PAC-Bayes Mini-tutorial: A Continuous Union Bound

Tim van Erven

March 5, 2014

Abstract

When I first encountered PAC-Bayesian concentration inequalities they seemed to me to be rather disconnected from good old-fashioned results like Hoeffding's and Bernstein's inequalities. But, at least for one flavour of the PAC-Bayesian bounds, there is actually a very close relation, and the main innovation is a continuous version of the union bound, along with some ingenious applications. Here's the gist of what's going on, presented from a machine learning perspective.

1 The Cramér-Chernoff Method

I will start by outlining the Cramér-Chernoff method, from which Hoeffding's and Bernstein's inequalities and many others follow. This method is incredibly well explained in Appendix A of the textbook by Cesa-Bianchi and Lugosi [3], but I will have to change the presentation a little to easily connect with the PAC-Bayesian bounds later on.

Let $D = ((X_1, Y_1), \ldots, (X_n, Y_n))$ be *independent, identically distributed* (i.i.d.) examples, and let h be a *hypothesis* from a set of hypotheses $\mathcal{H}$, which gets loss $\ell(X_i, Y_i, h)$ on the i-th example. For example, we might think of the squared loss $\ell(X_i, Y_i, h) = (Y_i - h(X_i))^2$. We also define the *empirical error*[1] of h

$$R_n(D, h) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, Y_i, h),$$

and our goal is to prove that the empirical error is close to the *generalisation error*

$$R(h) = \mathbb{E}[\ell(X, Y, h)]$$

with high probability. To do this, we define the function

$$M_\eta(h) = -\tfrac{1}{\eta} \ln \mathbb{E}\left[e^{-\eta \ell(X,Y,h)}\right] \qquad \text{for } \eta > 0,$$

which will act as a surrogate for $R(h)$. Now the Cramér-Chernoff method tells us that:

[1] Called the empirical *risk* in statistics; hence the notation with 'R'.

Lemma 1. *For any $\eta > 0$, $\delta \in (0, 1]$,*

$$M_\eta(h) \leq R_n(h, D) + \frac{1}{\eta n} \ln \frac{1}{\delta} \tag{1}$$

with probability at least $1 - \delta$.

Proof. By Markov's inequality the probability that

$$e^{-\eta n R_n(D,h)} \geq \mathbb{E}_{D'} \left[e^{-\eta n R_n(D',h)} \right] / \delta \tag{2}$$

is at most δ. Now, as the examples are i.i.d., we have

$$\mathbb{E}_{D'} \left[e^{-\eta n R_n(D',h)} \right] = \mathbb{E} \left[e^{-\eta \ell(X,Y,h)} \right]^n. \tag{3}$$

Plugging this in and rewriting, we find that (2) is the complement of the event (1), from which the result follows. □

It remains to relate $M_\eta(h)$ to $R(h)$, which can be done in different ways, and then to optimize η.

1.1 Specialisations

Hoeffding's Inequality To get Hoeffding's inequality, we use Hoeffding's bound [3, Lemma A.1]:

Lemma 2 (Hoeffding)**.** *Suppose $\ell(X, Y, h) \in [a, b]$. Then*

$$R(h) \leq M_\eta(h) + \eta \frac{(b-a)^2}{8}.$$

Plugging this into (1) gives

$$R(h) \leq R_n(D, h) + \eta \frac{(b-a)^2}{8} + \frac{1}{\eta n} \ln \frac{1}{\delta},$$

with probability at least $1 - \delta$. Then plugging in the choice $\eta = \sqrt{\frac{8 \ln(1/\delta)}{n(b-a)^2}}$, which optimizes the bound, yields

$$R(h) \leq R_n(D, h) + \sqrt{\frac{\ln(1/\delta)(b-a)^2}{2n}}$$

with probability at least $1 - \delta$. This is Hoeffding's inequality stated 'inside out'; to recover the usual formulation, define $\epsilon = n\sqrt{\frac{\ln(1/\delta)(b-a)^2}{2n}}$ and solve for δ in terms of ϵ/n, leading to

$$R(h) \leq R_n(D, h) + \frac{\epsilon}{n}$$

with probability at least $1 - \exp\left\{ -2\epsilon^2 / \big(n(b-a)^2\big) \right\}$.

