On the Optimization Landscape of Neural Collapse Under MSE loss: Global Optimality With Uncontrained Features

Jinxin Zhou[+], Xiao Li[*], Tianyu Ding[#], Chong You[$], Qing Qu[*], Zhihui Zhu[+]

[+]University of Denver

[*]University of Michigan, Ann Arbor

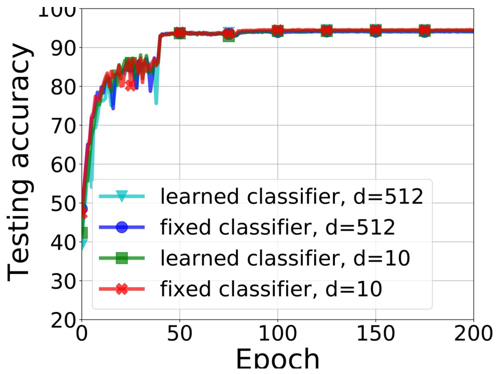
[#]Microsoft

[$]Google Research

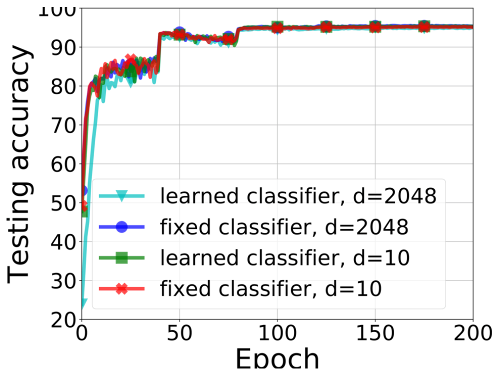
Previous work

- In machine learning community, MSE is often not suggested for classification problems since it does not strongly penalize misclassification, does this hold for deep learning?

- No, [Hui & Belkin] finds that the (rescaled) mean-square-error (MSE) loss performs comparably as cross-entropy (CE) loss across a range of tasks: NLP, Speech Recognition and computer vision.



Testing Accuracy (MSE)

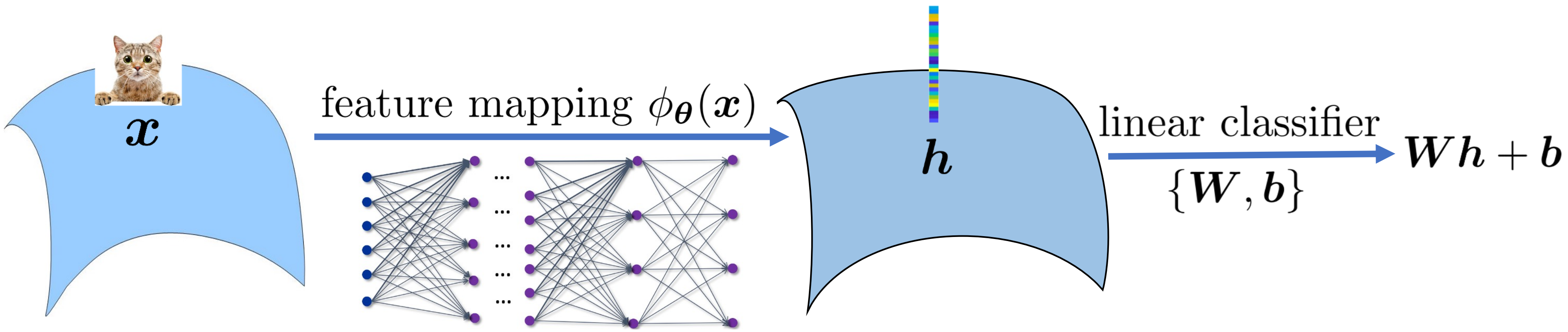


Testing Accuracy(CE)

1. Can we understand learned features and classifier?
2. Can we use them to explain why MSE is successful in training deep networks for classification?

Hui, Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. ICML, 2021.

