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] Univ Lyon, UJM-Saint-Etienne, CNRS, Laboratoire Hubert Curien UMR 5516, F-42023 Saint-Etienne, France; Département d’informatique et de génie logiciel, Université Laval (Québec); INRIA - Sierra Project-Team, Ecole Normale Supérieure, Paris, France. 2015. hal-01163722v3

HAL Id: hal-01163722
https://hal.science/hal-01163722v3
Submitted on 14 Mar 2016

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\textsuperscript{1}  Amaury Habrard\textsuperscript{2}  François Laviolette\textsuperscript{3}  Emilie Morvant\textsuperscript{2}

\textsuperscript{1}INRIA - Sierra Project-Team, École Normale Supérieure, Paris, France
\textsuperscript{2}Univ Lyon, UJM-Saint-Etienne, CNRS, Laboratoire Hubert Curien UMR 5516, F-42023 Saint-Etienne, France
\textsuperscript{3}Département d’informatique et de génie logiciel, Université Laval, Québec, Canada

March 14, 2016

Abstract

We study the issue of PAC-Bayesian domain adaptation: We want to learn, from a source domain, a majority vote model dedicated to a target one. Our theoretical contribution brings a new perspective by deriving an upper-bound on the target risk where the distributions’ divergence—expressed as a ratio—controls the trade-off between a source error measure and the target voters’ disagreement. Our bound suggests that one has to focus on regions where the source data is informative. From this result, we derive a PAC-Bayesian generalization bound, and specialize it to linear classifiers. Then, we infer a learning algorithm and perform experiments on real data.

1 Introduction

Machine learning practitioners are commonly exposed to the issue of domain adaptation\textsuperscript{1} (Jiang, 2008; Margolis, 2011): One usually learns a model from a corpus, i.e., a fixed yet unknown source distribution, 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; Ganin & Lempitsky, 2015), bioinformatics (Liu et al., 2008), natural language processing (Blitzer, 2007; Daumé III, 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 to address domain adaptation, 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 reweighting the importance of labeled data (Huang et al., 2006; Sugiyama et al., 2007; Cortes et al., 2010; 2015), and/or by learning a new common representation for the source and target distributions (Chen et al., 2012; Ganin & Lempitsky, 2015), 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 (e.g., Ben-David et al., 2006; 2010; Mansour et al., 2009; Li & Bilmes, 2007; Zhang et al., 2012). Recently, this issue has been studied through the PAC-Bayesian framework (Germain et al., 2013), which focuses on learning weighted majority votes\textsuperscript{2}, without any target label. Even if their result clearly opens the door to tackle domain adaptation in a PAC-Bayesian fashion, it shares the same philosophy as the seminal works of Ben-David et al. (2006; 2010) and Mansour et al. (2009): The risk of the target model is upper-bounded jointly by the risk of the model on the source distribution, the divergence between the marginal distributions, and a non-estimable term\textsuperscript{3} related to the ability to adapt in the current space. Note that Li & Bilmes (2007) previously proposed a PAC-Bayesian generalization bound for domain adaptation but they considered target labels.

In this paper, we derive a novel domain adaptation bound for the weighted majority vote framework. Concretely, the risk of the target model is still upper-bounded by three terms, but these terms differ in the information they capture. The first term is estimable from unlabeled data and relies on a notion of expected voters’ disagreement on the target domain. The second term depends on the expected accuracy of the voters on the source domain. Interestingly, this latter is weighted by a divergence between the

\textsuperscript{1}Domain adaptation is associated with transfer learning (Pan & Yang, 2010; Quionero-Candela et al., 2009).
\textsuperscript{2}This setting is not too restrictive since many machine learning algorithms can be seen as a majority vote learning. For instance, ensemble learning and kernel methods output classifiers which can be interpreted as majority votes.
\textsuperscript{3}More precisely, this term can only be estimated in the presence of labeled data from both the source and the target domains.
source and the target domains that enable controlling the relationship between domains. The third term estimates the “volume” of the target domain living apart from the source one, which has to be small for ensuring adaptation. From our bound, we deduce that a good adaptation strategy consists in finding a weighted majority vote leading to a suitable trade-off—controlled by the domains’ divergence—between the first two terms: Minimizing the first one corresponds to look for voters that disagree on the target domain, and minimizing the second one to seek accurate voters on the source. Thereafter, we provide PAC-Bayesian generalization guarantees to justify the empirical minimization of our new domain adaptation bound, and specialize it to linear classifiers (following a methodology known to give rise to tight bound values). This allows to design DALC, a learning algorithm that improves the performances of the previous PAC-Bayesian domain adaptation algorithm.

The rest of the paper is organized as follows. Section 2 presents the PAC-Bayesian domain adaptation setting. Section 3 reviews previous theoretical results on domain adaptation. Section 4 states our new PAC-Bayesian domain adaptation algorithm.

2 Unsupervised Domain Adaptation Setting

We tackle domain adaptation in a binary classification setting, from a $d$-dimensional input space $\mathbf{X} \subseteq \mathbb{R}^d$ to an output space $Y = \{-1, 1\}$. Our goal is to perform domain adaptation from a distribution $\mathcal{D}$—the source domain—to another (related) distribution $\mathcal{T}$—the target domain—on $\mathbf{X} \times Y$; $\mathcal{S} \mathbf{X}$ and $\mathcal{T} \mathbf{X}$ being the associated marginal distributions on $\mathbf{X}$. Given a distribution $\mathcal{D}$, we denote $(\mathcal{D})^m$ the distribution of a $m$-sample constituted by $m$ elements drawn 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 $T = \{\mathbf{x}_i\}_{i=1}^{m_t} \sim (\mathcal{T})^{m_t}$.

PAC-Bayesian domain adaptation. Our work is inspired by the PAC-Bayesian theory (first introduced by McAllester, 1999). More precisely, we adopt the PAC-Bayesian domain adaptation setting studied by Germain et al. (2013). Given $\mathcal{H}$, a set of voters $h : \mathbf{X} \rightarrow Y$, the elements of this approach are a prior distribution $\pi$ on $\mathcal{H}$, a pair of source-target learning samples $(S, T)$ and a posterior distribution $\rho$ on $\mathcal{H}$. The prior distribution $\pi$ models an a priori belief—before observing $(S, T)$—of the voters’ accuracy. Then, given the information provided by $(S, T)$, we aim at learning a posterior distribution $\rho$ leading to a $\rho$-weighted majority vote over $\mathcal{H}$,

$$ B_\rho(\cdot) = \text{sign} \left[ \sum_{h \sim \rho} I[h(\cdot) = \cdot] \right] , $$

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

$$ R_T(B_\rho) = \mathbb{E}_{(x, y) \sim \mathcal{T}} I[B_\rho(x) \neq y] , $$

where $I[a] = 1$ if $a$ is true, and 0 otherwise. However, in most PAC-Bayesian analyses one does not directly focus on this majority vote risk; but studies the expectation of the risks over $\mathcal{H}$ according to $\rho$, designed as the Gibbs risk:

$$ R_\mathcal{D}(G_\rho) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \mathbb{E}_{h \sim \rho} I[h(x) \neq y] . \quad (1) $$

