Introduction: why study memorization?
Interpolation in modern machine learning

Classical statistical wisdom: bigger models tend to overfit, so we need to limit model capacity.

Modern empirical wisdom: overparameterized models can interpolate data.
Interpolation in modern machine learning

Classical statistical wisdom
Interpolation in modern machine learning

Classical statistical wisdom

- bigger models tend to overfit
Interpolation in modern machine learning

Classical statistical wisdom

- bigger models tend to overfit
- need to limit model capacity
Interpolation in modern machine learning

### Classical statistical wisdom

- **bigger models tend to overfit**
- **need to limit model capacity**

### Modern empirical wisdom
Interpolation in modern machine learning

**Classical statistical wisdom**

- bigger models tend to overfit
- need to limit model capacity

**Modern empirical wisdom**

- overparameterized model
Interpolation in modern machine learning

### Classical statistical wisdom

- bigger models tend to overfit
- need to limit model capacity

### Modern empirical wisdom

- overparameterized model
- interpolate data
Sufficiency of interpolation

When is it sufficient to overfit?

Benign overfitting

Surprises in high-dimensional ridgeless least squares interpolation

Hastie et al., 2018.

The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime

Montanari et al., 2019.

Just interpolate: Kernel "ridgeless" regression can generalize


Two models of double descent for weak features

Belkin et al., 2020.
Sufficiency of interpolation

When is it sufficient to overfit?

Benign overfitting

Algorithmic bias tends to generalize well

$\theta_0 \in \text{Parameter space } \Theta$

$\text{ERM}_0 = \{ \theta : \text{Training error under } \theta \text{ is zero} \}$
When is it sufficient to overfit?

Benign overfitting

Literatures

Surprises in high-dimensional ridgeless least squares interpolation
Hastie et al., 2018.

The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime
Montanari et al., 2019.

Just interpolate: Kernel "ridgeless" regression can generalize

Two models of double descent for weak features
Belkin et al., 2020.
Sufficiency of interpolation

When is it sufficient to overfit?

Benign overfitting

Algorithmic bias tends to generalize well

$\theta_0 \in \text{Parameter space } \Theta$

$\text{ERM}_0 = \{ \theta : \text{Training error under } \theta \text{ is zero}\}$

Literatures

- Surprises in high-dimensional ridgeless least squares interpolation
  *Hastie et al., 2018.*
Sufficiency of interpolation

When is it sufficient to overfit?

Benign overfitting

Literatures

- Surprises in high-dimensional ridgeless least squares interpolation
  *Hastie et al., 2018.*

- The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime
  *Montanari et al., 2019.*
Sufficiency of interpolation

When is it sufficient to overfit?

Benign overfitting

Literatures

- Surprises in high-dimensional ridgeless least squares interpolation
  Hastie et al., 2018.

- The generalization error of max-margin linear classifiers: High-dimensional
  asymptotics in the overparametrized regime
  Montanari et al., 2019.

- Just interpolate: Kernel “ridgeless” regression can generalize
Sufficiency of interpolation

When is it sufficient to overfit?

Benign overfitting

Literatures

- Surprises in high-dimensional ridgeless least squares interpolation
  Hastie et al., 2018.
- The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime
  Montanari et al., 2019.
- Just interpolate: Kernel “ridgeless” regression can generalize
- Two models of double descent for weak features
  Belkin et al., 2020.
Sufficiency of interpolation

When is it sufficient to overfit?

Benign overfitting

Algorithmic bias tends to generalize well

\[ \theta_0 \in \text{Parameter space } \Theta \]

\[ \text{ERM}_0 = \{ \theta : \text{Training error under } \theta \text{ is zero} \} \]

Literatures

- Surprises in high-dimensional ridgeless least squares interpolation
  
  Hastie et al., 2018.

- The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime
  
  Montanari et al., 2019.

- Just interpolate: Kernel “ridgeless” regression can generalize
  

- Two models of double descent for weak features
  
  Belkin et al., 2020.

- ...

Necessity of interpolation

When is it necessary to overfit?

Competing considerations

Privacy and security.

Figure from Carlini et al., 2021

Can we generalize well without memorization?

Inspiring line of works


Takeaways

Heavy-tailed distributions.

Need to memorize each class.

Combinatorial setup.
Necessity of interpolation

When is it necessary to overfit?

Competing considerations
Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

Figure from Carlini et al., 2021
Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

Can we generalize well without memorization?

Figure from Carlini et al., 2021
Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

Inspiring line of works

- Does Learning Require Memorization? A Short Tale about a Long Tail
  Feldman, 2019.

- When Is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?

Figure from Carlini et al., 2021

- Can we generalize well without memorization?
Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

Inspiring line of works

- Does Learning Require Memorization? A Short Tale about a Long Tail
  Feldman, 2019.

- When Is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?

Takeaways

- Can we generalize well without memorization?
Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

Inspiring line of works

- Does Learning Require Memorization? A Short Tale about a Long Tail
  Feldman, 2019.

- When Is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?

Takeaways

- Heavy-tailed distributions.

Can we generalize well without memorization?

Figure from Carlini et al., 2021
Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

Inspiring line of works

- Does Learning Require Memorization? A Short Tale about a Long Tail
  Feldman, 2019.

- When Is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?

Takeaways

- Heavy-tailed distributions.
- Need to memorize each class.

Can we generalize well without memorization?

Figure from Carlini et al., 2021
Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

Inspiring line of works

- Does Learning Require Memorization? A Short Tale about a Long Tail
  Feldman, 2019.

- When Is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?

Takeaways

- Heavy-tailed distributions.

- Need to memorize each class.

- Combinatorial setup.

Can we generalize well without memorization?

