A New PAC-Bayesian Perspective on Domain Adaptation

Pascal Germain, Amaury Habrard, François Laviolette, Emilie Morvant

▶ **To cite this version:**

Pascal Germain, Amaury Habrard, François Laviolette, Emilie Morvant. A New PAC-Bayesian Perspective on Domain Adaptation. [Research Report] Laboratoire Hubert Curien, Université Jean Monnet, Saint-Etienne; Département d'informatique et de génie logiciel, Université Laval (Québec). 2015. hal-01163722v1

HAL Id: hal-01163722

https://hal.science/hal-01163722v1

Submitted on 15 Jun 2015 (v1), last revised 14 Mar 2016 (v3)

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.

A New PAC-Bayesian Perspective on Domain Adaptation

Pascal Germain[1] Amaury Habrard[2] François Laviolette[1] Emilie Morvant[2]
[1] Département d'informatique et de génie logiciel, Université Laval, Québec, Canada
[2] Lab. Hubert Curien, Université Jean Monnet, UMR CNRS 5516, St-Etienne, France

June 15, 2015

Abstract

We study the issue of domain adaptation: we want to adapt a model from a source distribution to a target one. We focus on models expressed as a majority vote. Our main contribution is a novel theoretical analysis of the target risk that is formulated as an upper bound expressing a trade-off between only two terms: *(i)* the voters' joint errors on the source distribution, and *(ii)* the voters' disagreement on the target one; both easily estimable from samples. Hence, this new study is more precise than other analyses that usually rely on three terms (including a hardly controllable term). Moreover, we derive a PAC-Bayesian generalization bound, and specialize the result to linear classifiers to propose a learning algorithm.

1 Introduction

Machine learning practitioners are commonly exposed to the issue of *domain adaptation*[1](Jiang, 2008; Margolis, 2011): One usually learns a model from a corpus, *i.e.*, a fixed yet unknown source distribution, and then wants to apply it on a new corpus, *i.e.*, a related but slightly different target distribution. Therefore, domain adaptation is widely studied in a lot of application fields like computer vision (Patel et al., 2015), natural language processing (Blitzer, 2007), etc. A simple example is the common spam filtering problem, where a model needs to be adapted from one user mailbox to another receiving significantly different emails.

Several approaches exist in the literature to address domain adaptation, but often with the same idea: If we are able to apply a transformation in order to "move closer" the distributions, then we can learn a model with the available labels. This process is generally performed by iterative procedures (Bruzzone & Marconcini, 2010; Chen et al., 2011), and/or by reweighting the importance of labeled data (Huang et al., 2006; Sugiyama et al., 2007; Cortes et al., 2010), and/or by minimizing a measure of divergence between the distributions (Cortes & Mohri, 2014; Germain et al., 2013). The divergence-based approach has especially been explored to derive generalization bounds for domain adaptation (Ben-David et al., 2006; 2010; Mansour et al., 2009; Zhang et al., 2012; Germain et al., 2013). Recently, this issue has been studied for the first time through the PAC-Bayesian framework (Germain et al., 2013), which focuses on learning weighted majority votes[2]. Even if their result clearly opens the door to tackle domain adaptation in a PAC-Bayesian fashion, it shares the same philosophy than the seminal works of Ben-David et al. (2006; 2010) and Mansour et al. (2009): The error of the target model is upper-bounded by a trade-off between the error of the model on the source distribution, the divergence between the marginal distributions, and a non-estimable term related to the ability to adapt in the current space.

In this paper, we derive a novel domain adaptation bound that is expressed as a simpler trade-off: The error of the target model is upper-bounded by the half of the voters' disagreement on the target distribution, and the voters' joint errors on the source distribution weighted by a divergence between the source and the target distributions. In other words, this leads to an original bound where the relation between the source and target distributions weights directly the trade-off as opposed to an additional term. Another crucial point is that this relationship can be dealt as a constant when no labeled information

[1]Domain adaptation is associated with *transfer learning* (Pan & Yang, 2010; Quionero-Candela et al., 2009).

[2]This setting is not too restrictive since many machine learning approaches can be seen as a majority vote learning. Think for example to ensemble learning, or to support vector machines which output classifiers that can be interpreted as majority votes.

is available in the target distribution and thus seen as a hyperparameter to control the trade-off. Along with this original domain adaptation bound, we provide a PAC-Bayesian generalization bound to justify its empirical minimization. Finally, we specialize it to linear classifiers to design DALC, an algorithm that clearly improves the performances of the first PAC-Bayesian domain adaptation algorithm proposed by Germain et al. (2013) on a popular dataset in the domain adaptation community.

The rest of the paper is organized as follows. Section 2 presents the PAC-Bayesian domain adaptation setting, for which we recall the seminal results of Germain et al. (2013) in Section 3. Section 4 deals with our new theoretical results leading to a new domain adaptation algorithm in Section 5. Before concluding, we experiment this latter in Section 6.

2 PAC-Bayesian Domain Adaptation Setting and Notations

The PAC-Bayesian theory was first introduced by McAllester (1999). In this paper, we stand in the following domain adaptation PAC-Bayesian setting studied for the first time by Germain et al. (2013).

We tackle *domain adaptation binary classification tasks* from a d-dimensional input space $\mathbf{X} \subseteq \mathbb{R}^d$ to the output space $Y = \{-1, 1\}$. Our objective is to perform domain adaptation from a distribution $\mathcal{S}$—the *source domain*—to another but related distribution $\mathcal{T}$—the *target domain*—on $\mathbf{X} \times Y$. The associated marginal distributions on $\mathbf{X}$ are respectively denoted by $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$. Given a distribution $\mathcal{D}$, we denote $(\mathcal{D})^m$ the distribution of a m-sample constituted by m elements *i.i.d.* from $\mathcal{D}$. We consider the *unsupervised domain adaptation* setting in which the algorithm is provided with a *labeled source* m_s*-sample* $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_s} \sim (\mathcal{S})^{m_s}$, and with an *unlabeled target* m_t*-sample* $T = \{\mathbf{x}_i\}_{i=1}^{m_t} \sim (\mathcal{T}_{\mathbf{X}})^{m_t}$. Note that all the results presented in this paper are still true for (semi-)supervised domain adaptation, *i.e.*, when we have access to (some) target labels.

Let $\mathcal{H}$ be a set of voters $h : \mathbf{X} \to Y$. Given $\mathcal{H}$, the ingredients of the *PAC-Bayesian domain adaptation* approach are a *prior* distribution π on $\mathcal{H}$, a pair of source-target learning samples (S, T) and a *posterior* distribution ρ on $\mathcal{H}$. The prior distribution π models an *a priori* belief—*i.e.*, before observing (S, T)—of the voters' accuracy. Then, given the information provided by (S, T), the learner aims at finding/learning a posterior distribution ρ leading to a *ρ-weighted majority vote* over $\mathcal{H}$,

$$B_\rho(\cdot) \;=\; \text{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\cdot)\right],$$

with nice generalization guarantees on the target domain $\mathcal{T}$. In other words, we want to find the posterior distribution ρ minimizing the true target risk of B_ρ:

$$\mathbf{R}_{\mathcal{T}}(B_\rho) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \mathbf{I}\left[B_\rho(\mathbf{x}) \neq y\right],$$

where $\mathbf{I}[a] = 1$ if a is true, and 0 otherwise.

