



HAL
open science

PAC-Bayesian Domain Adaptation Bounds for Multi-view learning

Mehdi Hennequin, Khalid Benabdeslem, Haytham Elghazel

► **To cite this version:**

Mehdi Hennequin, Khalid Benabdeslem, Haytham Elghazel. PAC-Bayesian Domain Adaptation Bounds for Multi-view learning. PAC-Bayes Meets Interactive Learning, Jul 2023, Hawaii-Honolulu, United States. hal-04382619

HAL Id: hal-04382619

<https://hal.science/hal-04382619>

Submitted on 9 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PAC-Bayesian Domain Adaptation Bounds for Multi-view learning

Mehdi Hennequin^{1,2} Khalid Benabdeslem¹ Haytham Elghazel¹

Abstract

This paper presents a series of new results for domain adaptation in the multi-view learning setting. The incorporation of multiple views in the domain adaptation was paid little attention in the previous studies (Germain et al., 2013; 2015a). In this way, we propose an analysis of generalization bounds with Pac-Bayesian theory to consolidate the two paradigms, which are currently treated separately.

1. Introduction

The objective of training predictive models is to train algorithms that fit training data well but generalize well on different distributions (e.g. test sets). In fact, in the theory of statistical learning, the strong hypothesis that training and test data are to be drawn from the same probability distribution (Gareth et al., 2013). However, this assumption is often too restrictive to be used in practice or in many real-life applications. Indeed, a hypothesis is learned and deployed in different and significantly changing environments. Due to that, we obtain a shift in the data distributions. A typical solution for addressing this issue is to retrain the models. This process of retraining can result in both time and financial expenses. Thus, we need to design methods for adapting a model from learning (source) data to test (target) data. In the field of machine learning, this scenario is commonly known as *domain adaptation (DA)* or *covariate shift* (Moreno-Torres et al., 2012). Essentially, DA techniques aim to address the challenge of learning when the learning task is the same but the domains exhibit variations in their feature spaces or marginal conditional probabilities.

On the other hand, data can be expressed through multiple independent feature sets, as stated in Xu et al. (2013). As a result, the data can be partitioned into independent groups,

*Equal contribution ¹Université Lyon 1, LIRIS, UMR CNRS 5205, F-69622, France ²Galilé Group, 8 Bd de la République 71100 Chalon-sur-Saône, France. Correspondence to: Mehdi Hennequin <mehdi.hennequin@univ-lyon1.fr>, Khalid Benabdeslem <khalid.benabdeslem@univ-lyon1.fr>, Haytham Elghazel <haytham.elghazel@univ-lyon1.fr>.

known as views (Xu et al., 2013). In domain adaptation, these views are usually merged into a single view to align with the learning objective. Nonetheless, this process of merging can potentially result in negative transfer, as highlighted in Zhang et al. (2020), whereby the integration of each view’s unique statistical characteristics could lead to the introduction of unwanted data or knowledge. We find little research on *multi-view domain adaptation (MVDA)* (Yang et al., 2012; Hennequin et al., 2022a;b; Munro et al., 2021) where considerable attention has been paid to algorithm, while analysis of generalization bound remains largely understudied. In this way, we propose a theoretical analysis by means of Pac-Bayesian theory, with the aim of unifying the two paradigms that have conventionally been treated as separate entities (Germain et al., 2013; 2015a; Goyal et al., 2017).

2. Related Works

In this section, we present theoretical studies of multi-view learning and domain adaptation related to the Pac-Bayesian theory. First, we introduce the theoretical concepts needed for the following sections. In a second phase, we present the work done by Germain et al. (2013) on the PAC-Bayesian domain adaptation. Finally, we present the works achieved to Pac-Bayesian theory and multi-view learning.

