Theories of Generalization
Generalization

- So far in the course, we’ve seen many ways to estimate parameters:
  - Maximum-likelihood estimation in language modeling, probabilistic context-free grammars, etc.
  - Smoothing of maximum-likelihood estimates:
    \[
P(dog \mid \text{the, green}) = \lambda_1 P_{ML}(dog \mid \text{the, green}) + \lambda_2 P_{ML}(dog \mid \text{the}) + \lambda_3 P_{ML}(dog)
    \]
  - Perceptron, boosting, log-linear models for global linear models (feature selection, penalties for large parameter values)

- Today’s lecture: theory and intuition behind various estimates
Overview

- A statistical framework
- Properties of maximum-likelihood estimates
- A first result, through Chernoff/Hoeffding bounds
- Generalization bounds for finite hypothesis spaces
- Structural Risk Minimization
- Generalization bounds for boosting
- Generalization bounds based on margins
The Basic Framework

- We have an input domain $\mathcal{X}$ and output domain $\mathcal{Y}$.  
  e.g., $\mathcal{X}$ is a set of possible sentences, $\mathcal{Y}$ is set of possible parse trees.

- The task is to learn a function $F : \mathcal{X} \rightarrow \mathcal{Y}$

- We have a training set $(x_i, y_i)$ where for $i = 1 \ldots m$ with $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$
Loss functions

- Say we have a new test example \( x \), whose true label is \( y \)

- The function \( F(x) \) has the output \( \hat{y} \)

- A **loss function** is a function \( L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R} \)

- \( L(\hat{y}, y) \) is cost of proposing \( \hat{y} \) for an example \( x \) when \( y \) is the true label

- One example loss function: “0-1 loss”

\[
L(\hat{y}, y) = \begin{cases} 
0 & \text{If } y = \hat{y} \\
1 & \text{otherwise}
\end{cases}
\]

- Another example: percentage of correct dependency relations in a parse

From now on, we’ll assume \( L(\hat{y}, y) \) is 0-1 loss
We can now define the *empirical loss* of the function $F$, as

$$\hat{E}_r(F) = \frac{1}{m} \sum_{i} L(F(x_i), y_i)$$

- $\hat{E}_r(F)$ is the average loss on the training samples
- If $L$ is the 0-1 loss, then $\hat{E}_r(F)$ is the percentage of errors on the training sample
A Statistical Assumption

- We assume that both training and test samples are generated from some distribution $D(x, y)$

- $D(x, y)$ is fixed, but also unknown

- Crucial point: both training and test samples are drawn from the same distribution $D(x, y)$. This allows us to learn properties/functions from the training data which generalize to new, test examples
Expected Loss

- We now define the *expected loss* for a function $F$ as

$$Er(F) = \sum_{x, y} D(x, y)L(F(x), y)$$

- If $L$ is 0-1 loss, then $Er(F)$ is the *probability of an error* on a newly drawn test example.

- $Er(F)$ is the measure of how “good” a function is: our aim is to find an $F$ such that $Er(F)$ is as low as possible.
We have input/output domains $\mathcal{X}$ and $\mathcal{Y}$

We assume there is some distribution $D(x, y)$ generating examples

$(x_1, y_1) \ldots (x_m, y_m)$ is a training sample drawn from $D$
this is the only evidence we have about $D$

For any function $F : \mathcal{X} \rightarrow \mathcal{Y}$, we define

$$Er(F) = \sum_{x,y} D(x, y)L(F(x), y)$$

$$\hat{Er}(F) = \frac{1}{n} \sum_i L(F(x_i), y_i)$$

Our aim is to find a function $F$ with a low value for $Er(F)$
The Bayes Optimal Hypothesis

- The *bayes optimal* function is

\[ F_B(x) = \arg\max_y D(x, y) \]

- Intuition: for an input \( x \), simply return the most likely label

- It can be shown that \( F_B \) has the lowest possible value for \( Er(F) \)

- We can never construct this function: it is a function of \( D \), which is unknown. But it is a useful theoretical construct.
Overview

- A statistical framework
- Properties of maximum-likelihood estimates
- A first result, through Chernoff/Hoeffding bounds
- Generalization bounds for finite hypothesis spaces
- Structural Risk Minimization
- Generalization bounds for boosting
- Generalization bounds based on margins
Maximum-Likelihood Estimates

• In these approaches, we attempt to model the underlying distribution $D(x, y)$ or $D(y \mid x)$.

