# Statistical Machine Learning

https://cvml.ist.ac.at/courses/SML_W20


Christoph Lampert

I|S|T AUSTRIA

*Institute of Science and Technology*


Fall Semester 2020/2021

Lecture 8

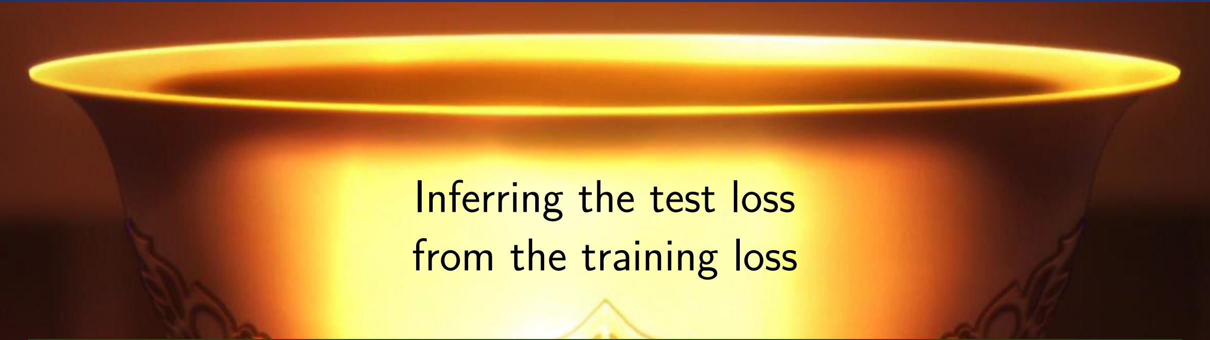
# Overview (tentative)

| Date | | no. | Topic |
|---|---|---|---|
| Oct 05 | Mon | 1 | A Hands-On Introduction |
| Oct 07 | Wed | 2 | Bayesian Decision Theory, Generative Probabilistic Models |
| Oct 12 | Mon | 3 | Discriminative Probabilistic Models |
| Oct 14 | Wed | 4 | Maximum Margin Classifiers, Generalized Linear Models |
| Oct 19 | Mon | 5 | Estimators; Overfitting/Underfitting, Regularization, Model Selection |
| Oct 21 | Wed | 6 | Bias/Fairness, Domain Adaptation |
| Oct 26 | Mon | - | no lecture (public holiday) |
| Oct 28 | Wed | 7 | Learning Theory I, Concentration of Measure |
| Nov 02 | Mon | 8 | Learning Theory II |
| Nov 04 | Wed | 9 | Deep Learning I |
| Nov 09 | Mon | 10 | Deep Learning II |
| Nov 11 | Wed | 11 | Unsupervised Learning |
| Nov 16 | Mon | 12 | project presentations |
| Nov 18 | Wed | 13 | buffer |

# The Holy Grail of Statistical Machine Learning

Inferring the test loss
from the training loss

# The Holy Grail of Statistical Machine Learning



Inferring the test loss
from the training loss

## Generalization Bound

For every $f \in \mathcal{H}$ it holds:

$$\underbrace{\mathop{\mathbb{E}}_{(x,y)} \ell(y, f(x))}_{\text{generalization loss}} \leq \underbrace{\frac{1}{n}\sum_i \ell(y_i, f(x_i))}_{\text{training loss}} + \text{something}$$

## Standard learning setting:

- input data $\mathcal{X}$, output set $\mathcal{Y}$, data distribution $p$ over $\mathcal{X} \times \mathcal{Y}$,
- loss function, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ (with some assumption),
- hypothesis set $\mathcal{H} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$,

**Generalization bounds: generic structure**

For any $\delta > 0$, the following statement holds with probablity at least $1 - \delta$ over the (random) training set $\mathcal{D}_n = \{(x^1, y^1), \dots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p$.

For all $f \in \mathcal{H}$:

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) \quad + \quad \text{something}$$

where the "something" typically increases for $\delta \to 0$ and decreases for $n \to \infty$.

**Observation:** if the inequality holds, it holds uniformly for all $f$.

$\to$ by minimizing the right hand side, we can find the "most promising" $f$

## Example: SVM radius/margin bound

Let $\ell(x, y; w) := \mathbf{max}\{0, 1 - y\langle w, x\rangle\}$ be the *hinge loss*. Let $p$ be a distribution on $\mathbb{R}^d \times \mathcal{Y}$ such that $\Pr\{\, \|x\| \leq R \,\} = 1$ and let $\mathcal{H} = \{f(x) = w^\top x : \quad w \in \mathbb{R}^d \wedge \|w\| \leq B\}$.

Then, with prob. at least $1 - \delta$ over $\mathcal{D}_m \overset{i.i.d.}{\sim} p$ the following inequality holds for all $w \in \mathcal{H}$:

$$\mathop{\mathbb{E}}_{(x,y)\sim p} [\![\langle w, x\rangle \neq y]\!] \leq \frac{1}{m}\sum_{i=1}^{m} \ell(x_i, y_i, w) + \frac{2RB}{\sqrt{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{1}$$

## Example: SVM radius/margin bound

Let $\ell(x,y;w) := \mathbf{max}\{0, 1 - y\langle w, x\rangle\}$ be the *hinge loss*. Let $p$ be a distribution on $\mathbb{R}^d \times \mathcal{Y}$ such that $\Pr\{\,\|x\| \leq R\,\} = 1$ and let $\mathcal{H} = \{f(x) = w^\top x : \quad w \in \mathbb{R}^d \wedge \|w\| \leq B\}$.

