The Implicit Bias of Benign Overfitting

Y Gao

August 26, 2025

Outline

Implicit Bias in ML

Standard supervised ML:

- Set of predictors $\mathcal{H}$; distribution over examples $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$
- Goal: For some loss function ℓ,

$$\min_{h \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},\mathbf{y})} \ell(h; (\mathbf{x}, \mathbf{y}))$$

- Standard approach: Empirical Risk Minimization (ERM). Sample training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^m$, return

$$\arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h; (\mathbf{x}_i, \mathbf{y}_i))$$

Implicit Bias in ML

In modern ML (e.g. deep learning), often many empirical risk minimizers; Choice depends on algorithm used

- Same empirical risk, not same expected loss/other properties
- Properties of returned predictor known as the algorithm's implicit bias

Classical learning theory often doesn't distinguish between ERMs; Raises many new questions

This Talk

Implicit bias of gradient-based methods, in the context of benign overfitting

1. Linear Regression with the Square Loss
2. Linear Regression Beyond the Square Loss
3. Linear Binary Classification

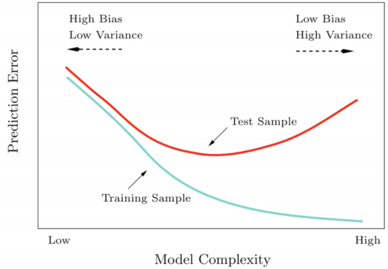
Benign Overfitting

Classical approach to explain learning with ERMs:

- Algorithm picks predictors from class $\mathcal{H}$
- $\mathcal{H}$ satisfies uniform convergence:
  - With high probability, average loss and expected loss are close, simultaneously for all $h \in \mathcal{H}$

$\Rightarrow$ ERM finds near-optimal predictor in $\mathcal{H}$

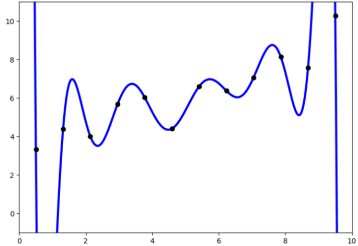
Conversely, if training/test performance differs, can overfit and get bad predictor



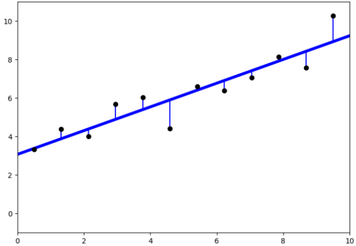
Benign Overfitting

Average/expected loss differs, overfitting, yet learned predictor is good

- Initially, strong empirical evidence from deep learning
- Later: Same in linear/kernel learning
- Emerging understanding under appropriate distributional assumptions



Overfitting



Benign Overfitting

Prior works

Reference	**Model**
Bartlett, Long, et al (2020)	Linear regression
Liang, Rakhlin (2018)	Kernel ridgeless regression
Mei, Montanari (2019)	Random feature regression
Belkin, Hsu, et al (2018)	Kernel smoothers / nearest neighbors
Rakhlin, Zhai (2019)	Laplace kernel interpolation
Koehler, Zhou, et al (2021)	High-dim linear regression
Ji, Li, et al(2021)	Early-stopped neural networks
Beaglehole, Belkin, et al (2022)	Shift-invariant kernel interpolators

Mallinar N, Simon J B, Abedsoltan A, et al. Benign, tempered, or catastrophic: A taxonomy of overfitting[J]. NIPS, 2022.

When will Occur?

Depends on learning algorithm + data distribution

- Linear predictors $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$, $\mathbf{x} \in \mathbb{R}^d$, squared loss

$$\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)^2$$

- dimension $d \gg m$.
- Gradient methods on above converge to

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)^2 = 0$$

When will Occur?

