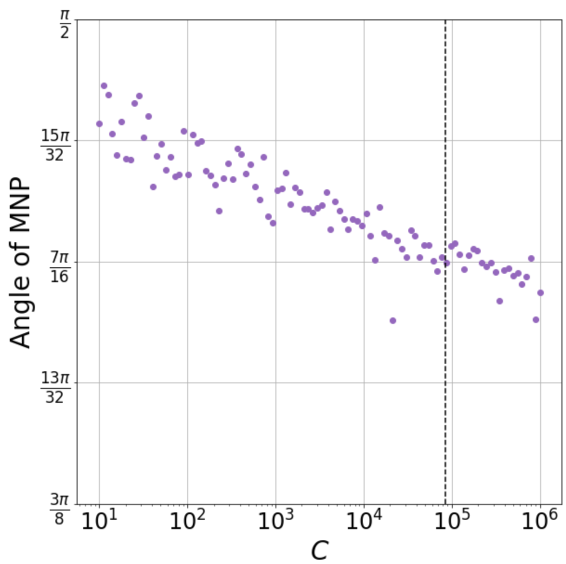


Figure 7: An empirical validation of Proposition 1. Under the assumption of uniform hyperspherical distribution, the expected angle of the nearest prototype (MNP) decreases only slightly while the total number of prototypes C increases exponentially. The vertical dashed line indicates $C = 85,742$, the total number of classes in C-MS1M, with the corresponding angle of MNP $= 78.64°$.

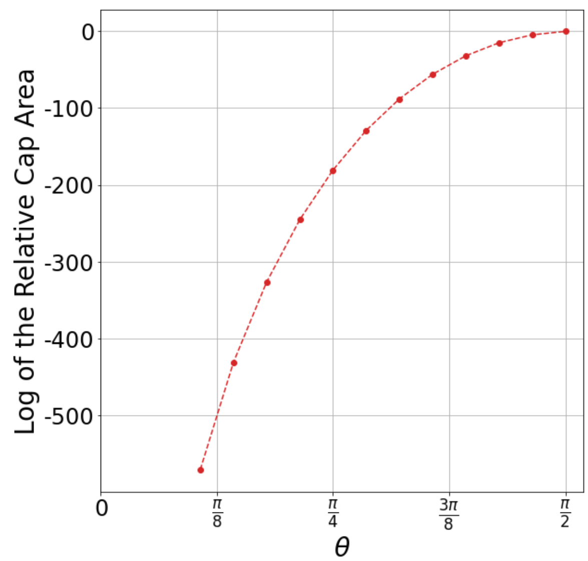


Figure 8: A corroboration of Proposition 2 showing the mass of a high-dimensional spherical cap concentrates on its boundary. The plot shows the relative area of a θ-cap of the $(d-1)$-sphere with $d = 512$, plotted in logarithmic scale, as a function of θ. The cap area decreases exponentially as its defining planar angle decreases.

A. Numerical Approximations

In Section 3.3, we present a theoretical framework for studying intra-class and inter-class distances under the assumption of uniform distribution of the prototypes over the unit hypersphere. In this appendix, we sample prototypes according to this assumption with dimension $d = 512$ and $C = 85,742$ training classes and show that the results closely match the analytical results obtained in the limit of large d and C.

Proposition 1

Proposition 1 expresses the nearest-neighbor inter-class distance in the limit of infinitely many classes $C \to \infty$. To see the behavior of the nearest-neighbor distance for finite C, we sample C prototypes uniformly on the surface of the 511-dimensional hypersphere and compute the mean nearest-neighbor angle across the C prototypes. Fig. 7 shows results for a range of C including the size of our training set C-MS1M (vertical dashed line).

Proposition 2

In Proposition 2, we argue that for large dimension d the mass of a $(d-1)$-dimensional spherical cap is highly concentrated in a tight annulus around its $(d-2)$-dimensional boundary. In Fig. 8, we compute the area of a θ-spherical cap of the $(d-1)$-sphere (with $\theta \in [0, \frac{\pi}{2}]$) given by

$$A_d^\theta(r) = \frac{1}{2} A_d(r) I_{\sin^2(\theta)}\left(\frac{d-1}{2}, \frac{1}{2}\right)$$

with $I_x(a, b)$ the regularized incomplete beta function. Notice the exponential decay of the area as a function of the θ-spherical cap planar angle θ, meaning that most of the area of a cap of angle θ is near angle θ.

