Neural Collapse에 대해, CE와 MSE에 대해 분석

MSE에 대해 새로운 decomposition.

MSE classifier에서 deviation capture

* Introduction.

MLP, SGD, BN, CE 등은 관행적으로 쓰이고 이에 대한 분석이 중요하다.

논문에서는 classifier의 마지막 layer에 대해 이론적으로 조사. (C개의 class, N개의 sample)

classifier의 input (앞에까지 architecture의 feature $h \in R^p$)

↳ $w_c \in R^p$, $b_c \in R$

이에 대해 error는 $\text{Error} = \underset{i,c}{\text{Ave}}\ 1\{c \neq \arg\max_{c'} (\langle w_{c'}, h_{i,c} \rangle + b_{c'})\}$,

CE는 이에 대해 $\text{CE} = -\underset{i,c}{\text{Ave}} \log \frac{\exp\{\langle w_c, h_{i,c} \rangle + b_c\}}{\sum_{c'=1}^{C} \exp\{\langle w_{c'}, h_{i,c} \rangle + b_{c'}\}}$,

이전 논문에서 overparameterized된 classifier는 test에 대해 만족시키지 않고 train을 memorize

memorize 이후 계속 훈련하면 성능 향상될 수 없음.

이에 대해 Error 0을 지나, zero-CE-loss를 향해 가는 것은 Terminal Phase of Training (TPT)

TPT 동안 Neural Collapse (NC) 나타남.

- Neural Collapse

Feature Global Mean : $\mu_G = \underset{i,c}{\text{Ave}}\ h_{i,c}$

Feature Class Mean : $\mu_c = \underset{i}{\text{Ave}}\ h_{i,c}$, $c = 1, \dots, C$

Feature within-Class Covariance : $\Sigma_W = \underset{i,c}{\text{Ave}} (h_{i,c} - \mu_c)(h_{i,c} - \mu_c)^T$

Feature between-Class Covariance : $\Sigma_B = \underset{i}{\text{Ave}} (\mu_c - \mu_G)(\mu_c - \mu_G)^T$

훈련 epoch 수가 지날 때마다 네개의 limiting behavior가 있다.

NC1. Within-class variability collapse

$\Sigma_B^{\dagger}\Sigma_W \to 0,$ †는 Moore-Penrose pseudoinverse

NC2. Convergence to Simplex ETF

$$\frac{\langle \mu_c - \mu_G, \mu_{c'} - \mu_G \rangle}{\|\mu_c - \mu_G\|_2 \|\mu_{c'} - \mu_G\|_2} \to \begin{cases} 1, & c = c' \\ \frac{-1}{C-1}, & c \neq c' \end{cases}$$

$$\|\mu_c - \mu_G\|_2 - \|\mu_{c'} - \mu_G\|_2 \to 0 \quad \forall c \neq c'$$

NC3. Convergence to self-duality

$$\frac{w_c}{\|w_c\|_2} - \frac{\mu_c - \mu_G}{\|\mu_c - \mu_G\|_2} \to 0$$

NC4. Simplification to nearest class center

$$\arg\max_{c'} \langle w_{c'}, h \rangle + b_{c'} \to \arg\min_{c'} \|h - \mu_{c'}\|_2$$

NC2는 ETF라는 단순한 기하학적 구조로 수렴을 capture

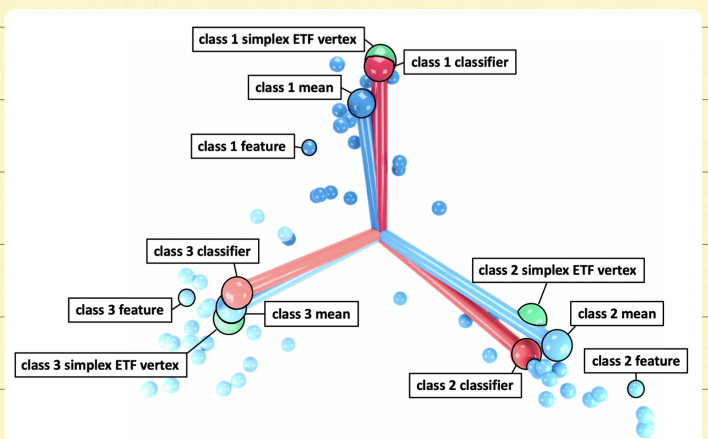
ETF는 각이나 같고 최대로 벌려진 pair-wise angle를 갖는 vector $\{v_c\}_{c=1}^{C}$의 collection이다.

