Covering Number Bounds and Statistical Learning Theory

Steve Kroon

Computer Science Division
Stellenbosch University

24 July 2009
1. Introduction

2. Model

3. Test sample vs training sample estimators

4. Weighted union bound

5. Covering number bounds

6. Statistical learning theory

7. More bounds
Examples: Classification

- Voice-activated dialing
- Good vs bad fruit
- Spam vs “real” e-mail (vs quarantine)

Kyocera Smartphone 7135

SybuData grape sorter component
More complex examples

- Automatic property valuation
- Ranked search results
- Weather prediction

Examples: Classification
- More complex examples
- Loss
- Risk and error
- Importance of error estimation

More bounds

Covering number bounds
- Statistical learning theory
- More bounds

Test sample vs training sample estimators
- Weighted union bound

Introduction
- Model
Loss

Loss measures the **cost of taking an action** in a given situation.

Examples:

- Voice-activated dialing: **zero-one loss**
- Spam vs “real” email: **asymmetric loss**
- Weather prediction: depends what the model is being used for!
Risk and error

Risk is simply average loss.

Error is another word for risk, used with the zero-one loss.

Average over future voice-dial attempts, future e-mails, future weather predictions, etc.

**Assumption**: future data points are independent and identically distributed (i.i.d.) from a distribution $D$. 
Importance of risk estimation

- No good having voice-activated dialing if it’s not reliable
- Will users’ emails be “thrown away” by the spam filter? How often?
- How sure can I be it won’t rain if a sunny day is predicted?
Accuracy of a point estimate

At the very least, we want a **point estimate** of the risk.

But **how accurate** is this estimate?

One solution: provide an **error bar**:

![Graph showing Antibodies and Joint Diameters over weeks of immunization.](image)
Better solution: provide a confidence interval.

Benefits:
- Need not be symmetric
- Associated probability
- One-sided intervals restrict attention to large risk
Core model components

- **inputs** $x$ from $\mathcal{X}$, e.g. representation of a spoken name, email, or collection of weather measurements.
- Associated **outputs** $y$ in $\mathcal{Y}$, e.g. correct name, correct label for e-mail, future weather system.
- $z$ is an input-output pair $(x, y)$ in $\mathcal{Z}$.
- These pairs come from an (unknown) distribution $D$.
- A **decision rule** $w$ is a rule for choosing an action based on $x$.
- The **decision class** $\mathcal{W}$ is a collection of decision rules.
Core model components

- **inputs** $x$ from $\mathcal{X}$, e.g. representation of a spoken name, email, or collection of weather measurements.

- **associated outputs** $y$ in $\mathcal{Y}$, e.g. correct name, correct label for e-mail, future weather system.

- $z$ is an input-output pair $(x, y)$ in $\mathcal{Z}$.

- These pairs come from an (unknown) distribution $D$.

- A **decision rule** $w$ is a rule for choosing an action based on $x$.

- The **decision class** $\mathcal{W}$ is a collection of decision rules.
Core model components

- **inputs** $x$ from $\mathcal{X}$, e.g. representation of a spoken name, email, or collection of weather measurements.
- **associated outputs** $y$ in $\mathcal{Y}$, e.g. correct name, correct label for e-mail, future weather system.
- $z$ is an input-output pair $(x, y)$ in $\mathcal{Z}$.
- These pairs come from an (unknown) distribution $D$.
- A **decision rule** $w$ is a rule for choosing an action based on $x$.
- The **decision class** $\mathcal{W}$ is a collection of decision rules.
Core model components

- **inputs** $x$ from $\mathcal{X}$, e.g. representation of a spoken name, email, or collection of weather measurements.
- associated **outputs** $y$ in $\mathcal{Y}$, e.g. correct name, correct label for e-mail, future weather system.
- $z$ is an input-output pair $(x, y)$ in $\mathcal{Z}$.
- These pairs come from an (unknown) distribution $D$.
- A **decision rule** $w$ is a rule for choosing an action based on $x$.
- The **decision class** $\mathcal{W}$ is a collection of decision rules.
Core model components

