Overfitting in Deep Networks

- 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.*

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

also (Belkin, Hsu, Ma, Mandal, 2018)
Statistical Wisdom and Overfitting

Classical approaches to prediction

Typically, we aim for a trade-off between

- Fit to the training data, e.g.,
  \[ \frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}(x_i) - y_i \right)^2, \]

- Complexity of a prediction rule, e.g.,
  - Number of parameters
  - Norm of parameter vector
  - Norm of function in a reproducing kernel Hilbert space,
  - Bandwidth of smoothing kernel,
  - ...

This is especially important for nonparametric methods, that is, those for which the number of parameters grows with the sample size.
Statistical Wisdom and Overfitting

**FIGURE 2.11.** *Test and training error as a function of model complexity.*

Figure 2.11 shows the typical behavior of the test and training error, as model complexity is varied. The training error tends to decrease whenever we increase the model complexity, that is, whenever we fit the data harder. However with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). In

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

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

A new statistical phenomenon: good prediction with zero 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.
Progress on Interpolating Prediction Rules

Simplicial interpolation

\[ \hat{f}(x) = \sum_{i=1}^{n} \frac{y_i K_h(x - x_i)}{\sum_{i=1}^{n} K_h(x - x_i)} \]

with \( K_h(x) = \begin{cases} 1[h\|x\| \leq 1] \\ \frac{1}{h\|x\|^\alpha} & \text{for } h\|x\| > 1 \end{cases} \).

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

Kernel smoothing with singular, compact kernels

Linear regression with \( d \asymp n \)

- Kernels defined in terms of the Euclidean inner product
  
  (Liang and Rakhlin, 2018)

- Linear regression with \( d, n \to \infty, d/n \to \gamma \) (Hastie, Montanari, Rosset, Tibshirani, 2019)
Outline

- Linear regression
- Characterizing benign overfitting
- Deep learning
- Adversarial examples
Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- $(x, y)$ Gaussian, mean zero.
- Define:

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

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

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

(or subgaussian, well-specified)
Definitions

Minimum norm estimator

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

\[
\begin{align*}
\min_{\theta \in \mathbb{H}} & \quad \|\theta\|^2 \\
\text{s.t.} & \quad \|X\theta - y\|^2 = \min_{\beta} \|X\beta - y\|^2.
\end{align*}
\]
Definitions

**Excess prediction error**

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

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

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

So \( \Sigma \) 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*
- Deep learning
- Adversarial examples
Interpolating Linear Regression

Overfitting regime

- We consider situations where \( \min_\beta \| X \beta - y \|^2 = 0 \).
- Hence, \( y_1 = x_1^T \hat{\theta}, \ldots, y_n = x_n^T \hat{\theta} \).
- When can the label noise be hidden in \( \hat{\theta} \) without hurting predictive accuracy?
Benign Overfitting: A Characterization

Theorem

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

1. With high probability,

$$R(\hat{\theta}) \leq c \left( \| \theta^* \|^2 \| \Sigma \| \sqrt{\frac{r_0(\Sigma)}{n}} + \sigma^2 \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right),$$

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

Also, $\frac{r_0(\Sigma)}{\ln(1 + r_0(\Sigma))} \geq \kappa n$ implies for some $\theta^*$, $\Pr(R(\hat{\theta}) \geq 1/c) \geq 1/4.$
Notions of Effective Rank

Definition (Effective Ranks)

Recall that \( \lambda_1 \geq \lambda_2 \geq \cdots \) are the eigenvalues of \( \Sigma \).

