Journal of Machine Learning Research 8 (2007) 1461-1487 Submitted 7/06; Revised 4/07; Published 7/07

PAC-Bayes Risk Bounds for Stochastic Averages and Majority Votes of Sample-Compressed Classifiers

François Laviolette FRANCOIS.LAVIOLETTE@IFT.ULAVAL.CA
Mario Marchand MARIO.MARCHAND@IFT.ULAVAL.CA
Département IFT-GLO
Université Laval
Québec (QC)
Canada, G1K 7P4

Editor: Manfred K. Warmuth

Abstract

We propose 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 setting, which is the appropriate one to describe many learning algorithms, strictly generalizes the usual data-independent setting where classifiers are represented only by data-independent message strings (or parameters taken from a continuous set). The proposed PAC-Bayes theorem for the sample-compression setting reduces to the PAC-Bayes theorem of Seeger (2002) and Langford (2005) when the compression subset of each classifier vanishes. 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 proposed in Laviolette et al. (2005). Finally, we extend our results to the case where each sample-compressed classifier of a data-dependent ensemble may abstain of predicting a class label.

Keywords: PAC-Bayes, risk bounds, sample-compression, set covering machines, decision list machines

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 might have over deterministic classifier selection.

©2007 François Laviolette and Mario Marchand.

A constraint normally 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 and Warmuth, 1986, Floyd and 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 (Graepel et al., 2005). Moreover, some sample-compression algorithms (Marchand and Shawe-Taylor, 2002, Marchand and Sokolova, 2005) have achieved very good performance in practice by deterministically choosing a sparse classifier making few training errors. It is therefore worthwhile to investigate how 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 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. As in the case for the usual data-independent setting, the PAC-Bayes theorem for the sample-compression setting states that a stochastic Gibbs classifier defined on a posterior over several sample-compressed classifiers generally has a smaller risk bound than any such single (deterministic) sample-compressed classifier. Nevertheless, in the limit where the posterior Q puts all its weight on a single sample-compressed classifier, the new PAC-Bayes risk bound reduces to a bound similar to the tight sample-compression bound of Laviolette et al. (2005) (which applies only to single sample-compressed classifiers).

Several "PAC-Bayesian sample-compression bounds" have recently been proposed by Graepel et al. (2005). However, all these bounds, except one (that concerns consistent SVM classifiers with fixed sparsity), deals with classifiers that use a fixed subset of the training examples. In contrast, we provide bounds that applies to a stochastic average (and a majority vote) of classifiers using different subsets (of different sizes) of the training examples. Finally, we extend our results to the important case where we have an ensemble of sample-compressed classifiers that can abstain of predicting a class label.

The paper is organized as follows. After providing a few definitions in Section 2, we review, in Section 3, the PAC-Bayes theorem for the data-independent setting. Section 4 is the "core" section of this paper. In that section, we present the sample-compression setting and show how it generalizes the usual data-independent setting. We then provide the main theorem of this paper, Theorem 3, which is a PAC-Bayes theorem for the sample-compression setting. In Section 5, we provide examples of learning algorithms that produce classifiers that are well-described within this sample-compression setting. We then show, in Section 6, that Theorem 3 reduces to a bound similar to the tight sample-compression bound of Laviolette et al. (2005) in the limit where the posterior Q puts all its weight on a single sample-compressed classifier. In that section, we also present a bound for the "intermediate" case where the posterior has all its weight on a single compression sequence and non-zero weight on several messages. We then show that the risk bound reduces to the one recently proposed in Laviolette et al. (2006) for the PAC-Bayes SCM. In Section 7, we provide an alternative formulation of Theorem 3 by including the training errors into the compression sequence. We then generalize, in Section 8, Theorem 3 to the case were the individual sample-compressed classifiers may abstain of predicting a class label. Finally, we conclude in Section 9.

This paper extends the preliminary work of Laviolette and Marchand (2005) and Laviolette et al. (2006).

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 $(\mathbf{x},y)$ 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 $(\mathbf{x},y)$ 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. Hence,

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

where $I(a) = 1$ if predicate a is true and 0 otherwise.

Given a training sequence $S = \langle(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_m,y_m)\rangle$ of m examples, the *empirical risk* $R_S(f)$ on S, of any classifier f, is defined according to

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

In this paper, we will distinguish the usual data-independent setting from the (more general) sample-compression setting. By the *data-independent setting*, we mean the “usual” setting where a space $\mathcal{H}$ of classifiers is defined without making any reference to the training data S. Examples of such a space $\mathcal{H}$ include the set of linear classifiers on $\mathbb{R}^n$, the set of radial-basis functions on $\mathbb{R}^n$, the set of k-CNF Boolean formulae (Valiant, 1984) on $\{0,1\}^n$, the set of decision lists (Rivest, 1987) on $\{0,1\}^n$. In contrast, the set of data-dependent balls (Marchand and Shawe-Taylor, 2002)—where each ball of this set is centered on a training example—is an example of a set of classifiers which is defined only after observing the training data S. Such a set of classifiers is qualified as being *data-dependent*. Moreover, since each data-dependent ball is constructed from a small subset of the training data S, it is an example of what we call a *sample-compressed classifier*. We define more formally the sample-compression setting in section 4 in such a way that it extends the usual data-independent setting. The next section presents the PAC-Bayes theorem within the (restricted) data-independent setting.

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

The PAC-Bayes theorem provides tight upper and lower bounds on the risk of a stochastic classifier called the *Gibbs classifier*. Given an input example $\mathbf{x}$, the label assigned to $\mathbf{x}$ by the Gibbs classifier G_Q is defined by the following process. We first choose randomly 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. Hence,

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

Similarly, the empirical risk $R_S(G_Q)$ of G_Q, on a training sequence S of m examples, is given by

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

The PAC-Bayes theorem was first proposed by McAllester (1999, 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 $\mathcal{H}$ and any $\delta \in (0,1]$, we have*

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

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:

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

It is rarely mentioned that this theorem provides both an upper bound and a lower bound on the true risk $R(G_Q)$ based on its empirical risk $R_S(G_Q)$. With probability at least $1-\delta$ over the random draws of S, $R(G_Q)$ is upper-bounded by

$$\sup\left(R\colon\ \mathrm{kl}(R_S(G_Q)\|R)\ \leq\ \frac{1}{m}\left[\mathrm{KL}(Q\,\|P) + \ln\frac{m+1}{\delta}\right]\right)$$

and lower-bounded by

$$\inf\left(R\colon\ \mathrm{kl}(R_S(G_Q)\|R)\ \leq\ \frac{1}{m}\left[\mathrm{KL}(Q\,\|P) + \ln\frac{m+1}{\delta}\right]\right)\,.$$

The bounds provided by Theorem 1 hold for any *fixed* prior P on $\mathcal{H}$ and also hold *uniformly* for all posteriors Q on $\mathcal{H}$; this includes any Q chosen by the learner *after* observing S. This is specified in the theorem by the fact that the quantifier $\forall Q$ occurs inside the probability over the random draws of S whereas the quantifier $\forall P$ occurs (textually) outside that probability.

The upper bound given by the PAC-Bayes theorem for the risk of Gibbs classifiers can be turned into an upper bound for the risk of majority-vote classifiers (often called 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 $R(B_Q) \leq 2R(G_Q)$. It has been shown (Langford and Shawe-Taylor, 2003, McAllester, 2003b, Germain et al., 2007, Lacasse et al., 2007) that there exists circumstances where this "factor-of-two" rule can be improved. However, for many [1] posteriors Q, one can often find a data-generating distribution where we have $R(B_Q) = 2R(G_Q) - \varepsilon$ for arbitrary small $\varepsilon > 0$.

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 $h^* \in \mathcal{H}$ such that $h^*(\mathbf{x}) = B_Q(\mathbf{x})\ \forall \mathbf{x} \in \mathcal{X}$. Such a classifier h^* is equivalent to B_Q since it performs the same classification $\forall \mathbf{x} \in \mathcal{X}$. For example, a linear classifier with weight vector $\mathbf{w}$ is equivalent to a Bayes classifier B_Q over linear classifiers with any distribution Q 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.

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

In the sample-compression setting, learning algorithms have access to a data-dependent set of classifiers defined as follows. Given a training sequence $S = \langle \mathbf{z}_1, \ldots, \mathbf{z}_m \rangle$ of m examples, each classifier is described entirely by two *complementary sources of information*: a subsequence $S_\mathbf{i}$ of S, called the *compression sequence*, and a *message* σ which represents the additional information needed to obtain a classifier from the compression sequence.

Given a training sequence S of m examples, the compression subsequence $S_\mathbf{i}$ of S is defined by the following 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 $\mathbf{i}$ of indices defined above. We will also use $\bar{\mathbf{i}}$ to denote the vector of indices not present in $\mathbf{i}$. Hence, the union of all the examples of $S_\mathbf{i}$ and $S_{\bar{\mathbf{i}}}$ gives all the m examples of S. Finally, we will denote by $\mathcal{I}$ the set of the 2^m possible realizations of $\mathbf{i}$.

The fact that each classifier is described by a compression sequence and a message implies that there exists a *reconstruction function* $\mathcal{R}$ that outputs a classifier $\mathcal{R}(\sigma, S_\mathbf{i})$ when given an arbitrary compression sequence $S_\mathbf{i}$ and a message σ chosen from the set $\mathcal{M}(S_\mathbf{i})$ of all distinct messages that can be supplied to $\mathcal{R}$ with the compression sequence $S_\mathbf{i}$. This set $\mathcal{M}(S_\mathbf{i})$ must be defined *a priori* (before observing S) for all possible sequences $S_\mathbf{i}$ of examples. For any sequence S of m examples, we will also use

$$\mathcal{M}_S \stackrel{\text{def}}{=} \bigcup_{\mathbf{i} \in \mathcal{I}} \mathcal{M}(S_\mathbf{i})\,.$$

[1] For example, if there exists $(\mathbf{x}, y)$ such that $B_Q(\mathbf{x}) \neq y$ and $\Pr_{h \sim Q}(h(\mathbf{x}) \neq y) = \frac{1}{2} + \varepsilon$, then $R(B_Q) = 2R(G_Q) - 2\varepsilon$ for the data-generating distribution that has all its weight on $(\mathbf{x}, y)$.

The perceptron learning rule and the SVM are examples of learning algorithms where the final classifier can be reconstructed solely from a compression sequence (Graepel et al., 2005). In contrast, the reconstruction functions for the set covering machine (Marchand and Shawe-Taylor, 2002) and the decision list machine (Marchand and Sokolova, 2005) need both a compression sequence and a message string. Furthermore, Marchand and Sokolova (2005) provide numerous examples where it is advantageous to have a set $\mathcal{M}(S_{\mathbf{i}})$ of possible messages that depend on the compression sequence $S_{\mathbf{i}}$. In these circumstances, the set of messages can be substantially reduced by using the information contained in $S_{\mathbf{i}}$. We will provide detailed examples below of data-dependent distributions of messages $\mathcal{M}(S_{\mathbf{i}})$.