An Alternative Variance-type Inequality There is another inequality that I want to highlight, which is closely related to Bernstein's inequality. It is derived by plugging in the following bound, which is essentially Lemma 10 from my NIPS 2012 paper [6]:

Lemma 3. *Suppose $\ell(X, Y, h) \geq a$ for some $a \leq 0$. Then, for any $\eta \in (0, v]$,*

$$R(h) \leq M_\eta(h) + \eta\phi(-va)\,\mathbb{E}[\ell(X, Y, h)^2].$$

where $\phi(x) = (e^x - x - 1)/x^2$ for $x \neq 0$ and $\phi(0) = 1/2$.

In particular, if $a = 0$, then $\phi(-va) = \phi(0) = 1/2$ for all v, so we can take v to be infinity.

Proof. Let $Z = \ell(X, Y, h)$. Then, by $-\ln x \geq 1 - x$, it is sufficient to show that

$$\mathbb{E}[Z] \leq \tfrac{1}{\eta}\Big(1 - \mathbb{E}\left[e^{-\eta Z}\right]\Big) + \eta\phi(-a\eta)\,\mathbb{E}[Z^2]. \tag{4}$$

Suppose that $\mathbb{E}[Z^2] = 0$. Then $Z = 0$ a.s., and (4) holds with equality. Otherwise, it may be rewritten as

$$\mathbb{E}\left[\frac{(\eta Z)^2}{\mathbb{E}[(\eta Z)^2]} \cdot \phi(-\eta Z)\right] \leq \phi(-a\eta)$$

Recognising the left-hand side as the expectation of $\phi(-\eta Z)$ under the distribution with density $\frac{(\eta Z)^2}{\mathbb{E}[(\eta Z)^2]}$ with respect to the original distribution of Z, we see that it can be bounded by $\max_z \phi(-\eta z)$. As ϕ is increasing, the maximum is achieved at the minimum $z = a$ and $\eta = v$, from which the desired result follows. □

Combining Lemma 3 with Lemma 1, we find that, if $\ell(X, Y, H) \geq a$ for some $a \leq 0$, then for any $\eta \in (0, v]$

$$R(h) \leq R_n(D, h) + \eta\phi(-va)\,\mathbb{E}[\ell(X, Y, h)^2] + \frac{1}{\eta n}\ln\frac{1}{\delta},$$

with probability at least $1 - \delta$. Optimizing η over its allowed range gives a bound with the flavour of Bernstein's inequality, except that we don't necessarily require $\ell(X, Y, h)$ to have mean 0.

Other Standard Inequalities As explained in Appendix A of Cesa-Bianchi and Lugosi [3], different bounds to relate $M_\eta(h)$ to $R(h)$ lead to other inequalities, like for example Bennett's inequality or the standard version of Bernstein's inequality.

2 The Union Bound

Let us get back to the big picture of Lemma 1 before its specialisations from the previous section. Now suppose we use an estimator $\hat{h} \equiv \hat{h}(D) \in \mathcal{H}$ to pick a hypothesis based on the data, for example using empirical risk minimization: $\hat{h} = \arg\min_{h\in\mathcal{H}} R_n(D, h)$. To get a bound for $\hat{h}$ instead of a fixed h, we want (1) to hold for all $h \in \mathcal{H}$ simultaneously. If $\mathcal{H}$ is countable, this can be done using the union bound:

Lemma 4. *Suppose $\mathcal{H}$ is countable. For $h \in \mathcal{H}$, let $\pi(h)$ be any numbers such that $\pi(h) \geq 0$ and $\sum_h \pi(h) = 1$. Then, for any $\eta > 0$, $\delta \in (0, 1]$,*

$$M_\eta(\hat{h}) \leq R_n(D, \hat{h}) + \frac{1}{\eta n} \ln \frac{1}{\pi(\hat{h})\delta} \tag{5}$$

with probability at least $1 - \delta$.

