Memorize to Generalize: on the Necessity of Interpolation in High Dimensional Linear Regression

Chen Cheng

Stanford University

Department of Statistics

Chen Cheng
Stanford Stat

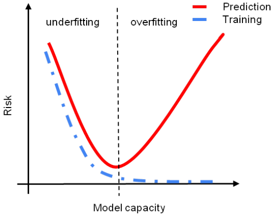
John Duchi
Stanford Stat & EE

Rohith Kuditipudi
Stanford CS

Introduction: why study memorization?

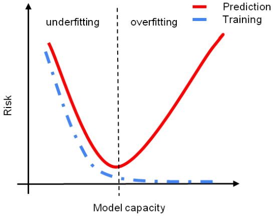
Interpolation in modern machine learning

Classical statistical wisdom



Interpolation in modern machine learning

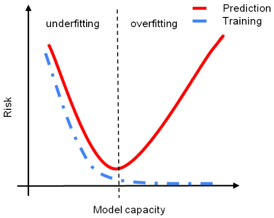
Classical statistical wisdom



- bigger models tend to overfit

Interpolation in modern machine learning

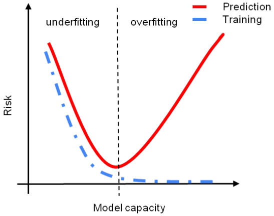
Classical statistical wisdom



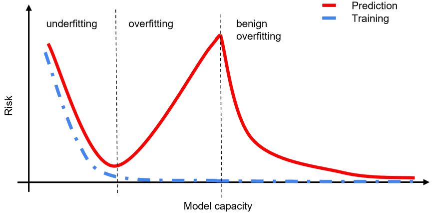
- bigger models tend to overfit
- need to limit model capacity

Interpolation in modern machine learning

Classical statistical wisdom

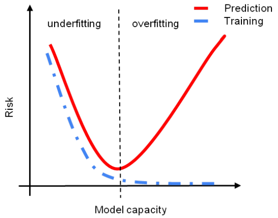


Modern empirical wisdom



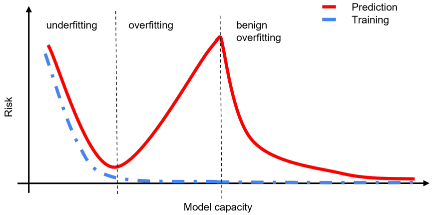
- bigger models tend to overfit
- need to limit model capacity

Classical statistical wisdom



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

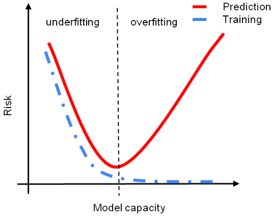
Modern empirical wisdom



- overparameterized model

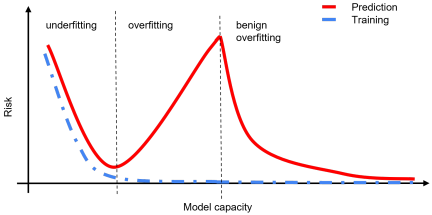
Interpolation in modern machine learning

Classical statistical wisdom



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

Modern empirical wisdom



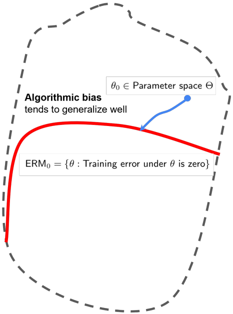
- overparameterized model
- interpolate data

Sufficiency of interpolation

When is it sufficient to overfit?

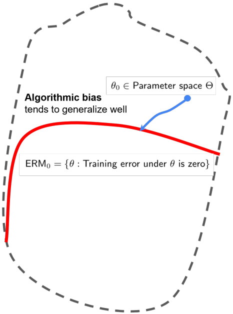
When is it sufficient to overfit?

Benign overfitting



When is it sufficient to overfit?

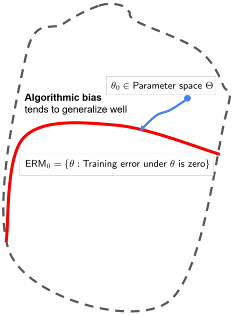
Benign overfitting



Literatures

When is it sufficient to overfit?

