The Proof of McAllester's PAC-Bayesian Theorem

Matthias Seeger
Institute for Adaptive and Neural Computation
University of Edinburgh
5 Forrest Hill, Edinburgh EH1 2QL
seeger@dai.ed.ac.uk

Abstract

Extends McAllester's PAC-Bayesian theorem [4], allowing for other tail bounds than Hoeffding's, and simplifies the proof somewhat.

1 McAllester's PAC-Bayesian theorem

McAllester's PAC-Bayesian theorem is an astonishing result. Although its proof requires only elementary techniques, it can be used to obtain very tight generalization error bounds for the Gibbs versions of (approximate) Bayesian classification models. In special cases, which are often practically relevant, we can obtain bounds for the corresponding Bayes versions as well.

Suppose we are given a concept space $\mathcal{C}$ and a space of labeled examples $\mathcal{X}$. We denote concepts by $c \in \mathcal{C}$, examples by $\boldsymbol{x} \in \mathcal{X}$. Suppose also we are given a loss function $l(c, \boldsymbol{x})$, mapping pairs $(c, \boldsymbol{x})$ to the interval $[a, b]$ or $[a, b)$ (here, $0 \le a < b$). There is an unknown data distribution D from which examples $\boldsymbol{x}$ can be drawn, and we do not make *any assumptions* on D. For each concept c, we can define the *expected loss* $l(c) = \mathrm{E}[l(c, \boldsymbol{x})]$, where the expectation is taken over $\boldsymbol{x} \sim D$. Given a sample $S = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\}$ of size m, drawn *independently and identically distributed (i.i.d.)* from D, we can also define the *empirical loss* $\hat{l}(c) = m^{-1} \sum_i l(c, \boldsymbol{x}_i)$. Note that $\mathrm{E}_S[\hat{l}(c)] = l(c)$.

Depending on the loss function, it is possible to bound the probability (over S) for large deviations between $\hat{l}(c)$ and its expected value $l(c)$. McAllester's technique takes such a large deviation bound for *single* concepts c and transforms it into a bound for *Gibbs concepts*, which in turn depend on a distribution over all concepts in $\mathcal{C}$ (this is introduced in detail below). Therefore, we need to assume the availability of a large deviation bound for single $\hat{l} = \hat{l}(c)$, $l = l(c)$. Namely, there have to exist constants $C > 0$, $\beta > 0$ and a function ϕ on $(a, b) \times (a, b)$ s.t.

$$\begin{aligned} \Pr\left\{\hat{l} \ge q\right\} &\le C e^{-\beta\phi(q,l)}, \quad q \ge l, \\ \Pr\left\{\hat{l} \le q\right\} &\le C e^{-\beta\phi(q,l)}, \quad q \le l. \end{aligned} \tag{1}$$

Here, $\phi(q, l)$ needs to behave nondecreasingly with increasing $|q - l|$. Furthermore, ϕ needs to be convex as a function of (q, l), and $\phi(\cdot, l)$ must be differentiable every-where (with possible exceptions at l, a, b), for all l. For simplicity, we also assume that $\phi(l,l)=0$. C, β and ϕ must be independent of the concept c.

If the loss function $l(c,\boldsymbol{x})$ is bounded, i.e. $b<\infty$, we can use Hoeffding's bound for (1). W.l.o.g. let $a=0$, $b=1$. Then, Hoeffding's bound is just (1) with $C=1$, $\beta=2m$ and $\phi(q,l)=(q-l)^2$. The convexity of ϕ is evident because of the convexity of the square function. If $l(c,\boldsymbol{x})$ is a zero-one loss, i.e. $l(c,\boldsymbol{x})\in\{0,1\}$ (and $a=0$, $b=1$), the Chernoff bound for sums of independent Bernoulli variables is often much tighter than Hoeffding's bound. It is obtained as (1) with $C=1$, $\beta=m$ and

$$\phi(q,l)=\mathrm{D}_{\mathrm{Ber}}[q\,\|\,l]=q\log\frac{q}{l}+(1-q)\log\frac{1-q}{1-l}. \tag{2}$$

This is simply the relative entropy between two biased coins with heads probabilities q and l respectively. It is well known that $\phi(q,l)$ is nonnegative, increases with $|q-l|$ and is convex (see [1]).

