A Note on the PAC Bayesian Theorem

Andreas Maurer
Adalbertstr. 55
D-80799 München
andreasmaurer@compuserve.com

August 8, 2006

Abstract

We prove general exponential moment inequalities for averages of [0,1]-valued iid random variables and use them to tighten the PAC Bayesian Theorem. The logarithmic dependence on the sample count in the enumerator of the PAC Bayesian bound is halved.

1 Introduction

The relative entropy or Kullback Leibler divergence of a Bernoulli variable with bias p to a Bernoulli variable with bias q is given by

$$KL(p,q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}.$$

Suppose that $\mathbf{X} = (X_1, ..., X_n)$ is a vector of iid random variables, $0 \leq X_i \leq 1$, $E[X_i] = \mu$ and let $M(\mathbf{X}) = (1/n)\sum X_i$ be the arithmetic mean. We will derive the following inequality, valid for $n \geq 8$:

$$E\left[e^{nKL(M(\mathbf{X}),\mu)}\right] \leq 2\sqrt{n} \tag{1}$$

We also show that the square root on the right side gives the optimal order in n because for Bernoulli ($\{0,1\}$-valued) variables X_i we have the additional inequality, valid for $n \geq 2$,

$$\sqrt{n} \leq E\left[e^{nKL(M(\mathbf{X}),\mu)}\right]. \tag{2}$$

We will also see that for Bernoulli variables the right side is independent of μ, so that the expectation $E\left[e^{nKL(M(\mathbf{X}),\mu)}\right]$ is the same for all Bernoulli variables and depends only on n.

It is likely that the inequalities (1) and (2) are known. The upper bound (1) can be applied to improve on the PAC-Bayesian Theorem (see e.g. [9],[11],[13])

in learning theory: Suppose one has a set of data $\mathcal{Z}$ with probability measure D and a set $\mathcal{H}$ of hypotheses $h : \mathcal{Z} \to [0,1]$ (this already includes the usual loss-function). Suppose further that there is a ('prior') probability measure P on $\mathcal{H}$ (assume $\mathcal{Z}$ and $\mathcal{H}$ to be finite to avoid questions of measurability). Then for any $\delta > 0$, with probability greater than $1-\delta$ a sample $\mathbf{S} = (Z_1, ..., Z_n) \in \mathcal{Z}^n$ is drawn from D^n such that for all ('posterior') probability measures Q on $\mathcal{H}$ we have for $n \geq 2$

$$KL\left(E_{h\sim Q}\left[M\left(h\left(\mathbf{S}\right)\right)\right], E_{h\sim Q}\left[E_{z\sim D}\left[h\left(z\right)\right]\right]\right) \leq \frac{KL\left(Q,P\right) + \ln\frac{1}{\delta} + \ln\left(2n\right)}{n-1}. \tag{3}$$

Here $h\left(\mathbf{S}\right)$ refers to the vector $h\left(\mathbf{S}\right) = \left(h\left(Z_1\right), ..., h\left(Z_n\right)\right)$, so that $M\left(h\left(\mathbf{S}\right)\right)$ is the empirical loss of the hypothesis h. The expression $KL\left(Q,P\right)$ refers to the relative entropy of the probability measures Q and P (see [5]). The importance to learning theory comes from the fact that Q may depend on $\mathbf{S}$. Note that (3) implies

$$E_{h\sim Q}\left[E_{z\sim D}\left[h\left(z\right)\right]\right] \leq \sup\left\{\epsilon : KL\left(E_{h\sim Q}\left[M\left(h\left(\mathbf{S}\right)\right)\right], \epsilon\right) \leq \frac{KL\left(Q,P\right) + \ln\frac{2n}{\delta}}{n-1}\right\},$$

which can drive a learning algorithm to select a posterior Q by minimizing the sample-dependent right side . Among other applications ([7], [13]) the PAC Bayesian bound has been applied to prove generalisation error bounds for large margin classifiers such as support vector machines ([8], [11]).

The right side of (3) has, with an overall factor of $1/\left(n-1\right)$, three terms: There is the relative entropy $KL\left(Q,P\right)$, which can be interpreted as the information gain in specializing from P to Q, an information normally extracted from the sample $\mathbf{S}$. The term $\ln\left(1/\delta\right)$ expresses the usual dependence on the confidence parameter δ, but the remaining $\ln\left(2n\right)$ is difficult to understand: Why do we need it, can't it be altogether eliminated or at least reduced?

