Trustworthy Machine Learning

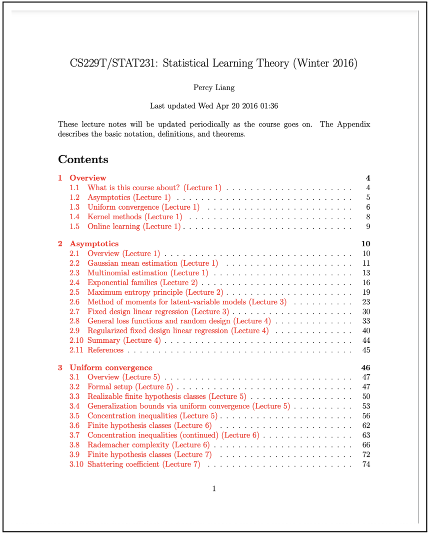
Beyond PAC Learning

Sangdon Park

POSTECH

October 8, 2024

Contents from



CS229T/STAT231: Statistical Learning Theory (Winter 2016)

Percy Liang

Last updated Wed Apr 20 2016 01:36

These lecture notes will be updated periodically as the course goes on. The Appendix describes the basic notation, definitions, and theorems.

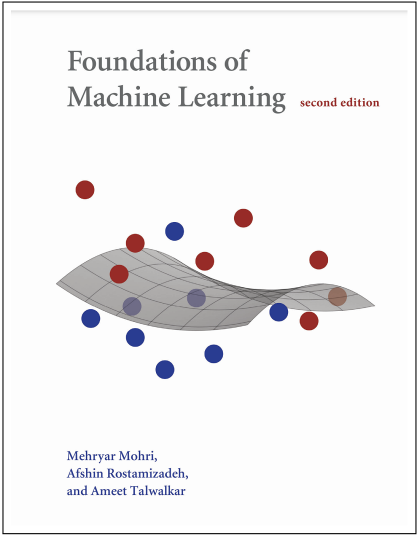
Contents

1 Overview 4
1.1 What is this course about? (Lecture 1) 4
1.2 Asymptotics (Lecture 1) 5
1.3 Uniform convergence (Lecture 1) 6
1.4 Kernel methods (Lecture 1) 8
1.5 Online learning (Lecture 1) 9

2 Asymptotics 10
2.1 Overview (Lecture 1) 10
2.2 Gaussian mean estimation (Lecture 1) 11
2.3 Multinomial estimation (Lecture 1) 13
2.4 Exponential families (Lecture 2) 16
2.5 Maximum entropy principle (Lecture 2) 19
2.6 Method of moments for latent-variable models (Lecture 3) 23
2.7 Fixed design linear regression (Lecture 3) 30
2.8 General loss functions and random design (Lecture 4) 33
2.9 Regularized fixed design linear regression (Lecture 4) 40
2.10 Summary (Lecture 4) 44
2.11 References 45

3 Uniform convergence 46
3.1 Overview (Lecture 5) 47
3.2 Formal setup (Lecture 5) 47
3.3 Realizable finite hypothesis classes (Lecture 5) 50
3.4 Generalization bounds via uniform convergence (Lecture 5) 53
3.5 Concentration inequalities (Lecture 5) 56
3.6 Finite hypothesis classes (Lecture 6) 62
3.7 Concentration inequalities (continued) (Lecture 6) 63
3.8 Rademacher complexity (Lecture 6) 66
3.9 Finite hypothesis classes (Lecture 7) 72
3.10 Shattering coefficient (Lecture 7) 74

1



Foundations of
Machine Learning second edition

Mehryar Mohri,
Afshin Rostamizadeh,
and Ameet Talwalkar

and various papers.

Is PAC Learning Okay?

Four Ingredients of Learning:

- Distribution $\mathcal{D}$
- Loss ℓ
- Hypothesis Space $\mathcal{H}$
- A Learning Algorithm $\mathcal{A}$

Problem?

Is PAC Learning Okay?

Four Ingredients of Learning:

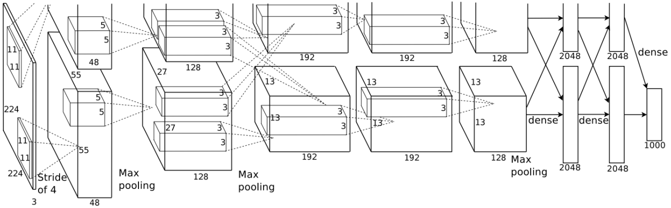
- Distribution $\mathcal{D}$
- Loss ℓ
- Hypothesis Space $\mathcal{H}$
- A Learning Algorithm $\mathcal{A}$

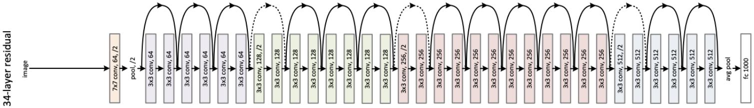
Problem?

The main assumption of PAC learning: $\mathcal{D}$ is separable by some $h^* \in \mathcal{H}$.

$\mathcal{D}$ Is Generally Not Separable

Usually we do not know a set of hypotheses $\mathcal{H}$ that has the true hypothesis h^*.





- What is the architecture of neural networks that perfectly classifies ImageNet?
- We mainly search for good hypothesis space $\mathcal{F}$ without any assumption on $\mathcal{D}$.

Contents

1. Concentration Inequalities

2. Generalization Bounds via Uniform Convergence

Contents

1. Concentration Inequalities

2. Generalization Bounds via Uniform Convergence

Why Concentration Inequalities?

- Understanding the expected loss is a key in statistical learning

$$\min_{f \in \mathcal{F}} \mathbb{E}\ell(x, y, f)$$

Why Concentration Inequalities?

- Understanding the expected loss is a key in statistical learning

$$\min_{f \in \mathcal{F}} \mathbb{E}\ell(x, y, f)$$

- Concentration inequalities
  - A concentration inequality provides a bound around an expected value.

Why Concentration Inequalities?

- Understanding the expected loss is a key in statistical learning

$$\min_{f \in \mathcal{F}} \mathbb{E}\ell(x, y, f)$$

- Concentration inequalities
  - A concentration inequality provides a bound around an expected value.
- An Example: Mean estimation
  - Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables with mean $\mu := \mathbb{E}[X_1]$
  - The empirical mean is defined as

$$\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

  - What is the relation between μ and $\hat{\mu}_n$?

Possible Argument 1

Consistency: Due to the law of large numbers,

$$\hat{\mu}_n - \mu \xrightarrow{P} 0$$

- $\xrightarrow{P}$: convergence “in probability”
- If we get more data, $\hat{\mu}_n$ reaches to μ

Possible Argument 1

Consistency: Due to the law of large numbers,

$$\hat{\mu}_n - \mu \xrightarrow{P} 0$$

- $\xrightarrow{P}$: convergence “in probability”
- If we get more data, $\hat{\mu}_n$ reaches to μ
- ✗ Asymptotic guarantee: it does not answer on the required number of samples to reach to the correct answer.

