Reconsidering Overfitting in the Age of Overparameterized Models

NeurIPS 2023 Tutorial, New Orleans

Speakers: Spencer Frei, Vidya Muthukumar, Fanny Yang, Moderator: Daniel Hsu
Textbook wisdom on overfitting

\[ f^*(\mathbf{x}) \]

random \( x_i \), \( y_i \) noisy version of \( f^*(x_i) \)

true \( f^*(x) \)

\( n = 20 \) samples
Textbook wisdom on overfitting

• cannot express function of interest (high statistical bias)
• largely ignores noise → does not fluctuate a lot (small variance)

\[
p(\mathbf{x})\quad \text{predicted } \hat{f}(x)
\]

\[
f(\mathbf{x}) = f^*(\mathbf{x})
\]

\[\begin{align*}
\text{n = 20 samples} \\
\text{polynomial fit degree } d = 2
\end{align*}\]
Textbook wisdom on overfitting

Small models cannot fit perfectly:
• cannot express function of interest (high statistical bias)

- n = 20 samples
- polynomial fit degree d = 2
- random $x_i$, $y_i$ noisy version of $f^*(x_i)$
- predicted $\hat{f}(x)$
- true $f^*(x)$
Textbook wisdom on overfitting

- cannot express function of interest (high statistical bias)
- largely ignores noise → does not fluctuate a lot (small variance)

Small models cannot fit perfectly:

- random $x_i$, $y_i$ noisy version of $f^*(x_i)$
- predicted $\hat{f}(x)$
- true $f^*(x)$
Textbook wisdom on overfitting

n = 20 samples
polynomial fit
degree d = 20

$\hat{f}(x)$

true $f^*(x)$

random $x_i$, $y_i$ noisy version
of $f^*(x_i)$
Textbook wisdom on overfitting

- \( n = 20 \) samples
- Polynomial fit
  - Degree \( d = 20 \)
  - Flexible and can express function of interest (small bias)
  - Fits too much of the noise (overfit) \( \rightarrow \) fluctuates a lot (high variance)

Large models fit perfectly (overfit):

- flexible and can express function of interest (small bias)
- Fits too much of the noise (overfit) \( \rightarrow \) fluctuates a lot (high variance)
Textbook wisdom: Avoid fitting noise

Classical theory: Improve generalization by optimizing expressivity via bias-variance trade-off

n = 20 samples

polynomial fit
degree d = 5

random $x_i$, $y_i$ noisy version of $f^*(x_i)$

predicted $\hat{f}(x)$

true $f^*(x)$

plot showing polynomial fit with noisy data points
Textbook wisdom: Avoid fitting noise

Classical theory: Improve generalization by optimizing expressivity via bias-variance trade-off
Textbook wisdom on overfitting

n = 20 samples
polynomial fit degree d = 20

What happens if we increase the polynomial degree even further without regularizing?
Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ‘20]
Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

[ResNet18 width parameter - Error graph with underparameterized标注]

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ’20]
Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ‘20]
Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

interpolation threshold:
training 0-1 error $\approx 0$

underparameterized

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ’20]
Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

interpolation threshold: training 0-1 error $\approx 0$

underparameterized

overparameterized: models flexible enough to partially fit noise

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ‘20]
Double descent on neural networks

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

interpolation threshold: training 0-1 error ≈ 0

underparameterized

overparameterized: models flexible enough to partially fit noise

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ‘20]
Obs. I: Second descent beyond interpolation

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

After interpolation threshold, we have a second "descent" (double descent) for interpolators

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ‘20]
Obs. II: Harmless interpolation for large models

Classification using neural networks and Adam on CIFAR-10 with 15\% additional label noise

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ‘20]
Obs. II: Harmless interpolation for large models

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

[0.5, 0.35, 0.65, 0.7] Diagram: Test error vs. ResNet18 width parameter. Optimal early stopping or regularization is indicated by a vertical line.

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ’20]
Obs. II: Harmless interpolation for large models

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

Trained # of epochs

interpolation threshold: training 0-1 error ≈ 0

compare dark blue (at convergence) with red dashed (best stopping time)
Obs. II: Harmless interpolation for large models

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

interpolation threshold: training 0-1 error ≈ 0

Trained # of epochs

Test error

Optimal Early Stopping/regularization

ResNet18 Width Parameter

compare dark blue (at convergence) with red dashed (best stopping time)

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ‘20]
Obs. II: Harmless interpolation for large models

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

For large models, interpolation is not worse than regularization (harmless interpolation)

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ’20]
Obs. III: Good generalization for large models

Classification using neural networks and Adam on CIFAR-10 with 15% additional label noise

For large models, we achieve reasonably good test accuracy

[Nakkiran, Kaplun, Bansal, Yang, Barak, Sutskever ‘20]
Textbooks need an update…

Understanding Deep Learning (Still) Requires Rethinking Generalization
By Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals

Communications of the ACM, 2021

*and many more papers that expressed the need for “rethinking”
Textbooks need an update…

Understanding Deep Learning (Still) Requires Rethinking Generalization

By Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals

Communications of the ACM, 2021

*and many more papers that expressed the need for “rethinking”
What the field set out to understand…

Try to understand *when* the following happens:

1. **Second “descent”** as model size grows beyond interpolation threshold

2. **Harmless interpolation** for large models, i.e. interpolation $\sim$ opt. regularization

3. **Good test performance** for large models, close to best possible prediction error
What the field set out to understand…

Try to understand **when** the following happens:

1. **Second “descent”** as model size grows beyond interpolation threshold
2. **Harmless interpolation** for large models, i.e., interpolation \( \sim \) opt. regularization
3. **Good test performance** for large models, close to best possible prediction error
What the field set out to understand...

Try to understand when the following happens:

1. Second “descent” as model size grows beyond interpolation threshold
2. Harmless interpolation for large models, i.e. interpolation ~ opt. regularization
3. Good test performance for large models, close to best possible prediction error
What the field set out to understand...

Try to understand *when* the following happens:

1. Second “descent” as model size grows grows beyond interpolation threshold
2. Harmless interpolation for large models, i.e. interpolation $\sim$ opt. regularization
3. Good test performance for large models, close to best possible prediction error

As overparameterization $\uparrow$:

- variance decays
- bias stays low
What the field set out to understand...

Try to understand when the following happens:

1. Second "descent" as model size grows beyond interpolation threshold

2. Harmless interpolation for large models, i.e. interpolation $\sim$ opt. regularization

3. Good test performance for large models, close to best possible prediction error

As overparameterization $\uparrow$:

- Variance decays
- Bias stays low

when is this the case?
Which factors govern... when we have this picture...
Which factors govern... when we have this picture... rather than this picture...
Seeking answers using theoretical analysis...

Neural network interpolators

- feature learning with
  \textit{overparameterization} \triangleq 
  e.g. \textit{width} of hidden layers

- found w/ 1st order methods to minimize \textit{non-convex losses}
Seeking answers using theoretical analysis…

Neural network interpolators \(\rightarrow\) Kernel / random features

- feature learning with overparameterization \(\triangleq\)
e.g. width of hidden layers
- found w/ 1st order methods to minimize non-convex losses

- using \(p\) nonlinear features w/
  overparameterization \(\triangleq\)
  number of features \(p \gg n\)
- found w/ 1st order methods
  to minimize a convex loss

complexity to analyze model
Seeking answers using theoretical analysis...

Neural network interpolators

- feature learning with overparameterization \( \triangleq \)
- e.g. width of hidden layers
- found w/ 1st order methods to minimize non-convex losses

Kernel / random features

- using \( p \) nonlinear features w/ overparameterization \( \triangleq \)
- number of features \( p \gg n \)
- found w/ 1st order methods to minimize a convex loss

Linear interpolators

- using \( d \) input features with overparameterization \( \triangleq \)
- dimension \( d \gg n \)
- found w/ 1st order methods to minimize a convex loss

complexity to analyze model
Seeking answers using theoretical analysis...

Neural network interpolators
- feature learning with overparameterization $\triangleq$ e.g. width of hidden layers
- found w/ 1st order methods to minimize non-convex losses

Kernel / random features
- using $p$ nonlinear features w/ overparameterization $\triangleq$ number of features $p \gg n$
- found w/ 1st order methods to minimize a convex loss

Linear interpolators
- using $d$ input features with overparameterization $\triangleq$ dimension $d \gg n$
- found w/ 1st order methods to minimize a convex loss
Seeking answers using theoretical analysis…

**Neural network interpolators**
- feature learning with overparameterization ∆ e.g. width of hidden layers
- found w/ 1st order methods to minimize **non-convex losses**

**Kernel / random features**
- using \( p \) nonlinear features w/ overparameterization ∆ number of features \( p \gg n \)
- found w/ 1st order methods to minimize a **convex loss**

**Linear interpolators**
- using \( d \) input features with overparameterization ∆ dimension \( d \gg n \)
- found w/ 1st order methods to minimize a **convex loss**

complexity to analyze model
Plan today...

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
Plan today...

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the
- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data
Plan today…

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the
- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data
Goal is **not to find** better interpolators in practice

but **to understand when** interpolation is benign
Plan today…

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the
- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data
Benefits of overparameterization and interpolation in linear models

We run gradient descent on $\|Y - X\theta\|_2^2$ at $\theta_0 = 0$ for $Y = X\theta^* + W$
(where $X, W$ are comprised of iid standard Gaussian entries)
Benefits of overparameterization and interpolation in linear models

We run gradient descent on $\|Y - X\theta\|^2$ at $\theta_0 = 0$ for $Y = X\theta^* + W$ (where $X, W$ are comprised of iid standard Gaussian entries)

$\mathbf{n = 500, } \theta^* = \hat{e}_1, \sigma^2 = 0.25$
Benefits of overparameterization and interpolation in **linear models**

We run gradient descent on $\|Y - X\theta\|_2^2$ at $\theta_0 = 0$ for $Y = X\theta^* + W$ (where $X, W$ are comprised of iid standard Gaussian entries)

\[ n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25 \]
Benefits of overparameterization and interpolation in **linear models**

We run gradient descent on \( \| \mathbf{Y} - \mathbf{X} \theta \|_2^2 \) at \( \theta_0 = 0 \) for \( \mathbf{Y} = \mathbf{X} \theta^* + \mathbf{W} \) (where \( \mathbf{X}, \mathbf{W} \) are comprised of iid standard Gaussian entries)

\[
\begin{align*}
n & = 500, \
\theta^* & = \hat{\mathbf{e}}_1, \
\sigma^2 & = 0.25
\end{align*}
\]

1. **Second Descent after interpolation**

2. **Harmless interpolation for large \( d/n \)**
Formal setup: overparameterized linear regression
Formal setup: overparameterized linear regression

\[
Y = X^\top \theta^* + W
\]

- **Output**: \( Y \)
- **Input features**, dimension = \( d \)
- **Noise**, variance = \( \sigma^2 \)
- **True parameter/signal** (unknown)
Formal setup: overparameterized linear regression

\[ Y = X^\top \theta^* + W \]

true parameter/signal (unknown)

output

input features, dimension = \( d \)

noise, variance = \( \sigma^2 \)

\[ \mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma \]

(data covariance)

e.g. “isotropic covariance” means \( \Sigma = I \)
Formal setup: overparameterized linear regression

\[ Y = X^\top \theta^* + W \]

- True parameter/signal (unknown)
- Output
- True parameter/signal (unknown)
- True parameter/signal (unknown)
- Input features, dimension = \( d \)
- Noise, variance = \( \sigma^2 \)
- Data covariance
- \( \mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma \)
- \( \mathbb{E}[XX^\top] = \Sigma \)
- E.g. “isotropic covariance” means \( \Sigma = I \)

\( \Rightarrow \)

\[ X\hat{\theta} = Y \] has infinitely many interpolating solutions!

\[ \hat{\theta} \approx \begin{bmatrix} Y \\ \end{bmatrix} \]
Formal setup: overparameterized linear regression

true parameter/signal (unknown)

\[ Y = X^\top \theta^* + W \]

output

input features, dimension = \( d \)

noise, variance = \( \sigma^2 \)

\[ \mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma \]

data covariance

e.g. “isotropic covariance” means \( \Sigma = I \)

\( d > n \) (no. of features) \( d > n \) (no. of samples)

(input) features

(input) samples

(output) samples

\[ X\hat{\theta} = Y \text{ has infinitely many interpolating solutions!} \]

Solutions of study today:

The minimum-\( p \)-norm interpolator

\[ \hat{\theta}_p = \arg \min ||\theta||_p \text{ subject to } X\theta = Y. \]

(beginning with \( p = 2 \))
Formal setup: overparameterized linear regression

true parameter/signal (unknown)

\[ Y = X^\top \theta^* + W \]

output
input features, dimension = \( d \)
noise, variance = \( \sigma^2 \)

\[ \mathbb{E}[X] = 0, \mathbb{E}[XX^\top] = \Sigma \]
(data covariance)
e.g. “isotropic covariance” means \( \Sigma = I \)

\[ \mathbb{E}[Y] = X^\top \theta^* + \mathbb{E}[W] \]

\[ \mathbb{E}[W] = 0, \mathbb{E}[WW^\top] = \sigma^2 I \]

\[ \text{Error metric is mean-squared-error: } \mathcal{E}_{\text{MSE}} := \mathbb{E} \left[ (X^\top(\hat{\theta} - \theta^*))^2 \right] \]

Solutions of study today:
The minimum-lp-norm interpolator
\[ \hat{\theta}_p = \arg \min_{\theta} ||\theta||_p \text{ subject to } X\theta = Y. \]
(beginning with \( p = 2 \))

\[ X\hat{\theta} = Y \] has infinitely many interpolating solutions!
Analysis framework

**Non-asymptotic:** we consider $d = n^\beta$, $\beta > 1$ (or $d \gg n$) and state results as:

- **Consistency:** goal is to have $\mathcal{E}_{\text{MSE}} \to 0$ as $n \to \infty$

- **Rates:** upper and lower bounds on $\mathcal{E}_{\text{MSE}}$ as a function of $n$ that match up to universal constants (not depending on $n, d, \theta^*, \Sigma$)
Analysis framework

**Non-asymptotic:** we consider $d = n^\beta, \beta > 1$ (or $d \gg n$) and state results as:

- **Consistency:** goal is to have $\mathcal{E}_{\text{MSE}} \to 0$ as $n \to \infty$

- **Rates:** upper and lower bounds on $\mathcal{E}_{\text{MSE}}$ as a function of $n$ that match up to universal constants (not depending on $n, d, \theta^*, \Sigma$)

An alternative asymptotic analysis framework (not the focus of this tutorial):

Considers $d \propto n, \frac{d}{n} = \gamma$.

