Statistical Learning Theory

Jun Zhu
dcszj@mail.tsinghua.edu.cn
http://bigml.cs.tsinghua.edu.cn/~jun
State Key Lab of Intelligent Technology & Systems
Tsinghua University

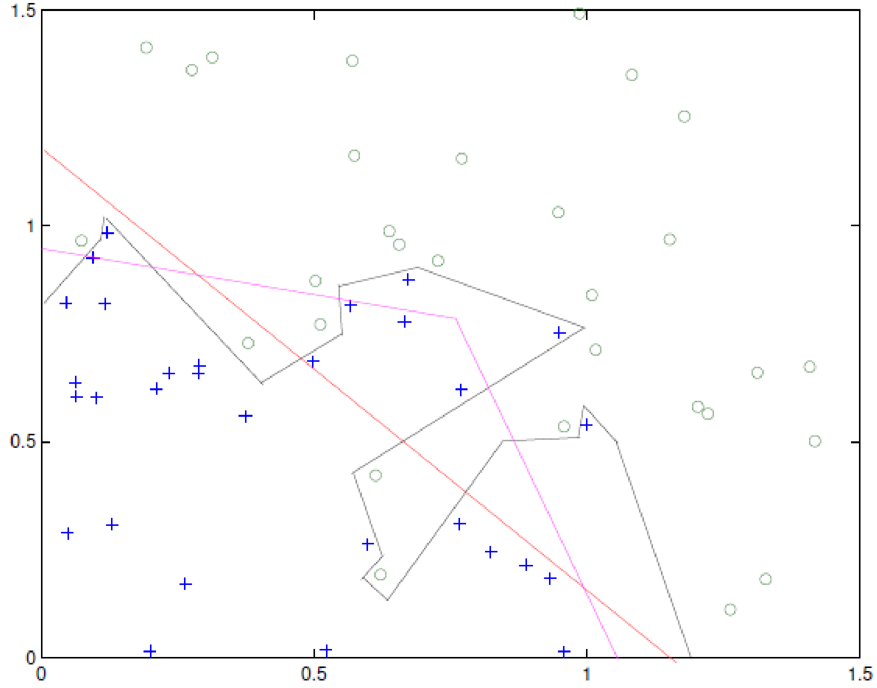
April 24, 2018

Supervised learning

- The basic framework:
  - Data consists of pairs (instance, label)
  - Label is +1 or -1
  - Algorithm construct a function (instance -> label)
  - Goal: make few mistakes on future unseen instances

Approximation/Interpolation

- It is always possible to build a function that fits exactly the data



- But is it reasonable?

Occam's Razor

- **Idea**: look for regularities in the observed phenomenon; these can be generalized from the observed past to the future
  - Choose the simplest consistent model

- How to measure simplicity?
  - Number of parameters
  - Description length
  - …

No Free Lunch

- **No free lunch theorem**:
  - If there is no assumption on how the past is related to the future, prediction is impossible
  - If there is no restriction on the possible phenomena, generalization is impossible

- We need to make assumptions
- Simplicity is not absolute
- Data will never replace knowledge
- Generalization = data + knowledge

Assumptions

- Two types of assumptions
  - Future observations related to past ones
    - Stationarity of the phenomenon
  - Constraints on the phenomenon
    - Notion of simplicity

Goals of SLT

How can we make predictions from the past?

What are the assumptions?

- Give a formal definition of learning, generalization, overfitting
- Characterize the performance of learning algorithms
- Design better algorithms

Probabilistic Model

- Relationship between past and future observations
  - Sampled independently from the same distribution

- **Independence**: each new observation yields maximum information
- **Identical distribution**: the observations give information about the underlying distribution

Probabilistic Model

- We consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$
  - For classification, we have $\mathcal{Y} = \{+1, -1\}$

- **Assumption**: the pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are distributed according to P (unknown)
- **Data**: we observe a sequence of n i.i.d. pairs (X_i, Y_i) sampled from P
- **Goal**: construct a function $g : \mathcal{X} \to \mathcal{Y}$ which predicts Y from X

Probabilistic Model

- Criterion to choose our function:
  - Low probability of error

$$P(g(X) \neq Y)$$

- **Risk**:

$$R(g) = P(g(X) \neq Y) = \mathbb{E}[1_{[g(X) \neq Y]}]$$

  - P is unknown so that we cannot directly measure the risk
  - Can only measure the agreement on the data

- **Empirical Risk**:

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} 1_{[g(X_i) \neq Y_i]}$$

Target Function

- P can be decomposed as $P_X \times P(Y|X)$
- Regression function:

$$\eta(x) = \mathbb{E}[Y|X = x] = 2\mathbb{P}[Y = 1|X = x] - 1$$

- Target function:

$$t(x) = \operatorname{sgn} \eta(x)$$

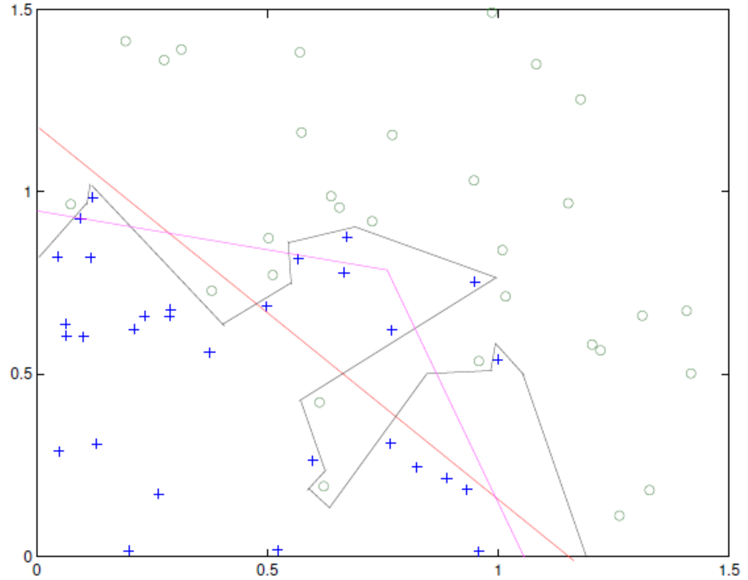
- In the deterministic case $Y = t(X)$ $(\mathbb{P}[Y = 1|X] \in \{0, 1\})$
- In general, the noise level

$$n(x) = \min(\mathbb{P}[Y = 1|X = x], 1 - \mathbb{P}[Y = 1|X = x])$$
$$= (1 - \eta(x))/2$$

Assumptions about P

- Need assumptions about P
  - If t(x) is totally chaotic, there is no possible generalization from finite data
- Assumptions can be
  - Preference (e.g., a prior distribution on possible functions)
  - Restriction (e.g., set of possible functions)
- Treating lack of knowledge
  - Bayesian approach: uniform distribution
  - Learning theory approach: worst-case analysis

Overfitting/Underfitting



- The data can mislead you
  - **Underfitting** model too small to fit the data
  - **Overfitting** artificially good agreement with the data
- No way to detect them from the data! Need extra validation data

Empirical Risk Minimization

- Choose a model $\mathcal{G}$ (set of functions)

- Minimize the empirical risk in the model

$$\min_{g \in \mathcal{G}} R_n(g)$$

- What if the Bayes classifier is not in the model?

