Advanced Topics in Data Sciences

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 11: Uniform Convergences in Statistical Learning Theory

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-731 (Spring 2016)

Outline

This lecture: Reducing the error bounds to **complexity measures** of the hypothesis class

1. Preliminaries
2. Classical VC Theory for Binary Classification
3. Uniform Convergence and Rademacher Complexity
4. A Brief View of Modern Statistical Learning

Recommended reading materials

Binary Classification:

1. Section 2.2, Section 5.3 in R. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*, The MIT Press, 2012.

2. Chapters 1–3 in S. Boucheron *et al.*, "Theory of classification: A survey of some recent advances," *ESIAM: Probab. Stat.*, 2005.

Modern Statistical Learning Theory:

1. S. Mendelson, "Learning without Concentration", Journal of ACM, 2015.

Preliminaries

The standard statistical learning model

- **Training Data:** $\mathcal{D}_n := \{Z_i\}_{i=1}^n$ i.i.d. unknown P on $\mathcal{Z}$
- **Hypothesis Class:** $\mathcal{H}$ a set of hypotheses h
- **Loss Function:** $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$
- **Risk:** $L(h) \coloneqq \mathbb{E}_{Z\sim P}\ell(h, Z)$, where Z is independent of $\mathcal{D}_n$
- **Empirical Risk Minimization:**

$$\hat{h}_n = \arg\min_{h\in\mathcal{H}} L_n(h) \coloneqq \arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h, Z_i)$$

Performance Measures

With high probability, we have

1. **Generalization Error:** $L(\hat{h}_n) \leq L_n(\hat{h}_n) + \epsilon_1$
2. **Excess Risk:** $L(\hat{h}_n) - \inf_{h\in\mathcal{H}} L(h) \leq \epsilon_2$

Uniform convergence

Definition (Uniform Convergence [1])

A hypothesis class $\mathcal{H}$ has the uniform convergence property, if there exists a function $n_{\mathcal{H}}(\varepsilon, \delta)$, such that for every $\varepsilon, \delta \in (0,1)$ and any probability distribution P, if $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$, we have

$$\sup_{h \in \mathcal{H}} |L_n(h) - L(h)| \leq \varepsilon,$$

with probability at least $1 - \delta$.

Proposition [1]

For any $\varepsilon > 0$, if

$$\sup_{h \in \mathcal{H}} |L_n(h) - L(h)| \leq \varepsilon,$$

then for any $h^\star \in \arg\min_{h \in \mathcal{H}} L(h)$, we have

1. $L(\hat{h}_n) \leq L_n(\hat{h}_n) + \varepsilon$.
2. $L(\hat{h}_n) - L(h^\star) \leq 2\varepsilon$.

Recall: Hoeffding's Lemma

Theorem (Hoeffding's Lemma [2])

Let Y *be a random variable with* $\mathbb{E}[Y] = 0$*, taking values in a bounded interval* $[a, b]$*.*
Let $\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$*. Then* $\psi_Y''(\lambda) \leq \frac{(b-a)^2}{4}$ *and* $Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$*.*
In particular, for all $Y \in [a, b]$*,*

$$\Pr\Big(|Y - \mathbb{E}Y| > t\Big) \leq 2\exp\left(-\frac{2t^2}{(b-a)^2}\right).$$

Recall: Bounded Difference Inequality

Definition (Bounded Difference Functions)

A function $f : \mathcal{X}^n \to \mathbb{R}$ has the bounded differences property if for some positive $c_1, .., c_n$,

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} |f(x_1, .., x_i, ..., x_n) - f(x_1, ..., x_i', ..., x_n)| \leq c_i.$$

Theorem (Bounded Differences Inequality [2])

Let $X_1, ..., X_n$ be independent random variables, and let f satisfy the bounded differences property with c_i's. Then

$$P\Big(|f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)| > t\Big) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Classical VC Theory for Binary Classification

Binary classification

- **Training Data:** $\mathcal{D}_n = \{Z_i = (X_i, Y_i) : 1 \leq i \leq n\}$
- **Hypothesis Class:** $\mathcal{H}$ a set of classifiers $h : \mathcal{X} \to \{0, 1\}$
- **Loss Function:** Binary loss $\ell(h, Z_i) := \mathbb{1}_{\{Y_i \neq h(X_i)\}}$
- **Risk:** $L(h) \coloneqq \mathbb{E}_{Z \sim P} \ell(h, Z) = P\left(Y \neq h(X)\right)$
- **Empirical Risk:** $L_n(h) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \neq h(X_i)\}}$

Key Question

How do we bound $\sup_{h \in \mathcal{H}} |L_n(h) - L(h)|$?