For \( k \geq 0 \), if \( \lambda_{k+1} > 0 \), define the effective ranks

\[
\begin{align*}
  r_k(\Sigma) &= \frac{\sum_{i > k} \lambda_i}{\lambda_{k+1}}, \\
  R_k(\Sigma) &= \frac{(\sum_{i > k} \lambda_i)^2}{\sum_{i > k} \lambda_i^2}.
\end{align*}
\]

Lemma

\[ 1 \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma). \]
Notions of Effective Rank

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

Examples

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

\[ r_0(\Sigma) = \text{rank}(\Sigma) s(\Sigma), \quad R_0(\Sigma) = \text{rank}(\Sigma) S(\Sigma), \]

with

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

Both \( s \) and \( S \) lie between \( 1/p \) (\( \lambda_2 \approx 0 \)) and \( 1 \) (\( \lambda_i \) 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 $k^* = \min \{k \geq 0 : r_k(\Sigma) \geq bn\}$,

1. With high probability,
   
   $$R(\hat{\theta}) \leq c \left( \|\theta^*\|^2 \|\Sigma\| \sqrt{\frac{r_0(\Sigma)}{n}} + \sigma^2 \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right),$$

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

Also, $\frac{r_0(\Sigma)}{\ln(1 + r_0(\Sigma))} \geq \kappa n$ implies for some $\theta^*$, $\Pr(R(\hat{\theta}) \geq 1/c) \geq 1/4$. 
Intuition

The mix of eigenvalues of $\Sigma$ 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 non-zero eigenvalues: large compared to $n$,
- Their sum: small compared to $n$,
- 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).
Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^T \theta^*$ and $y - x^T \theta^*$)
  1. $\hat{\theta}$ is a distorted version of $\theta^*$, 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?
Define the noise vector $\epsilon$ by $y = X\theta^* + \epsilon$.

**Estimator:**
$$\hat{\theta} = (X^TX)^\dagger X^Ty = (X^TX)^\dagger X^T(X\theta^* + \epsilon),$$

**Excess risk:**
$$R(\hat{\theta}) = (\hat{\theta} - \theta^*)^\top \Sigma (\hat{\theta} - \theta^*)$$
$$\approx \theta^*^\top \left( I - \hat{\Sigma}\hat{\Sigma}^\dagger \right) \left( \Sigma - \hat{\Sigma} \right) \left( I - \hat{\Sigma}^\dagger \hat{\Sigma} \right) \theta^*$$
$$+ \sigma^2 \text{tr} \left( \left( X^TX \right)^\dagger \Sigma \right).$$
Benign Overfitting: Proof Ideas

The excess risk

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

Write \( \Sigma = \sum_i \lambda_i v_i v_i^\top \).

Split the \( v_i \) into “heavy” directions (corresponding to \( \lambda_1 \geq \cdots \geq \lambda_k \)) and “light” ones (corresponding to \( \lambda_{k+1}, \ldots \)).

If \( r_k(\Sigma) \geq n \), the smallest positive \(((k + 1)\text{-th to } n\text{-th})\) eigenvalues of \( X^\top X \) are all concentrated (around \( \rho := \sum_{i > k} \lambda_i \)).

So \( XX^\top \succeq \rho I \).

c.f. ridge regression: \( \hat{\theta} = (X^\top X + \rho I)^{-1} X^\top y \)
Benign Overfitting: Proof Ideas

The minimum norm estimator

\[ \hat{\theta} = (X^T X)^\dagger X^T y = (X^T X)^\dagger X^T \epsilon + \cdots. \]

\[ R(\hat{\theta}) = (\hat{\theta} - \theta^*)^T \Sigma (\hat{\theta} - \theta^*). \]

Where does the energy from the noise go?

- A direction \( v_i \) sees noise energy (from \( X^T \epsilon \)) proportional to \( n \lambda_i \).
- This is scaled by no more than \( \rho^{-2} \).
- Its impact on the prediction error is scaled by another factor of \( \lambda_i \).
- Bound on prediction error: \( n \lambda_i^2 \rho^{-2} \).
- (We can do better in the “heavy” directions: \( \leq 1/n \).)

