EE-731: Advanced Topics in Data Sciences
Laboratory for Information and Inference Systems
Spring 2016

Instructor: Volkan Cevher

Scribers: Ya-Ping Hsieh, Ilija Bogunovic

Uniform Convergence in Statistical Learning Theory

Index terms: *Statistical Learning Theory, Uniform Convergence, VC Theory, Rademacher Complexity*

We have introduced the framework of Statistical Learning in previous lectures. The purpose of this lecture is to introduce some of the most basic error bounds in Statistical Learning Theory. We follow the traditional path of *Uniform Convergence*, and briefly explain how this notion should be improved for modern applications.

1 Preliminaries

Recall the following ingredients of Statistical Learning Theory:

- **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:** $L(h) := \mathbb{E}_{Z\sim P}\ell(h, Z)$, where Z is independent of $\mathcal{D}_n$.
- **Empirical Risk Minimization:**

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

The goal is to learn from data, in the sense that we have guarantees in

- Generalization error: $L(h) \le L(\hat{h}_n) + \epsilon_1$
- Excess risk: $L(\hat{h}_n) - \inf_{h\in\mathcal{H}} L(h) \le \epsilon_2$

for some small numbers ϵ_1 and ϵ_2, ideally decreasing in the number of samples. We have seen that this can be achieved by studying the uniform convergence property.

Definition 1 (Uniform Convergence [10])**.** *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 $n \ge n_{\mathcal{H}}(\varepsilon, \delta)$, we have*

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

with probability at least $1 - \delta$.

The following theorem explains why uniform convergence is useful in providing guarantees.

Theorem 1.1. *For any $\varepsilon > 0$, if $\sup_{h\in\mathcal{H}} |L_n(h) - L(h)| \le \varepsilon$, then for any $h^\star \in \arg\min_{h\in\mathcal{H}} L(h)$, we have*

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

Proof. The first inequality is simply $|L_n(\hat{h}_n) - L(\hat{h}_n)| \le \sup_{h\in\mathcal{H}} |L_n(h) - L(h)|$. The second inequality follows since

$$\begin{aligned} L(\hat{h}_n) - L(h^\star) &= L(\hat{h}_n) - L_n(\hat{h}_n) + L_n(\hat{h}_n) - L(h^\star) \\ &\le L(\hat{h}_n) - L_n(\hat{h}_n) + L_n(h^\star) - L(h^\star) \\ &\le 2\sup_{h\in\mathcal{H}} |L_n(h) - L(h)|. \end{aligned}$$

where we have used the fact that $\hat{h}_n$ minimizes $L_n(\cdot)$.

□

In the sequel we shall need the following two concentration of measure inequalities.

Theorem 1.2 (Hoeffding's inequality [3]). *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 $\psi''_Y(\lambda) \le \frac{(b-a)^2}{4}$ and $Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$.*

In particular, for all $Y \in [a, b]$,

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

Stated differently but equivalently, with probability at least $1 - \delta$, we have

$$|Y - \mathbb{E}Y| \le (b-a)\sqrt{\frac{\ln(2/\delta)}{2}}. \tag{1}$$

Definition 2 (Bounded Difference Functions). *A function $f : \mathcal{X}^n \to \mathbb{R}$ has the bounded differences property if for some positive $c_1, .., c_n$,*

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

Theorem 1.3 (Bounded Differences Inequality [3]). *Let $X_1, ..., X_n$ be independent random variables, and let f satisfy the bounded differences property with c_i's. Then*

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

A similar statement of (1) holds for the Bounded Differences Inequality.

2 Classical VC Theory for Binary Classification

Historically, the first bound of uniform convergence type is given by the VC bound for binary classification. In this section, we illustrate the key ideas of VC bound, and we refer the readers to Section 12 of [5] for proofs.

Consider the problem of binary classification; that is, we consider the following setup:

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

In view of **Theorem 1.1**, we can focus on bounding $\sup_{h\in\mathcal{H}} |L_n(h) - L(h)|$. To start, consider the case where $\mathcal{H}$ is a single, fixed hypothesis h. Applying Hoeffding's bound (1) to $L_n(h) = \frac{1}{n}\sum_{i=1}^n \ell(h, Z_i)$, we immediately see that, with probability at least $1 - \delta$,

$$\sup_{h\in\mathcal{H}} |L_n(h) - L(h)| = |L_n(h) - L(h)| \le \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{2}$$

An equally simple case is when $\mathcal{H}$ is finite, where Hoeffding's lemma followed by a union bound implies that, with probability at least $1 - \delta$,

$$\sup_{h\in\mathcal{H}} |L_n(h) - L(h)| \le \sqrt{\frac{\ln|\mathcal{H}| + \ln(2/\delta)}{2n}}. \tag{3}$$

We next consider infinitely large hypothesis spaces $\mathcal{H}$. Although the bound (3) is useless for such $\mathcal{H}$, it turns out it is possible to replace the cardinality in (3) by the **effective cardinality** of $\mathcal{H}$. This important fact lies at the heart of VC theory, and we illustrate how this can be done by the following example.

