UniGe | MaLGa

The Implicit Bias of Benign Overfitting

Ohad Shamir
Weizmann Institute of Science

Outline

Introduction

Regression

Beyond square loss

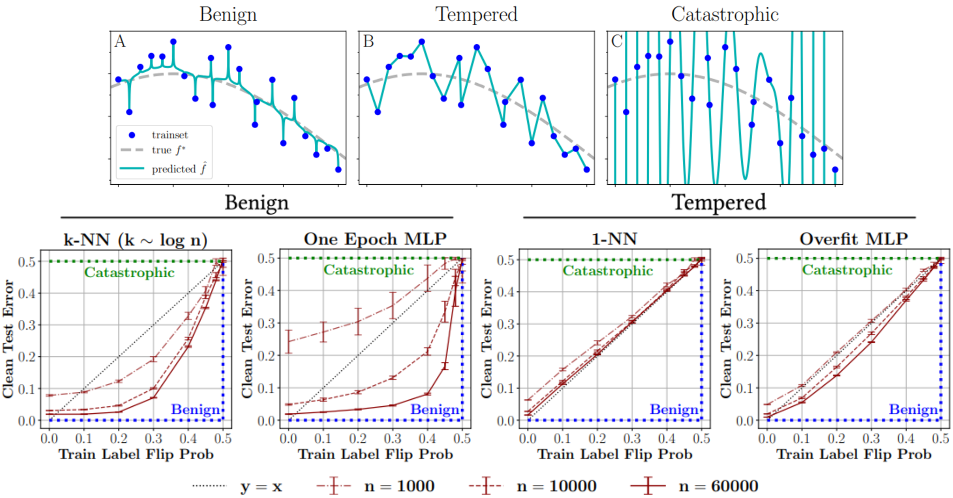
Classification

Zhang, Chiyuan et al. “Understanding deep learning requires rethinking generalization.” ArXiv abs/1611.03530 (2016): n. pag.

Zhang, Chiyuan, et al. ”Understanding deep learning (still) requires rethinking generalization.” Communications of the ACM 64.3 (2021): 107-115.

- Deep architectures are able to (and often do) memorize whole datasets...
- ...but traditional generalization bounds fail to explain their performance.
- Sparked research: tighter generalization bounds, interpolating regime, double descent, benign overfitting, implicit versus explicit regularization

Mallinar, Neil, et al. "Benign, tempered, or catastrophic: A taxonomy of overfitting." arXiv preprint arXiv:2207.06569 (2022).



Notation and setting

- $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ sequence of distributions on $\mathbb{R}^d \times Y$
  - $Y = \mathbb{R}$ for regression and $Y = \{-1, 1\}$ for classification
- If $(\mathbf{x}, y) \sim \mathcal{D}_d$, then
  - $\mathbf{x}_{|k} \in \mathbb{R}^k$ follows a (fixed) arbitrary distribution
  - $\mathbf{x}_{|d-k} \in \mathbb{R}^{d-k}$ follows a (changing) high-dimensional distribution; "junk features"
- sample sizes $\{m_d\}_{d=k+1}^{\infty}$ and IID samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m_d}$ from $\mathcal{D}_d$ for each d
- Consider linear predictors $\mathbf{w} \in \mathbb{R}^d$ evaluated as $\mathbf{x} \mapsto \mathbf{x}^T\mathbf{w}$ or $\mathbf{x} \mapsto \mathrm{sign}(\mathbf{x}^T\mathbf{w})$

Outline

Introduction

Regression

Beyond square loss

Classification

Min-norm interpolating predictors

- empirical loss $L_d(\mathbf{w}) := \frac{1}{m_d} \sum_{i=1}^{m_d} \left(\mathbf{x}_i^T \mathbf{w} - y_i\right)^2$
- true risk $R_d(\mathbf{w}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d} \left[\left(\mathbf{x}^T \mathbf{w} - y\right)^2\right]$
- Under mild conditions, iterative training methods with interpolating output converge to the min-norm predictor

$$\hat{\mathbf{w}} := \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\| : L_d(\mathbf{w}) = 0$$

UniGe | MaLGa

Benign overfitting

A sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ satisfies benign overfitting if there is a sequence of sample sizes $\{m_d\}_{d=k+1}^{\infty}$ such that

- $\Pr\{L_d(\mathbf{w}) = 0\} \to 1$ as $d \to \infty$,
- $\inf_d \inf_{\mathbf{w}\in\mathbb{R}^d} R_d(\mathbf{w}) > 0$, and
- with the min-norm predictor $\hat{\mathbf{w}}_d$ we have

$$R_d(\hat{\mathbf{w}}_d) - \inf_{\mathbf{w}\in\mathbb{R}^d} R_d(\mathbf{w}) \xrightarrow{P} 0$$

UniGe | MaLGa

The well-specified case

Bartlett, Peter L., et al. "Benign overfitting in linear regression." Proceedings of the National Academy of Sciences 117.48 (2020): 30063-30070.