- **inputs** $x$ from $\mathcal{X}$, e.g. representation of a spoken name, email, or collection of weather measurements.
- **associated outputs** $y$ in $\mathcal{Y}$, e.g. correct name, correct label for e-mail, future weather system.
- $z$ is an input-output pair $(x, y)$ in $\mathcal{Z}$.
- These pairs come from an (unknown) distribution $D$.
- A **decision rule** $w$ is a rule for choosing an action based on $x$.
- The **decision class** $\mathcal{W}$ is a collection of decision rules.
### Example: e-mail classification

<table>
<thead>
<tr>
<th>Feature</th>
<th>“real” e-mail</th>
<th>spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi Steve</td>
<td>Good meeting you over the YOU about, so when you get</td>
<td>Subject: BUY CIALLIS GENERIC,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Store this medication at room</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and heat. What happens if I m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eat this medication? It is a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tab, not a prescription (i.e.,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>it's not reported to the FDA).</td>
</tr>
<tr>
<td>CAPMAX</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>ch!</td>
<td>0</td>
<td>0.436</td>
</tr>
<tr>
<td>ch</td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>free</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>guaranteed</td>
<td>0</td>
<td>2.54</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>steve</td>
<td>1.05</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ z = (19, 0.436, 0.8, 0, 2.54, \ldots, 0, 1) \]
Choosing a decision rule

We begin with a sample of \( l \) independent points \( \sim D \).

Split the sample into a training sample \( S \in \mathcal{Z}^m \) and a test sample \( T \in \mathcal{Z}^k \) \((m + k = l)\).

Pick a decision rule \( w_S \) using \( S \).
1 Introduction

2 Model

3 Test sample vs training sample estimators

4 Weighted union bound

5 Covering number bounds

6 Statistical learning theory

7 More bounds

Steve Kroon

Covering Number Bounds and Statistical Learning Theory
More on risk

Risk \((r)\) or error \((e)\) is **average loss**

- \(r_D(w)\): **(true) risk** — don’t know
- \(r_S(w)\): **training risk** — do know
- \(r_T(w)\): **test risk** — do know

<table>
<thead>
<tr>
<th></th>
<th>genuine</th>
<th>spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>genuine</td>
<td>544</td>
<td>7</td>
</tr>
<tr>
<td>spam</td>
<td>146</td>
<td>223</td>
</tr>
</tbody>
</table>

**Table:** Confusion matrix for test sample of 920 e-mails

Other distributions define other risks
For any decision rule $w$: 

- $r_T(w)$ is an **unbiased estimator** of $r_D(w)$ (i.e. average of $r_T(w)$ over test samples $T \sim D^k$ is true risk of $w$).

- Hoeffding’s interval (1963):

  $$\text{Conf}_{1-\delta}(r_D(w)) = \left[0, r_T(w) + \sqrt{\frac{\ln \frac{1}{\delta}}{2k}}\right]$$

- $100(1 - \delta)$% of test samples $T$ will generate an interval containing the true risk of $w$.

Also for $w_S$. 
Choosing $k$ and $m$

- Larger $m$ generally means choosing a better $w_S$; but
  - larger $m$ means smaller $k$; and
  - smaller $k$ implies wider confidence intervals.

Conclusion: if we can get good confidence intervals for $r_D(w_S)$ without a test set, we can improve performance.
Choosing $k$ and $m$

- Larger $m$ generally means choosing a better $w_S$; but
- larger $m$ means smaller $k$; and
- smaller $k$ implies wider confidence intervals.

Conclusion: if we can get good confidence intervals for $r_D(w_S)$ without a test set, we can improve performance.
Choosing $k$ and $m$

- Larger $m$ generally means choosing a better $w_S$; but
- larger $m$ means smaller $k$; and
- smaller $k$ implies wider confidence intervals.