2.1. Notation and Assumptions

This section introduces the definitions and concepts needed for the following sections. Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$, denote respectively input space of dimension d and output space. In the *scenario of unsupervised multi-view domain adaptation (UMVDA)* we consider, the learner receives two samples: a labeled sample from a source domain \mathbb{D}_S , defined by a distribution Q over $\mathcal{X} \times \mathcal{Y}$; and unlabelled sample according to the target domain \mathbb{D}_T , defined by a distribution P over $\mathcal{X} \times \mathcal{Y}$; $Q_{\mathcal{X}}, P_{\mathcal{X}}$ being the respective marginal distributions over \mathcal{X} . We denote by $\mathbb{S} = \{(x_i, y_i)\}_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$ the labeled sample of size m received from the source domain, which is drawn i.i.d. from Q . In addition, we consider that the data instances can be represented or partitioned in V different views. More formally, for $v \in \{1, \dots, V\}$, and $V \geq 2$ is the number of views of not-necessarily the same dimension. Through-

out the paper, we will abbreviate $v \in \{1, \dots, V\}$ with $v \in \llbracket \mathcal{V} \rrbracket$. The labeled samples multi-view is defined as follows, $\forall v \in \llbracket \mathcal{V} \rrbracket$, $\mathbb{S} = \{(x_i^v, y_i)\}_{i=1}^m \in (\mathcal{X}^v \times \mathcal{Y})^m$, with $\{x_i^v\}_{i=1}^m$ supposed to be drawn i.i.d. according to distribution Q . Note that the multi-view observations $\{x_i\}_{i=1}^m = \{(x_i^{(1)}, \dots, x_i^{(v)})\}_{i=1}^m$ belong to a multi-view input set $\mathcal{X} = \mathcal{X}^1 \times \dots \times \mathcal{X}^v$, with $\mathcal{X}^v \in \mathbb{R}^{d_v}$, and d_v denote the dimension of the v^{th} view, where $d = d_1 \times \dots \times d_v$. In the same way, we define unlabeled samples from the target domain, $\forall v \in \llbracket \mathcal{V} \rrbracket$, $\mathbb{T}_{\mathcal{X}} = \{x_i^{v'}\}_{i=1}^n$ of size n drawn i.i.d. according to $P_{\mathcal{X}}$ (note that, $\mathbb{T} = \{(x_i^{v'}, y_i)\}_{i=1}^n$ drawn i.i.d. according to P). In our context, we consider that we have no labels in the target domain, however we have prior knowledge about the views in both domains.

We define a hypothesis class \mathcal{H} of hypotheses $h : \mathcal{X} \rightarrow \mathcal{Y}$. Besides, for the concept of multi-view learning, we consider for each view $v \in \llbracket \mathcal{V} \rrbracket$, a set \mathcal{H}_v of hypothesis $h : \mathcal{X}_v \rightarrow \mathcal{Y}$. The expected source risk or true source risk of $h \in \mathcal{H}$ over the distribution Q , are the probability that h errs on the entire source domain, $\mathcal{R}_Q(h) = \mathbb{E}_{(x,y) \sim Q}[\mathcal{L}_{01}(h(x), y)]$, where $\mathcal{L}_{01}(a, b) = \mathbb{I}[a \neq b]$ is the 0-1-loss function and where $\mathbb{I}[a \neq b]$ is the indicator function which returns 1 if $a \neq b$ and 0 otherwise. For any two functions $(h, h') \in \mathcal{H}$, we denote by $\mathcal{R}_{Q_{\mathcal{X}}}(h, h')$ the expected disagreement of $h(x)$ and $h'(x)$, which measures the probability that h and h' do not agree on the entire marginal distributions $Q_{\mathcal{X}}$ over \mathcal{X} , $\mathcal{R}_{Q_{\mathcal{X}}}(h, h') = \mathbb{E}_{x \sim Q_{\mathcal{X}}}[\mathcal{L}_{01}(h(x), h'(x))]$. The empirical source risk $\mathcal{R}_{\mathbb{S}}(h)$ for a given hypothesis $h \in \mathcal{H}$ and a training sample $\mathbb{S} = \{(x_i, y_i)\}_{i=1}^m$ where each example is drawn i.i.d. from Q is defined as, $\mathcal{R}_{\mathbb{S}}(h) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}_{01}(h(x_i), y_i)$. In the same way, we define the empirical source disagreement by $\mathcal{R}_{\mathbb{S}_{\mathcal{X}}}(h, h') = \frac{1}{m} \sum_{i=1}^m \mathcal{L}_{01}(h(x_i), h'(x_i))$, where $\mathbb{S}_{\mathcal{X}} = \{(x_i)\}_{i=1}^m$ where each example is drawn i.i.d. from $Q_{\mathcal{X}}$. The expected target risk $\mathcal{R}_P(\cdot)$ over P , the expected target disagreement $\mathcal{R}_{P_{\mathcal{X}}}(\cdot, \cdot)$ over $P_{\mathcal{X}}$, the empirical target risk $\mathcal{R}_{\mathbb{T}}(\cdot)$ over P , the empirical target disagreement $\mathcal{R}_{\mathbb{T}_{\mathcal{X}}}(\cdot, \cdot)$ over $P_{\mathcal{X}}$ are defined in a similar way.

