PAC-Bayesian Theorems for Multiview Learning

Anil Goyal, Emilie Morvant, Pascal Germain, Massih-Reza Amini

PAC-Bayesian Theorems for Multiview Learning

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] INRIA, SIERRA Project-Team, 75589 Paris, France, et D.I.,
École Normale Supérieure, 75230 Paris, France

November 17, 2016

Abstract

We tackle the issue of multiview learning which aims to take advantage of multiple representations/views of the data. In this context, many machine learning algorithms exist. However, the majority of the theoretical studies focus on learning with exactly two representations. In this paper, we consider two level hierarchy of distributions over views and propose a general PAC-Bayesian theory for multiview learning with potentially more than two views. We concentrate our study on binary classification models that take the form of a majority vote. We derive PAC-Bayesian generalization bounds involving different relations between empirical and true risks by taking into account a notion of diversity of the voters and views, and that can be naturally extended to semi-supervised learning.

1 Introduction

The PAC-Bayesian approach introduced by McAllester [24] provides Probably Approximately Correct (PAC) generalization guarantees for models expressed as a weighted majority vote[1] over a set of classifiers/voters $\mathcal{H}$. In this framework one assumes a prior distribution P over $\mathcal{H}$ which models the *a priori* weights associated with each classifier[2] in $\mathcal{H}$. After observing the learning sample S, the learner aims at finding a posterior distribution Q over $\mathcal{H}$ that leads to a well-performing majority vote. Many PAC-Bayesian studies have been conducted to characterize the error of majority votes [7, 13, 19, 30] and also to derive theoretically grounded learning algorithms, namely in the supervised learning setting (*e.g.*, Alquier et al. [1], Germain et al. [11], Laviolette et al. [21], Parrado-Hernández et al. [27]) and the domain adaptation setting (*e.g.*, Germain et al. [14]).

In this paper, we make use of this PAC-Bayesian framework to study the issue of multiview learning [3, 32] that has been expanding over the past decade, spurred by the seminal work of Blum and Mitchell [6] on co-training. Most of the existing methods try to combine multimodal information, either by directly merging the descriptions or by combining models learned from the different descriptions[3] [31], in order to produce a model more reliable for the considered task. It is worth noting that multiview learning algorithms have obtained good empirical results in real-life applications, such as by Morvant et al. [26], where the PAC-Bayesian algorithm MinCq [20] has been extended to deal with multimodal data.

However, despite numerous successes of multiview learning, few theoretical studies have been conducted for the common case where observations are described by more than two views. Amini et al. [2] proposed a multiview learning generalization bound for (semi-)supervised setting based on the error of a majority vote model restricted to uniform distribution over the views, and on empirical Rademacher complexity [4]. Note that, Sun et al. [33] have recently proposed a different PAC-Bayesian framework for multiview learning when considering two views. In this work, we derive two general PAC-Bayesian theorems that allow to study the generalization ability with data described from more than two views. The originality of these

[1] Note that the majority vote setting is not too restrictive since many machine learning approaches can be considered as majority vote learning, notably ensemble methods [9, 29].

[2] For example, the classifiers expected to be the most accurate for the task can have the largest weights under P.

[3] The fusion of descriptors, respectively of models, is sometimes called Early Fusion, respectively Late Fusion.

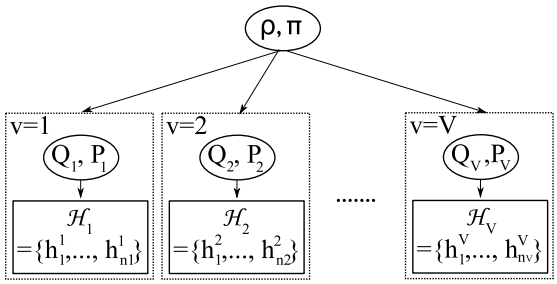

Figure 1: Distributions hierarchy for our PAC-Bayesian multiview approach. For all views $v \in \{1, \ldots, V\}$, we have a set of voters $\mathcal{H}_v = \{h_1^v, \ldots, h_{n_v}^v\}$ on which we consider prior P_v view-specific distribution, and "above" we consider a hyper-prior π distribution over the set of all the views. The objective is to learn a posterior Q_v view-specific distributions and a hyper-posterior ρ distribution leading to a good model.

results is that, given a set of base voters for each view, we define a hierarchy of posterior and prior distributions over the views: *(i)* for each view v, we consider prior P_v and posterior Q_v distributions over each view-specific voters' set, and *(ii)* we consider prior π and posterior ρ distributions over the set of views (see Figure 1 for an illustration), respectively called hyper-prior and hyper-posterior[4]. It is important to point out that this two-level hierarchy leads to a more natural multiview learning framework than the one proposed by Amini et al. [2]. The first PAC-Bayesian theorem is expressed as a general generalization bound involving an additional term taking into account the expectation, according to the hyper-posterior distribution, of view-specific Kullback-Leibler divergences between the posterior and prior distributions over the views. The second PAC-Bayesian theorem goes one step further by proposing a mechanism to control the trade-off between this expectation of view-specific Kullback-Leibler divergences and the Kullback-Leibler divergence between ρ and π. These bounds can be easily extended to semi-supervised learning thanks to a notion of disagreement between all the voters that appears in our bounds, allowing to take into account a notion of diversity between voters which is known as a key element in multiview learning [2, 8, 16, 22, 32].

The rest of the paper is organized as follows. Section 2 recalls the usual general PAC-Bayesian theorem as presented by Germain et al. [13]. Our contributions, consisting in PAC-Bayesian theorems for multiview data, are stated in Section 3. Note that the proofs of our results are provided in the supplementary material. In section 4, we discuss the relation between our analysis and those provided by Amini et al. [2] and Sun et al. [33]. Finally, we conclude and discuss future works in Section 5.