Given a distribution Q over $\mathcal{C}$, the corresponding *Gibbs concept* is defined probabilistically. Namely, it is evaluated on some $\boldsymbol{x}\in\mathcal{X}$ by first sampling a $c\sim Q$, then outputting $(c,\boldsymbol{x})$. If the Gibbs concept is to be evaluated over several points, the corresponding elementary concepts are sampled independently. If we define the *averaged expected loss* as $\mathrm{E}_{c\sim Q}[l(c)]$, the *averaged empirical loss* as $\mathrm{E}_{c\sim Q}[\hat{l}(c)]$, then McAllester's theorem bounds the probability of large deviations between these averaged quantities. The setup of the theorem is inspired by the *Bayesian* framework, in which we have the possibility to encode prior knowledge or assumptions about concepts in a *prior distribution* P. The interpretation of P is that it reflects our belief in which concepts are *a-priori* more sensible in light of what we know or assume about the problem. It is important to understand that, at least in the pragmatic PAC-Bayesian context of McAllester's theorem, it is not required for P to be in any sense the "correct" distribution, i.e. we do not require the unknown data distribution to be generated (in some sense) from a latent "target concept" distributed according to P. PAC-Bayesian is PAC: we are not allowed to make *any* assumptions on D. Furthermore, it is also not required that we encode *all* our prior knowledge into P, we may as well settle for a prior distribution which is computationally convenient or easy to assess. However, P (as well as the concept space $\mathcal{C}$) has to be chosen *before* the data sample S is observed. A second distribution, the *posterior distribution* Q, can be chosen depending on the sample S, and it is this distribution the expected and empirical losses are averaged over in the statement of McAllester's result. Note that in this paper, we allow for rather general spaces $\mathcal{C}$ and distributions Q and P. If Q is absolutely continuous w.r.t. P, i.e. every null set of P is also one of Q, then Q has a density w.r.t. P: the *Radon-Nikodym derivative* $dQ(c)/dP(c)$ (see e.g. [3], chapter 1.4). If $f(c)=dQ(c)/dP(c)$, then f is measurable w.r.t. P, and for every event A:

$$\Pr_{c\sim Q}\{A\}=\mathrm{E}_{c\sim P}\left[f(c)\,\mathrm{I}_{\{c\in A\}}\right].$$

Note that if $\mathcal{C}$ is a finite-dimensional Euclidean space and both Q and P have densities w.r.t. the Lebesgue measure, then $dQ(c)/dP(c)$ is simply the ratio of the densities of Q and P. A well-known divergence measure between Q,P is the *relative entropy*

$$\mathrm{D}[Q\,\|\,P]=\mathrm{E}_{c\sim Q}\left[\log\frac{dQ(c)}{dP(c)}\right] \tag{3}$$

(see [1]). If Q is not absolutely continuous w.r.t. P, we set $\mathrm{D}[Q\,\|\,P]=\infty$. If $\mathcal{Q}$ is finite or countable, we have

$$\mathrm{D}[Q\,\|\,P]=\sum_{c\in\mathcal{C}}\Pr_Q(c)\log\frac{\Pr_Q(c)}{\Pr_P(c)}.$$

