A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers

Pascal Germain, Amaury Habrard, François Laviolette, Emilie Morvant

A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers

Pascal Germain PASCAL.GERMAIN@IFT.ULAVAL.CA
Département d'informatique et de génie logiciel, Université Laval, Québec, Canada

Amaury Habrard AMAURY.HABRARD@UNIV-ST-ETIENNE.FR
Laboratoire Hubert Curien UMR CNRS 5516, Université Jean Monnet, 42000 St-Etienne, France

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

Emilie Morvant EMILIE.MORVANT@LIF.UNIV-MRS.FR
Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, 13013, Marseille, France

Abstract

We provide a first PAC-Bayesian analysis for domain adaptation (DA) which arises when the learning and test distributions differ. It relies on a novel distribution pseudodistance based on a disagreement averaging. Using this measure, we derive a PAC-Bayesian DA bound for the stochastic Gibbs classifier. This bound has the advantage of being directly optimizable for any hypothesis space. We specialize it to linear classifiers, and design a learning algorithm which shows interesting results on a synthetic problem and on a popular sentiment annotation task. This opens the door to tackling DA tasks by making use of all the PAC-Bayesian tools.

1. Introduction

In machine learning, many classifier learning approaches suppose that the learning and test data are drawn from the same probability distribution. However, this strong hypothesis may be irrelevant for a lot of real tasks. For instance, a spam filtering system suitable for one user can be poorly adapted to another who receives significantly different emails. In other words, the learning data associated with one user could be unrepresentative of the test data coming from another one. This enhances the need to design methods for adapting a classifier from learning (source) data to test (target) data. One solution to tackle this issue is to consider the *Domain Adaptation* (DA) framework[1], which arises when the distribution generating the target data (the *target domain*) differs from the one generating the source data (the *source domain*). In such a situation, it is well known that DA is a hard and challenging task even under strong assumptions[2] (Ben-David & Urner, 2012; Ben-David et al., 2010b). A major issue in DA is to define a measure allowing one to quantify how much the domains are related. Concretely, when they are close under this measure, the generalization guarantees over the target domain may be "easier" to provide. For example, in the context of binary classification with the 0-1 loss function, Ben-David et al. (2010a); Ben-David et al. (2006) have considered the $\mathcal{H}\Delta\mathcal{H}$-divergence between the marginal distributions. This quantity is based on the maximal disagreement between two classifiers, allowing them to deduce a DA generalization bound based on the *VC-dim* theory. The discrepancy distance (Mansour et al., 2009a) generalizes this divergence to real-valued functions and more general losses, and is used to obtain a generalization bound based on the Rademacher complexity. In this context, Cortes & Mohri (2011) have specialized the minimization of the discrepancy to regression with kernels. In these situations, DA can be viewed as a multiple trade-off between the complexity of the hypothesis class $\mathcal{H}$, the adaptation ability of $\mathcal{H}$

[1] Surveys: Jiang (2008); Quionero-Candela et al. (2009).

[2] As the *covariate-shift*, where source and target domains diverge only in their marginals (*i.e.*, they have the same labeling function).

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

according to the divergence between the marginals, and the empirical source risk. Moreover, other measures have been exploited under different assumptions, such as the Rényi divergence suitable for importance weighting (Mansour et al., 2009b), or the measure proposed by C. Zhang (2012) which takes into account the source and target true labeling, or the Bayesian "divergence prior" (Li & Bilmes, 2007) which favors classifiers closer to the best source model.

The novelty of our contribution is to explore the PAC-Bayesian framework to tackle DA in a binary classification situation without target labels (sometimes called *unsupervised domain adaptation*). Given a prior distribution over a family of classifiers $\mathcal{H}$, PAC-Bayesian theory (introduced by McAllester (1999)) focuses on algorithms that output a posterior distribution ρ over $\mathcal{H}$ (*i.e.*, a ρ-average over $\mathcal{H}$) rather than just a single classifier $h \in \mathcal{H}$. Following this principle, we propose a pseudometric which evaluates the domain divergence according to the ρ-average disagreement of the classifiers over the domains. This disagreement measure shows many advantages. First, it is ideal for the PAC-Bayesian setting, since it is expressed as a ρ-average over $\mathcal{H}$. Second, we prove that it is always lower than the popular $\mathcal{H}\Delta\mathcal{H}$-divergence. Last but not least, our measure can be easily estimated from samples. From this pseudometric, we derive a first PAC-Bayesian DA generalization bound expressed as a ρ-averaging.

