Benign Overfitting in Linear and Nonlinear Settings

Peter Bartlett
UC Berkeley

LMS Invited Lectures on the Mathematics of Deep Learning
Newton Institute, Cambridge
3 March, 2022

SIMONS FOUNDATION


Niladri Chatterji


Spencer Frei


Phil Long


Gábor Lugosi


Andrea Montanari


Alexander Rakhlin


Alexander Tsigler

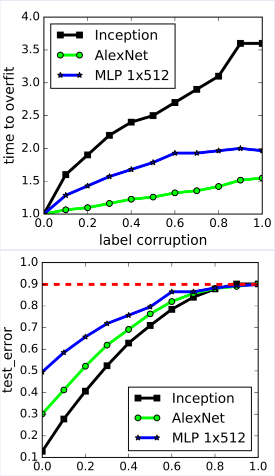
Overfitting in Deep Networks

- Deep networks can be trained to zero training error (for *regression* loss)

Overfitting in Deep Networks

- Deep networks can be trained to zero training error (for *regression* loss)
- ... with near state-of-the-art performance

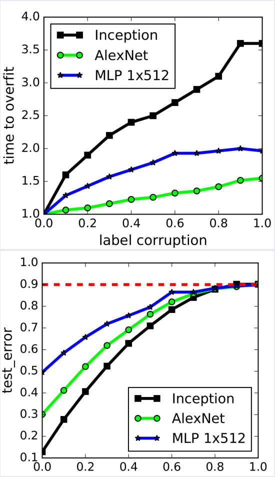
Overfitting in Deep Networks



(Zhang, Bengio, Hardt, Recht, Vinyals, 2017)

- Deep networks can be trained to zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for *noisy* problems.

Overfitting in Deep Networks

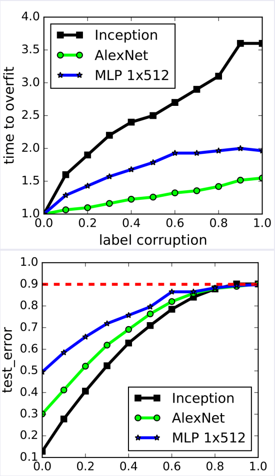


(Zhang, Bengio, Hardt, Recht, Vinyals, 2017)

- Deep networks can be trained to zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for *noisy* problems.

also (Belkin, Hsu, Ma, Mandal, 2018)

Overfitting in Deep Networks

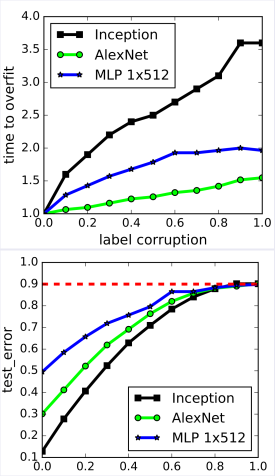


(Zhang, Bengio, Hardt, Recht, Vinyals, 2017)

- Deep networks can be trained to zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for *noisy* problems.
- No tradeoff between fit to training data and complexity!

also (Belkin, Hsu, Ma, Mandal, 2018)

Overfitting in Deep Networks



(Zhang, Bengio, Hardt, Recht, Vinyals, 2017)

- Deep networks can be trained to zero training error (for *regression* loss)
- ... with near state-of-the-art performance
- ... even for *noisy* problems.
- No tradeoff between fit to training data and complexity!
- *Benign overfitting.*

also (Belkin, Hsu, Ma, Mandal, 2018)

Statistical Wisdom and Overfitting

"... interpolating fits... [are] unlikely to predict future data well at all."

22 *2. How to Construct Nonparametric Regression Estimates?*

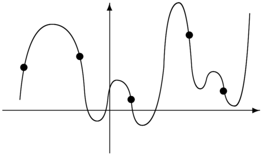
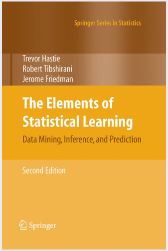
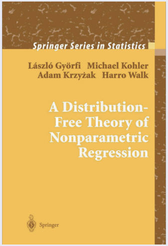


Figure 2.3. The estimate on the right seems to be more reasonable than the estimate on the left, which interpolates the data.

over $\mathcal{F}_n$. Least squares estimates are defined by minimizing the empirical L_2 risk over a general set of functions $\mathcal{F}_n$ (instead of (2.7)). Observe that it doesn't make sense to minimize (2.9) over all (measurable) functions f, because this may lead to a function which interpolates the data and hence is not a reasonable estimate. Thus one has to restrict the set of functions over





see also (B. and Rakhlin, Simons Institute, May 2019)

Benign Overfitting

A new statistical phenomenon:
good prediction with very small training error for regression loss

- Statistical wisdom says a prediction rule should not fit too well.
- But deep networks are trained to fit noisy data perfectly, and they predict well.

Belkin, Hsu and Mitra, 2018; Belkin, Rakhlin and Tsybakov, 2018

Liang and Rakhlin, 2018;

Benign Overfitting

A new statistical phenomenon:
good prediction with very small training error for regression loss

- Statistical wisdom says a prediction rule should not fit too well.
- But deep networks are trained to fit noisy data perfectly, and they predict well.

Belkin, Hsu and Mitra, 2018; Belkin, Rakhlin and Tsybakov, 2018

Liang and Rakhlin, 2018;

Belkin, Hsu, Ma and Mandal, 2019; Belkin, Hsu and Xu, 2019; Bibas, Fogel and Feder, 2019; Hastie, Montanari, Rosset and Tibshirani, 2019;

Benign Overfitting

A new statistical phenomenon:
good prediction with very small training error for regression loss

- Statistical wisdom says a prediction rule should not fit too well.
- But deep networks are trained to fit noisy data perfectly, and they predict well.

Belkin, Hsu and Mitra, 2018; Belkin, Rakhlin and Tsybakov, 2018

Liang and Rakhlin, 2018;

Belkin, Hsu, Ma and Mandal, 2019; Belkin, Hsu and Xu, 2019; Bibas, Fogel and Feder, 2019; Hastie, Montanari, Rosset and Tibshirani, 2019; Dereziński, Liang and Mahoney, 2019; Liang, Rakhlin and Zhai, 2019; Mei and Montanari, 2019; Mitra, 2019; Muthukumar, Vodrahalli and Sahai, 2019; Nakkiran, 2019; Bunea, Strimas-Mackey, Wegkamp, 2020; Chinot and Lerasle, 2020; Chinot, Löffler, van de Geer, 2020; Kobak, Lomond and Sanchez, 2020; Nakkiran, Venkat, Kakade and Ma, 2020; Hastie, Montanari, Rosset and Tibshirani, 2020; Mei, Misiakiewicz, Montanari, 2021; Celentano, Misiakiewicz, Montanari, 2021; Zou, Wu, Braverman, Gu and Kakade, 2021; Li, Zhou, Gretton, 2021; Minsker, Ndaoud, Shen, 2021;

Benign Overfitting

**A new statistical phenomenon:
good prediction with very small training error for regression loss**

- Statistical wisdom says a prediction rule should not fit too well.
- But deep networks are trained to fit noisy data perfectly, and they predict well.

