Generalization Bounds for Weighted Binary Classification with Applications to Statistical Verification

Vu Ha **Tariq Samad**
Honeywell Labs
3660 Technology Dr.
Minneapolis, MN 55418
{*vu.ha,tariq.samad*}*@honeywell.com*

Abstract

We describe an approach to statistically verifying complex controllers. This approach is based on deriving practical Vapnik-Chervonenkis-style (VC) generalization bounds for binary classifiers with *weighted* loss. An important case is deriving bounds on the probability of false positive. We show how existing methods to derive bounds on classification error can be extended to derive similar bounds on the probability of false positive, as well as bounds in a decision-theoretic setting that allows tradeoffs between false negatives and false positives. We describe experiments using these bounds in statistically verifying computational properties of an iterative controller for an Organic Air Vehicle (OAV).

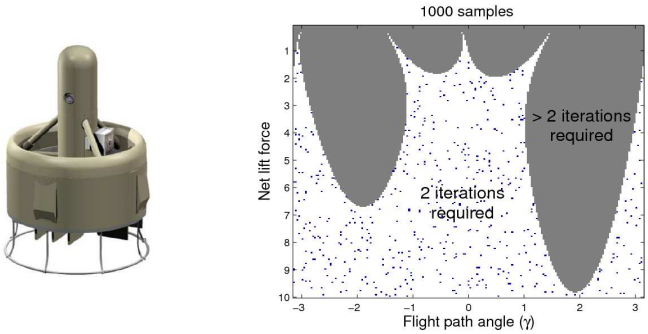

Figure 1: The Organic Air Vehicle (left), and the complexity of an iterative equilibrium angle-of-attack computation as determined by two factors: The flight path angle γ and the net lift force $\frac{\bar{q}S}{mg}$ (right). The computational complexity here is equated with the number of iterations. The safe operating envelope consists of those inputs that require at most two iterations. The decision boundary is empirically drawn using 1000 random samples.

1 Introduction

The computational requirements for high-performance complex control algorithms can vary considerably. The variation arises because the computation depends on a number of factors such as the sensed/estimated state of the system under control, environmental disturbances, and the operational mode of the system. Furthermore, the variation in general can not be determined analytically. In hard real-time systems where the computation must complete in time for a command to be issued to the actuator at the next sample instant, these uncertainties pose a significant challenge. This is the reason that PID (Proportional-Integral-Derivative) controllers, with their deterministic execution time, are still the preferred choice in many applications despite their lesser performance.

In order to bring practical acceptance to high-performance complex control algorithms, we propose a compromise that makes use of both types of controller algorithms. High-performance algorithms will be used within a *safe operational envelope* (SOE) where they are guaranteed to complete within the allocated time. Outside of the SOE, lower performance and computationally simpler algorithms will be used. The SOE is determined based on simulation data, and hence is only *safe with some statistical guarantees*.

The problem of identifying the SOE is a binary classification problem where a false negative merely means a conservative use of a low performance controller and a false positive may have drastic consequences such as loss of the vehicle. Thus we would like to obtain an SOE that has some statistical guarantee that it is indeed safe, i.e., a classifier with provably small *probability of false positive*. This however is not the only criterion, for otherwise the trivial classifier that classifies everything as negative, namely the current state of the art, would be the top candidate. Instead, the goal is to push the boundaries of the SOE as far as possible while keeping a cap on the probability of false positive.

As an example, let us consider the computational property of a high-performance controller for an OAV (Figure 1). The OAV has a ducted fan propulsion unit, with control provided by movable vanes in the propwash. The vanes are situated in the propulsion airflow and consequently the interactions between the propulsion and the control surfaces are highly non-linear. The trim calculation for the OAV is an iterative algorithm whose computational time depends on several factors [Elgersma and Morton, 2000]. We are interested in conditions under which this calculation can be reliably used.

Our approach is based on statistical learning theory (SLT). Specifically, we derive statistical guarantees for SOEs using Vapnik-Chervonenkis-style generalization bounds for classification problems with *weighted* loss, of which the classification problem with false positive loss is a special case. We are

interested *practical* bounds–bounds that are asymptotically competitive *and* have small pre-constants. While the SLT literature contains a vast collection of VC-style bounds, the majority of these bounds are stated and proved only for the *probability of misclassification* and only a few are directly applicable to our problem. Furthermore, SLT bounds are often derived with little emphasis on obtaining optimal pre-constants. In addition to the emphasis on finding small pre-constants, our analysis has two unique aspects. First, we assume that it is possible to achieve small empirical loss (which is true for the case of false positive loss). Second, our analysis is upper-tail-oriented, as we are only interested in deriving *upper* bounds for the expected loss.

2 Preliminaries

Notations: We use the symbols $\mathbb{P}$, $\mathbb{E}$, and $\mathbb{V}$ to denote the probability, expectation, and variance, respectively. $\sigma = \{\sigma_i\}_{i=1}^n$ denotes a Rademacher sequence–a sequence of independent, symmetric $-1/1$-valued random variables.