The practical optimization of this bound relies on multiple trade-offs between three quantities. The first two quantities being, as usual in the PAC-Bayesian approach, the complexity of the majority vote measured by a Kullback-Leibler divergence and the empirical risk measured by the ρ-average errors on the source sample. The third quantity corresponds to our domain divergence and assesses the capacity of the posterior distribution to distinguish some structural difference between the source and target samples. An interesting property of our analysis is that these quantities can be jointly optimized. Finally, we design an algorithm for optimizing our bound, tailored to linear classifiers.

The paper is structured as follow: Section 2 deals with the notation and the two seminal works on DA. The PAC-Bayesian framework is then recalled in Section 3. Our main contribution, which consists in a DA-bound suitable for PAC-Bayesian learning, is presented in Section 4. Then, we derive our new algorithm for PAC-Bayesian DA in Section 5. Before concluding in Section 7, we experiment our approach in Section 6.

2. Notation and DA Related Works

We consider DA for binary classification tasks where $X \subseteq \mathbb{R}^d$ is the input space of dimension d and $Y = \{-1,1\}$ is the label set. The *source domain* P_S and the *target domain* P_T are two different distributions over $X \times Y$, D_S and D_T being the respective marginal distributions over X. We tackle the challenging task where we have no target labels. A learning algorithm is then provided with a *labeled source sample* $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ drawn *i.i.d.* from P_S, and an *unlabeled target sample* $T = \{\mathbf{x}_j^t\}_{j=1}^{m'}$ drawn *i.i.d.* from D_T. Let $h : X \to Y$ be a hypothesis function. The *expected source error* of h over P_S is the probability that h errs,

$$R_{P_S}(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x}^s, y^s) \sim P_S} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}^s), y^s\big),$$

where $\mathcal{L}_{0\text{-}1}(a,b) \stackrel{\text{def}}{=} \mathbf{I}[a \neq b]$ is the 0-1 loss function which returns 1 if $a \neq b$ and 0 otherwise. The *expected target error* $R_{P_T}(\cdot)$ over P_T is defined in a similar way. $R_S(\cdot)$ is the *empirical source error*. The main objective in DA is to learn – without target labels – a classifier leading to the lowest expected target error $R_{P_T}(h)$.

We also introduce the *expected source disagreement* of h' and h, which measures the probability that two classifiers h and h' do not agree,

$$R_{D_S}(h, h') \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{x}^s \sim D_S} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}^s), h'(\mathbf{x}^s)\big).$$

The *expected target disagreement* $R_{D_T}(\cdot,\cdot)$ over D_T is similarly defined. $R_S(\cdot,\cdot)$ and $R_T(\cdot,\cdot)$ are the *empirical source and target disagreements* on S and T. Depending on the context, S denotes either the source labeled sample $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ or its unlabeled part $\{\mathbf{x}_i^s\}_{i=1}^m$.

2.1. Necessity of a Domain Divergence

The DA objective is to find a low-error target hypothesis, even if no target labels are available. Even under strong assumptions, this task can be impossible to solve (Ben-David & Urner, 2012; Ben-David et al., 2010b). However, for deriving generalization ability in a DA situation (with the help of a DA-bound), it is critical to make use of a divergence between the source and the target domains: the more similar the domains, the easier the adaptation appears. Some previous works (C. Zhang, 2012; Ben-David et al., 2010a; Mansour et al., 2009a;b; Ben-David et al., 2006; Li & Bilmes, 2007) have proposed different quantities to estimate how a domain is close to another one. Concretely, two domains P_S and P_T differ if their marginals D_S and D_T are different, or if the source labeling function differs from the target one, or if both happen. This suggests to take into account two divergences: one between D_S and D_T and one between the labeling. If we have some target labels, we can combine the two distances as C. Zhang (2012). Otherwise, we preferably consider two separate measures, since it

is impossible to estimate the best target hypothesis in such a situation. Usually, we suppose that the source labeling function is somehow related to the target one, then we look for a representation where the marginals D_S and D_T appear closer without losing performances on the source domain.

2.2. DA-Bounds for Binary Classification

We now review the first two seminal works which propose DA-bounds based on the marginal divergence.

First, under the assumption that there exists a hypothesis in $\mathcal{H}$ that performs well on both the source and the target domain, Ben-David et al. (2010a); Ben-David et al. (2006) have provided the following DA-bound.