Possible Argument 2

Asymptotic normality: Assuming Var$(X_1) = \sigma^2$, due to the central limit theorem,

$$\sqrt{n}(\hat{\mu}_n - \mu) \overset{D}{\to} \mathcal{N}(0, \sigma^2)$$

- $\overset{D}{\to}$: convergence "in distribution"
- If we get more data, $\hat{\mu}_n$ reaches to μ, where the variance is decreasing at a rate of $1/n$.

Possible Argument 2

Asymptotic normality: Assuming Var$(X_1) = \sigma^2$, due to the central limit theorem,

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

- $\xrightarrow{D}$: convergence "in distribution"
- If we get more data, $\hat{\mu}_n$ reaches to μ, where the variance is decreasing at a rate of $1/n$.

✗ Asymptotic guarantee: it does not answer on the required number of samples to reach to the correct answer.

Possible Argument 3

Tail bound: we wish to have a statement as follows:

$$\mathbb{P}\left\{|\hat{\mu}_n - \mu| \geq \varepsilon\right\} \leq \mathsf{SomeFunctionOf}(n, \varepsilon) = \delta.$$

- ε: a desired error level
- $1 - \delta$: the confidence of the error statement

Possible Argument 3

Tail bound: we wish to have a statement as follows:

$$\mathbb{P}\left\{|\hat{\mu}_n - \mu| \geq \varepsilon\right\} \leq \mathsf{SomeFunctionOf}(n, \varepsilon) = \delta.$$

- ε: a desired error level
- $1 - \delta$: the confidence of the error statement

✓ "$\mathsf{SomeFunctionOf}(n, \varepsilon) = \delta$" provides the required number of samples to reach a desired level of error with a desired level of confidence.

Hoeffding's Inequality

Theorem

Let $X_1, \ldots, X_n$ *be independent random variables with* $X_i \in [a_i, b_i]$ *for all* $i \in \{1, \ldots, n\}$*. Then, for any* $\varepsilon > 0$*, the following inequality holds for* $S_n \coloneqq \sum_{i=1}^n X_i$*:*

$$\mathbb{P}\left\{\mathbb{E}\{S_n\} - S_n \geq \varepsilon\right\} \leq \exp\left\{\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

Hoeffding's Inequality

Theorem

Let $X_1, \ldots, X_n$ *be independent random variables with* $X_i \in [a_i, b_i]$ *for all* $i \in \{1, \ldots, n\}$*. Then, for any* $\varepsilon > 0$*, the following inequality holds for* $S_n \coloneqq \sum_{i=1}^n X_i$*:*

$$\mathbb{P}\left\{\mathbb{E}\{S_n\} - S_n \geq \varepsilon\right\} \leq \exp\left\{\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

- Why is it called a tail bound?

Hoeffding's Inequality

Theorem

Let $X_1, \ldots, X_n$ *be independent random variables with* $X_i \in [a_i, b_i]$ *for all* $i \in \{1, \ldots, n\}$. *Then, for any* $\varepsilon > 0$, *the following inequality holds for* $S_n := \sum_{i=1}^n X_i$*:*

$$\mathbb{P}\left\{\mathbb{E}\{S_n\} - S_n \geq \varepsilon\right\} \leq \exp\left\{\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

- Why is it called a tail bound?
- What's the effect of n? Suppose $a_i = 0$ and $b_i = 1$,

$$\mathbb{P}\left\{\mathbb{E}\left\{\frac{S_n}{n}\right\} - \frac{S_n}{n} \geq \varepsilon'\right\} \leq \exp\left\{-2n\varepsilon'^2\right\}$$

Hoeffding's Inequality

Theorem

Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [a_i, b_i]$ for all $i \in \{1, \ldots, n\}$. Then, for any $\varepsilon > 0$, the following inequality holds for $S_n := \sum_{i=1}^n X_i$:

$$\mathbb{P}\left\{\mathbb{E}\{S_n\} - S_n \geq \varepsilon\right\} \leq \exp\left\{\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

- Why is it called a tail bound?
- What's the effect of n? Suppose $a_i = 0$ and $b_i = 1$,

$$\mathbb{P}\left\{\mathbb{E}\left\{\frac{S_n}{n}\right\} - \frac{S_n}{n} \geq \varepsilon'\right\} \leq \exp\left\{-2n\varepsilon'^2\right\}$$

- $X_1, \ldots, X_n$ need not to follow the same distribution

Binomial Distribution Tail Bound

A special version of the Hoeffding's inequality.

Theorem

Let $X_1, \ldots, X_n$ *be i.i.d. random variables with* $X_i \in \{0, 1\}$ *and* $\mathbb{P}\{X_i = 1\} = p \in [0, 1]$ *for all* $i \in \{1, \ldots, n\}$*. Then, for any* $\varepsilon > 0$*, the following inequality holds for* $S_n = \sum_{i=1}^{n} X_i$*:*

$$\mathbb{P}\left\{p \leq \hat{p}\right\} \geq 1 - \delta,$$

where $F(k; n, p)$ *is the CDF of a binomial distribution with* n *trials and success probability* p *and* $\hat{p} \coloneqq \inf\left\{p' \in [0, 1] \mid F(S_n; n, p') \leq \delta\right\}$*.*

- p is what we want to estimate and $\hat{p}$ is the smallest upper bound of $\hat{p}$ “described” by observations S_n.

Binomial Distribution Tail Bound

A special version of the Hoeffding's inequality.

Theorem

Let $X_1, \ldots, X_n$ *be i.i.d. random variables with* $X_i \in \{0, 1\}$ *and* $\mathbb{P}\{X_i = 1\} = p \in [0, 1]$ *for all* $i \in \{1, \ldots, n\}$. *Then, for any* $\varepsilon > 0$, *the following inequality holds for* $S_n = \sum_{i=1}^{n} X_i$:

$$\mathbb{P}\left\{p \leq \hat{p}\right\} \geq 1 - \delta,$$

where $F(k; n, p)$ *is the CDF of a binomial distribution with* n *trials and success probability* p *and* $\hat{p} := \inf\left\{p' \in [0, 1] \mid F(S_n; n, p') \leq \delta\right\}$.

- p is what we want to estimate and $\hat{p}$ is the smallest upper bound of $\hat{p}$ “described” by observations S_n.
- This is from the Clopper-Pearson interval for estimating binomial confidence intervals.

Binomial Distribution Tail Bound

A special version of the Hoeffding's inequality.

Theorem

Let $X_1, \ldots, X_n$ *be i.i.d. random variables with* $X_i \in \{0, 1\}$ *and* $\mathbb{P}\{X_i = 1\} = p \in [0, 1]$ *for all* $i \in \{1, \ldots, n\}$*. Then, for any* $\varepsilon > 0$*, the following inequality holds for* $S_n = \sum_{i=1}^{n} X_i$*:*

$$\mathbb{P}\left\{p \leq \hat{p}\right\} \geq 1 - \delta,$$

where $F(k; n, p)$ *is the CDF of a binomial distribution with* n *trials and success probability* p *and* $\hat{p} := \inf\left\{p' \in [0, 1] \mid F(S_n; n, p') \leq \delta\right\}$.

- p is what we want to estimate and $\hat{p}$ is the smallest upper bound of $\hat{p}$ “described” by observations S_n.
- This is from the Clopper-Pearson interval for estimating binomial confidence intervals.
- From the Hoeffding's inequality, $\mathbb{P}\left\{\frac{S_n}{n} - p > \varepsilon\right\} \leq \exp\left\{-2n\varepsilon^2\right\}$

Binomial Distribution Tail Bound

A special version of the Hoeffding's inequality.

