PAC-Bayes Risk Bounds for Sample-Compressed Gibbs Classifiers

François Laviolette FRANCOIS.LAVIOLETTE@IFT.ULAVAL.CA
Mario Marchand MARIO.MARCHAND@IFT.ULAVAL.CA
Département d'informatique et de génie logiciel, Université Laval, Québec, Canada, G1K-7P4

Abstract

We extend the PAC-Bayes theorem to the sample-compression setting where each classifier is represented by two independent sources of information: a *compression set* which consists of a small subset of the training data, and a *message string* of the additional information needed to obtain a classifier. The new bound is obtained by using a prior over a data-independent set of objects where each object gives a classifier only when the training data is provided. The new PAC-Bayes theorem states that a Gibbs classifier defined on a posterior over sample-compressed classifiers can have a smaller risk bound than any such (deterministic) sample-compressed classifier.

1. Introduction

The PAC-Bayes approach, initiated by McAllester (1999), aims at providing PAC guarantees to "Bayesian-like" learning algorithms. These algorithms are specified in terms of a *prior distribution* P over a space of classifiers that characterizes our prior belief about good classifiers (before the observation of the data) and a *posterior distribution* Q (over the same space of classifiers) that takes into account the additional information provided by the training data. A remarkable result that came out from this line of research, known as the "PAC-Bayes theorem", provides a tight upper bound on the risk of a stochastic classifier (defined on the posterior Q) called the *Gibbs classifier*.

This PAC-Bayes bound (see Theorem 1) depends both on the empirical risk (*i.e.*, training errors) of the Gibbs classifier and on "how far" is the data-dependent posterior Q from the data-independent prior P. Consequently, a Gibbs classifier with a posterior Q having all its weight on a single classifier will have a larger risk bound than another Gibbs classifier, making the same amount of training errors, using a "broader" posterior Q that gives weight to many classifiers. Hence, the PAC-Bayes theorem quantifies the additional predictive power that stochastic classifier selection has over deterministic classifier selection.

A constraint currently imposed by the PAC-Bayes theorem is that the prior P must be defined without reference to the training data. Consequently, we cannot directly use the PAC-Bayes theorem to bound the risk of sample-compression learning algorithms (Littlestone & Warmuth, 1986; Floyd & Warmuth, 1995) because the set of classifiers considered by these algorithms are those that can be reconstructed from various subsets of the training data. However, this is an important class of learning algorithms since many well known learning algorithms, such as the support vector machine (SVM) and the perceptron learning rule, can be considered as sample-compression learning algorithms. Moreover, some sample-compression algorithms (Marchand & Shawe-Taylor, 2002) have achieved very good performance in practice by deterministically choosing the sparsest possible classifier: the one described by the smallest possible subset of the training set. It is therefore worthwhile to investigate if the stochastic selection of sample-compressed classifiers provides an additional predictive power over the deterministic selection of a single sample-compressed classifier.

In this paper, we extend the PAC-Bayes theorem in such a way that it applies now to both the usual data-independent setting and the more general sample-compression setting. In the sample-compression setting, each classifier is represented by two independent sources of information: a *compression set* which consists of a small subset of the training data, and a *message string* of the additional information needed to obtain a classifier. In the limit where the compression set vanishes, each classifier is identified only by a message string and the new PAC-Bayes theorem reduces

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

to the "usual" PAC-Bayes theorem of Seeger (2002) and Langford (2005). However, new quantities appear in the risk bound when classifiers are also described by their compression sets. The new PAC-Bayes theorem states that a stochastic Gibbs classifier defined on a posterior over several sample-compressed classifiers can have a smaller risk bound than any such single (deterministic) sample-compressed classifier. Finally, the new PAC-Bayes risk bound reduces to the usual sample-compression risk bounds (Littlestone & Warmuth, 1986; Floyd & Warmuth, 1995; Langford, 2005) in the limit where the posterior Q puts all its weight on a single sample-compressed classifier.

2. Basic Definitions

We consider binary classification problems where the input space $\mathcal{X}$ consists of an arbitrary subset of $\mathbb{R}^n$ and the output space $\mathcal{Y} = \{-1, +1\}$. An example $\mathbf{z} \stackrel{\text{def}}{=} (\mathbf{x}, y)$ is an input-output pair where $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Throughout the paper, we adopt the PAC setting where each example $\mathbf{z}$ is drawn according to a fixed, but unknown, probability distribution D on $\mathcal{X} \times \mathcal{Y}$. The risk $R(f)$ of any classifier f is defined as the probability that it misclassifies an example drawn according to D:

$$R(f) \stackrel{\text{def}}{=} \Pr_{(\mathbf{x},y)\sim D}\Big(f(\mathbf{x}) \neq y\Big) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(f(\mathbf{x}) \neq y)$$

where $I(a) = 1$ if predicate a is true and 0 otherwise. Given a training set $S = \langle \mathbf{z}_1, \ldots, \mathbf{z}_m \rangle$ of m examples, the *empirical risk* $R_S(f)$ on S, of any classifier f, is defined according to:

$$R_S(f) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} I(f(\mathbf{x}_i) \neq y_i) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim S} I(f(\mathbf{x}) \neq y)$$

3. The PAC-Bayes Theorem in the Data-Independent Setting

The PAC-Bayes theorem provides a tight upper bound on the risk of a stochastic classifier called the *Gibbs classifier*. Given an input example $\mathbf{x}$, the label $G_Q(\mathbf{x})$ assigned to $\mathbf{x}$ by the Gibbs classifier is defined by the following process. We first choose a classifier h according to the posterior distribution Q and then use h to assign the label to $\mathbf{x}$. The risk of G_Q is defined as the expected risk of classifiers drawn according to Q:

$$R(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} R(h) = \mathop{\mathbf{E}}_{h\sim Q} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(f(\mathbf{x}) \neq y)$$

The PAC-Bayes theorem was first proposed by McAllester (2003a). The version presented here is due to Seeger (2002) and Langford (2005).