However, in usual PAC-Bayesian analyses (McAllester, 1999; Langford & Shawe-Taylor, 2002; Catoni, 2007), one does not directly focus on the risk of the deterministic classifier B_ρ, but studies the risk of the closely related stochastic Gibbs classifier G_ρ. Given a domain $\mathcal{D}$, the classifier G_ρ predicts the label of an example $\mathbf{x}$ by first drawing a voter h from $\mathcal{H}$ according to the posterior ρ, and then returning $h(\mathbf{x})$. Thus, the risk of G_ρ on a domain $\mathcal{D}$ (also called the *Gibbs risk*) corresponds to the expectation of the risks over $\mathcal{H}$ according to ρ:

$$\mathbf{R}_{\mathcal{D}}(G_\rho) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}\left[h(\mathbf{x}) \neq y\right]. \tag{1}$$

It is well-know in the PAC-Bayesian literature that the deterministic vote B_ρ and the stochastic classifier G_ρ are related by:

$$\mathbf{R}_{\mathcal{D}}(B_\rho) \leq 2\,\mathbf{R}_{\mathcal{D}}(G_\rho).$$

Furthermore, Lacasse et al. (2006) have exhibited that one can obtain a tighter bound on $\mathbf{R}_{\mathcal{D}}(B_\rho)$ by studying the *expected disagreement* $\mathbf{d}_{\mathcal{D}_{\mathbf{X}}}(\rho)$ and the *expected joint error* $\mathbf{e}_{\mathcal{D}}(\rho)$ of the pairs of voters, respectively defined as:

$$\mathbf{d}_{\mathcal{D}_{\mathbf{X}}}(\rho) \;=\; \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right], \tag{2}$$

$$\text{and} \quad \mathbf{e}_{\mathcal{D}}(\rho) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{I}\left[h(\mathbf{x}) \neq y\right] \mathbf{I}\left[h'(\mathbf{x}) \neq y\right], \tag{3}$$

with $\rho^2(h,h') = \rho(h) \times \rho(h')$. Given a m-sample $S \sim (\mathcal{D})^m$, we use $\widehat{\mathbf{R}}_S(G_\rho)$, $\widehat{\mathbf{d}}_S(\rho)$ and $\widehat{\mathbf{e}}_S(\rho)$ to denote the empirical estimation of the Gibbs risk, the disagreement and the joint error respectively.

Finally, note that, given a domain $\mathcal{D}$ on $\mathbf{X}\times Y$, the starting point of our work is the following simple observation:

$$\begin{aligned}\forall \rho \text{ on } \mathcal{H},\ \mathbf{R}_{\mathcal{D}}(G_\rho) &= \frac{1}{2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \Big(\mathbf{I}\left[h(\mathbf{x}) \neq y\right] + \mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \Big) \\ &= \frac{1}{2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \Big(\mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right] + 2 \times \mathbf{I}\left[h(\mathbf{x}) \neq y\right] \mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \Big) \\ &= \frac{1}{2}\, \mathbf{d}_{\mathcal{D}_\mathbf{X}}(\rho) + \mathbf{e}_{\mathcal{D}}(\rho)\,. \end{aligned} \tag{4}$$

3 The Previous PAC-Bayesian Domain Adaptation Analysis

Inspired by the seminal domain adaptation analyses of Ben-David et al. (2006; 2010) and Mansour et al. (2009), a first PAC-Bayesian domain adaptation bound was derived by Germain et al. (2013). This bound is based on a divergence between distributions—called the domain disagreement (see Equation (5))—suitable for the PAC-Bayesian analysis, *i.e.,* for the stochastic Gibbs classifier. Their main result is stated in the following theorem.

Theorem 1. *Let $\mathcal{H}$ be a set of voters. For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}\times Y$, we have:*

$$\forall \rho \text{ on } \mathcal{H}, \quad \mathbf{R}_{\mathcal{T}}(G_\rho) \ \leq\ \mathbf{R}_{\mathcal{S}}(G_\rho) + \mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) + \lambda(\rho, {\rho_\mathcal{T}}^*)\,,$$

where $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ is the domain disagreement between the marginals $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$:

$$\begin{aligned}\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) &= \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_\mathbf{X}} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_\mathbf{X}} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right] \right) \right| \\ &= \Big| \mathbf{d}_{\mathcal{S}_\mathbf{X}}(\rho) - \mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) \Big|\,, \end{aligned} \tag{5}$$

and $\lambda(\rho, {\rho_\mathcal{T}}^*) = \mathbf{R}_{\mathcal{T}}(G_{\rho_\mathcal{T}^*}) + \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{h'\sim\rho_\mathcal{T}^*} \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_\mathbf{X}} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_\mathbf{X}} \mathbf{I}\left[h(\mathbf{x}) \neq h'(\mathbf{x})\right] \right)$,
with ${\rho_\mathcal{T}}^* = \mathrm{argmin}_\rho\, \mathbf{R}_{\mathcal{T}}(G_\rho)$ *the best posterior distribution on the target domain.*

This bound reflects the usual philosophy in domain adaptation (Ben-David et al., 2006; 2010; Mansour et al., 2009). Indeed, assuming that the last term $\lambda(\rho, {\rho_\mathcal{T}}^*)$—which is not estimable from unlabeled target samples—is low, a favorable situation for domain adaptation arises when the deviation between the domains with respect to $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ is small and the accuracy on the source domain $\mathbf{R}_{\mathcal{S}}(G_\rho)$ is good.

Along with the above theorem, Germain et al. (2013) provide the following PAC-Bayesian generalization bound (based on the PAC-Bayesian analysis of non-adaptative learning of Catoni (2007)).

Theorem 2. *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}\times Y$, any set of voters $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta \in (0,1]$, any real numbers $a > 0$ and $c > 0$, with a probability at least $1-\delta$ over the choice of $S \times T \sim (\mathcal{S} \times \mathcal{T}_\mathbf{X})^m$, we have for every posterior distribution ρ on $\mathcal{H}$:*

$$\mathbf{R}_{\mathcal{T}}(G_\rho) \ \leq\ c'\, \widehat{\mathbf{R}}_S(G_\rho) + a'\, \widehat{\mathrm{dis}}_\rho(S,T) + \left(\frac{c'}{c} + \frac{2a'}{a}\right) \left(\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{\delta}}{m}\right) + \lambda(\rho, {\rho_\mathcal{T}}^*) + \alpha' - 1\,,$$

where $\widehat{\mathbf{R}}_S(G_\rho)$, respectively $\widehat{\mathrm{dis}}_\rho(S,T)$, is the empirical estimation of the source risk, respectively of the domain disagreement between $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$, and $c' = \frac{c}{1-e^{-c}}$, and $a' = \frac{2a}{1-e^{-2a}}$, and $\mathrm{KL}(\rho\|\pi)$ is the Kullback-Leibler divergence between ρ and π.

