with bias $e_i = \Pr_D(h_i(x) \neq y)$. This is true regardless of how we choose h_i. Consequently, we have:

$$= \prod_{i=1}^{m} E_D \left[e^{\lambda(e_i - \hat{e}_i - \epsilon)} \right] \tag{9.2.1}$$

Since this is true for all λ, we can choose the optimal λ. In general, this is an optimization problem which can be worst-case simplified with the inequality:

$$E_D e^{\lambda(e_i - \hat{e}_i)} \leq e^{\frac{\lambda^2}{8}}$$

which implies:

$$\forall \lambda \ \leq \prod_{i=1}^{m} e^{\frac{\lambda^2}{8} - \lambda\epsilon)}$$

This is optimized for $\frac{2\lambda}{8} = \epsilon$ which implies $\lambda = 4\epsilon$ giving:

$$\leq \prod_{i=1}^{m} e^{-2\epsilon^2}$$

$$= e^{-2m\epsilon^2}$$

□

Open Problem: Starting with 9.2.1, derive a bound similar to the relative entropy chernoff bound which holds for progressive validation

9.3. Combining sample complexity and holdout bounds

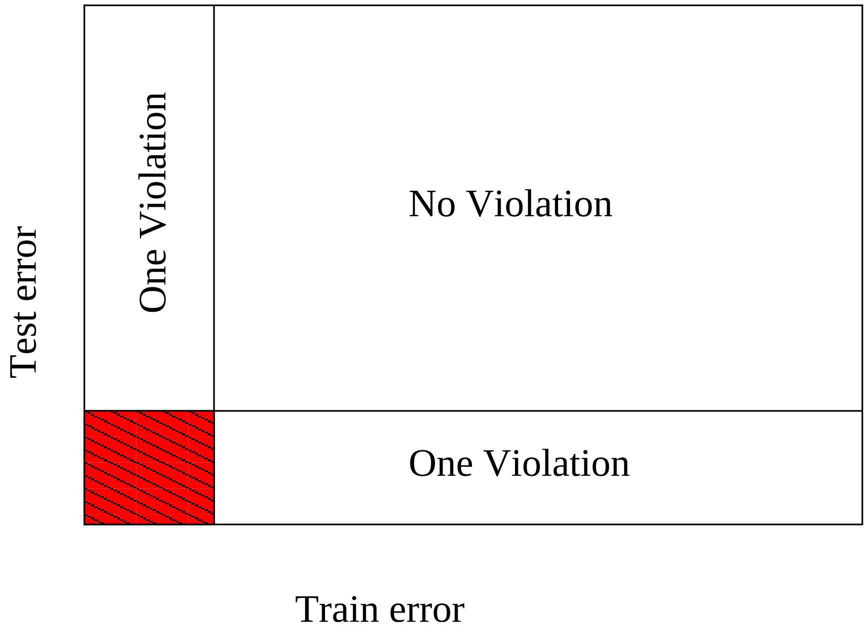
We have two forms of bound, one which uses training set errors and one which uses holdout set errors. The obvious question to ask is: can we combine the information from both bounds? Presumably, if we use both the training error and the test error, we should be able to construct a better confidence interval for the location of the true error rate.

Given a fixed hypothesis and learning problem, we know that the test error will Binomially distributed. Given a fixed learning algorithm and learning problem, the training error will have a considerably more complicated distribution. We can nonetheless, regard the training error as a fixed random variable which has some cumulative distribution parameterized by many parameters, one of which is the true error rate of the output hypothesis (which is itself a random variable).

How can we construct a confidence interval based upon information from both the training and testing sets? There are several possibilities.