2 The Classical PAC-Bayesian Theorem in a Single-View Setting

In this section, we recall the general PAC-Bayesian theorem proposed by Germain et al. [13]. 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 the 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. We consider a set $\mathcal{H}$ of classifiers (or voters) from $\mathcal{X}$ to $\mathcal{Y}$. In addition, the PAC-Bayesian approach requires a prior distribution P over $\mathcal{H}$ that models an *a priori* belief on the voters from $\mathcal{H}$ before the observation of the learning sample S. Given S, the learner objective is then to find a posterior distribution Q leading to an accurate Q-weighted majority vote $B_Q(x)$ defined as:

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

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

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

where $\mathbb{1}_{[\pi]} = 1$ if predicate π holds, and 0 otherwise.
However, PAC-Bayesian generalization bounds do not directly focus on the risk of the deterministic

[4]Note that our notion of hyper-prior and hyper-posterior distributions is different than the one proposed for Lifelong learning by Pentina and Lampert [28], 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.

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 of the Gibbs classifier on a data distribution $\mathcal{D}$ is given by:

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

From multiview learning standpoint where the notion of diversity between voters is known to be important, it is worth noting that this above risk can be rewritten [13] in terms of *expected disagreement* $d_{\mathcal{D}}(Q)$ and *expected joint error* $e_{\mathcal{D}}(Q)$ between all the pair of voters as:

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

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

Indeed, according to the above equations, the expected disagreement $d_{\mathcal{D}}(Q)$ and expected joint error $e_{\mathcal{D}}(Q)$ compare the output of pairwise voters. Hence, they directly capture the diversity between the voters in the output space while $e_{\mathcal{D}}(Q)$ takes into account the errors.

An important behavior of the above Gibbs classifier is that it 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 a prediction error on x. Therefore, we have:

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

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 [13, 17, 19], such as the C-Bound (first proved by Lacasse et al. [17]) which can be expressed as follows:

$$R_{\mathcal{D}}(B_Q) \leq 1 - \frac{\left(1 - 2R_{\mathcal{D}}(G_Q)\right)^2}{1 - 2d_{\mathcal{D}}(Q)}\,. \tag{3}$$

PAC-Bayesian generalization bounds also take into account the prior distribution P on $\mathcal{H}$ through the Kullback-Leibler divergence between the learned posterior distribution Q and the given prior P:

$$\mathrm{KL}(Q\|P) \;=\; \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}. \tag{4}$$

Finally, such bounds rely obviously on the empirical risk, that is here defined as:

$$R_S(G_Q) = \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbf{E}}_{h\sim Q} \mathbb{1}_{[h(x_i)\neq y]}.$$

The following theorem is a general PAC-Bayesian theorem which takes the form of an upper bound on the "deviation" between the true and empirical risks of the Gibbs classifier, according to a convex function $D : [0,1]\times[0,1] \to \mathbb{R}$.

Theorem 1 (Germain et al. [11, 13])**.** *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 $\delta \in (0,1]$, for any convex function $D : [0,1]\times[0,1] \to \mathbb{R}$, with a probability at least $1-\delta$ over the random choice of $S \sim (\mathcal{D})^m$, we have for all posterior distribution Q on $\mathcal{H}$:*

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

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

By selecting a well-suited deviation function D and by upper-bounding $\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\, D(R_S(h))}$, we can retrieve the classical versions of the PAC-Bayesian theorem (*i.e.*, Catoni [7], McAllester [24], Seeger [30]).

3 New PAC-Bayesian Theorems for Multiview Learning

In this section, we provide our contribution: An extension of the PAC-Bayesian framework to multiview learning with more than two views.

3.1 Notations and Setting

We consider binary classification problems where, given a multiview observation, our observations $\mathbf{x} = (x^1, \ldots, x^V)$ belong to the input set $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_V$. In binary classification, we assume that examples are pairs $(\mathbf{x}, y)$, with $y \in \mathcal{Y} = \{-1, +1\}$, drawn according to a fixed, but unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.

For each view $v \in \mathcal{V}$, we propose to consider a view-specific set $\mathcal{H}_v$ of voters $h : \mathcal{X}_v \to \{-1, +1\}$. Given a prior distribution P_v on $\mathcal{H}_v$ for each view $v \in \mathcal{V}$, an "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:

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

In other words, 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:

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

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:

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

Similarly to Equation (1), we can express the Gibbs classifier's risk in the terms of *expected disagreement* $d^{\text{MV}}_{\mathcal{D}}(\rho)$ and *expected joint error*, $e^{\text{MV}}_{\mathcal{D}}(\rho)$ as the following:

$$R_{\mathcal{D}}(G^{\text{MV}}_\rho) = \frac{1}{2} d^{\text{MV}}_{\mathcal{D}}(\rho) + e^{\text{MV}}_{\mathcal{D}}(\rho) \tag{5}$$

where

$$d^{\text{MV}}_{\mathcal{D}}(\rho) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{v'\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} \mathop{\mathbf{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq h'(x^{v'})]},$$
$$e^{\text{MV}}_{\mathcal{D}}(\rho) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{v'\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} \mathop{\mathbf{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq y]} \mathbb{1}_{[h'(x^{v'})\neq y]}.$$

As for Equation (1), these two terms compare the output of the voters by considering the probability of disagreement between them for all views, meaning that they capture a notion of diversity between voters and between views. As in classical 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 $R_{\mathcal{D}}(B^{\text{MV}}_\rho) \leq 2R_{\mathcal{D}}(G^{\text{MV}}_\rho)$.

