E0 370 Statistical Learning Theory **Lecture 3 (Aug 16, 2011)**

Uniform Convergence and Growth Function/VC-Entropy

Lecturer: Shivani Agarwal *Scribe: Shivani Agarwal*

1 Introduction

In the previous lecture we reviewed the SVM learning algorithm, which when given a training sample $S \in (\mathcal{X} \times \{-1,1\})^m$, selects a (linear or kernel-based) classifier $h_S : \mathcal{X} \to \{-1,1\}$ that maximizes the margin on S, possibly with some errors. We know (by design) exactly what the algorithm does on the training sample, but how will it perform on future data, which is what we are really interested in? In particular, if the examples in S are drawn randomly and independently according to some distribution D, what can we say about how the learned function will perform on a new example drawn from D?

In this lecture we will begin our study of the generalization error $\mathrm{er}_D[h_S]$ of a function $h_S : \mathcal{X} \to \mathcal{Y}$ learned by an algorithm from a training sample $S \in (\mathcal{X} \times \mathcal{Y})^m \sim D^m$; in particular we will see a classical technique, namely that of *uniform convergence*, for obtaining confidence bounds on $\mathrm{er}_D[h_S]$ when the function h_S is selected from a function class $\mathcal{H}$ of sufficiently limited 'capacity'.[1,2] (We will define what we mean by capacity over the course of this and the next few lectures.) For the time being, we will focus on binary classification ($\mathcal{Y} = \{-1,1\}, \ell = \ell_{0\text{-}1}$).

It is worth pointing out that if one has a sufficiently large number (say $m+n$) of labeled examples available, all drawn independently from D, one can use part of these, say $S \in (\mathcal{X} \times \mathcal{Y})^m$, for training, and test the learned function h_S on the remaining data, say $T \in (\mathcal{X} \times \mathcal{Y})^n$. In this case, it is easy to obtain a confidence bound on $\mathrm{er}_D[h_S]$ in terms of the empirical error of h_S on T, $\mathrm{er}_T[h_S]$, via a straightforward application of Hoeffding's inequality:[3]

$$\mathbf{P}_{T\sim D^n}\Big(\Big|\mathrm{er}_D[h_S] - \mathrm{er}_T[h_S]\Big| \geq \epsilon\Big) \;\leq\; 2\,e^{-2n\epsilon^2}\,. \tag{1}$$

In general, however, labeled data is limited, and one takes the view that it is beneficial to use all the available examples for training (with the expectation that this will result in a function with better generalization performance); the generalization error $\mathrm{er}_D[h_S]$ then needs to be estimated using the same training sample S from which the function h_S is learned. This is what makes things complicated; in particular, note that

$$\mathbf{P}_{S\sim D^m}\Big(\Big|\mathrm{er}_D[h_S] - \mathrm{er}_S[h_S]\Big| \geq \epsilon\Big) \tag{2}$$

can no longer be bounded using Hoeffding's inequality directly as above, since $\mathrm{er}_D[h_S]$ is now a random variable that depends on the random sample S. We will see how the notion of uniform convergence can help us solve this problem when the learning algorithm selects h_S from a function class $\mathcal{H}$ of limited capacity. Specifically, if h_S is selected from $\mathcal{H}$, then

$$\mathbf{P}_{S\sim D^m}\Big(\Big|\mathrm{er}_D[h_S] - \mathrm{er}_S[h_S]\Big| \geq \epsilon\Big) \;\leq\; \mathbf{P}_{S\sim D^m}\Big(\sup_{h\in\mathcal{H}}\Big|\mathrm{er}_D[h] - \mathrm{er}_S[h]\Big| \geq \epsilon\Big)\,. \tag{3}$$

Thus if we can obtain an appropriate upper bound on the latter probability, then with high probability, *all* functions in $\mathcal{H}$ *uniformly* will have generalization errors close to their empirical errors on S (and in particular, h_S will have generalization error $\mathrm{er}_D[h_S]$ close to $\mathrm{er}_S[h_S]$). We will see how to bound the latter probability for function classes $\mathcal{H}$ of limited capacity. Let's start first with the simple case of a finite function class $\mathcal{H}$.