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 (considered 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 σ taken from a set $\mathcal{M} \stackrel{\text{def}}{=} \mathcal{M}(\emptyset)$. 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. More recently, Catoni (2004) has introduced priors where some data-dependence is allowed. Here, we derive here a new PAC-Bayes theorem for priors that are more natural for sample-compression algorithms. These are priors defined over $I \times \mathcal{M}_S$ for any possible $S \in (\mathcal{X} \times \mathcal{Y})^m$. More precisely, for each $S \in (\mathcal{X} \times \mathcal{Y})^m$, we will only consider priors P_S on $I \times \mathcal{M}_S$ that can be be written as the product

$$P_S(\mathbf{i},\sigma) \;=\; P_I(\mathbf{i}) P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma), \tag{1}$$

where $P_I(\mathbf{i})$ is the prior probability of using the vector $\mathbf{i}$ of indices (defined above) and where $P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)$ is the prior probability of using the message string σ given that we use the compression sequence $S_{\mathbf{i}}$ (i.e., a vector $\mathbf{i}$ with a sequence S). The message string σ could also be a *parameter* chosen from a continuous set $\mathcal{M}(S_{\mathbf{i}})$. In this case, $P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)$ specifies a probability density function. Throughout the paper, a distribution on $I \times \mathcal{M}_S$, prior or posterior, will always mean a distribution that factorizes as Equation 1.

We consider learning algorithms that output a posterior distribution Q on $I \times \mathcal{M}_S$ after observing some training sequence S. The posterior Q has the same form $Q_I(\mathbf{i})Q_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)$ as the one given for the prior P_S but both $Q_I(\mathbf{i})$ and $Q_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)$ can be chosen *after* observing the training data S, that is, they can both depend on S in any way. In contrast, $P_I(\mathbf{i})$ cannot depend on S at all and $P_{\mathcal{M}(S_{\mathbf{i}})}$ can only depend on S through $\mathcal{M}(S_{\mathbf{i}})$. This implies that $P_I(\mathbf{i})$ must be defined *before* observing S and $P_{\mathcal{M}(S_{\mathbf{i}})}$ defined[2] for all possible values of S. Consequently, the set of messages $\mathcal{M}(S_{\mathbf{i}})$ must be defined *a priori* for any compression sequence $S_{\mathbf{i}}$ (we will provide examples in the next section).

Since we do not allow any dependence on S for $P_I(\mathbf{i})$, we cannot discriminate a priori between two vectors of indicies $\mathbf{i}, \mathbf{i}' \in I$ that have same size. Hence, we propose to assign the same prior probability to every vector $\mathbf{i}$ having the same size, that is, we choose

$$P_I(\mathbf{i}) = \zeta(|\mathbf{i}|) \cdot \binom{m}{|\mathbf{i}|}^{-1}, \tag{2}$$

[2] As we will precisely see later, the allowed dependence on $S_{\mathbf{i}}$ of the prior comes from the fact that the empirical risk of the classifiers will be computed only on the examples of S that are not in the compression sequence $S_{\mathbf{i}}$.

where ζ can be any function satisfying $\sum_{d=0}^{m}\zeta(d)=1$. However, since the risk upper bound will deteriorate as we put more weight on classifiers with large compression sizes $|\mathbf{i}|$, it will be preferable to choose a function $\zeta(d)$ that puts more weight on small values of d.

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 $R_{S_{\bar{\mathbf{i}}}}\left(\mathcal{R}(\sigma,S_{\mathbf{i}})\right)$ 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, 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 $S_{\bar{\mathbf{i}}}$.

Given a training sequence S and a distribution Q, and given a new (testing) input example $\mathbf{x}$, a *sample-compressed Gibbs classifier* G_Q chooses randomly $\mathbf{i}$ according to Q_I and then chooses σ according to $Q_{\mathcal{M}(S_{\mathbf{i}})}$ to obtain classifier $\mathcal{R}(\sigma,S_{\mathbf{i}})$ which is then used to determine the class label of $\mathbf{x}$. Therefore, given a training sequence S and a distribution Q, the true risk $R(G_Q)$ of the sample-compressed Gibbs classifier G_Q is given by

$$R(G_Q)=\mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\ \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} R(\sigma,S_{\mathbf{i}})\,.$$

Furthermore, its empirical risk $R_S(G_Q)$ is given by

$$R_S(G_Q)=\mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\ \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\,.$$

Note that these expectations are defined only within the context of a training sequence S.

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

Definition 2 *Given a distribution Q on $I\times\mathcal{M}_S$, we will denote by $\overline{Q}_I$ the distribution defined as*

$$\overline{Q}_I(\mathbf{i})\stackrel{\text{def}}{=}\frac{Q_I(\mathbf{i})}{|\bar{\mathbf{i}}|\mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\frac{1}{|\bar{\mathbf{i}}|}}\quad\forall\mathbf{i}\in I\,,\tag{3}$$

where $|\bar{\mathbf{i}}|\stackrel{\text{def}}{=}m-|\mathbf{i}|$. We will also denote by $\overline{Q}$, the distribution on $I\times\mathcal{M}_S$ given by the product

$$\overline{Q}_I(\mathbf{i})Q_{\mathcal{M}_{S_{\mathbf{i}}}}(\sigma)\,.$$

Furthermore, let

$$d_{\overline{Q}}\stackrel{\text{def}}{=}\mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{Q}_I}|\mathbf{i}|\,.\tag{4}$$

It follows directly from these definitions that

$$\mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\frac{1}{|\bar{\mathbf{i}}|}=\frac{1}{\mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{Q}_I}|\bar{\mathbf{i}}|}=\frac{1}{m-d_{\overline{Q}}}\,.\tag{5}$$

Let $\mathbf{i}_f\stackrel{\text{def}}{=}(1,2,\ldots,m)$ be the (full) vector $\mathbf{i}$ that contains all the m indicies. Since $|\bar{\mathbf{i}}_f|=0$, we might think that $\overline{Q}_I$ is undefined whenever $Q_I(\mathbf{i}_f)>0$. However, we can simply show that definition 2 implies that we must have $\overline{Q}_I(\mathbf{i}_f) = 1$ (and $\overline{Q}_I(\mathbf{i}) = 0\ \forall \mathbf{i} \neq \mathbf{i}_f$) whenever $Q_I(\mathbf{i}_f) > 0$. This claim simply follows from the fact that for all $\mathbf{i}$ we can write

$$|\bar{\mathbf{i}}| \mathop{\mathbf{E}}_{\mathbf{j}\sim Q_I} \frac{1}{|\bar{\mathbf{j}}|} \;=\; Q_I(\mathbf{i}) + |\bar{\mathbf{i}}| \sum_{\mathbf{j}\neq\mathbf{i}} Q_I(\mathbf{j}) \frac{1}{|\bar{\mathbf{j}}|}\,.$$

Consequently, we have

$$\overline{Q}_I(\mathbf{i}_f) \;=\; \frac{Q_I(\mathbf{i}_f)}{|\bar{\mathbf{i}}_f| \mathop{\mathbf{E}}_{\mathbf{j}\sim Q_I} \frac{1}{|\bar{\mathbf{j}}|}} \;=\; \frac{Q_I(\mathbf{i}_f)}{Q_I(\mathbf{i}_f) + |\bar{\mathbf{i}}_f| \sum_{\mathbf{j}\neq\mathbf{i}_f} Q_I(\mathbf{j}) \frac{1}{|\bar{\mathbf{j}}|}} \;=\; 1\,.$$

And for all $\mathbf{i} \neq \mathbf{i}_f$, we have

$$\begin{aligned}
\overline{Q}_I(\mathbf{i}) \;&=\; \frac{Q_I(\mathbf{i})}{|\bar{\mathbf{i}}| \mathop{\mathbf{E}}_{\mathbf{j}\sim Q_I} \frac{1}{|\bar{\mathbf{j}}|}} \;=\; \frac{Q_I(\mathbf{i})}{Q_I(\mathbf{i}) + |\bar{\mathbf{i}}| \sum_{\mathbf{j}\neq\mathbf{i}} Q_I(\mathbf{j}) \frac{1}{|\bar{\mathbf{j}}|}} \\
&\leq\; \frac{Q_I(\mathbf{i})}{Q_I(\mathbf{i}) + |\bar{\mathbf{i}}| Q_I(\mathbf{i}_f) \frac{1}{|\bar{\mathbf{i}}_f|}} \;=\; 0\,,
\end{aligned}$$

which proves the claim.

The next theorem constitutes our main result.

Theorem 3 *For any* $\delta \in (0,1]$, *for any reconstruction function mapping compression sequences and messages to classifiers, for any* $T \in (\mathcal{X}\times\mathcal{Y})^m$ *and for any prior* P_T *on* $I \times \mathcal{M}_T$, *we have*

$$\Pr_{S\sim D^m}\Bigg(\forall Q \text{ on } I\times\mathcal{M}_S\colon\ \mathrm{kl}(R_S(G_Q)\|R(G_Q)) \leq \frac{1}{m-d_{\overline{Q}}}\left[\mathrm{KL}(\overline{Q}\|P_S) + \ln\frac{m+1}{\delta}\right]\Bigg) \geq 1-\delta.$$

Similarly as Theorem 1, Theorem 3 provides both an upper bound and a lower bound on the true risk $R(G_Q)$ based on the empirical risk $R_S(G_Q)$.

Note that

$$\begin{aligned}
\mathrm{KL}(\overline{Q}\|P_S) \;&=\; \mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{Q}_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_\mathbf{i})}} \ln\frac{\overline{Q}_I(\mathbf{i}) Q_{\mathcal{M}(S_\mathbf{i})}(\sigma)}{P_I(\mathbf{i}) P_{\mathcal{M}(S_\mathbf{i})}(\sigma)} \\
&=\; \mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{Q}_I} \ln\frac{\overline{Q}_I(\mathbf{i})}{P_I(\mathbf{i})} \;+\; \mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{Q}_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_\mathbf{i})}} \ln\frac{Q_{\mathcal{M}(S_\mathbf{i})}(\sigma)}{P_{\mathcal{M}(S_\mathbf{i})}} \\
&=\; \mathrm{KL}(\overline{Q}_I\|P_I) \;+\; \mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{Q}_I} \mathrm{KL}(Q_{\mathcal{M}(S_\mathbf{i})}\|P_{\mathcal{M}(S_\mathbf{i})})\,.
\end{aligned}$$

Although we must define *a priori* a continuous family of priors (one prior P_T on $I\times\mathcal{M}_T$ per possible sequence $T \in (\mathcal{X}\times\mathcal{Y})^m$), only the prior on the observed training sequence S will contribute to the bounds.