Figure from Carlini et al., 2021
Necessity of interpolation

Real life data have heavy-tailed distributions
Necessity of interpolation

Real life data have heavy-tailed distributions

(a) The number of examples by object class in SUN dataset

(b) Distributions of the visibility patterns for bus and person
Necessity of interpolation

Real life data have heavy-tailed distributions

(a) The number of examples by object class in SUN dataset

(b) Distributions of the visibility patterns for bus and person
Necessity of interpolation

Real life data have heavy-tailed distributions

(a) The number of examples by object class in SUN dataset

(b) Distributions of the visibility patterns for bus and person

Have to memorize for each class
Necessity of interpolation

Carefully constructed combinatorial settings
Necessity of interpolation

Carefully constructed combinatorial settings

Figure from Brown et al., 2021
Necessity of interpolation

Carefully constructed combinatorial settings

Instance (distribution on labeled examples) $P \sim q$

$n$ labeled examples $X \sim p^{\otimes n}$

$d$ bits per example

Knows $q$ but not $P$

Learning Algorithm $\Rightarrow$ Model $M$ $\Rightarrow$ Predict

Test point $Z$

True label $Y$

Correct/Incorrect

Predicted label

Figure from Brown et al., 2021

Can we do it for something simpler? I don’t know, like $y = \theta + w$?

Average John
Necessity of interpolation

A general formulation

Data pairs $(x_i, y_i)$ from $y_i = f(x_i; \theta, w_i)$ and a hypothesis class $H$ containing $\theta$.

The cost of not-fitting

$$\min_{\theta \in H} \text{Pred}$$

subject to

$$\text{Train} \ b_\theta \geq \epsilon$$

A simpler model

Linear model for $X \in \mathbb{R}^{n \times d}$

$y = X\theta + w$

d $\geq n$ so we can interpolate

“memorization”: if we have to fit substantially below the inherent noise floor
Necessity of interpolation

A general formulation

- Data pairs \((x_i, y_i)\) from

\[ y_i = f(x_i; \theta, w_i) \]

and a hypothesis class \(\mathcal{H}\) containing \(\theta\).
Necessity of interpolation

A general formulation

- Data pairs \((x_i, y_i)\) from

\[ y_i = f(x_i; \theta, w_i) \]

and a hypothesis class \(\mathcal{H}\) containing \(\theta\).

- The cost of not-fitting

\[
\begin{align*}
\text{minimize} & \quad \text{Pred} \left( \hat{\theta} \right) \\
\text{subject to} & \quad \text{Train} \left( \hat{\theta} \right) \geq \epsilon^2
\end{align*}
\]
Necessity of interpolation

A general formulation

- Data pairs \((x_i, y_i)\) from

\[ y_i = f(x_i; \theta, w_i) \]

and a hypothesis class \(\mathcal{H}\) containing \(\theta\).

- The cost of not-fitting

\[
\begin{align*}
\text{minimize} & \quad \text{Pred}\left(\hat{\theta}\right) \\
\text{subject to} & \quad \text{Train}\left(\hat{\theta}\right) \geq \epsilon^2 \\
\end{align*}
\]

A simpler model

- Linear model for \(X \in \mathbb{R}^{n \times d}\)

\[ y = X\theta + w \]

so we can interpolate

"memorization": if we have to fit substantially below the inherent noise floor
Necessity of interpolation

A general formulation

- Data pairs \((x_i, y_i)\) from

\[ y_i = f(x_i; \theta, w_i) \]

and a hypothesis class \(\mathcal{H}\) containing \(\theta\).

- The cost of not-fitting

\[
\begin{align*}
\text{minimize} & \quad \text{Pred}(\hat{\theta}) \\
\text{subject to} & \quad \text{Train}(\hat{\theta}) \geq \epsilon^2
\end{align*}
\]

A simpler model

- Linear model for \(X \in \mathbb{R}^{n \times d}\)

\[ y = X\theta + w \]
Necessity of interpolation

A general formulation

- Data pairs \((x_i, y_i)\) from
  \[ y_i = f(x_i; \theta, w_i) \]
  and a hypothesis class \(H\) containing \(\theta\).

- The cost of not-fitting
  \[
  \begin{align*}
  &\text{minimize} \\
  &\hat{\theta} \in \mathcal{H} \\
  \text{subject to} \\
  &\text{Train}(\hat{\theta}) \geq \epsilon^2
  \end{align*}
  \]

A simpler model

- Linear model for \(X \in \mathbb{R}^{n \times d}\)
  \[ y = X\theta + w \]

- \(d \geq n\) so we can interpolate
Necessity of interpolation

A general formulation

- Data pairs \((x_i, y_i)\) from
  \[ y_i = f(x_i; \theta, w_i) \]

  and a hypothesis class \(\mathcal{H}\) containing \(\theta\).

- The cost of not-fitting

  \[
  \begin{align*}
  \text{minimize} & \quad \operatorname{Pred}(\hat{\theta}) \\
  \text{subject to} & \quad \operatorname{Train}(\hat{\theta}) \geq \epsilon^2
  \end{align*}
  \]

A simpler model

- Linear model for \(X \in \mathbb{R}^{n \times d}\)
  \[ y = X\theta + w \]

- \(d \geq n\) so we can interpolate

- “memorization”: if we have to fit substantially below the inherent noise floor
Main results: necessity of memorization in linear regression
Let’s start from the isotropic Gaussian case
Let’s start from the isotropic Gaussian case

**Problem setup**
Consider the standard overparameterized linear model $y = X\theta + w$, with
Let’s start from the isotropic Gaussian case

**Problem setup**
Consider the standard overparameterized linear model $y = X\theta + w$, with
- random isotropic i.i.d. design matrix $X = \mathbb{R}^{n \times d}$ ($d \geq n$)
Let’s start from the isotropic Gaussian case

**Problem setup**

Consider the standard overparameterized linear model $y = X\theta + w$, with

- random isotropic i.i.d. design matrix $X = \mathbb{R}^{n \times d}$ ($d \geq n$)
- Bayesian setup, with unknown signal $\theta \sim \mathcal{N}(0, I_{d/d})$
Let’s start from the isotropic Gaussian case