Assume $\mathbb{E}[y \mid \mathbf{x}] = \mathbf{x}^T \mathbf{w}^*$ and let $\mathbf{x}$ have zero mean. Let $\Sigma = \Sigma_k \oplus \Sigma_{d-k}$ be the covariance matrix split along eigenvectors. Then, benign overfitting necessarily implies that

$$m \cdot \frac{\|\Sigma_{d-k}\|_F^2}{\mathrm{Tr}^2(\Sigma_{d-k})} \to 0$$

Lemma

If $\mathbf{z}, \mathbf{z}'$ are IID random vectors such that $\mathbb{E}[\mathbf{z}\mathbf{z}^T] = \Sigma$, then $\mathbb{E}[\mathbf{z}^T \mathbf{z}'] = \|\Sigma\|_F^2$.
Consequently, if $m \cdot \|\Sigma\|_F^2 / \mathrm{Tr}^2(\Sigma) \to 0$, then also $m \cdot \frac{\mathbb{E}[(\mathbf{z}\mathbf{z}')^2]}{\mathbb{E}^2[\|\mathbf{z}\|^2]} \to 0$.

- Even in a well-specified setting we need high-dimensional distributions producing samples with almost orthogonal directions.

Deterministic perturbation bound

Let $\{(\mathbf{x}_i, y_i\}_{i=1}^m$ be a deterministic sample of size m. Define the perturbation matrix

$$E_{ij} := \mathbf{x}_{i|d-k}^T \mathbf{x}_{j|d-k} \cdot \mathbf{1}\{i \neq j\} \in \mathbb{R}^{m \times m}.$$

Assume that the $\{\mathbf{x}_i\}_{i=1}^m$ are linearly independent,

$$\frac{\|E\|}{\min_i \|\mathbf{x}_{i|d-k}\|^2} \leq \frac{1}{2}, \quad \text{and} \quad \frac{1}{2} \leq \frac{1}{2}\lambda_{\min}\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^T}{\|\mathbf{x}_{|d-k}\|^2}\right]\right).$$

Then, the min-norm predictor $\hat{\mathrm{w}}$ exists and we have the following two bounds:

$$\left\| \hat{\mathbf{w}}_{|k} - \left(\hat{\mathbb{E}} \left[\frac{\mathbf{x}_{|k} \mathbf{x}_{|k}^T}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^{-1} \hat{\mathbb{E}} \left[\frac{y \mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\|$$

$$\leq \frac{2 \left\| \hat{\mathbb{E}} \left[\frac{y \mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2} \right] \right\|}{\lambda_{\min} \left(\hat{\mathbb{E}} \left[\frac{\mathbf{x}_{|k} \mathbf{x}_{|k}^T}{\|\mathbf{x}_{|d-k}\|^2} \right] \right)^2} \cdot \frac{1}{m} + \frac{2 \sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|k}\|^2] \cdot \hat{\mathbb{E}}[y^2]}}{\min_i \|\mathbf{x}_{i|d-k}\|^4} \cdot m \cdot \|E\|,$$

and

$$\|\hat{\mathbf{w}}_{|d-k}\| \leq \frac{\sqrt{\hat{\mathbb{E}}[\|\mathbf{x}_{|d-k}\|^2] \cdot \hat{\mathbb{E}}[y^2]}}{\min_i \|\mathbf{x}_{i|d-k}\|^2} \cdot \left(1 + \frac{2\|E\|}{\min_i \|\mathbf{x}_{i|d-k}\|^2} \right) \cdot m.$$

Moving to statistical setting

Make sure that the previous bounds are well-defined there. Let $\mathbb{E}_d$ be shorthand for $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_d}$. Assumptions:

1. The quantities $\mathbb{E}_d[\|\mathbf{x}\|^4]$, $\mathbb{E}_d[y^4]$, $\mathbb{E}_d[\|y\mathbf{x}_{|k}\|^2/\|\mathbf{x}_{|d-k}\|^4]$, and $\mathbb{E}_d[\|\mathbf{x}_{|k}\|^4/\|\mathbf{x}_{|d-k}\|^4]$ are all bounded in the supremum over d,
2. $\inf_d \lambda_{\text{min}}(\mathbb{E}_d[\mathbf{x}_{|k}\mathbf{x}_{|k}^T/\|\mathbf{x}_{|d-k}\|^2]) > 0$,
3. If $\{\mathbf{x}_i, y_i\}_{i=1}^{m_d}$ are IID samples from $\mathcal{D}_d$, the family $\{\mathbf{x}_i\}_{i=1}^{m_d}$ is linear independent with probability approaching 1,
4. $\min_i \|\mathbf{x}_{i|d-k}\| > c > 0$ with c independent of d,
5. $m_d \cdot \|E\| \xrightarrow{P} 0$, and
6. $m_d^2 \cdot \|\mathbb{E}_d[\mathbf{x}_{|d-k}\mathbf{x}_{|d-k}^T]\| \to 0$.

UniGe | MaLGa

Overfitting is generally not benign in regression

Instead of finding the true unique minimizer (benign overfitting)

$$\mathbf{w} = \mathbb{E}_d[\mathbf{x}\mathbf{x}^T]^{-1}\mathbb{E}_d[y\mathbf{x}], \quad \text{i.e.} \quad \mathbf{w}_{|k} = \left(\mathbb{E}_d[\mathbf{x}\mathbf{x}^T]^{-1}\right)_k \mathbb{E}_d[y\mathbf{x}]$$

we find asymptotic behavior of the form

$$\mathbb{E}_d\left[\left(\mathbf{x}^T\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^T\hat{\mathbf{w}}_{d|k}\right)^2\right] \xrightarrow{P} 0, \quad \left\|\hat{\mathbf{w}}_{|k} - \underbrace{\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^T}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1}\hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]}\right\| \xrightarrow{P} 0.$$

- Even in ”textbook benign” settings, benign overfitting does not generally occur in misspecified case.

Outline

Trick

Let $\ell_y(\cdot)$ non-negative loss function with unique zero for any y at $\ell_y^{-1}(0)$. Observation:

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^{m} \ell_{y_i}(\mathbf{x}_i^T\mathbf{w}) = 0$$

equals

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^{m} \left(\mathbf{x}_i^T\mathbf{w} - \ell_{y_i}^{-1}(0)\right)^2 = 0$$

Generalized linear model

Suppose $\sigma : \mathbb{R} \to \mathbb{R}$ is a function with Lipschitz inverse and $y = \sigma(\mathbf{x}_{|k}^T \mathbf{w}^*) + \xi$. Then,

$$\mathbb{E}_d\left[(\mathbf{x}^T \hat{\mathbf{w}}_d - \mathbf{x}_{|k}^T \hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{P} 0, \quad \left\| \hat{\mathbf{w}}_{|k} - \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^T}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} \hat{\mathbb{E}}\left[\frac{\sigma^{-1}(y)\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right] \right\| \xrightarrow{P} 0.$$

- Even in the well-specified case we cannot expect benign overfitting to occur.
- Example: $\|\mathbf{x}_{|d-k}\| = 1$ a.s, $\sigma(0) = 0$, $\mathbf{w}^* = 0$, but $\mathbb{E}[\mathbf{x}_{|k}] \neq 0$, then $\hat{\mathbf{w}}_{|d-k} \approx \left(\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^T]\right)^{-1} \mathbb{E}[\mathbf{x}_{|k}] \cdot \mathbb{E}[\sigma^{-1}(\xi)]$, so necessarily $\mathbb{E}[\sigma^{-1}(\xi)] = 0$ but ...

Lemma

If $\sigma^{-1}(\cdot)$ is such that $\mathbb{E}[\sigma^{-1}(\xi)] = 0$ for all zero-mean RVs ξ with support of size at most 2, then σ (and hence σ^{-1}) must be linear.

Convex losses

Assume $\ell_y(\mathbf{x}^T\mathbf{w}) = f(\mathbf{x}^T\mathbf{w} - y)$ for some non-negative f with unique root at 0 and L_d is defined in terms of this. Then for the min-norm predictor $\hat{\mathbf{w}}$,

$$\mathbb{E}_d\left[(\mathbf{x}^T\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^T\hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{P} 0, \qquad \left\| \hat{\mathbf{w}}_{|k} - \underbrace{\left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^T}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1} \hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]} \right\| \xrightarrow{P} 0.$$

- $\hat{\mathbf{w}}_{|k}$ has same asymptotic characterization as before
- we can't expect this to be optimal for the objective $\mathbb{E}[f(\mathbf{x}^T\mathbf{w} - y)]$