**Exact error expressions** derived as a function of $\gamma$ as $n, d \to \infty$ together.
Why these types of “low-norm” interpolators?

Popular optimization algorithms converge to “low-norm” solutions!

Gradient descent on squared loss

(Folklore, see e.g. Engl et al 1996)

Minimum-$\mathbf{l}_2$-norm interpolation

\[
\hat{\theta}_2 = \arg \min ||\theta||_2 \\
\text{subject to} \\
X_i^\top \theta = Y_i, i \in [n].
\]
Why these types of “low-norm” interpolators?

Popular optimization algorithms converge to “low-norm” solutions!

Mirror descent on squared loss,
Potential $= \| \cdot \|_p$

(Gunasekar et al, 2018)

Minimum-$\| \cdot \|_p$-norm interpolation

$\hat{\theta}_p = \arg \min_{\theta} ||\theta||_p$
subject to
$X_i^\top \theta = Y_i, i \in [n].$

Coordinate descent/least-angle regression

(Efron et al, 2004)

Minimum-$\| \cdot \|_1$-norm interpolation

$\hat{\theta}_1 = \arg \min_{\theta} ||\theta||_1$
subject to
$X_i^\top \theta = Y_i, i \in [n].$
Why these types of “low-norm” interpolators?

Popular optimization algorithms converge to “low-norm” solutions!

- Mirror descent on squared loss, Potential = $\| \cdot \|_p$
  - Minimum-$l_p$-norm interpolation
    \[ \hat{\theta}_p = \text{arg min} \| \theta \|_p \text{ subject to } X_i^\top \theta = Y_i, i \in [n]. \]

- Coordinate descent/least-angle regression
  - Minimum-$l_1$-norm interpolation
    \[ \hat{\theta}_1 = \text{arg min} \| \theta \|_1 \text{ subject to } X_i^\top \theta = Y_i, i \in [n]. \]

(Gunasekar et al, 2018)

(Efron et al, 2004)

Implicit bias theory is a useful “sanity check” but not the full picture: do these solutions always generalize well?
Recall: what was observed for min-l2-norm interpolator

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\theta} - \theta^*)^2
\]

1. Second Descent after interpolation

2. Harmless interpolation for large \( d/n \)

At convergence: \( n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25 \)
Recall: what was observed for min-l2-norm interpolator

Second Descent after interpolation

Harmless interpolation for large $d/n$

(1) and (2) are implied by variance reduction with increased overparameterization!

Theorem (isotropic covariance)*: Variance term $\lesssim \frac{\sigma^2 n}{d}$.


$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$
Plan today…

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the
- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data
Variance reduction: main proof ideas

- **Step 1:** minimum-$l_2$-norm interpolator can be expressed in closed form

\[
\hat{\theta}_2 = X^\top (XX^\top)^{-1}Y = X^\top (XX^\top)^{-1}X\theta^* + X^\top (XX^\top)^{-1}W
\]
Variance reduction: main proof ideas

- **Step 1**: minimum-l2-norm interpolator can be expressed in closed form

\[
\hat{\theta}_2 = X^T (XX^T)^{-1} Y = X^T (XX^T)^{-1} X \theta^* + X^T (XX^T)^{-1} W
\]

Ideally: have this be close to \( \theta^* \) (error = bias)
Variance reduction: main proof ideas

• **Step 1:** minimum-l2-norm interpolator can be expressed in closed form

\[
\hat{\theta}_2 = X^\top (XX^\top)^{-1} Y = X^\top (XX^\top)^{-1} X\theta^* + X^\top (XX^\top)^{-1} W
\]

Ideally: have this be close to \(\theta^*\) (error = bias)  
Ideally: have this be close to 0 (error = variance)
Variance reduction: main proof ideas

- **Step 1:** minimum-l2-norm interpolator can be expressed in closed form
  \[
  \hat{\theta}_2 = X^T (XX^T)^{-1} Y = X^T (XX^T)^{-1} X \theta^* + X^T (XX^T)^{-1} W
  \]

  Ideally: have this be close to \( \theta^* \) (error = bias)
  Ideally: have this be close to 0 (error = variance)

- **Step 2:** variance term can also be expressed in closed form
  \[
  \text{Variance} = \| X^T (XX^T)^{-1} W \|^2_2 = W^T (XX^T)^{-1} XX^T (XX^T)^{-1} W
  \]

*Note:* this calculation is simplified for isotropic data covariance, but works more generally (Bartlett et al, 2020)
Variance reduction: main proof ideas

• **Step 1:** minimum-l2-norm interpolator can be expressed in closed form

\[ \hat{\theta}_2 = X^T (XX^T)^{-1} Y = X^T (XX^T)^{-1} X\theta^* + X^T (XX^T)^{-1} W \]

Ideally: have this be close to \( \theta^* \) (error = bias)

Ideally: have this be close to 0 (error = variance)

• **Step 2:** variance term can also be expressed in closed form

\[ \text{Variance} = \| X^T (XX^T)^{-1} W \|_2^2 = W^T (XX^T)^{-1} XX^T (XX^T)^{-1} W \]

\[ = W^T (XX^T)^{-1} W \]

**Note:** this calculation is simplified for isotropic data covariance, but works more generally (Bartlett et al, 2020)
Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when $d \gg n$ (with high prob.)

$$\langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|_2^2 \approx d$$
Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when \( d \gg n \) (with high prob.)

\[
\langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|_2^2 \approx d
\]

\[
\Rightarrow XX^\top \approx dI
\]
Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when $d \gg n$ (with high prob.)
  \[
  \langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|_2^2 \approx d
  \]
  \[
  \implies XX^\top \approx dI
  \]
  \[
  \implies \text{Variance} = W^\top (XX^\top)^{-1} W \approx \frac{\|W\|_2^2}{d}
  \]
Variance reduction: main proof ideas

- **Step 3:** data is **approximately orthogonal** when \( d \gg n \) (with high prob.)

\[
\langle X_i, X_j \rangle \approx 0 \quad \text{for} \quad i \neq j \quad \text{and} \quad \|X_i\|_2^2 \approx d
\]

\[
\implies XX^\top \approx dI
\]

\[
\implies \text{Variance} = W^\top (XX^\top)^{-1} W \approx \frac{\|W\|_2^2}{d}
\]

Total "noise energy"
Variance reduction: main proof ideas

- **Step 3**: data is approximately orthogonal when \( d \gg n \) (with high prob.)

\[
\langle X_i, X_j \rangle \approx 0 \text{ for } i \neq j \text{ and } \|X_i\|^2_2 \approx d
\]

\[
\Rightarrow XX^\top \approx dI
\]

\[
\Rightarrow \text{Variance} = W^\top (XX^\top)^{-1} W \approx \frac{\|W\|^2_2}{d}
\]

\[
\approx \frac{n\sigma^2}{d}
\]

**Intuition**: noise energy is "spread out" along \( d \) feature dimensions, contributes more harmlessly as \( d \) increases

**Note**: can show corresponding precise results when \( d \propto n, d, n \rightarrow \infty \) (Hastie et al, 2022)
So is min-l2-norm interpolation always a good idea?

Interpolator $\hat{\theta}_2 = \arg\min \|\theta\|_2$ subject to $X\theta = Y$ vs.

regularized estimator: $\arg\min \|X\theta - Y\|_2^2 + \lambda\|\theta\|_2^2$  

$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$
So is min-l2-norm interpolation always a good idea?

Interpolator $\hat{\theta}_2 = \arg \min \| \theta \|_2$ subject to $X\theta = Y$ vs.

regularized estimator: $\arg \min \|X\theta - Y\|_2^2 + \lambda \|\theta\|_2^2$

$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$

1. second descent ✓
2. harmless interpolation ✓
3. good generalization ❌
So is min-$l_2$-norm interpolation always a good idea?

Interpolator $\hat{\theta}_2 = \arg\min \|\theta\|_2$ subject to $X\theta = Y$ vs.

regularized estimator: $\arg\min \|X\theta - Y\|_2^2 + \lambda \|\theta\|_2^2$

$n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25$

1. second descent
2. harmless interpolation
3. good generalization

Core issue: bias increases with $d$, eventually dominates
Issues with isotropy and min-l2 inductive bias

**Recall:** minimum-l2-norm interpolator can be expressed in closed form

\[ \hat{\theta}_2 = X^T (XX^T)^{-1} Y = X^T (XX^T)^{-1} X\theta^* + X^T (XX^T)^{-1} W \]

Ideally: have this be close to \( \theta^* \) (error = bias)
Issues with isotropy and min-l2 inductive bias

**Recall:** minimum-l2-norm interpolator can be expressed in closed form

\[
\hat{\theta}_2 = X^\top (XX^\top)^{-1} Y = X^\top (XX^\top)^{-1} X \theta^* + X^\top (XX^\top)^{-1} W
\]

Ideally: have this be close to \( \theta^* \) (error = bias)

\[
\text{Bias} = \|(X^\top (XX^\top)^{-1} X - I) \theta^*\|^2_2
\]
Issues with isotropy and min-l2 inductive bias

Recall: minimum-l2-norm interpolator can be expressed in closed form

\[
\hat{\theta}_2 = X^\top (XX^\top)^{-1} Y = X^\top (XX^\top)^{-1} X \theta^* + X^\top (XX^\top)^{-1} W
\]

Ideally: have this be close to \( \theta^* \) (error = bias)

\[
\text{Bias} = \| (X^\top (XX^\top)^{-1} X - I) \theta^* \|_2^2
\]

Theorem*: \( \text{Bias} \approx \left(1 - \frac{n}{d}\right) \| \theta^* \|_2^2 \)

*included in results of Hastie et al (2022), Bartlett et al (2020)
Issues with isotropy and min-l2 inductive bias

**Recall:** minimum-l2-norm interpolator can be expressed in closed form

\[
\hat{\theta}_2 = X^T (XX^T)^{-1} Y = X^T (XX^T)^{-1} X \theta^* + X^T (XX^T)^{-1} W
\]

Ideally: have this be close to \( \theta^* \) (error = bias)

\[
\text{Bias} = \| (X^T (XX^T)^{-1} X - I) \theta^* \|^2_2
\]

**Theorem**: \( \text{Bias} \approx \left( 1 - \frac{n}{d} \right) \| \theta^* \|^2_2 \)

**Intuition**: under isotropy, **true parameter energy** also spread out across \( d \) features!

*included in results of Hastie et al (2022), Bartlett et al (2020)
Isotropy and min-l2-norm bias visualized at feature-by-feature level

**Theorem:** $\text{Bias} \approx \left(1 - \frac{n}{d}\right) \|\theta^*\|_2^2$

**Intuition:** under isotropy, **true parameter energy** also spread out across $d$ features

This signal attenuation observed in classical statistical signal processing (e.g. Chen, Donoho, Saunders 2001)
Isotropy and min-l2-norm bias visualized at feature-by-feature level

**Theorem:** $\text{Bias} \approx \left(1 - \frac{n}{d}\right) \|\theta^*\|^2_2$

**Intuition:** under isotropy, **true parameter energy** also spread out across $d$ features

**Canonical setting:** k-sparse signal

$$Y = X^\top \theta^* + W$$

$\theta^*_j \neq 0$ for $j \in [k]$, 0 otherwise

$k \ll n$

This signal attenuation observed in classical statistical signal processing (e.g. Chen, Donoho, Saunders 2001)
Isotropy and min-l2-norm bias visualized at feature-by-feature level

**Theorem:** \( \text{Bias} \propto \left(1 - \frac{n}{d}\right) \|\theta^*\|^2_2 \)

**Intuition:** under isotropy, **true parameter energy** also spread out across \( d \) features

**Canonical setting:** \( k \)-sparse signal

\[
Y = X^\top \theta^* + W \\
\theta^*_j \neq 0 \text{ for } j \in [k], 0 \text{ otherwise} \\
k \ll n
\]

This signal attenuation observed in classical statistical signal processing (e.g. Chen, Donoho, Saunders 2001)
Isotropy and min-l2-norm bias visualized at feature-by-feature level

Theorem: $\text{Bias} \propto \left(1 - \frac{n}{d}\right) \|\theta^*\|_2^2$

Intuition: under isotropy, true parameter energy also spread out across $d$ features

Canonical setting: k-sparse signal

$$Y = X^\top \theta^* + W$$
$$\theta^*_j \neq 0 \text{ for } j \in [k], 0 \text{ otherwise}$$
$$k \ll n$$

Core issue for bias: $|\hat{\theta}_j| \ll |\theta^*_j|$ for all $j \in [k]$!