We do not know the answer to the first two questions, but using (1) we can essentially cut the term in half, replacing $\ln\left(2n\right)$ by $\ln\left(2\sqrt{n}\right)$ for $n \geq 8$ and reduce the overall factor to $1/n$. Our substitute for (3) then reads

$$KL\left(E_{h\sim Q}\left[M\left(\mathbf{S}\right)\right], E_{h\sim Q}\left[E_{z\sim D}\left[h\left(z\right)\right]\right]\right) \leq \frac{KL\left(Q,P\right) + \ln\frac{1}{\delta} + \ln\left(2\sqrt{n}\right)}{n}. \tag{4}$$

Our improvement is not spectacular, but significant when viewed in terms of the confidence parameter δ. It gives a slightly smaller generalisation error bound (factor $\left(n-1\right)/n$) than (3) with a failure probability δ decreased by the factor $1/\sqrt{n}$. For example, if $n = 10000$ and (3) gives a fixed bound with a failure probability of $1/100$, our result will give the same bound with failure probability less than $1/10000$.

It is possible to prove bounds similar to the above (see [3] and [1]), where the $\ln\left(n\right)$ dependence is replaced by $\ln\left(\ln\left(n\right)\right)$ or eliminated alltogether, at the expense of multiplying $KL\left(Q,P\right)$ with a constant larger than unity. The

relative entropy $KL(Q,P)$ however is dependent on the posterior Q and thus implicitely on the sample and the sample-size n. In all cases where $KL(Q,P)$ grows faster than logarithmically in n (the generic case in machine learning) these bounds will therefore be weaker than (4) above.

We will prove the principal bounds (1) and (2) in section 2. We will then apply them to the PAC Bayesian Theorem in section 3.

2 Main Inequalities

Throughout this note $X_1, ..., X_n$ are assumed to be IID random variables with values in $[0,1]$ and expectation $E[X_i] = \mu$. We use $\mathbf{X}$ to denote the corresponding random vector $\mathbf{X} = (X_1, ..., X_n)$ with values in $[0,1]^n$ and $M(\mathbf{X})$ to denote its arithmetic mean

$$M(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

For any $[0,1]$-valued random variables X use X' to denote the unique Bernoulli ($\{0,1\}$-valued) random variable with $\Pr\{X' = 1\} = E[X'] = E[X]$. Evidently $X'' = X'$, $\forall X$. For $\mathbf{X} = (X_1, ..., X_n)$ we denote $\mathbf{X}' = (X_1', ..., X_n')$.

We restate our principal bounds in a slightly more general way.

Theorem 1 *For all $n \geq 2$*

$$E\left[e^{nKL(M(\mathbf{X}),\mu)}\right] \leq E\left[e^{nKL\left(M\left(\mathbf{X}'\right),\mu\right)}\right] \leq e^{\frac{1}{12n}}\sqrt{\frac{\pi n}{2}} + 2. \tag{5}$$

If the X_i are nontrivial Bernoulli variables (i.e. if $\mu \in (0,1)$) then there is a sequence c_n such that $1 \leq c_n \to \pi$ as $n \to \infty$ and

$$e^{-\frac{1}{6}}\sqrt{\frac{n}{2\pi}}c_n + 2 \leq E\left[e^{nKL(M(\mathbf{X}),\mu)}\right]. \tag{6}$$

In this case the expectation on the right is independent of μ.

The right side of (5) is bounded above by $2\sqrt{n}$ for $n \geq 8$ and the left side of (6) is bounded below by $\sqrt{n}$ for $n \geq 2$, thus giving the simpler bounds (1) and (2) of the introduction.

To prove Theorem 1 we need some auxilliary results. The first is Stirling's Formula:

Theorem 2 *For $n \in \mathbb{N}$*

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{g(n)}{12n}} \tag{7}$$

with $0 < g(n) < 1$.

For a proof see e.g. [2]. We will use this Theorem in form of the following inequalities

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}. \tag{8}$$

The following simple lemma shows that the expectation of a convex function of iid variables can always be bounded by the expectation of the corresponding Bernoulli variables.