**Problem setup**

Consider the standard overparameterized linear model $y = X\theta + w$, with

- random isotropic i.i.d. design matrix $X = \mathbb{R}^{n \times d}$ ($d \geq n$)
- Bayesian setup, with unknown signal $\theta \sim \mathcal{N}(0, I_d/d)$
- noise $w \sim \mathcal{N}(0, \sigma^2 I_n)$
Linear regression

Let’s start from the isotropic Gaussian case

**Problem setup**

Consider the standard overparameterized linear model $y = X\theta + w$, with

- random isotropic i.i.d. design matrix $X = \mathbb{R}^{n \times d}$ ($d \geq n$)
- Bayesian setup, with unknown signal $\theta \sim \mathcal{N}(0, I_d/d)$
- noise $w \sim \mathcal{N}(0, \sigma^2 I_n)$

$\ell_2$ error

$$\text{Train}_X \left( \hat{\theta} \right) = \frac{1}{n} \mathbb{E}_{w, \theta} \left[ \left\| X\hat{\theta} - y \right\|_2^2 \right]$$

$$\text{Pred}_X \left( \hat{\theta} \right) = \mathbb{E}_{x, w, \theta} \left[ \left\| x^\top \theta - x^\top \hat{\theta} \right\|_2^2 \right]$$
Cost of not-fitting for linear regression
Cost of not-fitting for linear regression

We want to solve

$$\min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X \left( \hat{\theta} \right) \quad \text{s.t.} \quad \text{Train}_X \left( \hat{\theta} \right) \geq \epsilon^2,$$

where $\mathcal{H} = \left\{ \hat{\theta}(X, y) \text{ square integrable} \right\}$. 
We want to solve
\[
\min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}) \quad \text{s.t.} \quad \text{Train}_X (\hat{\theta}) \geq \epsilon^2,
\]

where \( \mathcal{H} = \{ \hat{\theta}(X, y) \text{ square integrable} \} \). Let \( \mathcal{H}(\epsilon) = \{ \text{Train}_X (\hat{\theta}) \geq \epsilon^2 \} \) and

\[
\text{Cost}_X (\epsilon) := \min_{\hat{\theta} \in \mathcal{H}(\epsilon)} \text{Pred}_X (\hat{\theta}) - \min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}).
\]
Cost of not-fitting for linear regression

We want to solve

\[
\min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}) \quad \text{s.t.} \quad \text{Train}_X (\hat{\theta}) \geq \epsilon^2 ,
\]

where \( \mathcal{H} = \{ \hat{\theta}(X, y) \text{ square integrable} \} \). Let \( \mathcal{H}(\epsilon) = \{ \text{Train}_X (\hat{\theta}) \geq \epsilon^2 \} \) and

\[
\text{Cost}_X (\epsilon) := \min_{\hat{\theta} \in \mathcal{H}(\epsilon)} \text{Pred}_X (\hat{\theta}) - \min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}).
\]

To generalize well I need to fit to the noise level? Unavoidable?
Cost of not-fitting for linear regression

We want to solve

$$\min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}) \quad \text{s.t.} \quad \text{Train}_X (\hat{\theta}) \geq \epsilon^2,$$

where $\mathcal{H} = \{ \hat{\theta}(X, y) \text{ square integrable} \}$. Let $\mathcal{H}(\epsilon) = \{ \text{Train}_X (\hat{\theta}) \geq \epsilon^2 \}$ and

$$\text{Cost}_X (\epsilon) := \min_{\hat{\theta} \in \mathcal{H}(\epsilon)} \text{Pred}_X (\hat{\theta}) - \min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}).$$
Cost of not-fitting for linear regression

Theorem 1 (Cheng, Duchi, Kuditipudi ’22)

Under proportional asymptotics, namely $d/n \to \gamma$ as $n \to \infty$ for some $\gamma > 1$,

- (no-cost phase) $\lim_{n \to \infty} \text{Cost}_X(\epsilon) > 0$ iff $\epsilon^2 > \epsilon_\sigma^2 := \frac{\sigma^4}{\frac{\sigma^2}{\epsilon^2} + 1 - 1/\gamma} + o(\sigma^4)$

- (linear-growth phase) $\lim_{n \to \infty} \text{Cost}_X(\epsilon) \geq C\gamma \epsilon^2$ for $\epsilon^2 \geq c_\gamma \sigma^4$. 

I must fit with accuracy $>>$ inherent noise level.

John Duchi on an average day
Cost of not-fitting for linear regression

**Theorem 1 (Cheng, Duchi, Kuditipudi ’22)**

*Under proportional asymptotics* $d/n \to \gamma > 1$,

- **(no-cost phase)** $\lim_{n \to \infty} \text{Cost}_X(\epsilon) > 0$ *iff* $\epsilon^2 > \epsilon^2_\sigma := \frac{\sigma^4}{\sigma^4 + 1 - 1/\gamma} + o(\sigma^4)$
- **(linear-growth phase)** $\lim_{n \to \infty} \text{Cost}_X(\epsilon) \geq C_\gamma \epsilon^2$ *for* $\epsilon^2 \geq c_\gamma \sigma^4$.

---

**Diagram:**

- **No-cost phase**
- **Linear-growth phase**

---

*I must fit with accuracy >> inherent noise level.*

John Duchi

on an average day
Proof sketch: strong duality and random matrix theory
The proof consists of three parts.
Proof outline

The proof consists of three parts.

- **Strong duality for linear estimators.** Starting from linear hypothesis class $\hat{\theta} = Ay$, we solve the nonconvex minimization problem

  $$\min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}) \quad \text{s.t.} \quad \text{Train}_X (\hat{\theta}) \geq \epsilon^2,$$

  using strong duality.
The proof consists of three parts.