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

$$\forall h \in \mathcal{H},\ R_{P_T}(h) \leq R_{P_S}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + R_{P_S}(h^*) + R_{P_T}(h^*)\,, \quad (1)$$

with $\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \stackrel{\text{def}}{=} \sup_{(h,h') \in \mathcal{H}^2} \left| R_{D_T}(h,h') - R_{D_S}(h,h') \right|$ *is the $\mathcal{H}\Delta\mathcal{H}$-distance between the marginals D_S and D_T,* $h^* \stackrel{\text{def}}{=} \operatorname{argmin}_{h \in \mathcal{H}} \big(R_{P_S}(h) + R_{P_T}(h)\big)$ *is the best hypothesis overall.*

This bound depends on four terms. $R_{P_S}(h)$ is the classical source domain expected error. $\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$ depends on $\mathcal{H}$ and corresponds to the maximum disagreement between two hypothesis of $\mathcal{H}$. In other words, it quantifies how hypothesis from $\mathcal{H}$ can "detect" differences between these marginals: the lower this measure is for a given $\mathcal{H}$, the better are the generalization guarantees. The last terms $R_{P_S}(h^*)$ and $R_{P_T}(h^*)$ are related to the best hypothesis h^* over the domains and act as a quality measure of $\mathcal{H}$ in terms of labeling information. If h^* performs poorly, then it is hard to find a low-error hypothesis on the target domain. Hence, as pointed out by the authors, Equation (1), together with the usual VC-bound theory, expresses a multiple trade-off between the accuracy of some particular hypothesis h, the complexity of $\mathcal{H}$, and the "incapacity" of hypothesis of $\mathcal{H}$ to detect difference between the source and the target domain.

Second, Mansour et al. (2009a) have extended the $\mathcal{H}\Delta\mathcal{H}$-distance to the discrepancy divergence for regression and any symmetric loss $\mathcal{L}$ fulfilling the triangle inequality. Given $\mathcal{L} : [-1, 1]^2 \mapsto \mathbb{R}^+$ such a loss, the discrepancy $\operatorname{disc}_{\mathcal{L}}$ between D_S and D_T is: $\operatorname{disc}_{\mathcal{L}}(D_S, D_T) \stackrel{\text{def}}{=} \sup_{(h,h') \in \mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}^t \sim D_T} \mathcal{L}(h(\mathbf{x}^t), h'(\mathbf{x}^t)) - \mathop{\mathbf{E}}_{\mathbf{x}^s \sim D_S} \mathcal{L}(h(\mathbf{x}^s), h'(\mathbf{x}^s)) \right|$. Note that with the 0-1 loss in binary classification, we have: $\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) = \operatorname{disc}_{\mathcal{L}_{0\text{-}1}}(D_S, D_T)$. Even if these two divergences coincide, the DA-bound of Mansour et al. (2009a) differs from Theorem 1 and is,

$$\forall h \in \mathcal{H},\ R_{P_T}(h) - R_{P_T}(h^*_T) \leq R_{D_S}(h^*_S, h) + R_{D_S}(h^*_S, h^*_T) + \operatorname{disc}_{\mathcal{L}_{0\text{-}1}}(D_S, D_T)\,, \quad (2)$$

where $h^*_T \stackrel{\text{def}}{=} \operatorname{argmin}_{h \in \mathcal{H}} R_{P_T}(h)$ and $h^*_S \stackrel{\text{def}}{=} \operatorname{argmin}_{h \in \mathcal{H}} R_{P_S}(h)$ are respectively the ideal hypothesis on the target and source domains. In this context, Equation (2) can be tighter[3] since it bounds the difference between the target error of a classifier and the one of the optimal h^*_T. This bound expresses a trade-off between the disagreement (between h and the best source hypothesis h^*_S), the complexity of $\mathcal{H}$ (with the Rademacher complexity), and – again – the "incapacity" of hypothesis to detect differences between the domains.

To conclude, the DA-bounds (1) and (2) suggest that if the divergence between the domains is low, a low-error classifier over the source domain might perform well on the target one. These divergences compute the worst case of the disagreement between a pair of hypothesis. We propose an average case approach by making use of the essence of the PAC-Bayesian theory, which is known to offer tight generalization bounds (McAllester, 1999; Ambroladze et al., 2006).