Approximation/Estimation

- Bayes risk

$$R^* = \inf_g R(g)$$

  - Best risk a deterministic function can have (risk of the target function, or Bayes classifier)
- Decomposition: $R(g^*) = \inf_{g \in \mathcal{G}} R(g)$

$$R(g_n) - R^* = \underbrace{R(g^*) - R^*}_{\text{Approximation}} + \underbrace{R(g_n) - R(g^*)}_{\text{Estimation}}$$

  - Only the estimation error is random (i.e. depends on the data)
  - In statistics, this is known as bias-variance decomposition

Bias-Variance Decomposition

- Originally coined in regression with squared error loss

$$Y_i = f(X_i) + \epsilon,\ \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

  - Goal: find a function $\hat{f}(X)$ to approximate the truth $Y = f(X)$ from some training dataset
  - Expected square error for an unseen X:

$$\mathbb{E}[(Y - \hat{f}(X))^2] = \mathrm{Bias}[\hat{f}(X)]^2 + \mathrm{Var}[\hat{f}(X)] + \sigma^2$$

  - where

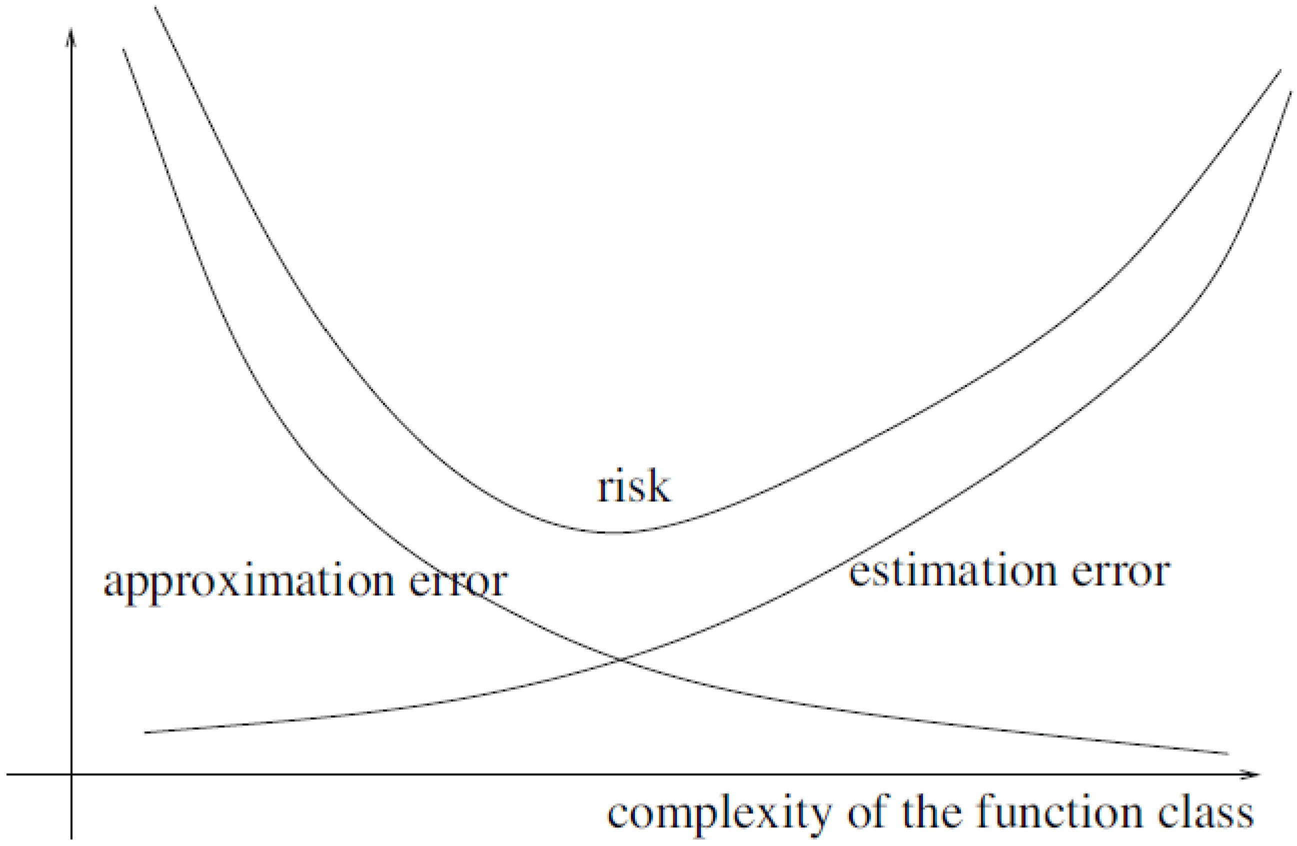
$$\mathrm{Bias}[\hat{f}(X)] = \mathbb{E}[\hat{f}(X)] - f(X),\ \ \mathrm{Var}[\hat{f}(X)] = \mathbb{E}[(\hat{f}(X) - \mathbb{E}[\hat{f}(X)])^2]$$

    - Note: expectation is taken over different choices of training set

- Now used in more general settings

Bias-Variance Decomposition

Bias-Variance Tradeoff



An example that ERM can fail

- Assume a data space $\mathcal{X} = [0, 1]$ with a uniform distribution
- Define the true label deterministically

$$Y = \begin{cases} -1 & \text{if } X < 0.5 \\ 1 & \text{if } X \geq 0.5. \end{cases}$$

- Given a set of training data, consider the classifier

$$g_n(X) = \begin{cases} Y_i & \text{if } X = X_i \text{ for some } i = 1, \ldots, n \\ 1 & \text{otherwise.} \end{cases}$$

- Then, we have $R_n(g) = 0$
- But $R(g) = 1/2$
- The classifier doesn't learn anything! Overfitting!

Structural risk Minimization

- Choose a collection of models $\{\mathcal{G}_d : d = 1, 2, \ldots\}$
- Minimize the empirical risk in each model
- Minimize the penalized empirical risk

$$\min_{d} \min_{g \in \mathcal{G}_d} R_n(g) + \mathrm{pen}(d, n)$$

  - pen(d,n) gives preference to models where estimation error is small
  - pen(d,n) measures the size or capacity of the model

Regularization

- Choose a large model $\mathcal{G}$ (possibly dense)
- Choose a regularizer $||g||$
- Minimize the regularized empirical risk

$$\min_{g \in \mathcal{G}} R_n(g) + \lambda \, ||g||^2$$

- Choose an optimal trade-off λ (regularization parameter)

- Most methods can be thought of as regularization methods, e.g., SVMs

Bounds

- A learning algorithm
  - Takes as input the data $(X_1, Y_1), \ldots, (X_n, Y_n)$
  - Produces a function g_n
- Can we estimate the risk of g_n ?

