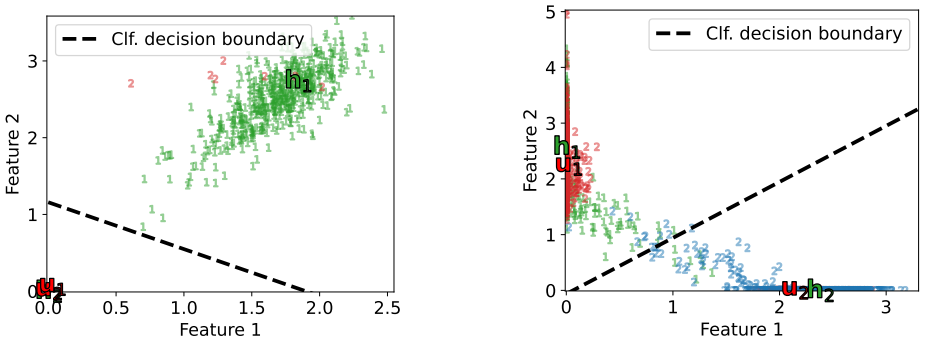
A EXPERIMENTAL DETAILS

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.



(a) (b)

Figure 4: 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 $M = N = 2$. 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 4 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 $M = 2$ 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 5, 6 and 7 show the results. Albeit not perfect, a similar dependence can be observed for multi-class settings.

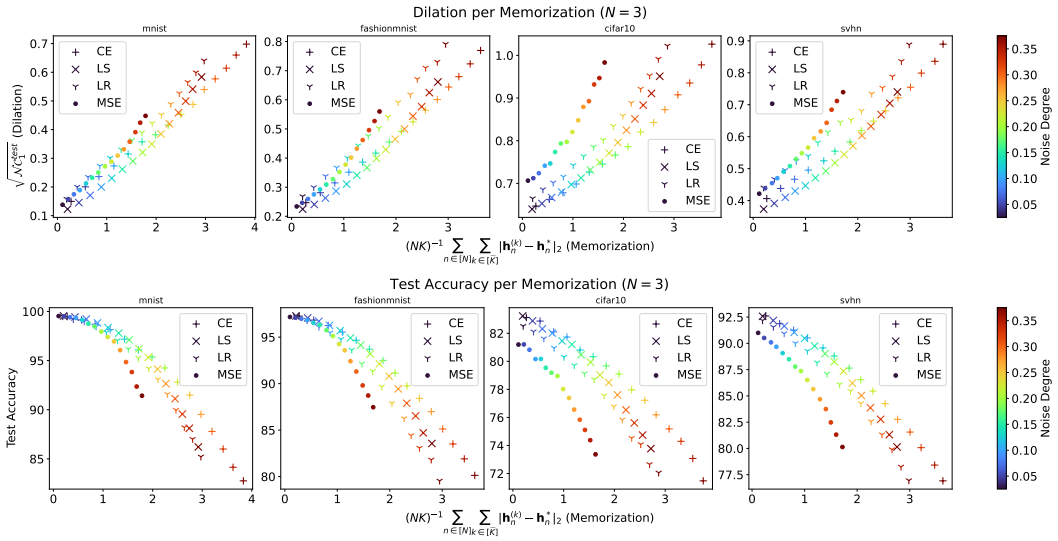


Figure 5: Feature collapse of the test instances in terms of $\sqrt{\mathcal{N}C_1^{\text{test}}}$ per memorization and the resulting test accuracies (averaged over ten seeds) for $N = 3$.

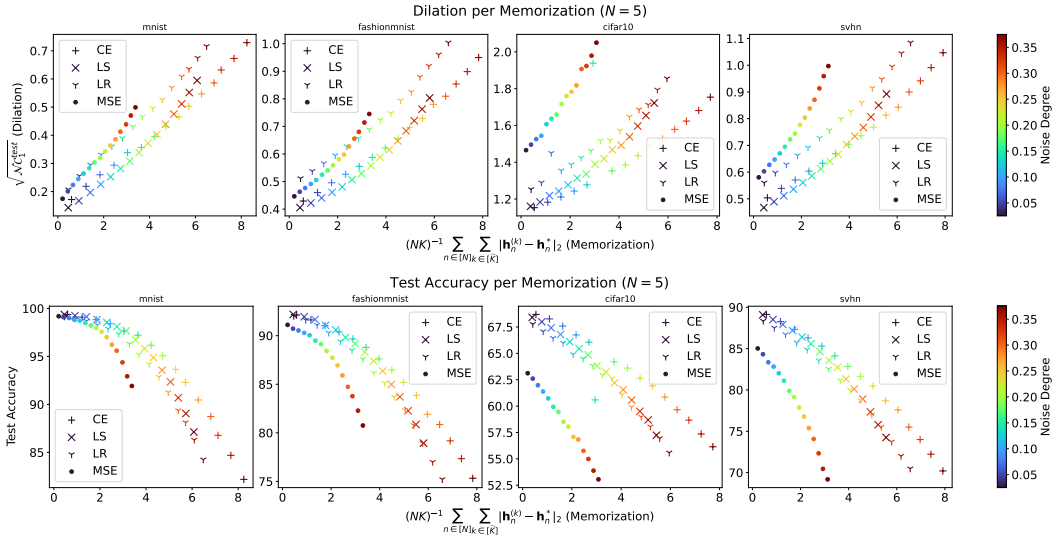


Figure 6: Feature collapse of the test instances in terms of $\sqrt{\mathcal{N}C_1^{\text{test}}}$ per memorization and the resulting test accuracies (averaged over ten seeds) for $N = 5$.

Additionally, we show results for $M > N$ to illustrate that the correspondence also holds for higher dimensions. Figure 8 shows the resulting dilation per memorization for $N \in \{5, 10\}$ on MNIST and CIFAR-10.

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

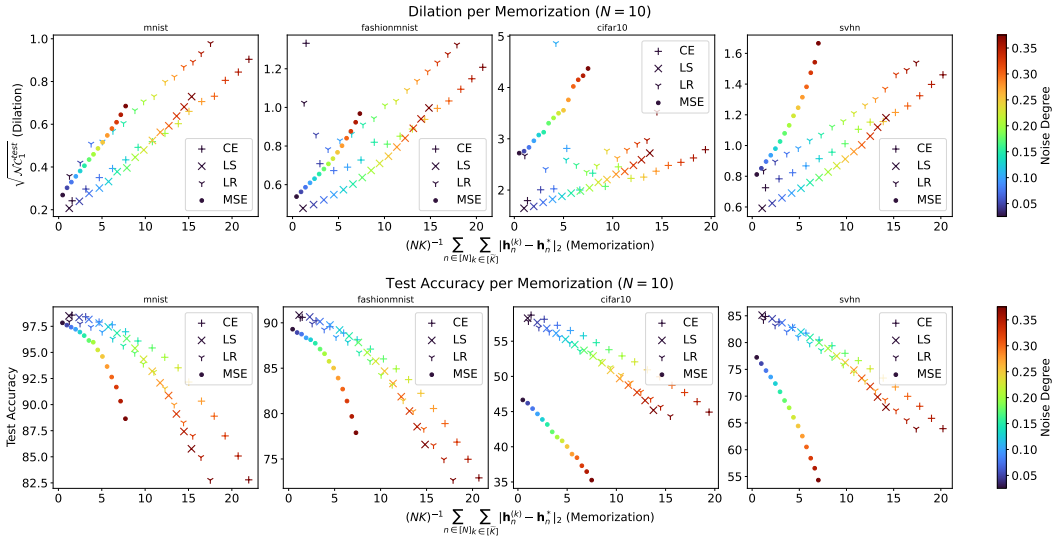


Figure 7: Feature collapse of the test instances in terms of $\sqrt{\mathcal{N}C_1^{\text{test}}}$ per memorization and the resulting test accuracies (averaged over ten seeds) for $N = 10$.

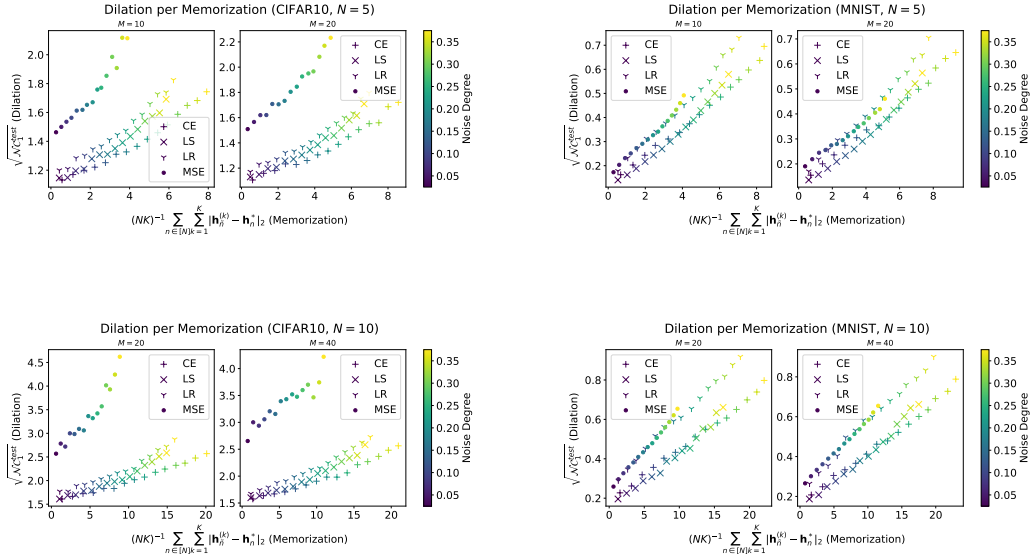


Figure 8: Feature collapse of the test instances in terms of $\sqrt{\mathcal{N}C_1^{\text{test}}}$ per memorization for higher feature dimensions M (averaged over five seeds) for $N \in \{5, 10\}$.

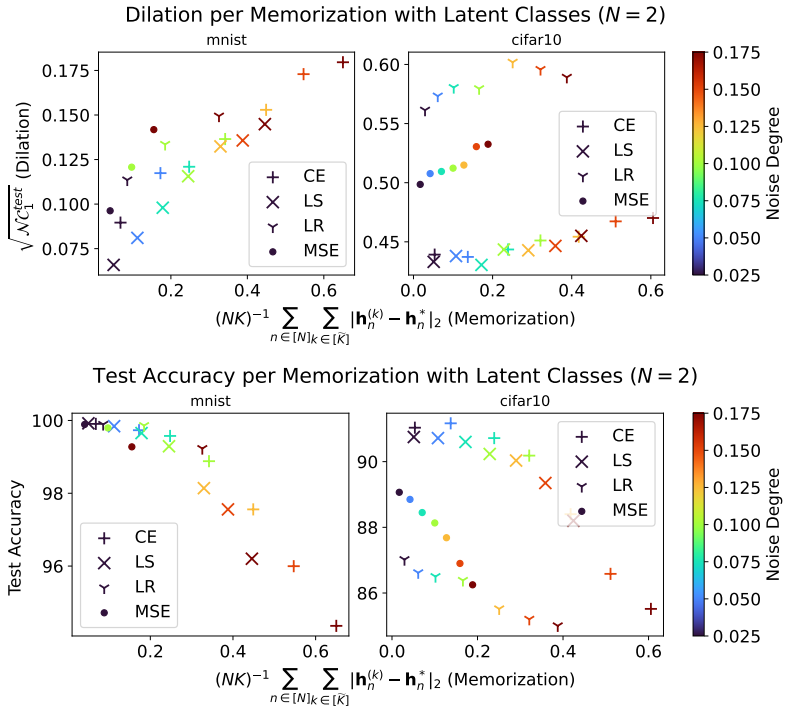


Figure 9: Dilation and test accuracy per memorization for label noise in form of latent noise classes, where we considered $N = 2$ original classes. The results are averaged over 5 random seeds.