Given this measure concentration property of the high-dimensional sphere, we claim that sampling feature points uniformly in a hyperspherical cap generates points that are tightly concentrated near the spherical cap boundary. In Fig. 9, we display the mean and standard deviation of the angle between points sampled on a θ-spherical cap and the axis of the cap, as a function of θ. As expected, with high embedding space dimension $d = 512$, these statistics lie very close to the line $y = \theta$. For comparison, we display the corresponding statistics for $d = 8$, which show a much lower concentration of sampled points near θ. To generate the samples for Fig. 9, we used Arun's linear time algorithm for sampling points from a section of the surface of a hypersphere [25].

The gradient of the margin softmax losses with respect to the correct-class angle nearly vanishes at a hyperparameter-dependent "termination angle", from which we assume that all correct-class angles will be smaller than this termination angle for a trained model. Proposition 2 tells us that angles

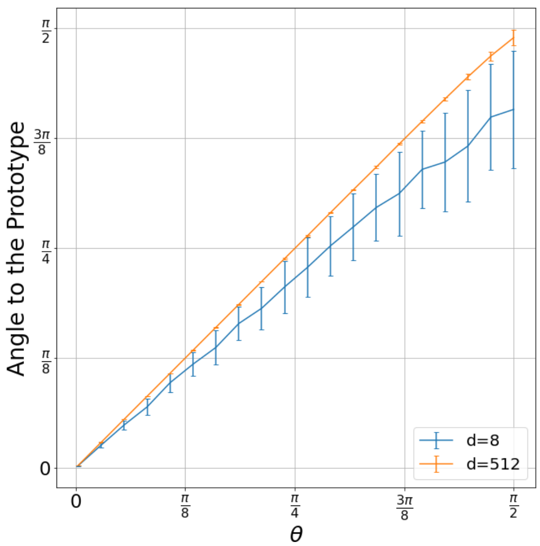


Figure 9: An empirical validation of Proposition 2. Vectors are sampled uniformly from within the hyper-spherical cap defined by angle θ, and the mean and standard deviation of angles between sampled vectors and the center axis (prototype) are plotted. The orange line corresponds to $d = 512$ and the blue line corresponds to $d = 8$. For a high-dimensional spherical cap, uniformly sampled vectors have a high probability to lay at the boundary of the cap as that part dominates the area.

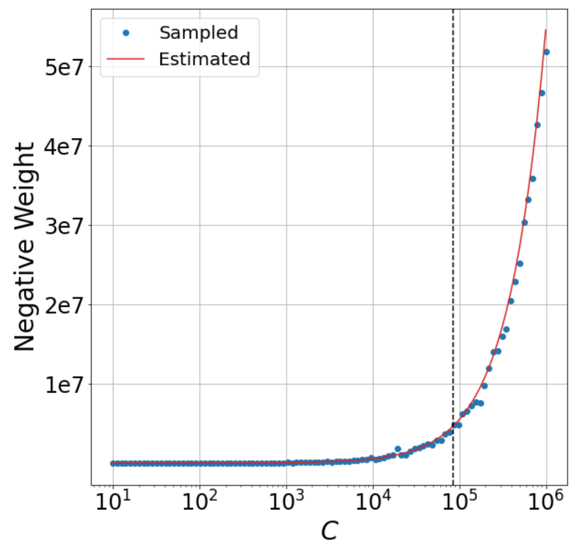


Figure 10: An empirical validation of Proposition 3. The red line represents the approximation from (10) of the negative weight $\sum_{j \neq y_i} e^{s z_j}$ against the number of classes C, while the blue dots indicate the negative weight from C vectors uniformly sampled on a $(d-1)$-sphere. The vertical dashed line indicates $C = 85,742$, the number of classes of C-MS1M, with corresponding negative weight $4.4562 \cdot 1e6$.

are unlikely to be significantly smaller than the termination angle by chance, which justifies approximating the average correct-class angle as *equal* to the termination angle (rather than less than or equal).

Proposition 3

Proof To make the dependency of C, d as a function of s explicit in the limit (11), recall the error bound given by the law of large numbers for the sum of C i.i.d. random variables $\overline{X}_C = X_1 + \cdots + X_C$ with mean μ and variance σ^2

$$\mathsf{P}\left(\frac{|\overline{X}_C - \mu|}{\mu} > \epsilon\right) < \frac{\sigma^2}{C\epsilon^2\mu^2} \tag{13}$$

Replacing in the right hand side of equation (13) the moments of the lognormal distribution $X_j = e^{s z_j}$, with $z_j \sim \mathcal{N}(0, 1/\sqrt{d})$ (namely $\mu = e^{s^2/2d}$ and $\sigma = e^{2s^2/d}$) we get an approximation error of $\dfrac{1}{\epsilon^2}\dfrac{e^{s^2/d}}{C}$, which can be made arbitrarily small, as long as $C \gg e^{s^2/d}$. □

Numerical Approximation Proposition 3 approximates the expected value of the total wrong-class weight in the softmax denominator under our assumptions of hyperspherical uniform distribution of the prototypes and C sufficiently large. Notice that for the typical setting of the feature representation space ($d = 512$, $s = 64$, $C = 85,742$), the quantity $e^{s^2/d}/C \leq 0.0348 \ll 1$, so for this setting and smaller s we expect small errors. To show the accuracy of this approximation for finite C, in Fig. 10 we compare the predicted expected value for the total negative weight against the numerical approximation obtained from sampling and normalizing C points from a d-dimensional Gaussian distribution.