Theorem

Let $X_1, \ldots, X_n$ *be i.i.d. random variables with* $X_i \in \{0,1\}$ *and* $\mathbb{P}\{X_i = 1\} = p \in [0,1]$ *for all* $i \in \{1, \ldots, n\}$. *Then, for any* $\varepsilon > 0$, *the following inequality holds for* $S_n = \sum_{i=1}^{n} X_i$*:*

$$\mathbb{P}\left\{p \leq \hat{p}\right\} \geq 1 - \delta,$$

where $F(k; n, p)$ *is the CDF of a binomial distribution with* n *trials and success probability* p *and* $\hat{p} := \inf\left\{p' \in [0,1] \mid F(S_n; n, p') \leq \delta\right\}$.

- p is what we want to estimate and $\hat{p}$ is the smallest upper bound of $\hat{p}$ “described” by observations S_n.
- This is from the Clopper-Pearson interval for estimating binomial confidence intervals.
- From the Hoeffding's inequality, $\mathbb{P}\left\{\frac{S_n}{n} - p > \varepsilon\right\} \leq \exp\left\{-2n\varepsilon^2\right\}$
- A tighter bound than the Hoeffding's inequality.

McDiarmid's Inequality

A generalized version of the Heoffding's inequality.

Theorem

Let $(X_1, \ldots, X_n) \in \mathcal{X}^n$ *be a list of* $n \geq 1$ *independent random variables and assume that there exist* $c_1, \ldots, c_n > 0$ *such that* $f : \mathcal{X}^n \to \mathbb{R}$ *satisfies the following conditions:*

$$\left| f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_m) \right| \leq c_i,$$

for all $i \in \{1, \ldots, n\}$ *and any* $x_1, \ldots, x_n, x_i \in \mathcal{X}$. *Let* $f(S)$ *denote* $f(X_1, \ldots, X_n)$, *then, for all* $\varepsilon > 0$, *the following inequality holds:*

$$\mathbb{P}\left\{ f(S) - \mathbb{E}\{f(S)\} \geq \varepsilon \right\} \leq \exp\left\{ \frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2} \right\}.$$

McDiarmid's Inequality

A generalized version of the Heoffding's inequality.

Theorem

Let $(X_1, \ldots, X_n) \in \mathcal{X}^n$ *be a list of* $n \geq 1$ *independent random variables and assume that there exist* $c_1, \ldots, c_n > 0$ *such that* $f : \mathcal{X}^n \to \mathbb{R}$ *satisfies the following conditions:*

$$\left| f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_m) \right| \leq c_i,$$

for all $i \in \{1, \ldots, n\}$ *and any* $x_1, \ldots, x_n, x_i \in \mathcal{X}$. *Let* $f(S)$ *denote* $f(X_1, \ldots, X_n)$, *then, for all* $\varepsilon > 0$, *the following inequality holds:*

$$\mathbb{P}\left\{ f(S) - \mathbb{E}\{f(S)\} \geq \varepsilon \right\} \leq \exp\left\{ \frac{-2\varepsilon^2}{\sum_{i=1}^{n} c_i^2} \right\}.$$

- Useful concentration inequality for a more complex function than a mean value under the "bounded difference".

McDiarmid's Inequality

A generalized version of the Heoffding's inequality.

Theorem

Let $(X_1, \ldots, X_n) \in \mathcal{X}^n$ *be a list of* $n \geq 1$ *independent random variables and assume that there exist* $c_1, \ldots, c_n > 0$ *such that* $f : \mathcal{X}^n \to \mathbb{R}$ *satisfies the following conditions:*

$$\left| f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_m) \right| \leq c_i,$$

for all $i \in \{1, \ldots, n\}$ *and any* $x_1, \ldots, x_n, x_i \in \mathcal{X}$*. Let* $f(S)$ *denote* $f(X_1, \ldots, X_n)$*, then, for all* $\varepsilon > 0$*, the following inequality holds:*

$$\mathbb{P}\left\{ f(S) - \mathbb{E}\{f(S)\} \geq \varepsilon \right\} \leq \exp\left\{ \frac{-2\varepsilon^2}{\sum_{i=1}^{n} c_i^2} \right\}.$$

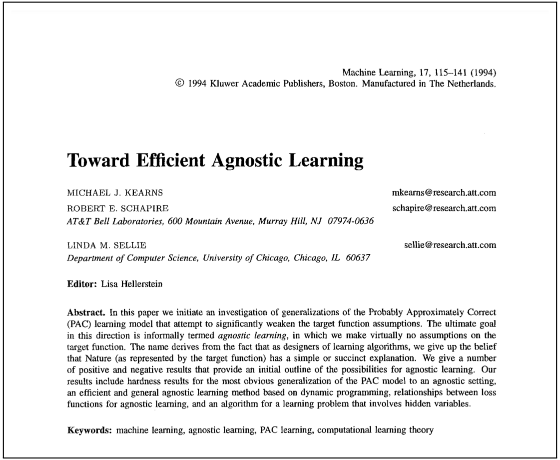
- Useful concentration inequality for a more complex function than a mean value under the "bounded difference".
- The main concentration inequality for a generalization bound.

Contents

1. Concentration Inequalities

2. Generalization Bounds via Uniform Convergence

Agnostic PAC Learning Algorithm



Machine Learning, 17, 115–141 (1994)
© 1994 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Toward Efficient Agnostic Learning

MICHAEL J. KEARNS mkearns@research.att.com
ROBERT E. SCHAPIRE schapire@research.att.com
AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974-0636

LINDA M. SELLIE sellie@research.att.com
Department of Computer Science, University of Chicago, Chicago, IL 60637

Editor: Lisa Hellerstein

Abstract. In this paper we initiate an investigation of generalizations of the Probably Approximately Correct (PAC) learning model that attempt to significantly weaken the target function assumptions. The ultimate goal in this direction is informally termed *agnostic learning*, in which we make virtually no assumptions on the target function. The name derives from the fact that as designers of learning algorithms, we give up the belief that Nature (as represented by the target function) has a simple or succinct explanation. We give a number of positive and negative results that provide an initial outline of the possibilities for agnostic learning. Our results include hardness results for the most obvious generalization of the PAC model to an agnostic setting, an efficient and general agnostic learning method based on dynamic programming, relationships between loss functions for agnostic learning, and an algorithm for a learning problem that involves hidden variables.

Keywords: machine learning, agnostic learning, PAC learning, computational learning theory

- For the smooth transition from PAC learning, I will introduce agnostic PAC learning.
- Later, we will mainly use languages from statistical learning theory.

Agnostic PAC Learning Algorithm

Definition (simplified definition)

An algorithm $\mathcal{A}$ is an agnostic PAC-learning algorithm for $\mathcal{H}$ if for any $\varepsilon > 0$, $\delta > 0$, ~~$h^* \in \mathcal{H}$~~, and $\mathcal{D}$ ~~separable by h^*~~, and for some minimum sample size n' (which depends on $\varepsilon, \delta, \mathcal{D}$), the following holds with any sample size $n \geq n'$:

$$\mathbb{P}\left\{L(\mathcal{A}(\mathcal{S})) - \min_{h \in \mathcal{H}} L(h) \leq \varepsilon\right\} \geq 1 - \delta,$$

where $\mathcal{S} := ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n$.