3. PAC-Bayesian Theory

Let us now review the classical supervised binary classification framework called the PAC-Bayesian theory, first introduced by McAllester (1999). Traditionally, the PAC-Bayesian theory considers weighted majority votes over a set $\mathcal{H}$ of binary hypothesis. Given a prior distribution π over $\mathcal{H}$ and a training set S, the learner aims at finding the posterior distribution ρ over $\mathcal{H}$ leading to a ρ-weighted majority vote B_ρ (also called the Bayes classifier) with good generalization guarantees and defined by,

$$B_\rho(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname{sign}\left[\mathop{\mathbf{E}}_{h \sim \rho} h(\mathbf{x}) \right].$$

Minimizing the risk of B_ρ is known to be NP-hard. In the PAC-Bayesian approach, it is replaced by the risk of the stochastic Gibbs classifier G_ρ associated with ρ. In order to predict the label of an example $\mathbf{x}$, the Gibbs classifier first draws a hypothesis h from $\mathcal{H}$ according to ρ, then returns $h(\mathbf{x})$ as label. Note that the error of the Gibbs classifier on a domain P_S corresponds to the expectation of the errors over ρ,

$$R_{P_S}(G_\rho) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h \sim \rho} R_{P_S}(h). \quad (3)$$

[3] Equation (1) can lead to an error term 3 times higher than Equation (2) in some cases (Mansour et al., 2009a).

In this setting, if B_ρ misclassifies $\mathbf{x}$, then at least half of the classifiers (under ρ) errs on $\mathbf{x}$. Hence we have: $R_{P_S}(B_\rho) \leq 2R_{P_S}(G_\rho)$. Another result on $R_{P_S}(B_\rho)$ is the C-bound (Lacasse et al., 2006) defined by,

$$R_{P_S}(B_\rho) \;\leq\; 1 - \frac{\left(1 - 2R_{P_S}(G_\rho)\right)^2}{1 - 2R_{D_S}(G_\rho, G_\rho)}\,, \qquad (4)$$

where $R_{D_S}(G_\rho, G_\rho)$ corresponds to the disagreement of the classifiers over ρ and is defined by,

$$R_{D_S}(G_\rho, G_\rho) \;\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_S}(h, h')\,. \qquad (5)$$

Equation (4) suggests that for a fixed numerator, the best majority vote is the one with the lowest denominator, *i.e.*, with the greatest disagreement between its voters (see Laviolette et al. (2011) for further analysis).

The PAC-Bayesian theory allows one to bound the expected error $R_{P_S}(G_\rho)$ in terms of two major quantities: the empirical error $R_S(G_\rho)=\mathbf{E}_{h\sim\rho} R_S(h)$ estimated on a sample S *i.i.d.* from P_S and the Kullback-Leibler divergence $\mathrm{KL}(\rho\|\pi) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)}$. In this paper we use the following PAC-Bayesian bound of Catoni (2007) in a simplified form suggested by Germain et al. (2009b).

Theorem 2 (Catoni (2007))**.** *For any domain P_S over $X\times Y$, for any set of hypothesis $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta \in (0,1]$, and any real number $c > 0$, with a probability at least $1-\delta$ over the choice of $S\sim(P_S)^m$, for every ρ on $\mathcal{H}$, we have,*

$$R_{P_S}(G_\rho) \;\leq\; \frac{c}{1-e^{-c}}\left[R_S(G_\rho) + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{1}{\delta}}{m\times c}\right].$$

This bound has two interesting characteristics. First, its minimization is closely related to the minimization problem associated with the SVM when ρ is an isotropic Gaussian over the space of linear classifiers (Germain et al., 2009a). Second, the value c allows to control the trade-off between the empirical risk $R_S(G_\rho)$ and the complexity term $\frac{\mathrm{KL}(\rho\|\pi)}{m}$. Moreover, putting $c = \frac{1}{\sqrt{m}}$, this bound becomes consistent: it converges to $1\times[R_S(G_\rho)+0]$ as m grows.

While the DA-bounds presented in Section 2 focus on a single classifier, we now define a ρ-average disagreement measure to compare the marginals. This leads us to derive our DA-bound suitable for PAC-Bayes.

4. A DA-Bound for the Gibbs Classifier

The originality of our contribution is to theoretically design a DA framework for PAC-Bayesian approach. In Section 4.1, we propose a domain comparison pseudometric suitable in this context. We then derive a PAC-Bayesian DA-bound in Section 4.2.

4.1. A Domain Divergence for PAC-Bayes

As seen in Section 2.1, the derivation of generalization ability in DA critically needs a divergence measure between the source and target marginals.