• We have parameters $\Theta$, and a model $P(x, y \mid \Theta)$ or $P(y \mid x, \Theta)$. e.g.,
  
  – In probabilistic context-free grammars, the parameters are rule probabilities, and $P(x, y \mid \Theta)$ is a product of rule probabilities
  – In global log-linear models, we take
    
    $$P(y \mid x, \Theta) = \frac{e^{\Phi(x, y) \cdot \Theta}}{\sum_{y' \in \text{GEN}(x)} e^{\Phi(x, y') \cdot \Theta}}$$

• Given training samples $(x_i, y_i)$, we maximize the log-likelihood

  $$L(\Theta) = \sum_i \log P(x_i, y_i \mid \Theta) \quad \text{or} \quad L(\Theta) = \sum_i \log P(y_i \mid x_i, \Theta)$$
Justification for Maximum-Likelihood Estimates

- **Assumption:** There is some parameter setting $\Theta^*$ such that $D(x, y) = P(x, y \mid \Theta^*)$ or $D(y \mid x) = P(y \mid x, \Theta^*)$

- Define the maximum-likelihood estimates:

$$\Theta_{ML} = \text{argmax}_\Theta L(\Theta)$$

- A usual property of maximum-likelihood estimates: as the training sample size goes to $\infty$, then $P(x, y \mid \Theta_{ML})$ converges to $D(x, y)$ (or, $P(y \mid x, \Theta_{ML})$ converges to $D(y \mid x)$)
Justification for Maximum-Likelihood Estimates

It follows that:

- Given that
  
  - **Assumption 1:** There is some parameter setting $\Theta^*$ such that $D(x, y) = P(x, y \mid \Theta^*)$ or $D(y \mid x) = P(y \mid x, \Theta^*)$
  
  - **Assumption 2:** we have enough training data for the maximum likelihood estimates to converge

- Then $P(x, y \mid \Theta_{ML})$ converges to $D(x, y)$, and

  $$\arg\max_y P(x, y \mid \Theta_{ML})$$

  converges to the Bayes-optimal function

  $$F_B = \arg\max_y D(x, y)$$

  (and similar properties follow for conditional models $P(y \mid x, \Theta)$)
Overview

- A statistical framework
- Properties of maximum-likelihood estimates
- A first result, through Chernoff/Hoeffding bounds
- Generalization bounds for finite hypothesis spaces
- Structural Risk Minimization
- Generalization bounds for boosting
- Generalization bounds based on margins
Estimating the Expected Loss

- We’d like to know what

\[ Er(F) = \sum_{x,y} D(x, y) L(F(x), y) \]

is for some function \( F \)

- A natural estimate of \( Er(F) \) is

\[ \hat{Er}(F) = \frac{1}{n} \sum_{i} L(F(x_i), y_i) \]

Question: how good an estimate is \( \hat{Er}(F) \)?
Chernoff/Hoeffding Bounds

- Say we have a coin with (unknown) probability of heads $= p$

- We derive an estimate of $p$ by the following procedure:
  - Toss the coin $m$ times
  - If we see heads $h$ times, our estimate is
    $$\hat{p} = \frac{h}{m}$$

- How good is this estimate? **Answer: for all $\epsilon, p, m$$$
  $$P[|p - \hat{p}| > \epsilon] \leq 2e^{-2m\epsilon^2}$$

  where the probability $P$ is taken over the generation of the training sample of $m$ coin tosses
• Additional bounds:

\[ P[p - \hat{p} > \varepsilon] \leq e^{-2m\varepsilon^2} \]

\[ P[\hat{p} - p > \varepsilon] \leq e^{-2m\varepsilon^2} \]