Agnostic PAC Learning Algorithm

Definition (simplified definition)

An algorithm $\mathcal{A}$ is an agnostic PAC-learning algorithm for $\mathcal{H}$ if for any $\varepsilon > 0$, $\delta > 0$, ~~$h^* \in \mathcal{H}$~~, and $\mathcal{D}$ ~~separable by h^*~~, and for some minimum sample size n' (which depends on $\varepsilon, \delta, \mathcal{D}$), the following holds with any sample size $n \geq n'$:

$$\mathbb{P}\left\{ L(\mathcal{A}(\mathcal{S})) - \min_{h \in \mathcal{H}} L(h) \leq \varepsilon \right\} \geq 1 - \delta,$$

where $\mathcal{S} := ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n$.

- $\arg\min_{h \in \mathcal{H}} L(h)$: the best hypothesis

Agnostic PAC Learning Algorithm

Definition (simplified definition)

An algorithm $\mathcal{A}$ is an agnostic PAC-learning algorithm for $\mathcal{H}$ if for any $\varepsilon > 0$, $\delta > 0$, ~~$h^* \in \mathcal{H}$~~, and $\mathcal{D}$ ~~separable by h^*~~, and for some minimum sample size n' (which depends on $\varepsilon, \delta, \mathcal{D}$), the following holds with any sample size $n \geq n'$:

$$\mathbb{P}\left\{L(\mathcal{A}(\mathcal{S})) - \min_{h \in \mathcal{H}} L(h) \leq \varepsilon\right\} \geq 1 - \delta,$$

where $\mathcal{S} := ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n$.

- $\arg\min_{h \in \mathcal{H}} L(h)$: the best hypothesis
- Vapnik notations on generalization bounds are more widely used.

Agnostic PAC Learning Algorithm

Definition (simplified definition)

An algorithm $\mathcal{A}$ is an agnostic PAC-learning algorithm for $\mathcal{H}$ if for any $\varepsilon > 0$, $\delta > 0$, ~~$h^* \in \mathcal{H}$~~, and $\mathcal{D}$ ~~separable by h^*~~, and for some minimum sample size n' (which depends on $\varepsilon, \delta, \mathcal{D}$), the following holds with any sample size $n \geq n'$:

$$\mathbb{P}\left\{L(\mathcal{A}(\mathcal{S})) - \min_{h\in\mathcal{H}} L(h) \leq \varepsilon\right\} \geq 1 - \delta,$$

where $\mathcal{S} := ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n$.

- $\arg\min_{h\in\mathcal{H}} L(h)$: the best hypothesis
- Vapnik notations on generalization bounds are more widely used.
- Please check out the original agnostic PAC learning definition.

Definitions

Definition (best hypothesis)

$$h^* := \arg\min_{h \in \mathcal{H}} L(h)$$

Definition (empirical risk minimizer)

$$\hat{h} := \arg\min_{h \in \mathcal{H}} \hat{L}(h)$$

Goal: Find Generalization Bounds

An Interesting Quantity:

$$L(h) - \hat{L}(h)$$

Goal: Find Generalization Bounds

An Interesting Quantity:

$$L(h) - \hat{L}(h)$$

- Why?

Goal: Find Generalization Bounds

An Interesting Quantity:

$$L(h) - \hat{L}(h)$$

- Why?
  - Generally the bound of the following is called a "generalization bound":

$$L(\hat{h}) - L(h^*)$$

Goal: Find Generalization Bounds

An Interesting Quantity:

$$L(h) - \hat{L}(h)$$

- Why?
  - Generally the bound of the following is called a “generalization bound”:

    $$L(\hat{h}) - L(h^*)$$

  - It is bounded as follows (will see later):

    $$\mathbb{P}\left\{L(\hat{h}) - L(h^*) \geq \varepsilon\right\} \leq \mathbb{P}\left\{\sup_{h \in \mathcal{H}} \left|L(h) - \hat{L}(h)\right| \geq \frac{\varepsilon}{2}\right\}$$

Goal: Find Generalization Bounds

An Interesting Quantity:

$$L(h) - \hat{L}(h)$$

- Why?
  - Generally the bound of the following is called a "generalization bound":
    $$L(\hat{h}) - L(h^*)$$
  - It is bounded as follows (will see later):
    $$\mathbb{P}\left\{L(\hat{h}) - L(h^*) \geq \varepsilon\right\} \leq \mathbb{P}\left\{\sup_{h\in\mathcal{H}} \left|L(h) - \hat{L}(h)\right| \geq \frac{\varepsilon}{2}\right\}$$
  - We also call a bound of $L(h) - \hat{L}(h)$ a generalization bound — The term "generalization bound" is used in multiple ways.

Goal: Find Generalization Bounds

An Interesting Quantity:

$$L(h) - \hat{L}(h)$$

- Why?
  - Generally the bound of the following is called a "generalization bound":
    $$L(\hat{h}) - L(h^*)$$
  - It is bounded as follows (will see later):
    $$\mathbb{P}\left\{L(\hat{h}) - L(h^*) \geq \varepsilon\right\} \leq \mathbb{P}\left\{\sup_{h \in \mathcal{H}} \left|L(h) - \hat{L}(h)\right| \geq \frac{\varepsilon}{2}\right\}$$
  - We also call a bound of $L(h) - \hat{L}(h)$ a generalization bound — The term "generalization bound" is used in multiple ways.
  - I'll introduce the philosophy on "From Theory to Algorithm", where $L(h) - \hat{L}(h)$ is more directly related.

Goal: Find Generalization Bounds

An Interesting Quantity:

$$L(h) - \hat{L}(h)$$

- Why?
  - Generally the bound of the following is called a "generalization bound":

    $$L(\hat{h}) - L(h^*)$$

  - It is bounded as follows (will see later):

    $$\mathbb{P}\left\{L(\hat{h}) - L(h^*) \geq \varepsilon\right\} \leq \mathbb{P}\left\{\sup_{h\in\mathcal{H}} \left|L(h) - \hat{L}(h)\right| \geq \frac{\varepsilon}{2}\right\}$$

  - We also call a bound of $L(h) - \hat{L}(h)$ a generalization bound — The term "generalization bound" is used in multiple ways.
  - I'll introduce the philosophy on "From Theory to Algorithm", where $L(h) - \hat{L}(h)$ is more directly related.
- The generalization bound will depend on the complexity of $\mathcal{H}$, which is harder to measure if $\mathcal{H}$ is an infinite set (than the finite case).

Example: A Learning Bound for a Finite Hypothesis Set I

Setup:

- $\mathcal{H}$: a *finite* set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$
- $\mathcal{D}$: any distribution — no assumption!
- $\mathcal{S}$: labeled examples
- $\mathcal{A}$: any algorithm — no assumption to use!

Example: A Learning Bound for a Finite Hypothesis Set II

Theorem

Let $\ell(\cdot) \in [0,1]$. For any $\varepsilon > 0$, $\delta > 0$, and $\mathcal{D}$, we have

$$\forall h \in \mathcal{H}, \quad L(h) \leq \hat{L}(h) + \sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{2n}}$$

with probability at least $1 - \delta$.

