A Proofs of Section 3.1

A.1 Proof of Theorem 2 and corresponding results

Recall an ETF defined by

$$M = \sqrt{\frac{K}{K-1}} P \left(I_K - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^T \right) = \sqrt{\frac{K}{K-1}} \left(P - \frac{1}{K} \sum_{k=1}^{K} P_k \mathbf{1}_K^T \right),$$

where $P = [P_1, \cdots, P_K] \in \mathbb{R}^{p \times K}$ is a partial orthogonal matrix with $P^T P = I_K$. Rewrite $M = [M_1, \cdots, M_K]$. Let the label $y = (y_1, \cdots, y_K)^T \in \{0, 1\}^K$ be represented by the one-hot encoding, that is, $y_k = 1$ and $y_j = 0$ for $j \neq k$ if y belongs to the k-th class.

Definition 6 (Classification problem under Neural Collapse)**.** *Let there be K classes. The distribution $\mathbb{P}[x = M_k | y_k = 1] = 1$ for $k = 1, ..., K$.*

Proof of Theorem 2. Let $W = [W_1, \cdots, W_K]^T \in \mathbb{R}^{K \times p}$. Consider the output function $f_W(x) = Wx \in \mathbb{R}^K$. Suppost that $y_k = 1$. Then, the cross-entropy loss is defined by

$$\ell(f_W(x), y) = -\log \left(\frac{e^{W_k^T x}}{\sum_{k'=1}^{K} e^{W_{k'}^T x}} \right).$$

The corresponding empirical risk is

$$R_n(M, W) = \sum_{k=1}^{K} -n_k \log \left(\frac{e^{W_k^T M_k}}{\sum_{k'=1}^{K} e^{W_{k'}^T M_k}} \right).$$

Note that

$$\nabla_W \ell(f_W(x), y) = (\mathrm{SoftMax}(f_W(x)) - y)\, x^T,$$

where $\mathrm{SoftMax} : \mathbb{R}^K \to \mathbb{R}^K$ is the SoftMax function defined by

$$\mathrm{SoftMax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \qquad \text{for all } z \in \mathbb{R}^K.$$

We obtain

$$\nabla_W R_n(M, W) = \sum_{k=1}^{K} n_k \left(\mathrm{SoftMax}(f_W(M_k)) - y^k \right) M_k^T,$$

where y^k is the label of the k-th class. For zero initialization, we have

$$\mathrm{SoftMax}(f_{\mathbf{0}}(M_k)) = \frac{1}{K} \mathbf{1}_K$$

and

$$\nabla_W (R_n(M, W)) \Big|_{W=\mathbf{0}} = \sum_{k=1}^{K} n_k \left(\frac{1}{K} \mathbf{1}_K - y^k \right) M_k^T. \tag{2}$$

Now we consider one step NoisyGD from 0 with learning rate $\eta = 1$:

$$\widehat{W} = -\sum_{k=1}^{K} n_k \left(\frac{1}{K} \mathbf{1}_K - y^k \right) M_k^T + \Xi,$$

where $\Xi \in \mathbb{R}^{K \times p}$ with Ξ_{ij} drawn independently from a normal distribution $\mathcal{N}(0, \sigma^2)$.

Consider $x = M_k$. It holds

$$f_{\widehat{W}}(x) = \widehat{W} M_k = -\sum_{k'=1}^{K} n_{k'} \left(\frac{1}{K} \mathbf{1}_K - y^{k'} \right) M_{k'}^T M_k + \Xi M_k.$$