Let X and Y be non-empty sets and $Z = X \times Y$. A pair of $(x, y) \in X \times Y$ is denoted as z. Let (Z, μ) be a fixed probability measure. A *training set* is a finite sample $\mathcal{S}_n = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ drawn independently according to μ. The probability and expectation with respect to $\mathcal{S}_n$ are written as $\mathbb{P}_n$ and $\mathbb{E}_n$. A *hypothesis space* is a set H of functions from X to Y. A *loss function* is a function $l : Y \times Y \to \mathbb{R}$. The loss of a hypothesis $h \in H$ on z is $l(h, z) = l(h(x), y)$. The *expected loss* of h is $l(h) = \mathbb{E}l(h, z)$. The *empirical loss* of h on the training set $\mathcal{S}_n$ is $l_n(h) = l(h, \mathcal{S}_n) = \frac{1}{n}\sum_{i=1}^n l(h, z_i)$. We assume that all loss functions have range $[0, 1]$. When Y is finite, we have a *classification* problem. When $|Y| = 2$, the classification is *binary*, and let $Y = \{-1, 1\}$. Here we only deal with binary classification.

If we assume that correct classifications incur zero loss, i.e. $l(-1, -1) = l(1, 1) = 0$, then what emerges is a loss function that we refer to as *weighted classification error*, defined as:

$$l^{(\rho)}(y, y') = \begin{cases} 0 & \text{if } y = y' \\ \rho & \text{if } y = -1, y' = 1 \text{ (false negative)} \\ 1 & \text{if } y' = -1, y = 1 \text{ (false positive).} \end{cases}$$

Here ρ is a number between 0 and 1. The idea is that false positives (bad errors) are more costly than false negatives (good errors). When $\rho = 1$, i.e. when no difference is made between the two types of errors, the loss is called the *misclassification error*. When $\rho = 0$, $l^{(0)}(h)$ is the probability of the classifier h making a false positive error.

Given a training set $\mathcal{S}_n$, a hypothesis $h \in H$, and a loss function l, how can we bound the expected loss $l(h)$? Statistical learning theory (SLT) provides a probabilistic answer to this question. A typical SLT result of the form $\mathbb{P}_n(l(h) > \epsilon) < \delta$ or, succinctly, $l(h) <_\delta \epsilon$ provides an upper bound ϵ on $l(h)$ with confidence at least $1 - \delta$, where ϵ and δ are positive, reasonably small numbers. The upper bound ϵ typically depends on δ, the sample size n, the empirical loss $l_n(h)$, and a complexity measure of the hypothesis space H. The most important complexity measure in SLT is *VC dimension*. Let $\mathcal{C} \subseteq 2^X$ be a set of subsets of X. Note that H, as a set of binary classifiers on X, is an example of such sets. For any $A \subseteq X$, define $\mathcal{C} \cap A = \{C \cap A : C \in \mathcal{C}\}$. The *growth function* of $\mathcal{C}$ is defined as

$$\Delta_k(\mathcal{C}) = \max\{|\mathcal{C} \cap A| : A \subseteq X, |A| = k\}.$$

Clearly, $\Delta_k(\mathcal{C}) \leq 2^k$. The *VC Dimension* of $\mathcal{C}$ is defined as

$$d(\mathcal{C}) = \sup\{k \in \mathbb{Z} : \Delta_k(\mathcal{C}) = 2^k\}.$$

Let $d = d(H)$. We assume that $d < \infty$.

Theorem 2.1. *[Vapnik, 2000, Equations 3.26]*

$$l(h) <_\delta l_n(h) + \sqrt{\frac{1}{n}\left(d \ln \frac{2en}{d} + \ln \frac{4}{\delta}\right)}. \tag{2.1}$$

The bound (2.1) is obtained by applying a very general result of Vapnik to the special setting of learning binary classifiers with a specific loss function $l^{(\rho)}$. It is thus natural to ask if this bound can be improved. While many SLT results have been obtained that address this issue, the majority of are formulated and proved for $l = l^{(1)}$ only. One of the goals of this paper is to examine if these results can be extended to the loss functions $l^{(\rho)}, \rho \in [0, 1)$. It turns out that these results fall into two categories: those that are applicable for all $\rho \in [0, 1]$, and those that are applicable for $\rho = 0, 1$ only.

We conclude the preliminaries with the statement of the often-used Sauer's lemma.

Sauer's Lemma. *[Sauer, 1972]* $\Delta_n(H) \leq (en/d)^d$.

3 VC Dimension-Based Bounds

Suppose that h is a hypothesis with "small" empirical loss: $l_n(h) \leq \epsilon_1$. We would like to bound the probability that $l(h)$ is "large": $l(h) > \epsilon$ for some $\epsilon > \epsilon_1$. This amounts to bounding $\mathbb{P}_n(Q)$ where Q is defined as

$$Q = \{\mathcal{S}_n : \exists h : l_n(h) \leq \epsilon_1, l(h) > \epsilon\}.$$

There are two major approaches to do this: The classical approach of Vapnik and Chervonenkis [1971], and the approach based on abstract concentration inequalities developed by Talagrand and others.