2.2. Simple Pac-Bayesian Bounds

The PAC-Bayesian approach abbreviated Pac-Bayes is analysis techniques of generalization in the theory of statistical learning. PAC-Bayes inequalities were introduced by [Shawe-Taylor & Williamson \(1997\)](#), and [McAllester \(1998; 1999\)](#); and further formalised [Catoni \(2007; 2004b;a\)](#) and other (see [Guedj \(2019\)](#) for a recent survey and [Alquier \(2021\)](#) for an introduction to the field). It provides PAC (probably approximately correct, Valiant, 1984) generalization bounds by expressing a trade-off between the empirical risk on the training set and a measure of complexity of the predictors class as a weighted majority vote over a set of functions from the hypothesis space \mathcal{H} .

In this section, we recall the general PAC-Bayesian generalization bounds in the setting of binary classification with the 0-1-loss defined in the above section. To derive such a generalization bound, one assumes a prior distribution \mathcal{P} over \mathcal{H} , which models an a priori belief on the hypothesis from \mathcal{H} before the observation of the training sample $\mathbb{S} \sim Q^m$. Given the training sample \mathbb{S} , the learner aims at finding a posterior distribution \mathcal{Q} over \mathcal{H} that leads to a well-performing \mathcal{Q} -weighted majority vote $\mathcal{B}_{\mathcal{Q}}$ (also called the bayes classifier) defined as: $\mathcal{B}_{\mathcal{Q}}(x) = \text{sign} \left[\mathbb{E}_{h \sim \mathcal{Q}} h(x) \right]$.

We want to learn \mathcal{Q} over \mathcal{H} such that it minimizes the true risk $\mathcal{R}_Q(\mathcal{B}_{\mathcal{Q}})$ of the \mathcal{Q} -weighted majority vote. However, the risk of $\mathcal{B}_{\mathcal{Q}}$ is known to be NP-hard, therefore PAC-Bayesian generalization bounds do not directly focus on the risk of $\mathcal{B}_{\mathcal{Q}}$. Instead, it gives an upper bound over the expectation over \mathcal{Q} of all the individual hypothesis true risk called the expected/true Gibbs risk: $\mathcal{R}_Q(G_{\mathcal{Q}}) = \mathbb{E}_{h \sim \mathcal{Q}} [\mathcal{R}_Q(h)]$.

The expected Gibbs risk is closely related to the deterministic \mathcal{Q} -weighted majority vote. Indeed, if $\mathcal{B}_{\mathcal{Q}}(\cdot)$ misclassifies $x \in \mathcal{X}$, then at least half of the classifiers (under measure \mathcal{Q}) make a prediction error on x . Therefore, we have $\mathcal{R}_Q(\mathcal{B}_{\mathcal{Q}}) \leq 2 \mathcal{R}_Q(G_{\mathcal{Q}})$. Another result on the relation between $\mathcal{R}_Q(\mathcal{B}_{\mathcal{Q}})$ and $\mathcal{R}_Q(G_{\mathcal{Q}})$ know as C-bound ([Lacasse et al., 2006](#)).

The PAC-Bayesian theory, suggests that minimizing the expected Gibbs risk $\mathcal{R}_Q(G_{\mathcal{Q}})$ can be done by minimizing the trade-off between the empirical Gibbs risk $\mathcal{R}_{\mathbb{S}}(G_{\mathcal{Q}})$ and Kullback-Leibler divergence minimization $D_{KL}(\mathcal{Q} \parallel \mathcal{P})$. Note that PAC-Bayesian generalization bounds do not directly take into account the complexity of the hypothesis class \mathcal{H} , but measure the deviation between the prior distribution \mathcal{P} and the posterior distribution \mathcal{Q} on \mathcal{H} through the Kullback-Leibler divergence. In the literature, we find three main PAC-Bayesian bound proposed by [McAllester \(1999\)](#); [Seeger \(2002\)](#); [Langford \(2005\)](#) and [Catoni \(2007\)](#).

