A New PAC-Bayesian Perspective on Domain Adaptation

Pascal Germain PASCAL.GERMAIN@INRIA.FR
INRIA, SIERRA Project-Team, 75589, Paris, France, and D.I., Ecole Normale Superieure, 75230 Paris, France

Amaury Habrard AMAURY.HABRARD@UNIV-ST-ETIENNE.FR
Univ Lyon, UJM-Saint-Etienne, CNRS, IOGS, Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France

François Laviolette FRANCOIS.LAVIOLETTE@IFT.ULAVAL.CA
Département d'informatique et de génie logiciel, Université Laval, Québec, Canada

Emilie Morvant EMILIE.MORVANT@UNIV-ST-ETIENNE.FR
Univ Lyon, UJM-Saint-Etienne, CNRS, IOGS, Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France

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*[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 common example is the spam filtering problem where a model needs to be adapted from one user mailbox to another receiving significantly different emails. Many approaches exist to address domain adaptation, often with the same idea: If we can apply a transformation to "move closer" the distributions, then we can learn a model with the available labels. This 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 common representation for the source and target distributions (Chen et al., 2012; Ganin et al., 2016), and/or by minimizing a measure of divergence between the distributions (Morvant et al., 2012; Germain et al., 2013; Cortes & Mohri, 2014). 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[2] without target label. Even the latter result opened 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); Mansour et al. (2009): The risk of the target model is upper-bounded jointly by the model's risk on the source distribution, the divergence between the marginal distributions, and a non-estimable term[3] related to the ability to adapt in the current space. Note that Li & Bilmes (2007) proposed a PAC-Bayesian generalization bound for domain adaptation but they considered target labels.

[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 algorithms can be seen as a majority vote learning. *E.g.,* ensemble learning and kernel methods output models interpretable as majority votes.

[3] More precisely, this term can only be estimated in the presence of labeled data from both the source and the target domains.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

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 they 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 source and the target domains that enables controlling the relationship between domains. The third term estimates the "volume" of the target domain living apart from the source one[4], 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 analysis of domain adaptation for majority votes, that we relate to other works in Section 5. Then, Section 6 provides generalization bounds, specialized to linear classifiers in Section 7 to motivate the DALC learning algorithm, evaluated in Section 8.

2. Unsupervised Domain Adaptation Setting

We tackle *domain adaptation* for *binary classification*, 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 $\mathcal{S}$—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* $T = \{\mathbf{x}_i\}_{i=1}^{m_t} \sim (\mathcal{T}_{\mathbf{X}})^{m_t}$.

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 previously studied in Germain et al. (2013). Given $\mathcal{H}$, a set of voters $h : \mathbf{X} \to Y$, the elements of this 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—before observing (S, T)—of the voters' accuracy. Then, given the information provided by (S, T), we aim at learning a posterior distribution ρ leading to a *ρ-weighted majority vote* over $\mathcal{H}$,

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

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 :

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

where $\mathrm{I}[a] = 1$ 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 ρ, designed as the *Gibbs risk* :

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

It is well-known in the PAC-Bayesian literature that $\mathrm{R}_{\mathcal{D}}(B_\rho) \leq 2\,\mathrm{R}_{\mathcal{D}}(G_\rho)$ (*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) (refined in Germain et al., 2015) have exhibited that one can obtain a tighter bound on $\mathrm{R}_{\mathcal{D}}(B_\rho)$ by studying the *expected disagreement* $\mathrm{d}_{\mathcal{D}}(\rho)$ of pairs of voters, defined as

$$\mathrm{d}_{\mathcal{D}}(\rho) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{h'\sim\rho} \mathrm{I}\big[h(\mathbf{x}) \neq h'(\mathbf{x})\big], \qquad (2)$$

as $\mathrm{R}_{\mathcal{D}}(B_\rho) \leq 1 - \frac{(1-2\,\mathrm{R}_{\mathcal{D}}(G_\rho))^2}{1-2\,\mathrm{d}_{\mathcal{D}}(\rho)}$. Note that, although relying on $\mathrm{d}_{\mathcal{D}}(\rho)$, our present work does not reuse the latter result.[5] Instead, we adopt another well-known strategy to obtain tight majority vote bounds, by 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

[4] Here we do not focus on learning a new representation to help the adaptation: We directly aim at adapting in the current space.

[5] The quantity $\mathrm{d}_{\mathcal{D}}(\rho)$ is also used in the domain adaptation bound of Germain et al. (2013) to measure divergence between distributions. See forthcoming Theorem 2.

bounds) express a similar trade-off between three terms: **(i)** the source risk, **(ii)** the distance between source and target marginal distributions over $\mathbf{X}$, **(iii)** a non-estimable term (without target label) quantifying the difficulty of the task. Ben-David et al. (2006) assumed that the domains are related in the sense that there exists a (unknown) model performing well on both domains. Formally, their domain adaptation bound depends on the error $\mu_{h^*} = \mathrm{R}_{\mathcal{S}}(h^*) + \mathrm{R}_{\mathcal{T}}(h^*)$ of the best hypothesis overall $h^* = \operatorname{argmin}_{h\in\mathcal{H}} \big(\mathrm{R}_{\mathcal{S}}(h) + \mathrm{R}_{\mathcal{T}}(h)\big)$. In practice, when no target label is available, μ_{h^*} is non-estimable and is assumed to be low when domain adaptation is achievable (or at least 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 $\mathcal{H}$ of possible models that behave "similarly" on both the source and target data, and to learn a model in $\mathcal{H}$ with a good accuracy on the source data. This similarity, called the *$\mathcal{H}\Delta\mathcal{H}$-distance*,

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = 2 \sup_{(h,h')\in\mathcal{H}^2} \Big| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}} \mathrm{I}\big[h(\mathbf{x}) \neq h'(\mathbf{x})\big] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}} \mathrm{I}\big[h(\mathbf{x}) \neq h'(\mathbf{x})\big] \Big|,$$

gives rise to the following domain adaptation bound.