P>C인 setting에서 maximal angle은

$$\frac{\langle v_c, v_{c'} \rangle}{\|v_c\|_2 \|v_{c'}\|_2} = \begin{cases} 1, & \text{for } c = c' \\ -\frac{1}{C-1}, & \text{for } c \neq c' \end{cases}$$

여기서 ETF는 simplex ETF로 불림.

C가 충분하면 orthogonal의 성질.



- DeepNet classification with MSE Loss

분류 net은 보통 CE로 훈련되지만 몇몇은 MSE로 훈련함.

$$\mathcal{L}(\boldsymbol{W},\boldsymbol{b},\boldsymbol{H}) = \frac{1}{2}\operatorname*{Ave}_{i,c}\|\boldsymbol{W}\boldsymbol{h}_{i,c}+\boldsymbol{b}-\boldsymbol{y}_{i,c}\|_2^2+\frac{\lambda}{2}(\|\boldsymbol{W}\|_F^2+\|\boldsymbol{b}\|_2^2)$$
$$= \frac{1}{2CN}\|\boldsymbol{W}\boldsymbol{H}+\boldsymbol{b}\mathbb{1}_{CN}^\top-\boldsymbol{Y}\|_F^2+\frac{\lambda}{2}(\|\boldsymbol{W}\|_F^2+\|\boldsymbol{b}\|_2^2),$$

$H \in R^{P\times CN}$, $Y \in R^{C\times CN}$, $W \in R^{C\times P}$, $b \in R^C$

- Contribution.

1. MSE에 대해서 새로운 decomposition. $\mathcal{L} = \mathcal{L}_{NC1} + \mathcal{L}_{NC2/3} + \mathcal{L}_{LS}^{\perp}$
   - classifier W는 MSE에 대한 optimal classifier W_{LS}라고 가정
2. real dataset에서 decomposition의 정확한 측정과 $\mathcal{L}_{LS}^{\perp}$가 훈련 중에 무시될 수 없음을 보임
   - central path: $\mathcal{L}_{LS}^{\perp}$ = zero
3. central path에 대해 invariant property 임을 밝힘
   - $X = \Sigma_W^{-\frac{1}{2}} H$ (renormalized features)
4. central에 대해 renormalized feature의 gradient flow를 연구하고

   NC를 의미하는 closed-form dynamics 도출

   dynamics는 init에서 renormalized feature class-mean의 singular value decomposition 측면에서 명시적

* Decomposition of MSE Loss

MSE에 대해 NC 분석을 위한 decomposition

W에 b를 합치면, $\widetilde{W} = [W, b] \in R^{C \times (P+1)}$, $\tilde{h}_{i,c} = [h_{i,c}; 1] \in R^{P+1}$

$$\mathcal{L}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) = \frac{1}{2} \operatorname*{Ave}_{i,c} \|\widetilde{\boldsymbol{W}} \tilde{\boldsymbol{h}}_{i,c} - \boldsymbol{y}_{i,c}\|_2^2 + \frac{\lambda}{2} \|\widetilde{\boldsymbol{W}}\|_F^2.$$

$\tilde{h}$로 확장된 NC로 보면,

$$\widetilde{\boldsymbol{\Sigma}}_T = \operatorname*{Ave}_{i,c} (\tilde{\boldsymbol{h}}_{i,c} - \tilde{\boldsymbol{\mu}}_G)(\tilde{\boldsymbol{h}}_{i,c} - \tilde{\boldsymbol{\mu}}_G)^\top \in \mathbb{R}^{(P+1)\times(P+1)}$$

$$\widetilde{\boldsymbol{M}} = [\tilde{\boldsymbol{\mu}}_1, \ldots, \tilde{\boldsymbol{\mu}}_C] \in \mathbb{R}^{(P+1)\times C}.$$

Proposition 1.

