PAC-Bayesian Analysis for a two-step Hierarchical Multiview Learning Approach

Anil Goyal[1,2] Emilie Morvant[1] Pascal Germain[3] Massih-Reza Amini[2]

[1] Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d'Optique Graduate School,
Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France

[2] Univ. Grenoble Alps, Laboratoire d'Informatique de Grenoble, AMA,
Centre Equation 4, BP 53, F-38041 Grenoble Cedex 9, France

[3] Département d'informatique de l'ENS, École Normale Supérieure,
CNRS, PSL Research University, 75005 Paris, France
INRIA

July 13, 2017

Abstract

We study a two-level multiview learning with more than two views under the PAC-Bayesian framework. This approach, sometimes referred as late fusion, consists in learning sequentially multiple view-specific classifiers at the first level, and then combining these view-specific classifiers at the second level. Our main theoretical result is a generalization bound on the risk of the majority vote which exhibits a term of diversity in the predictions of the view-specific classifiers. From this result it comes out that controlling the trade-off between diversity and accuracy is a key element for multiview learning, which complements other results in multiview learning. Finally, we experiment our principle on multiview datasets extracted from the Reuters RCV1/RCV2 collection.

1 Introduction

With the ever-increasing observations produced by more than one source, multiview learning has been expanding over the past decade, spurred by the seminal work of Blum and Mitchell [1998] on co-training. Most of the existing methods try to combine multimodal information, either by directly merging the views or by combining models learned from the different views[1] [Snoek et al., 2005], in order to produce a model more reliable for the considered task. Our goal is to propose a theoretically grounded criteria to "correctly" combine the views. With this in mind we propose to study multiview learning through the PAC-Bayesian framework (introduced in [McAllester, 1999]) that allows to derive generalization bounds for models that are expressed as a combination over a set of voters. When faced with learning from one view, the PAC-Bayesian theory assumes a prior distribution over the voters involved in the combination, and aims at learning—from the learning sample—a posterior distribution that leads to a well-performing combination expressed as a weighted majority vote. In this paper we extend the PAC-Bayesian theory to multiview with more than two views. Concretely, given a set of view-specific classifiers, we define a hierarchy of posterior and prior distributions over the views, such that *(i)* for each view v, we consider prior P_v and posterior Q_v distributions over each view-specific voters' set, and *(ii)* a prior π and a posterior ρ distribution over the set of views (see Figure 1), respectively called hyper-prior and hyper-posterior[2]. In this way, our proposed approach encompasses the one of Amini et al. [2009] that considered uniform distribution to combine the view-specific classifiers' predictions. Moreover, compared to the PAC-Bayesian work of Sun et al. [2016], we are interested here to the more general and natural case of multiview learning with more than two views. Note also that Lecué and Rigollet [2014] proposed a non-PAC-Bayesian theoretical analysis of a combination of voters (called Q-Aggregation) that is able to take into account a prior and a posterior distribution but in a single-view setting.

Our theoretical study also includes a notion of disagreement between all the voters, allowing to take into account a notion of diversity between them which is known as a key element in multiview learning [Kuncheva, 2004, Chapelle

[1] The fusion of descriptions, *resp.* of models, is sometimes called Early Fusion, *resp.* Late Fusion.

[2] Our notion of hyper-prior and hyper-posterior distributions is different than the one proposed for lifelong learning [Pentina and Lampert, 2014], where they basically consider hyper-prior and hyper-posterior over the set of possible priors: The prior distribution P over the voters' set is viewed as a random variable.

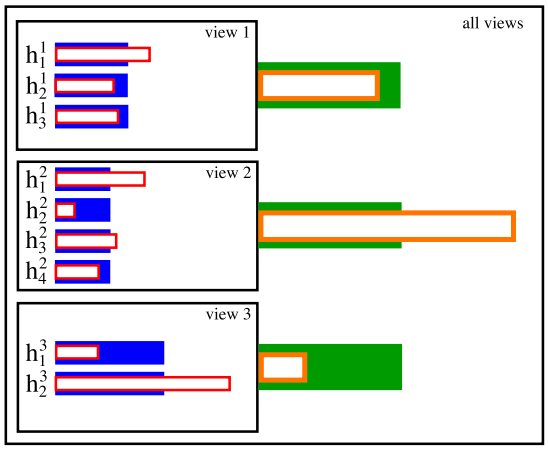

Figure 1: Example of the multiview distributions hierarchy with 3 views. For all views $v \in \{1, 2, 3\}$, we have a set of voters $\mathcal{H}_v = \{h_1^v, \ldots, h_{n_v}^v\}$ on which we consider prior P_v view-specific distribution (in blue), and we consider a hyper-prior π distribution (in green) over the set of 3 views. The objective is to learn a posterior Q_v (in red) view-specific distributions and a hyper-posterior ρ distribution (in orange) leading to a good model. The length of a rectangle represents the weight (or probability) assigned to a voter or a view.

et al., 2010, Maillard and Vayatis, 2009, Amini et al., 2009]. Finally, we empirically evaluate a two-level learning approach on the Reuters RCV1/RCV2 corpus to show that our analysis is sound.

In the next section, we recall the general PAC-Bayesian setup, and present PAC-Bayesian *expectation bounds*—while most of the usual PAC-Bayesian bounds are *probabilistic bounds*. In Section 3, we then discuss the problem of multiview learning, adapting the PAC-Bayesian expectation bounds to the specificity of the two-level multiview approach. In Section 4, we discuss the relation between our analysis and previous works. Before concluding in Section 6, we present experimental results obtained on a collection of the Reuters RCV1/RCV2 corpus in Section 5.

2 The Single-View PAC-Bayesian Theorem

In this section, we state a *new* general mono-view PAC-Bayesian theorem, inspired by the work of Germain et al. [2015], that we extend to multiview learning in Section 3.

2.1 Notations and Setting

We consider binary classification tasks on data drawn from a fixed yet unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a d-dimensional input space and $\mathcal{Y} = \{-1, +1\}$ the label/output set. A learning algorithm is provided with a training sample of m examples denoted by $S = \{(x_i, y_i)\}_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$, that is assumed to be independently and identically distributed (*i.i.d.*) according to $\mathcal{D}$. The notation $\mathcal{D}^m$ stands for the distribution of such a m-sample, and $\mathcal{D}_\mathcal{X}$ for the marginal distribution on $\mathcal{X}$. We consider a set $\mathcal{H}$ of classifiers or voters such that $\forall h \in \mathcal{H},\ h : \mathcal{X} \to \mathcal{Y}$. In addition, PAC-Bayesian approach requires a prior distribution P over $\mathcal{H}$ that models *a priori* belief on the voters from $\mathcal{H}$ before the observation of the learning sample S. Given $S \sim \mathcal{D}^m$, the learner objective is then to find a posterior distribution Q over $\mathcal{H}$ leading to an accurate Q-weighted majority vote $B_Q(x)$ defined as

$$B_Q(x) = \text{sign}\left[\mathop{\mathbb{E}}_{h\sim Q} h(x)\right].$$

In other words, one wants to learn Q over $\mathcal{H}$ such that it minimizes the true risk $R_\mathcal{D}(B_Q)$ of $B_Q(x)$:

$$R_\mathcal{D}(B_Q) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathbb{1}_{[B_Q(x)\neq y]}\,,$$

where $\mathbb{1}_{[\pi]} = 1$ if predicate π holds, and 0 otherwise. However, a PAC-Bayesian generalization bound does not directly focus on the risk of the deterministic Q-weighted majority vote B_Q. Instead, it upper-bounds the risk of the stochastic Gibbs classifier G_Q, which predicts the label of an example x by drawing h from $\mathcal{H}$ according to the posterior distribution Q and predicts $h(x)$. Therefore, the true risk $R_D(G_Q)$ of the Gibbs classifier on a data

distribution $\mathcal{D}$, and its empirical risk $R_S(G_Q)$ estimated on a sample $S \sim \mathcal{D}^m$ are respectively given by

$$R_{\mathcal{D}}(G_Q) \;=\; \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{h\sim Q} \mathbb{1}_{[h(x)\neq y]}\,,$$

$$\text{and} \quad R_S(G_Q) \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim Q} \mathbb{1}_{[h(x_i)\neq y_i]}\,.$$