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 9 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 & 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 α 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 $\eta > 0$.

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 `skopt`[1] 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).

[1] `https://scikit-optimize.github.io/`, BSD license

Parameter	Space
Initial learning rate	$[1e^{-5}, 0.5]$
Learning rate multiplier	$\{0.01, 0.1, 0.5\}$
Smoothing parameter α (LS)	$[0.01, 0.25]$
Relaxation parameter α (LR)	$[0.01, 0.25]$

Table 1: Search space considered within the hyperparameter optimization in the large-scale experiments.

		MNIST		FashionMNIST		CIFAR-10		SVHN	
η	Loss	ResNet18	VGG13	ResNet18	VGG13	ResNet18	VGG13	ResNet18	VGG13
0.0	CE	99.58 ± 0.05	99.22 ± 0.18	92.64 ± 0.20	91.87 ± 0.10	78.46 ± 0.44	80.40 ± 1.15	93.44 ± 0.05	94.55 ± 0.10
	LS	99.56 ± 0.01	99.34 ± 0.08	92.36 ± 0.13	91.94 ± 0.56	78.85 ± 0.29	81.33 ± 0.38	93.55 ± 0.21	92.50 ± 0.50
0.1	CE	98.32 ± 0.11	97.01 ± 0.28	89.35 ± 1.65	89.81 ± 0.30	70.86 ± 1.54	76.12 ± 1.58	88.89 ± 0.79	90.51 ± 0.24
	LS	98.51 ± 0.14	98.16 ± 0.03	89.62 ± 0.53	90.41 ± 0.52	71.82 ± 1.24	76.13 ± 1.74	89.68 ± 0.40	90.60 ± 0.40
0.2	CE	96.34 ± 0.20	92.72 ± 1.21	85.20 ± 2.67	85.39 ± 0.37	61.39 ± 1.30	70.09 ± 0.46	83.40 ± 2.19	86.03 ± 0.03
	LS	96.48 ± 0.13	95.00 ± 0.04	86.27 ± 1.24	87.54 ± 0.68	63.69 ± 3.06	70.97 ± 1.58	85.40 ± 0.49	86.81 ± 0.70
0.3	CE	91.82 ± 0.22	87.87 ± 0.64	79.91 ± 3.26	80.01 ± 0.65	52.52 ± 0.05	63.22 ± 1.26	77.10 ± 1.55	59.71 ± 28.37
	LS	92.05 ± 0.31	88.07 ± 0.42	81.18 ± 2.41	82.59 ± 1.25	55.31 ± 2.09	63.63 ± 2.88	79.62 ± 0.44	80.71 ± 1.56

Table 2: Generalization performances and their standard deviations in terms of test accuracies for different label noise degrees η (average over 3 seeds).

We repeated each run 3 times with different seeds and report the averaged results including their standard deviations.

A.2.2 Results

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 suggests 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)[2] and augmented it by further baselines, models and our evaluation metrics. This implementation leverages `PyTorch`[3] as deep learning framework and obtains data and models from `torchvision`[4]. 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 Theoretical supports for Theorem 3.2

In this appendix we introduce the theoretical supports for our theorem on the layer-peeled model, i.e. Theorem 3.2. Before going to the proof, we will shortly discuss in Subsection B.1 about how this finding differs from several related ones while still entailing the NC properties NC1-NC3. Then we will introduce some auxiliary results that are helpful for the proof in Subsections B.2 and B.3, and the proof itself in Subsection B.4.

Note that the concurrent work Zhou et al. (2022b) studies a very similar problem as ($\mathcal{P}_\alpha$), where the loss function L_α is replaced by a more general one that satisfies the so-called *contrastive property*. This covers both CE and LS loss, and hence the proof in this work is quite similar to ours. The main difference between the two models is namely the positivity constraint on the features, which leads to the main technical difference in the proof.

[2] `https://github.com/tding1/Neural-Collapse`

[3] `https://pytorch.org/`, BSD license

[4] `https://pytorch.org/vision/`, BSD license

Shortly speaking, the ultimate goal in the proof in Zhou et al. (2022b) (or several works of the same type, e.g. Zhu et al. (2021)) and ours is to show that the loss is lower bounded by some constant, and equality occurs only if the NC configuration is satisfied. However, if the minimization problem is unconstrained, one just needs to consider critical points, which might have a nice form. Although this is a clever trick, it cannot (at least not directly) be applied to the constrained problem, because the first-order optimality condition does not give as much information as in the unconstrained case (since it does not involve a simple equality form).

To overcome this, we will find a lower bound of the loss function in terms of $\|\boldsymbol{W}\|$ and $\|\boldsymbol{H}\|$, instead of the variables $\boldsymbol{W}$ and $\boldsymbol{H}$ themselves. This is possible in some certain region of the set of all feasible values of $\boldsymbol{W}$ and $\boldsymbol{H}$ (case (c) in Step 2 in our proof), while for other regions, we will show that either the minimizer does not belong to the region, or when it does, it must be also at NC configuration. Finally, the lower bound in terms of $\|\boldsymbol{W}\|$ and $\|\boldsymbol{H}\|$ is easier to deal with than the original one in terms of $\boldsymbol{W}$ and $\boldsymbol{H}$. In fact, we will see that despite the positivity of $\|\boldsymbol{W}\|$ and $\|\boldsymbol{H}\|$, one could consider the mentioned lower bound as a function of two variables that can take any value in $\mathbb{R}$.

B.1 Discussion about the result

The configurations defined in Definition 3.1 above differ from the ones specified in other works, e.g. Fang et al. (2021); Zhu et al. (2021); Zhou et al. (2022b) and describe more precisely the empirically observed NC phenomena. Indeed, it has been shown in those papers that the minimizers of $(\mathcal{P}_\alpha)$ without positivity constraint on the features must satisfy the following conditions

- (i′) The feature representations $\boldsymbol{h}_n^{(k)}$ within every class $n \in [N]$ are equal for all $k \in [K]$, and thus equal to their class mean $\boldsymbol{h}_n := \frac{1}{K}\sum_{k=1}^K \boldsymbol{h}_n^{(k)}$.
- (ii′) The class means $\{\boldsymbol{h}_n\}_{n=1}^N$ have equal norms and form an N-simplex equiangular tight frame (ETF) up to some rescaling.
- (iii′) The weight matrix $\boldsymbol{W}$ satisfies $\boldsymbol{w}_n = C\boldsymbol{h}_n$ for some constant $C > 0$.

Although the configuration defined by (i')-(iii') is different from the one defined in Def. 3.1, both entail the limit of the NC1-NC3 properties. Indeed, it is straightforward to observe the relation between NC1 and (i) or (i'). Moreover, the limit of NC2 is directly implied from (ii'), namely because the global mean $\boldsymbol{h} = \frac{1}{N}\sum_{n=1}^N \boldsymbol{h}_n$ must lie at the origin, i.e. $\boldsymbol{h} = 0$, while

$$\|\boldsymbol{h}_m\| = \|\boldsymbol{h}_n\| \quad \text{and} \quad \left\langle \frac{\boldsymbol{h}_m}{\|\boldsymbol{h}_m\|}, \frac{\boldsymbol{h}_n}{\|\boldsymbol{h}_n\|} \right\rangle = -\frac{1}{N-1} \quad \text{for any } m \neq n$$

hold as simple properties of a simplex ETF. The limit of NC3 follows also directly from the duality condition (iii') and the above observation that $\boldsymbol{h} = 0$. On the other hand, it is not as straightforward, but also not difficult to see the connection between the conditions (ii) and (iii) in Def. 3.1. Indeed, from (ii) it follows

$$\begin{aligned}
\|\boldsymbol{h}_m - \boldsymbol{h}\|^2 &= \|\boldsymbol{h}_m\|^2 - 2\langle \boldsymbol{h}_m, \boldsymbol{h}\rangle + \|\boldsymbol{h}\|^2 \\
&= \|\boldsymbol{h}_m\|^2 - \frac{2}{N}\|\boldsymbol{h}_m\|^2 + \frac{1}{N}\|\boldsymbol{h}_m\|^2 \\
&= \frac{N-1}{N}\|\boldsymbol{h}_m\|^2,
\end{aligned}$$

which combining with $\|\boldsymbol{h}_m\| = \|\boldsymbol{h}_n\|$ gives $\|\boldsymbol{h}_m - \boldsymbol{h}\| = \|\boldsymbol{h}_n - \boldsymbol{h}\|$ for any $m \neq n$. Also,

$$\begin{aligned}
\langle \boldsymbol{h}_m - \boldsymbol{h}, \boldsymbol{h}_n - \boldsymbol{h}\rangle &= -\langle \boldsymbol{h}, \boldsymbol{h}_n\rangle - \langle \boldsymbol{h}, \boldsymbol{h}_m\rangle + \|\boldsymbol{h}\|^2 \\
&= \frac{-2}{N}\|\boldsymbol{h}_n\|^2 + \frac{1}{N}\|\boldsymbol{h}_n\|^2 \\
&= \frac{-1}{N}\|\boldsymbol{h}_n\|^2,
\end{aligned}$$

which combining with the above finding shows

$$\left\langle \frac{\boldsymbol{h}_m - \boldsymbol{h}}{\|\boldsymbol{h}_n - \boldsymbol{h}\|_2}, \frac{\boldsymbol{h}_m - \boldsymbol{h}}{\|\boldsymbol{h}_n - \boldsymbol{h}\|_2} \right\rangle = -\frac{1}{N-1}.$$

Finally, the duality condition (iii) in Def. 3.1 involves the projection $P_{\boldsymbol{h}^\perp}\boldsymbol{h}_n$, which is the same as $\boldsymbol{h}_n - \boldsymbol{h}$. To see this, observe that $\boldsymbol{h}_n - (\boldsymbol{h}_n - \boldsymbol{h}) = \boldsymbol{h} \perp \boldsymbol{h}^\perp$ and $\langle \boldsymbol{h}_n - \boldsymbol{h}, \boldsymbol{h}\rangle = \frac{1}{N}\|\boldsymbol{h}_n\| - \frac{1}{N^2}\sum_{i=1}^N \|\boldsymbol{h}_i\|^2 = 0$.

Hence, the condition (iii) simply means that $\boldsymbol{w}_n$ is proportional to $\boldsymbol{h}_n - \boldsymbol{h}$, which corresponds with the limit of the property NC3.

The main difference between our definition of NC configuration (i.e. Def. 3.1) and the one described by the conditions (i')-(iii') above is that we require the *centralized* class means $\{\boldsymbol{h}_n - \boldsymbol{h}\}_{n=1}^N$ to form a simplex ETF, not the class means $\{\boldsymbol{h}_n\}_{n=1}^N$ themselves. Similarly, the duality in our definition involves the weights $\boldsymbol{w}_n$ and the *centralized* class means $\boldsymbol{h}_n - \boldsymbol{h}$, and not directly the class means or class features. Concerning the class means $\{\boldsymbol{h}_n\}_{n=1}^N$, we require them in (ii) to be an equinorm orthogonal system, which differs from a simplex ETF. In practice, if ReLU operation is applied, the features must be positive and hence the class means cannot form or approximate a simplex ETF, which must center at the origin.