Theorem 1 *Given any space $\mathcal{H}$ of classifiers. For any data-independent prior distribution P over $\mathcal{H}$:*

$$\Pr_{S\sim D^m}\left(\forall Q:\ \mathrm{kl}(R_S(G_Q)\|R(G_Q)) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\tfrac{m+1}{\delta}\right]\right) \geq 1-\delta$$

where $\mathrm{KL}(Q\|P)$ is the Kullback-Leibler divergence between distributions Q and P:

$$\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln\frac{Q(h)}{P(h)}$$

and where $\mathrm{kl}(q\|p)$ is the Kullback-Leibler divergence between the Bernoulli distributions with probability of success q and probability of success p:

$$\mathrm{kl}(q\|p) \stackrel{\text{def}}{=} q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$$

The bound given by the PAC-Bayes theorem for the risk of Gibbs classifiers can be turned into a bound for the risk of Bayes classifiers in the following way. Given a posterior distribution Q, the Bayes classifier B_Q performs a majority vote (under measure Q) of binary classifiers in $\mathcal{H}$. Then B_Q misclassifies an example $\mathbf{x}$ iff at least half of the binary classifiers (under measure Q) misclassifies $\mathbf{x}$. It follows that the error rate of G_Q is at least half of the error rate of B_Q. Hence $R(B_Q) \leq 2R(G_Q)$.

Finally, for certain distributions Q, a bound for $R(B_Q)$ can be turned into a bound for the risk of a single classifier whenever there exists $h^*_Q \in \mathcal{H}$ such that $h^*_Q(\mathbf{x}) = B_Q(\mathbf{x})\ \forall \mathbf{x} \in \mathcal{X}$. Such a classifier h^*_Q is said to be Bayes-equivalent under Q since it performs the same classification as B_Q. For example, a linear classifier with weight vector $\mathbf{w}$ is Bayes-equivalent under any distribution Q which is rotationally invariant around $\mathbf{w}$. By choosing a Gaussian (or a rectified Gaussian tail) centered on $\mathbf{w}$ for Q and Gaussian centered at the origin for P, Langford (2005), Langford and Shawe-Taylor (2003), and McAllester (2003b) have been able to derived tight risk bounds for the SVM from the PAC-Bayes theorem in terms of the "margin errors" achieved on the training data. These are the tightest risk bounds currently known for SVMs.

4. A PAC-Bayes Theorem for the Sample-Compression Setting

An algorithm A is said to be a *sample-compression learning algorithm* if, given a training set S =

$\langle \mathbf{z}_1, \ldots, \mathbf{z}_m \rangle$ of m examples, the classifier $A(S)$ returned by algorithm A is described entirely by two *complementary sources of information*: a subset $S_{\mathbf{i}}$ of S, called the *compression set*, and a *message string* σ which represents the additional information needed to obtain a classifier from the compression set.

Given a training set S (considered as an m-tuple), the compression set $S_{\mathbf{i}} \subseteq S$ is defined by a vector $\mathbf{i}$ of indices:

$$\begin{array}{rcl} \mathbf{i} & \stackrel{\text{def}}{=} & (i_1, i_2, \ldots, i_{|\mathbf{i}|}) \\ \text{with} & : & i_j \in \{1, \ldots, m\} \ \forall j \\ \text{and} & : & i_1 < i_2 < \ldots < i_{|\mathbf{i}|} \end{array}$$

where $|\mathbf{i}|$ denotes the number of indices present in $\mathbf{i}$. Hence, $S_{\mathbf{i}}$ denotes the $|\mathbf{i}|$-tuple of examples of S that are pointed by the vector of indices $\mathbf{i}$ defined above. We will use $\bar{\mathbf{i}}$ to denote the vector of indices not present in $\mathbf{i}$. Hence, abusing slightly of the notation, we write $S = S_{\mathbf{i}} \cup S_{\bar{\mathbf{i}}}$ for any vector $\mathbf{i} \in \mathcal{I}$ where $\mathcal{I}$ denotes the set of the 2^m possible realizations of $\mathbf{i}$.

The fact that any classifier returned by algorithm A is described by a compression set and a message string implies that there exists a *reconstruction function* $\mathcal{R}$, associated to A, that outputs a classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$ when given an arbitrary compression set $S_{\mathbf{i}}$ and a message string σ chosen from the set $\mathcal{M}(\mathbf{i})$ of all distinct messages that can be supplied to $\mathcal{R}$ with the index vector $\mathbf{i}$. In other words, $\mathcal{R}$ is of the form:

$$\mathcal{R} : \quad (\mathcal{X} \times \mathcal{Y})^{|\mathbf{i}|} \times \mathcal{K} \quad \longrightarrow \quad \mathcal{H}$$

where $\mathcal{H}$ is a set of classifiers and where $\mathcal{K}$ is some subset of $\mathcal{I} \times \mathcal{M}$ for $\mathcal{M} = \cup_{\mathbf{i} \in \mathcal{I}} \mathcal{M}(\mathbf{i})$. Hence:

$$\mathcal{M}(\mathbf{i}) = \{\sigma \in \mathcal{M} \mid (\mathbf{i}, \sigma) \in \mathcal{K}\}.$$

The perceptron learning rule and the SVM are examples of learning algorithms where the final classifier can be reconstructed solely from a compression set (Graepel et al., 2000; Graepel et al., 2001). In contrast, the reconstruction function for the set covering machine (Marchand & Shawe-Taylor, 2002) needs both a compression set and a message string.