The above Gibbs classifier is closely related to the Q-weighted majority vote B_Q. Indeed, if B_Q misclassifies $x \in \mathcal{X}$, then at least half of the classifiers (under measure Q) make an error on x. Therefore, we have

$$R_{\mathcal{D}}(B_Q) \leq 2R_{\mathcal{D}}(G_Q). \tag{1}$$

Thus, an upper bound on $R_{\mathcal{D}}(G_Q)$ gives rise to an upper bound on $R_{\mathcal{D}}(B_Q)$. Other tighter relations exist [Langford and Shawe-Taylor, 2002, Lacasse et al., 2006, Germain et al., 2015], such as the so-called C-Bound [Lacasse et al., 2006] that involves the *expected disagreement* $d_{\mathcal{D}}(Q)$ between all the pair of voters, and that can be expressed as follows (when $R_{\mathcal{D}}(G_Q) \leq \frac{1}{2}$):

$$R_{\mathcal{D}}(B_Q) \;\leq\; 1 - \frac{\left(1-2R_{\mathcal{D}}(G_Q)\right)^2}{1-2d_{\mathcal{D}}(Q)}\,, \tag{2}$$

$$\text{where} \quad d_{\mathcal{D}}(Q) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbb{E}}_{(h,h')\sim Q^2} \mathbb{1}_{[h(x)\neq h'(x)]}\,.$$

Moreover, Germain et al. [2015] have shown that the Gibbs classifier's risk can be rewritten in terms of $d_{\mathcal{D}}(Q)$ and *expected joint error* $e_{\mathcal{D}}(Q)$ between all the pair of voters as

$$R_{\mathcal{D}}(G_Q) \;=\; \frac{1}{2}d_{\mathcal{D}}(Q) + e_{\mathcal{D}}(Q)\,, \tag{3}$$

$$\text{where} \quad e_{\mathcal{D}}(Q) \;=\; \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \mathop{\mathbb{E}}_{(h,h')\sim Q^2} \mathbb{1}_{[h(x)\neq y]}\, \mathbb{1}_{[h'(x)\neq y]}\,.$$

It is worth noting that from multiview learning standpoint where the notion of diversity among voters is known to be important [Amini et al., 2009, Maillard and Vayatis, 2009, Sun et al., 2016, Atrey et al., 2010, Kuncheva, 2004], Equations (2) and (3) directly capture the trade-off between diversity and accuracy. Indeed, $d_{\mathcal{D}}(Q)$ involves the diversity between voters [Morvant et al., 2014], while $e_{\mathcal{D}}(Q)$ takes into account the errors. Note that the principle of controlling the trade-off between diversity and accuracy through the C-bound of Equation (2) has been exploited by Laviolette et al. [2011] and Roy et al. [2016] to derive well-performing PAC-Bayesian algorithms that aims at minimizing it. For our experiments in Section 5, we make use of CqBoost [Roy et al., 2016]—one of these algorithms—for multiview learning.

Last but not least, PAC-Bayesian generalization bounds take into account the given prior distribution P on $\mathcal{H}$ through the Kullback-Leibler divergence between the learned posterior distribution Q and P:

$$\mathrm{KL}(Q\|P) \;=\; \mathop{\mathbb{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}\,.$$

2.2 A New PAC-Bayesian Theorem as an Expected Risk Bound

In the following we introduce a new variation of the general PAC-Bayesian theorem of Germain et al. [2009, 2015]; it takes the form of an upper bound on the "deviation" between the true risk $R_{\mathcal{D}}(G_Q)$ and empirical risk $R_S(G_Q)$ of the Gibbs classifier, according to a convex function $D{:}[0,1]\times[0,1]\to\mathbb{R}$. While most of the PAC-Bayesian bounds are probabilistic bounds, we state here an *expected risk bound*. More specifically, Theorem 1 below is a tool to upper-bound $\mathbb{E}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{Q_S})$—where Q_S is the posterior distribution outputted by a given learning algorithm after observing the learning sample S—while PAC-Bayes usually bounds $R_{\mathcal{D}}(G_Q)$ uniformly for all distribution Q, but with high probability over the draw of $S \sim \mathcal{D}^m$. Since by definition posterior distributions are data dependent, this different point of view on PAC-Bayesian analysis has the advantage to involve an expectation over all the possible learning samples (of a given size) in bounds itself.

Theorem 1. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any set of voters $\mathcal{H}$, for any prior distribution P on $\mathcal{H}$, for any convex function $D : [0,1]\times[0,1]\to\mathbb{R}$, we have*

$$D\Big(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_S(G_{Q_S}), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{Q_S})\Big) \leq \frac{1}{m}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathrm{KL}(Q_S\|P) + \ln\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{h\sim P} e^{m\, D(R_S(h),R_{\mathcal{D}}(h))}\right)\right],$$

where $R_{\mathcal{D}}(h)$ and $R_S(h)$ are respectively the true and the empirical risks of individual voters.

Similarly to Germain et al. [2009, 2015], by selecting a well-suited deviation function D and by upper-bounding $\mathbb{E}_S \mathbb{E}_h e^{m\, D(R_S(h), R_{\mathcal{D}}(h))}$, we can prove the *expected bound* counterparts of the classical PAC-Bayesian theorems of McAllester [1999], Seeger [2002], Catoni [2007]. The proof presented below borrows the straightforward proof technique of Bégin et al. [2016]. Interestingly, this approach highlights that the expectation bounds are obtained simply by replacing the *Markov inequality* by the *Jensen inequality* (respectively Theorems 5 and 6, in Appendix).

Proof of Theorem 1. The last three inequalities below are obtained by applying Jensen's inequality on the convex function D, the change of measure inequality [as stated by Bégin et al., 2016, Lemma 3], and Jensen's inequality on the concave function $\ln$.

$$\begin{aligned} mD\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_S(G_{Q_S}), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{Q_S})\right) &= mD\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim Q_S} R_S(h), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim Q_S} R_{\mathcal{D}}(h)\right) \\ &\le \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim Q_S} mD\left(R_S(h), R_{\mathcal{D}}(h)\right) \\ &\le \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\mathrm{KL}(Q_S\|P) + \ln\left(\mathop{\mathbb{E}}_{h\sim P} e^{m\, D(R_S(h), R_{\mathcal{D}}(h))}\right)\right] \\ &\le \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathrm{KL}(Q_S\|P) + \ln\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{h\sim P} e^{m\, D(R_S(h), R_{\mathcal{D}}(h))}\right). \end{aligned}$$

□

Since the C-bound of Equation (2) involves the expected disagreement $d_{\mathcal{D}}(Q)$, we also derive below the expected bound that upper-bounds the deviation between $\mathbb{E}_{S\sim\mathcal{D}^m} d_S(Q_S)$ and $\mathbb{E}_{S\sim\mathcal{D}^m} d_{\mathcal{D}}(Q_S)$ under a convex function D. Theorem 2 can be seen as the *expectation* version of probabilistic bounds over $d_S(Q_S)$ proposed by Lacasse et al. [2006], Germain et al. [2015].