In this sense, the NC configuration defined in Def. 3.1 can be seen as capturing more closely the NC phenomena in practice. Meanwhile, the usual configuration (i')-(iii') is obtained by a translation by the global mean $\boldsymbol{h}$ from our NC configuration. This translation may drop several interesting properties of the configuration, for example the property that the class means $\boldsymbol{h}_n$ are orthogonal and therefore (as they are positive) have separate supports (i.e. the indices of the nonzero entries in $\boldsymbol{h}_m$ and $\boldsymbol{h}_n$ do not overlap).

Notably, our NC configuration defined in Definition 3.1 and the *orthogonal frame* configuration in Theorem 3.1 in Tirer & Bruna (2022) appear to be similar, but have certain differences. Despite having an equivalent description for $\boldsymbol{H}$, our work considers positive features, which requires the feature vectors of different classes to have separate supports, i.e. their entries are supported on disjoint dimensions. Moreover, the weights $\boldsymbol{w}_n$ in our approach, as stated in (iii) in Def. 3.1, are not proportional to the class means $\boldsymbol{h}_n$ but to the centralized ones $\boldsymbol{h}_n - \boldsymbol{h}$, and hence form a simplex ETF and not an equinorm orthogonal system. Moreover, even in the presence of bias terms, the resulting configuration in our paper remains unchanged, which is different to Theorem 3.2 in Tirer & Bruna (2022). Note that all these differences come from the fact that we consider the CE or LS loss, while the authors of Tirer & Bruna (2022) study the MSE loss.

B.2 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,

$$
\begin{aligned}
L_\alpha(\boldsymbol{W}, \boldsymbol{H}) &= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \ell_\alpha\Big(\boldsymbol{W}, \boldsymbol{h}_n^{(k)}, \boldsymbol{y}_n^{(\alpha)}\Big) \\
&= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \Bigg[(1 - \frac{N-1}{N}\alpha) \log\Big(\sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) \\
&\quad + \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \log\Big(\sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} \rangle} \Big) \Bigg] \\
&= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \Bigg[(1 - \frac{N-1}{N}\alpha) \log\Big(\sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) \\
&\quad + \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \log\Big(e^{\langle \boldsymbol{w}_n - \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} \rangle} \sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) \Bigg] \\
&= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \Bigg[(1 - \frac{N-1}{N}\alpha) \log\Big(\sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) \\
&\quad + \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \log\Big(\sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) + \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \Big\langle \boldsymbol{w}_n - \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} \Big\rangle \Bigg]
\end{aligned}
$$

Hence,

$$
\begin{aligned}
L_\alpha(\boldsymbol{W}, \boldsymbol{H}) &= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \Bigg[(1 - \frac{N-1}{N}\alpha) \log\Big(\sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) \\
&\quad + \frac{N-1}{N}\alpha \log\Big(\sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) + \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \Big\langle \boldsymbol{w}_n - \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} \Big\rangle \Bigg] \\
&= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \Bigg[\log\Big(\sum_{i=1}^{N} e^{\langle \boldsymbol{w}_i - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) + \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \Big\langle \boldsymbol{w}_n - \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} \Big\rangle \Bigg] \\
&= \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \Bigg[\log\Big(1 + \sum_{\substack{m=1 \\ m \neq n}}^{N} e^{\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \rangle} \Big) - \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \Big\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \Big\rangle \Bigg].
\end{aligned}
$$

Shortly speaking, this differs from the conventional CE loss just by an additional bilinear term $\frac{\alpha}{N} \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m \neq n} \Big\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \Big\rangle$.

B.3 TECHNICAL LEMMATA

Lemma B.1. *We define*

$$
P(\boldsymbol{W}, \boldsymbol{H}) := \frac{1}{KN(N-1)} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{\substack{m=1 \\ m \neq n}}^{N} \Big\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \Big\rangle .
$$

Then under the condition $\boldsymbol{H} \geq 0$ it holds

$$
P(\boldsymbol{W}, \boldsymbol{H}) \geq -\frac{1}{\sqrt{KN(N-1)}} \|\boldsymbol{W}\| \, \|\boldsymbol{H}\| . \tag{7}
$$

The inequality (7) becomes an equality if and only if the following conditions hold simultaneously

$$\sum_{n=1}^{N} \boldsymbol{w}_n = 0 \tag{8}$$

$$\left\langle \boldsymbol{h}_n^{(k)}, \boldsymbol{h}_m^{(k)} \right\rangle = 0 \quad \textit{for all } m, n \in [N], k \in [K], m \neq n, \tag{9}$$

$$\left\| \boldsymbol{h}_n^{(k)} \right\| \quad \textit{is independent of } n, k, \tag{10}$$

$$\boldsymbol{w}_m - \boldsymbol{w}_n = c'(\boldsymbol{h}_m^{(k)} - \boldsymbol{h}_n^{(k)}) \quad \textit{for some } c' > 0 \textit{ not depending on } m, n, k. \tag{11}$$

Proof. Using the Cauchy-Schwarz inequality we get

$$\begin{aligned}
P(W,H) &:= \frac{1}{KN(N-1)} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{\substack{m=1 \\ m\neq n}}^{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle \\
&= \frac{1}{KN(N-1)} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{m=n+1}^{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} - \boldsymbol{h}_m^{(k)} \right\rangle \\
&\geq -\frac{1}{KN(N-1)} \sum_{k=1}^{K} \sqrt{\Big(\sum_{n=1}^{N} \sum_{m=n+1}^{N} \|\boldsymbol{w}_n - \boldsymbol{w}_m\|^2 \Big) \Big(\sum_{n=1}^{N} \sum_{m=n+1}^{N} \left\| \boldsymbol{h}_n^{(k)} - \boldsymbol{h}_m^{(k)} \right\|^2 \Big)} \\
&= -\frac{1}{KN(N-1)} \underbrace{\sqrt{\sum_{n=1}^{N} \sum_{m=n+1}^{N} \|\boldsymbol{w}_n - \boldsymbol{w}_m\|^2}}_{=:P_1} \underbrace{\sum_{k=1}^{K} \sqrt{\sum_{n=1}^{N} \sum_{m=n+1}^{N} \left\| \boldsymbol{h}_n^{(k)} - \boldsymbol{h}_m^{(k)} \right\|^2}}_{=:P_2}.
\end{aligned}$$

Further application of Cauchy-Schwarz inequality yields

$$\begin{aligned}
P_1 &= \sqrt{\sum_{n} \sum_{m=n+1}^{N} \|\boldsymbol{w}_n - \boldsymbol{w}_m\|^2} \\
&= \sqrt{N \sum_{n=1}^{N} \|\boldsymbol{w}_n\|^2 - \left\| \sum_{n=1}^{N} \boldsymbol{w}_n \right\|^2} \leq \sqrt{N}\, \|\boldsymbol{W}\|
\end{aligned}$$

and

$$\begin{aligned}
P_2 &= \sum_{k=1}^{K} \sqrt{\Big(\sum_{n=1}^{N} \sum_{m=n+1}^{N} \left\| \boldsymbol{h}_n^{(k)} - \boldsymbol{h}_m^{(k)} \right\|^2 \Big)} \\
&= \sum_{k=1}^{K} \sqrt{(N-1) \sum_{n=1}^{N} \left\| \boldsymbol{h}_n^{(k)} \right\|^2 - \sum_{n=1}^{N} \sum_{m=n+1}^{N} \left\langle \boldsymbol{h}_n^{(k)}, \boldsymbol{h}_m^{(k)} \right\rangle} \\
&\leq \sqrt{N-1} \sum_{k=1}^{K} \sum_{n=1}^{N} \left\| \boldsymbol{h}_n^{(k)} \right\| \\
&\leq \sqrt{KN(N-1)} \sqrt{\sum_{k=1}^{K} \sum_{n=1}^{N} \left\| \boldsymbol{h}_n^{(k)} \right\|^2} = \sqrt{KN(N-1)}\, \|\boldsymbol{H}\|
\end{aligned}$$

Therefore

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

This becomes an equality if and only if

- The upper bound on P_1 becomes equality, i.e.

$$\sum_{n=1}^{N} \boldsymbol{w}_n = 0$$

- The upper bound on P_2 becomes equality, i.e.

$$\left\langle \boldsymbol{h}_n^{(k)}, \boldsymbol{h}_m^{(k)} \right\rangle = 0 \quad \text{for all } m, n \in [N], k \in [K], m \neq n,$$

$$\left\| \boldsymbol{h}_n^{(k)} \right\| \quad \text{is independent of } n, k.$$

- The estimate $P \geq -\frac{1}{KN(N-1)} P_1 P_2$ becomes an equality, i.e.

$$\boldsymbol{w}_m - \boldsymbol{w}_n = c(\boldsymbol{h}_m^{(k)} - \boldsymbol{h}_n^{(k)}) \quad \text{for some } c' > 0 \text{ not depending on } m, n, k$$

□

Lemma B.2. *Assume that the inequality (7) 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*

$$\left\langle \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} \right\rangle = c_{n,k} \quad \textit{for every } m \in [N] \setminus \{n\}, \tag{12}$$

$$\sum_{\substack{m=1 \\ m \neq n}}^{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle = c \textit{ (not depending on } n, k), \tag{13}$$

for all $n \in [N]$ and $k \in [K]$. Then, the pair $(\boldsymbol{W}, \boldsymbol{H})$ must form a neural collapse configuration. Conversely, if $(\boldsymbol{W}, \boldsymbol{H})$ is a neural collapse configuration, then (7) 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

$$0 = \sum_{m=1}^{N} \left\langle \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} \right\rangle = \underbrace{\sum_{m=1}^{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle}_{=c} + N \left\langle \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle,$$

so

$$\left\langle \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle = \frac{-c}{N}. \tag{14}$$

Combining this with (13,12) gives

$$c = \sum_{\substack{m=1 \\ m \neq n}}^{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle = (N-1)c_{n,k} - (N-1)\frac{-c}{N},$$

and hence

$$\left\langle \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} \right\rangle = c_{n,k} = \frac{c}{N(N-1)}. \tag{15}$$

Combining (14,15) with (9,11) gives

$$\frac{-2c}{N-1} = \left\langle \boldsymbol{w}_n - \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} - \boldsymbol{h}_m^{(k)} \right\rangle = c' \left\| \boldsymbol{h}_n^{(k)} - \boldsymbol{h}_m^{(k)} \right\|^2 = c' \Big(\left\| \boldsymbol{h}_n^{(k)} \right\|^2 + \left\| \boldsymbol{h}_n^{(k)} \right\|^2 \Big), \tag{16}$$

Combining (16) with (10) shows that for every $n \in [N]$ and $k \in [K]$, it holds

$$\left\| \boldsymbol{h}_n^{(k)} \right\|^2 = \frac{-c}{c'(N-1)}. \tag{17}$$

On the other hand, it follows also from (14,15) that

$$\|\boldsymbol{w}_n\|^2 - \|\boldsymbol{w}_m\|^2 = \langle \boldsymbol{w}_n - \boldsymbol{w}_m, \boldsymbol{w}_n + \boldsymbol{w}_m \rangle = c' \left\langle \boldsymbol{h}_n^{(k)} - \boldsymbol{h}_m^{(k)}, \boldsymbol{w}_n + \boldsymbol{w}_m \right\rangle = 0, \tag{18}$$

and hence the vectors $\boldsymbol{w}_n$, $n \in [N]$ have the same length, which can be computed via

$$\begin{aligned} N^2 \|\boldsymbol{w}_i\|^2 &= N \sum_{n=1}^{N} \|\boldsymbol{w}_n\|^2 = \sum_{n>m} \|\boldsymbol{w}_n - \boldsymbol{w}_m\|^2 \\ &= c' \sum_{n>m} \left\langle \boldsymbol{w}_n - \boldsymbol{w}_m, \boldsymbol{h}_n^{(k)} - \boldsymbol{h}_m^{(k)} \right\rangle \\ &= c' \cdot \frac{N(N-1)}{2} \cdot \frac{-2c}{N-1} \\ &= -cc'N. \end{aligned}$$