Then, with prob. at least $1-\delta$ over $\mathcal{D}_m \overset{i.i.d.}{\sim} p$ the following inequality holds for all $w \in \mathcal{H}$:

$$\mathop{\mathbb{E}}_{(x,y)\sim p} [\![\langle w, x\rangle \neq y]\!] \leq \frac{1}{m}\sum_{i=1}^{m} \ell(x_i, y_i, w) + \frac{2RB}{\sqrt{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}. \qquad (1)$$

This results provides a good justification for using SVMs:

- (1) holds uniformly in $w$, including for the $w$ that minimizes the right hand side
  $\rightarrow$ hinge loss on training set should be small
  $\rightarrow$ we should only consider $w$ with small $\|w\|$, such that $B$ can be chosen small

---

Reminder: (soft-margin) support vector machine (SVM):

$$\min_{w} \; \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i} \mathbf{max}\{0, 1 - y_i\langle w, x_i\rangle\}$$

## Example: Finite Hypothesis Sets

Setup:

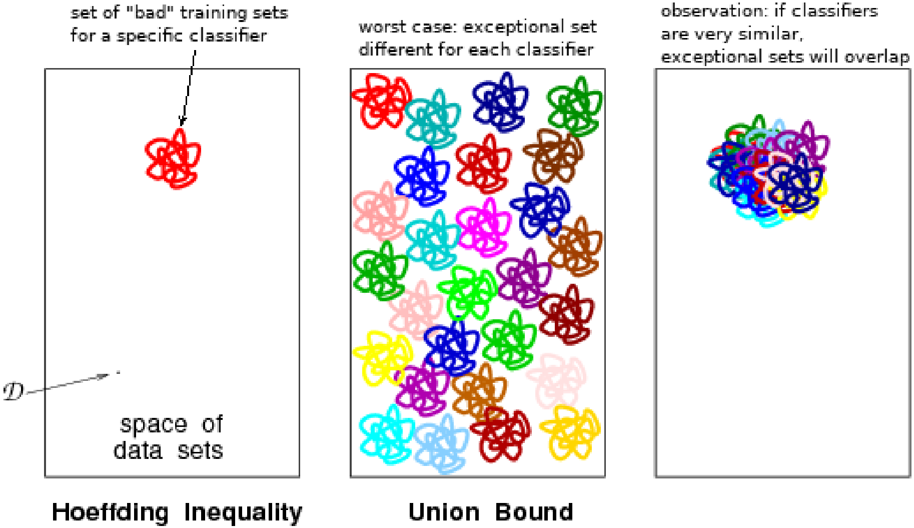
- $\ell(y, \bar{y}) = [\![ y \neq \bar{y} ]\!]$ (0-1 loss)
- finite number of possible classifiers $\mathcal{H} = \{f_1, \dots, f_T\} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$

For any $\delta > 0$, the following statement holds with probability at least $1 - \delta$ over the training set $\mathcal{D} = \{(x^1, y^1) \dots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p(x, y)$:

For all $f \in \mathcal{H}$: $$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{\log |\mathcal{H}| + \log 1/\delta}{2n}}.$$

This is essentially the lemma about uniform approximation we proved in lecture 7.

- Bound prob. of undesired outcome, $\mathcal{R}(f) - \hat{\mathcal{R}}(f) > \epsilon$, separately for each classifier $f$
- Combine by union bound $\to$ factor $|\mathcal{H}|$ (but ultimately enters only logarithmically)




Union bound is "worst case": usually overly pessimistic

Image: https://work.caltech.edu/library/

Union bound will only work for finite $\mathcal{H}$, otherwise even logarithm will not save us.

**Can we find a better way to characterize hypothesis classes than simply the number of their elements? Can we benefit from redundancy among hypotheses?**

Suggested complexity measures:

- covering numbers
- growth function
- VC dimension
- Rademacher complexity

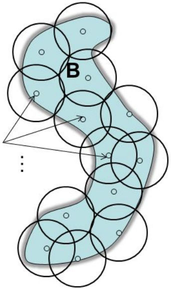
In particular, these work also for infinitely large (continuous) hypothesis sets.

**Definition (Covering)**

Let $\mathcal{F}$ be a set of functions. We say $\mathcal{F}$ is $\epsilon$-**covered** by $\mathcal{F}'$ with respect to a norm $\|\cdot\|$:

$$\forall f \in \mathcal{F} \quad \exists f' \in \mathcal{F}' \quad \|f - f'\| \leq \epsilon$$

$\mathcal{F}'$ is called an $\epsilon$-**cover** of $\mathcal{F}$.




**Definition (Covering Number)**

Let $\mathcal{F}$ be a set of functions. The $\epsilon$-**covering number**, $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$, is the size of the smallest $\epsilon$-cover of $\mathcal{F}$ with respect to $\|\cdot\|$.

Main idea: $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$ can be small (finite), even if $\mathcal{F}$ is large (infinite). We can use the cover $\mathcal{F}'$ for everything, yet still only make a small error.