It is well-known in the PAC-Bayesian literature that $R_\mathcal{D}(B_\rho) \leq 2 R_\mathcal{D}(G_\rho)$ (e.g., Herbrich & Graepel, 2000). Unfortunately, this worst case bound often leads to poor generalization guarantees on the majority vote risk. To address this issue, Lacasse et al. (2006) have exhibited that one can obtain a tighter bound on $R_\mathcal{D}(B_\rho)$ by studying the expected disagreement $d_\mathcal{D}(\rho)$ of pairs of voters, defined as

$$ d_\mathcal{D}(\rho) = \mathbb{E}_{x \sim \mathcal{D} \mathbf{X}} \mathbb{E}_{h \sim \rho, h' \sim \rho} I[h(x) \neq h'(x)] , \quad (2) $$

as $R_\mathcal{D}(B_\rho) \leq 1 - \left( \frac{1 - 2 R_\mathcal{D}(G_\rho)}{1 - 2 d_\mathcal{D}(\rho)} \right)^2$. Note that, although relying on $d_\mathcal{D}(\rho)$, our present work does not reuse the latter result. Instead, we adopt another well-known strategy to obtain tight majority vote bounds, by

---

4Note that in this paper, we do not focus on learning a new representation to help the adaptation: We directly aim at adapting in the current representation space.

5The quantity $d_\mathcal{D}(\rho)$ is also used in the domain adaptation bound of Germain et al. (2013) to measure divergence between distributions. See forthcoming Theorem 2.
specializing our PAC-Bayesian bound to linear classifiers. We describe this approach, and refer to related works, in Section 7.

3 Some Previous Domain Adaptation Bounds

Many approaches tackling domain adaptation share the same underlying “philosophy”, pulling its origins in the work of Ben-David et al. (2006; 2010) which proposed a domain adaptation bound (Theorem 1, below). To summarize, the domain adaptation bounds reviewed in this section (see Zhang et al., 2012; Cortes et al., 2010; 2015, for other bounds) express a similar trade-off between three terms: (i) the source risk, (ii) the distance between source and target marginal distributions over $X$, (iii) a non-estimable term (without target label) quantifying the difficulty of the task.

In Ben-David et al. (2006), the domains are first assumed to be related, in the sense that there exists a representation space in which this assumption can be verified. In such a scenario, the domain adaptation strategy is then to look for a set $H$ of possible models that behave “similarly” on both the source and target data, and to learn a model in $H$ with a good accuracy on the source data.

Theorem 1 (Ben-David et al., 2006; 2010). Let $H$ be a (symmetric\textsuperscript{*}) hypothesis class. We have,

$$\forall h \in H, \quad R_T(h) \leq R_S(h) + \frac{1}{2} d_{\Delta H}(S_X, T_X) + \mu_{\ast}.$$  \hfill (3)

Pursuing in the same line of research, Mansour et al. (2009) generalizes the $\Delta H$-distance to real-valued loss functions $L : [-1,1]^2 \rightarrow \mathbb{R}^+$, to express a similar theorem for regression. Their discrepancy $\text{disc}_L(S_X, T_X)$ is defined as

$$\sup_{(h, h')} \left| \mathbf{E}_{x \sim S_X} L(h(x), h'(x)) - \mathbf{E}_{x \sim T_X} L(h(x), h'(x)) \right|.$$ 

The accuracy of the Mansour et al. (2009)'s bound also relies on a non-estimable term assumed to be low when adaptation is achievable. Roughly, this term depends on the risk of the best target hypothesis and its agreement with the best source hypothesis on the source domain.

Building on previous domain adaptation analyses, Germain et al. (2013) derived a PAC-Bayesian domain adaptation bound. This bound is based on a divergence suitable for PAC-Bayes, i.e., for the risk of a $\rho$-weighted majority vote of the voters of $H$ (instead of a single classifier $h \in H$). This domain disagreement $\text{dis}_\rho(S_X, T_X)$ is defined as

$$\text{dis}_\rho(S_X, T_X) = \left| d_S(\rho) - d_T(\rho) \right|.$$  \hfill (4)

In favorable adaptation situations, they make the strong assumption that the learned posterior agrees with the best target one $\rho_T = \arg\min \rho, R_T(G_\rho)$. Indeed, their bound (Theorem 2, below) relies on the following non-estimable term: $\lambda(\rho) = R_T(G_{\rho_T}) + \mathbf{E}_{h \sim \rho} \mathbf{E}_{h' \sim \rho_T} \mathbf{E}_{x \sim S_X} I[h(x) \neq h'(x)] + \mathbf{E}_{h \sim \rho} \mathbf{E}_{h' \sim \rho_T} \mathbf{E}_{x \sim T_X} I[h(x) \neq h'(x)]$.

Theorem 2 (Germain et al., 2013). Let $H$ be a set of voters. For any domains $S$ and $T$ over $X \times Y$, we have,

$$\forall \rho \text{ on } H, \quad R_T(G_\rho) \leq R_S(G_\rho) + \text{dis}_\rho(S_X, T_X) + \lambda(\rho).$$

A compelling aspect of this PAC-Bayesian analysis is the suggested trade-off, which is function of $\rho$. Indeed, given a fixed instance space $X$ and a fixed class $H$, apart from using importance weighting methods, the only way to minimize the bound of Theorem 1 is to find $h \in H$ that minimizes $R_S(h)$. Conversely,\textsuperscript{*}In a symmetric $H$, for all $h \in H$, its inverse $-h$ is also in $H$. 

Technical Report V 3 3
in Germain et al. (2013), the bound of Theorem 2 inspired an algorithm—named PBDA—selecting \( \rho \) over \( \mathcal{H} \) that achieves a trade-off between \( R_\mathcal{T}(G_\rho) \) and \( \text{dis}_\rho(S_\mathcal{X}, T_\mathcal{X}) \). However, the term \( \lambda(\rho) \) does not appear in the optimization process of PBDA, even if it relies on the learned weight distribution \( \rho \). Germain et al. (2013) argued that the value of \( \lambda(\rho) \) should be negligible (uniformly for all \( \rho \)) when adaptation is achievable. Nevertheless, this strong assumption cannot be verified because the best target posterior distribution \( \rho_\mathcal{T}^* \) is unknown. We claim this is a major weakness of the work of Germain et al. (2013) that the approach of this paper overcomes.