Theorem 2. *For any distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$, for any set of voters $\mathcal{H}$, for any prior distribution P on $\mathcal{H}$, for any convex function $D : [0,1]\times[0,1]\to\mathbb{R}$, we have*

$$D\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_S(Q_S), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_{\mathcal{D}}(Q_S)\right) \le \frac{2}{m}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathrm{KL}(Q_S\|P) + \ln\sqrt{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{(h,h')\sim P^2} e^{mD(d_S(h,h'), d_{\mathcal{D}}(h,h'))}}\right],$$

where $d_{\mathcal{D}}(h,h') = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\, \mathbb{1}_{[h(x)\neq h'(x)]}$ is the disagreement of voters h and h' on the distribution $\mathcal{D}$, and $d_S(h,h')$ is its empirical counterpart.

Proof. First, we apply the exact same steps as in the proof of Theorem 1:

$$\begin{aligned} mD\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_S(Q_S), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_{\mathcal{D}}(Q_S)\right) &= mD\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{(h,h')\sim Q_S^2} d_S(h,h'), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{(h,h')\sim Q_S^2} d_{\mathcal{D}}(h,h')\right) \\ &\vdots \\ &\le \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathrm{KL}(Q_S^2\|P^2) + \ln \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{(h,h')\sim P^2} e^{mD\left(d_S(h,h'), d_{\mathcal{D}}(h,h')\right)}. \end{aligned}$$

Then, we use the fact that $\mathrm{KL}(Q_S^2\|P^2) = 2\,\mathrm{KL}(Q_S\|P)$ [see Germain et al., 2015, Theorem 25]. □

In the following we provide an extension of this PAC-Bayesian framework to multiview learning with more than two views.

3 Multiview PAC-Bayesian Approach

3.1 Notations and Setting

We consider binary classification problems where the multiview observations $\mathbf{x} = (x^1, \ldots, x^V)$ belong to a multiview input set $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_V$, where $V \ge 2$ is the number of views of not-necessarily the same dimension. We denote $\mathcal{V}$ the set of the V views. In binary classification, we assume that examples are pairs $(\mathbf{x}, y)$, with $y \in \mathcal{Y} = \{-1, +1\}$, drawn according to an unknown distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$. To model the two-level multiview approach, we follow the next setting. For each view $v \in \mathcal{V}$, we consider a view-specific set $\mathcal{H}_v$ of voters $h : \mathcal{X}_v \to \mathcal{Y}$, and a prior distribution P_v on $\mathcal{H}_v$. Given a *hyper-prior* distribution π over the views $\mathcal{V}$, and a multiview learning sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{D})^m$, our PAC-Bayesian learner objective is twofold: *(i)* finding a posterior

distribution Q_v over $\mathcal{H}_v$ for all views $v \in \mathcal{V}$; *(ii)* finding a *hyper-posterior* distribution ρ on the set of views $\mathcal{V}$. This hierarchy of distributions is illustrated by Figure 1. The learned distributions express a multiview weighted majority vote B^{MV}_ρ defined as

$$B^{\text{MV}}_\rho(\mathbf{x}) = \text{sign}\left[\mathop{\mathbb{E}}_{v\sim\rho}\mathop{\mathbb{E}}_{h\sim Q_v} h(x^v)\right].$$

Thus, the learner aims at constructing the posterior and hyper-posterior distributions that minimize the true risk $R_{\mathcal{D}}(B^{\text{MV}}_\rho)$ of the multiview weighted majority vote:

$$R_{\mathcal{D}}(B^{\text{MV}}_\rho) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathbb{1}_{[B^{\text{MV}}_\rho(\mathbf{x})\neq y]}.$$

As pointed out in Section 2, the PAC-Bayesian approach deals with the risk of the stochastic Gibbs classifier G^{MV}_ρ defined as follows in our multiview setting, and that can be rewritten in terms of *expected disagreement* $d^{\text{MV}}_{\mathcal{D}}(\rho)$ and *expected joint error* $e^{\text{MV}}_{\mathcal{D}}(\rho)$:

$$\begin{aligned}
R_{\mathcal{D}}(G^{\text{MV}}_\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\mathop{\mathbb{E}}_{v\sim\rho}\mathop{\mathbb{E}}_{h\sim Q_v} \mathbb{1}_{[h(x^v)\neq y]} \\
&= \tfrac{1}{2}\, d^{\text{MV}}_{\mathcal{D}}(\rho) + e^{\text{MV}}_{\mathcal{D}}(\rho)\,, \qquad (4)\\
\text{where}\quad d^{\text{MV}}_{\mathcal{D}}(\rho) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}}\mathop{\mathbb{E}}_{v\sim\rho}\mathop{\mathbb{E}}_{v'\sim\rho}\mathop{\mathbb{E}}_{h\sim Q_v}\mathop{\mathbb{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq h'(x^{v'})]}, \\
\text{and}\quad e^{\text{MV}}_{\mathcal{D}}(\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\mathop{\mathbb{E}}_{v\sim\rho}\mathop{\mathbb{E}}_{v'\sim\rho}\mathop{\mathbb{E}}_{h\sim Q_v}\mathop{\mathbb{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq y]}\mathbb{1}_{[h'(x^{v'})\neq y]}.
\end{aligned}$$

Obviously, the empirical counterpart of the Gibbs classifier's risk $R_{\mathcal{D}}(G^{\text{MV}}_\rho)$ is

$$\begin{aligned}
R_S(G^{\text{MV}}_\rho) &= \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{v\sim\rho}\mathop{\mathbb{E}}_{h\sim Q_v} \mathbb{1}_{[h(x^v_i)\neq y_i]} \\
&= \frac{1}{2} d^{\text{MV}}_S(\rho) + e^{\text{MV}}_S(\rho)\,,
\end{aligned}$$

where $d^{\text{MV}}_S(\rho)$ and $e^{\text{MV}}_S(\rho)$ are respectively the empirical estimations of $d^{\text{MV}}_{\mathcal{D}}(\rho)$ and $e^{\text{MV}}_{\mathcal{D}}(\rho)$ on the learning sample S. As in the single-view PAC-Bayesian setting, the multiview weighted majority vote B^{MV}_ρ is closely related to the stochastic multiview Gibbs classifier G^{MV}_ρ, and a generalization bound for G^{MV}_ρ gives rise to a generalization bound for B^{MV}_ρ. Indeed, it is easy to show that $R_{\mathcal{D}}(B^{\text{MV}}_\rho) \leq 2R_{\mathcal{D}}(G^{\text{MV}}_\rho)$, meaning that an upper bound over $R_{\mathcal{D}}(G^{\text{MV}}_\rho)$ gives an upper bound for the majority vote. Moreover the C-Bound of Equation (2) can be extended to our multiview setting by Lemma 1 below. Equation (5) is a straightforward generalization of the single-view C-bound of Equation (2). Afterward, Equation (6) is obtained by rewriting $R_{\mathcal{D}}(G^{\text{MV}}_\rho)$ as the ρ-average of the risk associated to each view, and lower-bounding $d^{\text{MV}}_{\mathcal{D}}(\rho)$ by the ρ-average of the disagreement associated to each view.