This result justifies the learning algorithm PBDA (Germain et al., 2013). Given a source sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_s}$ and a target sample $T = \{(\boldsymbol{x}_i)\}_{i=1}^{m_t}$, the goal of PBDA is to minimize the bound of Theorem 2. However, the term $\lambda(\rho, {\rho_\mathcal{T}}^*)$ does not appear in the optimization process of PBDA, even if it relies on ρ. Germain et al. (2013) argued that the value of $\lambda(\rho, {\rho_\mathcal{T}}^*)$ should be negligible (uniformly for all

ρ) when adaptation to the target distribution is achievable[3]. Therefore, given the hyperparamters $A > 0$ and $C > 0$, the algorithm PBDA minimizes the trade-off

$$C\,\widehat{\mathbf{R}}_S(G_\rho) + A\,\widehat{\mathrm{dis}}_\rho(S,T) + \mathrm{KL}(\rho\|\pi) \tag{6}$$

specialized to linear classifiers, as detailed below.

Let $\mathcal{H}$ be a set of linear classifiers. Each $h_{\mathbf{w}'} \in \mathcal{H}$ is defined by a weight vector $\mathbf{w}' \in \mathbb{R}^d$:

$$h_{\mathbf{w}'}(\mathbf{x}) = \mathrm{sign}\,(\mathbf{w}' \cdot \mathbf{x})\ ,$$

where $\cdot$ denotes the dot product. Building on previous PAC-Bayesian analyses for linear classifiers (Langford & Shawe-Taylor, 2002; Ambroladze et al., 2006; Parrado-Hernández et al., 2012; Germain et al., 2009a), Germain et al. (2013) consider that prior and posterior distributions are Gaussian distributions. Indeed, if the posterior distribution $\rho_{\mathbf{w}}$, respectively the prior distribution $\pi_{\mathbf{0}}$, is defined as a spherical Gaussian with identity covariance matrix centered on the vector $\mathbf{w}$, respectively $\mathbf{0}$, then we have:

$$\begin{aligned}\forall h_{\mathbf{w}'} \in \mathcal{H}, \quad \rho_{\mathbf{w}}(h_{\mathbf{w}'}) &= \left(\frac{1}{\sqrt{2\boldsymbol{\pi}}}\right)^d e^{-\frac{1}{2}\|\mathbf{w}'-\mathbf{w}\|^2}\,,\\ \text{and} \quad \pi_{\mathbf{0}}(h_{\mathbf{w}'}) &= \left(\frac{1}{\sqrt{2\boldsymbol{\pi}}}\right)^d e^{-\frac{1}{2}\|\mathbf{w}'\|^2},\end{aligned}$$

and the KL-divergence between $\rho_{\mathbf{w}}$ and $\pi_{\mathbf{0}}$ simply is

$$\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}}) = \frac{1}{2}\|\mathbf{w}\|^2\,.$$

Moreover, it is easy to verify that the prediction of the majority vote $B_{\rho_{\mathbf{w}}}$ corresponds to the one of the linear classifier $h_{\mathbf{w}}$:

$$\begin{aligned}\forall\,\mathbf{x} \in X,\ \forall\,\mathbf{w} \in \mathcal{H}, \quad h_{\mathbf{w}}(\mathbf{x}) &= \mathrm{sign}\left[\mathop{\mathbf{E}}_{h_{\mathbf{w}'}\sim\rho_{\mathbf{w}}} h_{\mathbf{w}'}(\mathbf{x})\right]\\ &= B_{\rho_{\mathbf{w}}}(\mathbf{x})\,.\end{aligned}$$

Finally, by rewriting Equation (6) in the case of linear classifiers, the algorithm PBDA consists in minimizing the following function of $\mathbf{w}$:

$$C\sum_{i=1}^{m_s}\Phi_{\mathrm{cvx}}\left(y_i\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right) + A\left|\sum_{i=1}^{m_s}\left[\Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right) - \sum_{i=1}^{m_t}\Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right)\right]\right| + \frac{1}{2}\|\mathbf{w}\|^2\,, \tag{7}$$

where

$$\begin{aligned}\Phi_{\mathrm{cvx}}(x) &= \max\left\{\Phi(x),\ \frac{1}{2}-\frac{x}{\sqrt{2\boldsymbol{\pi}}}\right\}\,,\\ \Phi_{\mathrm{dis}}(x) &= 2\times\Phi(x)\times\Phi(-x)\,,\\ \text{and} \quad \Phi(x) &= \tfrac{1}{2}\left[1-\mathbf{Erf}\left(\frac{x}{\sqrt{2}}\right)\right]\,,\end{aligned}$$

with $\mathbf{Erf}(x) = \frac{2}{\sqrt{\boldsymbol{\pi}}}\displaystyle\int_0^x \exp(-t^2)\mathrm{d}t$ the Gauss error function.

As pointed out by Germain et al. (2013), the kernel trick applies to Equation (7). That is, given a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, one can express a linear classifier in a *RKHS* by a dual weight vector $\boldsymbol{\alpha}\in\mathbb{R}^{m_s+m_t}$:

$$h_{\mathbf{w}}(\cdot) = \mathrm{sign}\left[\sum_{i=1}^{m_s}\alpha_i k(\mathbf{x}_i,\cdot) + \sum_{i=1}^{m_t}\alpha_{i+m_s}k(\mathbf{x}_i,\cdot)\right]. \tag{8}$$

[3]This strong assumption cannot be verified because $\rho_{\mathcal{T}}{}^*$ is unknown. We claim that this is a major weakness of the work of Germain et al. (2013) that our new approach overcomes.

4 A New PAC-Bayesian Domain Adaptation Bound

We now derive a simpler and more precise analysis of PAC-Bayesian domain adaptation. Inspired by the idea of Lacasse et al. (2006), we first decompose the risk $\mathbf{R}_{\mathcal{T}}(G_\rho)$ into the expected disagreement $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and the expected joint error $\mathbf{e}_{\mathcal{T}}(\rho)$, as exhibited by Equation (4). In the present domain adaptation context, we are able to estimate the quantity $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ using a target sample, since the voters' disagreement does not rely on label. However, the expected joint error can only be estimated on the labeled source sample. Theorem 3 below presents our domain adaptation bound and links the target joint error $\mathbf{e}_{\mathcal{T}}(\rho)$ with the source one $\mathbf{e}_{\mathcal{S}}(\rho)$ by weighting the latter by a divergence measure between the two domains. This domain divergence $\beta_q(\mathcal{T}\|\mathcal{S})$ is parametrized by a real value $q > 0$:

$$\beta_q(\mathcal{T}\|\mathcal{S}) \;=\; \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right)^q\right]^{\frac{1}{q}}. \tag{9}$$

In particular, we denote the limit case $q \to \infty$ by:

$$\beta_\infty(\mathcal{T}\|\mathcal{S}) = \sup_{(\mathbf{x},y)\in\mathbf{X}\times Y} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right).$$

It is worth noting that considering some q values allows recovering well-known divergences. For instance, choosing $q = 2$ relates our result to the χ^2-distance between the two domains, as

$$\beta_2(\mathcal{T}\|\mathcal{S}) = \sqrt{\chi^2(\mathcal{T}\|\mathcal{S}) + 1}\,.$$

Moreover, we can relate $\beta_q(\mathcal{T}\|\mathcal{S})$ to the Rényi divergence[4], which has already led to generalization bounds in the specific context of importance weighting by Cortes et al. (2010).