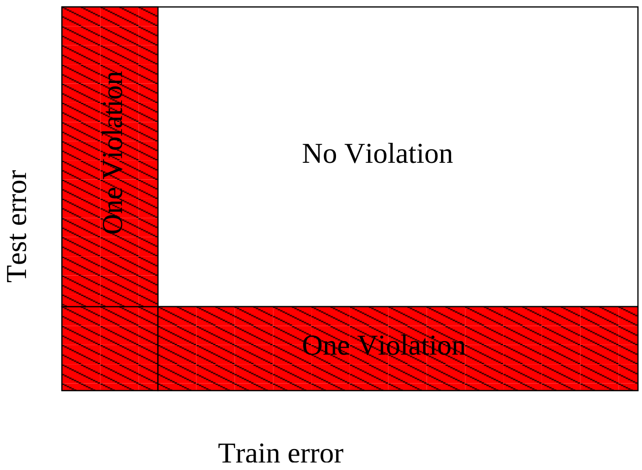
(1) Construct an interval based upon the probability that *both* lower bounds are violated.
(2) Construct an interval based upon the probability that at most one of the lower bounds are violated.
(3) Are there better, more complicated techniques?

Technique (1) can be seen visually by graphing training error vs test error and marking the regions that are bounded away.



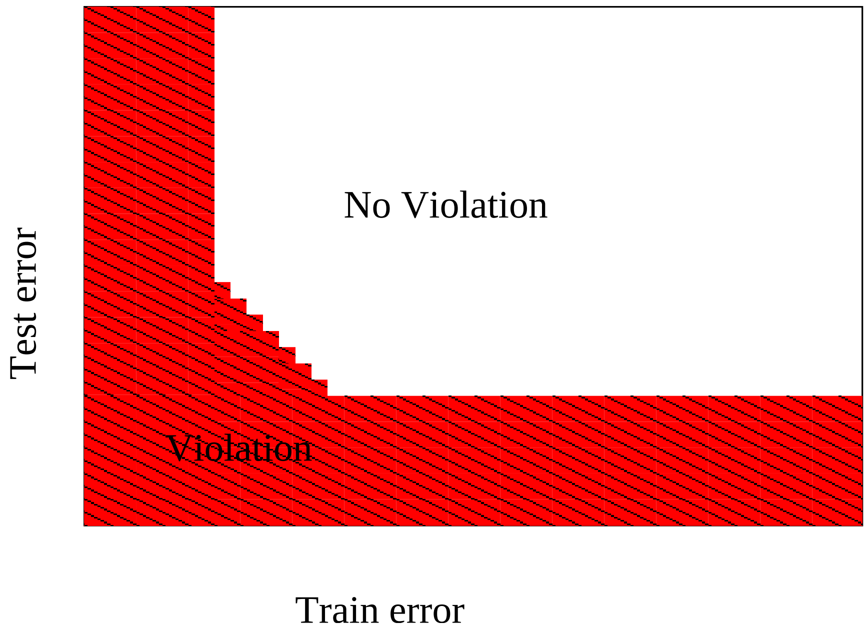
The essential problem with technique (1) is that the resulting true error bound essentially takes the *maximum* (minus a small amount) of the bounds based upon both the test set and the training set. Given that we don't trust either bound to always return tight information, we expect the maximum will not behave well.

Technique (2) can be seen visually in a similar way:



Technique (2) works moderately well. Mathematically, we can calculate the minimum of the two error bounds and add a small amount. This is essentially equivalent to taking a union bound. While this approach allows us to combine the bounds, it does not let us achieve an improvement over either which is intuitively possible.

A better approach may be possible. We would like to construct a rejection region bound of the following form:



Such a rejection region has two important properties:

(1) If one bound is loose, it does not greatly harm the final true error bound.
(2) The final true error bound can be tighter than either individual true error bound.

Showing that technique (2) works is just an application of the union bound. Given any two bounds on the true error rate, we can apportion $\frac{\delta}{2}$ confidence to each bound. Then both bounds will hold with probability δ which implies that the smaller of the two true error bounds holds.

Showing that technique (3) works is more difficult and so we must discuss confidence intervals in greater depth. Fundamentally, a bound can be viewed as a set of outcomes. Let X be space of outcomes, then a bound $\phi \subseteq X$ is a subset. The probability that this generalized bound is violated is given by:

$$\Pr_{x \sim P}(x \in \phi)$$

Typically, we parameterize ϕ with both P and δ to get:

$$\Pr_{x \sim P}(x \in \phi_P(\delta)) \leq \delta$$

We can expand the definition of a high probability set to include a *randomized* high probability set. In particular, let $\phi_P(w, \delta)$ satisfy:

$$\forall w : \Pr_{x \sim P}(x \in \phi_P(w, \delta)) \leq \delta$$

Then, statement such as:

$$\Pr_{w \sim Q, x \sim P}(x \in \phi_P(w, \delta)) \leq \delta$$

In fact, we can make a stronger statement. If

$$E_{w \sim Q} \Pr_{x \sim P}(x \in \phi_P(w, \delta)) \leq \delta \tag{9.3.1}$$

then

$$\Pr_{w\sim Q, x\sim P}(x \in \phi_P(w,\delta)) \leq \delta \tag{9.3.2}$$

.