It is also noteworthy that the C-bound of Equation (3) can be extended to our multiview setting as in Lemma 1 below. Equation (6) is a straightforward generalization of the "single-view" C-bound of Equation (3). Afterward, Equation (7) 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 distributions,*

$$R_{\mathcal{D}}(B^{\text{MV}}_\rho) \leq 1 - \frac{\left(1 - 2R_{\mathcal{D}}(G^{\text{MV}}_\rho)\right)^2}{1 - 2d^{\text{MV}}_{\mathcal{D}}(\rho)} \tag{6}$$

$$\leq 1 - \frac{\left(1 - 2\,\mathbf{E}_{v\sim\rho} R_{\mathcal{D}}(G_{Q_v})\right)^2}{1 - 2\,\mathbf{E}_{v\sim\rho} d_{\mathcal{D}}(Q_v)}\,. \tag{7}$$

Proof. Eq. (6) follows from the Chebyshev inequality (Appendix A). Moreover, note 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 therefore

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

In the other hand,

$$\begin{aligned} d^{\text{MV}}_{\mathcal{D}}(\rho) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{v'\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} \mathop{\mathbf{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq h'(x^{v'})]} \\ &= \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{v'\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} \mathop{\mathbf{E}}_{h\sim Q_{v'}} h(x^v) \times h'(x^{v'})\right) \\ &= \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \left[\mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} h(x^v)\right]^2\right) \end{aligned} \tag{8}$$

From the Jensen's inequality it comes

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

The two above results allow obtaining the result, by upper bounding Equation (6):

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

□

The trade-off of Equation (6)—between $d^{\text{MV}}_{\mathcal{D}}(\rho)$ and $R_{\mathcal{D}}(G^{\text{MV}}_\rho)$—suggests that the diversity between voters should be increased while preserving a low Gibbs classifier's risk. Equation (7) exhibits the role of diversity among the views. Consequently, this separation can be very useful for multiview learning where the diversity between views and/or voters plays an important role [2, 16, 22, 32, 33].

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

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

where $d^{\text{MV}}_S(\rho)$ and $e^{\text{MV}}_S(\rho)$ respectively are the empirical estimations on S of $d^{\text{MV}}_{\mathcal{D}}(\rho)$ and $e^{\text{MV}}_{\mathcal{D}}(\rho)$.

3.2 Multiview PAC-Bayesian Theorems

We state now our two PAC-Bayesian theorems suitable for the above multiview learning setting, both in a general form.

First bound. A key step in PAC-Bayesian proofs is the use of a *change of measure inequality* [25], based on the Donsker-Varadhan inequality [10]. Lemma 2 below extends the usual change of measure to our multiview setting.

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

$$\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}\phi(h) \leq \mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v)+\mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{\phi(h)}\right). \tag{9}$$

Proof. Deferred to Appendix B. □

Based on the previous lemma, the following theorem can be seen as a generalization of Theorem 1 to our multiview setting. Note that we still rely on a 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 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 any convex function $D : [0,1]\times[0,1]\to\mathbb{R}$, for any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the random choice of $S\sim(\mathcal{D})^m$, for all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior ρ distributions, we have:*

$$D\big(\tfrac{1}{2}d_S^{\mathrm{MV}}(\rho)+e_S^{\mathrm{MV}}(\rho), R_{\mathcal{D}}(G_\rho^{\mathrm{MV}})\big) \leq \frac{1}{m}\left[\mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v)+\mathrm{KL}(\rho\|\pi)+\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right)\right].$$

Proof. First, note that $\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{mD(R_S(h),R_{\mathcal{D}}(h))}$ is a non-negative random variable. Using Markov's inequality, with $\delta\in(0,1]$, and a probability at least $1-\delta$ over the random choice of the multiview learning sample $S\sim(\mathcal{D})^m$, we have:

$$\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{m\,D(R_S(h),R_{\mathcal{D}}(h))} \leq \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{mD(R_S(h),R_{\mathcal{D}}(h))}.$$

By taking the logarithm on both sides, with a probability at least $1-\delta$ over $S\sim(\mathcal{D})^m$, we have:

$$\ln\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}\right] \leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right]. \tag{10}$$

We now apply Lemma 2 on the left hand-side of the Inequality (10) with $\phi(h) = m\,D(R_S(h), R_{\mathcal{D}}(h))$. Therefore, for any Q_v on $\mathcal{H}_v$ for all views $v\in\mathcal{V}$, and for any ρ on views $\mathcal{V}$, with a probability at least $1-\delta$ over $S\sim(\mathcal{D})^m$, we have:

$$\begin{aligned}\ln\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}\right] &\geq m\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}D(R_S(h),R_{\mathcal{D}}(h)) - \mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v) - \mathrm{KL}(\rho\|\pi)\\ &\geq m\,D\left(\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}R_S(h), \mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}R_{\mathcal{D}}(h)\right) - \mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v) - \mathrm{KL}(\rho\|\pi),\end{aligned}$$

where the last inequality is obtain by applying Jensen's inequality on the convex function D. By rearranging the terms, we have:

$$\begin{aligned}D\left(\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}R_S(h), \mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}R_{\mathcal{D}}(h)\right) \leq \frac{1}{m}\Bigg[&\mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v) + \mathrm{KL}(\rho\|\pi)\\ &+\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}\right)\Bigg].\end{aligned}$$

Finally, the theorem statement is obtained by rewriting:

$$\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}R_S(h) = \tfrac{1}{2}d_S^{\mathrm{MV}}(\rho) + e_S^{\mathrm{MV}}(\rho)\,, \tag{11}$$

$$\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}R_{\mathcal{D}}(h) = R_{\mathcal{D}}(G_\rho^{\mathrm{MV}})\,. \tag{12}$$

□

It is interesting to compare this generalization bound to the classical one of Theorem 1. The main difference relies on the introduction of view-specific prior and posterior distributions, which mainly leads to an additional term $\mathbf{E}_{v\sim\rho}\,\mathrm{KL}(Q_v\|P_v)$ 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.

Second bound. We now derive another PAC-Bayesian generalization bound (Theorem 3), where the above expectation over view-specific prior and posterior distributions is expressed over the hyper-prior distribution π. To this end, we propose to control the influence of the expected *view-specific* Kullback-Leibler divergences by taking into account the deviation between ρ and π with the following measure:

$$\forall q>0,\qquad \beta_q(\rho\|\pi)=\left[\mathop{\mathbf{E}}_{v\sim\pi}\left(\frac{\rho(v)}{\pi(v)}\right)^q\right]^{\frac{1}{q}}.$$

The measure $\beta_q(\rho\|\pi)$ has been proposed by [14] in the context of domain adaptation to quantify the divergence between two data distributions[5]. Here, we make use of it to infer a new change of measure inequality, stated as Lemma 3 below.