Image: Lee Wee Sun. https://slideplayer.com/slide/7277867/

**Definition (Growth function)**

Let $\mathcal{H} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$ be a set of binary-valued hypotheses. The **growth function** $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ of $\mathcal{H}$ is defined as:

$$\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \left| \left\{ \big( h(x_1), \ldots, h(x_n) \big) : h \in \mathcal{H} \right\} \right|$$

For any $n \in \mathbb{N}$, $\Pi_{\mathcal{H}}(n)$ is the largest number of different labelings that can be produced with functions in $\mathcal{H}$.

Growth function: $$\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \left| \{ \big( h(x_1), \ldots, h(x_n) \big) : h \in \mathcal{H} \} \right|$$

**Examples:** growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$

Growth function: $\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \left| \{ \big( h(x_1), \ldots, h(x_n) \big) : h \in \mathcal{H} \} \right|$

**Examples:** growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$
- $\mathcal{H} = \{f_1, \ldots, f_T\}$ $\quad\rightarrow\quad$ $\Pi_{\mathcal{H}}(n) \leq \mathbf{min}\{2^n, |\mathcal{H}|\}$

Growth function: $$\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \left| \{ (h(x_1),\ldots,h(x_n)) : h \in \mathcal{H} \} \right|$$

**Examples:** growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$
- $\mathcal{H} = \{f_1, \ldots, f_T\}$ $\rightarrow$ $\Pi_{\mathcal{H}}(n) \leq \min\{2^n, |\mathcal{H}|\}$
- $\mathcal{H} = \{f : \mathcal{X} \rightarrow \{\pm 1\}\}$ (all binary values functions) and $|\mathcal{X}| = \infty$ $\rightarrow \Pi_{\mathcal{H}}(n) = 2^n$

Growth function: $$\Pi_{\mathcal{H}}(n) = \max_{x_1,\dots,x_n \in \mathcal{X}} \Big| \{ \big( h(x_1), \dots, h(x_n) \big) : h \in \mathcal{H} \} \Big|$$

**Examples:** growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$
- $\mathcal{H} = \{f_1, \dots, f_T\}$ $\rightarrow$ $\Pi_{\mathcal{H}}(n) \leq \mathbf{min}\{2^n, |\mathcal{H}|\}$
- $\mathcal{H} = \{f : \mathcal{X} \rightarrow \{\pm 1\}\}$ (all binary values functions) and $|\mathcal{X}| = \infty$ $\rightarrow \Pi_{\mathcal{H}}(n) = 2^n$
- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\text{sign}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ all linear classifiers
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2^n$ for $n \leq d + 1$, but $\Pi_{\mathcal{H}}(n) < 2^n$ for $n > d + 1$.

Growth function: $\Pi_{\mathcal{H}}(n) = \max_{x_1,\dots,x_n \in \mathcal{X}} \Big| \{ \big( h(x_1), \dots, h(x_n) \big) : h \in \mathcal{H} \} \Big|$

**Examples:** growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$
- $\mathcal{H} = \{f_1, \dots, f_T\}$ $\rightarrow$ $\Pi_{\mathcal{H}}(n) \leq \mathbf{min}\{2^n, |\mathcal{H}|\}$
- $\mathcal{H} = \{f : \mathcal{X} \to \{\pm 1\}\}$ (all binary values functions) and $|\mathcal{X}| = \infty$ $\rightarrow \Pi_{\mathcal{H}}(n) = 2^n$
- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathrm{sign}(\langle w, x \rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ all linear classifiers
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2^n$ for $n \leq d+1$, but $\Pi_{\mathcal{H}}(n) < 2^n$ for $n > d+1$.
- $\mathcal{X} = [0,1]$, $\mathcal{H} = \{\mathrm{sign}(\sin(\omega x)), \quad \omega \in \mathbb{R}\}$ $\rightarrow \Pi_{\mathcal{H}}(n) = 2^n$

## Growth Function Generalization Bound

Setup:

- $\ell(y, \bar{y}) = [\![ y \neq \bar{y} ]\!]$ (0-1 loss)
- $\mathcal{H} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$

For any $\delta > 0$, the following statement holds with probability at least $1 - \delta$ over the training set $\mathcal{D} = \{(x^1, y^1), \dots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p(x, y)$:

For all $f \in \mathcal{H}$:

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}}{n}} + \sqrt{\frac{\log 1/\delta}{2n}}$$

## Growth Function Generalization Bound

Setup:

- $\ell(y, \bar{y}) = [\![y \neq \bar{y}]\!]$ (0-1 loss)
- $\mathcal{H} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$

For any $\delta > 0$, the following statement holds with probability at least $1 - \delta$ over the training set $\mathcal{D} = \{(x^1, y^1) \dots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p(x, y)$:

For all $f \in \mathcal{H}$:

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}}{n}} + \sqrt{\frac{\log 1/\delta}{2n}}$$

- for $|\mathcal{H}| < \infty$, we (almost) recover the bound for finite hypothesis sets
- bound is vacuous for $\Pi_{\mathcal{H}}(n) = 2^n$, but interesting for $\Pi_{\mathcal{H}}(n) \ll 2^n$

Problem: growth function (for all $n \in \mathbb{N}$) can be hard to determine precisely
Easier: at what value does it change from $\Pi_{\mathcal{H}}(n) = 2^n$ to $\Pi_{\mathcal{H}}(n) < 2^n$ ?