2.3. Analysis of Domain Adaptation Pac-Bayesian Bounds

In this section, we recall the work done by [Germain et al. \(2013; 2015a\)](#) on how the PAC-Bayesian theory can help to theoretically understand domain adaptation through the weighted majority vote learning point of view.

The first Pac-Bayesian generalization bound for domain adaptation was introduced in [Germain et al. \(2013\)](#). The authors defined a divergence measure that follows the idea of C-bound. Thus, [Germain et al.](#) underlined that the domains $\mathbb{D}_{\mathcal{S}}$ and $\mathbb{D}_{\mathcal{T}}$ are close according to \mathcal{Q} if the expected disagreement over the two domains tends to be close. More formally, if $\mathcal{R}_Q(G_{\mathcal{Q}})$ and $\mathcal{R}_P(G_{\mathcal{Q}})$ are similar, then

and $\mathcal{R}_Q(\mathcal{B}_Q)$ and $\mathcal{R}_P(\mathcal{B}_Q)$ are similar when $d_{Q_X}(\mathcal{Q})$ and $d_{P_X}(\mathcal{Q})$ are also similar. In this way, the authors introduced the following domain disagreement pseudometric ¹.

Note that $dis_Q(\cdot, \cdot)$ is symmetric and fulfills the triangle inequality (Germain et al., 2013). Note that for the sake of simplicity, we suppose that $m = n$, i.e., the size of \mathbb{S}/\mathbb{S}_X and \mathbb{T}/\mathbb{T}_X are equal. The authors showed that $dis_Q(Q_X, P_X)$ can be bounded in terms of the classical PAC-Bayesian quantities and propose the following theorem (the theorem and its proof can be found in Germain et al. (2013; 2015a)).

Note that the authors in Germain et al. (2013; 2015a) propose the theorem as "Catoni's type". It present interesting characteristics. First, its minimization is closely related to the minimization problem associated with the SVM when \mathcal{Q} is an isotropic Gaussian over the space of linear classifiers (Germain et al., 2009). Second, the value $C = 2\alpha$ allows to control the trade-off between the empirical risk and the complexity term $D_{KL}(\cdot || \cdot)$.

Thereby, from this domain's divergence, the authors proved the following domain adaptation bound (the theorem and its proof can be found in Germain et al. (2015a)).

In this section, we presented the principal bounds for Pac-Bayesian domain adaptation, in the next section we will discuss about Pac-Bayesian bounds in the multi-view learning setting.

2.4. Analysis of Pac-Bayesian Multi-view Bounds

First, the authors in Sun et al. (2017) provided PAC-Bayesian bounds over the concatenation of the views, using priors that reflect how well the views agree on average over all examples, and deduced a SVM-like learning algorithm from this framework. However, this concatenation is designed for two views and kernel method, it is not generalizable to other methods. A more general framework of Pac-Bayesian bounds for multi-views was introduced in Goyal et al. (2017). In the paper, the authors introduced the two-level multiview approach. For each view $v \in \llbracket \mathcal{V} \rrbracket$, they consider a view-specific set \mathcal{H}_v of voters $h : \mathcal{X}^v \rightarrow Y$, and a prior distribution \mathcal{P}_v on \mathcal{H}_v . Given a hyper-prior distribution π over the views $\llbracket \mathcal{V} \rrbracket$. In the paper, PAC-Bayesian learner objective has two parts. The first part is finding a posterior distribution \mathcal{Q}_v over $\mathcal{H}_v, \forall v \in \llbracket \mathcal{V} \rrbracket$; The second is finding a hyper-posterior ρ distribution on the set of views $\llbracket \mathcal{V} \rrbracket$. Thereby, Goyal et al. (2017) defined the multi-view weighted majority vote $\mathcal{B}_\rho^{MV}(x) = \text{sign} \left[\mathbb{E}_{v \sim \rho} \mathbb{E}_{h \sim \mathcal{Q}_v} h(x^v) \right]$.