B. Optimal Margin Values

In Section 4.2, we report test set performance for models trained with each of the four different margin-based softmax losses using the margin values proposed in the corresponding literature (SphereFace $m_1 = 1.35$, ArcFace $m_2 = 0.5$, CosFace $m_3 = 0.35$). A more thorough analysis, when using the same setup for the different losses (same architecture, model hyperparameters, optimization schedule) and with several runs per margin value, reveals that improved performance is achieved for slightly different best margin values. In particular, this unveils equivalent performance between the additive margins of ArcFace and CosFace under our training scheme. Our multiplicative margin m_0 (AmpFace) has performance comparable to ArcFace and CosFace but slightly inferior. The lower performance may be due to the gradient with respect to the positive class angle scaling linearly with m_0, which may slow down opti-

AmpFace		SphereFace		ArcFace		CosFace	
m_0	MF Id	m_1	MF Id	m_2	MF Id	m_3	MF Id
1.0	88.95	1.0	88.95	0.0	88.58	0.0	88.58
0.9	90.05$_{\pm 0.58}$	–	–	0.1	92.97$_{\pm 0.12}$	0.1	92.44
0.8	92.07$_{\pm 0.55}$	1.2	95.51$_{\pm 0.04}$	0.2	95.96$_{\pm 0.06}$	0.2	96.42$_{\pm 0.10}$
0.7	93.69$_{\pm 0.17}$	1.3	96.25	0.3	97.35$_{\pm 0.17}$	0.3	97.72$_{\pm 0.04}$
0.6	95.39$_{\pm 0.61}$	**1.4**	**96.30**$_{\pm 0.06}$	0.4	97.93$_{\pm 0.04}$	0.4	98.27$_{\pm 0.07}$
0.5	96.91$_{\pm 0.06}$	–	–	0.5	98.21$_{\pm 0.03}$	**0.5**	**98.43**$_{\pm 0.09}$
0.4	97.58$_{\pm 0.05}$	1.6	93.04$_{\pm 0.11}$	**0.6**	**98.26**$_{\pm 0.09}$	0.6	98.30$_{\pm 0.03}$
0.3	**97.82**$_{\pm 0.06}$	–	–	0.7	98.26$_{\pm 0.06}$	0.7	98.20$_{\pm 0.12}$
0.2	97.40	1.8	79.26$_{\pm 0.99}$	0.8	98.20$_{\pm 0.13}$	0.8	98.31$_{\pm 0.11}$
0.1	96.85	–	–	0.9	98.07$_{\pm 0.06}$	0.9	98.18$_{\pm 0.12}$

Table 3: An extensive margin hyperparameter search for AmpFace, SphereFace, ArcFace, and CosFace. Models are all trained under the same protocol as described in section 4.1 and performance on MegaFace Id is reported. AmpFace and SphereFace are trained with WC-ReLU to avoid the polar collapse mode described in Section 3.2.

mization and impede final performance with a fixed training schedule. SphereFace has inferior performance and suffers from difficult optimization as observed in [14]. Table 3 shows the hyperparameter grid search for the optimal values for each of the proposed margin penalties.

The resulting best margins $m_0 = 0.35$, $m_2 = 0.6$ and $m_3 = 0.5$ have the interesting property visualized in Fig. 5. In each case, the gradient of the corresponding loss with respect to the correct-class angle approximately vanishes at the same value of the angle. Moreover this shared termination angle is approximately half the predicted inter-class angle. We hope that future research will elucidate this relationship and perhaps enable analytical predictions of the best margin values, replacing the laborious hyperparameter search. However, as described in Section 4.3, a precise argument for why optimization should terminate with intra-class distance half of inter-class distance is beyond our current understanding.

For SphereFace, the best performance is attained for a model trained with $m_1 = 1.35$, with MF Id score 96.39 ± 0.18 significantly lower than its alternative margin penalty formulations. We notice that SphereFace suffers from difficult optimization, with the gradient of the loss w.r.t. the positive class scaling quadratically with m_1 (equation 12). In Fig. 5, we plot the value of $m_1 = 1.85$ that would lead to the same optimization termination angle as AmpFace, ArcFace, and CosFace. In Section 4.3, we discuss recent work [14] that shows higher performance can be achieved in this setting under more carefully regularized model training.

