On Bayesian Bounds

Arindam Banerjee BANERJEE@CS.UMN.EDU
Dept of Computer Science & Engg, University of Minnesota, Twin Cities

Abstract

We show that several important Bayesian bounds studied in machine learning, both in the batch as well as the online setting, arise by an application of a simple compression lemma. In particular, we derive (i) PAC-Bayesian bounds in the batch setting, (ii) Bayesian log-loss bounds and (iii) Bayesian bounded-loss bounds in the online setting using the compression lemma. Although every setting has different semantics for prior, posterior and loss, we show that the core bound argument is the same. The paper simplifies our understanding of several important and apparently disparate results, as well as brings to light a powerful tool for developing similar arguments for other methods.

1. Introduction

Prediction is widely studied under two settings: batch and online. In the batch setting, data is assumed to be generated from a fixed but unknown distribution (Langford, 2005). Prediction performance of a specific algorithm, such as support vector machine or boosting, is theoretically evaluated using PAC bounds. The PAC-Bayesian bound (McAllester, 2003a) is currently one of the most widely used results for proving algorithm specific bounds (McAllester, 2003b). In the online setting, prediction proceeds in iterations and no assumptions are made regarding how the data is being generated. Data can be generated by an adversary and the performance of a specific algorithm, such as weighted majority (Littlestone & Warmuth, 1994) or hedging (Freund & Schapire, 1997), is evaluated using cumulative loss in the worst case. Typically, one computes a bound on the cumulative loss relative to a fixed "expert" or a fixed distribution over experts. Although the settings are rather different, some powerful ideas from online learning have been successfully used in the batch setting. This includes direct applications such as that of the leave-one-out method (Helmbold & Warmuth, 1995) to voted perceptrons (Freund & Schapire, 1999b), as well as indirect applications such as weighted majority (Littlestone & Warmuth, 1994) motivating adaboost (Freund & Schapire, 1997).

In this paper, we focus on the theoretical methods for performance evaluation in the two settings. In particular, we focus on PAC-Bayesian bounds (McAllester, 2003a) in the batch setting, and log-loss (Freund et al., 1997; Kakade & Ng, 2004) and bounded-loss bounds (Freund & Schapire, 1997; Freund & Schapire, 1999a) in the online setting. We show that all these bounds result from an application of a simple compression lemma, along with some additional setting specific arguments. From a Bayesian perspective, each setting has different semantics for prior, posterior and loss. However, the core bound argument is exactly the same.

There are two desirable aspects to our treatment:

1. It gives a unified and simple understanding of the most widely studied and apparently disparate bounds in machine learning.

2. It explicitly brings to light a powerful tool for developing related inequalities and bounds.

We note that the original arguments for each of the bounds considered in this paper were ingeniously derived, and all of them implicitly or explicitly had to prove and use the compression lemma that unify these bounds.

The rest of the paper is organized as follows. In Section 2, we discuss the compression lemma that will be applied throughout the rest of the paper. In Section 3, we derive the PAC-Bayesian bound using the compression lemma. Using the same result, in Section 4 we prove log-loss bounds for adaptive probabilistic prediction models in the online setting. In Section 5, we present a similar analysis for arbitrary bounded loss

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

functions, the most common setting for studying online learning algorithms. We highlight further connections in Section 6, and conclude in Section 7.

2. A Compression Lemma

In this section, we present and discuss a simple compression lemma that will be repeatedly used to derive the batch as well as the online Bayesian bounds in subsequent sections. The lemma can also be viewed as a special case of Fenchel's inequality, which is a powerful class of "best" inequalities using the concept of a conjugate of a convex function (see Appendix A for an exposition).

Since we are interested in Bayesian prediction, let $\mathcal{H}$ be a set of predictors under consideration. Our results hold without the semantics that $h \in \mathcal{H}$ are predictors, but it serves our purpose well. The compression lemma can be simply stated as follows:

Lemma 1 (Compression Lemma) *For any measurable function $\phi(h)$ on $\mathcal{H}$, and any distributions P and Q on $\mathcal{H}$, we have*

$$E_Q[\phi(h)] - \log E_P[\exp(\phi(h))] \leq KL(Q\|P) \ . \qquad (1)$$

Further,

$$\sup_{\phi} \left(E_Q[\phi(h)] - \log E_P[\exp(\phi(h))]\right) = KL(Q\|P) \ . \qquad (2)$$

Proof: For any measurable function $\phi(h)$, we have

$$\begin{aligned} E_Q[\phi(h)] &= E_Q\left[\log\left(\frac{dQ(h)}{dP(h)}\exp(\phi(h))\frac{dP(h)}{dQ(h)}\right)\right] \\ &= KL(Q\|P) + E_Q\left[\log\left(\exp(\phi(h))\frac{dP(h)}{dQ(h)}\right)\right] \\ &\overset{(a)}{\leq} KL(Q\|P) + \log E_Q\left[\exp(\phi)\frac{dP(h)}{dQ(h)}\right] \\ &= KL(Q\|P) + \log E_P[\exp(\phi(h))] \ , \end{aligned}$$

where (a) follows from Jensen's inequality. Rearranging terms give (1).