Thus, the authors propose to build a learner that intends to construct posterior and hyperposterior distributions that minimize the actual risk $\mathcal{R}_D(\mathcal{B}_\rho^{MV})$ of the mul-

tiview weighted majority vote defined as: $\mathcal{R}_Q(\mathcal{B}_\rho^{MV}) = \mathbb{E}_{(x^v, y) \sim Q} \left[\mathcal{L}_{0-1}(\mathcal{B}_\rho^{MV}(x^v), y) \right]$.

The authors in the paper provide general multi-view PAC-Bayesian theorems and derive also a generalization bound with the approaches of McAllester (1998); Seeger (2002); Langford (2005); and Catoni (2007) introduce in section 2.2. The main difference between Goyal et al.'s bounds to theorems McAllester (1998); Seeger (2002); Langford (2005); and Catoni (2007) relies on the introduction of view-specific prior and posterior distributions, which mainly leads to an additional term $\mathbb{E}_{v \sim \rho} D_{KL}(\mathcal{Q}_v || \mathcal{P}_v)$, expressed as the expectation of the view-specific Kullback-Leibler divergence term over the views $\llbracket \mathcal{V} \rrbracket$ according to the hyper-posterior distribution ρ .

3. Analysis of Unsupervised Multi-view Domain Adaptation

In this section we propose to introduce the concept of multi-view learning in the DA with generalization Pac-Bayesian guarantees. Then, we adapt the divergence proposed by Germain et al., Germain et al. (2013; 2016) with the concept of multi-view weighted majority vote introduced in Goyal et al. (2017). In a second phase, we propose a Pac-Bayesian domain adaptation bound in the multi-view setting.

3.1. Multi-view Domain Disagreement

Germain et al. (2013) and Mansour et al. (2009) propose a divergence measure that is based on the expected disagreement over the two domains. In the idea of measure disagreement we propose to adapt the definition proposed by Germain et al. (2013) to multi-view learning. Thus, we define the multi-view domain disagreement as follows:

Definition 3.1. (Multi-view domain disagreement) $\forall v \in \llbracket \mathcal{V} \rrbracket$, for any set of voters \mathcal{H}_v for any marginal distributions Q_X and P_X over \mathcal{X} , any set of posterior distribution $\{\mathcal{Q}_v\}_{v=1}^V$ on \mathcal{H}_v , for any hyper-posterior distribution ρ over $\llbracket \mathcal{V} \rrbracket$, the multi-view domain disagreement $dis_\rho^{MV}(Q_X, P_X)$ between Q_X and P_X is defined by:

$$dis_\rho^{MV}(Q_X, P_X) = \left| d_{P_X}^{MV}(\rho) - d_{Q_X}^{MV}(\rho) \right|, \quad (1)$$

where $d_{Q_X}^{MV}(\rho), d_{P_X}^{MV}(\rho)$, are expected disagreement defined in Goyal et al. (2017).

Theorem 3.2. $\forall v \in \llbracket \mathcal{V} \rrbracket$, for any distributions Q and P over $\mathcal{X} \times \mathcal{Y}$, for any set of voters \mathcal{H}_v , for any marginal distributions Q_X and P_X over \mathcal{X} , any set of posterior distribution $\{\mathcal{Q}_v\}_{v=1}^V$ on \mathcal{H}_v , for any hyper-posterior distribution ρ over $\llbracket \mathcal{V} \rrbracket$, we have:

$$\mathcal{R}_P(G_\rho^{MV}) \leq \mathcal{R}_Q(G_\rho^{MV}) + \frac{1}{2} dis_\rho^{MV}(Q_X, P_X) + \lambda_\rho, \quad (2)$$

¹A pseudometric d is a metric for which the property $d(x, y) = 0 \Leftrightarrow x = y$ is relaxed to $d(x, y) = 0 \Leftarrow x = y$

where λ_ρ is the deviation between the expected joint errors between pairs for voters ad pairs of views defined in section 2.4 on the target and source domains, which is defined as

$$\lambda_\rho = \left| e_P^{\text{MV}}(\rho) - e_Q^{\text{MV}}(\rho) \right| \text{ and where}$$

$$e_P^{\text{MV}}(\rho) = \mathbb{E}_{(x,y) \sim P} \mathbb{E}_{(v,v') \sim \rho^2} \mathbb{E}_{(h,h') \sim \mathcal{Q}_v^2} [\mathcal{L}_{0-1}(h(x), y) \times \mathcal{L}_{0-1}(h'(x), y)],$$

$$e_Q^{\text{MV}}(\rho) = \mathbb{E}_{(x,y) \sim Q} \mathbb{E}_{(v,v') \sim \rho^2} \mathbb{E}_{(h,h') \sim \mathcal{Q}_v^2} [\mathcal{L}_{0-1}(h(x), y) \times \mathcal{L}_{0-1}(h'(x), y)].$$