Benign overfitting

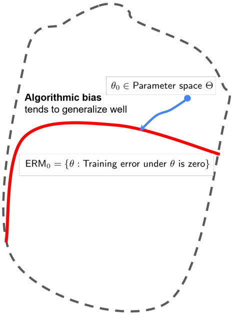


Literatures

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

When is it sufficient to overfit?

Benign overfitting



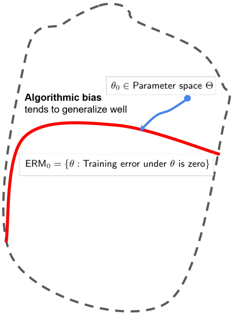
Literatures

- 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

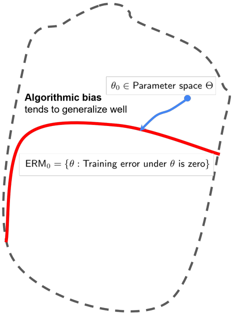


Literatures

- 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
  Liang and Rakhlin, 2020.

When is it sufficient to overfit?

Benign overfitting

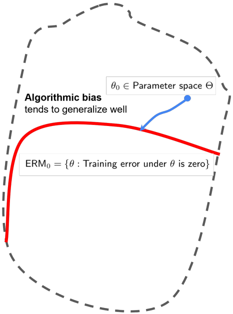


Literatures

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

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

Necessity of interpolation

When is it necessary to overfit?

When is it necessary to overfit?

Competing considerations

When is it necessary to overfit?

Competing considerations

- Privacy and security.

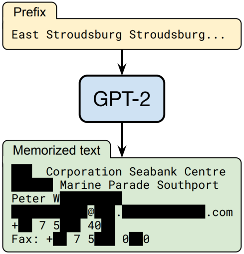


Figure from **Carlini et al., 2021**

Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

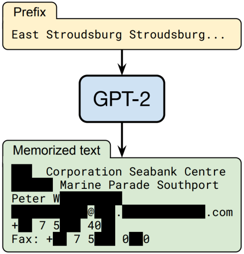


Figure from **Carlini et al., 2021**

- Can we generalize well without memorization?

Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

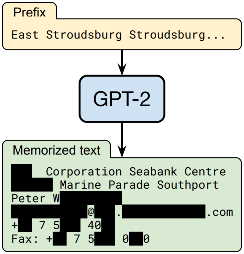


Figure from **Carlini et al., 2021**

- Can we generalize well without memorization?

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?
  Brown, Bun, Feldman, Smith, Talwar, 2021.

Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

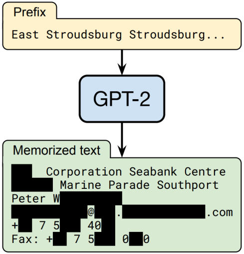


Figure from **Carlini et al., 2021**

- Can we generalize well without memorization?

Inspiring line of works

- 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?
  Brown, Bun, Feldman, Smith, Talwar, 2021.

Takeaways

Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

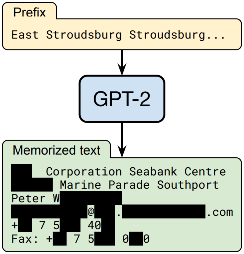


Figure from **Carlini et al., 2021**

- Can we generalize well without memorization?

Inspiring line of works

- 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?
  Brown, Bun, Feldman, Smith, Talwar, 2021.

Takeaways

- Heavy-tailed distributions.

Necessity of interpolation

When is it necessary to overfit?

Competing considerations

- Privacy and security.

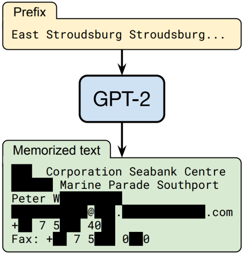


Figure from **Carlini et al., 2021**

- Can we generalize well without memorization?

Inspiring line of works

- 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?
  Brown, Bun, Feldman, Smith, Talwar, 2021.

Takeaways

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

When is it necessary to overfit?

Competing considerations

- Privacy and security.

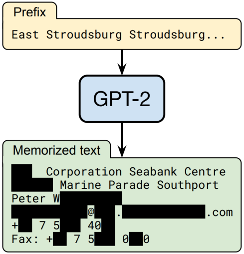


Figure from **Carlini et al., 2021**

- Can we generalize well without memorization?

Inspiring line of works

- 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?
  Brown, Bun, Feldman, Smith, Talwar, 2021.