Theorem 1 (Ben-David et al., 2006; 2010)**.** *Let $\mathcal{H}$ be a (symmetric[6]) hypothesis class. We have,*

$$\forall h \in \mathcal{H},\ \mathrm{R}_{\mathcal{T}}(h) \leq \mathrm{R}_{\mathcal{S}}(h) + \tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \mu_{h^*}. \quad (3)$$

Pursuing in the same line of research, Mansour et al. (2009) generalizes the $\mathcal{H}\Delta\mathcal{H}$-distance to real-valued loss functions $\mathcal{L} : [-1, 1]^2 \to \mathbb{R}^+$, to express a similar theorem for regression. Their *discrepancy* $\mathrm{disc}_{\mathcal{L}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ is defined as

$$\sup_{(h,h')\in\mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}} \mathcal{L}\big(h(\mathbf{x}), h'(\mathbf{x})\big) - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}} \mathcal{L}\big(h(\mathbf{x}), h'(\mathbf{x})\big) \right|.$$

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 ρ-weighted majority vote of the voters of $\mathcal{H}$ (instead of a single classifier $h \in \mathcal{H}$). This *domain disagreement* $\mathrm{dis}_{\rho}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ is defined as

$$\mathrm{dis}_{\rho}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = \big|\, \mathrm{d}_{\mathcal{S}}(\rho) - \mathrm{d}_{\mathcal{T}}(\rho) \,\big|. \quad (4)$$

Theorem 2 (below) needs the strong assumption that, in favorable adaptation situations, the learned posterior agrees with the best target one $\rho_{\mathcal{T}}{}^* = \operatorname{argmin}_{\rho} \mathrm{R}_{\mathcal{T}}(G_\rho)$. Indeed, it relies on the following non-estimable term: $\lambda(\rho) = \mathrm{R}_{\mathcal{T}}(G_{\rho_{\mathcal{T}}{}^*}) + \mathbf{E}_{h\sim\rho}\, \mathbf{E}_{h'\sim\rho_{\mathcal{T}}{}^*}\, \mathbf{E}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}}\, \mathrm{I}\big[h(\mathbf{x}) \neq h'(\mathbf{x})\big] + \mathbf{E}_{h\sim\rho}\, \mathbf{E}_{h'\sim\rho_{\mathcal{T}}{}^*}\, \mathbf{E}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}}\, \mathrm{I}\big[h(\mathbf{x}) \neq h'(\mathbf{x})\big]$.

[6] In a symmetric $\mathcal{H}$, for all $h \in \mathcal{H}$, its inverse $-h$ is also in $\mathcal{H}$.

Theorem 2 (Germain et al., 2013)**.** *Let $\mathcal{H}$ be a set of voters. For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X} \times Y$, we have,* $\forall \rho$ *on* $\mathcal{H}$, $\mathrm{R}_{\mathcal{T}}(G_\rho) \leq \mathrm{R}_{\mathcal{S}}(G_\rho) + \mathrm{dis}_{\rho}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \lambda(\rho)$.

A compelling aspect of this PAC-Bayesian analysis is the suggested trade-off, which is function of ρ. Indeed, given a fixed instance space $\mathbf{X}$ and a fixed class $\mathcal{H}$, apart from using importance weighting methods, the only way to minimize the bound of Theorem 1 is to find $h \in \mathcal{H}$ that minimizes $\mathrm{R}_{\mathcal{S}}(h)$. In Germain et al. (2013), the bound of Theorem 2 inspired an algorithm—named PBDA—selecting ρ over $\mathcal{H}$ that achieves a trade-off between $\mathrm{R}_{\mathcal{S}}(G_\rho)$ and $\mathrm{dis}_{\rho}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$. However, the term $\lambda(\rho)$ does not appear in the optimization process of PBDA, even if it relies on the learned weight distribution ρ. It is assumed that the value of $\lambda(\rho)$ should be negligible (uniformly for all ρ) when adaptation is achievable. Nevertheless, this strong assumption cannot be verified because the best target posterior distribution $\rho_{\mathcal{T}}{}^*$ is unknown. This is a major weakness of the previous PAC-Bayesian work that our new approach overcomes.