Theorem 3 is a generalization of Theorem 1 because the latter corresponds to the case where the probability distribution Q has non-zero weight only for $|\mathbf{i}| = 0$. Indeed, in this case we have $\frac{1}{m-d_{\overline{Q}}} = \frac{1}{m}$ and $\overline{Q} = Q$.

Note also that, when Q_I is non-zero only for one compression size $|\mathbf{i}| = d$, we have $\overline{Q}_I = Q_I$ and $d_{\overline{Q}} = d$. Hence, for a stochastic average of sample-compressed classifiers of fixed compression size d, the risk bounds depend only on the "original" posterior Q_I.

More generally, note that $\overline{Q}_I(\mathbf{i})$ is smaller than $Q_I(\mathbf{i})$ for classifiers having a compression size $|\mathbf{i}|$ smaller than the Q-average. This, combined with the fact that $\mathrm{KL}(\overline{Q}\|P_S)$ favors $\overline{Q}$'s close to P_S, implies that there will be a specialization performed by Q on classifiers having small compression sizes. As an example, in the case where $\overline{Q} = P_S$, it is easy to see that Q will put more weight than P_S on "small" classifiers. The specialization suggested by Theorem 3 is therefore stronger than what it would have been if $\mathrm{KL}(Q\|P_S)$ would have been in the risk bound instead of $\mathrm{KL}(\overline{Q}\|P_S)$. Thus, Theorem 3 reinforces Occam's principle of parsimony.

Note also that, since $R(B_Q) \le 2R(G_Q)$, Theorem 3 provides an upper bound for the true risk of the (deterministic) majority vote B_Q. Consider, for example, a majority vote of m classifiers, each having a compression size $|\mathbf{i}| = 1$. In that case, this majority vote uses all the m training examples of S. However, the upper bound given by Theorem 3 will be small (whenever $\mathrm{KL}(\overline{Q}\|P_S)$ and $R_S(G_Q)$ are both small) since $d_{\overline{Q}} = 1$.

The rest of this section is devoted to the proof of Theorem 3. We first provide a lemma about the following quantity.

Definition 4 *Let $S \in (\mathcal{X}\times\mathcal{Y})^m$ and D be a distribution on $\mathcal{X}\times\mathcal{Y}$. 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 $|\bar{\mathbf{i}}|R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})$ errors on $S'_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|}$. Hence, equivalently, we have*

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

Lemma 5 *For any $\delta \in (0,1]$, for any reconstruction function mapping compression sequences and messages to classifiers, for any $T \in (\mathcal{X}\times\mathcal{Y})^m$ and for any prior P_T on $I \times \mathcal{M}_T$, we have*

$$\Pr_{S\sim D^m}\left(\mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)} \le \frac{m+1}{\delta} \right) \ge 1-\delta .$$

Proof First observe that (for any $\mathbf{i} \in I$, $S_{\mathbf{i}} \in (\mathcal{X}\times\mathcal{Y})^{|\mathbf{i}|}$, and $\sigma \in \mathcal{M}(S_{\mathbf{i}})$)

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

Since the expectation over $S_{\bar{\mathbf{i}}}$ is independent of $S_{\mathbf{i}}$, for any P_I and $P_{\mathcal{M}(S_{\mathbf{i}})}$ we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)} &= \mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim D^{|\bar{\mathbf{i}}|}} \frac{1}{B_S(\mathbf{i},\sigma)} \\
&= \mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} m-|\mathbf{i}|+1 \\
&= m-|\mathbf{i}|+1 \le m+1 .
\end{aligned}$$

In the first equation above, note that $\mathcal{M}(S_{\mathbf{i}})$ must be defined for all possible values of $S_{\mathbf{i}}$ since the expectation on the left-hand side is performed for all possible values of S. Note also that the dependence of the prior P on S comes only through $S_{\mathbf{i}}$. Finally, since $\mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)}$ is a non-negative random variable (function of S) having an expectation of at most $m+1$, we can use Markov's inequality to obtain the lemma. ■

The next step is to transform the expectation over P_S into an expectation over Q to obtain the following lemma.

Lemma 6 *For any $\delta \in (0,1]$, for any reconstruction function mapping compression sequences and messages to classifiers, for any $T \in (\mathcal{X} \times \mathcal{Y})^m$ and for any prior P_T on $I \times \mathcal{M}_T$, we have*

$$\Pr_{S\sim D^m}\left(\forall Q \text{ on } I\times\mathcal{M}_S\colon \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{|\overline{\mathbf{i}}|}\ln\frac{1}{B_S(\mathbf{i},\sigma)} \right.$$
$$\left. \leq \frac{1}{m-d_{\overline{Q}}}\left[\mathrm{KL}(\overline{Q}\|P_S)+\ln\frac{m+1}{\delta}\right]\right) \geq 1-\delta\,.$$

Proof Lemma 5 gives us

$$\Pr_{S\sim D^m}\left(\ln\left[\mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)}\right] \leq \ln\frac{m+1}{\delta}\right) \geq 1-\delta\,.$$

Now, for any distribution Q (possibly dependent on S), we have

$$\mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)} = \mathop{\mathbf{E}}_{\mathbf{i}\sim \overline{Q}_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \frac{P_I(\mathbf{i})P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)}{\overline{Q}_I(\mathbf{i})Q_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)}\frac{1}{B_S(\mathbf{i},\sigma)}\,.$$

Since $\ln x$ is concave, we can use Jensen's inequality to obtain

$$\begin{aligned}
\ln\left(\mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)}\right) &\geq \mathop{\mathbf{E}}_{\mathbf{i}\sim \overline{Q}_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \ln\left(\frac{P_I(\mathbf{i})P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)}{\overline{Q}_I(\mathbf{i})Q_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)}\frac{1}{B_S(\mathbf{i},\sigma)}\right)\\
&= \mathop{\mathbf{E}}_{\mathbf{i}\sim \overline{Q}_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \ln\left(\frac{P_I(\mathbf{i})P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)}{\overline{Q}_I(\mathbf{i})Q_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)}\right)\\
&\qquad + \mathop{\mathbf{E}}_{\mathbf{i}\sim \overline{Q}_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \ln\left(\frac{1}{B_S(\mathbf{i},\sigma)}\right)\\
&= -\mathrm{KL}(\overline{Q}\|P_S) + \mathop{\mathbf{E}}_{\mathbf{i}\sim \overline{Q}_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \ln\left(\frac{1}{B_S(\mathbf{i},\sigma)}\right).
\end{aligned}$$

Consequently, we have

$$\Pr_{S\sim D^m}\left(\forall Q \text{ on } I\times\mathcal{M}_S: \mathop{\mathbf{E}}_{\mathbf{i}\sim \overline{Q}_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \ln\left[\frac{1}{B_S(\mathbf{i},\sigma)}\right] \leq \mathrm{KL}(\overline{Q}\|P_S)+\ln\frac{m+1}{\delta}\right) \geq 1-\delta\,.$$

The lemma is obtained from this last equation by using Equations 3, 4, and 5 to transform the expectation with respect to $\overline{Q}_I$ into an expectation with respect to Q_I. ■

We can now prove that Theorem 3 is a direct consequence of Lemma 6, of the convexity of $\mathrm{kl}(q\|p)$, and of a trivial upper-bound on the Binomial.

Proof of Theorem 3

For all non-negative integers n and k such that $k \le n$ and $n \ge 1$, we have

$$\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} \le 1\,.$$

From Definition 4, we then have (for any $\mathbf{i}$, σ, and S)

$$B_S(\mathbf{i},\sigma) \le \left(\frac{R(\sigma,S_\mathbf{i})}{R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})}\right)^{|\bar{\mathbf{i}}|R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})}\left(\frac{1-R(\sigma,S_\mathbf{i})}{1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})}\right)^{|\bar{\mathbf{i}}|-|\bar{\mathbf{i}}|R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})}\,.$$

Consequently, for any $\mathbf{i}$, σ, and S, we have

$$\begin{aligned}\frac{1}{|\bar{\mathbf{i}}|}\ln\frac{1}{B_S(\mathbf{i},\sigma)} &\ge R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})\ln\frac{R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})}{R(\sigma,S_\mathbf{i})}+(1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i}))\ln\frac{1-R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})}{1-R(\sigma,S_\mathbf{i})}\\ &\overset{\mathrm{def}}{=} \mathrm{kl}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})\|R(\sigma,S_\mathbf{i})\right)\,. \qquad (6)\end{aligned}$$

We now exploit the fact that $\mathrm{kl}(q\|p)$ is a convex function of the pair (q,p) of variables. Indeed, from the log-sum inequality (Cover and Thomas, 1991), we can show that for any $(q,p)\in[0,1]\times[0,1]$, any $(r,s)\in[0,1]\times[0,1]$, and any $\alpha\in[0,1]$, we have

$$\mathrm{kl}\big(\alpha q+(1-\alpha)r\|\alpha p+(1-\alpha)s\big) \le \alpha\mathrm{kl}(q\|p)+(1-\alpha)\mathrm{kl}(r\|s)\,.$$

Hence, from Equation 6 and Jensen's inequality applied to $\mathrm{kl}(q\|p)$, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_\mathbf{i})}}\frac{1}{|\bar{\mathbf{i}}|}\ln\frac{1}{B_S(\mathbf{i},\sigma)} &\ge \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_\mathbf{i})}}\mathrm{kl}\left(R_{S_{\bar{\mathbf{i}}}}(\sigma,S_\mathbf{i})\|R(\sigma,S_\mathbf{i})\right)\\ &\ge \mathrm{kl}\left(R_S(G_Q)\|R(G_Q)\right)\,.\end{aligned}$$

Theorem 3 then directly follows from this equation and Lemma 6. ■

5. Learning Algorithms for Stochastic Averages and Majority-Votes of Sample-Compressed Classifiers

There exists numerous learning algorithms for producing a single sample-compressed classifier. The perceptron learning rule, for example, produces a linear classifier which can be reconstructed from the subsequence of training examples that have been used to update the weight vector and the bias of the linear separator (Graepel et al., 2005). This subsequence of examples then constitutes the compression sequence of the sample-compressed classifier and the reconstruction function just consists of the perceptron learning rule executed on the compression sequence. Another example, also studied by Graepel et al. (2005), is the support vector machine (SVM). Here, the compression sequence consists of the set of support vectors and the reconstruction function, again, just consists of running the original learning algorithm on the compression sequence.

Theorem 3 bounds the risk a stochastic average (and the associated majority-vote) of sample-compressed classifiers. Hence, to apply Theorem 3 to algorithms producing a single classifier (as the ones described above), we need to use a posterior Q that has all its weight on a single classifier. In that case, Theorem 3 reduces to Theorem 7 of the next section. However, due to the presence of $\mathrm{KL}(\overline{Q}\|P_S)$ in Theorem 3, and because the prior P_I in Equation 2 gives an equal *a priori* weight to every vector $\mathbf{i}$ having the same number $|\mathbf{i}|$ of indices, Theorem 3 can provide a smaller risk upper-bound for posteriors Q having non-zero weight on several sample-compressed classifiers than for posteriors having all their weight on a single sample-compressed classifier. In short, the guarantee provided by Theorem 3 might be better for a stochastic average of sample-compressed classifiers than for any single sample-compressed classifier. This observation motivates the consideration of learning algorithms for producing posteriors having non-zero weight over several classifiers.