Lemma 1. *Let $V \geq 2$ be the number of views. For all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior ρ distribution, if $R_{\mathcal{D}}(G^{\text{MV}}_\rho) < \frac{1}{2}$, then we have*

$$\begin{aligned}
R_{\mathcal{D}}(B^{\text{MV}}_\rho) &\leq 1 - \frac{\left(1 - 2R_{\mathcal{D}}(G^{\text{MV}}_\rho)\right)^2}{1 - 2d^{\text{MV}}_{\mathcal{D}}(\rho)} \qquad (5)\\
&\leq 1 - \frac{\left(1 - 2\,\mathbb{E}_{v\sim\rho} R_{\mathcal{D}}(G_{Q_v})\right)^2}{1 - 2\,\mathbb{E}_{v\sim\rho} d_{\mathcal{D}}(Q_v)}\,. \qquad (6)
\end{aligned}$$

Proof. Equation (5) follows from the Cantelli-Chebyshev's inequality (Theorem 7, in Appendix). To prove Equation (6), we first notice that in the binary setting where $y \in \{-1, 1\}$ and $h : \mathcal{X} \to \{-1, 1\}$, we have $\mathbb{1}_{[h(x^v)\neq y]} = \frac{1}{2}(1 - y\, h(x^v))$, and

$$\begin{aligned}
R_{\mathcal{D}}(G^{\text{MV}}_\rho) &= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\mathop{\mathbb{E}}_{v\sim\rho}\mathop{\mathbb{E}}_{h\sim Q_v} \mathbb{1}_{[h(x^v)\neq y]} \\
&= \frac{1}{2}\left(1 - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\mathop{\mathbb{E}}_{v\sim\rho}\mathop{\mathbb{E}}_{h\sim Q_v} y\, h(x^v)\right) \\
&= \mathop{\mathbb{E}}_{v\sim\rho} R_{\mathcal{D}}(G_{Q^v})\,.
\end{aligned}$$

Moreover, we have

$$\begin{aligned} d_{\mathcal{D}}^{\text{MV}}(\rho) &= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{v'\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \mathop{\mathbb{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq h'(x^{v'})]} \\ &= \frac{1}{2}\left(1 - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{v'\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \mathop{\mathbb{E}}_{h\sim Q_{v'}} h(x^v) \times h'(x^{v'})\right) \\ &= \frac{1}{2}\left(1 - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \left[\mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} h(x^v)\right]^2\right). \end{aligned}$$

From Jensen's inequality (Theorem 6, in Appendix) it comes

$$\begin{aligned} d_{\mathcal{D}}^{\text{MV}}(\rho) &\geq \frac{1}{2}\left(1 - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbb{E}}_{v\sim\rho} \left[\mathop{\mathbb{E}}_{h\sim Q_v} h(x^v)\right]^2\right) \\ &= \mathop{\mathbb{E}}_{v\sim\rho}\left[\frac{1}{2}\left(1 - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \left[\mathop{\mathbb{E}}_{h\sim Q_v} h(x^v)\right]^2\right)\right] \\ &= \mathop{\mathbb{E}}_{v\sim\rho} d_{\mathcal{D}}(Q_v)\,. \end{aligned}$$

By replacing $R_{\mathcal{D}}(G_\rho^{\text{MV}})$ and $d_{\mathcal{D}}^{\text{MV}}(\rho)$ in Equation (5), we obtain

$$1 - \frac{\left(1 - 2R_{\mathcal{D}}(G_\rho^{\text{MV}})\right)^2}{1 - 2d_{\mathcal{D}}^{\text{MV}}(\rho)} \leq 1 - \frac{\left(1 - 2\,\mathbb{E}_{v\sim\rho}\, R_{\mathcal{D}}(G_{Q^v})\right)^2}{1 - 2\,\mathbb{E}_{v\sim\rho}\, d_{\mathcal{D}}(Q_v)}\,. \qquad \square$$

Similarly than for the mono-view setting, Equations (4) and (5) suggest that a good trade-off between the risk of the Gibbs classifier G_ρ^{MV} and the disagreement $d_{\mathcal{D}}^{\text{MV}}(\rho)$ between pairs of voters will lead to a well-performing majority vote. Equation (6) exhibits the role of diversity among the views thanks to the disagreement's expectation over the views $\mathbb{E}_{v\sim\rho}\, d_{\mathcal{D}}(Q_v)$.

3.2 General Multiview PAC-Bayesian Theorems

Now we state our general PAC-Bayesian theorem suitable for the above multiview learning setting with a two-level hierarchy of distributions over views (or voters). A key step in PAC-Bayesian proofs is the use of a *change of measure inequality* [McAllester, 2003], based on the Donsker-Varadhan inequality [Donsker and Varadhan, 1975]. Lemma 2 below extends this tool to our multiview setting.

Lemma 2. *For any set of priors $\{P_v\}_{v=1}^V$ and any set of posteriors $\{Q_v\}_{v=1}^V$, for any hyper-prior distribution π on views $\mathcal{V}$ and hyper-posterior distribution ρ on $\mathcal{V}$, and for any measurable function $\phi : \mathcal{H}_v \to \mathbb{R}$, we have*

$$\mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \phi(h) \leq \mathop{\mathbb{E}}_{v\sim\rho} \text{KL}(Q_v\|P_v) + \text{KL}(\rho\|\pi) + \ln\left(\mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} e^{\phi(h)}\right).$$

Proof. We have

$$\begin{aligned} \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \phi(h) &= \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \ln e^{\phi(h)} \\ &= \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \ln\left(\frac{Q_v(h)}{P_v(h)}\frac{P_v(h)}{Q_v(h)}e^{\phi(h)}\right) \\ &= \mathop{\mathbb{E}}_{v\sim\rho}\left[\mathop{\mathbb{E}}_{h\sim Q_v} \ln\left(\frac{Q_v(h)}{P_v(h)}\right) + \mathop{\mathbb{E}}_{h\sim Q_v} \ln\left(\frac{P_v(h)}{Q_v(h)}e^{\phi(h)}\right)\right]. \end{aligned}$$

According to the Kullback-Leibler definition, we have

$$\mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \phi(h) = \mathop{\mathbb{E}}_{v\sim\rho}\left[\text{KL}(Q_v\|P_v) + \mathop{\mathbb{E}}_{h\sim Q_v} \ln\left(\frac{P_v(h)}{Q_v(h)}e^{\phi(h)}\right)\right].$$

By applying Jensen's inequality (Theorem 6, in Appendix) on the concave function $\ln$, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{v\sim\rho} \mathop{\mathbb{E}}_{h\sim Q_v} \phi(h) &\leq \mathop{\mathbb{E}}_{v\sim\rho}\left[\text{KL}(Q_v\|P_v) + \ln\left(\mathop{\mathbb{E}}_{h\sim P_v} e^{\phi(h)}\right)\right] \\ &= \mathop{\mathbb{E}}_{v\sim\rho} \text{KL}(Q_v\|P_v) + \mathop{\mathbb{E}}_{v\sim\rho} \ln\left(\frac{\rho(v)}{\pi(v)}\frac{\pi(v)}{\rho(v)} \mathop{\mathbb{E}}_{h\sim P_v} e^{\phi(h)}\right) \\ &= \mathop{\mathbb{E}}_{v\sim\rho} \text{KL}(Q_v\|P_v) + \text{KL}(\rho\|\pi) + \mathop{\mathbb{E}}_{v\sim\rho} \ln\left(\frac{\pi(v)}{\rho(v)} \mathop{\mathbb{E}}_{h\sim P_v} e^{\phi(h)}\right). \end{aligned}$$

Finally, we apply again the Jensen inequality (Theorem 6) on $\ln$ to obtain the lemma. $\square$

Based on Lemma 2, the following theorem can be seen as a generalization of Theorem 1 to multiview. Note that we still rely on a general convex function $D : [0,1] \times [0,1] \to \mathbb{R}$, that measures the "deviation" between the empirical disagreement/joint error and the true risk of the Gibbs classifier.