- **Strong duality for linear estimators.** Starting from linear hypothesis class \( \hat{\theta} = Ay \), we solve the nonconvex minimization problem

\[
\min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}) \quad \text{s.t.} \quad \text{Train}_X (\hat{\theta}) \geq \epsilon^2,
\]

using strong duality. Bayes optimal estimator is linear without constraint—with constraint? Still linear!
Proof outline

The proof consists of three parts.

- **Strong duality for linear estimators.** Starting from linear hypothesis class \( \hat{\theta} = Ay \), we solve the nonconvex minimization problem

  \[
  \min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X (\hat{\theta}) \quad \text{s.t.} \quad \text{Train}_X (\hat{\theta}) \geq \epsilon^2 ,
  \]

  using strong duality. Bayes optimal estimator is linear without constraint— with constraint? Still linear!

- **Derive asymptotics using RMT.** With the exact form of the (approximate) minimizer, we derive asymptotic limits of threshold value \( \epsilon_{\sigma} \), cost of not-fitting \( \text{Cost}_X (\epsilon) \) by random matrix theory.
Proof outline

The proof consists of three parts.

- **Strong duality for linear estimators.** Starting from linear hypothesis class \( \hat{\theta} = Ay \), we solve the nonconvex minimization problem

  \[
  \min_{\hat{\theta} \in H} \text{Pred}_X \left( \hat{\theta} \right) \quad \text{s.t.} \quad \text{Train}_X \left( \hat{\theta} \right) \geq \epsilon^2 ,
  \]

  using strong duality. Bayes optimal estimator is linear without constraint—with constraint? Still linear!

- **Derive asymptotics using RMT.** With the exact form of the (approximate) minimizer, we derive asymptotic limits of threshold value \( \epsilon_\sigma \), cost of not-fitting \( \text{Cost}_X (\epsilon) \) by random matrix theory.

- **Upgrade by functional strong duality.** Finally, we upgrade to any square integrable estimator \( \hat{\theta}(X, y) \) by showing a functional strong duality result.
Strong duality for linear hypothesis class

For linear estimator $\hat{\theta} = Ay$, let $\mathcal{P}(A) := \text{Pred}_X \left(\hat{\theta}\right)$ and $\mathcal{T}(A) := \text{Train}_X \left(\hat{\theta}\right)$. 

Reduction to QCQP 

$$\minimize_{A \in \mathbb{R}^{d \times n}} \mathcal{P}(A) = 1$$

$$d \|AX - I\|_2^2 + \sigma^2 \|A\|_F^2$$

subject to

$$\mathcal{T}(A) = 1$$

$$\|XAX - X\|_2^2 + \sigma^2 n \|XA - I\|_2^2 \geq \epsilon^2.$$ 

Strong duality

The problem—while nonconvex—has quadratic objective and a single quadratic constraint. Strong duality holds!

Optimality condition with $\rho_n := \rho_n(\epsilon)$

$$A(\rho_n) = I - \rho_n \sigma^2 I - \rho_n d X^\top X - 1 (X^\top X + d \sigma^2 I - 1) X^\top.$$ 

Ridge estimator when $\rho = 0$, optimal with $\epsilon^2 \sigma$ training error.
Strong duality for linear hypothesis class

For linear estimator $\hat{\theta} = Ay$, let $\mathcal{P}(A) := \text{Pred}_X \left( \hat{\theta} \right)$ and $\mathcal{T}(A) := \text{Train}_X \left( \hat{\theta} \right)$.

Reduction to QCQP

$$\text{minimize} \quad \mathcal{P}(A) = \frac{1}{d} \| AX - I \|_F^2 + \sigma^2 \| A \|_F^2$$

subject to

$$\mathcal{T}(A) = \frac{1}{nd} \| XAX - X \|_F^2 + \frac{\sigma^2}{n} \| XA - I \|_F^2 \geq \epsilon^2.$$
Strong duality for linear hypothesis class

For linear estimator \( \hat{\theta} = Ay \), let \( \mathcal{P}(A) := \text{Pred}_X (\hat{\theta}) \) and \( \mathcal{T}(A) := \text{Train}_X (\hat{\theta}) \).

Reduction to QCQP

\[
\begin{align*}
\text{minimize}_{A \in \mathbb{R}^{d \times n}} & \quad \mathcal{P}(A) = \frac{1}{d} \| AX - I \|_F^2 + \sigma^2 \| A \|_F^2 \\
\text{subject to} & \quad \mathcal{T}(A) = \frac{1}{nd} \| XAX - X \|_F^2 + \frac{\sigma^2}{n} \| XA - I \|_F^2 \geq \epsilon^2.
\end{align*}
\]

Strong duality

- The problem—while nonconvex—has \textbf{quadratic objective and a single quadratic constraint}. Strong duality holds!
For linear estimator $\hat{\theta} = Ay$, let $\mathcal{P}(A) := \text{Pred}_X (\hat{\theta})$ and $\mathcal{T}(A) := \text{Train}_X (\hat{\theta})$.

**Reduction to QCQP**

$$\minimize_{A \in \mathbb{R}^{d \times n}} \quad \mathcal{P}(A) = \frac{1}{d} \| AX - I \|_F^2 + \sigma^2 \| A \|_F^2$$

subject to

$$\mathcal{T}(A) = \frac{1}{nd} \| XAX - X \|_F^2 + \frac{\sigma^2}{n} \|XA - I\|_F^2 \geq \epsilon^2.$$

**Strong duality**

- The problem—while nonconvex—has *quadratic objective and a single quadratic constraint*. Strong duality holds!

- Optimality condition with $\rho_n := \rho_n(\epsilon)$

$$A(\rho_n) = \left( I - \rho_n \sigma^2 \left( I - \frac{\rho_n}{d} X^\top X \right)^{-1} \right) (X^\top X + d\sigma^2 I)^{-1} X^\top$$
For linear estimator $\hat{\theta} = Ay$, let $\mathcal{P}(A) := \text{Pred}_X \left( \hat{\theta} \right)$ and $\mathcal{T}(A) := \text{Train}_X \left( \hat{\theta} \right)$.