It is important to realize that the sample-compression setting is strictly more general than the usual data-independent setting where the space $\mathcal{H}$ of possible classifiers (returned by learning algorithms) is defined without reference to the training data. Indeed, we recover this usual setting when each classifier is identified only by a message string σ. In that case, for each $\sigma \in \mathcal{M}$, we have a classifier $\mathcal{R}(\sigma)$. Hence, in this limit, we have a data-independent set $\mathcal{H}$ of classifiers given by $\mathcal{R}$ and $\mathcal{M}$ such that: $\mathcal{H} = \{\mathcal{R}(\sigma) \mid \sigma \in \mathcal{M}\}$.

However, the validity of theorem 1 has been established only in the usual data-independent setting where the priors are defined without reference to the training data S. We now derive a new PAC-Bayes theorem for priors that are more natural for sample-compression algorithms. These are priors defined over $\mathcal{K}$, the set of all the parameters needed by the reconstruction function $\mathcal{R}$, once a training set S is given. The prior will therefore be written as:

$$P(\mathbf{i}, \sigma) = P_{\mathcal{I}}(\mathbf{i}) P_{\mathcal{M}}(\sigma|\mathbf{i}) \tag{1}$$

Definition 2 *We will denote by $\mathcal{P}_{\mathcal{K}}$ the set of all distributions P on $\mathcal{K}$ that satisfy Equation (1).*

For $P_{\mathcal{M}}(\sigma|\mathbf{i})$, we could choose a uniform distribution over $\mathcal{M}(\mathbf{i})$. However it should generally be better to choose $1/2^{|\sigma|}$ for some prefix-free code (Cover & Thomas, 1991). More generally, the message string could be a *parameter* chosen from a continuous set $\mathcal{M}$. In this case, $P_{\mathcal{M}}(\sigma|\mathbf{i})$ specifies a probability density function.

For $P_{\mathcal{I}}(\mathbf{i})$, there is no a priori information that can help us to differentiate two index $\mathbf{i}, \mathbf{i}' \in \mathcal{I}$ that have same size. Hence we should choose:

$$P_{\mathcal{I}}(\mathbf{i}) = \frac{p(|\mathbf{i}|)}{\binom{m}{|\mathbf{i}|}}$$

where p is any function satisfying $\sum_{d=0}^{m} p(d) = 1$.

To shorten the notation, we will denote the true risk $R\left(\mathcal{R}(\sigma, S_{\mathbf{i}})\right)$ of classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$ simply by $R(\sigma, S_{\mathbf{i}})$. Similarly, we will denote the empirical risk $R_{S_{\bar{\mathbf{i}}}}(\mathcal{R}(\sigma, S_{\mathbf{i}}))$ of classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$ simply by $R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})$. Recall that $S_{\bar{\mathbf{i}}}$ is the set of training examples which are *not* in the compression set $S_{\mathbf{i}}$. Indeed, in the following, it will become obvious that the bound on the risk of classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$ depends only on its empirical risk on $S_{\bar{\mathbf{i}}}$.

Our main theorem is a PAC-Bayes risk bound for *sample-compressed Gibbs classifiers*. Therefore, we consider learning algorithms that output a posterior distribution $Q \in \mathcal{P}_{\mathcal{K}}$ after observing some training set S. Hence, given a training set S and given a new (testing) input example $\mathbf{x}$, a sample-compressed Gibbs classifier G_Q chooses randomly $(\mathbf{i}, \sigma)$ according to Q to obtain classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$ which is then used to determine the class label of $\mathbf{x}$.

Hence, *given a training set* S, the true risk $R(G_Q)$ of G_Q is defined by:

$$R(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{i}, \sigma) \sim Q} R(\sigma, S_{\mathbf{i}})$$

and its empirical risk $R_S(G_Q)$ is defined by:

$$R_S(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})$$

Given a posterior Q, some expectations below will be performed on another distribution defined by the following.

Definition 3 *Given a distribution $Q \in \mathcal{P}_{\mathcal{K}}$ we will denote by $\overline{Q}$ the distribution of $\mathcal{P}_{\mathcal{K}}$ defined as follows:*

$$\overline{Q}(\mathbf{i},\sigma) \stackrel{\text{def}}{=} \frac{Q(\mathbf{i},\sigma)}{|\bar{\mathbf{i}}| \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \frac{1}{|\bar{\mathbf{i}}|}} \quad \forall (\mathbf{i},\sigma) \in \mathcal{K}$$

where $|\bar{\mathbf{i}}| \stackrel{\text{def}}{=} m - |\mathbf{i}|$.

To simplify some formulas, let us also define:

$$d_{\overline{Q}} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim \overline{Q}} |\mathbf{i}| \tag{2}$$

Then, it follows directly from the definitions that:

$$\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \frac{1}{|\bar{\mathbf{i}}|} = \frac{1}{\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim \overline{Q}} |\bar{\mathbf{i}}|} = \frac{1}{m - d_{\overline{Q}}} \tag{3}$$

The next theorem constitutes our main result:

Theorem 4 *For any reconstruction function*

$$\mathcal{R} : (\mathcal{X} \times \mathcal{Y})^m \times \mathcal{K} \longrightarrow \mathcal{H}$$

and for any prior distribution $P \in \mathcal{P}_{\mathcal{K}}$, we have:

$$\Pr_{S\sim D^m}\left(\begin{array}{l} \forall\, Q \in \mathcal{P}_{\mathcal{K}}: \\ \mathrm{kl}(R_S(G_Q) \| R(G_Q)) \leq \\ \quad \frac{1}{m - d_{\overline{Q}}} \left[\mathrm{KL}(\overline{Q}\,\|P) + \ln \frac{m+1}{\delta} \right] \end{array} \right) \geq 1 - \delta$$

This theorem is a generalization of Theorem 1 because the latter correspond to the case where the probability distribution Q has its weight only when $|\mathbf{i}| = 0$, and clearly in this case: $\frac{1}{m - d_{\overline{Q}}} = \frac{1}{m}$ and $\overline{Q} = Q$. We have also obtained, under some restrictions (see Theorem 12), a risk bound that does not depend on the transformed posterior $\overline{Q}$.