This signal attenuation observed in classical statistical signal processing (e.g. Chen, Donoho, Saunders 2001)
Plan today…

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the
- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data
Anisotropy to the rescue: “upweighting” features aligned with signal

• A special case \( \Sigma = \begin{bmatrix} RI_k & 0 \\ 0 & I_{d-k} \end{bmatrix} \), \( R \gg 1 \) (spiked-covariance)
Anisotropy to the rescue: “upweighting” features aligned with signal

- A special case $\Sigma = \begin{bmatrix} R I_k & 0 \\ 0 & I_{d-k} \end{bmatrix}$, $R \gg 1$ (spiked-covariance)

Effective “upweighting” on top k features
Anisotropy to the rescue: “upweighting” features aligned with signal

- A special case: \( \Sigma = \begin{bmatrix} R I_k & 0 \\ 0 & I_{d-k} \end{bmatrix} \), \( R \gg 1 \) (spiked-covariance)

Effective “upweighting” on top \( k \) features

\( (k = 500, n = 5000, d = 30000, R = 100) \)
Anisotropy to the rescue: “upweighting” features aligned with signal

- A special case $\Sigma = \begin{bmatrix} RI_k & 0 \\ 0 & I_{d-k} \end{bmatrix}$, $R \gg 1$ (spiked-covariance)

Effective “upweighting” on top $k$ features

(k = 500, n = 5000, d = 30000, R = 100)

Intuition: under near-orthogonality, $\hat{\theta}_j \propto \sum_{i=1}^{n} y_i x_{i,j}$ - attenuation mitigated for larger R as $x_{i,j} \sim \mathcal{N}(0,R)$ for $j \in [k]$
A sensible model for l2: the **spiked-covariance** ensemble

\[ \Sigma = \text{diag}(\Lambda) = \]

**Spiked covariance:** \((n, d, k, R)\)

- Feature magnitude \((\lambda_j)\)
- Feature index \((j)\)

\[ n \leq d \gg n \]

Conditions for **general anisotropic covariances** in terms of “effective ranks” by Bartlett et al (2020)
A sensible model for l2: the **spiked-covariance** ensemble

\[ \Sigma = \text{diag}(\Lambda) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{pmatrix} \]

**(sparsity level)**

Feature magnitude \( (\lambda_j) \)

Feature index \( (j) \)

\( n \)

\( d \gg n \)

\( k \ll n \)

Conditions for **general anisotropic covariances** in terms of “effective ranks” by Bartlett et al (2020)
A sensible model for \( l_2 \): the **spiked-covariance** ensemble

\[
\Sigma = \text{diag}(\Lambda) = \begin{array}{c}
\begin{array}{c}
\text{Feature magnitude} (\lambda_j) \\
\text{Feature index} (j)
\end{array}
\end{array}
\]

**Spiked covariance:** \((n, d, k, R)\)

\( k \ll n \)  
\( n \)  
\( d \gg n \)

Conditions for **general anisotropic covariances** in terms of “effective ranks” by Bartlett et al (2020)
A sensible model for l2: the **spiked-covariance** ensemble

\[ \Sigma = \text{diag}(\Lambda) = \]

(sparsity level)

\[ \lambda_j \]

Feature magnitude

\[ \mathbf{n} \]

Feature index (j)

\[ k \ll n \quad d \gg n \]

\[ R \gg 1 \]

Additionally assume \( \theta_j^* = 0 \) for all \( j = k + 1, \ldots, d \)

**Spiked covariance**: \((n, d, k, R)\)

**Conditions for general anisotropic covariances** in terms of “effective ranks” by Bartlett et al (2020)
A sensible model for l2: the **spiked-covariance** ensemble

Spiked covariance: $(n, d, k, R)$

$\Sigma = \text{diag}(\Lambda) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{pmatrix}$

(Sparsity level)

$\lambda_j \approx \begin{cases} \sqrt{\frac{d}{n}} & \text{if } j \leq k \\ 0 & \text{if } j > k \end{cases}$

Feature magnitude $(\lambda_j)$

Feature index $(j)$

$n \ll k \ll n \ll d \gg n$

Additionally assume $\theta_j^* = 0$ for all $j = k + 1, \ldots, d$

Will **always** achieve

Variance $\to 0$ as $n, d \to \infty$:

Noise hidden along $(d-k)$ directions!

Conditions for **general anisotropic covariances** in terms of “effective ranks” by Bartlett et al (2020)
A sensible model for l2: the **spiked-covariance** ensemble

Spiked covariance: \((n, d, k, R)\)

Will **always** achieve

Variance \(\rightarrow 0\) as \(n, d \rightarrow \infty\):

Noise hidden along \((d-k)\) directions!

Additionally assume \(\theta_j^* = 0\) for all \(j = k + 1, \ldots, d\)

Also achieves Bias \(\rightarrow 0\) as \(n, d \rightarrow \infty\)

provided that \(R \gg \frac{d}{n}\)

Conditions for **general anisotropic covariances** in terms of “effective ranks” by Bartlett et al (2020)
Summary: Uniform benefits of overparameterization with spiked covariance

\[ n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25 \]

Isotropic covariance
Summary: Uniform benefits of overparameterization with spiked covariance

\[ n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25 \]

Isotropic covariance

Spiked covariance, \( R = 10 \)
Summary: Uniform benefits of overparameterization with spiked covariance

\( n = 500, \theta^* = \hat{e}_1, \sigma^2 = 0.25 \)

Isotropic covariance

Spiked covariance, \( R = 10 \)

For spiked covariance:  
1. second descent  
2. harmless interpolation  
3. good generalization
For fixed interpolator...
For fixed interpolator…

varying distribution: covariance “spikiness”

test error

bias

variance

overparameterization
For fixed interpolator...

- test error
- bias
- variance

Varying distribution: covariance "spikiness"

- Overparameterization
For fixed interpolator...

For a fixed distribution (e.g. isotropic), different algorithms → different interpolators
how do bias and variance behave?
Plan today…

**Part I:** For linear regression, we discuss how

• variance can decay as overparameterization increases (simple math)
• Two factors can govern variance decay vs. bias increase
  • For fixed interpolator, certain problem instances/distributions are more benign
  • For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the

• effect of loss function choices
• implicit bias of optimization algorithms for neural networks
• generalization of neural networks on noisy, high-dimensional data
Implicit bias → inductive bias

opt. algorithm
minimizing loss
Implicit bias $\rightarrow$ inductive bias

opt. algorithm minimizing loss

has implicit bias towards

certain interpolator
Implicit bias $\rightarrow$ inductive bias

- Optimal algorithm minimizing loss has implicit bias towards certain interpolator

- E.g., 1st order method on $||y - Xw||^2_2$
Implicit bias → inductive bias

Opt. algorithm minimizing loss

Towards certain interpolator

E.g. 1st order method on $\|y - Xw\|_2^2$

E.g. for $p \in [1,2]$

$\hat{w}_p = \arg\min_w \|w\|_p$

s.t. $y = Xw$
Implicit bias $\rightarrow$ inductive bias

- Optimal algorithm minimizing loss has implicit bias towards certain interpolator has certain strength of inductive bias towards certain structure

- E.g. 1st order method on $||y - Xw||_2^2$

- E.g. for $p \in [1,2]$
  \[ \hat{w}_p = \text{argmin}_w ||w||_p \]
  \[ s.t. y = Xw \]
Implicit bias $\rightarrow$ inductive bias

opt. algorithm minimizing loss

\[ \text{has implicit bias towards} \]

certain interpolator

\[ \text{has certain strength of inductive bias towards} \]

certain structure

e.g. 1st order method on $\|y - Xw\|_2^2$

\[ \hat{w}_p = \arg\min_w \|w\|_p \]

\[ s.t. y = Xw \]

e.g. for $p \in [1,2]$

e.g. sparsity, invariances
Implicit bias → inductive bias

opt. algorithm minimizing loss

has implicit bias towards

e.g. 1st order method

on \( \|y - Xw\|_2^2 \)

has certain interpolator towards

e.g. for \( p \in [1,2] \)
\[ \hat{w}_p = \arg\min_w \|w\|_p \]
\[ s.t. y = Xw \]

has certain strength of inductive bias towards

e.g. sparsity, invariances

has certain structure
Implicit bias $\rightarrow$ inductive bias

opt. algorithm minimizing loss

has implicit bias towards
certain interpolator

has certain strength of inductive bias towards
certain structure

e.g. 1st order method on $\|y - Xw\|_2^2$

e.g. for $p \in [1,2]$

$\hat{w}_p = \arg\min_w \|w\|_p$

$s.t. y = Xw$

e.g. sparsity, invariances

Next: Recall how as $p \rightarrow 1$ has an inductive bias towards sparse solutions
Recall: Inductive bias for sparse linear models

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with **sparse** $w^*$, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$
Recall: Inductive bias for sparse linear models

Fixed distribution: \( y_i = \langle w^*, x_i \rangle + \xi_i \) with \textbf{sparse} \( w^* \), i.e. \( |w|_0 = k \ll d \), i.i.d. noise \( \xi_i \) and \( x_i \sim N(0, I) \)

- \( \|w\|_0 - \)norm encourages sparsity \( \rightarrow \) aligns with \( w^* \) structure (strong inductive bias)
- \( \|w\|_2 - \)norm \( \rightarrow \) does not restrict search space in right way! (weak inductive bias)

Subspace of all linear interpolators \( \{ w: Xw = y = Xw^* \} \)

For noiseless \( \xi_i = 0 \)
Recall: Inductive bias for sparse linear models

Fixed distribution: \( y_i = \langle w^*, x_i \rangle + \xi_i \) with \textbf{sparse} \( w^* \), i.e. \( \|w\|_0 = k \ll d \), i.i.d. noise \( \xi_i \) and \( x_i \sim N(0, I) \)

\[
\text{Min-} \ell_p \text{-norm interpolation } \hat{w}_p = \arg\min_w \|w\|_p \text{ s.t. } y = Xw
\]

Subspace of all linear interpolators \( \{w: Xw = y = Xw^*\} \)

For noiseless \( \xi_i = 0 \)
Recall: Inductive bias for sparse linear models

Fixed distribution: \( y_i = \langle w^*, x_i \rangle + \xi_i \) with **sparse** \( w^* \), i.e. \( |w|_0 = k \ll d \), i.i.d. noise \( \xi_i \) and \( x_i \sim N(0, I) \)

\[
\text{Min-} \ell_p \text{-norm interpolation } \hat{w}_p = \arg\min_w |w|_p \text{ s.t. } y = Xw
\]

- small \( |w|_1 \)-norm encourages sparsity → aligns with \( w^* \) structure (**strong inductive bias**)

subspace of all linear interpolators \( \{w: Xw = y = Xw^*\} \)

for noiseless \( \xi_i = 0 \)
Recall: Inductive bias for sparse linear models

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with sparse $w^*$, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0,I)$

Min-$\ell_p$-norm interpolation $\hat{w}_p = \text{argmin}_w |w|_p \text{ s.t. } y = Xw$

- small $|w|_1$-norm encourages sparsity → aligns with $w^*$ structure (strong inductive bias)
- small $|w|_2$-norm → does not restrict search space in right way! (weak inductive bias)