UniGe | MaLGa

Implicit bias towards weighted square loss problem

We found asymptotically

$$\hat{\mathbf{w}}_{d|k} \approx \left(\hat{\mathbb{E}}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^T}{\|\mathbf{x}_{|d-k}\|^2}\right]\right)^{-1}\hat{\mathbb{E}}\left[\frac{y\mathbf{x}_{|k}}{\|\mathbf{x}_{|d-k}\|^2}\right]$$

Thus, the first k coordinates actually optimize the objective function

$$\mathbb{E}_d\left[\left(\frac{\mathbf{x}^T}{\|\mathbf{x}_{|d-k}\|}\mathbf{w} - \frac{y}{\|\mathbf{x}_{i|d-k}\|}\right)^2\right]$$

- $\hat{\mathrm{w}}$ is consistent w.r.t. the statistical problem that is using the empirical version of the above as loss.

Outline

Introduction

Regression

Beyond square loss

Classification

Max-margin classifier

- Now consider RVs $(\mathbf{x}, y) \sim \mathbb{R}^d \times \{-1, 1\}$
- Same conventions for the sequence $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ of distributions as before
- Now, true risk $R_d = \Pr_{(\mathbf{x},y)}\{y\mathbf{x}^T\mathbf{w} \leq 0\}$
- Standard gradient methods run on standard convex classification loss (e.g. logistic or cross-entropy) converge in direction of the max-margin predictor

$$\hat{\mathbf{w}} := \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{w}\| : \ \min_i y_i \mathbf{x}_i^T \mathbf{w} \geq 1$$

Benign overfitting

A sequence of distributions $\{\mathcal{D}_d\}_{d=k+1}^{\infty}$ satisfies benign overfitting if there is a sequence of sample sizes $\{m_d\}_{d=k+1}^{\infty}$ such that

- $\hat{\mathbf{w}}_d$ exists with probability approaching 1,
- $\inf_d \inf_{\mathbf{w}\in\mathbb{R}^d} R_d(\mathbf{w}) > 0$, and
- $R_d(\hat{\mathbf{w}}_d) - \inf_{\mathbf{w}\in\mathbb{R}^d} R_d(\mathbf{w}) \xrightarrow{P} 0$

UniGe | MaLGa

Deterministic perturbation bound

Let $\{(\mathbf{x}_i, y_i\}_{i=1}^m$ be a deterministic sample of size m. Define the perturbation matrix

$$E_{ij} := y_i y_j \mathbf{x}_{i|d-k}^T \mathbf{x}_{j|d-k} \cdot \mathbf{1}\{i \neq j\} \in \mathbb{R}^{m \times m}.$$

Assume that the max-margin predictor exists and suppose

$$\epsilon_0 := \frac{2\|E\| \cdot \max_i \|\mathbf{x}_{i|d-k}\|^2}{\min_i \|\mathbf{x}_{i|d-k}\|^4} \leq \frac{1}{2}.$$

Then,

$$\hat{\mathbf{w}}_{|k} = \arg\min_{\mathbf{v} \in \mathbb{R}^k} (1 + \epsilon_{\mathbf{v}}) \cdot \hat{\mathbb{E}} \left[\frac{[1 - y\mathbf{x}_{|k}^T \mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2} \right] + \frac{\|\mathbf{v}\|^2}{m} \quad \text{with } \sup_{\mathbf{v}} |\epsilon_{\mathbf{v}}| \leq \epsilon_0,$$

and

$$\|\mathbf{w}_{|d-k}\|^2 \leq \frac{5m}{\min_i \|\mathbf{x}_{i|d-k}\|^2}.$$

Moving to statistical setting

Assumptions:

1. $\sup_d \mathbb{E}_d \left[\frac{1+\|y\mathbf{x}_k\|^4}{\|\mathbf{x}_{|d-k}\|^4}\right] < \infty$,
2. With probability approaching 1 (as $d \to \infty$), the max-margin predictor $\hat{\mathrm{w}}_d$ exists and $0 > c \geq \max_i \max\{\|\mathbf{x}_{i|d-k}\|^{-1}, \|\mathbf{x}_{i|d-k}\|\}$,
3. $m_d \cdot \|E\| \xrightarrow{P} 0$,
4. $m_d \cdot \|\mathbb{E}_d[\mathbf{x}_{|d-k}\mathbf{x}_{|d-k}^T]\| \to 0$.