One way to produce a (hopefully) good posterior over several sample-compressed classifiers is to exploit some inherent randomness, or variability, present in the base learning algorithm for single classifiers. The perceptron learning rule is a good example of a learning algorithm that naturally presents some variability that can be exploited. Indeed, the linear classifier produced by the perceptron learning rule is generally very sensitive to the order of the training sequence of examples. Different permutations of the training sequence are likely to produce different linear classifiers. This variability has been exploited by Herbrich et al. (2001) to produce a large-scaled Bayes point machine. If we are not concerned by the space occupied by a large population of classifiers, it is clear from the work of Herbrich et al. (2001) that we could equally well produce a uniformly-weighted majority vote of perceptrons obtained from a large number of permutations of the training sequence.[3] In that case, the stochastic average of these classifiers would represent the typical perceptron that we would obtain by choosing at random a permutation of the training sequence S. We therefore expect that this stochastic Gibbs classifier would be less sensitive to S than any single perceptron obtained from S. Theorem 3 confirms this intuition with an upper-bound on the risk that increases with the amount of the KL-divergence between the posterior and the prior.

Often, the base learning algorithm has no obvious randomness or variability that can be exploited. The SVM provides an obvious example of this type of algorithm since, given any training sequence S, the maximum soft-margin classifier is unique (and the same for any permutation of S). In these cases, a population of distinct classifiers can be obtained by training the base learning algorithm on several training sequences sampled from the bootstrap distribution defined on the original training sequence S; as done in bagging (Breiman, 1996). Another possibility, is to boost (Freund and Schapire, 1997) the base learning algorithm by adaptively re-weighting a distribution defined on the training sequence S.

5.1 Stochastic Averages and Majority-Votes of Set Covering Machines

The perceptron learning rule and the SVM are examples of learning algorithms that produce sample-compressed classifiers $\mathcal{R}(S_\mathbf{i})$ that can be reconstructed *solely* from a compression sequence $S_\mathbf{i}$.

[3] For the case where the training sequence is not linearly-separable, we could simply add a correction to the diagonal of the kernel matrix as was done by Herbrich et al. (2001).

Hence, they are not examples illustrating all the potential of the machinery of data-dependent messages and priors that was developed in Section 4. We therefore present here an example of a learning algorithm, called the set covering machine (SCM) (Marchand and Shawe-Taylor, 2002), that does make use of data-dependent messages to represent sample-compressed classifiers.

As described in Marchand and Shawe-Taylor (2002) and Marchand and Sokolova (2005), a SCM classifier is a conjunction of data-dependent Boolean-valued features.[4] For simplicity, let us limit ourselves to the case where this set of features consists of data-dependent balls and holes; as described in Marchand and Sokolova (2005).

Each such feature g is identified by a *center* $\mathbf{x}_c$ and a radius ρ. Let $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ be any metric. In the case where feature g is a *ball*, its output $g(\mathbf{x})$, for any $\mathbf{x} \in \mathcal{X}$, is given by

$$g(\mathbf{x}) \quad = \quad \begin{cases} 1 & \text{if} \quad d(\mathbf{x}, \mathbf{x}_c) \leq \rho \\ 0 & \text{if} \quad d(\mathbf{x}, \mathbf{x}_c) > \rho \end{cases} \quad \text{(for balls)} \ .$$

When feature g is a *hole*, its output $g(\mathbf{x})$, for any $\mathbf{x} \in \mathcal{X}$, is given by

$$g(\mathbf{x}) \quad = \quad \begin{cases} 0 & \text{if} \quad d(\mathbf{x}, \mathbf{x}_c) \leq \rho \\ 1 & \text{if} \quad d(\mathbf{x}, \mathbf{x}_c) > \rho \end{cases} \quad \text{(for holes)} \ .$$

Each possible pair $(\mathbf{z}_c, \mathbf{z}_b)$ of training examples taken from S defines a feature (ball or hole) that could potentially be included in the (final) conjunction classifier. The first example, $\mathbf{z}_c = (\mathbf{x}_c, y_c)$, defines the center of the feature and the second example, $\mathbf{z}_b = (\mathbf{x}_b, y_b)$, called the *border*, determines its radius $\rho = d(\mathbf{x}_c, \mathbf{x}_b) \pm \varepsilon$; where ε is some very small positive constant chosen *a priori*. To obtain a conjunction consistent with each center of its features, we limit ourselves to the case where a feature centered on a positive example must be a ball and a feature centered on a negative example must be a hole. Moreover, because we want to obtain a conjunction that makes few training errors and that contains a small number of features, it is a good strategy to try to use features that, individually, assigns 0 to a large number of negative examples and to a small number of positive examples.[5] In order to achieve this goal reasonably rapidly and to reduce as far as possible the expected size of the message strings, we will here limit ourselves to conjunctions of features satisfying the following conditions:

1. no two concentric balls or holes can belong to the same conjunction;
2. border points are only chosen among the positive examples of S;
3. the conjunction makes no error on its compression sequence $S_\mathbf{i}$.

The algorithm proposed by Marchand and Shawe-Taylor (2002) greedily constructs conjunctions of balls and holes that respect these conditions. Condition 1 is motivated by the fact that a conjunction of two concentric balls gives the same classifier as the single inner ball (and, symmetrically, a conjunction of two concentric holes can be replaced by the outer hole). Condition 2 follows from our strategy of looking for features that, individually, assign 0 to a large number of negative examples

[4] Here we use the usual convention that the truth values “false” and “true” of Boolean-valued classifiers are mapped, respectively, to the output values of -1 and $+1$ of binary-valued classifiers.

[5] We exploit here some properties of a conjunction. Namely, a conjunction that makes no training errors consists of features that, individually, classify correctly all the positive examples in S. In addition, each negative example of S must be correctly classified by at least one feature in the conjunction.

and to a small number of positive examples. Indeed, a hole, that can assign 0 to one more negative example by increasing its radius, without assigning 0 to an extra positive example, will be a "better" hole to include in the conjunction (a similar observation applies to balls). For Condition 3, we will see, in the next paragraph, how it helps in reducing the expected size of the messages strings. For now, note that, to make each feature consistent with its border point, we have to choose $\rho = d(\mathbf{x}_c, \mathbf{x}_b) + \varepsilon$ for a ball and $\rho = d(\mathbf{x}_c, \mathbf{x}_b) - \varepsilon$ for a hole.

Even under these three conditions, the compression sequence $S_\mathbf{i}$ alone, does not give enough information to reconstruct the conjunction of features. From the previous paragraph, we do know that each negative example in $S_\mathbf{i}$ must be the center of a hole. However, each positive example in $S_\mathbf{i}$ could either be a center (of a ball) or a border point (or both). Hence, we will use messages to identify centers from among the positive examples in $S_\mathbf{i}$. For this task, let $P(S_\mathbf{i})$ denote the set of positive examples among $S_\mathbf{i}$. It follows from Condition 1 that, once we use a message that specifies which are the examples among $P(S_\mathbf{i})$ that are used for centers, we automatically know how many balls to reconstruct from $S_\mathbf{i}$. Moreover, since, by Condition 2, each negative example of $S_\mathbf{i}$ must be the center of one hole, the number of holes to reconstruct from $S_\mathbf{i}$ is equal to the number of negative examples in $S_\mathbf{i}$. What remains to be determined, to specify the (final) conjunction classifier, is the border point in $P(S_\mathbf{i})$ for each center in $S_\mathbf{i}$. Let us now recall that a conjunction that makes no errors on $S_\mathbf{i}$ consists of features that, individually, classify correctly each example in $P(S_\mathbf{i})$. Thus, Condition 3 implies that the border point of each ball center $\mathbf{x}_c$ must be the example in $P(S_\mathbf{i})$ which is located furthest from $\mathbf{x}_c$ and the border point of each hole center $\mathbf{x}'_c$ must be the example in $P(S_\mathbf{i})$ which is located closest from $\mathbf{x}'_c$. Hence, the additional information message $\sigma \in \mathcal{M}(S_\mathbf{i})$ only needs to specify which are the examples in $P(S_\mathbf{i})$ that are centers. For this task, we can simply use a single bit for each example in $P(S_\mathbf{i})$ to indicate if that example is a center. The set $\mathcal{M}(S_\mathbf{i})$ then consists of the set of all such possible bit sequences that we can use, given $S_\mathbf{i}$. The total number $|\mathcal{M}(S_\mathbf{i})|$ of possible messages is then equal to $2^{p(S_\mathbf{i})}$, where $p(S_\mathbf{i}) = |P(S_\mathbf{i})|$.

Now, to use Theorem 3 as a risk bound for a stochastic average (i.e., Gibbs) of SCMs, we need to specify the prior $P_I(\mathbf{i})P_{\mathcal{M}(S_\mathbf{i})}(\sigma)$ over sets of indices and messages. For $P_I(\mathbf{i})$, we use the form proposed in Equation 2. For $P_{\mathcal{M}(S_\mathbf{i})}(\sigma)$, we can simply[6] assign the same probability to each message $\sigma \in \mathcal{M}(S_\mathbf{i})$. In that case we have

$$P_{\mathcal{M}(S_\mathbf{i})}(\sigma) \; = \; 2^{-p(S_\mathbf{i})} \quad \forall \sigma \in \mathcal{M}(S_\mathbf{i}) \, .$$

The task of the learner will then be to produce a posterior $Q_I(\mathbf{i})Q_{\mathcal{M}(S_\mathbf{i})}(\sigma)$ such that, hopefully, $R(G_Q)$ (and thus, indirectly, $R(B_Q)$) will be minimal. For this task we could either bag (Breiman, 1996) or boost (Freund and Schapire, 1997) the SCM learning algorithm proposed by Marchand and Shawe-Taylor (2002). In these cases, the message part of the posterior, $Q_{\mathcal{M}(S_\mathbf{i})}(\sigma)$, will be non-zero only for the particular message that is actually used for each compression sequence $S_\mathbf{i}$ that occurs in the majority-vote of SCMs. Hence, for each such $S_\mathbf{i}$, we have

$$\mathrm{KL}(Q_{\mathcal{M}(S_\mathbf{i})} \| P_{\mathcal{M}(S_\mathbf{i})}) = p(S_\mathbf{i}) \ln 2 \, .$$

Another version of the SCM, called PAC-Bayes SCM, has been proposed recently by Laviolette et al. (2006). In this version each ball radius is specified by a real-valued "message" instead of