Conclusion: if we can get good confidence intervals for $r_D(w_S)$ without a test set, we can improve performance.
Choosing \( k \) and \( m \)

- Larger \( m \) generally means choosing a better \( w_S \); but
- larger \( m \) means smaller \( k \); and
- smaller \( k \) implies wider confidence intervals.

**Conclusion:** if we can get good confidence intervals for \( r_D(w_S) \) without a test set, we can improve performance.
Training sample estimators

Aim: to estimate $r_D(w_S)$ without using $T$.

Focus on interval estimators.

We want $100(1 - \delta)\%$ of training samples $S$ to generate an interval containing the true risk of $w_S$.

$$\text{Conf}_{1-\delta}(r_D) = [0, U(S)] \iff \mathbb{P}_{S \sim D^m} \{r_D(w_S) \leq U(S)\} \geq 1 - \delta$$

Complication: dependence of $w_S$ on $S$. 

Introduction

Model

Test sample vs training sample estimators

Weighted union bound

Covering number bounds

Statistical learning theory

More bounds

Weighted union bound

Bonferroni’s inequality

Applying Bonferroni’s inequality

Using a test sample interval

Putting it together

Interval estimator for $w_S$

Eliminating waste
Any vs every

Test sample intervals: any decision rule \( w \).

Training sample intervals: every decision rule \( w \).

Combine probability statements with Bonferroni’s inequality.
Bonferroni’s inequality

\[ P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}. \]

Generally, \( P\{\bigcup_i A_i\} \leq \sum_i P\{A_i\} \).
Bonferroni’s inequality

\[ P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}. \]

Generally, \[ P\{\bigcup_i A_i\} \leq \sum_i P\{A_i\}. \]
Applying Bonferroni’s inequality

Finite decision class \( \mathcal{W} = \{ w_1, w_2, \cdots, w_N \} \)

\( A_i = \) “true risk of decision \( w_i \) lies outside interval”.

Define \( \delta_i = P\{A_i\} \).

\[
P \left\{ \bigcup_{i=1}^{N} A_i \right\} \leq \sum_{i=1}^{N} P\{A_i\} = \sum_{i=1}^{N} \delta_i
\]

Strategy: choose \( \delta_i \) to sum to \( \delta \) for a desired confidence level \( 100(1 - \delta)\% \).

Simplest: \( \delta_i = \frac{\delta}{N} \).
Using a test sample interval

Given $\delta_i$, we can use Hoeffding’s interval for $w_i$:

No dependence problem — each $i$ is considered independent of $S$.

$$\text{Conf}_{1-\delta_i}(r_D(w_i)) = \left[0, r_S(w_i) + \sqrt{\frac{\ln \frac{1}{\delta_i}}{2m}}\right] : \mathbb{P}\{A_i\} = \delta_i$$
Putting it together

Probability all true risks are in their intervals, i.e. no true risk is outside its corresponding interval.

\[
P \left\{ \forall i \in [1: N] : r_D(w_i) \leq r_S(w_i) + \sqrt{\frac{\ln \frac{1}{\delta_i}}{2k}} \right\} 
= P \left\{ \bigcap_{i=1}^{N} \bar{A}_i \right\} = 1 - P \left\{ \bigcup_{i=1}^{N} A_i \right\} 
\geq 1 - \sum_{i=1}^{N} \delta_i = 1 - \delta
\]
Putting it together

Probability all true risks are in their intervals, i.e. no true risk is outside its corresponding interval.

\[
\mathbb{P} \left\{ \forall i \in [1 : N] : r_D(w_i) \leq r_S(w_i) + \sqrt{\frac{\ln \frac{1}{\delta_i}}{2k}} \right\} \\
= \mathbb{P} \left\{ \bigcap_{i=1}^{N} \overline{A_i} \right\} = 1 - \mathbb{P} \left\{ \bigcup_{i=1}^{N} A_i \right\} \\
\geq 1 - \sum_{i=1}^{N} \delta_i = 1 - \delta
\]
Interval estimator for $w_S$

Suppose we use $S$ to pick $w_S$.