Depends on learning algorithm + data distribution

- Linear predictors $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$, $\mathbf{x} \in \mathbb{R}^d$, squared loss

$$\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)^2$$

- dimension $d \gg m$.
- Gradient methods on above converge to

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)^2 = 0$$

Benign Overfitting: as $d, m \to \infty$,

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}(\mathbf{x}^\top \hat{\mathbf{w}} - \mathbf{y})^2 \to \min_{\mathbf{w}} \mathbb{E}_{(\mathbf{x},\mathbf{y})}(\mathbf{x}^\top \mathbf{w} - \mathbf{y})^2 \quad (> 0)$$

- Bartlett et al. 2019: Benign overfitting if
  - $\mathbf{y} = \mathbf{x}^\top \mathbf{w}^* +$ noise (well-specified/realizable setting)
  - Covariance matrix of $\mathbf{x}$ has "many small positive eigenvalues"

Intuition

- Distributional assumption: $\mathbf{x} = (\mathbf{x}_{|k}, \mathbf{x}_{|d-k})$
  - $\mathbf{x}_{|k}$: k "important" coordinates ($\mathbf{y}$ depends on $\mathbf{x}_{|k}$)
  - $\mathbf{x}_{|d-k}$: $d-k$ small "junk" coordinates (e.g. $\sim \mathcal{N}(0, \frac{1}{d-k} I_{d-k})$ independently)
- If $d \gg k$, can show that $\hat{\mathbf{w}} = (\hat{\mathbf{w}}_{|k}, \hat{\mathbf{w}}_{|d-k})$ where
  - $\hat{\mathbf{w}}_{|k} \approx$ optimum on first k coordinates w.r.t. expected loss
  - $\hat{\mathbf{w}}_{|d-k}$ used to fit training examples
  - On new $\mathbf{x} \sim \mathcal{D}$, $\mathbf{x}^\top \hat{\mathbf{w}} \approx \mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{|k}$

Most existing results for regression are extensions of this idea. But, has proven difficult to generalize

- Agnostic/misspecified setting: $\mathbf{y} \neq \mathbf{x}^\top \mathbf{w}^* +$ noise
- Non-linear predictors...

An Observation

- Slight extension of previous setting: $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$. Non-negative loss $\ell(\mathbf{x}^\top \mathbf{w}, \mathbf{y})$, equals 0 for unique prediction value $\ell_{\mathbf{y}}^{-1}(0)$

An Observation

- Slight extension of previous setting: $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$. Non-negative loss $\ell(\mathbf{x}^\top \mathbf{w}, \mathbf{y})$, equals 0 for unique prediction value $\ell_{\mathbf{y}}^{-1}(0)$
- **Theorem**: Gradient methods will still converge to

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{x}_i^\top \mathbf{w}, \mathbf{y}_i) = 0.$$

An Observation

- Slight extension of previous setting: $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$. Non-negative loss $\ell(\mathbf{x}^\top \mathbf{w}, \mathbf{y})$, equals 0 for unique prediction value $\ell_{\mathbf{y}}^{-1}(0)$
- **Theorem**: Gradient methods will still converge to

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{x}_i^\top \mathbf{w}, \mathbf{y}_i) = 0.$$

- **Observation**: Also equals

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \frac{1}{m}\sum_{i=1}^{m} (\mathbf{x}_i^\top \mathbf{w} - \ell_{\mathbf{y}_i}^{-1}(0))^2 = 0.$$

An Observation

- Slight extension of previous setting: $\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}$. Non-negative loss $\ell(\mathbf{x}^\top \mathbf{w}, \mathbf{y})$, equals 0 for unique prediction value $\ell_{\mathbf{y}}^{-1}(0)$
- **Theorem**: Gradient methods will still converge to

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{x}_i^\top \mathbf{w}, \mathbf{y}_i) = 0.$$

- **Observation**: Also equals

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \frac{1}{m}\sum_{i=1}^{m} (\mathbf{x}_i^\top \mathbf{w} - \ell_{\mathbf{y}_i}^{-1}(0))^2 = 0.$$

This is ERM w.r.t. two different statistical learning problems:

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}[\ell(\mathbf{x}^\top \mathbf{w}; \mathbf{y})] \quad \text{vs.} \quad \mathbb{E}_{(\mathbf{x},\mathbf{y})}[(\mathbf{x}^\top \mathbf{w} - \ell_{\mathbf{y}}^{-1}(0))^2]$$

But algorithm converges to same point $\Rightarrow$
Generally can't have consistency/benign overfitting w.r.t. to both!