**Definition (VC Dimension)**

The **VC dimension** of a hypothesis class $\mathcal{H}$, denoted VCdim$(\mathcal{H})$, is the maximal value $n$, for which $\Pi_{\mathcal{H}}(n) = 2^n$. If no such value exists, we say that VCdim$(\mathcal{H}) = \infty$.

Problem: growth function (for all $n \in \mathbb{N}$) can be hard to determine precisely
Easier: at what value does it change from $\Pi_{\mathcal{H}}(n) = 2^n$ to $\Pi_{\mathcal{H}}(n) < 2^n$ ?

## Definition (VC Dimension)

The **VC dimension** of a hypothesis class $\mathcal{H}$, denoted VCdim$(\mathcal{H})$, is the maximal value $n$, for which $\Pi_{\mathcal{H}}(n) = 2^n$. If no such value exists, we say that VCdim$(\mathcal{H}) = \infty$.

**Examples:**

- $\mathcal{H} = \{f_+, f_-\}$ for $f_+(x) = +1$ and $f_-(x) = -1$. $\rightarrow$ VCdim$(\mathcal{H}) = 1$
- $\mathcal{H} = \{f_1, \dots, f_T\}$ $\rightarrow$ VCdim$(\mathcal{H}) \leq \lfloor \log_2 |\mathcal{H}| \rfloor$
- $\mathcal{H} = \{f : \mathcal{X} \rightarrow \{\pm 1\}\}$ (all binary values functions) and $|\mathcal{X}| = \infty$
  $\rightarrow$ VCdim$(\mathcal{H}) = \infty$
- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathrm{sign}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ (linear classifiers)
  $\rightarrow$ VCdim$(\mathcal{H}) = d + 1$
- $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{\mathrm{sign}(\sin(\omega x)), \quad \omega \in \mathbb{R}\}$ $\rightarrow$ VCdim$(\mathcal{H}) = \infty$

**Reminder:**
VCdim($\mathcal{H}$) is the maximal value $n$, for which $\Pi_{\mathcal{H}}(n) = 2^n$, or $\infty$ if no such $n$ exists.

**Lemma (Sauer's Lemma)**

*For any $\mathcal{H}$ with VCdim($\mathcal{H}$) $< \infty$, for any $m$:* $\Pi_{\mathcal{H}}(n) \leq \sum_{k=0}^{\text{VCdim}(\mathcal{H})} \binom{n}{k}$.

Consequence:

- up to $n =$ VCdim($\mathcal{H}$), growth function grows **exponentially**
- for $n \geq$ VCdim($\mathcal{H}$)$+1$, growth function grows only **polynomially**:

$$\Pi_{\mathcal{H}}(n) \leq (en/d)^d = O(n^d). \qquad \text{(proof: textbook)}$$

- for $n >$ VCdim($\mathcal{H}$), complexity term $\sqrt{\frac{2 \log \Pi_{\mathcal{H}}(n)}{n}}$ starts decreasing like $O(\sqrt{\frac{\log n}{n}})$

## VC-Dimension Generalization Bound

Setup: inputs $\mathcal{X}$, outputs $\mathcal{Y} = \{\pm 1\}$, $\ell(y, \bar{y}) = [\![ y \neq \bar{y} ]\!]$, $\mathcal{H} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$.

For any $\delta > 0$, the following statement holds with probability at least $1 - \delta$ over the training set $\mathcal{D} = \{(x^1, y^1) \dots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p(x, y)$:

For all $f \in \mathcal{H}$: $$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{2d \log \frac{en}{d}}{n}} + \sqrt{\frac{\log 1/\delta}{2n}}$$ where $d = \text{VCdim}(\mathcal{H})$

**Observations:**

- Dimension of $\mathcal{X}$ plays no role, only $d = \text{VCdim}(\mathcal{H})$
- Crucial quantity: $\frac{d}{n}$. Non-trivial bound only for $n > d$.

1) **polynomial classifiers**,

$$\mathcal{H} = \{h(x) = \operatorname{sign} f(x), \text{for } f \text{ any polynomial of degree } k \text{ in } \mathbb{R}^d\}.$$

$\mathrm{VCdim}(\mathcal{H}) = \sum_{i=0}^{k} \binom{d+1}{i}$

# More examples: VC dimension (from the literature)

1) **polynomial classifiers**,

$$\mathcal{H} = \{h(x) = \text{sign}\, f(x), \text{for } f \text{ any polynomial of degree } k \text{ in } \mathbb{R}^d\}.$$

$$\text{VCdim}(\mathcal{H}) = \sum_{i=0}^{k} \binom{d+1}{i}$$

2) **boosting**: base set, $\mathcal{F}$, of weak classifiers with VCdim $D$.

$$\mathcal{H} = \Big\{ f(x) = \sum_{t=1}^{T} \alpha_t g_t(x), \text{ for } g_1, \dots, g_T \in \mathcal{F} \text{ and } \alpha_1, \dots, \alpha_T \in \mathbb{R} \Big\}$$

$$\text{VCdim}(\mathcal{H}) \leq T(D+1) \cdot (3 \log(T(D+1)) + 2)$$

# More examples: VC dimension (from the literature)

1) **polynomial classifiers**,

$$\mathcal{H} = \{h(x) = \operatorname{sign} f(x), \text{for } f \text{ any polynomial of degree } k \text{ in } \mathbb{R}^d\}.$$

$$\text{VCdim}(\mathcal{H}) = \sum_{i=0}^{k} \binom{d+1}{i}$$

2) **boosting**: base set, $\mathcal{F}$, of weak classifiers with VCdim $D$.