Example 1: 1-Dimensional Linear Classifiers Consider the set of hypotheses formed by picking any real number r, and declare $+1$ if $x > r$, and -1 if $x < r$ (breaking ties arbitrarily); see Figure 1. The cardinality of such hypothesis set is infinite. However, for any given n points, we can produce *at most* $n + 1$ different output labels. This suggests that we may think of the hypothesis class as containing effectively $n + 1$ different hypotheses when the number of samples is n.

The above rough statements can be made precise. Toward this end, we need some definitions.

Definition 3 (Dichotomies). *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 $\mathcal{H}$:*

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

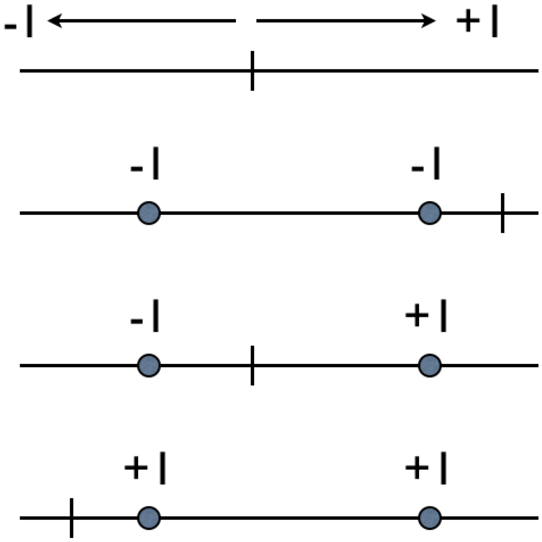

Figure 1: One dimensional linear classifier. Given 2 input points, one can produce at most 3 different output labels using such hypothesis class, although the number of hypotheses is infinite.

Definition 4 (Growth Function)**.**

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

In other words, given n input points, the growth function is the maximum number of different output labels that $\mathcal{H}$ can produce. For example, $\Pi_{\mathcal{H}}(n) = n + 1$ in **Example 1**.

We are now ready to state a key result in VC theory. Ignoring constants, the next theorem states that one can replace $|\mathcal{H}|$ in (3) with $\Pi_{\mathcal{H}}(n)$.

Theorem 2.1. *With probability at least* $1 - \delta$, *we have*

$$\sup_{h \in \mathcal{H}} |L_n(h) - L(h)| \leq \sqrt{\frac{32 \ln \Pi_{\mathcal{H}}(n) + \ln(16/\delta)}{n}}. \tag{4}$$

So our attention naturally turns to the growth function $\Pi_{\mathcal{H}}(n)$. In "nice" cases, such as in **Example 1**, the growth function can be expressed in a close form. However, the growth function needs not always admit such expressions. Moreover, it is easy to construct examples where $\mathcal{H}$ can perfectly reproduce all possible combinations of output labels, in which case $\Pi_{\mathcal{H}}(n) = 2^n$ and (4) becomes a constant (consider the hypotheses class formed by declaring $+1$ on union of arbitrary intervals, and -1 on the complement). All these observations seem to suggest that (4) cannot be useful in full generality. Surprisingly, a remarkable result of combinatorics says that *controlling* $\Pi_{\mathcal{H}}(n)$ *is easy*: All we need is to compute one characteristic of the hypothesis class, the VC dimension.

Definition 5 (Shattering coefficient)**.** *The shattering coefficient of a hypothesis class* $\mathcal{H}$ *is defined as*

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

Definition 6 (Vapnik-Chervonenkis (VC) dimension)**.** *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* $d := \infty$*.*

In short, the VC dimension is the maximum number d such that we can perfectly recover 2^d output labels on d input points, and produce strictly $< 2^{d+1}$ output labels on $d + 1$ samples. For example, in **Example 1**, $d = 1$ since we can produce only $3 < 2^{1+1}$ output labels on 2 input points.

There exist some obvious connections between $\Pi_{\mathcal{H}}(n)$ and the VC dimension d. For instance, $\Pi_{\mathcal{H}}(n) < 2^n$ if and only if $d < \infty$ by definition. The following fundamental theorem gives a much stronger statement: $\Pi_{\mathcal{H}}(n)$ is upper bounded by a polynomial of degree d.

Lemma 2.2 (Sauer-Shelah)**.** *Let the VC dimension be d. The growth function is bounded by*

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

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

For a proof, see, e.g., [12].

The VC bound is obtained by substituting the above into (4).