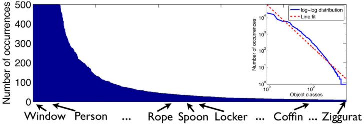
Takeaways

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

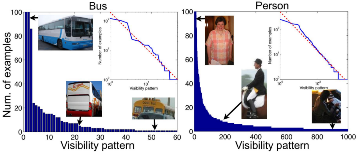
Necessity of interpolation

Real life data have heavy-tailed distributions

Real life data have heavy-tailed distributions

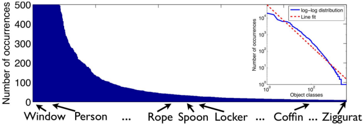


(a) The number of examples by object class in SUN dataset

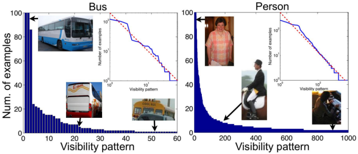


(b) Distributions of the visibility patterns for bus and person

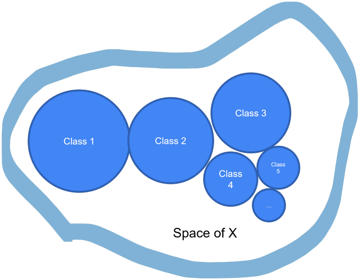
Real life data have heavy-tailed distributions



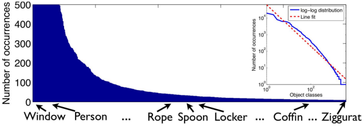
(a) The number of examples by object class in SUN dataset



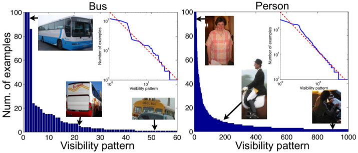
(b) Distributions of the visibility patterns for bus and person



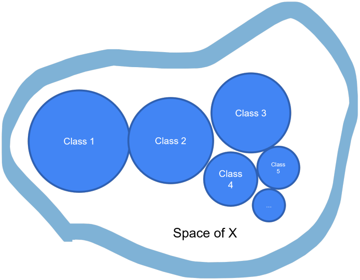
Real life data have heavy-tailed distributions



(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

Carefully constructed combinatorial settings

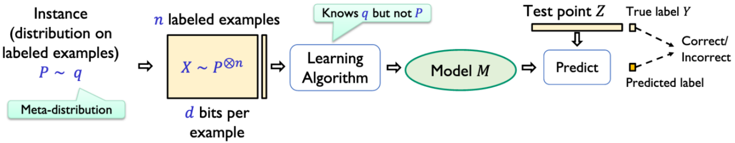


Figure from **Brown et al., 2021**

Carefully constructed combinatorial settings

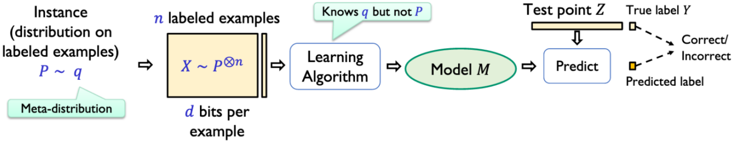
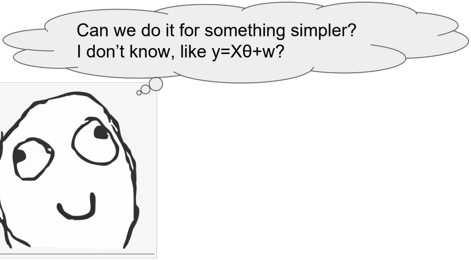


Figure from **Brown et al., 2021**



Average John

Necessity of interpolation

A general formulation

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 θ.

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 θ.

- The cost of not-fitting

$$\begin{aligned} \underset{\widehat{\theta} \in \mathcal{H}}{\text{minimize}} \quad & \text{Pred}\left(\widehat{\theta}\right) \\ \text{subject to} \quad & \text{Train}\left(\widehat{\theta}\right) \geq \epsilon^2 \end{aligned}$$

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 θ.

- The cost of not-fitting

$$\begin{aligned} \underset{\widehat{\theta} \in \mathcal{H}}{\text{minimize}} \quad & \text{Pred}\left(\widehat{\theta}\right) \\ \text{subject to} \quad & \text{Train}\left(\widehat{\theta}\right) \geq \epsilon^2 \end{aligned}$$

A simpler model

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 θ.

