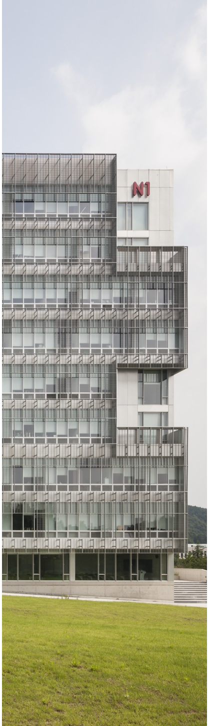

Provable Weak-to-Strong Generalization via Benign Overfitting

ICLR 2025 (Submitted)
David X. Wu, Anant Sahai

Preliminaries

Model

- $f_w \in \mathbb{R}^{d_w}$: Train on n datapoints using weak features and ground-truth labels.
- $f_{w2s} \in \mathbb{R}^{d_s}$: Train on $m \gg n$ datapoints using strong features and pseudo-labels from f_w.

Feature and Label

- i.i.d. Gaussian features $\boldsymbol{x}_i \sim N(0, \Sigma)$: $\Sigma = U\Lambda U^\top$, where $\Sigma, U, \Lambda \in \mathbb{R}^d$.
- $\boldsymbol{x}_i$: Linear transformation of the $\boldsymbol{g}_i \sim N(0, I_D)$.
- Label $y = sgn(\langle \boldsymbol{g}, \boldsymbol{v}^* \rangle)$ for unknown unit-norm direction $\boldsymbol{v}^*$.

Assumption

- (1-sparse assumption) $\boldsymbol{v}^*$ is aligned with a top eigenvector, i.e., $\boldsymbol{v}^* = \boldsymbol{e}_1$.

Bi-Level Ensemble

Bi-Level Ensemble

Bi-level ensemble parameterizes $\Lambda = \Lambda(p, q, r)$, where $p > 1$, $0 \leq r < 1$, and $0 < q < (p - r)$. The number of features(d), the number of spiked directions(s), the degree of favoring(a) all scale with the number of training points(n) as follows:

$$d = \lfloor n^p \rfloor, s = \lfloor n^r \rfloor, a = n^{-q}$$

Then $\lambda = diag(\lambda_i)_{i \in [d]}$, where

$$\lambda_j = \begin{cases} \frac{ad}{s} := \lambda_F, & 1 \leq j \leq s; \\ \frac{(1-a)d}{d-s} := \lambda_U, & \text{otherwise.} \end{cases}$$

Observations

- $\sum_j \lambda_j = d$, where and $\sum_{j \in [s]} \lambda_j = ad$ and $\sum_{j \neq [s]} \lambda_j = (1 - a)d$.
- Total features $d = n^p \gg n$, while the spiked features $s = n^r \ll n$.
- $\lambda_F = n^{p-(q+r)} >$

Weak-to-Strong Subset Ensemble

Weak-to-Strong Subset Ensemble

Let $\Lambda = \Lambda(p, q, r) \in \mathbb{R}^{d \times d}$ denote the strong eigenvalues and $\Lambda_w = \Lambda_w(p_w, q_w, r_w) \in \mathbb{R}^{d_w \times d_w}$ denote the weak eigenvalues, both drawn from the bi-level ensemble. Let U be any distinguished eigenbasis of Σ where $\boldsymbol{v}^* = \boldsymbol{e}_1$. The weak and strong features in the basis U are related as follows:

1. Strong feature: $\boldsymbol{x}_s \sim N(0, \Lambda)$, where $\Lambda = \lambda_F \boldsymbol{I}_{[s]} + \lambda_U \boldsymbol{I}_{[d] \setminus [s]}$.
2. Weak feature: There exists subsets of coordinates $S \in [s]$, $T \in [d] \setminus [s]$, such that

$$\boldsymbol{x}_w \sim N(0, \lambda_{F,w} \boldsymbol{I}_S + \lambda_{U,w} \boldsymbol{I}_T),$$

where $1 \in S$ and $|S| = s_w$.