**Reduction to QCQP**

$$
\begin{align*}
\min_{A \in \mathbb{R}^{d \times n}} & \quad \mathcal{P}(A) = \frac{1}{d} \|AX - I\|_F^2 + \sigma^2 \|A\|_F^2 \\
\text{subject to} & \quad \mathcal{T}(A) = \frac{1}{nd} \|XAX - X\|_F^2 + \frac{\sigma^2}{n} \|XA - I\|_F^2 \geq \epsilon^2.
\end{align*}
$$

**Strong duality**

- The problem—while nonconvex—has **quadratic objective and a single quadratic constraint**. Strong duality holds!
- Optimality condition with $\rho_n := \rho_n(\epsilon)$

$$
A(\rho_n) = \left( I - \rho_n \sigma^2 \left( I - \frac{\rho_n}{d} X^\top X \right)^{-1} \right) (X^\top X + d\sigma^2 I)^{-1} X^\top
$$

Ridge estimator when $\rho = 0$, optimal with $\epsilon^2_\sigma$ training error.
Let $X$ have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The empirical spectral distribution of $X X^\top$ is $\mu_n$ with its c.d.f. $H_n(s) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i^2} \leq s$.

Marchenko-Pastur law $\mu_n \Rightarrow \mu$, $H_n(s) \rightarrow H(s)$.

d$H(s) = \gamma^2 \pi \rho(\lambda_+ - s)(s - \lambda_-)$, $s \in [\lambda_-, \lambda_+]$. 

$\lambda_{\pm} = \frac{1}{\gamma^2} \pm \frac{1}{\sqrt{\gamma^2}}$. 

Let $X$ have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The empirical spectral distribution of $\frac{1}{d}XX^\top$ is $\mu_n$ with its c.d.f. $H_n(s) := \frac{1}{n} \sum_{i=1}^{n} 1_{\lambda_i^2/d \leq s}$. 

Marchenko-Pastur law $\mu_n \Rightarrow \mu$, $H_n(s) \to H(s)$.

$$dH(s) = \frac{\gamma_2}{\pi} p(\lambda_+ - s)(s - \lambda_-) ds,$$

with $\lambda_\pm := \frac{1}{\gamma_2^2}$. 

Asymptotics by RMT
Asymptotics by RMT

Let $X$ have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The empirical spectral distribution of $\frac{1}{d}XX^\top$ is $\mu_n$ with its c.d.f. $H_n(s) := \frac{1}{n} \sum_{i=1}^{n} 1_{\lambda_i^2/d \leq s}$.

Marchenko-Pastur law

$\mu_n \Rightarrow \mu$, $H_n(s) \to H(s)$. 
Let $X$ have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The empirical spectral distribution of $\frac{1}{d} XX^\top$ is $\mu_n$ with its c.d.f. $H_n(s) := \frac{1}{n} \sum_{i=1}^{n} 1_{\lambda_i^2/d \leq s}$.

**Marchenko-Pastur law**

$\mu_n \Rightarrow \mu$, $H_n(s) \to H(s)$.

$$dH(s) = \frac{\gamma}{2\pi} \frac{\sqrt{(\lambda_+ - s)(s - \lambda_-)}}{s} 1_{s \in [\lambda_-, \lambda_+]} ds,$$

with $\lambda_\pm := (1 \pm 1/\sqrt{\gamma})^2$. 
Let $X$ have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The empirical spectral distribution of $\frac{1}{d}XX^\top$ is $\mu_n$ with its c.d.f. $H_n(s) := \frac{1}{n} \sum_{i=1}^{n} 1_{\lambda_i^2/d \leq s}$.

**Marchenko-Pastur law**

$\mu_n \Rightarrow \mu, H_n(s) \to H(s)$.

$$dH(s) = \frac{\gamma}{2\pi} \frac{\sqrt{(\lambda_+ - s)(s - \lambda_-)}}{s} 1_{s \in [\lambda_-, \lambda_+]} ds,$$

with $\lambda_{\pm} := \left(1 \pm \frac{1}{\sqrt{\gamma}}\right)^2$. 
Prediction and training errors

\[ P(A(\rho)) - P(A(0)) = \rho^2 \sigma^4 d \text{Tr} I - \rho dX^\top X - 2X^\top X dX + \sigma^2 I - \frac{1}{n} \int T(A(\rho)) \]

Prediction and training errors in ESD

\[ P(A(\rho)) - P(A(0)) = \rho^2 n d Z \sigma^4 s (1 - \rho s) \frac{2}{(s + \sigma^2)} dH n(s) \]

Limit of Lagrange multiplier

Since \[ T(A(\rho n)) = \epsilon^2 \], would expect \[ \rho n \rightarrow \rho \epsilon \]

\[ Z \sigma^4 (1 - \rho \epsilon s)^2 (s + \sigma^2) dH(s) = \epsilon^2 \]
Prediction and training errors

\[ \mathcal{P}(A(\rho)) - \mathcal{P}(A(0)) = \frac{\rho^2 \sigma^4}{d} \text{Tr} \left( \left( I - \frac{\rho}{d} X^\top X \right)^{-2} \frac{X^\top X}{d} \left( \frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right) \]

\[ \mathcal{T}(A(\rho)) = \frac{\sigma^4}{n} \text{Tr} \left( \left( I - \frac{\rho}{d} X^\top X \right)^{-2} \left( \frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right) \]
Prediction and training errors

\[ P(A(\rho)) - P(A(0)) = \frac{\rho^2 \sigma^4}{d} \text{Tr} \left( \left( I - \frac{\rho}{d} X^\top X \right)^{-2} \frac{X^\top X}{d} \left( \frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right) \]

\[ T(A(\rho)) = \frac{\sigma^4}{n} \text{Tr} \left( \left( I - \frac{\rho}{d} X^\top X \right)^{-2} \left( \frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right) \]

