GENERALIZATION BOUNDS

MARK D. REID — THE AUSTRALIAN NATIONAL UNIVERSITY

Synonyms: Sample Complexity, Inequalities

Definition

In the theory of statistical machine learning, a generalization bound—or, more precisely, a generalization error bound—is a statement about the predictive performance of a learning algorithm or class of algorithms. Here, a learning algorithm is viewed as a procedure that takes some finite training sample of labelled instances as input and returns a hypothesis regarding the labels of all instances, including those which may not have appeared in the training sample. Assuming labelled instances are drawn from some fixed distribution, the quality of a hypothesis can be measured in terms of its risk—that is, its incompatibility with the distribution. The performance of a learning algorithm can then be expressed in terms of the expected risk of its hypotheses given randomly generated training samples.

Under these assumptions a generalization bound is a theorem which holds for any distribution and states that, with high probability, applying the learning algorithm to a randomly drawn sample will result in a hypothesis with risk no greater than some value. This bounding value typically depends on the size of the training sample, an empirical assessment of the risk of the hypothesis on the training sample as well as the "richness" or "capacity" of the class of predictors that can be output by the learning algorithm.

Motivation and Background

Suppose we have built an email classifier and then collected a random sample of email labelled as "spam" or "not spam" to test it on. We notice that that the classifier incorrectly labels 5% of the sample. What can be said about the accuracy of this classifier when it is applied to new, previously unseen email? If we make the reasonable assumption that the mistakes made on future emails are independent of mistakes made on the sample, basic results from statistics tell us that the classifier's true error rate will also be around 5%.

Now suppose that instead of building a classifier by hand we use a learning algorithm to *infer* one from the sample. What can be said about the future error rate of the inferred classifier if it also misclassifies 5% of the training sample? In general, the answer is "nothing" since we can no longer assume future mistakes are independent of those made on the training sample. As an extreme case, consider a learning algorithm that outputs a classifier that just "memorizes" the training sample—that is, predicts labels for email in the sample according to what appears in the sample—and predicts randomly otherwise. Such a classifier will have a 0% error rate on the sample, however if most future email does not appear in the training sample the classifier will have a true error rate around 50%.

To avoid the problem of memorizing or over-fitting the training data it is necessary to restrict the "flexibility" of the hypotheses a learning algorithm can output. Doing so forces predictions made off the training set to be related to those made on the training set so that some form of generalization takes place. However, doing this can limit the ability of the learning algorithm to output a hypothesis with small risk. Thus, there is a classic bias and variance trade-off: the bias being the limits placed on how flexible the hypotheses can be versus the variance between the training and true error rates.

By quantifying the notion of hypothesis flexibility in various ways, generalization bounds provide inequalities that show how the flexibility and empirical error rate can be traded off to control the true error rate. Importantly, these statements are typically probabilistic but distribution-independent—that is, they hold for nearly all sets of training data drawn from a fixed but unknown distribution. When such a bound holds for a learning algorithm it means that, unless the choice of training sample was very unlucky, we can be confident that some form of generalization will take place. The first results of this kind were established by Vapnik and Chervonenkis about 40 years ago [15] and the measure of hypothesis flexibility they introduced—the VC dimension (see below)—now bears their initials. A similar style of results were obtained independently by Valiant in 1984 in the Probably Approximately Correct, or PAC learning framework [14]. These two lines of work were drawn together by Blumer *et al.* in 1989 [5] and now form the basis of what is known today as statistical learning theory.

Details

For simplicity we restrict our attention to generalization bounds for binary classification problems such as the spam classification example above. In this setting *instances* (*e.g.*, email) from a set $\mathcal{X}$ are associated with *labels* from a set $\mathcal{Y} = \{-1, 1\}$ (*e.g.*, indicating not-spam/spam) and an *example* $z = (x, y)$ is a labelled instance from $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. The association of instances to labels is assumed to be governed by some unknown distribution P over $\mathcal{Z}$.

A *hypothesis* h is a function that assigns labels $h(x) \in \mathcal{Y}$ to instances. The quality of a hypothesis is assessed via a *loss* function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ which assigns penalty $\ell(y, y')$ when h predicts the label $y' = h(x)$ for the example (x, y). For convenience, we will often combine the loss and hypothesis evaluation on an example $z = (x, y)$ by defining $\ell_h(z) = \ell(y, h(x))$. When examples are sampled from P the expected penalty, or *risk*

$$L_P(h) := \mathbb{E}_P\left[\ell_h(z)\right]$$

can be interpreted as a measure of how well h models the distribution P. A loss that is prevalent in classification is the *0-1 loss* $\ell^{0-1}(y, y') = [\![y \neq y']\!]$ where $[\![p]\!]$ is the indicator function for the predicate p. This loss simply assigns a penalty of 1 for an incorrect prediction and 0 otherwise. The associated 0-1 risk for h is the probability the prediction $h(x)$ disagrees with a randomly drawn sample (x, y) from P. Unless stated otherwise, the bounds discussed below are for the 0-1 loss only but, with care, can usually be made to hold with more general losses also.