Hence, for each $n \in [N]$, it holds

$$\|\boldsymbol{w}_n\|^2 = \frac{-cc'}{N}. \tag{19}$$

Now let $\boldsymbol{h}^{(k)} := \frac{1}{N} \sum_{m=1}^{N} \boldsymbol{h}_m^{(k)}$ for each $k \in [K]$. Observe that it holds

$$\begin{aligned} \left\langle \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} - \boldsymbol{h}^{(k)} \right\rangle &= \frac{N-1}{N} \left\langle \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle - \frac{1}{N} \sum_{\substack{m=1 \\ m \neq n}}^{N} \left\langle \boldsymbol{w}_n, \boldsymbol{h}_m^{(k)} \right\rangle \\ &= -\frac{(N-1)c}{N^2} - \frac{c}{N^2} \\ &= \frac{c}{N}. \end{aligned} \tag{20}$$

On the other hand, from (17) we have for each $n \in [N]$ and $k \in [K]$ that

$$\begin{aligned} \left\| \boldsymbol{h}_n^{(k)} - \boldsymbol{h}^{(k)} \right\|^2 &= \left\| \boldsymbol{h}_n^{(k)} \right\|^2 - 2 \left\langle \boldsymbol{h}_n^{(k)}, \frac{1}{N} \sum_m \boldsymbol{h}_m^{(k)} \right\rangle + \frac{1}{N^2} \left\| \sum_m \boldsymbol{h}_m \right\|^2 \\ &= \frac{N-1}{N} \left\| \boldsymbol{h}_n^{(k)} \right\|^2 \\ &= \frac{N-1}{N} \frac{-c}{c'(N-1)} \\ &= \frac{-c}{Nc'}. \end{aligned} \tag{21}$$

From (19-21) it follows that

$$\left\langle \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} - \boldsymbol{h}^{(k)} \right\rangle = \|\boldsymbol{w}_n\| \left\| \boldsymbol{h}_n^{(k)} - \boldsymbol{h}^{(k)} \right\|,$$

which implies that $\boldsymbol{w}_n$ is parallel to $\boldsymbol{h}_n^{(k)} - h^{(k)}$ for every $n \in [N]$ and $k \in [K]$. More precisely, by combining this finding with the above calculation of $\|\boldsymbol{w}_n\|$ and $\left\| \boldsymbol{h}_n^{(k)} - h^{(k)} \right\|$ in (19, 21) we obtain

$$\boldsymbol{w}_n = c' \Big(\boldsymbol{h}_n^{(k)} - \boldsymbol{h}^{(k)} \Big). \tag{22}$$

Finally it is left to show that $\boldsymbol{h}_n^{(k)} = \boldsymbol{h}_n^{(\ell)}$ for any $k, \ell \in [K]$. For this observe that $\boldsymbol{h}_n^{(k)} - h^{(k)} = \boldsymbol{h}_n^{(\ell)} - h^{(\ell)} = \boldsymbol{w}_n$ implies

$$\begin{aligned} \left\| \boldsymbol{h}_n^{(k)} \right\|^2 &= \left\| \boldsymbol{h}_n^{(\ell)} + \boldsymbol{h}^{(k)} - \boldsymbol{h}^{(\ell)} \right\|^2 \\ &= \left\| \boldsymbol{h}_n^{(\ell)} \right\|^2 + 2 \left\langle \boldsymbol{h}^{(k)} - \boldsymbol{h}^{(\ell)}, \boldsymbol{h}_n^{(k)} \right\rangle + \left\| \boldsymbol{h}^{(k)} - \boldsymbol{h}^{(\ell)} \right\|^2, \end{aligned}$$

and thus

$$2\left\langle \boldsymbol{h}^{(k)} - \boldsymbol{h}^{(\ell)}, \boldsymbol{h}_n^{(k)} \right\rangle + \left\| \boldsymbol{h}^{(k)} - \boldsymbol{h}^{(\ell)} \right\|^2 = 0.$$

Similarly

$$2\left\langle \boldsymbol{h}^{(\ell)} - \boldsymbol{h}^{(k)}, \boldsymbol{h}_n^{(\ell)} \right\rangle + \left\| \boldsymbol{h}^{(k)} - \boldsymbol{h}^{(\ell)} \right\|^2 = 0.$$

Combining the two equalities and taking the sum over n we obtain

$$\left\| \boldsymbol{h}^{(k)} - \boldsymbol{h}^{(\ell)} \right\|^2 = 0,$$

which means that $\boldsymbol{h}^{(k)} = \boldsymbol{h}^{(\ell)}$ and therefore $\boldsymbol{h}_n^{(k)} = \boldsymbol{h}_n^{(\ell)}$. □

B.4 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

$$\sum_{\substack{m=1 \\ m \neq n}}^{N} e^{\left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle} \geq (N-1) e^{\frac{1}{N-1} \sum_{\substack{m=1 \\ m \neq n}}^{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle},$$

with equality if and only if $\langle \boldsymbol{w}_m, \boldsymbol{h}_n \rangle = c_n$ for every $m \neq n$, independently of m, for some constant c_n. Inserting this into the formulation of L_α in Section B.2 we get

$$L_\alpha(\boldsymbol{W}, \boldsymbol{H}) \geq \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \left[\log\left(1 + (N-1) e^{\frac{1}{N-1} \sum_{\substack{m=1 \\ m \neq n}}^{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle}\right) - \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle \right].$$

Observe that the function $t \mapsto \log\left(1 + (N-1) e^{\frac{t}{N-1}}\right)$ is also convex, hence applying again Jensen's inequality we can lower bound the right-hand side in the estimate above, and obtain

$$\begin{aligned} L_\alpha(W, H) \geq \log &\left(1 + (N-1) e^{\frac{1}{KN(N-1)} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{\substack{m=1 \\ m \neq n}}^{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle}\right) \\ &- \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{\substack{m=1 \\ m \neq n}}^{N} \frac{\alpha}{N} \left\langle \boldsymbol{w}_m - \boldsymbol{w}_n, \boldsymbol{h}_n^{(k)} \right\rangle. \end{aligned} \tag{23}$$

Equality in (23) occurs if and only if the conditions (12, 13) (see Lemma B.2) hold simultaneously.

Step 2. Recall that with the notation $P = P(\boldsymbol{W}, \boldsymbol{H})$ from Lemma B.1, the inequality (23) becomes

$$\mathcal{L}_\alpha(\boldsymbol{W}, \boldsymbol{H}) \geq \log\left(1 + (N-1)e^P\right) - \beta P + \lambda_W \|\boldsymbol{W}\|^2 + \frac{\lambda_H}{K} \|\boldsymbol{H}\|^2 =: \tilde{L}(\boldsymbol{W}, \boldsymbol{H}), \tag{24}$$

with $\beta := \frac{N-1}{N}\alpha > 0$. Consider the function $g \colon \mathbb{R} \to \mathbb{R}$,

$$g(t) := \log\left(1 + (N-1)e^t\right) - \beta t.$$

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 [1]

$$t_0 := \log\Big(\frac{1}{N-1}\cdot\frac{\beta}{1-\beta}\Big) < 0$$

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 $\tilde{L}(\boldsymbol{W},\boldsymbol{H})$ of (24). We consider the following three cases, corresponding to three different regions of the feasible set of $(\boldsymbol{W},\boldsymbol{H})$:

(a) Case $t_0 > P(\boldsymbol{W},\boldsymbol{H})$: We will show that the minimizers of $\mathcal{L}_\alpha$ cannot be in this region. Toward a contradiction, assume that there is a minimizer $(\boldsymbol{W}_0,\boldsymbol{H}_0)$ s.t. $P(\boldsymbol{W}_0,\boldsymbol{H}_0) < t_0$. We construct $(\boldsymbol{W}_1,\boldsymbol{H}_1)$ to be a NC configuration (according to Definition 3.1) satisfying $\|\boldsymbol{W}_0\| = \|\boldsymbol{W}_1\|$ and $\|\boldsymbol{H}_0\| = \|\boldsymbol{H}_1\|$. Then we have

$$\begin{aligned}
P(\boldsymbol{W}_1,\boldsymbol{H}_1) &= -\frac{1}{\sqrt{K(N-1)}}\|\boldsymbol{W}_1\|\,\|\boldsymbol{H}_1\| \\
&= -\frac{1}{\sqrt{K(N-1)}}\|\boldsymbol{W}_0\|\,\|\boldsymbol{H}_0\| \\
&\le P(\boldsymbol{W}_0,\boldsymbol{H}_0) \\
&< t_0 < 0.
\end{aligned}$$

By rescaling $\boldsymbol{W}_1,\boldsymbol{H}_1$ (with a constant smaller than 1) we obtain a pair $(\boldsymbol{W},\boldsymbol{H})$ with $P(\boldsymbol{W},\boldsymbol{H}) = t_0$ and $\|\boldsymbol{W}\| < \|\boldsymbol{W}_0\|$, $\|\boldsymbol{H}\| < \|\boldsymbol{H}_0\|$. Thus it holds

$$\begin{aligned}
\mathcal{L}_\alpha(\boldsymbol{W}_0,\boldsymbol{H}_0) &\ge \log\Big(1+(N-1)e^{P(\boldsymbol{W}_0,\boldsymbol{H}_0)}\Big) - \beta P(\boldsymbol{W}_0,\boldsymbol{H}_0) \\
&\quad + \lambda_W\|\boldsymbol{W}_0\|^2 + \lambda_H\|\boldsymbol{H}_0\|^2 \\
&> \log\Big(1+(N-1)e^{t_0}\Big) - \beta t_0 + \lambda_W\|\boldsymbol{W}\|^2 + \frac{\lambda_H}{K}\|\boldsymbol{H}\|^2 \\
&= \mathcal{L}_\alpha(\boldsymbol{W},\boldsymbol{H}),
\end{aligned}$$

which means that $(\boldsymbol{W}_0,\boldsymbol{H}_0)$ cannot be a minimizer of $\mathcal{L}_\alpha$. Note that the last equality holds because the inequality (23) equalizes when $(\boldsymbol{W},\boldsymbol{H})$ is a NC configuration (see Lemma B.2).

(b) Case $P(\boldsymbol{W},\boldsymbol{H}) \ge t_0 \ge -\frac{1}{\sqrt{K(N-1)}}\|\boldsymbol{W}\|\,\|\boldsymbol{H}\|$: We will show that at the minimizers in this region, P must be t_0. Assume that $(\boldsymbol{W}_0,\boldsymbol{H}_0)$ is a minimizer of $\tilde{L}$ in this region with $P(\boldsymbol{W}_0,\boldsymbol{H}_0) \ne t_0$. Then we consider all pairs $(\boldsymbol{W},\boldsymbol{H})$ with $\|\boldsymbol{W}\| \le \|\boldsymbol{W}_0\|$ and $\|\boldsymbol{H}\| \le \|\boldsymbol{H}_0\|$. By continuity we have that $P(\boldsymbol{W},\boldsymbol{H})$ can take all values in the interval

$$\Big[-\frac{1}{\sqrt{K(N-1)}}\|\boldsymbol{W}_0\|\,\|\boldsymbol{H}_0\|\,,\,\frac{1}{\sqrt{K(N-1)}}\|\boldsymbol{W}_0\|\,\|\boldsymbol{H}_0\|\Big],$$

which also includes t_0. It follows that $\tilde{L}(\boldsymbol{W},\boldsymbol{H}) < \tilde{L}(\boldsymbol{W}_0,\boldsymbol{H}_0)$, so $(\boldsymbol{W}_0,\boldsymbol{H}_0)$ cannot be a minimizer of $\tilde{L}$, meaning that a minimizer $(\boldsymbol{W},\boldsymbol{H})$ of $\tilde{L}$ must satisfy $P(\boldsymbol{W},\boldsymbol{H}) = t_0$. The minimization of $\tilde{L}$ then reduces to

$$\min_{W,H}\quad \lambda_W\|W\|^2 + \lambda_H\|H\|^2 \quad \text{s.t.} \quad -\frac{1}{\sqrt{K(N-1)}}\|W\|\,\|H\| = t_0.$$

[1] 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).