Since

$$\Xi M_k \sim \mathcal{N}\left(0, \sigma^2 \|M_k\|_2^2 I_K\right) \qquad \text{and} \qquad \|M_k\|_2^2 = 1$$

we have

$$\widehat{W} M_k \sim \mathcal{N}\left(\boldsymbol{\mu}_{n,K}, \sigma^2 I_K\right),$$

where $\boldsymbol{\mu}_{n,K} = -\sum_{k'=1}^{K} n_{k'} \left(\frac{1}{K}\mathbf{1}_K - y^{k'}\right) M_{k'}^T M_k$. Note that

$$M_{k'}^T M_k = \frac{K}{K-1}\left(\delta_{k,k'} - \frac{1}{K}\right).$$

We obtain

$$(\boldsymbol{\mu}_{n,K})_j = \begin{cases} n/K, & j = k, \\ -\frac{n(K-2)}{K^2(K-1)}, & j \neq k, \end{cases}$$

for $n_{k'} = n/K$ (balanced data). By the union bound, the mis-classification error is

$$(K-1)\mathbb{P}\left[\mathcal{N}(n/K, \sigma^2) < \mathcal{N}(-\frac{n(K-2)}{K^2(K-1)}, \sigma^2)\right] = (K-1)\Phi\left(-\frac{n}{K\sigma}\left(1 + \frac{K-2}{K(K-1)}\right)\right)$$

□

Proof sketches of the insights. Note that in Equation equation 2, the gradient is a linear function of the feature map thanks to the zero-initialization while for least-squares loss, one can derive a similar gradient as Equation equation 2. Thus, the proof can be extended to the least squares loss directly. Moreover, by replacing n_k with $n_k\eta$ in equation 2, one can extend the results to any η.

A.2 Proof of Theorem 3

Recall the re-parameterization for $K = 2$. Precisely, an equivalent neural collapse case gives $M = [-e_1, e_1]$ with $e_1 = [1, 0, \ldots, 0]^T$. Furthermore, we consider the re-parameterization with $y \in \{-1, 1\}$, $\theta \in \mathbb{R}^p$ and the decision rule being $\hat{y} = \text{sign}(\theta^T x)$. Then, the logistic loss is $\log(1 + e^{-y \cdot \theta^T x})$.

Proof of Theorem 3. According to the re-parameterization, for the class imbalanced case, we have

$$\hat{\theta} = -\eta\left(\frac{n}{2} \cdot 0.5 \cdot (-\begin{bmatrix} -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}) + \frac{n}{2} \cdot 0.5 \cdot \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \mathcal{N}(0, \frac{G^2}{2\rho} I_p)\right) = -\eta\left(\begin{bmatrix} n/2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \mathcal{N}(0, \frac{G^2}{2\rho} I_p)\right).$$

The rest of the proof is similar to that of Theorem 2.

For the class-imbalanced case, assume that we have αn data points have with label $+1$ while the rest $(1-\alpha)n$ points have label -1. Then, the gradient is

$$\hat{\theta} = -\eta\left(\frac{n\alpha}{2} \cdot \cdot (-\begin{bmatrix} -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}) + \frac{n(1-\alpha)}{2} \cdot \cdot \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \mathcal{N}(0, \frac{G^2}{2\rho} I_p)\right) = -\eta\left(\begin{bmatrix} n/2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \mathcal{N}(0, \frac{G^2}{2\rho} I_p)\right).$$

Thus, the same conclusion holds. □

A.3 Proof of Theorem 4

In this section, we consider a broad pre-training on a gigantic dataset with K_0 classes. The downstream task is a K-class classification problem with $K \le K_0$. Let $P = [P_1, \cdots, P_{K_0}] \in \mathbb{R}^{p\times K_0}$ be a partial orthogonal matrix with $P^T P = I_{K_0}$. Let

$$M_0 = \sqrt{\frac{K_0}{K_0-1}} P\left(I_{K_0} - \frac{1}{K_0}\mathbf{1}_{K_0}\mathbf{1}_{K_0}^T\right) = \sqrt{\frac{K_0}{K_0-1}}\left(P - \frac{1}{K_0}\sum_{k=1}^{K_0} P_k \mathbf{1}_{K_0}^T\right).$$

Denote $M = [M_1, \cdots, M_K]$ with each M_k being a column of M_0. Note that

$$M_{k'}^T M_k = \frac{K_0}{K_0-1}\left(\delta_{k,k'} - \frac{1}{K_0}\right).$$

We have

$$\boldsymbol{\mu}_{n,K} := -\sum_{k'=1}^{K} n_{k'}\left(\frac{1}{K}\mathbf{1}_K - y^{k'}\right) M_{k'}^T M_k.$$