- Key points:
  - random quantity (depends on the data)
  - need probabilistic bounds

Bounds

- Error bounds

$$R(g_n) \leq R_n(g_n) + B$$

  - Estimation from the data

- Relative error bounds
  - Best in a class

$$R(g_n) \leq R(g^*) + B$$

  - Bayes risk

$$R(g_n) \leq R^* + B$$

- => theoretical guarantees

Basic Bounds

Probability Tools

- Basic facts:
  - Union:
    $$\mathbb{P}\left[A \text{ or } B\right] \leq \mathbb{P}\left[A\right] + \mathbb{P}\left[B\right]$$
  - Inclusion:
    $$\text{If } A \Rightarrow B, \text{ then } \mathbb{P}\left[A\right] \leq \mathbb{P}\left[B\right]$$
  - Inversion:
    $$\text{If } \mathbb{P}\left[X \geq t\right] \leq F(t) \text{ then with probability at least } 1-\delta,$$
    $$X \leq F^{-1}(\delta)$$

Probability Tools

- Basic Inequalities
  - Jensen:

$$\text{for } f \text{ convex, } f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

  - Markov:

$$\text{If } X \geq 0 \text{ then for all } t > 0,\ \mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

  - Chebyshev:

$$\text{for } t > 0,\ \mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathsf{Var}[X]}{t^2}$$

  - Chernoff:

$$\text{for all } t \in \mathbb{R},\ \mathbb{P}[X \geq t] \leq \inf_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(X-t)}\right]$$

Error Bounds

- Recall that we want to bound $R(g_n) = \mathbb{E}\left[1_{[g_n(X) \neq Y]}\right]$ where g_n has been constructed from $(X_1, Y_1), \ldots, (X_n, Y_n)$

  - Cannot be observed (*P* is unknown)
  - Random (depends on the data)

- We want to bound

$$\mathbb{P}\left[R(g_n) - R_n(g_n) > \varepsilon\right]$$

Loss class

- For convenience, let $Z = (X, Y)$. Given $\mathcal{G}$ define the loss class

$$\mathcal{F} = \{f : (x, y) \mapsto 1_{[g(x) \neq y]} : g \in \mathcal{G}\}$$

- Denote

$$Pf = \mathbb{E}\left[f(X, Y)\right] \qquad P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$$

- Quantity of interest:

$$Pf - P_n f$$

The Law of Large Numbers

$$R(g) - R_n(g) = \mathbb{E}\left[f(Z)\right] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)$$

- Difference between the expectation and the empirical average of r.v. $f(Z)$

- Law of large numbers:

$$\mathbb{P}\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right] = 0\right] = 1$$

  - Can we quantify it for a finite n?

Hoeffding's Inequality

- A quantitative version of law of large numbers
- Assumes bounded random variables

Theorem 1. *Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables. If $f(Z) \in [a, b]$. Then for all $\varepsilon > 0$, we have*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right]\right| > \varepsilon\right] \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) .$$

Hoeffding's Inequality

- We can rewrite it to better understand
- Let $$\delta = 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)$$
- Then

$$\mathbb{P}\left[|P_n f - Pf| > (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right] \leq \delta$$

or [Inversion] with probability at least $1-\delta$,

$$|P_n f - Pf| \leq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Hoeffding's Inequality

- Let's apply to $f(Z) = 1_{[g(X) \neq Y]}$
- For any g and any $\delta > 0$, with probability at least $1 - \delta$

$$R(g) \leq R_n(g) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}\,.$$

- Notice that one has to consider a fixed function g and the probability is respect to the sampling of data
- If the function depends on the data, this does not apply!

Limitations

- For each fixed function $f \in \mathcal{F}$, there is a set S of samples for which
$$Pf - P_n f \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \quad (\mathbb{P}[S] \geq 1-\delta)$$
- They may be different for different functions
- The function chosen by the algorithm depends on the sample

- For the observed sample, only some of the functions in $\mathcal{F}$ will satisfy this inequality!

Limitations

- What we need to bound is

$$P f_n - P_n f_n$$

  - where f_n is the function chosen by the algorithm on the data

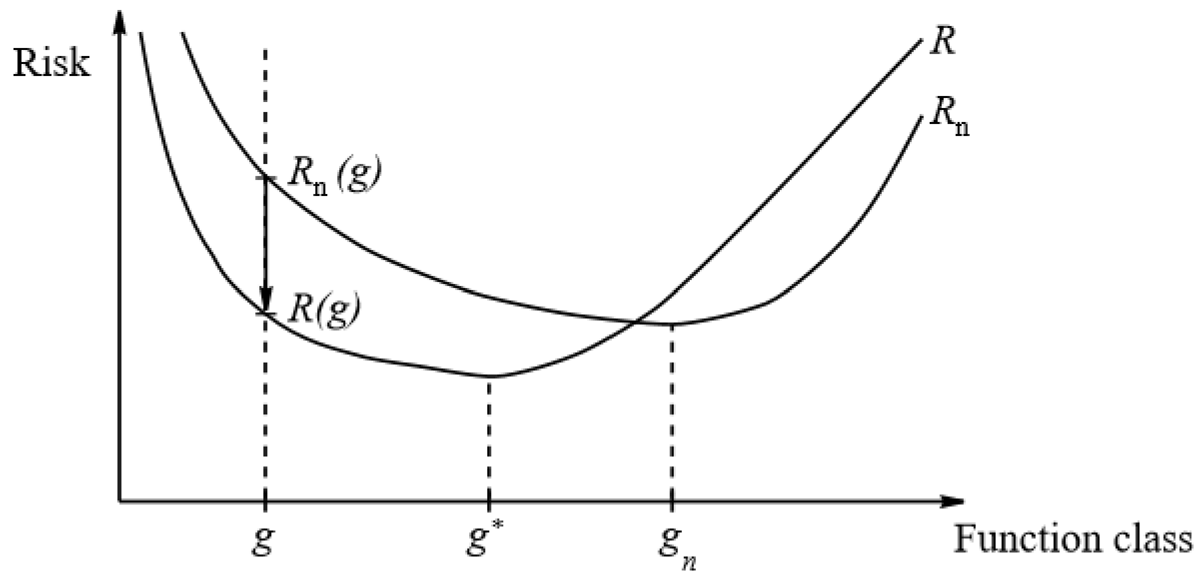
- For any fixed sample, there exists a function f such that

$$P f - P_n f = 1$$

  - E.g.: take the function which is $f(X_i) = Y_i$ on the data; and $f(X) = -Y$ elsewhere
  - This does not contradict Hoeffding but shows it is not enough!

Limitations

- Hoeffding's inequality quantifies differences for ***a fixed function***



Uniform Deviations

- Before seeing the data, we don't know which function the algorithm will choose

- The trick is to consider uniform deviations

$$R(f_n) - R_n(f_n) \leq \sup_{f \in \mathcal{F}} (R(f) - R_n(f))$$

- We need a bound which holds simultaneously for all functions in a class

Union Bound

- Consider two functions f_1, f_2 and define

$$C_i = \{(x_1, y_1), \ldots, (x_n, y_n) : Pf_i - P_n f_i > \varepsilon\}$$

- From Hoeffding's inequality, for each i:

$$\mathbb{P}\left[C_i\right] \leq \delta$$

- We want to bound the probability of being "bad" for i=1 or i=2

$$\mathbb{P}\left[C_1 \cup C_2\right] \leq \mathbb{P}\left[C_1\right] + \mathbb{P}\left[C_2\right]$$

Union Bound – finite case

- In general, for the finite case

$$\mathbb{P}\left[C_1 \cup \ldots \cup C_N\right] \leq \sum_{i=1}^{N} \mathbb{P}\left[C_i\right]$$

- We have

$$\begin{aligned} &\mathbb{P}\left[\exists f \in \{f_1, \ldots, f_N\} : Pf - P_n f > \varepsilon\right] \\ &\leq \sum_{i=1}^{N} \mathbb{P}\left[Pf_i - P_n f_i > \varepsilon\right] \\ &\leq N \exp\left(-2n\varepsilon^2\right) \end{aligned}$$

Union Bound – finite case

- We obtain, for $\mathcal{G} = \{g_1, \ldots, g_N\}$, for any $\delta > 0$ with probability at least $1 - \delta$

$$\forall g \in \mathcal{G}, \;\; R(g) \le R_n(g) + \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}$$

  - This is a generalization bound!