- We have logarithmic dependence on $|\mathcal{H}|$ and $1 \ /\delta$ – this bound is not “sensitive” to them.
- This is a uniform convergence bound: “$\forall h$” is inside of the probability.

$$(\text{✗}) \quad \forall h \in \mathcal{H}, \quad \mathbb{P}\left\{ L(h) \leq \hat{L}(h) + \sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{2n}} \right\} \geq 1 - \delta$$

- Conservative (=data-independent): even though some h is “bad”, we need the convergence guarantee.

Example: A Learning Bound for a Finite Hypothesis Set III

Proof Sketch:

$$\begin{aligned}
\mathbb{P}\left\{\exists h \in \mathcal{H},\ L(h) - \hat{L}(h) > \varepsilon\right\} &= \mathbb{P}\left\{\bigvee_{h\in\mathcal{H}} L(h) - \hat{L}(h) > \varepsilon\right\} \\
&\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left\{L(h) - \hat{L}(h) > \varepsilon\right\} \qquad (1) \\
&\leq |\mathcal{H}| \exp\left\{-2n\varepsilon^2\right\} \qquad (2)
\end{aligned}$$

- (1): Uniform convergence via the union bound
- (2): A “point” convergence via the Hoeffding’s inequality

From the Previous Learning Bound to an Algorithm

Learning bound:

$$\forall h \in \mathcal{H}, \quad L(h) \leq \hat{L}(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{2n}}$$

- This bound holds for any h, including $\mathcal{A}(\mathcal{S})$ for any $\mathcal{A}$.
- If $\mathcal{A}$ minimizes the upper bound, $\mathcal{A}(\mathcal{S})$ minimizes the expected error.
- One such algorithm is the empirical risk minimizer!

Algorithm: Given $\mathcal{H}$ and labeled examples $\mathcal{S}$,

$$\min_{h \in \mathcal{H}} \hat{L}(h)$$

- As the learning bound holds for any h, our algorithm can be more general, *e.g.*, a regularized ERM.
- For this distribution-free setup, the sample complexity is not very meaningful.

ERM is Agnostic-PAC

Example: Under Finite Hypotheses

Why?

$$\begin{aligned} L(\mathcal{A}(\mathcal{S})) - L(h^*) &= \left\{ L(\mathcal{A}(\mathcal{S})) - \hat{L}(\mathcal{A}(\mathcal{S})) \right\} + \left\{ \hat{L}(\mathcal{A}(\mathcal{S})) - \hat{L}(h^*) \right\} + \left\{ \hat{L}(h^*) - L(h^*) \right\} \\ &\leq \underbrace{\left\{ L(\mathcal{A}(\mathcal{S})) - \hat{L}(\mathcal{A}(\mathcal{S})) \right\}}_{\text{uniform convergence}} + \underbrace{\left\{ \hat{L}(h^*) - L(h^*) \right\}}_{\text{concentration inequality}} \\ &\leq \sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta_1}}{2n}} + \sqrt{\frac{\ln \frac{1}{\delta_2}}{2n}} \end{aligned}$$

with probability at least $1 - (\delta_1 + \delta_2)$.

Separable $\mathcal{D}$ v.s. $\mathcal{D}$

A bound under the separability assumption

$$L(\mathcal{A}(\mathcal{S})) \leq \frac{1}{n}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)$$

A bound without separability

$$\forall h \in \mathcal{H}, \quad L(h) \leq \hat{L}(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

- This is not an apple-and-apple comparison, but let's try to compare.

Separable $\mathcal{D}$ v.s. $\mathcal{D}$

A bound under the separability assumption

$$L(\mathcal{A}(\mathcal{S})) \leq \frac{1}{n}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)$$

A bound without separability

$$\forall h \in \mathcal{H}, \quad L(h) \leq \hat{L}(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

- This is not an apple-and-apple comparison, but let's try to compare.
- A bound that exploits more information is tighter.
  - A distribution is separable ($\approx$ no noise).

Separable $\mathcal{D}$ v.s. $\mathcal{D}$

A bound under the separability assumption

$$L(\mathcal{A}(\mathcal{S})) \leq \frac{1}{n}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)$$

A bound without separability

$$\forall h \in \mathcal{H}, \quad L(h) \leq \hat{L}(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

- This is not an apple-and-apple comparison, but let's try to compare.
- A bound that exploits more information is tighter.
  - A distribution is separable ($\approx$ no noise).
- Under the additional information, we can learn faster (*i.e.*, $\frac{1}{n}$ vs $\frac{1}{\sqrt{n}}$).

A More General Bound

- In general, $\mathcal{H}$ is infinite (*e.g.*, a set of neural networks)

A More General Bound

- In general, $\mathcal{H}$ is infinite (*e.g.*, a set of neural networks)
- The related bound is one of the key results of statistical learning theory (via Vapnik)

A More General Bound

- In general, $\mathcal{H}$ is infinite (*e.g.*, a set of neural networks)
- The related bound is one of the key results of statistical learning theory (via Vapnik)
- Related keywords include
    - McDiarmid's Inequality
    - Rademacher Complexity
    - VC dimension
    - A learning bound for SVM

A More General Bound

- In general, $\mathcal{H}$ is infinite (*e.g.*, a set of neural networks)
- The related bound is one of the key results of statistical learning theory (via Vapnik)
- Related keywords include
  - McDiarmid's Inequality
  - Rademacher Complexity
  - VC dimension
  - A learning bound for SVM
- Caution: this "data-independent" bound cannot not explain the learnability of deep networks!

Rademacher Complexity

A way to measure the complexity of $\mathcal{H}$ (when $\mathcal{H}$ is infinite)!

Rademacher Complexity

A way to measure the complexity of $\mathcal{H}$ (when $\mathcal{H}$ is infinite)!

Definition

Let $\mathcal{F}$ be a set of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ (*e.g.*, $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$). The Rademacher complexity of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right\},$$

where $Z_1, \ldots, Z_n$ are drawn i.i.d. from a distribution and $\sigma_1, \ldots, \sigma_n$ are drawn i.i.d. from the uniform distribution over $\{-1, +1\}$ (a.k.a. Rademacher variables).

Rademacher Complexity

A way to measure the complexity of $\mathcal{H}$ (when $\mathcal{H}$ is infinite)!

Definition

Let $\mathcal{F}$ be a set of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ (*e.g.*, $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$). The Rademacher complexity of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}\left\{\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i)\right\},$$

where $Z_1, \ldots, Z_n$ are drawn i.i.d. from a distribution and $\sigma_1, \ldots, \sigma_n$ are drawn i.i.d. from the uniform distribution over $\{-1, +1\}$ (a.k.a. Rademacher variables).

- Previously, "concentration inequalities" + "union bound" provides a generalization bound.

Rademacher Complexity

A way to measure the complexity of $\mathcal{H}$ (when $\mathcal{H}$ is infinite)!