Belkin, Hsu and Mitra, 2018; Belkin, Rakhlin and Tsybakov, 2018

Liang and Rakhlin, 2018;

Belkin, Hsu, Ma and Mandal, 2019; Belkin, Hsu and Xu, 2019; Bibas, Fogel and Feder, 2019; Hastie, Montanari, Rosset and Tibshirani, 2019; Dereziński, Liang and Mahoney, 2019; Liang, Rakhlin and Zhai, 2019; Mei and Montanari, 2019; Mitra, 2019; Muthukumar, Vodrahalli and Sahai, 2019; Nakkiran, 2019; Bunea, Strimas-Mackey, Wegkamp, 2020; Chinot and Lerasle, 2020; Chinot, Löffler, van de Geer, 2020; Kobak, Lomond and Sanchez, 2020; Nakkiran, Venkat, Kakade and Ma, 2020; Hastie, Montanari, Rosset and Tibshirani, 2020; Mei, Misiakiewicz, Montanari, 2021; Celentano, Misiakiewicz, Montanari, 2021; Zou, Wu, Braverman, Gu and Kakade, 2021; Li, Zhou, Gretton, 2021; Minsker, Ndaoud, Shen, 2021;

Deep learning: a statistical viewpoint. B., Montanari, Rakhlin. *Acta Numerica*. 2021. arXiv:2103.09177

Benign Overfitting

Intuition

- Benign overfitting prediction rule $\hat{f}$ decomposes as

$$\hat{f} = \hat{f}_0 + \Delta.$$

- $\hat{f}_0$ = simple component useful for *prediction*.
- Δ = spiky component useful for *benign overfitting*.
- Classical statistical learning theory applies to $\hat{f}_0$.
- Δ is not useful for prediction, but it is benign.

(Deep learning: a statistical viewpoint. B., Montanari, Rakhlin. *Acta Numerica*. 2021)

Benign Overfitting

Example: kernel smoothing

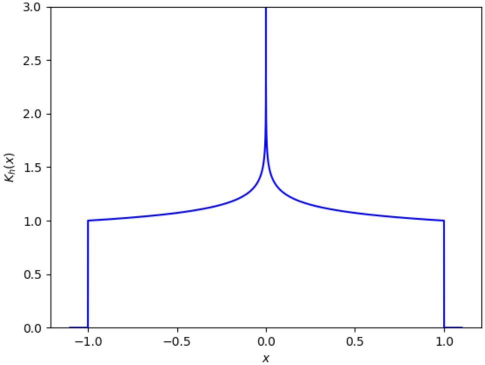
$$\hat{f}(x) = \sum_{i=1}^{n} \frac{y_i K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)}$$

Benign Overfitting

Example: kernel smoothing with singular, compact kernels

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{y_i K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)} \qquad \text{e.g., with } K_h(x) = \frac{1\left[h\|x\| \le 1\right]}{h\|x\|^{\alpha}}.$$

Minimax rates (with suitable h). (Belkin, Rakhlin, Tsybakov, 2018), (Belkin, Hsu, Mitra, 2018)



Benign Overfitting

Example: kernel smoothing with singular, compact kernels

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{y_i K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)} \qquad \text{e.g., with } K_h(x) = \frac{1\left[h\|x\| \le 1\right]}{h\|x\|^{\alpha}}.$$

Minimax rates (with suitable h). (Belkin, Rakhlin, Tsybakov, 2018), (Belkin, Hsu, Mitra, 2018)

- Benign overfitting prediction rule $\hat{f}$ decomposes as

$$\hat{f} = \hat{f}_0 + \Delta.$$

- $\hat{f}_0$ = simple component useful for *prediction*: standard (e.g., constant) compact kernel
- Δ = spiky component useful for *benign overfitting*: spiky piece (with small norm in $L_2(P)$).

Outline

- *Linear regression*
- Characterizing benign overfitting
- Ridge regression
- Beyond linear settings

Definitions

Simple Prediction Setting: Linear Regression

Definitions

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.

Definitions

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- Assumptions:
  (x, y) subgaussian, mean zero, well-specified: $\mathbb{E}[y|x] = x^\top \theta^*$.
  x satisfies a small ball condition: $\exists c > 0$, $\Pr\left(\|x\|^2 < c\mathbb{E}\|x\|^2\right) \leq \delta$.

Definitions

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- Assumptions:
  (x, y) subgaussian, mean zero, well-specified: $\mathbb{E}[y|x] = x^\top \theta^*$.
  x satisfies a small ball condition: $\exists c > 0$, $\Pr\left(\|x\|^2 < c\mathbb{E}\|x\|^2\right) \leq \delta$.
- Define:

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad (\text{assume } \lambda_1 \geq \lambda_2 \geq \cdots)$$

Definitions

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- Assumptions:
  (x, y) subgaussian, mean zero, well-specified: $\mathbb{E}[y|x] = x^\top \theta^*$.
  x satisfies a small ball condition: $\exists c > 0$, $\Pr\left(\|x\|^2 < c\mathbb{E}\|x\|^2\right) \leq \delta$.
- Define:

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad (\text{assume } \lambda_1 \geq \lambda_2 \geq \cdots)$$

$$\theta^* := \arg\min_\theta \mathbb{E}\left(y - x^\top \theta\right)^2,$$

Definitions

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- Assumptions:
  (x, y) subgaussian, mean zero, well-specified: $\mathbb{E}[y|x] = x^\top \theta^*$.
  x satisfies a small ball condition: $\exists c > 0$, $\Pr\left(\|x\|^2 < c\mathbb{E}\|x\|^2\right) \leq \delta$.
- Define:

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad (\text{assume } \lambda_1 \geq \lambda_2 \geq \cdots)$$

$$\theta^* := \arg\min_\theta \mathbb{E}\left(y - x^\top \theta\right)^2,$$

$$\sigma^2 := \mathbb{E}(y - x^\top \theta^*)^2.$$

Definitions

Minimum norm estimator

Definitions

Minimum norm estimator

- Data: $X \in \mathbb{H}^n$, $y \in \mathbb{R}^n$.

$$X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Definitions

Minimum norm estimator

- Data: $X \in \mathbb{H}^n$, $y \in \mathbb{R}^n$.
- Estimator $\hat{\theta} = \left(X^\top X\right)^\dagger X^\top y$, which solves

$$
\begin{aligned}
\min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\
\text{s.t.} \quad & \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 .
\end{aligned}
$$

$$
X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}
$$

Definitions

Minimum norm estimator

- Data: $X \in \mathbb{H}^n$, $y \in \mathbb{R}^n$.
- Estimator $\hat{\theta} = \left(X^\top X\right)^\dagger X^\top y$, which solves

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{s.t.} \quad & \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 . \end{aligned}$$

$$X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Notice that gradient flow, initialized at 0:

$$\theta_0 = 0, \qquad \dot{\theta}_t = -\nabla_\theta \|X\theta - y\|^2$$

converges to the minimum norm solution.