Theorem 3. *Let $V \geq 2$ be the number of views. For any distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of prior distributions $\{P_v\}_{v=1}^V$, for any hyper-prior distribution π over $\mathcal{V}$, for any convex function $D : [0,1] \times [0,1] \to \mathbb{R}$, we have*

$$D\Big(\tfrac{1}{2} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_S^{\text{MV}}(\rho_S) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e_S^{\text{MV}}(\rho_S), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{\rho_S}^{\text{MV}})\Big) \leq \frac{1}{m}\Bigg[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\rho_S} \text{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \text{KL}(\rho_S\|\pi) + \ln\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right)\Bigg].$$

Proof. We follow the same steps as in Theorem 1 proof.

$$\begin{aligned}
&mD\Big(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_S(G_{\rho_S}^{\text{MV}}), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{\rho_S}^{\text{MV}})\Big)\\
&= mD\Big(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\rho_S} \mathop{\mathbb{E}}_{h\sim Q_{v,S}} R_S(h), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\rho_S} \mathop{\mathbb{E}}_{h\sim Q_{v,S}} R_{\mathcal{D}}(h)\Big)\\
&\leq \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\rho_S} \mathop{\mathbb{E}}_{h\sim Q_{v,S}} mD\left(R_S(h), R_{\mathcal{D}}(h)\right)\\
&\leq \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\Bigg[\mathop{\mathbb{E}}_{v\sim\rho_S} \text{KL}(Q_{v,S}\|P_v) + \text{KL}(\rho_S\|\pi) + \ln\left(\mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right)\Bigg],
\end{aligned}$$

where the last inequality is obtained using Lemma 2. After distributing the expectation of $S \sim \mathcal{D}^m$, the final statement follows from Jensen's inequality (Theorem 6)

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \ln\left(\mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right) \leq \ln\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right),$$

and from Equation (3): $R_S(G_{\rho_S}^{\text{MV}}) = \frac{1}{2} d_S^{\text{MV}}(\rho_S) + e_S^{\text{MV}}(\rho_S)$. □

It is interesting to compare this generalization bound to Theorem 1. The main difference relies on the introduction of view-specific prior and posterior distributions, which mainly leads to an additional term $\mathbf{E}_{v\sim\rho} \text{KL}(Q_v\|P_v)$, expressed as the expectation of the view-specific Kullback-Leibler divergence term over the views $\mathcal{V}$ according to the hyper-posterior distribution ρ. We also introduce the empirical disagreement allowing us to directly highlight the presence of the diversity between voters and between views. As Theorem 1, Theorem 3 provides a tool to derive PAC-Bayesian generalization bounds for a multiview supervised learning setting. Indeed, by making use of the same trick as Germain et al. [2009, 2015], the generalization bounds can be derived from Theorem 3 by choosing a suitable convex function D and upper-bounding $\mathbb{E}_S \mathbb{E}_v \mathbb{E}_h e^{m D(R_S(h),R_{\mathcal{D}}(h))}$. We provide the specialization to the three most popular PAC-Bayesian approaches McAllester [1999], Catoni [2007], Seeger [2002], Langford [2005] in the next section.

Following the same approach, we can obtain a mutiview bound for the expected disagreement.

Theorem 4. *Let $V \geq 2$ be the number of views. For any distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of prior distributions $\{P_v\}_{v=1}^V$, for any hyper-prior distribution π over $\mathcal{V}$, for any convex function $D : [0,1] \times [0,1] \to \mathbb{R}$, we have*

$$\begin{aligned}
&D\Big(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_S^{\text{MV}}(\rho_S), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_{\mathcal{D}}^{\text{MV}}(\rho_S)\Big)\\
&\leq \frac{2}{m}\Bigg[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\rho_S} \text{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \text{KL}(\rho_S\|\pi) + \ln\sqrt{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{(h,h')\sim P^2} e^{mD(d_S(h,h'),d_{\mathcal{D}}(h,h'))}}\Bigg].
\end{aligned}$$

Proof. The result is obtained straightforwardly by following the proof steps of Theorem 3, using the disagreement instead of the Gibbs risk. Then, similarly at what we have done to obtain Theorem 2, we substitute $\text{KL}(Q_{v,S}^2\|P_v^2)$ by $2\,\text{KL}(Q_{v,S}\|P_v)$, and $\text{KL}(\rho_S^2\|\pi^2)$ by $2\,\text{KL}(\rho_S\|\pi)$. □

3.3 Specialization of our Theorem to the Classical Approaches

In this section, we provide specialization of our multiview theorem to the most popular PAC-Bayesian approaches [McAllester, 1999, Catoni, 2007, Seeger, 2002, Langford, 2005]. To do so, we follow the same principles as Germain et al. [2009, 2015].

3.3.1 A McAllester-Like Theorem

We derive here the specialization of our multiview PAC-Bayesian theorem to the McAllester [2003]'s point of view.

Corollary 1. *Let $V \geq 2$ be the number of views. For any distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of prior distributions $\{P_v\}_{v=1}^V$, for any hyper-prior distribution π over $\mathcal{V}$, we have*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G^{\text{MV}}_{\rho_S}) \leq \frac{1}{2}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d^{\text{MV}}_S(\rho_S) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e^{\text{MV}}_S(\rho_S) + \sqrt{\frac{\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\rho_S} \text{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\text{KL}(\rho_S\|\pi) + \ln\frac{2\sqrt{m}}{\delta}}{2m}}\,.$$

Proof. To prove the above result, we apply Theorem 3 with $D(a,b) = 2(a-b)^2$.
Then, we upper-bound $\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}$. According to Pinsker's inequality, we have

$$D(a,b) \;\leq\; \text{kl}(a,b) \;=\; a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}.$$

By considering $R_S(h)$ as a random variable which follows a binomial distribution of m trials with a probability of success $R(h)$, we obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{m\,D(R_S(h),R_{\mathcal{D}}(h))} &\leq \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{m\ \text{kl}(R_S(h),R_{\mathcal{D}}(h))}\\
&= \mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\frac{R_S(h)}{R_{\mathcal{D}}(h)}\right]^{mR_S(h)}\left[\frac{1-R_S(h)}{1-R_{\mathcal{D}}(h)}\right]^{m(1-R_S(h))}\\
&= \mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v}\sum_{k=0}^{m}\Pr_{S\sim\mathcal{D}^m}\left[R_S(h)=\tfrac{k}{m}\right]\left[\frac{k/m}{R_{\mathcal{D}}(h)}\right]^{k}\left[\frac{1-k/m}{1-R_{\mathcal{D}}(h)}\right]^{m-k}\\
&= \sum_{k=0}^{m}\binom{m}{k}\left[\frac{k}{m}\right]^{k}\left[1-\frac{k}{m}\right]^{m-k}\\
&\leq 2\sqrt{m}\,.
\end{aligned}$$

□

3.3.2 A Catoni-Like Theorem

To derive a generalization bound with the Catoni [2007]'s point of view—given a convex function $\mathcal{F}$ and a real number $C > 0$—we define the measure of deviation between the empirical disagreement/joint error and the true risk as $D(a,b) = \mathcal{F}(b) - C\,a$ [Germain et al., 2009, 2015]. We obtain the following generalization bound.