Once a loss is specified, the goal of a learning algorithm is to produce a low risk hypothesis based on a finite number of examples. Formally, a *learning algorithm* $\mathcal{A}$ is a procedure that takes a *training sample* $\mathbf{z} = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ as input and returns a hypothesis $h = \mathcal{A}(\mathbf{z})$ with an associated *empirical risk*

$$\hat{L}_{\mathbf{z}}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell_h(z_i).$$

In order to relate the empirical and true risks, a common assumption made in statistical learning theory is that the examples are drawn independently from P. In this case, a sample $\mathbf{z} = (z_1, \ldots, z_n)$ is a random variable from the product distribution P^n over $\mathcal{Z}^n$. Since the sample can be of arbitrary but finite size a learning algorithm can be viewed as a function $\mathcal{A} : \bigcup_{n=1}^{\infty} \mathcal{Z}^n \to \mathcal{H}$ where $\mathcal{H}$ is the algorithm's hypothesis space.

A generalization bound typically comprises several quantities: an empirical estimate of a hypothesis's performance $\hat{L}_{\mathbf{z}}(h)$; the actual (and unknown) risk of the hypothesis $L_P(h)$; a confidence term $\delta \in [0, 1]$; and some measure of the flexibility or *complexity* C of the hypotheses that can be output by learning algorithm. The majority of the bounds found in the literature fit the following template.

> **A Generic Generalization Bound** Let $\mathcal{A}$ be a learning algorithm, P some unknown distribution over $\mathcal{X} \times \mathcal{Y}$, and $\delta > 0$. Then, with probability at least $1 - \delta$ over randomly drawn samples $\mathbf{z}$ from P^n, the hypothesis $h = \mathcal{A}(\mathbf{z})$ has risk $L_P(h)$ no greater than $\hat{L}_{\mathbf{z}}(h) + \epsilon(\delta, C)$.

Of course, there are many variations, refinements and improvements of the bounds presented below and not all fit this template. The bounds discussed below are only intended to provide a survey of some of the key ideas and main results.

Basic Bounds. The penalties $\ell_h(z_i) := \ell(y_i, h(x_i))$ made by a fixed hypothesis h on a sample $\mathbf{z} = (z_1, \ldots, z_n)$ drawn from P^n are independent random variables. The law of large numbers guarantees (under some mild conditions) that their mean $\hat{L}_{\mathbf{z}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_h(z_i)$ converges to the true risk $L_P(h) = \mathbb{E}_P[\ell_h(z)]$ for h as the sample size increases and several inequalities from probability theory can be used to quantify this convergence. A key result is McDiarmid's inequality which can be used to bound the deviation of a function of independent random variables from its mean. Since the 0-1 loss takes values in $[0, 1]$, applying this result to the random variables $\ell_h(Z_i)$ gives

$$P^n\left(L_P(h) > \hat{L}_{\mathbf{z}}(h) + \epsilon\right) \le \exp\left(-2n\epsilon^2\right) \tag{1}$$

We can invert this and obtain an upper bound for the true risk that will hold on a given proportion of samples. That is, if we want $L_P(h) \le \hat{L}_{\mathbf{z}}(h) + \epsilon$ to hold on at least $1 - \delta$ of the time on randomly drawn samples we can solve $\delta = \exp(-2n\epsilon^2)$ for ϵ and obtain $\epsilon = \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$ so that

$$P^n\left(L_P(h) \le \hat{L}_{\mathbf{z}}(h) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\right) \ge 1 - \delta. \tag{2}$$

This simple bound lays the foundation for many of the subsequent bounds discussed below and is the reason for the ubiquity of the $\sqrt{\frac{\ln \frac{1}{\delta}}{n}}$-like terms.

A crucial observation to make about the above bound is that while it holds for any hypothesis h it does *not* hold for all $h \in \mathcal{H}$ *simultaneously.* That is, the samples for which the bounds hold for h_1 may be completely different to those which make the bound hold for h_2. Since a generalization bound must hold for all possible hypotheses output by a learning algorithm we need to extend the above analysis by exploiting additional properties of the hypothesis space $\mathcal{H}$.

In the simple case when there are only finitely many hypothesis we use the *union bound.* This states that for any distribution P and any finite or countably infinite sequence of events $A_1, A_2 \ldots$ we have $P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i)$. For $\mathcal{H} = \{h_1, \ldots, h_m\}$ we consider the events $Z_h = \{\mathbf{z} \in \mathcal{Z}^n : L_P(h) > \hat{L}_{\mathbf{z}}(h) + \epsilon\}$ when samples of size n give empirical risks for h that are least ϵ smaller than its true risk. Using the union bound and (1) on these events gives

$$P^n\left(\bigcup_{h\in\mathcal{H}} Z_h(n,\epsilon)\right) \leq \sum_{i=1}^{m} P^n\left(Z_h(n,\epsilon)\right) = m.\exp(-2n\epsilon^2).$$

This is a bound on the probability of drawing a training sample from P^n such that *every* hypothesis has a true risk that is ϵ larger than its empirical risk. Inverting this inequality by setting $\delta = m\exp(-2n\epsilon^2)$ yields the following bound.