• An example: say we take \( m = 1000 \), and \( \varepsilon = 0.05 \). Then

\[
e^{-2m\varepsilon^2} = e^{-5} \approx \frac{1}{148}
\]

• Then if we repeatedly take samples of size 1000, for (roughly) 147/148 samples we will have \( (p - \hat{p}) \leq 0.05 \), for 147/148 samples we will have \( (\hat{p} - p) \leq 0.05 \), for 146/148 samples we will have \( |\hat{p} - p| \leq 0.05 \).
• Put another way: our estimation procedure has probability \(2e^{-5} \approx 2/148\) of returning a value of \(\hat{p}\) that is not within 0.05 of \(p\).

• We derive an estimate of \(p\) by the following procedure:
  
  – Toss the coin \(m\) times
  – If we see heads \(h\) times, our estimate is

\[
\hat{p} = \frac{h}{n}
\]
Estimating the Expected Loss

- We’d like to know value of $\mathbb{E}_r(F) = \sum_{x,y} D(x, y) L(F(x), y)$ for some function $F$

- A natural estimate of $\mathbb{E}_r(F)$ is $\hat{\mathbb{E}}_r(F) = \frac{1}{m} \sum_i L(F(x_i), y_i)$

- From Chernoff/Hoeffding bounds:

$$P[\hat{\mathbb{E}}_r(F) - \mathbb{E}_r(F) > \epsilon] \leq e^{-2m\epsilon^2}$$
Converting this Result to a Confidence Interval

- Introduce a variable $0 < \delta < 1$, which is
  \[
  \delta = e^{-2m\epsilon^2}
  \]
  next, solve for $\epsilon$, giving
  \[
  \epsilon = \sqrt{\log \frac{1}{\delta}} \quad \frac{2m}{2m}
  \]

**Theorem:** For a single hypothesis $F$, for any distribution $D(x, y)$, for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of the training sample,

\[
Er(F) \leq \hat{Er}(F) + \sqrt{\log \frac{1}{\delta}} \quad \frac{2m}{2m}
\]
An Example

- Say we measure $\hat{Er}(F) = 0.25$ from a sample of size 1000. We take $\delta = 0.01$. Then with probability at least $1 - \delta = 99\%$,

$$Er(F) \leq 0.25 + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} = 0.25 + 0.048 = 0.298$$
We have to be careful!

- It's tempting to choose (train) a function $F$ using the training sample, then use the previous bound to estimate its error.

- **But** in this case the function $F$ depends on the training sample, and the bound isn't valid.

- The bound is only valid if $Er(F)$ and $\hat{Er}(F)$ are calculated from a sample that is independent of $F$. 

Overview

- A statistical framework
- Properties of maximum-likelihood estimates
- A first result, through Chernoff/Hoeffding bounds
- Generalization bounds for finite hypothesis spaces
- Structural Risk Minimization
- Generalization bounds for boosting
- Generalization bounds based on margins
Hypothesis Spaces

- A hypothesis space $\mathcal{H}$ is a set of functions mapping $\mathcal{X}$ to $\mathcal{Y}$

- Learning from a training set $\equiv$ choosing a member of $\mathcal{H}$ based on the training set

- We’ll first consider finite hypothesis spaces
• Example of an infinite hypothesis space:
given $\text{GEN}, \Phi, W$, define

$$ F_W(x) = \operatorname{argmax}_{y \in \text{GEN}(x)} \Phi(x, y) \cdot W $$

• For every member of $W \in \mathbb{R}^d$, we have a different function

• An infinite hypothesis space is

$$ \mathcal{H} = \{ F_W : W \in \mathbb{R}^d \} $$

• Note that if we store each element of $W$ to $b$ bits of precision, then this becomes a finite hypothesis class of size $|\mathcal{H}| = 2^{db}$
Choosing Between the Members of $\mathcal{H}$

- An obvious choice: choose

$$F_{ERM} = \arg \min_{F \in \mathcal{H}} \hat{E}_r(F)$$

i.e., choose the member of $\mathcal{H}$ which has lowest training error

- This method is called “Empirical Risk Minimization” ([Vapnik, 1995])