C. Visualization of Loss Parameters

The assumption of uniform distribution of the negative classes on the unit hypersphere detailed in Proposition 3 allows us to visualize the action of each hyperparameter in the softmax loss from (3): the number of classes C, the dimensionality d, the scale hyperparameter s and all four margins m_0, m_1, m_2 and m_3. We visualize the effect of each of these parameters in isolation on the loss L_i as a function of the correct-class angle θ_{y_i} (Fig. 11), the gradient for the correct-class angle $\partial L_i / \partial \theta_{y_i}$ (Fig. 12), and the gradient for an arbitrary wrong-class angle $\partial L_i / \partial \theta_j$, with the latter shown as a function of both θ_{y_i} (Fig. 13) and θ_j (Fig. 14).

Fig. 11 and Fig. 12 reveal a surprising similarity between certain parameters. For example, increasing the CosFace margin m_3 has a similar effect on the loss as decreasing the dimensionality d. By rearranging the terms in (3) to remove m_3 from the correct class logits and using the result from Proposition 3, the negative weight can be approximated by $(C-1)e^{s(s/(2d)+m_3)}$ which illustrates how d and m_3 might be inversely related. We also observe a similar connection between the scale factor s and the margin m_0, which is less surprising as they are both multiplicative coefficients for correct-class logits.

These similarities do not extend to the wrong-class gradients illustrated in Fig. 13 and Fig. 14, where decreasing d and s simply shrinks gradients for all angles while applying m_3 and m_0 changes the point at which gradients become non-zero (and, in the case of θ_j, increases the gradients for larger θ_j). An interesting observation in these plots is the similarity of negative gradients induced by all margin hyperparameters, indicating that the variation in the resulting classification performance is largely a function of the gradients for correct-class angles.

D. Visualization of Angular Decision Margins

Prior literature on margin-based softmax losses typically illustrates the differences between margins in a binary classification setting by showing the *decision margin* surrounding the decision boundary. Increasing the margin size shrinks the region of positive classification for each

