}) - \frac{\alpha}{2} t$ is monotonically increasing on $[0, \infty)$, moving $u_1$ in this direction decreases the objective in $f(P_{u_1})$. Therefore, the best candidate on $\partial F_3$ is the lowest possible point on $\partial F_3$, i.e. $0$ in case $C$ has intersection point with the $x$-axis, or is the lower intersection point of $C$ with the $y$-axis otherwise.

2. On $\partial F_2$ (in case it is not empty): Here, the objective becomes $f(||u_1||)$ where the function $f$ is defined by

$$f(t) = \log(1 + e^{\alpha t}) - \frac{\alpha}{2} t + \lambda t^2,$$

with $c_1 := \frac{\langle w_2 - w_1, h_1 \rangle}{\|h_1\|^2}$. Observe that $f$ is convex (this can be seen by directly computing the 2nd derivative of $f$) and achieves its minimum at $t = ||h_1||$ (because $h_1$ is the minimum of $f(P_{u_1})$ without the side constraint, see Remark 2). Hence on the interval $[0, ||h_1||]$ it is monotonically decreasing. It follows that $u_1^*$ is the best candidate on $\partial F_2$.

In summary we have shown that the optimal position of $u_1$ must be on $\partial F_1$, $\partial F_2$ or $\partial F_3$. On the other hand, all candidates on $\partial F_3$ are worse than a point in $\partial F_2$ and all candidates on $\partial F_2$ are worse than a point in $\partial F_1$ (in case $\partial F_2 = \emptyset$ we have that all candidates on $\partial F_3$ are worse than a point in $\partial F_1$). Therefore the minimizer must be a point on $\partial F_1$.

Having said that the optimal position of $u_1$ with respect to $f(P_{u_1})$ must be on the circle of radius $\frac{C_{M,D_r}}{\|h_1 - h_2\|}$, we are now interested in the case where $r$ is close to $r_{\text{max}}$, and hence the circle $\partial F_1$ has intersection with the segment $(0, h_1)$ (see Figure 8). In this case we can even restrict the possible optimal positions to a smaller subset of the circle.

**Lemma C.2.** Suppose that $r_{\text{max}} \geq r \geq r_{\text{max}}/\sqrt{2}$, so that the circle $C$ of radius $\frac{C_{M,D_r}}{\|h_1 - h_2\|}$ around $h_2$ has intersection $u_1^*$ with the line segment $(0, h_1)$ and intersection $u_1^{**}$ with the line segment $(h_2, h_1)$. Then, the minimizer of $f(P_{u_1})$ lies on the arc between $u_1^*$ and $u_1^{**}$.

**Proof.** First we rewrite the objective of $f(P_{u_1})$ as

$$f(\langle w_2 - w_1, u_1 \rangle) + \lambda \|u_1\|^2.$$

where $f : \mathbb{R} \to \mathbb{R}$ is the function defined by

$$f(t) = \log \left(1 + e^t\right) - \frac{\alpha}{2} t.$$

Next, we parameterize the circle $C$ by the polar coordinate. Let $R := \frac{C_{M,D_r}}{\|h_1 - h_2\|}$ and $\theta$ be the angle between $(h_2, u_1)$ and $(h_2, h_1)$. Then, since $w_2 - w_1$ is proportional to $h_2 - h_1$ we have

$$\langle w_2 - w_1, u_1 \rangle = \langle w_2 - w_1, h_2 \rangle + \langle w_2 - w_1, u_1 - h_2 \rangle = \langle w_2 - w_1, h_2 \rangle - R \|w_2 - w_1\| \cos \theta.$$ 

Note that $u_1$ can be on both sides of the line $(h_2, h_1)$, but for the calculation of $\langle w_2 - w_1, u_1 \rangle$ it is not necessary to distinguish between these two cases. In general, when $\theta$ increases, $\cos \theta$ decreases (we can exclude the case $\theta > \pi/2$ because in this case $\langle w_2 - w_1, u_1 \rangle$ becomes positive and the norm of $u_1$ is also large, so the objective becomes large and $u_1$ cannot be the minimizer), and thus $\langle w_2 - w_1, u_1 \rangle$ increases.