3.1 The Classical Approach

The classical VC analysis begins with the observation that $\mathbb{P}_n(Q) \leq \mathbb{P}(R)/\mathbb{P}(R|Q)$ for any R that satisfies $\mathbb{P}(R|Q) > 0$. We then define R based on $\mathcal{S}_n$ and an additional independent sample whose size may or may not be equal to n. In this analysis we take the former approach: Let $\mathcal{S}'_n = \{z_i\}_{i=n+1}^{2n}$ be an independent sample of size n, commonly referred to as the *ghost sample* and let $l'_n(h) = l(\mathcal{S}'_n, h)$, the empirical loss on the ghost sample. Denote $\mathcal{S}_{2n} = (\mathcal{S}_n, \mathcal{S}'_n)$. Let $0 \leq \epsilon_1 < \epsilon_2 \leq \epsilon$. Define R as

$$R = \{\mathcal{S}_{2n} : \exists h : l_n(h) \leq \epsilon_1, l'_n(h) > \epsilon_2\}. \tag{3.1}$$

Upper bounding $\mathbb{P}_n(Q)$ now reduces to upper bounding $\mathbb{P}_{2n}(R)$ (the covering step) and lower bounding $\mathbb{P}_{2n}(R|Q)$ (the symmetrization step).

The Covering Step: Upper Bounding $\mathbb{P}_{2n}(R)$. Intuitively, $\mathbb{P}_{2n}(R)$ is small because if h has small empirical loss on a sample, its empirical loss on a ghost sample should also be small. We can change the definition of R in (3.1) as

$$R = \left\{\mathcal{S}_{2n} : \exists h : l'_n(h) - l_n(h) > \eta, \text{ where } \eta = \epsilon_2 - \epsilon_1\right\}.$$

The next step uses the so-called *permutation technique*.

$$\mathbb{P}_{2n}(R) = \mathbb{E}_{2n}\left(\exists h : l'_n(h) - l_n(h) > \eta\right)$$
$$= \mathbb{E}_{2n}\mathbb{E}_\sigma\left(\exists h : \sum_{i=1}^{n} \sigma_i(l(h, z_{i+n}) - l(h, z_i)) > n\eta\right).$$

Next, we fix $\mathcal{S}_{2n}$ and bound the inner expectation. Let

$$H_{2n} = \{h_{2n} = (h(x_1), \ldots, h(x_{2n}))\,|h \in H\} \subseteq \{-1, 1\}^{2n}.$$

The mapping $h \mapsto h_{2n}$ is many-to-one. Denote $l_i(h_{2n}) = l(h, z_i), 1 \le i \le 2n$. The inner expectation can be written as

$$\mathbb{E}_\sigma\left(\exists h_{2n} \in H_{2n} : \sum_{i=1}^{n} \sigma_i\left(l_{i+n}(h_{2n}) - l_i(h_{2n})\right) > n\eta\right)$$

which, by the union bound, is bounded by

$$\sum_{h_{2n} \in H_{2n}} \mathbb{E}_\sigma\left(\sum_{i=1}^{n} \sigma_i\left(l_{i+n}(h_{2n}) - l_i(h_{2n})\right) > n\eta\right). \quad (3.2)$$

The cardinality of H_{2n} is at most $\Delta_{2n}(H)$, which is at most $(2en/d)^d$ by Sauer's lemma. For a fix $h_{2n} \in H_{2n}$, the summand in (3.2) can be written as

$$\mathbb{P}_\sigma\left(\sum_{i=1}^{n} \sigma_i\left(l_{i+n}(h_{2n}) - l_i(h_{2n})\right) > n\eta\right)$$

which, by Höeffding's right-tail inequality, is bounded by

$$\exp\left(\frac{-2n^2\eta^2}{4\sum_i (l_{i+n}(h_{2n}) - l_i(h_{2n}))^2}\right) \le \exp\left(-\frac{n\eta^2}{2}\right).$$

Thus we have arrived at the following result.

Lemma 3.1. *[Vapnik and Chervonenkis, 1971]*

$$\mathbb{P}_{2n}\left(l_n(h) \le \epsilon_1, l'_n(h) > \epsilon_2\right) \le \left(\frac{2en}{d}\right)^d \exp\left(-.5n(\epsilon_2 - \epsilon_1)^2\right).$$

When $\epsilon_1 \ll \epsilon_2$, it is possible to improve upon Lemma 3.1. The idea is, instead of bounding the probability that the *absolute discrepancy* $l'_n(h) - l_n(h)$ is large, we bound the probability that the *relative discrepancy* $\frac{l'_n(h) - l_n(h)}{\sqrt{l'_n(h) + l_n(h)}}$ is large. We "weaken" the definition of R in (3.1) as

$$R = \left\{\mathcal{S}_{2n} : \exists h : l'_n(h) - l_n(h) > \eta'\right\},$$
$$\eta' = (\epsilon_2 - \epsilon_1)\sqrt{\frac{l'_n(h) + l_n(h)}{\epsilon_2 + \epsilon_1}} = \eta\sqrt{\frac{l'_n(h) + l_n(h)}{\epsilon_2 + \epsilon_1}}.$$

Now, proceed identically as before, except that η is now replaced with η', we arrive at the following results.