Observe that

$$\lambda_W \|\boldsymbol{W}\|^2 + \frac{\lambda_H}{K}\|\boldsymbol{H}\|^2 \geq 2\sqrt{\frac{\lambda_W\lambda_H}{K}}\|\boldsymbol{W}\|\|\boldsymbol{H}\| \geq -2t_0\sqrt{(N-1)\lambda_W\lambda_H}.$$

Therefore we have $\tilde{L}(W,H) \geq g(t_0) - 2t_0\sqrt{(N-1)\lambda_W\lambda_H}$ and this equalizes if and only if the following conditions hold:

- $P(\boldsymbol{W},\boldsymbol{H}) = t_0$
- $\lambda_W\|\boldsymbol{W}\|^2 = \lambda_H\|\boldsymbol{H}\|^2$ and $\|\boldsymbol{W}\|\|\boldsymbol{H}\| = -\sqrt{K(N-1)}t_0$.

(c) Case $P(\boldsymbol{W},\boldsymbol{H}) \geq -\frac{1}{\sqrt{K(N-1)}}\|\boldsymbol{W}\|\|\boldsymbol{H}\| \geq t_0$:

In this region, it holds $g\Big(P(\boldsymbol{W},\boldsymbol{H})\Big) \geq g\Big(-\frac{1}{\sqrt{K(N-1)}}\|\boldsymbol{W}\|\|\boldsymbol{H}\|\Big)$, so

$$\tilde{L}(\boldsymbol{W},\boldsymbol{H}) \geq f(\|\boldsymbol{W}\|,\|\boldsymbol{H}\|),$$

with $f\colon \mathbb{R}^2 \to \mathbb{R}$,

$$f(w,h) := \log\Big(1+(N-1)e^{-Cwh}\Big) + \beta Cwh + \lambda_W w^2 + \frac{\lambda_H}{K}h^2$$

where we set $C := \frac{1}{\sqrt{K(N-1)}}$ to shorten notation. Observe that even though w and h, as representatives for $\|\boldsymbol{W}\|$ and $\|\boldsymbol{H}\|$ 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 $\lambda_W w^2 = \frac{\lambda_H}{K}h^2$, 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

$$-\frac{(N-1)e^{-C'a^2}}{1+(N-1)e^{-C'w^2}}C'w + 2\beta C'w + 2\lambda_W w = 0.$$

Excluding the trivial solution $(w,h) = (0,0)$, so that we can multiply both sides with $1/w$, we get

$$w^2 = \frac{1}{C}\sqrt{\frac{\lambda_H}{K\lambda_W}}\log\left((N-1)\frac{1-\beta-2\sqrt{(N-1)\lambda_W\lambda_H}}{\beta+2\sqrt{(N-1)\lambda_W\lambda_H}}\right) \tag{25}$$

and

$$h^2 = \frac{1}{C}\sqrt{\frac{K\lambda_W}{\lambda_H}}\log\left((N-1)\frac{1-\beta-2\sqrt{(N-1)\lambda_W\lambda_H}}{\beta+2\sqrt{(N-1)\lambda_W\lambda_H}}\right) \tag{26}$$

Finally, it is easy to check that

$$-Cwh = \log\Big(\frac{1}{N-1}\frac{\beta+2\sqrt{(N-1)\lambda_W\lambda_H}}{1-\beta-2\sqrt{(N-1)\lambda_W\lambda_H}}\Big) > \log\Big(\frac{1}{N-1}\cdot\frac{\beta}{1-\beta}\Big) = t_0,$$

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 $\tilde{L}(\boldsymbol{W},\boldsymbol{H}) \geq f(w_0,h_0)$ with (w_0,h_0) specified as in (25, 26), and this becomes equality if and only if $P(\boldsymbol{W},\boldsymbol{H}) = -\frac{1}{\sqrt{K(N-1)}}\|\boldsymbol{W}\|\|\boldsymbol{H}\|$ and $\|\boldsymbol{W}\| = w_0$, $\|\boldsymbol{H}\| = h_0$.

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

□

C THEORETICAL SUPPORTS FOR 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

$$\min_{U\geq 0, r\geq 0} \quad \mathcal{R}_{\lambda,\eta,\alpha}(\boldsymbol{U}, r) := F_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{H}, r) + \eta G_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{U}, r)$$

under the constraints

$$\eta \left\|\boldsymbol{h}_1 - \boldsymbol{u}_2\right\| \leq \frac{C_{MD} r}{\left\|\boldsymbol{h}_1 - \boldsymbol{h}_2\right\|},$$
$$\eta \left\|\boldsymbol{h}_2 - \boldsymbol{u}_1\right\| \leq \frac{C_{MD} r}{\left\|\boldsymbol{h}_1 - \boldsymbol{h}_2\right\|}.$$

Observe that $F_{\lambda,\alpha}$ does not depend on $\boldsymbol{U}$. Hence, for each $r \geq 0$ we can first solve the problem

$$\min_{\boldsymbol{U}\geq 0} \quad G_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{U}, r)$$

under the same constraints to obtain the optimal configuration of $\boldsymbol{U} = \boldsymbol{U}(r)$, and then solve

$$\min_{r\geq 0} \quad \mathcal{R}_{\lambda,\eta,\alpha}\Big(\boldsymbol{U}(r), r\Big).$$

The problem of optimizing $G_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{U}, r)$ over $\boldsymbol{U}$ can be separated into two subproblems over $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, which are independent and symmetric. We hence consider only the problem over $\boldsymbol{u}_1$, namely

$$\begin{aligned} \min_{\boldsymbol{u}_1 \in \mathbb{R}_+^M} \quad & \log\left(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1\rangle}\right) - \frac{\alpha}{2}\left\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1\right\rangle + \lambda \left\|\boldsymbol{u}_1\right\|^2 \\ \text{s.t.} \quad & \eta \left\|\boldsymbol{h}_2 - \boldsymbol{u}_1\right\| \leq \frac{C_{MD} r}{\left\|\boldsymbol{h}_1 - \boldsymbol{h}_2\right\|}. \end{aligned} \tag{$\mathcal{P}_{\boldsymbol{u}_1}$}$$

Remark 1. Without its constraint, the minimization of $G_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{U}, r)$ over $\boldsymbol{U}$ becomes a reduction of the problem

$$\min_{\boldsymbol{W},\boldsymbol{H}} \quad \ell_\alpha(\boldsymbol{W}, \boldsymbol{h}_1, y_1^{(\alpha)}) + \ell_\alpha(\boldsymbol{W}, \boldsymbol{h}_2, y_2^{(\alpha)}) + \lambda_W \left\|\boldsymbol{W}\right\|^2 + \lambda_H \left\|\boldsymbol{H}\right\|^2$$

where $\boldsymbol{W}$ is restricted to be in the optimal NC configuration (see Definition 3.1). Hence the problem $(\mathcal{P}_{\boldsymbol{u}_1})$ without its constraint has the minimizer at $\boldsymbol{u}_1 = \boldsymbol{h}_1$. Furthermore the problem $(\mathcal{P}_{\boldsymbol{u}_1})$ itself also has its minimizer at $\boldsymbol{h}_1$ if $\boldsymbol{h}_1$ is feasible under the side constraint. Namely, when

$$\eta \left\|\boldsymbol{h}_2 - \boldsymbol{h}_1\right\| \leq \frac{C_{MD} r}{\left\|\boldsymbol{h}_1 - \boldsymbol{h}_2\right\|},$$

or equivalently when

$$r \geq \frac{\eta \left\|\boldsymbol{h}_1 - \boldsymbol{h}_2\right\|^2}{C_{MD}} =: r_{\max}.$$

Thus we only need to study the problem $(\mathcal{P}_{\boldsymbol{u}_1})$ in case $r < r_{\max}$.

The rest of the proof can be summarized as follows: First, in Subsection C.2 we show that the solution $u_1(r)$ to $(\mathcal{P}_{\boldsymbol{u}_1})$ 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 $u_1(r) - u_1(r_{\max})$ on the distance $r_{\max} - r$ (see Lemma C.3). Next, we study the behavior of $r \mapsto G_{\lambda,\alpha}\Big(\boldsymbol{W}, \boldsymbol{U}(r), r\Big)$ locally around $r_{\max}$ and the behavior of $r \mapsto F_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{H}, r)$ around 0. This lets us show that the decay of the former function near $r_{\max}$ dominates the increasing of the latter one, and hence the optimal dilation r_* must be close to $r_{\max}$. 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 $(\mathcal{P}_{\boldsymbol{u}_1})$

For convenience, in this section we introduce several notations. Let

$$\mathcal{S} := \operatorname{span}\{\boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\} = \operatorname{span}\{\boldsymbol{h}_2 - \boldsymbol{h}_1, \boldsymbol{h}_1\} = \operatorname{span}\{\boldsymbol{h}_2, \boldsymbol{h}_1\}$$

be the two-dimensional subspace spanned by $\boldsymbol{w}_2 - \boldsymbol{w}_1$ and $\boldsymbol{h}_1$. Furthermore, let $\mathcal{B}$ be the ball of radius $\frac{C_{MD}r}{\eta\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|}$ around $\boldsymbol{h}_2$. Let $\mathcal{C} = \partial\mathcal{B} \cap \mathcal{S}$ be the circle that is the intersection of the ball $\mathcal{B}$ and the subspace $\mathcal{S}$. We will show that the minimizer of $(\mathcal{P}_{\boldsymbol{u}_1})$ must lie on the circle $\mathcal{C}$. Note that the feasiblity of a vector $x \in \mathbb{R}^M$ for the problem $(\mathcal{P}_{\boldsymbol{u}_1})$ can be expressed as $x \in \mathbb{R}^M_+ \cap \mathcal{B}$.

Lemma C.1. *Let $r < r_{\max}$, then the minimizer of $(\mathcal{P}_{\boldsymbol{u}_1})$ lies on the circle $\mathcal{C}$, i.e. it lies on the subspace $\mathcal{S}$ and the inequality constraint in $(\mathcal{P}_{\boldsymbol{u}_1})$ must equalize at the minimizer.*

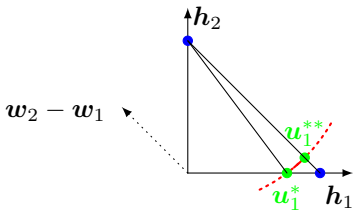


Figure 10: Illustration for Lemma C.1 and C.2. The feasible set of $(\mathcal{P}_{\boldsymbol{u}_1})$ is the intersection of the positive quadrant and the disc with boundary given by the red circle $\mathcal{C}$. We consider the case $\mathcal{C}$ has an intersection $\boldsymbol{u}_1^*$ with the segment $(0, \boldsymbol{h}_1)$ and $\boldsymbol{u}_1^{**}$ with the segment $(\boldsymbol{h}_1, \boldsymbol{h}_2)$. The minimizer of $(\mathcal{P}_{\boldsymbol{u}_1})$ must lie on the red arc between $\boldsymbol{u}_1^*$ and $\boldsymbol{u}_1^{**}$.