Finite Class Bound Suppose $\mathcal{A}$ has finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_m\}$. Then with probability at least $1-\delta$ over draws of $\mathbf{z}$ from P^n the hypothesis $h = \mathcal{A}(\mathbf{z})$ satisfies

$$L_P(h) \leq \hat{L}_{\mathbf{z}}(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{2n}}. \tag{3}$$

It is instructive to compare this to the single hypothesis bound in (2) and note the bound is weakened by the additional term $\ln|\mathcal{H}|$.

Since the union bound also holds for countable sets of events this style of bound can be extended from finite hypothesis classes to countable ones. To do this requires a slight modification of the above argument and the introduction of a distribution π over a countable hypothesis space $\mathcal{H} = \{h_1, h_2, \ldots\}$ which is chosen before any samples are seen. This distribution can be interpreted as a prior belief or preference over the hypotheses in $\mathcal{H}$. Letting $\delta(h) = \delta.\pi(h)$ in the bound (2) implies that for each $h \in \mathcal{H}$ we have

$$P^n\left(L_P(h) > \hat{L}_{\mathbf{z}}(h) + \sqrt{\frac{\ln\frac{1}{\delta.\pi(h)}}{2n}}\right) < \delta.\pi(h).$$

Thus, applying the countable union bound to the union of these events over all of $\mathcal{H}$, and noting that $\sum_{h\in\mathcal{H}} \delta.\pi(h) = \delta$ since π is a distribution over $\mathcal{H}$, gives use the following bound:

Countable Class Bound Suppose μ is a probability distribution over a finite or countably infinite hypothesis space $\mathcal{H}$. Then with probability at least $1-\delta$ over draws of $\mathbf{z}$ from P^n the following bound holds for all $h \in \mathcal{H}$

$$L_P(h) \leq \hat{L}_{\mathbf{z}}(h) + \sqrt{\frac{\ln\frac{1}{\pi(h)} + \ln\frac{1}{\delta}}{2n}}. \tag{4}$$

Although the finite and countable class bounds are proved using very similar techniques (indeed, the former can be derived from the latter by choosing $\pi(h) = \frac{1}{|\mathcal{H}|}$), they differ in the type of penalty they introduce for simultaneously bounding all the hypotheses in $\mathcal{H}$. In (3) the penalty $\ln|\mathcal{H}|$ is purely a function of the size of the class whereas in (4) the penalty $\ln\frac{1}{\pi(h)}$ varies with h. These two different styles of bound can be seen as templates for the two main classes of bounds discussed below: the hypothesis-independent bounds of the next section and the hypothesis-dependent bounds in the section on PAC-Bayesian bounds. The main conceptual leap from here is the extension of the arguments above to non-countable hypothesis classes.

Class Complexity Bounds. A key result in extending the notion of size or complexity in the above bounds to more general classes of hypotheses is the *symmetrization lemma*. Intuitively, it is based on the observation that if the empirical risks for different samples are frequently near the true risk then they will also be near each other. Formally, it states that for any $\epsilon > 0$ such that $n\epsilon^2 \geq 2$ we have

$$P^n\left(\sup_{h\in\mathcal{H}}\left|L_P(h) - \hat{L}_{\mathbf{z}}(h)\right| > \epsilon\right) \leq 2P^{2n}\left(\sup_{h\in\mathcal{H}}\left|\hat{L}_{\mathbf{z}'}(h) - \hat{L}_{\mathbf{z}}(h)\right| > \frac{\epsilon}{2}\right). \quad (5)$$

Thus, to obtain a bound on the difference between empirical and true risk it suffices to bound the difference in empirical risks on two independent samples $\mathbf{z}$ and $\mathbf{z}'$, both drawn from P^n. This is useful since the maximum difference $\sup_{h\in\mathcal{H}}\left|\hat{L}_{\mathbf{z}'}(h) - \hat{L}_{\mathbf{z}}(h)\right|$ is much easier to handle than the difference involving $L_P(h)$ as the former term only evaluates losses on the points in $\mathbf{z}$ and $\mathbf{z}'$ while the latter takes into account the entire space $\mathcal{Z}$.