In this context, the function π is often referred to as a *prior distribution*, even though it need not have anything to do with prior beliefs.

Proof. By the union bound and Lemma 1 we have

$$\begin{aligned}
&\Pr\left(M_\eta(\hat{h}) > R_n(D, \hat{h}) + \frac{1}{\eta n} \ln \frac{1}{\pi(\hat{h})\delta}\right) \\
&\qquad \leq \Pr\left(\exists h : M_\eta(h) > R_n(D, h) + \frac{1}{\eta n} \ln \frac{1}{\pi(h)\delta}\right) \\
&\qquad \leq \sum_h \Pr\left(M_\eta(h) > R_n(D, h) + \frac{1}{\eta n} \ln \frac{1}{\pi(h)\delta}\right) \leq \sum_h \pi(h)\delta = \delta. \,\square
\end{aligned}$$

Just like for Lemma 1, we can then again relate $M_\eta(h)$ to $R(h)$ to obtain a bound on the generalisation error, but there is now a slight complication: when we want to optimize η, we find that we are not allowed to, because the optimal choice of η depends on $\hat{h}$, which depends on the data, whereas Lemma 1 only allows a fixed choice of η. In some applications using a fixed η may be good enough, but this does limit the applicability of the result. Luckily, it turns out that we can optimize η "for free":

Lemma 5. *Suppose $\mathcal{H}$ is countable. For $h \in \mathcal{H}$, let $\pi(h)$ be any numbers such that $\pi(h) \geq 0$ and $\sum_h \pi(h) = 1$. Then, for any $\delta \in (0, 1]$,*

$$M_\eta(\hat{h}) \leq R_n(D, \hat{h}) + \frac{1}{\eta n} \ln \frac{1}{\pi(\hat{h})\delta} \qquad \textit{for all } \eta > 0 \tag{6}$$

with probability at least $1 - \delta$.

Proof. Let $\eta(h) = \arg\min_{\eta > 0} \frac{1}{\eta n} \ln \frac{1}{\pi(h)\delta} - M_\eta(\hat{h})$ be the optimal value for η if $\hat{h} = h$. Now apply Lemma 4 with $\eta = 1$ and the scaled loss $\ell'(X, Y, h) = \eta(h)\ell(X, Y, h)$ to obtain

$$-\ln \mathbb{E}\left[e^{-\eta(\hat{h})\ell(X,Y,\hat{h})}\right] \leq \eta(\hat{h}) R_n(D, \hat{h}) + \frac{1}{n} \ln \frac{1}{\pi(\hat{h})\delta} \tag{7}$$

with probability at least $1 - \delta$. Dividing both sides by $\eta(\hat{h})$ gives the result for the choice of η that optimizes the bound. It follows that the bound holds simultaneously for all other η as well. $\square$

This shows, in a nutshell, how one can combine the Cramér-Chernoff method with the union bound to obtain concentration inequalities for estimators $\hat{h}$. The use of the union bound, however, is quite crude when there are multiple hypotheses in $\mathcal{H}$ with very similar losses, and the current proof breaks down completely if we want to extend it to continuous classes $\mathcal{H}$. This is where PAC-Bayesian bounds come to the rescue: in the next section I will explain the PAC-Bayesian generalisation of Lemmas 4 and 5 to continuous hypothesis classes $\mathcal{H}$, which will require replacing $\hat{h}$ by a randomized estimator.