- Next question: how good is $\hat{E}_r(F_{ERM})$ as an estimate of $E_r(F_{ERM})$?
Choosing Between the Members of \( \mathcal{H} \)

- Chernoff/Hoeffding bounds for a single hypothesis:
  \[
P[\hat{E}r(F) - Er(F) > \epsilon] \leq e^{-2m\epsilon^2}
  \]

- A new bound for finite hypothesis spaces:
  \[
P[\max_{F \in \mathcal{H}} (\hat{E}r(F) - Er(F)) > \epsilon] \leq |\mathcal{H}| e^{-2m\epsilon^2}
  \]

Intuition: if we have \( |\mathcal{H}| \) functions, there is \( |\mathcal{H}| \) times the probability that at least one of them will have a value for \( \hat{E}r(F) \) that deviates by at least \( \epsilon \) from \( Er(F) \)
A Proof

- The **union bound** says that for any events \(A_1, A_2, \ldots A_n\),

\[
P(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_n) \leq \sum_{i=1}^{n} P(A_i)
\]

- Say we have \(n\) functions in \(\mathcal{H}\), numbered \(F_1, F_2, \ldots F_n\)

- Note that \[
\left[ \max_{F \in \mathcal{H}} \left( \hat{E}r(F) - Er(F) \right) \right] > \epsilon \text{ if and only if }
\]

\[
\hat{E}r(F_1) - Er(F_1) > \epsilon \text{ or } \hat{E}r(F_2) - Er(F_2) > \epsilon \text{ or } \cdots \\
\hat{E}r(F_n) - Er(F_n) > \epsilon
\]

- By the union bound, the probability of at least one of these events happening is at most

\[
\sum_{i} P(\hat{E}r(F_i) - Er(F_i) > \epsilon) = |\mathcal{H}| e^{-2m\epsilon^2}
\]
**Theorem:** For any finite hypothesis class \( \mathcal{H} \), distribution \( D(x, y) \), and \( \delta > 0 \), with probability at least \( 1 - \delta \) over the choice of training sample, for all \( F \in \mathcal{H} \),

\[
Er(F) \leq \hat{Er}(F) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2m}}
\]
**Theorem:** For any finite hypothesis class $\mathcal{H}$, distribution $D(x, y)$, and $\delta > 0$, with probability at least $1 - \delta$ over the choice of training sample, for all $F \in \mathcal{H}$,

\[
Er(F) \leq \hat{Er}(F) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2m}}
\]

An example: say we have a hypothesis class $\mathcal{H}$ of size 1000. We have 10,000 training examples. For each function in $\mathcal{H}$, we measure the error on the training examples, $\hat{Er}(F)$. Say we choose $\delta = 0.01$. In this scenario we have

\[
\sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2m}} = 0.00239
\]

and for $1 - \delta = 99\%$ of all experiments with a sample of size 10,000, we will have

\[
Er(F) \leq \hat{Er}(F) + 0.00239
\]

for all members of $\mathcal{H}$
**Corollary:** For any finite hypothesis class $\mathcal{H}$, distribution $D(x, y)$, and $\delta > 0$, with probability at least $1 - \delta$ over the choice of training sample,

$$Er(F_{ERM}) \leq \hat{Er}(F_{ERM}) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2m}}$$

where

$$F_{ERM} = \arg \min_{F \in \mathcal{H}} \hat{Er}(F)$$
Overview

- A statistical framework
- Properties of maximum-likelihood estimates
- A first result, through Chernoff/Hoeffding bounds
- Generalization bounds for finite hypothesis spaces
- **Structural Risk Minimization**
- Generalization bounds for boosting
- Generalization bounds based on margins
Another Look at the Bound

**Corollary:** For any finite hypothesis class $\mathcal{H}$, distribution $D(x, y)$, and $\delta > 0$, with probability at least $1 - \delta$ over the choice of training sample,

$$
E_r(F_{ERM}) \leq \hat{E}_r(F_{ERM}) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2m}}
$$

where $F_{ERM} = \arg \min_{F \in \mathcal{H}} \hat{E}_r(F)$

- Crucial point: as $|\mathcal{H}|$ becomes larger, the number of training examples required for $F_{ERM}$ to be reliable increases.
Comparison to the Bayes Optimal Hypothesis