Subspace of all linear interpolators $\{w: Xw = y = Xw^*\}$

for noiseless $\xi_i = 0$
Recall: small $\ell_1$-norm $\rightarrow$ small statistical bias

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with \textbf{sparse} $w^*$, i.e. $||w||_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$ for $i = 1, \ldots, n$ samples and input and parameter dimension $d \gg n$
Recall: small $\ell_1$-norm $\rightarrow$ small statistical bias

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with **sparse** $w^*$, i.e. $\|w\|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$ for $i = 1, \ldots, n$ samples and input and parameter dimension $d \gg n$

Noiseless
$y = Xw^*$

Basis pursuit: $\hat{w}_1 = \text{argmin}_w \|w\|_1$ s.t. $y = Xw$

Perfect recovery
w.h.p. for $n \sim k \log d$
Recall: small $\ell_1$-norm $\rightarrow$ small statistical bias

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with \textbf{sparse} $w^*$, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$ for $i = 1, \ldots, n$ samples and input and parameter dimension $d \gg n$

- Noiseless $y = Xw^*$
  - Basis pursuit: $\hat{w}_1 = \text{argmin}_w \|w\|_1$ s.t. $y = Xw$

- Noisy $y = Xw^* + \xi$
  - Lasso: $\hat{w}_\lambda = \text{argmin}_w \|y - Xw\|_2^2 + \lambda \|w\|_1$

Perfect recovery w.h.p. for $n \sim k \log d$

when observations are noisy

Estimation error achieves minimax optimal rate $O\left(\frac{k \log d}{n}\right)$ for best $\lambda$

e.g. BP: [Candes, Tao ’05, Donoho ‘06], Lasso: [Bunea, Tsybakov, Wegkamp ’07, vandeGeer ‘08], [Wainwright ‘09]
Recall: small $\ell_1$-norm $\rightarrow$ small statistical bias

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with \textbf{sparse $w^*$}, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$ for $i = 1, \ldots, n$ samples and input and parameter dimension $d \gg n$

\begin{align*}
\text{Noiseless } & \quad y = Xw^* \\
\text{Basis pursuit: } & \quad \hat{w}_1 = \text{argmin}_w |w|_1 \text{ s.t. } y = Xw \\
& \quad \downarrow \text{when observations are noisy} \\
\text{Noisy } & \quad y = Xw^* + \xi \\
\text{Lasso: } & \quad \hat{w}_\lambda = \text{argmin}_w |y - Xw|_2^2 + \lambda |w|_1 \\
& \quad \downarrow \text{Estimation error achieves minmax optimal rate}
\end{align*}

$p = 1$ has a strong inductive bias towards sparse solutions $\rightarrow$ small \textit{statistical} bias!

Perfect recovery w.h.p. for $n \sim k \log d$

\[ O \left( \frac{k \log d}{n} \right) \] for best $\lambda$

e.g. BP: [Candes, Tao ‘05, Donoho ‘06], Lasso: [Bunea, Tsybakov, Wegkamp ‘07, vandeGeer ‘08], [Wainwright ‘09]
Recall: small $\ell_1$-norm $\rightarrow$ small statistical bias

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with sparse $w^*$, i.e. $||w||_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$ for $i = 1, \ldots, n$ samples and input and parameter dimension $d \gg n$

Noiseless $y = Xw^*$

Basis pursuit: $\hat{w}_1 = \arg\min_w ||w||_1$ s.t. $y = Xw$

Perfect recovery w.h.p. for $n \sim k \log d$

Noisy $y = Xw^* + \xi$

Lasso: $\hat{w}_\lambda = \arg\min_w ||y - Xw||_2^2 + \lambda ||w||_1$

when observations are noisy

Estimation error achieves optimal minimax rate $O \left( \frac{k \log d}{n} \right)$ for best $\lambda$

Previously unknown: prediction/estimation error of min-$\ell_1$ interpolation for noisy data

e.g. BP: [Candes, Tao ‘05, Donoho ‘06], Lasso: [Bunea, Tsybakov, Wegkamp ‘07, vandeGeer ‘08], [Wainwright ‘09]
For fixed distribution…

- **Overparameterization**
- **Test Error**
- **Bias**
- **Variance**

Varying interpolator: strength of inductive bias

Overparameterization
For fixed distribution…

When interpolating noise, how strong of an inductive bias leads to good generalization.

varying interpolator: strength of inductive bias

overparameterization

bias

variance

test error

overparameterization
Inductive bias for noisy sparse linear models

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with sparse $w^*$, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$

Min-$\ell_p$-norm interpolation $\hat{w}_p = \arg\min_w |w|_p$ s.t. $y = Xw$

Subspace of all linear interpolators
$
\{ w: Xw = y = Xw^* + \xi \}$

for i.i.d noise $\xi_i$
Inductive bias for noisy sparse linear models

Fixed distribution: \( y_i = \langle w^*, x_i \rangle + \xi_i \) with 

\textbf{sparse} \( w^* \), i.e. \( |w|_0 = k \ll d \), i.i.d. noise \( \xi_i \) and \( x_i \sim N(0, I) \)

\[
\text{Min-} \ell_p \text{-norm interpolation } \hat{w}_p = \arg\min_w |w|_p \text{ s.t. } y = Xw
\]

- small \( |w|_1 \)-norm encourages sparsity \( \Rightarrow \) aligns with \( w^* \) structure \( \text{(strong inductive bias)} \)

subspace of all linear interpolators 
\[ \{w: Xw = y = Xw^* + \xi\} \]

for i.i.d noise \( \xi_i \)
Inductive bias for noisy sparse linear models

Fixed distribution: \( y_i = \langle w^*, x_i \rangle + \xi_i \) with \textbf{sparse} \( w^* \), i.e. \( |w|_0 = k \ll d \), i.i.d. noise \( \xi_i \) and \( x_i \sim N(0, I) \)

\[
\text{Min-}\ell_p\text{-norm interpolation } \hat{w}_p = \arg\min_w |w|_p \text{ s.t. } y = Xw
\]

- small \( |w|_1 \text{-norm} \) encourages sparsity → aligns with \( w^* \) structure \((\text{strong inductive bias})\)
- small \( |w|_2 \text{-norm} \) → does not restrict search space in right way! \((\text{weak inductive bias})\)

Subspace of all linear interpolators \( \{w: Xw = y = Xw^* + \xi\} \)

for i.i.d noise \( \xi_i \)
Inductive bias for noisy sparse linear models

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with sparse $w^*$, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$

$$\text{Min-}\ell_p\text{-norm interpolation } \hat{w}_p = \arg\min_{w} |w|_p \text{ s.t. } y = Xw$$

- small $|w|_1$-norm encourages sparsity $\rightarrow$ aligns with $w^*$ structure (strong inductive bias)
- small $|w|_2$-norm $\rightarrow$ does not restrict search space in right way! (weak inductive bias)

subspace of all linear interpolators
{$w: Xw = y = Xw^* + \xi$}

for i.i.d noise $\xi_i$
Varying inductive bias via $p \in [1,2]$

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with **sparse** $w^*$, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$

Min-$\ell_p$-norm interpolation $\hat{w}_p = \arg\min_w |w|_p$ s.t. $y = Xw$

- Consider overparameterized regime $d \gg n$, think of $d \propto n^\beta$ with $\beta > 1$ (high-dimensional)
Varying inductive bias via $p \in [1,2]$

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with \textbf{sparse} $w^*$, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0,I)$

Min-$\ell_p$-norm interpolation $\hat{w}_p = \arg\min_w |w|_p$ s.t. $y = Xw$

- Consider overparameterized regime $d \gg n$, think of $d \propto n^\beta$ with $\beta > 1$ (high-dimensional)
- Compare estimators using tight, high-probability, non-asymptotic \textit{statistical rates} of prediction error

$$\mathbb{E}_{x\sim N(0,I)} \left( x^T \hat{w}_p - x^T w^* \right)^2 = \left\| \hat{w}_p - w^* \right\|^2 = \Theta(h(n,d)) \text{ as } n \to \infty \text{ for some function } h \downarrow$$
Varying inductive bias via $p \in [1, 2]$

Fixed distribution: $y_i = \langle w^*, x_i \rangle + \xi_i$ with sparse $w^*$, i.e. $|w|_0 = k \ll d$, i.i.d. noise $\xi_i$ and $x_i \sim N(0, I)$

Min-$\ell_p$-norm interpolation $\hat{w}_p = \arg\min_w |w|_p$ s.t. $y = Xw$

- Consider overparameterized regime $d \gg n$, think of $d \propto n^\beta$ with $\beta > 1$ (high-dimensional)
- Compare estimators using tight, high-probability, non-asymptotic statistical rates of prediction error

$$
\mathbb{E}_{x \sim N(0, I)} \left( x^T \hat{w}_p - x^T w^* \right)^2 = \left\| \hat{w}_p - w^* \right\|^2 = \Theta(h(n, d)) \text{ as } n \to \infty \text{ for some function } h \downarrow
$$

strong inductive bias towards sparsity

$p=1$

no inductive bias towards sparsity

$p=2$
Strong inductive bias: $p = 1$

- Tight bounds for adversarial noise vectors $\boldsymbol{\xi}$ but $O(\sigma \#)$ for $\boldsymbol{\xi}$ i.i.d. with variance $\sigma^2$.
- Lower bound for i.i.d. noise for sub-Gaussians $\Omega_{9!}$.
- Tight bounds for i.i.d. noise for Gaussian covariates $\Omega_{9!}$.
- $O_{9!}$ for $d \approx n_0$ with $\beta > 1$ we obtain the rate $\Theta(1)$.

Inconsistent but harmless interpolation.
Strong inductive bias: \( p = 1 \)
Strong inductive bias: $p = 1$

- Tight bounds for adversarial noise vectors $\xi$ but $O(\sigma^2)$ for $\xi_i$ i.i.d. with variance $\sigma^2$
  
  [Chinot, Loeffler, vandeGeer ‘20], [Wojtaszczyk ‘10]

- Tight bounds for i.i.d. noise for sub-Gaussians $\Omega_9$:

  $\sigma^2$

  [Muthukumar, Vodrahalli, Subramanian, and Sahai ‘20]

- Tight bounds for i.i.d. noise for Gaussian covariates $\Omega_9$:

  $\sigma^2$

  [Wang, Donhauser, Yang ‘22]

For $d \approx n_0$ with $\beta > 1$ we obtain the rate $\Theta(0)$ but decreasing statistical bias

\[\text{Inconsistent but harmless interpolation}\]
Strong inductive bias: $p = 1$

- Tight bounds for adversarial noise vectors $\xi$ but $O(\sigma^2)$ for $\xi_i$ i.i.d. with variance $\sigma^2$
  [Chinot, Loeffler, vandeGeer ’20], [Wojtaszczyk ’10]

- Lower bound for i.i.d. noise for sub-Gaussians $\Omega\left(\frac{\sigma^2}{\log(\frac{d}{n})}\right)$
  [Muthukumar, Vodrahalli, Subramanian, and Sahai ’20]

$p=1$ decreasing statistical bias $\Theta(1)$

$p=2$ but inconsistent, harmless interpolation
Strong inductive bias: $p = 1$

- Tight bounds for adversarial noise vectors $\xi$ but $O(\sigma^2)$ for $\xi_i$ i.i.d. with variance $\sigma^2$ [Chinot, Loeffler, vandeGeer ’20], [Wojtaszczyk ’10]
- Lower bound for i.i.d. noise for sub-Gaussians $\Omega\left(\frac{\sigma^2}{\log(d/n)}\right)$ [Muthukumar, Vodrahalli, Subramanian, and Sahai ’20]
- Tight bounds for i.i.d. noise for Gaussian covariates $\frac{\sigma^2}{\log(d/n)} + O\left(\frac{\sigma^2}{\log^{3/2}(d/n)}\right)$ [Wang, Donhauser, Yang ’22]

for $d = n^\beta$ with $\beta > 1$ we obtain the rate $\Theta\left(\frac{1}{(\beta+1)\log n}\right)$

\[\text{decreasing statistical bias}\]

\[\text{Inconsistent but harmless interpolation}\]
**Strong inductive bias: \( p = 1 \)**

- Tight bounds for adversarial noise vectors \( \xi \) but \( O(\sigma^2) \) for \( \xi_i \) i.i.d. with variance \( \sigma^2 \)  
  [Chinot, Loeffler, vandeGeer ’20], [Wojtaszczyk ’10]

- Lower bound for i.i.d. noise for sub-Gaussians \( \Omega \left( \frac{\sigma^2}{\log(d/n)} \right) \)  
  [Muthukumar, Vodrahalli, Subramanian, and Sahai ’20]

- Tight bounds for i.i.d. noise for Gaussian covariates \( \frac{\sigma^2}{\log(d/n)} + O \left( \frac{\sigma^2}{\log^{3/2}(d/n)} \right) \)  
  [Wang, Donhauser, Yang ’22]

  for \( d = n^\beta \) with \( \beta > 1 \) we obtain the rate \( \Theta \left( \frac{1}{(\beta-1)\log n} \right) \)

---

**Consistent**

but harmful interpolation: opt. regularized \( O \left( \frac{k \log n}{n} \right) \)

\[ \text{rate } \Theta \left( \frac{1}{\log n} \right) = \Theta(1) \]

**Inconsistent**

but harmless interpolation

\[ \text{rate } \Theta(1) \]
The problem of $p = 1$ lies in the variance...