[6] Marchand and Sokolova (2005) proposed a different data-dependent set of messages and prior which can give a (slightly) tighter risk bound than the solution we present here. For pedagogical reasons, we have chosen to present here a simpler (and easier to understand) example of $\mathcal{M}(S_\mathbf{i})$ and $P_{\mathcal{M}(S_\mathbf{i})}(\sigma)$.

a border point. Hence, each compression sequence $S_{\mathbf{i}}$ consists only of the centers. Given $S_{\mathbf{i}}$, the message $\boldsymbol{\sigma}$ consists of a $|\mathbf{i}|$-tuple of radius values. Given some large distance R defined *a priori*, the prior is given by

$$P_{\mathcal{M}(S_{\mathbf{i}})}(\boldsymbol{\sigma}) \;=\; \prod_{i\in\mathbf{i}} \frac{1}{R} \quad \forall\sigma_i \in [0,R]\,, \tag{7}$$

and the posterior is given by

$$Q_{\mathcal{M}(S_{\mathbf{i}})}(\boldsymbol{\sigma}) \;=\; \prod_{i\in\mathbf{i}} \frac{1}{b_i - a_i} \quad \forall\sigma_i \in [a_i,b_i] \subseteq [0,R]\,, \tag{8}$$

where each a_i and b_i values are selected by the learner. This gives

$$\mathrm{KL}(Q_{\mathcal{M}(S_{\mathbf{i}})} \| P_{\mathcal{M}(S_{\mathbf{i}})}) = \sum_{i\in\mathbf{i}} \ln\left(\frac{R}{b_i - a_i}\right)\,.$$

A posterior over several SCMs of this type could be constructed by bagging (Breiman, 1996) or boosting (Freund and Schapire, 1997) the soft-greedy algorithm proposed by Laviolette et al. (2006). Theorem 3 then provides an upper bound for the risk of these stochastic averages (and associated majority-votes) of PAC-Bayes SCMs.

6. Single Sample-Compressed Classifiers

In this section, we examine the case when the posterior has all its weight on a single sample-compressed classifier and show that the risk upper-bound for this case is competitive with the currently-known tightest sample-compression risk bounds (Langford, 2005, Laviolette et al., 2005).

Let us examine the case when, given a training sequence S, the (stochastic) sample-compressed Gibbs classifier becomes a deterministic classifier with a posterior having all its weight on a single sample-compressed classifier $\mathcal{R}(S_{\mathbf{i}},\sigma)$. In that case, $\overline{Q}_I = Q_I$, $d_{\overline{Q}} = |\mathbf{i}|$, and $\mathrm{KL}(\overline{Q}\|P_S) = -\ln(P_I(\mathbf{i})P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma))$. Consequently, Lemma 6 gives the following inequality for any prior P_S and any reconstruction function.

$$\Pr_{S\sim D^m}\Bigg(\forall\mathbf{i}\in I, \forall\sigma\in\mathcal{M}(S_{\mathbf{i}})\colon \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\; \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \ln\frac{1}{B_S(\mathbf{i},\sigma)} \le \ln\frac{1}{P_I(\mathbf{i})P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)} + \ln\frac{m+1}{\delta}\Bigg) \ge 1-\delta\,. \tag{9}$$

Now, let us use the binomial distribution

$$\mathrm{Bin}\,(m,k,r) \;\stackrel{\mathrm{def}}{=}\; \binom{m}{k} r^k (1-r)^{m-k}\,,$$

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

$$B_S(\mathbf{i},\sigma) \;=\; \mathrm{Bin}\left(m, mR_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}), R(\sigma,S_{\mathbf{i}})\right)\,.$$

Let us now define the *binomial inversion* as

$$\overline{\mathrm{Bin}}\,(m,k,\delta) \;\stackrel{\mathrm{def}}{=}\; \sup\left\{r\colon \mathrm{Bin}\,(m,k,r) \ge \delta\right\}\,.$$

Equation 9 then gives the following theorem.

Theorem 7 *For any $\delta \in (0,1]$, for any reconstruction function mapping compression sequences and messages to classifiers, for any $T \in (\mathcal{X} \times \mathcal{Y})^m$ and for any prior P_T on $I \times \mathcal{M}_T$, we have*

$$\Pr_{S\sim D^m}\left(\forall \mathbf{i} \in I, \forall \sigma \in \mathcal{M}(S_{\mathbf{i}})\colon\ R(\sigma, S_{\mathbf{i}}) \leq\ \overline{\mathrm{Bin}}\left(m, mR_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}), \frac{P_I(\mathbf{i})P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)\delta}{(m+1)}\right)\right) \geq 1-\delta\,.$$

Let us now compare this risk bound with the tightest currently known sample-compression risk bounds. The bound proposed in Laviolette et al. (2005) generalizes the the bound proposed by Langford (2005) to the case where message strings are also used to identify classifiers. With the current notation, the bound proposed by Laviolette et al. (2005) can be written as

$$\Pr_{S\sim D^m}\left(\forall \mathbf{i} \in I, \forall \sigma \in \mathcal{M}(S_{\mathbf{i}})\colon\ R(\sigma, S_{\mathbf{i}}) \leq\ \overline{\mathrm{BinT}}\left(m, mR_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}), P_I(\mathbf{i})P_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)\delta\right)\right) \geq 1-\delta\,,$$

where, instead of the binomial inversion, we use the binomial *tail* inversion defined as

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

Consequently, for all values of m, k, δ, we have

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

When both m and δ are non zero, the equality is realized only for $k=0$. Hence, the bound of Theorem 7 would be tighter than the bound of Laviolette et al. (2005) if Theorem 7 would hold for δ instead of $\delta/(m+1)$. The bound of Theorem 7 is therefore “competitive” with the currently-known tightest sample-compression risk bound.

Let us now examine the “intermediate” case where Q_I has all its weight on a single vector of indices $\mathbf{i}$ and $Q_{\mathcal{M}(S_{\mathbf{i}})}$ has non-zero weight on several messages $\sigma \in \mathcal{M}(S_{\mathbf{i}})$. In this case we have $\overline{Q} = Q$ and $\mathrm{KL}(Q\|P_S) = -\ln(P_I(\mathbf{i})) + \mathrm{KL}(Q_{\mathcal{M}(S_{\mathbf{i}})}\|P_{\mathcal{M}(S_{\mathbf{i}})})$. Moreover, given a training sequence S of m examples and a vector $\mathbf{i}$ selected by the learner, the Gibbs classifier $G_{Q_{\mathcal{M}(S_{\mathbf{i}})}}$ just chooses randomly according to $Q_{\mathcal{M}(S_{\mathbf{i}})}$ a message σ to classify any new example $\mathbf{x}$ with classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$. Consequently, Theorem 3 reduces to the following corollary.

Corollary 8 *For any $\delta \in (0,1]$, for any reconstruction function mapping compression sequences and messages to classifiers, for any $T \in (\mathcal{X} \times \mathcal{Y})^m$ and for any prior P_T on $I \times \mathcal{M}_T$, we have*

$$\Pr_{S\sim D^m}\Bigg(\forall \mathbf{i} \in I, \forall Q_{\mathcal{M}(S_{\mathbf{i}})}\colon \mathrm{kl}(R_S(G_{Q_{\mathcal{M}(S_{\mathbf{i}})}})\|R(G_{Q_{\mathcal{M}(S_{\mathbf{i}})}})$$

$$\leq\ \frac{1}{m-|\mathbf{i}|}\left[\ln\frac{1}{P_I(\mathbf{i})} + \mathrm{KL}(Q_{\mathcal{M}(S_{\mathbf{i}})}\|P_{\mathcal{M}(S_{\mathbf{i}})}) + \ln\frac{m+1}{\delta}\right]\Bigg) \geq 1-\delta\,.$$

This corollary gives the risk bound proposed in Laviolette et al. (2006) for the PAC-Bayes SCM when the prior is given by Equation 7 and the posterior is given by Equation 8.

7. Compression Sequences that Include Training Errors

In their pioneering work on sample compression, Littlestone and Warmuth (1986) have handled the case of non-zero training errors (also called the “lossy” compression case) by including the training error points in the sample compression sequence $S_{\mathbf{i}}$. Within this methodology, a part of the message string σ is used to indicate which indices in $\mathbf{i}$ are pointing to training error examples. The other indices in $\mathbf{i}$ are then automatically pointing to the training examples actually used for constructing the classifier. We can also use this methodology for deriving another upper-bound on the risk of a stochastic average of sample-compressed classifiers. The resulting upper-bound is generally slightly looser than the one given by Theorem 3 but it has the advantage of being simpler (and easier to interpret). In addition, it becomes slightly *tighter* than the bound of Theorem 3 in the limit of a consistent Gibbs classifier (i.e., when G_Q makes no training errors). We will thus present (and prove) this other upper-bound.

Since each sample-compressed classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$ is still given by $\mathbf{i}$ and σ (once we have a training sequence S), we can still use all the definitions up to Theorem 3. However, $\mathbf{i}$ points also to training errors in S and σ specifies also the indices of $\mathbf{i}$ pointing to training errors. In particular, this implies that we still use posteriors of the form $Q_I(\mathbf{i})Q_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)$ but now $R_S(G_Q)$ is always zero. With this important difference in mind, we have the following theorem.

Theorem 9 *For any* $\delta \in (0,1]$, *for any reconstruction function mapping compression sequences and messages to classifiers, for any* $T \in (\mathcal{X} \times \mathcal{Y})^m$ *and for any prior* P_T *on* $I \times \mathcal{M}_T$, *we have*

$$\Pr_{S\sim D^m}\left(\forall Q \text{ on } I \times \mathcal{M}_S \text{ such that } R_S(G_Q) = 0 : \right.$$
$$\left. R(G_Q) \;\leq\; 1 - \exp\left[\frac{-1}{m - d_{\overline{Q}}}\left(\mathrm{KL}(\overline{Q}\|P_S) + \ln\frac{1}{\delta}\right)\right]\right) \geq 1 - \delta\,.$$

Before we prove this theorem, let us compare it with Theorem 3 in the consistent case (when G_Q makes no training errors). For Theorem 9, this means that no part of the message σ is needed to indicate the indices in $\mathbf{i}$ that point to training errors. Hence, the messages used for the reconstruction function are identical for both theorems in the consistent case. Theorem 3, however, applies for $R_S(G_Q) = 0$. Since

$$\mathrm{kl}(0\|R(G_Q)) = \ln\left(\frac{1}{1 - R(G_Q)}\right),$$

the upper bound of Theorem 3 becomes identical to the bound of Theorem 9 except for the presence of a $\ln(m+1)$ term in the argument of the exponential in Theorem 3. Consequently, the bound of Theorem 9 is slightly tighter in the consistent case.