[1] Note that the notation h_m or $\hat{h}_m$ is often used in the literature to denote the (random) output of a learning algorithm when trained on a (random) training sample $S \sim D^m$. As discussed in class, we will find it convenient to use the notation h_S that makes explicit the dependence of the learned function on the input sample S; this notation will be especially helpful in later parts of the course (e.g. when we talk about algorithmic stability).

[2] For a brief review of confidence bounds/intervals, see Appendix A.

[3] For a brief review of some basic concentration inequalities, see Appendix B.

2 Uniform Convergence in a Finite Function Class $\mathcal{H}$

If $\mathcal{H}$ is finite, the supremum in Eq. (3) becomes a maximum, and we have

$$\begin{aligned}
\mathbf{P}_{S\sim D^m}\Big(\big|\mathrm{er}_D[h_S]-\mathrm{er}_S[h_S]\big|\geq\epsilon\Big) &\leq \mathbf{P}_{S\sim D^m}\Big(\max_{h\in\mathcal{H}}\big|\mathrm{er}_D[h]-\mathrm{er}_S[h]\big|\geq\epsilon\Big) && (4)\\
&= \mathbf{P}_{S\sim D^m}\Bigg(\bigcup_{h\in\mathcal{H}}\Big\{\big|\mathrm{er}_D[h]-\mathrm{er}_S[h]\big|\geq\epsilon\Big\}\Bigg) && (5)\\
&\leq \sum_{h\in\mathcal{H}}\mathbf{P}_{S\sim D^m}\Big(\big|\mathrm{er}_D[h]-\mathrm{er}_S[h]\big|\geq\epsilon\Big)\,, \quad \text{by union bound} && (6)\\
&\leq \sum_{h\in\mathcal{H}} 2\,e^{-2m\epsilon^2}\,, \quad \text{by Hoeffding's inequality} && (7)\\
&= 2|\mathcal{H}|\,e^{-2m\epsilon^2}\,. && (8)
\end{aligned}$$

Equivalently, with probability at least $1-\delta$ (over the draw of $S\sim D^m$), we have

$$\mathrm{er}_S[h_S]-\sqrt{\frac{\ln|\mathcal{H}|+\ln(\frac{2}{\delta})}{2m}} \;\leq\; \mathrm{er}_D[h_S] \;\leq\; \mathrm{er}_S[h_S]+\sqrt{\frac{\ln|\mathcal{H}|+\ln(\frac{2}{\delta})}{2m}}\,. \tag{9}$$

Typically, one cares mainly about obtaining (high confidence) *upper* bounds on $\mathrm{er}_D[h_S]$; we will follow this convention below.

The above gives a simple uniform convergence bound that can be used to bound (with high confidence) the generalization error of a function learned from a finite function class. However in many learning situations, the function learned by an algorithm is selected from an infinite function class (this is the case, for example, when learning a linear classifier over $\mathbb{R}^n$ using the SVM algorithm); in such situations, the above bound does not apply. Next we will see how to obtain a uniform convergence bound for function classes that are infinite but have limited capacity. For this we will need the notion of *growth function*; the related notion of *VC-entropy* is also of interest.

3 Growth Function and VC-Entropy

We will use the notation x_1^m for a sequence $(x_1,\ldots,x_m)\in\mathcal{X}^m$, and for a class $\mathcal{H}$ of functions $h:\mathcal{X}\to\mathcal{Y}$ (denoted $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$), we will denote by $\mathcal{H}_{|x_1^m}$ the *restriction* of $\mathcal{H}$ to x_1^m:

$$\mathcal{H}_{|x_1^m}=\{(h(x_1),\ldots,h(x_m))\mid h\in\mathcal{H}\}\,. \tag{10}$$

For a class of binary-valued functions $\mathcal{H}\subseteq\{-1,1\}^{\mathcal{X}}$, define the *growth function* $\Pi_{\mathcal{H}}:\mathbb{N}\to\mathbb{N}$ associated with $\mathcal{H}$ as follows:[4]

$$\Pi_{\mathcal{H}}(m)=\max_{x_1^m\in\mathcal{X}^m}\big|\mathcal{H}_{|x_1^m}\big|\,. \tag{11}$$

The growth function of $\mathcal{H}$ evaluated at m, $\Pi_{\mathcal{H}}(m)$, is sometimes referred to as the *m-th shatter coefficient* of $\mathcal{H}$.