4 A New Domain Adaptation Perspective

In this section, we introduce an original approach to upper-bound the non-estimable risk of a \( \rho \)-weighted majority vote on a target distribution \( \mathcal{T} \) thanks to a term depending on its marginal distribution \( T_\mathcal{X} \), another one on a related source domain \( \mathcal{S} \), and a term capturing the “volume” of the source distribution uninformative for the target task. We base our bound on the expected disagreement \( d_\mathcal{D}(\rho) \) of Equation (2) and the expected joint error \( e_\mathcal{D}(\rho) \), defined as

\[
e_\mathcal{D}(\rho) = E_{(x,y) \sim \mathcal{D}} E_{h \sim \rho} E_{h' \sim \rho} [I[h(x) \neq y] 1[h'(x) \neq y]].
\]

(5)

Indeed, Lacasse et al. (2006) observed that, given a domain \( \mathcal{D} \) on \( \mathcal{X} \times \mathcal{Y} \) and a distribution \( \rho \) on \( \mathcal{H} \), we can decompose the Gibbs risk as

\[
R_\mathcal{D}(G_\rho) = \frac{1}{2} E_{(x,y) \sim \mathcal{D}} E_{h \sim \rho} E_{h' \sim \rho} [I[h(x) \neq y] 1[h'(x) \neq y]] \]

\[= E_{(x,y) \sim \mathcal{D}} E_{h \sim \rho} E_{h' \sim \rho} [I[h(x) \neq h'(x)] + 2 I[h(x) \neq y \land h'(x) \neq y]] \]

\[= \frac{1}{2} d_\mathcal{D}(\rho) + e_\mathcal{D}(\rho).
\]

(6)

A key observation is that the voters’ disagreement does not rely on labels; we can compute \( d_\mathcal{D}(\rho) \) using the marginal distribution \( P_\mathcal{X} \). Thus, in the present domain adaptation context, we have access to \( d_\mathcal{T}(\rho) \) even if the target labels are unknown. However, the expected joint error can only be computed on the labeled source domain.

Domains’ divergence. In order to link the target joint error \( e_\mathcal{T}(\rho) \) with the source one \( e_\mathcal{S}(\rho) \), we weight the latter thanks to a divergence measure between the domains \( \beta_q(\mathcal{T}||\mathcal{S}) \) parametrized by a real value \( q > 0 \):

\[
\beta_q(\mathcal{T}||\mathcal{S}) = \left[ E_{(x,y) \sim \mathcal{S}} \left( \frac{T(x,y)}{S(x,y)} \right)^q \right]^{\frac{1}{q}}.
\]

(7)

It is worth noting that considering some \( q \) values allows us to recover well-known divergences. For instance, choosing \( q = 2 \) relates our result to the \( \chi^2 \)-distance, as \( \beta_2(\mathcal{T}||\mathcal{S}) = \sqrt{\chi^2(\mathcal{T}||\mathcal{S}) + 1} \). Moreover, we can link \( \beta_q(\mathcal{T}||\mathcal{S}) \) to the Rényi divergence\(^5\), which has led to generalization bounds in the context of importance weighting (Cortes et al., 2010). We denote the limit case \( q \rightarrow \infty \) by

\[
\beta_\infty(\mathcal{T}||\mathcal{S}) = \sup_{(x,y) \in \text{supp}(\mathcal{S})} \left( \frac{T(x,y)}{S(x,y)} \right),
\]

(8)

where \( \text{supp}(\mathcal{S}) \) is the support of domain \( \mathcal{S} \).

This divergence \( \beta_q(\mathcal{T}||\mathcal{S}) \) handles the input space areas where the source distribution support \( \text{supp}(\mathcal{S}) \) is included in the target one \( \text{supp}(\mathcal{T}) \). It seems reasonable to assume that, when adaptation is achievable, such areas are fairly large. However, it is also likely that \( \text{supp}(\mathcal{T}) \) is not entirely included in \( \text{supp}(\mathcal{S}) \). We denote \( \mathcal{T} \setminus \mathcal{S} \) the distribution of \( (x,y) \sim \mathcal{T} \) conditional to \( (x,y) \in \text{supp}(\mathcal{T}) \setminus \text{supp}(\mathcal{S}) \). Since it is hardly conceivable to estimate the joint error \( e_{\mathcal{T} \setminus \mathcal{S}}(\rho) \) without making extra assumptions, we consider the worst possible risk for this unknown area. To do so, we define

\[
\eta_{\mathcal{T} \setminus \mathcal{S}} = \sup_{(x,y) \notin \text{supp}(\mathcal{S})} \left( \frac{P_{\mathcal{T}}(x,y)}{\text{supp}(\mathcal{S})} \right) \sup_{h \in \mathcal{H}} R_{\mathcal{T} \setminus \mathcal{S}}(h).
\]

(9)

Even if we cannot evaluate \( \sup_{h \in \mathcal{H}} R_{\mathcal{T} \setminus \mathcal{S}}(h) \), the value of \( \eta_{\mathcal{T} \setminus \mathcal{S}} \) is necessarily lower than \( P_{\mathcal{T}}((x,y) \notin \text{supp}(\mathcal{S})) \).

---

\(^5\)For \( q \geq 0 \), we can show \( \beta_q(\mathcal{T}||\mathcal{S}) = 2^{\frac{1}{q}} - D_q(\mathcal{T}||\mathcal{S}) \), where \( D_q(\mathcal{T}||\mathcal{S}) \) is the Rényi divergence between \( \mathcal{T} \) and \( \mathcal{S} \).
The domain adaptation bound. Let us state the result underlying the domain adaptation perspective of this paper.

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

$$R_T(G_\rho) \leq \frac{1}{2} d_T(\rho) + \beta_q(T\|\mathcal{S}) \times \left[ e_\mathcal{S}(\rho) \right]^{1 - \frac{1}{q}} + \eta_{T\setminus \mathcal{S}},$$

where $d_T(\rho)$, $e_\mathcal{S}(\rho)$, $\beta_q(T\|\mathcal{S})$ and $\eta_{T\setminus \mathcal{S}}$ are respectively defined by Equations (2), (5), (7) and (9).

**Proof.** Let us define $t = E_{(x, y) \sim T} [(x, y) \notin \text{supp}(\mathcal{S})]$, then

$$\eta_q = E_{(x, y) \sim T} I[(x, y) \notin \text{supp}(\mathcal{S})] E_{h \sim \rho, h' \sim \rho} I[h(x) \neq y] I[h'(x) \neq y]$$

$$= E_{(x, y) \sim T \setminus \mathcal{S}, h \sim \rho, h' \sim \rho} I[h(x) \neq y] I[h'(x) \neq y] = t e_{T \setminus \mathcal{S}}(\rho)$$

$$= t \left( R_{T \setminus \mathcal{S}}(G_\rho) - \frac{1}{2} d_{T \setminus \mathcal{S}}(\rho) \right) \leq t \sup_{h \in \mathcal{H}} R_{T \setminus \mathcal{S}}(h) = \eta_{T \setminus \mathcal{S}}.$$

Then, with $\beta_q = \beta_q(T\|\mathcal{S})$ and $p$ such that $\frac{1}{p} = 1 - \frac{1}{q}$,

$$e_T(\rho) = E_{(x, y) \sim T} E_{h \sim \rho, h' \sim \rho} I[h(x) \neq y] I[h'(x) \neq y]$$

$$= E_{(x, y) \sim \mathcal{S}, (x', y) \sim \mathcal{S}, h \sim \rho, h' \sim \rho} I[h(x) \neq y] I[h'(x) \neq y] + \eta_q$$

$$\leq \beta_q E_{h \sim \rho, h' \sim \rho} E_{(x, y) \sim \mathcal{S}} \left[ (1 - I[h(x) \neq y] I[h'(x) \neq y])^q \right]^{\frac{1}{q}} + \eta_q. (10)$$

Last line is due to H"{o}lder inequality. Finally, we remove the exponent from expression $(1 - I[h(x) \neq y] I[h'(x) \neq y])^p$ without affecting its value, which is either 1 or 0, and the final result follows from Equation (6).