However, it is important to note that $\overline{Q}(\mathbf{i},\sigma)$ is smaller than $Q(\mathbf{i},\sigma)$ for classifiers $\mathcal{R}(\mathbf{i},\sigma)$ having a compression set size smaller than the Q-average. This, combined with the the fact that $\mathrm{KL}(\overline{Q}\,\|P)$ favors $\overline{Q}$'s close to P, implies that there will be a specialization performed by Q on classifiers having small compression set sizes. As example, in the case where $\overline{Q} = P$, it is easy to see that Q will put more weight than P on "small" classifiers. The specialization suggested by Theorem 4 is therefore stronger than what it would have been if $\mathrm{KL}(Q\|P)$ would have been in the risk bound instead of $\mathrm{KL}(\overline{Q}\,\|P)$. Thus, Theorem 4 reinforces Occam's principle of parsimony.

The rest of the section is devoted to the proof of Theorem 4.

Definition 5 *Let $S \in (\mathcal{X} \times \mathcal{Y})^m$, D a distribution on $\mathcal{X} \times \mathcal{Y}$, and $(\mathbf{i},\sigma) \in \mathcal{K}$. We will denote by $B_S(\mathbf{i},\sigma)$, the probability that the classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$ of (true) risk $R(\sigma, S_{\mathbf{i}})$ makes exactly $|\bar{\mathbf{i}}| R(\sigma, S_{\mathbf{i}})$ errors on $S'_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|}$. Hence, equivalently:*

$$B_S(\mathbf{i},\sigma) \stackrel{\text{def}}{=} \binom{|\bar{\mathbf{i}}|}{|\bar{\mathbf{i}}| R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})} (R(\sigma, S_{\mathbf{i}}))^{|\bar{\mathbf{i}}| R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})} (1 - R(\sigma, S_{\mathbf{i}}))^{|\bar{\mathbf{i}}| - |\bar{\mathbf{i}}| R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})}$$

Lemma 6 *For any prior distribution $P \in \mathcal{P}_{\mathcal{K}}$, we have:*

$$\Pr_{S\sim D^m}\left(\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim P} \frac{1}{B_S(\mathbf{i},\sigma)} \leq \frac{m+1}{\delta} \right) \geq 1 - \delta$$

Proof First observe that:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim D^m} \frac{1}{B_S(\mathbf{i},\sigma)} \\
&\quad = \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}} \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim D^{|\bar{\mathbf{i}}|}} \frac{1}{B_S(\mathbf{i},\sigma)} \\
&\quad = \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}} \sum_{k=0}^{|\bar{\mathbf{i}}|} \frac{\Pr_{S_{\bar{\mathbf{i}}}\sim D^{|\bar{\mathbf{i}}|}} \left(|\bar{\mathbf{i}}| R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = k \right)}{\binom{|\bar{\mathbf{i}}|}{k} (R(\sigma, S_{\mathbf{i}}))^k (1 - R(\sigma, S_{\mathbf{i}}))^{|\bar{\mathbf{i}}| - k}} \\
&\quad = \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}} \sum_{k=0}^{|\bar{\mathbf{i}}|} \frac{\Pr_{S_{\bar{\mathbf{i}}}\sim D^{|\bar{\mathbf{i}}|}} \left(|\bar{\mathbf{i}}| R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = k \right)}{\Pr_{S_{\bar{\mathbf{i}}}\sim D^{|\bar{\mathbf{i}}|}} \left(|\bar{\mathbf{i}}| R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = k \right)} \\
&\quad = \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}} \sum_{k=0}^{m-|\mathbf{i}|} 1 \;\; = \;\; m - |\mathbf{i}| + 1
\end{aligned}$$

Then, for any distribution $P \in \mathcal{P}_{\mathcal{K}}$ (independent of S):

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim P} \frac{1}{B_S(\mathbf{i},\sigma)} &= \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \frac{1}{B_S(\mathbf{i},\sigma)} \\
&\leq m + 1
\end{aligned}$$

By Markov's inequality, we are done. □

Lemma 7 *Given $S \in (\mathcal{X} \times \mathcal{Y})^m$ and any two distributions $P, Q \in \mathcal{P}_{\mathcal{K}}$, we have:*

$$\mathrm{KL}(Q\|P) \;\geq\; \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \ln\!\left(\frac{1}{B_S(\mathbf{i},\sigma)}\right) - \ln\!\left(\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim P} \frac{1}{B_S(\mathbf{i},\sigma)}\right)$$

Proof Let $P' \in \mathcal{P}_{\mathcal{K}}$, be the following distribution:

$$P'(\mathbf{i},\sigma) = \frac{\frac{1}{B_S(\mathbf{i},\sigma)} P(\mathbf{i},\sigma)}{\mathop{\mathbf{E}}_{(\mathbf{i}',\sigma')\sim P}\left(\frac{1}{B_S(\mathbf{i}',\sigma')}\right)} \qquad \forall (\mathbf{i},\sigma) \in \mathcal{K}$$