In order to prove (2), for a given P and Q, we simply give a function $\phi(h)$ that achieves the upper bound. In particular, let

$$\phi(h) = \log\left(\frac{dQ(h)}{dP(h)}\right) \ . \qquad (3)$$

With this choice of $\phi(h)$,

$$E_P[\exp(\phi(h))] = E_P\left[\left(\frac{dQ(h)}{dP(h)}\right)\right] = E_Q[1] = 1 \ .$$

Hence,

$$\begin{aligned} &E_Q[\phi(h)] - \log E_P[\exp(\phi(h))] \\ &\qquad = E_Q\left[\log\left(\frac{dQ(h)}{dP(h)}\right)\right] - \log 1 \\ &\qquad = KL(Q\|P) \ . \end{aligned}$$

That completes the proof. ■

The first part of the result has earlier explicitly appeared in the literature (McAllester, 2003b, Lemma 8). The second part essentially follows from the observation that only one application of Jensen's inequality is used to prove the first part. If one claims a stronger bound, then it can always be falsified by a proper choice of ϕ, P, Q. This argument simply carries over from the exact same property of Jensen's inequality—it cannot be tightened any further in the general case. The property of Jensen's inequality, in turn, follows from the fact that a (closed) convex function is the point-wise supremum of all affine functions majorized by the convex function.

One can view the compression lemma as a very special case of Fenchel's inequality (Appendix A). To see this, for any measurable function $\phi : \mathcal{H} \mapsto \mathbb{R}$, let

$$f(\phi) = \log E_{h\sim P}[\exp(\phi(h))] \ ,$$

for any fixed distribution P on $\mathcal{H}$. Now, for any ϕ_1, ϕ_2 and $\forall \lambda \in [0,1]$,

$$\begin{aligned} &\lambda f(\phi_1) + (1-\lambda) f(\phi_2) \\ &\quad = \log\left(E_P[\exp(\phi_1(h))]^{\lambda} E_P[\exp(\phi_2(h))]^{(1-\lambda)}\right) \\ &\quad \overset{(a)}{\geq} \log E_P[\exp(\lambda\phi_1(h) + (1-\lambda)\phi_2(h))] \\ &\quad = \ f(\lambda\phi_1 + (1-\lambda)\phi_2) \ , \end{aligned}$$

where (a) follows from Hölder's inequality. Thus, f is a convex function. We choose ϕ^* to be the density corresponding to a distribution Q on $\mathcal{H}$ so that

$$\langle \phi, \phi^* \rangle = E_{h\sim Q}[\phi(h)] \ .$$

Then, the conjugate of f is

$$\begin{aligned} f^*(\phi^*) &= \sup_{\phi} \left(\langle \phi, \phi^* \rangle - f(\phi)\right) \\ &= \sup_{\phi} \left(E_Q[\phi(h)] - \log E_P[\exp(\phi(h))]\right) \\ &= KL(Q\|P) \ , \end{aligned}$$

which follows from (2) in Lemma 1. Hence, from Fenchel's inequality, we have

$$\begin{aligned} \langle \phi, \phi^* \rangle - f(\phi) &\leq f^*(\phi^*) \\ \Rightarrow \quad E_Q[\phi(h)] - \log E_P[\exp(\phi(h))] &\leq KL(Q\|P) \ . \end{aligned}$$

We present yet another viewpoint of Lemma 1 from the compression perspective. Consider the Gibbs density

$$\frac{dG(h)}{dP(h)} = g(h) = \exp(\phi(h) - f(\phi)) \ .$$

The fact that $g(h)$ is a valid probability density with respect to P follows from the observation that $f(\phi)$ is simply the cumulant function, thereby ensuring the integral over P is 1. Recall that the expected description length $E_Q[-\log \pi(h)]$ of encoding Q using a density $\pi(h)$ is minimized when $\pi(h) = dQ(h)$, and the corresponding minimum description length is simply $H(Q)$, the Shannon entropy of Q (Cover & Thomas, 1991). Then, if one chooses $\pi(h) = dG(h)$, we have

$$E_Q[-\log dG(h)] \geq E_Q[-\log dQ(h)] = H(Q) \ .$$

Putting in the expression for $dG(h)$ and rearranging sides, we obtain

$$\begin{array}{rcl} E_Q[\phi(h) - f(\phi)] + E_Q[\log dP(h)] & \leq & E_Q[\log dQ(h)] \\ \Rightarrow \qquad E_Q[\phi(h)] - f(\phi) & \leq & KL(Q\|P) \ , \end{array}$$

which is exactly (1) in Lemma 1. This information theoretic interpretation justifies the name of the compression lemma.