Theorem 2.3 (The VC Bound for Binary Classification [13])**.** *Let $\mathcal{H}$ be a hypothesis class with VC dimension d. Assume that $n \geq d$. Then with probability at least $1 - \delta$,*

$$\sup_{h\in\mathcal{H}} |L_n(h) - L(h)| \leq O\left(\sqrt{\frac{d\ln(n/d) + \ln(1/\delta)}{n}}\right). \tag{5}$$

In short, ignoring constants and log factors, the effective cardinality of a hypothesis class is 2^d, where d is the VC dimension. Hence the VC dimension is a natural measure of the complexity of a hypothesis class.

3 Uniform Convergence and Rademacher Complexity

Although the VC theory is extremely important in revealing qualitative behaviors of learning, several drawbacks make it inapplicable in practice. First, the VC bound (5) is very loose in practice. This is due to the fact that it holds for *all* distributions and *all* possible realization of data. Such a doubly worst-case scenario, however, is never encountered in real world applications. Another significant disadvantage of the arguments in **Section 2** is that they do not easily generalize to other learning problems, such as regression or density estimation, which are equally important as binary classification.

In this section, we prove the uniform convergence through another important notion, the *Rademacher complexity*, that can be viewed again as a complexity measure of the hypothesis class. The benefits we gain by adopting Rademacher complexity will become obvious later. Specifically, the Rademacher complexity improves upon VC bound by

1. Providing data-dependent bounds, which are usually much tighter than (5) in practice.
2. Allowing immediate generalization to other learning problems.

To illustrate the intuition of Rademacher complexity, let us first consider the binary classification problem. Let the sample be $(x_1, y_1), ..., (x_n, y_n)$, where $y_i \in \{1, -1\}$. By the identity $\mathbb{1}_{\{h(x_i)\neq y_i\}} = \frac{1-h(x_i)y_i}{2}$, we can rewrite the empirical risk minimization procedure as

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

Now, suppose that, instead of the true labels $\{y_i\}_i$, the learner receives a set of random noise $\{\sigma_i\}_i$ where σ_i takes probability $\frac{1}{2}$ on both $+1$ and -1. If one can still achieve small training error in this setting, then we expect some sort of overfitting phenomenon has occurred, a sign that the hypothesis class $\mathcal{H}$ might be too complex for the sake of learning. Formally, we define the Rademacher complexity as follows.

Definition 7 (Rademacher Complexity, Binary Classification [9])**.** *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*

$$\mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i). \tag{6}$$

That is, in binary classification, the Rademacher complexity measures how well $\mathcal{H}$ can fit pure noise.

In the above definition, the function class is restricted to be binary. The importance of Rademacher complexity is that, unlike VC dimension (or growth function related notions), the Rademacher complexity naturally extends to arbitrary type of functions.

Definition 8 (Rademacher Complexity, General Cases [9])**.** *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 arbitrary functions $\mathcal{F}$ with respect to S is defined as*

$$R_S(\mathcal{F}) := \mathbb{E}\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right|. \tag{7}$$

We remark that the presence of the absolute value above, as oppose to the definition in (6), is insignificant since the term $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(z_i)$ is usually positive. However, adding absolute value greatly simplifies the analysis to come.

One has to be careful that, when extending to general hypothesis classes, we lose the interpretation of Rademacher complexity maximization being equivalent to empirical risk minimization over pure noise. However, one can still interpret the Rademacher complexity as measuring the *correlation* between $\mathcal{F}$ and pure noise. It turns out that, in a much wider class of learning problems, characterizing the uniform convergence can be done through the Rademacher complexity.

Theorem 3.1 ([2]). *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 $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} \left| \mathbb{E} f(Z) - \frac{1}{n}\sum_{i=1}^n f(Z_i) \right| \le 2\mathbb{E}_S R_S(\mathcal{F}) + \sqrt{\frac{2\ln(1/\delta)}{n}}. \tag{8}$$

We also have

$$\sup_{f\in\mathcal{F}} \left| \mathbb{E} f(Z) - \frac{1}{n}\sum_{i=1}^n f(Z_i) \right| \le 2R_S(\mathcal{F}) + \sqrt{\frac{2\ln(2/\delta)}{n}}. \tag{9}$$

Since the proof of **Theorem 3.1** is simple and elegant, we include it here for completeness.

We first need the so-called "symmetrization lemma", which is very useful in many fields of mathematics and is arguably more important than the **Theorem 3.1** itself.