To study these restricted evaluations, we define the restriction of a function class $\mathcal{F}$ to the sample $\mathbf{z}$ by $\mathcal{F}_{\mathbf{z}} = \{(f(z_1),\ldots,f(z_n)) : f \in \mathcal{F}\}$. Since the empirical risk $\hat{L}_{\mathbf{z}}(h) = \frac{1}{n}\sum_{i=1}^n \ell_h(z_i)$ only depends on the values of the loss functions ℓ_h on samples from $\mathbf{z}$ we define the *loss class* $\mathcal{L} = \ell_{\mathcal{H}} = \{\ell_h : h \in \mathcal{H}\}$ and consider its restriction $\mathcal{L}_{\mathbf{z}}$ as well as the restriction $\mathcal{H}_{\mathbf{z}}$ of the hypothesis class it is built upon. As we will see, the measures of complexity of these two classes are closely related.

One such complexity measure is arrived at by examining the size of a restricted function class $\mathcal{F}_{\mathbf{z}}$ as the size of the sample $\mathbf{z}$ increases. The *growth function* or shattering coefficient for the function class $\mathcal{F}$ is defined as the maximum number of distinct values the vectors in $\mathcal{F}_{\mathbf{z}}$ can take given a sample of size n: $S_n(\mathcal{F}) = \sup_{\mathbf{z}\in\mathcal{Z}^n}|\mathcal{F}_{\mathbf{z}}|$. In the case of binary classification with a 0-1 loss, it is not hard to see that the growth functions for both $\mathcal{L}$ and $\mathcal{H}$ are equal, that is $S_n(\mathcal{L}) = S_n(\mathcal{H})$, and so they can be used interchangeably. Applying a union bound argument to (1) as in the previous bounds guarantees that $P^n\left(\sup_{h\in\mathcal{H}}\left|L_P(h) - \hat{L}_{\mathbf{z}}(h)\right| > \epsilon\right) \leq 2S_n(\mathcal{H})\exp(-n\epsilon^2/8)$ and by inversion we obtain the following generalization bound for arbitrary hypothesis classes $\mathcal{H}$:

> **Growth Function Bound** For all $\delta > 0$, a draw of $\mathbf{z}$ from P^n will, with probability at least $1-\delta$, satisfy for all $h \in \mathcal{H}$
>
> $$L_P(h) \leq \hat{L}_{\mathbf{z}}(h) + 2\sqrt{\frac{2\ln S_n(\mathcal{H}) + 2\ln\frac{2}{\delta}}{n}}. \quad (6)$$

One conclusion that can be immediately drawn from this bound is that the shattering coefficient must grow sub-exponentially for the bound to provide any meaningful guarantee. If the class $\mathcal{H}$ is so rich that hypotheses from it can fit all 2^n possible label combinations – that is, if $S_n(\mathcal{H}) = 2^n$ for all n – then the term $\sqrt{2\ln S_n(\mathcal{H})/n} > 1$ and so (6) just states $L_P(h) \leq 1$. Therefore, to get non-trivial bounds from (6) there needs to exist some value d for which $S_n(\mathcal{H}) < 2^n$ whenever $n > d$.

VC Dimension. This desired property of the growth function is exactly what is captured by the VC dimension $VC(\mathcal{H})$ of a hypothesis class $\mathcal{H}$. Formally, it is defined as $VC(\mathcal{H}) = \max\{n \in \mathbb{N} : S_n(\mathcal{H}) = 2^n\}$ and is infinite if no finite maximum exists. Whether or not the VC dimension is finite plays a central role in the consistency of empirical risk minimization techniques. Indeed, it is possible to show that using ERM on a hypothesis class $\mathcal{H}$ is consistent if and only if $VC(\mathcal{H}) < \infty$. This is partly due to *Sauer's lemma* which shows that when a hypothesis class $\mathcal{H}$ has finite VC dimension $VC(\mathcal{H}) = d_{\mathcal{H}} < \infty$ its growth function is eventually polynomial in the sample size. Specifically, for all $n \geq d_{\mathcal{H}}$ the growth function satisfies $S_n(\mathcal{H}) \leq \left(\frac{en}{d_{\mathcal{H}}}\right)^{d_{\mathcal{H}}}$. By substituting this result into the Growth Function Bound (6) we obtain the following bound which shows how the VC dimension plays a role that is analogous to the size a hypothesis class in the finite case.

> **VC Dimension Bound** Suppose $\mathcal{A}$ has hypothesis class $\mathcal{H}$ with finite VC dimension $d_{\mathcal{H}}$. Then with probability at least $1 - \delta$ over draws of $\mathbf{z}$ from P^n the hypothesis $h = \mathcal{A}(\mathbf{z})$ satisfies
>
> $$L_P(h) \leq \hat{L}_{\mathbf{z}}(h) + 2\sqrt{\frac{2d_{\mathcal{H}} \ln\left(\frac{2en}{d_{\mathcal{H}}}\right) + 2\ln\frac{2}{\delta}}{n}}. \tag{7}$$

There are many other bounds in the literature that are based on the VC dimension. See the Recommended Reading for pointers to these.