Proof.

1. Let $x \in \mathbb{R}^M_+ \cap \mathcal{B}$ be a feasible solution. According to $\mathbb{R}^M = \mathcal{S} \oplus \mathcal{S}^\perp$ we can decompose x into

$$x = \boldsymbol{x}_{\mathcal{S}} + x - \boldsymbol{x}_{\mathcal{S}},$$

where $\boldsymbol{x}_{\mathcal{S}}$ is the orthogonal projection of x on the subspace $\mathcal{S}$ and $x - \boldsymbol{x}_{\mathcal{S}}$ is orthogonal to $\mathcal{S}$. We will show that $\boldsymbol{x}_{\mathcal{S}}$ is a better candidate for $(\mathcal{P}_{\boldsymbol{u}_1})$ than x, which means that $\boldsymbol{x}_{\mathcal{S}}$ is also feasible and leads to smaller objective value. The second point is straightforward, because one can observe that

$$\langle x, w_2 - w_1 \rangle = \langle \boldsymbol{x}_{\mathcal{S}}, w_2 - w_1 \rangle,$$

and

$$\|x\|^2 = \|\boldsymbol{x}_{\mathcal{S}}\|^2 + \|x - \boldsymbol{x}_{\mathcal{S}}\|^2 \geq \|\boldsymbol{x}_{\mathcal{S}}\|^2.$$

Thus it is left to show the feasibility of $\boldsymbol{x}_{\mathcal{S}}$. For this, observe that $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ form an (entrywise) nonnegative orthogonal basis of $\mathcal{S}$ (remark: not necessarily an orthonormal basis, because $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are not necessarily normalized). Thus $\boldsymbol{x}_{\mathcal{S}}$ can be written as

$$\boldsymbol{x}_{\mathcal{S}} = a_1\boldsymbol{h}_1 + a_2\boldsymbol{h}_2$$

in which the coefficients $a_1, a_2 \in \mathbb{R}$ satisfy

$$a_i = \left\langle \boldsymbol{x}_{\mathcal{S}}, \frac{\boldsymbol{h}_i}{\|\boldsymbol{h}_i\|^2} \right\rangle = \left\langle x, \frac{\boldsymbol{h}_i}{\|\boldsymbol{h}_i\|^2} \right\rangle \geq 0,$$

where the last inequality holds because both vectors $\boldsymbol{x}$ and $\boldsymbol{h}_i$ are nonnegative.

Finally we need to show that $\boldsymbol{x}_{\mathcal{S}} \in \mathcal{B}$. For this consider again the decomposition

$$\boldsymbol{x} = \boldsymbol{x}_{\mathcal{S}} + \boldsymbol{x} - \boldsymbol{x}_{\mathcal{S}} = a_1\boldsymbol{h}_1 + a_2\boldsymbol{h}_2 + (\boldsymbol{x} - \boldsymbol{x}_S).$$

We obtain

$$\begin{aligned}
\|\boldsymbol{h}_2 - \boldsymbol{x}\|^2 &= \|a_1\boldsymbol{h}_1 + (a_2 - 1)\boldsymbol{h}_2 + \boldsymbol{x} - \boldsymbol{x}_{\mathcal{S}}\|^2 \\
&= a_1^2 \|\boldsymbol{h}_1\|^2 + (a_2 - 1)^2 \|\boldsymbol{h}_2\|^2 + \|\boldsymbol{x} - \boldsymbol{x}_{\mathcal{S}}\|^2 \\
&\geq a_1^2 \|\boldsymbol{h}_1\|^2 + (a_2 - 1)^2 \|\boldsymbol{h}_2\|^2 \\
&= \|a_1\boldsymbol{h}_1 + (a_2 - 1)\boldsymbol{h}_2\|^2 \\
&= \|\boldsymbol{h}_2 - \boldsymbol{x}_{\mathcal{S}}\|^2,
\end{aligned}$$

and hence the claim $\boldsymbol{x}_{\mathcal{S}} \in \mathcal{B}$ follows from the feasibility of $\boldsymbol{x}$.

2. Now we reduce to the subspace $\mathcal{S}$. Since $\boldsymbol{h}_1/\|\boldsymbol{h}_1\|$ and $\boldsymbol{h}_2/\|\boldsymbol{h}_1\|$ form a nonnegative orthonormal basis for this subspace, it is equivalent to consider the space $\mathbb{R}^2$ of the coefficients. Note that the positivity of a vector $\boldsymbol{x} \in \mathcal{S} \subseteq \mathbb{R}^M$ is also equivalent to the positivity of its coefficients in $\mathbb{R}^2$. Thus for convenience we may assume without loss of generality that $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are orthogonal vectors in $\mathbb{R}^2$, more precisely $\boldsymbol{h}_1$ lies on the x-axis and $\boldsymbol{h}_2$ lies on the y-axis as in Figure 10.

   We denote by $\mathcal{F} \subset \mathbb{R}^2_+$ the set of all points in $\mathcal{S}$ that is feasible to $(\mathcal{P}_{\boldsymbol{u}_1})$, i.e. inside the ball $\mathcal{B}$ around $\boldsymbol{h}_2$. Note that on the two-dimensional subspace $\mathcal{S}$ the ball $\mathcal{B}$ reduces to a disc, whose boundary is the circle $\mathcal{C}$ as defined above. We will show that the solution to $(\mathcal{P}_{\boldsymbol{u}_1})$ must lie on the boundary of $\mathcal{F}$ (w.r.t. the topology in $\mathcal{S} \cong \mathbb{R}^2$), i.e either on the axes (due to positivity constraints) or on the circle $\mathcal{C}$.

   Indeed, let $\boldsymbol{u}_1$ be an arbitrary feasible point in the interior of $\mathcal{F}$, we prove that $\boldsymbol{u}_1$ is not the minimum of the problem $(\mathcal{P}_{\boldsymbol{u}_1})$. Since $\boldsymbol{u}_1$ is an interior point, there exists a disc $\mathcal{B}'$ around $\boldsymbol{u}_1$ which lies completely inside $\mathcal{F}$. Let $\mathcal{A}$ be the intersection of the disc $\mathcal{B}'$ and the circle of radius $\|\boldsymbol{u}_1\|$ around the origin. Observe that moving $\boldsymbol{u}_1$ along the arc $\mathcal{A}$ keeps its norm unchanged, but can both increase and decrease the value of $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle$. Hence the objective in $(\mathcal{P}_{\boldsymbol{u}_1})$ cannot reach its minimum at $\boldsymbol{u}_1$, unless $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle$ is equal to the minimizer t_0 of the function

$$t \mapsto \log(1 + e^t) - \frac{\alpha}{2}t.$$

   However, in case $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle = t_0$, the point $\boldsymbol{u}_1$ lies on a line that is orthogonal to $\boldsymbol{w}_2 - \boldsymbol{w}_1$, and one can choose another point $\boldsymbol{u}_1'$ on the intersection of this line and the disc $\mathcal{B}'$ such that $\|\boldsymbol{u}_1'\| < \|\boldsymbol{u}_1\|$. In particular, $\boldsymbol{u}_1'$ is a better feasible candidate in comparison to $\boldsymbol{u}_1$.

3. We have shown above that the interior point $\boldsymbol{u}_1$ cannot be the solution of $(\mathcal{P}_{\boldsymbol{u}_1})$. Excluding all interior points, we now consider the boundary set $\partial\mathcal{F}$, which consists of points on the circle $\mathcal{C}$ that lie in the positive quadrant (denoted by $\partial\mathcal{F}_1$), points on the x-axis between 0 and the intersection point $\boldsymbol{u}_1^*$ of $\mathcal{C}$ with the x-axis (denoted by $\partial\mathcal{F}_2$), as well as points on the y-axis between 0 and the intersection point of $\mathcal{C}$ with the y-axis (denoted by $\partial\mathcal{F}_3$). Note that the circle $\mathcal{C}$ may have no intersection with the x-axis, in that case we simply consider $\partial\mathcal{F}_2$ as the empty set.

   We will show that the solution must be a point on $\partial\mathcal{F}_1$, in which we find the possible optimal positions of $\boldsymbol{u}_1$ on each of the other boundary subsets, i.e. $\partial\mathcal{F}_2$ and $\partial\mathcal{F}_3$.

   (a) On $\partial\mathcal{F}_3$: Observe that moving a point $\boldsymbol{u}_1$ along $\partial\mathcal{F}_3$ in the direction toward the origin will decrease both the scalar product $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle > 0$ (as the angle is kept

unchanged while the length of $\boldsymbol{u}_1$ is decreased) and the regularization term $\|\boldsymbol{u}_1\|^2$. Since the function $t \mapsto \log(1+e^t) - \frac{\alpha}{2}t$ is monotonically increasing on $[0, \infty)$, moving $\boldsymbol{u}_1$ in this direction decreases the objective in $(\mathcal{P}_{\boldsymbol{u}_1})$. Therefore, the best candidate on $\partial\mathcal{F}_3$ is the lowest possible point on $\partial\mathcal{F}_3$, i.e. 0 in case $\mathcal{C}$ has intersection point with the x-axis, or is the lower intersection point of $\mathcal{C}$ with the y-axis otherwise.

(b) On $\partial\mathcal{F}_2$ (in case it is not empty): Here, the objective becomes $f(\|\boldsymbol{u}_1\|)$ where the function f is defined by

$$f(t) = \log(1 + e^{c_1 t}) - \frac{\alpha}{2} c_1 t + \lambda t^2,$$

with $c_1 := \frac{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}{\|\boldsymbol{h}_1\|}$. Observe that f is convex (this can be seen by directly computing the 2nd derivative of f) and achieves its minimum at $t = \|\boldsymbol{h}_1\|$ (because $\boldsymbol{h}_1$ is the minimum of $(\mathcal{P}_{\boldsymbol{u}_1})$ without the side constraint, see Remark 1). Hence on the interval $[0, \|\boldsymbol{h}_1\|]$ it is monotonically decreasing. It follows that $\boldsymbol{u}_1^*$ is the best candidate on $\partial\mathcal{F}_2$.

In summary we have shown that the optimal position of $\boldsymbol{u}_1$ must be on $\partial\mathcal{F}_1$, $\partial\mathcal{F}_2$ or $\partial\mathcal{F}_3$. On the other hand, all candidates on $\partial\mathcal{F}_3$ are worse than a point in $\partial\mathcal{F}_2$ and all candidates on $\partial\mathcal{F}_2$ are worse than a point in $\partial\mathcal{F}_1$ (in case $\partial\mathcal{F}_2 = \emptyset$ we have that all candidates on $\partial\mathcal{F}_3$ are worse than a point in $\partial\mathcal{F}_1$). Therefore the minimizer must be a point on $\partial\mathcal{F}_1$, in particular on the circle $\mathcal{C}$.

□

Having said that the optimal position of $\boldsymbol{u}_1$ with respect to $(\mathcal{P}_{\boldsymbol{u}_1})$ must be on the circle $\mathcal{C}$, we are now interested in the case where r is close to $r_{\max}$, in which the circle $\mathcal{C}$ has intersection with the segment $(0, \boldsymbol{h}_1)$ (see Figure 10). In this case we can even restrict the possible optimal positions to a smaller subset of the circle.