3 PAC-Bayesian Concentration

Let $\hat{\pi} \equiv \hat{\pi}(D)$ be a distribution on $\mathcal{H}$ that depends on the data D, which we will interpret as a randomized estimator: instead of choosing $\hat{h}$ deterministically, we will sample $h \sim \hat{\pi}$ randomly. The distribution $\hat{\pi}$ is often called the PAC-Bayesian *posterior distribution*. Now the result that the PAC-Bayesians have, may be expressed as follows:

Lemma 6. *Let π be a (prior) distribution on $\mathcal{H}$ that does not depend on D, and let $\hat{\pi}$ be a randomized estimator that is allowed to depend on D. Then, for any $\eta > 0$, $\delta \in (0, 1]$,*

$$\mathbb{E}_{h\sim\hat{\pi}}[M_\eta(h)] \leq \mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] + \frac{1}{\eta n}\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta}\Big) \tag{8}$$

with probability at least $1-\delta$. Moreover,

$$\mathbb{E}_D\,\mathbb{E}_{h\sim\hat{\pi}}[M_\eta(h)] \leq \mathbb{E}_D\,\Big[\,\mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] + \frac{1}{\eta n}D(\hat{\pi}\|\pi)\Big]. \tag{9}$$

Here $D(\hat{\pi}\|\pi) = \int \hat{\pi}(h)\ln\frac{\hat{\pi}(h)}{\pi(h)}\mathrm{d}h$ denotes the Kullback-Leibler divergence of $\hat{\pi}$ from π.

Proof of Lemma 6. By (3), we have

$$e^{-\eta n M_\eta(h)} = \mathbb{E}_D\left[e^{-\eta n R_n(D,h)}\right].$$

Hence

$$\begin{aligned}
1 &= \mathbb{E}_{h\sim\pi}\,\mathbb{E}_D\left[\exp\Big\{-\eta n R_n(D,h) + \eta n M_\eta(h)\Big\}\right] \\
&= \mathbb{E}_D\,\mathbb{E}_{h\sim\pi}\left[\exp\Big\{-\eta n R_n(D,h) + \eta n M_\eta(h)\Big\}\right] \\
&= \mathbb{E}_D\,\mathbb{E}_{h\sim\hat{\pi}}\left[\exp\Big\{-\eta n R_n(D,h) + \eta n M_\eta(h) - \ln\frac{\hat{\pi}(h)}{\pi(h)}\Big\}\right] \\
&\geq \mathbb{E}_D\left[\exp\Big\{\underbrace{-\eta n\,\mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] + \eta n\,\mathbb{E}_{h\sim\hat{\pi}}[M_\eta(h)] - D(\hat{\pi}\|\pi)}_{A}\Big\}\right],
\end{aligned}$$

where the inequality is Jensen's. Now notice that (8) is equivalent to $A \leq \ln(1/\delta)$, whereas (9) is equivalent to $\mathbb{E}[A] \leq 0$, and that we have derived that $\mathbb{E}[e^A] \leq 1$. (8) therefore follows by Markov's inequality:

$$\Pr\Big(A > \ln\frac{1}{\delta}\Big) = \Pr(e^A > 1/\delta) \leq \mathbb{E}[e^A]\delta \leq \delta,$$

and (9) follows by another application of Jensen's inequality:

$$e^{\mathbb{E}[A]} \leq \mathbb{E}\left[e^A\right] \leq 1 \qquad \Longrightarrow \qquad \mathbb{E}[A] \leq 0. \qquad \square$$

To see that Lemma 6 generalises Lemma 4, suppose that $\hat{\pi}$ is a point-mass on $\hat{h}$. Then $D(\hat{\pi}\|\pi) = \ln(1/\pi(\hat{h}))$, and we recover Lemma 4 as a special case of (8). An important difference with Lemma 4, however, is that Lemma 6 does not require $\mathcal{H}$ to be countable, and in fact in many PAC-Bayesian applications it is not.