Total prediction error bound:

\[ \frac{k}{n} + n \sum_{i > k} \lambda_i^2 \rho^{-2} = \frac{k}{n} + \frac{n}{R_k(\Sigma)}. \]
Benign Overfitting: Proof Ideas

Lower bound

- The excess expected loss is at least as big as the same trace term, \( \text{tr} \left( (X^\top X)^\dagger \Sigma \right) \).

- When the eigenvalues of \( XX^\top \) are concentrated, the same split gives a lower bound within a constant factor of the upper bound.

- And otherwise, the excess expected loss is at least a constant.
Benign Overfitting: A Characterization

**Theorem**

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

1. With high probability,

\[
R(\hat{\theta}) \leq c \left( \|\theta^*\|^2 \|\Sigma\| \sqrt{\frac{r_0(\Sigma)}{n}} + \sigma^2 \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right),
\]

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

Also, \( \frac{r_0(\Sigma)}{\ln(1 + r_0(\Sigma))} \geq \kappa n \) implies for some \( \theta^* \), \( \Pr(R(\hat{\theta}) \geq 1/c) \geq 1/4 \).
What kinds of eigenvalues?

We say $\Sigma$ is *asymptotically benign* if

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

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

**Example**

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

The $\lambda_i$ must be almost diverging!!?!
What kinds of eigenvalues?

Example: Finite dimension, 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 \( \Sigma_n \) 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 \( \epsilon_n p_n = \omega(ne^{-n}) \),

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

Universal phenomenon: fast converging \( \lambda_i, p_n \gg n \), noise in all directions.
## Extensions

### Beyond Gaussian

1. **Linear model:**
   \[ \mathbb{E}[y|x] = x^\top \theta^*. \]

2. **Noise is subgaussian:**
   \[ \mathbb{E}\left[ \exp\left(\lambda (y - x^\top \theta^*)\right) \mid x \right] \leq \exp\left(\sigma_y^2 \lambda^2 / 2\right). \]

3. **Components of \( \Sigma^{-1/2} x \) are independent subgaussian:**
   \[ \mathbb{E}\left[ \exp(\lambda^\top \Sigma^{-1/2} x) \right] \leq \exp\left(\sigma_x^2 \|\lambda\|^2 / 2\right). \]

### Open questions

- Misspecified?
- Less independence?
  
  e.g., \( k(x, \cdot) \in \mathbb{H} \)?
  
  e.g., see (Rakhlin and Zhai, 2018)
Outline

- Linear regression
- Characterizing benign overfitting
- *Deep learning*
- Adversarial examples
Implications for deep learning

Neural networks versus linear prediction

For wide enough randomly initialized neural networks, gradient descent dynamics quickly converge to (approximately) a *min-norm interpolating solution* with respect to a certain kernel. 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).

(Generalization results in these papers: only no noise.)
Implications for deep learning

Neural networks versus linear prediction

- What can we say about realistic deep networks?
- The characterization of benign overfitting in linear regression requires $x = \Sigma^{1/2}z$ for a vector $z$ with independent components.
Outline

- Linear regression
- Characterizing benign overfitting
- Deep learning
- Adversarial examples
Label noise appears in \( \hat{\theta} \)

We can find a unit norm \( \Delta \)

such that perturbing an input \( x \) by \( \Delta \) changes the output enormously: even if \( \Delta^T \theta^* = 0 \),

\[
\| (x + \Delta)^T \hat{\theta} - x^T \hat{\theta} \|^2 \geq \frac{\sigma}{\sqrt{\lambda_{k^*+1}}} \geq \sqrt{\frac{n}{\text{tr}(\Sigma)}} \sigma.
\]

Benign overfitting leads to huge sensitivity.
Interpolating prediction: Future directions

- Between interpolation and regularization?
- Can we extend these results to interpolating deep networks?
  - What is the analog of the minimum norm linear prediction rule?
  - What role does the optimization method play?
  - Implications for regularization methods?
  - Implications for robustness?
Benign Overfitting in Linear Regression

- Interpolation: 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 parameters
- But it leads to huge sensitivity to (adversarial) perturbations.