Corollary 2. *Let $V \geq 2$ be the number of views. For any distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of prior distributions $\{P_v\}_{v=1}^V$, for any hyper-prior distributions π over $\mathcal{V}$, for all $C > 0$, we have:*

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G^{\text{MV}}_{\rho_S}) \;\leq\; \frac{1}{1-e^{-C}}\Bigg(1-\exp\Bigg[-\Bigg(C\left(\frac{1}{2}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d^{\text{MV}}_S(\rho_S) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e^{\text{MV}}_S(\rho_S)\right)+\\
\frac{1}{m}\Big[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\rho_S}\text{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\text{KL}(\rho_S\|\pi) + \ln\tfrac{1}{\delta}\Big]\Bigg)\Bigg]\Bigg)
\end{aligned}$$

Proof. The result comes from Theorem 3 by taking $D(a,b) = \mathcal{F}(b) - Ca$, for a convex $\mathcal{F}$ and $C > 0$, and by upper-bounding $\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{mD(R_S(h),R_{\mathcal{D}}(h))}$. We consider $R_S(h)$ as a random variable following a binomial distribution of m trials with a probability of success $R(h)$. We have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{m\,D(R_S(h),R_{\mathcal{D}}(h))} &= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{m\,\mathcal{F}(R_{\mathcal{D}}(h)-C\,m\,R_S(h))}\\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{m\,\mathcal{F}(R_{\mathcal{D}}(h))}\sum_{k=0}^{m}\Pr_{S\sim(\mathcal{D})^m}\left(R_S(h)=\frac{k}{m}\right)e^{-Ck}\\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{m\mathcal{F}(R_{\mathcal{D}}(h))}\sum_{k=0}^{m}\tbinom{m}{k}R_{\mathcal{D}}(h)^k(1-R_{\mathcal{D}}(h))^{m-k}e^{-Ck}\\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{v\sim\pi}\mathop{\mathbb{E}}_{h\sim P_v} e^{m\mathcal{F}(R_{\mathcal{D}}(h))}\big(R_{\mathcal{D}}(h)\,e^{-\,C}+(1-R_{\mathcal{D}}(h))\big)^m.
\end{aligned}$$

The corollary is obtained with

$$\mathcal{F}(p) = \ln \frac{1}{(1 - p[1 - e^{-C}])}.$$

□

3.3.3 A Langford/Seeger-Like Theorem.

If we make use, as function $D(a, b)$ between the empirical risk and the true risk, of the Kullback-Leibler divergence between two Bernoulli distributions with probability of success a and b, we can obtain a bound similar to Seeger [2002], Langford [2005]. Concretely, we apply Theorem 3 with:

$$D(a, b) \;=\; \mathrm{kl}(a, b) \;=\; a \ln \frac{a}{b} + (1 - a) \ln \frac{1 - a}{1 - b}.$$

Corollary 3. *Let $V \geq 2$ be the number of views. For any distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for any set of prior distributions $\{P_v\}_{v=1}^V$, for any hyper-prior distributions π over views $\mathcal{V}$, we have:*

$$\mathrm{kl}\left(\tfrac{1}{2} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} d_S^{\mathrm{MV}}(\rho_S) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e_S^{\mathrm{MV}}(\rho_S), \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} R_{\mathcal{D}}(G_{\rho_S}^{\mathrm{MV}})\right)$$
$$\leq \frac{1}{m}\left[\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\rho_S} \mathrm{KL}(Q_{v,S}\|P_v) + \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathrm{KL}(\rho_S\|\pi) + \ln \frac{2\sqrt{m}}{\delta}\right].$$

where $\xi(m) = \displaystyle\sum_{k=0}^{m} \binom{m}{k} \left(\frac{k}{m}\right)^k \left(1 - \frac{k}{m}\right)^{m-k} \leq 2\sqrt{m}.$

Proof. The result follows from Theorem 3 by taking $D(a, b) = \mathrm{kl}(a, b)$, and upper-bounding $\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} e^{m\ \mathrm{kl}(R_S(h), R_{\mathcal{D}}(h))}$. By considering $R_S(h)$ as a random variable which follows a binomial distribution of m trials with a probability of success $R(h)$, we can prove:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} e^{m\ \mathrm{kl}(R_S(h), R_{\mathcal{D}}(h))} &= \mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \left[\frac{R_S(h)}{R_{\mathcal{D}}(h)}\right]^{mR_S(h)} \left[\frac{1 - R_S(h)}{1 - R_{\mathcal{D}}(h)}\right]^{m(1-R_S(h))} \\
&= \mathop{\mathbb{E}}_{v\sim\pi} \mathop{\mathbb{E}}_{h\sim P_v} \sum_{k=0}^{m} \Pr_{S\sim\mathcal{D}^m}\left(R_S(h) = \tfrac{k}{m}\right) \left[\frac{k/m}{R_{\mathcal{D}}(h)}\right]^k \left[\frac{1 - k/m}{1 - R_{\mathcal{D}}(h)}\right]^{m-k} \\
&= \sum_{k=0}^{m} \binom{m}{k} \left[\frac{k}{m}\right]^k \left[1 - \frac{k}{m}\right]^{m-k} \\
&= \xi(m).
\end{aligned}$$

□

4 Discussion on Related Work

In this section, we discuss two related theoretical studies of multiview learning related to the notion of Gibbs classifier.

Amini et al. [2009] proposed a Rademacher analysis of the risk of the stochastic Gibbs classifier over the view-specific models (for more than two views) where the distribution over the views is restricted to the uniform distribution. In their work, each view-specific model is found by minimizing the empirical risk: $h_v^* = \operatorname*{argmin}_{h\in\mathcal{H}_v} \dfrac{1}{m} \displaystyle\sum_{(\mathbf{x},y)\in S} \mathbb{1}_{[h(x^v)\neq y]}$. The prediction for a multiview example $\mathbf{x}$ is then based over the stochastic Gibbs classifier defined according to the uniform distribution, *i.e.,* $\forall v \in V,\ \rho(v) = \frac{1}{V}$. The risk of the multiview classifier Gibbs is hence given by

$$R_{\mathcal{D}}(G_{\rho=1/V}^{\mathrm{MV}}) = \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \frac{1}{V} \sum_{v=1}^{V} \mathbb{1}_{[h_v^*(x^v)\neq y]}.$$

Moreover, Sun et al. [2016] proposed a PAC-Bayesian analysis for multiview learning over the concatenation of the views, where the number of views is set to two, and deduced a SVM-like learning algorithm from this framework.

Table 1: Accuracy and $F1$-score averages for all the classes over 20 random sets. Note that the results are obtained for different sizes m of the learning sample and are averaged over the six *one-vs-all* classification problems. Along the columns, best results are in bold. $^{\downarrow}$ indicates statistically significantly worse performance than the best result, according to Wilcoxon rank sum test ($p < 0.02$) [Lehmann, 1975].