• Say $F^*$ is the best function in the hypothesis space

$$F^* = \arg \min_{F \in \mathcal{H}} Er(F)$$

• How close are we to the Bayes optimal hypothesis?

$$Er(F_{ERM}) - Er(F_B) = \underbrace{(Er(F_{ERM}) - Er(F^*))}_{\text{Variance term}} + \underbrace{(Er(F^*) - Er(F_B))}_{\text{Bias term}}$$

• Tension:
  
  – If $\mathcal{H}$ is too large, variance term is likely to be large
  – If $\mathcal{H}$ is too small, bias term is likely to be large
    (less chance of a “good” function being in our hypothesis space)
A Compromise: Structural Risk Minimization

- First step: pick a series of hypothesis classes of increasing size, \( \mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3 \ldots \mathcal{H}_s \), where \( |\mathcal{H}_1| < |\mathcal{H}_2| < \cdots < |\mathcal{H}_s| \).
  (This step must be done independently from the training sample)

**Theorem:** Assume a set of finite hypothesis classes \( \mathcal{H}_1, \mathcal{H}_2 \ldots \mathcal{H}_s \), and some distribution \( D(x, y) \). For all \( i = 1 \ldots s \), for all hypotheses \( F \in \mathcal{H}_i \), with probability at least \( 1 - \delta \) over the choice of training set of size \( m \) drawn from \( D \),

\[
Er(F) \leq \hat{Er}(F) + \sqrt{\frac{\log |\mathcal{H}_i| + \log \frac{1}{\delta} + \log s}{2m}}
\]
A Compromise: Structural Risk Minimization

- Pick the hypothesis that minimizes the bound, i.e.,

\[ F_{SRM} = \arg \min_F \left( \hat{E}r(F) + \sqrt{\frac{\log |\mathcal{H}_i| + \log \frac{1}{\delta} + \log s}{2m}} \right) \]

- The bound has two components

\[ \hat{E}r(F) \quad \text{Fit to training data} \quad \sqrt{\frac{\log |\mathcal{H}_i| + \log \frac{1}{\delta} + \log s}{2m}} \quad \text{Penalty for complexity} \]

- Some points:
  - The “complexity” of a function is related to the size of the hypothesis space of which it is a member
  - The complexity of \( F \) is also related to the reliability of \( \hat{E}r(F) \) as an estimate of \( E_r(F) \)
Overview

- A statistical framework
- Properties of maximum-likelihood estimates
- A first result, through Chernoff/Hoeffding bounds
- Generalization bounds for finite hypothesis spaces
- Structural Risk Minimization
- Generalization bounds for boosting
- Generalization bounds based on margins
Back to Global Linear Models

• Example of an infinite hypothesis space: given $\text{GEN}$, $\Phi$, $W$, define

$$F_W(x) = \arg\max_{y \in \text{GEN}(x)} \Phi(x, y) \cdot W$$

• For every member of $W \in \mathbb{R}^d$, we have a different function

• An infinite hypothesis space is

$$\mathcal{H} = \{ F_W : W \in \mathbb{R}^d \}$$
Back to Global Linear Models

- For now, we’ll “cheat” by considering $\mathcal{H}$ to be finite

- We do this by assuming that we store each element of $\mathbf{W}$ to $b$ bits of precision, then this becomes a finite hypothesis class of size $|\mathcal{H}| = 2^{db}$

- We can then apply our previous theorem:

**Theorem:** For a global linear model with finite hypothesis class $\mathcal{H}$ ($d$ parameters at $b$ bits of precision), distribution $D(x, y)$, and $\delta > 0$, with probability at least $1 - \delta$ over the choice of training sample, for all $F \in \mathcal{H}$,

$$\hat{E}r(F) \leq \hat{E}r(F) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2m}} = \hat{E}r(F) + \sqrt{\frac{db \log 2 + \log \frac{1}{\delta}}{2m}}$$
• The theorem implies that $d \times b \times \log 2$ must be small compared to $2m$ where $m$ is the sample size.