In the PAC-Bayesian theorem, $\mathrm{D}[Q \,\|\, P]$ turns out to dominate the expression which bounds the difference between averaged expected and averaged empirical loss in the PAC sense. We can see McAllester's theorem as a theoretical justification of the Bayesian approach to learning. A small averaged expected loss can be guaranteed if a favourable tradeoff between a small averaged empirical loss and a small divergence $\mathrm{D}[Q \,\|\, P]$ can be realized. For a fixed prior P, the *Bayesian posterior distribution* Q is a very good candidate for a close-to-minimal tradeoff. Furthermore, P should be chosen a-priori s.t. later on, we are more likely to be able to realize a minimal tradeoff, given by a posterior Q which is close to P as well as attains a small averaged empirical loss. Thus, we are urged to encode available prior knowledge into P. However, note that P and the concept space $\mathcal{C}$ have to be chosen a-priori, independent of S, therefore popular methods for choosing free prior parameters or model selection, such as empirical Bayes cannot be used[1] if the range of possible parameter values is infinite.

We can now state McAllester's PAC-Bayesian theorem. We choose an arbitrary prior distribution P over concepts and a confidence parameter $\delta \in (0,1)$. Suppose that the large deviation bound of (1) holds, parameterized by C, β and the function $\phi(q,l)$.

Theorem 1 (PAC-Bayesian theorem (McAllester,[4])) *For any data distribution D over $\mathcal{X}$, we have that the following bound holds, where the probability is over random i.i.d. samples $S = \{\boldsymbol{x}_i \mid i = 1,\dots,m\}$ of size m drawn from D:*

$$\Pr_S \left\{ \phi\left(\mathrm{E}_{c\sim Q}[\hat{l}(c)], \mathrm{E}_{c\sim Q}[l(c)]\right) \geq \frac{\mathrm{D}[Q \,\|\, P] + \log 2C\beta + \log \delta^{-1}}{\beta - 1} \right\} < \delta. \qquad (4)$$

Here, Q is an arbitrary posterior distribution over concepts, which may depend on the sample S and on the prior P.

Note that for large deviation bounds typically used for (1), β grows linearly with the sample size m. Now, it is straightforward to compute the upper bound on $\mathrm{E}_{c\sim Q}[l(c)]$ implied by the theorem. Consider $f(l) = \phi(\mathrm{E}_{c\sim Q}[\hat{l}(c)], l) - \varepsilon$ for $l \geq \mathrm{E}_{c\sim Q}[\hat{l}(c)]$. Here, ε is the r.h.s. in the inequality of the "bad" event in (4). $f(l)$ is convex and monotonically nondecreasing, thus a straightforward one-dimensional search reveals a zero l_0 of f. If $f(l)$ is monotonically increasing, l_0 is unique. Otherwise, use the supremum of all zeros. l_0 is an upper bound on $\mathrm{E}_{c\sim Q}[l(c)]$ (with confidence $1-\delta$).

The proof of this theorem, as given in [4], proceeds in two steps, the first of which is given below in lemma 1. Let $\Delta(c) = \phi(\hat{l}(c), l(c))$. In the first step, we try to find a tight upper bound on the expression

$$\mathrm{E}_{S,c\sim P}\left[e^{\alpha\Delta(c)}\right], \quad \alpha > 0.$$

We choose an α s.t. the bound is linear in β. This choice is in a certain sense maximal, because for larger α and general ϕ, the expression may become infinite, and in general the upper bound is at least exponential in β. We then use Markov's inequality to obtain a large deviation inequality with confidence $1-\delta$. Note that this technique is closely related to Chernoff's bounding technique: in order to achieve tightness in the application of Markov's inequality, we take the expectation over the *exponential* of the variable $\Delta(c)$ of interest. In the second step, we assume that we are given a sample S s.t. the inequality holds (for the prior P), thus

$$\mathrm{E}_{c\sim P}\left[e^{\alpha\,\Delta(c)}\right] \leq K. \qquad (5)$$

[1] If the selection ranges over a finite number of models, we can allow model selection by using an obvious union bound argument.