Since $\mathrm{KL}(Q\|P') \geq 0$, we have:

$$\begin{aligned}
&\mathrm{KL}(Q\|P) \\
&\geq \mathrm{KL}(Q\|P) - \mathrm{KL}(Q\|P') \\
&= \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \ln\!\left(\frac{Q(\mathbf{i},\sigma)}{P(\mathbf{i},\sigma)}\right) \\
&\qquad - \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \ln\!\left(\frac{Q(\mathbf{i},\sigma)}{P(\mathbf{i},\sigma)} \frac{\mathop{\mathbf{E}}_{(\mathbf{i}',\sigma')\sim P}\left(\frac{1}{B_S(\mathbf{i}',\sigma')}\right)}{\frac{1}{B_S(\mathbf{i},\sigma)}}\right) \\
&= \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \ln\!\left(\frac{1}{B_S(\mathbf{i},\sigma)}\right) - \ln\left(\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim P} \frac{1}{B_S(\mathbf{i},\sigma)}\right)
\end{aligned}$$

□

Lemma 8 *For any prior distribution $P \in \mathcal{P}_{\mathcal{K}}$, we have:*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall Q\in\mathcal{P}_{\mathcal{K}}:\\ \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \ln\frac{1}{B_S(\mathbf{i},\sigma)} \leq \mathrm{KL}(Q\|P) + \ln\frac{m+1}{\delta}\end{array}\right) \geq 1-\delta$$

Proof It follows from Lemma 6 that:

$$\Pr_{S\sim D^m}\left(\ln\left(\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim P} \frac{1}{B_S(\mathbf{i},\sigma)}\right) \leq \ln\left(\tfrac{m+1}{\delta}\right)\right) \geq 1-\delta$$

This implies that:

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall Q\in\mathcal{P}_{\mathcal{K}}:\\ \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \ln\frac{1}{B_S(\mathbf{i},\sigma)} \leq \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \ln\frac{1}{B_S(\mathbf{i},\sigma)} \\ \qquad - \ln\left(\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim P} \frac{1}{B_S(\mathbf{i},\sigma)}\right) + \ln\left(\frac{m+1}{\delta}\right)\end{array}\right) \geq 1-\delta$$

By Lemma 7, we then obtain the result. □

Lemma 9 *For any $f : \mathcal{K} \longrightarrow \mathbb{R}^+$, and for any $Q, Q' \in \mathcal{P}_{\mathcal{K}}$ related by*

$$Q'(\mathbf{i},\sigma)\, f(\mathbf{i},\sigma) \;=\; \frac{1}{\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \frac{1}{f(\mathbf{i},\sigma)}}\, Q(\mathbf{i},\sigma) \quad ,$$

we have:

$$\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q'}\Big(f(\mathbf{i},\sigma)\ \mathrm{kl}(R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\|R(\sigma,S_{\mathbf{i}}))\Big) \;\geq\; \frac{1}{\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q}\left(\frac{1}{f(\mathbf{i},\sigma)}\right)}\ \mathrm{kl}(R_S(G_Q)\|R(G_Q))$$

Note that, by Definition 3, we have $Q' = \overline{Q}$ when $f(\mathbf{i},\sigma) = |\bar{\mathbf{i}}|$.

We provide here a proof for the countable case. Because of the convexity of $\mathrm{kl}(\cdot\|\cdot)$, the lemma holds for the uncountable case as well.

Proof By the log-sum inequality (Cover & Thomas, 1991) that we apply twice at line 4, we have:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q'}\Big(f(\mathbf{i},\sigma)\ \mathrm{kl}(R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\|R(\sigma,S_{\mathbf{i}}))\Big) \\
&= \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q'}\Bigg(f(\mathbf{i},\sigma)\ \Bigg[R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}) \ln\frac{R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{R(\sigma,S_{\mathbf{i}})} \\
&\qquad\qquad + (1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))\ln\frac{1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{1-R(\sigma,S_{\mathbf{i}})}\Bigg]\Bigg) \\
&= \sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma) f(\mathbf{i},\sigma)\Bigg[R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\ln\frac{R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{R(\sigma,S_{\mathbf{i}})} \\
&\qquad\qquad + (1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))\ln\frac{1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{1-R(\sigma,S_{\mathbf{i}})}\Bigg] \\
&= \sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma) f(\mathbf{i},\sigma) R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}) \ln\frac{Q'(\mathbf{i},\sigma)f(\mathbf{i},\sigma)R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{Q'(\mathbf{i},\sigma)f(\mathbf{i},\sigma)R(\sigma,S_{\mathbf{i}})} \\
&\qquad + \sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma) f(\mathbf{i},\sigma)(1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})) \\
&\qquad\qquad \cdot\ \ln\frac{Q'(\mathbf{i},\sigma)f(\mathbf{i},\sigma)(1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))}{Q'(\mathbf{i},\sigma)f(\mathbf{i},\sigma)(1-R(\sigma,S_{\mathbf{i}}))} \\
&\geq \left(\sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma) f(\mathbf{i},\sigma) R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\right) \\
&\qquad \cdot \ln\frac{\sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma)f(\mathbf{i},\sigma)R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{\sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma)f(\mathbf{i},\sigma)R(\sigma,S_{\mathbf{i}})} \\
&\qquad + \left(\sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma) f(\mathbf{i},\sigma)(1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))\right) \\
&\qquad\qquad \cdot\ \ln\frac{\sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma)f(\mathbf{i},\sigma)(1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))}{\sum_{(\mathbf{i},\sigma)} Q'(\mathbf{i},\sigma)f(\mathbf{i},\sigma)(1-R(\sigma,S_{\mathbf{i}}))}
\end{aligned} \tag{4}$$