Proof. The proof borrows the straightforward proof technique of Theorem 9 in Germain et al. (2015a). \square

3.2. Specialization of multi-view domain disagreement to the Classical Approaches

In this section, we follow the same principles as Germain et al. (2009; 2015b). Selecting a well-suited deviation function Δ , we can derive easily the classical PAC-Bayesian theorem of Catoni (2007) presented in the section 2.2.

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 as $\Delta(a, b) = \mathcal{F}(b) - c$ Germain et al. (2009; 2015a;b). We obtain the following generalization bound:

Corollary 3.3. $\forall v \in \llbracket \mathcal{V} \rrbracket$, for any set of voters \mathcal{H}_v for any marginal distributions $Q_{\mathcal{X}}$ and $P_{\mathcal{X}}$ over \mathcal{X} , any set of posterior distribution $\{\mathcal{Q}_{v,\mathbb{S}}\}_{v=1}^V$ on \mathcal{H}_v , for any hyper-posterior distribution $\rho_{\mathbb{S}}$ over $\llbracket \mathcal{V} \rrbracket$, for any set of prior distributions $\{\mathcal{P}_v\}_{v=1}^V$ on \mathcal{H}_v , for any hyper-prior distribution π over $\llbracket \mathcal{V} \rrbracket$, for any $\delta \in (0, 1]$, $\forall \alpha > 0$, with probability at least $1 - \delta$, we have:

$$\mathbb{E}_{\mathbb{S} \sim Q^m} \text{dis}_{\rho_{\mathbb{S}}}^{\text{MV}}(Q_{\mathcal{X}}, P_{\mathcal{X}}) \leq \frac{2\alpha}{1 - e^{-2\alpha}} \left[\mathbb{E}_{\mathbb{S} \sim Q^m} \text{dis}_{\rho_{\mathbb{S}}}^{\text{MV}}(\mathbb{S}_{\mathcal{X}}, \mathbb{T}_{\mathcal{X}}) + \frac{\mathbb{E}_{\mathbb{S} \sim Q^m} \mathbb{E}_{v \sim \rho_{\mathbb{S}}} D_{\text{KL}}(\mathcal{Q}_{v,\mathbb{S}} \| \mathcal{P}_v) + \mathbb{E}_{\mathbb{S} \sim Q^m} D_{\text{KL}}(\rho_{\mathbb{S}} \| \pi) + \ln \sqrt{\frac{1}{\delta}}}{m \times \alpha} \right]. \quad (3)$$

Proof. The proof uses the ideas of the techniques and tricks of Germain et al. (2015a;b).

3.3. The PAC-Bayesian DA-Bound

Finally, the Theorem 3.2 leads to a PAC-Bayesian bound based on both the empirical source error of the Gibbs classifier and the empirical Multi-view domain disagreement pseudometric estimated on a source and target samples. The following bound is based on Catoni's approach 3.3:

Theorem 3.4. $\forall v \in \llbracket \mathcal{V} \rrbracket$, for any set of voters \mathcal{H}_v for any marginal distributions $Q_{\mathcal{X}}$ and $P_{\mathcal{X}}$ over \mathcal{X} , any set of posterior distribution $\{\mathcal{Q}_{v,\mathbb{S}}\}_{v=1}^V$ on \mathcal{H}_v , for any hyper-posterior distribution $\rho_{\mathbb{S}}$ over $\llbracket \mathcal{V} \rrbracket$, for any set of prior distributions $\{\mathcal{P}_v\}_{v=1}^V$ on \mathcal{H}_v , for any hyper-prior distribution π over $\llbracket \mathcal{V} \rrbracket$, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, we have:

$$\mathbb{E}_{\mathbb{S} \sim Q^m} \mathcal{R}_P(G_{\rho_{\mathbb{S}}}^{\text{MV}}) \leq \mathbb{E}_{\mathbb{S} \sim Q^m} c' \mathcal{R}_{\mathbb{S}}(G_{\rho_{\mathbb{S}}}^{\text{MV}}) + \mathbb{E}_{\mathbb{S} \sim Q^m} \alpha' \frac{1}{2} \text{dis}_{\rho_{\mathbb{S}}}^{\text{MV}}(\mathbb{S}_{\mathcal{X}}, \mathbb{T}_{\mathcal{X}}) + \left(\frac{c'}{c} + \frac{\alpha'}{\alpha} \right) \frac{\mathbb{E}_{\mathbb{S} \sim Q^m} \mathbb{E}_{v \sim \rho_{\mathbb{S}}} D_{\text{KL}}(\mathcal{Q}_{v,\mathbb{S}} \| \mathcal{P}_v) + \mathbb{E}_{\mathbb{S} \sim Q^m} D_{\text{KL}}(\rho_{\mathbb{S}} \| \pi)}{m} + \frac{\ln \frac{1}{\delta}}{m} + \lambda_\rho + \frac{1}{2}(\alpha' - 1), \quad (4)$$

where $c' = \frac{c}{1 - e^{-c}}$ and $\alpha' = \frac{2\alpha}{1 - e^{-2\alpha}}$.

Proof. In Theorem 3.2, replace $\mathbb{E}_{\mathbb{S} \sim Q^m} \mathcal{R}_Q(G_{\rho_{\mathbb{S}}}^{\text{MV}})$ and $\mathbb{E}_{\mathbb{S} \sim Q^m} \text{dis}_{\rho_{\mathbb{S}}}^{\text{MV}}(Q_{\mathcal{X}}, P_{\mathcal{X}})$ by their upper bound, obtained from Corollary 2 in Goyal et al. (2017) and Corollary 3.3. \square

4. Discussions and Conclusion

The primary contrast between our bounds 3.3; 3.4, and Germain et al. (2015a)'s bounds lies in the incorporation of view-specific prior and posterior distributions. This results in an extra term, $\mathbb{E}_{v \sim \rho} D_{\text{KL}}(\mathcal{Q}_v \| \mathcal{P}_v)$, which represents the expected value of the view-specific Kullback-Leibler divergence term over the views $\llbracket \mathcal{V} \rrbracket$, based on the hyper-posterior distribution ρ . The second difference comes from the expectation over all the possible learning samples in bounds itself (Goyal et al., 2017). In this way, the expectation $\mathbb{E}_{\mathbb{S} \sim Q^m}$ is distributed for the all terms in the bounds. Thereby, the $D_{\text{KL}}(\cdot \| \cdot)$ terms take account of the the posterior and hyper-posterior distribution $\mathcal{Q}_{v,\mathbb{S}}/\rho_{\mathbb{S}}$ outputted by a given learning algorithm after observing the learning sample \mathbb{S} .

Finally in this paper, we propose a first PAC-Bayesian analysis of weighted majority vote classifiers for domain adaptation with the concept of multi-view learning. Our works is based on theoretical results and for the future we aim to derive from introduced bounds a new domain adaptation multi-view algorithm. We will build on the work of Germain et al. (2009; 2013) to propose a specialized algorithm for linear classifiers or to propose a specialized algorithm for neural networks Sicilia et al. (2022).