Designing the Divergence. We define a *domain disagreement pseudometric*[4] to measure the structural difference between domain marginals in terms of posterior distribution ρ over $\mathcal{H}$. Since we are interested in learning a ρ-weighted majority vote B_ρ leading to good generalization guarantees, we propose to follow the idea behind Equation (4): Given P_S, P_T, and ρ, if $R_{P_S}(G_\rho)$ and $R_{P_T}(G_\rho)$ are similar, then $R_{P_S}(B_\rho)$ and $R_{P_T}(B_\rho)$ are similar when $\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_S}(h,h')$ and $\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_T}(h,h')$ are also similar. Thus, the domains P_S and P_T are close according to ρ if the divergence between $\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_S}(h,h')$ and $\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_T}(h,h')$ tends to be low. Our pseudometric is defined as follows.

Definition 1. *Let $\mathcal{H}$ be a hypothesis class. For any marginal distributions D_S and D_T over X, any distribution ρ on $\mathcal{H}$, the domain disagreement* $\mathrm{dis}_\rho(D_S, D_T)$ *between D_S and D_T is defined by,*

$$\mathrm{dis}_\rho(D_S, D_T) \;\stackrel{\text{def}}{=}\; \left|\mathop{\mathbf{E}}_{h,h'\sim\rho^2}[R_{D_T}(h,h') - R_{D_S}(h,h')]\right|.$$

Note that $\mathrm{dis}_\rho(\cdot,\cdot)$ is symmetric and fulfills the triangle inequality. The following theorem shows that $\mathrm{dis}_\rho(D_S, D_T)$ can be bounded in terms of the classical PAC-Bayesian quantities: the empirical disagreement $\mathrm{dis}_\rho(S,T)$ estimated on the source and target samples, and the KL-divergence between the prior and posterior distribution on $\mathcal{H}$. For the sake of simplicity, we suppose that $m=m'$, *i.e.*, the size of S and T are equal[5].

Theorem 3. *For any distributions D_S and D_T over X, any set of hypothesis $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, and any real number $\alpha>0$, with a probability at least $1-\delta$ over the choice of $S\times T\sim(D_S\times D_T)^m$, for every ρ on $\mathcal{H}$, we have,*

$$\mathrm{dis}_\rho(D_S, D_T) \leq \frac{2\alpha\left[\,\mathrm{dis}_\rho(S,T) + \frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{2}{\delta}}{m\times\alpha} + 1\right] - 1}{1 - e^{-2\alpha}},$$

where $\mathrm{dis}_\rho(S,T)$ *is the empirical domain disagreement.*

Similarly to the empirical risk bound of Catoni (2007) shown by Theorem 2, the above domain disagreement bound is consistent if one puts $\alpha = \frac{1}{2\sqrt{m}}$. Indeed, it converges to $1\times[\mathrm{dis}_\rho(S,T)+0+1]-1$ as m grows.

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

[5] The Supplementary Material gives other DA PAC-Bayesian bounds, notably for the case where $m\neq m'$.

Proof of Theorem 3. (details given in Supp. Material) Firstly, we bound $d^{(1)} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h,h'\sim\rho^2}[R_{D_S}(h,h') - R_{D_T}(h,h')]$. Consider an "abstract" classifier $\hat{h} \stackrel{\text{def}}{=} (h,h') \in \mathcal{H}^2$ chosen from a distribution $\hat{\rho}$, with $\hat{\rho}(\hat{h}) = \rho(h)\rho(h')$. Notice that with $\hat{\pi}(\hat{h}) = \pi(h)\pi(h')$, we obtain that $\mathrm{KL}(\hat{\rho}\|\hat{\pi}) = 2\mathrm{KL}(\rho\|\pi)$. Let us define the "abstract" loss of $\hat{h}$ on a pair of examples $(\mathbf{x}^s,\mathbf{x}^t) \sim D_S \times D_T$ by,

$$\mathcal{L}_{d^{(1)}}(\hat{h},\mathbf{x}^s,\mathbf{x}^t) \stackrel{\text{def}}{=} \frac{1+\mathcal{L}_{0\text{-}1}(h(\mathbf{x}^s),h'(\mathbf{x}^s)) - \mathcal{L}_{0\text{-}1}(h(\mathbf{x}^t),h'(\mathbf{x}^t))}{2}.$$