Lemma C.2. *Suppose that $r_{\max} \geq r \geq r_{\max}/\sqrt{2}$, so that the circle $\mathcal{C}$ has intersection $\boldsymbol{u}_1^*$ with the line segment $(0, \boldsymbol{h}_1)$ and intersection $\boldsymbol{u}_1^{**}$ with the line segment $(\boldsymbol{h}_2, \boldsymbol{h}_1)$. Then, the minimizer of $(\mathcal{P}_{\boldsymbol{u}_1})$ lies on the arc between $\boldsymbol{u}_1^*$ and $\boldsymbol{u}_1^{**}$.*

Proof. First we rewrite the objective of $(\mathcal{P}_{\boldsymbol{u}_1})$ as

$$f\big(\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle\big) + \lambda \|\boldsymbol{u}_1\|^2 .$$

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

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

Next, we parameterize the circle $\mathcal{C}$ by the polar coordinate. Let $R := \frac{C_{MD} r}{\eta \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|}$ and θ be the angle between $(\boldsymbol{h}_2, \boldsymbol{u}_1)$ and $(\boldsymbol{h}_2, \boldsymbol{h}_1)$. Then, since $\boldsymbol{w}_2 - \boldsymbol{w}_1$ is proportional to $\boldsymbol{h}_2 - \boldsymbol{h}_1$ we have

$$\begin{aligned}\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle &= \langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_2 \rangle + \langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 - \boldsymbol{h}_2 \rangle \\ &= \langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_2 \rangle - R \|\boldsymbol{w}_2 - \boldsymbol{w}_1\| \cos\theta.\end{aligned}$$

Note that $\boldsymbol{u}_1$ can be on both sides of the line $(\boldsymbol{h}_2, \boldsymbol{h}_1)$, but for the calculation of $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle$ it is not necessary to distinguish between these two cases. In general, when θ increases, $\cos\theta$ decreases (we can exclude the case $\theta > \pi/2$ because in this case $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle$ becomes positive and the norm of $\boldsymbol{u}_1$ is also large, so the objective becomes large and $\boldsymbol{u}_1$ cannot be the minimizer), and thus $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle$ increases.

Now we claim that the optimal position of $\boldsymbol{u}_1$ must be on the arc between $\boldsymbol{u}_1^*$ and its reflection $\boldsymbol{u}_1'$ about the line $(\boldsymbol{h}_1, \boldsymbol{h}_2)$. To show this we consider a point $\boldsymbol{u}_1$ that lies on the other part of the circle $\mathcal{C}$. By the above observation on the monotonicity of $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle$ with respect to θ we see that

$$\begin{aligned}\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle &> \langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1' \rangle \\ &= \langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1^* \rangle \\ &\geq \langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle .\end{aligned}$$

Recall from the proof of Theorem 3.2 that $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle$ 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

$$f\Big(\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle \Big) > f\Big(\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1' \rangle \Big)$$

On the other hand, by the law of cosines applied to the triangle $(0, \boldsymbol{h}_2, \boldsymbol{u}_1)$ we obtain

$$\|\boldsymbol{u}_1\|^2 = \|\boldsymbol{h}_2\|^2 + R^2 - 2R\,\|\boldsymbol{h}_2\| \cos\Big(\frac{\pi}{4} + \theta\Big),$$

which is increasing in θ (again we exclude the case $\theta > \pi/2$ as discussed above). This shows that $\|\boldsymbol{u}_1\| > \|\boldsymbol{u}_1'\|$. Combining the above two inequalities we see that $\boldsymbol{u}_1'$ is a better candidate than $\boldsymbol{u}_1$.

Finally, the desired statement follows from the observation that we can exclude all points on the arc between $\boldsymbol{u}_1^{**}$ and $\boldsymbol{u}_1'$, because each point on this arc can be reflected about the line $(\boldsymbol{h}_1, \boldsymbol{h}_2)$ to a point with the same value of $\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 \rangle$, but with smaller norm and this gives a better value of the objective.

□

Note that similar to the optimal position of $\boldsymbol{U}$, the points $\boldsymbol{u}_1^*$, $\boldsymbol{u}_1^{**}$ from Lemma C.2 also depend on r. Hence to be clear, we may write $\boldsymbol{u}_1 = \boldsymbol{u}_1(r)$, $\boldsymbol{u}_1^* = \boldsymbol{u}_1^*(r)$ and $\boldsymbol{u}_1^{**} = \boldsymbol{u}_1^{**}(r)$ for $r \in [r_{\max}/\sqrt{2}, r_{\max}]$. Observe that

$$\boldsymbol{u}_1(r_{\max}) = \boldsymbol{u}_1^*(r_{\max}) = \boldsymbol{u}_1^{**}(r_{\max}) = \boldsymbol{h}_1.$$

The following lemma shows that for r close to $r_{\max}$, the distance between $\boldsymbol{u}_1(r)$ and $\boldsymbol{h}_1$ behaves almost linearly with respect to the distance between r and $r_{\max}$.

Lemma C.3. *Let $\boldsymbol{u}_1(r)$ be the minimizer of ($\mathcal{P}_{\boldsymbol{u}_1}$) with input $r \in [r_{\max}/\sqrt{2}, r_{\max}]$. Then, there exists constants $c, C > 0$ (depending on $\|\boldsymbol{h}_1\| = \|\boldsymbol{h}_2\|$ and C_{MD}, but not on other parameters such as r, η, etc) such that*

$$\|\boldsymbol{u}_1(r) - \boldsymbol{h}_1\| \in \Big(c\frac{r_{\max} - r}{\eta}, C\frac{r_{\max} - r}{\eta} \Big).$$

Proof. From Lemma C.2 we know that $\boldsymbol{u}_1(r)$ lies on the arc between $\boldsymbol{u}_1^*(r)$ and $\boldsymbol{u}_1^{**}(r)$, hence its distance to $\boldsymbol{h}_1$ is lower bounded by the distance from $\boldsymbol{u}_1^{**}(r)$ to $\boldsymbol{h}_1$ and is upper bounded by the

distance from $\boldsymbol{u}_1^*(r)$ to $\boldsymbol{h}_1$. Hence we have

$$\begin{aligned}
\|\boldsymbol{u}_1(r)-\boldsymbol{h}_1\| &\le \|\boldsymbol{u}_1^*(r)-\boldsymbol{u}_1^*(r_{\max})\| \\
&= \|\boldsymbol{u}_1^*(r_{\max})\| - \|\boldsymbol{u}_1^*(r)\| \\
&= \sqrt{\|\boldsymbol{h}_2-\boldsymbol{u}_1^*(r_{\max})\|^2 - \|\boldsymbol{h}_2\|^2} - \sqrt{\|\boldsymbol{h}_2-\boldsymbol{u}_1^*(r)\|^2 - \|\boldsymbol{h}_2\|^2} \\
&= \sqrt{\frac{C_{MD}^2 r_{\max}^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \|\boldsymbol{h}_2\|^2} - \sqrt{\frac{C_{MD}^2 r^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \|\boldsymbol{h}_2\|^2} \\
&= \frac{\frac{C_{MD}^2(r_{\max}^2-r^2)}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2}}{\sqrt{\frac{C_{MD}^2 r_{\max}^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \|\boldsymbol{h}_2\|^2} + \sqrt{\frac{C_{MD}^2 r^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \|\boldsymbol{h}_2\|^2}} \\
&= \frac{C_{MD}(r_{\max}-r)}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|} \cdot \frac{\frac{C_{MD}(r_{\max}+r)}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2}}{\sqrt{\frac{C_{MD}^2 r_{\max}^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \|\boldsymbol{h}_2\|^2} + \sqrt{\frac{C_{MD}^2 r^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \|\boldsymbol{h}_2\|^2}} \\
&\le \frac{C_{MD}(r_{\max}-r)}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|} \cdot \frac{\frac{2C_{MD}r_{\max}}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|}}{\sqrt{\frac{C_{MD}^2 r_{\max}^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \|\boldsymbol{h}_2\|^2}} \\
&= \frac{C_{MD}(r_{\max}-r)}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|} \cdot \frac{\frac{2C_{MD}r_{\max}}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|}}{\sqrt{\frac{C_{MD}^2 r_{\max}^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \frac{1}{2}\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2}} \\
&= \frac{C_{MD}(r_{\max}-r)}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|} \cdot \frac{\frac{2C_{MD}r_{\max}}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|}}{\sqrt{\frac{C_{MD}^2 r_{\max}^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2} - \frac{1}{2}\frac{C_{MD}^2 r_{\max}^2}{\eta^2\|\boldsymbol{h}_1-\boldsymbol{h}_2\|^2}}} \\
&= \frac{2\sqrt{2}C_{MD}}{\|h_1-h_2\|} \cdot \frac{r_{\max}-r}{\eta}.
\end{aligned}$$

On the other hand it also holds

$$\begin{aligned}
\|\boldsymbol{u}_1(r)-\boldsymbol{h}_1\| &\ge \|\boldsymbol{u}_1^{**}(r)-\boldsymbol{h}_1\| \\
&= \|\boldsymbol{h}_2-\boldsymbol{u}_1^{**}(r_{\max})\| - \|\boldsymbol{h}_2-\boldsymbol{u}_1^{**}(r)\| \\
&= \frac{C_{MD}r_{\max}}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|} - \frac{C_{MD}r}{\eta\|\boldsymbol{h}_1-\boldsymbol{h}_2\|} \\
&= \frac{C_{MD}}{\|\boldsymbol{h}_1-\boldsymbol{h}_2\|} \cdot \frac{r_{\max}-r}{\eta}.
\end{aligned}$$

Combining the above estimates yields the desired statement. □

C.3 THE BEHAVIOR OF $G_{\lambda,\alpha}$ NEAR $r_{\max}$

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

Lemma C.4. *Let $\boldsymbol{u}_1(r)$ be the minimizer of ($\mathcal{P}_{\boldsymbol{u}_1}$) with input $r\in[r_{\max}/\sqrt{2},r_{\max}]$. Then for any r such that $\frac{r_{\max}-r}{\eta}<1$, it holds*

$$G_{\lambda,\alpha}\Big(\boldsymbol{W},\boldsymbol{U}(r),r\Big) - G_{\lambda,\alpha}\Big(\boldsymbol{W},\boldsymbol{U}(r_{\max}),r_{\max}\Big) \ge C_1\left(\frac{r-r_{\max}}{\eta}\right)^2$$

for some constant $C_1>0$.