References

- Alquier, P. User-friendly introduction to pac-bayes bounds, 2021. URL <https://arxiv.org/abs/2110.11216>.
- Catoni, O. A pac-bayesian approach to adaptive classification. 2004a.
- Catoni, O. *Statistical learning theory and stochastic optimization. Ecole d'été de probabilités de Saint-Flour XXXI-2001*. Springer, 2004b. URL <https://hal.science/hal-00104952>. Collection : Lecture notes in mathematics n°1851.
- Catoni, O. Pac-bayesian supervised classification: The thermodynamics of statistical learning. *Lecture Notes-Monograph Series*, 56:i-163, 2007. ISSN 07492170. URL <http://www.jstor.org/stable/20461499>.
- Gareth, J., Daniela, W., Trevor, H., and Robert, T. *An introduction to statistical learning: with applications in R*. Spinger, 2013.
- Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pp. 353–360, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553419. URL <https://doi.org/10.1145/1553374.1553419>.
- Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 738–746, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL <https://proceedings.mlr.press/v28/germain13.html>.
- Germain, P., Habrard, A., Laviolette, F., and Morvant, E. Pac-bayesian theorems for domain adaptation with specialization to linear classifiers, 2015a. URL <https://arxiv.org/abs/1503.06944>.
- Germain, P., Lacasse, A., Laviolette, F., March, M., and Roy, J.-F. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015b. URL <http://jmlr.org/papers/v16/germain15a.html>.
- Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A new pac-bayesian perspective on domain adaptation. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 859–868, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <https://proceedings.mlr.press/v48/germain16.html>.
- Goyal, A., Morvant, E., Germain, P., and Amini, M.-R. Pac-bayesian analysis for a two-step hierarchical multiview learning approach. In Ceci, M., Hollmén, J., Todorovski, L., Vens, C., and Džeroski, S. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 205–221, Cham, 2017. Springer International Publishing. ISBN 978-3-319-71246-8.
- Guedj, B. A primer on pac-bayesian learning. 2019. doi: 10.48550/ARXIV.1901.05353. URL <https://arxiv.org/abs/1901.05353>.
- Hennequin, M., Benabdeslem, K., and Elghazel, H. Adversarial multi - view domain adaptation for regression. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022a. doi: 10.1109/IJCNN55064.2022.9892148.
- Hennequin, M., Benabdeslem, K., Elghazel, H., Ranvier, T., and Michoux, E. Multi-view self-attention for regression domain adaptation with feature selection. In *Neural Information Processing: 29th International Conference, ICONIP 2022, Virtual Event, November 22–26, 2022, Proceedings, Part I*, pp. 177–188, 2022b.
- Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL <https://proceedings.neurips.cc/paper/2006/file/779efbd24d5a7e37ce8dc93e7c04d572-Paper.pdf>.
- Langford, J. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6(10): 273–306, 2005. URL <http://jmlr.org/papers/v6/langford05a.html>.
- Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL <http://www.cs.mcgill.ca/%7Ecolt2009/papers/003.pdf#page=1>.
- McAllester, D. A. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT' 98*, pp. 230–234,

- New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570. doi: 10.1145/279943.279989. URL <https://doi.org/10.1145/279943.279989>.
- McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT '99*, pp. 164–170, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131674. doi: 10.1145/307400.307435. URL <https://doi.org/10.1145/307400.307435>.
- Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012. ISSN 0031-3203. doi: <https://doi.org/10.1016/j.patcog.2011.06.019>. URL <https://www.sciencedirect.com/science/article/pii/S0031320311002901>.
- Munro, J., Wray, M., Larlus, D., Csurka, G., and Damen, D. Domain adaptation in multi-view embedding for cross-modal video retrieval. *CoRR*, abs/2110.12812, 2021. URL <https://arxiv.org/abs/2110.12812>.
- Seeger, M. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.
- Shawe-Taylor, J. and Williamson, R. C. A pac analysis of a bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT '97*, pp. 2–9, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918916. doi: 10.1145/267460.267466. URL <https://doi.org/10.1145/267460.267466>.
- Sicilia, A., Atwell, K., Alikhani, M., and Hwang, S. J. Pac-bayesian domain adaptation bounds for multiclass learners, 2022. URL <https://arxiv.org/abs/2207.05685>.
- Sun, S., Shawe-Taylor, J., and Mao, L. Pac-bayes analysis of multi-view learning. *Information Fusion*, 35:117–131, 2017. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2016.09.008>. URL <https://www.sciencedirect.com/science/article/pii/S156625351630080X>.
- Xu, C., Tao, D., and Xu, C. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013. URL <http://arxiv.org/abs/1304.5634>.
- Yang, P., Gao, W., Tan, Q., and Wong, K.-F. Information-theoretic multi-view domain adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 270–274, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL <https://aclanthology.org/P12-2053>.
- Zhang, W., Deng, L., and Wu, D. Overcoming negative transfer: A survey. *CoRR*, abs/2009.00909, 2020. URL <https://arxiv.org/abs/2009.00909>.