Now we claim that the optimal position of $u_1$ must be on the arc between $u_1^*$ and its reflection $u_1'$ about the line $(h_1, h_2)$. To show this we consider a point $u_1$ that lies on the other part of the circle $C$. By the above observation on the monotonicity of $\langle w_2 - w_1, u_1 \rangle$ with respect to $\theta$ we see that

$$\langle w_2 - w_1, u_1 \rangle > \langle w_2 - w_1, u_1' \rangle = \langle w_2 - w_1, u_1^* \rangle \geq \langle w_2 - w_1, h_1 \rangle.$$
Recall from the proof of Theorem 3.2 that $\langle w_2 - w_1, h_1 \rangle$ is not smaller than the minimizer $t_0$ of $f$, and due to convexity $f$ is monotone increasing on $[t_0, \infty)$. Therefore we obtain

$$f\left( \langle w_2 - w_1, u_1 \rangle \right) > f\left( \langle w_2 - w_1, u_1' \rangle \right)$$

On the other hand, by the law of cosines applied to the triangle $(0, h_2, u_1)$ we obtain

$$\|u_1\|^2 = \|h_2\|^2 + R^2 - 2R \|h_2\| \cos \left( \frac{\pi}{4} + \theta \right),$$

which is increasing in $\theta$ (again we exclude the case $\theta > \pi/2$ as discussed above). This shows that $\|u_1\| > \|u_1'\|$. Combining the above two inequalities we see that $u_1'$ is a better candidate than $u_1$.

Finally, the desired statement follows from the observation that we can exclude all points on the arc between $u_1^{**}$ and $u_1'$, because each point on this arc can be reflected about the line $(h_1, h_2)$ to a point with the same value of $\langle w_2 - w_1, u_1 \rangle$, but with smaller norm and this gives a better value of the objective.

Note that similar to the optimal position of $u$, the points $u_1^*, u_1^{**}$ from Lemma C.2 also depend on $r$. Hence to be clear, we may write $u_1 = u_1(r)$, $u_1^* = u_1^*(r)$ and $u_1^{**} = u_1^{**}(r)$ for $r \in \left[ r_{\text{max}}/\sqrt{2}, r_{\text{max}} \right]$. Observe that

$$u_1(r_{\text{max}}) = u_1^*(r_{\text{max}}) = u_1^{**}(r_{\text{max}}) = h_1.$$

The following lemma shows that for $r$ close to $r_{\text{max}}$, the distance between $u_1(r)$ and $h_1$ behaves almost linearly with respect to the distance $|r - r_{\text{max}}|$.

**Lemma C.3.** Let $u_1(r)$ be the minimizer of $\mathcal{P}_{u_1}$ with input $r \in \left[ r_{\text{max}}/\sqrt{2}, r_{\text{max}} \right]$. Then, there exists constants $c, C > 0$ such that

$$\|u_1(r) - h_1\| \in \left( c \frac{r_{\text{max}} - r}{\eta}, C \frac{r_{\text{max}} - r}{\eta} \right).$$

**Proof.** From Lemma C.2 we know that $u_1(r)$ lies on the arc between $u_1^*(r)$ and $u_1^{**}(r)$, hence its distance to $h_1$ is lower bounded by the distance from $u_1^{**}(r)$ to $h_1$ and is upper bounded by the
distance from $u^*_1(r)$ to $h_1$. Hence we have
\[
\|u_1(r) - h_1\| \leq \|u^*_1(r) - u^*_1(r_{\text{max}})\| \\
= \|u^*_1(r_{\text{max}})\| - \|u^*_1(r)\| \\
= \sqrt{\frac{C^2_{MD}}{\eta^2} \|h_1 - h_2\|^2 - \|h_2\|^2} - \sqrt{\frac{C^2_{MD}}{\eta^2} \|h_1 - h_2\|^2 - \|h_2\|^2} \\
= \sqrt{\frac{C^2_{MD}}{\eta^2} \|h_1 - h_2\|^2 - \|h_2\|^2} - \sqrt{\frac{C^2_{MD}}{\eta^2} \|h_1 - h_2\|^2 - \|h_2\|^2} \\
= \frac{C_{MD}(r_{\text{max}} - r)}{\eta \|h_1 - h_2\|}. \\
\]

On the other hand it also holds
\[
\|u_1(r) - h_1\| \geq \|u^*_1(r) - h_1\| \\
= \|h_2 - u^*_1(r_{\text{max}})\| - \|h_2 - u^*_1(r)\| \\
= \frac{C_{MD}r_{\text{max}}}{\eta \|h_1 - h_2\|} - \frac{C_{MD}r}{\eta \|h_1 - h_2\|} \\
= \frac{C_{MD}}{\eta \|h_1 - h_2\|} (r_{\text{max}} - r) \\
= \frac{C}{\eta \|h_1 - h_2\|}. \\
\]

Combining the above estimates yields the desired statement. □

C.3 The behavior of $G_{\lambda, \alpha}$ near $r_{\text{max}}$

We study the behavior of $G_{\lambda, \alpha}(W, U, r)$ as a function of $r$, where $W$ is fixed as in Assumption 4.2, $U = U(r)$ is the optimal position discussed in Subsection C.2 and $r$ lies near $r_{\text{max}}$.