The error of the Gibbs classifier associated with this loss is $R^{(1)}_{D_S\times D_T}(G_{\hat{\rho}}) = \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} \mathop{\mathbf{E}}_{\mathbf{x}^s\sim D_S} \mathop{\mathbf{E}}_{\mathbf{x}^t\sim D_T} \mathcal{L}_{d^{(1)}}(\hat{h},\mathbf{x}^s,\mathbf{x}^t)$. As $\mathcal{L}_{d^{(1)}}$ lies in $[0,1]$, following the principle of the proof of Theorem 2 (with $c=2\alpha$), one can bound the true $R^{(1)}_{D_S\times D_T}(G_{\hat{\rho}})$ (see Supp. Material). Thereafter, we obtain a bound on $d^{(1)}$ from its empirical counterpart (denoted by $d^{(1)}_{S\times T}$), because $d^{(1)} = 2R^{(1)}_{D_S\times D_T}(G_{\hat{\rho}}) - 1$. Hence, we obtain with probability at least $1-\frac{\delta}{2}$ over the choice of $S\times T \sim (D_S\times D_T)^m$,

$$\frac{d^{(1)}+1}{2} \le \frac{2\alpha}{1-e^{-2\alpha}}\left[\frac{d^{(1)}_{S\times T}+1}{2} + \frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{2}{\delta}}{m\times 2\alpha}\right].$$

Then, we bound $d^{(2)} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h,h'\sim\rho^2}[R_{D_T}(h,h') - R_{D_S}(h,h')]$ from $d^{(2)}_{S\times T}$ using the same method. Note that $|d^{(1)}| = |d^{(2)}| = \mathrm{dis}_\rho(D_S,D_T)$. Thus, the maximum of the bound on $d^{(1)}$ and the bound on $d^{(2)}$ gives a bound on $\mathrm{dis}_\rho(D_S,D_T)$. Using the union bound, we obtain with probability $1-\delta$ over the choice of $S\times T \sim (D_S\times D_T)^m$,

$$\frac{|d^{(1)}|+1}{2} \le \frac{\alpha}{1-e^{-2\alpha}}\left[|d^{(1)}_{S\times T}| + 1 + \frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{2}{\delta}}{m\times\alpha}\right]. \quad \square$$

Before deriving a DA-bound for ρ-average of classifiers, we compare our dis_ρ with the $\mathcal{H}\Delta\mathcal{H}$-divergence.

Comparison of $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}$ and dis_ρ. While estimating the $\mathcal{H}\Delta\mathcal{H}$-divergence of Theorem 1 is NP-hard (Ben-David et al., 2010a; Ben-David et al., 2006), our empirical disagreement measure is easier to assess, since we simply have to compute the ρ-average of the classifiers disagreement instead of finding the pair of classifiers that maximizes the disagreement. Indeed, dis_ρ depends on the majority vote, which suggests that we can directly minimize it via the empirical $\mathrm{dis}_\rho(S,T)$ and the KL-divergence. This can be done without instance reweighting, space representation changing or family of classifiers modification. On the contrary, $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}$ is a supremum over all $h\in\mathcal{H}$ and hence, does not depend on the h on which the risk is considered. Moreover, dis_ρ (the ρ-average) is lower than the $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}$ (the worst case). Indeed, for every $\mathcal{H}$ and ρ over $\mathcal{H}$, we have,

$$\begin{aligned}\tfrac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T) &= \sup_{(h,h')\in\mathcal{H}^2} |R_{D_T}(h,h') - R_{D_S}(h,h')| \\ &\ge \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} |R_{D_T}(h,h') - R_{D_S}(h,h')| \ge \mathrm{dis}_\rho(D_S,D_T).\end{aligned}$$

4.2. The PAC-Bayesian DA-Bound

We now derive our main result in the following theorem. Note that for the sake of readability, we prefer to use the notations $R_P(G_\rho)$ and $R_D(G_\rho,\cdot)$, we recall that they correspond to the respective ρ-averages $\mathop{\mathbf{E}}_{h\sim\rho} R_P(h)$ and $\mathop{\mathbf{E}}_{h\sim\rho} R_D(h,\cdot)$ (see Equations (3) and (5)).

Theorem 4. *Let $\mathcal{H}$ be a hypothesis class. We have,*

$$\begin{aligned}\forall\rho \text{ on } \mathcal{H},\ R_{P_T}(G_\rho) - R_{P_T}(G_{\rho^*_T}) \ \le\ & R_{P_S}(G_\rho) \\ &+ \mathrm{dis}_\rho(D_S,D_T) + R_{D_T}(G_\rho,G_{\rho^*_T}) + R_{D_S}(G_\rho,G_{\rho^*_T}),\end{aligned}$$

with $\rho^*_T = \mathrm{argmin}_\rho\, R_{P_T}(G_\rho)$ is the best target posterior, and $R_D(G_\rho,G_{\rho^*_T}) = E_{h\sim\rho}E_{h'\sim\rho^*_T} R_D(h,h')$.