The divergence measure $\beta_q(\mathcal{T}\|\mathcal{S})$ between the two domains is the only term that cannot be estimated from samples (since we do not consider target labels) in the statement of Theorem 3 below.

Theorem 3. *Let $\mathcal{H}$ be a hypothesis space, let $\mathcal{S}$ and $\mathcal{T}$ respectively be the source and the target domains on $\mathbf{X}\times Y$. Let $q > 0$ be a constant. We have:*

$$\forall \rho \text{ on } \mathcal{H}, \quad \mathbf{R}_{\mathcal{T}}(G_\rho) \;\le\; \frac{1}{2}\,\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) + \beta_q(\mathcal{T}\|\mathcal{S}) \times \Big[\mathbf{e}_{\mathcal{S}}(\rho)\Big]^{1-\frac{1}{q}}.$$

where $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$, $\mathbf{e}_{\mathcal{S}}(\rho)$ and $\beta_q(\mathcal{T}\|\mathcal{S})$ are respectively defined by Equations (2), (3) and (9).

Proof. Starting from Equation (4). we have, for every ρ on $\mathcal{H}$,

$$\begin{aligned}
\mathbf{R}_{\mathcal{T}}(G_\rho) \;&=\; \frac{1}{2}\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \mathbf{I}\left[h(\mathbf{x}) \neq y\right] \\
&=\; \frac{1}{2}\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \mathbf{I}\left[h(\mathbf{x}) \neq y\right]\right) \\
&\le\; \frac{1}{2}\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) + \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right)^q\right]^{\frac{1}{q}} \left[\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left(\mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \mathbf{I}\left[h(\mathbf{x}) \neq y\right]\right)^p\right]^{\frac{1}{p}}.
\end{aligned} \tag{10}$$

Last line is due to Hölder inequality, with p such that $\frac{1}{p} = 1-\frac{1}{q}$. Finally, we remove the exponent from expression $(\mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \mathbf{I}\left[h(\mathbf{x}) \neq y\right])^p$ without affecting its value, which is either 1 or 0. □

It is instructive to compare Theorem 3 statement with the previous PAC-Bayesian domain adaptation bound of Theorem 1. In our bound, the only non-estimable term is the domain divergence $\beta_q(\mathcal{T}\|\mathcal{S})$, and contrary to the non-controllable term $\lambda(\rho, \rho_\mathcal{T}{}^*)$ of Theorem 1, it does not depend on the posterior distribution ρ learned: For every ρ on $\mathcal{H}$, $\beta_q(\mathcal{T}\|\mathcal{S})$ is a constant measuring the relation between the two domains. Moreover, this latter domain divergence is not an additive term but a multiplicative one (as opposed to $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) + \lambda(\rho, \rho_\mathcal{T}{}^*)$ in Theorem 1). This is a contribution of our analysis, since $\beta_q(\mathcal{T}\|\mathcal{S})$ can be considered as a hyperparameter that allows tuning the trade-off between the target voters' disagreement and the source joint error[5]. Consequently, we do not need to make assumptions on

[4] For every $q \ge 0$, we can easily prove that: $\beta_q(\mathcal{T}\|\mathcal{S}) = d_q(\mathcal{T}\|\mathcal{S})^{\frac{q-1}{q}}$, where $d_q(T\|S) = 2^{D_q(\mathcal{T}\|\mathcal{S})}$ with $D_q(\mathcal{T}\|\mathcal{S})$ the Rényi divergence between $\mathcal{T}$ and $\mathcal{S}$.

[5] Experiments of Section 6 show that this hyperparameter can be successfully selected by reverse validation.

its value, while usual domain adaptation approaches require that such non-estimable terms are negligible (even in non-PAC-Bayesian bounds, similar additional terms also appear (Ben-David et al., 2006; 2010; Mansour et al., 2009)).

Another very attractive point of Theorem 3 comes from the parameter q, that allows considering different relationships between $\beta_q(\mathcal{T}\|\mathcal{S})$ and $\mathbf{e}_{\mathcal{S}}(\rho)$. In particular, the case $q \to \infty$ exhibits an interesting analysis: Whenever the two domains are equals (*i.e.*, $\mathcal{S} = \mathcal{T}$) then $\beta_\infty(\mathcal{T}\|\mathcal{S}) = 1$, and the bound becomes an equality. Therefore, when adaptation is not necessary, our analysis is still sound:

$$\forall \rho \text{ on } \mathcal{H}, \quad \mathbf{R}_{\mathcal{T}}(G_\rho) \;\le\; \frac{1}{2}\,\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) + \mathbf{e}_{\mathcal{S}}(\rho) \;=\; \frac{1}{2}\,\mathbf{d}_{\mathcal{S}_\mathbf{X}}(\rho) + \mathbf{e}_{\mathcal{S}}(\rho) \;=\; \mathbf{R}_{\mathcal{S}}(G_\rho) \;=\; \mathbf{R}_{\mathcal{T}}(G_\rho)\,.$$

Furthermore, under the *covariate-shift* (Shimodaira, 2000) assumption, that is the domains only diverge in their marginals, (*i.e.*, $\mathcal{T}_{Y|\mathbf{x}}(y) = \mathcal{S}_{Y|\mathbf{x}}(y)$), one may estimate the value of $\beta_q(\mathcal{T}_\mathbf{X}\|\mathcal{S}_\mathbf{X})$ using unsupervised density estimation methods. Interestingly, from Line (10), we also can obtain the following equality:

$$\forall \rho \text{ on } \mathcal{H}, \quad \mathbf{R}_{\mathcal{T}}(G_\rho) \;=\; \frac{1}{2}\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_\mathbf{X}} \left(\frac{\mathcal{T}_\mathbf{X}(\mathbf{x})}{\mathcal{S}_\mathbf{X}(\mathbf{x})} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \mathbf{I}\left[h(\mathbf{x}) \neq y\right] \right), \tag{11}$$

which suggests a way to correct the *shift* between the two distributions by reweighting the labeled source information captured by the joint error. Note that we consider this as future work studies.