Note that the bound of Theorem 3 is reached whenever the domains are equal ($\mathcal{S} = \mathcal{T}$). Thus, when adaptation is not necessary, our analysis is still sound and non-degenerated:

$$R_{\mathcal{S}}(G_\rho) = R_T(G_\rho) \leq \frac{1}{2} d_T(\rho) + 1 \times [e_\mathcal{S}(\rho)]^{\frac{1}{q}} + 0$$

$$= \frac{1}{2} d_\mathcal{S}(\rho) + e_\mathcal{S}(\rho) = R_{\mathcal{S}}(G_\rho).$$

**Meaningful quantities.** Similarly to the previous results recalled in Section 3, our domain adaptation theorem bounds the target risk by a sum of three terms. However, our approach breaks the problem into atypical quantities: (i) The expected disagreement $d_T(\rho)$ captures second degree information about the target domain. (ii) The domains’ divergence $\beta_q(T\|\mathcal{S})$ weights the influence of the expected joint error $e_\mathcal{S}(\rho)$ of the source domain; the parameter $q$ allows us to consider different relationships between $\beta_q(T\|\mathcal{S})$ and $e_\mathcal{S}(\rho)$. (iii) The term $\eta_{T \setminus \mathcal{S}}$ quantifies the worst feasible target error on the regions where the source domain is uninformative for the target one. In the current work, we assume that this area is small.

## 5 Comparison With Related Works

In this section, we discuss how our domain adaptation bound can be related to some previous works.

### 5.1 On the previous PAC-Bayesian bound.

It is instructive to compare our Theorem 3 with the previous PAC-Bayesian domain adaptation bound of Theorem 2. In our bound, the non-estimable terms are the domain divergence $\beta_q(T\|\mathcal{S})$ and the term $\eta_{T \setminus \mathcal{S}}$. Contrary to the non-controllable term $\lambda(\rho)$ of Theorem 2, these terms do not depend on the learned posterior distribution $\rho$: For every $\rho$ on $\mathcal{H}$, $\beta_q(T\|\mathcal{S})$ and $\eta_{T \setminus \mathcal{S}}$ are constant values measuring the relation between the domains. Moreover, the fact that the domain divergence $\beta_q(T\|\mathcal{S})$ is not an additive term but a multiplicative one (as opposed to $d_{\text{dis}}(\mathbf{x}_T, \mathcal{T}_S) + \lambda(\rho)$ in Theorem 2) is a contribution of our analysis. Consequently, $\beta_q(T\|\mathcal{S})$ can be viewed as a hyperparameter allowing us to tune the trade-off between the target voters’ disagreement and the source joint error. Experiments of Section 8 confirm that this hyperparameter can be successfully selected.
5.2 On some domain adaptation assumptions

In order to characterize which domain adaptation task may be learnable, Ben-David et al. (2012) presented three assumptions that can help domain adaptation. Our Theorem 3 does not rely on these assumptions, but they can be interpreted in our framework as discussed below.

On the covariate shift. A domain adaptation task fulfills the covariate shift assumption (Shimodaira, 2000) if the source and target domains only differ in their marginals according to the input space, i.e., $T_{Y|X}(y) = S_{Y|X}(y)$. In this scenario, one may estimate $\beta_0(T||S)$, and even $\eta_{T\setminus S}$, by using unsupervised density estimation methods. Interestingly, by also assuming that the domains share the same support, we have $\eta_{T\setminus S} = 0$. Then from Line (10) we obtain

$$R_T(G_\rho) = \frac{1}{2} d_T(\rho) + \frac{1}{Z} \sum_{x \sim S_X} \mathbb{E}_{h \sim \rho} \mathbb{E}[I[h(x) \neq y]I[h'(x) \neq y]],$$

which suggests a way to correct the shift between the domains by reweighting the labeled source distribution, while considering the information from the target disagreement.

On the weight ratio. The weight ratio (Ben-David et al., 2012) of source and target domains, with respect to a collection of input space subsets $B \subseteq 2^X$, is given by

$$C_B(S, T) = \inf_{\beta \in B} \frac{S_X(b)}{T_X(b)}.$$

When $C_B(S, T)$ is bounded away from 0, adaptation should be achievable under covariate shift. In this context, and when $\text{supp}(S) = \text{supp}(T)$, the limit case of $\beta_0(T||S)$ is equal to the inverse of the point-wise weight ratio obtained by letting $B = \{\{x\} : x \in X\}$ in $C_B(S, T)$. Indeed, both $\beta_0$ and $C_B$ compare the densities of source and target domains, but provide distinct strategies to relax the point-wise weight ratio; the former by lowering the value of $\rho$ and the latter by considering larger subspaces $B$.

On the cluster assumption. A target domain fulfills the cluster assumption when examples of the same labels belong to a common “area” of the input space, and when the differently labeled “areas” are well separated by low-density regions (this concept has been mathematically formalized by the probabilistic Lipschitzness notion in Urner et al., 2011). As we shall see in Section 7, once specialized to linear classifiers, $d_T(\rho)$ behaves nicely in this context.

5.3 On representation learning

The main assumption underlying our domain adaptation algorithm exhibited in Section 7 is that the support of the target domain is mostly included in the support of the source domain, i.e., the value of the term $\eta_{T \setminus S}$ is small. When $T \setminus S$ is sufficiently large to prevent proper adaptation, one could try to reduce its volume, while taking care to preserve a good compromise between $d_T(\rho)$ and $e_S(\rho)$, using a representation learning approach, i.e., by projecting source and target examples into a new common input space, as done for example by Chen et al. (2012); Ganin & Lempitsky (2015).

6 PAC-Bayesian Generalization Guarantees

To compute our domain adaptation bound, one needs to know the distributions $S$ and $T_X$, which is never the case in real life tasks. The PAC-Bayesian theory provides tools to convert the bound of Theorem 3 into a generalization bound on the target risk computable from a pair of source-target samples $(S, T) \sim (S)^m \times (T_X)^m$. To achieve this, we first provide generalization guarantees for $d_T(\rho)$ and $e_S(\rho)$. These results are presented as corollaries of Theorem 4 below, that generalizes a PAC-Bayesian theorem of Catoni (2007) to arbitrary loss functions. Indeed, Theorem 4, with $\ell(h, x, y) = I[h(x) \neq y]$ and Equation (1), gives the usual bound on the Gibbs risk.