VC Theory for Binary Classification

- Single Hypothesis $\mathcal{H} = \{h\}$: $\sup_{h \in \mathcal{H}} |L_n(h) - L(h)| = |L_n(h) - L(h)|$.
  Hoeffding's lemma applied to $\ell(h, Z_i) = \mathbb{1}_{\{Y_i \neq h(X_i)\}}$ implies, with probability at least $1 - \delta$,

$$|L_n(h) - L(h)| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

VC Theory for Binary Classification

- Single Hypothesis $\mathcal{H} = \{h\}$: $\sup_{h\in\mathcal{H}} |L_n(h) - L(h)| = |L_n(h) - L(h)|$.
  Hoeffding's lemma applied to $\ell(h, Z_i) = \mathbb{1}_{\{Y_i \neq h(X_i)\}}$ implies, with probability at least $1-\delta$,
$$|L_n(h) - L(h)| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}.$$
- Finite Hypotheses: Union bound + Hoeffding's lemma implies, with probability at least $1-\delta$,
$$\sup_{h\in\mathcal{H}} |L_n(h) - L(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}}$$

VC Theory for Binary Classification

- Single Hypothesis $\mathcal{H} = \{h\}$: $\sup_{h\in\mathcal{H}} |L_n(h) - L(h)| = |L_n(h) - L(h)|$.
  Hoeffding's lemma applied to $\ell(h, Z_i) = \mathbb{1}_{\{Y_i \neq h(X_i)\}}$ implies, with probability at least $1-\delta$,
$$|L_n(h) - L(h)| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}.$$
- Finite Hypotheses: Union bound + Hoeffding's lemma implies, with probability at least $1-\delta$,
$$\sup_{h\in\mathcal{H}} |L_n(h) - L(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}}$$
- Infinite Hypotheses: VC theory.

VC Theory for Binary Classification

Key Insight

Instead of considering the number of hypothesis, consider the number of **effective hypothesis**.

VC Theory for Binary Classification

Key Insight

Instead of considering the number of hypothesis, consider the number of **effective hypothesis**.

Examples

1. Linear classifiers.
2. Rectangle classifiers.

VC Theory for Binary Classification

Definition (Dichotomies [7])

For any finite sample $S = \langle x_1, ..., x_n \rangle$, the set of dichotomies is defined to be all possible labelings of S by the functions in $\mathcal{H}$:

$$\Pi_{\mathcal{H}}(S) \coloneqq \{\langle h(x_1), ..., h(x_n)\rangle : h \in \mathcal{H}\}.$$

Definition (Growth Function [7])

$$\Pi_{\mathcal{H}}(n) \coloneqq \max_{S \in \mathcal{X}^n} |\Pi_{\mathcal{H}}(S)|$$

VC Theory for Binary Classification

Definition (Dichotomies [7])

For any finite sample $S = \langle x_1, ..., x_n \rangle$, the set of dichotomies is defined to be all possible labelings of S by the functions in $\mathcal{H}$:

$$\Pi_{\mathcal{H}}(S) \coloneqq \{\langle h(x_1), ..., h(x_n) \rangle : h \in \mathcal{H}\}.$$

Definition (Growth Function [7])

$$\Pi_{\mathcal{H}}(n) \coloneqq \max_{S \in \mathcal{X}^n} |\Pi_{\mathcal{H}}(S)|$$

Theorem ([7])

With probability at least $1 - \delta$, *we have*

$$\sup_{h \in \mathcal{H}} |L_n(h) - L(h)| \leq \sqrt{\frac{32 \ln \Pi_{\mathcal{H}}(n) + \ln(16/\delta)}{n}}.$$

Bounding the Growth Function: The VC Dimension

Definition (Shattering coefficient [7])

The shattering coefficient of a hypothesis class $\mathcal{H}$ is defined as

$$S_n(\mathcal{H}) := \sup_{x_1,\ldots,x_n \in \mathcal{X}} \left| \{(h(x_i))_{1 \le i \le n} : h \in \mathcal{H}\} \right|.$$

Definition (Vapnik-Chervonenkis (VC) dimension [7])

The VC dimension of a hypothesis class $\mathcal{H}$, denoted by d, is defined as the largest integer k such that $S_k(\mathcal{H}) = 2^k$. If $S_k(\mathcal{H}) = 2^k$ for all k, then $d \coloneqq \infty$.