- The cost of not-fitting

$$\begin{aligned} \underset{\widehat{\theta}\in\mathcal{H}}{\text{minimize}} \quad & \text{Pred}\left(\widehat{\theta}\right) \\ \text{subject to} \quad & \text{Train}\left(\widehat{\theta}\right) \geq \epsilon^2 \end{aligned}$$

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 $\mathcal{H}$ containing θ.

- The cost of not-fitting

$$\begin{aligned} \underset{\widehat{\theta} \in \mathcal{H}}{\text{minimize}} \quad & \text{Pred}\left(\widehat{\theta}\right) \\ \text{subject to} \quad & \text{Train}\left(\widehat{\theta}\right) \geq \epsilon^2 \end{aligned}$$

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 θ.

- The cost of not-fitting

$$\begin{aligned} \underset{\widehat{\theta} \in \mathcal{H}}{\text{minimize}} \quad & \text{Pred}\left(\widehat{\theta}\right) \\ \text{subject to} \quad & \text{Train}\left(\widehat{\theta}\right) \geq \epsilon^2 \end{aligned}$$

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

Linear regression

Let's start from the isotropic Gaussian case

Linear regression

Let's start from the isotropic Gaussian case

Problem setup

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

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 $X = \mathbb{R}^{n\times d}$ $(d \geq n)$

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 $\theta \sim \mathcal{N}(0, I_d/d)$

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 $w \sim \mathcal{N}(0, \sigma^2 I_n)$

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 $w \sim \mathcal{N}(0, \sigma^2 I_n)$

ℓ_2 **error**

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

$$\text{Pred}_X\left(\widehat{\theta}\right) = \mathbb{E}_{x,w,\theta}\left[\left\|x^\top\theta - x^\top\widehat{\theta}\right\|_2^2\right]$$

Cost of not-fitting for linear regression

Cost of not-fitting for linear regression

We want to solve

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

where $\mathcal{H} = \left\{\widehat{\theta}(X, y) \text{ square integrable}\right\}$.

Cost of not-fitting for linear regression

We want to solve

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

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

$$\mathsf{Cost}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}} \mathsf{Pred}_X\left(\widehat{\theta}\right).$$

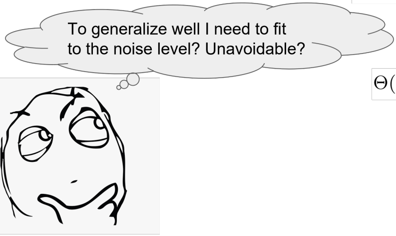
Cost of not-fitting for linear regression

We want to solve

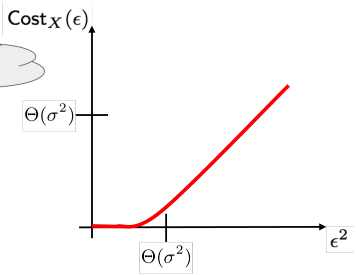
$$\min_{\widehat{\theta}\in\mathcal{H}} . \quad \text{Pred}_X\left(\widehat{\theta}\right) \quad \text{s.t.} \quad \text{Train}_X\left(\widehat{\theta}\right) \geq \epsilon^2 ,$$

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

$$\text{Cost}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \text{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}} \text{Pred}_X\left(\widehat{\theta}\right) .$$



Average John



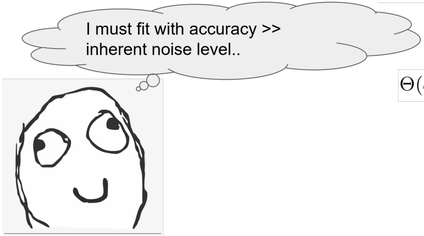
Cost of not-fitting for linear regression

We want to solve

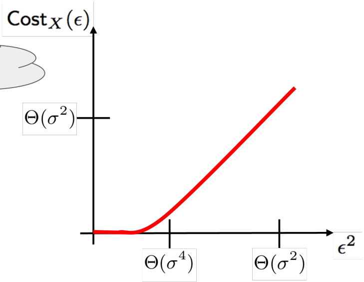
$$\min_{\widehat{\theta}\in\mathcal{H}}. \ \text{Pred}_X\left(\widehat{\theta}\right) \quad \text{s.t.} \quad \text{Train}_X\left(\widehat{\theta}\right) \geq \epsilon^2,$$

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

$$\text{Cost}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \text{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}} \text{Pred}_X\left(\widehat{\theta}\right).$$