If we can show that

$$\mathrm{E}_{c\sim Q}[\alpha\,\Delta(c)] \le \mathrm{D}[Q\,\|\,P] + \log \mathrm{E}_{c\sim P}\left[e^{\alpha\,\Delta(c)}\right], \tag{6}$$

then we have that

$$\mathrm{E}_{c\sim Q}[\Delta(c)] \le \frac{\mathrm{D}[Q\,\|\,P] + \log K}{\alpha}. \tag{7}$$

In order to show (6), define the Gibbs measure

$$dP_G(c) = \frac{e^{\alpha\,\Delta(c)}}{\mathrm{E}_{c\sim P}\left[e^{\alpha\,\Delta(c)}\right]}\,dP(c),$$

which is a probability measure relative to $P(c)$ (the definition is proper because $\exp(\alpha\,\Delta(c))$ is measurable w.r.t. $P(c)$ and (5) holds). It is a well-known fact that the relative entropy between two distributions is always non-negative (for a proof, see e.g. [3], theorem 1.4.1). Thus,

$$\begin{aligned}
0 \le \mathrm{D}[Q\,\|\,P_G] &= \int \log\left(\frac{\mathrm{E}_{c\sim P}\left[e^{\alpha\,\Delta(c)}\right]}{e^{\alpha\,\Delta(c)}}\,\frac{dQ(c)}{dP(c)}\right) dQ(c) \\
&= \mathrm{D}[Q\,\|\,P] + \log \mathrm{E}_{c\sim P}\left[e^{\alpha\,\Delta(c)}\right] - \mathrm{E}_{c\sim Q}[\alpha\,\Delta(c)].
\end{aligned}$$

From (7), another application of Jensen's inequality (using the convexity of ϕ) will be enough to obtain the proof of theorem 1.

Lemma 1 *For any prior distribution P and any $\delta \in (0,1)$, we have that*

$$\Pr_S\left\{\mathrm{E}_{c\sim P}\left[e^{(\beta-1)\Delta(c)}\right] \ge \frac{2C\beta}{\delta}\right\} < \delta. \tag{8}$$

Here, $\Delta(c) = \phi(\hat{l}(c), l(c))$.

As for the proof, fix an arbitrary $c \in \mathcal{C}$. If we can show that

$$\mathrm{E}_S\left[e^{(\beta-1)\Delta(c)}\right] \le 2C\beta,$$

then the rest follows by averaging c over P and applying Markov's inequality. Write $\hat{l} = \hat{l}(c)$, $l = l(c)$. $\hat{l}$ has a distribution induced by the one on S. Let us try to maximize $\mathrm{E}[\exp((\beta-1)\phi(\hat{l}, l))]$ w.r.t. distributions over $\hat{l}$, subject to the constraints induced from (1), namely:

$$\begin{aligned}
\Pr\{\hat{l} \ge q\} &\le C e^{-\beta\phi(q,l)}, \quad q \ge l, \\
\Pr\{\hat{l} \le q\} &\le C e^{-\beta\phi(q,l)}, \quad q \le l.
\end{aligned}$$

From these constraints and the fact that $\phi(q,l)$ grows with $|q-l|$, it is obvious that a maximizer should fulfil these constraints with equality for all[2] q. The more general problem of finding a measure over $\hat{l}$ with a density $f(\hat{l})$ (which does not necessarily integrates to 1) fulfilling all constraints as equalities is easily solved, by simply differentiating w.r.t. q. We obtain

$$f(\hat{l}) = \mathrm{sgn}(\hat{l} - l)C\beta\frac{d\phi(\hat{l}, l)}{d\hat{l}}e^{-\beta\phi(\hat{l},l)}.$$

[2]This is a simple *water-filling* argument. The density function $f(\hat{l})$ fulfilling all constraints with equality may integrate to more than 1. In this case, the maximizing probability density is obtained by shrinking the initial part of $f(\hat{l})$ to zero. Since this tightens the inequality only insignificantly, we however stick with $f(\hat{l})$ here.