Bounding the Growth Function: The VC Dimension

Definition (Shattering coefficient [7])

The shattering coefficient of a hypothesis class $\mathcal{H}$ is defined as

$$S_n(\mathcal{H}) := \sup_{x_1,\ldots,x_n \in \mathcal{X}} \left| \{(h(x_i))_{1 \le i \le n} : h \in \mathcal{H}\} \right| .$$

Definition (Vapnik-Chervonenkis (VC) dimension [7])

The VC dimension of a hypothesis class $\mathcal{H}$, denoted by d, is defined as the largest integer k such that $S_k(\mathcal{H}) = 2^k$. If $S_k(\mathcal{H}) = 2^k$ for all k, then $d := \infty$.

Lemma (Sauer-Shelah [9])

The growth function is bounded by

$$\Pi_{\mathcal{H}}(n) \le \sum_{i=0}^{d} \binom{n}{i}.$$

In particular, $\Pi_{\mathcal{H}}(n) \le \left(\frac{en}{d}\right)^d$

VC Theory for Binary Classification

Theorem (The VC Bound for Binary Classification [10])

Let $\mathcal{H}$ be a hypothesis class with VC dimension d. Assume that $n \geq d$. Then with probability at least $1-\delta$,

$$\sup_{h \in \mathcal{H}} |L_n(h) - L(h)| \leq \mathcal{O}\left(\sqrt{\frac{d\ln(n/d) + \ln(1/\delta)}{n}}\right).$$

Important Implications 1: Learnability and VC Dimension

Sauer-Shelah lemma implies that there can be only two cases for the growth function $\Pi_{\mathcal{H}}(n)$:

- Case1: $\Pi_{\mathcal{H}}(n) = 2^n \Leftrightarrow d = \infty \Leftrightarrow$ the function class $\mathcal{H}$ is not learnable.

Important Implications 1: Learnability and VC Dimension

Sauer-Shelah lemma implies that there can be only two cases for the growth function $\Pi_{\mathcal{H}}(n)$:

- Case1: $\Pi_{\mathcal{H}}(n) = 2^n \Leftrightarrow d = \infty \Leftrightarrow$ the function class $\mathcal{H}$ is not learnable.
- Case2: $\Pi_{\mathcal{H}}(n)$ grows polynomially $\Leftrightarrow d < \infty \Leftrightarrow$ the function class $\mathcal{H}$ is learnable and the VC bound holds.

Important Implications 2: Fast Rates

Definition (Realizable/Consistent Hypotheses)

A set of training samples $S = \{Z_i\}_{i=1}^n$ is said to be *consistent* with the hypothesis class $\mathcal{H}$ if there is a $h \in \mathcal{H}$ such that $L_n(h) = 0$ on S.

Theorem (The VC Bound for Binary Classification [7])

Let $\mathcal{H}$ be a hypothesis class with VC dimension d. Let S be a set of training samples with size n and assume that $n \geq d$. Then with probability at least $1 - \delta$,

$$L(h) \leq \mathcal{O}\left(\frac{d \ln(n/d) + \ln(1/\delta)}{n}\right)$$

for every $h \in \mathcal{H}$ that is consistent with S.

Drawbacks of VC Theory

Although the VC bound reveals many important phenomena in learning, it has some serious drawbacks:

1. It is very loose in practice (holds for *all* data and *all* distributions).
2. Generalization to regression problems is not straightforward.

Uniform Convergence and Rademacher Complexity

Rademacher Complexity: Another Measure of Complexity

Motivation: Consider a binary classification problem. Let the sample be $(x_1, y_1), ..., (x_n, y_n)$, where $y_i \in \{1, -1\}$. We can rewrite the empirical risk minimization procedure as

$$\max_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} y_i h(x_i).$$

Rademacher Complexity: Another Measure of Complexity

Motivation: Consider a binary classification problem. Let the sample be $(x_1, y_1), ..., (x_n, y_n)$, where $y_i \in \{1, -1\}$. We can rewrite the empirical risk minimization procedure as

$$\max_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} y_i h(x_i).$$

Definition (Rademacher Complexity, Binary Classification [7])

Let $S = \langle x_1, ..., x_n \rangle$ be a given set of input instances, and let σ_i be a Rademacher random variable (-1 or $+1$ with equal probability). The Rademacher complexity of a class of binary functions $\mathcal{H}$ with respect to S is defined as

$$\mathbb{E} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i).$$

Remark: Rademacher complexity measures how well $\mathcal{H}$ can fit pure noise.