Definitions

Excess prediction error

$$R(\hat{\theta}) := \mathbb{E}_{(x,y)}\left(y - x^\top \hat{\theta}\right)^2 - \underbrace{\min_{\theta} \mathbb{E}\left(y - x^\top \theta\right)^2}_{\text{optimal prediction error}}$$

Definitions

Excess prediction error

$$R(\hat{\theta}) := \mathbb{E}_{(x,y)} \left(y - x^\top \hat{\theta}\right)^2 - \underbrace{\min_{\theta} \mathbb{E} \left(y - x^\top \theta\right)^2}_{\text{optimal prediction error}}$$

$$= \mathbb{E}_{(x,y)} \left[\left(y - x^\top \hat{\theta}\right)^2 - \left(y - x^\top \theta^*\right)^2 \right]$$

Definitions

Excess prediction error

$$R(\hat{\theta}) := \mathbb{E}_{(x,y)} \left(y - x^\top \hat{\theta} \right)^2 - \underbrace{\min_\theta \mathbb{E} \left(y - x^\top \theta \right)^2}_{\text{optimal prediction error}}$$

$$= \mathbb{E}_{(x,y)} \left[\left(y - x^\top \hat{\theta} \right)^2 - \left(y - x^\top \theta^* \right)^2 \right]$$

$$= \left(\hat{\theta} - \theta^* \right)^\top \Sigma \left(\hat{\theta} - \theta^* \right).$$

Definitions

Excess prediction error

$$R(\hat{\theta}) := \mathbb{E}_{(x,y)} \left(y - x^\top \hat{\theta}\right)^2 - \underbrace{\min_{\theta} \mathbb{E} \left(y - x^\top \theta\right)^2}_{\text{optimal prediction error}}$$

$$= \mathbb{E}_{(x,y)} \left[\left(y - x^\top \hat{\theta}\right)^2 - \left(y - x^\top \theta^*\right)^2 \right]$$

$$= \left(\hat{\theta} - \theta^*\right)^\top \Sigma \left(\hat{\theta} - \theta^*\right).$$

So Σ determines the importance of parameter directions.

(Recall that $\Sigma = \sum_i \lambda_i v_i v_i^\top$ for orthonormal v_i, $\lambda_1 \geq \lambda_2 \geq \cdots$.)

Outline

- Linear regression
- *Characterizing benign overfitting*
- Ridge regression
- Beyond linear settings

From regularization to overfitting

Regularized linear regression

$$\min \qquad \lambda\|\theta\|^2 + \frac{1}{n}\|X\theta - y\|^2,$$

From regularization to overfitting

Regularized linear regression

$$\min \quad \lambda\|\theta\|^2 + \frac{1}{n}\|X\theta - y\|^2,$$

$$\begin{aligned} \min \quad & \|X\theta - y\|^2 \\ \text{s.t.} \quad & \|\theta\| \leq b, \end{aligned}$$

From regularization to overfitting

Regularized linear regression

$$\min \quad \lambda\|\theta\|^2 + \frac{1}{n}\|X\theta - y\|^2,$$

$$\begin{aligned} \min \quad & \|X\theta - y\|^2 \\ \text{s.t.} \quad & \|\theta\| \le b, \end{aligned}$$

$$\begin{aligned} \min \quad & \|\theta\| \\ \text{s.t.} \quad & \frac{1}{n}\|X\theta - y\|^2 \le c. \end{aligned}$$

From regularization to overfitting

Regularized linear regression

$$\min \quad \lambda\|\theta\|^2 + \frac{1}{n}\|X\theta - y\|^2,$$

$$\begin{aligned} \min \quad & \|X\theta - y\|^2 \\ \text{s.t.} \quad & \|\theta\| \leq b, \end{aligned}$$

$$\begin{aligned} \min \quad & \|\theta\| \\ \text{s.t.} \quad & \frac{1}{n}\|X\theta - y\|^2 \leq c. \end{aligned}$$

- The overfitting regime: $c \ll \min_\theta \mathbb{E}\left(y - x^\top\theta\right)^2.$

Interpolating Linear Regression

Overfitting regime

- We consider situations where $\min_\beta \|X\beta - y\|^2 = 0$.

Interpolating Linear Regression

Overfitting regime

- We consider situations where $\min_\beta \|X\beta - y\|^2 = 0$.
- Estimator $\hat{\theta} = (X^\top X)^\dagger X^\top y$ solves

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{s.t.} \quad & \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 = 0. \end{aligned}$$

Interpolating Linear Regression

Overfitting regime

- We consider situations where $\min_\beta \|X\beta - y\|^2 = 0$.
- Estimator $\hat{\theta} = (X^\top X)^\dagger X^\top y$ solves

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{s.t.} \quad & \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 = 0. \end{aligned}$$

- Hence, $y_1 = x_1^\top \hat{\theta}, \ldots, y_n = x_n^\top \hat{\theta}$.

Interpolating Linear Regression

Overfitting regime

- We consider situations where $\min_\beta \|X\beta - y\|^2 = 0$.
- Estimator $\hat{\theta} = (X^\top X)^\dagger X^\top y$ solves

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{s.t.} \quad & \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 = 0. \end{aligned}$$

- Hence, $y_1 = x_1^\top \hat{\theta}, \ldots, y_n = x_n^\top \hat{\theta}$.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?