Next, we show that the PAC-Bayesian bound in the batch setting, and the Bayesian log-loss and bounded-loss bounds in the online setting follow by a direct application of (1), along with some setting specific calculations. Although the exact semantics associated with ϕ, P, Q are different for each setting, ϕ always quantifies loss, P is the prior and Q is the reference distribution or posterior. The setting specific additional calculations are always focused on the cumulant $f(\phi) = \log E_P[\exp(\phi(h))]$, taking the form of concentration bounds in the batch setting, and convex bounds in the online setting. We revisit these similarities in Section 6.

3. PAC-Bayesian Bounds

In the PAC-Bayesian setting, the domain $\mathcal{H}$ is the set of possible classifiers h, and P, Q are the prior and posterior distributions on $\mathcal{H}$. We consider a batch classification task using the predictors from $\mathcal{H}$. Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ be the train set drawn independently according to a fixed (but unknown) distribution D. By abuse of notation, we denote the sample distribution of the train set by S as well. For any classifier h, let $\ell(h, W) \in [0, 1]$ denote the error-rate of h given the samples are drawn according to W. If the atomic loss function is simply the 0-1 classification error-rate, then $\ell(h, W) = E_{(\mathbf{x},y)\sim W}[h(\mathbf{x}) \neq y]$. Of course, more general loss functions are admissible (Bartlett et al., 2004). The Bayesian prediction scheme proceeds as follows: let P be the prior distribution over $\mathcal{H}$ that gets updated to Q after observing S. One is interested in quantifying the performance of the Bayesian classifier based on Q when samples are drawn following D. With $\ell(Q, W) = E_Q[\ell(h, W)]$, and for $p, q \in [0, 1], KL_{\text{B}}(p\|q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$, the PAC-Bayesian bound can be stated as follows:

Theorem 1 (PAC-Bayesian Bound) *With probability at least $(1-\delta)$ over the choice of $S \sim D^m$,*

$$KL_{\text{B}}(\ell(Q,S)\|\ell(Q,D))] \leq \frac{KL(Q\|P) + \log\frac{m+1}{\delta}}{m} \ . \tag{4}$$

The result, due to (McAllester, 2003a), originally appeared in a more explicit, albeit weaker, form. The implicit form, due to (Langford, 2005), is quantitatively tighter. The proof of the PAC-Bayesian theorem is well-known (McAllester, 2003b, Appendix), and was simplified by (Seeger, 2002). In the context of our current paper, we highlight the fact that the "simple" proof is just an application of Lemma 1.

Proof: From the compression lemma, we know that for any measurable function $\phi(h)$, we have

$$E_Q[\phi(h)] \leq KL(Q\|P) + \log E_P[\exp(\phi(h))] \ .$$

Let $\phi(h) = mKL_{\text{B}}(\ell(h,S)\|\ell(h,D))$ where S is the sample distribution and D is the true (unknown) distribution. Then,

$$\begin{aligned} &mE_Q[KL_{\text{B}}(\ell(h,S)\|\ell(h,D))] \\ &\leq KL(Q\|P) + \log E_P[\exp(mKL_{\text{B}}(\ell(h,S)\|\ell(h,D)))] \ . \end{aligned} \tag{5}$$

Since relative entropy is convex in both arguments (Cover & Thomas, 1991, Theorem 2.7.2), from Jensen's inequality, we have

$$KL_{\text{B}}(\ell(Q,S)\|\ell(Q,D)) \leq E_Q[KL_{\text{B}}(\ell(h,S)\|\ell(h,D))] \ . \tag{6}$$

Now, since $m\ell(h,S)$ is binomially distributed with probability $\pi = \ell(h,D)$, by definition we have

$$\begin{aligned} &E_{S\sim D^m}[\exp(mKL_{\text{B}}(\ell(h,S)\|\pi))] \\ &\quad = \sum_{s\sim \text{Bin}(\pi,m)} p(s)\exp(mKL_{\text{B}}(\ell(h,s)\|\pi)) \\ &\quad = \sum_{n=0}^{m} \binom{m}{n} \pi^n (1-\pi)^{m-n} \exp(mKL_{\text{B}}(n/m\|\pi)) \\ &\quad = \sum_{n=0}^{m} \binom{m}{n} \exp(-mH(n/m)) \overset{(a)}{\leq} \sum_{n=0}^{m} 1 = m+1 \ , \end{aligned}$$

where (a) follows from the fact that $\binom{m}{n} \leq \exp(-mH(n/m))$ (Cover & Thomas, 1991, Chapter 12). Taking expectations with respect to P and applying Fubini's theorem (Williams, 1991), we have

$$E_{S\sim D^m}[E_{h\sim P}[\exp(mKL_{\mathrm{B}}(\ell(h,S)\|\ell(h,D)))] \leq m+1 \ .$$

Then, from Markov's inequality, with probability at least $(1-\delta)$ over $S \sim D^m$, we have

$$E_P[\exp(mKL_{\mathrm{B}}(\ell(h,S)\|\ell(h,D)))] \leq \frac{m+1}{\delta} \ . \qquad (7)$$

Plugging (6) and (7) back into (5) completes the proof. ■

Application of the bound for deterministic classifiers involves a carefully constructed derandomization argument. For example, dimension independent margin bounds can be derived by choosing the prior and posterior to be shifted versions of an identity covariance Gaussian distribution (Langford & Shawe-Taylor, 2002; McAllester, 2003b; Langford, 2005).