Bound above holds for all $w_i$ including $w_S$!

Write $\delta(S)$ for $\delta_i$ corresponding to $w_S$. Then:

$$\text{Conf}_{1-\delta}(r_D(w_S)) = \left[ 0, r_S(w_S) + \sqrt{\frac{\ln \frac{1}{\delta(S)}}{2m}} \right]$$

Bigger $\delta(S) \Rightarrow$ narrower interval.
Unused $\delta_i$ are wasted — lost confidence.

“Prior” $\alpha$ over $\mathcal{W}$: $\delta_i = \delta\alpha(i)$ — weighted union bound.

Bound also applies to countable $\mathcal{W}$.

Wasteful: highly correlated decision rules have big overlap of $A_i$!
Introduction

Model

Test sample vs training sample estimators

Weighted union bound

Covering number bounds

Statistical learning theory

More bounds

Introduction to covers

Symmetrization

Sighting the target

A dual sample bound

Union bound over a cover

Adjust for approximation

Final bound

Steve Kroon

Covering Number Bounds and Statistical Learning Theory
Question

How many places are there in the world?

- Infinitely many
- Distinguish by satellite
- Number of towns
- Number of states/provinces
- Number of countries
- Number of continents
Question

How many places are there in the world?

- Infinitely many
- Distinguish by satellite
- Number of towns
- Number of states/provinces
- Number of countries
- Number of continents
Overview

Similar decision rules have similar confidence intervals.

Use union bound on a representative set of decision rules, and adjust for approximation:

- Coarse approximation: small set with bigger adjustment.
- Fine approximation: large set with smaller adjustment.

Benefits:

- Handles infinite decision classes
- Less wasteful (in principle)
Covers

A cover is a representative set of decision rules.

For a given level of approximation $\gamma$, we want to use the smallest possible cover.

Smallest cover size is the covering number: $\mathcal{N}(\gamma, \mathcal{W}, d)$.

$d$ is a pseudometric: way of measuring closeness of decision rules.

Problem: closeness depends on the distribution $D$. 
Symmetrization lemma

Allows us to reconstruct intervals for $r_D(w_S)$ from intervals for risk on a ghost sample, $r_P(w_S)$

$P$ is another sample of size $u$.

Example: For any $0 < \beta \leq 1$,

$$\mathbb{P}_{S \sim D^m} \left\{ \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] > \epsilon \right\} \leq \beta^{-1} \mathbb{P}_{S \oplus P \sim D^{m+u}} \left\{ \sup_{w \in \mathcal{W}} [r_P(w) - r_S(w)] > \epsilon - \alpha(u, \beta) \right\}$$
Symmetrization by permutation

$$\mathbb{P}_{S \oplus P \sim D^{m+u}} \{ \mathcal{E}(S \oplus P) \} = \mathbb{E}_{Q \sim D^{m+u}} \mathbb{P}_{\tau \sim \text{Unif} S_{m+u}} \{ \mathcal{E}(\tau(Q)) \mid Q \}$$

- $\mathcal{E}(Q)$ denotes whether an event takes place for the $m + u$-sample $Q$;
- $\tau$ is uniformly distributed on the symmetric group $S_{m+u}$ of permutations of $Q$.

For any given $Q$, the right hand side probability does not depend on $D$!
Sighting the target

**Aim:** for a given $Q$ we want to bound

$$\mathbb{P}_{\tau \sim \text{Unif } S_{m+u}} \{ \mathcal{E}(\tau(Q)) | Q \}$$

with $\mathcal{E}(S \oplus P)$ set to:

$$\sup_{w \in \mathcal{W}} [r_P(w) - r_S(w)] > \epsilon.$$  

**Strategy:**

- bound probability of deviation for a **single** $w$;
- combine such bounds over a cover with a union bound;
- adjust for approximation by cover.