Proof. Let $\mathcal{H}$ be a hypothesis set. Let ρ over $\mathcal{H}$. Let $\rho^*_T = \mathrm{argmin}_\rho\, R_{P_T}(G_\rho)$ be the distribution leading to the best Gibbs classifier on P_T. With the triangle inequality, and since for every h and any marginal D,

$$R_D(G_\rho,h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{I}[G_\rho(\mathbf{x})\neq h(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathop{\mathbf{E}}_{h'\sim\rho} I[h'(\mathbf{x})\neq h(\mathbf{x})],$$

we can write,

$$\begin{aligned}R_{P_T}(G_\rho) &\le \mathop{\mathbf{E}}_{h\sim\rho}\Big[R_{P_T}(G_{\rho^*_T}) + R_{D_T}(G_{\rho^*_T},G_\rho) + R_{D_T}(G_\rho,h)\Big] \\ &\le R_{P_T}(G_{\rho^*_T}) + R_{D_T}(G_{\rho^*_T},G_\rho) \\ &\quad + \mathop{\mathbf{E}}_{h\sim\rho}\left[R_{D_T}(G_\rho,h) - R_{D_S}(G_\rho,h) + R_{D_S}(G_\rho,h)\right] \\ &\le R_{P_T}(G_{\rho^*_T}) + R_{D_T}(G_\rho,G_{\rho^*_T}) + \mathop{\mathbf{E}}_{h\sim\rho} R_{D_S}(G_\rho,h) \\ &\quad + \Big|\mathop{\mathbf{E}}_{h,h'\sim\rho^2}\left[R_{D_T}(h,h') - R_{D_S}(h,h')\right]\Big| \\ &\le R_{P_T}(G_{\rho^*_T}) + R_{D_T}(G_\rho,G_{\rho^*_T}) + R_{D_S}(G_\rho,G_{\rho^*_T}) \\ &\quad + \Big|\mathop{\mathbf{E}}_{h,h'\sim\rho^2}\left[R_{D_T}(h,h') - R_{D_S}(h,h')\right]\Big| + \mathop{\mathbf{E}}_{h\sim\rho} R_{P_S}(h) \\ &= R_{P_T}(G_{\rho^*_T}) + R_{P_S}(G_\rho) + \mathrm{dis}_\rho(D_S,D_T) \\ &\quad + R_{D_T}(G_\rho,G_{\rho^*_T}) + R_{D_S}(G_\rho,G_{\rho^*_T}). \quad \square\end{aligned}$$

Our bound is, in general, incomparable with Equations (1) and (2). However, similarly to the DA-bound of Equation (2) (Mansour et al., 2009a), we directly bound the difference between the ρ-average target errors and the optimal one. Our bound can be seen as a trade-off between different quantities. $R_{P_S}(G_\rho)$ and $\mathrm{dis}_\rho(D_S,D_T)$ are similar to the first two terms of the DA-bound of Ben-David et al. (2010a) (Equation (1)): $R_{P_S}(G_\rho)$ is the ρ-average risk over $\mathcal{H}$ on the source domain, and $\mathrm{dis}_\rho(D_T,DS)$ measures the ρ-average disagreement between the marginals but is specific to the current ρ. The other terms $R_{D_T}(G_\rho,G_{\rho^*_T})$ and $R_{D_S}(G_\rho,G_{\rho^*_T})$ measure how much the considered distribution ρ is close (in terms of disagreements) to the optimal target Gibbs classifier both on P_S and P_T. According to this theory, a good DA is possible if the optimal distribution ρ^*_T has a low-error on the target domain (which is an usual assumption). Moreover, the quantity $R_{D_T}(G_\rho,G_{\rho^*_T}) + R_{D_S}(G_\rho,G_{\rho^*_T})$, which can be

seen as a measure of adaptation capability in terms of labeling functions, has to be low: G_ρ has to agree with the optimal solution on both domains.

Finally, our Theorem 4 leads to a PAC-Bayesian bound based on both the empirical source error of the Gibbs classifier and the empirical domain disagreement pseudometric estimated on a source and target samples.

Theorem 5. *For any domains P_S and P_T (resp. with marginals D_S and D_T) over $X\times Y$, any set of hypothesis $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, any real numbers $\alpha>0$ and $c>0$, with a probability at least $1-\delta$ over the choice of $S\times T\sim(P_S\times D_T)^m$, we have,*

$$\forall\rho\sim\mathcal{H},\ R_{P_T}(G_\rho)-R_{P_T}(G_{\rho_T^*})\ \leq\ \lambda_\rho+\alpha'-1 + c'R_S(G_\rho)+\alpha'\mathrm{dis}_\rho(S,T)+\left(\frac{c'}{c}+\frac{2\,\alpha'}{\alpha}\right)\frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{3}{\delta}}{m},$$

where $\lambda_\rho \stackrel{\text{def}}{=} R_{D_T}(G_\rho,G_{\rho_T^*})+R_{D_S}(G_\rho,G_{\rho_T^*})$, $c'\stackrel{\text{def}}{=}\frac{c}{1-e^{-c}}$, *and* $\alpha'\stackrel{\text{def}}{=}\frac{2\alpha}{1-e^{-2\alpha}}$.