For $p = 1$ and $k = 1$, "sensitivity to noise" and variance larger than for $p = 2$
The problem of $p = 1$ lies in the variance...

For $p = 1$ and $k = 1$, “sensitivity to noise” and variance larger than for $p = 2$

for $d = 5000, n = 100$
The problem of $p = 1$ lies in the variance...

For $p = 1$ and $k = 1$, “sensitivity to noise” and variance larger than for $p = 2$ for $d = 5000, n = 100$

as overparameterization increases, variance decay is slower for $p = 1$ than for $p = 2$!
A bias-variance trade-off for $p \in [1,2]$

Min-$\ell_p$-norm interpolation $\widehat{w}_p = \arg\min_w ||w||_p$ s. t. $y = Xw$

- strong inductive bias towards sparsity
  - $p=1$
  - test error
  - variance
  - statistical bias
- no inductive bias towards sparsity
  - $p=2$
  - rate $\Theta(1)$
A bias-variance trade-off for $p \in [1,2]$ 

Min-$\ell_p$-norm interpolation $\hat{w}_p = \arg\min_w |w|_p$ s. t. $y = Xw$
A bias-variance trade-off for $p \in [1,2]$

$$\text{Min-}$l_p\text{-norm interpolation } \hat{w}_p = \arg\min_w ||w||_p \text{ s.t. } y = Xw$$

- **strong inductive bias towards sparsity**
  - $p=1$
  - rate $\theta\left(\frac{1}{\log n}\right)$
- **no inductive bias towards sparsity**
  - $p=2$
  - rate $\theta(1)$
- **statistical bias**
- **variance**

Trade-off between bias and variance for interpolators via strength of inductive bias!
A bias-variance trade-off for $p \in [1,2]$ 

Min-$\ell_p$-norm interpolation $\hat{w}_p = \arg\min_w |w|_p$ s.t. $y = Xw$ 

Trade-off between bias and variance for interpolators via strength of inductive bias!
A bias-variance trade-off for $p \in [1,2]$

Min-$\ell_p$-norm interpolation $\hat{w}_p = \arg\min_w \|w\|_p$ s.t. $y = Xw$

Which rates?

- $p=1$ rate $\Theta\left(\frac{1}{\log n}\right)$
- $p=2$ rate $\Theta(1)$

- strong inductive bias towards sparsity
- no inductive bias towards sparsity

Trade-off between bias and variance for interpolators via strength of inductive bias!
Tight bounds for $p \in [1, 2]$
Tight bounds for $p \in [1, 2]$

We plot $\alpha$ where $\left\| \hat{w}_p - w^* \right\|^2 = \tilde{\Theta}(n^\alpha)$ w.h.p.
Tight bounds for $p \in [1, 2]$

- minimax optimal rate, e.g. for (best) regularized estimator with $p = 1$ (LASSO) $||\hat{w}_p - w^*||^2 = \tilde{\Theta}(n^\alpha)$ w.h.p.

We plot $\alpha$ where $||\hat{w}_p - w^*||^2 = \tilde{\Theta}(n^\alpha)$ w.h.p.

$||\hat{w}_{\lambda} - w^*||^2 = \tilde{\Theta}(n^{-1}) \rightarrow \alpha = -1$
Tight bounds for $p \in [1, 2]$ 

- minimax optimal rate, e.g. for (best) regularized estimator with $p = 1$ (LASSO) $||\hat{w}_\lambda - w^*||^2 = \tilde{O}(n^{-1}) \rightarrow \alpha = -1$

- Interpolators with $p = 1, 2$: $||\hat{w}_p - w^*||^2 = \tilde{O}(1) \rightarrow \alpha = 0$

For $p \geq 1$: [Wang, Donhauser, Yang ‘22]
Tight bounds for $p \in [1, 2]$

- minimax optimal rate, e.g. for (best) regularized estimator with $p = 1$ (LASSO)
  \[ ||\hat{w}_p - w^*||^2 = \tilde{\Theta}(n^{-1}) \rightarrow \alpha = -1 \]

- Interpolators with $p = 1, 2$:
  \[ ||\hat{w}_p - w^*||^2 = \tilde{\Theta}(1) \rightarrow \alpha = 0 \]

- Interpolators for $p \in (1, 2)$:
  \[ ||\hat{w}_p - w^*||^2 = \tilde{\Theta}(n^\alpha) \text{ with } \alpha < 0 \]

We plot $\alpha$ where $||\hat{w}_p - w^*||^2 = \tilde{\Theta}(n^\alpha)$ w.h.p.

For $p \in [1,2)$: [Wang, Donhauser, Yang ‘22], [Donhauser, Ruggeri, Stojanovic, Yang ‘22]
Tight bounds for $p \in [1, 2]$ for $p \in [1, 2)$: [Wang, Donhauser, Yang '22], [Donhauser, Ruggeri, Stojanovic, Yang '22]

We plot $\alpha$ where $\|\hat{w}_p - w^*\|^2 = \tilde{\Theta}(n^\alpha)$ w.h.p.

"second" descent: decrease due to variance decay

For $p \in [1, 2)$: [Wang, Donhauser, Yang ‘22], [Donhauser, Ruggeri, Stojanovic, Yang ‘22]
Tight bounds for $p \in [1, 2]$

We plot $\alpha$ where $\|\hat{w}_p - w^*\|^2 = \tilde{\Theta}(n^\alpha)$ w.h.p.

For $p \in [1, 2)$: [Wang, Donhauser, Yang ‘22], [Donhauser, Ruggeri, Stojanovic, Yang ‘22]
Tight bounds for $p \in [1, 2]$

For $p \in [1,2)$: [Wang, Donhauser, Yang ‘22], [Donhauser, Ruggeri, Stojanovic, Yang ‘22]

We plot $\alpha$ where $\|\hat{w}_p - w^*\|^2 = \tilde{\Theta}(n^\alpha)$ w.h.p.

"second" descent: decrease due to variance decay

eventual increase due to bias increase

good generalization & $\approx$ harmless interpolation
A new bias-variance trade-off for interpolators

Min-$\ell_p$-norm interpolation $\hat{w}_p = \arg\min_w ||w||_p$ s.t. $y = Xw$

- strong inductive bias towards sparsity
  - $p=1$ rate $\theta\left(\frac{1}{\log n}\right)$
- statistical bias
  - $1 < p < 2$ rate $\theta(n^\alpha)$ $-1 < \alpha < 0$
- variance
  - $p=2$ rate $\theta(1)$
- no inductive bias towards sparsity

Take-away: medium strength of inductive bias is best when interpolating noise
How transferable is this “new” intuition?

How transferable is this “new” intuition?


Synthetic experiment:
Isotropic Gaussians with $d \sim 5000, n \sim 100$
How transferable is this “new” intuition?


![Diagram showing synthetic experiment results](image)

**Synthetic experiment:**
Isotropic Gaussians with $d \sim 5000, n \sim 100$

[Stojanovic, Donhauser, Yang ‘22], [Donhauser, Ruggeri, Stojanovic, Yang ‘22]
How transferable is this “new” intuition?


Real-world experiment:
Leukemia dataset with $d \sim 7000, n \sim 70$

Synthetic experiment:
Isotropic Gaussians with $d \sim 5000, n \sim 100$
How transferable is this “new” intuition?


[open: theory is still incomplete and restricted to Gaussians!]
How transferable is this “new” intuition?


  - open: theory is still incomplete and restricted to Gaussians!

- Intuition carries over to high-dimensional kernel learning with convolutional kernels where bias and variance vary with inductive bias [Aerni, Milanta, Donhauser, Yang ’23].
How transferable is this “new” intuition?


  open: theory is still incomplete and restricted to Gaussians!

- Intuition carries over to high-dimensional kernel learning with convolutional kernels where bias and variance vary with inductive bias [Aerni, Milanta, Donhauser, Yang ‘23].

- Preliminary experiments for neural networks also suggest this behavior for rotational invariance and filter size…
Nonlinear structure: Rotational invariance for WideResNet

- Satellite images (EuroSAT) to be classified in terms of type of land usage

- Strength of rotational invariance via "amount of" data augmentation

[Aerni, Milanta, Donhauser, Yang ’23]
Nonlinear structure: Rotational invariance for WideResNet

- Satellite images (EuroSAT) to be classified in terms of type of land usage
- Strength of rotational invariance via "amount of" data augmentation

[Aerni, Milanta, Donhauser, Yang ’23]
Nonlinear structure: Rotational invariance for WideResNet

- Satellite images (EuroSAT) to be classified in terms of type of land usage

- Strength of rotational invariance via “amount of” data augmentation

Confirmed: medium strength of inductive bias is best when interpolating noise

[Aerni, Milanta, Donhauser, Yang ‘23]
Open: How transferable is this “new” intuition?


- Intuition carries over to high-dimensional kernel learning with convolutional kernels where bias and variance vary with inductive bias [Aerni, Milanta, Donhauser, Yang ’23].

- Preliminary experiments for neural networks also suggest this behavior for rotational invariance and filter size.

* open: theory is still incomplete and restricted to Gaussians!

** open: comprehensive experimental NN study!
Plan today…

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the
- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data
### Classification-vs-regression: A tale of two loss functions

<table>
<thead>
<tr>
<th></th>
<th>0-1 loss</th>
<th>Squared loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Squared loss</td>
<td></td>
<td>Regression</td>
</tr>
</tbody>
</table>
Classification-vs-regression: A tale of two loss functions

<table>
<thead>
<tr>
<th></th>
<th>0-1 loss</th>
<th>Squared loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic loss</td>
<td>Classification, most popular</td>
<td></td>
</tr>
<tr>
<td>Squared loss</td>
<td>Classification, less popular</td>
<td>Regression</td>
</tr>
</tbody>
</table>
Differences in training loss functions

• Closed-form
• Linked to MLE under additive noise

Squared loss

Gradient descent
(Folklore, see e.g. Engl et al 1996)

Minimum-l2-norm interpolation

\[ \hat{\theta}_2 = \arg \min \| \theta \|_2 \]
subject to
\[ X_i^T \theta = Y_i, i \in [n]. \]

- Closed-form
- Linked to MLE under additive noise
Differences in training loss functions

**Logistic loss**

\[ \hat{\theta}_{SVM} = \arg \min \|\theta\|_2 \]

subject to

\[ Y_i \cdot X_i^T \theta \geq 1, i \in [n]. \]

**Squared loss**

\[ \hat{\theta}_2 = \arg \min \|\theta\|_2 \]

subject to

\[ X_i^T \theta = Y_i, i \in [n]. \]

- Closed-form
- Linked to MLE under additive noise
Differences in training loss functions

**Logistic loss**
- Not closed-form
- Linked to MLE under logistic noise

\[ \hat{\theta}_{\text{SVM}} = \arg \min \| \theta \|_2 \text{ subject to } Y_i \cdot X_i^T \theta \geq 1, i \in [n]. \]

**Squared loss**
- Closed-form
- Linked to MLE under additive noise

\[ \hat{\theta}_2 = \arg \min \| \theta \|_2 \text{ subject to } X_i^T \theta = Y_i, i \in [n]. \]

**Gradient descent**
(Soudry et al, Ji & Telgarsky, 2018)
(Folklore, see e.g. Engl et al 1996)

**Minimum-I2-norm interpolation**
Differences in test loss functions

Regression: Test MSE
\[ \mathcal{E}_{\text{MSE}} = \mathbb{E} \left[ (X^\top (\hat{\theta} - \theta^*))^2 \right] \]

Classification: Test 0-1 error
\[ \mathcal{E}_{0-1} = \mathbb{E} \left[ \mathbb{I}[\text{sgn}(X^\top \hat{\theta}) \neq \text{sgn}(X^\top \theta^*]) \right] \]
Differences in test loss functions

**Regression: Test MSE**

\[ \mathcal{E}_{\text{MSE}} = \mathbb{E} \left[ (X^\top (\hat{\theta} - \theta^*))^2 \right] \]

**Classification: Test 0-1 error**

\[ \mathcal{E}_{0-1} = \mathbb{E} \left[ \mathbb{I}[\text{sgn}(X^\top \hat{\theta}) \neq \text{sgn}(X^\top \theta^*)] \right] \]