Lemma 3. *For any set of prior distributions $\{P_v\}_{v=1}^V$ and posterior distribution $\{Q_v\}_{v=1}^V$, for any hyper-prior distributions π over views $\mathcal{V}$ and hyper-posterior distributions ρ over $\mathcal{V}$, for any $p,q>0$ such that $\frac{1}{q}+\frac{1}{p}=1$, and for any measurable function $\phi:\mathcal{H}_v\to\mathbb{R}$, we have:*

$$\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}\phi(h)\le\ \beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}}+\mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{\phi(h)}\right).$$

Proof. Deferred to Appendix C. □

Lemma 3 leads us to the next theorem.

Theorem 3. *Let $V\ge 2$ 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 $\mathcal{V}$, and for any convex function $D:[0,1]\times[0,1]\to\mathbb{R}$, for any $p,q>0$ such that $\frac{1}{q}+\frac{1}{p}=1$, for any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the random choice of $S\sim(\mathcal{D})^m$ for all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior ρ distributions, we have:*

$$D\left(\tfrac{1}{2}d_S^{\mathrm{MV}}(\rho)+e_S^{\mathrm{MV}}(\rho),R_{\mathcal{D}}(G_\rho^{\mathrm{MV}})\right)\le\ \frac{1}{m}\Bigg[\beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}}+\mathrm{KL}(\rho\|\pi)$$
$$+\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}\right)\Bigg],$$

where

$$\beta_q(\rho\|\pi)\ \ge\ \exp\left(\frac{1}{p}\,\mathrm{KL}(\rho\|\pi)\right).\tag{13}$$

Proof. The first steps of the proof are the same than the ones of Theorem 2. We start from Equation (10). Let's apply Lemma 3 on the left hand-side of Inequality (10) with $\phi(h)=m\,D\left(R_S(h),R_{\mathcal{D}}(h)\right)$. Therefore, for all $p,q>0$ such that $\frac{1}{q}+\frac{1}{p}=1$, with a probability at least $1-\delta$ over $S\sim(\mathcal{D})^m$, for all Q_v on $\mathcal{H}_v$ for every views $v\in\mathcal{V}$ and $\forall\rho$ on $\mathcal{V}$, we have:

$$\ln\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{mD(R_S(h),R_{\mathcal{D}}(h))}\right]\ge m\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}D\left(R_S(h),R_{\mathcal{D}}(h)\right)-\mathrm{KL}(\rho\|\pi)-\beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}}$$
$$\ge m\,D\left(\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}R_S(h),\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}R_{\mathcal{D}}(h)\right)$$
$$-\mathrm{KL}(\rho\|\pi)-\beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}}.$$

[5]The term $\beta_q(\rho\|\pi)$ appears in the proof of Lemma 3 through the use of Hölder's inequality. Note that the latter plays a central role in the PAC-Bayesian work of Bégin et al. [5], where Hölder's inequality allows to replace the KL divergence by Rényi's one in a classical supervised classification setting.

where the last inequality is obtain by applying Jensen's inequality on the convex function D. By rearranging the terms, we have:

$$D\left(\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v} R_S(h), \mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v} R_{\mathcal{D}}(h)\right) \leq \frac{1}{m}\left[\beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}} + \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v} e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}\right)\right].$$

Theorem 3 statement is obtained by the substitution of Equations (11) and (12) in above inequality.

Finally, we prove the relation between $\beta_q(\rho\|\pi)$ and $\mathrm{KL}(\rho\|\pi)$ of Equation (13). We have:

$$\begin{aligned}
\mathrm{KL}(\rho\|\pi) &= \mathop{\mathbf{E}}_{v\sim\rho}\ln\frac{Q(v)}{P(v)} \\
&= \mathop{\mathbf{E}}_{v\sim\rho}\frac{1}{q-1}\ln\left(\left[\frac{Q(v)}{P(v)}\right]^{q-1}\right) \\
&\leq \frac{1}{q-1}\ln\left(\mathop{\mathbf{E}}_{v\sim\rho}\left[\frac{Q(v)}{P(v)}\right]^{q-1}\right) \qquad (14)\\
&= \frac{1}{q-1}\ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\left[\left(\frac{Q(v)}{P(v)}\right)^{q-1}\times\frac{\rho(v)}{\pi(v)}\right]\right) \\
&= \frac{1}{q-1}\ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\left[\frac{Q(v)}{P(v)}\right]^{q}\right).
\end{aligned}$$

Line (14) is obtained thanks to the Jensen's inequality. Now, by multiplying by $\frac{1}{q}$ on the both side of the inequality and since $p=\frac{q}{q-1}$, we have:

$$\begin{aligned}
&\frac{1}{q}\mathrm{KL}(\rho\|\pi) \leq \frac{1}{q}\times\frac{1}{q-1}\ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\left[\frac{Q(v)}{P(v)}\right]^{q}\right) \\
\iff &\frac{1}{p}\mathrm{KL}(\rho\|\pi) \leq \ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\left[\frac{Q(v)}{P(v)}\right]^{q}\right)^{q} \\
\iff &\exp\left(\frac{1}{p}\mathrm{KL}(\rho\|\pi)\right) \leq \beta_q\,.
\end{aligned}$$

□

It is worth noting that dealing with different values of q and p leads to different bounds. For example, if $q=\infty$ then $p=1$ the bound of Theorem 3 becomes:

$$D\left(\tfrac{1}{2}d_S^{\mathrm{MV}}(\rho)+e_S^{\mathrm{MV}}(\rho), R_{\mathcal{D}}(G_\rho^{\mathrm{MV}})\right) \leq \frac{1}{m}\left[\beta_\infty(\rho\|\pi)\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v) + \mathrm{KL}(\rho\|\pi) + \ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v} e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}\right)\right],$$

with

$$\beta_\infty(\rho\|\pi) \;=\; \sup_{v\in\mathcal{V}}\frac{\rho(v)}{\pi(v)} \;\geq\; e^{\mathrm{KL}(\rho\|\pi)}. \qquad (15)$$

Therefore, $\beta_\infty(\rho\|\pi)$ clearly weights the influence of the expectation over all the view-specific prior and posterior distributions according to the relation between the hyper-prior and hyper-posterior distributions. We believe that this behavior can be useful for algorithmic purposes. Moreover, the relation of Equation (15) suggests that if $\mathrm{KL}(\rho\|\pi)$ increases then $\beta_q(\rho\|\pi)$ increases exponentially: The higher the deviation between ρ and π is, the higher the importance of the expectation of Kullback-Leibler divergence between the view-specific prior and posterior distributions would be.