Conclusion

- The fact that we have benign overfitting on one learning problem precludes benign overfitting on other learning problems
- Implicit bias in the space of learning problems!
- In what follows, use this to prove positive + negative results, going well-specified linear regression

Baseline result

Model: $\mathbf{x} = (\mathbf{x}_{|k}, \mathbf{x}_{|d-k})$, $\mathbf{x}_{|d-k}$ distributed as $\mathcal{N}(\mathbf{0}, \frac{1}{d-k} \cdot I_{d-k})$

Theorem

As $d, m \to \infty$, min-norm predictor $\hat{\mathbf{w}}$ satisfies

$$\hat{\mathbf{w}}_{|k} \to \mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]^{-1} \cdot \mathbb{E}[\mathbf{y}\mathbf{x}_{|k}]$$

and

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}[(\mathbf{x}^\top \hat{\mathbf{w}} - \mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{|k})^2] \to 0.$$

We have benign overfitting

Negative example 1

Tweak: $\mathbf{x}_{|d-k}$ distributed as $\mathcal{N}(\mathbf{0}, \frac{g(\mathbf{x}_{|k})}{d-k} \cdot I_{d-k})$ for some bounded positive function $g(\cdot)$

Theorem

As $d, m \to \infty$, min-norm predictor $\hat{\mathbf{w}}$ satisfies

$$\hat{\mathbf{w}}_{|k} \to \mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{g(\mathbf{x}_{|k})}\right]^{-1} \cdot \mathbb{E}\left[\frac{\mathbf{y}\mathbf{x}_{|k}}{g(\mathbf{x}_{|k})}\right]$$

and

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})}[(\mathbf{x}^\top\hat{\mathbf{w}} - \mathbf{x}_{|k}^\top\hat{\mathbf{w}}_{|k})^2] \to 0.$$

$\hat{\mathbf{w}}$ no longer consistent!

Negative example 1

Proof: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)^2 = 0$ is also

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\mathbf{x}_i^\top}{\sqrt{g(\mathbf{x}_{i|k})}}\mathbf{w} - \frac{\mathbf{y}_i}{\sqrt{g(\mathbf{x}_{i|k})}}\right)^2 = 0,$$

which now falls into baseline model

Negative example 1

Proof: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)^2 = 0$ is also

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\mathbf{x}_i^\top}{\sqrt{g(\mathbf{x}_{i|k})}}\mathbf{w} - \frac{\mathbf{y}_i}{\sqrt{g(\mathbf{x}_{i|k})}}\right)^2 = 0,$$

which now falls into baseline model

- Needs misspecified setting! Otherwise $\hat{\mathbf{w}}_{|k}$ converges to

$$\mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top}{g(\mathbf{x}_{|k})}\right]^{-1} \cdot \mathbb{E}\left[\frac{\mathbf{x}_{|k}\mathbf{x}_{|k}^\top \mathbf{w}^*}{g(\mathbf{x}_{|k})}\right] = \mathbf{w}^*$$

Implication: Cannot generally expect benign overfitting in misspecified/agnostic linear regression

Negative example 2

Generalized linear model / single neuron, well-specified setting:

- $\mathbf{y} = \sigma(\mathbf{x}_{|k}^\top \mathbf{w}^*) + \xi$, $\sigma(\cdot)$ strictly monotonic
- Want to solve $\min_{\mathbf{w}} \mathbb{E}[(\sigma(\mathbf{x}^\top \mathbf{w}) - \mathbf{y})^2]$
- Apply gradient method on $\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} (\sigma(\mathbf{x}_i^\top \mathbf{w}) - \mathbf{y}_i)^2$

Negative example 2

Generalized linear model / single neuron, well-specified setting:

- $\mathbf{y} = \sigma(\mathbf{x}_{|k}^\top \mathbf{w}^*) + \xi$, $\sigma(\cdot)$ strictly monotonic
- Want to solve $\min_{\mathbf{w}} \mathbb{E}[(\sigma(\mathbf{x}^\top \mathbf{w}) - \mathbf{y})^2]$
- Apply gradient method on $\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} (\sigma(\mathbf{x}_i^\top \mathbf{w}) - \mathbf{y}_i)^2$