Clearly, for a finite function class $\mathcal{H}$, $\Pi_{\mathcal{H}}(m)\leq|\mathcal{H}|$; in general, $\Pi_{\mathcal{H}}(m)\leq 2^m$. (In the next lecture we will see tighter upper bounds on $\Pi_{\mathcal{H}}(m)$ for certain classes $\mathcal{H}$.)

A related notion is that of the *VC-entropy* of $\mathcal{H}$ w.r.t. a probability distribution μ over $\mathcal{X}$, which we will denote by $\text{VC-entropy}_{\mathcal{H},\mu}:\mathbb{N}\to[0,\infty)$:[5]

$$\text{VC-entropy}_{\mathcal{H},\mu}(m)=\log_2\mathbf{E}_{x_1^m\sim\mu^m}\Big[\big|\mathcal{H}_{|x_1^m}\big|\Big]\,. \tag{12}$$

Clearly, $\text{VC-entropy}_{\mathcal{H},\mu}(m)\leq\log_2\Pi_{\mathcal{H}}(m)$.

[4] Sometimes, the term 'growth function' is used to refer to the logarithm of the function $\Pi_{\mathcal{H}}$ defined here, but the definition given here is more standard in the literature.

[5] VC here stands for Vapnik and Chervonenkis, who pioneered the technique of uniform covergence. In the next few lectures, we will see the related notions of VC-dimension and metric entropy.

4 Uniform Convergence in a General/Infinite Function Class $\mathcal{H}$

Consider now a general (possibly infinite) function class $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$. We will obtain a uniform convergence result for $\mathcal{H}$ by effectively reducing $\mathcal{H}$ to a finite size; this will lead to a result similar to that obtained for the finite case above, but with $|\mathcal{H}|$ replaced with $\Pi_{\mathcal{H}}(2m)$.

Theorem 4.1 (Vapnik and Chervonenkis, 1971)**.** Let $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$. Let D be any distribution on $\mathcal{X} \times \{-1,1\}$. For any $\epsilon > 0$,

$$\mathbf{P}_{S\sim D^m}\left(\sup_{h\in\mathcal{H}}\Big|\mathrm{er}_D[h] - \mathrm{er}_S[h]\Big| \geq \epsilon\right) \;\leq\; 4\Pi_{\mathcal{H}}(2m)\, e^{-m\epsilon^2/8}\,. \tag{13}$$

Proof. The proof consists of 4 main steps: (1) symmetrization through a ghost sample; (2) introduction of swapping permutations; (3) reduction to a finite class; and (4) an application of Hoeffding's inequality to bound a probability involving random swapping permutations.

Step 1: Symmetrization. We first reduce the probability above to a probability involving two samples, $S, \tilde{S}$, both drawn independently according to D^m; in particular we claim that for $m\epsilon^2 \geq 2$,

$$\mathbf{P}_{S\sim D^m}\left(\sup_{h\in\mathcal{H}}\Big|\mathrm{er}_D[h] - \mathrm{er}_S[h]\Big| \geq \epsilon\right) \;\leq\; 2\,\mathbf{P}_{(S,\tilde{S})\sim D^m\times D^m}\left(\sup_{h\in\mathcal{H}}\Big|\mathrm{er}_S[h] - \mathrm{er}_{\tilde{S}}[h]\Big| \geq \frac{\epsilon}{2}\right). \tag{14}$$