Proof Since the index vector $\mathbf{i}$, used by classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$, now contains pointers to error points, all the classifiers having non-zero posterior weight will have $R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = 0$. Consequently, instead of using $B_S(\mathbf{i},\sigma)$ (see Definition 4), we will now use $C_S(\mathbf{i},\sigma)$ defined as

$$C_S(\mathbf{i},\sigma) \;\stackrel{\text{def}}{=}\; \frac{1}{(1 - R(\sigma, S_{\mathbf{i}}))^{m-|\mathbf{i}|}} I\left(R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = 0\right).$$

For any $\mathbf{i} \in I$, $S_{\mathbf{i}} \in (\mathcal{X} \times \mathcal{Y})^{|\mathbf{i}|}$, and $\sigma \in \mathcal{M}(S_{\mathbf{i}})$, we have

$$\mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|}} C_S(\mathbf{i},\sigma) \;=\; \mathop{\Pr}_{S_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|}} \big(R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = 0\big) \left[\mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|} | R_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})=0} C_S(\mathbf{i},\sigma) \right] \;=\; 1\,.$$

Then for any P_I and $P_{\mathcal{M}(S_{\mathbf{i}})}$, we have

$$\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} C_S(\mathbf{i},\sigma) \;=\; \mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}}\sim D^{|\bar{\mathbf{i}}|}} C_S(\mathbf{i},\sigma) \;=\; 1\,.$$

Since $\mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} C_S(\mathbf{i},\sigma)$ is a non-negative random variable (function of S) having an expectation of 1, we can use Markov's inequality to obtain

$$\mathop{\Pr}_{S\sim D^m} \left(\mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} C_S(\mathbf{i},\sigma) \;\le\; \frac{1}{\delta} \right) \ge 1-\delta\,.$$

Thus

$$\mathop{\Pr}_{S\sim D^m} \left(\ln\left[\mathop{\mathbf{E}}_{\mathbf{i}\sim P_I} \mathop{\mathbf{E}}_{\sigma\sim P_{\mathcal{M}(S_{\mathbf{i}})}} C_S(\mathbf{i},\sigma) \right] \;\le\; \ln\frac{1}{\delta} \right) \ge 1-\delta\,.$$

Given this last result, we can use the same technique as in the proof of Lemma 6 to convert the expectation over P_S into an expectation over Q. With this technique, we find that, for any prior P_S, we have

$$\mathop{\Pr}_{S\sim D^m} \Bigg(\forall Q \text{ on } I \times \mathcal{M}_S \text{ such that } R_S(G_Q) = 0\colon$$
$$\mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{|\bar{\mathbf{i}}|} \ln C_S(\mathbf{i},\sigma) \;\le\; \frac{1}{m-d_{\overline{Q}}} \left[\mathrm{KL}(\overline{Q}\|P_S) + \ln\frac{1}{\delta} \right] \Bigg) \ge 1-\delta\,. \quad (10)$$

Since the posterior $Q_I(\mathbf{i})Q_{\mathcal{M}(S_{\mathbf{i}})}(\sigma)$ is non-zero only when $R_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = 0$, we have

$$\mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{|\bar{\mathbf{i}}|} \ln C_S(\mathbf{i},\sigma) \;=\; \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \ln\left(\frac{1}{1-R(\sigma,S_{\mathbf{i}})} \right) \;\ge\; \ln\left(\frac{1}{1-R(G_Q)} \right),$$

where the last inequality results from Jensen's inequality applied to the convex function $\ln(1/(1-x))$.

The theorem is obtained by including this last result into Equation 10. ■

8. A PAC-Bayes Theorem for Classifiers that Can Abstain

Many commercial learning systems are producing a “meta-classifier” that consists of an ensemble of rules. In these cases, each rule is basically a classifier that abstains of predicting the class label of an example **x** whenever its premiss (which often consists of a conjunction of Boolean-valued features) is false on **x**. If several rules predict a class label for **x**, the assigned class label will be the one which is predicted by the largest number of rules in the ensemble. When there is a tie, the whole ensemble may abstain or predict the most frequently encountered class in the training sequence S. Hence, an ensemble of rules is just a majority-vote of classifiers that may abstain.

To bound the risk of majority-votes and stochastic averages of classifiers that can abstain, it is “natural” to consider loss functions that may take values $\notin \{0,1\}$. If we limit ourselves to loss functions taking values in the $[0,1]$ interval (including the loss value for abstaining), we can use the risk bound of Theorem 1 of McAllester (2003a). However, that bound can be considerably higher than the trivial upper-bound of 1. A better approach would be to use the bound of Theorem 3.2 of Seeger (2003)—which is valid for classifiers predicting a class among k possible values (for any integer $k > 1$). In this section, we propose to generalize this latter approach to the sample-compression setting. [7] Consequently, we will generalize Theorem 3 for a stochastic average of sample-compressed classifiers that may abstain. As before, the theorem will also apply to the usual data-independent setting in the limit where each classifier is only described by a data-independent message (and an empty compression sequence).

Each classifier h has now three possible outcomes $h(\mathbf{x}) \in \{-1,0,+1\}$ on any $\mathbf{x} \in \mathcal{X}$. Classifier h abstains on $\mathbf{x}$ whenever $h(\mathbf{x}) = 0$. Therefore, each classifier h is now characterized by two different probabilities with respect to the random draws of an example $(\mathbf{x},y) \in \mathcal{X} \times \mathcal{Y}$ (where $\mathcal{Y}$ is still equal to $\{-1,+1\}$). First, we are concerned with the probability $a(h)$ that classifier h abstains on a new example, where

$$a(h) \stackrel{\text{def}}{=} \Pr_{(\mathbf{x},y)\sim D} \left(h(\mathbf{x}) = 0\right) .$$

And we are also concerned with the probability $e(h)$ that classifier h predicts the wrong class label of a new example, where

$$e(h) \stackrel{\text{def}}{=} \Pr_{(\mathbf{x},y)\sim D} \left(h(\mathbf{x}) \neq y \wedge h(\mathbf{x}) \neq 0\right) .$$

The probability that classifier h predicts the correct class label of a new example is then equal to $1 - e(h) - a(h)$. In contrast with the case where classifiers cannot abstain, each classifier is now characterized by a trivalent random variable (instead of a Bernoulli random variable).

The empirical estimates (of these probabilities) on a training sequence S of m examples are defined as

$$\begin{aligned} e_S(h) &\stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} I(h(\mathbf{x}_i) \neq 0) \cdot I(h(\mathbf{x}_i) \neq y_i)\,, \\ a_S(h) &\stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} I(h(\mathbf{x}_i) = 0)\,. \end{aligned}$$

Similarly as before, each classifier $\mathcal{R}(\sigma,S_{\mathbf{i}})$ is described by a compression sequence $S_{\mathbf{i}}$ and a message σ taken from a data-dependent set $\mathcal{M}(S_{\mathbf{i}})$. Given a training sequence S, the prior and the posterior have the same form as before. Moreover, $a(\sigma,S_{\mathbf{i}})$ and $e(\sigma,S_{\mathbf{i}})$ will denote, respectively, the probability of abstaining and the probability of incorrectly predicting a label for classifier $\mathcal{R}(\sigma,S_{\mathbf{i}})$. We will also denote by $a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})$ and $e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})$ the empirical estimates (of the corresponding probabilities) on the subsequence $S_{\bar{\mathbf{i}}}$ of examples in S that are *not* used for constructing classifier $\mathcal{R}(\sigma,S_{\mathbf{i}})$.

[7] We consider here the particular case where each classifier either predicts -1 or +1 or abstains of predicting. The generalization to k classes is straightforward.

The stochastic sample-compressed Gibbs classifier is the same as before. Namely, given a training sequence S and a distribution Q, and given a new (testing) input example $\mathbf{x}$, a sample-compressed Gibbs classifier G_Q chooses randomly $\mathbf{i}$ according to Q_I and then chooses σ according to $Q_{\mathcal{M}(S_{\mathbf{i}})}$ to obtain classifier $\mathcal{R}(\sigma, S_{\mathbf{i}})$ which is then used to determine the class label of $\mathbf{x}$. Therefore, given a training sequence S and a distribution Q, $e(G_Q)$ and $a(G_Q)$ are given by

$$\begin{aligned} e(G_Q) &= \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} e(\sigma, S_{\mathbf{i}})\,, \\ a(G_Q) &= \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} a(\sigma, S_{\mathbf{i}})\,. \end{aligned}$$

Furthermore, their empirical estimates (on a training sequence S of m examples) are given by

$$\begin{aligned} e_S(G_Q) &= \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} e_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})\,, \\ a_S(G_Q) &= \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I} \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} a_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})\,. \end{aligned}$$

Note that these expectations are defined only within the context of a training sequence S.

Given a posterior Q, the re-scaled posterior $\overline{Q}$ is still given by Definition 2.

Theorem 10 *For any $\delta \in (0,1]$, for any reconstruction function mapping compression sequences and messages to classifiers that may abstain, for any $T \in (\mathcal{X}\times\mathcal{Y})^m$ and for any prior P_T on $I \times \mathcal{M}_T$, we have*

$$\Pr_{S\sim D^m}\Bigg(\forall Q \text{ on } I\times\mathcal{M}_S\colon\ \mathrm{kl}(a_S(G_Q), e_S(G_Q) \| a(G_Q), e(G_Q)) \\ \leq \frac{1}{m-d_{\overline{Q}}}\left[\mathrm{KL}(\overline{Q}\|P_S) + \ln\frac{(m+1)(m+2)}{2\delta}\right]\Bigg) \geq 1-\delta,$$

where $\mathrm{kl}(q_1,q_2\|p_1,p_2)$ is the Kullback-Leibler divergence between the distributions of two trivalent random variables Y_q and Y_p with probabilities (q_1,q_2) and (p_1,p_2) respectively. Hence,

$$\mathrm{kl}(q_1,q_2\|p_1,p_2) \;=\; q_1\ln\frac{q_1}{p_1} + q_2\ln\frac{q_2}{p_2} + (1-q_1-q_2)\ln\frac{1-q_1-q_2}{1-p_1-p_2}\,.$$

Proof The proof essentially parallels the one given for Theorem 3 but with the important difference that the empirical estimates $e_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})$ and $a_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}})$ are distributed according to a trinomial (instead of a binomial) with respect to the random draws of S. Consequently, we now define $B_S(\mathbf{i},\sigma)$ as

$$B_S(\mathbf{i},\sigma) \stackrel{\text{def}}{=} \binom{|\bar{\mathbf{i}}|}{|\bar{\mathbf{i}}|a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}\binom{|\bar{\mathbf{i}}|(1-a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))}{|\bar{\mathbf{i}}|e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}(a(\sigma,S_{\mathbf{i}}))^{|\bar{\mathbf{i}}|a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})} \\ (e(\sigma,S_{\mathbf{i}}))^{|\bar{\mathbf{i}}|e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}(1-a(\sigma,S_{\mathbf{i}})-e(\sigma,S_{\mathbf{i}}))^{|\bar{\mathbf{i}}|(1-a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})-e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))}\,.$$