Estimation Error

- Let (best in a class)

$$g^* = \arg\min_{g \in \mathcal{G}} R(g)$$

- If g_n minimizes the empirical risk in G, we have

$$R_n(g^*) - R_n(g_n) \geq 0$$

- Thus

$$\begin{aligned} R(g_n) &= R(g_n) - R(g^*) + R(g^*) \\ &\leq R_n(g^*) - R_n(g_n) + R(g_n) - R(g^*) + R(g^*) \\ &\leq 2 \sup_{g \in \mathcal{G}} |R(g) - R_n(g)| + R(g^*) \end{aligned}$$

- We obtain with probability at least $1 - \delta$

$$R(g_n) \leq R(g^*) + 2\sqrt{\frac{\log N + \log \frac{2}{\delta}}{2n}}$$

Summary

- Inference requires assumptions
  - Data sampled i.i.d from P
  - Restrict the possible functions to G
  - Choose a sequence of models to have more flexibility/control
- Bounds are valid w.r.t. repeated sampling
  - For a fixed function g, for most of the samples

$$R(g) - R_n(g) \approx 1/\sqrt{n}$$

  - For most of samples if $|\mathcal{G}| = N$

$$\sup_{g \in \mathcal{G}} R(g) - R_n(g) \approx \sqrt{\log N/n}$$

Improvements

- We obtained

$$\sup_{g \in \mathcal{G}} R(g) - R_n(g) \leq \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2n}}$$

- Can be improved
  - Hoeffding only uses boundedness, not the variance
  - Union bound as bad as if independent
  - Supremum is not what the algorithm chooses

Refined Union Bound

- For each $f \in \mathcal{F}$, apply Hoeffding's inequality

$$\mathbb{P}\left[Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}} \right] \leq \delta(f)$$

- Thus, if we have a countable set $\mathcal{F}$, the union bound yields

$$\mathbb{P}\left[\exists f \in \mathcal{F} : Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}} \right] \leq \sum_{f \in \mathcal{F}} \delta(f)$$

- Choose $\delta(f) = \delta p(f)$ with $\sum_{f \in \mathcal{F}} p(f) = 1$
- With probability at least $1 - \delta$

$$\forall f \in \mathcal{F},\ Pf \leq P_n f + \sqrt{\frac{\log \frac{1}{p(f)} + \log \frac{1}{\delta}}{2n}}$$

  - Can put prior knowledge about the algorithm into $p(f)$!

Infinite Case: VC Theory

Infinite Case

- Measure of the size of an infinite class?
- Consider

$$\mathcal{F}_{z_1,\ldots,z_n} = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$$

  - The size of this set is the number of possible ways in which the data $(z_1, \ldots, z_n)$ can be classified

- Growth function

$$S_{\mathcal{F}}(n) = \sup_{(z_1,\ldots,z_n)} |\mathcal{F}_{z_1,\ldots,z_n}|$$

  - Note that $S_{\mathcal{F}}(n) = S_{\mathcal{G}}(n)$

Infinite Case

- Result (Vapnik-Chervonenkis): with probability at least $1-\delta$

$$\forall g \in \mathcal{G},\ \ R(g) \leq R_n(g) + 2\sqrt{2\frac{\log S_{\mathcal{G}}(2n) + \log \frac{\cancel{2}}{\delta}}{n}}\ \ 4$$

  - Always better than N in the finite case ($S_{\mathcal{G}}(n) \leq N$)
  - How to compute $S_{\mathcal{G}}(n)$ in general?

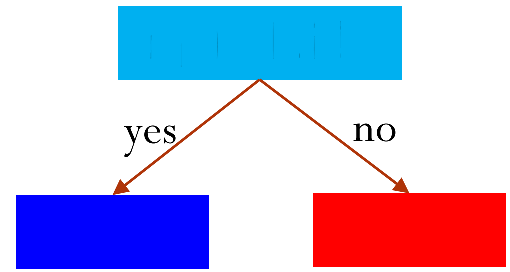
  - => use **VC dimension**!

VC Dimension

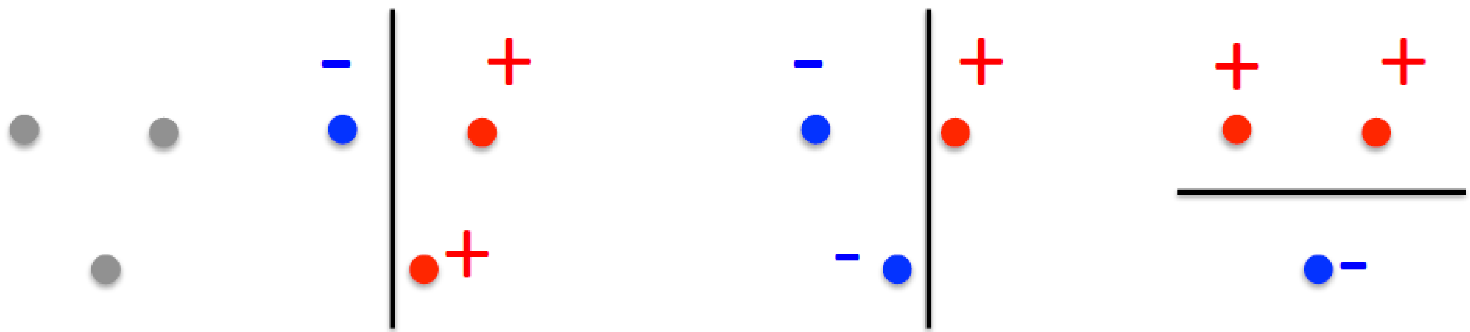
- Note that since $g \in \{-1, 1\}$, $S_{\mathcal{G}}(n) \leq 2^n$
- If $S_{\mathcal{G}}(n) = 2^n$, the class of functions can generate any classification on *n* points (shattering)
- **Definition**: The VC-dimension of $\mathcal{G}$ is the largest *n* such that