To see this, for each S, let $h_S^{\mathrm{bad}} \in \mathcal{H}$ be a function for which $\big|\mathrm{er}_D[h_S^{\mathrm{bad}}] - \mathrm{er}_S[h_S^{\mathrm{bad}}]\big| \geq \epsilon$ if such a function exists, and any fixed function in $\mathcal{H}$ otherwise. Then

$$\mathbf{P}_{(S,\tilde{S})\sim D^m\times D^m}\left(\sup_{h\in\mathcal{H}}\Big|\mathrm{er}_S[h] - \mathrm{er}_{\tilde{S}}[h]\Big| \geq \frac{\epsilon}{2}\right) \tag{15}$$

$$\geq\; \mathbf{P}_{(S,\tilde{S})\sim D^m\times D^m}\left(\Big|\mathrm{er}_S[h_S^{\mathrm{bad}}] - \mathrm{er}_{\tilde{S}}[h_S^{\mathrm{bad}}]\Big| \geq \frac{\epsilon}{2}\right) \tag{16}$$

$$\geq\; \mathbf{P}_{(S,\tilde{S})\sim D^m\times D^m}\left(\Big\{\Big|\mathrm{er}_S[h_S^{\mathrm{bad}}] - \mathrm{er}_D[h_S^{\mathrm{bad}}]\Big| \geq \epsilon\Big\} \bigcap \Big\{\Big|\mathrm{er}_{\tilde{S}}[h_S^{\mathrm{bad}}] - \mathrm{er}_D[h_S^{\mathrm{bad}}]\Big| \leq \frac{\epsilon}{2}\Big\}\right) \tag{17}$$

$$=\; \mathbf{E}_{S\sim D^m}\left[\mathbf{1}\left(\Big|\mathrm{er}_S[h_S^{\mathrm{bad}}] - \mathrm{er}_D[h_S^{\mathrm{bad}}]\Big| \geq \epsilon\right)\cdot \mathbf{P}_{\tilde{S}|S}\left(\Big|\mathrm{er}_{\tilde{S}}[h_S^{\mathrm{bad}}] - \mathrm{er}_D[h_S^{\mathrm{bad}}]\Big| \leq \frac{\epsilon}{2}\right)\right]. \tag{18}$$

Now the conditional probability inside can be bounded using Chebyshev's inequality:

$$\mathbf{P}_{\tilde{S}|S}\left(\Big|\mathrm{er}_{\tilde{S}}[h_S^{\mathrm{bad}}] - \mathrm{er}_D[h_S^{\mathrm{bad}}]\Big| \leq \frac{\epsilon}{2}\right) \;\geq\; 1 - \frac{\mathbf{Var}_{\tilde{S}|S}\left[\mathrm{er}_{\tilde{S}}[h_S^{\mathrm{bad}}]\right]}{\frac{\epsilon^2}{4}}\,. \tag{19}$$

Since $\tilde{S}|S \sim D^m$ and $\mathrm{er}_{\tilde{S}}[h_S^{\mathrm{bad}}]$ is $\frac{1}{m}$ times a Binomial random variable with parameters $(m, \mathrm{er}_D[h_S^{\mathrm{bad}}])$, we have $\mathbf{Var}_{\tilde{S}|S}\left[\mathrm{er}_{\tilde{S}}[h_S^{\mathrm{bad}}]\right] = \frac{\mathrm{er}_D[h_S^{\mathrm{bad}}](1-\mathrm{er}_D[h_S^{\mathrm{bad}}])}{m} \leq \frac{1}{4m}$. This gives

$$\mathbf{P}_{\tilde{S}|S}\left(\Big|\mathrm{er}_{\tilde{S}}[h_S^{\mathrm{bad}}] - \mathrm{er}_D[h_S^{\mathrm{bad}}]\Big| \leq \frac{\epsilon}{2}\right) \;\geq\; 1 - \frac{1}{m\epsilon^2} \;\geq\; \frac{1}{2} \tag{20}$$

whenever $m\epsilon^2 \geq 2$. Thus we get

$$\mathbf{P}_{(S,\tilde{S})\sim D^m\times D^m}\left(\sup_{h\in\mathcal{H}}\Big|\mathrm{er}_S[h] - \mathrm{er}_{\tilde{S}}[h]\Big| \geq \frac{\epsilon}{2}\right) \tag{21}$$