Benign Overfitting: A Characterization

Theorem (B., Long, Lugosi, Tsigler, 2019), (Tsigler, B., 2020)

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$,

Benign Overfitting: A Characterization

Theorem (B., Long, Lugosi, Tsigler, 2019), (Tsigler, B., 2020)

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

Benign Overfitting: A Characterization

Theorem (B., Long, Lugosi, Tsigler, 2019), (Tsigler, B., 2020)

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

Benign Overfitting: A Characterization

Theorem (B., Long, Lugosi, Tsigler, 2019), (Tsigler, B., 2020)

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

2. If $X = \Sigma^{1/2}Z$ where Z has independent components and θ^* is symmetrized (random sign flips of components),
$$\mathbb{E}R(\hat{\theta}) \geq \frac{1}{c}\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \min\left\{\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1\right\}\right).$$

Benign Overfitting: A Characterization

Theorem (B., Long, Lugosi, Tsigler, 2019), (Tsigler, B., 2020)

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

1. With high probability,
$$R(\hat{\theta}) \leq c \left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right),$$
2. If $X = \Sigma^{1/2} Z$ where Z has independent components and θ^* is symmetrized (random sign flips of components),
$$\mathbb{E} R(\hat{\theta}) \geq \frac{1}{c} \left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \min\left\{ \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1 \right\} \right).$$

Here, $\text{bias}(\theta^*, \Sigma, n) = \|\theta^*_{k+1:\infty}\|^2_{\Sigma_{k+1:\infty}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}} \left(\frac{\sum_{i>k} \lambda_i}{n} \right)^2.$

Notions of Effective Rank

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

Notions of Effective Rank

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

Lemma

$$1 \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma).$$

Notions of Effective Rank

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

Examples

Notions of Effective Rank

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

Examples

1. $r_0(I_p) = R_0(I_p) = p$.

Notions of Effective Rank

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

Examples

1. $r_0(I_p) = R_0(I_p) = p$.
2. If $\mathrm{rank}(\Sigma) = p$, we can write

$$r_0(\Sigma) = \mathrm{rank}(\Sigma) s(\Sigma), \qquad R_0(\Sigma) = \mathrm{rank}(\Sigma) S(\Sigma),$$

$$\text{with} \quad s(\Sigma) = \frac{1/p \sum_{i=1}^p \lambda_i}{\lambda_1}, \qquad S(\Sigma) = \frac{\left(1/p \sum_{i=1}^p \lambda_i\right)^2}{1/p \sum_{i=1}^p \lambda_i^2}.$$

Both s and S lie between $1/p$ ($\lambda_2 \approx 0$) and 1 (λ_i all equal).

Benign Overfitting: A Characterization

Theorem

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

2. With some independence properties,

$$\mathbb{E}R(\hat{\theta}) \geq \frac{1}{c}\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \min\left\{\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1\right\}\right).$$

Benign Overfitting: A Characterization

Theorem

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

2. With some independence properties,

$$\mathbb{E}R(\hat{\theta}) \geq \frac{1}{c}\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \min\left\{\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1\right\}\right).$$

$$\text{bias}(\theta^*, \Sigma, n) = \|\theta^*_{k+1:\infty}\|^2_{\Sigma_{k+1:\infty}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{\sum_{i>k}\lambda_i}{n}\right)^2.$$

Benign Overfitting

- Benign overfitting prediction rule $\hat{f}$ decomposes as

$$\hat{f} = \hat{f}_0 + \Delta.$$

- $\hat{f}_0$ = *prediction* component:
  k^*-dim subspace corresponding to $\lambda_1, \ldots, \lambda_{k^*}$.
- Δ = *benign overfitting* component:
  orthogonal subspace. Δ is benign only if $R_{k^*} \gg n$.

Benign Overfitting: A Characterization

Intuition

- The mix of eigenvalues of Σ determines:
  1. how the label noise is distributed in $\hat{\theta}$, and

Benign Overfitting: A Characterization

Intuition

- The mix of eigenvalues of Σ determines:
  1. how the label noise is distributed in $\hat{\theta}$, and
  2. how errors in $\hat{\theta}$ affect prediction accuracy.

Benign Overfitting: A Characterization

Intuition

- The mix of eigenvalues of Σ determines:
  1. how the label noise is distributed in $\hat{\theta}$, and
  2. how errors in $\hat{\theta}$ affect prediction accuracy.
- To avoid harming prediction accuracy, the noise energy must be distributed across many unimportant directions.

Benign Overfitting: A Characterization

Intuition

- The mix of eigenvalues of Σ determines:
  1. how the label noise is distributed in $\hat{\theta}$, and
  2. how errors in $\hat{\theta}$ affect prediction accuracy.
- To avoid harming prediction accuracy, the noise energy must be distributed across many unimportant directions.
- Overparameterization is essential for benign overfitting

Benign Overfitting: A Characterization

Intuition

- The mix of eigenvalues of Σ determines:
  1. how the label noise is distributed in $\hat{\theta}$, and
  2. how errors in $\hat{\theta}$ affect prediction accuracy.
- To avoid harming prediction accuracy, the noise energy must be distributed across many unimportant directions.
- Overparameterization is essential for benign overfitting
  - Number of 'small' eigenvalues: large compared to n,

Benign Overfitting: A Characterization

Intuition

- The mix of eigenvalues of Σ determines:
  1. how the label noise is distributed in $\hat{\theta}$, and
  2. how errors in $\hat{\theta}$ affect prediction accuracy.
- To avoid harming prediction accuracy, the noise energy must be distributed across many unimportant directions.
- Overparameterization is essential for benign overfitting
  - Number of 'small' eigenvalues: large compared to n,
  - Small eigenvalues: roughly equal (but they can be more assymmetric if there are many more than n of them).

Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top\theta^*$ and $y - x^\top\theta^*$)

Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.

Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.

Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.

     Not a problem, even in high dimensions ($p > n$).
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.

Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.

     Not a problem, even in high dimensions ($p > n$).
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.

     Problematic.

Benign Overfitting: Proof Ideas

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top\theta^*$ and $y - x^\top\theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.

     Not a problem, even in high dimensions ($p > n$).
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.

     Problematic.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?

Benign Overfitting: Proof Ideas

Bias-variance decomposition

Define the noise vector ϵ by $y = X\theta^* + \epsilon$.

Benign Overfitting: Proof Ideas

Bias-variance decomposition

Define the noise vector ϵ by $y = X\theta^* + \epsilon$.

Estimator: $\hat{\theta} = (X^\top X)^\dagger X^\top y$

Benign Overfitting: Proof Ideas

Bias-variance decomposition

Define the noise vector ϵ by $y = X\theta^* + \epsilon$.

Estimator: $$\hat{\theta} = (X^\top X)^\dagger X^\top y = (X^\top X)^\dagger X^\top (X\theta^* + \epsilon),$$

Benign Overfitting: Proof Ideas

Bias-variance decomposition

Define the noise vector ϵ by $y = X\theta^* + \epsilon$.