4. Bayesian Log-Loss Bounds

The online-Bayes setting is seeming very different but, as before, the main relative loss bounds result from a direct application of Lemma 1. In the online setting, the domain $\mathcal{H}$ is a set of predictors h, known as "experts" when $\mathcal{H}$ is finite, and one assumes a prior distribution P_0 over $\mathcal{H}$. The predictors are assumed to be stochastic, so that for any input $\mathbf{x}$, each h generates a distribution $p(y|\mathbf{x},h)$ over the output. Prediction proceeds in iterations, where in iteration t, an input $\mathbf{x}_t$ is presented, and each of the predictors generates a predicted distribution. If y_t is the true label, then each predictor h incurs a loss of

$$\ell_t(h) = -\log p(y_t|\mathbf{x}_t,h) \ .$$

Moreover, depending on the loss incurred by individual predictors, the distribution over $\mathcal{H}$ is updated. If P_{t-1} is the distribution on $\mathcal{H}$ after seeing $S_{t-1} = \{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_{t-1},y_{t-1})\}$, the combined prediction from the ensemble on $\mathbf{x}_t$ is simply $E_{h\sim P_{t-1}}[p(y|\mathbf{x}_t,h)]$. There are two important questions that are of interest: first, how to update the distribution P_{t-1} on $\mathcal{H}$ depending on the performance of the predictors; second, how to get a bound on the performance of the ensemble.

The "obvious" update of P_{t-1} involves an application of Bayes rule. In particular,

$$\begin{aligned} P_t \equiv p(h|S_t) &= p(h|S_{t-1},\mathbf{x}_t,y_t) \\ &= \frac{p(y_t|\mathbf{x}_t,h,S_{t-1})p(h|\mathbf{x}_t,S_{t-1})}{p(S_t)} \\ &\overset{(a)}{=} \frac{p(y_t|\mathbf{x}_t,h)p(h|S_{t-1})}{p(S_t)} = \frac{p(y_t|\mathbf{x}_t,h)P_{t-1}}{p(S_t)} \ , \qquad (8) \end{aligned}$$

where (a) follows since y_t is independent of S_{t-1} given $\mathbf{x}_t,h$, i.e., individual predictors or "experts" are memory-less, and since $p(h|\mathbf{x}_t,S_{t-1}) = p(h|S_{t-1})$ as we update the distribution over $\mathcal{H}$ only after getting the label on the current $\mathbf{x}_t$. Hence, (8) is a simple application of Bayes rule. However, note that

$$P_t(h) = \frac{\exp(-\ell_t(h))P_{t-1}(h)}{Z_t} \ , \qquad (9)$$

where $Z_t = p(S_t) = E_{P_{t-1}}[\exp(-\ell_t(h))]$ is the normalization term. Note that this is the strategy used by exponentiated gradient methods (Kivinen & Warmuth, 1997) with well studied properties.

We now focus on the regret bound of the adaptive ensemble predictor compared to a predictor that uses a fixed distribution Q over $\mathcal{H}$. While such bounds have been well studied in the literature (Freund et al., 1997), we simply show that similar to the PAC-Bayesian case, a direct application of Lemma 1 gives such bounds in a straightforward way.

At iteration t, on receiving input $\mathbf{x}_t$, the Bayesian model predicts

$$p(y|\mathbf{x}_t,S_{t-1}) = E_{h\sim P_{t-1}}[p(y|\mathbf{x}_t,h)] \ .$$

If y_t is the true label, the model incurs a log-loss of $-\log p(y_t|\mathbf{x}_t,S_{t-1})$. Hence, after such iterative prediction on S_T, the total log-loss incurred by the Bayesian model is

$$L_{BLL}(S_T) = \sum_{t=1}^{T} -\log p(y_t|\mathbf{x}_t,S_{t-1}) \ .$$

Let Q be any fixed distribution over the predictors $h \in \mathcal{H}$. Consider a prediction scheme that samples $h \sim Q$, and then predicts on $\mathbf{x}_t$ based on h. The expected loss incurred by Q at iteration t is simply

$$\ell_t(Q) = E_{h\sim Q}[\ell_t(h)] = E_{h\sim Q}[-\log p(y_t|\mathbf{x}_t,h)] \ ,$$

and the total loss

$$L_Q(S_T) = \sum_{t=1}^{T}\ell_t(Q) = E_Q\left[\sum_{t=1}^{T}\ell_t(h)\right] \ .$$

Then, with $S = S_T$, we have the following result.