Definition

Let $\mathcal{F}$ be a set of real-valued functions $f : \mathcal{Z} \to \mathbb{R}$ (*e.g.*, $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$). The Rademacher complexity of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}\left\{\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i)\right\},$$

where $Z_1, \dots, Z_n$ are drawn i.i.d. from a distribution and $\sigma_1, \dots, \sigma_n$ are drawn i.i.d. from the uniform distribution over $\{-1, +1\}$ (a.k.a. Rademacher variables).

- Previously, "concentration inequalities" + "union bound" provides a generalization bound.
- This term will be upper-bounded by a term with "VC dimension" later.

Rademacher Complexity: Interpretation

$$R_n(\mathcal{F}) := \mathbb{E}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right\}$$

- This term considers an "imaginary binary classification" problem with randomly labeled examples (Z_i, σ_i).
    - If $\sigma_i = \mathrm{sign}(f(Z_i))$, f is correct on (Z_i, σ_i).
    - Solving $\sup$ = finding a "best" binary classifier.
    - Fix n and $\mathcal{F}$ → draw Z_i and σ_i → find f.

Rademacher Complexity: Interpretation

$$R_n(\mathcal{F}) := \mathbb{E}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right\}$$

- This term considers an “imaginary binary classification” problem with randomly labeled examples (Z_i, σ_i).
  - If $\sigma_i = \mathrm{sign}(f(Z_i))$, f is correct on (Z_i, σ_i).
  - Solving $\sup$ = finding a “best” binary classifier.
  - Fix n and $\mathcal{F}$ → draw Z_i and σ_i → find f.
- $R_n(\mathcal{F})$ captures how well the “best classifier” from $\mathcal{F}$ can align with random labels.
  - Large $R_n(\mathcal{F})$ means that there is some $f \in \mathcal{F}$, “flexible” enough to learn randomly labeled examples.
  - *e.g.*, linear functions v.s. neural networks

Generalization Bound via Rademacher Complexity

Theorem

Let $\mathcal{F} := \{z \mapsto \ell(z,h) \mid h \in \mathcal{H}\}$ *and* $\ell(\cdot) \in [0,1]$*. For all* $h \in \mathcal{H}$*,*

$$L(h) \leq \hat{L}(h) + 2R_n(\mathcal{F}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

with probability at least $1 - \delta$*.*

- $f \in \mathcal{F}$ is a composition of h and ℓ.

Proof Sketch: A Bird's-eye View

1. Define a random variable G_n
    - $G_n := \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$
    - A maximum difference between the expected and empirical error (*i.e.*, the worse case $= \sup$).
    - The bound of this term is a generalization bound.

Proof Sketch: A Bird's-eye View

1. Define a random variable G_n
    - $G_n := \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$
    - A maximum difference between the expected and empirical error (*i.e.*, the worse case $= \sup$).
    - The bound of this term is a generalization bound.
2. Show that G_n concentrates to $\mathbb{E}\{G_n\}$.
    - We will use the McDiarmid's inequality.

Proof Sketch: A Bird's-eye View

1. Define a random variable G_n
   - $G_n := \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$
   - A maximum difference between the expected and empirical error (*i.e.*, the worse case $= \sup$).
   - The bound of this term is a generalization bound.
2. Show that G_n concentrates to $\mathbb{E}\{G_n\}$.
   - We will use the McDiarmid's inequality.
3. Use a technique called "symmetrization" to bound $\mathbb{E}\{G_n\}$ using the Rademacher complexity.

Proof Sketch

1. Setup

Define an interesting quantity to us!

- Consider the maximum difference between $L(h)$ and $\hat{L}(h)$.

$$G_n := \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$$

  - G_n is a random variable that depends on $Z_1, \ldots, Z_n$.

Proof Sketch

1. Setup

Define an interesting quantity to us!

- Consider the maximum difference between $L(h)$ and $\hat{L}(h)$.

$$G_n := \sup_{h \in \mathcal{H}} L(h) - \hat{L}(h)$$

  - G_n is a random variable that depends on $Z_1, \ldots, Z_n$.
- We will consider the following tail bound:

$$\mathbb{P}\left\{G_n \geq \varepsilon\right\}.$$

  - What should we do?

Proof Sketch I

2. Concentration

Derive a tail bound via a concentration inequality!

- Let g be the deterministic function such that $G_n = g(Z_1, \ldots, Z_n)$.
- Then, the following holds:

$$\left|g(Z_1, \ldots, Z_i, \ldots, Z_n) - g(Z_1, \ldots, Z_i', \ldots, Z_n)\right| \le \frac{1}{n}.$$

- Why?
  - Recall $\hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, h)$.
  - Recall $\ell(\cdot) \in [0, 1]$.
  - We have

$$\left| \underbrace{\sup_{h \in \mathcal{H}} \left[L(h) - \hat{L}(h)\right]}_{g(Z_1, \ldots, Z_i, \ldots, Z_n)} - \underbrace{\sup_{h \in \mathcal{H}} \left[L(h) - \hat{L}(h) + \frac{1}{n}\left(\ell(Z_i, h) - \ell(Z_i', h)\right)\right]}_{g(Z_1, \ldots, Z_i', \ldots, Z_n)} \right| \le \frac{1}{n}.$$

Proof Sketch II

2. Concentration

- Apply the McDiarmid's inequality:

$$\mathbb{P}\left\{G_n \geq \mathbb{E}\{G_n\} + \varepsilon'\right\} \leq \exp\left(-2n\varepsilon'^2\right).$$

  - g is a non-trivial function, including $\sup$ over $h \in \mathcal{H}$; thus, we cannot use the usual concentration inequality (*e.g.*, the Hoeffding's inequality).
  - But, we can still use the McDiarmid's inequality due to the bounded difference.
  - We can find our generalization bound if we can bound $\mathbb{E}\{G_n\}$. But how?
  - Note that $\mathbb{E}\{G_n\}$ is related to the complexity of $\mathcal{F}$ (will see soon).

Proof Sketch I

3. Symmetrization

Bound $\mathbb{E}\{G_n\}$!

- $\mathbb{E}\{G_n\}$ is not easy to analysis as it depends on $L(h)$, an expectation of an unknown distribution $\mathcal{D}$.
- We will replace this to depend on $\mathcal{D}$ only through samples $Z_1, \ldots, Z_n$.
- The key idea of "symmetrization" is to introduce "ghost" samples $Z_1', \ldots, Z_n'$, drawn i.i.d. from $\mathcal{D}$ to rewrite $\mathbb{E}\{G_n\}$.
  - Let $\hat{L}'(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(Z_i', h)$.
  - Rewrite $L(h)$ in terms of the ghost samples, *i.e.*,

$$\mathbb{E}\{G_n\} = \mathbb{E}\left\{\sup_{h\in\mathcal{H}} L(h) - \hat{L}(h)\right\} = \mathbb{E}\left\{\sup_{h\in\mathcal{H}} \mathbb{E}\{\hat{L}'(h)\} - \hat{L}(h)\right\}$$

Proof Sketch II

3. Symmetrization

- Simplify and bound this rewritten $\mathbb{E}\{G_n\}$:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{Z}}\{G_n\} &= \mathbb{E}_{\mathcal{Z}}\left\{\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{Z}'}\{\hat{L}'(h)\} - \hat{L}(h)\right\} \\
&= \mathbb{E}_{\mathcal{Z}}\left\{\sup_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{Z}'}\left\{\hat{L}'(h) - \hat{L}(h)\right\}\right\} \\
&\le \mathbb{E}_{\mathcal{Z}}\left\{\mathbb{E}_{\mathcal{Z}'}\left\{\sup_{h\in\mathcal{H}} \hat{L}'(h) - \hat{L}(h)\right\}\right\} \\
&= \mathbb{E}_{\mathcal{Z},\mathcal{Z}'}\left\{\sup_{h\in\mathcal{H}} \hat{L}'(h) - \hat{L}(h)\right\} \\
&= \mathbb{E}_{\mathcal{Z},\mathcal{Z}'}\left\{\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\big(\ell(Z_i', h) - \ell(Z_i, h)\big)\right\}
\end{aligned}
$$

where $\mathcal{Z} := \{Z_1, \ldots, Z_n\}$ and $\mathcal{Z}' := \{Z_1', \ldots, Z_n'\}$.