Lemma 3.2. *[Vapnik and Chervonenkis, 1971]*

$$\mathbb{P}_{2n}\left(l_n(h) \le \epsilon_1, l'_n(h) > \epsilon_2\right) \le \left(\frac{2en}{d}\right)^d \exp\left(\frac{-n(\epsilon_2 - \epsilon_1)^2}{2(\epsilon_2 + \epsilon_1)}\right).$$

Corollary 3.3. *[Vapnik and Chervonenkis, 1971]*

$$\mathbb{P}_{2n}\left(l_n(h) = 0, l'_n(h) > \epsilon_2\right) \le \left(\frac{2en}{d}\right)^d \exp\left(-.5n\epsilon_2\right).$$

By considering the relative discrepancy, we have managed to insert the term $\epsilon_2 + \epsilon_1$, resulting in a tighter bound in Lemma (3.2). Further tightening is possible when $\epsilon_1 = 0$. Instead of using Hoeffding's inequality, Blumer *et al.* [1989] use a combinatorial argument that leads to the following improvement of Corollary 3.3.

Lemma 3.4. *[Blumer* et al.*, 1989]*

$$\mathbb{P}_{2n}\left(l_n(h) = 0, l'_n(h) > \epsilon_2\right) \le \left(\frac{2en}{d}\right)^d \exp\left(-n\epsilon_2 \ln 2\right).$$

The Symmetrization Step: Lower Bounding $\mathbb{P}_{2n}(R|Q)$. To lower bound $\mathbb{P}_{2n}(R|Q)$, we can fix $\mathcal{S}_n$, ignore the condition $l_n(h) \le \epsilon_1$, and bound the following conditional probability:

$$\mathbb{P}_{\mathcal{S}'_n}\left(\{\mathcal{S}'_n : \exists h : l'_n(h) > \epsilon_2\}|\exists h : l(h) > \epsilon\right),$$

or, equivalently, $\mathbb{P}_n\left(\{\exists h : l_n(h) > \epsilon_2\}|\exists h : l(h) > \epsilon\right)$.

Let h be a hypothesis such that $l(h) > \epsilon$. It suffices to lower bound $\mathbb{P}_n(l_n(h) > \epsilon_2)$. Intuitively, this quantity is large because the empirical loss should be large ($> \epsilon_2$) wherever the expected loss is large ($> \epsilon$). Since $\epsilon_2 < \epsilon$, we have

$$\mathbb{P}_n(l_n(h) \le \epsilon_2) \le \mathbb{P}(l(h) - l_n(h) > \epsilon - \epsilon_2)$$
$$\le \exp\left(-\frac{6n(\epsilon - \epsilon_2)^2}{4(\epsilon - \epsilon_2) + 3}\right) := \iota(n, \epsilon, \epsilon_2), \quad (3.3)$$

by Bernstein's left-tail inequality. Coupled this with Lemma 3.2, we obtain the following result.

Corollary 3.5.

$$\mathbb{P}_n\left(l_n(h) \le \epsilon_1, l(h) > \epsilon\right)$$
$$\le \left(1 \vee (1 - \iota(n, \epsilon, \epsilon_2))^{-1}\right)\left(\frac{2en}{d}\right)^d \exp\left(\frac{-n(\epsilon_2 - \epsilon_1)^2}{2(\epsilon_2 + \epsilon_1)}\right).$$

In this inequality, the parameter ϵ_2 is unspecified, and we can minimize the bound over $\epsilon_2 \in (\epsilon_1, \epsilon)$.

When $\rho = 0, 1$, the loss function $l^{(\rho)}$ is binary, and $nl_n(h)$ is a binomial random variable with parameters n and $l(h)$. We can thus use several lower bounds on the right-tails of the binomial to obtain

$$n\epsilon > 1 \Rightarrow \mathbb{P}_n\left(\{l_n(h) > \epsilon/4\}|\exists h : l(h) > \epsilon\right) > 1/4 \quad (3.4)$$
$$n\epsilon > 2 \Rightarrow \mathbb{P}_n\left(\{l_n(h) > \epsilon/2\}|\exists h : l(h) > \epsilon\right) > 1/2 \quad (3.5)$$

For the case $\epsilon_1 > 0$, we can set $\epsilon_2 = \epsilon$ and combine (3.4) with Lemma 3.2 to obtain the following result.