Theorem 2 (Log Loss Bound) *In the Bayesian log-loss setting, for any distribution Q on $\mathcal{H}$, we have*

$$L_{BLL}(S) \leq L_Q(S) + KL(Q \| P_0) \ .$$

The result has appeared in different forms in the literature (Freund et al., 1997; Kakade & Ng, 2004). We give a proof using Lemma 1.

Proof: From the compression lemma, we know for any measurable function $\phi(h)$,

$$E_Q[\phi(h)] - \log E_{P_0}[\exp(\phi(h))] \leq KL(Q \| P_0) \ .$$

To complete the proof, we simple choose $\phi(h)$ such that $E_Q[\phi(h)] = -L_Q(S)$ and $-\log E_{P_0}[\exp(\phi(h))] = L_{BLL}(S)$. Note that since

$$-L_Q(S) = E_Q\left[-\sum_{t=1}^{T} \ell_t(h)\right] = E_Q[\phi(h)] \ ,$$

we have $\phi(h) = -\sum_{t=1}^{T} \ell_t(h)$. By definition,

$$\begin{aligned} L_{BLL}(S) &= \sum_{t=1}^{T} -\log p(y_t|\mathbf{x}_t, S_{t-1}) \\ &= -\log \prod_{t=1}^{T} p(y_t|\mathbf{x}_t, S_{t-1}) \ . \end{aligned}$$

Note that $p(y_t|\mathbf{x}_t, S_{t-1}) = E_{P_{t-1}}[\exp(-\ell_t(h))] = Z_t$, so that $L_{BLL}(S) = -\log \prod_{t=1}^{T} Z_t$. Now, repeatedly using (9), we have

$$\begin{aligned} Z_T &= \int_{\mathcal{H}} \exp(-\ell_T(h)) p_{T-1}(h) dh \\ &= \frac{1}{Z_{T-1} \cdots Z_1} \int_{\mathcal{H}} \exp(-\sum_{t=1}^{T} \ell_t(h)) p_0(h) \ . \end{aligned}$$

Hence, we have

$$\prod_{t=1}^{T} Z_t \ = \ E_{P_0}[\exp(-\sum_{t=1}^{T} \ell_t(h))] \ = \ E_{P_0}[\exp(\phi(h))] \ .$$

That completes the proof. ■

5. Bayesian Bounded-Loss Bounds

The online bounded-loss bounds are similar in essence to the online log-loss bounds, but works with arbitrary bounded losses. Further, the updates can be more general than a direct application of Bayes rule as in Section 4. As before, let $\mathcal{H}$ be set of predictors h, better known as "experts" when $\mathcal{H}$ is finite. Prediction proceeds in iterations, where at iteration t every h predicts the outcome of an event and receives loss $\ell_t(h) \in [0,1]$. Hence, the loss received by experts in $\mathcal{H}$ at iteration t is a function $\ell_t : \mathcal{H} \mapsto [0,1]$. Let $S_t = \{\ell_1, \ldots, \ell_t\}$. In a Bayesian setting, one starts with a prior distribution P_0 over $\mathcal{H}$. At any iteration t, since the individual predictors receive a loss of $\ell_t(h)$, the Bayesian predictor receives the expected loss

$$L_t = E_{P_{t-1}}[\ell_t(h)] \in [0,1] \ .$$

Further, the distribution P_{t-1} over $\mathcal{H}$ is updated to P_t such that

$$\begin{aligned} P_t \ \equiv \ p(h|S_t) \ &= \ \frac{\beta^{\ell_t(h)} p_{t-1}(h)}{Z_t(\beta)} \\ &= \ \frac{\exp(-k_\beta \ \ell_t(h)) P_{t-1}}{Z_t(\beta)} \ , \end{aligned}$$

where $\beta \in (0,1), k_\beta = \log(1/\beta) > 0$, and

$$Z_t(\beta) = E_{P_{t-1}}[\beta^{\ell_t(h)}] = E_{P_{t-1}}[\exp(-k_\beta \ \ell_t(h))]$$

is the normalization. Note that the update is similar to that in (9), other than the scaling factor k_β. Since there is no explicit notion of input or output, the performance of the Bayesian predictor is measured by the cumulative loss over T iterations,

$$L_{BLB}(S_T) = \sum_{t=1}^{T} L_t = \sum_{t=1}^{T} E_{P_{t-1}}[\ell_t(h)] \ .$$

As before, we focus on the performance of the Bayesian predictor relative to a predictor based on any fixed distribution Q on $\mathcal{H}$. The cumulative loss incurred by such a predictor is

$$L_Q(S_T) = \sum_{t=1}^{T} E_Q[\ell_t(h)] = E_Q\left[\sum_{t=1}^{T} \ell_t(h)\right] \ .$$

Then, we have the following result.