**Two core challenges when analyzing classification:**

1. Hard-margin SVM does not have a closed-form solution, unlike minimum-l2-norm interpolation
2. 0-1 error metric challenging to sharply analyze as compared to MSE
Plan today…

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the
- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data
One analysis path for l2, step 1: showing that $\text{SVM} = \text{interpolation}$

Fourier features on 1-dimensional data, isotropic covariance

$n = 32$, 
$d = 1000$
One analysis path for $l_2$, step 1: showing that $\text{SVM} = \text{interpolation}$

Fourier features on 1-dimensional data, isotropic covariance

$n = 32$, $d = 1000$

**Result** (Hsu, Muthukumar and Xu 2021): hard margin SVM = minimum-$l_2$-norm interpolation on binary labels in spiked covariance ensemble if $d \gg n \log n$ and $R \ll \frac{d}{n}$

Conditions for general anisotropic covariances also provided in terms of "effective ranks" in Hsu et al (2021)
One analysis path for l2, step 1: showing that \( \text{SVM} = \text{interpolation} \)

Fourier features on 1-dimensional data, isotropic covariance

\[ n = 32, \quad d = 1000 \]

**Result** (Hsu, Muthukumar and Xu 2021): **hard margin SVM** = minimum-l2-norm interpolation on binary labels in spiked covariance ensemble if \( d \gg n \log n \) and \( R \ll \frac{d}{n} \)

**Implication:** SVM has a closed-form expression, can be more easily analyzed!

Conditions for general anisotropic covariances also provided in terms of “effective ranks” in Hsu et al (2021)
One analysis path for l2, step 2: analyzing 0-1 error of interpolator

Spiked covariance: $(n, d, k, R)$

$\Sigma = \text{diag}(\Lambda) = \ldots$

Limiting test error, $n \to \infty$

Ratio $R \gg 1$

Feature index $(j)$

Feature magnitude

(sparsity level)

$(k \ll n)$

$(d \gg n)$

Ratio $R$

[Muthukumar, Narang, Subramaniam, Belkin, Hsu, Sahai JMLR’21]
One analysis path for l2, step 2: analyzing 0-1 error of interpolator

\[ \Sigma = \text{diag}(\Lambda) = \]

Spiked covariance: \((n, d, k, R)\)

Feature magnitude

Feature index \((j)\)

(sparsity level) \(k \ll n\)

Ratio \(R \gg 1\)

Regression and classification work \(\mathcal{E}_{\text{MSE}} \to 0, \mathcal{E}_{0-1} \to 0\)

Limiting test error, \(n\)

Ratio \(R\)

\[
\frac{d}{n}
\]

[\text{Muthukumar, Narang, Subramaniam, Belkin, Hsu, Sahai JMLR'21}]
One analysis path for l2, step 2: analyzing 0-1 error of interpolator

\[
\Sigma = \text{diag}(\Lambda) =
\]

Spiked covariance: \((n, d, k, R)\)

Classification works, regression does not!

\[
\mathcal{E}_{\text{MSE}} \to \|\theta^*\|_2^2
\]

\[
\mathcal{E}_{0-1} \to 0
\]

Regression and classification work

\[
\mathcal{E}_{\text{MSE}} \to 0, \mathcal{E}_{0-1} \to 0
\]

[\text{Muthukumar, Narang, Subramaniam, Belkin, Hsu, Sahai JMLR'21}]
One analysis path for l2, step 2: analyzing 0-1 error of interpolator

\[ \Sigma = \text{diag}(\Lambda) = \]

Spiked covariance: \((n, d, k, R)\)

Limiting test error, \(n\)

Neither work

\[ \mathcal{E}_{\text{MSE}} \rightarrow \|\theta^*\|_2^2 \]
\[ \mathcal{E}_{0-1} \rightarrow 1/2 \]

Classification works, regression does not!

\[ \mathcal{E}_{\text{MSE}} \rightarrow \|\theta^*\|_2^2 \]
\[ \mathcal{E}_{0-1} \rightarrow 0 \]

Regression and classification work

\[ \mathcal{E}_{\text{MSE}} \rightarrow 0, \mathcal{E}_{0-1} \rightarrow 0 \]

[\text{Muthukumar, Narang, Subramaniam, Belkin, Hsu, Sahai JMLR’21}]
Takeaways for classification with l2-minimizing solutions

- Different training loss functions could yield similar or even identical solutions
Takeaways for classification with l2-minimizing solutions

- Different training loss functions could yield similar or even identical solutions

- Classification 0-1 test loss is much more benign than regression MSE; so l2-inductive bias could work better for classification tasks
Plan today…

**Part I:** For linear regression, we discuss how
- variance can decay as overparameterization increases (simple math)
- Two factors can govern variance decay vs. bias increase
  - For fixed interpolator, certain problem instances/distributions are more benign
  - For fixed problem instance, certain interpolators generalize better

**Part II:** For classification, we discuss the
- effect of loss function choices
- implicit bias of optimization algorithms for neural networks
- generalization of neural networks on noisy, high-dimensional data
Benign overfitting in neural networks

• Most theoretical works on benign overfitting focus on linear/kernel setting.
• We’ll discuss recent works in neural networks and open questions.
Benign overfitting in neural networks

- Most theoretical works on benign overfitting focus on linear/kernel setting.
- We’ll discuss recent works in neural networks and open questions.
- Notably: all results on benign overfitting in neural nets require ambient dimension $d \gg n$
- Very unsatisfying: neural nets can be overparameterized in $d \ll n$ regime, when is overfitting benign in this setting?
Which estimators do we care about?

<table>
<thead>
<tr>
<th>Model</th>
<th>Algorithm</th>
<th>Setting</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>Gradient descent</td>
<td>Classification</td>
<td>$\ell_2$ max-margin</td>
</tr>
<tr>
<td>Linear</td>
<td>Gradient descent</td>
<td>Regression</td>
<td>$\ell_2$ min-norm interpolator</td>
</tr>
<tr>
<td>Linear</td>
<td>Adaboost</td>
<td>Classification</td>
<td>$\ell_1$ max-margin</td>
</tr>
<tr>
<td>Linear</td>
<td>Basis pursuit</td>
<td>Regression</td>
<td>$\ell_1$ min-norm interpolator</td>
</tr>
<tr>
<td>Neural nets</td>
<td>Gradient descent</td>
<td>Classification</td>
<td>?</td>
</tr>
<tr>
<td>Neural nets</td>
<td>Gradient descent</td>
<td>Regression</td>
<td>?</td>
</tr>
</tbody>
</table>

• Next: implicit bias of GD in neural net classification.
• After: "trajectory analysis", directly analyzing properties of neural networks trained by GD, Telgarsky ’13, Soudry-Hoffer-Nacson-Gunasekar-Srebro ’18, Ji-Telgarsky ’18, …
Which estimators do we care about?

<table>
<thead>
<tr>
<th>Model</th>
<th>Algorithm</th>
<th>Setting</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>Gradient descent</td>
<td>Classification</td>
<td>$\ell_2$ max-margin</td>
</tr>
<tr>
<td>Linear</td>
<td>Gradient descent</td>
<td>Regression</td>
<td>$\ell_2$ min-norm interpolator</td>
</tr>
<tr>
<td>Linear</td>
<td>Adaboost</td>
<td>Classification</td>
<td>$\ell_1$ max-margin</td>
</tr>
<tr>
<td>Linear</td>
<td>Basis pursuit</td>
<td>Regression</td>
<td>$\ell_1$ min-norm interpolator</td>
</tr>
<tr>
<td>Neural nets</td>
<td>Gradient descent</td>
<td>Classification</td>
<td>?</td>
</tr>
<tr>
<td>Neural nets</td>
<td>Gradient descent</td>
<td>Regression</td>
<td>?</td>
</tr>
</tbody>
</table>

- Next: implicit bias of GD in neural net classification.
- After: “trajectory analysis”, directly analyzing properties of neural nets trained by GD

Telgarsky’13, Soudry-Hoffer-Nacson-Gunasekar-Srebro’18, Ji-Telgarsky’18, …
Implicit bias in neural networks

- Which interpolators do we care about for neural nets?
- We’ll focus on classification tasks, training by GD/GF on logistic loss.
  - Very little known about implicit bias of GD for neural nets in regression setting.

**Theorem**

For large class of neural nets, if $\theta(t)$ reaches a small enough loss, then $\theta(t)$ converges in direction to a first-order stationary point (KKT point) of the $\ell^2$-max margin problem,

$$
\min_{\theta} \|\theta\|_2 \quad \text{s.t.} \quad y_i f(x_i; \theta) \geq 1, \quad \forall i \in [n]. 
$$

- KKT point does not imply even local optimality in general.
- In general, very little is known about KKT points of (1).

Lyu-Li'20, Ji-Telgarsky'20
Implicit bias in neural networks

- Which interpolators do we care about for neural nets?
- We’ll focus on classification tasks, training by GD/GF on logistic loss.
  - Very little known about implicit bias of GD for neural nets in regression setting.

**Theorem**

For large class of neural nets, if GD/GF $\theta(t)$ reaches a small enough loss, then $\theta(t)$ converges in direction to a first-order stationary point (KKT point) of the $\ell^2$-max margin problem,

$$\min_{\theta} \|\theta\|^2 \quad \text{s.t.} \quad y_i f(x_i; \theta) \geq 1, \ \forall i \in [n].$$  \hspace{1cm} (1)
Implicit bias in neural networks

- Which interpolators do we care about for neural nets?
- We’ll focus on classification tasks, training by GD/GF on logistic loss.
  - Very little known about implicit bias of GD for neural nets in regression setting.

**Theorem**

For large class of neural nets, if GD/GF $\theta(t)$ reaches a small enough loss, then $\theta(t)$ converges in direction to a first-order stationary point (KKT point) of the $\ell^2$-max margin problem,

$$
\min_{\theta} \|\theta\|^2 \quad \text{s.t.} \quad y_i f(x_i; \theta) \geq 1, \forall i \in [n].
$$

- KKT point does not imply even local optimality in general.
- In general, very little is known about KKT points of (1).

Lyu-Li’20, Ji-Telgarsky’20
Implicit bias in neural networks

- A setting where we understand KKT points of max-margin: two-layer leaky ReLU nets with nearly-orthogonal data. \( \phi(q) = \max(\gamma q, q) \)
Implicit bias in neural networks

- A setting where we understand KKT points of max-margin: two-layer leaky ReLU nets with nearly-orthogonal data. ($\phi(q) = \max(\gamma q, q)$)

$$f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad a_j \in \{\pm 1/\sqrt{m}\},$$

$$\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|.$$

- Satisfied in many settings w.h.p. when $d \gg n^2$ and $(x_i, y_i) \sim \text{i.i.d.} P$ (e.g., $x \sim N(0, I_d)$)
Implicit bias in neural networks

- A setting where we understand KKT points of max-margin: two-layer leaky ReLU nets with nearly-orthogonal data. \( \phi(q) = \max(\gamma q, q) \)

\[
f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad a_j \in \{\pm 1/\sqrt{m}\},
\]

\[
\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|.
\]

- Satisfied in many settings w.h.p. when \( d \gg n^2 \) and \((x_i, y_i) \overset{i.i.d.}{\sim} P\) (e.g., \( x \sim \mathcal{N}(0, I_d) \))

**Theorem**

Suppose data is nearly orthogonal. If \( \theta \) satisfies KKT conditions for \( \ell^2 \)-max-margin, then \( \exists s_i > 0 \) s.t.

\[
\text{for any } x \in \mathbb{R}^d, \quad \text{sgn}(f(x; \theta)) = \text{sgn}(\langle \sum_{i=1}^{n} s_i y_i x_i, x \rangle),
\]

where \( s_i > 0 \) satisfy \( \max_{i,j} s_i/s_j = O(1) \).
Theorem

Suppose data satisfies $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$. If $\theta$ satisfies KKT conditions for $\ell^2$-max-margin for 2-layer leaky nets, then $\exists s_i > 0$ s.t.

$$\text{for any } x \in \mathbb{R}^d, \quad \text{sgn}(f(x; \theta)) = \text{sgn}(\langle \sum_{i=1}^n s_i y_i x_i, x \rangle),$$

where $s_i > 0$ satisfy $\max_{i,j} s_i / s_j = O(1)$. 

Implicit bias in neural networks

- Although two-layer nets are universal approximators, KKT points for margin maximization have linear decision boundaries under near-orthogonality.
- Decision boundary is very simple, $\approx$ uniform average of data.
- Linear model can capture behavior of non-linear net, trained beyond NTK.

Frei-Vardi-Bartlett-Srebro-Hu'23
Implicit bias in neural networks

Theorem

Suppose data satisfies $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$. If $\theta$ satisfies KKT conditions for $\ell^2$-max-margin for 2-layer leaky nets, then $\exists s_i > 0$ s.t.

$$\text{for any } x \in \mathbb{R}^d, \quad \text{sgn}(f(x; \theta)) = \text{sgn}(\langle \sum_{i=1}^n s_i y_i x_i, x \rangle),$$

where $s_i > 0$ satisfy $\max_{i,j} s_i/s_j = O(1)$.