Rademacher Averages. Rademacher averages are a second kind of measure of complexity for uncountable function classes and can be used to derive more refined bounds than those above. These averages arise naturally by treating as a random variable the sample-dependent quantity $M_{\mathcal{F}}(\mathbf{z}) = \sup_{f\in\mathcal{F}} [\mathbb{E}_P[f] - \mathbb{E}_{\mathbf{z}}[f]]$. This is just the largest difference taken over all $f \in \mathcal{F}$ between its true mean $\mathbb{E}_P[f]$ and its empirical mean $\mathbb{E}_{\mathbf{z}}[f] := \frac{1}{|\mathbf{z}|}\sum_{i=1}^{|\mathbf{z}|} f(z_i)$. For a loss class $\mathcal{L} = \ell_{\mathcal{H}}$ a bound on this maximum difference—*e.g.*, $M_{\mathcal{L}}(\mathbf{z}) \leq B$—immediately gives a generalization bound of the form $L_P(h) \leq \hat{L}_{\mathbf{z}}(h) + B$. Since $M_{\mathcal{F}}(\mathbf{z})$ is a random variable, McDiarmid's inequality can be used to bound its value in terms of its expected value plus the usual $\sqrt{\frac{\ln\frac{1}{\delta}}{2n}}$ term. Applying symmetrization it can then be shown that this expected value satisfies

$$\mathbb{E}_{P^n}\left[M_{\mathcal{F}}(\mathbf{z})\right] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \rho_i\left(f(z_i') - f(z_i)\right)\right] \leq 2R_n(\mathcal{F})$$

where the right-hand expectation is taken over two independent samples $\mathbf{z}, \mathbf{z}' \sim P^n$ and the *Rademacher variables* $\rho_1, \ldots, \rho_n$. These are independent random variables, each with equal probability of taking the values -1 or 1, that give their name to the *Rademacher average*

$$R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \rho_i f(z_i)\right].$$

Intuitively, this quantity measures how well the functions in $\mathcal{F}$ can be chosen to align with randomly chosen labels ρ_i. The Rademacher averages for the loss class $\mathcal{L}$ and the hypothesis class $\mathcal{H}$ are closely related. For 0-1 loss, it can be shown they satisfy $R_n(\mathcal{L}) = \frac{1}{2}R_n(\mathcal{H})$.

Putting all the above steps together gives the following bounds.

> **Rademacher Bound** Suppose $\mathcal{A}$ has hypothesis class $\mathcal{H}$. Then with probability at least $1-\delta$ over draws of $\mathbf{z}$ from P^n the hypothesis $h = \mathcal{A}(\mathbf{Z})$ satisfies

$$L_P(h) \leq \hat{L}_{\mathbf{z}}(h) + R_n(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}. \tag{8}$$

This bound is qualitatively different to the Growth Function and VC bounds above as the Rademacher average term is *distribution-dependent* whereas the other complexity terms are purely a function of the hypothesis space. Indeed, it is possible to bound the Rademacher average in terms of the VC dimension and obtain the VC bound (7) from (8). Furthermore, the Rademacher average is closely related to the minimum empirical risk via $R_n(\mathcal{H}) = 1 - 2\mathbb{E}[\inf_{h\in\mathcal{H}} \hat{L}_{\mathbf{x},\rho}(h)]$ where $\hat{L}_{\mathbf{x},\rho}(h)$ is the empirical risk of h for the randomly labelled sample $\mathbf{z} = ((x_1,\rho_1),\ldots,(x_n,\rho_n))$. Thus, in principle, $R_n(\mathcal{H})$ could be estimated for a given learning problem using standard ERM methods.

The Rademacher bound can be further refined so that the complexity term is *data-dependent* rather than distribution-dependent. This is done by noting that the Rademacher average $R_n(\mathcal{F}) = \mathbb{E}\left[\hat{R}_{\mathbf{z}}(\mathcal{F})\right]$ where $\hat{R}_{\mathbf{z}}(\mathcal{F})$ is the *empirical Rademacher average* for $\mathcal{F}$ conditioned on the sample $\mathbf{z}$. Applying McDiarmid's inequality to the difference between $\hat{R}_{\mathbf{z}}(\mathcal{F})$ and its mean gives a sample-dependent bound:

> **Empirical Rademacher Bound** Under the same conditions as the Rademacher bound, the following holds with probability $1-\delta$:

$$L_P(h) \leq \hat{L}_{\mathbf{z}}(h) + \hat{R}_{\mathbf{z}}(\mathcal{H}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}. \tag{9}$$

PAC-Bayesian Bounds. All the bounds in the previous section provide bounds on deterministic hypotheses which include complexity terms that are functions of the entire hypothesis space. PAC-Bayesian bounds differ from these in two ways: they provide bounds on non-deterministic hypotheses—that is, labels may be predicted for instances stochastically; and their complexity terms are *hypothesis-dependent.* The term "Bayesian" given to these bounds refers to the use of a distribution over hypotheses that is used to define the complexity term. This distribution can be interpreted as a prior belief over the efficacy of each hypothesis before any observations are made.