Estimator: $$\hat{\theta} = (X^\top X)^\dagger X^\top y = (X^\top X)^\dagger X^\top (X\theta^* + \epsilon),$$

Excess risk: $$R(\hat{\theta}) = \left(\hat{\theta} - \theta^*\right)^\top \Sigma \left(\hat{\theta} - \theta^*\right)$$

Benign Overfitting: Proof Ideas

Bias-variance decomposition

Define the noise vector ϵ by $y = X\theta^* + \epsilon$.

$$\text{Estimator:} \quad \hat{\theta} = (X^\top X)^\dagger X^\top y = (X^\top X)^\dagger X^\top (X\theta^* + \epsilon),$$

$$\begin{aligned}\text{Excess risk:} \quad R(\hat{\theta}) &= \left(\hat{\theta} - \theta^*\right)^\top \Sigma \left(\hat{\theta} - \theta^*\right) \\ &\approx \theta^{*\top} \left(I - \hat{\Sigma}\hat{\Sigma}^\dagger\right) \left(\Sigma - \hat{\Sigma}\right) \left(I - \hat{\Sigma}^\dagger \hat{\Sigma}\right) \theta^* \\ &\quad + \sigma^2 \operatorname{tr}\left(\left(X^\top X\right)^\dagger \Sigma\right).\end{aligned}$$

Benign Overfitting: Two Examples

1. Low dimension

Suppose $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = I_k$ and $k \ll n$.

Then $X^\top X = n\hat{\Sigma} \approx n\Sigma$, and

$$R(\hat{\theta}) \approx \theta^{*\top} \left(I - \hat{\Sigma}\hat{\Sigma}^\dagger\right) \left(\Sigma - \hat{\Sigma}\right) \left(I - \hat{\Sigma}^\dagger\hat{\Sigma}\right) \theta^* + \sigma^2 \mathrm{tr}\left(\left(X^\top X\right)^\dagger \Sigma\right),$$

Benign Overfitting: Two Examples

1. Low dimension

Suppose $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = I_k$ and $k \ll n$.

Then $X^\top X = n\hat{\Sigma} \approx n\Sigma$, and

$$R(\hat{\theta}) \approx \theta^{*\top} \left(I - \hat{\Sigma}\hat{\Sigma}^\dagger\right) \left(\Sigma - \hat{\Sigma}\right) \left(I - \hat{\Sigma}^\dagger\hat{\Sigma}\right) \theta^* + \sigma^2 \mathrm{tr}\left(\left(X^\top X\right)^\dagger \Sigma\right),$$

$$\sigma^2 \mathrm{tr}\left(\left(X^\top X\right)^\dagger \Sigma\right) \approx \sigma^2 \mathrm{tr}\left((n\Sigma)^{-1}\Sigma\right) = \frac{k}{n}\sigma^2.$$

Benign Overfitting: Two Examples

2. High dimension, isotropic

Suppose $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = I_p$ and $p \gg n$.
Then $\hat{\Sigma}^\dagger \hat{\Sigma}$ is the projection on the span of the data in $\mathbb{R}^p$. This is an n-dimensional subspace that's almost orthogonal to θ^*, so

$$R(\hat{\theta}) \approx \left\| \left(I - \hat{\Sigma}^\dagger \hat{\Sigma} \right) \theta^* \right\|^2 + \sigma^2 \mathrm{tr}\left(\left(X X^\top \right)^{-1} \right)$$
$$\approx \left(1 - \frac{n}{p} \right) \|\theta^*\|^2 + \frac{n}{p} \sigma^2.$$

i.e., $\hat{\theta}$ is a low variance estimate of 0.

Benign Overfitting: A Characterization

Theorem

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

Here, $\text{bias}(\theta^*, \Sigma, n) = \|\theta^*_{k+1:\infty}\|^2_{\Sigma_{k+1:\infty}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{\sum_{i>k}\lambda_i}{n}\right)^2$.

If $\lambda_1 = \cdots = \lambda_k = 1$ and $\lambda_{k+1} = \cdots = \lambda_p = \epsilon$ with $k \ll n \ll p \ll n/\epsilon$, then $k^* = k$ and $r_k(\Sigma) = R_k(\Sigma) = p - k$.

Low-dimension example: the heaviest k-dimensional subspace.

High-dimension example: the $p - k$-dimensional tail.

Benign Overfitting: What kinds of eigenvalues?

Theorem

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$,

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

2. With some independence properties,
$$\mathbb{E}R(\hat{\theta}) \geq \frac{1}{c}\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \min\left\{\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1\right\}\right).$$

What kinds of eigenvalues?

We say $\{\Sigma_n\}$ is *asymptotically benign* if

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

where $k_n^* = \min\{k \geq 0 : r_k(\Sigma_n) \geq bn\}$.

What kinds of eigenvalues?

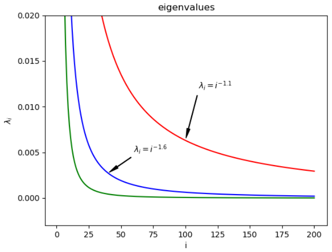
We say $\{\Sigma_n\}$ is *asymptotically benign* if

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

where $k_n^* = \min\{k \geq 0 : r_k(\Sigma_n) \geq bn\}$.

Example

If $\lambda_i = i^{-\alpha}\ln^{-\beta}(i+1)$,
Σ is benign iff $\alpha = 1$ and $\beta > 1$.



What kinds of eigenvalues?

We say $\{\Sigma_n\}$ is *asymptotically benign* if

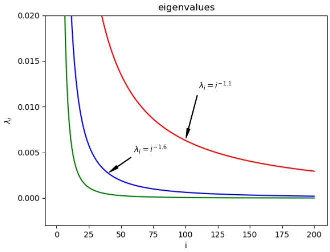
$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

where $k_n^* = \min\{k \geq 0 : r_k(\Sigma_n) \geq bn\}$.

Example

If $\lambda_i = i^{-\alpha}\ln^{-\beta}(i+1)$,
Σ is benign iff $\alpha = 1$ and $\beta > 1$.

The $\sum_i \lambda_i$ must almost diverge!!?!



What kinds of eigenvalues?

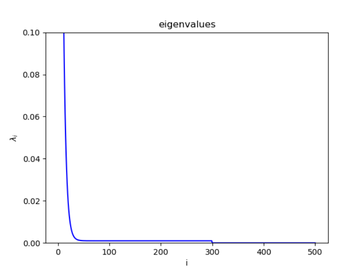
Example: *Finite dimension*, fast λ_i decay, plus isotropic noise

If

$$\lambda_{k,n} = \begin{cases} e^{-k} + \epsilon_n & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then Σ_n is benign iff

- $p_n = \omega(n)$,
- $\epsilon_n p_n = o(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$.