Lemma 3 *Suppose that $f : [0,1]^n \to R$ is convex. Then*

$$E\left[f\left(\mathbf{X}\right)\right] \leq E\left[f\left(\mathbf{X}'\right)\right].$$

If f is permutation symmetric in its arguments and $\boldsymbol{\theta}(k)$ denotes the vector $\boldsymbol{\theta}(k) = (1,...,1,0,...,0)$ in $\{0,1\}^n$, whose first k coordinates are 1 and whose remaining $n-k$ coordinates are zero, we also have

$$E\left[f\left(\mathbf{X}'\right)\right] = \sum_{k=0}^{n}\binom{n}{k}\left(1-\mu\right)^{n-k}\mu^k f\left(\boldsymbol{\theta}\left(k\right)\right).$$

Proof. A straightforward argument by induction shows that we can write any point $\mathbf{x} = (x_1,...,x_n) \in [0,1]^n$ as a convex combination of the extremepoints $\boldsymbol{\eta} = (\eta_1,...,\eta_n) \in \{0,1\}^n$ of $[0,1]^n$ in the following way:

$$\mathbf{x} = \sum_{\boldsymbol{\eta}\in\{0,1\}^n}\left(\prod_{i:\eta_i=0}\left(1-x_i\right)\prod_{i:\eta_i=1}x_i\right)\boldsymbol{\eta}.$$

Convexity of f therefore implies

$$f\left(\mathbf{x}\right) \leq \sum_{\boldsymbol{\eta}\in\{0,1\}^n}\left(\prod_{i:\eta_i=0}\left(1-x_i\right)\prod_{i:\eta_i=1}x_i\right)f\left(\boldsymbol{\eta}\right),$$

with equality if $\mathbf{x} \in \{0,1\}^n$. Taking the expectation and using independence and $E\left[X_i\right] = \mu$ we get

$$E\left[f\left(\mathbf{X}\right)\right] \leq \sum_{\boldsymbol{\eta}\in\{0,1\}^n}\left(\prod_{i:\eta_i=0}\left(1-\mu\right)\prod_{i:\eta_i=1}\mu\right)f\left(\boldsymbol{\eta}\right).$$

This becomes an equality if $\mathbf{X}$ is Bernoulli, for then $\mathbf{X}$ takes values only in $\{0,1\}^n$. In particular $E\left[f\left(\mathbf{X}\right)\right] \leq E\left[f\left(\mathbf{X}'\right)\right]$, which gives the first assertion. If f is permutation symmetric then $f\left(\boldsymbol{\eta}\right) = f\left(\boldsymbol{\theta}\left(\left|\{i:\eta_i=1\}\right|\right)\right)$ and we can rewrite the sum above as

$$\begin{aligned}&\sum_{\boldsymbol{\eta}\in\{0,1\}^n}\left(1-\mu\right)^{\left|\{i:\eta_i=0\}\right|}\mu^{\left|\{i:\eta_i=1\}\right|}f\left(\theta\left(\left|\{i:\eta_i=1\}\right|\right)\right)\\ &= \sum_{k=0}^{n}\binom{n}{k}\left(1-\mu\right)^{n-k}\mu^k f\left(\theta\left(k\right)\right).\end{aligned}$$

■

The next lemma is concerned with a series which can be viewed as a Rieman sum approximating an instance of the Beta-function.

Lemma 4 *For $n \geq 2$ the sequence*

$$c_n = \sum_{k=1}^{n-1} \frac{1}{\sqrt{k\,(n-k)}}$$

satisfies $1 \leq c_n \leq \pi$, and $c_n \to \pi$ as $n \to \infty$.

Proof. Define a function ψ on $(0,1)$ by

$$\psi\,(t) = \frac{1}{\sqrt{t\,(1-t)}}.$$

The change of variables $t \to \cos^2\theta$ shows that

$$\int_0^1 \psi\,(t)\,dt = \pi.$$

It follows from elementary calculus that ψ has a unique minimum at $t = 1/2$ with minimal value 2. This implies that $1/\sqrt{k\,(n-k)} \geq 2/n$ and therefore $c_n \geq 2\,(n-1)\,/n \geq 1$ for $n \geq 2$. It also implies that the functions ψ_n defined on $(0,1)$ by

$$\psi_n\,(t) = \begin{cases} \frac{1}{\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)}} & \text{if} \quad t \in \left[\frac{k-1}{n}, \frac{k}{n}\right) \quad \text{and} \quad k \leq n/2 \\ 0 & \text{if} \quad t \in \left[\frac{k-1}{n}, \frac{k}{n}\right) \quad \text{and} \quad k-1 \leq n/2 < k \\ \frac{1}{\sqrt{\frac{k-1}{n}\left(1-\frac{k-1}{n}\right)}} & \text{if} \quad t \in \left[\frac{k-1}{n}, \frac{k}{n}\right) \quad \text{and} \quad n/2 < k-1 \end{cases}$$