Simplification: Unconstrained Features



$$\min_{\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b}} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}_{\mathrm{MSE}}\left(\boldsymbol{W}\left(\phi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{k,i}\right)\right) + \boldsymbol{b}, \boldsymbol{y}_k\right) + \lambda \left\|\left(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{b}\right)\right\|_F^2$$

- This training problem is highly nonconvex!
- Treat $\boldsymbol{h}_{k,i} = \phi_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i})$ as a **free** optimization variable
- Called unconstrained features model [Mixon et al'20, Fang et al'21, E et al'20, Lu et al'22]

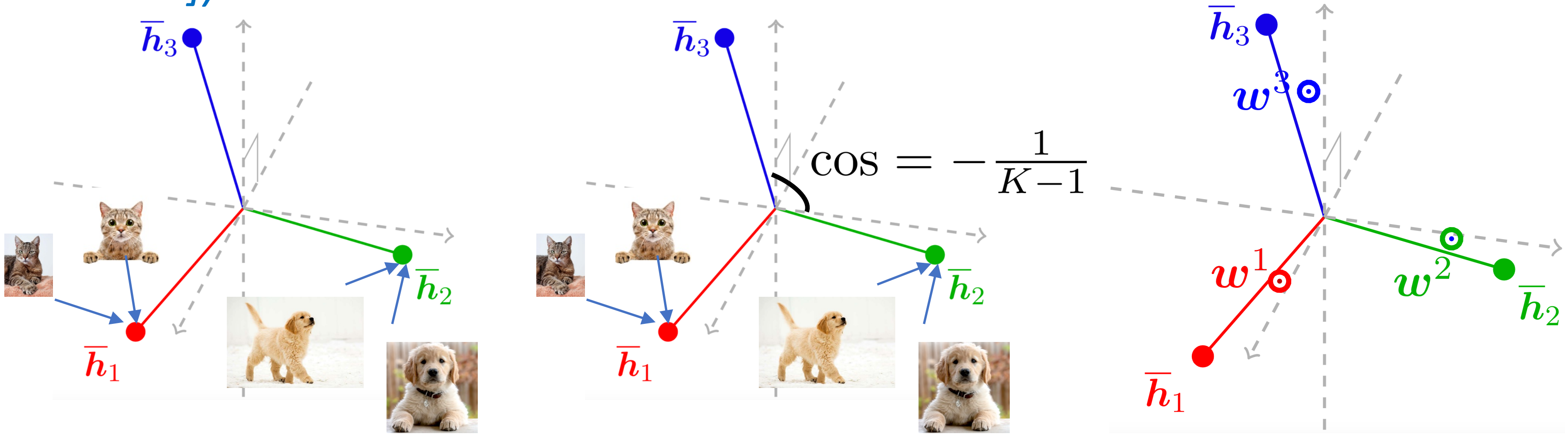
$$\min_{\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{b}} \frac{1}{Kn} \sum_{k=1}^{K} \sum_{i=1}^{n} \mathcal{L}_{\mathrm{MSE}}\left(\boldsymbol{W}\boldsymbol{h}_{k,i} + \boldsymbol{b}, \boldsymbol{y}_k\right) + \lambda \left\|\boldsymbol{H}, \boldsymbol{W}, \boldsymbol{b}\right\|_F^2$$

$$\text{where } \boldsymbol{H} := \begin{bmatrix} \boldsymbol{h}_{1,1} & \cdots & \boldsymbol{h}_{K,n} \end{bmatrix}$$

Summary of contributions

1. Characterization of global solutions with unconstrained features ($d \geq K$-1)
   - *All the global solutions satisfy the NC properties with certain choice of regularization parameters*

(Note that MSE learns identical NC features as CE loss in the work by [Papyan et al' 2020])



NC1

Within-Class Variability Collapse

$$\boldsymbol{h}_{k,i} \rightarrow \bar{\boldsymbol{h}}_k$$

NC2

Convergence to Simplex ETF

$$\frac{\langle \bar{\boldsymbol{h}}_k, \bar{\boldsymbol{h}}_{k'} \rangle}{\|\bar{\boldsymbol{h}}_k\| \|\bar{\boldsymbol{h}}_{k'}\|} \rightarrow \begin{cases} 1, & k = k' \\ -\frac{1}{K-1}, & k \neq k' \end{cases}$$

NC3

Convergence to Self-Duality

$$\frac{\boldsymbol{w}^k}{\|\boldsymbol{w}^k\|} \rightarrow \frac{\bar{\boldsymbol{h}}_k}{\|\bar{\boldsymbol{h}}_k\|},$$

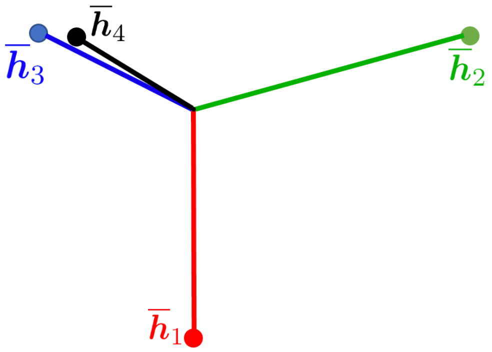
Papyan, Han, Donoho, Prevalence of neural collapse during the terminal phase of deep learning training, PNAS, 2020.