John Duchi
on an average day

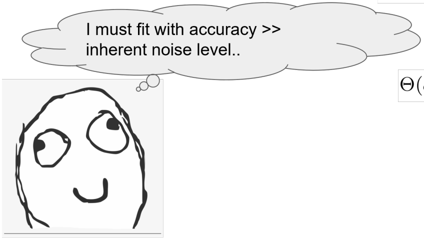


Cost of not-fitting for linear regression

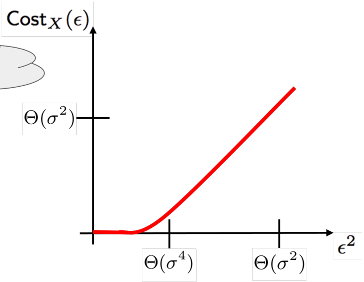
Theorem 1 (Cheng, Duchi, Kuditipudi '22)

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

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



John Duchi
on an average day

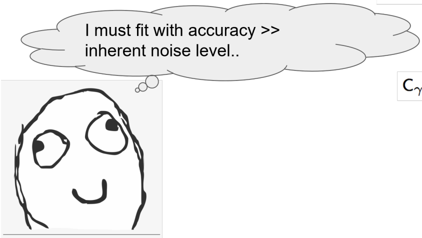


Cost of not-fitting for linear regression

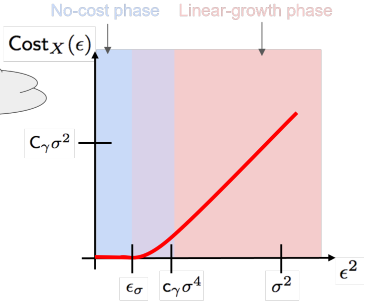
Theorem 1 (Cheng, Duchi, Kuditipudi '22)

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

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



John Duchi
on an average day



Proof sketch: strong duality and random matrix theory

Proof outline

The proof consists of three parts.

Proof outline

The proof consists of three parts.

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

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

  using strong duality.

Proof outline

The proof consists of three parts.

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

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

  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 $\widehat{\theta} = Ay$, we solve the nonconvex minimization problem

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

  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 ϵ_σ, cost of not-fitting $\mathsf{Cost}_X(\epsilon)$ by random matrix theory.

Proof outline

The proof consists of three parts.

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

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

  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 ϵ_σ, cost of not-fitting $\mathsf{Cost}_X(\epsilon)$ by random matrix theory.

- **Upgrade by functional strong duality.** Finally, we upgrade to any square integrable estimator $\widehat{\theta}(X, y)$ by showing a functional strong duality result.

Strong duality for linear hypothesis class

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

Strong duality for linear hypothesis class

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

Reduction to QCQP

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

Strong duality for linear hypothesis class

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

Reduction to QCQP

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

Strong duality

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

Strong duality for linear hypothesis class

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

Reduction to QCQP

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

Strong duality

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

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

Strong duality for linear hypothesis class

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

Reduction to QCQP

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

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 ϵ_σ^2 training error.

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 $\frac{1}{d} X X^\top$ is μ_n with its c.d.f. $H_n(s) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\lambda_i^2/d \leq s}$.

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 μ_n with its c.d.f. $H_n(s) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\lambda_i^2/d \leq s}$.

Marchenko-Pastur law

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

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 μ_n with its c.d.f. $H_n(s) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\lambda_i^2/d \leq s}$.

Marchenko-Pastur law

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

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

with $\lambda_\pm := \left(1 \pm 1/\sqrt{\gamma}\right)^2$.

Asymptotics by RMT

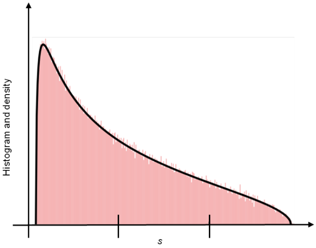
Let X have singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. The empirical spectral distribution of $\frac{1}{d}XX^\top$ is μ_n with its c.d.f. $H_n(s) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\lambda_i^2/d \leq s}$.

Marchenko-Pastur law

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

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

with $\lambda_\pm := \left(1 \pm 1/\sqrt{\gamma}\right)^2$.