Non-deterministic hypotheses are modelled by assuming that a distribution μ over $\mathcal{H}$ is used to randomly draw a deterministic hypothesis $h \in \mathcal{H}$ to predict $h(x)$ each time a new instance x is seen. Such a strategy is called a *Gibbs hypothesis* for μ. Since its behaviour is defined by the distribution μ we will abuse our notation slightly and define its loss on the example z to be $\ell_\mu(z) := \mathbb{E}_{h\sim\mu}[\ell_h(z)]$. Similarly, the true risk and empirical risk for a Gibbs hypothesis are, respectively, defined to be $L_P(\mu) := \mathbb{E}_{h\sim\mu}[L_P(h)]$ and $\hat{L}_{\mathbf{z}}(\mu) := \mathbb{E}_{h\sim\mu}\left[\hat{L}_{\mathbf{z}}(h)\right]$. As with the earlier generalization bounds, the aim is to provide guarantees about the difference between $L_P(\mu)$ and $\hat{L}_{\mathbf{z}}(\mu)$. In the case of 0-1 loss, $p := L_P(\mu) \in [0,1]$ is just the probability of the Gibbs hypothesis for μ misclassifying an example and $q := \hat{L}_{\mathbf{z}}(\mu) \in [0,1]$ can be thought of as an estimate of p. However, unlike the earlier bounds on the difference between the true and estimated risk, PAC-Bayesian bounds are expressed in terms the *Kullback-Leibler (KL) divergence*. For the values $p, q \in [0,1]$ this is defined as $kl(q\|p) := q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}$ and for distributions μ and π over the hypothesis space $\mathcal{H}$ we write $KL(\mu\|\pi) := \int_{\mathcal{H}} \ln \frac{d\mu}{d\pi}\, d\mu$. Using these definitions, the most common PAC-Bayesian bound states the following.

> **Theorem (PAC-Bayesian Bound)** For all choices of the distribution π over $\mathcal{H}$ made prior to seeing any examples, the Gibbs hypothesis defined by μ satisfies
>
> $$kl(L_P(\mu), \hat{L}_{\mathbf{z}}(\mu)) \leq \frac{KL(\mu\|\pi) + \ln \frac{n+1}{\delta}}{n} \tag{10}$$
>
> with probability at least $1-\delta$ over draws of $\mathbf{z}$ from P^n

This says that the difference (as measured by *kl*) between the true and empirical risk for the Gibbs hypothesis based on μ is controlled by two terms: a *complexity* term $\frac{KL(\mu\|\pi)}{n}$ and a *sampling* term $\frac{\ln \frac{n+1}{\delta}}{n}$, both of which converge to zero as n increases. To make connections with the previous bounds more apparent, we can weaken (10) using the inequality $kl(q\|p) \geq 2(p-q)^2$ to get the following bound which holds under the same assumptions:

$$L_P(\mu) \leq \hat{L}_{\mathbf{z}}(\mu) + \sqrt{\frac{KL(\mu\|\pi) + \ln \frac{n+1}{\delta}}{2n}}$$

The sampling term is similar to the ubiquitous estimation penalty in the earlier bounds but with an additional $\ln(n+1)/n$. The complexity term is a measure of the complexity of the Gibbs hypothesis for μ *relative* to the distribution π. Intuitively, $KL(\cdot\|\pi)$ can be thought of as a parametrized family of complexity measures where hypotheses from a region where π is large are “cheap” and those where π is small are “expensive”. Information theoretically, it is the expected number of extra bits required to code hypotheses drawn from μ using a code based on π instead of a code based on μ. It is for these reasons the PAC-Bayes bound is said to demonstrate the importance of choosing a good prior. If the Gibbs hypothesis μ which minimises $\hat{L}_{\mathbf{z}}(\mu)$ is also “close” to π then the bound will be tight.

Unlike the other bounds discussed above, PAC-Bayesian bounds are in terms of the complexity of single meta-classifiers rather than the complexity of classes. Furthermore, for specific base hypothesis classes such as margin classifiers used by SVMs it is possible to get hypothesis-specific bounds via the PAC-Bayesian bounds. These are typically much tighter than the VC or Rademacher bounds.

Other Bounds. While the above bounds are landmarks in statistical learning theory there is obviously much more territory that has not been covered here. For starters, the VC bounds for classification can be refined by using more sophisticated results from empirical process theory such as the Bernstein and Variance-based bounds. These are discussed in Section 5 of [6]. There are also other distribution- and sample-dependent complexity measures that are motivated differently to Rademacher averages. For example, the *VC entropy* (see Section 4.5 of [7]) is a distribution-dependent measure obtained by averaging $|\mathcal{F}_{\mathbf{z}}|$ with respect to the sample distribution rather than taking supremum in the definition of the shattering coefficient.