Prediction and training errors in ESD

\[ P(A(\rho)) - P(A(0)) = \frac{\rho^2 n}{d} \int \frac{\sigma^4 s}{(1 - \rho s)^2 (s + \sigma^2)} dH_n(s) \]

\[ T(A(\rho)) = \int \frac{\sigma^4}{(1 - \rho s)^2 (s + \sigma^2)} dH_n(s) \]
Asymptotics by RMT

**Prediction and training errors**

\[
P(A(\rho)) - P(A(0)) = \frac{\rho^2 \sigma^4}{d} \text{Tr} \left( \left( I - \frac{\rho}{d} X^\top X \right)^{-2} \frac{X^\top X}{d} \left( \frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right)
\]

\[
\mathcal{T}(A(\rho)) = \frac{\sigma^4}{n} \text{Tr} \left( \left( I - \frac{\rho}{d} X^\top X \right)^{-2} \left( \frac{X^\top X}{d} + \sigma^2 I \right)^{-1} \right)
\]

**Prediction and training errors in ESD**

\[
P(A(\rho)) - P(A(0)) = \frac{\rho^2 n}{d} \int \frac{\sigma^4 s}{(1 - \rho s)^2 (s + \sigma^2)} dH_n(s)
\]

\[
\mathcal{T}(A(\rho)) = \int \frac{\sigma^4}{(1 - \rho s)^2 (s + \sigma^2)} dH_n(s)
\]

**Limit of Lagrange multiplier**

Since \( \mathcal{T}(A(\rho_n)) = \epsilon^2 \), would expect \( \rho_n \to \rho_\epsilon \)

\[
\int \frac{\sigma^4}{(1 - \rho_\epsilon s)^2 (s + \sigma^2)} dH(s) = \epsilon^2
\]
Taking $\rho \epsilon = 0$ gives $\epsilon^2 \sigma = T(A(0)) \rightarrow Z \sigma^4 s + \sigma^2 dH(s) = \sigma^4 \sigma^2 + 1 - 1/\gamma + o(\sigma^4)$.

Limit of cost of not-fitting

$$\text{Cost } X(\epsilon) = P(A(\rho n)) - P(A(0)) = P(A(\rho \epsilon)) - P(A(0)) + o(n)$$

$$\rightarrow \rho^2 \epsilon \gamma Z \sigma^4 s (1 - \rho \epsilon s) \frac{1}{s + \sigma^2} dH(s)$$

Theorem 1 (Cheng, Duchi, Kuditipudi '22)

Under proportional asymptotics $\frac{d}{n} \rightarrow \gamma > 1$,

$$(\text{no-cost phase}) \lim_{n \to \infty} \text{Cost } X(\epsilon) > 0 \iff \epsilon^2 > \epsilon^2 \sigma := \sigma^4 \sigma^2 + 1 - 1/\gamma + o(\sigma^4).$$

$$(\text{linear-growth phase}) \lim_{n \to \infty} \text{Cost } X(\epsilon) \geq C \gamma \epsilon^2$$ for $\epsilon^2 \geq c \gamma \sigma^4$. 
Limit of threshold

Taking $\rho_\epsilon = 0$ gives

$$\epsilon^2 = T(A(0)) \rightarrow \int \frac{\sigma^4}{s + \sigma^2} dH(s) = \frac{\sigma^4}{\sigma^2 + 1 - 1/\gamma} + o(\sigma^4)$$
Limit of threshold

Taking $\rho_{\epsilon} = 0$ gives

$$
\epsilon^2_{\sigma} = \mathcal{T}(A(0)) \to \int \frac{\sigma^4}{s + \sigma^2} dH(s) = \frac{\sigma^4}{\sigma^2 + 1 - 1/\gamma} + o(\sigma^4)
$$

Limit of cost of not-fitting

$$
\text{Cost}_X(\epsilon) = \mathcal{P}(A(\rho_n)) - \mathcal{P}(A(0)) = \mathcal{P}(A(\rho_{\epsilon})) - \mathcal{P}(A(0)) + o_n(1)
$$

$$
\to \frac{\rho_{\epsilon}^2}{\gamma} \int \frac{\sigma^4 s}{(1 - \rho_{\epsilon} s)^2 (s + \sigma^2)} dH(s)
$$
Asymptotics by RMT

Limit of threshold

Taking $\rho_\epsilon = 0$ gives

$$\epsilon_\sigma^2 = T(A(0)) \rightarrow \int \frac{\sigma^4}{s + \sigma^2} dH(s) = \frac{\sigma^4}{\sigma^2 + 1 - 1/\gamma} + o(\sigma^4)$$

Limit of cost of not-fitting

$$\text{Cost}_X(\epsilon) = \mathcal{P}(A(\rho_n)) - \mathcal{P}(A(0)) = \mathcal{P}(A(\rho_\epsilon)) - \mathcal{P}(A(0)) + o_n(1)$$

$$\rightarrow \frac{\rho_\epsilon^2}{\gamma} \int \frac{\sigma^4 s}{(1 - \rho_\epsilon s)^2(s + \sigma^2)} dH(s)$$

Theorem 1 (Cheng, Duchi, Kuditipudi ’22)

Under proportional asymptotics $d/n \to \gamma > 1$,

- (no-cost phase) $\lim_{n \to \infty} \text{Cost}_X(\epsilon) > 0$ iff $\epsilon^2 > \epsilon_\sigma^2 := \frac{\sigma^4}{\sigma^2 + 1 - 1/\gamma} + o(\sigma^4)$
- (linear-growth phase) $\lim_{n \to \infty} \text{Cost}_X(\epsilon) \geq C_\gamma \epsilon^2$ for $\epsilon^2 \geq c_\gamma \sigma^4$. 
It only remains to show the same conclusion holds for $\hat{\theta}(X, y)$ square integrable given Gaussianity.
Upgrade to general hypothesis class