**Theorem 4.** For any domain $D$ over $X \times Y$, any set of voters $\mathcal{H}$, any prior $\pi$ over $\mathcal{H}$, any loss $\ell : \mathcal{H} \times X \times Y \rightarrow [0, 1]$, any real number $c > 0$, with a probability at least $1 - \delta$ over the choice of $\{(x_i, y_i)\}_{i=1}^m \sim (D)^m$, we have for all $\rho$ on $\mathcal{H}$:

$$\mathbb{E}_{(x, y) \sim D \sim \rho} \mathbb{E}[\ell(h, x, y)] \leq \frac{c}{1 - e^{-c}} \left[ \frac{1}{m} \sum_{i=1}^m \mathbb{E}[\ell(h, x_i, y_i)] + \frac{\text{KL}(\rho||\pi) + \ln \frac{2}{\delta}}{m \times c} \right].$$

**Proof.** Reported in Appendix A. \(\square\)
Note that, similarly to McAllester & Keshet (2011), we could choose to restrict $c \in (0, 2)$ to obtain a slightly looser but simpler bound. Using $e^{-c} \leq 1 - c - \frac{1}{2}c^2$, an upper bound on the right hand side of above equation is given by
\[
\frac{1}{1 - e^{-c}} \frac{1}{m} \sum_{i=1}^{m} E_{h \sim \rho} \ell(h, x_i, y_i) + \frac{\text{KL}(\rho \| \pi) + \ln \frac{1}{c}}{m \times c}.
\]
We now exploit Theorem 4 to obtain generalization guarantees on the expected disagreement and the expected joint error. PAC-Bayesian bounds on these quantities appeared in Lacasse et al. (2006), but under different forms. In Corollary 5 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 $\text{KL}(\rho \| \pi)$—with the help of parameters $b$ and $c$.

**Corollary 5.** For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X} \times \mathbf{Y}$, any set of voters $\mathcal{H}$, any prior $\pi$ over $\mathcal{H}$, any $\delta \in (0, 1]$, any real numbers $b > 0$ and $c > 0$, we have:
- with a probability at least $1 - \delta$ over $\mathcal{T} \sim (\mathcal{T}_X)^{m_t}$,
  \[
  \forall \rho \text{ on } \mathcal{H}, \hat{d}_T(\rho) \leq \frac{c}{1 - e^{-c}} \left[ \hat{d}_T(\rho) + \frac{2\text{KL}(\rho \| \pi) + \ln \frac{1}{c}}{m_t \times c} \right],
  \]
- with a probability at least $1 - \delta$ over $\mathcal{S} \sim (\mathcal{S}_X)^{m_s}$,
  \[
  \forall \rho \text{ on } \mathcal{H}, \hat{e}_S(\rho) \leq \frac{b}{1 - e^{-c}} \left[ \hat{e}_S(\rho) + \frac{2\text{KL}(\rho \| \pi) + \ln \frac{1}{c}}{m_s \times b} \right],
  \]
where $\hat{d}_T(\rho)$ and $\hat{e}_S(\rho)$ are the empirical estimations of the target voters’ disagreement and the source joint error.

**Proof.** Given $\pi$ and $\rho$ over $\mathcal{H}$, we consider a new prior $\pi^2$ and a new posterior $\rho^2$, both over $\mathcal{H}^2$, such that:
\[\forall h_{ij} = (h_i, h_j) \in \mathcal{H}^2, \pi^2(h_{ij}) = \pi(h_i)\pi(h_j), \text{ and } \rho^2(h_{ij}) = \rho(h_i)\rho(h_j)\]. Thus, $\text{KL}(\rho^2 \| \pi^2) = 2\text{KL}(\rho \| \pi)^2)$ (see Lacasse et al., 2006). Let us define two new loss functions for a “paired voter” $h_{ij} \in \mathcal{H}^2$:
\[
\ell_d(h_{ij}, x, y) = I[h_i(x) \neq h_j(x)],
\]
and
\[
\ell_e(h_{ij}, x, y) = I[h_i(x) \neq y] \times I[h_j(x) \neq y].
\]
Then, the bound on $d_T(\rho)$ is obtained from Theorem 4 with $\ell := \ell_d$, and Equation (2). The bound on $e_S(\rho)$ is similarly obtained with $\ell := \ell_e$ and using Equation (5).\]

For algorithmic simplicity, we deal with Theorem 3 when $q \to \infty$. Thanks to Corollary 5, we obtain the following generalization bound defined with respect to the empirical estimates of the target disagreement and the source joint error.

**Theorem 6.** For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X} \times \mathbf{Y}$, any set of voters $\mathcal{H}$, any prior $\pi$ over $\mathcal{H}$, any $\delta \in (0, 1]$, any $b > 0$ and $c > 0$, with a probability at least $1 - \delta$ over the choices of $\mathcal{S} \sim (\mathcal{S}_X)^{m_s}$ and $\mathcal{T} \sim (\mathcal{T}_X)^{m_t}$, we have
\[
\forall \rho \text{ on } \mathcal{H}, R_T(G_\rho) \leq c' \frac{1}{m} \frac{1}{1 - e^{-c}} \left[ \hat{d}_T(\rho) + b' \hat{e}_S(\rho) + \eta_{\mathcal{T} \setminus \mathcal{S}} \right] + \left( c' \frac{1}{m_s \times c} + b' \frac{1}{m_s \times b} \right) \left( 2\text{KL}(\rho \| \pi) + \ln \frac{1}{c} \right),
\]
where $\hat{d}_T(\rho)$ and $\hat{e}_S(\rho)$ are the empirical estimations of the target voters’ disagreement and the source joint error, and $b' = \frac{b}{1 - e^{-c}} \beta_\infty(\mathcal{T} \| \mathcal{S})$, and $c' = \frac{c}{1 - e^{-c}}$.

**Proof.** We bound separately $d_T(\rho)$ and $e_S(\rho)$ using Corollary 5 (with probability $1 - \frac{\delta}{2}$ each), and then combine the two upper bounds according to Theorem 3.\]

From an optimization perspective, the problem suggested by the bound of Theorem 6 is much more convenient to minimize than the PAC-Bayesian bound that Germain et al. (2013) derived from Theorem 2. The former is smoother than the latter: the absolute value related to the domain disagreement $d_\rho(\mathbf{S}_x, \mathbf{T}_x)$ of Equation (4) disappears in benefit of the domain divergence $\beta_\infty(\mathcal{T} \| \mathcal{S})$, which is constant and can be considered as an hyperparameter of the algorithm. Additionally, Theorem 2 requires equal source and target sample sizes while Theorem 6 allows $m_s \neq m_t$. Moreover, recall that Germain et al. (2013) choose to ignore the $\rho$-dependent non-constant term $\lambda(\rho)$. In our case such compromise is not mandatory in order to apply the theoretical result to real problems, since our non-estimable term $\eta_{\mathcal{T} \setminus \mathcal{S}}$ is constant and does not depend on the learned $\rho$. Hence, we can neglect $\eta_{\mathcal{T} \setminus \mathcal{S}}$ without any impact on the optimization problem described in the next section. Moreover, it is realistic to consider this is a small quantity in situations where the source and target supports are similar.
Then, \( R_y \) is related to the linear classifier margin \( x_\text{classifier} \) where \( \Phi(\cdot) \) is a surrogate—sometimes called the probit loss (e.g., McAllester & Keshet, 2011)—of the zero-one loss function \( I(x \leq 0) \) relying on \( y \frac{w \cdot x}{\|w\|} \). Note that \( \|w\| \) plays an important role on the value of \( R_D(G_{\rho_w}) \), but not on \( R_D(h_w) \). Indeed, \( R_D(G_{\rho_w}) \) tends to \( R_D(h_w) \) as \( \|w\| \) grows, which can provide very tight bounds (see the empirical analyses of Ambroladze et al., 2006; Germain et al., 2009b). In the PAC-Bayesian context, \( \|w\| \) turns out to be a measure of complexity of the learned classifier, as Equation (11) shows.