Summary of contributions

2. Characterization of global solutions with unconstrained features($d < K$-1)
   - *All global solutions satisfy the d-dim projection of K-dim Simplex ETF*

Distribution of class-mean features

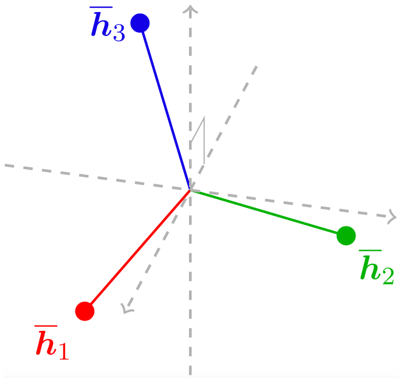
feature dim $d <$ #class $K - 1$



Class-means of features have

- *different lengths*
- *different angles with each other*
- *angles could be* $< 90^o$

feature dim $d \geq$ #class $K - 1$

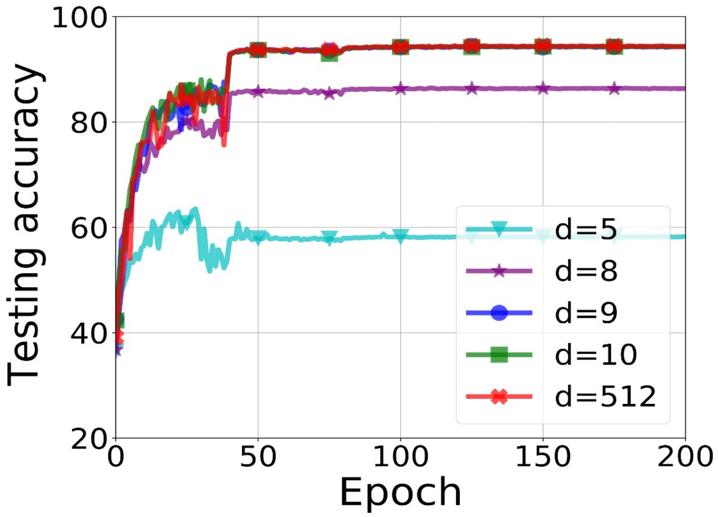


Class-means of features form an ETF:

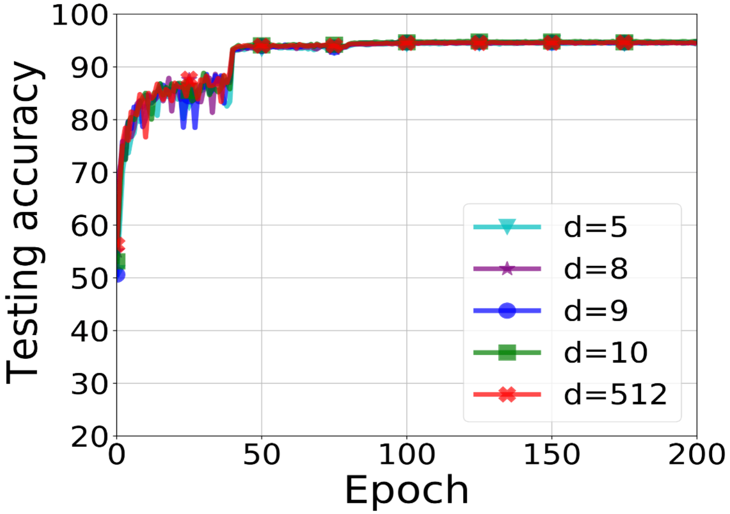
- *same lengths*
- *same angles with each other*
- *angles always* $\geq 90^o$

Experiment: choice of feature dim *d*

- ResNet18, CIFAR10, **Comparison of performance under different *d***



Testing Accuracy (MSE)



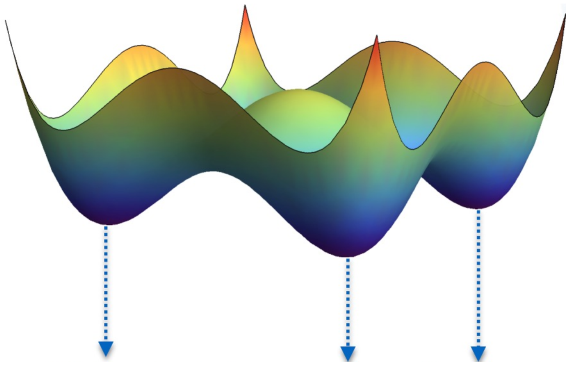
Testing Accuracy(CE)