$$= \left(\frac{1}{\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q}\left(\frac{1}{f(\mathbf{i},\sigma)}\right)} \sum_{(\mathbf{i},\sigma)} Q(\mathbf{i},\sigma) R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) \right) \cdot \ln \frac{\sum_{(\mathbf{i},\sigma)} Q(\mathbf{i},\sigma) R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})}{\sum_{(\mathbf{i},\sigma)} Q(\mathbf{i},\sigma) R(\sigma, S_{\mathbf{i}})}$$

$$+ \left(\frac{1}{\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q}\left(\frac{1}{f(\mathbf{i},\sigma)}\right)} \sum_{(\mathbf{i},\sigma)} Q(\mathbf{i},\sigma)(1 - R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})) \right) \cdot \ln \frac{\sum_{(\mathbf{i},\sigma)} Q(\mathbf{i},\sigma)(1 - R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}))}{\sum_{(\mathbf{i},\sigma)} Q(\mathbf{i},\sigma)(1 - R(\sigma, S_{\mathbf{i}}))}$$

$$= \frac{1}{\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q}\left(\frac{1}{f(\mathbf{i},\sigma)}\right)} \left[R_S(G_Q) \ln \frac{R_S(G_Q)}{R(G_Q)} + (1 - R_S(G_Q)) \ln \frac{1 - R_S(G_Q)}{1 - R(G_Q)} \right]$$

$$= \frac{1}{\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q}\left(\frac{1}{f(\mathbf{i},\sigma)}\right)} \ \mathrm{kl}(R_S(G_Q) \| R(G_Q))$$

□

Proof of Theorem 4: By the relative entropy Chernoff bound (see Langford (2005) for example):

$$\sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} \leq \exp\left[-n \cdot \mathrm{kl}\left(\frac{k}{n} \middle\| p \right) \right]$$

for all $\frac{k}{n} \leq p$. Since $\binom{n}{j} = \binom{n}{n-j}$ and $\mathrm{kl}\left(\frac{k}{n}\|p\right) = \mathrm{kl}\left(1 - \frac{k}{n}\|1-p\right)$, the following inequality therefore holds for every $k \in \{0, 1, \ldots, n\}$:

$$\binom{n}{k} p^k (1-p)^{n-k} \leq \exp\left[-n \cdot \mathrm{kl}\left(\frac{k}{n} \middle\| p \right) \right]$$

In our setting this means that $\forall (\mathbf{i}, \sigma) \in \mathcal{K}$ and $\forall S \in (\mathcal{X} \times \mathcal{Y})^m$, we have:

$$B_S(\mathbf{i}, \sigma) \leq \exp\left[-|\bar{\mathbf{i}}| \cdot \mathrm{kl}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) \middle\| R(\sigma, S_{\mathbf{i}}) \right) \right]$$

For any distribution $Q \in \mathcal{P}_{\mathcal{K}}$, we therefore have:

$$\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} \ln\left(\frac{1}{B_S(\mathbf{i},\sigma)} \right) \geq \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q} |\bar{\mathbf{i}}| \cdot \mathrm{kl}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) \| R(\sigma, S_{\mathbf{i}}) \right) \quad (5)$$

Hence, Lemma 8 together with Lemma 9 [with $f(\mathbf{i},\sigma) \stackrel{\text{def}}{=} |\bar{\mathbf{i}}|$ and, therefore, with $Q' = \overline{Q}$] gives the result. □

In the next sections, we derive new PAC-Bayes bounds by restricting the set of possible posteriors on $\mathcal{K}$.

5. Single Sample-compressed Classifiers

Let us examine the case when the (stochastic) sample-compressed Gibbs classifier becomes a deterministic classifier with a posterior having with all its weight on a single $(\mathbf{i}, \sigma)$. In that case, Lemma 8 gives the following risk bound for any prior $P \in \mathcal{P}_{\mathcal{K}}$:

$$\Pr_{S\sim D^m}\left(\begin{array}{l} \forall (\mathbf{i},\sigma) \in \mathcal{K}: \\ \ln \frac{1}{B_S(\mathbf{i},\sigma)} \leq \ln\left(\frac{1}{P(\mathbf{i},\sigma)}\right) + \ln \frac{m+1}{\delta} \end{array} \right) \geq 1 - \delta$$

If we now use the binomial distribution:

$$\mathrm{Bin}\left(\frac{k}{m}, r\right) \stackrel{\text{def}}{=} \binom{m}{k} r^k (1-r)^{m-k}$$

to express $B_S(\mathbf{i}, \sigma)$ as:

$$B_S(\mathbf{i}, \sigma) \ = \ \mathrm{Bin}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}), R(\sigma, S_{\mathbf{i}}) \right)$$

and use the *binomial inversion* defined as:

$$\overline{\mathrm{Bin}}\left(\frac{k}{m}, \delta\right) \stackrel{\text{def}}{=} \sup\left\{ r \colon \mathrm{Bin}\left(\frac{k}{m}, r\right) \geq \delta \right\}$$

the previous risk bound gives the following:

Theorem 10 *For any reconstruction function*

$$\mathcal{R} : (\mathcal{X} \times \mathcal{Y})^m \times \mathcal{K} \longrightarrow \mathcal{H}$$

and for any prior distribution $P \in \mathcal{P}_{\mathcal{K}}$, we have:

$$\Pr_{S\sim D^m}\left(\begin{array}{l} \forall (\mathbf{i},\sigma) \in \mathcal{K}: \\ R(\sigma, S_{\mathbf{i}}) \leq \overline{\mathrm{Bin}}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}), \frac{P(\mathbf{i},\sigma)\delta}{(m+1)} \right) \end{array} \right) \geq 1 - \delta$$