Theorem

As $d, m \to \infty$, returned $\hat{\mathbf{w}}$ satisfies

$$\begin{aligned}\hat{\mathbf{w}}_{|k} &\to \mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]^{-1} \cdot \mathbb{E}[\mathbf{x}_{|k} \cdot \sigma^{-1}(\mathbf{y})] \\ &= \mathbb{E}[\mathbf{x}_{|k}\mathbf{x}_{|k}^\top]^{-1} \cdot \mathbb{E}[\mathbf{x}_{|k} \cdot \sigma^{-1}(\sigma(\mathbf{x}_{|k}^\top \mathbf{w}^*) + \xi)]\end{aligned}$$

expression is generally $\neq \mathbf{w}^*$ for non-linear σ

Negative example 2

Proof: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^{m}(\sigma(\mathbf{x}_i^\top \mathbf{w}) - \mathbf{y}_i)^2 = 0$ is also

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i^\top \mathbf{w} - \sigma^{-1}(\mathbf{y}_i))^2 = 0$$

which now falls into baseline model with target values $\sigma^{-1}(\mathbf{y})$

Negative example 3

Linear regression w.r.t. loss other than the squared loss

- Want to minimize $\mathbb{E}[f(\mathbf{x}^\top \mathbf{w} - \mathbf{y})]$ for some non-negative $f(\cdot)$ minimized at 0 (e.g. absolute loss)
- Apply gradient method on $\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} f(\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)$

Theorem

As $d, m \to \infty$, returned $\hat{\mathbf{w}}$ satisfies

$$\hat{\mathbf{w}}_{|k} \to \mathbb{E}[\mathbf{x}_{|k} \mathbf{x}_{|k}^\top]^{-1} \cdot \mathbb{E}[\mathbf{y} \mathbf{x}_{|k}]$$

This is optimum w.r.t. $\mathbb{E}[(\mathbf{x}^\top \mathbf{w} - \mathbf{y})^2]$ not $\mathbb{E}[f(\mathbf{x}^\top \mathbf{w} - \mathbf{y})]$! **Proof**: $\hat{\mathbf{w}}$ is

also $\arg\min_{\mathbf{w}} \|\mathbf{w}\| : \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)^2 = 0$

Is Benign Overfitting Bogus?

- Still, does not accord with practice...

Is Benign Overfitting Bogus?

- Still, does not accord with practice...
- One option: Only happens with hyper-transformer-convnets with 10000000 layers (+ batchnorm)
  - Representation learning: Misspecified $\Rightarrow$ well-specified
  - Implicit bias: Flat minima

Is Benign Overfitting Bogus?

- Still, does not accord with practice...
- One option: Only happens with hyper-transformer-convnets with 10000000 layers (+ batchnorm)
  - Representation learning: Misspecified $\Rightarrow$ well-specified
  - Implicit bias: Flat minima
- Another option: Regression is the wrong setting to look at
  - All negative examples relied on prediction value exactly matching target value

Classification

Focus on linear predictors + binary classification: $\mathbf{y} \in \{-1, +1\}$, want to minimize

$$\min_{\mathbf{w}} \Pr(\text{sign}(\mathbf{x}^\top \mathbf{w}) \neq \mathbf{y}) = \Pr(\mathbf{y}\mathbf{x}^\top \mathbf{w} \leq 0)$$

- Value of $\mathbf{x}^\top \mathbf{w}$ doesn't matter, only sign!
- Gradient methods with standard losses known to return max-margin predictor

$$\hat{\mathbf{w}} := \arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \min_i \mathbf{y}_i \mathbf{x}_i^\top \mathbf{w} \geq 1$$

Classification

Focus on linear predictors + binary classification: $\mathbf{y} \in \{-1, +1\}$, want to minimize

$$\min_{\mathbf{w}} \Pr(\operatorname{sign}(\mathbf{x}^\top \mathbf{w}) \neq \mathbf{y}) = \Pr(\mathbf{y}\mathbf{x}^\top \mathbf{w} \leq 0)$$

- Value of $\mathbf{x}^\top \mathbf{w}$ doesn't matter, only sign!
- Gradient methods with standard losses known to return max-margin predictor

$$\hat{\mathbf{w}} := \arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \min_i \mathbf{y}_i \mathbf{x}_i^\top \mathbf{w} \geq 1$$

- Several previous papers studied benign overfitting for classification
- Challenge: max-margin predictor has no closed-form solution (unlike min-norm predictor in regression)
- Most results considered specific settings where max-margin and min-norm predictors coincide