$$\geq\; \frac{1}{2}\mathbf{P}_{S\sim D^m}\left(\Big|\mathrm{er}_S[h_S^{\mathrm{bad}}] - \mathrm{er}_D[h_S^{\mathrm{bad}}]\Big| \geq \epsilon\right) \tag{22}$$

$$=\; \frac{1}{2}\mathbf{P}_{S\sim D^m}\left(\sup_{h\in\mathcal{H}}\Big|\mathrm{er}_S[h] - \mathrm{er}_D[h]\Big| \geq \epsilon\right), \quad \text{by definition of } h_S^{\mathrm{bad}}. \tag{23}$$

Step 2: Swapping permutations. Let Γ_{2m} be the set of all permutations on $[2m] = \{1,\ldots,2m\}$ which swap some of the elements in the first half with the corresponding elements in the second half:[6,7]

$$\Gamma_{2m} = \{\sigma \in S_{2m} \mid \sigma(i) = i \text{ or } m+i\ \forall i \in [m]; \sigma(i) = j \Leftrightarrow \sigma(j) = i\ \forall i,j \in [2m]\}\,.$$

[6] We will frequently use the notation $[n] = \{1,\ldots,n\}$.

[7] S_{2m} here refers to the permutation group over $2m$ elements.

Clearly $|\Gamma_{2m}| = 2^m$. For $\sigma \in \Gamma_{2m}$ and $S = ((x_1, y_1), \ldots, (x_{2m}, y_{2m})) \in (\mathcal{X} \times \{-1, 1\})^{2m}$, denote $\sigma(S) = ((x_{\sigma(1)}, y_{\sigma(1)}), \ldots, (x_{\sigma(2m)}, y_{\sigma(2m)}))$, and for any subsequence S' of S, denote by $\sigma(S')$ the corresponding subsequence of $\sigma(S)$. Then clearly, if $(S, \tilde{S}) \sim D^m \times D^m$, then for any $\sigma \in \Gamma_{2m}$, the random variable $\sup_{h\in\mathcal{H}} |\mathrm{er}_S[h] - \mathrm{er}_{\tilde{S}}[h]|$ has the same distribution as the random variable $\sup_{h\in\mathcal{H}} |\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]|$. Therefore we have[8]

$$\mathbf{P}_{(S,\tilde{S})\sim D^m\times D^m}\left(\sup_{h\in\mathcal{H}}\left|\mathrm{er}_S[h] - \mathrm{er}_{\tilde{S}}[h]\right| \geq \frac{\epsilon}{2}\right) \tag{24}$$

$$= \frac{1}{2^m}\sum_{\sigma\in\Gamma_{2m}} \mathbf{P}_{(S,\tilde{S})\sim D^m\times D^m}\left(\sup_{h\in\mathcal{H}}\left|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]\right| \geq \frac{\epsilon}{2}\right) \tag{25}$$

$$= \frac{1}{2^m}\sum_{\sigma\in\Gamma_{2m}} \mathbf{E}_{(S,\tilde{S})\sim D^m\times D^m}\left[\mathbf{1}\left(\sup_{h\in\mathcal{H}}\left|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]\right| \geq \frac{\epsilon}{2}\right)\right] \tag{26}$$

$$= \mathbf{E}_{(S,\tilde{S})\sim D^m\times D^m}\left[\frac{1}{2^m}\sum_{\sigma\in\Gamma_{2m}} \mathbf{1}\left(\sup_{h\in\mathcal{H}}\left|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]\right| \geq \frac{\epsilon}{2}\right)\right] \tag{27}$$

$$= \mathbf{E}_{(S,\tilde{S})\sim D^m\times D^m}\left[\mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\sup_{h\in\mathcal{H}}\left|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]\right| \geq \frac{\epsilon}{2}\right)\right] \tag{28}$$

$$\leq \sup_{(S,\tilde{S})\in(\mathcal{X}\times\{-1,1\})^{2m}}\left[\mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\sup_{h\in\mathcal{H}}\left|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]\right| \geq \frac{\epsilon}{2}\right)\right] \tag{29}$$