PAC-Bayesian Generalization Guarantees

In order to justify the empirical minimization of our bound of Theorem 3, we first provide here PAC-Bayesian generalization guarantees for $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and $\mathbf{e}_{\mathcal{S}}(\rho)$. These results are presented as a corollary of Theorem 4 below, that generalizes the PAC-Bayesian theorem of Catoni (2007) (more precisely, the simplified form of Germain et al. (2009b)), to arbitrary loss functions. Indeed, Theorem 4, with $\ell(h, \mathbf{x}, y) = \mathbf{I}\left[h(\mathbf{x}) \neq y\right]$ and Equation (1), gives the usual bound on the Gibbs risk.

Theorem 4. *For any domain $\mathcal{D}$ over $\mathbf{X} \times Y$, any set of voters $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any function $\ell : \mathcal{H} \times \mathbf{X} \times Y \to [0,1]$, any real number $c > 0$, with a probability at least $1-\delta$ over the choice of $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{D})^m$, we have for every posterior distribution ρ on $\mathcal{H}$:*

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho} \ell(h,\mathbf{x},y) \;\le\; \frac{1}{1-e^{-c}} \left[\frac{c}{m} \sum_{i=1}^{m} \mathop{\mathbf{E}}_{h\sim\rho} \ell(h,\mathbf{x}_i,y_i) + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{m} \right].$$

Proof. We use the following shorthand notation:

$$\mathcal{L}_{\mathcal{D}}(h) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(h,\mathbf{x},y)\,, \quad \text{and} \quad \mathcal{L}_S(h) = \frac{1}{m} \sum_{(\mathbf{x},y)\in S} \ell(h,\mathbf{x},y)\,.$$

Consider any convex function $\Delta : [0,1]\times[0,1] \to \mathbb{R}$. Applying consecutively Jensen's Inequality and the Change of measure inequality (see Seldin & Tishby (2010, Lemma 4) and McAllester (2013, Equation (20))), we obtain:

$$\begin{aligned} \forall \rho \text{ on } \mathcal{H} : \quad m\times\Delta\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathcal{L}_S(h), \mathop{\mathbf{E}}_{h\sim\rho} \mathcal{L}_{\mathcal{D}}(h) \right) &\le \mathop{\mathbf{E}}_{h\sim\rho} m\times\Delta\left(\mathcal{L}_S(h), \mathcal{L}_{\mathcal{D}}(h)\right) \\ &\le \mathrm{KL}(\rho\|\pi) + \ln\Big[X_\pi(S)\Big]\,, \end{aligned}$$

with

$$X_\pi(S) = \mathop{\mathbf{E}}_{h\sim\pi} e^{m\times\Delta(\mathcal{L}_S(h),\,\mathcal{L}_{\mathcal{D}}(h))}.$$

Then, Markov's Inequality gives

$$\Pr_{S\sim\mathcal{D}^m}\left(X_\pi(S) \le \tfrac{1}{\delta} \mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m} X_\pi(S') \right) \;\le\; 1-\delta\,,$$

and

$$\begin{aligned} \mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m} X_\pi(S') &= \mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m} \mathop{\mathbf{E}}_{h\sim\pi} e^{m\times\Delta(\mathcal{L}_{S'}(h),\,\mathcal{L}_{\mathcal{D}}(h))} \\ &= \mathop{\mathbf{E}}_{h\sim\pi} \mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m} e^{m\times\Delta(\mathcal{L}_{S'}(h),\,\mathcal{L}_{\mathcal{D}}(h))} \\ &\le \mathop{\mathbf{E}}_{h\sim\pi} \sum_{k=0}^{m} \binom{k}{m} (\mathcal{L}_{\mathcal{D}}(h))^k (1-\mathcal{L}_{\mathcal{D}}(h))^{m-k} e^{m\times\Delta\left(\frac{k}{m},\,\mathcal{L}_{\mathcal{D}}(h)\right)}, \end{aligned} \tag{12}$$

where the last inequality is due to Maurer (2004, Lemma 3) (we have an equality when the output of ℓ is in $\{0,1\}$). As shown in Germain et al. (2009a, Corollary 2.2), by fixing

$$\Delta(q,p) = -c{\times}q - \ln[1{-}p\,(1{-}e^{-c})]\,,$$

Line 12 becomes equal to 1, and then $\mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m} X_\pi(S') \leq 1$. Hence,

$$\Pr_{S\sim\mathcal{D}^m}\left(\forall\rho \text{ on } \mathcal{H}\,:\, -c\mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_S(h) - \ln[1{-}\mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_\mathcal{D}(h)\,(1{-}e^{-c})] \leq \frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m}\right) \;\leq\; 1{-}\delta\,.$$

By reorganizing the terms, we have, with probability $1{-}\delta$ over the choice of $S\in\mathcal{D}^m$,

$$\forall\rho \text{ on } \mathcal{H}\,:\, \mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_\mathcal{D}(h) \leq \frac{1}{1{-}e^{-c}}\left[1-\exp\left(-c\mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_S(h) - \frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m}\right)\right]\,.$$

The final result is obtained by using the inequality $1-\exp(-z)\leq z$. □

We now extend this result to the expected disagreement and the expected joint error. PAC-Bayesian bounds on these quantities already appeared in Lacasse et al. (2006), but under different forms. In the statement of Corollary 1 below, we are especially interested in the possibility of controlling the trade-off—between the empirical estimate computed on the samples and the complexity term captured by $\mathrm{KL}(\rho\|\pi)$—with the help of parameters a and c.

Corollary 1. *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}\times Y$, any set of voters $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, any real numbers $a>0$ and $c>0$, we have:*

$$\Pr_{T\sim(\mathcal{T}_\mathbf{X})^{m_t}}\left(\forall\rho \text{ on } \mathcal{H},\quad \mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) \;\leq\; c'\,\widehat{\mathbf{d}}_T(\rho) + \frac{c'}{c}\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m_t}\right) \;\geq\; 1-\delta\,,$$

$$\textit{and}\quad \Pr_{S\sim(\mathcal{S})^{m_s}}\left(\forall\rho \text{ on } \mathcal{H},\quad \mathbf{e}_\mathcal{S}(\rho) \;\leq\; a'\,\widehat{\mathbf{e}}_S(\rho) + \frac{a'}{a}\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m_s}\right) \;\geq\; 1-\delta\,.$$

where $\widehat{\mathbf{R}}_T(\rho)$, respectively $\widehat{\mathbf{e}}_S(\rho)$, is the empirical estimation of the target voters' disagreement, respectively of source joint error, and $c'=\frac{c}{1-e^{-c}}$, and $a'=\frac{a}{1-e^{-a}}$.