3.1 Optimizing η

Lemma 6 has the same issue as Lemma 4; namely that it does not allow us to optimize η based on $\hat{\pi}$. For the result in expectation (9) I do not really know how to introduce optimization over η in a satisfying way, and we are stuck with a fixed η. For the result in probability (8) we cannot use the same trick that allowed us to optimize η "for free" in Lemma 5, but we can still optimize η at very small cost using the union bound as long as we can find a good lower bound on its range:

Lemma 7. *For any constants $\alpha > 1$ and $0 < u < v$, and any $\delta \in (0, 1]$,*

$$\mathbb{E}_{h\sim\hat{\pi}}[M_\eta(h)] \le \mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] + \frac{\alpha}{\eta n}\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta} + \ln\Big\lceil \log_\alpha \frac{v}{u}\Big\rceil\Big) \qquad \text{for all } \eta \in [u,v] \quad (10)$$

with probability at least $1-\delta$.

Proof. For $i = 0, \dots, \Big\lceil \log_\alpha \frac{v}{u}\Big\rceil - 1$, let $\eta_i = u\alpha^i$. Then for every $\eta \in [u,v]$, there exists an η_i such that $\eta_i \le \eta \le \alpha\eta_i$. Using the union bound to extend (8) to hold uniformly over all η_i, we find that

$$\mathbb{E}_{h\sim\hat{\pi}}[M_{\eta_i}(h)] \le \mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] + \frac{1}{\eta_i n}\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta} + \ln\Big\lceil \log_\alpha \frac{v}{u}\Big\rceil\Big) \qquad \text{for all } \eta_i$$

with probability at least $1-\delta$. Now we use that $M_\eta(h)$ is nonincreasing in η, so that, for any $\eta \in [u,v]$ and η_i such that $\eta_i \le \eta \le \alpha\eta_i$, we have $M_\eta(h) \le M_{\eta_i}(h)$ and $\frac{1}{\eta_i} \le \frac{\alpha}{\eta}$, from which the lemma follows. □

Having an upper bound on the range of η is not an issue, because

$$\min_{0<\eta\le v}\Big(\eta A + \frac{B}{\eta}\Big) \le \min_{\eta>0}\Big(\eta A + \frac{B}{\eta}\Big) + \frac{2B}{v} \qquad \text{for } A, B > 0,$$

which only adds the term $\frac{2B}{v}$, which is always negligible in our case. So it remains to find a good lower bound u for η to plug into Lemma 7. I don't know of a general procedure to do that, but after applying the specialisations from Section 1.1 it actually becomes easy:

Lemma 8 (PAC-Hoeffding)**.** *Suppose $\ell(X,Y,h) \in [a,b]$. Then, for any constants $\alpha > 1$ and $v > 0$, and any $\delta \in (0,1]$,*

$$\mathbb{E}_{h\sim\hat{\pi}}[R(h)] \le \mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] + \eta\frac{(b-a)^2}{8} + \frac{\alpha}{\eta n}\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta} + \ln(\tfrac{1}{2}\log_\alpha n + C)\Big) \qquad \text{for all } \eta \in (0,v] \quad (11)$$

with probability at least $1-\delta$, where $C = \max\{\log_\alpha\big(\frac{v(b-a)}{\sqrt{8\alpha}}\big), 0\} + e$.

Proof. Combining Lemma 7 with Lemma 2, we find for any $u \in (0, v)$

$$\mathbb{E}_{h\sim\hat{\pi}}[R(h)] \leq \mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] + \eta\frac{(b-a)^2}{8} + \frac{\alpha}{\eta n}\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta} + \ln\Big\lceil \log_\alpha \frac{v}{u}\Big\rceil\Big) \quad \text{for all } \eta \in [u, v]$$

with probability at least $1-\delta$. Using that $C \geq e$, the unconstrained value for η that optimizes (11) can be bounded from below by

$$\eta = \sqrt{\frac{8\alpha\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta} + \ln(\frac{1}{2}\log_\alpha n + C)\Big)}{n(b-a)^2}} \geq \sqrt{\frac{8\alpha}{n(b-a)^2}},$$

which does not depend on h. So now we choose $u = \frac{1}{\sqrt{n}}\min\{\sqrt{\frac{8\alpha}{(b-a)^2}}, v\}$, from which the desired result follows. □