고정된 $\tilde{H}$에서 MSE에 대한 optimal classifier는

$$\widetilde{\boldsymbol{W}}_{LS} = \frac{1}{C} \widetilde{\boldsymbol{M}}^\top (\widetilde{\boldsymbol{\Sigma}}_T + \tilde{\boldsymbol{\mu}}_G \tilde{\boldsymbol{\mu}}_G^\top + \lambda \boldsymbol{I})^{-1},$$

↳ $\tilde{H}$에만 dependent.

Theorem 1. - Appendix B

MSE Loss $\mathcal{L}(\tilde{W}, \tilde{H}) = \mathcal{L}_{LS}(\tilde{H}) + \mathcal{L}_{LS}^{\perp}(\tilde{W}, \tilde{H})$로 분해 가능.

$$\mathcal{L}_{LS}(\widetilde{\boldsymbol{H}}) = \frac{1}{2} \operatorname*{Ave}_{i,c} \|\widetilde{\boldsymbol{W}}_{LS} \tilde{\boldsymbol{h}}_{i,c} - \boldsymbol{y}_{i,c}\|_2^2 + \frac{\lambda}{2} \|\widetilde{\boldsymbol{W}}_{LS}\|_F^2,$$

→ $\tilde{W}$에 independent

$$\mathcal{L}_{LS}^{\perp}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) = \frac{1}{2} \operatorname{tr}\left\{ (\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{W}}_{LS}) \left(\widetilde{\boldsymbol{\Sigma}}_T + \tilde{\boldsymbol{\mu}}_G \tilde{\boldsymbol{\mu}}_G^\top + \lambda \boldsymbol{I} \right) (\widetilde{\boldsymbol{W}} - \widetilde{\boldsymbol{W}}_{LS})^\top \right\}.$$

직관적으로 $\tilde{W}_{LS}$는 $\tilde{W}$ 보단 $\tilde{H}$에 연관이 있다.

$\mathcal{L}_{LS}^{\perp}$는 non-negative고, $\tilde{W} = \tilde{W}_{LS}$일 때만 0 이다.

∴ $\mathcal{L}_{LS}^{\perp}$는 $\tilde{W}$와 $\tilde{W}_{LS}$ 간의 distance로 볼 수 있음

$\mathcal{L}_{LS}$는 network의 행동을 capture

$\mathcal{L}_{LS}$를 추가적으로 decompose. → NC 1,2,3을 보기 위해.

Theorem 2 - Appendix C.

Theorem 1의 $\mathcal{L}_{LS}(\tilde{H})$는 그대로 decompose 가능.

$\mathcal{L}_{LS}(\tilde{H}) = \mathcal{L}_{NC1}(\tilde{H}) + \mathcal{L}_{NC2/3}(\tilde{H})$

$$\mathcal{L}_{NC1}(\widetilde{\boldsymbol{H}}) = \frac{1}{2}\operatorname{tr}\left\{\widetilde{\boldsymbol{W}}_{LS}\left[\widetilde{\boldsymbol{\Sigma}}_W + \lambda \boldsymbol{I}\right]\widetilde{\boldsymbol{W}}_{LS}^\top\right\},$$
$$\mathcal{L}_{NC2/3}(\widetilde{\boldsymbol{H}}) = \frac{1}{2C}\|\widetilde{\boldsymbol{W}}_{LS}\widetilde{\boldsymbol{M}} - \boldsymbol{I}\|_F^2.$$

$\mathcal{L}_{NC2/3}$이 class-mean과 MSE-optimal classifier에 관한 합수

Minimizing $\mathcal{L}_{NC2/3}$은 class-mean과 classifier가 동일한 Simplex ETF matrix로 이동한다.

NC1은 mean과는 independent.

따라서 classifier가 potentially large ETF matrix에 수렴하지만 Σ_W를 0으로 유도해, $\mathcal{L}_{NC1}$을 줄일 수 있다.