Proof. Given π and ρ over $\mathcal{H}$, we consider a new prior π^2 and a new posterior ρ^2, both over $\mathcal{H}^2$, such that: $\forall\, h_{ij}=(h_i,h_j)\in\mathcal{H}^2$, $\pi^2(h_{ij})=\pi(h_i)\pi(h_j)$ and $\rho^2(h_{ij})=\rho(h_i)\rho(h_j)$. Thus, $\mathrm{KL}(\rho^2\|\pi^2)=2\mathrm{KL}(\rho\|\pi^2)$ (see (Germain et al., 2013; Lacasse et al., 2006)). Let us define two new loss functions for a "paired voter" $h_{ij}\in\mathcal{H}^2$:

$$\ell_d(h_{ij},\mathbf{x},y) = \mathbf{I}\left[h_i(\mathbf{x})\neq h_j(\mathbf{x})\right]\,,\quad\text{and}\quad \ell_e(h_{ij},\mathbf{x},y) = \mathbf{I}\left[h_i(\mathbf{x})\neq y\right]\times\mathbf{I}\left[h_j(\mathbf{x})\neq y\right]\,.$$

Then, the bound on $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ is obtained from Theorem 4 with $\ell\coloneqq\ell_d$, and Equation (2). The bound on $\mathbf{e}_\mathcal{S}(\rho)$ is similarly obtained with $\ell\coloneqq\ell_e$ and using Equation (3). □

For algorithmic reasons, we are going to deal with our bound of Theorem 3 when $q\to\infty$. Thanks to Theorem 1, minimizing this bound is equivalent to optimize the following generalization bound defined with respect to the empirical estimates of the target disagreement and the source joint error.

Theorem 5. *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}\times Y$, any set of voters $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, any real numbers $a>0$ and $c>0$, with a probability at least $1-\delta$ over the choice of $S\times T\sim(\mathcal{S}\times\mathcal{T}_\mathbf{X})^m$, we have for every posterior distribution ρ on $\mathcal{H}$:*

$$\mathbf{R}_\mathcal{T}(G_\rho) \;\leq c'\,\frac{1}{2}\,\widehat{\mathbf{d}}_T(\rho) + b'\,\widehat{\mathbf{e}}_S(\rho) + \left(\frac{c'}{c}+\frac{b'}{b\,a}\right)\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{2}{\delta}}{m}\,,$$

where $\widehat{\mathbf{R}}_T(\rho)$, respectively $\widehat{\mathbf{e}}_S(\rho)$, is the empirical estimation of the target voters' disagreement, respectively of source joint error, and $b=\beta_\infty(\mathcal{T}\|\mathcal{S})$, and $b'=b\,\frac{a}{1-e^a}$, and $c'=\frac{c}{1-e^{-c}}$.

Proof. The result is obtained by bounding separately $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and $\mathbf{e}_\mathcal{S}(\rho)$ using Corollary 1 (with probability $1{-}\frac{\delta}{2}$ each), and combining the two upper bounds according to Theorem 3. □

From an optimization perspective, the problem suggested by the bound of Theorem 5 is much more convenient to minimize than the bound of Theorem 2. The former is *smoother* than the latter that contains an absolute value required by the domain disagreement $\widehat{\mathrm{dis}}_\rho(S,T)$. Moreover, recall that Germain et al. (2013) choose to ignore the non-constant term $\lambda(\rho,\rho_\mathcal{T}{}^*)$ of Theorem 2. In our case, such compromise is not mandatory to apply the theoretical result to real domain adaptation problems.

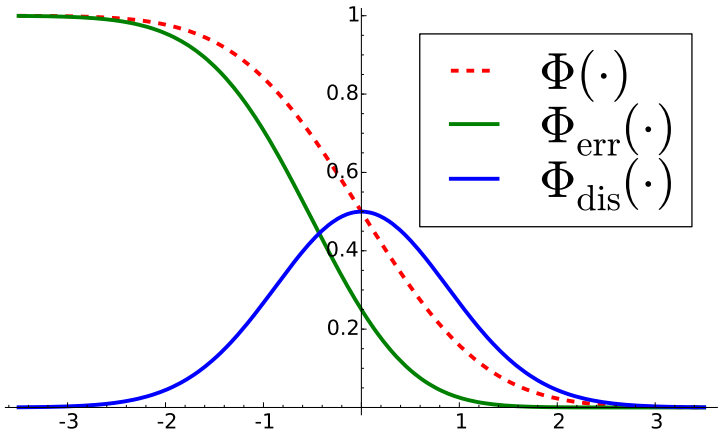

Figure 1: Graphical representation of the loss functions given by the specialization to linear classifiers.

5 From the Bound to an Algorithm Specialized to Linear Classifiers

As observed in the previous section, $\beta_q(\mathcal{T}\|\mathcal{S})$ is a constant, whatever is the value of q. Then it can be considered as a hyperparameter to tune. According to Theorem 5, given $C > 0$ and $B > 0$ the hyperparameters of the algorithm, we propose to minimize the following trade-off:

$$C\,\widehat{\mathbf{d}}_T(\rho) + B\,\widehat{\mathbf{e}}_S(\rho) + \mathrm{KL}(\rho\|\pi)\,, \tag{13}$$

where B models the compromise between the target and the source domains. Contrary to the trade-off optimized by PBDA (Germain et al., 2013) of Equation (6), we do not neglect any term of our bound.

We now follow the setting presented in Section 3 for specializing Equation (13) to linear classifiers. Therefore, we consider $\mathcal{H}$ as a set of linear classifiers in a d-dimensional space, and we use Gaussians posterior $\rho_{\mathbf{w}}$ and prior $\pi_{\mathbf{0}}$ with identity covariance matrix (respectively centered on vectors $\mathbf{w}$ and $\mathbf{0}$). With $\Phi_{\text{dis}}(x) = 2{\times}\Phi(x){\times}\Phi(-x)$, Germain et al. (2013) showed :

$$\forall \rho_{\mathbf{w}} \text{ on } \mathcal{H}, \quad \mathbf{d}_{\mathcal{T}_{\mathbf{X}}}(\rho_{\mathbf{w}}) \;=\; \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}} \Phi_{\text{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right).$$

Following a similar approach, with $\Phi_{\text{err}}(x) = \big[\Phi(x)\big]^2$, we obtain:

$$\begin{aligned}
\forall \rho_{\mathbf{w}} \text{ on } \mathcal{H}, \quad \mathbf{e}_{\mathcal{S}}(\rho_{\mathbf{w}}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \mathbf{I}\left[h(\mathbf{x}) \neq y\right] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}\left[h'(\mathbf{x}) \neq y\right] \mathop{\mathbf{E}}_{h'\sim\rho} \mathbf{I}\left[h(\mathbf{x}) \neq y\right] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \Phi_{\text{err}}\left(y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right).
\end{aligned}$$

Figure 1 illustrates the behavior of the loss functions Φ, Φ_{err} and Φ_{dis}. Finally, by specializing Equation (13) to linear classifiers, our new algorithm consists in minimizing

$$G(\mathbf{w}) = C \times \Phi \sum_{i=1}^{m_t} \left(\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right) \Phi\left(-\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right) + B \times \sum_{i=1}^{m_s} \left[\Phi\left(y_i\,\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right)\right]^2 + \frac{1}{2}\|\mathbf{w}\|^2\,. \tag{14}$$

We call this algorithm DALC for Domain Adaptation of Linear Classifiers.