• In many of our experiments (e.g., parse reranking) we have many features, so $d$ is huge $\Rightarrow$ the bound implies that choosing $F_{ERM}$ is bad.

• Instead, we balanced fit to the training data against some penalty for “complexity”:
  
  – In boosting, minimize an upper bound on the training error while using a small number of features.
  – In log-linear models, maximize likelihood while keeping parameter values small.
A Bound for Feature-Selection Methods

• Before, our hypothesis class was

\[ \mathcal{H} = \{ F_w : W \in \mathbb{R}^d \} \]

which has \(2^{bd}\) members if we store parameters to \(b\) bits of precision

• Now, consider a restricted hypothesis space:

\[ \mathcal{H}_k = \{ F_w : W \in \mathbb{R}^d, \text{only } k \text{ parameters have non-zero values} \} \]

• What is the size of \(\mathcal{H}_k\) under precision \(b\) for the parameters?
A Bound for Feature-Selection Methods

• There are

\[ C_k^d = \binom{d}{k} = \frac{d!}{(d-k)!k!} \]

ways of choosing \( k \) features out of \( d \) features in total

• For each choice of \( k \) features, there are \( 2^{kb} \) ways of setting their parameters given \( b \) bits of precision

• It follows that

\[ |\mathcal{H}_k| = C_k^d \times 2^{kb} \]

and

\[ \log |\mathcal{H}_k| = \log C_k^d + kb \log 2 \]
A Bound for Feature-Selection Methods

- Also, note that

\[ C_k^d < d^k \]

\[ \Rightarrow \log C_k^d < k \log d \]

- Giving:

\[
\log |\mathcal{H}_k| = \log C_k^d + kb \log 2
\]

\[
< k \log d + kb \log 2
\]

\[ = k(\log d + b \log 2) \]

**Theorem:** For a global linear model with finite hypothesis class \( \mathcal{H}_k \) (\( d \) parameters at \( b \) bits of precision, with \( k \) non-zero parameters), distribution \( D(x, y) \), and \( \delta > 0 \), with probability at least \( 1 - \delta \) over the choice of training sample, for all \( F \in \mathcal{H} \),

\[
\hat{E}r(F) \leq \hat{E}r(F) + \sqrt{\frac{\log |\mathcal{H}_k| + \log \frac{1}{\delta}}{2m}} = \hat{E}r(F) + \sqrt{\frac{k(\log d + b \log 2) + \log \frac{1}{\delta}}{2m}}
\]
Theorem: For a global linear model with finite hypothesis class $\mathcal{H}_k$ ($d$ parameters at $b$ bits of precision, with $k$ non-zero parameters), distribution $D(x,y)$, and $\delta > 0$, with probability at least $1 - \delta$ over the choice of training sample, for all $F \in \mathcal{H}_k$, 

$$Er(F) \leq \hat{Er}(F) + \sqrt{\frac{k(\log d + b \log 2) + \log \frac{1}{\delta}}{2m}}$$

Fit to training data \hspace{3cm} \text{Complexity penalty}

- Complexity penalty is \textit{linear} in $k$, but \textit{logarithmic} in $d$: $\Rightarrow$ we can have a very large number of features ($d$ can be large) as long as only a small number are selected ($k$ is small)

- One justification for \textbf{Boosting} is that it minimizes this kind of bound
Overview

- A statistical framework
- Properties of maximum-likelihood estimates
- A first result, through Chernoff/Hoeffding bounds
- Generalization bounds for finite hypothesis spaces
- Structural Risk Minimization
- Generalization bounds for boosting
- Generalization bounds based on margins
We can think of the training data \((x_i, y_i)\), and \textbf{GEN}, providing a set of good/bad parse pairs

\[(x_i, y_i, z_{i,j}) \text{ for } i = 1 \ldots n, j = 1 \ldots n_i\]