Lemma C.4. Let $u_1(r)$ be the minimizer of $(P_\alpha)$ with input $r \in [r_{\text{max}}/\sqrt{\eta}, r_{\text{max}}]$. Then locally around $r_{\text{max}}$ the function $r \mapsto G_{\lambda, \alpha}(W, U(r), r)$ satisfies
\[
G_{\lambda, \alpha}(W, U(r), r) - G_{\lambda, \alpha}(W, U(r_{\text{max}}), r_{\text{max}}) \geq c \left(\frac{r - r_{\text{max}}}{\eta}\right)^2
\]
for some constant $c > 0$.

Proof. Due to symmetry, we only need to consider the half of $G_{\lambda, \alpha}$ that involves $u_1$, i.e. the function $g : \mathbb{R}^2 \to \mathbb{R}$, which is defined by
\[
g(u_1) = \log \left(1 + e^{\langle w_2 - w_1, u_1 \rangle}\right) - \frac{\alpha}{2} \langle w_2 - w_1, u_1 \rangle + \lambda \|u_1\|^2. \quad (27)
\]
We approximate \( g(u_1(r)) \) using the second-order Taylor approximation around \( h_1 = u_1(r_{\max}) \).

\[
g(u_1) = g(h_1) + \langle \nabla u_1 g(h_1), u_1 - h_1 \rangle + \frac{1}{2} \langle H_{u_1} g(h_1)(u_1 - h_1), u_1 - h_1 \rangle + O(\|u_1 - h_1\|^3),
\]

where the derivatives of \( g \) (at \( h_1 \)) are given by

\[
\nabla u_1 g(h_1) = \frac{e^{(w_2 - w_1, h_1)}}{1 + e^{(w_2 - w_1, h_1)}} (w_2 - w_1) - \frac{\alpha}{2} (w_2 - w_1) + 2\lambda h_1,
\]

\[
H_{u_1} g(h_1) = \frac{e^{(w_2 - w_1, h_1)}}{(1 + e^{(w_2 - w_1, h_1)})^2} (w_2 - w_1) (w_2 - w_1)^\top + 2\lambda I.
\]

Since \( u_1 = h_1 \) is the minimum of \( g(u_1) \) under the constraint \( u_1 \geq 0 \) and \( u_1(r) \) is always feasible for any \( r \), the linear term in (28) is non-negative, i.e.

\[
\langle \nabla u_1 g(h_1), u_1 - h_1 \rangle \geq 0.
\]

Next, we consider the second-order term in (28). We have

\[
\langle H_{u_1} g(h_1)(u_1 - h_1), u_1 - h_1 \rangle > \frac{e^{(w_2 - w_1, h_1)}}{(1 + e^{(w_2 - w_1, h_1)})^2} \|w_2 - w_1\|^2 \|u_1 - h_1\|^2,
\]

where the last inequality holds because the angle between \( w_2 - w_1 \) and \( u_1 - h_1 \) lies between 0 and \( \pi/4 \), which follows directly from Lemma C.2.

Inserting the above observations back into the Taylor expansion (28) and applying Lemma C.3 we obtain

\[
g(u_1(r)) - g(h_1) > c\|u_1(r) - h_1\|^2 \geq c \left( \frac{r - r_{\max}}{\eta} \right)^2
\]

for some constants \( c, c' > 0 \).

\[
\square
\]

**C.4 The behavior of \( F_{\lambda, \alpha} \)**

In this subsection we study the behavior of the function \( F_{\lambda, \alpha} \) under the assumptions in Assumption 4.2.

**Lemma C.5.** For \( r \ll 1 \), the function \( r \mapsto F_{\lambda, \alpha}(W, H, r) \) satisfies

\[
F_{\lambda, \alpha}(W, H, r) - F_{\lambda, \alpha}(W, H, 0) \leq cr^2
\]

for some constant \( c > 0 \).