4. A New Domain Adaptation Perspective

In this section, we introduce an original approach to upper-bound the non-estimable risk of a ρ-weighted majority vote on a target distribution $\mathcal{T}$ thanks to a term depending on its marginal distribution $\mathcal{T}_{\mathbf{X}}$, 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 $\mathrm{d}_{\mathcal{D}}(\rho)$ of Equation (2) and the expected joint error $\mathrm{e}_{\mathcal{D}}(\rho)$, defined as

$$\mathrm{e}_{\mathcal{D}}(\rho) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{h'\sim\rho} \mathrm{I}\big[h(\mathbf{x}) \neq y\big]\, \mathrm{I}\big[h'(\mathbf{x}) \neq y\big]. \quad (5)$$

Indeed, Lacasse et al. (2006); Germain et al. (2015) observed that, given a domain $\mathcal{D}$ on $\mathbf{X} \times Y$ and a distribution ρ on $\mathcal{H}$, we can decompose the Gibbs risk as

$$\begin{aligned} \mathrm{R}_{\mathcal{D}}(G_\rho) &= \tfrac{1}{2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{h'\sim\rho} \mathrm{I}\big[h(\mathbf{x})\neq y\big] + \mathrm{I}\big[h'(\mathbf{x})\neq y\big] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{h'\sim\rho} \frac{\mathrm{I}\big[h(\mathbf{x})\neq h'(\mathbf{x})\big] + 2\,\mathrm{I}\big[h(\mathbf{x})\neq y \wedge h'(\mathbf{x})\neq y\big]}{2} \\ &= \tfrac{1}{2}\, \mathrm{d}_{\mathcal{D}}(\rho) + \mathrm{e}_{\mathcal{D}}(\rho)\,. \end{aligned} \quad (6)$$

A key observation is that the *voters' disagreement does not rely on labels*; we can compute $\mathrm{d}_{\mathcal{D}}(\rho)$ using the marginal distribution $\mathcal{D}_{\mathbf{X}}$. Thus, in the present domain adaptation context, we have access to $\mathrm{d}_{\mathcal{T}}(\rho)$ 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 $\mathrm{e}_{\mathcal{T}}(\rho)$ with the source one $\mathrm{e}_{\mathcal{S}}(\rho)$, we weight the latter thanks to a divergence measure between the domains

$\beta_q(\mathcal{T}\|\mathcal{S})$ parametrized by a real value $q>0$:

$$\beta_q(\mathcal{T}\|\mathcal{S}) \;=\; \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}}\left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right)^q\right]^{\frac{1}{q}}. \qquad (7)$$

It is worth noting that considering some q values allow us to recover well-known divergences. For instance, choosing $q=2$ relates our result to the χ^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[7], which has led to generalization bounds in the context of importance weighting (Cortes et al., 2010). We denote the limit case $q\to\infty$ by

$$\beta_\infty(\mathcal{T}\|\mathcal{S}) \;=\; \sup_{(\mathbf{x},y)\in\textsc{supp}(\mathcal{S})}\left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right),$$

with $\textsc{supp}(\mathcal{S})$ the support of $\mathcal{S}$. The divergence $\beta_q(\mathcal{T}\|\mathcal{S})$ handles the input space areas where the source domain support $\textsc{supp}(\mathcal{S})$ is included in the target one $\textsc{supp}(\mathcal{T})$. It seems reasonable to assume that, when adaptation is achievable, such areas are fairly large. However, it is likely that $\textsc{supp}(\mathcal{T})$ is *not entirely* included in $\textsc{supp}(\mathcal{S})$. We denote $\mathcal{T}\backslash\mathcal{S}$ the distribution of $(\mathbf{x},y)\sim\mathcal{T}$ conditional to $(\mathbf{x},y)\in\textsc{supp}(\mathcal{T})\backslash\textsc{supp}(\mathcal{S})$. Since it is hardly conceivable to estimate the joint error $\mathrm{e}_{\mathcal{T}\backslash\mathcal{S}}(\rho)$ without making extra assumptions, we define the worst risk for this *unknown* area

$$\eta_{\mathcal{T}\backslash\mathcal{S}} = \Pr_{(\mathbf{x},y)\sim\mathcal{T}}\Big((\mathbf{x},y)\notin\textsc{supp}(\mathcal{S})\Big)\sup_{h\in\mathcal{H}} R_{\mathcal{T}\backslash\mathcal{S}}(h). \quad (8)$$

Even if we cannot evaluate $\sup_{\mathcal{H}} R_{\mathcal{T}\backslash\mathcal{S}}(h)$, the value of $\eta_{\mathcal{T}\backslash\mathcal{S}}$ is necessarily lower than $\Pr_{\mathcal{T}}((\mathbf{x},y)\notin\textsc{supp}(\mathcal{S}))$.

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 $\mathbf{X}\times Y$. Let $q>0$ be a constant. We have, for all ρ on $\mathcal{H}$,*

$$\mathrm{R}_{\mathcal{T}}(G_\rho) \;\le\; \frac{1}{2}\,\mathrm{d}_{\mathcal{T}}(\rho) + \beta_q(\mathcal{T}\|\mathcal{S})\times\Big[\mathrm{e}_{\mathcal{S}}(\rho)\Big]^{1-\frac{1}{q}} + \eta_{\mathcal{T}\backslash\mathcal{S}}\,,$$

where $\mathrm{d}_{\mathcal{T}}(\rho)$, $\mathrm{e}_{\mathcal{S}}(\rho)$, $\beta_q(\mathcal{T}\|\mathcal{S})$ and $\eta_{\mathcal{T}\backslash\mathcal{S}}$ are respectively defined by Equations (2), (5), (7) and (8).