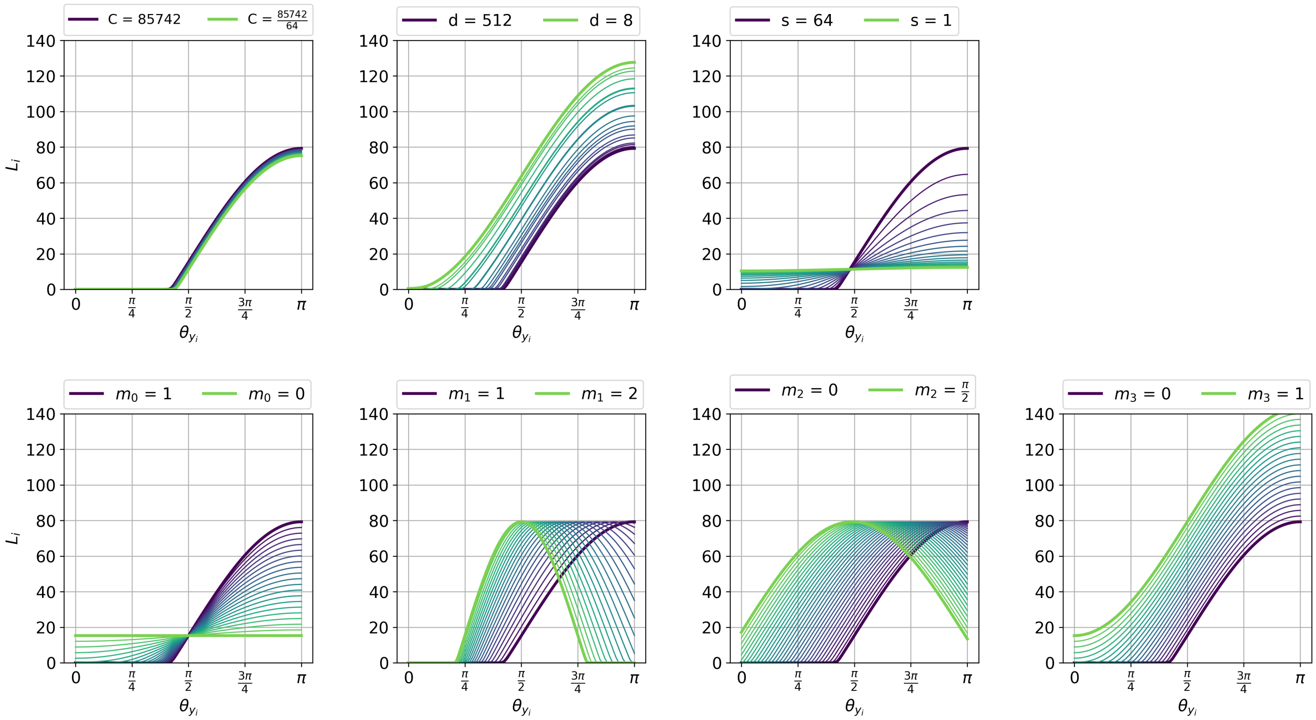


Figure 11: The loss L_i as a function of the correct-class angle θ_{y_i} for a range of values chosen to illustrate the effect of the following hyperparameters: C, the number of prototypes; d, the number of dimensions of the feature embedding space; s, the scale; and the m_0, m_1, m_2, and m_3 margins. Note the surprising similarity in the effect of varying d (row 1, column 2) and m_3 (row 2, column 4). Best viewed in color.

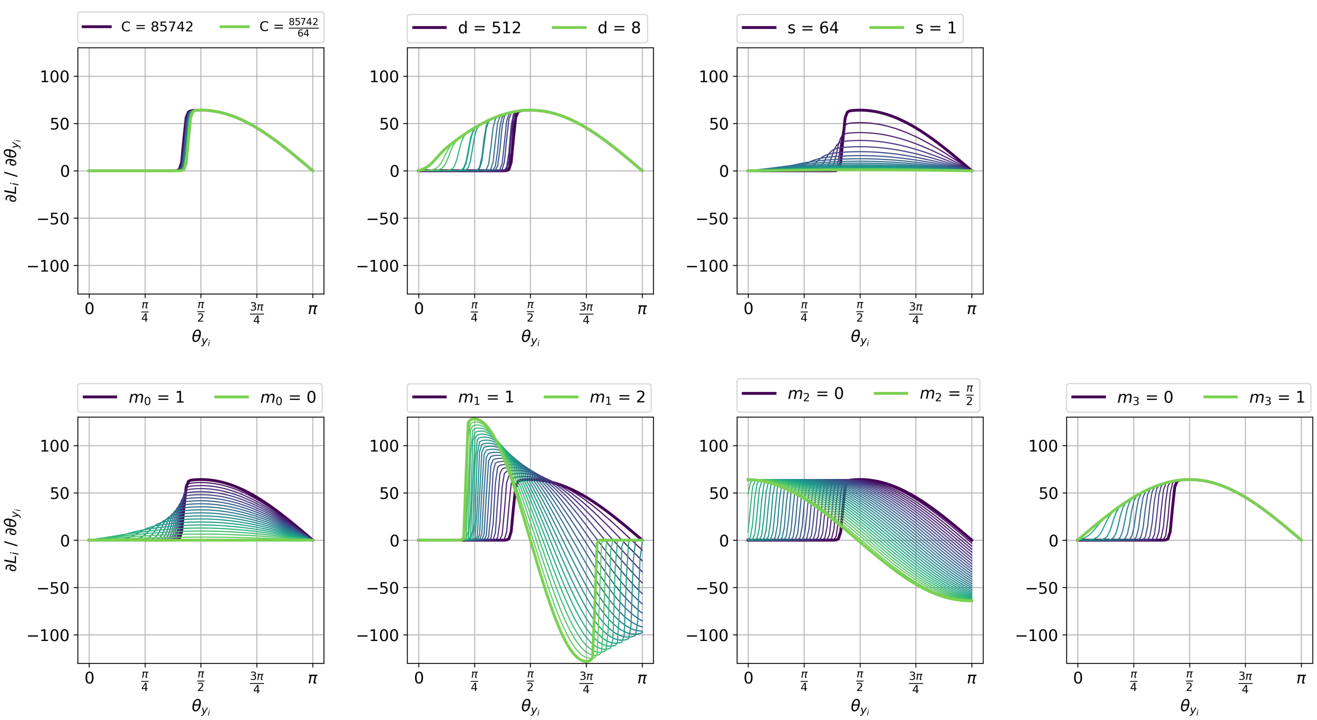


Figure 12: The gradient of the loss $\partial L_i / \partial \theta_{y_i}$ with respect to the correct-class angle θ_{y_i} as θ_{y_i} is varied. Best viewed in color.