Asymptotics by RMT

Prediction and training errors

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

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

Prediction and training errors

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

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

Prediction and training errors in ESD

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

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

Prediction and training errors

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

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

Prediction and training errors in ESD

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

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

Limit of Lagrange multiplier

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

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

Asymptotics by RMT

Asymptotics by RMT

Limit of threshold

Taking $\rho_\epsilon = 0$ gives

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

Asymptotics by RMT

Limit of threshold

Taking $\rho_\epsilon = 0$ gives

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

Limit of cost of not-fitting

$$\begin{aligned} \text{Cost}_X(\epsilon) &= \mathcal{P}(A(\rho_n)) - \mathcal{P}(A(0)) = \mathcal{P}(A(\rho_\epsilon)) - \mathcal{P}(A(0)) + o_n(1) \\ &\to \frac{\rho_\epsilon^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho_\epsilon s)^2 (s+\sigma^2)} dH(s) \end{aligned}$$

Asymptotics by RMT

Limit of threshold

Taking $\rho_\epsilon = 0$ gives

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

Limit of cost of not-fitting

$$\begin{aligned} \mathsf{Cost}_X(\epsilon) &= \mathcal{P}(A(\rho_n)) - \mathcal{P}(A(0)) = \mathcal{P}(A(\rho_\epsilon)) - \mathcal{P}(A(0)) + o_n(1) \\ &\to \frac{\rho_\epsilon^2}{\gamma} \int \frac{\sigma^4 s}{(1-\rho_\epsilon s)^2 (s+\sigma^2)} dH(s) \end{aligned}$$

Theorem 1 (Cheng, Duchi, Kuditipudi '22)

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

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

Upgrade to general hypothesis class

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

Upgrade to general hypothesis class

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

$$\begin{aligned} \underset{\widehat{\theta}(X,y) \in L^2}{\text{minimize}} \quad & \int \left\| \widehat{\theta} - \left(X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 d\mu \\ \text{subject to} \quad & \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu \geq \epsilon^2 \end{aligned}$$

where $\mu \overset{d}{=} \mathsf{N}(0, \frac{1}{d} XX^\top + \sigma^2 I)$.

Upgrade to general hypothesis class

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

$$\begin{aligned} \underset{\widehat{\theta}(X,y)\in L^2}{\text{minimize}} \quad & \int \left\| \widehat{\theta} - \left(X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 d\mu \\ \text{subject to} \quad & \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu \geq \epsilon^2 \end{aligned}$$

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

Upgrade to general hypothesis class

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

$$\begin{aligned} \underset{\widehat{\theta}(X,y)\in L^2}{\text{minimize}} \quad & \int \left\| \widehat{\theta} - \left(X^\top X + d\sigma^2 I\right)^{-1} X^\top y \right\|_2^2 d\mu \\ \text{subject to} \quad & \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu \geq \epsilon^2 \end{aligned}$$

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

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

Upgrade to general hypothesis class

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

$$\begin{aligned} \underset{\widehat{\theta}(X,y)\in L^2}{\text{minimize}} \quad & \int \left\| \widehat{\theta} - \left(X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 d\mu \\ \text{subject to} \quad & \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu \geq \epsilon^2 \end{aligned}$$

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

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

We exactly have

$$\widehat{\theta} = A(X)y$$

Upgrade to general hypothesis class

Functional strong duality

Upgrade to general hypothesis class

Functional strong duality

$$\begin{aligned} \underset{\widehat{\theta}(X, y_i) \in \mathbb{R}^d, 1 \leq i \leq m}{\text{minimize}} \quad & \int \left\| \widehat{\theta} - \left(X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 d\mu_m \\ \text{subject to} \quad & \int \left\| X \widehat{\theta} - y \right\|_2^2 d\mu_m \geq \epsilon^2 \end{aligned}$$

where μ_m are empirical distributions for i.i.d. samples of $y \mid X$.

Upgrade to general hypothesis class

Functional strong duality

$$\begin{aligned} \underset{\widehat{\theta}(X, y_i) \in \mathbb{R}^d, 1 \le i \le m}{\text{minimize}} \quad & \int \left\| \widehat{\theta} - \left(X^\top X + d\sigma^2 I \right)^{-1} X^\top y \right\|_2^2 d\mu_m \\ \text{subject to} \quad & \int \left\| X\widehat{\theta} - y \right\|_2^2 d\mu_m \ge \epsilon^2 \end{aligned}$$

where μ_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

Cost of not-interpolating

Cost of not-fitting

$$\mathsf{Cost}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}} \mathsf{Pred}_X\left(\widehat{\theta}\right).$$