Note that even if the true distribution of $\hat{l}$ does not have a density, it cannot give rise to a larger criterion value. Thus, we have

$$\begin{aligned} \mathrm{E}\left[e^{(\beta-1)\phi(\hat{l},l)}\right] &\le C\beta\int_a^b \operatorname{sgn}(\hat{l}-l)\frac{d\phi}{d\hat{l}}e^{-\phi(\hat{l},l)}\,d\hat{l} = C\beta\int_a^b \operatorname{sgn}(l-\hat{l})\frac{d}{d\hat{l}}\left(e^{-\phi(\hat{l},l)}\right)d\hat{l} \\ &= C\beta\left(2e^{-\phi(l,l)}-e^{-\phi(a,l)}-e^{-\phi(b,l)}\right) \le 2C\beta. \end{aligned}$$

Note that if we replace $(\beta-1)$ in the exponent by $(\beta+\varepsilon)$, $\varepsilon \ge 0$, we do not end up with a bound linear in β. $(\beta-\varepsilon)$ for $\varepsilon \in (0,1)$ is possible, but does not lead to improvement. This completes the proof of the lemma.

Now fix a sample S for which the "unlucky" case in lemma 1 does not occur, i.e. (5) is true with $K = 2C\beta\delta^{-1}$, $\alpha = \beta - 1$. We have already shown above that for such a sample S, (7) must be true as well, thus

$$\phi\left(\mathrm{E}_{c\sim Q}[\hat{l}(c)], \mathrm{E}_{c\sim Q}[l(c)]\right) \le \mathrm{E}_{c\sim Q}[\phi(\hat{l}(c), l(c))] \le \frac{\mathrm{D}[Q \,\|\, P] + \log K}{\beta - 1}.$$

Here, the first inequality is simply an application of Jensen's inequality, using the fact that ϕ is convex. This completes the proof of theorem 1.

We finally note that McAllester's theorem follows directly from lemma 1 which may be possible to prove, under restrictions on the prior distribution P, even if the large deviation inequalities (1) do not hold for all concepts c. For example, we might be interested in potentially unbounded loss functions, such as the classification margin or the squared loss in regression estimation. Typically, in regularized algorithms, while extreme loss values are always possible, they are a-priori extremely unlikely, so that for special cases it may be possible to prove that lemma 1 holds.

Note that in the special case of using $\phi(q,l) = \mathrm{D}_{\mathrm{Ber}}[q \,\|\, l]$, a more direct argument than lemma 1 can be used, leading to a slight tightening of the final result. A direct computation gives

$$\mathrm{E}_S\left[e^{n\mathrm{D}_{\mathrm{Ber}}[\hat{l}\,\|\,l]}\right] = \sum_{m=0}^{n}\binom{n}{m}\left(\frac{m}{n}\right)^m\left(1-\frac{m}{n}\right)^{n-m} = \sum_{m=0}^{n}\binom{n}{m}e^{-n\mathrm{H}[m/n]},$$

where $\mathrm{H}[p] = -p\log p - (1-p)\log(1-p)$ is the entropy of a p-Bernoulli variable. But $\binom{n}{m} \le e^{n\mathrm{H}[m/n]}$ (e.g. [1], chapter 12.1), therefore

$$\mathrm{E}_S\left[e^{n\mathrm{D}_{\mathrm{Ber}}[\hat{l}\,\|\,l]}\right] \le n+1.$$

Thus, in case of the zero-one loss l and $\phi = \mathrm{D}_{\mathrm{Ber}}$, we can replace (4) in theorem 1 by

$$\Pr_S\left\{\phi\left(\mathrm{E}_{c\sim Q}[\hat{l}(c)], \mathrm{E}_{c\sim Q}[l(c)]\right) \ge \frac{\mathrm{D}[Q \,\|\, P] + \log(n+1) + \log\delta^{-1}}{n}\right\} < \delta.$$