### 7 Specialization to Linear Classifiers

In order to derive an algorithm, we now specialize the bounds of Theorems 3 and 6 to the risk of a linear classifier \( h_w \), defined by a weight vector \( w \in \mathbb{R}^d \):

\[
\forall x \in X, \quad h_w(x) = \text{sign}(w \cdot x).
\]

The taken approach is the one privileged in numerous PAC-Bayesian works (e.g., Langford & Shawe-Taylor, 2002; Ambroladze et al., 2006; McAllester & Keshet, 2011; Parrado-Hernández et al., 2012; Germain et al., 2009a; 2013), as it makes the risk of the linear classifier \( h_w \) and the risk of a (properly parametrized) majority vote coincide, while in the same time promoting large margin classifiers. To this end, let \( \mathcal{H} \) be the set of all linear classifiers over the input space, \( \mathcal{H} = \{h_w \mid w \in \mathbb{R}^d\} \), and let \( \rho_w \) over \( \mathcal{H} \) be a posterior distribution, respectively a prior distribution \( \pi_0 \), that is constrained to be a spherical Gaussian with identity covariance matrix centered on vector \( w \), respectively \( 0 \),

\[
\forall h_w' \in \mathcal{H}, \quad \rho_w(h_w') = \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\frac{1}{2} \|w'-w\|^2},
\]

\[
\pi_0(h_w) = \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\frac{1}{2} \|w\|^2}.
\]

The KL-divergence between \( \rho_w \) and \( \pi_0 \) simply is

\[
\text{KL}(\rho_w \| \pi_0) = \frac{1}{2} \|w\|^2. \tag{11}
\]

Thanks to this parameterization, the majority vote classifier \( B_{\rho_w} \) corresponds to the one of the linear classifier \( h_w \) (see above cited PAC-Bayesian works). That is,

\[
\forall x \in X, w \in \mathcal{H}, \quad h_w(x) = \text{sign} \left[ E_{h_w' \sim \rho_w} (x) \right] = B_{\rho_w}(x).
\]

Then, \( R_D(h_w) = R_D(B_{\rho_w}) \) for any data distribution \( D \).

Moreover, Langford & Shawe-Taylor (2002) showed that the closely related Gibbs risk (Equation (1)) is related to the linear classifier margin \( y \frac{w \cdot x}{\|w\|} \), as follows:

\[
R_D(G_{\rho_w}) = E_{(x,y) \sim D} \Phi \left(y \frac{w \cdot x}{\|w\|} \right), \tag{12}
\]

where \( \Phi(x) = \frac{1}{2} - \frac{1}{2} \text{Erf}(x \sqrt{2}) \), and \( \text{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \) is the Gauss error function. Here, \( \Phi(x) \) can be seen as a smooth surrogate—sometimes called the probit loss (e.g., McAllester & Keshet, 2011)—of the zero-one loss function \( I(x \leq 0) \) relying on \( y \frac{w \cdot x}{\|w\|} \).
We now seek to express the expected disagreement \(d_D(\rho_w)\) and the expected joint error \(e_D(\rho_w)\) of Equations (2) and (5) related to the parameterized distribution \(\rho_w\). As shown by Germain et al. (2013) the former is given by

\[
d_D(\rho_w) = \mathbb{E}_{x \sim \mathcal{D}_x} \Phi_{\text{dis}}(w \cdot x / \|x\|),
\]

where \(\Phi_{\text{dis}}(x) = 2 \times \Phi(x) \times \Phi(-x)\). Following a similar approach, we obtain, for all \(w \in \mathbb{R}\),

\[
e_D(\rho_w) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{E}_{h \sim \mathcal{P}_{\rho_w}} \mathbb{E}_{h' \sim \mathcal{P}_{\rho_w}} [1|h(x) \neq y|1|h'(x) \neq y]
\]

\[
= \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{E}_{h \sim \mathcal{P}_{\rho_w}} [1|h(x) \neq y|] \mathbb{E}_{h' \sim \mathcal{P}_{\rho_w}} [1|h'(x) \neq y] + \mathbb{E}_{(x,y) \sim \mathcal{D}} \Phi_{\text{err}} \left(y \frac{w \cdot x}{\|x\|}\right),
\]

with \(\Phi_{\text{err}}(x) = [\Phi(x)]^2\). As function \(\Phi\) in Equation (12), functions \(\Phi_{\text{err}}\) and \(\Phi_{\text{dis}}\) defined above can be interpreted as loss functions for linear classifiers (illustrated by Figure 1).

**Domain adaptation bound.** Theorem 3 specialized to linear classifiers gives the following corollary. Note that, as mentioned above, \(R_T(h_w) = R_T(B_{\rho_w}) \leq 2 R_T(G_{\rho_w})\).

**Corollary 7.** Let \(\mathcal{S}\) and \(\mathcal{T}\) respectively be the source and the target domains on \(X \times Y\). For all \(w \in \mathbb{R}\), we have:

\[
R_T(h_w) \leq d_D(\rho_w) + 2 \beta_\infty(T\|S\) \times e_S(\rho_w) + 2 \eta_{\mathcal{T}\setminus\mathcal{S}}.
\]

Figure 1 leads to an insightful geometric interpretation of the domain adaptation trade-off promoted by Corollary 7. For fixed values of \(\beta_\infty(T\|S)\) and \(\eta\), the target risk \(R_T(h_w)\) is upper-bounded by a (\(\beta_\infty\)-weighted) sum of two losses. The expected \(\Phi_{\text{err}}\)-loss (i.e., the joint error) is computed on the (labeled) source domain; it aims to label the source examples correctly, but is more permissive on the required margin than the \(\Phi\)-loss (i.e., the Gibbs risk). The expected \(\Phi_{\text{dis}}\)-loss (i.e., the disagreement) is computed on the target (unlabeled) domain; it promotes large unsigned target margins. Thus, if a target domain fulfills the cluster assumption (described in Section 5.2), \(d_D(\rho_w)\) will be low when the decision boundary crosses a low-density region between the homogeneous labeled clusters. Hence, Corollary 7 reflects that some source errors may be allowed if, doing so, the separation of the target domain is improved.