Strategy	$m = 150$		$m = 200$		$m = 250$		$m = 300$	
	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1
$\texttt{Mono}_v$	.8516±.0031$^{\downarrow}$	.1863±.0299$^{\downarrow}$	.8424±.0272$^{\downarrow}$	.3056±.0233$^{\downarrow}$	.8691±.0017$^{\downarrow}$	.3352±.0164$^{\downarrow}$	.8770±.0018$^{\downarrow}$	.4103±.0158$^{\downarrow}$
$\texttt{Concat}_{\texttt{SVM}}$	.8507±.0051$^{\downarrow}$	.1577±.0403$^{\downarrow}$	.8615±.0018$^{\downarrow}$	.2505±.0182$^{\downarrow}$	.8674±.0026$^{\downarrow}$	.3006±.0267$^{\downarrow}$	.8746±.0022$^{\downarrow}$	.3647±.0258$^{\downarrow}$
$\texttt{Aggreg}_{\texttt{P}}$	.8521±.0041$^{\downarrow}$	.1810±.0305$^{\downarrow}$	.8420±.0385$^{\downarrow}$	.2852±.0339$^{\downarrow}$	.8676±.0023$^{\downarrow}$	.3027±.0234$^{\downarrow}$	.8774±.0021$^{\downarrow}$	.3945±.0185$^{\downarrow}$
$\texttt{Aggreg}_{\texttt{L}}$	.8507±.0043$^{\downarrow}$	.1653±.0336$^{\downarrow}$	.8477±.0377$^{\downarrow}$	.2806±.0244$^{\downarrow}$	.8682±.0022$^{\downarrow}$	.3116±.0210$^{\downarrow}$	.8773±.0024$^{\downarrow}$	.3943±.0204$^{\downarrow}$
$\texttt{Fusion}^{\texttt{all}}_{\texttt{SVM}}$	.8568±.0087$^{\downarrow}$	.3899±.0789$^{\downarrow}$	.8527±.0406$^{\downarrow}$	.5027±.0780	.8490±.0716$^{\downarrow}$	**.5399±.0585**	.8422±.0526$^{\downarrow}$	**.5779±.0422**
$\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$	**.8692±.0059**	**.4298±.0570**	**.8768±.0082**	**.5066±.0402**	**.8846 ±.0047**	.5365±.0371	**.8881± .0060**	.5705±.0286

The key idea of their approach is to define a prior distribution that promotes similar classification among the two views, and the notion of diversity among the views is handled by a different strategy than ours. We believe that the two approaches are complementary, as in the general case of more than two views that we consider in our work, we can also use a similar informative prior as the one proposed by Sun et al. [2016] for learning.

5 Experiments

In this section, we present experiments to highlight the usefulness of our theoretical analysis by following a two-level hierarchy strategy. To do so, we learn a multiview model in two stages by following a classifier late fusion approach [Snoek et al., 2005] (sometimes referred as stacking [Wolpert, 1992]). Concretely, we first learn view-specific classifiers for each view at the base level of the hierarchy. Each view-specific classifier is expressed as a majority vote of kernel functions. Then, we learn weighted combination based on predictions of view-specific classifiers. It is worth noting that this is the procedure followed by Morvant et al. [2014] in a PAC-Bayesian fashion, but without any theoretical justifications and in a ranking setting.

We consider a publicly available multilingual multiview text categorization corpus extracted from the Reuters RCV1/RCV2 corpus [Amini et al., 2009][3], which contains more than $110,000$ documents from five different languages (English, German, French, Italian, Spanish) distributed over six classes. To transform the dataset into a binary classification task, we consider six *one-versus-all* classification problems: For each class, we learn a multiview binary classification model by considering all documents from that class as positive examples and all others as negative examples. We then split the dataset into training and testing sets: we reserve a test sample containing 30% of total documents. In order to highlight the benefits of the information brought by multiple views, we train the models with small learning sets by randomly choosing the learning sample S from the remaining set of the documents; the number of learning examples m considered are: 150, 200, 250 and 300. For each fusion-based approach, we split the learning sample S into two parts: S_1 for learning the view-specific classifier at the first level and S_2 for learning the final multiview model at the second level; such that $|S_1| = \frac{3}{5}m$ and $|S_2| = \frac{2}{5}m$ (with $m = |S|$). In addition, the reported results are averaged on 20 runs of experiments, each run being done with a new random learning sample. Since the classes are highly unbalanced, we report in Table 1 the accuracy along with the $F1$-measure, which is the harmonic average of precision and recall, computed on the test sample.

To assess that multiview learning with late fusion makes sense for our task, we consider as baselines the four following one-step learning algorithms (provided with the learning sample S). First, we learn a view-specific model on each view and report, as $\texttt{Mono}_v$, their average performance. We also follow an early fusion procedure, referred as $\texttt{Concat}_{\texttt{SVM}}$, consisting of learning one single model using SVM [Cortes and Vapnik, 1995] over the simple concatenation of the features of five views. Moreover, we look at two simple voters' combinations, respectively denoted by $\texttt{Aggreg}_{\texttt{P}}$ and $\texttt{Aggreg}_{\texttt{L}}$, for which the weights associated with each view follow the uniform distribution. Concretely, $\texttt{Aggreg}_{\texttt{P}}$, respectively $\texttt{Aggreg}_{\texttt{L}}$, combines the real-valued prediction, respectively the labels, returned

[3] https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,+Multiview+Text+Categorization+Test+collection

by the view-specific classifiers. In other words, we have

$$\texttt{Aggreg}_{\texttt{P}}(\mathbf{x}) \;=\; \frac{1}{5}\sum_{v=1}^{5} h^v(x^v)\,,$$

$$\text{and} \quad \texttt{Aggreg}_{\texttt{L}}(\mathbf{x}) \;=\; \frac{1}{5}\sum_{v=1}^{5} \text{sign}\left[h^v(x^v)\right]\,,$$

where $h^v(x^v)$ is the real-valued prediction of the view-specific classifier learned on view v.

We compare the above one-step methods to the two following late fusion approaches that only differ at the second level. Concretely, at the first level we construct from S_1 different view-specific majority vote expressed as linear SVM models[4] with different hyperparameter C values (12 values between 10^{-8} and 10^3): We do not perform cross-validation at the first level. This has the advantage to *(i)* lighten the first level learning process, since we do not need to validate models, and *(ii)* to potentially increase the expressivity of the final model.

At the second level, as it is often done for late fusion, we learn from S_2 the final weighted combination over the view specific voters using a RBF kernel. The methods referred as $\texttt{Fusion}^{\texttt{all}}_{\texttt{SVM}}$, respectively $\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$, make use of SVM, respectively the PAC-Bayesian algorithm CqBoost [Roy et al., 2016]. Note that, as recalled in Section 2, CqBoost is an algorithm that tends to minimize the C-Bound of Equation (2): it directly captures a trade-off between accuracy and disagreement.

We follow a 5-fold cross-validation procedure for selecting the hyperparameters of each learning algorithm. For $\texttt{Mono}_v$, $\texttt{Concat}_{\texttt{SVM}}$, $\texttt{Aggreg}_{\texttt{P}}$ and $\texttt{Aggreg}_{\texttt{L}}$ the hyperparameter C is chosen over a set of 12 values between 10^{-8} and 10^3. For $\texttt{Fusion}^{\texttt{all}}_{\texttt{SVM}}$ and $\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$ the hyperparameter γ of the RBF kernel is chosen over 9 values between 10^{-6} and 10^2. For $\texttt{Fusion}^{\texttt{all}}_{\texttt{SVM}}$, the hyperparameter C is chosen over a set of 12 values between 10^{-8} and 10^3. For $\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$, the hyperparameter μ is chosen over a set of 8 values between 10^{-8} and 10^{-1}. Note that we made use of the *scikit-learn* [Pedregosa et al., 2011] implementation for learning our SVM models.