For $j \neq k$, we have

$$(\mu_{n,K})_j = -\frac{n}{K}\left[\frac{1}{K} + \frac{K-1}{K(K_0-1)} - \frac{K-2}{K(K_0-1)}\right] = -\frac{n(K_0-2)}{K^2(K_0-1)}.$$

For $j = k$, it holds

$$(\mu_{n,K})_j = -\frac{n}{K}\left[\frac{1}{K} - 1 - \frac{K-1}{K(K_0-1)}\right] = \frac{n(K-1)K_0}{K^2(K_0-1)}.$$

By the union bound, the mis-classification error is

$$(K-1)\mathbb{P}\left[\mathcal{N}((\mu_{n,K})_k, \sigma^2) < \mathcal{N}((\mu_{n,K})_1, \sigma^2)\right] = (K-1)\Phi\left(\frac{nC_{K,K_0}}{\sigma}\right)$$

with $C_{K,K_0} = \frac{1}{K}\left[\frac{K\cdot K_0 - 2}{K^2(K_0-1)}\right]$.

B Results for purturbing the testing data

B.1 Fixed perturbation

Recall that the output of DP-GD has the form $\widehat{\theta} = \mathcal{N}(-\frac{\eta n}{2}, \sigma^2)$. One has

$$\hat{\theta}^T(e+v) = \frac{n}{2} + \mathcal{N}(0, \frac{G^2(p\epsilon^2+1)}{2\rho}).$$

The sample complexity can be derived similarly as previous sections, which is dimension dependent.

B.2 Random Perturbation

Let's say in prediction time, the input data point can be perturbed by a small value in ℓ_∞. If we allow the perturbation to be adversarial chosen, then there exits v satisfying $\|v\|_\infty \le \beta$ such that

$$\hat{\theta}^T(x+v) = \frac{n}{2} + \frac{G}{\sqrt{2\rho}}Z_1 - \sum_{i=1}^{p}|Z_i|\frac{G\beta}{\sqrt{2\rho}}$$

where $Z_1, ..., Z_n \sim \mathcal{N}(0,1)$ i.i.d. Note that the additional term scales as $O(p\frac{G\beta}{\sqrt{\rho}})$, which can alter the prediction if $p \asymp n$ even if ρ is a constant (weak privacy).

The number of data points needed to achieve $1-\delta$ robust classification under neural collapse is $O\left(\frac{G\max\{p\epsilon, 1\}\sqrt{\log(1/\delta)}}{\sqrt{2\rho}}\right)$.

C Results for perturbing the training data

C.1 Fixed perturbation

Without loss of generality, we assume $0 < \alpha < 1/2$ Consider the class imbalanced case with $n_{-1} = \alpha n$ and $n_{+1} = (1-\alpha)n$. The gradient for $\theta_0 = 0$ is

$$\nabla \mathcal{L}(\theta_0) = \alpha n \cdot 0.5 \cdot -(-e_1 + v) + (1-\alpha) n \cdot 0.5 \cdot (e_1 + v) = \frac{n}{2} e_1 + \frac{(1-2\alpha)n}{2} v.$$

Thus, the output is

$$\widehat{\theta} = -\eta \left(\frac{n}{2} e_1 + \frac{(1-2\alpha)n}{2} v + \mathcal{N}(0, \sigma^2) \right)$$

The sensitivity is $G = \sqrt{1 + \|v\|_2}$ and σ^2 is taken to be $G^2/2\rho$ to achieve ρ-zCDP. Moreover, we have

$$\widehat{\theta}^T e_1 = -\frac{n}{2} - \frac{(1-2\alpha)n}{2} v_1 + \mathcal{N}(0, \sigma^2).$$

Thus, the mis-classification error is

$$\mathbb{P}[\widehat{\theta} e_1 > 0] = \Phi \left(\frac{n\left[1 - (1-2\alpha)\, v_1\right]}{2\sigma} \right) \leq e^{-\frac{n^2(1-\beta+2\alpha\beta)^2 \rho}{4G^2}}.$$

As a result, the sample complexity to achieve $1 - \gamma$ accuracy is

$$n = O\left(\sqrt{\frac{4G^2 \log \frac{1}{\delta}}{(1-\beta+2\beta\alpha)^2 \cdot \rho}} \right)$$

The sensitivity $G = \sqrt{1 + \epsilon^2 p}$ here is dimension-dependent.