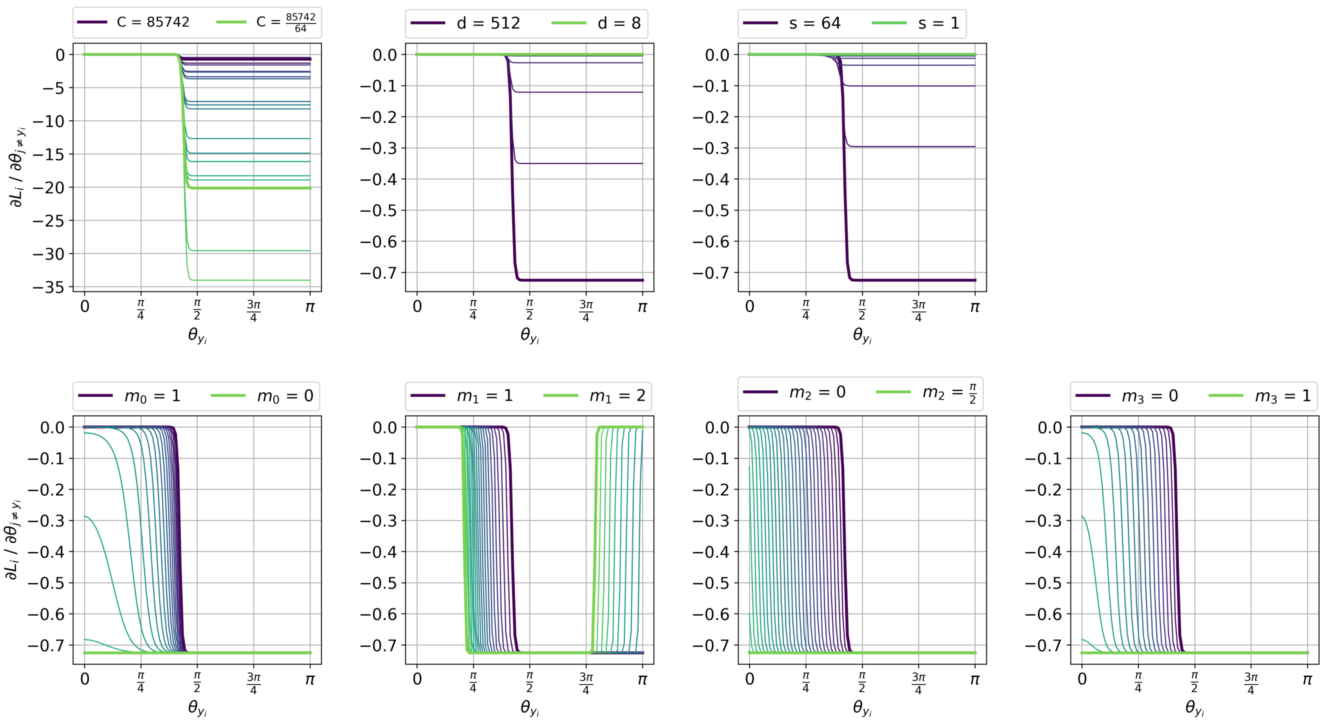


Figure 13: The gradient of the loss $\partial L_i/\partial\theta_j$ with respect to an arbitrary wrong-class angle $\theta_j, j \neq y_i$ as the correct-class angle θ_{y_i} is varied and θ_j is fixed to $80°$. Best viewed in color.