$$\mathcal{H} = \Big\{ f(x) = \sum_{t=1}^{T} \alpha_t g_t(x), \text{ for } g_1, \dots, g_T \in \mathcal{F} \text{ and } \alpha_1, \dots, \alpha_T \in \mathbb{R} \Big\}$$

$$\text{VCdim}(\mathcal{H}) \leq T(D+1) \cdot (3 \log(T(D+1)) + 2)$$

3) **neural networks** with threshold activation functions,
VCdim$(\mathcal{H}) \leq O(W \log W)$ where $W$ is number of network weights

4) **neural networks** with ReLU activation functions,
VCdim$(\mathcal{H}) \leq O(W L \log W)$ where $L$ is the number of network layers

# From classical to modern generalization bounds

# Towards Modern Generalization Bounds

Generalization bounds so far: with probability at least $1-\delta$:

$$\forall f \in \mathcal{H}: \quad \mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + B(\mathcal{H}, n, \delta)$$

Observation:

- $B(\mathcal{H}, n, \delta)$ is data-independent
- data distribution does not show up anywhere
  $\rightarrow$ holds for "easy" as well as "hard" learning problems

- minimizing right hand side is just ERM

More interesting: data-dependent or distribution-dependent bounds

- $\mathcal{Z}$: input set (later: $\mathcal{Z} = \mathcal{X}$ or $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$), $p(z)$: probability distribution over $\mathcal{Z}$
- $\mathcal{F} \subseteq \{f : \mathcal{Z} \to \mathbb{R}\}$: set of real-valued functions

**Definition (Empirical Rademacher Complexity)**

Let $\mathcal{F} = \{f : \mathcal{Z} \to \mathbb{R}\}$ be a set of real-valued functions and $\mathcal{D}_m = \{z_1, \ldots, z_m\}$ a finite set. The **empirical Rademacher complexity** of $\mathcal{F}$ with respect to $\mathcal{D}_m$ is defined as

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} \sigma_i f(z_i)\right)\right]$$

where $\sigma_1, \ldots, \sigma_m$ are independent binary random variables with $p(+1) = p(-1) = \frac{1}{2}$ (called **Rademacher variables**).

Intuition: think of $\sigma_i$ as random noise. The $\mathbf{sup}$ measures how well functions in $\mathcal{F}$ can correlate to arbitrary values (=memorize random noise).

Note: $\hat{\mathfrak{R}}_{\mathcal{D}_m}$ is data-dependent, it depends on $\mathcal{D}_m$.

**Example**

Let $\mathcal{F} = \{f\}$ (a single function). Then, for any $m$,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right) = \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\sigma}[\sigma_i]f(z_i) = 0$$

**Example**

Let $\mathcal{F} = \{f : \mathcal{Z} \to [-B, B]\}$ all bounded functions. Then, when there are no duplicates in $\mathcal{D}$,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right) \overset{f(z_i)=B\sigma_i}{=} \mathop{\mathbb{E}}_{\sigma}\frac{1}{m}\sum_{i=1}^{m}B = \mathop{\mathbb{E}}_{\sigma}B = B$$

(same argument would work also, e.g., for piecewise linear functions)

**Example**

Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ with $f_i : \mathcal{X} \to [-B, B]$ for $i = 1, \ldots, K$ (finitely many bounded functions). Then

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \leq B\sqrt{\frac{2\log K}{m}}$$

Proof: textbook

**Example**

Let $\mathcal{F} = \{f = w^\top z : \mathbb{R}^d \to \mathbb{R}\}$ with $\|w\| \leq B$ all *linear* functions with bounded slope. If $m > d$, then $z_1, \ldots, z_m$ are linearly dependent and $\mathbf{sup}$ can't fit all possible signs $\to$ $\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$ will decrease with $m$.

(we'll prove a more rigorous statement later)

**Definition**

The **Rademacher complexity** of $\mathcal{F}$ is defined as

$$\mathfrak{R}_m(\mathcal{F}) = \underset{\mathcal{D}_m \sim p^{\otimes m}}{\mathbb{E}} [\ \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})\ ]$$

Note: $\mathfrak{R}_m$ is a distribution-dependent quantity (w.r.t. $p$).

In some cases, more convenient to compute than the empirical one.