The \textbf{Margin} on example \(z_{i,j}\) under parameters \(W\) is

\[m_{i,j}(W) = \Phi(x_i, y_i) \cdot W - \Phi(x_i, z_{i,j}) \cdot W\]
• A couple more definitions:

\[ m_i(W) = \min_j m_{i,j}(W) \]

\[ \hat{E}r(W, \gamma) = \frac{1}{m} \sum_i [m_i(W) < \gamma] \]

• So, \( m_i(W) \) is the minimum margin on the \( i \)'th example

• \( \hat{E}r(W, \gamma) \) is the percentage of examples whose minimum margin is less than \( \gamma \)
**Theorem:** Assume the hypothesis class $\mathcal{H}$ is as defined above, and that there is some distribution $D(x, y)$ generating examples. For all $F_w \in \mathcal{H}$, for all $\gamma > 0$, with probability at least $1 - \delta$ over the choice of training set of size $m$ drawn from $D$,

$$Er(F_w) \leq \hat{Er}(W, \gamma) + O\left(\sqrt{\frac{1}{m} \left( \frac{R^2 \|W\|^2}{\gamma^2} (\log m + \log N) + \log \frac{1}{\delta} \right)}\right)$$

where $R$ is a constant such that $\forall x \in \mathcal{X}, \forall y \in \text{GEN}(x), \forall z \in \text{GEN}(x), \|\Phi(x, y) - \Phi(x, z)\| \leq R$. The variable $N$ is the smallest positive integer such that $\forall x \in \mathcal{X}, |\text{GEN}(x)| - 1 \leq N$. 
Notes on the bound

\[
E_r(F_W) \leq \hat{E}_r(W, \gamma) + O \left( \sqrt{\frac{1}{m} \left( \frac{R^2 \|W\|^2}{\gamma^2} (\log m + \log N) + \log \frac{1}{\delta} \right)} \right)
\]

Fit to the data

Complexity Penalty

- The complexity penalty does not (directly) depend on the number of parameters in the model.

- The bound has two conflicting terms: keep the margin \( m_i(W) \) high on as many examples as possible, but keep \( \|W\|^2 \) low.

- The dependence on \( \log N \) is bad: perhaps the bound can be improved?
Notes on the bound

\[ Er(F_W) \leq \hat{Er}(W, \gamma) + O\left(\sqrt{\frac{1}{m} \left( \frac{R^2 ||W||^2}{\gamma^2} (\log m + \log N) + \log \frac{1}{\delta} \right)}\right) \]

Fit to the data

Complexity Penalty

- Note the relationship to global log-linear models with a gaussian prior:

\[ W_{MAP} = \arg\max_W \left( L(W) - C ||W||^2 \right) \]

where

\[ L(W) = \sum_i \log P(y_i | x_i, W) \]

\[ = - \sum_i \log \left( 1 + \sum_j e^{-m_{i,j}(W)} \right) \]
Summary

- One assumption: the same distribution $D(x, y)$ is generating training and test examples

- $E_r(F)$ is the error rate w.r.t. this distribution: we would like to find an $F$ which minimizes this. $\hat{E}_r(F)$ is the error rate on the training sample

- Started considering how good an estimate $\hat{E}_r(F)$ is of $E_r(F)$. This depends on the complexity of $F$.

- “Structural risk minimization” means we search for a function which has a low value for $\hat{E}_r(F)$, but is also not too “complex”

- Several measures of complexity have been considered:
  - Size of hypothesis class the function comes from
  - Number of non-zero parameter values
  - Size of the margins on training examples vs. $||W||^2$
Some Final Points

- Advantage of these bounds is that they make very few assumptions (for example, no assumptions about $D(x, y)$).

- Disadvantage is that they can be very pessimistic, or “loose”

- A great deal of current research on how to get “tighter” bounds

- The bounds were originally developed for classification problems: several important issues remain for NLP, e.g.,
  - Results for loss functions other than 0 – 1 loss
  - Dependence on $\log |\text{GEN}(x)|$ in margin bounds
  - How to optimize the bounds in practice