It only remains to show the same conclusion holds for $\hat{\theta}(X, y)$ square integrable given Gaussianity.

$$\minimize_{\hat{\theta}(X, y) \in L^2} \int \left\| \hat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 d\mu$$

subject to

$$\int \left\| X\hat{\theta} - y \right\|_2^2 d\mu \geq \epsilon^2$$

where $\mu \overset{d}{=} N(0, \frac{1}{d}XX^\top + \sigma^2 I)$. 
It only remains to show the same conclusion holds for \( \hat{\theta}(X, y) \) square integrable given Gaussianity.

\[
\text{minimize}_{\hat{\theta}(X, y) \in L^2} \int \left\| \hat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|^2 d\mu
\]

subject to \( \int \left\| X\hat{\theta} - y \right\|^2 d\mu \geq \epsilon^2 \)

where \( \mu = N \left( 0, \frac{1}{d}XX^\top + \sigma^2 I \right) \). Strong duality in Hilbert space?
It only remains to show the same conclusion holds for \( \hat{\theta}(X, y) \) square integrable given Gaussianity.

\[
\begin{align*}
\text{minimize} & \quad \int \left\| \hat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|^2_2 d\mu \\
\text{subject to} & \quad \int \left\| X\hat{\theta} - y \right\|^2_2 d\mu \geq \epsilon^2
\end{align*}
\]

where \( \mu \overset{d}{=} N(0, \frac{1}{d} XX^\top + \sigma^2 I) \). Strong duality in Hilbert space?

\[
\hat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y - \rho X^\top (X\hat{\theta} - y)/d = 0
\]
It only remains to show the same conclusion holds for $\hat{\theta}(X, y)$ square integrable given Gaussianity.

\[
\begin{align*}
\text{minimize} & \quad \int \left\| \hat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 \, d\mu \\
\text{subject to} & \quad \int \left\| X\hat{\theta} - y \right\|_2^2 \, d\mu \geq \epsilon^2
\end{align*}
\]

where $\mu \overset{d}{=} N(0, \frac{1}{d} XX^\top + \sigma^2 I)$. Strong duality in Hilbert space?

\[
\hat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y - \rho X^\top (X\hat{\theta} - y)/d = 0
\]

We exactly have

\[
\hat{\theta} = A(X)y
\]
Upgrade to general hypothesis class

Functional strong duality
Upgrade to general hypothesis class

**Functional strong duality**

\[
\begin{align*}
\text{minimize} & \quad \int \left\| \hat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 \, d\mu_m \\
\text{subject to} & \quad \int \left\| X\hat{\theta} - y \right\|_2^2 \, d\mu_m \geq \epsilon^2
\end{align*}
\]

where \( \mu_m \) are empirical distributions for i.i.d. samples of \( y \mid X \).
Functional strong duality

\[
\begin{align*}
\text{minimize} & \quad \int \left\| \hat{\theta} - \left( X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 d\mu_m \\
\text{subject to} & \quad \int \left\| X\hat{\theta} - y \right\|_2^2 d\mu_m \geq \epsilon^2
\end{align*}
\]

where \( \mu_m \) are empirical distributions for i.i.d. samples of \( y \mid X \). Strong duality applies to finite dimensional problems! Take \( m \to \infty \) and conclude by SLLN.
Cost of not-interpolating

The optimal interpolant is the OLS estimator $b_{\theta_{ols}} = X^\top (XX^\top)^{-1}y$.

**Theorem 2 (Cheng, Duchi, Kuditipudi '22)**

Under proportional asymptotics $d/n \to \gamma > 1$, (no-cost phase) $\lim_{n \to \infty} \text{Cost}_X(\epsilon) > 0$ iff $\epsilon_{2}^2 > \epsilon_{2}^2 \sigma_{ols}$.

(linear-growth phase) $\lim_{n \to \infty} \text{Cost}_X(\epsilon) \geq C_{\gamma} \epsilon_{2}^2$ for $\epsilon_{2}^2 \geq c_{\gamma} \sigma_{4}^4$.

(threshold value) $\epsilon_{\sigma} < \epsilon_{\sigma},\text{ols} \leq \kappa_{\gamma} \epsilon_{\sigma}$.
Cost of not-interpolating

Cost of not-fitting

\[ \text{Cost}_X (\epsilon) := \min_{\hat{\theta} \in \mathcal{H}(\epsilon)} \text{Pred}_X \left( \hat{\theta} \right) - \min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X \left( \hat{\theta} \right). \]
Cost of not-interpolating

Cost of not-fitting

\[ \text{Cost}_X(\epsilon) := \min_{\hat{\theta} \in \mathcal{H}(\epsilon)} \text{Pred}_X(\hat{\theta}) - \min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X(\hat{\theta}). \]

Cost of not-interpolating

\[ \overline{\text{Cost}}_X(\epsilon) := \min_{\hat{\theta} \in \mathcal{H}(\epsilon)} \text{Pred}_X(\hat{\theta}) - \min_{\hat{\theta} \in \mathcal{H}_0} \text{Pred}_X(\hat{\theta}). \]

The optimal interpolant is the OLS estimator \( \hat{\theta}_{\text{ols}} = X^\top (XX^\top)^{-1} y. \)
Cost of not-interpolating

Cost of not-fitting

\[
\text{Cost}_X(\epsilon) := \min_{\hat{\theta} \in \mathcal{H}(\epsilon)} \text{Pred}_X(\hat{\theta}) - \min_{\hat{\theta} \in \mathcal{H}} \text{Pred}_X(\hat{\theta}).
\]

Cost of not-interpolating

\[
\overline{\text{Cost}}_X(\epsilon) := \min_{\hat{\theta} \in \mathcal{H}(\epsilon)} \text{Pred}_X(\hat{\theta}) - \min_{\hat{\theta} \in \mathcal{H}_0} \text{Pred}_X(\hat{\theta}).
\]

The optimal interpolant is the OLS estimator \(\hat{\theta}_{\text{ols}} = X^\top (XX^\top)^{-1} y\).