• Although two-layer nets are universal approximators, KKT points for margin maximization have linear decision boundaries under near-orthogonality.
Implicit bias in neural networks

**Theorem**

Suppose data satisfies $\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|$. If $\theta$ satisfies KKT conditions for $\ell^2$-max-margin for 2-layer leaky nets, then $\exists s_i > 0$ s.t.

$$\text{for any } x \in \mathbb{R}^d, \quad \text{sgn}(f(x; \theta)) = \text{sgn}(\langle \sum_{i=1}^{n} s_i y_i x_i, x \rangle),$$

where $s_i > 0$ satisfy $\max_{i,j} s_i / s_j = O(1)$.

- Although two-layer nets are universal approximators, KKT points for margin maximization have linear decision boundaries under near-orthogonality.
- Decision boundary is very simple, $\approx$ uniform average of data.
Implicit bias in neural networks

**Theorem**

Suppose data satisfies \( \|x_i\|^2 \gg n \max_k \max_{k \neq j} |\langle x_j, x_k \rangle| \). If \( \theta \) satisfies KKT conditions for \( \ell^2 \)-max-margin for 2-layer leaky nets, then \( \exists s_i > 0 \) s.t.

\[
\text{for any } x \in \mathbb{R}^d, \quad \text{sgn}(f(x; \theta)) = \text{sgn}(\langle \sum_{i=1}^n s_i y_i x_i, x \rangle),
\]

where \( s_i > 0 \) satisfy \( \max_i s_i/s_j = O(1) \).

- Although two-layer nets are universal approximators, KKT points for margin maximization have linear decision boundaries under near-orthogonality.
- Decision boundary is very simple, \( \approx \) uniform average of data.
- Linear model can capture behavior of nonlinear net, trained beyond NTK.

Frei-Vardi-Bartlett-Srebro-Hu’23
Benign overfitting of neural nets in mixture model

- KKT points for 2-layer leaky nets $\approx \sum_{i=1}^{n} y_i x_i$, when training data is nearly-orthogonal $\left(\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|\right)$.
Benign overfitting of neural nets in mixture model

- KKT points for 2-layer leaky nets $\approx \sum_{i=1}^{n} y_i x_i$, when training data is nearly-orthogonal $\left( \left\| x_i \right\|^2 \gg n \max_{k \neq j} \left| \langle x_j, x_k \rangle \right| \right)$.

- Near-orthogonality typically holds in low-SNR, $d \gg n$ settings, e.g. mixture model:

  $\tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y}$ w.p. $p$.

- Holds if $\|\mu\| = O(d^{1/2})$ and $d \gg n^2$. 
Benign overfitting of neural nets in mixture model

- KKT points for 2-layer leaky nets $\approx \sum_{i=1}^{n} y_i x_i$, when training data is nearly-orthogonal $(\|x_i\|^2 \gg n \max_{k \neq j} |\langle x_j, x_k \rangle|)$.

- Near-orthogonality typically holds in low-SNR, $d \gg n$ settings, e.g. mixture model:
  
  $\tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y}$ w.p. $p$.

- **Holds** if $\|\mu\| = O(d^{1/2})$ and $d \gg n^2$.
- Following results will only hold in this low-SNR, high-dimensional regime
  - We’ll see consistency is still possible in this setting
Benign overfitting of neural nets in mixture model

\[ \tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \]

**Theorem (informal)**

Suppose labels flipped w.p. \( p < 1/2 \), low SNR and \( d \gg n^2 \). Then w.h.p., any KKT point \( \theta \) of 2-layer leaky ReLU net \( \ell_2 \)-max-margin problem satisfies

\[
\forall k \in [n], \quad y_k = \text{sgn}(f(x_k; \theta)), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta)) \leq p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right).\]"
Benign overfitting of neural nets in mixture model

\[ \tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \]

**Theorem (informal)**

Suppose labels flipped w.p. \( p < \frac{1}{2} \), low SNR and \( d \gg n^2 \). Then w.h.p., any KKT point \( \theta \) of 2-layer leaky ReLU net \( \ell_2 \)-max-margin problem satisfies

\[ \forall k \in [n], \quad y_k = \text{sgn}(f(x_k; \theta)), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta))) \leq p + \exp \left( -\Omega \left( \frac{n\|\mu\|^4}{d} \right) \right). \]

- No dependence on number of neurons in network.
Benign overfitting of neural nets in mixture model

\( \tilde{y} \sim \text{Unif}(\lbrace \pm 1 \rbrace), \quad x = \tilde{y} \mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \)

**Theorem (informal)**

Suppose labels flipped w.p. \( p < \frac{1}{2} \), low SNR and \( d \gg n^2 \). Then w.h.p., any KKT point \( \theta \) of 2-layer leaky ReLU net \( \ell_2 \)-max-margin problem satisfies

\[
\forall k \in [n], \quad y_k = \text{sgn}(f(x_k; \theta)), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta))) \leq p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right).
\]

- No dependence on number of neurons in network.
- **Overfitting**: perfectly fit training data, even though \( \approx pn \) labels are flipped.
Benign overfitting of neural nets in mixture model

\( \tilde{y} \sim \text{Unif}(\{\pm 1\}), \ x = \tilde{y}\mu + z, \ z \sim \mathcal{N}(0, I_d), \ y = -\tilde{y} \text{ w.p. } p. \)

**Theorem (informal)**

Suppose labels flipped w.p. \( p < \frac{1}{2} \), low SNR and \( d \gg n^2 \). Then w.h.p., any KKT point \( \theta \) of 2-layer leaky ReLU net \( \ell_2 \)-max-margin problem satisfies

\[
\forall k \in [n], \ y_k = \text{sgn}(f(x_k; \theta)), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta)) \leq p + \exp\left(-\Omega\left(\frac{n\|\mu\|^4}{d}\right)\right).
\]

- No dependence on number of neurons in network.
- **Overfitting**: perfectly fit training data, even though \( \approx pn \) labels are flipped
- **Benign overfitting**: if \( n\|\mu\|^4 \gg d \), test error \( \approx \) noise rate.
Benign overfitting of neural nets in mixture model

\[ \tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \]

**Theorem (informal)**

Suppose labels flipped w.p. $p < \frac{1}{2}$, low SNR and $d \gg n^2$. Then w.h.p., any KKT point $\theta$ of 2-layer leaky ReLU net $\ell_2$-max-margin problem satisfies

\[
\forall k \in [n], \quad y_k = \text{sgn}(f(x_k; \theta)), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta)) \leq p + \exp \left( -\Omega \left( \frac{n\|\mu\|^4}{d} \right) \right). \]

- No dependence on number of neurons in network.
- **Overfitting**: perfectly fit training data, even though $\approx pn$ labels are flipped
- **Benign overfitting**: if $n\|\mu\|^4 \gg d$, test error $\approx$ noise rate.
- Low-SNR requires $\|\mu\| = O(d^{1/2})$, so results hold for $\|\mu\| = \Theta(d^\varepsilon)$ for $\varepsilon \in (1/4, 1/2)$
Benign overfitting of neural nets in mixture model

\[ \tilde{y} \sim \text{Unif}(\{\pm1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \text{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \]

**Theorem (informal)**

Suppose labels flipped w.p. \( p < \frac{1}{2} \), low SNR and \( d \gg n^2 \). Then w.h.p., any KKT point \( \theta \) of 2-layer leaky ReLU net \( \ell_2 \)-max-margin problem satisfies

\[
\forall k \in [n], \quad y_k = \text{sgn}(f(x_k; \theta)), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta)) \leq p + \exp \left( -\Omega \left( \frac{n\|\mu\|^4}{d} \right) \right). 
\]

- No dependence on number of neurons in network.
- **Benign overfitting**: if \( n\|\mu\|^4 \gg d \), test error \( \approx \) noise rate.
- **Overfitting**: perfectly fit training data, even though \( \approx pn \) labels are flipped.
- Low-SNR requires \( \|\mu\| = O(d^{1/2}) \), so results hold for \( \|\mu\| = \Theta(d^\varepsilon) \) for \( \varepsilon \in (1/4, 1/2) \).
- \( \exp(-\Omega(n\|\mu\|^4/d)) \) is minimax-optimal!
Benign overfitting of neural nets in mixture model

Recall $\text{sgn}(f(x; \theta)) = \text{sgn}(\langle \sum_{i=1}^{n} y_i x_i, x \rangle)$. What does this estimator look like? Since $x_i = \tilde{y}_i \mu + z_i$, 

• Signal and overfitting component balanced to allow both interpolation + generalization.
Benign overfitting of neural nets in mixture model

Recall \( \text{sgn}(f(x; \theta)) = \text{sgn}(\langle \sum_{i=1}^{n} y_i x_i, x \rangle) \). What does this estimator look like? Since \( x_i = \tilde{y}_i \mu + z_i \),

\[
\begin{align*}
\sum_{i=1}^{n} y_i x_i &= \sum_{i \in \text{clean}} \tilde{y}_i (\tilde{y}_i \mu + z_i) + \sum_{i \in \text{noisy}} -\tilde{y}_i (\tilde{y}_i \mu + z_i) \\
&= (|\text{clean}| - |\text{noisy}|) \mu + \sum_{i=1}^{n} \tilde{y}_i z_i \\
&\approx (1 - 2p)n \cdot \mu + \sum_{i=1}^{n} \tilde{y}_i z_i
\end{align*}
\]

- **Signal and overfitting component balanced to allow both interpolation + generalization**
Benign overfitting of neural nets in mixture model

Recall $\text{sgn}(f(x; \theta)) = \text{sgn}\left(\langle \sum_{i=1}^n y_i x_i, x \rangle\right)$. What does this estimator look like? Since $x_i = \tilde{y}_i \mu + z_i$,

$$\sum_{i=1}^n y_i x_i = \sum_{i \in \text{clean}} \tilde{y}_i (\tilde{y}_i \mu + z_i) + \sum_{i \in \text{noisy}} -\tilde{y}_i (\tilde{y}_i \mu + z_i)$$

$$= (|\text{clean}| - |\text{noisy}|) \mu + \sum_{i=1}^n \tilde{y}_i z_i$$

$$\approx (1 - 2p)n \cdot \mu + \sum_{i=1}^n \tilde{y}_i z_i$$

Overfitting component helps interpolation, signal helps generalization:

<table>
<thead>
<tr>
<th>Training data</th>
<th>classify $(x, y)$ correctly</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle y_i x_i, \sum_{i=1}^n \tilde{y}_i z_i \rangle$</td>
<td>is large, positive,</td>
</tr>
<tr>
<td>$\langle y_i x_i, n \mu \rangle$</td>
<td>is small, noisy labels make ±,</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test data</th>
<th>classify $(x, \tilde{y})$ correctly</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle \tilde{y} x, \sum_{i=1}^n \tilde{y}_i z_i \rangle$</td>
<td>is small, random ±,</td>
</tr>
<tr>
<td>$\langle \tilde{y} x, n \mu \rangle$</td>
<td>is (optimally) large, positive.</td>
</tr>
</tbody>
</table>

• Signal and overfitting component balanced to allow both interpolation + generalization
Other approaches for benign overfitting in neural nets

- Analysis of implicit bias (KKT conditions, minimum norm interpolation, …)
  Frei-Vardi-Bartlett-Srebro'23; Kornowski-Yehudai-Shamir'23; Kou-Chen-Gu'23; …
  - Kornowski-Yehudai-Shamir’23 look at local and global minima of margin-maximization problems (rather than just KKT points)
  - Only applies to $\infty$-time limit of training

- "Trajectory analysis": directly track the weights of neural net trained by GD/GF from random initialization on noisy data, show that it achieves small train and test error
  Frei-Chatterji-Bartlett'22; Xu-Gu'23; Kou-Chen-Chen-Gu ICML'23; Xu-Wang-Frei-Vardi-Hu'23; Meng-Zou-Cao'23; …
  - Characterizes finite-time performance
  - More narrow, less clear "why" benign overfitting happens
Other approaches for benign overfitting in neural nets

- Analysis of implicit bias (KKT conditions, minimum norm interpolation, …)
  Frei-Vardi-Bartlett-Srebro’23; Kornowski-Yehudai-Shamir’23; Kou-Chen-Gu’23; …
  - Kornowski-Yehudai-Shamir’23 look at local and global minima of margin-maximization problems (rather than just KKT points)
  - Only applies to $\infty$-time limit of training