Theorem 3 (Bounded Loss Bound) *For any sequence of loss functions S and any $\beta \in (0,1)$*

$$(1-\beta) L_{BLB}(S) \leq k_\beta L_Q(S) + KL(Q \| P_0) \ , \qquad (10)$$

where $k_\beta = \log(1/\beta)$.

The bounded-loss bounds have been primarily studied in the case when the number of predictors is finite (Littlestone & Warmuth, 1994; Freund & Schapire, 1997; Freund & Schapire, 1999a; Cesa-Bianchi et al., 1997). In such cases, each predictor is called an expert. The above result is a simple extension to a general Bayesian setting. More interestingly, the proof is a simple application of the compression lemma, exactly like the PAC-Bayesian and log-loss bounds discussed earlier. For convenience, we denote $L(h) = \sum_{t=1}^{T} \ell_t(h)$.

Proof: From the compression lemma, for any measurable function $\phi(h)$, we have

$$E_Q[\phi(h)] - \log E_{P_0}[\exp(\phi(h))] \leq KL(Q\|P_0) \ .$$

As before, we choose $\phi(h)$ such that $E_Q[\log \phi(h)] = -k_\beta L_Q(S)$. Note that since

$$-k_\beta L_Q(S) \ = \ E_Q[-k_\beta \sum_{t=1}^{T} \ell_t(h)] \ = \ E_Q[\phi(h)] \ ,$$

we have $\phi(h) = -k_\beta \sum_{t=1}^{T} \ell_t(h) = -k_\beta L(h)$. Now,

$$\begin{aligned} Z_T(\beta) &= \int_h \exp(-k_\beta \ \ell_T(h)) \ p_{T-1}(h) dh \\ &= \frac{1}{Z_{T-1}(\beta) \cdots Z_1(\beta)} \int_h \exp(-k_\beta \sum_{t=1}^{T} \ell_t(h)) \ p_0(h) \ , \end{aligned}$$

which implies

$$\prod_{t=1}^{T} Z_t(\beta) = E_{P_0}[\exp(-k_\beta L(h))] = E_{P_0}[\exp(\phi(h))] \ .$$

Since $\exp(-k_\beta L(h)) = \beta^{L(h)}$, from the compression lemma we have

$$-\log E_{P_0}[\beta^{L(h)}] \leq k_\beta \ L_Q(S) + KL(Q\|P_0) \ .$$

We complete the proof by showing that $\log E_{P_0}[\beta^{L(h)}] = \sum_{t=1}^{T} \log Z_t \leq -(1-\beta) L_{BLB}(S)$. Treating $q = \ell_t(h) \in [0,1]$ as a probability, from Jensen's inequality, we have $\beta^q \leq (1-q)\beta^0 + q\beta^1 = 1 - (1-\beta)q$. Hence,

$$\begin{aligned} \log Z_t &= \log E_{P_{t-1}}[\beta^{\ell_t(h)}] \\ &\leq \log(1 - (1-\beta) E_{P_{t-1}}[\ell_t(h)]) \\ &\leq -(1-\beta) E_{P_{t-1}}[\ell_t(h)] \ . \end{aligned}$$

Summing over $t = 1, \ldots, T$, we have $\sum_{t=1}^{T} \log Z_t \leq -(1-\beta) L_{BLB}(S)$. That completes the proof. ■

6. Discussion

In this section, we revisit some aspects of the bounds in Sections 3, 4 and 5, discuss properties and highlight connections with other existing work.

6.1. A Simplified Understanding

We have tried to develop a simplified understanding of the batch and online Bayesian bounds that are applicable to significantly disparate settings. Somewhat surprisingly, and with elementary arguments, we have shown that all these bounds follow from a simple compression lemma. In particular, there is only one term, the cumulant $f(\phi) = \log E_{P_0}[\exp(\phi(h))]$, that needed separate treatment depending on the setting. In the PAC-Bayesian setting,

$$\phi(h) = mKL_{\mathrm{B}}(\ell(h,S) \| \ell(h,D)) \ ,$$

where $\ell(h,W)$ is the error-rate of h when samples are drawn following W, measures the discrepancy between the error rate on the sample distribution S and the true (unknown) distribution D. In the log-loss and bounded-loss settings, we respectively have

$$\phi(h) = -\sum_{t=1}^{T} \ell_t(h) \quad \text{and} \quad \phi(h) = -k_\beta \sum_{t=1}^{T} \ell_t(h) \ ,$$

where $\ell_t(h)$ is the log-loss of a probabilistic prediction in the first case, and any bounded-loss in the second case. In case of log-loss bounds, no setting specific calculations are necessary and the main result (Theorem 2) follows directly from the compression lemma. In PAC-Bayesian setting, since the sample S is random, the term $f(\phi) = \log E_{P_0}[\exp(\phi(h))]$ is random, and one uses concentration bounds (McAllester, 2003a) to get the main result (Theorem 1). In the bounded-loss setting, elementary convexity arguments on the $f(\phi) = \log E_{P_0}[\exp(\phi(h))]$ term are required to get the desired bound (Theorem 3).