Let us now compare this risk bound with the tightest currently known sample-compression risk bound. This bound, due to Langford (2005), is currently restricted to the case when no message string σ is used to identify classifiers. If we generalize it to the case when message strings are used, we get (in our notation):

$$\Pr_{S\sim D^m}\left(\begin{array}{l} \forall (\mathbf{i},\sigma) \in \mathcal{K}: \\ R(\sigma, S_{\mathbf{i}}) \leq \overline{\mathrm{BinT}}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}), P(\mathbf{i},\sigma)\delta \right) \end{array} \right) \geq 1 - \delta$$

Note that, instead of using the binomial inversion, the bound of Langford (2005) uses the binomial *tail* inversion defined by:

$$\overline{\mathrm{BinT}}\left(\frac{k}{m}, \delta\right) \stackrel{\text{def}}{=} \sup\left\{ r \colon \sum_{i=0}^{k} \mathrm{Bin}\left(\frac{i}{m}, r\right) \geq \delta \right\}$$

We therefore have (for all values of m, k, δ):

$$\overline{\mathrm{Bin}}\left(\frac{k}{m}, \delta\right) \ \leq \ \overline{\mathrm{BinT}}\left(\frac{k}{m}, \delta\right)$$

When both m and δ are non zero, the equality is realized only for $k = 0$. Hence, if we did not have the denominator of $(m + 1)$, our bound would be tighter than the bound of Langford (2005).

Hence, for a single sample-compressed classifier, the bound of Theorem 10 is "competitive" with the currently tightest sample-compression risk bound.

Let us now investigate, through some special cases, the additional predictive power that can be obtained by using a (randomized) sample-compressed Gibbs classifier instead of a deterministic one.

6. The Consistent Case

An interesting case is when we restrict ourselves to posteriors having non zero weight only on consistent classifiers (i.e., classifiers having zero empirical risk). For these cases, we have:

$$\mathrm{kl}(R_S(G_Q)\|R(G_Q)) = \ln\left(\frac{1}{1-R(G_Q)}\right)$$

Given a training set S, let us consider the *version space* $\mathcal{V}(S)$ defined as:

$$\mathcal{V}(S) \stackrel{\text{def}}{=} \left\{(\mathbf{i},\sigma)\in\mathcal{K} \mid R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})=0\right\}$$

Suppose that the posterior Q has non zero weight only in some subset $\mathcal{U}\subseteq\mathcal{V}(S)$. From Definition 3, it is clear that $\overline{Q}$ has also non zero weight on the same subset $\mathcal{U}$.

Denote by $P(\mathcal{U})$ the weight assigned to $\mathcal{U}$ by the prior P, and define $P_{\mathcal{U}}$ and $Q_{\mathcal{U}}\in\mathcal{P}_{\mathcal{K}}$ as:

$$P_{\mathcal{U}}(\mathbf{i},\sigma)\stackrel{\text{def}}{=}\frac{P(\mathbf{i},\sigma)}{P(\mathcal{U})} \quad\text{and}\quad Q_{\mathcal{U}}(\mathbf{i},\sigma)\stackrel{\text{def}}{=}\frac{|\bar{\mathbf{i}}|\,P_{\mathcal{U}}(\mathbf{i},\sigma)}{\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim P_{\mathcal{U}}}|\bar{\mathbf{i}}|}$$

for all $(\mathbf{i},\sigma)\in\mathcal{U}$.

Then, from Definition 3 and Equation 3), it follows that

$$\overline{Q_{\mathcal{U}}}=P_{\mathcal{U}} \quad\text{and}\quad \mathrm{KL}(\overline{Q_{\mathcal{U}}}\,\|\,P) \;=\; \ln\left(\frac{1}{P(\mathcal{U})}\right).$$

Hence, in this case, Theorem 4 reduces to:

Corollary 11 *For any reconstruction function*

$$\mathcal{R}:\,(\mathcal{X}\times\mathcal{Y})^m\times\mathcal{K}\longrightarrow\mathcal{H}$$

and for any prior $P\in\mathcal{P}_{\mathcal{K}}$, we have:

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\,\mathcal{U}\in\mathcal{V}(S):\\ R(G_{Q_{\mathcal{U}}})\leq 1-\exp\left[-\dfrac{\ln\left(\dfrac{m+1}{P(\mathcal{U})\delta}\right)}{m-d_{\overline{Q_{\mathcal{U}}}}}\right]\end{array}\right)\geq 1-\delta$$

This bound exhibits a non trivial tradeoff between $|\mathcal{U}|$ and $d_{\overline{Q_{\mathcal{U}}}}$. The bound gets smaller as $P(\mathcal{U})$ increases, and larger as $d_{\overline{Q_{\mathcal{U}}}}$ increases. However, an increase in $|\mathcal{U}|$ should normally be accompanied by an increase in $d_{\overline{Q_{\mathcal{U}}}}$ and the minimum of the bound should generally be reached for some non trivial value of $|\mathcal{U}|$. Only on rare occasions we expect the minimum to be reached for the single classifier case of $|\mathcal{U}|=1$.

Therefore, the risk bound supports the theory that it is generally preferable to randomize the predictions over several (empirically good) sample-compressed classifiers than to predict only with a single (empirically good) sample-compressed classifier.

7. Bounded Compression Set Sizes

Another interesting case is when we restrict Q to have a non zero weight only on classifiers having a compression set size $|\mathbf{i}|\leq d$ for some $d\in\{0,1,\ldots,m\}$.

For these cases, let us define:

$$\mathcal{P}_{\mathcal{K}}^{|\mathbf{i}|\leq d} \stackrel{\text{def}}{=} \{Q\in\mathcal{P}_{\mathcal{K}} \mid Q(\mathbf{i},\sigma)=0 \quad \text{if } |\mathbf{i}|>d\}$$

We then have the following theorem.