Bound for the first step is a **dual sample bound**.
A dual sample bound

For any $w$, write $\mathcal{E}_w(S \oplus P, \epsilon)$ for

$$r_P(w) - r_S(w) > \epsilon .$$

Then:

$$\mathbb{P}_{\tau \sim \text{Unif } S_{m+u}} \left\{ \mathcal{E}_w(\tau(Q), \epsilon) | Q \right\} < \exp \left( -2m \left( \frac{\epsilon u}{m + u} \right)^2 \right) $$
We choose a minimal cover \( \mathcal{W}^* \), of size \( \mathcal{N}(\gamma, \mathcal{W}, d) \).

Applying the union bound, we get

\[
P_{\tau \sim \text{Unif } S_{m+u}} \left\{ \exists w \in \mathcal{W}^*: \mathcal{E}_w(\tau(Q), \epsilon) \mid Q \right\}
< \mathcal{N}(\gamma, \mathcal{W}, d) \exp \left(-2m \left(\frac{\epsilon u}{m + u}\right)^2\right).
\]
Adjust for approximation I

Choice of $d$ in cover, determines the adjustment needed.

- Average difference in loss of $w$ over $Q$: $d_{1,Q}$
- Maximum difference in loss of $w$ over $Q$: $d_{\infty,Q}$

$$\mathcal{N}(\gamma, \mathcal{W}, d_{1,Q}) \leq \mathcal{N}(\gamma, \mathcal{W}, d_{\infty,Q})$$

but $d_{1,Q}$ uses a bigger adjustment.
Adjust for approximation II

\[ \mathbb{P}_{\tau \sim \text{Unif } S_{m+u}} \{ \exists w \in \mathcal{W} : \mathcal{E}_w(\tau(Q), \epsilon) \mid Q \} \]

\[ < \mathcal{N}(\gamma, \mathcal{W}, d_{1,Q}) \exp \left( -2m \left( \frac{(\epsilon - \frac{(2m+u)(m+u)}{um})\gamma}{m+u} \right) u \right)^2 \].

\[ \mathbb{P}_{\tau \sim \text{Unif } S_{m+u}} \{ \exists w \in \mathcal{W} : \mathcal{E}_w(\tau(Q), \epsilon) \mid Q \} \]

\[ < \mathcal{N}(\gamma, \mathcal{W}, d_{\infty,Q}) \exp \left( -2m \left( \frac{(\epsilon - 2\gamma)u}{m+u} \right)^2 \right) \].
Substituting this bound into earlier results, we obtain \((d_{\infty,Q})\): for any \(0 < \beta \leq 1\), \(u > 0\), \(\gamma > 0\), and \(\alpha(u, \beta) < \epsilon < 1\) satisfying 
\[
\epsilon - \alpha(u, \beta) - 2\gamma > 0,
\]

\[
\mathbb{P}_{S \sim D^m} \left\{ \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] > \epsilon \right\} \leq \frac{1}{\beta} \mathbb{E}_{Q \sim D^{m+u}} \mathcal{N}(\gamma, \mathcal{W}, d_{\infty,Q}) \exp \left( -2m \left( \frac{(\epsilon - \alpha(u, \beta) - 2\gamma)u}{m + u} \right)^2 \right). 
\]
Applying the bound

To apply the bound, one needs to:

- get a formula for/bound on mean covering numbers;
- select $\beta$, $\gamma$, $u$, $\alpha$ (defaults exist, but not optimal — future research);
- solve for $\epsilon$ given $\delta$ (straightforward).
Introduction