6.2. The Online Setting

The similarities between the two online settings are quite apparent. Consider the log-loss setting. If $\exists \epsilon > 0$ such that $p(y|\mathbf{x}, h) \geq \epsilon, \forall \mathbf{x}, y, h$, then the loss incurred by the Bayesian model at any stage is bounded above by $\log(1/\epsilon)$. With proper normalization, this reduces to the bounded loss setting. In general, the prediction $p(y_t|\mathbf{x}_t, S_{t-1})$ from the Bayesian model can be arbitrarily close to 0. In such a case, the log-loss bound inequality is trivialized, i.e., although the inequality still holds, both sides become arbitrarily large and are not of practical interest (Kakade & Ng, 2004; Kakade et al., 2005).

6.3. Tightness of Bounds

The compression lemma is a special case of Fenchel's inequality, which gives the best bounds of the form $f(\mathbf{x}) + g(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle$. In particular, being an instance of Fenchel's inequality, Lemma 1 and any of its direct applications (such as Theorem 2) cannot be improved in the general case. Note that such tightness properties of some of the bounds discussed here have been rigorously established in the literature, e.g., see (Freund & Schapire, 1997, Appendix) and (Vovk, 1995). In fact, Vovk used conjugacy and Fenchel's inequality in order to establish large deviation bounds (Vovk, 1995, Section 5). All such analysis have been primarily targeted to the online setting with bounded loss and a constant number of experts. In fact, it is interesting to note that the PAC-Bayesian bound was independently discovered (McAllester, 2003a). Since we show that all these bounds are consequences of the compression lemma, it seems that they cannot be improved in the general case, other than possibly some better arguments for the setting specific calculations.

6.4. Maximum Entropy Learning

Like any other application of Fenchel's inequality, Lemma 1 is a mathematical result, irrespective of any semantics. In other words, it is not the property of any particular algorithm. Of course, several (conservative) algorithms have been designed to get "best" worst-case performance as dictated by the inequality. From the results, we note that the best worst-case performance implies maximizing entropy. Several learning algorithms—such as the entropy projection viewpoint of boosting (Kivinen & Warmuth, 1999) as well as direct maximum entropy discrimination in the batch (Jaakkola et al., 1998) and online settings (Long & Wu, 2004)—follow the maximum entropy "principle." Since entropy is the minimum description length, maximizing entropy leads to a maximin problem. However, as our results indicate, such updates try to minimize the worst case (relative) loss, i.e., it solves a minimax problem. Such a duality connection has been well studied in the literature (Haussler, 1997; Topsoe, 1979) and has been significantly generalized (Grünwald & Dawid, 2004) in recent years.

7. Conclusion

In this paper, we show that several of the popular and widely used Bayesian bounds in the batch as well as online settings are consequences of a simple compression lemma. While there are no "new" results, we hope that the analysis will result in a simplified and accessible understanding of the existing results. We note that the original arguments of all the results discussed here were ingeniously developed without making use of the compression lemma explicitly. In fact, the development of the PAC-Bayesian theorem, which historically followed the relevant developments in online learning, never made use of that literature and was independently established. We hope that our analysis will bridge any seeming conceptual gaps in the understanding of some of the most important bounds in the batch and online settings.

Acknowledgements: We want to thank the reviewers for their valuable comments that significantly improved the presentation.

References

Bartlett, P., Collins, M., Taskar, B., & McAllester, D. (2004). Exponentiated gradient algorithms for large-margin structured classification. *Proceedings of the 18th Annual Conference on Neural Information Processing Systems.*

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., & Warmuth, M. K. (1997). How to use expert advice. *Journal of the ACM*, *44*, 427–485.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley-Interscience.

Freund, Y., & Schapire, R. (1999a). Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, *29*, 79–103.

Freund, Y., & Schapire, R. (1999b). Large margin classification using the perceptron algorithm. *Machine Learning Journal*, *37*, 277–296.

Freund, Y., Schapire, R., Singer, Y., & Warmuth, M. (1997). Using and combining predictors that specialize. *Proceedings of the 29th Annual ACM Symposium on the Theory of Computing* (pp. 334–343).

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Grünwald, P. D., & Dawid, A. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, *32*.

Haussler, D. (1997). A general minimax result for relative entropy. *IEEE Transactions of Information Theory*, *43*, 1276–1280.

Helmbold, D., & Warmuth, M. (1995). On weak learning. *Journal of Computer and System Sciences*, *50*, 551–573.

Jaakkola, T., Meila, M., & Jebara, T. (1998). Maximum entropy discrimination. *Proceedings of the 12th*

Annual Conference on Neural Information Processing Systems.