Remarks. An interesting behavior of our theorems is that they can be easily extended to *semi-supervised* multiview learning, where we have 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, in this situation the bounds 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 above results is that a factor 2 on the Kullback-Leibler divergences appear[6].

As Theorem 1 for classical supervised learning, Theorems 2 and 3 provide tools to derive PAC-Bayesian generalization bounds for a multiview supervised learning setting. Indeed, by making use of the same trick as Germain et al. [11, 13], the generalization bounds can be derived from Theorems 2 and 3 by choosing a suitable convex function D and upper-bounding $\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v} e^{m\,D(R_S(h),R_{\mathcal{D}}(h))}$. In the next section we give an example of this kind of deviation through the approach of Catoni [7], that is one of the three classical PAC-Bayesian approaches [7, 18, 24, 30]. Note that we provide the specialization to the two other approaches in supplementary material.

3.3 An Example of Specialization of the General Theorems for Multi-View

To derive a generalization bound with the Catoni [7]'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 $D(a, b) = \mathcal{F}(b) - C\,a$ [11, 13]. We obtain the following generalization bound.

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 distributions π over $\mathcal{V}$, for any $p, q>0$ such that $\frac{1}{q}+\frac{1}{p}=1$, for all $C > 0$, for any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the random choice of $S\sim(\mathcal{D})^m$ for all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior ρ distributions, we have:*

$$R_{\mathcal{D}}(G^{\text{MV}}_\rho) \;\leq\; \frac{1}{1-e^{-C}}\left(1-\exp\left[-\left(C\left(\tfrac{1}{2}d^{\text{MV}}_S(\rho)+e^{\text{MV}}_S(\rho)\right)+\frac{1}{m}\left[\mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v)+\mathrm{KL}(\rho\|\pi)+\ln\tfrac{1}{\delta}\right]\right)\right]\right),$$

and

$$R_{\mathcal{D}}(G^{\text{MV}}_\rho) \leq \frac{1}{1-e^{-C}}\left(1-\exp\left[-\left(C\left(\tfrac{1}{2}d^{\text{MV}}_S(\rho)+e^{\text{MV}}_S(\rho)\right)+\frac{1}{m}\left[\beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}}+\mathrm{KL}(\rho\|\pi)+\ln\tfrac{1}{\delta}\right]\right)\right]\right).$$

Proof. Deferred to Appendix D (similar to the proof of Germain et al. [11, Corollary 2.2]). □

The advantage of the bounds of Corollary 1 is that they are expressed as a trade-off between the empirical disagreement, empirical joint error and the Kullback-Leibler divergences. This trade-off can be controlled to have a high diversity between the views. From a practical standpoint, it could help to learn a (hyper-)posterior distribution performing well by taking into consideration all the views along with the performances of individual view specific majority vote and diversity.

4 Discussion on Related Works

In this section, we discuss two related theoretical studies of multiview learning related to the notion of Gibbs classifier [2, 33].

Amini et al. [2] present a Rademacher analysis of the risk of the stochastic Gibbs classifier over the view-specific model (for more than two views) where the distribution over the views is restricted to the uniform distribution. In this work, each view-specific model is found by minimizing the empirical risk:

$$h^*_v \;=\; \mathop{\mathrm{argmin}}_{h\in\mathcal{H}_v}\frac{1}{m}\sum_{(\mathbf{x},y)\in S}\mathbb{1}_{[h(x^v)\neq y]}.$$

[6] We refer the reader to Germain et al. [13, Section 5.4] to have insights on the derivation of PAC-Bayesian bounds for single-view disagreement and joint errors. The generalization to our multiview setting is straightforward.

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 as:

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

Moreover, in the semi-supervised situation, the authors propose to reduce the set of possible view-specific models by estimating the diversity on the unlabeled data through the variance between the classifiers. This is done, by selecting the set of good view-specific classifiers which have small expected variance, given by:

$$\mathbb{V}(h^*_1,\ldots,h^*_V) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\frac{1}{V}\sum_{v=1}^{V} h^*_v(x^v)^2 - \left(\frac{1}{V}\sum_{v=1}^{V} h^*_v(x^v)\right)^2\right].$$

This can be seen as an *a priori* on what models could be able to take into account the diversity between views, since this diversity measure does not appear in the generalization bound. In our case, for each view v, the model h^*_V can be seen as the best empirical view-specific Gibbs classifier defined as

$$\operatorname*{argmin}_{Q_v} \frac{1}{m}\sum_{(\mathbf{x},y)\in S} \mathop{\mathbf{E}}_{h\sim Q_v} \mathbb{1}_{[h(x^v)\neq y]}.$$

Our setting is then less constraining since thanks to the hyper-posterior distribution over the views we do not restrict it to be uniform. Our generalization bounds suggest that in order to be able to find the best multiview model we have to achieve a small multiview Gibbs classifier's risk, while jointly controlling the diversity (according to the disagreement and joint error).