**Theorem 2 (Cheng, Duchi, Kuditipudi ’22)**

Under proportional asymptotics \(d/n \to \gamma > 1\),

- *(no-cost phase)* \(\lim_{n \to \infty} \overline{\text{Cost}}_X(\epsilon) > 0\) iff \(\epsilon^2 > \epsilon_{\sigma,\text{ols}}^2\).
- *(linear-growth phase)* \(\lim_{n \to \infty} \overline{\text{Cost}}_X(\epsilon) \geq \overline{C}_\gamma \epsilon^2\) for \(\epsilon^2 \geq \overline{\epsilon}_\gamma \sigma^4\).
- *(threshold value)* \(\epsilon_{\sigma} < \epsilon_{\sigma,\text{ols}} \leq \kappa_{\gamma} \epsilon_{\sigma}\).
Relax assumptions

- General covariance
- The empirical spectral distribution of $\Sigma$ converges.
- The condition number of $\Sigma$ is bounded.

- General prior and noise distributions
- Gaussianity ensures model complexity.
- A counterexample when memorization does not happen is $\theta = \frac{e^j}{\sqrt{d}}$ with equal probability.

- For $\theta \sim (0, I_d/d)$ and $w \sim (0, \sigma^2 I_n)$, we restrict to linear estimators $H = n b_{\theta}(X, y)$: $b_{\theta} = A(X) y$.

**Theorem 3 (Cheng, Duchi, Kuditipudi '22)**

(Informal)

Under above conditions, we have to train till below $O(\sigma^4)$ error to generalize well.
Relax assumptions

General covariance

The empirical spectral distribution of $\Sigma$ converges. The condition number of $\Sigma$ is bounded.

General prior and noise distributions

Gaussianity ensures model complexity. A counterexample when memorization does not happen is $\theta = e^{ij}/\sqrt{d}$ with equal probability.

For $\theta \sim (0, I_d/d)$ and $w \sim (0, \sigma^2 I_n)$, we restrict to linear estimators $H = n b (\theta (X, y)) : b (\theta (X)) = A (X) y_0$.

Theorem 3 (Cheng, Duchi, Kuditipudi '22)

(Informal) Under above conditions, we have to train till below $O(\sigma^4)$ error to generalize well.
General covariance

- The empirical spectral distribution of $\Sigma$ converges.

Theorem 3 (Cheng, Duchi, Kuditipudi '22)
(Informal)
Under above conditions, we have to train till below $O(\sigma^4)$ error to generalize well.
Relax assumptions

General covariance

- The empirical spectral distribution of $\Sigma$ converges.
- The condition number of $\Sigma$ is bounded.
Relax assumptions

**General covariance**
- The empirical spectral distribution of $\Sigma$ converges.
- The condition number of $\Sigma$ is bounded.

**General prior and noise distributions**
Relax assumptions

General covariance
- The empirical spectral distribution of $\Sigma$ converges.
- The condition number of $\Sigma$ is bounded.

General prior and noise distributions
General covariance

- The empirical spectral distribution of $\Sigma$ converges.
- The condition number of $\Sigma$ is bounded.

General prior and noise distributions

- Gaussianity ensures model complexity. A counterexample when memorization does not happen is $\theta = e_j / \sqrt{d}$ with equal probability.
Relax assumptions

**General covariance**
- The empirical spectral distribution of \( \Sigma \) converges.
- The condition number of \( \Sigma \) is bounded.

**General prior and noise distributions**
- Gaussianity ensures model complexity. A counterexample when memorization does not happen is \( \theta = e_j / \sqrt{d} \) with equal probability.
- For \( \theta \sim (0, I_d/d) \) and \( w \sim (0, \sigma^2 I_n) \), we restrict to linear estimators

\[
\mathcal{H} = \left\{ \hat{\theta}(X, y) : \hat{\theta} = A(X)y \right\}.
\]
General covariance
- The empirical spectral distribution of $\Sigma$ converges.
- The condition number of $\Sigma$ is bounded.

General prior and noise distributions
- Gaussianity ensures model complexity. A counterexample when memorization does not happen is $\theta = e_j / \sqrt{d}$ with equal probability.
- For $\theta \sim (0, I_d/d)$ and $w \sim (0, \sigma^2 I_n)$, we restrict to linear estimators
  \[ H = \left\{ \hat{\theta}(X, y) : \hat{\theta} = A(X)y \right\}. \]

Theorem 3 (Cheng, Duchi, Kuditipudi ’22)
(Informal) Under above conditions, we have to train till below $O(\sigma^4)$ error to generalize well.
Concluding remarks
Conclusions

Necessity of memorization in linear regression

I must fit with accuracy >> inherent noise level.

John Duchi on an average day

For more details: arXiv:2202.09889
Conclusions

Necessity of memorization in linear regression

Similar results for other problems?
Conclusions

Necessity of memorization in linear regression

I must fit with accuracy >> inherent noise level.

Similar results for other problems?

Implications for data cleaning and security. Can we have both?
Conclusions

Necessity of memorization in linear regression

I must fit with accuracy $\gg$ inherent noise level.

John Duchi on an average day

Similar results for other problems?

Implications for data cleaning and security. Can we have both?

Motivation to construct datasets with multiple labels

- Theory of dataset with multiple labels. Hilal Asi, Chen Cheng, John Duchi.
- Surrogate consistency with data aggregation. Chen Cheng, John Duchi.

For more details: arXiv:2202.09889