Similarly to Germain et al. (2013), we can apply the kernel trick to DALC, using the dual vector $\boldsymbol{\alpha}$ of Equation (8). Even if the objective function is highly non-convex, we achieved good empirical results by minimizing the "kernelized" version of Equation (14) by gradient descent, with a uniform weight vector as a starting point. More details are given in the supplementary material.

Let $S{=}\{(\mathbf{x}_i, y_i)\}_{i=1}^{m_s}$, $T{=}\{\mathbf{x}_i\}_{i=1}^{m_t}$ and $\mathrm{M} = m_s + m_t$. We will denote

$$\mathbf{x}_{\#} \;=\; \begin{cases} \mathbf{x}_i & \text{if } \# \leq m_s \quad \text{(source examples)} \\ \mathbf{x}_{\#-m_s} & \text{otherwise.} \quad \text{(target examples)} \end{cases}$$

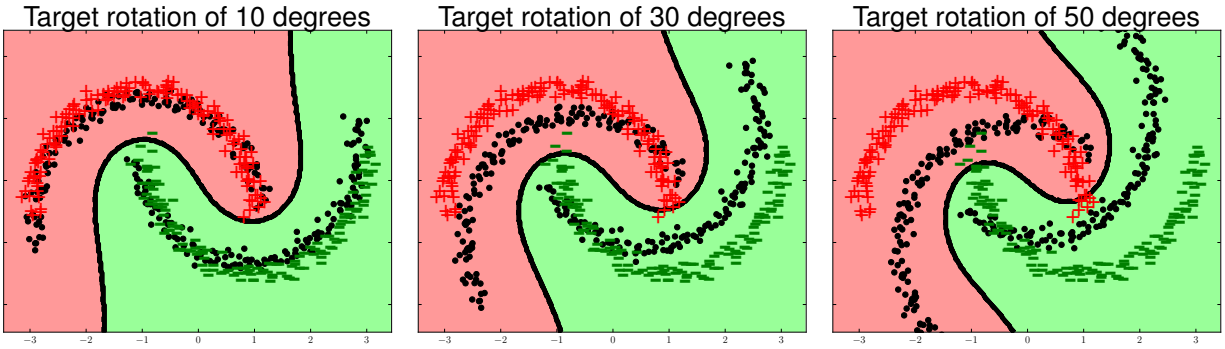

Figure 2: Decision boundaries of DALC on the *intertwining moons* toy problem, for fixed parameters B=C=1, and a RBF kernel $k(\mathbf{x}, \mathbf{x}') = e^{-\|\mathbf{x}-\mathbf{x}'\|^2}$. The target points are black. The positive, *resp.* negative, source points are red, *resp.* green.

The kernel trick allows us to work with dual weight vector $\boldsymbol{\alpha} \in \mathbb{R}^{\mathrm{M}}$ that is a linear classifier in an augmented space. Given a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we have

$$h_{\mathbf{w}}(\cdot) \;=\; \mathrm{sign}\left[\sum_{i=1}^{\mathrm{M}} \alpha_i k(\mathbf{x}_i, \cdot)\right].$$

Let us denote K the kernel matrix of size $\mathrm{M} \times \mathrm{M}$ such as $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$.

In that case, the objective function $G(\mathbf{w})$—Equation (14)—can be rewritten in term of the vector

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots \alpha_{\mathrm{M}})$$

as

$$G(\boldsymbol{\alpha}) = C \times \sum_{i=m_s}^{\mathrm{M}} \Phi\left(\frac{\sum_{j=1}^{\mathrm{M}} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right) \Phi\left(-\frac{\sum_{j=1}^{\mathrm{M}} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right) + B \times \sum_{i=1}^{m_s} \left[\Phi\left(y_i \frac{\sum_{j=1}^{\mathrm{M}} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right)\right]^2 + \frac{1}{2} \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} \alpha_i \alpha_j K_{i,j}\,. \tag{15}$$

6 Experimental Results

Firstly, Figure 2 illustrates the behavior of the decision boundary of our algorithm DALC on an intertwining moons toy problem[6], where each moon corresponds to a label. The target domain, for which we have no label, is a rotation of the source domain. The figure shows clearly that DALC succeeds to adapt to the target domain, even for a rotation angle of 50°. We see that DALC does not rely on the restrictive *covariate-shift* assumption, as some source examples are misclassified. This behavior illustrates the trade-off proposed by DALC in action, by conceding some errors on the source sample to improve the disagreement on the target sample.

Secondly, we evaluate DALC[7] on the classical *Amazon.com Reviews* benchmark (Blitzer et al., 2006) according to the setting used by Chen et al. (2011) and Germain et al. (2013). This dataset contains reviews of four types of products (books, DVDs, electronics, and kitchen appliances) described with about 100,000 attributes. Originally, the reviews were labeled with a rating from 1 to 5. Chen et al. (2011) proposed a simplified binary setting by regrouping ratings in two classes (products rated lower than 3 and products rated higher than 4). Moreover, they reduced the dimensionality to about 40,000 by only keeping the features appearing at least ten times for a given domain adaptation task. Finally, the data are preprocessed with a tf-idf re-weighting. A domain corresponds to a kind of product. Therefore, we perform twelve domain adaptation tasks. For instance, "books→DVD's" is the task for which the source domain

[6] We generate each pair of moons with the `make_moons` function provided in `scikit-learn` (Pedregosa et al., 2011).

[7] In these experiments, we minimize the objective function (Equation (15)) using a *Broyden-Fletcher-Goldfarb-Shanno method (BFGS)* implemented in the *scipy* python library (Jones et al., 2001–). We initialize the optimization procedure at $\alpha_i = \frac{1}{\mathrm{M}}$ for all $i \in \{1, \ldots, \mathrm{M}\}$.

Table 1: Error rates on *Amazon* dataset. Best risks appear in **bold** and seconds are in *italic*.