Theorem 12 *For any reconstruction function*

$$\mathcal{R}:\,(\mathcal{X}\times\mathcal{Y})^m\times\mathcal{K}\longrightarrow\mathcal{H}$$

and for any prior distribution $P\in\mathcal{P}_{\mathcal{K}}$, we have:

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\,Q\in\mathcal{P}_{\mathcal{K}}^{|\mathbf{i}|\leq d}:\\ \mathrm{kl}(R_S(G_Q)\|R(G_Q))\leq\\ \qquad\frac{1}{m-d}\left[\mathrm{KL}(Q\|P)+\ln\frac{m+1}{\delta}\right]\end{array}\right)\geq 1-\delta$$

Proof As in the proof of Theorem 4, for any distribution $Q\in\mathcal{P}_{\mathcal{K}}^{|\mathbf{i}|\leq d}$, Equation 5 holds. Hence, by Lemma 9 [with $Q'=Q$ and $f(\mathbf{i},\sigma)=1$ for all $(\mathbf{i},\sigma)$], and since $|\bar{\mathbf{i}}|\geq m-d$ for any $(\mathbf{i},\sigma)$ that has weight in Q, we have:

$$\begin{aligned}&\mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q}\ln\left(\frac{1}{B_S(\mathbf{i},\sigma)}\right)\\ &\geq \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q}|\bar{\mathbf{i}}|\cdot\mathrm{kl}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\|R(\sigma,S_{\mathbf{i}})\right)\\ &\geq \mathop{\mathbf{E}}_{(\mathbf{i},\sigma)\sim Q}(m-d)\cdot\mathrm{kl}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\|R(\sigma,S_{\mathbf{i}})\right)\\ &\geq (m-d)\cdot\mathrm{kl}\left(R_S(G_Q)\|R(G_Q)\right)\end{aligned}$$

Then, by Lemma 8, we are done. □

Again, $\mathrm{KL}(Q\|P)$ is small when Q has all its weight on a single $(\mathbf{i}_0, \sigma_0)$ and increases as we add to Q more classifiers having $|\mathbf{i}| \leq d$. If we find a set U containing several such classifiers, each having (roughly) the same empirical risk $R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})$, the risk bound for $R(G_Q)$ with a posterior having non zero weight on all U, like

$$Q(\mathbf{i}, \sigma) \quad = \quad \frac{P(\mathbf{i}, \sigma)}{\sum_{(\mathbf{i}', \sigma') \in U} P(\mathbf{i}', \sigma')}$$

for all $(\mathbf{i}, \sigma) \in U$, will be smaller than the risk bound for a posterior Q having all its weight on a single $(\mathbf{i}_0, \sigma_0)$.

Therefore, for these cases, the risk bound supports the theory that it is preferable to randomize the predictions over several (empirically good) sample-compressed classifiers than to predict only with a single (empirically good) sample-compressed classifier.

8. Conclusion

We have derived a PAC-Bayes theorem for the sample-compression setting where each classifier is described by a compression subset of the training data and a message string of additional information. This theorem reduces to the PAC-Bayes theorem of Seeger (2002) and Langford (2005) in the usual data-independent setting when classifiers are represented only by data-independent message strings (or parameters taken from a continuous set). For posteriors having all their weights on a single sample-compressed classifier, the general risk bound reduces to a bound similar to the tight sample-compression bound of Langford (2005). The PAC-Bayes risk bound of Theorem 4 is, however, valid for sample-compressed Gibbs classifiers with arbitrary posteriors. We have shown, both in the consistent case and in the case of posteriors on classifiers with bounded compression set sizes, that a stochastic Gibbs classifier defined on a posterior over several sample-compressed classifiers can have a smaller risk bound than any such single (deterministic) sample-compressed classifier.

Since the risk bounds derived in this paper are tight, it is hoped that they will be effective at guiding learning algorithms for choosing the optimal tradeoff between the empirical risk, the sample compression set size, and the "distance" between the prior and the posterior.

Acknowledgements

Work supported by NSERC Discovery grants 262067 and 0122405.

References

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*, chapter 12. Wiley.

Floyd, S., & Warmuth, M. (1995). Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, *21*, 269–304.

Graepel, T., Herbrich, R., & Shawe-Taylor, J. (2000). Generalisation error bounds for sparse linear classifiers. *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory* (pp. 298–303).

Graepel, T., Herbrich, R., & Williamson, R. C. (2001). From margin to sparsity. *Advances in neural information processing systems* (pp. 210–216).

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

Langford, J., & Shawe-Taylor, J. (2003). PAC-Bayes & margins. In S. T. S. Becker and K. Obermayer (Eds.), *Advances in neural information processing systems 15*, 423–430. Cambridge, MA: MIT Press.

Littlestone, N., & Warmuth, M. (1986). *Relating data compression and learnability* (Technical Report). University of California Santa Cruz, Santa Cruz, CA.

Marchand, M., & Shawe-Taylor, J. (2002). The set covering machine. *Journal of Machine Learning Reasearch*, *3*, 723–746.

McAllester, D. (1999). Some PAC-Bayesian theorems. *Machine Learning*, *37*, 355–363.

McAllester, D. (2003a). PAC-Bayesian stochastic model selection. *Machine Learning*, *51*, 5–21. A priliminary version appeared in proceedings of COLT'99.

McAllester, D. (2003b). Simplified PAC-Bayesian margin bounds. *Proceedings of the 16th Annual Conference on Learning Theory, Lecture Notes in Artificial Intelligence*, *2777*, 203–215.

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