Rademacher Complexity: Another Measure of Complexity

Definition (Rademacher Complexity, General Cases [7])

Let $S = \langle z_1, ..., z_n \rangle$ be a given set of input instances, and let σ_i be a Rademacher random variable (-1 or $+1$ with equal probability). The Rademacher complexity of a class of binary functions $\mathcal{F}$ with respect to S is defined as

$$R_S(\mathcal{F}) \coloneqq \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i).$$

Remark: Rademacher complexity measures the *correlation* between $\mathcal{F}$ and pure noise.

Uniform Convergence through Rademacher Complexity

Theorem ([1])

Let $\mathcal{F}$ be any family of functions $\mathcal{Z} \to [-1, +1]$. Let $S = \{Z_i\}_{i=1}^n$ be random samples of size n. Then, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E} f(Z) - \frac{1}{n} \sum_{i=1}^{n} f(Z_i) \right| \leq 2 \mathbb{E}_S R_S(\mathcal{F}) + \sqrt{\frac{2 \ln(1/\delta)}{n}}.$$

We also have

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E} f(Z) - \frac{1}{n} \sum_{i=1}^{n} f(Z_i) \right| \leq 2 R_S(\mathcal{F}) + \sqrt{\frac{2 \ln(2/\delta)}{n}}.$$

Uniform Convergence through Rademacher Complexity

Theorem ([1])

Let $\mathcal{F}$ be any family of functions $\mathcal{Z} \to [-1, +1]$. Let $S = \{Z_i\}_{i=1}^n$ be random samples of size n. Then, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E} f(Z) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right| \leq 2 \mathbb{E}_S R_S(\mathcal{F}) + \sqrt{\frac{2 \ln(1/\delta)}{n}}.$$

We also have

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E} f(Z) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right| \leq 2 R_S(\mathcal{F}) + \sqrt{\frac{2 \ln(2/\delta)}{n}}.$$

Remark: For binary classification, let $\mathcal{F} = \ell \circ \mathcal{H}$; that is, let $f(Z_i) = \ell(h, Z_i)$. Then the following holds with probability at least $1 - \delta$:

$$\sup_{h \in \mathcal{H}} |L_n(h) - L(h)| \leq R_S(\mathcal{H}) + \sqrt{\frac{2 \ln(2/\delta)}{n}}.$$

Uniform Convergence through Rademacher Complexity

The same analysis can be applied to **bounded and Lipschitz loss**, due to the following important property of Rademacher complexity:

Theorem (Contraction Principle [4])

Suppose that $\phi : \mathbb{R} \to \mathbb{R}$ is a L-Lipschitz function with $\phi(0) = 0$. Then, for any function class $\mathcal{F}$ and any sample S,

$$R_S(\phi \circ \mathcal{F}) := \mathbb{E} \sup_{h \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i (\phi \circ f)(z_i) \leq L R_S(\mathcal{F}).$$

Uniform Convergence through Rademacher Complexity

We consider the problem of bounded regression:

- **Training Data:** $\mathcal{D}_n = \{Z_i = (X_i, Y_i) : 1 \leq i \leq n\}$, where $Y_i \in [-\frac{1}{2}, +\frac{1}{2}]$
- **Hypothesis Class:** $\mathcal{H}$ a set of regression function $h : \mathcal{X} \to [-\frac{1}{2}, +\frac{1}{2}]$
- **Loss Function:** Squared loss $\ell(h, Z_i) := (h(X_i) - Y_i)^2$

Uniform Convergence through Rademacher Complexity

For bounded regression,

Theorem ([1])

Let $\mathcal{F}$ be any family of functions $\mathcal{Z} \to [-1, +1]$. Let $S = \{Z_i\}_{i=1}^n$ be random samples of size n. Then, with probability at least $1 - \delta$,

$$\sup_{f \in \boldsymbol{\mathcal{F}}} \left| \mathbb{E} f(Z) - \frac{1}{n} \sum_{i=1}^{n} f(Z_i) \right| \leq 2 R_S(\boldsymbol{\mathcal{F}}) + \sqrt{\frac{2 \ln(2/\delta)}{n}}.$$

$+$

Contraction Principle

$\Downarrow$

$$\sup_{f \in \mathcal{H}} |L_n(h) - L(h)| \leq 2 R_S(\mathcal{H}) + \sqrt{\frac{2 \ln(2/\delta)}{n}}$$

with probability at least $1 - \delta$.