Slightly more general notation than before:

- hypothesis set $\mathcal{H} \subset \{\mathcal{X} \to \mathbb{R}\}$ (can be real-valued)
- loss $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$, e.g. $\ell(x, y, h) = \mathbf{max}\{0, 1 - yh(x)\}$,
- $\mathcal{R}(h) = \mathbb{E}_{(x,y)\sim p}\, \ell(x, y, h)$, $\hat{\mathcal{R}}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(x_i, y_i, h)$

Slightly more general notation than before:

- hypothesis set $\mathcal{H} \subset \{\mathcal{X} \to \mathbb{R}\}$ (can be real-valued)
- loss $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$, e.g. $\ell(x, y, h) = \mathbf{max}\{0, 1 - yh(x)\}$,
- $\mathcal{R}(h) = \mathbb{E}_{(x,y)\sim p}\, \ell(x, y, h)$, $\hat{\mathcal{R}}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(x_i, y_i, h)$

**Theorem (Rademacher-based generalization bound)**

*Let $\ell(x, y, h) \le c$ be a bounded loss function and set*

$$\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\} \quad = \{\ell(x, y, h(x)) : h \in \mathcal{H}\} \subset \{f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$$

*Then, with prob. at least $1 - \delta$ over $\mathcal{D}_m \overset{i.i.d.}{\sim} p$, it holds for all $h \in \mathcal{H}$:*

$$\mathcal{R}(h) \le \hat{\mathcal{R}}(h) + 2\mathfrak{R}_m(\mathcal{F}) + c\sqrt{\frac{\log(1/\delta)}{2m}}.$$

*Also, with prob. at least $1 - \delta$, it holds for all $h \in \mathcal{H}$:*

$$\mathcal{R}(h) \le \hat{\mathcal{R}}(h) + 2\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) + 3c\sqrt{\frac{2\log(4/\delta)}{m}}.$$

**Proof.** textbook/notes □

Useful properties:

**Lemma**

*For* $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ *let* $\mathcal{F}' := \{f + f_0 : f \in \mathcal{F}\}$ *be a translated version for some* $f_0 : \mathcal{X} \to \mathbb{R}$ *. Then, for any* $m$*,*

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') = \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

**Lemma**

*For* $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ *let* $\mathcal{F}' := \{\lambda f : f \in \mathcal{F}\}$ *be scaled by a constant* $\lambda \in \mathbb{R}$*. Then, for any* $m$*,*

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') = \lambda \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

**Lemma**

*For* $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ *and* $\phi : \mathbb{R} \to \mathbb{R}$ *let* $\mathcal{F}' := \{\phi \circ f : f \in \mathcal{F}\}$*. If* $\phi$ *is* $L$*-Lipschitz continuous, i.e.* $|\phi(t) - \phi(t')| \leq L|t - t'|$*, then for any* $m$*,*

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') \leq L \cdot \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

**Lemma**

*Let $\mathcal{Z}$ be an inner-product space (e.g. $\mathbb{R}^d$ with $\langle \cdot, \cdot \rangle$). Let $\mathcal{F} = \{f = \langle w, z \rangle : \mathcal{X} \to \mathbb{R}\}$ be the set of linear functions with $\|w\| \leq B$. Then, for any $\mathcal{D}_m = \{z_1, \ldots, z_m\}$,*

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \leq \frac{B}{m}\sqrt{\sum_i \|z_i\|^2}$$

**Proof:** textbook/notes

**Lemma**

*Let $\mathcal{F} = \{f = \langle w, z \rangle : \mathcal{X} \to \mathbb{R}\}$ be linear functions with $\|w\| \leq B$ and let $p$ be such that $\Pr\{\|z\| < R\} = 1$ Then*

$$\mathfrak{R}_m(\mathcal{F}) \leq BR\sqrt{\frac{1}{m}}$$

Proof: $\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \leq \frac{B}{m}\sqrt{mR^2}$ with prob. 1, so $\mathbb{E}_{\mathcal{D}}\hat{\mathfrak{R}} \leq \frac{B}{m}\sqrt{mR^2}$, too.

Reminder: (soft-margin) support vector machine (SVM):

$$\min_{w} \; \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i} \mathbf{max}\{0, 1 - y_i\langle w, x_i\rangle\}$$

Reminder: (soft-margin) support vector machine (SVM):

$$\min_w \; \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_i \mathbf{max}\{0, 1 - y_i\langle w, x_i\rangle\}$$

## Example: SVM "radius/margin" bound

Let $\ell(x, y; w) := \mathbf{max}\{0, 1 - y\langle w, x\rangle\}$ be the *hinge loss*. Let $p$ be a distribution on $\mathbb{R}^d \times \mathcal{Y}$ such that $\Pr\{\,\|x\| \leq R\,\} = 1$ and let $\mathcal{H} = \{h(x) = \langle w, x\rangle : \; w \in \mathbb{R}^d \wedge \|w\| \leq B\}$.

Then, with prob. at least $1 - \delta$ over $\mathcal{D}_m \overset{i.i.d.}{\sim} p$ the following inequality holds for all $w \in \mathcal{H}$:

$$\mathop{\mathbb{E}}_{(x,y)\sim p} [\![\text{sign}\langle w, x\rangle \neq y]\!] \leq \frac{1}{m}\sum_{i=1}^{m} \mathbf{max}\{0, 1 - y^i\langle w, x^i\rangle\} + \frac{2BR}{\sqrt{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Properties:

- complexity terms decrease with rate $O(\sqrt{\frac{1}{m}})$
- short $\|w\|$ is better than long $\|w\|$
- dimensionality of $x$ does not show up, no curse of dimensionality!