Corollary 3.6. *[Vapnik and Chervonenkis, 1971] For* $l = l^{(\rho)}, \rho = 0, 1$,

$$l(h) <_\delta 2l_n(h) + \frac{4}{n}\left(d\ln\frac{2en}{d} + \ln\frac{4}{\delta}\right). \quad (3.6)$$

For the case $\epsilon_1 = 0$, we can set $\epsilon_2 = \epsilon/2$ and combine (3.5) with Lemma 3.4 to obtain the following result.

Corollary 3.7. *[Blumer* et al.*, 1989] For* $l = l^{(\rho)}, \rho = 0, 1$,

$$l_n(h) = 0 \Rightarrow l(h) <_\delta \frac{2}{n\ln 2}\left(d\ln\frac{2en}{d} + \ln\frac{2}{\delta}\right). \quad (3.7)$$

Shawe-Taylor *et al.* [1993] further improve (3.7), using an argument that uses a second sample $\mathcal{S}'_k$ of size k, and $\epsilon_2 = r\epsilon$, where $r = 1 - \sqrt{2/(\epsilon k)}, k = n\left(ern/d - 1\right)$.

Theorem 3.8. *[Shawe-Taylor* et al.*, 1993] For* $l = l^{(\rho)}, \rho = 0, 1$,

$$n\epsilon > 4d \Rightarrow \mathbb{P}_n\left(l_n(h) = 0, l(h) > \epsilon\right)$$
$$\le 2\exp\left(2\sqrt{2d} - \epsilon n + d\ln\epsilon + 2d\ln\frac{en}{d}\right). \quad (3.8)$$

Compared with (3.7), the sample complexity derived from (3.8) is smaller by a factor of $\frac{4}{\ln 2}(1-\sqrt{\epsilon}) \approx 5.7$ for typical values ϵ (say, < 0.05).

We point out that Corollary 3.6, 3.7, and Theorem 3.8 were previously stated and proved for the loss function $l = l^{(1)}$ only. Our analysis extends them to the case $l = l^{(0)}$ (and shows that they do *not* hold when $0 < \rho < 1$). The covering argument remains the same, while the symmetrization argument uses the simple observation that $nl_n^{(\rho)}(h)$ is a binomial random variable with parameters n and $l(h)$, regardless of whether $\rho = 0$ or $\rho = 1$.

3.2 Talagrand's Method

Observe that $l(h) \leq l_n(h) + \sup_{h\in H}(l(h) - l_n(h))$. The supremum is a random function of $\mathcal{S}_n$, where changing a single element z_i results in a change of at most $1/n$, and thus is $<_\delta$-bounded by $\mathbb{E}_n\left(\sup_{h\in H}(l(h) - l_n(h))\right) + \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}$ by McDiarmid's inequality [McDiarmid, 1997]. The next quantity to bound is the expectation of the supremum. This is accomplished using the concept of Rademacher average. Let G be a class of functions from Z to the reals $\mathbb{R}$. The *Rademacher average* of G is defined as $R_nG = R(G, \mathcal{S}_n, \sigma) = \sup_{g\in G}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i)\right)$. In this analysis, the role of G is played by the loss class associated with H: $G = l_H = \{z \mapsto l(h, z) : h \in H\}$. The symmetrization inequality (e.g. [Bartlett *et al.*, 2005]) states that $\mathbb{E}_n\left(\sup_{h\in H}(l(h) - l_n(h))\right) \leq 2\mathbb{E}R_n l_H$. Thus it remains to bound the Rademacher average. The technique is well-established and based on the concepts of *covering number*. Denote by $N(u, l_H, L_2(\mu^n))$ the u-covering number of l_H with respect to the metric $L_2(\mu^n)$.

Lemma 3.9. *[Dudley, 1999]*

$$\mathbb{E}_\sigma R_n l_H \leq \frac{12}{\sqrt{n}}\int_0^1 \sqrt{\ln N(u, l_H, L_2(\mu^n))}du. \tag{3.9}$$

The last piece of the puzzle reveals a bound on $N(u, l_H, L_2(\mu^n)))$, defined based on the VC dimension d.

Lemma 3.10. *[Haussler, 1995] For all $\mathcal{S}_n$:*

$$N(u, l_H, L_2(\mu^n)) \leq e(d+1)\left(\frac{2e}{u^2}\right)^d. \tag{3.10}$$

Combining (3.9) and (3.10), with some algebra we obtain the following result.