Lemma 3.2 (Rademacher Symmetrization). *With the same notation as **Theorem 3.1**, we have*

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

Proof. Introduce a set of "ghost" samples $S' = \{Z_i'\}_{i=1}^n$ which are identical and independent of the original samples S. Then

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}} \left| \mathbb{E} f(Z) - \frac{1}{n}\sum_{i=1}^n f(Z_i) \right| &\overset{(1)}{=} \mathbb{E}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \left(f(Z_i) - \mathbb{E}_{Z_i'} f(Z_i') \right) \right| \\
&\overset{(2)}{=} \mathbb{E}\sup_{f\in\mathcal{F}} \left| \mathbb{E}_{S'} \frac{1}{n}\sum_{i=1}^n \left(f(Z_i) - f(Z_i') \right) \right| \\
&\overset{(3)}{\le} \mathbb{E}_{S,S'}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \left(f(Z_i) - f(Z_i') \right) \right| \\
&\overset{(4)}{=} \mathbb{E}_{S,S',\sigma}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i \left(f(Z_i) - f(Z_i') \right) \right| \\
&\overset{(5)}{\le} 2\mathbb{E}_{S,\sigma}\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i) \right|
\end{aligned}$$

where

(1) follows since Z and Z_i' have the same distribution,

(2) is because of the independence of Z_i and Z_i',

(3) uses Jensen's inequality,

(4) follows since the distribution of $f(Z_i) - f(Z_i')$ is in variant to sign change, and

(5) uses triangle inequality.

□

*proof of **Theorem 3.1**:* In view of **Lemma 3.2**, to prove (8), it suffices to note that the function $\sup_{f\in\mathcal{F}} \left| \mathbb{E} f(Z) - \frac{1}{n}\sum_{i=1}^n f(Z_i) \right|$ satisfies the assumption of bounded differences inequality.

For (9), we simply use the bounded differences inequality again, this time on the Rademacher complexity itself. □

To use **Theorem 3.1**, define $\mathcal{F} = \ell \circ \mathcal{H} := \{\ell(h, \cdot) : h \in \mathcal{H}\}$, and apply (9) to $\mathcal{F}$. The left-hand side of (9) then becomes the desired form in uniform convergence. The second term on the right-hand is simply a constant, so it remains to calculate the Rademacher complexity of $\ell \circ \mathcal{H}$ (**not** $\mathcal{H}$!). Evaluating the Rademacher complexity of the above form is a profound area in its own, and we refer the readers to [11] for a comprehensive account. Nonetheless, it is worth mentioning that the Rademacher complexity of many important function classes admit simple upper bounds, so one simply needs to plug in those existing results. For example, if the loss function is Lipschitz continuous, then one can reduce the tasking of bounding $R_S(\ell \circ \mathcal{H})$ to $R_S(\mathcal{H})$, where many results are readily available.

4 A Brief View of Modern Statistical Learning

Since the advent of compressive sensing [4], [6], the so-called "high-dimensional" phenomena have been a research area of intensive study. In the language of Statistical Learning Theory, the high-dimensional phenomena say that when the best hypotheses are typically *structured*, as is usually the case in modern applications, learning problems become much easier.

As a result, 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 in classical Statistical Learning Theory).

The central idea here is that, instead of demanding uniform convergence, a property that treats all the members in a function class equally, we consider a local neighborhood of the best member in the class, and require the worst case of *only* that local neighborhood to converge. If it happens that the best member has some structures (such as sparsity, etc.), then typically the local convergence is much easier than the global (uniform) convergence.

It turns out that we need to impose more assumptions on the distribution that generates the data (that is, we can no longer afford the luxury of distribution-free assertions). In a seminal paper [8], Mendelson has introduced two parameters that involve the [1]localized Rademacher complexity. Denote $\|f\|^2 := \int f^2 dP$ where P is the distribution generating the data. We define:

Definition 9 ([8])**.** *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 10 ([8])**.** *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

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

Mendelson showed that, under very mild assumptions on the distribution, it is possible to reveal high-dimensional phenomena for the least squared regression problems:

- **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 4.1 ([8])**.** *Under mild assumptions, there exist constants $c_1, c_2, c_3 > 0$ such that, with probability $1 - \delta - \exp(-nc_1)$,*

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

Very roughly speaking, instead of computing the Rademacher complexity for the whole function class, it suffices to consider only the Rademacher complexity in a neighborhood of the best hypothesis. Evaluating these localized Rademacher complexity will automatically reveal the desired high-dimensional phenomena.

References

[1] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Stat.*, 33(4):1497–1537, 2005.

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

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

[4] Emmanuel Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, December 2005.

[5] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[1] We remark that the localized Rademacher complexity appeared several years before the appearance of compressive sensing; see, e.g., [7] or [1]. The main novelty of [8] is to introduce a new set of assumptions that allow us to the achieve goals mentioned in the beginning of the section. For example, previously the results only apply to bounded loss.

[6] David L. Dohono. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Commun. Pure Appl. Math.*, LIX:797–829, 2006.

[7] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. 2004. arXiv:math/0405338v1 [math.PR].

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

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

[10] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning*. Cambridge Univ. Press, Cambridge, UK, 2014.

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

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

[13] 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.