First of all, from Table 1, the two-step approaches provide the best results on average. Secondly, according to a Wilcoxon rank sum test [Lehmann, 1975] with $p < 0.02$, the PAC-Bayesian late fusion based approach $\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$ is significantly the best method—in terms of accuracy, and except for the smallest learning sample size ($m = 150$), $\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$ and $\texttt{Fusion}^{\texttt{all}}_{\texttt{SVM}}$ produce models with similar $F1$-measure. We can also remark that $\texttt{Fusion}^{\texttt{all}}_{\texttt{Cq}}$ is more "stable" than $\texttt{Fusion}^{\texttt{all}}_{\texttt{SVM}}$ according to the standard deviation values. These results confirm the potential of using PAC-Bayesian approaches for multiview learning where we can control a trade-off between accuracy and diversity among voters.

6 Conclusion and Future Work

In this paper, we proposed a first PAC-Bayesian analysis of weighted majority vote classifiers for multiview learning when observations are described by more than two views. Our analysis is based on a hierarchy of distributions, *i.e.* weights, over the views and voters: *(i)* for each view v a posterior and prior distributions over the view-specific voter's set, and *(ii)* a hyper-posterior and hyper-prior distribution over the set of views. We derived a general PAC-Bayesian theorem tailored for this setting, that can be specialized to any convex function to compare the empirical and true risks of the stochastic Gibbs classifier associated with the weighted majority vote. We also presented a similar theorem for the expected disagreement, a notion that turns out to be crucial in multiview learning. Moreover, while usual PAC-Bayesian analyses are expressed as probabilistic bounds over the random choice of the learning sample, we presented here bounds in expectation over the data, which is very interesting from a PAC-Bayesian standpoint where the posterior distribution is data dependent.

According to the distributions' hierarchy, we evaluated a simple two-step learning algorithm (based on late fusion) on a multiview benchmark. We compared the accuracies while using SVM and the PAC-Bayesian algorithm CqBoost for weighting the view-specific classifiers. The latter revealed itself as a better strategy, as it deals nicely with accuracy and the disagreement trade-off promoted by our PAC-Bayesian analysis of the multiview hierarchical approach.

We believe that our theoretical and empirical results are a first step toward the goal of theoretically understanding the multiview learning issue through the PAC-Bayesian point of view, and toward the objective of deriving new multiview learning algorithms. It gives rise to exciting perspectives.
Among them, we would like to specialize our result to linear classifiers for which PAC-Bayesian approaches are known to lead to tight bounds and efficient learning algorithms [Germain et al., 2009]. This clearly opens the door to derive theoretically founded algorithms for multiview learning.

[4] We use linear SVM model as it is usually done for text classification tasks [*e.g.*, Joachims, 1998].

Another possible algorithmic direction is to take into account a second statistical moment information to link it explicitly to important properties between views, such as diversity or agreement Kuncheva [2004], Amini et al. [2009]. A first direction is to deal with our multiview PAC-Bayesian C-Bound of Lemma 1—that already takes into account such a notion of diversity Morvant et al. [2014]—in order to derive an algorithm as done in a mono-view setting by Laviolette et al. [2011], Roy et al. [2016].
Another perspective is to extend our bounds to diversity-dependent priors, similarly to the approach used by Sun et al. [2016], but for more than two views. This would allow to additionally consider an *a priori* knowledge on the diversity.
Moreover, we would like to explore the *semi-supervised* multiview learning where one has access to unlabeled data $S_u = \{\mathbf{x}_j\}_{j=1}^{m_u}$ along with labeled data $S_l = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_l}$ during training. Indeed, an interesting behaviour of our theorem is that it can be easily extended to this situation: the bound will be a concatenation of a bound over $\frac{1}{2}d^{\text{MV}}_{S_u}(\rho)$ (depending on m_u) and a bound over $e^{\text{MV}}_{S_l}(\rho)$ (depending on m_s). The main difference with the supervised bound is that the Kullback-Leibler divergence will be multiplied by a factor 2.

Appendix—Mathematical Tools

Theorem 5 (Markov's ineq.). *For any random variable X s.t. $\mathbb{E}(|X|) = \mu$, for any $a > 0$, we have*

$$\mathbb{P}(|X| \geq a) \leq \frac{\mu}{a}.$$

Theorem 6 (Jensen's ineq.). *For any random variable X, for any concave function g, we have*

$$g(\mathbb{E}[X]) \;\geq\; \mathbb{E}[g(X)].$$

Theorem 7 (Cantelli-Chebyshev ineq.). *For any random variable X s.t. $\mathbb{E}(X) = \mu$ and $\mathbf{Var}(X) = \sigma^2$, and for any $a > 0$, we have*

$$\mathbb{P}(X - \mu \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

Acknowledgments.

This work is partially funded by the French ANR project LIVES ANR-15-CE23-0026-03, the "Région Rhône-Alpes", and by the CIFAR program in Learning in Machines & Brains.

References

Massih-Reza Amini, Nicolas Usunier, and Cyril Goutte. Learning from Multiple Partially Observed Views - an Application to Multilingual Text Categorization. In *NIPS*, pages 28–36, 2009.

Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El-Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Syst.*, 16(6):345–379, 2010.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *AISTATS*, pages 435–444, 2016.

Avrim Blum and Tom M. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *COLT*, pages 92–100, 1998.

Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56. Inst. of Mathematical Statistic, 2007.

Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010. ISBN 0262514125, 9780262514125.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *ICML*, pages 353–360, 2009.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: from a PAC-Bayesian analysis to a learning algorithm. *JMLR*, 16:787–860, 2015.

Thorsten Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *ECML*, pages 137–142, 1998. ISBN 3-540-64417-2.

Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004. ISBN 0471210781.

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 *NIPS*, pages 769–776, 2006.

John Langford. Tutorial on practical prediction theory for classification. *JMLR*, 6:273–306, 2005.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *NIPS*, pages 423–430. MIT Press, 2002.

Francois Laviolette, Mario Marchand, and Jean-Francis Roy. From PAC-Bayes bounds to quadratic programs for majority votes. In *ICML*, 2011.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *Ann. Statist.*, 42(1):211–224, 02 2014. doi: 10.1214/13-AOS1190. URL `http://dx.doi.org/10.1214/13-AOS1190`.

E. Lehmann. *Nonparametric Statistical Methods Based on Ranks*. McGraw-Hill, 1975.

Odalric-Ambrym Maillard and Nicolas Vayatis. Complexity versus agreement for many views. In *ALT*, pages 232–246, 2009.

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

David A. McAllester. PAC-Bayesian stochastic model selection. In *Machine Learning*, pages 5–21, 2003.

Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Majority Vote of Diverse Classifiers for Late Fusion. In *S+SSPR*, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Anastasia Pentina and Christoph H. Lampert. A PAC-Bayesian bound for lifelong learning. In *ICML*, pages 991–999, 2014.

Jean-Francis Roy, Mario Marchand, and Franois Laviolette. A column generation bound minimization approach with PAC-Bayesian generalization guarantees. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1241–1249, 2016.

Matthias W. Seeger. PAC-Bayesian generalisation error bounds for gaussian process classification. *JMLR*, 3:233–269, 2002.

Cees Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *ACM Multimedia*, pages 399–402, 2005.

Shiliang Sun, John Shawe-Taylor, and Liang Mao. PAC-Bayes analysis of multi-view learning. *CoRR*, abs/1406.5614, 2016.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.