C.2 Random perturbation

Now we consider the random perturbation. Denote $\{v_i\}_{i=1}^n \subseteq \mathbb{R}^p$ a sequence of i.i.d. copies of a random vector v. Consider the binary classification problem with training set $\{(x_i, y_i)\}_{i=1}^n$. Here $x_i = e_1 + v_i$ if $y_i = 1$ and $x_i = -e_1 + v_i$ if $y_i = -1$. Then, the loss function is $\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n \log\left(1 + e^{-y_i \theta^T x_i}\right)$. The one-step iterate of DP-GD from 0 outputs

$$\widehat{\theta} = -\eta \sum_{i=1}^n (-y_i x_i) + \mathcal{N}(0, \sigma^2 I_p)$$

with $\sigma^2 = G^2/2\rho$ and $G = \sup_{v_i} \sqrt{1 + \|v_i\|^2}$ Assume that v_i is symmetric, that is $y_i v_i$ has the same distribution as $-y_i v_i$. Then, it holds

$$\sum_{i=1}^n y_i x_i = n e_1 + \sum_{i=1}^n v_i =: \mu_n.$$

The mis-classification error is now given by

$$\mathbb{P}[\widehat{\theta}^T e_1 < 0] = \mathbb{P}[\mathcal{N}(\mu_n^T e_1, \sigma^2) < 0].$$

Assume that $\|v_i\|_\infty = \epsilon < 1$. Then, we have $\mu_n^T e_1 \geq n - \epsilon n$ and the sample complexity is $O\left(\sqrt{\frac{4G^2 \log(1/\delta)}{(1-\epsilon)^2 \rho}} \right)$. with $G = \sqrt{1 + \epsilon^2 p}$.

D REMEDY TO NON-ROBUSTNESS

D.1 DETAILS OF THE NORMALIZATION

Consider the case where the feature is shifted by a constant offset v. The feature of the k-th class is

$$\widetilde{x}_i = x_i - \frac{1}{n}\sum_{i=1}^{n} x_i = \widetilde{M}_k = M_k + v$$

with M_k being the k-th column of the ETF M.

The offset v can be canceled out by considering the differences between the features. That is, we train with the feature $\widetilde{M}_k - \frac{1}{K}\sum_{j=1}^{K}\widetilde{M}_j$ for the k-th class. In fact, let P_k be the k-th column of P and we have

$$\begin{aligned}\widetilde{M}_k - \frac{1}{K}\sum_{j=1}^{K}\widetilde{M}_j &= M_k - \frac{1}{K}\sum_{j=1}^{K} M_j \\ &= \sqrt{\frac{K}{K-1}}\left[\left(P_k - \frac{1}{K}\sum_{i=1}^{K} P_i\right) - \frac{1}{K}\sum_{j=1}^{K}\left(P_j - \frac{1}{K}\sum_{i=1}^{K} P_i\right)\right] \\ &= \sqrt{\frac{K}{K-1}}\left(P_k - \frac{1}{K}\sum_{j=1}^{K} P_j\right) = M_k.\end{aligned}$$