Then, for any $\mathbf{i} \in I$, $S_{\mathbf{i}} \in (\mathcal{X} \times \mathcal{Y})^{|\mathbf{i}|}$ and $\sigma \in \mathcal{M}(S_{\mathbf{i}})$, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|}} \frac{1}{B_S(\mathbf{i},\sigma)} &= \sum_{j=0}^{|\bar{\mathbf{i}}|} \sum_{k=0}^{|\bar{\mathbf{i}}|-j} \Pr_{S_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|}} \left(a_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = \frac{j}{|\bar{\mathbf{i}}|} \wedge e_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = \frac{k}{|\bar{\mathbf{i}}|} \right) \\
&\quad \mathop{\mathbf{E}}_{\left[S_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|} \,\middle|\, \begin{array}{l} a_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = \frac{j}{|\bar{\mathbf{i}}|}, \\ e_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = \frac{k}{|\bar{\mathbf{i}}|} \end{array} \right]} \left(\frac{1}{B_S(\mathbf{i},\sigma)} \right) \\
&= \sum_{j=0}^{|\bar{\mathbf{i}}|} \sum_{k=0}^{|\bar{\mathbf{i}}|-j} \frac{\Pr_{S_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|}} \left(a_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = \frac{j}{|\bar{\mathbf{i}}|} \wedge e_{S_{\bar{\mathbf{i}}}}(\sigma, S_{\mathbf{i}}) = \frac{k}{|\bar{\mathbf{i}}|} \right)}{\binom{|\bar{\mathbf{i}}|}{j} \binom{|\bar{\mathbf{i}}|-j}{k} (a(\sigma, S_{\mathbf{i}}))^j \, (e(\sigma, S_{\mathbf{i}}))^k \, (1 - a(\sigma, S_{\mathbf{i}}) - e(\sigma, S_{\mathbf{i}}))^{|\bar{\mathbf{i}}|-j-k}} \\
&= \frac{(|\bar{\mathbf{i}}|+1)(|\bar{\mathbf{i}}|+2)}{2} \,.
\end{aligned}$$

Since the expectation over $S_{\bar{\mathbf{i}}}$ is independent of $S_{\mathbf{i}}$, for any P_I and $P_{\mathcal{M}(S_{\mathbf{i}})}$ we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{S \sim D^m} \mathop{\mathbf{E}}_{\mathbf{i} \sim P_I} \mathop{\mathbf{E}}_{\sigma \sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)} &= \mathop{\mathbf{E}}_{\mathbf{i} \sim P_I} \mathop{\mathbf{E}}_{S_{\mathbf{i}} \sim D^{|\mathbf{i}|}} \mathop{\mathbf{E}}_{\sigma \sim P_{\mathcal{M}(S_{\mathbf{i}})}} \mathop{\mathbf{E}}_{S_{\bar{\mathbf{i}}} \sim D^{|\bar{\mathbf{i}}|}} \frac{1}{B_S(\mathbf{i},\sigma)} \\
&= \mathop{\mathbf{E}}_{\mathbf{i} \sim P_I} \mathop{\mathbf{E}}_{S_{\mathbf{i}} \sim D^{|\mathbf{i}|}} \mathop{\mathbf{E}}_{\sigma \sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{(|\bar{\mathbf{i}}|+1)(|\bar{\mathbf{i}}|+2)}{2} \\
&= \frac{(|\bar{\mathbf{i}}|+1)(|\bar{\mathbf{i}}|+2)}{2} \le \frac{(m+1)(m+2)}{2} \,.
\end{aligned}$$

Since $\mathop{\mathbf{E}}_{\mathbf{i} \sim P_I} \mathop{\mathbf{E}}_{\sigma \sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)}$ is a non-negative random variable (function of S) having an expectation of at most $(m+1)(m+2)/2$, we can use Markov's inequality to obtain

$$\Pr_{S \sim D^m} \left(\mathop{\mathbf{E}}_{\mathbf{i} \sim P_I} \mathop{\mathbf{E}}_{\sigma \sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)} \le \frac{(m+1)(m+2)}{2\delta} \right) \ge 1 - \delta \,.$$

Hence,

$$\Pr_{S \sim D^m} \left(\ln \left[\mathop{\mathbf{E}}_{\mathbf{i} \sim P_I} \mathop{\mathbf{E}}_{\sigma \sim P_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{B_S(\mathbf{i},\sigma)} \right] \le \ln \frac{(m+1)(m+2)}{2\delta} \right) \ge 1 - \delta \,.$$

Similarly as in the proof of Lemma 6, we can transform the expectation over P_S into an expectation over Q to obtain

$$\begin{aligned}
\Pr_{S \sim D^m} \Bigg(\forall Q \text{ on } I \times \mathcal{M}_S \colon \mathop{\mathbf{E}}_{\mathbf{i} \sim Q_I} \mathop{\mathbf{E}}_{\sigma \sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{|\bar{\mathbf{i}}|} \ln \frac{1}{B_S(\mathbf{i},\sigma)} \\
\le \frac{1}{m - d_{\overline{Q}}} \left[\mathrm{KL}(\overline{Q} \| P_S) + \ln \frac{(m+1)(m+2)}{2\delta} \right] \Bigg) \ge 1 - \delta \,. \quad (11)
\end{aligned}$$

For all non-negative integers n, j, k such that $j + k \le n$ and $n \ge 1$, we have

$$\binom{n}{j} \binom{n-j}{k} \left(\frac{j}{n} \right)^j \left(\frac{k}{n} \right)^k \left(1 - \frac{j}{n} - \frac{k}{n} \right)^{n-j-k} \le 1 \,.$$

Then, for any $\mathbf{i}$, σ, and S, we have

$$\begin{aligned} B_S(\mathbf{i},\sigma) \;\leq\; & \left(\frac{a(\sigma,S_{\mathbf{i}})}{a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}\right)^{|\bar{\mathbf{i}}|a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})} \cdot \left(\frac{e(\sigma,S_{\mathbf{i}})}{e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}\right)^{|\bar{\mathbf{i}}|e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})} \\ & \cdot \left(\frac{1-a(\sigma,S_{\mathbf{i}})-e(\sigma,S_{\mathbf{i}})}{1-a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})-e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}\right)^{|\bar{\mathbf{i}}|(1-a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})-e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))} . \end{aligned}$$

Consequently, for any $\mathbf{i}$, σ, and S, we have

$$\begin{aligned} \frac{1}{|\bar{\mathbf{i}}|}\ln\frac{1}{B_S(\mathbf{i},\sigma)} \quad & \geq \quad a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\ln\frac{a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{a(\sigma,S_{\mathbf{i}})} + e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\ln\frac{e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{e(\sigma,S_{\mathbf{i}})} \\ & \quad\; +(1-a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})-e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}))\ln\frac{1-a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})-e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})}{1-a(\sigma,S_{\mathbf{i}})-e(\sigma,S_{\mathbf{i}})} \\ & \overset{\text{def}}{=} \quad \mathrm{kl}\left(a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}),e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\|a(\sigma,S_{\mathbf{i}}),e(\sigma,S_{\mathbf{i}})\right) . \end{aligned}$$

Since $\mathrm{kl}(q_1,q_2\|p_1,p_2)$ is a convex function of (q_1,q_2,p_1,p_2), we can use Jensen's inequality to obtain

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\; \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \frac{1}{|\bar{\mathbf{i}}|}\ln\frac{1}{B_S(\mathbf{i},\sigma)} \quad & \geq \quad \mathop{\mathbf{E}}_{\mathbf{i}\sim Q_I}\; \mathop{\mathbf{E}}_{\sigma\sim Q_{\mathcal{M}(S_{\mathbf{i}})}} \mathrm{kl}\left(a_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}}),e_{S_{\bar{\mathbf{i}}}}(\sigma,S_{\mathbf{i}})\|a(\sigma,S_{\mathbf{i}}),e(\sigma,S_{\mathbf{i}})\right) \\ & \geq \quad \mathrm{kl}\left(a_S(G_Q),e_S(G_Q)\|a(G_Q),e(G_Q)\right) . \end{aligned}$$

The theorem directly follows from this equation and Equation 11. ■

All PAC-Bayes theorems, including the above, are statements about probabilities and their empirical estimates on a sample—no loss functions are involved. Here, let us consider that the loss $\ell(h(\mathbf{x}),y)$ suffered by classifier h on an example $(\mathbf{x},y)$ is given by

$$\ell(h(\mathbf{x}),y) \;=\; \begin{cases} 1 & \text{if} \quad h(\mathbf{x})\neq y \,\wedge\, h(\mathbf{x})\neq 0 \\ 0 & \text{if} \quad h(\mathbf{x})=y \\ c & \text{if} \quad h(\mathbf{x})=0\,, \end{cases}$$

for some constant $c\in[0,1]$. Since the *risk* $R(h)$ of classifier h is its expected loss, we have

$$R(h) \;=\; e(h)+c\cdot a(h)\,.$$

Hence, for a sample-compressed Gibbs classifier G_Q, we have $R(G_Q)=e(G_Q)+c\cdot a(G_Q)$ (with a similar relation for the empirical estimates on a training sequence S). Therefore, to upper-bound $R(G_Q)$, we simply need to find the largest value of $e(G_Q)+c\cdot a(G_Q)$ permitted by Theorem 10 given that we know $e_S(G_Q)$ and $a_S(G_Q)$. Consequently, Theorem 10 has the following corollary.

Corollary 11 *For any $\delta\in(0,1]$, for any reconstruction function mapping compression sequences and messages to classifiers that may abstain, for any $T\in(\mathcal{X}\times\mathcal{Y})^m$ and for any prior P_T on $I\times\mathcal{M}_T$, we have*

$$\begin{aligned} \Pr_{S\sim D^m}\Bigg(\forall Q \text{ on } I\times\mathcal{M}_S\colon R(G_Q) \;\leq\; \sup\Bigg\{ & e+ca \;\Big|\; \mathrm{kl}(a_S(G_Q),e_S(G_Q)\|a,e) \\ & \leq\; \frac{1}{m-d_{\overline{Q}}}\left[\mathrm{KL}(\overline{Q}\|P_S)+\ln\frac{(m+1)(m+2)}{2\delta}\right]\Bigg\}\Bigg) \geq 1-\delta. \end{aligned}$$

To upper-bound the risk of the majority-vote B_Q with Theorem 10, we need to redefine the risk $R(B_Q)$ (in terms of the loss function ℓ defined above) and related it to $e(G_Q)$ and $a(G_Q)$. For this task, let us adopt the convention that $B_Q(\mathbf{x})$ abstains of predicting the label of $\mathbf{x}$ whenever the Q-weight of classifiers predicting $+1$ is equal to the Q-weight of classifiers predicting -1 (this includes the case when all the classifiers having non-zero posterior weight abstain).