**Generalization bound and learning algorithm.** Theorem 6 specialized to linear classifiers gives the following.

**Corollary 8.** For any domains \(\mathcal{S}\) and \(\mathcal{T}\) over \(X \times Y\), any \(\delta \in (0,1]\), any \(a > 0\) and \(b > 0\), with a probability at least \(1 - \delta\) over the choices of \(\mathcal{S} \sim (\mathcal{S})^m_\mathcal{S}\) and \(\mathcal{T} \sim (\mathcal{T}_\mathcal{S})^m_\mathcal{S}\), we have

\[
\forall w \in \mathbb{R} : R_T(h_w) \leq c' \tilde{d}_T(\rho_w) + 2b' \tilde{e}_S(\rho_w) + 2 \eta_{\mathcal{T}\setminus\mathcal{S}}
\]

\[
+ 2 \left( \frac{c}{m_{\mathcal{T}\setminus\mathcal{S}}} + \frac{b}{m_{\mathcal{T}}}ight) \left(\|w\|^2 + \ln \frac{2}{\delta}\right).
\]

For a source \(\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{m_\mathcal{S}}\) and a target \(\mathcal{T} = \{(x'_i)\}_{i=1}^{m_\mathcal{T}}\) samples of potentially different size, and some hyperparameters \(C > 0\), \(B > 0\), minimizing the next objective function w.r.t. \(w \in \mathbb{R}\) is equivalent to minimize the above bound.

\[
C \tilde{d}_T(\rho_w) + B \tilde{e}_S(\rho_w) + \|w\|^2
\]

\[
= C \sum_{i=1}^{m_\mathcal{S}} \Phi_{\text{dis}} \left(\frac{w \cdot x_i}{\|x_i\|}\right) + B \sum_{i=1}^{m_\mathcal{T}} \Phi_{\text{err}} \left(y_i \frac{w \cdot x'_i}{\|x'_i\|}\right) + \|w\|^2.
\]

We call the optimization of Equation (13) by gradient descent the Dalc algorithm, for Domain Adaptation of Linear Classifiers. The kernel trick applies to Dalc. That is, given a kernel \(k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}\), one can express a linear classifier in a RKHS\(^8\) by a dual weight vector \(a \in \mathbb{R}^{m_\mathcal{S} + m_\mathcal{T}}\):

\[
h_w(\cdot) = \text{sign} \left[ \sum_{i=1}^{m_\mathcal{T}} \alpha_i k(x_i, \cdot) + \sum_{i=1}^{m_\mathcal{S}} \alpha_{i+m_\mathcal{T}} k(x'_i, \cdot) \right].
\]

\(^8\)It is non-trivial to show that the kernel trick holds when \(\pi_0\) and \(\rho_w\) are Gaussian over infinite-dimensional feature space. As mentioned by McAllester & Keshet (2011), it is, however, the case provided we consider Gaussian processes as measure of distributions \(\pi_0\) and \(\rho_w\) over (infinite) \(H\).
Table 1: Error rates on Amazon dataset. Best risks appear in **bold** and seconds are in *italic*.

<table>
<thead>
<tr>
<th></th>
<th>SVM (CV)</th>
<th>DASVM (RCV)</th>
<th>CODA (RCV)</th>
<th>PBDA (RCV)</th>
<th>DALC (RCV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>books→ DVDs</td>
<td>0.179</td>
<td>0.193</td>
<td>0.181</td>
<td>0.183</td>
<td><strong>0.178</strong></td>
</tr>
<tr>
<td>books→ electro</td>
<td>0.290</td>
<td>0.226</td>
<td>0.232</td>
<td>0.263</td>
<td><strong>0.212</strong></td>
</tr>
<tr>
<td>books→ kitchen</td>
<td>0.251</td>
<td><strong>0.179</strong></td>
<td>0.215</td>
<td>0.229</td>
<td><strong>0.194</strong></td>
</tr>
<tr>
<td>DVDs→ books</td>
<td>0.203</td>
<td>0.202</td>
<td>0.217</td>
<td>0.197</td>
<td><strong>0.186</strong></td>
</tr>
<tr>
<td>DVDs→ electro</td>
<td>0.269</td>
<td><strong>0.186</strong></td>
<td>0.214</td>
<td>0.241</td>
<td>0.245</td>
</tr>
<tr>
<td>DVDs→ kitchen</td>
<td>0.232</td>
<td>0.183</td>
<td>0.181</td>
<td>0.186</td>
<td><strong>0.175</strong></td>
</tr>
<tr>
<td>electro→ books</td>
<td>0.287</td>
<td>0.305</td>
<td>0.275</td>
<td><strong>0.232</strong></td>
<td><strong>0.240</strong></td>
</tr>
<tr>
<td>electro→ DVDs</td>
<td>0.267</td>
<td><strong>0.214</strong></td>
<td>0.239</td>
<td>0.221</td>
<td>0.256</td>
</tr>
<tr>
<td>electro→ kitchen</td>
<td>0.129</td>
<td>0.149</td>
<td>0.134</td>
<td>0.141</td>
<td><strong>0.123</strong></td>
</tr>
<tr>
<td>kitchen→ books</td>
<td>0.267</td>
<td>0.259</td>
<td>0.247</td>
<td><strong>0.234</strong></td>
<td><strong>0.236</strong></td>
</tr>
<tr>
<td>kitchen→ DVDs</td>
<td>0.253</td>
<td><strong>0.198</strong></td>
<td>0.223</td>
<td>0.233</td>
<td><strong>0.225</strong></td>
</tr>
<tr>
<td>kitchen→ electro</td>
<td>0.149</td>
<td>0.157</td>
<td>0.153</td>
<td><strong>0.129</strong></td>
<td><strong>0.131</strong></td>
</tr>
</tbody>
</table>

Average | 0.231 | 0.204 | 0.210 | 0.208 | **0.200** |

Even though the objective function is highly non-convex, we achieved good empirical results by minimizing the “kernelized” version of Equation (13) by gradient descent, with a uniform weight vector as a starting point. More details are given in Appendix B.

8 Experimental Results

Firstly, Figure 2 illustrates the behavior of the decision boundary of our algorithm DALC on an intertwining moons toy problem\(^9\), 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\(^\circ\). We see that DALC does not rely on the restrictive covariate shift assumption, as some source examples are misclassified. This behavior illustrates the DALC trade-off in action, that concedes some errors on the source sample to lower the disagreement on the target sample.

Secondly, we evaluate DALC on the classical Amazon.com Reviews benchmark (Blitzer et al., 2006) according to the setting used by Chen et al. (2011); 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 into 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 pre-processed 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 is “books” and the target one is “DVDs”. We compare DALC with the classical non-adaptive algorithm SVM (trained only on the source sample), the adaptive algorithm DASVM (Bruzzone & Marconcini, 2010), the adaptive co-training CODA (Chen et al., 2011), and the PAC-Bayesian domain adaptation algorithm PBDA (Germain et al., 2013) based on Theorem 2. Note that, in Germain et al. (2013), DASVM has shown better accuracy than SVM, CODA and PBDA. Each parameter is selected with a grid search thanks to an usual cross-validation (CV) on the source sample for SVM, and thanks to a reverse validation procedure\(^10\) (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 adaptive 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 adaptive tasks between “electronics” and “DVDs”, DALC is either the best one (six times), or the second one (four times). Moreover, DALC clearly increases the performance over the previous PAC-Bayesian algorithm (PBDA), which confirms that our new bound improves the analysis done by Germain et al. (2013).