**Proof sketch:**

- $\|x\| \leq R$ (with probability 1)
- "ramp loss": $\ell(x,y,h) = \mathbf{min}\{\ \mathbf{max}\{0, 1-yh(x)\}, 1\ \} \ \in [0,1]$
- $\mathcal{H} = \{h(x) = \langle w, x\rangle : \|w\| \leq B\}, \quad \mathcal{F} = \{\ell \circ h,\ h \in \mathcal{H}\}$

With prob. $1-\delta$: $\forall h \in \mathcal{H} : \mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m}}$

---

- $\ell$ is 1-Lipschitz, i.e. for $\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\}$:

$$\mathfrak{R}_m(\mathcal{F}) \overset{\text{1-Lip.}}{\leq} \mathfrak{R}_m(\mathcal{H}) \overset{\text{Lemma}}{\leq} BR\sqrt{\frac{1}{m}}$$

- $\ell$ is upper bounds to 0/1 error and lower bound to hinge loss

$$\Pr\{h(x) \neq y\} \leq \mathcal{R}(h) \qquad \hat{\mathcal{R}}(h) \leq \frac{1}{m}\sum_{i=1}^{m} \mathbf{max}\{0, 1 - y_i h(x_i)\}$$

---

With prob. $1-\delta$ for every $h = \langle w, x\rangle \in \mathcal{H}$:

$$\Pr\{\operatorname{sign}\langle w, x\rangle \neq y\} \leq \frac{1}{m}\sum_{i=1}^{m} \mathbf{max}\{0, 1 - y_i\langle w, x_i\rangle\} + \frac{2RB}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

### Theorem (Connections to other complexity measures)

*Let* $\mathcal{H} = \{h : \mathcal{X} \to \{\pm 1\}\}$ *be a hypothesis class. Then*

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2\log|\mathcal{H}|}{m}} \qquad \text{if } |\mathcal{H}| \text{ is finite,}$$

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2\log\Pi_{\mathcal{H}}(m)}{m}} \qquad \text{where } \Pi_{\mathcal{H}}(m) \text{ is the growth function,}$$

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2d\log m}{m}} \qquad \text{where } d = \text{VCdim}(\mathcal{H}).$$

### Theorem (Connections to covering numbers)

*Let* $\mathcal{H} \subset \{\mathcal{X} \to [-1,1]\}$ *and* $\mathcal{D} \overset{i.i.d.}{\sim} p(x,y)$ *with* $|\mathcal{D}| = m$. *Then*

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leq \inf_{\alpha} \left[\alpha + \sqrt{\frac{\mathcal{N}(\alpha, \mathcal{H}|_{\mathcal{D}}, \|\cdot\|_{L_1})}{m}}\ \right]$$

*where* $\mathcal{N}$ *are covering numbers of the set of values that* $\mathcal{H}$ *assigns to* $\mathcal{D}$.

# Beyond Complexity Measures

# Algorithm-dependent bounds

Generalization bounds so far: with probability at least $1-\delta$:

$$\forall f \in \mathcal{H}: \quad \mathcal{R}(f) \le \hat{\mathcal{R}}(f) + \text{"something"}$$

Observation:

- holds simultaneous for all hypotheses in $\mathcal{H}$, we can pick any we like

  but: in practice, we have some algorithm that choses the hypothesis and we really only need the result for that

# Algorithm-dependent bounds

Generalization bounds so far: with probability at least $1-\delta$:

$$\forall f \in \mathcal{H}: \quad \mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \text{"something"}$$

Observation:

- holds simultaneous for all hypotheses in $\mathcal{H}$, we can pick any we like

  but: in practice, we have some algorithm that choses the hypothesis and we really only need the result for that

**Goal: algorithm-dependent bounds**

Instead of

- *"For which hypothesis sets does learning not overfit?"*

ask

- *"Which learning algorithms do not overfit?"*

- hypothesis set $\mathcal{H}$,     write loss function in form $L(x, y, h) = \ell(y, h(x))$

**Definition (Learning algorithm)**

A **learning algorithm**, $A$, is a function that takes as input a finite subset, $\mathcal{D}_m \subset \mathcal{Z}$, and outputs a hypothesis $A[\mathcal{D}_m] \in \mathcal{H}$.

- hypothesis set $\mathcal{H}$,    write loss function in form $L(x, y, h) = \ell(y, h(x))$

**Definition (Learning algorithm)**

A **learning algorithm**, $A$, is a function that takes as input a finite subset, $\mathcal{D}_m \subset \mathcal{Z}$, and outputs a hypothesis $A[\mathcal{D}_m] \in \mathcal{H}$.

**Definition (Uniform stability)**

For a training set, $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, we define the training set with the $i$-th element removed

$$\mathcal{D}^{\setminus i} = \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_m, y_m)\}.$$

A learning algorithm, $A$, has **uniform stability** $\beta$ with respect to the loss $\ell$ if the following holds,

$$\forall \mathcal{D}_m \subset \mathcal{X} \times \mathcal{Y} \; \forall i \in \{1, 2, \ldots, m\} \quad \| L(\cdot, \cdot, A[\mathcal{D}]\,) - L(\cdot, \cdot, A[\mathcal{D}^{\setminus i}]\,) \|_\infty \leq \beta$$

A small change to the training does not affect on the quality of the learned function much.