Cost of not-interpolating

Cost of not-fitting

$$\mathsf{Cost}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}} \mathsf{Pred}_X\left(\widehat{\theta}\right).$$

Cost of not-interpolating

$$\overline{\mathsf{Cost}}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}_0} \mathsf{Pred}_X\left(\widehat{\theta}\right).$$

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

Cost of not-interpolating

Cost of not-fitting

$$\mathsf{Cost}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}} \mathsf{Pred}_X\left(\widehat{\theta}\right).$$

Cost of not-interpolating

$$\overline{\mathsf{Cost}}_X(\epsilon) := \min_{\widehat{\theta}\in\mathcal{H}(\epsilon)} \mathsf{Pred}_X\left(\widehat{\theta}\right) - \min_{\widehat{\theta}\in\mathcal{H}_0} \mathsf{Pred}_X\left(\widehat{\theta}\right).$$

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

Theorem 2 (Cheng, Duchi, Kuditipudi '22)

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

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

Relax assumptions

Relax assumptions

General covariance

Relax assumptions

General covariance

- The empirical spectral distribution of Σ converges.

Relax assumptions

General covariance

- The empirical spectral distribution of Σ converges.
- The condition number of Σ is bounded.

Relax assumptions

General covariance

- The empirical spectral distribution of Σ converges.
- The condition number of Σ is bounded.

General prior and noise distributions

Relax assumptions

General covariance

- The empirical spectral distribution of Σ converges.
- The condition number of Σ is bounded.

General prior and noise distributions

Relax assumptions

General covariance

- The empirical spectral distribution of Σ converges.
- The condition number of Σ 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 Σ converges.
- The condition number of Σ 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

$$\mathcal{H} = \left\{ \widehat{\theta}(X, y) : \widehat{\theta} = A(X)y \right\}.$$

Relax assumptions

General covariance

- The empirical spectral distribution of Σ converges.
- The condition number of Σ 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

$$\mathcal{H} = \left\{ \widehat{\theta}(X, y) : \widehat{\theta} = A(X)y \right\}.$$

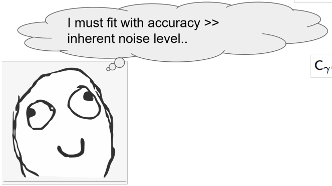
Theorem 3 (Cheng, Duchi, Kuditipudi '22)

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

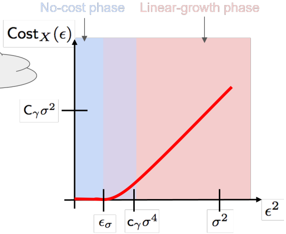
Concluding remarks

Conclusions

Necessity of memorization in linear regression

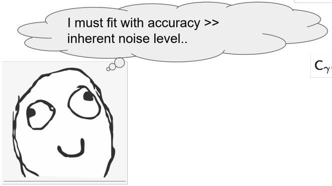


John Duchi
on an average day

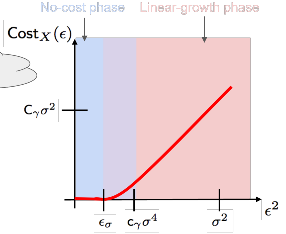


Conclusions

Necessity of memorization in linear regression



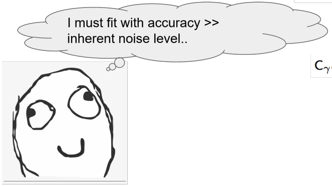
John Duchi
on an average day



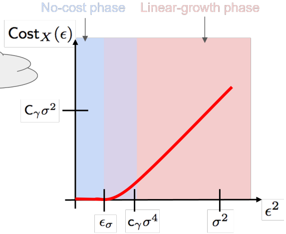
Similar results for other problems?

Conclusions

Necessity of memorization in linear regression



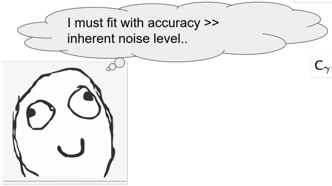
John Duchi
on an average day



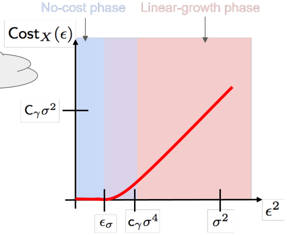
Similar results for other problems?

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

Necessity of memorization in linear regression



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