satisfy $\psi_n \leq \psi$. Since

$$c_n = \sum_{k=1}^{n-1} \frac{1}{\sqrt{k\,(n-k)}} = \sum_{k=1}^{n-1} \frac{1}{n\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)}} = \int_0^1 \psi_n\,(t)\,dt$$

this implies that $c_n \leq \pi$. Also $\psi_n \to \psi$ a.e. so that by dominated convergence

$$c_n = \int_0^1 \psi_n\,(t)\,dt \to \int_0^1 \psi\,(t)\,dt = \pi.$$

■

Proof of Theorem 1. If X_i is trivial (i.e. if $\mu \in \{0,1\}$) (5) is evident, so we can assume $\mu \in (0,1)$. Define

$$f : \mathbf{x} \in [0,1]^n \mapsto \exp\left(nKL\left(\frac{1}{n}\sum_{i=1}^{n} x_i, \mu\right)\right).$$

Since the average is linear and KL is convex (see [5]) and the exponential function is nondecreasing and convex, the function f is also convex. f is clearly permutation symmetric in its arguments. Lemma 3 immediately gives

$$E\left[e^{nKL(M(\mathbf{X}),\mu)}\right] \leq E\left[e^{nKL\left(M(\mathbf{X}'),\mu\right)}\right] = \sum_{k=0}^{n} \binom{n}{k} (1-\mu)^{n-k} \mu^k f\left(\boldsymbol{\theta}\left(k\right)\right). \tag{9}$$

establishing also the first inequality in (5). Using the special form of the function f we find

$$f\left(\boldsymbol{\theta}\left(k\right)\right) = \exp\left(nKL\left(\frac{k}{n},\mu\right)\right) = \left(\frac{n-k}{n\left(1-\mu\right)}\right)^{n-k} \left(\frac{k}{n\mu}\right)^{k}.$$

Substitution in (9) leads to cancellation of the dependence in μ (proving the last statement of the theorem) and gives

$$\begin{aligned} E\left[e^{nKL\left(M(\mathbf{X}'),\mu\right)}\right] &= \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^{k} \left(\frac{n-k}{n}\right)^{n-k} \\ &= \frac{n!}{n^n} \sum_{k=1}^{n-1} \frac{k^k}{k!} \frac{(n-k)^{n-k}}{(n-k)!} + 2. \end{aligned}$$

Using Stirling's formula (8) and Lemma 4 on the last expression we obtain

$$\begin{aligned} &E\left[e^{nKL\left(M(\mathbf{X}'),\mu\right)}\right] \\ &\leq \sqrt{2\pi n}\left(\frac{1}{e}\right)^{n} e^{\frac{1}{12n}} \sum_{k=1}^{n-1} \frac{1}{\sqrt{2\pi k}\left(\frac{1}{e}\right)^{k}} \frac{1}{\sqrt{2\pi\left(n-k\right)}\left(\frac{1}{e}\right)^{n-k}} + 2 \\ &= e^{\frac{1}{12n}} \sqrt{\frac{n}{2\pi}} \sum_{k=1}^{n-1} \frac{1}{\sqrt{k\left(n-k\right)}} + 2 \\ &\leq e^{\frac{1}{12n}} \sqrt{\frac{\pi n}{2}} + 2, \end{aligned}$$

which gives (5). Similarly

$$\begin{aligned} E\left[e^{nKL\left(M(\mathbf{X}'),\mu\right)}\right] &\geq e^{-\frac{1}{6}} \sqrt{\frac{n}{2\pi}} \sum_{k=0}^{n} \frac{1}{\sqrt{k\left(n-k\right)}} + 2 \\ &= e^{-\frac{1}{6}} \sqrt{\frac{n}{2\pi}} c_n + 2, \end{aligned}$$

which gives (6) for Bernoulli variables. ■

3 Application to the PAC-Bayesian Theorem

Consider an unknown probability distribution D on a set $\mathcal{Z}$, and a set $\mathcal{H}$ of hypotheses $h : \mathcal{Z} \to [0,1]$ (includes the loss function). To avoid a discussion

of measurability Z and H are both assumed to be finite: Their cardinality is otherwise irrelevant and will not appear in our results. The sample $\mathbf{S} = (Z_1, ..., Z_n)$ is a $\mathcal{Z}^n$-valued random vector drawn from $\Pr = D^n$. For $h \in \mathcal{H}$ we use $h(\mathbf{S})$ to denote the $[0,1]$-valued random vector $h(\mathbf{S}) = (h(Z_1), ..., h(Z_n))$. We write

$$h(D) = E_{z\sim D}[h(z)] \text{ and } M(h(\mathbf{S})) = \frac{1}{m}\sum_{i=1}^{m} h(Z_i).$$