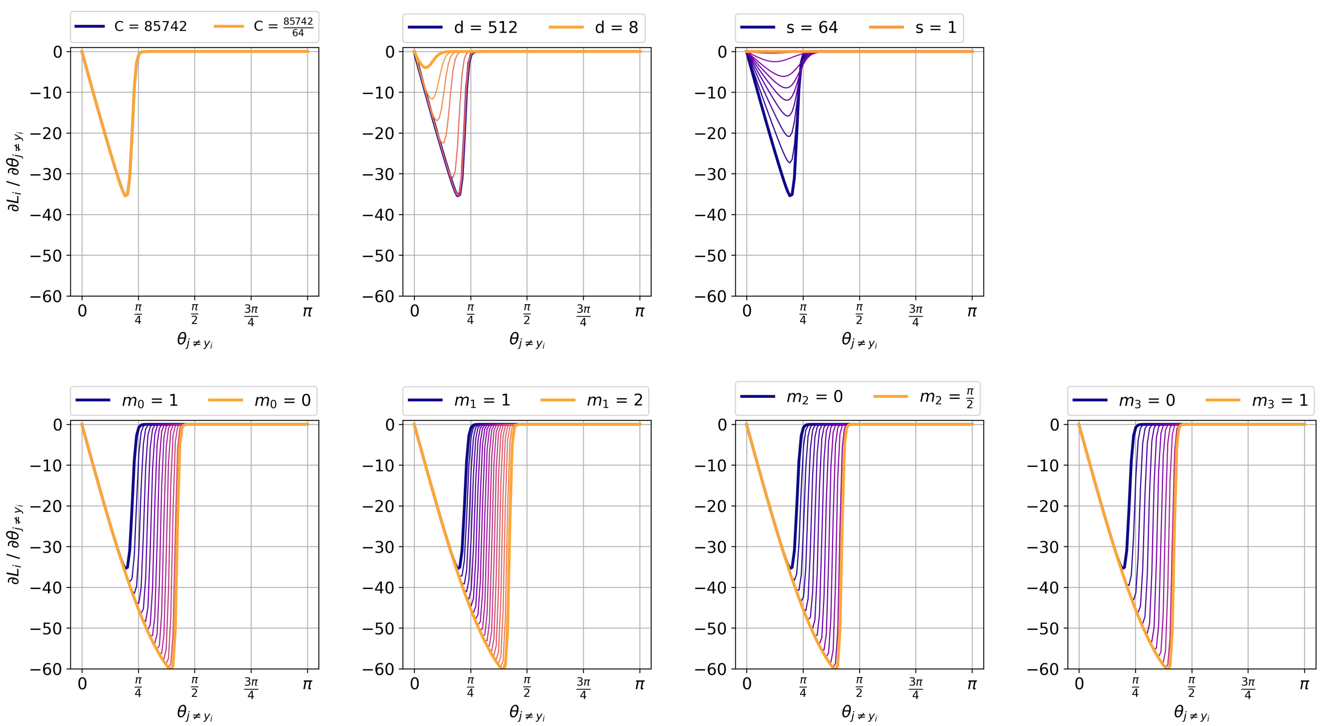


Figure 14: The gradient of the loss $\partial L_i/\partial\theta_j$ with respect to an arbitrary wrong-class angle $\theta_j, j \neq y_i$ as θ_j is varied and the correct-class angle θ_{y_i} is fixed to $40°$. Best viewed in color.

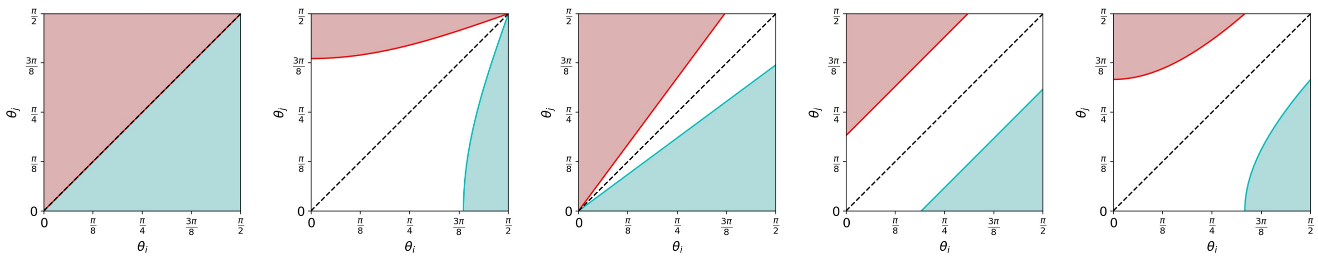
class. Fig. 15 visualizes binary decision margins in angular space for the four margin-based loss formulations studied here, with parameters set to optimal tuned values from Appendix B.

The leftmost Fig. 15a corresponds to the settings $m_0 = 1$, $m_1 = 1$, $m_2 = 0$, $m_3 = 0$ and only one parameter deviates from these settings in each other plot. ArcFace [3] and CosFace [27] attribute the success of their margin-based loss formulations to the additive nature of their margins, which lead to constant linear decision margins in angular space and cosine space (not pictured) respectively.

However, the clear differences between margin formulations in this low-dimensional visualization do not well characterize their behavior in practical settings with high C and d. To better approximate this, Fig. 16 translates these angular decision margins to a 3-d sphere, with prototypes defined at $(0, 0, 1)$ and $(1, 0, 0)$. We observe that when prototypes are sufficiently distant, the decision margins are similar, with all but AmpFace resembling the linear margin of ArcFace. Further differences emerge only when prototypes are much closer or when m_1 is set to large values, resulting in discontinuous classification regions.