	SVM^{CV}	DASVM^{RCV}	CODA^{RCV}	PBDA^{RCV}	DALC ^{RCV}
books→DVDs	*0.179*	0.193	0.181	0.183	**0.178**
books→electronics	0.290	*0.226*	0.232	0.263	**0.212**
books→kitchen	0.251	**0.179**	0.215	0.229	*0.194*
DVDs→books	0.203	0.202	0.217	*0.197*	**0.186**
DVDs→electronics	0.269	**0.186**	*0.214*	0.241	0.245
DVDs→kitchen	0.232	0.183	*0.181*	0.186	**0.175**
electronics→books	0.287	0.305	0.275	**0.232**	*0.240*
electronics→DVDs	0.267	**0.214**	0.239	*0.221*	0.256
electronics→kitchen	*0.129*	0.149	0.134	0.141	**0.123**
kitchen→books	0.267	0.259	*0.247*	*0.247*	**0.236**
kitchen→DVDs	0.253	**0.198**	0.238	0.233	*0.225*
kitchen→electronics	0.149	0.157	0.153	**0.129**	*0.131*
Average	0.231	*0.204*	0.210	0.208	**0.200**

is "books" and the target one is "DVDs". We compare DALC with the classical non-adaptative algorithm SVM (trained only on the source sample), the adaptative algorithm DASVM (Bruzzone & Marconcini, 2010), the adaptative co-training CODA (Chen et al., 2011), and the PAC-Bayesian domain adaptation algorithm PBDA (Germain et al., 2013) of Equation (7). Note that, in Germain et al. (2013), DASVM has shown the best results in average on this *Amazon.com Reviews* dataset. Each parameter is selected with a grid search thanks to a usual cross-validation (^{CV}) on the source sample for SVM, and thanks to a reverse validation procedure[8] (^{RCV}) for CODA, DASVM, PBDA, and DALC. The algorithms use a linear kernel and consider $2,000$ labeled source examples and $2,000$ unlabeled target examples. Table 1 reports the error rates of all the methods evaluated on the same separate target test sets proposed by Chen et al. (2011).

Above all, we observe that the adaptative approaches show the best result, implying that tackling this problem with a domain adaptation method is reasonable. Then, our new method DALC is the best algorithm overall on this task. Except for the two adaptative tasks between "electronics" and "DVDs", DALC is either the best one (six times), or the second one (four times). Moreover, DALC clearly increases the performances of the previous PAC-Bayesian algorithm (PBDA), which confirms that our novel bound improves the analysis done by Germain et al. (2013).

7 Conclusion

In this paper, we derive a novel and original analysis of domain adaptation in the context of majority vote learning. This analysis relies on an upper bound over the target risk, expressed as a simple trade-off between the voters' disagreement measured on the target domain and the voters' joint errors measured on the source one. A crucial point is that the divergence between the two domains is not an additive term (as in many domain adaptation bounds), but is a factor that controls the trade-off given by our bound. To the best of our knowledge, this latter point is a major contribution in domain adaptation, and thus gives a new point of view to tackle it. Moreover, our bound has the clear advantage to lead to a non-degenerated analysis when the two domain are the same. This analysis, combining with a PAC-Bayesian generalization bound, leads to a new domain adaptation algorithm for linear classifiers (named DALC). We provide an experiment on a popular domain adaptation dataset where we showed that our new algorithm can lead to better results.

As future work, we aim at extending our approach to the case in which some target labels are available to accurately estimate the divergence $\beta_q(\mathcal{T}\|\mathcal{S})$. Besides, we would like to explore in details the covariate-shift issue (Shimodaira, 2000), when we suppose that the two domains only differ on their marginals according to the input space. Actually, we believe that decomposing the risk as a trade-off between target voters' disagreement and the weighted source joint errors gives another point of view of this issue which may improve basics methods based only on a reweighting of the source risk. A first step towards this goal

[8] For more details on the reverse validation procedure, the reader can refer to (Bruzzone & Marconcini, 2010; Zhong et al., 2010). For obtaining the DALC^{RCV} results of Table 1, the reverse validation procedure searches on a 20×20 parameter grid for a C between 0.01 and 10^6 and a parameter B between 1.0 and 10^8, both on a logarithm scale. The results of the other algorithms are reported from Germain et al. (2013).

is to study the relationships of our bound with the work of Cortes et al. (2010) on importance weighting algorithms. Indeed, they derived bounds depending on the Rényi divergence between $\mathcal{S}$ and $\mathcal{T}$ which can be related to our divergence $\beta_q(\mathcal{T}\|\mathcal{S})$. A second approach will be to take advantage of Equation (11) which is not a bound but an equality that directly relates the target risk to the disagreement on the unlabeled data and the joint error on the labeled examples.

References

Ambroladze, A., Parrado-Hernández, E., and Shawe-Taylor, J. Tighter PAC-Bayes bounds. In *NIPS*, pp. 9–16, 2006.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *NIPS*, pp. 137–144, 2006.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. Wortman. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010.

Blitzer, J. *Domain adaptation of natural language processing systems*. PhD thesis, UPenn, 2007.

Blitzer, J., McDonald, R., and Pereira, F. Domain adaptation with structural correspondence learning. In *EMNLP*, pp. 120–128, 2006.

Bruzzone, L. and Marconcini, M. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Trans. Pattern Anal. Mach. Intel.*, 32(5):770–787, 2010.

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

Chen, M., Weinberger, K.Q., and Blitzer, J. Co-training for domain adaptation. In *NIPS*, pp. 2456–2464, 2011.

Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theor. Comput. Sci.*, 519:103–126, 2014.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *NIPS*, pp. 442–450, 2010.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *ICML*, pp. 353–360, 2009a.

Germain, P., Lacasse, A., Laviolette, F., Marchand, M., and Shanian, S. From PAC-Bayes bounds to KL regularization. In *NIPS*, pp. 603–610. 2009b.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *ICML*, pp. 738–746, 2013.

Huang, J., Smola, A., Gretton, A., Borgwardt, K., and Schölkopf, B. Correcting sample selection bias by unlabeled data. In *NIPS*, pp. 601–608, 2006.

Jiang, J. A literature survey on domain adaptation of statistical classifiers, 2008.

Jones, Eric, Oliphant, Travis, Peterson, Pearu, et al. SciPy: Open source scientific tools for Python, 2001–. URL `http://www.scipy.org/`.

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 *NIPS*, pp. 769–776, 2006.

Langford, J. and Shawe-Taylor, J. PAC-Bayes & margins. In *NIPS*, pp. 439–446, 2002.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *COLT*, 2009.

Margolis, A. A literature review of domain adaptation with unlabeled data, 2011.

Maurer, A. A note on the PAC-Bayesian theorem. *CoRR*, cs.LG/0411099, 2004.

McAllester, D. A PAC-Bayesian tutorial with a dropout bound. *CoRR*, abs/1307.2118, 2013.

McAllester, D. A. Some PAC-Bayesian theorems. *Mach. Learn.*, 37:355–363, 1999.

Pan, S. J. and Yang, Q. A survey on transfer learning. *T. Knowl. Data En.*, 22(10):1345–1359, 2010.

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. PAC-Bayes bounds with data dependent priors. *JMLR*, 13:3507–3531, 2012.

Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. Visual domain adaptation: A survey of recent advances. *IEEE Signal Proc. Mag.*, 32(3):53–69, 2015.

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

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Seldin, Y. and Tishby, N. PAC-Bayesian analysis of co-clustering and beyond. *JMLR*, 11:3595–3646, 2010.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, 90(2):227–244, 2000.

Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2007.

Zhang, C., Zhang, L., and Ye, J. Generalization bounds for domain adaptation. In *NIPS*, pp. 3320–3328, 2012.

Zhong, E., Fan, W., Yang, Q., Verscheure, O., and Ren, J. Cross validation framework to choose amongst models and datasets for transfer learning. In *ECML-PKDD*, pp. 547–562, 2010.