Proof Sketch III

3. Symmetrization

- Remove the dependence on the ghost samples.
  - Introduce the i.i.d. Rademacher variables $\sigma_1, \ldots, \sigma_n$, where σ_i is uniform over $\{-1, 1\}$.
  - Observe that $\ell(Z_i', h) - \ell(Z_i, h)$ is symmetric around 0.
  - Thus, we have

$$
\begin{aligned}
\mathbb{E}\{G_n\} &\leq \mathbb{E}\left\{\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} (\ell(Z_i', h) - \ell(Z_i, h))\right\} \\
&= \mathbb{E}\left\{\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \left(\ell(Z_i', h) - \ell(Z_i, h)\right)\right\} \\
&\leq \mathbb{E}\left\{\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(Z_i', h) + \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} (-\sigma_i)\ell(Z_i, h)\right\} \\
&= 2\mathbb{E}\left\{\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(Z_i, h)\right\} = 2R_n(\mathcal{F})
\end{aligned}
$$

Proof Sketch

4. Combine

- From concentration, we have

$$\mathbb{P}\left\{G_n \geq \mathbb{E}\{G_n\} + \varepsilon'\right\} \leq \exp\left(-2n\varepsilon'^2\right).$$

- From symmetrization, we have

$$\mathbb{E}\{G_n\} \leq 2R_n(\mathcal{F}).$$

Proof Sketch

4. Combine

- From concentration, we have

$$\mathbb{P}\left\{G_n \geq \mathbb{E}\{G_n\} + \varepsilon'\right\} \leq \exp\left(-2n\varepsilon'^2\right).$$

- From symmetrization, we have

$$\mathbb{E}\{G_n\} \leq 2R_n(\mathcal{F}).$$

- Our goal is to bound the following tail probability:

$$\begin{aligned}\mathbb{P}\{G_n \geq \varepsilon\} &\leq \exp\left(-2n\left(\varepsilon - \mathbb{E}\{G_n\}\right)^2\right) \\ &\leq \exp\left(-2n\left(\varepsilon - 2R_n(\mathcal{F})\right)^2\right)\end{aligned}$$

Proof Sketch

4. Combine

- From concentration, we have

$$\mathbb{P}\left\{G_n \geq \mathbb{E}\{G_n\} + \varepsilon'\right\} \leq \exp\left(-2n\varepsilon'^2\right).$$

- From symmetrization, we have

$$\mathbb{E}\{G_n\} \leq 2R_n(\mathcal{F}).$$

- Our goal is to bound the following tail probability:

$$\begin{aligned}\mathbb{P}\{G_n \geq \varepsilon\} &\leq \exp\left(-2n\left(\varepsilon - \mathbb{E}\{G_n\}\right)^2\right)\\ &\leq \exp\left(-2n\left(\varepsilon - 2R_n(\mathcal{F})\right)^2\right)\end{aligned}$$

- This shows the claim, as

$$\delta = \exp\left(-2n\left(\varepsilon - 2R_n(\mathcal{F})\right)^2\right) \quad \Rightarrow \quad \varepsilon = 2R_n(\mathcal{F}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}.$$

Connection to the VC Generalization Bound

$$R_n(\mathcal{F}) \leq \sqrt{\frac{2\mathsf{VC}(\mathcal{H})(\ln n + 1)}{n}}$$

- $\mathsf{VC}(\mathcal{H})$: VC dimension of $\mathcal{H}$
- Related concepts:
  - Empirical Rademacher Complexity
  - A shattering coefficient or growth function
  - Sauer's lemma

Application: Support Vector Machine (SVM)

Setup:

- $\mathcal{X} \in \mathbb{R}^d$: example space

Application: Support Vector Machine (SVM)

Setup:

- $\mathcal{X} \in \mathbb{R}^d$: example space
- $\mathcal{Y} := \{-1, 1\}$: binary label space

Application: Support Vector Machine (SVM)

Setup:

- $\mathcal{X} \in \mathbb{R}^d$: example space
- $\mathcal{Y} := \{-1, 1\}$: binary label space
- $\mathcal{H}$: a set of linear functions (without a bias term for simplicity), *i.e.*,

$$\mathcal{H} := \{x \mapsto w \cdot x \mid w \in \mathbb{R}^d,\ \|w\|_2 \leq 1\}$$

or equivalently $\mathcal{H} := \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 1\}$.

Application: Support Vector Machine (SVM)

Setup:

- $\mathcal{X} \in \mathbb{R}^d$: example space
- $\mathcal{Y} := \{-1, 1\}$: binary label space
- $\mathcal{H}$: a set of linear functions (without a bias term for simplicity), *i.e.*,

$$\mathcal{H} := \{x \mapsto w \cdot x \mid w \in \mathbb{R}^d,\ \|w\|_2 \leq 1\}$$

  or equivalently $\mathcal{H} := \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 1\}$.
- ℓ_γ: margin loss

$$\ell_\gamma(v) := \min\left\{1, \max\left\{0, 1 - \frac{v}{\gamma}\right\}\right\},$$

Application: Support Vector Machine (SVM)

Setup:

- $\mathcal{X} \in \mathbb{R}^d$: example space
- $\mathcal{Y} := \{-1, 1\}$: binary label space
- $\mathcal{H}$: a set of linear functions (without a bias term for simplicity), *i.e.*,

$$\mathcal{H} := \{x \mapsto w \cdot x \mid w \in \mathbb{R}^d,\ \|w\|_2 \leq 1\}$$

  or equivalently $\mathcal{H} := \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 1\}$.

- ℓ_γ: margin loss

$$\ell_\gamma(v) := \min\left\{1, \max\left\{0, 1 - \frac{v}{\gamma}\right\}\right\},$$

- $L_\gamma / \hat{L}_\gamma$: the expected/empirical margin loss

$$L_\gamma(w) := \mathbb{E}\left\{\ell_\gamma(y(w \cdot x))\right\} \quad \text{and} \quad \hat{L}_\gamma(w) := \frac{1}{n}\sum_{i=1}^{n} \ell_\gamma(y_i(w \cdot x_i))\}$$

A Generalization Bound of Large-margin Classifiers

Theorem

For all $w \in \mathcal{H}$ *and* $\gamma > 0$,

$$L(w) \leq \hat{L}_{\gamma}(w) + \frac{2R_n(\mathcal{H})}{\gamma} + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

with probability at least $1 - \delta$.