Proof. Let us define $t=\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{T}}\,\mathrm{I}\big[(\mathbf{x},y)\notin\textsc{supp}(\mathcal{S})\big]$, then

$$\begin{aligned}
\eta_\rho &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}}\mathrm{I}\big[(\mathbf{x},y)\notin\textsc{supp}(\mathcal{S})\big]\mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{h'\sim\rho}\mathrm{I}\big[h(\mathbf{x})\neq y\big]\mathrm{I}\big[h'(\mathbf{x})\neq y\big]\\
&= t\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}\backslash\mathcal{S}}\mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{h'\sim\rho}\mathrm{I}\big[h(\mathbf{x})\neq y\big]\,\mathrm{I}\big[h'(\mathbf{x})\neq y\big] = t\,\mathrm{e}_{\mathcal{T}\backslash\mathcal{S}}(\rho)\\
&= t\Big(R_{\mathcal{T}\backslash\mathcal{S}}(G_\rho)-\tfrac{1}{2}\mathrm{d}_{\mathcal{T}\backslash\mathcal{S}}(\rho)\Big) \le t\sup_{h\in\mathcal{H}} R_{\mathcal{T}\backslash\mathcal{S}}(h) = \eta_{\mathcal{T}\backslash\mathcal{S}}\,.
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{e}_{\mathcal{T}}(\rho) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}}\mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{h'\sim\rho}\mathrm{I}\big[h(\mathbf{x})\neq y\big]\,\mathrm{I}\big[h'(\mathbf{x})\neq y\big]\\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}}\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{h'\sim\rho}\mathrm{I}\big[h(\mathbf{x})\neq y\big]\mathrm{I}\big[h'(\mathbf{x})\neq y\big]+\eta_\rho \qquad (9)\\
&\le \beta_q\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{h'\sim\rho}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}}\big(\mathrm{I}\big[h(\mathbf{x})\neq y\big]\mathrm{I}\big[h'(\mathbf{x})\neq y\big]\big)^p\right]^{\frac{1}{p}}+\eta_\rho\,.
\end{aligned}$$

Last line is due to Hölder inequality. Finally, we remove the exponent from expression $(\mathrm{I}\big[h(\mathbf{x})\neq y\big]\mathrm{I}\big[h'(\mathbf{x})\neq y\big])^p$ 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:

$$\begin{aligned}
\mathrm{R}_{\mathcal{S}}(G_\rho) \;=\; \mathrm{R}_{\mathcal{T}}(G_\rho) \;&\le\; \tfrac{1}{2}\,\mathrm{d}_{\mathcal{T}}(\rho) + 1\times[\mathrm{e}_{\mathcal{S}}(\rho)]^1 + 0\\
&=\; \tfrac{1}{2}\,\mathrm{d}_{\mathcal{S}}(\rho) + \mathrm{e}_{\mathcal{S}}(\rho) =\; \mathrm{R}_{\mathcal{S}}(G_\rho)\,.
\end{aligned}$$

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 $\mathrm{d}_{\mathcal{T}}(\rho)$ captures *second degree* information about the target domain. **(ii)** The domains' divergence $\beta_q(\mathcal{T}\|\mathcal{S})$ weights the influence of the expected joint error $\mathrm{e}_{\mathcal{S}}(\rho)$ of the source domain; the parameter q allows us to consider different relationships between $\beta_q(\mathcal{T}\|\mathcal{S})$ and $\mathrm{e}_{\mathcal{S}}(\rho)$. **(iii)** The term $\eta_{\mathcal{T}\backslash\mathcal{S}}$ 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 the new bound of Theorem 3 with the previous PAC-Bayesian domain adaptation bound of Theorem 2. In Theorem 3, the non-estimable terms are the domain divergence $\beta_q(\mathcal{T}\|\mathcal{S})$ and the term $\eta_{\mathcal{T}\backslash\mathcal{S}}$. Contrary to the non-controllable term $\lambda(\rho)$ of Theorem 2, these terms do not depend on the *learned* posterior distribution ρ: For every ρ on $\mathcal{H}$, $\beta_q(\mathcal{T}\|\mathcal{S})$ and $\eta_{\mathcal{T}\backslash\mathcal{S}}$ are constant values measuring the relation between the domains. Moreover, the fact that the domain divergence $\beta_q(\mathcal{T}\|\mathcal{S})$ is not an additive term but a multiplicative one (as opposed to $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X})+\lambda(\rho)$ in Theorem 2) is a contribution of our new analysis. Consequently, $\beta_q(\mathcal{T}\|\mathcal{S})$ 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.

[7] For $q\ge 0$, we can show $\beta_q(\mathcal{T}\|\mathcal{S})=2^{\frac{q-1}{q}D_q(\mathcal{T}\|\mathcal{S})}$, where $D_q(\mathcal{T}\|\mathcal{S})$ is the Rényi divergence between $\mathcal{T}$ and $\mathcal{S}$.