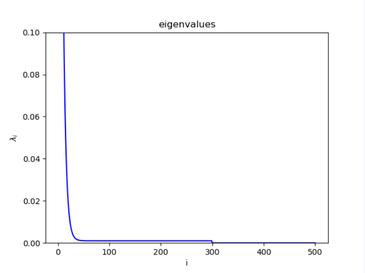


What kinds of eigenvalues?

Example: *Finite dimension*, fast λ_i decay, plus isotropic noise

If

$$\lambda_{k,n} = \begin{cases} e^{-k} + \epsilon_n & \text{if } k \leq p_n, \\ 0 & \text{otherwise,} \end{cases}$$



then Σ_n is benign iff

- $p_n = \omega(n)$,
- $\epsilon_n p_n = o(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$.

Furthermore, for $p_n = \Omega(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$,

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

What kinds of eigenvalues?

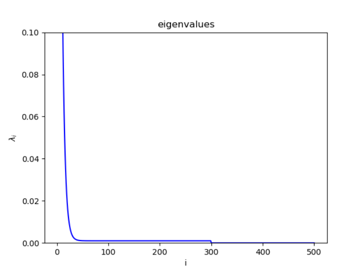
Example: *Finite dimension*, fast λ_i decay, plus isotropic noise

If

$$\lambda_{k,n} = \begin{cases} e^{-k} + \epsilon_n & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then Σ_n is benign iff

- $p_n = \omega(n)$,
- $\epsilon_n p_n = o(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$.



Furthermore, for $p_n = \Omega(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$,

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

Generic phenomenon:

quickly converging λ_i plus noise in all directions, $p_n \gg n$.

What kinds of eigenvalues?

Example: *Finite dimension*, slow eigenvalue decay

If

$$\lambda_{k,n} = \begin{cases} k^{-\alpha} & \text{if } k \leq p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then Σ_n is benign iff either

- $0 < \alpha < 1$, $p_n = \omega(n)$ and $p_n = o\left(n^{1/(1-\alpha)}\right)$, or
- $\alpha = 1$, $p_n = e^{\omega(\sqrt{n})}$ and $p_n = e^{o(n)}$.

What kinds of eigenvalues?

Example: *Finite dimension*, slow eigenvalue decay

If

$$\lambda_{k,n} = \begin{cases} k^{-\alpha} & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then Σ_n is benign iff either

- $0 < \alpha < 1$, $p_n = \omega(n)$ and $p_n = o\left(n^{1/(1-\alpha)}\right)$, or
- $\alpha = 1$, $p_n = e^{\omega(\sqrt{n})}$ and $p_n = e^{o(n)}$.

Universal phenomenon:

slowly converging λ_i, truncated at $p_n \gg n$.

Outline

- Linear regression
- Characterizing benign overfitting
- *Ridge regression*
- Beyond linear settings

Ridge Regression

Minimum norm ridge regression

$$\begin{aligned}\hat{\theta}_\lambda = \arg\min \quad & \|\theta\| \\ \text{s.t.} \quad & \theta \in \arg\min_\beta \left\{\|X\beta - y\|^2 + \lambda\|\beta\|_2^2\right\} \\ = X^\top \left(XX^\top + \lambda I\right)^{-1} y.\end{aligned}$$

Ridge Regression

Minimum norm ridge regression

$$\begin{aligned}\hat{\theta}_\lambda = \arg\min \quad & \|\theta\| \\ \text{s.t.} \quad & \theta \in \arg\min_\beta \left\{ \|X\beta - y\|^2 + \lambda \|\beta\|_2^2 \right\} \\ = X^\top \left(XX^\top + \lambda I \right)^{-1} y. &\end{aligned}$$

- Covers the range of solutions, from overfitting to regularized.

Ridge Regression

Minimum norm ridge regression

$$
\begin{aligned}
\hat{\theta}_\lambda = \arg\min \quad & \|\theta\| \\
\text{s.t.} \quad & \theta \in \arg\min_\beta \left\{ \|X\beta - y\|^2 + \lambda \|\beta\|_2^2 \right\} \\
= X^\top \left(XX^\top + \lambda I \right)^{-1} & y.
\end{aligned}
$$

- Covers the range of solutions, from overfitting to regularized.
- Tight bounds on bias and variance for $\lambda \in \mathbb{R}$.

Ridge Regression

Minimum norm ridge regression

$$\begin{aligned}\hat{\theta}_\lambda = \arg\min \quad & \|\theta\| \\ \text{s.t.} \quad & \theta \in \arg\min_\beta \left\{ \|X\beta - y\|^2 + \lambda \|\beta\|_2^2 \right\} \\ = X^\top \left(XX^\top + \lambda I \right)^{-1} y. &\end{aligned}$$

- Covers the range of solutions, from overfitting to regularized.
- Tight bounds on bias and variance for $\lambda \in \mathbb{R}$.
- Effective ranks, r_k and R_k, replaced by

$$r_k^\lambda(\Sigma) = \frac{\lambda + \sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k^\lambda(\Sigma) = \frac{\left(\lambda + \sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

- In some cases ($r_{k^*}(\Sigma) \gg n$), the optimal λ is *negative*: this decreases bias without significantly affecting variance.

Ridge Regression

Theorem
(Tsigler and B., 2020)

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\left\{k \geq 0 : r_k^\lambda(\Sigma) \geq bn\right\}$, the ridge regression estimate $\hat{\theta}_\lambda$ satisfies

1. With high probability,
$$R(\hat{\theta}_\lambda) \leq c\left(\mathrm{bias}(\theta^*, \Sigma, n, \lambda) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}^\lambda(\Sigma)}\right)\right),$$
2. If $X = \Sigma^{1/2}Z$ where Z has independent components and the components of θ^* are subject to random sign flips,
$$\mathbb{E}R(\hat{\theta}_\lambda) \geq \frac{1}{c}\left(\mathrm{bias}(\theta^*, \Sigma, n, \lambda) + \sigma^2 \min\left\{\frac{k^*}{n} + \frac{n}{R_{k^*}^\lambda(\Sigma)}, 1\right\}\right).$$

Here, $\mathrm{bias}(\theta^*, \Sigma, n, \lambda) = \|\theta^*_{k+1:\infty}\|^2_{\Sigma_{k+1:\infty}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{\lambda + \sum_{i>k}\lambda_i}{n}\right)^2$.