Model

Test sample vs training sample estimators

Weighted union bound

Covering number bounds

Statistical learning theory

More bounds
Mean covering numbers

- Key is relationship between growth of mean covering number and decay of exponential term as $m$ and $u$ increase.
- Generally, $D$ is unknown, and could be arbitrarily bad.
- Hence, we often bound mean covering number by

$$
N_\infty(\gamma, \mathcal{V}, m + u) \equiv \sup_{Q \in \mathcal{Z}^{m+u}} N(\gamma, \mathcal{V}, d_\infty, Q).
$$

- Restrict ourselves to zero-one loss in what follows: $N_\infty(\mathcal{V}, m + u)$.
- This can grow exponentially in $m + u$ for some $\mathcal{V}$. 
Suppose $\mathcal{N}_\infty(\mathcal{W}, m) = 2^m$.

- Common for small $m$

- Then there is an $m$-sample which can be fitted perfectly regardless of class labels.

- In a sense, class can model noise on samples up to size $m$.

- If it holds for all $m$, minimizing empirical risk is doomed!

- Still doomed for smaller exponential base.

- Corresponds to a nonfalsifiable scientific theory.
For many function classes, after a certain $m^*$, $\mathcal{N}_\infty(\mathcal{W}, m)$ grows polynomially.

VCSS lemma: if $\mathcal{N}_\infty(\mathcal{W}, m) < 2^m$, it’s polynomial bounded.

Polynomial degree = $m^*$ called the VC dimension.

For these classes, empirical risk minimization is consistent and has fast convergence.
Convergence rate

- Question: how fast is the convergence?
- Bound on VC dimension implies bound on covering numbers, but this is weak.
- Directly bounding covering numbers for function classes much better (but more difficult).
- Newer approach: concentration results for expected covering numbers.
Many classes in practice have infinite VC dimension. What then?

- Distribution-specific bounds.
- Non-uniform approaches (e.g. structural risk minimization).
- Algorithm-specific bounds (e.g. regularization, adaptive classification).
- Data-dependent bounds (e.g. sample compression bounds).
- Bounding expected covering number using concentration inequalities.
ULLNs

- Law of large numbers (LLN): average converges to mean.
- Glivenko-Cantelli (GC) theorem: empirical distribution converges (pointwise) to true distribution.
- GC theorem is a uniform LLN for a specific set of events.
- General ULLNs are for more general sets of events.
- Under some cases, we get uniform central limit theorems (e.g. Donsker’s theorem)
Zero-one loss functions: tighter dual sample bounds.

Realizable/realistic case: tighter dual sample bound.

Margin bounds: for thresholded classifiers, use covering numbers of unthresholded function class.

Random subsample lemma and bound.

Chaining and generic chaining.
Improvements II

- Stratification by complexity (non-uniform union bound).
- Stratification by data-dependent complexity: luckiness bounds.
- Algorithm-specific bounds: algorithmic luckiness bounds; compression schemes.
Other measures of deviation

Similar generalized bounds for other measures of deviation.

Examples:

- relative deviation: \( \frac{r_D(w) - r_S(w)}{\sqrt{r_D(w)}} \);
- Bartlett-Lugosi \( \nu \)-deviation: \( \frac{r_D(w) - r_S(w) - \nu}{\sqrt{r_D(w)}} \);
- Pollard-Haussler \( \nu \)-deviation: \( \frac{r_D(w) - r_S(w)}{\nu + r_D(w) + r_S(w)} \).
Margin bounds

- Predictions often more trustworthy when unthresholded output is far from decision boundary.
- Margin bounds uses covering numbers of unthresholded functions to construct bounds.
- Ingredient 1: proxy loss — loss when low-confidence predictions are treated as errors.
- Ingredient 2: triangle inequality.
Example

- Thresholded real value.
- Margin loss $L_{\gamma}$, and intermediary loss $L_{\frac{\gamma}{2}}$.
- Let thresholded $h$ be $g_h$. Then, for any $(x, y)$

  $$L_{\gamma}(h(x), y) \geq L_{\frac{\gamma}{2}}(h^*(x), y) \geq L(g_h(x), y).$$

- Bound over cover of unthresholded class (middle term) using proxy loss (first term) to get a bound on actual loss (third term).
Rademacher bounds

Uses concentration inequalities.