Theorem 3.11. *(Dudley-Haussler)*

$$l(h) <_\delta l_n(h) + 30\sqrt{\frac{d}{n}} + \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{3.11}$$

The bound (3.11) is based on an analysis of the absolute discrepancy $l(h) - l_n(h)$. Its deviation term is $O(\sqrt{d/n})$, which should not come as a surprise. It is natural to ask if this approach can be used to analyze some form of relative discrepancy between the expected loss $l(h)$ and the empirical loss $l_n(h)$. Consider

$$\tilde{L}_n = \sup\left\{\frac{l(h) - l_n(h)}{\sqrt{l(h)}} : h \in H, l(h) > 0\right\}. \tag{3.12}$$

Simple algebra shows that for all $h \in H$, $l(h) \leq (1 + \tilde{L}_n)l_n(h) + \tilde{L}_n^2$. Consequently, an upper bound on $\tilde{L}_n$ can be translated into an upper bound on $l(h)$. Unlike the previous analysis, it appears that we can not use McDiarmid's bounded difference inequality, as the introduction of the term $\sqrt{l(h)}$ renders the "difference" unbounded. The solution to this problem originates from the work of Talagrand on abstract concentration inequalities and their applications to bounding the suprema of empirical processes [Talagrand, 1994, 1996]. Talagrand's inequalities were later improved using the so-called entropy method. The following version provides the best known bound.

Lemma 3.12. *[Bousquet, 2002] Let $\mathcal{F}$ be a countable set of functions from Z to $\mathbb{R}$. Let $b = \sup_{f\in\mathcal{F}}(\sup(\mathbb{E}(f) - f))$, $v = \sup_{f\in\mathcal{F}}\mathbb{V}(f)$ and $B_n = B(\mathcal{S}_n) = \sup_{f\in\mathcal{F}}(\mathbb{E}(f) - \frac{1}{n}\sum_{i=1}^n f(z_i))$. Then for any $\alpha > 0$,*

$$B_n <_\delta (1+\alpha)\mathbb{E}_n(B_n) + \sqrt{\frac{2v\ln(1/\delta)}{n}} + \left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{b\ln(1/\delta)}{n}.$$

We now apply Lemma 3.12 with $f(z) = l(h, z))/\sqrt{l(h)}$, $b = v = 1$, to obtain

$$\tilde{L}_n <_\delta (1+\alpha)\mathbb{E}_n(\tilde{L}_n) + \sqrt{\frac{2\ln(1/\delta)}{n}} + \left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{\ln(1/\delta)}{n}.$$

We then proceed to bound $\mathbb{E}_n(\tilde{L}_n)$ with a technique from Massart [2000] that is referred to as *peeling* and the concept of sub-root functions. A function $\psi : (0,\infty) \to (0,\infty)$ is called *sub-root* if ψ is non-decreasing and $\psi(r)/\sqrt{r}$ is non-increasing. Any sub-root function $\psi(r)$ is known to have a unique fix-point r^* (i.e. $\psi(r^*) = r^*$) [Bartlett *et al.*, 2005]. Now, suppose that ψ is a sub-root function with fixed point r^* such that

$$\mathbb{E}_n\left(\sup\{|l(h) - l_n(h)| : l(h) \leq r\}\right) \leq \psi(r), \forall r > 0.$$

Then we can show that $\forall\alpha > 0$, $\tilde{L}_n$ is $<_\delta$-bounded by

$$(1+\alpha)\left(\sqrt{r^*}\left(1 + \frac{e}{2}(1 + \ln(\frac{1}{r^*}))\right)\right) + \sqrt{\frac{2\ln\frac{1}{\delta}}{n}} + \frac{(3+\alpha)\ln\frac{1}{\delta}}{3\alpha n}.$$

The final step is to bound r^* using d. This can be done [Koltchinskii and Panchenko, 2000] by setting $\psi(r)$ to be Dudley's entropy integral. What we end up with is a $<_\delta$ bound on $l(h)$ that is $O(d\ln n/n)$, asymptotically comparable to those obtained using the classical approach such as Theorem 3.8, albeit with worse constants.

3.3 Summary of VC Dimension-Based Bounds

Given a sample $\mathcal{S}_n$ of size n and a hypothesis h that has empirical loss $l_n(h)$ on $\mathcal{S}_n$, what can we say about the expected loss $l(h)$ of h? The results in this section provide several answers to this question. They all have the general form of *"If $l_n(h)$ is small, then with high probability, $l(h)$ is small"*. The common assumption is that h is selected from a hypothesis space H with finite VC dimension d. In the general case when l is only assumed to have range $[0, 1]$, Corollary 3.5 seems most useful. When $l = l^{(\rho)}, \rho = 0, 1$, we can exploit several binomial tail inequalities to obtain much

simpler bounds. When $l_n(h) > 0$, Corollary 3.6 should be used. When $l_n(h) = 0$, Theorem 3.8 provides the best-known bound. These results are all based on the idea of uniform convergence of *relative* discrepancies (UCRD). Even Corollary 3.7 and Theorem 3.8 can be viewed as based on degenerate cases of UCRD, with special-purpose combinatorial arguments replacing the general-purpose Hoeffding's bound. Results that are based on uniform convergence of absolute discrepancies such as Lemmas 3.1 and Theorem 3.11 are not as useful for our purpose, as they have to cover situations that Vapnik and Chervonenkis [1971] refer to as *pessimistic cases*. Simply put, pessimistic bounds are loose since they need to account for hypotheses with expected losses close to 0.5. In contrast, we only need to concern ourselves with hypotheses with zero or small empirical losses.