- “Trajectory analysis”: directly track the weights of neural net trained by GD/GF from random initialization on noisy data, show that it achieves small train and test error
  Frei-Chatterji-Bartlett’22; Xu-Gu’23; Kou-Chen-Chen-Gu ICML’23; Xu-Wang-Frei-Vardi-Hu’23; Meng-Zou-Cao’23; …
  - Characterizes finite time performance
  - More narrow, less clear “why” benign overfitting happens
Benign overfitting from trajectory analysis

- Directly examine inductive bias of training by GD/GF, e.g. in 2 layer nets

\[ f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)), \]

\[ \theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = -\frac{\alpha}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta^{(t)})) \cdot y_i \nabla f(x_i; \theta^{(t)}). \]
Benign overfitting from trajectory analysis

- Directly examine inductive bias of training by GD/GF, e.g. in 2 layer nets

\[
f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)),
\]

\[
\theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta^{(t)})) \cdot y_i \nabla f(x_i; \theta^{(t)}).
\]

- Tasks:
  - Analyze weights \( \theta^{(t)} \) and empirical risk \( \hat{L}(\theta^{(t)}) \) (training example margins \( y_i f(x_i; \theta^{(t)}) \))
Benign overfitting from trajectory analysis

• Directly examine inductive bias of training by GD/GF, e.g. in 2 layer nets

\[ f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle) , \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)) , \]

\[ \theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta^{(t)})) \cdot y_i \nabla f(x_i; \theta^{(t)}) . \]

• Tasks:
  • Analyze weights \( \theta^{(t)} \) and empirical risk \( \hat{L}(\theta^{(t)}) \) (training example margins \( y_i f(x_i; \theta^{(t)}) \))
  • Track test error \( \mathbb{P}(y \neq \text{sgn}(f(x; \theta^{(t)}))) \) (test example margin \( y f(x; \theta^{(t)}) \))
Benign overfitting from trajectory analysis

- Directly examine inductive bias of training by GD/GF, e.g. in 2 layer nets

\[ f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)), \]

\[ \theta(t+1) - \theta(t) = -\alpha \nabla \hat{L}(\theta(t)) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta(t))) \cdot y_i \nabla f(x_i; \theta(t)). \]

- Tasks:
  - Analyze weights \( \theta(t) \) and empirical risk \( \hat{L}(\theta(t)) \) (training example margins \( y_i f(x_i; \theta(t)) \))
  - Track test error \( \mathbb{P}(y \neq \text{sgn}(f(x; \theta(t)))) \) (test example margin \( y f(x; \theta(t)) \))
  - These two must be very different for benign overfitting to occur
Benign overfitting from trajectory analysis

\[ \tilde{y} \sim \text{Unif}\{\pm 1\}, \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \]

**Theorem**

Suppose labels flipped w.p. \( p = O(1) \), low SNR and \( d \gg n^2 \). Then when training a two-layer leaky ReLU network by gradient descent (w/ appropriate random init \( \theta^{(0)} \), learning rate), for all \( t \geq 1 \),

\[ \hat{L}(\theta^{(t)}) \leq O\left(\frac{1}{t}\right), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta^{(t)}))) \leq p + \exp\left(-\Omega(n\|\mu\|_4^4/d)\right). \]
Benign overfitting from trajectory analysis

\[ \tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \]

**Theorem**

Suppose labels flipped w.p. \( p = O(1) \), low SNR and \( d \gg n^2 \). Then when training a two-layer leaky ReLU network by gradient descent (w/ appropriate random init \( \theta(0) \), learning rate), for all \( t \geq 1 \),

\[
\hat{L}(\theta(t)) \leq O\left(\frac{1}{t}\right), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta(t)))) \leq p + \exp\left(-\Omega\left(n\|\mu\|^4/d\right)\right).
\]

- No dependence on number of neurons in network.
Benign overfitting from trajectory analysis

\[ \tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \]

**Theorem**

Suppose labels flipped w.p. \( p = O(1) \), low SNR and \( d \gg n^2 \). Then when training a two-layer leaky ReLU network by gradient descent (w/ appropriate random init \( \theta^{(0)} \), learning rate), for all \( t \geq 1 \),

\[
\hat{L}(\theta^{(t)}) \leq O\left(\frac{1}{t}\right), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta^{(t)}))) \leq p + \exp\left(-\Omega\left(n\|\mu\|^4/d\right)\right).
\]

- No dependence on number of neurons in network.
- **Benign overfitting** if \( t \) is large and \( n\|\mu\|^4 \gg d \).
Benign overfitting from trajectory analysis

\[ \tilde{y} \sim \text{Unif}(\{\pm 1\}), \quad x = \tilde{y}\mu + z, \quad z \sim \mathcal{N}(0, I_d), \quad y = -\tilde{y} \text{ w.p. } p. \]

**Theorem**

Suppose labels flipped w.p. \( p = O(1) \), low SNR and \( d \gg n^2 \). Then when training a two-layer leaky ReLU network by gradient descent (w/ appropriate random init \( \theta(0) \), learning rate), for all \( t \geq 1 \),

\[
\hat{L}(\theta(t)) \leq O\left(\frac{1}{t}\right), \quad \text{and} \quad p \leq \mathbb{P}(y \neq \text{sgn}(f(x; \theta(t)))) \leq p + \exp\left(-\Omega(n\|\mu\|^4/d)\right).
\]

- No dependence on number of neurons in network.
- **Benign overfitting** if \( t \) is large and \( n\|\mu\|^4 \gg d \).
- Same generalization bound as KKT analysis, but now holds throughout GD trajectory.
  - Only tolerates \( p = O(1) \), rather than \( p < \frac{1}{2} \) from KKT analysis.

Frei-Chatterji-Bartlett’22; Xu-Gu’23
Benign overfitting from trajectory analysis

\[ f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)), \]

\[ \theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \alpha \frac{1}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta^{(t)})) \cdot y_i \nabla f(x_i; \theta^{(t)}). \]
Benign overfitting from trajectory analysis

\[ f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)), \]

\[ \theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta^{(t)})) \cdot y_i \nabla f(x_i; \theta^{(t)}). \]

- Difficulty arises: “clean label” examples (in principle) are easier, larger margin \( y_i f(x_i; \theta^{(t)}) \), while “noisy label” examples harder, smaller margin
Benign overfitting from trajectory analysis

\[ f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)), \]

\[ \theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta^{(t)})) \cdot y_i \nabla f(x_i; \theta^{(t)}). \]

- Difficulty arises: “clean label” examples (in principle) are easier, larger margin \( y_i f(x_i; \theta^{(t)}) \), while “noisy label” examples harder, smaller margin
- Since \(-\ell'\) is decreasing, implies noisy labels could have outsized influence on training dynamics \(\rightarrow\) hard for overfitting to be ‘benign’
Benign overfitting from trajectory analysis

\[ f(x; \theta) = \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \quad \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)), \]

\[ \theta^{(t+1)} - \theta^{(t)} = -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta^{(t)})) \cdot y_i \nabla f(x_i; \theta^{(t)}). \]

- Difficulty arises: “clean label” examples (in principle) are easier, larger margin \( y_i f(x_i; \theta^{(t)}) \), while “noisy label” examples harder, smaller margin
- Since \(-\ell'\) is decreasing, implies noisy labels could have outsized influence on training dynamics \( \rightarrow \) hard for overfitting to be ‘benign’
- Key technical lemma shown in most trajectory analyses: [loss ratio bound],

\[ \sup_{t \geq 0} \max_{i,j} \frac{-\ell'(y_i f(x_i; \theta^{(t)}))}{-\ell'(y_j f(x_j; \theta^{(t)}))} = O(1). \]
Benign overfitting from trajectory analysis

\begin{align*}
  f(x; \theta) &= \sum_{j=1}^{m} a_j \phi(\langle \theta_j, x \rangle), \\
  \hat{L}(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \theta)), \\
  \theta^{(t+1)} - \theta^{(t)} &= -\alpha \nabla \hat{L}(\theta^{(t)}) = \frac{\alpha}{n} \sum_{i=1}^{n} -\ell'(y_i f(x_i; \theta^{(t)})) \cdot y_i \nabla f(x_i; \theta^{(t)}). \\
\end{align*}

- Difficulty arises: “clean label” examples (in principle) are easier, larger margin $y_i f(x_i; \theta^{(t)})$, while “noisy label” examples harder, smaller margin
- Since $-\ell'$ is decreasing, implies noisy labels could have outsized influence on training dynamics $\rightarrow$ hard for overfitting to be ‘benign’
- Key technical lemma shown in most trajectory analyses: loss ratio bound,

\[
\sup_{t \geq 0} \max_{i,j} \frac{-\ell'(y_i f(x_i; \theta^{(t)}))}{-\ell'(y_j f(x_j; \theta^{(t)}))} = O(1).
\]

- Known proofs all rely on nearly-orthogonal data $(d \gg n)$ to show this
“Blessing of Dimensionality”

- $d/n \to \infty$ necessary for benign overfitting in linear models, but unknown if necessary for neural networks.
“Blessing of Dimensionality”

- $d/n \to \infty$ necessary for benign overfitting in linear models, but unknown if necessary for neural networks.

- Consider again the Gaussian mixture model, with $p = 0.15$ labels flipped (train and test), $m = 512$ neurons, vary $d/n$. 
“Blessing of Dimensionality”

- \( \frac{d}{n} \to \infty \) necessary for benign overfitting in linear models, but unknown if necessary for neural networks.
- Consider again the Gaussian mixture model, with \( p = 0.15 \) labels flipped (train and test), \( m = 512 \) neurons, vary \( \frac{d}{n} \).
- Learning dynamics different in \( n > d \) setting; overfitting less ‘benign’
  \( \to “Blessing of dimensionality”? \) See also:

[Kornowski-Yehudai-Shamir'23]
Benign, tempered, and catastrophic overfitting

- There is a spectrum of generalization behavior when overfitting.

\[
\begin{align*}
\text{Regression BinaryClassification} \\
\text{Benign } & & \lim_{n \to \infty} R_n = R^* \\
\text{Tempered } & & \lim_{n \to \infty} R_n \in (R^*, \infty) \\
\text{Catastrophic } & & \lim_{n \to \infty} R_n = \infty
\end{align*}
\]
Benign, tempered, and catastrophic overfitting

- There is a spectrum of generalization behavior when overfitting.
- Let $R_n$ be test error for interpolator (train error = 0) using $n$ samples, $R^*$ best possible test error.

<table>
<thead>
<tr>
<th></th>
<th>Regression</th>
<th>Binary Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>$\lim_{n \to \infty} R_n = R^*$</td>
<td>$\lim_{n \to \infty} R_n = R^*$</td>
</tr>
<tr>
<td>Tempered</td>
<td>$\lim_{n \to \infty} R_n \in (R^*, \infty)$</td>
<td>$\lim_{n \to \infty} R_n \in (R^*, 1/2)$</td>
</tr>
<tr>
<td>Catastrophic</td>
<td>$\lim_{n \to \infty} R_n = \infty$</td>
<td>$\lim_{n \to \infty} R_n = 1/2$</td>
</tr>
</tbody>
</table>
Benign, tempered, and catastrophic overfitting

- There is a spectrum of generalization behavior when overfitting.
- Let $R_n$ be test error for interpolator (train error = 0) using $n$ samples, $R^*$ best possible test error.

<table>
<thead>
<tr>
<th></th>
<th>Regression</th>
<th>Binary Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>$\lim_{n \to \infty} R_n = R^*$</td>
<td>$\lim_{n \to \infty} R_n = R^*$</td>
</tr>
<tr>
<td>Tempered</td>
<td>$\lim_{n \to \infty} R_n \in (R^*, \infty)$</td>
<td>$\lim_{n \to \infty} R_n \in (R^*, 1/2)$</td>
</tr>
<tr>
<td>Catastrophic</td>
<td>$\lim_{n \to \infty} R_n = \infty$</td>
<td>$\lim_{n \to \infty} R_n = 1/2$</td>
</tr>
</tbody>
</table>

- Neural net trained on high-dimensional mixture model: (provably) benign; low-dimensional: tempered?
Open questions

• Is benign overfitting in neural nets possible in low dimensions \((n \gg d)\)?
  • Overparameterization through wider nets could help, but does it? When? Why?
Open questions

• Is benign overfitting in neural nets possible in low dimensions ($n \gg d$)?
  • Overparameterization through wider nets could help, but does it? When? Why?
• Which neural net interpolators do we care about in regression?
Open questions

- Is benign overfitting in neural nets possible in low dimensions \((n \gg d)\)?
  - Overparameterization through wider nets could help, but does it? When? Why?
- Which neural net interpolators do we care about in regression?
- Necessary and sufficient conditions for benign overfitting in linear classification?
  - Fairly complete picture of min-\(\ell^2\) linear regression, but mostly sufficiency guarantees in classification.
  - Dream: data-dependent, signal-dependent, tight guarantees.
Thanks!