Maximal $r_D(w) - r_S(w)$ is highly concentrated around its mean.

Bounds the mean using Rademacher penalties — correlation of loss with noise!
Define $Y = \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)]$.

New approach — bound the average maximal deviation:

$$\mathbb{P}_{D^m} \{ Y > \mathbb{E}_{S \sim D^m} Y + \epsilon \} \leq \exp(-2\epsilon^2 m)$$

This follows from McDiarmid's inequality.
Symmetrization inequality

\[ E_{S \sim D^m} \ Y \leq 2 \ E_{S \sim D^m, \zeta \sim \text{Unif}\{-1,1\}^m} \mathcal{R}_S(\mathcal{W}) \ , \text{ where} \]

- \( \zeta \sim \text{Unif}\{-1,1\}^m \) are independent Rademacher variables;
- \( \mathcal{R}_S(\mathcal{W}) \) is the Rademacher penalty of \( \mathcal{W} \) for \( S \),

\[ \sup_{w \in \mathcal{W}} \left[ \frac{1}{m} \sum_{i=1}^{m} \zeta_i L(w(x_i), y_i) \right] \]

Rademacher penalty can be seen as a maximal covariance with noise over the class.
Again from McDiarmid’s inequality, we have:

\[
P_{S \sim D^m, \zeta \sim \text{Unif}\{-1,1\}} m \left\{ \mathbb{E}_{S \sim D^m, \zeta \sim \text{Unif}\{-1,1\}} R_S(\mathcal{W}) > R_S(\mathcal{W}) + \epsilon \right\} \leq \exp \left(-\frac{\epsilon^2 m}{2}\right).
\]

In principle, finding $R_S(\mathcal{W})$ is possible, but equivalent to minimizing empirical risk (NP-hard).
Resulting confidence interval

Putting these results together, we obtain (for any $0 < \delta_1 < \delta$):

$$\text{Conf}_{1-\delta}(r_D(w_S)) = \left[ 0, r_S(w_S) + 2 \left( \mathcal{R}_S(\mathcal{W}) + \sqrt{\frac{2 \ln \frac{1}{\delta-\delta_1}}{m}} + \sqrt{\frac{\ln \frac{1}{\delta_1}}{2m}} \right) \right]$$
Improvements/alternatives

- Single application of McDiarmid’s inequality;
- More refined concentration inequalities: functional Bennett’s inequality;
- Bounds in terms of mean Rademacher penalty: tighter, but more difficult to evaluate;
- **Local** Rademacher bounds: faster decay, but only effective for (almost) ERM
Specific to **averaging classifiers**: Bayesian approach, bagging.

“Prior” need not be correct!

Deviation controlled by:
- K-L divergence of posterior from “prior”; and
- Convex conjugate of K-L divergence from “prior”.
**Improvements**

PAC-Bayesian **margin bounds**: thresholded averages.

Exchangeable “priors”: generalized the covering number result presented earlier to averaging classifiers.

Algorithm- and data-dependent PAC-Bayesian bounds.

Much better than other bounds **when applicable**.

**Problem**: verifying specification of the “prior”.
Other bounds

Shell decomposition bounds: stratifying the decision class by true risk of decision rules, and estimating the layers.

Occam’s hammer (Blanchard and Fleuret, 2007):
- classifier sampled once from posterior distribution.
- confidence interval obtained from classical interval by modifying the required confidence level.

Many others for specific decision classes, algorithms, other scenarios: decision trees, boosting, set covering machine, semi-supervised learning, on-line learning, transductive learning.