**Proof.** By symmetry we only need to consider the half of \( F_{\lambda, \alpha} \) that involves \( h_1 \), and to simplify notations we denote this by \( \tilde{F} = \tilde{F}(r) \) with

\[
\tilde{F}(r) = \int \left( f_\lambda \left( W, p(h_1 + v), y_1^{(\alpha)} \right) + \lambda \|h_1 + v\|^2 \right) d\mu_r^\perp(v)
\]

\[
= \int \left( \log \left( 1 + e^{(w_2 - w_1, h_1 + v)} \right) - \frac{\alpha}{2} (w_2 - w_1, h_1 + v) + \lambda \|h_1 + v\|^2 \right) d\mu_r^\perp(v)
\]

\[
= \int \left( \log \left( 1 + e^{(w_2 - w_1, h_1 + v)} \right) + \lambda \|v\|^2 \right) d\mu_r^\perp(v) - \frac{\alpha}{2} \langle w_2 - w_1, h_1 \rangle + \lambda \|h_1\|^2,
\]

29
where the last equality comes from the second statement in Assumption 4.2. We denote the integrand in the above formulation by \( \tilde{f} \).

\[
\tilde{f}(v) = \log\left(1 + e^{(w_2-w_1,h_1+v)}\right) + \lambda \|v\|^2,
\]

Now we approximate \( \tilde{f} \) using its second-order Taylor expansion, which yields a rest of order \( O(\|v\|^3) \). From the second statement in Assumption 4.2 \( \|v\| \) is upper bounded by \( Ar \) and hence the rest of the Taylor approximation is of order \( O(r^3) \). Hence we obtain

\[
\hat{f}(v) = \tilde{f}(0) + \left\langle \nabla \tilde{f}(0), v \right\rangle + \frac{1}{2} \left\langle H \tilde{f}(0) v, v \right\rangle + O(r^3)
\]

\[
= \tilde{f}(0) + \frac{e^{(w_2-w_1,h_1)}}{1 + e^{(w_2-w_1,h_1)}} (w_2 - w_1, v)
\]

\[
+ \frac{1}{2} \frac{e^{(w_2-w_1,h_1)}}{1 + e^{(w_2-w_1,h_1)}} (w_2 - w_1, v)^2 + \lambda \|v\|^2 + O(r^3).
\]

Taking the integral \( \int d\mu_1^*(v) \) we see that again due to the second statement in Assumption 4.2, the first order term in the Taylor expansion of \( \hat{f} \) vanishes. Therefore we obtain

\[
\hat{F}(r) - \hat{F}(0) = \int \hat{f}(v) d\mu_1^*(v)
\]

\[
= \int \left( \frac{e^{(w_2-w_1,h_1)}}{2(1 + e^{(w_2-w_1,h_1)})^2} (w_2 - w_1, v)^2 + \lambda \|v\|^2 \right) d\mu_1^*(v) + O(r^3)
\]

\[
\leq \left( \frac{e^{(w_2-w_1,h_1)}}{2(1 + e^{(w_2-w_1,h_1)})^2} \|w_2 - w_1\|^2 + \lambda \right) \int \|v\|^2 d\mu_1^*(v) + O(r^3)
\]

\[
\leq A^2 \left( \frac{e^{(w_2-w_1,h_1)}}{2(1 + e^{(w_2-w_1,h_1)})^2} \|w_2 - w_1\|^2 + \lambda \right) r^2 + O(r^3),
\]

where \( A \) is the constant for which \( \|v\| \leq Ar \) holds (see Assumption 4.2). \( \square \)

### C.5 Estimation of the Optimal Dilation \( r_\star \)

In this subsection we come back to the \( MD \) problem, i.e. the minimization of the MD risk

\[
\min_{U,r} F_{\lambda,\alpha}(W, H, r) + \eta G_{\lambda,\alpha}(W; U, r) \quad \text{s.t.} \quad (4), (5).
\]

As discussed in Subsection A.1 we can simplify this problem by inserting into \( U \) the solution \( U(r) \) to the problem

\[
\min_U G_{\lambda,\alpha}(W; U, r) \quad \text{s.t.} \quad (4), (5).
\]

Then the MD problem is reduced to the minimization over the dilation \( r \), namely

\[
\min_r F_{\lambda,\alpha}(W, H, r) + \eta G_{\lambda,\alpha}(W; U(r), r).
\]

(29)

**Lemma C.6.** The solution \( r_\star \) to the problem (29) is upper bounded by \( r_{\max} \) and lies inside the ball of radius \( \theta^\prime (\eta^{1/2}) r_{\max} \) around \( r_{\max} \), i.e. \( r_{\max} \geq r_\star \geq r_{\max}(1 - C' \eta^{1/2}) \) where \( C' \) is an implied constant that depends however only on other absolute constants (in particular \( C' = O(1) \)).