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.,* $\mathcal{T}_{Y|\mathbf{x}}(y) = \mathcal{S}_{Y|\mathbf{x}}(y)$. In this scenario, one may estimate $\beta_q(\mathcal{T}_{\mathbf{X}} \| \mathcal{S}_{\mathbf{X}})$, and even $\eta_{\mathcal{T}\setminus\mathcal{S}}$, by using unsupervised density estimation methods. Interestingly, by also assuming that the domains share the same support, we have $\eta_{\mathcal{T}\setminus\mathcal{S}} = 0$. Then from Line (9) we obtain

$$\mathrm{R}_{\mathcal{T}}(G_\rho) = \tfrac{1}{2}\mathrm{d}_{\mathcal{T}}(\rho) + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}} \frac{\mathcal{T}_{\mathbf{X}}(\mathbf{x})}{\mathcal{S}_{\mathbf{X}}(\mathbf{x})} \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{h'\sim\rho} \mathrm{I}\big[h(\mathbf{x}) \neq y\big] \mathrm{I}\big[h'(\mathbf{x}) \neq y\big],$$

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 $\mathcal{B} \subseteq 2^{\mathbf{X}}$, is given by

$$C_{\mathcal{B}}(\mathcal{S}, \mathcal{T}) \;=\; \inf_{b\in\mathcal{B},\, \mathcal{T}_{\mathbf{X}}(b)\neq 0} \frac{\mathcal{S}_{\mathbf{X}}(b)}{\mathcal{T}_{\mathbf{X}}(b)} \,.$$

When $C_{\mathcal{B}}(\mathcal{S}, \mathcal{T})$ is bounded away from 0, adaptation should be achievable under covariate shift. In this context, and when $\textsc{supp}(\mathcal{S}) = \textsc{supp}(\mathcal{T})$, the limit case of $\beta_\infty(\mathcal{T}\|\mathcal{S})$ is equal to the inverse of the *point-wise weight ratio* obtained by letting $\mathcal{B} = \{\{\mathbf{x}\} : \mathbf{x} \in \mathbf{X}\}$ in $C_{\mathcal{B}}(\mathcal{S}, \mathcal{T})$. Indeed, both β_q and $C_{\mathcal{B}}$ 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 q and the latter by considering larger subspaces $\mathcal{B}$.

On the cluster assumption. A target domain fulfills the *cluster assumption* when examples of the same label belong to a common "area" of the input space, and the differently labeled "areas" are well separated by *low-density regions* (formalized by the *probabilistic Lipschitzness* of Urner et al., 2011). Once specialized to linear classifiers, $\mathrm{d}_{\mathcal{T}}(\rho)$ behaves nicely in this context (see Section 7).

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 $\eta_{\mathcal{T}\setminus\mathcal{S}}$ is small. When $\mathcal{T}\setminus\mathcal{S}$ is sufficiently large to prevent proper adaptation, one could try to reduce its volume while taking care to preserve a good compromise between $\mathrm{d}_{\mathcal{T}}(\rho)$ and $\mathrm{e}_{\mathcal{S}}(\rho)$, using a *representation learning* approach, *i.e.*, by projecting source and target examples into a new common space, as done for example by Chen et al. (2012); Ganin et al. (2016).

6. PAC-Bayesian Generalization Guarantees

To compute our domain adaptation bound, one needs to know the distributions $\mathcal{S}$ and $\mathcal{T}_{\mathbf{X}}$, 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 $(S, T) \sim (\mathcal{S})^{m_s} \times (\mathcal{T}_{\mathbf{X}})^{m_t}$. To achieve this, we first provide generalization guarantees for $\mathrm{d}_{\mathcal{T}}(\rho)$ and $\mathrm{e}_{\mathcal{S}}(\rho)$. These results are presented as corollaries of Theorem 4 below, that generalizes a PAC-Bayesian theorem of Catoni (2007) to arbitrary loss functions.[8] Indeed, Theorem 4, with $\ell(h, \mathbf{x}, y) = \mathrm{I}\big[h(\mathbf{x}) \neq y\big]$ 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 π over $\mathcal{H}$, any loss $\ell : \mathcal{H}\times\mathbf{X}\times Y \to [0,1]$, any real number $c > 0$, with a probability at least $1-\delta$ over the choice of $\{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{D})^m$, we have for all ρ on $\mathcal{H}$:*

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

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 $\frac{1}{1-\frac{1}{2}c}\left[\frac{1}{m}\sum_{i=1}^m \mathbf{E}_{h\sim\rho}\,\ell(h, \mathbf{x}_i, y_i) + \frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m\times c}\right]$.

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 Germain et al. (2015), 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 $\mathrm{KL}(\rho\|\pi)$—with the help of parameters b and c.

Corollary 5. *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}\times Y$, any set of voters $\mathcal{H}$, any prior π 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 $T \sim (\mathcal{T}_{\mathbf{X}})^{m_t}$,

$$\forall \rho \text{ on } \mathcal{H},\; \mathrm{d}_{\mathcal{T}}(\rho) \;\leq\; \frac{c}{1-e^{-c}} \left[\widehat{\mathrm{d}}_T(\rho) + \frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m_t \times c} \right],$$

— with a probability at least $1-\delta$ over $S \sim (\mathcal{S})^{m_s}$,

$$\forall \rho \text{ on } \mathcal{H},\; \mathrm{e}_{\mathcal{S}}(\rho) \;\leq\; \frac{b}{1-e^{-b}} \left[\widehat{\mathrm{e}}_S(\rho) + \frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m_s \times b} \right],$$

where $\widehat{\mathrm{d}}_T(\rho)$ and $\widehat{\mathrm{e}}_S(\rho)$ are the empirical estimations of the target voters' disagreement and the source joint error.