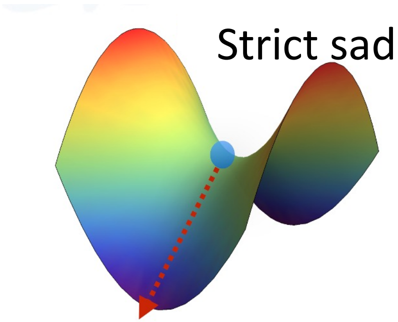
- Choosing $d \geqq K-1$ is crucial for MSE.
- Comparable performance for CE and MSE when $d \geqq K-1$
- Contrast to the phenomenon of CE when $d < k-1$.

Summary of contributions

$$\min_{\boldsymbol{H},\boldsymbol{W},\boldsymbol{b}} \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathcal{L}_{\mathrm{MSE}}\left(\boldsymbol{W}\boldsymbol{h}_{k,i}+\boldsymbol{b},\boldsymbol{y}_k\right)+\lambda\left\|\boldsymbol{H},\boldsymbol{W},\boldsymbol{b}\right\|_F^2$$

2. Landscape analysis of NC with unconstrained features
   - Benign global landscape: *deep networks always learn Neural Collapse features and classifiers* —negative curvature for non-global critical point.



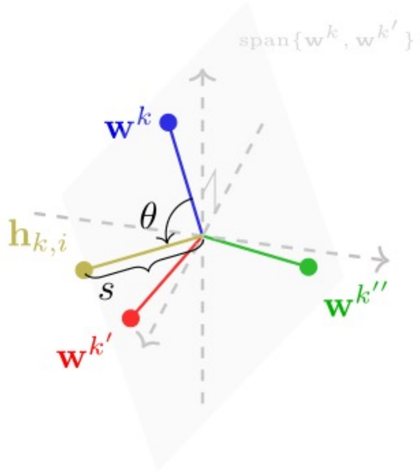


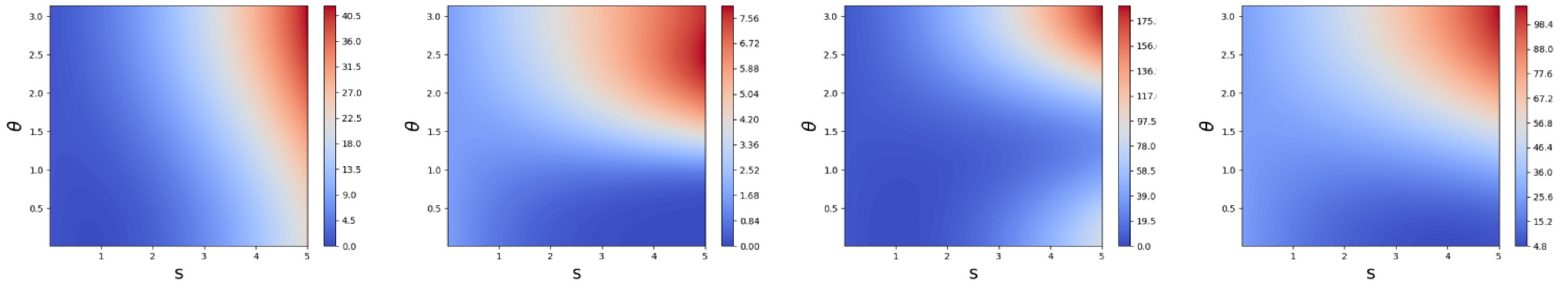
Summary of contributions

- Understanding effects of the rescaled MSE
  - Rescaled MSE leads to a "better" optimization landscape, which is steeper w.r.t. ϴ (angular distance) than w.r.t. s (length distance), like CE.

MSE : $\ell_{MSE}(\bar{\boldsymbol{y}}, \boldsymbol{y}_k) = (\bar{y}_k - 1)^2 + \sum_{j \neq k} \bar{y}_j$

Rescaled MSE: $\ell_{RMSE}(\bar{\boldsymbol{y}}, \boldsymbol{y}_k) = \alpha(\bar{y}_k - M)^2 + \sum_{j \neq k} \bar{y}_j$





(a) Vanilla MSE ($\alpha = 1, M = 1$) (b) Cross Entropy (c) Rescaled MSE ($\alpha = 5, M = 1$) (d) Rescaled MSE ($\alpha = 1, M = 5$)

Acknowledgements

Thank you for your attention!