$$S_{\mathcal{G}}(n) = 2^n$$

Example of VC Dimension



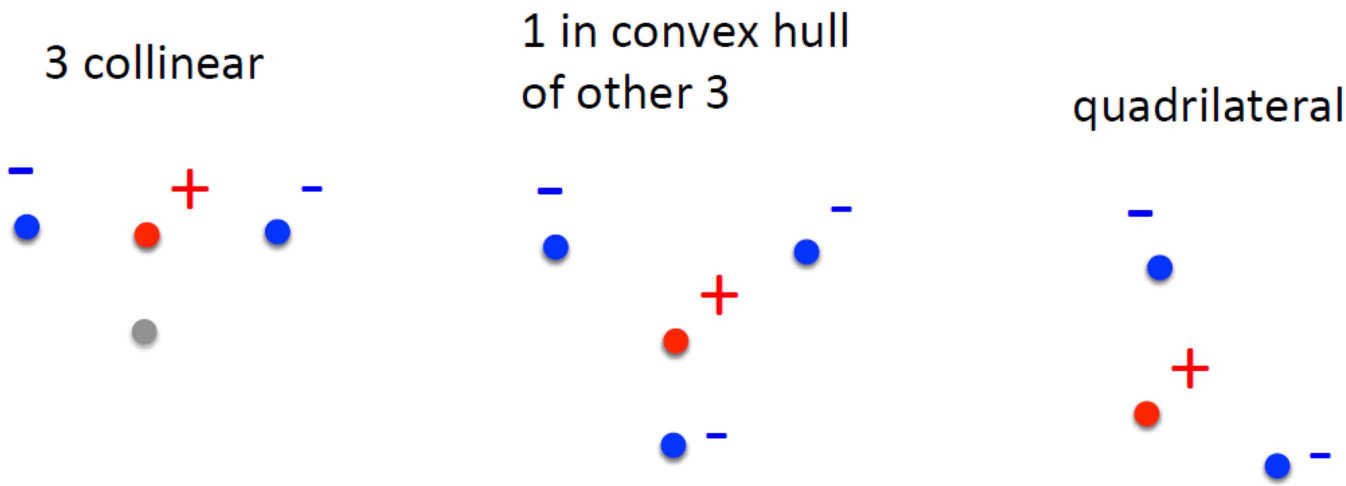
- Decision stumps in 2D
  - 3 data points can be shattered



  - How about 3 points in the same line?
    - Degenerating case (1D)!
  - VC >= 3

Example of VC Dimension

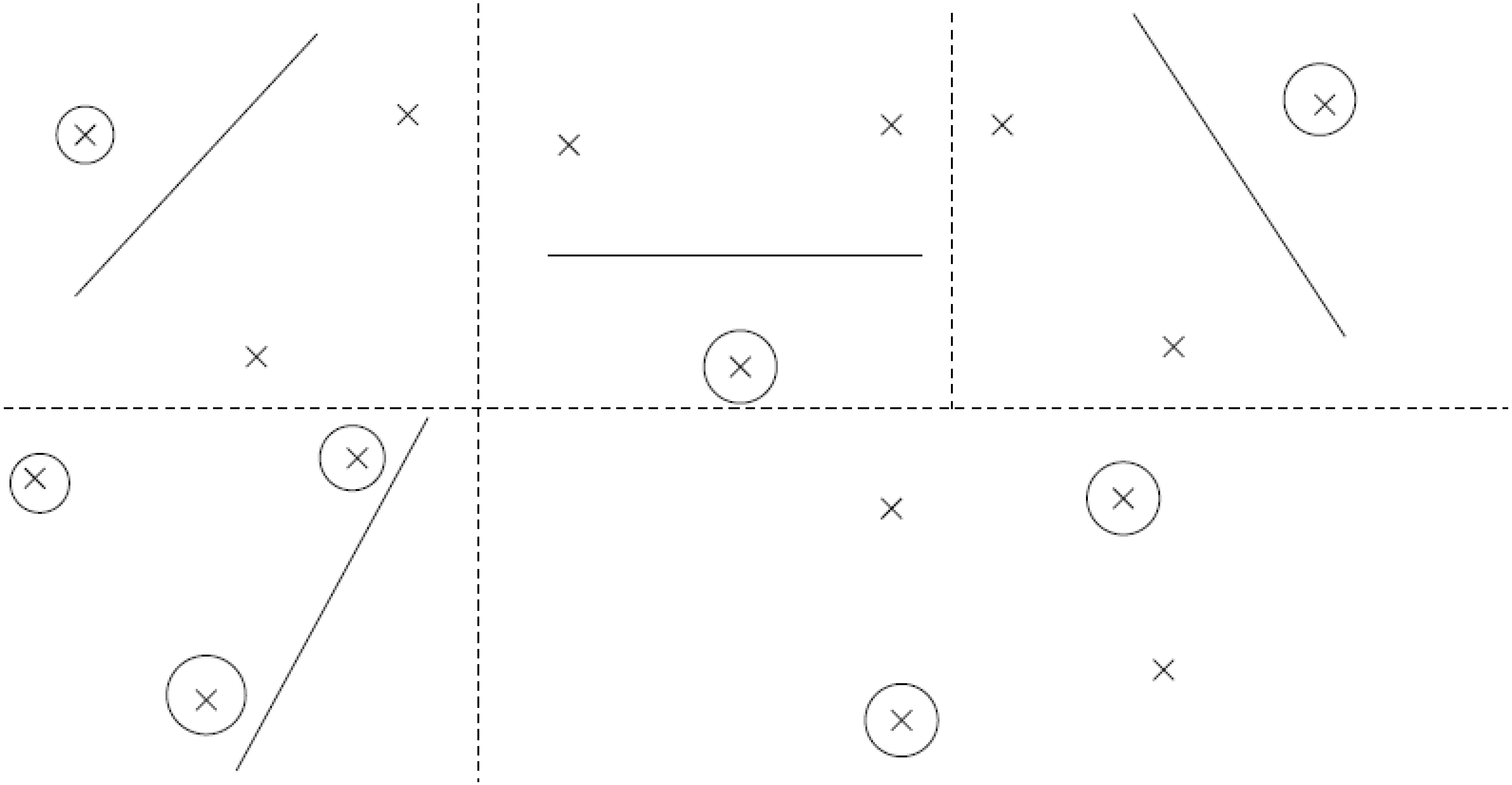
- Decision stumps in 2D
  - How about 4 data points?
    - For all placements of 4 pts, there exists a labeling that can't be shattered!



  - VC = 3
- In general, VC = d+1 (d-dim space)