Proof Sketch I

- Recall

$$\ell_\gamma(v) := \min\left\{1, \max\left\{0, 1 - \frac{v}{\gamma}\right\}\right\}, \quad L_\gamma(w) := \mathbb{E}\{\ell_\gamma(y(w \cdot x))\}, \quad \text{and} \quad \hat{L}_\gamma(w) := \frac{1}{n}\sum_{i=1}^{n} \ell_\gamma(y_i(w \cdot x_i))\}$$

- Our generalization bound via the Rademacher complexity:

$$L(h) \leq \hat{L}(h) + 2R_n(\mathcal{F}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

- As $\ell_{\text{0-1}} \leq \ell_\gamma$, for any $w \in \mathcal{H}$, we have

$$L(w) \leq L_\gamma(w)$$

Proof Sketch II

- Thus, we have

$$
\begin{aligned}
L(w) &\leq L_\gamma(w) \\
&\leq \hat{L}_\gamma(w) + 2R_n(\ell_\gamma \circ \mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} && (1) \\
&\leq \hat{L}_\gamma(w) + \frac{2R_n(\mathcal{H})}{\gamma} + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} && (2)
\end{aligned}
$$

  - (1) the generalization bound via Rademacher complexity.
  - (2) the Talagrand's lemma (check out our references!)

From Theory to Algorithm I

From the Large-margin Bound to the SVM Algorithm

Theory:

$$L(w) \leq \hat{L}_\gamma(w) + \frac{2R_n(\mathcal{H})}{\gamma} + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

Algorithm:

$$\min_w \; \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathsf{hinge}}(y_i(w \cdot x_i)) + \lambda \|w\|_2$$

We will see only a high-level connection (see our references for details).

From Theory to Algorithm II

From the Large-margin Bound to the SVM Algorithm

Connection?

- margin loss $\ell_\gamma(v)$ and hinge loss $\ell_{\text{hinge}}(v)$:

$$\ell_\gamma(v) \coloneqq \min\left\{1, \max\left\{0, 1 - \frac{v}{\gamma}\right\}\right\} \quad \text{and} \quad \ell_{\text{hinge}}(v) \coloneqq \max(0, 1 - v)$$

- the upper bound of $\ell_\gamma(v)$:

$$\begin{aligned}
\ell_\gamma(y(w \cdot x)) &= \min\left\{1, \max\left\{0, 1 - \frac{y(w \cdot x)}{\gamma}\right\}\right\} \\
&\leq \max\left\{0, 1 - \frac{y(w \cdot x)}{\gamma}\right\} \\
&= \max\left\{0, 1 - y\left(\frac{w}{\gamma} \cdot x\right)\right\} \\
&= \ell_{\text{hinge}}\left(y\left(\frac{w}{\gamma} \cdot x\right)\right)
\end{aligned}$$

From Theory to Algorithm III

From the Large-margin Bound to the SVM Algorithm

- The Rademacher complexity is (roughly) bounded as follows:

$$R_n(\mathcal{H}) \leq \mathcal{O}\left(\sqrt{\frac{1}{\gamma^2 n}}\right)$$

- An algorithm that minimizes the upper bound (given a hyper-parameter γ):

$$\min_{w:\|w\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathsf{hinge}}\left(y_i \left(\frac{w}{\gamma} \cdot x_i\right)\right)$$

- The change of a variable:

$$w' = \frac{w}{\gamma} \quad \Rightarrow \quad \|w'\|_2 \leq \frac{1}{\gamma}$$

From Theory to Algorithm IV

From the Large-margin Bound to the SVM Algorithm

- SVM algorithm:

$$\min_{w' : \|w'\|_2 \leq \frac{1}{\gamma}} \frac{1}{n} \sum_{i=1}^{n} \ell_{\textsf{hinge}} \left(y_i \left(w' \cdot x_i \right) \right) \quad \Longleftrightarrow \quad \min_{w' \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell_{\textsf{hinge}} \left(y_i \left(w' \cdot x_i \right) \right) + \lambda \|w'\|_2$$

  - Why? Check your convex optimization book.
- This algorithm minimizes the expected error (as we directly minimize the upper bound of the expected error).

SVM is Agnostic-PAC

Bound (again): Given γ

$$L(w) \leq \underbrace{\hat{L}_\gamma(w)}_{\text{minimized}} + \frac{2R_n(\mathcal{H})}{\gamma} + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

Why? — the same argument as in ERM.

$$\begin{aligned}
L(\mathcal{A}(\mathcal{S})) - L(h^*) &= \left\{L(\mathcal{A}_{\mathsf{SVM}}(\mathcal{S})) - \hat{L}(\mathcal{A}_{\mathsf{SVM}}(\mathcal{S}))\right\} + \left\{\hat{L}(\mathcal{A}_{\mathsf{SVM}}(\mathcal{S})) - \hat{L}(h^*)\right\} + \left\{\hat{L}(h^*) - L(h^*)\right\} \\
&\leq \underbrace{\left\{L(\mathcal{A}_{\mathsf{SVM}}(\mathcal{S})) - \hat{L}(\mathcal{A}_{\mathsf{SVM}}(\mathcal{S}))\right\}}_{\text{uniform convergence}} + \underbrace{\left\{\hat{L}(h^*) - L(h^*)\right\}}_{\text{concentration inequality}} \\
&\leq \frac{2R_n(\mathcal{H})}{\gamma} + \sqrt{\frac{\ln \frac{1}{\delta_1}}{2n}} + \sqrt{\frac{\ln \frac{1}{\delta_2}}{2n}}
\end{aligned}$$

with probability at least $1 - (\delta_1 + \delta_2)$.

Conclusion

1. We have explored generalization bounds via uniform convergence under various setups.

Conclusion

1. We have explored generalization bounds via uniform convergence under various setups.
    - $\mathcal{H}$: finite

Conclusion

1. We have explored generalization bounds via uniform convergence under various setups.
    - $\mathcal{H}$: finite
    - $\mathcal{H}$: infinite – Rademacher complexity

Conclusion

1. We have explored generalization bounds via uniform convergence under various setups.
    - $\mathcal{H}$: finite
    - $\mathcal{H}$: infinite – Rademacher complexity
    - ℓ: 0-1 loss

Conclusion

1. We have explored generalization bounds via uniform convergence under various setups.
    - $\mathcal{H}$: finite
    - $\mathcal{H}$: infinite – Rademacher complexity
    - ℓ: 0-1 loss
    - ℓ: margin loss

Conclusion

1. We have explored generalization bounds via uniform convergence under various setups.
    - $\mathcal{H}$: finite
    - $\mathcal{H}$: infinite – Rademacher complexity
    - ℓ: 0-1 loss
    - ℓ: margin loss
2. What are potential limitations of statistical learning theory?

Conclusion

1. We have explored generalization bounds via uniform convergence under various setups.
    - $\mathcal{H}$: finite
    - $\mathcal{H}$: infinite – Rademacher complexity
    - ℓ: 0-1 loss
    - ℓ: margin loss
2. What are potential limitations of statistical learning theory?
    - the i.i.d. assumption!
3. In online learning, we will learn a learning algorithm without the i.i.d. assumption.