**Proof.** To shorten notations, we define

\[
f(r) := F_{\lambda,\alpha}(W, H, r) \quad \text{and} \quad g(r) := G_{\lambda,\alpha}(W, U(r), r).
\]

30
Then the problem (29) can be rewritten as
\[
\min_r f(r) + \eta g(r).
\]

First we observe that \( r_* \leq r_{\text{max}} \) because for any \( r > r_{\text{max}} \) we have that \( f(r) > f(r_{\text{max}}) \) while \( g(r) = g(r_{\text{max}}) \) (see Remark 2). Thus we only need to show the lower bound on \( r_* \).

Let \( \epsilon \ll 1 \), we will show that the solution to the reduced \( MD \) problem (30) cannot be \( r \) for any \( r < (1 - \epsilon)r_{\text{max}} \), provided that \( \epsilon \) is sufficiently large (this condition will be later specified more precisely). To see this we will show that for any such \( r \) it holds
\[
f(r) + \eta g(r) > f(r_{\text{max}}) + \eta g(r_{\text{max}}).
\]

By Lemma C.5 we have that
\[
f(r_{\text{max}}) - f(r) \leq f(r_{\text{max}}) - f(0) \leq c_1 r_{\text{max}}^2
\]
for some constant \( c_1 > 0 \).

On the other hand, from Lemma C.4 it follows that
\[
g(r) - g(r_{\text{max}}) \geq g((1 - \epsilon)r_{\text{max}}) - g(r_{\text{max}}) \geq c_2 \frac{\epsilon^2 r_{\text{max}}^2}{\eta^2}
\]
holds for some constant \( c_2 > 0 \).

Combining the above observations we see that (31) will hold if
\[
c_2 \frac{\epsilon^2 r_{\text{max}}^2}{\eta^2} \geq c_1 r_{\text{max}}^2,
\]
or equivalently if \( \epsilon > C' \eta^{1/2} \) with where \( C' > 0 \) is some constant implied from \( c_1 \) and \( c_2 \).

Since any candidate outside the interval \( \left[r_{\text{max}}(1 - C' \eta^{1/2}), r_{\text{max}}\right] \) is worse than \( r_{\text{max}} \), we conclude that \( r_* \) must be in this interval.

\[\square\]

C.6 Finalizing the proof

In previous subsections we have approximately estimated the optimal dilation \( r_* \) of the \( MD \) problem in general, i.e. the LS parameter \( \alpha \) can take any value in \( \{0, \alpha_0\} \). Now we distinguish between the two values of \( \alpha \) by adding the superscripts \( LS \) (corresponding to \( \alpha = \alpha_0 \)) and \( CE \) (corresponding to \( \alpha = 0 \)), and we will finalize the proof of Theorem 4.3 by showing
\[
\frac{r_*^{CE}}{\|h_1^{CE} - h_2^{CE}\|} > \frac{r_*^{LS}}{\|h_1^{LS} - h_2^{LS}\|}.\tag{32}
\]

By Assumption 4.2 we have \( \|h_1^{CE} - h_2^{CE}\| = \gamma \|h_1^{LS} - h_2^{LS}\| \), hence
\[
\frac{r_{\text{max}}^{CE}}{\|h_1^{CE} - h_2^{CE}\|} = \eta \frac{\|h_1^{CE} - h_2^{CE}\|}{C_{MD}} = \eta \frac{\|h_1^{LS} - h_2^{LS}\|}{C_{MD}} = \eta \frac{r_{\text{max}}^{LS}}{\|h_1^{LS} - h_2^{LS}\|}.\]

Combining this and Lemma C.6 we obtain
\[
\frac{r_*^{CE}}{\|h_1^{CE} - h_2^{CE}\|} > \frac{(1 - C' \eta^{1/2}) r_{\text{max}}^{CE}}{\|h_1^{CE} - h_2^{CE}\|} = \gamma (1 - C' \eta^{1/2}) \frac{r_{\text{max}}^{LS}}{\|h_1^{LS} - h_2^{LS}\|} \]
\[
\geq \gamma (1 - C' \eta^{1/2}) \frac{r_*^{LS}}{\|h_1^{LS} - h_2^{LS}\|}.\]

31
Hence (32) holds provided that \( \gamma \left( 1 - C'\eta^{1/2} \right) \geq 1 \), which follows from the third statement in Assumption 4.2.