Lemma 9 (PAC-Variance)**.** *Suppose $\ell(X,Y,h) \in [a,b]$ with $a \leq 0$. Then, for any constants $\alpha > 1$ and $v > 0$, and any $\delta \in (0,1]$,*

$$\mathbb{E}_{h\sim\hat{\pi}}[R(h)] \leq \mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] + \eta\phi(-av)\,\mathbb{E}[\ell(X,Y,h)^2] + \frac{\alpha}{\eta n}\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta} + \ln(\tfrac{1}{2}\log_\alpha n + C)\Big) \quad \textit{for all } \eta \in (0, v]$$

with probability at least $1-\delta$, where $C = \max\{\frac{1}{2}\log_\alpha\big(\frac{v\max\{a^2,b^2\}\phi(-av)}{\alpha}\big), 0\} + e$.

Proof. Analogously to the proof of Lemma 8, combine Lemma 3 with Lemma 7, and now observe that the minimizing η is at least $\sqrt{\frac{\alpha}{n\phi(-av)\max\{a^2,b^2\}}}$. Then pick $u = \frac{1}{\sqrt{n}}\min\{\sqrt{\frac{\alpha}{\phi(-av)\max\{a^2,b^2\}}}, v\}$ to obtain the result. □

4 Corollaries

Because Lemma 6 works for any choice of loss, we may in particular plug in the relative loss $\ell'(X,Y,h) = \ell(X,Y,h) - \ell(X,Y,h^*)$, where $h^* = \arg\min_{h\in\mathcal{H}} R(h)$ is the hypothesis with smallest generalisation error in $\mathcal{H}$. Combining this, for example, with the PAC-Bayesian version of Hoeffding's lemma (Lemma 8), we obtain:

Corollary 1. *Suppose $\ell(X,Y,h) \in [0,b]$, so that $\ell'(X,Y,h) \in [-b,b]$. Then, for any constants $\alpha > 1$ and $v > 0$, and any $\delta \in (0,1]$,*

$$\mathbb{E}_{h\sim\hat{\pi}}[R(h)] - R(h^*) \leq \mathbb{E}_{h\sim\hat{\pi}}\big[R_n(D,h)\big] - R_n(D,h^*) + \eta\frac{b^2}{2} + \frac{\alpha}{\eta n}\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta} + \ln(\tfrac{1}{2}\log_\alpha n + C)\Big) \quad \textit{for all } \eta \in (0, v] \quad (12)$$

with probability at least $1-\delta$, where $C = \max\{\log_\alpha\big(\frac{2vb}{\sqrt{8\alpha}}\big), 0\} + e$.

And combining with Lemma 9, we get:

Lemma 10. *Suppose $\ell(X,Y,h) \in [0,b]$, so that $\ell'(X,Y,h) \in [-b,b]$. Then, for any constants $\alpha > 1$ and $v > 0$, and any $\delta \in (0,1]$,*

$$\begin{aligned}\mathbb{E}_{h\sim\hat{\pi}}[R(h)] - R(h^*) \leq \\ \mathbb{E}_{h\sim\hat{\pi}}[R_n(D,h)] - R_n(D,h^*) + \eta\phi(bv)\,\mathbb{E}[\ell'(X,Y,h)^2] \\ + \frac{\alpha}{\eta n}\Big(D(\hat{\pi}\|\pi) + \ln\frac{1}{\delta} + \ln(\tfrac{1}{2}\log_\alpha n + C)\Big) \qquad \textit{for all } \eta \in (0,v]\end{aligned}$$

with probability at least $1-\delta$, where $C = \max\{\frac{1}{2}\log_\alpha\big(vb^2\phi(bv)/\alpha\big), 0\} + e$.

