A New PAC-Bayesian View of Domain Adaptation

Pascal Germain, François Laviolette, Amaury Habrard, Emilie Morvant

A New PAC-Bayesian View of Domain Adaptation

Pascal Germain
IFT-GLO, Université Laval
Québec (QC), Canada
pascal.germain@ift.ulaval.ca

Amaury Habrard
LaHC, UMR CNRS 5516
Univ. of St-Etienne, France
amaury.habrard@univ-st-etienne.fr

François Laviolette
IFT-GLO, Université Laval
Québec (QC), Canada
francois.laviolette@ift.ulaval.ca

Emilie Morvant
LaHC, UMR CNRS 5516
Univ. of St-Etienne, France
emilie.morvant@univ-st-etienne.fr

Abstract

We propose a new theoretical study of domain adaptation for majority vote classifiers (from a source to a target domain). We upper bound the target risk by a trade-off between only two terms: The voters' joint errors on the source domain, and the voters' disagreement on the target one. Hence, this new study is simpler than other analyses that usually rely on three terms. We also derive a PAC-Bayesian generalization bound leading to a DA algorithm for linear classifiers.

1 Introduction

Machine learning practitioners are commonly exposed to *domain adaptation* (DA) [1, 2]: One usually learns a model from a corpus, *i.e.*, a fixed yet unknown source distribution/domain, and then wants to apply it on a new corpus, *i.e.*, a related but slightly different target distribution/domain. Several approaches exist in the literature to address DA, but often with the same idea: If the source domain is "close" to the target domain (possibly given a transformation), then one can learn a model from the source examples. This process is generally performed by iterative procedures [3, 4], and/or by reweighting the importance of labeled data [5, 6, 7], and/or by minimizing a measure of divergence between the domains [8, 9]. The divergence-based approach has especially been explored to derive generalization bounds for DA [10, 11, 12, 13, 14, 9]. Recently, this issue has been studied through the PAC-Bayesian framework [9], which focuses on learning weighted majority votes[1]. Even if this result opens the door to tackle DA in a PAC-Bayesian fashion, it shares the same philosophy than the seminal works of Ben-David et al. [11, 12] and Mansour et al. [13]: The target error is upper-bounded by a trade-off between the source error, the divergence between the marginal domains, and a non-estimable term related to the ability to adapt in the current space.

In this paper, we derive a novel DA bound relying on a simpler expression: The target error is upper-bounded by a trade-off between the voters' disagreement on the target domain, and the voters' joint errors on the source domain. The trade-off between these two terms is given by a notion of divergence between the source and the target domains. From an algorithm design perspective, an interesting characteristic of the new bound is that this trade-off expression can be dealt as a constant and thus seen as a hyperparameter to tune. Therefore, we provide a PAC-Bayesian generalization bound to justify the empirical minimization of this new DA trade-off and provide an algorithm that clearly improves the performances of the previous PAC-Bayesian DA algorithm [9].

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

2 PAC-Bayesian Domain Adaptation Setting and the Previous Analysis

Notations. We stand in the DA PAC-Bayesian setting studied in [9]. We tackle *DA binary classification tasks* from a d-dimensional input space $\mathbf{X}\subseteq\mathbb{R}^d$ to the output space $Y=\{-1,1\}$. Our goal is to perform DA from a source domain $\mathcal{S}$ on $\mathbf{X}\times Y$ to a different but related target domain $\mathcal{T}$ on $\mathbf{X}\times Y$; with marginal distribution on $\mathbf{X}$ respectively denoted $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$. Given a domain $\mathcal{D}$, we denote $(\mathcal{D})^m$ the distribution of a m-sample of m elements drawn *i.i.d.* from $\mathcal{D}$. We consider *unsupervised DA* for 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* $T=\{\mathbf{x}_i\}_{i=1}^{m_t}\sim(\mathcal{T}_\mathbf{X})^{m_t}$. Given $\mathcal{H}$, a set of voters $h:\mathbf{X}\to Y$, the "ingredients" of the *PAC-Bayesian DA* approach are a *prior* distribution π on $\mathcal{H}$, a pair of source-target learning sets (S,T) and a *posterior* distribution ρ on $\mathcal{H}$. The prior π models an *a priori* belief—before observing (S,T)—of the voters' accuracy. Then, given (S,T), the learner aims at finding a posterior ρ leading to a ρ*-weighted majority vote* over $\mathcal{H}$, $B_\rho(\cdot)=\operatorname{sign}\left[\mathbf{E}_{h\sim\rho}\,h(\cdot)\right]$, with a low true target risk: $\mathbf{R}_\mathcal{T}(B_\rho)=\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{T}}\,\mathbf{I}\left[B_\rho(\mathbf{x})\neq y\right]$, where $\mathbf{I}\left[a\right]=1$ if a is true, and 0 otherwise. However, in the PAC-Bayes framework [15, 16, 17], one does not directly focus on the risk of the deterministic B_ρ, but studies the risk of the closely related stochastic Gibbs classifier G_ρ. Given $\mathbf{x}\in\mathbf{X}$, the output of $G_\rho(\mathbf{x})$ is obtained by first drawing a voter $h\in\mathcal{H}$ according to ρ, and then returning $h(\mathbf{x})$. Thus, the risk of G_ρ on a domain $\mathcal{D}$ is the expectation of the risks according to ρ:

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

The basic relation between the deterministic B_ρ and the stochastic G_ρ is $\mathbf{R}_\mathcal{D}(B_\rho)\leq 2\mathbf{R}_\mathcal{D}(G_\rho)$. A tighter bound on $\mathbf{R}_\mathcal{D}(B_\rho)$ exists [18, 19], and depends on the *expected disagreement* $\mathbf{d}_{\mathcal{D}_\mathbf{X}}(\rho)$ and the *expected joint error* $\mathbf{e}_\mathcal{D}(\rho)$ of the pairs of voters, defined as

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

with $\rho^2(h,h')=\rho(h)\times\rho(h')$. Given $S\sim(\mathcal{D})^m$, we use $\widehat{\mathbf{R}}_S(G_\rho)$, $\widehat{\mathbf{d}}_S(\rho)$ and $\widehat{\mathbf{e}}_S(\rho)$ to denote the empirical estimation of the risk of the Gibbs classifier, the disagreement and the joint error respectively. Note that, given a domain $\mathcal{D}$, the starting point of our work is the following simple observation:

$$\forall\rho\text{ on }\mathcal{H},\ \mathbf{R}_\mathcal{D}(G_\rho)=\frac{1}{2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\;\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\Big(\mathbf{I}\left[h(\mathbf{x})\neq y\right]+\mathbf{I}\left[h'(\mathbf{x})\neq y\right]\Big)\;=\;\frac{1}{2}\,\mathbf{d}_{\mathcal{D}_\mathbf{X}}(\rho)+\mathbf{e}_\mathcal{D}(\rho)\,. \tag{3}$$

The Previous PAC-Bayesian DA Analysis and Algorithm. Inspired by the seminal DA analyses [11, 12, 13], the following PAC-Bayesian DA bound was derived by Germain et al. [9]. It is based on a divergence between distributions suitable for the stochastic Gibbs classifier (see Eq. (4)).

Theorem 1 (Germain et al. [9])**.** *Let $\mathcal{H}$ be a set of voters. For any domains $\mathcal{S}$ and $\mathcal{T}$, we have:*

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

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

$$\operatorname{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X})\;=\;\left|\,\mathbf{d}_{\mathcal{S}_\mathbf{X}}(\rho)-\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)\,\right|, \tag{4}$$

and $\lambda(\rho,{\rho_\mathcal{T}}^*)$ is a non-estimable term from unlabeled target samples, and ${\rho_\mathcal{T}}^*=\operatorname{argmin}_\rho\mathbf{R}_\mathcal{T}(G_\rho)$.

This bound reflects a prevalent DA philosophy [11, 12, 13]. Indeed, assuming that $\lambda(\rho,{\rho_\mathcal{T}}^*)$ is small, a nice situation for DA arises when the divergence $\operatorname{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X})$ and the source risk $\mathbf{R}_\mathcal{S}(G_\rho)$ are small. Along with the above theorem, Germain et al. [9] provide the following PAC-Bayesian generalization bound (based on the PAC-Bayes analysis of non-adaptative learning of Catoni [17]).

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

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

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

This justifies the DA algorithm PBDA [9], which aims at minimizing the previous bound given a source-target sample (S,T). However, $\lambda(\rho,{\rho_\mathcal{T}}^*)$ does not appear in the optimization process since it cannot be estimated from unlabeled target data. In [9], they argued that $\lambda(\rho,{\rho_\mathcal{T}}^*)$ is negligible when DA is achievable (which is a strong assumption as it relies on ρ). Thus, given $A>0$ and $C>0$, PBDA minimizes the trade-off $C\,\widehat{\mathbf{R}}_S(G_\rho)+A\,\widehat{\operatorname{dis}_\rho}(S,T)+\operatorname{KL}(\rho\|\pi)$, specialized to linear classifiers.