Classification

Data model $(\mathbf{x}, \mathbf{y})$:

- $\mathbf{x}_{|k}, \mathbf{y}$ arbitrary fixed distribution
- $\mathbf{x}_{|d-k} \sim \mathcal{N}(\mathbf{0}, \frac{1}{d-k} I_{d-k})$

Classification

Data model $(\mathbf{x}, \mathbf{y})$:

- $\mathbf{x}_{|k}, \mathbf{y}$ arbitrary fixed distribution
- $\mathbf{x}_{|d-k} \sim \mathcal{N}(\mathbf{0}, \frac{1}{d-k} I_{d-k})$

Theorem

Under mild assumptions, max-margin predictor $\hat{\mathbf{w}}$ satisfies:

- $\mathbb{E}_{(\mathbf{x},\mathbf{y})}[(\mathbf{x}^\top \hat{\mathbf{w}} - \mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{|k})^2] \to 0$
- $\hat{\mathbf{w}}_{|k}$ *asymptotically minimizes expected squared hinge loss*

$$g(\mathbf{w}) = \mathbb{E}[\max\{0, 1 - \mathbf{y}\mathbf{x}_{|k}^\top \mathbf{w}\}^2]$$

Important: this loss is not the one used for training! $\hat{\mathbf{w}}$ is implicitly biased in that manner

Classification

Data model $(\mathbf{x}, \mathbf{y})$:

- $\mathbf{x}_{|k}, \mathbf{y}$ arbitrary fixed distribution
- $\mathbf{x}_{|d-k} \sim \mathcal{N}(\mathbf{0}, \frac{1}{d-k} I_{d-k})$

Theorem

Under mild assumptions, max-margin predictor $\hat{\mathbf{w}}$ satisfies:

- $\mathbb{E}_{(\mathbf{x},\mathbf{y})}[(\mathbf{x}^\top \hat{\mathbf{w}} - \mathbf{x}_{|k}^\top \hat{\mathbf{w}}_{|k})^2] \to 0$
- $\hat{\mathbf{w}}_{|k}$ *asymptotically minimizes expected squared hinge loss*

$$g(\mathbf{w}) = \mathbb{E}[\max\{0, 1 - \mathbf{y}\mathbf{x}_{|k}^\top \mathbf{w}\}^2]$$

Important: this loss is not the one used for training! $\hat{\mathbf{w}}$ is implicitly biased in that manner

Corollary

Benign overfitting occurs if $g(\cdot)$ is a good surrogate for misclassification error

Implications

- Similar data model as before
- $\mathbf{y}$ equals $\mathrm{sign}(\mathbf{x}^\top \mathbf{w}^*)$ plus label noise w.p. p

Theorem

For any distribution on $\mathbf{x}_{|k}$, benign overfitting for some $p > 0$

Theorem

If $\mathbf{x}_{|k}$ mixture of symmetric distributions, benign overfitting for any $p \in (0, \frac{1}{2})$

Can study other settings as well (and no need to assume max-margin and min-norm predictors coincide)

Proof Intuition

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \min_i \mathbf{y}_i \mathbf{x}_i^\top \mathbf{w} \geq 1$$

- Suppose $\mathbf{y}_i = 1$, $\mathbf{x}_{i|d-k} = \mathbf{e}_i$ for all i

Proof Intuition

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \min_i \mathbf{y}_i \mathbf{x}_i^\top \mathbf{w} \geq 1$$

- Suppose $\mathbf{y}_i = 1$, $\mathbf{x}_{i|d-k} = \mathbf{e}_i$ for all i
- Rewrite problem as

$$\arg\min_{\mathbf{w}} \|\mathbf{w}_{|k}\|^2 + \|\mathbf{w}_{|d-k}\|^2 \quad : \quad (\mathbf{w}_{|d-k})_i \geq 1 - \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}$$

Proof Intuition

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\| \quad : \quad \min_i \mathbf{y}_i \mathbf{x}_i^\top \mathbf{w} \geq 1$$

- Suppose $\mathbf{y}_i = 1$, $\mathbf{x}_{i|d-k} = \mathbf{e}_i$ for all i
- Rewrite problem as

$$\arg\min_{\mathbf{w}} \|\mathbf{w}_{|k}\|^2 + \|\mathbf{w}_{|d-k}\|^2 \quad : \quad (\mathbf{w}_{|d-k})_i \geq 1 - \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}$$