5 Choosing the Prior and the Posterior

Even though the names prior and posterior for π and $\hat{\pi}$ suggest some kind of fixed relationship between the two, all the previous results actually hold for any way of choosing these two distributions. This is exploited in applications, in which there appear to be two main approaches:

Optimal Posterior In the first approach, the prior π is fixed, and the posterior $\hat{\pi}$ is chosen as the distribution that optimizes the bound. In Lemmas 7–10 this is always the *Gibbs distribution*

$$\hat{\pi}(h) = \frac{e^{-\frac{\eta}{\alpha}nR_n(D,h)}\pi(h)}{\int e^{-\frac{\eta}{\alpha}nR_n(D,h')}\pi(h')\mathrm{d}h'}. \tag{13}$$

Localised Priors By contrast, in the second approach the posterior $\hat{\pi}$ is fixed, and then the prior π is chosen to (almost) optimize the bound. This way of selecting π was developed by Catoni [2], who refers to such π as *localised priors*. For given $\hat{\pi}$, the prior that exactly optimizes the bound[2] is

$$\pi(h) = \mathbb{E}_D[\hat{\pi}(h)], \tag{14}$$

but when the posterior takes the form (13) another common choice, for which the bound becomes easier to manipulate, is the prior π' defined by

$$\pi'(h) = \frac{e^{-\frac{\eta}{\alpha}nR(h)}\pi(h)}{\int e^{-\frac{\eta}{\alpha}nR(h')}\pi(h')\mathrm{d}h'}$$

for π from the definition of $\hat{\pi}$.

Remark 1. For given prior, the posterior (13) minimizes the bound, and, for given posterior, the prior (14) minimizes the bound. Since both steps reduce the bound, we can conceivably iterate them until convergence. I wonder whether there exists any stability point π such that

$$\pi(h) = \mathbb{E}_D\Big[\frac{e^{-\frac{\eta}{\alpha}nR_n(D,h)}\pi(h)}{\int e^{-\frac{\eta}{\alpha}nR_n(D,h')}\pi(h')\mathrm{d}h'}\Big],$$

and, if so, whether it is unique.

[2] At a NIPS 2013 workshop David McAllester referred to this as "Langford's prior", because apparently John Langford already observed that it optimized the bound 13 years ago, but I don't have a reference.

6 Summary

We have seen how PAC-Bayesian inequalities naturally extend standard concentration inequalities based on the Cramér-Chernoff method by generalising the union bound to a continuous version. There are some technicalities involved if we want to optimize over η, but these can be managed if we can find a good lower bound on the value of the optimizing η. I have not discussed any applications, for which I will have to refer to the references discussed next.

7 Further Reading

I learned about PAC-Bayesian concentration inequalities by discussions with Peter Grünwald about papers by Zhang [7], and by reading the (quite technical) monograph of Catoni [2]. For a much more accessible presentation of Catoni's idea of localised priors and their applications, see the recent paper by Lever, Laviolette and Shawe-Taylor [4]. McAllester also has a recent tutorial [5], which includes an application to analysing drop-out. Except for the connection to standard concentration inequalities, which is probably well known, but which I have not seen emphasised before, all the results I have presented here can be found (more or less) in these references. For more advanced concentration inequalities based on the Cramér-Chernoff method, I also highly recommend the recent textbook by Boucheron, Lugosi and Massart [1], which I am sure will be a classic.

References

[1] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[2] O. Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. IMS Lecture Notes — Monograph Series, Volume 56, 2007.

[3] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[4] G. Lever, F. Laviolette, and J. Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

[5] D. McAllester. A PAC-Bayesian tutorial with a dropout bound. *Preprint posted on the CS arXiv, arXiv:1307.2118 [cs.LG]*, 2013.

[6] T. van Erven, P. D. Grünwald, M. D. Reid, and R. C. Williamson. Mixability in statistical learning. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012.

[7] T. Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.