Proof. Due to symmetry, we only need to consider the half of $G_{\lambda,\alpha}$ that involves $\boldsymbol{u}_1$, i.e. the function $g:\mathbb{R}^M\to\mathbb{R}$,

$$g(\boldsymbol{u}_1) = \log\left(1+e^{\langle \boldsymbol{w}_2-\boldsymbol{w}_1,\boldsymbol{u}_1\rangle}\right) - \frac{\alpha}{2}\langle \boldsymbol{w}_2-\boldsymbol{w}_1,\boldsymbol{u}_1\rangle + \lambda\|\boldsymbol{u}_1\|^2. \tag{27}$$

We approximate $g(\boldsymbol{u}_1(r))$ using the second-order Taylor approximation around $\boldsymbol{h}_1 = \boldsymbol{u}_1(r_{\max})$,

$$g(\boldsymbol{u}_1) = g(\boldsymbol{h}_1) + \langle \nabla_{\boldsymbol{u}_1} g(\boldsymbol{h}_1), \boldsymbol{u}_1 - \boldsymbol{h}_1 \rangle + \frac{1}{2} \langle \boldsymbol{H}_{\boldsymbol{u}_1} g(h_1)(\boldsymbol{u}_1 - \boldsymbol{h}_1), \boldsymbol{u}_1 - \boldsymbol{h}_1 \rangle + O(\|\boldsymbol{u}_1 - \boldsymbol{h}_1\|^3), \tag{28}$$

where the derivatives of g (at $\boldsymbol{h}_1$) are given by

$$\begin{aligned}
\nabla_{\boldsymbol{u}_1} g(\boldsymbol{h}_1) &= \frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}}{1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}} (\boldsymbol{w}_2 - \boldsymbol{w}_1) - \frac{\alpha}{2} (\boldsymbol{w}_2 - \boldsymbol{w}_1) + 2\lambda \boldsymbol{h}_1, \\
\boldsymbol{H}_{\boldsymbol{u}_1} g(\boldsymbol{h}_1) &= \frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}}{(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle})^2} (\boldsymbol{w}_2 - \boldsymbol{w}_1)(\boldsymbol{w}_2 - \boldsymbol{w}_1)^\top + 2\lambda \boldsymbol{I}.
\end{aligned}$$

Since $\boldsymbol{u}_1 = \boldsymbol{h}_1$ is the minimum of $g(\boldsymbol{u}_1)$ under the constraint $\boldsymbol{u}_1 \geq 0$ and $\boldsymbol{u}_1(r)$ is always feasible for any r, the linear term in (28) is non-negative, i.e.

$$\langle \nabla_{\boldsymbol{u}_1} g(\boldsymbol{h}_1), \boldsymbol{u}_1 - \boldsymbol{h}_1 \rangle \geq 0.$$

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

$$\begin{aligned}
\langle H_{\boldsymbol{u}_1} g(\boldsymbol{h}_1)(\boldsymbol{u}_1 - \boldsymbol{h}_1), \boldsymbol{u}_1 - \boldsymbol{h}_1 \rangle &> \frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}}{(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle})^2} \langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{u}_1 - \boldsymbol{h}_1 \rangle^2 \\
&\geq \frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}}{2(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle})^2} \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 \|\boldsymbol{u}_1 - \boldsymbol{h}_1\|^2,
\end{aligned}$$

where the last inequality holds because the angle between $\boldsymbol{w}_2 - \boldsymbol{w}_1$ and $\boldsymbol{u}_1 - \boldsymbol{h}_1$ 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

$$\begin{aligned}
g(\boldsymbol{u}_1(r)) - g(\boldsymbol{h}_1) &> \left(\frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}}{2(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle})^2} \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 + 2\lambda \right) \|\boldsymbol{u}_1(r) - \boldsymbol{h}_1\|^2 \\
&\geq C_1 \left(\frac{r - r_{\max}}{\eta} \right)^2
\end{aligned}$$

where the constant C_1 is given by

$$C_1 = \Big(\frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}}{2(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle})^2} \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 + 2\lambda \Big) c^2$$

with c from Lemma C.3, i.e.

$$C_1 = \Big(\frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle}}{2(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \rangle})^2} \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 + 2\lambda \Big) \frac{C_{MD}^2}{\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|^2}.$$

□

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 $r < 1$, the function $r \mapsto F_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{H}, r)$ satisfies*

$$F_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{H}, r) - F_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{H}, 0) \leq C_2 r^2$$

for some constant $C_2 > 0$.

Proof. By symmetry we only need to consider the half of $F_{\lambda,\alpha}$ that involves $\boldsymbol{h}_1$, and to simplify notations we denote this by $\tilde{F} = \tilde{F}(r)$ with

$$\begin{aligned}
\tilde{F}(r) &= \int \left(\ell_\alpha\Big(\boldsymbol{W}, \boldsymbol{h}_1 + \boldsymbol{v}, \boldsymbol{y}_1^{(\alpha)}\Big) + \lambda \left\|\boldsymbol{h}_1 + \boldsymbol{v}\right\|^2 \right) d\mu_r^1(\boldsymbol{v}) \\
&= \int \left(\log\Big(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 + \boldsymbol{v}\rangle}\Big) - \frac{\alpha}{2} \left\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 + \boldsymbol{v} \right\rangle + \lambda \left\|\boldsymbol{h}_1 + \boldsymbol{v}\right\|^2 \right) d\mu_r^1(\boldsymbol{v}) \\
&= \int \left(\log\Big(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 + \boldsymbol{v}\rangle}\Big) + \lambda \left\|\boldsymbol{v}\right\|^2 \right) d\mu_r^1(\boldsymbol{v}) - \frac{\alpha}{2} \left\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 \right\rangle + \lambda \left\|\boldsymbol{h}_1\right\|^2,
\end{aligned}$$

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

$$\tilde{f}(\boldsymbol{v}) = \log\Big(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1 + \boldsymbol{v}\rangle}\Big) + \lambda \left\|\boldsymbol{v}\right\|^2,$$

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

$$\begin{aligned}
\tilde{f}(\boldsymbol{v}) &= \tilde{f}(0) + \Big\langle \nabla \tilde{f}(0), \boldsymbol{v} \Big\rangle + \frac{1}{2} \Big\langle H\tilde{f}(0)\boldsymbol{v}, \boldsymbol{v} \Big\rangle + O(r^3) \\
&= \tilde{f}(0) + \frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}}{1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}} \left\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{v} \right\rangle \\
&\quad + \frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}}{2\Big(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}\Big)^2} \left\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{v} \right\rangle^2 + \lambda \left\|\boldsymbol{v}\right\|^2 + O(r^3).
\end{aligned}$$

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

$$\begin{aligned}
\tilde{F}(r) - \tilde{F}(0) &= \int \tilde{f}(\boldsymbol{v}) d\mu_1^r(\boldsymbol{v}) \\
&= \int \left(\frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}}{2\Big(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}\Big)^2} \left\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, v \right\rangle^2 + \lambda \left\|\boldsymbol{v}\right\|^2 \right) d\mu_1^r(\boldsymbol{v}) + O(r^3) \\
&\leq \left(\frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle} \left\|\boldsymbol{w}_2 - \boldsymbol{w}_1\right\|^2}{2\Big(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}\Big)^2} + \lambda \right) \int \left\|\boldsymbol{v}\right\|^2 d\mu_1^r(\boldsymbol{v}) + O(r^3) \\
&\leq A^2 \left(\frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle} \left\|\boldsymbol{w}_2 - \boldsymbol{w}_1\right\|^2}{2\Big(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}\Big)^2} + \lambda \right) r^2 + O(r^3),
\end{aligned}$$

where A is the constant for which $\|\boldsymbol{v}\| \leq Ar$ holds (see Assumption 4.2). Thus the desired statement follows with

$$C_2 = A^2 \left(\frac{e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle} \left\|\boldsymbol{w}_2 - \boldsymbol{w}_1\right\|^2}{2\Big(1 + e^{\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{h}_1\rangle}\Big)^2} + \lambda \right).$$

□

C.5 Estimation of the optimal dilation r_*

In this subsection we come back to the $\mathcal{MD}$ problem, i.e. the minimization of the MD risk

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

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

$$\min_{U} \quad G_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{U}, r) \quad \text{s.t.} \quad (4), (5).$$

Then the $\mathcal{MD}$ problem is reduced to the minimization over the dilation r, namely

$$\min_{r} \quad F_{\lambda,\alpha}(\boldsymbol{W}, \boldsymbol{H}, r) + \eta G_{\lambda,\alpha}\big(\boldsymbol{W}, \boldsymbol{U}(r), r\big). \tag{29}$$

Lemma C.6. *The solution r_* to the problem (29) satisfies $r_{\max} \geq r_* \geq r_{\max}(1 - C'\eta^{1/2})$, where C' is a constant given by*

$$C' := \frac{A\sqrt{2}\,\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|}{C_{MD}}.$$

Proof. To shorten notations, we define

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

Then the problem (29) can be rewritten as

$$\min_{r} \quad f(r) + \eta g(r). \tag{30}$$

First we observe that $r_* \leq r_{\max}$ because for any $r > r_{\max}$ we have that $f(r) > f(r_{\max})$ while $g(r) = g(r_{\max})$ (see Remark 1). Thus we only need to show the lower bound on r_*.

Let $\epsilon \in \left(0, \frac{C_{MD}}{\|h_1 - h_2\|^2}\right)$, we will show that the solution to the reduced $\mathcal{MD}$ problem (30) cannot be r for any $r < (1-\epsilon)r_{\max}$, provided that ϵ is sufficiently large (this condition will be later specified more precisely). To see this we will show that for any such r it holds

$$f(r) + \eta g(r) > f(r_{\max}) + \eta g(r_{\max}). \tag{31}$$

By Lemma C.5 we have that

$$f(r_{\max}) - f(r) \leq f(r_{\max}) - f(0) \leq C_2 r_{\max}^2.$$

On the other hand, from Lemma C.4 it follows that

$$g(r) - g(r_{\max}) \geq g\Big((1-\epsilon)r_{\max}\Big) - g(r_{\max}) \geq C_1 \frac{\epsilon^2 r_{\max}^2}{\eta^2}$$

holds for some constant $c_2 > 0$.

Combining the above observations we see that (31) will hold if

$$C_1 \frac{\epsilon^2 r_{\max}^2}{\eta} \geq C_2 r_{\max}^2,$$

which holds provided that

$$\epsilon \geq C'\eta^{1/2}$$

with

$$C' = \frac{A\sqrt{2}\,\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|}{C_{MD}}.$$

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

□

C.6 FINALIZING THE PROOF

In previous subsections we have approximately estimated the optimal dilation r_* of the $\mathcal{MD}$ problem in general, i.e. the LS parameter α can take any value in $\{0, \alpha_0\}$. Now we distinguish between the two values of α 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

$$\frac{r_*^{CE}}{\left\|\boldsymbol{h}_1^{CE} - \boldsymbol{h}_2^{CE}\right\|} > \frac{r_*^{LS}}{\left\|\boldsymbol{h}_1^{LS} - \boldsymbol{h}_2^{LS}\right\|}. \tag{32}$$

By Assumption 4.2 we have $\left\|\boldsymbol{h}_1^{CE} - \boldsymbol{h}_2^{CE}\right\| = \gamma \left\|\boldsymbol{h}_1^{LS} - \boldsymbol{h}_2^{LS}\right\|$, hence

$$\frac{r_{\max}^{CE}}{\left\|\boldsymbol{h}_1^{CE} - \boldsymbol{h}_2^{CE}\right\|} = \frac{\eta \left\|\boldsymbol{h}_1^{CE} - \boldsymbol{h}_2^{CE}\right\|}{C_{MD}} = \gamma \frac{\eta \left\|\boldsymbol{h}_1^{LS} - \boldsymbol{h}_2^{LS}\right\|}{C_{MD}} = \gamma \frac{r_{\max}^{LS}}{\left\|\boldsymbol{h}_1^{LS} - \boldsymbol{h}_2^{LS}\right\|}.$$

Combining this and Lemma C.6 we obtain

$$\begin{aligned}
\frac{r_*^{CE}}{\left\|\boldsymbol{h}_1^{CE} - \boldsymbol{h}_2^{CE}\right\|} &> \frac{\left(1 - C'\eta^{1/2}\right) r_{\max}^{CE}}{\left\|\boldsymbol{h}_1^{CE} - \boldsymbol{h}_2^{CE}\right\|} \\
&= \gamma\left(1 - C'\eta^{1/2}\right) \frac{r_{\max}^{LS}}{\left\|\boldsymbol{h}_1^{LS} - \boldsymbol{h}_2^{LS}\right\|} \\
&\geq \gamma\left(1 - C'\eta^{1/2}\right) \frac{r_*^{LS}}{\left\|\boldsymbol{h}_1^{LS} - \boldsymbol{h}_2^{LS}\right\|}.
\end{aligned}$$

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.