- For any fixed $\mathbf{w}_{|k}$, best to pick $(\mathbf{w}_{|d-k})_i = \max\{0, 1 - \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}\}$, leading to

$$\arg\min_{\mathbf{w}_{|k}} \|\mathbf{w}_{|k}\|^2 + \sum_{i=1}^{m} \max\{0, 1 - \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}\}^2$$

$$= \arg\min_{\mathbf{w}_{|k}} \frac{1}{m}\|\mathbf{w}_{|k}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - \mathbf{x}_{i|k}^\top \mathbf{w}_{|k}\}^2$$

$$m \to \infty \implies \mathbb{E}[\max\{0, 1 - \mathbf{y}\mathbf{x}_{|k}^\top \mathbf{w}_{|k}\}^2]$$

Recent Progress

- **Beyond linear:** Extensions to multi-class, kernels, two-layer nets
- **Rates & finite-sample:** Precise bounds
- **Geometry of noise:** Correlated / anisotropic effects
- **Algorithms:** How do optimizers change implicit bias?

Beyond Linear

- 2-layer CNN: First characterize the conditions under which benign overfitting can occur in training CNN [1]
- 2-layer ReLU CNN: Establish algorithm-dependent risk bounds for learning 2-layer ReLU CNN with noise [2]
- Sparse LR: A new implicit bias effect that combines the benefit of ℓ_1 and ℓ_2 interpolators [3]

[1] Cao Y, Chen Z, Belkin M, et al. Benign overfitting in two-layer convolutional neural networks[J]. NIPS, 2022.

[2] Kou Y, Chen Z, Chen Y, et al. Benign overfitting in two-layer relu convolutional neural networks[C]. PMLR, 2023.

[3] Zhou M, Ge R. Implicit regularization leads to benign overfitting for sparse linear regression[C]. PMLR, 2023.

Rates & Finite-Sample Guarantees

- Ridge regression: Sharp conditions for benign overfitting with arbitrary covariance, explicit variance and bias rates [4]
- Nonlinear networks: 2-layer neural networks achieve minimax optimal test error under noisy labels [5]
- Distribution shift: Characterizations under covariate shift in overparameterized regimes [6]

[4] Tsigler A, Bartlett P L. Benign overfitting in ridge regression[J]. JMLR, 2023.

[5] Frei S, Chatterji N S, Bartlett P. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data[C]. PMLR, 2022.

[6] Tang S, Wu J, Fan J, et al. Benign overfitting in out-of-distribution generalization of linear models[J]. arXiv preprint, 2024.

Geometry of Noise

- Linear models: Independence is not required! Benign overfitting can hold under correlated and anisotropic designs [7]
- Neural networks:
  - Incorporate class-dependent heterogeneous noise [8]
  - Conditions on signal-to-noise ratio critical for whether margin-maximization still leads to benign overfitting [9]

[7] Tsigler A, Bartlett P L. Benign overfitting in ridge regression[J]. JMLR, 2023.

[8] Xu R, Chen K. Rethinking benign overfitting in two-layer neural networks[J]. arXiv preprint, 2025.

[9] Karhadkar K, George E, Murray M, et al. Benign overfitting in leaky relu networks with moderate input dimension[J]. NIPS, 2024.

Algorithms

- Momentum-based: Consider heavy-ball and Nesterov's method of accelerated gradients [10]
- Adam: Iterate align with max ℓ_∞-margin classifier [11]
- Steepest descent family: Converge to solutions maximizing the margin with respect to the classifier matrix's p-norm [12]

[10] Lyu B, Wang H, Wang Z, et al. Effects of Momentum in Implicit Bias of Gradient Flow for Diagonal Linear Networks[C]. AAAI, 2025.

[11] Zhang C, Zou D, Cao Y. The implicit bias of adam on separable data[J]. NIPS, 2024.

[12] Fan C, Schmidt M, Thrampoulidis C. Implicit Bias of Spectral Descent and Muon on Multiclass Separable Data[J]. arXiv preprint, 2025.

Thank You For Listening.

Any Questions?