Randomized confidence intervals are useful here because we can regard the the draw of the "test" set as constructing a randomized interval for the "training" set. As long as constraint 9.3.1 is obeyed, the bound will hold with probability at least δ.

The inexact nature of bounds will force us to impose a monotonicity structure on the random set $\phi_P(w,\delta)$. We can consider any bound $\phi_Q(\delta)$ as inducing a cumulative distribution on the set of w events. This cumulative distribution is *not* the cumulative distribution of the underlying probability. To construct this distribution, let

$$\delta(w) = \min_{\delta}\{\delta : w \in \phi_Q(\delta)\}$$

Thus, $\delta(w)$ is the smallest δ such that the event w is rejected. The value of $\delta(w)$ varies between 0 and 1 so it is a cumulative distribution along a dimension defined by the bound $\phi_Q(\delta)$. Let $F_{\phi_Q}(w)$ be the "real" cumulative distribution along this dimension. Then, the following inequality must hold by the definition of an upper bound:

$$\forall w \;\delta(w) \geq F_{\phi}(w)$$

Given the upper bound cumulative, $\delta(x)$, we can look at distributions satisfying:

$$E_{w\sim\delta(w)} \Pr_{x\sim P}(x \in \phi_P(w,\delta)) \leq \delta$$

If we are guaranteed that $\phi_P(w,\delta)$ is monotonic in $\delta(w)$ (explicitly this is: $\delta(w_1) \geq \delta(w_2) \Rightarrow \phi_P(w_1,\delta) \subseteq \phi_P(w_2,\delta)$) then equation 9.3.2 will hold. This leads to our theorem:

THEOREM 9.3.1. *Let $\phi_P(\delta)$ be any bound satisfying $\Pr_P(x \in \phi_P(\delta)) \leq \delta$ and $\phi_Q(\delta)$ be any bound satisfying $\Pr_Q(w \in \phi_Q(\delta))$, then any randomized set satisfying $\phi_P(w_1,\delta) \subseteq \phi_P(w_2,\delta)$ and $E_{w\sim\delta(w)} \Pr_{x\sim P}(x \in \phi_P(w,\delta)) \leq \delta$ will satisfy:*

$$\Pr_{w\sim Q, x\sim P}(x \in \phi_P(w,\delta)) \leq \delta$$

PROOF.

$$\Pr_{w\sim Q, x\sim P}(x \in \phi_P(w,\delta))$$

Monotonicity and boundedness $\Rightarrow$

$$\begin{aligned} &\leq \Pr_{w\sim\delta(w), x\sim P}(x \in \phi_P(w,\delta)) \\ &= E_{w\sim\delta(w)} \Pr_{x\sim P}(x \in \phi_P(w,\delta)) \\ &\leq \delta \end{aligned}$$

□

The only constraint that we must check in applying a combined training and test set bound is the monotonicity constraint. Heuristically, this is satisfied for the pictures of techniques (1)-(3) because the set of excluded events increases monotonically along the x axis as it decreases along the y axis.

An explicit mathematical form for technique (3) can be given by considering the bound-based cumulative distributions, $\delta(x)$ and $\delta(w)$. In particular, we can define the rejection region to be $\{x, w : \delta(x)\delta(w) \leq t(\delta)\}$ where $t(\delta)$ satisfies

$\Pr_{x \sim \delta(x), w \sim \delta(w)}(\delta(x)\delta(w) \leq t) \leq \delta$. The monotonicity constraint is satisfied by this construction because $\delta(x)$ is implicitly monotonic decreasing with $\delta(w)$ given a constant t. We will use both techniques (2) and (3) for combining training and test sets in the following experiments.