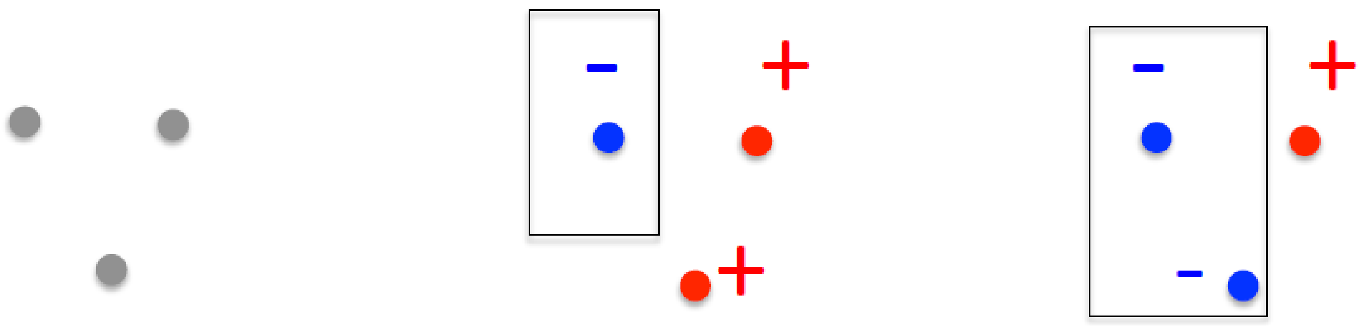
Example of VC Dimension

- Hyperplanes: In $\mathbb{R}^d$, $VC(\text{hyperplanes}) = d + 1$



Example of VC Dimension

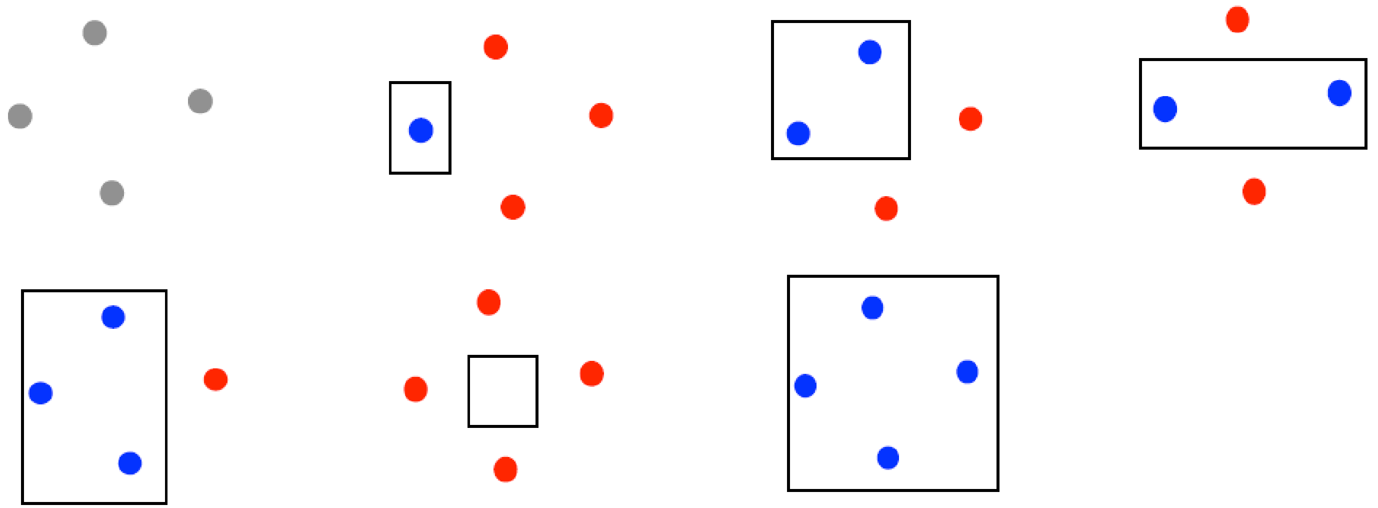
- Parallel rectangles in 2D



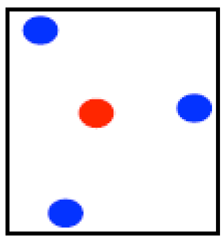
  - VC >= 3

Example of VC Dimension

- Parallel rectangles in 2D
  - How about 4 points?



  - Some placement of 4 points can't be shattered



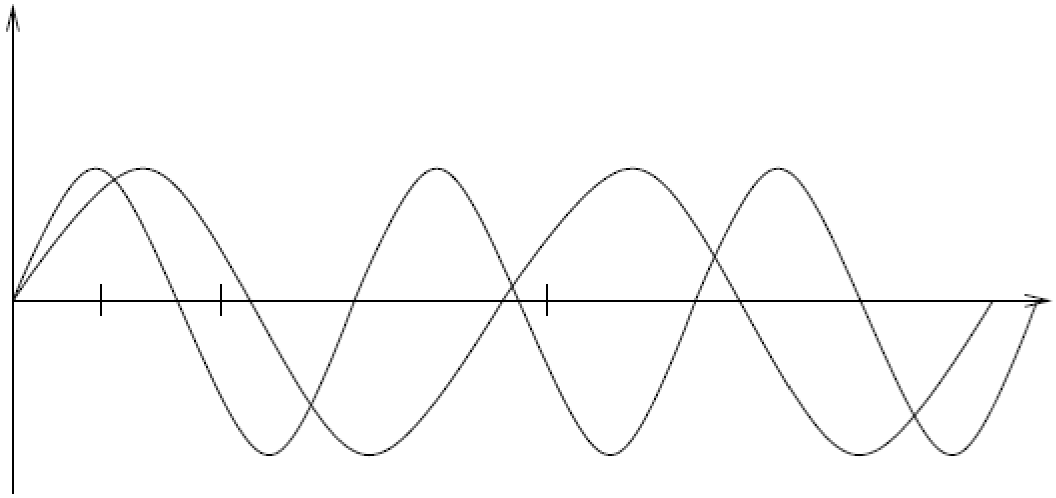
  - VC >= 4

Example of VC Dimension

- Parallel rectangles in 2D
  - How about 5 points? (homework)
  - Note: if VC = 4, then for all placements of 5 points, there exists a labeling that can't be shattered

VC Dimension

- Is VC-dimension equal to number of parameters?



  - One parameter $\{\mathrm{sgn}(\sin(tx)) : t \in \mathbb{R}\}$
  - Infinite VC dimension!

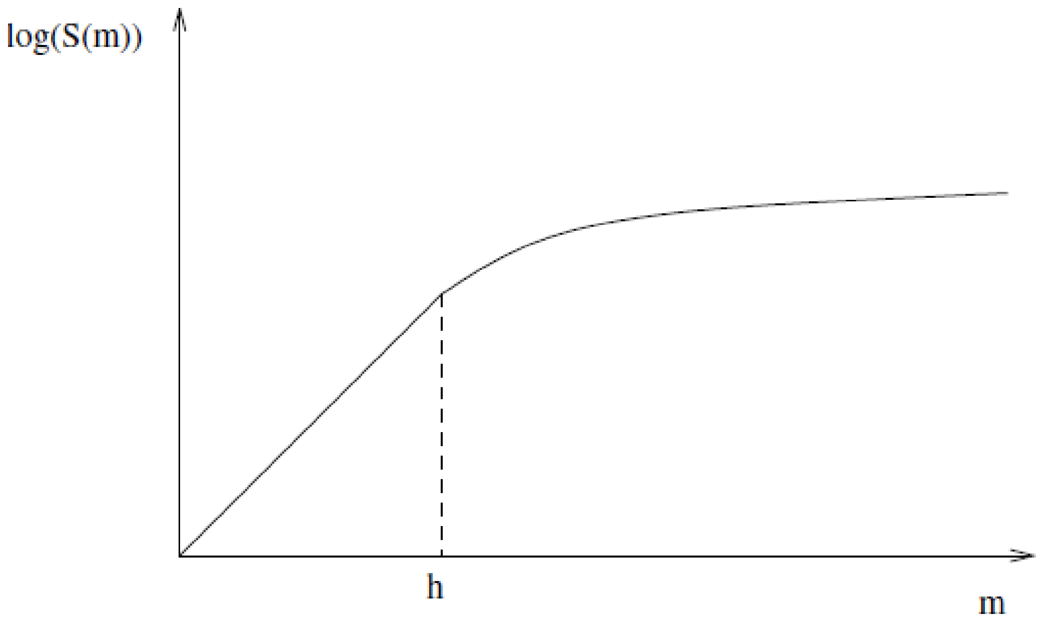
VC Dimension

- Is VC-dimension equal to number of parameters?

- 1 nearest neighbor:
  - Infinite dimension!