Setting

$$g_d(\mathbf{v}) := \mathbb{E}_d \left[\frac{[1 - y\mathbf{x}_{|k}^T\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right], \quad \text{and} \quad \hat{g}_d(\mathbf{v}) := \hat{\mathbb{E}}_d \left[\frac{[1 - y\mathbf{x}_{|k}^T\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right].$$

5. There exists $c' > 0$ independent of d such that with probability approaching 1, $\hat{g}_d$ has a minimizer of norm at most c',
6. $\inf_{\mathbf{v}} \limsup_d \left(g_d(\mathbf{v}) - \inf_{\mathbf{u}} g_d(\mathbf{u})\right) = 0$.

UniGe | MaLGa

Asymptotic characterization of max-margin predictor

Under these assumptions, with

$$g_d(\mathbf{v}) := \mathbb{E}_d\left[\frac{[1 - y\mathbf{x}_{|k}^T\mathbf{v}]_+^2}{\|\mathbf{x}_{|d-k}\|^2}\right], \quad \text{and} \quad \hat{\mathbf{w}}_d = \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \ \min_i y_i \mathbf{x}_i^T \mathbf{w} \geq 1,$$

the max-margin predictor satisfies

$$g_d(\hat{\mathbf{w}}_{d|k}) - \inf_{\mathbf{v}} g_d(\mathbf{v}) \xrightarrow{P} 0, \quad \text{and} \quad \mathbb{E}_d\left[(\mathbf{x}^T\hat{\mathbf{w}}_d - \mathbf{x}_{|k}^T\hat{\mathbf{w}}_{d|k})^2\right] \xrightarrow{P} 0.$$

- The junk features are asymptotically negligible, and the relevant ones actually minimize a scaled squared hinge loss.
- Whether minimizers of g_d also have small misclassification error is equivalent to asking whether this weighted squared hinge loss is a good surrogate.

Study setting

The characterization makes the following assumptions look mild:

1. The joint distribution of $(\mathbf{x}_{|k}, \|\mathbf{x}_{|d-k}\|, y)$ is the same under any d,
2. $\mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^T]$ is positive definite
3. $\Pr\{\|\mathbf{x}_{|d-k}\| \in I\} = 1$ for a closed interval $I \subset (0, \infty)$.

Let $(\mathbf{x}_{|k}, y)$ be distributed according to a linearly separable distribution $\mathcal{D}_{\text{clean}}$ but implement label flips via the loss function:

$$L_p(\mathbf{w}) := \mathbb{E}_{(\mathbf{x}_{|k}, z, y) \sim \mathcal{D}_{\text{clean}}}[z \cdot \ell_p(y\mathbf{x}_{|k}^T\mathbf{w})], \quad \text{where} \quad \ell_p(\beta) := (1-p) \cdot [1-\beta]_+^2 + p \cdot [1+\beta]_+^{-1}$$

- for $p \in (0, \frac{1}{2})$, L_p is strongly convex $\implies$ unique minimizer $\mathbf{w}_p^*$
- benign overfitting happens if $\Pr_{(\mathbf{x}_{|k}, y) \sim \mathcal{D}_{\text{clean}}}\{y\mathbf{x}_{|k}^T\mathbf{w}_p^* \le 0\} = 0$

Classification setting favorable for benign overfitting

Two cases in which we have benign overfitting

Theorem

$(\mathbf{x}_{|k}, z, y) \sim \mathcal{D}_{\text{clean}}$ any distribution so that $(\mathbf{x}_{|k}, y)$ linearly separable and $\mathbf{x}_{|k}$ has bounded support. Then, there is $a \in (0, \frac{1}{2}$ dependent on $\mathcal{D}_{\text{clean}}$ such that for all $p \in (0, a)$ we have

$$\Pr_{(\mathbf{x}_{|k}, y) \sim \mathcal{D}_{\text{clean}}} \{y \mathbf{x}_{|k}^T \mathbf{w}_p^* \leq 0\} = 0.$$

Theorem

$(\mathbf{x}_{|k}, z, y) \sim \mathcal{D}_{\text{clean}}$ any distribution so that $(\mathbf{x}_{|k}, y)$ linearly separable. Suppose that for some unit vector $\mathbf{u}$ and conditioned on any value of y,

- $\mathbf{u}^T \mathbf{x}$ and $(I - \mathbf{u}\mathbf{u}^T)\mathbf{x}$ are mutually independent, and
- the distributions of $(I - \mathbf{u}\mathbf{u}^T)\mathbf{x}$ and $-(I - \mathbf{u}\mathbf{u}^T)\mathbf{x}$ are identical.

Then, for all $p \in (0, \frac{1}{2})$ we have

$$\Pr_{(\mathbf{x}_{|k}, y) \sim \mathcal{D}_{\text{clean}}} \{y \mathbf{x}_{|k}^T \mathbf{w}_p^* \leq 0\} = 0.$$