1.1 Bayes classifiers

Suppose we are given a binary classification problem, i.e. examples are $\boldsymbol{x} = (x,t)$, $t \in \{-1,+1\}$. Many schemes for this problem are based on a space of *discriminants* $\mathcal{H}$. A discriminant $h \in \mathcal{H}$ maps inputs x to real values $h(x)$, with the interpretation that h classifies x as $\operatorname{sgn} h(x)$, and $|h(x)|$ can be interpreted as confidence h has in this classification. If we define the concept space $\mathcal{C} = \{\operatorname{sgn} h \mid h \in \mathcal{H}\}$, McAllester's theorem applies directly to the corresponding *Gibbs classifier*

$$c_{Gibbs}(x) = \operatorname{sgn} h(x), \ h \sim Q(h).$$

Another way of using Q results in the *Bayes classifier*

$$c_{Bayes}(x) = \operatorname{sgn} \mathrm{E}_{h\sim Q(h)}[h(x)].$$

The differences are evident. While c_{Gibbs} is probabilistic, c_{Bayes} is deterministic. Also, c_{Bayes} takes the size of the output of h into account, while c_{Gibbs} does not. Empirically, c_{Bayes} often outperforms c_{Gibbs}. Nevertheless, McAllester's theorem holds for c_{Gibbs} only. For the special case that all h are themselves binary (i.e. classifiers), one can use the following simple observation made by Herbrich and Gräpel (e.g. [2], lemma 5.3). Fix (x,t). If $\mathrm{E}_{h\sim Q(h)}[h(x)] \neq t$, then $\mathrm{E}_{h\sim Q(h)}[\mathrm{I}_{\{h(x)\neq t\}}] \geq 1/2$. Taking expectations over (x,t), we see that the error of c_{Bayes} can be bounded by twice the expected error of c_{Gibbs}. If $\mathcal{H}$ contains general real-valued hypotheses, we can define the *Bayes voting classifier* as

$$c_{Vote}(x) = \operatorname{sgn} \mathrm{E}_{h\sim Q(h)}[\operatorname{sgn} h(x)],$$

whose error can be bounded by twice the expected error of c_{Gibbs}. In general, c_{Bayes} and c_{Vote} will result in different predictions, but in the special case that for each fixed x, the distribution of $h(x)$, $h \sim Q(h)$ is symmetric around its mean, it is easy to see[3] that they are in fact identical. Namely, fix x and write $h = h(x)$ and $\langle h\rangle = \mathrm{E}_{h\sim Q}[h(x)]$. Then,

$$c_{Bayes}(x)\, c_{Vote}(x) = \operatorname{sgn} \mathrm{E}_{h\sim Q}\left[\operatorname{sgn}\left(\langle h\rangle h\right)\right] = \operatorname{sgn} \mathrm{E}_{u}\left[\operatorname{sgn}\left(\langle h\rangle^2 + \langle h\rangle u\right)\right],$$

where $u = h - \langle h\rangle$ is distributed symmetrically around 0. This shows that $c_{Bayes}(x)\, c_{Vote}(x) = 1$ if $\langle h\rangle \neq 0$, and $c_{Vote}(x) = \operatorname{sgn} 0$ if $\langle h\rangle = 0$, thus $c_{Bayes}(x) = c_{Vote}(x)$. This seemingly special case is frequently encountered in practice, where approximate posteriors are used which often *have* symmetry properties, even though the true intractable posterior is skew.

Acknowledgments

We thank Manfred Opper for many hints concerning simplifications in the proof of the PAC-Bayesian theorem. The author gratefully acknowledges support through a research studentship from *Microsoft Research Ltd.*

References

[1] Thomas Cover and Joy Thomas. *Elements of Information Theory.* Series in Telecommunications. John Wiley & Sons, 1st edition, 1991.

[2] Ralf Herbrich. *Learning Kernel Classifiers.* MIT Press, 1st edition, 2001.

[3] Shunsuke Ihara. *Information Theory for Continuous Systems.* World Scientific, 1st edition, 1993.

[4] David McAllester. PAC-Bayesian stochastic model selection. To appear in Machine Learning. See `www.autoreason.com`, 2002.

[3] Thanks to Manfred Opper for pointing this out.