VC Dimension

- We know that $S_{\mathcal{G}}(n) = 2^n \text{ for } n \leq h$
- What happens for $n \geq h$?



Vapnik-Chervonenkis-Sauer-Shelah Lemma

- Let $\mathcal{G}$ be a class of functions with finite VC-dimension h.
- Then for all $n \in \mathbb{N}$

$$S_{\mathcal{G}}(n) \leq \sum_{i=0}^{h} \binom{n}{i}$$

- and for all $n \geq h$

$$S_{\mathcal{G}}(n) \leq \left(\frac{en}{h}\right)^{h}$$

VC Bound

- Let $\mathcal{G}$ be a class with VC-dimension h.
- With probability at least $1-\delta$

$$\forall g \in \mathcal{G},\ \ R(g) \leq R_n(g) + 2\sqrt{2\frac{h \log \frac{2en}{h} + \log \frac{2}{\delta}}{n}}$$

- So the error is of order

$$\sqrt{\frac{h \log n}{n}}$$

Interpretation of VC Dimension

- It is a measure of effective dimension
  - Depends on the geometry of the class
  - Gives a natural definition of simplicity (by quantifying the potential overfitting)
  - Not related to the number of parameters
  - Finiteness guarantees learnability under any distribution

Symmetrization Lemma

- Key ingredient in VC bounds

- Let $Z'_1, \ldots, Z'_n$ be an independent (ghost) sample and P'_n be the corresponding empirical measure

- **Lemma**: for any $t > 0$, *such that* $nt^2 \geq 2$,

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P - P_n)f \geq t\right] \leq 2\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P'_n - P_n)f \geq t/2\right]$$

Proof

f_n the function achieving the supremum (depends on $Z_1, \ldots, Z_n$)

$$\begin{aligned}
1_{[(P-P_n)f_n > t]} 1_{[(P-P'_n)f_n < t/2]} &= 1_{[(P-P_n)f_n > t \,\wedge\, (P-P'_n)f_n < t/2]} \\
&\leq 1_{[(P'_n-P_n)f_n > t/2]}
\end{aligned}$$

Taking expectations with respect to the second sample gives

$$1_{[(P-P_n)f_n > t]} \mathbb{P}' \left[(P - P'_n) f_n < t/2 \right] \leq \mathbb{P}' \left[(P'_n - P_n) f_n > t/2 \right]$$

By Chebyshev inequality,

$$\mathbb{P}' \left[(P - P'_n) f_n \geq t/2 \right] \leq \frac{4\mathrm{Var}\left[f_n\right]}{nt^2} \leq \frac{1}{nt^2}$$

Proof

- We have

$$1_{[(P-P_n)f_n>t]}(1-\frac{1}{nt^2}) \leq \mathbb{P}'\left[(P'_n-P_n)f_n>t/2\right]$$

- Take expectation with respect to first sample

Proof of VC Bound

- Symmetrization allows to replace expectation by average on ghost sample
- Function class projected on the double sample

$$\mathcal{F}_{Z_1,\ldots,Z_n,Z_1',\ldots,Z_n'}$$

- Union bound on $\mathcal{F}_{Z_1,\ldots,Z_n,Z_1',\ldots,Z_n'}$
- Variant of Hoeffding's inequality

$$\mathbb{P}\left[P_n f - P_n' f > t\right] \leq 2e^{-nt^2/2}$$

Proof of VC Bound

- More details

$$
\begin{aligned}
&\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P - P_n)f \geq t\right] \\
&\quad \leq \quad 2\mathbb{P}\left[\sup_{f\in\mathcal{F}}(P'_n - P_n)f \geq t/2\right] \\
&\quad = \quad 2\mathbb{P}\left[\sup_{f\in\mathcal{F}_{Z_1,\ldots,Z_n,Z'_1,\ldots,Z'_n}}(P'_n - P_n)f \geq t/2\right] \\
&\quad \leq \quad 2S_F(2n)\mathbb{P}\left[(P'_n - P_n)f \geq t/2\right] \\
&\quad \leq \quad 4S_F(2n)e^{-nt^2/8}
\end{aligned}
$$