Similarly as our definition of $e(G_Q)$, let $e(B_Q)$ denote the probability that B_Q predicts incorrectly the label of $\mathbf{x}$ on a random draw of $(\mathbf{x},y)$. Furthermore, let $e_{(\mathbf{x},y)}(G_Q)$ denote the probability that G_Q predicts incorrectly the label of $(\mathbf{x},y)$ and let $a_{(\mathbf{x},y)}(G_Q)$ denote the probability that G_Q abstains on $(\mathbf{x},y)$, that is,

$$\begin{aligned} e_{(\mathbf{x},y)}(G_Q) &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} I(h(\mathbf{x}) \neq y \wedge h(\mathbf{x}) \neq 0) \\ a_{(\mathbf{x},y)}(G_Q) &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} I(h(\mathbf{x}) = 0)\,. \end{aligned}$$

Similarly, let $e_{(\mathbf{x},y)}(B_Q) = 1$ iff B_Q predicts incorrectly the label of $(\mathbf{x},y)$. Therefore,

$$e_{(\mathbf{x},y)}(B_Q) = 1 \iff e_{(\mathbf{x},y)}(G_Q) > \frac{1 - a_{(\mathbf{x},y)}(G_Q)}{2}\,.$$

Hence,

$$\begin{aligned} e(B_Q) &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} e_{(\mathbf{x},y)}(B_Q) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I\left(e_{(\mathbf{x},y)}(G_Q) > \frac{1 - a_{(\mathbf{x},y)}(G_Q)}{2} \right) \\ &= \mathop{\Pr}_{(\mathbf{x},y)\sim D} \left(2e_{(\mathbf{x},y)}(G_Q) + a_{(\mathbf{x},y)}(G_Q) > 1 \right) \\ &< 2e(G_Q) + a(G_Q)\,, \end{aligned}$$

where, for the last line, we have used Markov's inequality for the non-negative random variable $2e_{(\mathbf{x},y)}(G_Q) + a_{(\mathbf{x},y)}(G_Q)$ with expectation $2e(G_Q) + a(G_Q)$.

Since $R(B_Q) = e(B_Q) + c \cdot a(B_Q)$, we can obtain an upper bound on $R(B_Q)$ by upper-bounding $a(B_Q)$. However,

$$a(B_Q) = \mathop{\Pr}_{(\mathbf{x},y)\sim D} \left(e_{(\mathbf{x},y)}(G_Q) = \frac{1 - a_{(\mathbf{x},y)}(G_Q)}{2} \right)\,.$$

Since Theorem 10 gives non control on the domain of $(e(G_Q), a(G_Q))$ that is bounded with high probability, we cannot use it to find a tight upper-bound for $a(B_Q)$. Therefore, since the loss c of abstaining is at most 1, we will simply use

$$\begin{aligned} R(B_Q) &\leq e(B_Q) + a(B_Q) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I\left(e_{(\mathbf{x},y)}(G_Q) \geq \frac{1 - a_{(\mathbf{x},y)}(G_Q)}{2} \right) \\ &= \mathop{\Pr}_{(\mathbf{x},y)\sim D} \left(2e_{(\mathbf{x},y)}(G_Q) + a_{(\mathbf{x},y)}(G_Q) \geq 1 \right) \leq 2e(G_Q) + a(G_Q)\,, \end{aligned}$$

where we have, once again, used Markov's inequality. [8] Consequently, we have the following corollary to bound $R(B_Q)$.

[8] We might think that this upper bound for $R(B_Q)$ is worse than the the upper bound of $2R(G_Q)$ for classifiers that cannot abstain. However, the two upper bounds coincides whenever the cost c of abstaining is $1/2$ or, equivalently, if we force the abstaining classifiers to predict and if their predictions are correct with probability $1/2$.

Corollary 12 *For any* $\delta \in (0,1]$, *for any reconstruction function mapping compression sequences and messages to classifiers that may abstain, for any* $T \in (\mathcal{X} \times \mathcal{Y})^m$ *and for any prior* P_T *on* $I \times \mathcal{M}_T$, *we have*

$$\Pr_{S \sim D^m}\left(\forall Q \textit{ on } I \times \mathcal{M}_S \colon R(B_Q) \;\leq\; \sup\left\{ 2e + a \;\middle|\; \mathrm{kl}(a_S(G_Q), e_S(G_Q) \| a, e) \right.\right.$$
$$\left.\left. \leq\; \frac{1}{m - d_{\overline{Q}}} \left[\mathrm{KL}(\overline{Q} \| P_S) + \ln \frac{(m+1)(m+2)}{2\delta} \right] \right\} \right) \geq 1 - \delta\,.$$

Note that the values of e and a for which the supremum in Corollary 11 and 12 are attained are generally not upper bounds of both $e(G_Q)$ and $a(G_Q)$. Consequently, the risk bound given by Corollary 11 and 12 are tighter than those we would have obtained by bounding $e(G_Q)$ and $a(G_Q)$ separately.

8.1 Reduced Coulomb Energy Networks

Corollaries 8 and 11 can be used to bound the risk of stochastic averages and majority-votes of sample-compressed classifiers that can abstain. The reduced Coulomb energy (RCE) network (see Reilly et al. 1982 and Duda et al. 2000) provides a simple example of such a majority-vote. Indeed, a RCE network is basically a majority-vote of single balls. As for the SCM case, each ball is described by a training example called a *center* $(\mathbf{x}_c, y_c)$, and another training example called a *border* $(\mathbf{x}_b, y_b)$. Given any metric d, the output $h(\mathbf{x})$ on any input example $\mathbf{x}$ of a ball is given by y_c if $d(\mathbf{x}, \mathbf{x}_c) \leq d(\mathbf{x}, \mathbf{x}_b)$, otherwise (if $d(\mathbf{x}, \mathbf{x}_c) > d(\mathbf{x}, \mathbf{x}_b)$) it *abstains* of predicting a class label. Hence, each sample-compressed classifier has here a compression sequence $S_\mathbf{i}$ of only two examples. Given $S_\mathbf{i}$, the message string (which consists here of a single bit) just specifies which of the two examples of $S_\mathbf{i}$ is the center.

Consequently, the prior $P_I(\mathbf{i})$ will be non-zero only for $|\mathbf{i}| = 2$ and distributed uniformly over all pairs of (distinct) indices. The posterior $Q_I(\mathbf{i})$ will also be non-zero only for $|\mathbf{i}| = 2$ but only balls selected by the RCE network learning algorithm, described in Reilly et al. (1982) and Duda et al. (2000), will give pairs of indices of non-zero posterior weight. The message-part of the prior, $P_{\mathcal{M}(S_\mathbf{i})}(\sigma)$, assigns equal probability to the two possible single-bit messages and the message-part of the posterior, $Q_{\mathcal{M}(S_\mathbf{i})}(\sigma)$, assigns a weight of one to the single-bit message that is actually used with the two-example compression sequence $S_\mathbf{i}$. With this form for the prior and the posterior, Corollary 8 provides a tight bound for the risk of the stochastic average G_Q. However, the empirical error rate $e_S(G_Q)$ may be large on some S for simple classifiers that are constructed from only two examples in the RCE network. Hence, since $e_S(G_Q)$ may be large, Corollary 11 may only provide a loose bound for the majority-vote B_Q due to the looseness involved in upper-bounding $e(B_Q)$ by $2e(G_Q) + a(G_Q)$.

9. 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. We have emphasized that many learning algorithms are producing classifiers that are well-described within this setting. We have seen that the PAC-Bayes theorem for the sample-compression setting 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 Laviolette et al. (2005). The PAC-Bayes risk bound of Theorem 3 is, however, valid for sample-compressed Gibbs classifiers with arbitrary posteriors. Moreover, we have seen that the risk bound supports the strategy of randomizing the predictions over several sample-compressed classifiers instead of predicting with a single sample-compressed classifier. Indeed, 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, to obtain a performance guarantee for many "rule-based systems" and RCE networks, we have generalized the PAC-Bayes theorem to the case where each sample-compressed classifier in the ensemble can abstain of predicting a class label.

Since the risk bounds derived in this paper are tight for stochastic averages of classifiers, 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. However, given an ensemble of classifiers, we usually prefer to predict with the majority-vote B_Q instead of the stochastic average G_Q. In these cases, the PAC-Bayesian guarantee for B_Q only comes indirectly through the inequality $R(B_Q) \leq 2R(G_Q)$ (for ensemble of classifiers that cannot abstain). This is clearly inappropriate for many extensively-used learning algorithms, such as Adaboost (Freund and Schapire, 1997), that produce majority-votes having a large underlying $R(G_Q)$ and a very small $R(B_Q)$. Finding better guarantees in these circumstances, along the lines proposed by Lacasse et al. (2007) and Germain et al. (2007), is therefore an important open problem in machine learning.

Acknowledgments

Work supported by NSERC Discovery grants 262067 and 122405.

References

Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. Thecnical report, Université Paris 6, 2004.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*, chapter 12. Wiley, 1991.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2000.

Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

Pascal Germain, Alexandre Lacasse, Francois Laviolette, and Mario Marchand. A pac-bayes risk bound for general loss functions. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59:55–76, 2005.

Ralf Herbrich, Thore Graepel, and Colin Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, 2001.

Alexandre Lacasse, Francois Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

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

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 423–430. MIT Press, Cambridge, MA, 2003.

François Laviolette and Mario Marchand. PAC-Bayes risk bounds for sample-compressed Gibbs classifiers. *Proceedings of the 22nth International Conference on Machine Learning (ICML 2005)*, pages 481–488, 2005.

François Laviolette, Mario Marchand, and Mohak Shah. Margin-sparsity trade-off for the set covering machine. *Proceedings of the* 16^{th} *European Conference on Machine Learning (ECML 2005); Lecture Notes in Artificial Intelligence*, 3720:206–217, 2005.

François Laviolette, Mario Marchand, and Mohak Shah. A PAC-Bayes approach to the set covering machine. *Proceedings of the 2005 conference on Neural Information Processing Systems (NIPS 2005)*, 2006.

Nicholas Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, Santa Cruz, CA, 1986.

Mario Marchand and John Shawe-Taylor. The set covering machine. *Journal of Machine Learning Reasearch*, 3:723–746, 2002.

Mario Marchand and Marina Sokolova. Learning with decision lists of data-dependent features. *Journal of Machine Learning Reasearch*, 6:427–451, 2005.

David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37:355–363, 1999.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003a.

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

Douglas L. Reilly, Leon N. Cooper, and Charles Elbaum. A neural model for category learning. *Biological Cybernetics*, 45:35–41, 1982.

Ronald L. Rivest. Learning decision lists. *Machine Learning*, 2:229–246, 1987.

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

Matthias Seeger. Bayesian gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations. *PhD Thesis, University of Edinburgh*, 2003.

Leslie G. Valiant. A theory of the learnable. *Communications of the Association of Computing Machinery*, 27(11):1134–1142, November 1984.