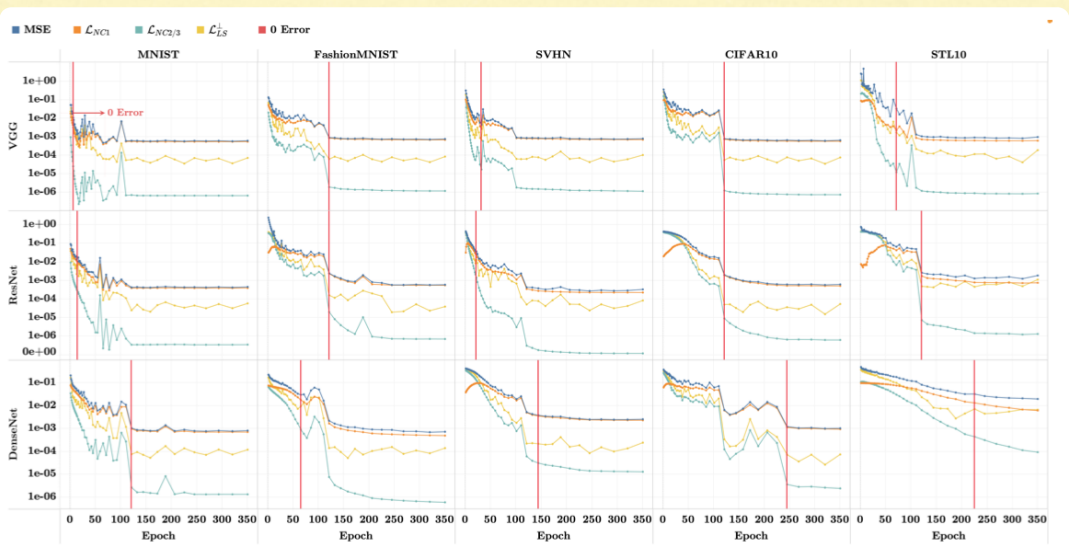


Figure 2: *Decomposition of MSE loss:* Each array column shows a benchmark image classification dataset while each row shows a canonical deep net architecture trained with MSE loss. The red vertical line indicates the epoch at which zero training error was achieved. In each array cell, we plot terms of the MSE loss decomposition $\mathcal{L}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) = \mathcal{L}_{\text{NC1}}(\widetilde{\boldsymbol{H}}) + \mathcal{L}_{\text{NC2/3}}(\widetilde{\boldsymbol{H}}) + \mathcal{L}_{\text{LS}}^{\perp}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ from Section 2. Starting from an early epoch in training, $\mathcal{L}_{\text{LS}}^{\perp}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}})$ becomes negligible compared to the dominant term, $\mathcal{L}_{\text{NC1}}(\widetilde{\boldsymbol{H}})$, implying $\mathcal{L}_{\text{LS}}^{\perp}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{H}}) \ll \mathcal{L}_{\text{LS}}(\widetilde{\boldsymbol{H}}) = \mathcal{L}_{\text{NC1}}(\widetilde{\boldsymbol{H}}) + \mathcal{L}_{\text{NC2/3}}(\widetilde{\boldsymbol{H}})$, i.e. the features and classifiers are *effectively on the central path* during TPT. Note that $\mathcal{L}_{\text{NC2/3}}(\widetilde{\boldsymbol{H}})$ diminishes the fastest among all the terms: Intuitively, this shows that the network primarily focuses on distributing the feature class-means into a "uniform" Simplex ETF configuration **(NC1)-(NC2)** early on and, from there, compresses the activations towards their class-means, i.e. **(NC1)** , as much as possible. Further experimental details are in Appendix A. Outlier behavior is discussed in Appendix A.7.

→ $\mathcal{L}_{\text{aug}}$는 $\mathcal{L}_{\text{LS}}$에 비해 무시할만한 수준

즉 $\mathcal{L}(\tilde{W}, \tilde{H}) \simeq \mathcal{L}_{\text{LS}}(\tilde{H})$

↓ central path

$$\mathcal{P} = \left\{ (\widetilde{\boldsymbol{W}}_{\text{LS}}(\widetilde{\boldsymbol{H}}), \widetilde{\boldsymbol{H}}) \mid \widetilde{\boldsymbol{H}} \in \mathbb{R}^{(P+1)\times CN} \right\}$$

$\tilde{W}_{\text{LS}}$가 $\tilde{H}$에만 dependent 하기 때문에

$\tilde{W}$가 central path에 있다는건 $\tilde{W}_{\text{LS}}$라 같다는것.

* Exact Closed-Form Analysis on Central Path.