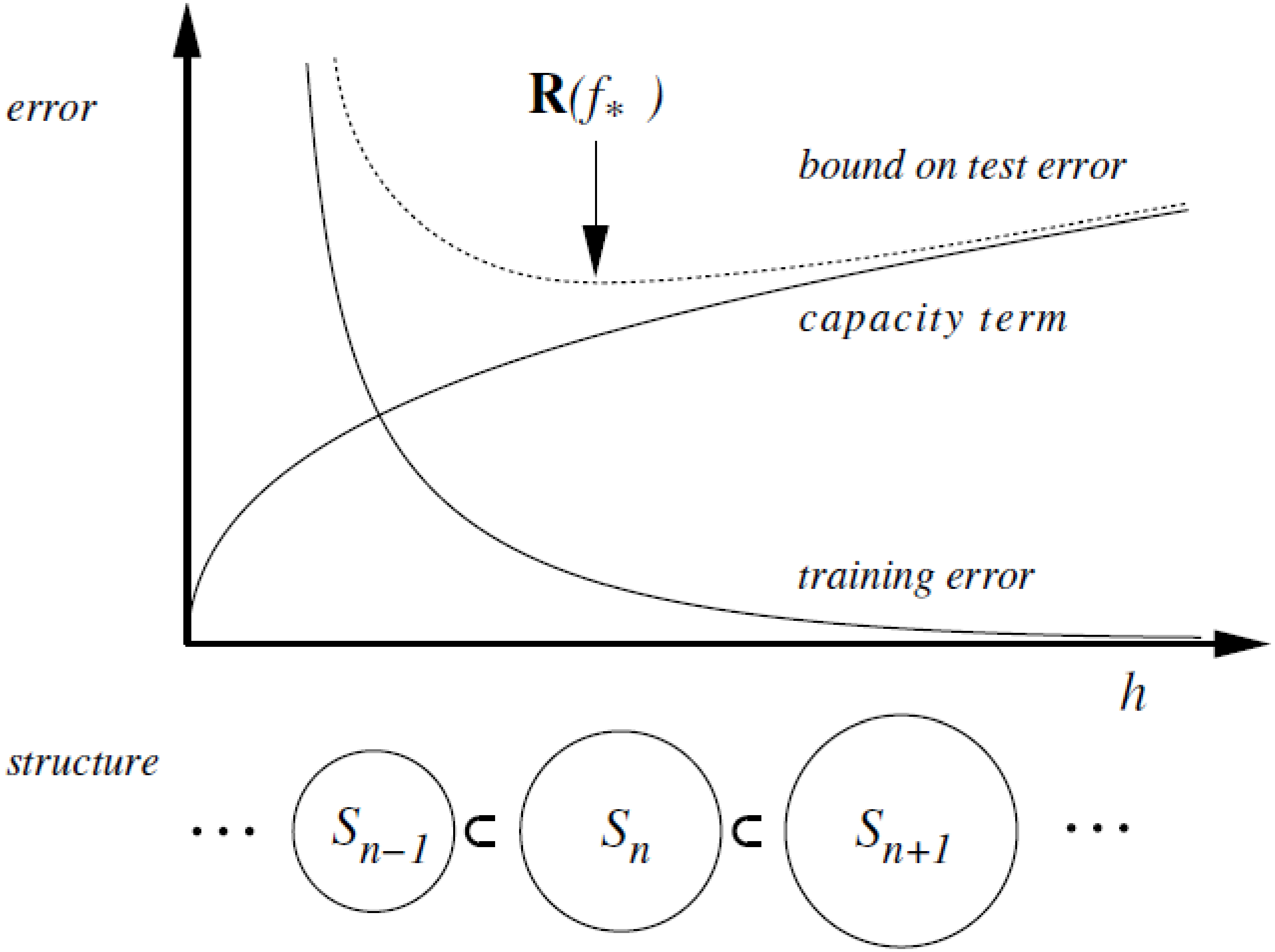
Structural Risk Minimization

- SRM (Vapnik, 1979): minimize the right hand side of

$$R(g) \leq R_n(g) + B(h, n)$$

  - To this end, introduce a structure on $\mathcal{G}$

SRM: the picture



Rademacher Complexity

- Rademacher variables: $\sigma_1, \ldots, \sigma_n$ independent r.v.s with

$$\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$$

- Randomized empirical fitness:

$$R_n g = \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(X_i)$$

- Consider all functions and define the Rademacher average:

$$\mathcal{R}(\mathcal{G}) = \mathbb{E}\left[\sup_{g \in \mathcal{G}} R_n g\right]$$

- Conditional Rademacher average (data fixed):

$$\mathcal{R}_n(\mathcal{G}) = \mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}} R_n g\right]$$

Error Bounds

- Distribution dependent
  - with high probability (at least 1-delta)

$$\forall g \in \mathcal{G},\ \ R(g) \le R_n(g) + \mathcal{R}(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- Data dependent
  - with high probability

$$\forall g \in \mathcal{G},\ \ R(g) \le R_n(g) + \mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{2\log(2/\delta)}{n}}$$

  - which depends solely on the data!

Computing Rademacher Average

- We have the rewritten form:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right] \\
&= \frac{1}{2}+\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}-\frac{1-\sigma_i g(X_i)}{2}\right] \\
&= \frac{1}{2}-\mathbb{E}\left[\inf_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\frac{1-\sigma_i g(X_i)}{2}\right] \\
&= \frac{1}{2}-\mathbb{E}\left[\inf_{g\in\mathcal{G}}R_n(g,\sigma)\right]
\end{aligned}$$

  - Computing $\mathcal{R}_n(\mathcal{G})$ is not harder than the empirical risk minimizer!
  - If the class is very large, $\mathcal{R}_n(\mathcal{G})=\frac{1}{2}$

Relationship with VC-dimension

- For a finite set $|\mathcal{G}| = N$

$$\mathcal{R}_n(\mathcal{G}) \leq 2\sqrt{\log N/n}$$

- For a class with VC-dimension h:

$$\mathcal{R}(\mathcal{G}) \leq 2\sqrt{\frac{h \log \frac{en}{h}}{n}}$$

- Recovers the VC bound with a concentration proof!
- One can improve the bound by a chaining technique (remove the $\log n$ factor!)

$$\mathcal{R}(\mathcal{G}) \leq C\sqrt{\frac{h}{n}}$$

Randomized Classifiers

- Given $\mathcal{G}$ a class of functions
- Deterministic: picks a function g_n and always use it to predict
- Randomized:
  - Construct a distribution ρ_n over $\mathcal{G}$
  - For each instance to classify, pick $g \sim \rho_n$
- Error is averaged over

Union Bound

- Let π be a (fixed) distribution over $\mathcal{G}$
- Recall the refined union bound

$$\forall g \in \mathcal{G},\ \ R(g) - R_n(g) \leq \sqrt{\frac{\log \frac{1}{\pi(g)} + \log \frac{1}{\delta}}{2n}}$$

- Take expectation with respect to ρ_n

$$\forall g \in \mathcal{G},\ \ R(\rho_n) - R_n(\rho_n) \leq \mathbb{E}_{\rho_n}\left[\sqrt{\frac{\log \frac{1}{\pi(g)} + \log \frac{1}{\delta}}{2n}}\right]$$

Union Bound

- We have

$$\forall g \in \mathcal{G},\ \ R(\rho_n) - R_n(\rho_n) \leq \mathbb{E}_{\rho_n}\left[\sqrt{\frac{\log\frac{1}{\pi(g)} + \log\frac{1}{\delta}}{2n}}\right]$$

$$\leq \sqrt{\frac{-\mathbb{E}_{\rho_n}\left[\log \pi(g)\right] + \log\frac{1}{\delta}}{2n}}$$

$$= \sqrt{\frac{KL(\rho_n, \pi) + H(\rho_n) + \log\frac{1}{\delta}}{2n}}$$

PAC-Bayesian Refinement

- It is possible to improve the previous bound
- With high probability at least $1 - \delta$

$$\forall g \in \mathcal{G}, \; R(\rho_n) - R_n(\rho_n) \leq \sqrt{\frac{KL(\rho_n, \pi) + \log 4n + \log \frac{1}{\delta}}{2n - 1}}$$

- Motivate development of classifiers by minimizing the bound!

PAC bound for SVMs

- SVMs use a linear classifier
  - For d features, VC = d+1

$$\forall g \in \mathcal{G}, \;\; R(g) \leq R_n(g) + \sqrt{\frac{(d+1)\log\frac{2en}{d+1} + \log\frac{4}{\delta}}{8n}}$$

- Problems!!!
  - Doesn't take margin into account
  - What about kernels?
    - Polynomials: number of features grows really fast = Bad bound!

$$\frac{(p+d-1)!}{p!(d-1)!}$$

d: input features
p: degree of polynomials

    - RBF kernels: can classify any set of points exactly!

Large Margin Bounds

- $\mathcal{G}$ class of linear functions with all have margin at least ρ
- R: radius of the smallest sphere enclosing the data points

$$VC(\mathcal{G}) \leq \min\left\{d, \frac{4R^2}{\rho}\right\} + 1$$

  - The larger the margin of functions in class $\mathcal{G}$, the smaller is its VC dimension!

Large Margin Bounds

$$\forall g \in \mathcal{G}, \;\; R(g) \leq R_n^{\rho}(g) + C\sqrt{\frac{\frac{R^2}{\rho^2}\log n + \log\frac{1}{\delta}}{n}}$$

- $R_n^{\rho}(g)$ the fraction of training examples which have margin smaller than ρ

- SVMs maximize margin ρ + minimize the hinge loss
  - Optimize tradeoff training error (bias) vs. margin ρ (variance)

What you need to know

- PAC bounds on true risk in terms of empirical risk (training error) and complexity of hypothesis space
- Complexity of the classifier depends on the number of points that can be classified exactly
  - Finite case – Number of hypothesis
  - Infinite case – VC dimension
- Bias-Variance tradeoff in learning theory
- Empirical and Structural Risk Minimization
  - But often bounds too loose in practice
- Other bounds – Margin based, PAC-Bayes, …

References

- O. Bousquet, S. Boucheron, & G. Lugosi. Introduction to Statistical Learning Theory, Lecture Notes in Artificial Intelligence, 3176:169-207, 2004.
- U. von Luxburg & B. Scholkopf. Statistical Learning Theory: Models, Concepts, and Results. Handbook of the History of Logic, 2009.