3 A New PAC-Bayesian Domain Adaptation Bound and Algorithm

We now derive a simpler and more precise analysis. Inspired by the idea of [18, 19], we separate the risk $\mathbf{R}_{\mathcal{T}}(G_\rho)$ into $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and $\mathbf{e}_{\mathcal{T}}(\rho)$ (see Eq. (3)). In the present DA scenario, we are able to estimate $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ using T since it does not rely on label. However, $\mathbf{e}_{\mathcal{T}}(\rho)$ cannot be estimated from T. Theorem 3 below presents our DA bound and links $\mathbf{e}_{\mathcal{T}}(\rho)$ with $\mathbf{e}_{\mathcal{S}}(\rho)$ by weighting the latter by a divergence measure $\beta_q(\mathcal{T}\|\mathcal{S})$ between the domains, parameterized by a real value $q>0$:

$$\beta_q(\mathcal{T}\|\mathcal{S}) \;=\; \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right)^{q}\right]^{\frac{1}{q}}. \tag{5}$$

We denote $\beta_\infty(\mathcal{T}\|\mathcal{S}) = \sup_{(\mathbf{x},y)\in\mathbf{X}\times Y} \frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}$ the limit case $q\to\infty$. With some q values, we can recover known divergences (*e.g.,* the χ^2-distance, with $q=2$). Moreover, we can relate $\beta_q(\mathcal{T}\|\mathcal{S})$ to the Rényi divergence[2], which has already led to generalization bounds in the specific context of importance weighting [7].

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

Theorem 3. *Let $\mathcal{H}$ be a set of voters, let $\mathcal{S}$ and $\mathcal{T}$ resp. be the source and the target domains on $\mathbf{X}\times Y$. Let $q>1$ be a constant. We have:*

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

where $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$, $\mathbf{e}_{\mathcal{S}}(\rho)$ and $\beta_q(\mathcal{T}\|\mathcal{S})$ are respectively defined by Eq. (2) *and* (5)*.*

Proof. Starting from Eq. (3), and thanks to Hölder inequality, with p such that $\frac{1}{p} = 1-\frac{1}{q}$, we have

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

We remove the exponent from $(\mathbf{I}\left[h(\mathbf{x})\neq y\right]\mathbf{I}\left[h'(\mathbf{x})\neq y\right])^{p}$ since its value is either 1 or 0. □

It is instructive to compare Th. 3 with the previous Th. 1. In the new bound, the only non-estimable term is $\beta_q(\mathcal{T}\|\mathcal{S})$; Contrary to $\lambda(\rho,\rho_{\mathcal{T}}{}^*)$ of Th. 1, it does not depend on the posterior ρ learned, *i.e.*, for all ρ on $\mathcal{H}$, $\beta_q(\mathcal{T}\|\mathcal{S})$ is a constant measuring the relation between domains. Moreover, $\beta_q(\mathcal{T}\|\mathcal{S})$ is not an additive term but a multiplicative one (as opposed to $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X})+\lambda(\rho,\rho_{\mathcal{T}}{}^*)$). This is a contribution of our new analysis, since $\beta_q(\mathcal{T}\|\mathcal{S})$ can be considered as a hyperparameter to tune the trade-off between $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and $\mathbf{e}_{\mathcal{S}}(\rho)$. Thus, we do not need to make assumptions on its value.

PAC-Bayesian Generalization Bounds. (we skip the proofs of this paragraph, the reader can find them in our research report [20]) For justifying the empirical minimization of our new bound, we first provide PAC-Bayesian theorems for $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and $\mathbf{e}_{\mathcal{S}}(\rho)$.

Theorem 4. *For any $\mathcal{S}$ and $\mathcal{T}$, any $\mathcal{H}$, any π on $\mathcal{H}$, any $\delta\in(0,1]$, any $a>0$ and $c>0$, we have:*

$$\begin{aligned}&\mathop{\mathbf{Pr}}_{T\sim(\mathcal{T}_\mathbf{X})^{m_t}}\left(\forall\rho \text{ on } \mathcal{H},\quad \mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho) \;\le\; c'\,\widehat{\mathbf{d}}_T(\rho) + \frac{c'}{c}\,\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m_t}\right) \;\ge\; 1-\delta\,,\\ \textit{and}\quad &\mathop{\mathbf{Pr}}_{S\sim(\mathcal{S})^{m_s}}\left(\forall\rho \text{ on } \mathcal{H},\quad \mathbf{e}_{\mathcal{S}}(\rho) \;\le\; a'\,\widehat{\mathbf{e}}_S(\rho) + \frac{a'}{a}\,\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m_s}\right) \;\ge\; 1-\delta\,.\end{aligned}$$

where $\widehat{\mathbf{d}}_T(\rho)$, resp. $\widehat{\mathbf{e}}_S(\rho)$, is the empirical estimation of $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$, resp. $\mathbf{e}_{\mathcal{S}}(\rho)$, $c'=\frac{c}{1-e^{-c}}$, and $a'=\frac{a}{1-e^{-a}}$.

For algorithmic reasons, we deal with Th. 3 when $q\to\infty$. Thanks to Th. 4, minimizing Th. 3 bound amounts to optimize Th. 5 bound below, defined *w.r.t.* the empirical estimates of $\mathbf{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and $\mathbf{e}_{\mathcal{S}}(\rho)$.