[8] To do so, we exploit a result of Maurer (2004) that allows to generalize PAC-Bayes theorems to arbitrary bounded loss function (see the proof of Theorem 4 in supplemental).

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, $\mathrm{KL}(\rho^2\|\pi^2) = 2\mathrm{KL}(\rho^2\|\pi^2)$ (see Germain et al., 2015). Let us define two new loss functions for a "paired voter" $h_{ij} \in \mathcal{H}^2$:

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

Then, the bound on $\mathrm{d}_{\mathcal{T}}(\rho)$ is obtained from Theorem 4 with $\ell \coloneqq \ell_d$, and Equation (2). The bound on $\mathrm{e}_{\mathcal{S}}(\rho)$ is similarly obtained with $\ell \coloneqq \ell_e$ 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 $\mathbf{X}\times Y$, any set of voters $\mathcal{H}$, any prior π 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 $S\sim(\mathcal{S})^{m_s}$ and $T\sim(\mathcal{T}_{\mathbf{X}})^{m_t}$, we have*

$$\forall \rho \text{ on } \mathcal{H},\; \mathrm{R}_{\mathcal{T}}(G_\rho) \leq\; c'\,\tfrac{1}{2}\,\widehat{\mathrm{d}}_T(\rho) + b'\,\widehat{\mathrm{e}}_S(\rho) + \eta_{\mathcal{T}\setminus\mathcal{S}} + \left(\tfrac{c'}{m_t \times c} + \tfrac{b'}{m_s \times b}\right)\left(2\,\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2}{\delta}\right),$$

where $\widehat{\mathrm{d}}_T(\rho)$ and $\widehat{\mathrm{e}}_S(\rho)$ are the empirical estimations of the target voters' disagreement and the source joint error, and $b' = \frac{b}{1-e^{-b}}\beta_\infty(\mathcal{T}\|\mathcal{S})$, and $c' = \frac{c}{1-e^{-c}}$.

Proof. We bound separately $\mathrm{d}_{\mathcal{T}}(\rho)$ and $\mathrm{e}_{\mathcal{S}}(\rho)$ 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 derived from Theorem 2 in Germain et al. (2013). The former is *smoother* than the latter: the absolute value related to the domain disagreement $\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ 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 in Germain et al. (2013) the ρ-dependent non-constant term $\lambda(\rho)$ is ignored. In our new analysis, such compromise is not mandatory in order to apply the theoretical result to real problems, since the non-estimable term $\eta_{\mathcal{T}\setminus\mathcal{S}}$ is constant and does not depend on the learned ρ. Hence, we can neglect $\eta_{\mathcal{T}\setminus\mathcal{S}}$ without any impact on the optimization problem described in the next section. Beside, it is realistic to consider $\eta_{\mathcal{T}\setminus\mathcal{S}}$ as a small quantity in situations where the source and target supports are similar.

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 $h_{\mathbf{w}}$, defined by a weight vector $\mathbf{w} \in \mathbb{R}^d$:

$$\forall \mathbf{x} \in \mathbf{X},\;\; h_{\mathbf{w}}(\mathbf{x}) \;=\; \mathrm{sign}\,(\mathbf{w}\cdot\mathbf{x})\,.$$

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., 2009; 2013), as it makes the risk of the linear classifier $h_{\mathbf{w}}$ 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, $\mathcal{H} = \left\{h_{\mathbf{w}'} \mid \mathbf{w}' \in \mathbb{R}^d\right\}$, and let $\rho_{\mathbf{w}}$ over $\mathcal{H}$ be a *posterior* distribution, *resp.* a prior distribution $\pi_{\mathbf{0}}$, that is constrained to be a spherical Gaussian with identity covariance matrix centered on vector $\mathbf{w}$, *resp.* $\mathbf{0}$,

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

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

$$\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}}) = \tfrac{1}{2}\|\mathbf{w}\|^2\,. \tag{10}$$

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

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

Then, $\mathrm{R}_{\mathcal{D}}(h_{\mathbf{w}}) = \mathrm{R}_{\mathcal{D}}(B_{\rho_{\mathbf{w}}})$ for any data distribution $\mathcal{D}$. Moreover, Langford & Shawe-Taylor (2002) showed that the closely related Gibbs risk (Equation 1) is related to the linear classifier margin $y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}$, as follows:

$$\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}}) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \Phi\left(y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right), \tag{11}$$

where $\Phi(x) = \frac{1}{2} - \frac{1}{2}\mathbf{Erf}\left(\frac{x}{\sqrt{2}}\right)$, and $\mathbf{Erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t$ 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 $\mathrm{I}\big[x \leq 0\big]$ relying on $y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}$. Note that $\|\mathbf{w}\|$ plays an important role on the value of $\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}})$, but not on $\mathrm{R}_{\mathcal{D}}(h_{\mathbf{w}})$. Indeed, $\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}})$ tends to $\mathrm{R}_{\mathcal{D}}(h_{\mathbf{w}})$ as $\|\mathbf{w}\|$ grows, which can provide *very tight bounds* (see the empirical analyses of Ambroladze et al., 2006; Germain et al., 2009). In the PAC-Bayesian context, $\|\mathbf{w}\|$ turns out to be a measure of *complexity* of the learned classifier, as Equation (10) shows. We now seek to express the *expected disagreement* $\mathrm{d}_{\mathcal{D}}(\rho_{\mathbf{w}})$ and the *expected joint error* $\mathrm{e}_{\mathcal{D}}(\rho_{\mathbf{w}})$ of Equations (2)

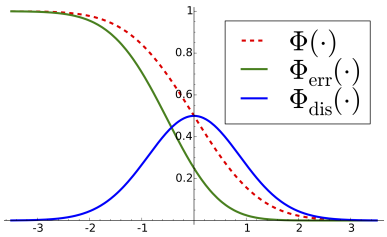

Figure 1. Graphical representation of the loss functions given by the specialization to linear classifiers.