If Q is a probability measure on $\mathcal{H}$, we write

$$Q(D) = E_{h\sim Q}[h(D)] \text{ and } Q(\mathbf{S}) = E_{h\sim Q}[M(h(\mathbf{S}))].$$

The relative entropy of two probability measures Q and P on a set $\mathcal{H}$, denoted $KL(Q,P)$, is defined to be infinite if Q is not absolutely continuous w.r.t. P. Otherwise, if $\frac{dQ}{dP}$ is the density of Q w.r.t. P, we set

$$KL(Q,P) = E_Q\left[\ln \frac{dQ}{dP}\right].$$

Theorem 5 *We have for any probability distribution P on $\mathcal{H}$, for $n \geq 8$ and $\forall \delta > 0$*

$$\Pr_S\left\{\exists Q : KL(Q(\mathbf{S}), Q(D)) > \frac{KL(Q,P) + \ln\frac{1}{\delta} + \ln(2\sqrt{n})}{n}\right\} \leq \delta. \tag{10}$$

Proof. For every hypothesis $h \in \mathcal{H}$, applying the bound (1) to the random vector $h(\mathbf{S})$ gives

$$E_{\mathbf{S}}\left[e^{nKL(M(h(\mathbf{S})),h(D))}\right] \leq 2\sqrt{n}.$$

Let $\mathbf{S} \mapsto Q_{\mathbf{S}}$ be any map from samples to the probability distributions on $\mathcal{H}$ (a learning algorithm for Gibbs classifiers). Using Jensen's inequality, convexity of the KL-divergence and of the exponential function, we have

$$\begin{aligned}
&E_{\mathbf{S}}\left[\exp\left(nKL(Q_{\mathbf{S}}(\mathbf{S}), Q_{\mathbf{S}}(D)) - KL(Q_{\mathbf{S}}, P)\right)\right] \\
&\leq E_{\mathbf{S}}\left[\exp\left(E_{h\sim Q_{\mathbf{S}}}\left[nKL(M(h(\mathbf{S})), h(D)) - \ln\frac{dQ_{\mathbf{S}}}{dP}(h)\right]\right)\right] \\
&\leq E_{\mathbf{S}}\left[E_{h\sim Q_{\mathbf{S}}}\left[\exp\left(nKL(M(h(\mathbf{S})), h(D)) - \ln\frac{dQ_{\mathbf{S}}}{dP}(h)\right)\right]\right] \\
&= E_{\mathbf{S}}\left[E_{h\sim P}\left[e^{nKL(M(h(\mathbf{S})),h(D))}\left(\frac{dQ_{\mathbf{S}}}{dP}\right)^{-1}\left(\frac{dQ_{\mathbf{S}}}{dP}\right)\right]\right] \\
&\leq E_{h\sim P}\left[E_{\mathbf{S}}\left[e^{nKL(M(h(\mathbf{S})),h(D))}\right]\right] \\
&\leq 2\sqrt{n}.
\end{aligned}$$

Finally, by Markov's inequality,

$$\begin{aligned}\delta &\geq \Pr_{\mathbf{S}}\left\{e^{nKL(Q_{\mathbf{S}}(\mathbf{S}),Q_{\mathbf{S}}(D))-KL(Q_{\mathbf{S}},P)} > \frac{2\sqrt{n}}{\delta}\right\} \\ &= \Pr_{\mathbf{S}}\left\{KL\left(Q_{\mathbf{S}}(\mathbf{S}),Q_{\mathbf{S}}(D)\right) > \frac{KL\left(Q_{\mathbf{S}},P\right)+\ln\frac{1}{\delta}+\ln\left(2\sqrt{n}\right)}{n}\right\}.\end{aligned}$$

An appropriate worst-case choice of the function $S \mapsto Q_{\mathbf{S}}$ gives (10). ■