The statistical/computational learning theory literature contains a vast collection of generalization bounds in the general case of function learning and in the special case of learning binary classifiers. To our knowledge no work has explicitly derived generalization bounds for binary classifiers with weighted error penalties as defined in this paper. The bounds in Section 3.1 originate from the seminal work of Vapnik and Chervonenkis [1971]. Our contribution here is the extensions to the case $\rho = 0$, and Corollary 3.5.

Talagrand's approach provides a completely different way to arrive at generalization bounds for all loss function $l^{(\rho)}, \rho \in [0, 1]$ that are asymptotically equivalent with classical bounds. This approach analyzes the mean (or median [Panchenko, 2002]) of the supremum of the (sometimes weighted) discrepancies between the expected and empirical losses using Talagrand's various concentration inequalities, completed invariably with the symmetrization inequality, Dudley's entropy integral bound, and Haussler's packing bound. The resulting bounds often have much larger constants and are not as useful for our non-asymptotic purpose.

4 Experiments

We now describe the applications of the bounds derived in Section 3 to our OAV experiment as described in the introduction. We identify four factors that affect the computational time of the iterative algorithm, and choose 4-dimensional axis-parallel hyper-rectangles as our hypothesis space. The VC dimension of this hypothesis space is 8, as it is known that the VC dimension of axis-parallel hyper-rectangles in $\mathbb{R}^m$ is $2m$. Thus for $\delta = .05, \epsilon = .05$, we need 34,000 samples using Theorem 2.1. But using Theorem 3.8, we need only 1810 samples. With 34,000 samples, if we set $\delta = .05$, ϵ can be as small as 0.0035. The improvement in generalization bound (ϵ) is about 14-fold and in samples complexity (n) is about 18-fold.

The search for the best hyper-rectangle in this experiment is rather simple. For each sampled input, we determine if the iterative algorithm converges. It turns out that in 32,114 instances (roughly 94%), the algorithm converges. Despite this high (empirical) rate of success, in current practice it still loses out to a PID-like controller with fixed deterministic computation time. Next, we randomly choose an axis-parallel 4-dim hyper-rectangle in the input ranges as a hypothesis. If the hyper-rectangle contains a sampled point for which the algorithm does not converge (an unsafe point), then we eliminate that hyper-rectangle (since the guarantees are based on Theorem 3.8 and require zero false positives). Otherwise, we count the number of safe points that lie outside the hyper-rectangle (false negatives), and choose the hyper-rectangle that has the fewest number of false negatives. After looking at 10,000 random hyper-rectangles, we are able to come up with one that contains 18,616 safe points and no unsafe points. This hypothesis thus has 32,114 - 18,616 = 13,498 false negatives. Note that the number of false negatives is still quite large. This is because we use hyper-rectangles which constitute a simple hypothesis space that does not approximate the decision surface very well (this SOE is nevertheless a big improvement over the trivial, empty SOE that is the current state of the art). The advantage to this is that the VC dimension is low, and thus only a small number of samples are required to obtain the statistical guarantee, which reads *"The found hyper-rectangle has probability of false positive bounded by* 0.0035*, and we have at least* 95% *confidence in this statement."*

There are a number of alternatives to the above procedure. For example, we can use Corollary 3.6 instead of Theorem 3.8, if the requirement of zero empirical loss is too restrictive. In the OAV example with 34,000 samples, this leads to a hypothesis with 111 false positives but only 3272 false negatives (a reduction of 75%!) while still maintaining 99% confidence that the probability of false positive is less than 0.01. Furthermore, we can replace the criterion "as few false negatives as possible" with other criteria, for example one that prefers hypotheses with large volumes. Finally, if we are willing to make a decision-theoretic tradeoff between false negatives and false positives (e.g. one false positive is as costly as one thousand false negatives), we can set $\rho = .001$ and apply Corollary 3.5.

5 Summary and Related Work

It has been said that the divide between SLT and practice is of Grand Canyon proportions, perhaps because VC bounds are often too loose to be useful in practice. This paper offers a counterargument in the form of an SLT-based approach to verifying complex controllers. We demonstrated this approach on a problem of significant industrial and military interest: Deriving a safe operating envelope for a complex control algorithm. This approach offers control engineers a principled way to increasingly replace low-performance, simple control algorithms with high-performance, complex ones while still maintaining a statistically high confidence in safety. A key to making this offer attractive lies in deriving practical VC-style generalization bounds for weighted binary classification (a problem that hitherto has not been given much attention). Our VC analysis, which builds upon standard VC analysis of unweighted binary classification, shows that such bounds are indeed possible. They are significantly better than a general bound by Vapnik. Our analysis precisely pointed to the place where the false negative penalty had an effect, namely the symmetrization argument. We have successfully applied this verification framework to several other

control applications, to be reported in an extended version of this paper. We expect these results to have applications outside of the controller verification problem.