Step 3: Reduction to a finite class. Fix any $(S, \tilde{S}) \in (\mathcal{X} \times \{-1, 1\})^{2m}$, and consider the random draw of $\sigma \in \Gamma_{2m}$. For each $h \in \mathcal{H}$, the quantity $|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]|$ is a random variable; clearly, the number of distinct random variables of this form (as h varies over $\mathcal{H}$) is at most $\Pi_{\mathcal{H}}(2m)$. Therefore, the supremum in the probability above can be written as a maximum over $\Pi_{\mathcal{H}}(2m)$ random variables, and so by the union bound one gets

$$\mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\sup_{h\in\mathcal{H}}\left|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]\right| \geq \frac{\epsilon}{2}\right) \tag{30}$$

$$\leq \Pi_{\mathcal{H}}(2m)\sup_{h\in\mathcal{H}}\mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\left|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]\right| \geq \frac{\epsilon}{2}\right). \tag{31}$$

Step 4: Hoeffding's inequality. Fix $h \in \mathcal{H}$. The probability on the right above can be bounded by viewing the random selection of $\sigma \in \Gamma_{2m}$ as a selection of m independent, uniform $\{-1, 1\}$-valued (Rademacher) random variables r_i $(i = 1, \ldots, m)$ and applying Hoeffding's inequality:

$$\mathbf{P}_{\sigma\in\Gamma_{2m}}\left(\left|\mathrm{er}_{\sigma(S)}[h] - \mathrm{er}_{\sigma(\tilde{S})}[h]\right| \geq \frac{\epsilon}{2}\right) \tag{32}$$

$$= \mathbf{P}_{\mathbf{r}\in\{-1,1\}^m}\left(\frac{1}{m}\left|\sum_{i=1}^m r_i\Big(\mathbf{1}(h(x_i)\neq y_i) - \mathbf{1}(h(\tilde{x}_i)\neq\tilde{y}_i)\Big)\right| \geq \frac{\epsilon}{2}\right) \tag{33}$$

$$\leq 2\,e^{-m\epsilon^2/8}. \tag{34}$$

Putting everything together yields the desired result for $m\epsilon^2 \geq 2$; for $m\epsilon^2 < 2$, the result holds trivially. □

In the above result, the growth function $\Pi_{\mathcal{H}}(2m)$ serves to measure the 'capacity' of $\mathcal{H}$ in place of the size $|\mathcal{H}|$ as we saw in the finite case. Note that the result is meaningful only if $\Pi_{\mathcal{H}}(2m)$ grows sufficiently slowly with m in comparison with $e^{m\epsilon^2/8}$; we will have more to say on this in the next lecture. As in the finite case, we can re-write the above result as follows (focusing now only on upper bounding $\mathrm{er}_D[h_S]$): if h_S is selected from $\mathcal{H}$, then for any $0 < \delta < 1$, with probability at least $1 - \delta$ (over the draw of $S \sim D^m$),

$$\mathrm{er}_D[h_S] \;\leq\; \mathrm{er}_S[h_S] + \sqrt{\frac{8\left(\ln\Pi_{\mathcal{H}}(2m) + \ln(\frac{4}{\delta})\right)}{m}}\,. \tag{35}$$

[8] Here and throughout the course, we will use the notation $\mathbf{P}_{x\in U}$ and $\mathbf{E}_{x\in U}$ to denote the probability or expectation over the draw of x uniformly at random from a set U.