Kakade, S. M., & Ng, A. (2004). Online bounds for Bayesian algorithms. *Proceedings of the 18th Annual Conference on Neural Information Processing Systems.*

Kakade, S. M., Seeger, M., & Foster, D. (2005). Worst-case bounds for Gaussian process models. *Proceedings of the 19th Annual Conference on Neural Information Processing Systems.*

Kivinen, J., & Warmuth, M. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, *132*, 1–64.

Kivinen, J., & Warmuth, M. K. (1999). Boosting as entropy projection. *Proceedings of the 12th Annual Conference on Learning Theory* (pp. 134–144).

Langford, J. (2005). Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, *6*, 273–306.

Langford, J., & Shawe-Taylor, J. (2002). PAC-Bayes and margins. *Proceedings of the 16th Annual Conference on Neural Information Processing Systems.*

Littlestone, N., & Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*, *108*, 212–261.

Long, P., & Wu, X. (2004). Mistake bounds for maximum entropy discrimination. *Proceedings of the 18th Annual Conference on Neural Information Processing Systems.*

McAllester, D. (2003a). PAC-Bayesian model averaging. *Machine Learning Journal*, *5*, 5–21.

McAllester, D. (2003b). Simplified PAC-Bayesian margin bounds. *Proceedings of the 16th Annual Conference on Learning Theory* (pp. 203–215).

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press.

Seeger, M. (2002). PAC-Bayesian generalization bounds for Gaussian processes. *Journal of Machine Learning Research*, *3*, 233–269.

Topsoe, F. (1979). Information theoretical optimization techniques. *Kybernetika*, *15*, 8–27.

Vovk, V. G. (1995). A game of prediction with expert advice. *Proceedings of the 8th Annual Conference on Computational Learning Theory* (pp. 51–60). ACM Press, New York, NY.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.

A. Fenchel's Inequality

A fundamental duality appears in convex analysis from the fact that any closed convex set can be equivalently specified as an intersection of half-spaces that contain the set (Rockafellar, 1970, Theorem 11.5). Extending the result to functions, any closed convex function f is the point-wise supremum of the collection of all affine functions h majorized by f, i.e., $h \leq f$ (Rockafellar, 1970, Theorem 12.1). This result leads to the concept of conjugacy and a powerful class of "best" inequalities collectively known as Fenchel's inequality. We briefly review the key concepts and results. Our presentation follows (Rockafellar, 1970, Section 12).

Let f be a closed convex function on $\mathbb{R}^d$. Let F^* be the set of all pairs $(\mathbf{x}^*, v^*) \in \mathbb{R}^{d+1}$ such that the affine function $h(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}^* \rangle - v^*$ is majorized by f. Now, we have $f(\mathbf{x}) \geq h(\mathbf{x})$, $\forall \mathbf{x}$, if and only if $v^* \geq \sup_{\mathbf{x} \in \mathbb{R}^d} (\langle \mathbf{x}, \mathbf{x}^* \rangle - f(\mathbf{x}))$. Hence, F^* is the epigraph of the function on $\mathbb{R}^n$ given by

$$f^*(\mathbf{x}^*) = \sup_{\mathbf{x} \in \mathbb{R}^n} (\langle \mathbf{x}, \mathbf{x}^* \rangle - f(\mathbf{x})) \ . \tag{11}$$

f^* is called the conjugate of f, and is a closed convex function itself as it is the point-wise supremum of affine functions $h^*(\mathbf{x}^*) = \langle \mathbf{x}, \mathbf{x}^* \rangle - v$, where $(\mathbf{x}, v)$ belongs to F, the epigraph of f. Further, the conjugate f^{**} of f^* is f.

The theory of conjugacy is regarded as the theory of the "best"inequalities of the type

$$f(\mathbf{x}) + g(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle, \quad \forall \mathbf{x}, \mathbf{y}.$$

Let W denote all the pairs (f, g) for which the inequality holds. The "best" pairs (f, g) in W are those for which the inequality cannot be tightened, i.e., if $(f', g') \in W$, $f \geq f'$, $g \geq g'$, then $f = f'$ and $g = g'$. For any (f, g) in W, we have

$$\begin{aligned} g(\mathbf{y}) &\geq \sup_{\mathbf{x}} (\langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x})) = f^*(\mathbf{y}), \ \forall \mathbf{y} \\ f(\mathbf{x}) &\geq \sup_{\mathbf{y}} (\langle \mathbf{x}, \mathbf{y} \rangle - g(\mathbf{y})) = g^*(\mathbf{x}), \ \forall \mathbf{x} \ . \end{aligned}$$

Hence, the "best" pairs in W are precisely those such that $g = f^*$ and $f = g^*$. In particular, we have

$$f(\mathbf{x}) + f^*(\mathbf{x}^*) \geq \langle \mathbf{x}, \mathbf{x}^* \rangle, \quad \forall \mathbf{x}, \mathbf{x}^* \ , \tag{12}$$

holds for any proper convex function f and its conjugate f^*. This relationship is known as *Fenchel's inequality.*