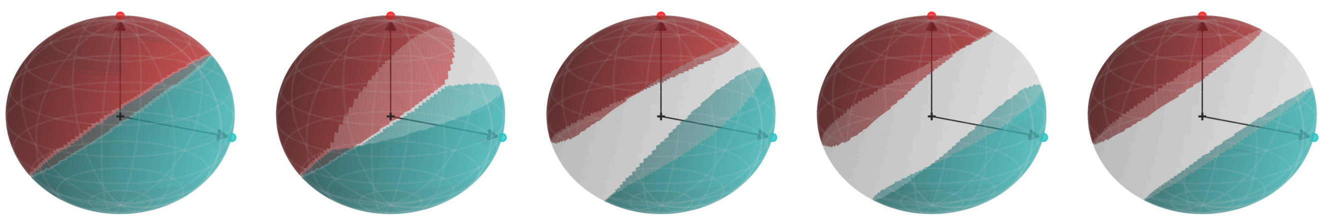
Practical face recognition problems typically feature a large number of prototypes, and their decision margins cannot be easily derived from the binary classification setting. Fig. 17 illustrates decision margins with the same parameter regime for 3-way classification, introducing an additional prototype defined at $(0, -1, 0)$. Decision margins appear more similar in this setting, with AmpFace as the biggest outlier, suggesting that the location of prototypes on the hypersphere is more relevant to the decision landscape than the margin type.

Fig. 18 attempts to visualize the effect of high C and d on decision margins in 3-d. With $C = 85,742$ and $d = 512$, which results in an inter-class distance M_{inter} of $78.64°$, we sample 200 prototypes around the equatorial region of the sphere and illustrate the effect of their presence on the classification region of the prototype at $(0, 0, 1)$ under different margin settings. We observe that all such regions take the form of spherical caps surrounding the prototype, as noted in Section 3.3. There is no visible effect of linear or non-linear decision margins, as the boundary of the cap is influenced by multiple prototypes located at M_{inter} with high probability.



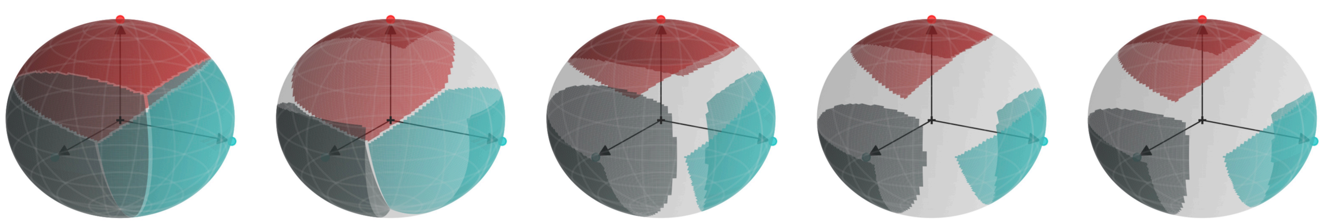
(a) No margins (b) $m_0 = 0.35$ (c) $m_1 = 1.35$ (d) $m_2 = 0.60$ (e) $m_3 = 0.50$

Figure 15: Angular decision margins (blank area) surrounding a decision boundary (dotted line) for binary classification.



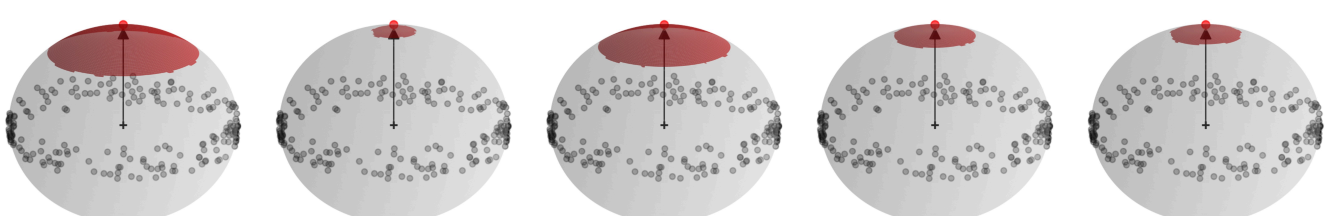
(a) No margins (b) $m_0 = 0.35$ (c) $m_1 = 1.35$ (d) $m_2 = 0.60$ (e) $m_3 = 0.50$

Figure 16: Decision margins for binary classification on a 3-d sphere with prototypes at $(0, 0, 1)$ and $(1, 0, 0)$.



(a) No margins (b) $m_0 = 0.35$ (c) $m_1 = 1.35$ (d) $m_2 = 0.60$ (e) $m_3 = 0.50$

Figure 17: Decision margins for 3-way classification on a 3-d sphere with prototypes at $(0, 0, 1)$, $(1, 0, 0)$ and $(0, -1, 0)$.



(a) No margins (b) $m_0 = 0.35$ (c) $m_1 = 1.35$ (d) $m_2 = 0.60$ (e) $m_3 = 0.50$

Figure 18: Approximation on a 3-d sphere of positive region for a prototype under our experimental conditions ($C = 85,742$, $d = 512$), where the expected inter-class distance is 80°. The positive region forms a hyperspherical cap regardless of the margin type or value.