We note that it is possible to tighten the above result using the VC-entropy; in particular, we can avoid the last inequality in Step 2 of the proof, and for each $(S, \tilde{S})$, bound the probability over $\sigma \in \Gamma_{2m}$ in Step 3 using $|\mathcal{H}_{|(x_1^m, \tilde{x}_1^m)}|$ in place of $\Pi_{\mathcal{H}}(2m)$; this will yield a similar uniform convergence result with $\Pi_{\mathcal{H}}(2m)$ replaced by $2^{\text{VC-entropy}_{\mathcal{H},\mu}(2m)}$, where μ here corresponds to the marginal distribution of D over $\mathcal{X}$. This leads to a *distribution-dependent* measure of capacity, which can potentially be tighter; however it is typically difficult to estimate the VC-entropy of a function class w.r.t. an unknown distribution, and hence one usually falls back on the bound in terms of the growth function, $2^{\text{VC-entropy}_{\mathcal{H},\mu}(2m)} \leq \Pi_{\mathcal{H}}(2m)$. However we will see similar ideas later (in a few lectures from now) when we talk about Rademacher averages, which provide an alternative distribution-dependent measure of the capacity (or complexity) of a function class $\mathcal{H}$, and which often *can* be estimated reliably; the proof there will rely on a more powerful concentration inequality due to McDiarmid, but will be simpler in structure and will yield a tighter uniform convergence result.

5 Next Lecture

In the next lecture, we will meet Sauer's lemma, which will allow us to bound the growth function of a function class $\mathcal{H}$ in terms of the VC-dimension of $\mathcal{H}$, a combinatorial parameter associated with $\mathcal{H}$.

A Confidence Bounds/Intervals

Confidence bounds/intervals are used for estimating the values of unknown quantities in statistics.

For example, say that a random variable X is known to be distributed according to a normal distribution of unit variance, $\mathcal{N}(\mu, 1)$, but we don't know its mean μ. We observe a value x drawn from this distribution. What can we say about μ based on this observation? We know from standard normal CDF tables that $\mathbf{P}(|X - \mu| \geq 1.96) \leq 0.05$. Thus, with probability at least 0.95 over the draw of X, $\mu \in [X - 1.96, X + 1.96]$; this is called a (95%) *confidence interval* for μ (note that the *interval* here is random). On observing a value x, we can then say that with *confidence* 0.95, $\mu \in [x - 1.96, x + 1.96]$; the interpretation is that, if x was truly drawn from the assumed distribution, then the chances of the constructed interval not containing μ would be at most 0.05. Sometimes one is interested in only upper (or lower) bounds on an unknown quantity, e.g. in the above example, we have that $\mathbf{P}(X - \mu \leq -1.96) \leq 0.025$; thus with probability at least 0.975, $\mu \leq X + 1.96$, which gives a (97.5%) *confidence bound* for μ.

A similar approach as above can be used for estimating the mean of a random variable X when a bound on the probability $\mathbf{P}(|X - \mathbf{E}[X]| \geq \epsilon)$ (for appropriate $\epsilon > 0$) is available via other means, for example via a concentration inequality (see Appendix B below).

B Some Basic Concentration Inequalities

Concentration inequalities (also called tail bounds) give us an idea of how tightly concentrated a random variable is around its mean. Some basic concentration inequalities are included below; proofs are standard and can be found in most probability texts.

Theorem B.1 (Markov's inequality)**.** Let X be a non-negative random variable. For any $t > 0$,

$$\mathbf{P}\left(X \geq t\right) \leq \frac{\mathbf{E}[X]}{t}\,.$$

In particular, for any $k > 0$, this gives $\mathbf{P}\left(X \geq k\mathbf{E}[X]\right) \leq \frac{1}{k}$.

Theorem B.2 (Chebyshev's inequality)**.** Let X be a random variable. For any $\epsilon > 0$,

$$\mathbf{P}\left(|X - \mathbf{E}[X]| \geq \epsilon\right) \leq \frac{\mathbf{Var}[X]}{\epsilon^2}\,.$$

Theorem B.3 (Hoeffding's inequality; Hoeffding, 1963)**.** Let $X_1, \ldots, X_n$ be independent random variables, with X_i taking values in $[a_i, b_i]$. For any $\epsilon > 0$,

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geq \epsilon\right) \leq e^{-2\epsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2} .$$

and

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \leq -\epsilon\right) \leq e^{-2\epsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2} .$$