Outline

- Linear regression
- Characterizing benign overfitting
- Ridge regression
- *Beyond linear settings*

Benign Overfitting

- Far from the regime of a tradeoff between fit to training data and complexity.

Benign Overfitting

- Far from the regime of a tradeoff between fit to training data and complexity.
- In linear regression, a long, flat tail of the covariance eigenvalues is necessary and sufficient for the minimum norm interpolant to predict well: The noise is hidden in many unimportant directions.

Benign Overfitting

- Far from the regime of a tradeoff between fit to training data and complexity.
- In linear regression, a long, flat tail of the covariance eigenvalues is necessary and sufficient for the minimum norm interpolant to predict well: The noise is hidden in many unimportant directions.
  - Relies on many (roughly equally) unimportant parameter directions

Benign Overfitting

- Far from the regime of a tradeoff between fit to training data and complexity.
- In linear regression, a long, flat tail of the covariance eigenvalues is necessary and sufficient for the minimum norm interpolant to predict well: The noise is hidden in many unimportant directions.
  - Relies on many (roughly equally) unimportant parameter directions
  - Finite dimensional data is important:
    infinite dimension requires specific eigenvalue decay;
    it is a generic phenomenon for truncated slow decay.

Benign Overfitting

- Far from the regime of a tradeoff between fit to training data and complexity.
- In linear regression, a long, flat tail of the covariance eigenvalues is necessary and sufficient for the minimum norm interpolant to predict well: The noise is hidden in many unimportant directions.
  - Relies on many (roughly equally) unimportant parameter directions
  - Finite dimensional data is important:
    infinite dimension requires specific eigenvalue decay;
    it is a generic phenomenon for truncated slow decay.
- From interpolation to ridge regression

Next steps

Benign Overfitting

Next steps

- Linear regression: beyond minimum Euclidean norm

(Koehler, Zhou, Sutherland, Srebro, 2021)

Next steps

- Linear regression: beyond minimum Euclidean norm

  (Koehler, Zhou, Sutherland, Srebro, 2021)

- Linear neural networks:
  neural tangent kernels, random feature models

  (Liang, Rakhlin, Zhai, 2020)

  (Mei, Misiakiewicz, Montanari, 2021)

Benign overfitting in deep networks?

Neural networks versus linear prediction

For wide enough randomly initialized neural networks, gradient descent dynamics quickly converge to a *min-norm interpolating solution* in a certain finite-dimensional reproducing kernel Hilbert space.

Benign overfitting in deep networks?

Neural networks versus linear prediction

For wide enough randomly initialized neural networks, gradient descent dynamics quickly converge to a *min-norm interpolating solution* in a certain finite-dimensional reproducing kernel Hilbert space.

For example, for

$$f(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma\left(\langle w_i, x \rangle\right),$$

the corresponding (random) kernel is

$$K^m(x, x_j) := \frac{1}{m} \sum_{i=1}^{m} a_i^2 \sigma'\left(\langle w_i, x \rangle\right) \sigma'\left(\langle w_i, x_j \rangle\right) \langle x, x_j \rangle.$$

(Xie, Liang, Song, '16), (Jacot, Gabriel, Hongler '18), (Li and Liang, 2018), (Du, Poczós, Zhai, Singh, 2018), (Du, Lee, Li, Wang, Zhai, 2018), (Arora, Du, Hu, Li, Wang, 2019).

Benign Overfitting

Next steps

- Linear regression: beyond minimum Euclidean norm

  (Koehler, Zhou, Sutherland, Srebro, 2021)

- Linear neural networks: neural tangent kernels, random feature models (fix random w_i, estimate a_i)

  (Liang, Rakhlin, Zhai, 2020)

  (Mei, Misiakiewicz, Montanari, 2021)

Next steps

- Linear regression: beyond minimum Euclidean norm

  (Koehler, Zhou, Sutherland, Srebro, 2021)

- Linear neural networks: neural tangent kernels, random feature models (fix random w_i, estimate a_i)

  (Liang, Rakhlin, Zhai, 2020)

  (Mei, Misiakiewicz, Montanari, 2021)

- High-dimensional logistic regression.

  (Chatterji and Long, 2020)

Next steps

- Linear regression: beyond minimum Euclidean norm
  (Koehler, Zhou, Sutherland, Srebro, 2021)
- Linear neural networks: neural tangent kernels,
  random feature models (fix random w_i, estimate a_i)
  (Liang, Rakhlin, Zhai, 2020)
  (Mei, Misiakiewicz, Montanari, 2021)
- High-dimensional logistic regression.
  (Chatterji and Long, 2020)
- Invariance to transformations of losses. (Shamir, 2022)

Benign Overfitting

Next steps — beyond linear settings

- Linear regression: beyond minimum Euclidean norm

  (Koehler, Zhou, Sutherland, Srebro, 2021)

- Linear neural networks: neural tangent kernels, random feature models (fix random w_i, estimate a_i)

  (Liang, Rakhlin, Zhai, 2020)

  (Mei, Misiakiewicz, Montanari, 2021)

- High-dimensional logistic regression.

  (Chatterji and Long, 2020)

- Invariance to transformations of losses. (Shamir, 2022)

Benign Overfitting

Next steps — beyond linear settings

- Linear regression: beyond minimum Euclidean norm

  (Koehler, Zhou, Sutherland, Srebro, 2021)

- Linear neural networks: neural tangent kernels, random feature models (fix random w_i, estimate a_i)

  (Liang, Rakhlin, Zhai, 2020)

  (Mei, Misiakiewicz, Montanari, 2021)

- High-dimensional logistic regression.

  (Chatterji and Long, 2020)

- Invariance to transformations of losses. (Shamir, 2022)
- Classification with two-layer ReLU networks.

Benign overfitting with two-layer ReLU networks

Classification with a linear signal with label noise

Benign overfitting with two-layer ReLU networks

Classification with a linear signal with label noise

- Clean data:
  Class conditionals are μ-separated, 1-subgaussian, log-concave distributions in $\mathbb{R}^d$.

Benign overfitting with two-layer ReLU networks

Classification with a linear signal with label noise

- Clean data:
  Class conditionals are μ-separated, 1-subgaussian, log-concave distributions in $\mathbb{R}^d$.
- Plus noise:
  Labels are flipped with probability $\eta(x)$, and $\mathbb{E}\eta(x) \leq \eta$.