\(^9\) We generate each pair of moons with the make_moons function provided in scikit-learn (Pedregosa et al., 2011).

\(^10\) For details on the reverse validation procedure, see Bruzzone & Marconcini (2010); Zhong et al. (2010). Other details on our experimental protocol are given in Appendix C.
9 Conclusion

We propose a new domain adaptation analysis for majority vote learning. It relies on an upper bound on the target risk, expressed as a trade-off between the voters’ disagreement on the target domain, the voters’ joint errors on the source one, and a term reflecting the worst case error in regions where the source domain is non-informative. To the best of our knowledge, a crucial novelty of our contribution is that the trade-off is controlled by the divergence $\beta_q(T\|S)$ (Equation 7) between the domains: The divergence is not an additive term (as in many domain adaptation bounds) but is a factor weighting the importance of the source information. Our analysis, combined with a PAC-Bayesian generalization bound, leads to a new domain adaptation algorithm for linear classifiers. The empirical experiments show that our new algorithm outperforms the previous PAC-Bayesian approach of Germain et al. (2013).

As future work, we first aim at investigating the case where the domains’ divergence $\beta_q(T\|S)$ can be estimated, i.e., when the covariate shift assumption holds or when some target labels are available. In these scenarios, $\beta_q(T\|S)$ might not be considered as a hyperparameter to tune.

Last but not least, the term $\eta_{T\setminus S}$ of our bound—suggesting that the two domains should live in the same regions—can be dealt with a representation learning approach. As mentioned in Section 5.3, this could be an incentive to combine our learning algorithm with existing representation learning techniques. In another vein, considering an active learning setup (as in Berlind & Urner, 2015), one could query the labels of target examples to estimate the value bounded by $\eta_{T\setminus S}$. We see this as a great source of inspiration for new algorithms for this learning paradigm.

Appendix

A Proof of Theorem 4

Proof. We use the following shorthand notation:

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

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

$$
\forall \rho \text{ on } \mathcal{H} : \quad m \times \Delta \left( \mathbb{E}_{h \sim \rho} \mathcal{L}_S(h), \mathbb{E}_{h \sim \rho} \mathcal{L}_D(h) \right) \leq \mathbb{E}_{h \sim \rho} m \times \Delta (\mathcal{L}_S(h), \mathcal{L}_D(h)) \\
\leq \text{KL}(\rho || \pi) + \ln \left[ X_\pi(S) \right],
$$

with

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

Then, Markov’s Inequality gives

$$
\Pr_{S \sim D^m} \left( X_\pi(S) \leq \frac{1}{\delta} \mathbb{E}_{S' \sim D^m} X_\pi(S') \right) \geq 1 - \delta,
$$

and

$$
\mathbb{E}_{S' \sim D^m} X_\pi(S') = \mathbb{E}_{S' \sim D^m} \mathbb{E}_{h \sim \pi} e^{m \times \Delta (\mathcal{L}_{S'}(h), \mathcal{L}_D(h))} \\
= \mathbb{E}_{h \sim \pi} \mathbb{E}_{S' \sim D^m} e^{m \times \Delta (\mathcal{L}_{S'}(h), \mathcal{L}_D(h))} \\
\leq \mathbb{E}_{h \sim \pi} \sum_{k=0}^m \frac{k^k}{m^k} (\mathcal{L}_D(h))^{k/2} \mathcal{L}_D(h)^{m-k} e^{m \times \Delta (\frac{1}{\pi}, \mathcal{L}_D(h))}, \quad (14)
$$

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

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

Technical Report V 3 11
Line 14 becomes equal to 1, and then $\mathbb{E}_{S' \sim \mathcal{D}'^m} X(S') \leq 1$. Hence,

$$\Pr_{S \sim \mathcal{D}^m} \left( \forall \rho \in \mathcal{H} : -c \mathbb{E}_{h \sim \rho} \mathcal{L}_S(h) - \ln \left( 1 - \mathbb{E}_{h \sim \rho} \mathcal{L}_D(h) \right) \leq \frac{\text{KL}(\rho \| \pi) + \ln \frac{1}{m}}{m} \right) \geq 1 - \delta.$$ 

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

$$\forall \rho \in \mathcal{H} : \mathbb{E}_{h \sim \rho} \mathcal{L}_D(h) \leq \frac{1}{1 - e^{-c}} \left[ 1 - \exp \left( -c \mathbb{E}_{h \sim \rho} \mathcal{L}_S(h) - \frac{\text{KL}(\rho \| \pi) + \ln \frac{1}{m}}{m} \right) \right].$$

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

\section{Using dalc with a kernel function}

Let $S=\{(x_i, y_i)\}_{i=1}^{m_s}$, $T=\{x'_i\}_{i=1}^{m_t}$ and $m = m_s + m_t$. We will denote $x_# = \begin{cases} x_i & \text{if } # \leq m_s \text{ (source examples)} \\ x'_{#-m_s} & \text{otherwise. (target examples)} \end{cases}$

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

$$h_{\mathbf{w}}(\cdot) = \text{sign} \left[ \sum_{i=1}^{m} \alpha_i k(x_i, \cdot) \right].$$

Let us denote $K$ the kernel matrix of size $m \times m$ such as $K_{i,j} = k(x_i, x_j)$. In that case, the objective function—Equation (13) of the main paper—can be rewritten in term of the vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ as

$$C \times \sum_{i=m_s+1}^{m} \Phi \left( \frac{\sum_{j=1}^{m_s} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) \Phi \left( -\frac{\sum_{j=1}^{m_s} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) + B \times \sum_{i=1}^{m_s} \left[ \Phi \left( y_i \sum_{j=1}^{m_s} \alpha_j K_{i,j} \right) \right]_2 + \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K_{i,j}.$$ 

For our experiments, we minimize this objective function using a Broyden-Fletcher-Goldfarb-Shanno method (BFGS) implemented in the scipy python library (Jones et al., 2001–).

We initialize the optimization procedure at $\alpha_i = \frac{1}{m}$ for all $i \in \{1, \ldots, m\}$.

\section{Experimental Protocol}

For obtaining the DALC$^{RCV}$ results of Table 1, the reverse validation procedure searches on a $20 \times 20$ 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).

\section{References}


Ben-David, S., Shalev-Shwartz, S., and Urner, R. Domain adaptation–can quantity compensate for quality? In ISAIM, 2012.


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


Urner, R., Shalev-Shwartz, S., and Ben-David, S. Access to unlabeled data can speed up prediction time. In ICML, pp. 641–648, 2011.