Rademacher Complexity

Advantages of Rademacher complexity:

1. It works for many learning problems.
2. It is tighter than the VC bound, both in practice and theory.
3. It allows us to derive data dependent bounds.

A Brief View of Modern Statistical Learning

A Brief View of Modern Statistical Learning

Modern statistical learning theory aims at:

1. Deriving bounds that reveal high-dimensional phenomena, such as distribution dependent bounds.
2. Getting rid of redundant assumptions (such as boundedness).

A Brief View of Modern Statistical Learning

Modern statistical learning theory aims at:

1. Deriving bounds that reveal high-dimensional phenomena, such as distribution dependent bounds.
2. Getting rid of redundant assumptions (such as boundedness).

To achieve these goals, we need to impose more assumptions on the distribution that generates the data.

A Brief View of Modern Statistical Learning

Two parameters that involve the localized Rademacher complexity:

Definition ([6])

Given a function class $\mathcal{F}$ and $\gamma > 0$. Set

$$\beta^*(\gamma) = \inf \left\{ r > 0 : \ \mathbb{E} \sup_{f \in \mathcal{F} \cap r D_{f^*}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i (f - f^*)(X_i) \right| \leq \gamma r \right\}$$

where $D_{f^*} = \{ f : \| f - f^* \| \leq 1 \}$.

Definition ([6])

Let $\xi_i = f^*(X_i) - Y_i$ and $\psi_n(s) = \sup_{f \in \mathcal{F} \cap s D_{f^*}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \xi_i (f - f^*)(X_i) \right|$.
Given $\gamma, \delta > 0$. Set

$$\alpha^*(\gamma, \delta) = \inf \left\{ s > 0 : \ P\left(\psi_n(s) \leq \gamma s^2 \right) \geq 1 - \delta \right\}.$$

A Brief View of Modern Statistical Learning

- **Training Data:** $\mathcal{D}_n = \{Z_i = (X_i, Y_i) : 1 \leq i \leq n\}$, where $Y_i \in \mathbb{R}$
- **Hypothesis Class:** $\mathcal{F}$ a set of convex regression function $f : \mathcal{X} \to \mathbb{R}$
- **Loss Function:** Squared loss $\ell(f, Z_i) := (f(X_i) - Y_i)^2$

Theorem ([6])

Under mild assumptions, there exist constants $c_1, c_2, c_3 > 0$ such that, with probability $1 - \delta - \exp(-nc_1)$,

$$\|\hat{f} - f^*\| \leq 2 \max\{\alpha^*(c_2, \delta/4), \beta^*(c_3)\}.$$

What's not covered in this lecture...

Not covered in this lecture...

1. Bounding the Rademacher complexity: **Gaussian complexity, Chaining/Generic Chaining** [8]
2. Missing assumptions in modern statistical learning: **Small-ball conditions** [6]
3. General convex loss functions [5]
4. Stability analysis [3]

References I

[1] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi.
Theory of classification: A survey of some recent advances.
ESAIM: Probab. Stat., 9:323–375, November 2005.

[2] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart.
Concentration Inequalities: A Nonasymptotic Theory of Independence.
Oxford Univ. Press, Oxford, 2013.

[3] Moritz Hardt, Benjamin Recht, and Yoram Singer.
Train faster, generalize better: Stability of stochastic gradient descent.
2015.
arXiv:1509.01240v1 [cs.LG].

[4] Vladimir Koltchinskii.
Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.
Springer-Verl., Berlin, 2011.

[5] Shahar Mendelson.
Learning without concentration for general loss functions.
2014.
arXiv:1410.3192v1 [stat.ML].

References II

[6] Shahar Mendelson.
Learning without concentration.
J. ACM, 62(3), 2015.

[7] Robert E. Schapire and Yoav Freund.
Boosting.
MIT Press, Cambridge, MA, 2012.

[8] Michel Talagrand.
Upper and Lower Bounds for Stochastic Processes.
Springer, Berlin, 2014.

[9] Ramon van Handel.
Probability in High Dimension.
June 2014.

[10] V. N. Vapnik and A. Ya. Chervonenkis.
On the uniform convergence of relative frequencies of events to their probabilities.
Theory Probab. Appl., XVI(2):264–280, 1971.