Secondly, Sun et al. [33] proposed a PAC-Bayesian analysis for two views learning. In this restricted context, they study the generalization ability of linear classifiers (possibly using kernels) over the concatenation of the two views, and propose a SVM-like learning algorithm. The key idea of their approach is to define a prior distribution that promotes similar classification among the two view. Therefore, the notion of diversity among the views is handled by a very different strategy than ours. We believe than the two approaches are complementary, as we could use a similar informative prior than Sun et al. [33]. A positive aspect of our hierarchical framework is that it naturally scales to an arbitrary number of views.

5 Conclusion and Future Works

In this paper, we proposed a first and novel PAC-Bayesian analysis for weighted majority vote classifiers for multiview learning in the case where 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. This allows us to derive two general PAC-Bayesian theorems that can be specialized to any convex function to compare the empirical and true risk of the stochastic Gibbs classifier associated with the weighted majority vote. Our theorems have the advantages to be very general and to directly involve a notion of diversity between views and between voters. We believe that these results are a first step toward the goal of theoretically understanding the multiview learning issue through the PAC-Bayesian point of view. 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 [11]. This clearly opens the door to derive theoretically founded algorithms for multiview learning in supervised and semi-supervised settings. Another perspective is to extend our bounds for diversity-dependant priors similar to the approach used by Sun et al. [33] for more than two views to additionally consider an *a priori* knowledge on the diversity. We also desire to extend these theoretical results to the domain adaptation issue [15, 23], for which one wants to adapt a model from a *source task* to a different, but related, *target task*. Indeed, the PAC-Bayesian approach has already led to promising theoretical and empirical results in a domain adaptation context while taking into account the diversity between voters [12, 14]. Moreover, the domain adaptation scenario is more realistic than usual supervised learning setting, since in real-life applications the target data—on which the model will be applied—may not come from the same data distribution as the source data used during the learning step.

Acknowledgments

This work is partially funded by the french ANR project LIVES ANR-15-CE23-0026-03 and the "Région Rhône-Alpes".

Appendix A Proof of the $\mathcal{C}$-Bound for Multiview Learning (Lemma 1)

In this section, we present the proof of Lemma 1, inspired by the proof provided by Germain et al. [13]. Firstly, we need to define the margin of the multiview weighted majority vote B_ρ^{MV} and its first and second statistical moments.

Definition 1. *Let M_ρ^{MV} is a random variable that outputs the margin of the multiview weighted majority vote on the example $(\mathbf{x}, y)$ drawn from distribution $\mathcal{D}$, given by:*

$$M_\rho(\mathbf{x}, y) = \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} y\, h(x^v).$$

The first and second statistical moments of the margin are respectively given by:

$$\mu_1(M_\rho^{\mathcal{D}}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} M_\rho(\mathbf{x}, y). \tag{12}$$

and,

$$\mu_2(M_\rho^{\mathcal{D}}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \big[M_\rho(\mathbf{x}, y)\big]^2 = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_X} y^2 \left[\mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} h(x^v)\right]^2 = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_X} \left[\mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} h(x^v)\right]^2. \tag{13}$$

According to this definition, the risk of the multiview weighted majority vote can be rewritten as follow:

$$R_{\mathcal{D}}(B_\rho^{\text{MV}}) = \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}\big(M_\rho(\mathbf{x}, y) \le 0\big).$$

Moreover, the risk of the multiview Gibbs classifier can be expressed thanks to the first statistical moment of the margin. Note 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 therefore

$$\begin{aligned} R_{\mathcal{D}}(G_\rho^{\text{MV}}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} \mathbb{1}_{[h(x^v)\neq y]} \\ &= \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} y\, h(x^v)\right) \\ &= \frac{1}{2}(1 - \mu_1(M_\rho^{\mathcal{D}}))\,. \end{aligned} \tag{14}$$

Similarly, the expected disagreement can be expressed thanks to the second statistical moment of the margin by

$$\begin{aligned} d_{\mathcal{D}}^{\text{MV}}(\rho) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{v'\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} \mathop{\mathbf{E}}_{h'\sim Q_{v'}} \mathbb{1}_{[h(x^v)\neq h'(x^{v'})]} \\ &= \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{v'\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} \mathop{\mathbf{E}}_{h\sim Q_{v'}} h(x^v) \times h'(x^{v'})\right) \\ &= \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \left[\mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} h(x^v)\right] \times \left[\mathop{\mathbf{E}}_{v'\sim\rho} \mathop{\mathbf{E}}_{h'\sim Q_{v'}} h'(x^{v'})\right]\right) \\ &= \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{X}}} \left[\mathop{\mathbf{E}}_{v\sim\rho} \mathop{\mathbf{E}}_{h\sim Q_v} h(x^v)\right]^2\right) \\ &= \frac{1}{2}(1 - \mu_2(M_\rho^{\mathcal{D}}))\,. \end{aligned} \tag{15}$$

From above, we can easily deduce that $0 \le d_{\mathcal{D}}^{\text{MV}}(\rho) \le 1/2$ as $0 \le \mu_2(M_\rho^{\mathcal{D}}) \le 1$. Therefore, the variance of the margin can be written as:

$$\begin{aligned} \text{Var}(M_\rho^{\mathcal{D}}) &= \mathop{\mathbf{Var}}_{(\mathbf{x},y)\sim\mathcal{D}}(M_\rho(\mathbf{x}, y)) \\ &= \mu_2(M_\rho^{\mathcal{D}}) - (\mu_1(M_\rho^{\mathcal{D}}))^2. \end{aligned} \tag{16}$$