D.2 PROOF OF THEOREM 5

Proof of Theorem 5. Consider the case with $K = 2$ and a projection vector $\widehat{P} = (e_1 + \Delta)$ with some perturbation $\Delta = (\Delta_1, \cdots, \Delta_p)$ such that $\|\Delta\|_\infty \le \beta_0$ for some $0 < \beta_0 \ll p$. $\widehat{P}$ can be generated by the pre-training dataset or the testing dataset. Consider training with features $\widetilde{x}_i = \widehat{P}x_i$. Then, the sensitivity of the NoisyGD is $G = \sup_v |\widehat{P}^T(e_1 + v)| = 1 + \beta + \beta|\Delta_1| + \beta(\sum_{j=1}^{p}|\Delta_j|) \le 1 + \beta(1 + \beta_0 + p\beta_0)$. The output of Noisy-GD is then given by

$$\widehat{\theta} = -\widehat{P}\cdot\left(\sum_{i=1}^{n} y_i\widetilde{x}_i\right) + \mathcal{N}(0, \sigma^2).$$

Moreover, for any testing data point $e_1 + v$, define

$$\widehat{\mu}_n = -\left(\sum_{i=1}^{n} y_i\widetilde{x}_i\right)\widehat{P}^T(e_1 + v) = (e_1 + V)^T\,\widehat{P}\widehat{P}^T(e_1 + v)$$

with $V = \frac{1}{n}\sum_{i=1}^{n} v_i =: (V_1, \cdots, V_p)$.

We now divide $\widehat{\mu}_n$ into four terms and bound each term separately.

For the first term $e_1^T\widehat{P}\widehat{P}^Te_1$, it holds

$$e_1^T\widehat{P}\widehat{P}^Te_1 = (1 + e_1^T\Delta_1)^2 \le (1 - \beta_0)^2.$$

For the second term $V^T\widehat{P}\widehat{P}^Te_1$, we have

$$V^T\widehat{P}\widehat{P}^Te_1 = V_1 + V^T\Delta$$

Note that V_1 is the average of n i.i.d. random variables bounded by β. By Hoeffding's inequality, we obtain

$$|V_1| \le \frac{\beta\log\frac{2}{\gamma}}{\sqrt{n}}, \text{ with probability at least } 1 - \gamma.$$

Similarly, with confidence $1-\gamma$, it holds

$$|V^T\Delta| \leq \frac{p\beta\beta_0 \log\frac{2}{\gamma}}{\sqrt{n}}.$$

The third term $e_1^T\widehat{P}\widehat{P}^T v$ can be bounded as

$$|e_1^T\widehat{P}\widehat{P}^T v| = (1+\Delta_1)\left(\sum_{j=1}^{p} v_i\,(1+\Delta_i)\right) \leq (1+\beta_0)\left(\beta+\beta_0\sqrt{p\log\frac{2}{\gamma}}\right),$$

where the last inequality is a result of the Hoeffding's inequality by assuming that each coordinate of v are independent of each others. Moreover, without further assumptions on the independence of each coordinate of v, we have

$$|e_1^T\widehat{P}\widehat{P}^T v| = (1+\Delta_1)\left(\sum_{j=1}^{p} v_i\,(1+\Delta_i)\right) \leq (1+\beta_0)\,(\beta+\beta_0 p)\,.$$

Using the Hoeffding's inequality again, for the last term $V^T\widehat{P}\widehat{P}^T(e_1+v)$, it holds

$$|V^T\widehat{P}\widehat{P}^T(e_1+v)| \leq \frac{(\beta+\beta_0\sqrt{p})(1+\beta+\beta_0+\beta\beta_0\sqrt{p})\log\frac{4}{\gamma}}{\sqrt{n}}$$

with confidence $1-\gamma$ if we assume that all coordinates of v are independent of each other. Without further assumptions on the independence of each coordinate of v, we have

$$|V^T\widehat{P}\widehat{P}^T(e_1+v)| \leq \frac{(\beta+\beta_0 p)(1+\beta+\beta_0+\beta\beta_0 p)\log\frac{2}{\gamma}}{\sqrt{n}}.$$