**Theorem (Stable algorithms generalize well [Bousquet *et al.*, 2002])**

*Let $A$ be a $\beta$-uniformly stable learning algorithm. For a training set $\mathcal{D}_m$ that consists of $m$ i.i.d. samples, denote by $f = A[\mathcal{D}_m]$ be the output of $A$ on $\mathcal{D}_m$. Let $\ell(y, \bar{y})$ be bounded by $M$.*

*Then, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + 2\beta + (4m\beta + M)\sqrt{\frac{\log(1/\delta)}{2m}}$$

Note: for the bound to be useful, the stability $\beta$ should decrease faster than $\sqrt{\frac{1}{m}}$ (but preferably least like $\frac{1}{m}$)

Reminder: stochastic gradient descent (SGD): minimize a function

$$f(\theta) = \frac{1}{m}\sum_{i=1}^{m} f(x_i, y_i; \theta)$$

**Theorem (Stability of Stochastic Gradient Descent [Hardt *et al.*, 2016])**

*Let $f(x, y; \cdot)$ be $\gamma$-smooth, convex and $L$-Lipschitz for every $(x, y)$. Suppose that we run SGD with step sizes $\alpha_t \leq 2/\gamma$ for $T$ steps. Then, SGD satisfies uniform stability with*

$$\beta \leq \frac{2L^2}{m}\sum_{t=1}^{T} \alpha_t.$$

*Let $f(x, y; \cdot)$ be $\gamma$-smooth and $L$-Lipschitz, but not necessarily convex. Assume we run SGD with monotonically non-increasing step sizes $\alpha_t \leq c/t$ for some $c$. Then, SGD satisfies uniform stability with*

$$\beta \leq \frac{1 + \frac{1}{\gamma c}}{m-1}(2cL^2)^{\frac{1}{\gamma c+1}} T^{\frac{\gamma c}{\gamma c+1}}.$$

# The Power of Compression

**Reminder:**

## Perceptron – Training

**input** training set $\mathcal{D} \subset \mathbb{R}^d \times \{-1, +1\}$
  initialize $w = (0, \dots, 0) \in \mathbb{R}^d$.
  **repeat**
    **for all** $(x, y) \in \mathcal{D}$: **do**
      compute $a := \langle w, x \rangle$  ('activation')
      **if** $ya \leq 0$ **then**
        $w \leftarrow w + yx$
      **end if**
    **end for**
  **until** $w$ wasn't updated for a complete pass over $\mathcal{D}$

Let's assume $\mathcal{D}$ is very large, so we don't need multiple passes.
Properties:

- sequential training, one pass over data
- only those examples matter, where perceptron made a mistake (only those affect $w$)

- Take training set as a sequence:

$$T = \big((x^1, y^1), (x^2, y^2), \dots, (x^n, y^n)\big)$$

- algorithm $A$ processes $T$ in order, producting output $f := A(T)$
- What if only a subset of examples influence the algorithm output?
- for increasing subsequence, $I \subset \{1, \dots, n\}$, with $|I| = l$, set

$$T_I = \big((x^{i_1}, y^{i_1}), (x^{i_2}, y^{i_2}), \dots, (x^{i_l}, y^{i_l})\big)$$

**Definition**

$I$ is a **compression set** for $T$, if $A(T) = A(T_I)$.

Example: $I = \{$set of examples where Perceptron made a mistake$\}$

## Definition (Compression scheme [Littlestone/Warmuth, 1986])

A learning algorithm $A$ is called **compression scheme**, if there is a pair of functions: $C$ (called compression function), and $L$ (called reconstruction function), such that:

- $C$ takes as input a finite dataset and outputs a subsequence of indices
- $L$ takes as input a finite dataset and outputs a predictor
- $A$ is the result of applying $L$ to the data selected by $C$

$$A = L(T_I) \text{ for } I = C(T)$$

**Examples:**

- $C$ selects half of the data from $T$ at random
- $C$ run a clustering algorithm on $T$ and returns the cluster centers as $I$

**Examples, where $A = L(T_I)$ equals $L(T)$:**

- Perceptron ($I =$ indices of examples where will be updated)
- SVMs ($I =$ set of support vectors)
- $k$-NN ($I =$ set of examples that support the decision boundaries)

$$\hat{\mathcal{R}}_I(h) = \frac{1}{|I|}\sum_{i\in I} \ell(\,y^i, h(x^i)\,) \quad \text{and} \quad \hat{\mathcal{R}}_{\neg I}(h) = \frac{1}{n-|I|}\sum_{i\notin I} \ell(\,y^i, h(x^i)\,)$$

**Theorem (Compression Bound [Littlestone/Warmuth, 1986; Graepel 2005] )**

*Let $A$ be a compression scheme with compression function $C$. Let the loss $\ell$ be bounded by $[0,1]$. Then, with probability at least $1-\delta$ over the random draw of $T$, we have that:*

*If $\hat{\mathcal{R}}_{\neg I}(A(T)) = 0$:*

$$\mathcal{R}(A(T)) \leq \frac{1}{n-l}\big(\,(l+1)\log n + \log\frac{1}{\delta}\big). \qquad \rightarrow O(\frac{1}{n})$$

*For general $\hat{\mathcal{R}}_{\neg I}(A(T))$:*

$$\mathcal{R}(A(T)) \leq \frac{n}{n-l}\hat{\mathcal{R}}_{\neg I}(A(T)) + \sqrt{\frac{(l+2)\log n + \log\frac{1}{\delta}}{2(n-l)}} \qquad \rightarrow O(\frac{1}{\sqrt{n}})$$

*where $I = C(T)$ and $l = |I|$.*