Moving beyond classification, bounds for regression problems have been studied in depth and have similar properties to those for classification. These bounds are obtained by essentially discretizing the function spaces. The growth function is replaced by what is known as a *covering number* but the essence of the bounds remain the same. The reader is referred to [8] for a brief discussion and [1] for more detail.

There are a variety of bounds that, unlike those above, are algorithm-specific. For example, the regularized empirical risk minimization performed by SVMs has been analysed within an *algorithmic stability* framework. As discussed in [6, 8], hypotheses are considered stable if their predictions are not varied too much when a single training example is perturbed. Two other algorithm-dependent frameworks include the *luckiness* and *compression* frameworks, both summarised in [8]. The former gives bounds in terms of an *a priori* measure of luckiness—how well a training sample aligns with biases encoded in an algorithm—while the latter considers algorithms, like SVMs, which base hypotheses on key examples within a training sample.

Recently, there has been work on a type of algorithm-dependent, relative bound called *reductions* (see [4] for an overview). By transforming inputs and outputs for one type of problem (*e.g.*, probability estimation) into a different type of problem (*e.g.*, classification), bounds for the former can be given in terms of bounds for the latter while making very few assumptions. This opens up a variety of avenues for applying existing results to new learning tasks.

See Also

Classification, Regression, Data Dependant Bounds, Empirical Risk Minimization, Hypothesis Space, Loss, PAC Learning, Probability Distribution, Regularization, Statistical Machine Learning, Structural Risk Minimization, VC Dimension.

Recommended Reading

As mentioned above, the uniform convergence bounds by Vapnik and Chervonenkis [15] and the PAC framework of Valiant [14] were the first generalization bounds for statistical learning. Ideas from both were synthesized and extended by Blumer *et al.* [5]. The book by [9] provides a good overview of the early PAC-style bounds while Vapnik's comprehensive book [16], and Antony and Bartlett's book [1] cover classification and regression bounds involving the VC dimension. Rademacher averages were first considered as an alternative to VC dimension in the context of learning theory by Koltchinskii and Panchenko [10] and were refined and extended by Bartlett and Mendelson [3] who provide a readable overview. Early PAC-Bayesian bounds were established by McAllester [12] based on an earlier PAC analysis of Bayesian estimators by Shaw-Taylor and Williamson [13]. Applications of the PAC-Bayesian bound to SVMs are discussed in Langford's tutorial on prediction theory [11] and recent paper by Banerjee [2] provides an information theoretic motivation, a simple proof of the bound in (10), as well as connections with similar bounds in online learning.

There are several well-written surveys of generalization bounds and learning theory in general. Herbrich and Williamson [8] present a unified view of VC, compression, luckiness, PAC-Bayesian and stability bounds. In a very readable introduction to statistical learning theory, Bousquet *et al.* [7] provide good intuition and concise proofs for all but the PAC-Bayesian bounds presented above. That introduction is a good companion for the excellent but more technical survey by Boucheron *et al.* [6] based on tools from the theory of empirical processes. The latter paper also provides a wealth of further references and a concise history of the development of main techniques in statistical learning theory.

References

[1] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 1999.

[2] Arindam Banerjee. On Bayesian bounds. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 81–88, 2006.

[3] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

[4] A. Beygelzimer, J. Langford, and B. Zadrozny. Machine learning techniques—reductions between prediction quality metrics. *Performance Modeling and Engineering. Springer*, pages 3–28, 2008.

[5] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[6] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM Probability and statistics*, 9:323–375, 2005.

[7] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. *Introduction to Statistical Learning Theory*, volume 3176 of *Lecture Notes in Artificial Intelligence*, pages 169–207. Springer, 2004.

[8] Ralf Herbrich and Robert C. Williamson. Learning and generalization: Theory and bounds. In Michael Arbib, editor, *Handbook of Brain Theory and Neural Networks.* MIT Press, 2nd edition, 2002.

[9] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, 1994.

[10] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[11] John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6(1):273–306, 2005.

[12] D.A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[13] J. Shawe-Taylor and R.C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, page 9. ACM, 1997.

[14] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1142, 1984.

[15] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[16] Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

Definitions of Key Terms

Rademacher Complexity. (Synonym: Rademacher Average)

Rademacher complexity is a measure used in generalization bounds to quantify the "richness" of a class of functions. Letting $\rho_1, \ldots, \rho_n$ denote *Rademacher variables*—independent random variables that take the values ± 1 with equal probability—the *empirical* or *conditional Rademacher complexity* of a class of real-valued functions $\mathcal{F}$ on the points $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ is the conditional expectation

$$\hat{R}_{\mathbf{x}}(\mathcal{F}) = \mathbb{E}_{\mathbf{x},\rho}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \rho_i f(x_i) \middle| \mathbf{x}\right].$$

Intuitively, the empirical Rademacher average $R_n(\mathcal{F})$ measures how well functions $f \in \mathcal{F}$ evaluated on $\mathbf{x} \in \mathcal{X}$ can align with randomly chosen labels. The (full)