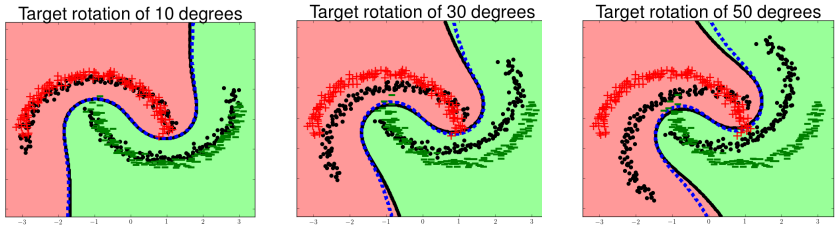

Figure 2. Decision boundaries of DALC on the *intertwining moons* toy problem, for fixed parameters $B{=}C{=}1$, and a RBF kernel $k(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2)$. The target points are black. The positive, *resp.* negative, source points are red, *resp.* green. The blue dashed line shows the decision boundaries of algorithm PBDA (Germain et al., 2013).

and (5) related to the parameterized distribution $\rho_{\mathbf{w}}$. As shown in Germain et al. (2013) the former is given by

$$\mathrm{d}_{\mathcal{D}}(\rho_{\mathbf{w}}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \Phi_{\text{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right),$$

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

$$\begin{aligned}\mathrm{e}_{\mathcal{D}}(\rho_{\mathbf{w}}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho_{\mathbf{w}}} \mathop{\mathbf{E}}_{h'\sim\rho_{\mathbf{w}}} \mathrm{I}\big[h(\mathbf{x})\neq y\big]\mathrm{I}\big[h'(\mathbf{x})\neq y\big] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho_{\mathbf{w}}} \mathrm{I}\big[h(\mathbf{x})\neq y\big] \mathop{\mathbf{E}}_{h'\sim\rho_{\mathbf{w}}} \mathrm{I}\big[h'(\mathbf{x})\neq y\big] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \Phi_{\text{err}}\left(y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right),\end{aligned}$$

with $\Phi_{\text{err}}(x) = \big[\Phi(x)\big]^2$. As function Φ in Equation (11), functions Φ_{err} and Φ_{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, $\mathrm{R}_{\mathcal{T}}(h_{\mathbf{w}}) = \mathrm{R}_{\mathcal{T}}(B_{\rho_{\mathbf{w}}}) \leq 2\,\mathrm{R}_{\mathcal{T}}(G_{\rho_{\mathbf{w}}})$.

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

$$\mathrm{R}_{\mathcal{T}}(h_{\mathbf{w}}) \;\leq\; \mathrm{d}_{\mathcal{T}}(\rho_{\mathbf{w}}) + 2\,\beta_\infty(\mathcal{T}\|\mathcal{S}) \times \mathrm{e}_{\mathcal{S}}(\rho_{\mathbf{w}}) + 2\,\eta_{\mathcal{T}\setminus\mathcal{S}}\,,$$

Figure 1 leads to an insightful geometric interpretation of the domain adaptation trade-off promoted by Corollary 7. For fixed values of $\beta_\infty(\mathcal{T}\|\mathcal{S})$ and $\eta_{\mathcal{T}\setminus\mathcal{S}}$, the target risk $\mathrm{R}_{\mathcal{T}}(h_{\mathbf{w}})$ is upper-bounded by a (β_∞-weighted) sum of two losses. The expected Φ_{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 Φ-loss (*i.e.*, the Gibbs risk). The expected Φ_{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), $\mathrm{d}_{\mathcal{T}}(\rho_{\mathbf{w}})$ 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 $\mathbf{X}\times Y$, any $\delta\in(0,1]$, any $a{>}0$ and $b{>}0$, with a probability at least $1{-}\delta$ over the choices of $S{\sim}(\mathcal{S})^{m_s}$ and $T{\sim}(\mathcal{T}_{\mathbf{X}})^{m_t}$, we have*

$$\begin{aligned}\forall\mathbf{w}\in\mathbb{R}:\ \mathrm{R}_{\mathcal{T}}(h_{\mathbf{w}}) \leq\; & c'\,\widehat{\mathrm{d}}_T(\rho_{\mathbf{w}}) + 2\,b'\,\widehat{\mathrm{e}}_S(\rho_{\mathbf{w}}) + 2\,\eta_{\mathcal{T}\setminus\mathcal{S}} \\ & + 2\left(\tfrac{c'}{m_t\times c} + \tfrac{b'}{m_s\times b}\right)\left(\|\mathbf{w}\|^2 + \ln\tfrac{2}{\delta}\right).\end{aligned}$$