The loosest step in the proof is the use of Markov's inequality. The lower bound (2) can be used to show that the other inequalities are rather tight: Let $\mathcal{Z}$ be any large finite set, $\mathcal{H}$ a set of functions $h : Z \to \{0,1\}$ and D a distribution such that the members of $\mathcal{H}$ all induce nontrivial Bernoulli variables i.e. $E_{z\sim D}[h] \in (0,1), \forall h \in \mathcal{H}$. Let P be uniform on $\mathcal{H}$. For a sample $\mathbf{S}$ we define the posterior $Q_{\mathbf{S}}$ by its density w.r.t. P:

$$\frac{dQ_{\mathbf{S}}}{dP}(h) = \frac{e^{mK(M(h(\mathbf{S})),h(D))}}{E_{h\sim P}\left[e^{mK(M(h(\mathbf{S})),h(D))}\right]}.$$

Then

$$\psi(h,\mathbf{S}) = mK\left(M\left(h\left(\mathbf{S}\right)\right),h\left(D\right)\right) - \ln\frac{dQ_{\mathbf{S}}}{dP} = \ln E_{h\sim P}\left[e^{mK(M(h(\mathbf{S})),h(D))}\right]$$

is independent of h. Therefore with (2)

$$\begin{aligned}E_{\mathbf{S}}\left[e^{E_{h\sim Q}\left[mK(M(h(\mathbf{S})),h(D))-\ln\frac{dQ_S}{dP}\right]}\right] &= E_{\mathbf{S}}\left[e^{E_{h\sim Q}[\psi(h,\mathbf{S})]}\right] \\ &= E_{\mathbf{S}}\left[E_{h\sim Q}\left[e^{\psi(h,\mathbf{S})}\right]\right] \\ &= E_{h\sim P}\left[E_{\mathbf{S}}\left[e^{mK(M(h(S)),h(D))}\right]\right] \\ &\geq \sqrt{n}.\end{aligned}$$

This can be rewritten as the statement: For every $\delta > 0$

$$E_{\mathbf{S}}\left[\exp\left(mE_{h\sim Q}\left[K\left(M\left(h\left(\mathbf{S}\right)\right),h\left(D\right)\right)\right] - KL\left(Q,P\right)+\ln\frac{1}{\delta}+\ln\sqrt{m}\right)\right] \geq \delta,$$

a very weak lower bound version of the PAC-Bayesian theorem, which nevertheless shows that an elimination or the $\sqrt{m}$ term, if it is at all possible, would have to follow a completely different path.

References

[1] Jean-Yves Audibert, Olivier Bousquet, "PAC-Bayesian generic chaining", http://www.kyb.mpg.de/publications/pdfs/pdf2341.pdf.

[2] Heinz Bauer, *Wahrscheinlichkeitstheorie,* De Gruyter New York, 2002

[3] Olivier Catoni, "A PAC-Bayesian Approach to adaptive classification", http://www.proba.jussieu.fr/mathdoc/textes/PMA-840.pdf.

[4] Herman Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations", *Annals of Mathematical Statistics*, 23:493-507, 1952.

[5] Thomas Cover, Joy Thomas, *Elements of Information Theory*, Wiley, 1991

[6] Wassily Hoeffding, "Probability inequalities for sums of bounded random variables", *Journal of the American Statistical Association*, 58:13-30, 1963.

[7] John Langford and Matthias Seger, "Bounds for averaging classifiers", *CMU Technical report*, CMU-CS-01-102, 2002

[8] John Langford and John Shawe-Taylor, "PAC Bayes and Margins", *Neural Information Processing Systems (NIPS)*, 2002

[9] David McAllester, "Some PAC-Bayesian Theorems", *Proceedings of the Eleventh Annual Conference In Computational Learning Theory*, 230-234, 1998.

[10] David McAllester, "PAC-Bayesian Stochastic Model Selection", *Machine Learning*, 5:5-21, 2003

[11] David McAllester, "Simplified PAC-Bayesian Margin Bounds", *COLT 03*, 2003

[12] Colin McDiarmid, "Concentration", in *Probabilistic Methods of Algorithmic Discrete Mathematics*, p. 195-248. Springer, Berlin, 1998.

[13] Matthias Seger, "PAC Bayesian generalisation bounds for Gaussian processes", *Journal of Machine Learning Research*, 3:233-269, 2002