Rademacher complexity $R_n(\mathcal{F})$ with respect to a distribution P over $\mathcal{X}$ is the average empirical complexity when the arguments $x_1, \ldots, x_n$ are independent random variables drawn from P. That is,

$$R_n(\mathcal{F}) = \mathbb{E}_{\mathbf{x}} \left[\hat{R}_{\mathbf{x}}(\mathcal{F}) \right].$$

There are several properties of the Rademacher average that that make it a useful quantity in analysis: for any two classes $\mathcal{F} \subseteq \mathcal{G}$ we have $R_n(\mathcal{F}) \leq R_n(\mathcal{G})$; when $c \cdot \mathcal{F} := \{cf : f \in \mathcal{F}\}$ for $c \in \mathbb{R}$ we have $R_n(c \cdot \mathcal{F}) = |c| R_n(\mathcal{F})$; when $\mathcal{F} + g := \{f + g : f \in \mathcal{G}\}$ for some fixed function g we have $R_n(\mathcal{F} + g) = R_n(\mathcal{F})$; and if $\mathrm{conv}(\mathcal{F})$ is the convex hull of $\mathcal{F}$ then $R_n(\mathrm{conv}(\mathcal{F})) = R_n(\mathcal{F})$.

McDiarmid's Inequality. (Synonym: Bounded Differences Inequality)

McDiarmid's inequality shows how the values of a bounded function of independent random variables concentrate about its mean. Specifically, suppose $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the bounded differences property. That is, for all $i = 1, \ldots, n$ there is a $c_i \geq 0$ such that for all $x_1, \ldots, x_n, x' \in \mathcal{X}$

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x', x_{i+1}, \ldots, x_n)| \leq c_i.$$

If $\mathbf{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$ is a random variable drawn according to P^n and $\mu = \mathbb{E}_{P^n}[f(\mathbf{X})]$ then for all $\epsilon > 0$

$$P^n \left(f(\mathbf{X}) - \mu \geq \epsilon \right) \leq \exp\left(\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2} \right).$$

McDiarmid's is a generalization of Hoeffding's inequality which can be obtained by assuming $\mathcal{X} = [a, b]$ and choosing $f(\mathbf{X}) = \sum_{i=1}^n X_i$. When applied to empirical risks this inequality forms the basis of many generalization bounds.

Shattering Coefficient. (Synonym: Growth Function)

The shattering coefficient $S_{\mathcal{F}}(n)$ is a function that measures the size of a function class $\mathcal{F}$ when its functions $f : \mathcal{X} \to \mathbb{R}$ are restricted to sets of points $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ of size n. Specifically, for each $n \in \mathbb{N}$ the shattering coefficient is the maximum size of the set of vectors $\mathcal{F}_{\mathbf{x}} = \{(f(x_1), \ldots., f(x_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n$ that can be realized for some choice of $\mathbf{x} \in \mathcal{X}^n$. That is,

$$S_{\mathcal{F}}(n) = \sup_{\mathbf{x} \in \mathcal{X}^n} |\mathcal{F}_{\mathbf{x}}|.$$

The shattering coefficient of a hypothesis class $\mathcal{H}$ is used in generalization bounds as an analogue to the class's size in the finite case.

Symmetrization Lemma. (Synonym: Basic Lemma)

Given a distribution P over a sample space $\mathcal{Z}$, a finite sample $\mathbf{z} = (z_1, \ldots, z_n)$ drawn i.i.d. from P and a function $f : \mathcal{Z} \to \mathbb{R}$ we define the shorthand $\mathbb{E}_P f = \mathbb{E}_P[f(z)]$ and $\mathbb{E}_{\mathbf{z}} f = \frac{1}{n} \sum_{i=1}^n f(z_i)$ to denote the true and empirical average of f. The symmetrization lemma is an important result in the learning theory as it allows the true average $\mathbb{E}_P f$ found in generalization bounds to be replaced by a second empirical average $\mathbb{E}_{\mathbf{z}'} f$ taken over an independent *ghost sample* $\mathbf{z}' = (z'_1, \ldots z'_n)$ drawn i.i.d. from P. Specifically, the symmetrization lemma states that for any $\epsilon > 0$ whenever $n\epsilon^2 \geq 2$

$$P^n \left(\sup_{f \in \mathcal{F}} |\mathbb{E}_P f - \mathbb{E}_{\mathbf{z}} f| > \epsilon \right) \leq 2P^{2n} \left(\sup_{f \in \mathcal{F}} |\mathbb{E}_{\mathbf{z}'} f - \mathbb{E}_{\mathbf{z}} f| > \frac{\epsilon}{2} \right).$$

This means the typically difficult to analyse behaviour of $\mathbb{E}_P f$ – which involves the entire sample space $\mathcal{Z}$ – can be replaced by the evaluation of functions from $\mathcal{F}$ over the points in $\mathbf{z}$ and $\mathbf{z}'$.