Reusing test sets

ImageNet

New test accuracy (top-1, %)

Original test accuracy (top-1, %)

Image credit: (Recht et al., 2019)
Capacity-generalization tradeoff

• With probability $1 - \delta$, for all $h \in \mathcal{H}$,

$$L_D(h) \leq L_S(h) + \sqrt{\frac{8d \log(en/d) + \log(1/\delta)}{n}}$$

(1)

• As the capacity of $\mathcal{H}$ increases, $\min_{h \in \mathcal{H}} L_S(h)$ drops but the second term goes up.
Capacity-generalization tradeoff

![Graph showing the tradeoff between capacity and error](image-url)
Failure case 2
Large hypothesis classes

• Compare

\[ \mathcal{H}_1 = \text{the set of two-layer neural networks with 512 hidden units} \quad (2) \]

\[ \mathcal{H}_2 = \text{the set of all two-layer neural networks} \quad (3) \]

• \( \mathcal{H}_1 \) has a finite VC dimension, while the VC dimension of \( \mathcal{H}_2 \) is infinite!

• It is much easier (and tempting) to reduce the training error by increasing the hypothesis class.
Failure case 2
Failure case 2

• Compare

\[ w_2 = [0.206, -0.317] \]

\[ w_9 = [-30.69, 93.27, -2.65, -3.29, -0.124, 0.0248, 0.0017, 0.0000245, -0.00000423, -0.0000000857] \]

• The learned weights are either too large or too small for degree 9.

• What if instead we optimize

\[ \min_{w \in \mathcal{H}} L_S(w) + \frac{\lambda}{2} \| w \|^2 \]  

(4)
Failure case 2

\[ \lambda = 0.002 \]
Failure case 2

$\lambda = 0.02$
Failure case 2

$\lambda = 0.1$
Failure case 2

\[ \lambda = 0.2 \]
**L₂ Regularization**

• The term $\frac{\lambda}{2} \| w \|^2$ is called an $L₂$ regularizer.

• It is also known as weight decay.

• The expression

$$L_S(w) + \frac{\lambda}{2} \| w \|^2$$

is the Lagrangian of

$$\min_w L_S(w)$$

s.t. $\| w \| \leq B$
L₂ Regularization

• The L₂ regularizer has an effect of controlling the capacity of the hypothesis class.

• Compare

\[ \mathcal{H} = \{ x \mapsto w^\top \phi(x) : w \in \mathbb{R}^d \} \]

\[ \mathcal{H} = \{ x \mapsto w^\top \phi(x) : \|w\| \leq B \} \]
Shattering

• Given $n$ data points, there are $2^n$ ways of label them $\{+1, -1\}$.

• A set of $n$ points is shattered by $\mathcal{H}$ if there is an arrangement of $n$ points such that classifiers in $\mathcal{H}$ can produce all $2^n$ ways of labeling.

• VC dimension is the largest number of points that $\mathcal{H}$ can shatter.
Rademacher complexity

- Rademacher complexity (in binary classification) on a data set $S$ is defined as

$$\mathcal{R}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[ \max_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i) \right],$$  \hspace{1cm} (10)

where $\sigma \in \{+1, -1\}^n$ is uniformly chosen.

- In words, Rademacher complexity measures how well a class of classifiers correlate with random noise.

- Rademacher complexity (in binary classification) for $n$ points is defined as

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{S \sim \mathcal{D}^n}[\mathcal{R}_S(\mathcal{H})].$$ \hspace{1cm} (11)
Rademacher generalization bounds

• With probability $1 - \delta$, for all $h \in \mathcal{H}$

$$L_D(h) \leq L_S(h) + \mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$  \hspace{1cm} (12)

• With probability $1 - \delta$, for all $h \in \mathcal{H}$

$$L_D(h) \leq L_S(h) + \mathcal{R}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$ \hspace{1cm} (13)
Linear classifiers with bounded norm

• If $S = \{ x : \|x\| \leq r \}$ and $\mathcal{H} = \{ x \mapsto w^\top x : \|w\| \leq B \}$,

$$\mathcal{R}_S(\mathcal{H}) \leq \sqrt{\frac{r^2 B^2}{n}}$$  \hspace{1cm} (14)
• If we replace a data point in the data set, do you get a very different classifier?

• We say that the learning algorithm is stable is changing a data point does not change the classifier by much.

• If $S$ is the data set, then $S^{(i)}$ is the same data set with the $i$-th data point replaced with another random data point.
Stability

- Stable learning algorithms don't overfit.

\[
\mathbb{E}_{S \sim \mathcal{D}^n}[L_D(A(S)) - L_S(A(S))] = \mathbb{E}_{i \sim U(n)}[[\ell(A(S^{(i)})(x_i), y_i) - \ell(A(S)(x_i), y_i)]
\]

(15)

- Proof

\[
\mathbb{E}_S[L_D(A(S))] = \mathbb{E}_S[\mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(A(S)(x), y)]] = \mathbb{E}_S[\mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(A(S^{(i)})(x_i), y_i)]]
\]

(16)

\[
\mathbb{E}_S[L_S(A(S))] = \mathbb{E}_S[\mathbb{E}_{i \sim U(n)}[\ell(A(S)(x_i), y_i)]]
\]

(17)
Lipschitz loss

- If the loss is $\rho$-Lipschitz continuous,

\[ \ell(A(S^{(i)})(x_i), y_i) - \ell(A(S)(x_i), y_i) \leq \rho \|A(S^{(i)}) - A(S)\|. \]  

(18)

- We only need a bound on $\|A(S^{(i)}) - A(S)\|$.
Lipschitz and strongly convex

• If a function is $\lambda$-strongly convex,

$$\lambda \frac{1}{2} \|x - x^*\|^2 \leq f(x) - f(x^*)$$

(19)

where $x^*$ is the minimizer.

• If we can bound $f(x) - f(x^*)$, then we can have bound on $\|x - x^*\|$.

• We will then let $x = A(S^{(i)})$ and $x^* = A(S)$. 
\( L_2 \) regularizer

- \( \frac{\lambda}{2} \| w \|^2 \) is \( \lambda \)-strongly convex.

- \( L_S(w) + \frac{\lambda}{2} \| w \|^2 \) is \( \lambda \)-strongly convex if \( L_S(w) \) is convex.

- Adding a \( L_2 \) regularizer makes learning stable.

- If we choose \( A(S) = \arg\min_{w \in \mathcal{H}} L_S(w) + \frac{\lambda}{2} \| w \|^2 \), we get
  \[
  \|A(S^{(i)}) - A(S)\| \leq \frac{2\rho}{\lambda n}.
  \]
  (20)

- In the end, we have
  \[
  \mathbb{E}_{S \sim \mathcal{D}^n}[L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda n}
  \]
  (21)
Hypothesis class limited by the learning algorithm

• Compare

\[ \mathcal{H}_1 = \text{the set of all two-layer neural networks} \]  
\[ \mathcal{H}_2 = \text{the set of all two-layer neural networks with bounded norm } B \]  
\[ \mathcal{H}_3 = \text{the set of all two-layer neural networks searched with } t \text{ gradient updates} \]

\[ (22) \]
\[ (23) \]
\[ (24) \]

• \( \mathcal{H}_1 \) has infinite VC dimension, while the last two has bounded Rademacher complexity.