A simpler proof of McAllester's PAC-Bayes result

For conceptual convenience, Let's just talk about the "unrealizable case". In other words, the "bad event" for a concept is the event that its empirical error is far from its true error. Specifically, let's define:

- $err(c)$ is the true error of c, and $\widehat{err}(c)$ is the empirical error of c.
- $\text{badness}(c) = |err(c) - \widehat{err}(c)|$.
- "c is ϵ-bad" if $\text{badness}(c) > \epsilon$.
- $\varepsilon(\delta, m) = \sqrt{\frac{\ln(2/\delta)}{2m}}$.

I will begin with a weaker version of McAllester's Theorem 2, which is easy to prove and I found to be very helpful in thinking conceptually about it. McAllester then uses a cool trick to get his stronger result. I will give a much uglier but conceptually easier way to do that part.

Theorem 1 *For any measure P over concepts, for any $\delta > 0$, for any $p > 0$, with probability $1 - \delta$ over the selection of an IID sample S of m instances, all measurable sets U with $P(U) = p$ satisfy*

$$\mathbf{E}_{c \in U}[\text{badness}(c)] < \sqrt{\frac{\ln(1/p) + \ln(2/\delta) + \ln(m)}{2m}} + \frac{1}{m}$$

Proof. Fix some concept c and some δ'. Hoeffding's bound tells us $\Pr_S[c \text{ is } \varepsilon(\delta', m)\text{-bad}] < \delta'$. Since this is true for any given c, it is certainly true for any fixed distribution over c's. Specifically, define the random variable $B_{\delta'}$ to be the weight (under P) of concepts c that are $\varepsilon(\delta', m)$-bad. Then, $\mathbf{E}_S[B_{\delta'}] < \delta'$.

Markov's inequality therefore says that for any given δ, $\Pr_S[B_{\delta'} > \delta'/\delta] < \delta$. In other words, we expect only a δ'-measure of concepts to be $\varepsilon(\delta', m)$-bad, and it is unlikely we will see a huge amount more than that.

Good. So, now all we need to do is set $\delta' = \delta p/m$. I.e., in words, the probability (over S) that the weight of $\varepsilon(\delta p/m, m)$-bad concepts is greater than p/m is at most δ. This means that with probability $1 - \delta$, *all* sets U of measure p behave well. Formally, with probability $1 - \delta$, for any set U with measure p, at most a measure $\delta p/m$ subset of U consists of $\varepsilon(\delta p/m, m)$-bad concepts. Therefore,

$$\mathbf{E}_{c \in U}[\text{badness}(c)] \leq 1/m + [1 - 1/m] \cdot \varepsilon(\delta p/m, m)$$

and we are done. ∎

The above theorem is weaker than McAllester's because it requires us to commit to the measure of U in advance. What we really want is a theorem that allows us to pick p after the sample is drawn. The essence of the difficulty is that we are examining a 0-1 random variable "is c bad or not". McAllester uses a neat trick of defining a smoother random variable $f_c = 2e^{-2\text{badness}(c)^2 m}$ with the property that $\Pr_S[f_c < \delta] < \delta$, and working directly with that.

However, here is an uglier thing we can do. We can make the above theorem hold simultaneously for the m values of p in $\{1, 1/2, 1/4, \ldots, 1/2^m\}$ by just replacing "δ" with "δ/m" in the bound (equivalently, this puts a factor of 2 in front of the "$\ln m$" term). Now I claim that we can make it hold simultaneously for *all* values of p with the cost of replacing the $\ln(1/p)$ term with $\ln(2/p)$. In particular, if U has $P(U) < 1/2^m$ then the theorem is vacuous so we don't care. On the other hand, if $1/2^i < P(U) < 1/2^{i-1}$ then just apply the theorem to the set $U' \subset U$ consisting of the measure-$(1/2^i)$ worst concepts in U (or actually any set U' of measure $1/2^i$ such that $\mathbf{E}_{c \in U'}[\text{badness}(c)] \geq \mathbf{E}_{c \in U}[\text{badness}(c)]$.