□

E Some calculations on random Initialization

E.1 Gaussian initialization without offset

For Gaussian initialization $\xi = (\xi_1, \cdots, \xi_p) \sim \mathcal{N}(0, I_p)$, we have

$$\begin{aligned}
\hat{\theta} &= \xi - \eta\left(\frac{n}{2}\cdot\frac{-e^{-\xi_1}}{1+e^{-\xi_1}}\cdot(-\begin{bmatrix}-1\\0\\\vdots\\0\end{bmatrix}) + \frac{n}{2}\cdot\frac{-e^{-\xi_1}}{1+e^{-\xi_1}}\cdot\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} + \mathcal{N}(0,\frac{G^2}{2\rho}I_p)\right)\\
&= \xi + \eta\left(\frac{e^{-\xi_1}}{1+e^{-\xi_1}}\cdot\begin{bmatrix}n\\0\\\vdots\\0\end{bmatrix} + \mathcal{N}(0,\frac{G^2}{2\rho}I_p)\right)
\end{aligned}$$

The sensitivity is $\frac{e^{-\xi_1}}{1+e^{-\xi_1}} < 1$. Consider $x = (-1, 0, \cdots, 0)^T$. We have

$$\widehat{\theta}^T x = -\xi_1 + \eta\left(-\frac{ne^{-\xi_1}}{1+e^{-\xi_1}}\right) + \mathcal{N}(0,\frac{G^2}{2\rho}) =: \mu_{\xi_1,n} + \mathcal{N}(0,\frac{G^2}{2\rho}).$$

The mis-classification error is

$$\begin{aligned}
\mathbb{P}[\widehat{\theta}^T x > 0] &= \mathbb{E}_{\xi_1\sim\mathcal{N}(0,1)}\mathbb{P}\left[\mathcal{N}\left(\mu_{\xi_1,n}, \frac{G^2}{2\rho}\right) > 0 \middle| \xi_1\right]\\
&= \mathbb{E}_{\xi_1\sim\mathcal{N}(0,1)}\left[\Phi\left(\frac{\sqrt{2\rho}\mu_{\xi_1,n}}{G}\right)\right]
\end{aligned}$$

E.2 Gaussian initialization with off-set

Denote $x_1 = -e_1 + v$ and $x_2 = e_1 + v$ with $\|v\|_\infty \leq \beta$. For the logistic loss $\ell(y, \theta^T x) = \log(1 + e^{-y\theta^T x})$, we have

$$g(\theta, y \cdot x) := \nabla_\theta \ell(y, \theta^T x) = \frac{e^{-y\theta^T x}}{1 + e^{-y\theta^T x}}(-yx).$$

Denote

$$g_1(\theta) = g(\theta, -1 \cdot x_1) = \frac{e^{\theta^T x_1}}{1 + e^{\theta^T x_1}} x_1$$

and

$$g_2(\theta) = g(\theta, 1 \cdot x_2) = \frac{e^{-\theta^T x_2}}{1 + e^{-\theta^T x_2}}(-x_2).$$

If we shift the feature by some vector v, then the loss function is

$$R_n = \frac{n}{2}\log(1 + e^{\theta^T x_1}) + \frac{n}{2}\log(1 + e^{-\theta^T x_2}).$$

And the gradient is

$$\nabla_\theta R_n(\theta) = \frac{n}{2}\left(g_1(\theta) + g_2(\theta)\right).$$

Thus, the output of one-step NoiseGD is given by

$$\widehat{\theta} = \theta_0 - \frac{\eta n}{2}\left[g_1(\theta_0) + g_2(\theta_0) + \mathcal{N}(0, \sigma^2)\right].$$

Let $\mu_\xi = \xi - \frac{\eta n}{2}\left[g_1(\xi) + g_2(\xi)\right]$. Then, we have

$$\mu_\xi^T e_1 = \xi_1 - \frac{\eta n e^{\xi^T x_1}}{2 + 2e^{\xi^T x_1}}(-1 + v_1) + \frac{\eta n e^{\xi^T x_2}}{2 + 2e^{\xi^T x_2}}(1 + v_1).$$

And the mis-classification error is

$$\mathbb{E}_\xi\left(\Phi\left(-\frac{\sqrt{2}\rho\mu_\xi^T e_1}{G}\right)\right).$$