Benign overfitting with two-layer ReLU networks

Classification with a linear signal with label noise

- Clean data:
  Class conditionals are μ-separated, 1-subgaussian, log-concave distributions in $\mathbb{R}^d$.
- Plus noise:
  Labels are flipped with probability $\eta(x)$, and $\mathbb{E}\eta(x) \leq \eta$.
- For sample size n, probability of failure δ:

Benign overfitting with two-layer ReLU networks

Classification with a linear signal with label noise

- Clean data:
  Class conditionals are μ-separated, 1-subgaussian, log-concave distributions in $\mathbb{R}^d$.
- Plus noise:
  Labels are flipped with probability $\eta(x)$, and $\mathbb{E}\eta(x) \leq \eta$.
- For sample size n, probability of failure δ:
  - $d = \tilde{\Omega}(n\|\mu\|^2 + n^2 \log(1/\delta))$,

Benign overfitting with two-layer ReLU networks

Classification with a linear signal with label noise

- Clean data:
  Class conditionals are μ-separated, 1-subgaussian, log-concave distributions in $\mathbb{R}^d$.
- Plus noise:
  Labels are flipped with probability $\eta(x)$, and $\mathbb{E}\eta(x) \leq \eta$.
- For sample size n, probability of failure δ:
  - $d = \tilde{\Omega}(n\|\mu\|^2 + n^2 \log(1/\delta))$,
  - $\|\mu\|^2 = \Omega(\log(n/\delta))$.

Benign overfitting with two-layer ReLU networks

Classification with a linear signal with label noise

- Clean data:
  Class conditionals are μ-separated, 1-subgaussian, log-concave distributions in $\mathbb{R}^d$.
- Plus noise:
  Labels are flipped with probability $\eta(x)$, and $\mathbb{E}\eta(x) \leq \eta$.
- For sample size n, probability of failure δ:
  - $d = \tilde{\Omega}(n\|\mu\|^2 + n^2 \log(1/\delta))$,
  - $\|\mu\|^2 = \Omega(\log(n/\delta))$.
  - $n = \Omega(\log(1/\delta))$.

Benign overfitting with two-layer ReLU networks

Two-layer network, gradient descent

- Smooth leaky ReLU: $0 < \gamma \leq \phi'(z) \leq 1$ and $\|\phi''\|_\infty \leq H$.

Benign overfitting with two-layer ReLU networks

Two-layer network, gradient descent

- Smooth leaky ReLU: $0 < \gamma \leq \phi'(z) \leq 1$ and $\|\phi''\|_\infty \leq H$.
- m hidden units with adjustable parameters, fixed output parameters.

Benign overfitting with two-layer ReLU networks

Two-layer network, gradient descent

- Smooth leaky ReLU: $0 < \gamma \leq \phi'(z) \leq 1$ and $\|\phi''\|_\infty \leq H$.
- m hidden units with adjustable parameters, fixed output parameters.
- Low variance random initialization (no NTK).

Benign overfitting with two-layer ReLU networks

Two-layer network, gradient descent

- Smooth leaky ReLU: $0 < \gamma \leq \phi'(z) \leq 1$ and $\|\phi''\|_\infty \leq H$.
- m hidden units with adjustable parameters, fixed output parameters.
- Low variance random initialization (no NTK).
- Gradient descent on logistic loss with suitably small step-size.

Benign overfitting with two-layer ReLU networks

Two-layer network, gradient descent

- Smooth leaky ReLU: $0 < \gamma \leq \phi'(z) \leq 1$ and $\|\phi''\|_\infty \leq H$.
- m hidden units with adjustable parameters, fixed output parameters.
- Low variance random initialization (no NTK).
- Gradient descent on logistic loss with suitably small step-size.

Theorem

(Chatterji, Frei, B., 2022)

After poly$(\|\mu\|, n, d, m, 1/\epsilon)$ steps, gradient descent finds weights with

- Training loss below ϵ,

Benign overfitting with two-layer ReLU networks

Two-layer network, gradient descent

- Smooth leaky ReLU: $0 < \gamma \leq \phi'(z) \leq 1$ and $\|\phi''\|_\infty \leq H$.
- m hidden units with adjustable parameters, fixed output parameters.
- Low variance random initialization (no NTK).
- Gradient descent on logistic loss with suitably small step-size.

Theorem (Chatterji, Frei, B., 2022)

After poly$(\|\mu\|, n, d, m, 1/\epsilon)$ steps, gradient descent finds weights with

- Training loss below ϵ,
- Test error within $\eta + 2\exp\left(-\frac{cn\|\mu\|^4}{d}\right)$ of the optimal test error for the clean distribution.

Benign overfitting with two-layer ReLU networks

Remarks

- The parameters change dramatically during training, even at the first step. This is an essentially nonlinear method.

Benign overfitting with two-layer ReLU networks

Remarks

- The parameters change dramatically during training, even at the first step. This is an essentially nonlinear method.
- The analysis tracks a proxy loss, $g(yf(x)) = -\ell'(yf(x))$, and exploits a PL-inequality (gradient bounded below by loss). (Frei, Cao, Gu, 2019)

Benign overfitting with two-layer ReLU networks

Remarks

- The parameters change dramatically during training, even at the first step. This is an essentially nonlinear method.
- The analysis tracks a proxy loss, $g(yf(x)) = -\ell'(yf(x))$, and exploits a PL-inequality (gradient bounded below by loss). (Frei, Cao, Gu, 2019)
- Notice that the covariance of x has a single dominant direction, and this is the signal direction (difference of class-conditional means).

Open Questions

Open Questions

- Nonlinear signal models?

Benign Overfitting in Linear and Nonlinear Settings

Open Questions

- Nonlinear signal models?
- Deep networks?

Open Questions

- Nonlinear signal models?
- Deep networks? $\hat{f} = \hat{f}_0 + \Delta$?

Benign Overfitting in Linear and Nonlinear Settings

SIMONS FOUNDATION

Niladri Chatterji

Spencer Frei

Phil Long

Gábor Lugosi

Andrea Montanari

Alexander Rakhlin

Alexander Tsigler

- Benign overfitting in linear regression. B., Long, Lugosi, Tsigler. PNAS 117(48):30063–30070, 2020. arXiv:1906.11300
- Benign overfitting in ridge regression. Tsigler, B. arXiv:2009.14286
- Failures of model-dependent generalization bounds for least-norm interpolation. B., Long. JMLR 22(204):1–15, 2021. arXiv:2010.08479
- Deep learning: a statistical viewpoint. B., Montanari, Rakhlin. Acta Numerica 30:87-201, 2021. arXiv:2103.09177
- The interplay between implicit bias and benign overfitting in two-layer linear networks. Chatterji, Long, B. arXiv:2108.11489
- Benign overfitting without linearity. Chatterji, Frei, B. arXiv:2202.05928