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

$$\begin{aligned}& C\,\widehat{\mathrm{d}}_T(\rho_{\mathbf{w}}) + B\,\widehat{\mathrm{e}}_S(\rho_{\mathbf{w}}) + \|\mathbf{w}\|^2 \qquad (12)\\ &= C\sum_{i=1}^{m_t}\Phi_{\text{dis}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}'_i}{\|\mathbf{x}'_i\|}\right) + B\sum_{i=1}^{m_s}\Phi_{\text{err}}\left(y_i\,\tfrac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right) + \|\mathbf{w}\|^2\,.\end{aligned}$$

We call the optimization of Equation (12) by gradient descent the DALC algorithm, for Domain Adaptation of Linear Classifiers. The kernel trick applies to DALC. That is, given a kernel $k:\mathbb{R}^d{\times}\mathbb{R}^d{\to}\mathbb{R}$, one can express a linear classifier in a *RKHS*[9] by a dual weight vector $\boldsymbol{\alpha}\in\mathbb{R}^{m_s+m_t}$:

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

Even though the objective function is highly non-convex, we achieved good empirical results by minimizing the "kernelized" version of Equation (12) by gradient descent, with a uniform weight vector as a starting point. More details are given in the supplementary material.

8. Experimental Results

Firstly, Figure 2 illustrates the behavior of the decision boundary of our algorithm DALC on an intertwining moons toy problem[10], where each moon corresponds to a label.

[9] It is non-trivial to show that the kernel trick holds when $\pi_{\mathbf{0}}$ and $\rho_{\mathbf{w}}$ 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 $\pi_{\mathbf{0}}$ and $\rho_{\mathbf{w}}$ over (infinite) $\mathcal{H}$.

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

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→electro	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→electro	0.269	**0.186**	*0.214*	0.241	0.245
DVDs→kitchen	0.232	0.183	*0.181*	0.186	**0.175**
electro→books	0.287	0.305	0.275	**0.232**	*0.240*
electro→DVDs	0.267	**0.214**	0.239	*0.221*	0.256
electro→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→electro	0.149	0.157	0.153	**0.129**	*0.131*
Average	0.231	*0.204*	0.210	0.208	**0.200**

The target domain, for which we have no label, is a rotation of the source one. 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 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 a usual cross-validation (CV) on the source sample for SVM, and thanks to a reverse validation procedure[11] (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, 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, according to a Wilcoxon signed rank test with a 5% significance level, we obtain a probability of 89.5% that DALC is better than PBDA. This test tends to confirm that our new bound improves the analysis done previously in Germain et al. (2013), in addition to being more interpretable.

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 $\beta_q(\mathcal{T}\|\mathcal{S})$ (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 (Germain et al., 2013).

As future work, we first aim at investigating the case where the domains' divergence $\beta_q(\mathcal{T}\|\mathcal{S})$ can be estimated, *i.e.*, when the covariate shift assumption holds or when some target labels are available. In these scenarios, $\beta_q(\mathcal{T}\|\mathcal{S})$ might not be considered as a hyperparameter to tune.
Last but not least, the term $\eta_{\mathcal{T}\setminus\mathcal{S}}$ 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 $\eta_{\mathcal{T}\setminus\mathcal{S}}$. We see this as a great source of inspiration for new algorithms for this learning paradigm.

[11] For details on the reverse validation procedure, see Bruzzone & Marconcini (2010); Zhong et al. (2010). Other details on our experimental protocol are given in supplementary material.

Acknowledgements

This work was supported in part by the French project LIVES ANR-15-CE23-0026-03, and in part by NSERC discovery grant 262067.

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.

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

Berlind, C. and Urner, R. Active nearest neighbors in changing environments. In *ICML*, pp. 1870–1879, 2015.

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.

Chen, M., Xu, Z. E., Weinberger, K. Q., and Sha, F. Marginalized denoising autoencoders for domain adaptation. In *ICML*, pp. 767–774, 2012.

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.

Cortes, C., Mohri, M., and Medina, A. Muñoz. Adaptation algorithm and theory based on generalized discrepancy. In *ACM SIGKDD*, pp. 169–178, 2015.

Daumé III, H. Frustratingly easy domain adaptation. In *ACL*, 2007.

Ganin, Y. and Lempitsky, V. S. Unsupervised domain adaptation by backpropagation. In *ICML*, pp. 1180–1189, 2015.

Ganin, Y., Ustinova, E., H, Ajakan, Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *JMLR*, 17(59):1–35, 2016.

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

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.

Germain, P., Lacasse, A., Laviolette, F., Marchand, ., and Roy, J.-F. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *JMLR*, 16:787–860, 2015.

Herbrich, R. and Graepel, T. A PAC-Bayesian margin bound for linear classifiers: Why svms work. In *NIPS*, pp. 224–230, 2000.

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.

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.

Li, X. and Bilmes, J. A Bayesian divergence prior for classiffier adaptation. In *AISTATS*, pp. 275–282, 2007.

Liu, Q., Mackey, A. J., Roos, D. S., and Pereira, F. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics*, 24(5): 597–605, 2008.

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

McAllester, D. A. and Keshet, J. Generalization bounds and consistency for latent structural probit and ramp loss. In *NIPS*, pp. 2205–2212, 2011.

Morvant, E., Habrard, A., and Ayache, S. Parsimonious Unsupervised and Semi-Supervised Domain Adaptation with Good Similarity Functions. *KAIS*, 33(2):309–349, 2012.

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.

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.

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

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.