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

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

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

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

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

where $b = \beta_\infty(\mathcal{T}\|\mathcal{S})$, and $b' = b\,\frac{a}{1-e^{a}}$, and $c' = \frac{c}{1-e^{-c}}$.

The Algorithm. (more details on the derivation of the algorithm are given in our research report [20]) From an optimization perspective, the problem suggested by Th. 5 is much more convenient to minimize than the one of Th. 2. The former is *smoother* than the latter that contains an absolute value required by $\widehat{\mathrm{dis}_\rho}(S,T)$. Germain et al. [9] choose also to ignore the non-constant term $\lambda(\rho, \rho_{\mathcal{T}}{}^*)$. In our case, such compromise is not mandatory to apply the theoretical result to real DA problems. Recalling that $\beta_q(\mathcal{T}\|\mathcal{S})$ is a constant that can tuned, and according to Th. 5, given hyperparameters $C{>}0$ and $B{>}0$—where B models the divergence between $\mathcal{T}$ and $\mathcal{S}$—we propose to minimize the trade-off $C\,\widehat{\mathbf{d}}_T(\rho) + B\,\widehat{\mathbf{e}}_S(\rho) + \mathrm{KL}(\rho\|\pi)$. We follow the setting of PBDA [9] for specializing this equation to linear classifiers. Therefore, we consider $\mathcal{H}$ as a set of linear classifiers in a d-dimensional space, and we use Gaussian posterior $\rho_{\mathbf{w}}$ and prior $\pi_{\mathbf{0}}$ with identity covariance matrix (respectively centered on vectors $\mathbf{w}$ and $\mathbf{0}$). Then, our new algorithm, called DALC (Domain Adaptation of Linear Classifiers), consists in minimizing

$$G(\mathbf{w}) = C \times \sum_{i=1}^{m_t} \Phi\left(\frac{\mathbf{w}\cdot\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}\right) \Phi\left(-\frac{\mathbf{w}\cdot\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|}\right) + B \times \sum_{i=1}^{m_s} \left[\Phi\left(y_i\,\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right)\right]^2 + \frac{1}{2}\|\mathbf{w}\|^2\,, \qquad (6)$$

where $\Phi(x) = \frac{1}{2}\Big[1 - \mathbf{Erf}\big(\frac{x}{\sqrt{2}}\big)\Big]$. Similarly to PBDA [9], we can apply the kernel trick. Even though the objective function is highly non-convex, we achieved good empirical results by minimizing the "kernelized" version of Eq. (6) by gradient descent, with a uniform weight vector as a starting point.

Experiments. We evaluate DALC[3] on the *Amazon.com Reviews* benchmark [21] according to the setting used in [4, 9] (See [9] for a complete description). This dataset contains reviews of 4 types of products (books, DVDs, electronics, and kitchen appliances), labeled into two classes: products rated ≤ 3 and products rated ≥ 4. The data are described with about $40,000$ attributes. A domain is a type of product, and a task consists in adapting from a kind to (→) another one. We compare DALC with the non-adaptative SVM (trained only on S), the adaptative DASVM [3], the DA co-training CODA [4], and the PAC-Bayesian DA algorithm PBDA [9]. Each parameter is selected with cross-validation (^{CV}) on the S for SVM, and thanks to a reverse validation procedure [3, 22](^{RCV}) for CODA, DASVM, PBDA, and DALC. The algorithms use a linear kernel, with $|S| = |T| = 2,000$. Tab. 1 reports the error rates of the methods evaluated on the same separate target test sets proposed in [4]. Above all, the DA approaches show the best result, implying that tackling this problem with a DA method is reasonable. Except for the two DA tasks between "electronics" and "DVDs", DALC is either the best one (6 times), or the second one (4 times). Moreover, DALC clearly increases the performances over the previous PAC-Bayesian PBDA, which confirms that our new bound improves the analysis done by Germain et al. [9].

[3] We released the source code at `http://graal.ift.ulaval.ca/dalc/`.

References

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

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

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

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

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

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

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

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

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

[10] X. Li and J. Bilmes. A bayesian divergence prior for classiffier adaptation. In *AISTATS*, pages 275–282, 2007.

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

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

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

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

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

[16] J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *NIPS*, pages 439–446, 2002.

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

[18] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *NIPS*, pages 769–776, 2006.

[19] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and J.-F. Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *JMLR*, 16:787–860, 2015. URL `http://jmlr.org/papers/v16/germain15a.html`.

[20] P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A new PAC-Bayesian perspective on domain adaptation. *arXiv:1506.04573 (research report)*, 2015. URL `http://arxiv.org/abs/1506.04573`.

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

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