The proof of the $\mathcal{C}$-bound

Proof of Equation (6) [13]. By making use of one-sided Chebyshev inequality (Lemma 6 of Appendix E), with $X = -M_\rho(\mathbf{x}, y)$, $\mu = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}(M_\rho(\mathbf{x}, y))$ and $a = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} M_\rho(\mathbf{x}, y)$, we have

$$
\begin{aligned}
R_{\mathcal{D}}(B^{\text{MV}}_\rho) &= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}\big(M_\rho(\mathbf{x}, y) \leq 0\big) \\
&= \mathop{\mathbb{P}}_{(\mathbf{x},y)\sim\mathcal{D}}\left(- M_\rho(\mathbf{x}, y) + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} M_\rho(\mathbf{x}, y) \geq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} M_\rho(\mathbf{x}, y)\right) \\
&\leq \frac{\mathop{\mathbf{Var}}_{(\mathbf{x},y)\sim\mathcal{D}}(M_\rho(\mathbf{x}, y))}{\mathop{\mathbf{Var}}_{(\mathbf{x},y)\sim\mathcal{D}}(M_\rho(\mathbf{x}, y)) + \left(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} M_\rho(\mathbf{x}, y)\right)^2} \\
&= \frac{\text{Var}(M^{\mathcal{D}}_\rho)}{\mu_2(M^{\mathcal{D}}_\rho) - \left(\mu_1(M^{\mathcal{D}}_\rho)\right)^2 + \left(\mu_1(M^{\mathcal{D}}_\rho)\right)^2} \\
&= \frac{\text{Var}(M^{\mathcal{D}}_\rho)}{\mu_2(M^{\mathcal{D}}_\rho)} \\
&= \frac{\mu_2(M^{\mathcal{D}}_\rho) - \left(\mu_1(M^{\mathcal{D}}_\rho)\right)^2}{\mu_2(M^{\mathcal{D}}_\rho)} \\
&= 1 - \frac{\left(\mu_1(M^{\mathcal{D}}_\rho)\right)^2}{\mu_2(M^{\mathcal{D}}_\rho)} \\
&= 1 - \frac{\left(1 - 2\, R_{\mathcal{D}}(G^{\text{MV}}_\rho)\right)^2}{1 - 2\, d^{\text{MV}}_{\mathcal{D}}(\rho)}
\end{aligned}
$$

□

Appendix B Proof of the First Multiview Change of Measure (Lemma 2)

The proof of Theorem 2 (presented in the paper) relies on the change of measure inequality of Lemma 2.

Proof of Lemma 2. We have:

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

According to the Kullback-Leibler definition, we have:

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

By applying Jensen's inequality (see Lemma 4 in Appendix E) on the concave function $z \mapsto \ln(z)$, we

have:

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

We apply again the Jensen's inequality on $z\mapsto\ln(z)$:

$$\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}\phi(h)\leq\mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v)+\mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{\phi(h)}\right). \tag{17}$$

□

Appendix C Proof of the Second Multiview Change of Measure (Lemma 3)

The proof of Theorem 3 (presented in the paper) relies on the change of measure inequality of Lemma 3.

Proof of Lemma 3. Let's start the proof from above Equation (17). We have:

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

Thanks to Hölder's inequality (Lemma 7 in Appendix E), for all $p,q>0$ *s.t.* $\frac{1}{p}+\frac{1}{q}=1$, we have:

$$\begin{aligned}\mathop{\mathbf{E}}_{v\sim\rho}\mathop{\mathbf{E}}_{h\sim Q_v}\phi(h) &\leq \left[\mathop{\mathbf{E}}_{v\sim\pi}\left(\frac{\rho(v)}{\pi(v)}\right)^q\right]^{\frac{1}{q}}\left[\mathop{\mathbf{E}}_{v\sim\pi}\left(\mathrm{KL}(Q_v\|P_v)\right)^p\right]^{\frac{1}{p}}+\mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{\phi(h)}\right)\\ &\leq \beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}}+\mathrm{KL}(\rho\|\pi)+\ln\left(\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v}e^{\phi(h)}\right),\end{aligned}$$

where

$$\forall q>0,\quad \beta_q(\rho\|\pi)=\left[\mathop{\mathbf{E}}_{v\sim\pi}\left(\frac{\rho(v)}{\pi(v)}\right)^q\right]^{\frac{1}{q}}.$$

□

Appendix D Specialization to Usual PAC-Bayesian Approaches

In this section, we provide specialization of our general theorems to the most popular PAC-Bayesian approaches [7, 18, 24, 30].

D.1 A Catoni-Like Theorem—Proof of Corollary 1

Now, we present the proof of Corollary 1 presented in Section 3.3, that follows the point of view of Catoni [7].

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 distributions π over $\mathcal{V}$, for any $p,q>0$ such that $\frac{1}{q}+\frac{1}{p}=1$, for*

all $C > 0$, for any $\delta \in (0,1]$, with a probability at least $1-\delta$ over the random choice of $S \sim (D)^m$ for all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior ρ distributions, we have:

$$R_{\mathcal{D}}(G_\rho^{\mathrm{MV}}) \;\leq\; \frac{1}{1-e^{-C}}\left(1-\exp\left[-\left(C\left(\frac{1}{2}d_{\mathcal{S}}^{\mathrm{MV}}(\rho)+e_{\mathcal{S}}^{\mathrm{MV}}(\rho)\right)+\frac{1}{m}\left[\mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v)+\mathrm{KL}(\rho\|\pi)+\ln\tfrac{1}{\delta}\right]\right)\right]\right),$$

and

$$R_{\mathcal{D}}(G_\rho^{\mathrm{MV}}) \;\leq\; \frac{1}{1-e^{-C}}\left(1-\exp\left[-\left(C.\left(\frac{1}{2}d_{\mathcal{S}}^{\mathrm{MV}}(\rho)+e_{\mathcal{S}}^{\mathrm{MV}}(\rho)\right)+\frac{1}{m}\left[\beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}}+\mathrm{KL}(\rho\|\pi)+\ln\tfrac{1}{\delta}\right]\right)\right]\right).$$

Proof. The result comes from Theorems 2 and 3 by taking $D(a,b)=\mathcal{F}(b)-Ca$, for a convex $\mathcal{F}$ and $C>0$, and by upper-bounding $\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m}\mathop{\mathbf{E}}_{v\sim\pi}\mathop{\mathbf{E}}_{h\sim P_v} e^{mD(R_S(h),R_{\mathcal{D}}(h))}$. 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:

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

The corollary is obtained with $\mathcal{F}(p)=\ln\frac{1}{(1-p[1-e^{-C}])}$. □

D.2 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 [18, 30]. Concretely, we apply Theorems 2 and 3 with:

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

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 views $\mathcal{V}$, for any $p,q>0$ such that $\frac{1}{q}+\frac{1}{p}=1$, for any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the random choice of $S\sim(D)^m$ for all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior ρ distributions, we have:*

$$\mathrm{kl}\left(\tfrac{1}{2}d_S^{\mathrm{MV}}(\rho)+e_S^{\mathrm{MV}}(\rho),R_{\mathcal{D}}(G_\rho^{\mathrm{MV}})\right)\leq\frac{1}{m}\left[\mathop{\mathbf{E}}_{v\sim\rho}\mathrm{KL}(Q_v\|P_v)+\mathrm{KL}(\rho\|\pi)+\ln\frac{\xi(m)}{\delta}\right],$$

and

$$\mathrm{kl}\left(\tfrac{1}{2}d_S^{\mathrm{MV}}(\rho)+e_S^{\mathrm{MV}}(\rho),R_{\mathcal{D}}(G_\rho^{\mathrm{MV}})\right)\leq\frac{1}{m}\left[\beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi}\mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}}+\mathrm{KL}(\rho\|\pi)+\ln\frac{\xi(m)}{\delta}\right],$$

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

Proof. The result follows from Theorems 2 and 3 by taking $D(a,b) = \mathrm{kl}(a,b)$, and upper-bounding $\mathop{\mathbf{E}}_{S\sim(\mathcal{D})^m} \mathop{\mathbf{E}}_{v\sim\pi} \mathop{\mathbf{E}}_{h\sim P_v} e^{m\ \mathrm{kl}(R_S(h),R_{\mathcal{D}}(h))}$. 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:

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

□

D.3 A McAllester-Like Theorem.

In order to "simplify" the above Corollary 2, we can make use of Pinsker's inequality,

$$2(a-b)^2 \le \mathrm{kl}(a,b), \tag{18}$$

to derive a theorem through the point of view of McAllester [25].

Corollary 3. *Let $V \ge 2$ 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 any $p,q>0$ such that $\frac{1}{q}+\frac{1}{p}=1$, for any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the random choice of $S\sim(D)^m$ for all posterior $\{Q_v\}_{v=1}^V$ and hyper-posterior ρ distributions, we have:*

$$R_{\mathcal{D}}(G_\rho^{\mathrm{MV}}) \ \le \ \tfrac{1}{2}d_S^{\mathrm{MV}}(\rho) + e_S^{\mathrm{MV}}(\rho) + \sqrt{\frac{1}{2m}\left[\mathop{\mathbf{E}}_{v\sim\rho} \mathrm{KL}(Q_v\|P_v) + \mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(m)}{\delta}\right]},$$

and

$$R_{\mathcal{D}}(G_\rho^{\mathrm{MV}}) \ \le \ \tfrac{1}{2}d_S^{\mathrm{MV}}(\rho) + e_S^{\mathrm{MV}}(\rho) + \sqrt{\frac{1}{2m}\left[\beta_q(\rho\|\pi)\left[\mathop{\mathbf{E}}_{v\sim\pi} \mathrm{KL}(Q_v\|P_v)^p\right]^{\frac{1}{p}} + \mathrm{KL}(\rho\|\pi) + \ln\frac{\xi(m)}{\delta}\right]}.$$

Proof. Directly derived from Corollary 2, and the Pinkser's Inequality of Equation (18). □

Appendix E Mathematical Tools

Lemma 4 (Jensen's inequality)**.** *For any random variable X and any concave function g, we have:*

$$g(\mathbf{E}[X]) \ \ge \ \mathbf{E}[g(X)].$$

Lemma 5 (Markov's inequality)**.** *For any random variable X s.t. $\mathbf{E}(X)=\mu$, and for any $a>0$, we have:*

$$\mathbb{P}(|X| \ge a) \le \frac{\mu}{a}.$$

Lemma 6 (One-sided Chebyshev inequality)**.** *For any random variable X s.t. $\mathbf{E}(X)=\mu$ and $\mathbf{Var}(X) = \sigma^2$, and for any $a>0$, we have:*

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

Lemma 7 (Hölder's inequality)**.** *For all $p,q>0$ such that $\frac{1}{p}+\frac{1}{q}=1$, we have:*

$$\int_a^b |f(x)g(x)|dx \le \left[\int_a^b |f(x)|^p dx\right]^{\frac{1}{p}} \left[\int_a^b |g(x)|^q dx\right]^{\frac{1}{q}}.$$

References

[1] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *ArXiv e-prints*, 2015. URL `http://arxiv.org/abs/1506.04091`.

[2] 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.

[3] 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.

[4] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, pages 463–482, 2002.

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

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

[7] Olivier Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

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

[9] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, 2000.

[10] 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.

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

[12] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *ICML*, pages 738–746, 2013.

[13] 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.

[14] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A New PAC-Bayesian Perspective on Domain Adaptation. In *ICML*, 2016.

[15] Jing Jiang. A literature survey on domain adaptation of statistical classifiers, 2008. `http://sifaka.cs.uiuc.edu/jiang4/domainadaptation/survey`.

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

[17] 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.

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

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

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

[21] Franois Laviolette, Mario Marchand, and Jean-Francis Roy. A column generation bound minimization approach with PAC-Bayesian generalization guarantees. In *AISTAT*, 2016.

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

[23] Anna Margolis. A literature review of domain adaptation with unlabeled data, 2011.

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

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

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

[27] Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-bayes bounds with data dependent priors. *JMLR*, 13:3507–3531, 2012.

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

[29] M. Re and G. Valentini. Ensemble methods: a review. *Advances in machine learning and data mining for astronomy*, pages 563–582, 2012.

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

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

[32] Shiliang Sun. A survey of multi-view machine learning. *Neural Comput Appl*, 23(7-8):2031–2038, 2013.

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