From a practical point of view, SLT-based generalization bounds have been used mostly in the model selection problem (see e.g. [Bartlett *et al.*, 2002]). Aside from this, they have also been used in several control systems applications, for example, in deriving randomized algorithms for robust control problems whose exact solution is NP-hard, and in the context of system identification [Vidyasagar, 2003, Chapter 11]. Also, machine learning researchers have long recognized the importance of learning classifiers with general loss function. The work in this area is generally referred to as *cost-sensitive learning* [Turney, 2000].

The bounds derived in Section 3 consist of the empirical loss and a VC confidence term that is independent of the probability measure μ and the particular sample $\mathcal{S}_n$. They are thus necessarily "loose" since they need to hold for "bad" distributions and "bad" samples $\mathcal{S}_n$. Recent research has focused on data-based measures of hypothesis space complexity such as Rademacher averages [Koltchinskii, 2001; Bartlett *et al.*, 2005]. This direction relies on Talagrand's approach as described in Section 3.2. Although this approach yields VC dimension-based bounds that have worse constants compared to bounds using the classical approach such as (3.6), it can be used to derive bounds based entirely on data (i.e. $\mathcal{S}_n$) without *a priori* information about the hypothesis space H (such as its VC dimension d). The following result is an example of such data-dependent bounds.

Theorem 5.1. *[Bartlett* et al.*, 2005, Corollary 6.2] Let $l = l^{(\rho)}, \rho = 0, 1$. Let the random function $\hat{\psi}_n$ be defined on $(0, \sqrt{1/2})$ as*

$$\hat{\psi}_n(r) = 20 \sup_{\beta \in [\sqrt{2r}, 1]} \beta \mathbb{E}_\sigma R_n\{z \mapsto l(h, z) : l_n(h) \leq \frac{2r}{\beta^2}\} + \frac{26}{n} \ln \frac{3}{\delta}.$$

Then $\hat{\psi}_n$ is sub-root with fixed point $\hat{r}^*$, and for any $\alpha > 0$,

$$l(h) <_\delta (1 + \alpha) l_n(h) + 6 \frac{\alpha + 1}{\alpha} \hat{r}^* + \frac{16\alpha + 5}{\alpha n} \ln \frac{3}{\delta}.$$

Compared to similar result in Section 3.2, the present one is completely data-dependent: In the definition of $\hat{\psi}_n(r)$, the bound $2r/\beta^2$ is on the empirical (as opposed to the expected) loss $l_n(h)$, and the empirical Rademacher average $\mathbb{E}_\sigma R_n$ (as opposed to $\mathbb{E} R_n$) is used. Thus in theory we can obtain a bound without a priori knowledge of the complexity (such as the VC dimension) of the hypothesis space H or of the underlying probability measure μ. However, in practice, computing or estimating $\hat{\psi}_n$ and its fixed point r^* is far from easy, although Bartlett *et al*. [2005, Section 6] had made some initial progress in this direction.

Acknowledgements

This work was supported in part by the U.S. Defense Advanced Research Projects Agency (DARPA) and the U.S. Air Force Research Laboratory under Contract No. F33615-01-C-1848. We thank Michael Elgersma for help with the OAV experiment, and XuanLong Nguyen for many useful comments.

References

P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Mach. Learn.*, 48(1-3):85–113, 2002.

P.L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 2005. To appear.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris, Ser. I*, 334:495–500, 2002.

R. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, MA, 1999.

M.R. Elgersma and B.G. Morton. Nonlinear six-degree-of-freedom aircraft trim. *Journal of Guidance, Control, and Dynamics*, 23(2):305–311, 2000.

D. Haussler. Sphere packing numbers for subsets of Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory (A)*, 69:217–232, 1995.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, 47:443–459, 2000.

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

P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX(2):245–303, 2000.

C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic method for algorithmic discrete mathematics*, pages 195–248. Springer-Verlag, New York, 1997.

D. Panchenko. Some extensions of an inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory*, 13:145–147, 1972.

John Shawe-Taylor, Martin Anthony, and N. L. Biggs. Bounding sample size with the Vapnik-Chervonenkis dimension. *Discrete Appl. Math.*, 42(1):65–73, 1993.

M. Talagrand. Sharper bounds for gaussian and empirical processes. *Annals of Probability*, 22:20–76, 1994.

M. Talagrand. New concentration inequalities for product spaces. *Inventiones Mathematicae*, 126(3):505–563, 1996.

P. Turney. Types of cost in inductive concept learning. In *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning*, pages 15–21, Stanford University, California, 2000.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16:264–280, 1971.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 2000.

M. Vidyasagar. *Learning and Generalization, with Applications to Neural Networks*. Springer, second edition, 2003.