Proof. In Theorem 4, replace $R_S(G_\rho)$ and $\mathrm{dis}_\rho(S,T)$ by their upper bound, obtained from Theorem 2 and Theorem 3, with δ chosen respectively as $\frac{\delta}{3}$ and $\frac{2\delta}{3}$ (in the latter case, we use $\ln\frac{2}{2\delta/3}=\ln\frac{3}{\delta}<2\ln\frac{3}{\delta}$). □

Under the assumption that the domains are somehow related in terms of labeling agreement on P_S and P_T (for every distribution ρ over $\mathcal{H}$), *i.e.*, a low $\mathrm{dis}_\rho(D_S,D_T)$ implies a negligible λ_ρ, a natural solution for a PAC-Bayesian DA algorithm without target labels is to minimize the bound of Theorem 5 by disregarding[6] λ_ρ. Notice that a major advantage of our DA-bound is that we can jointly optimize the risk and the divergence with a theoretical justification.

5. PAC-Bayesian Domain Adaptation Learning of Linear Classifiers

Now, let $\mathcal{H}$ be a set of linear classifiers $h_{\mathbf{v}}(\mathbf{x})\stackrel{\text{def}}{=}\mathrm{sgn}\,(\mathbf{v}\cdot\mathbf{x})$ such that $\mathbf{v}\in\mathbb{R}^d$ is a weight vector. By restricting the prior and the posterior to be Gaussian distributions, Langford & Shawe-Taylor (2002); Ambroladze et al. (2006) have specialized the PAC-Bayesian theory in order to bound the expected risk of any linear classifier $h_{\mathbf{w}}\in\mathcal{H}$ identified by a weight vector $\mathbf{w}$. More precisely, given a prior $\pi_{\mathbf{0}}$ and a posterior $\rho_{\mathbf{w}}$ defined as spherical Gaussians with identity covariance matrix respectively centered on vectors $\mathbf{0}$ and $\mathbf{w}$, for any $h_{\mathbf{v}}\in\mathcal{H}$, we have,

$$\pi_{\mathbf{0}}(h_{\mathbf{v}})\stackrel{\text{def}}{=}\left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\frac{1}{2}\|\mathbf{v}\|^2},\ \text{and}\ \rho_{\mathbf{w}}(h_{\mathbf{v}})\stackrel{\text{def}}{=}\left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\frac{1}{2}\|\mathbf{v}-\mathbf{w}\|^2}.$$

[6] With few target labels we can imagine to estimate λ_ρ.

The expected risk of the Gibbs classifier $G_{\rho_{\mathbf{w}}}$ on a domain P_S is then given by,

$$R_{P_S}(G_{\rho_{\mathbf{w}}})=\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S}\ \mathop{\mathbf{E}}_{h_{\mathbf{v}}\sim\rho_{\mathbf{w}}}\mathbf{I}(h_{\mathbf{v}}\neq y)=\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S}\Phi\left(y\tfrac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right),$$

where $\Phi(a)\stackrel{\text{def}}{=}\frac{1}{2}\left[1-\mathbf{Erf}\left(\frac{a}{\sqrt{2}}\right)\right]$, and $\mathbf{Erf}$ is the Gauss error function. In this situation, the KL-divergence between $\rho_{\mathbf{w}}$ and $\pi_{\mathbf{0}}$ becomes simply $\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}})=\frac{1}{2}\|\mathbf{w}\|^2$.

5.1. Supervised PAC-Bayesian Learning

Based on the specialization of the PAC-Bayesian theory to linear classifiers, Germain et al. (2009a) suggested to minimize the bound on $R_{P_S}(G_{\rho_{\mathbf{w}}})$ of Theorem 2. Given a sample $S=\{(\mathbf{x}_i^s,y_i^s)\}_{i=1}^m$ and an hyperparameter $C>0$, the resulting learning algorithm performs a gradient descent in order to find an optimal weight vector $\mathbf{w}$ that minimizes,

$$CmR_S(G_{\rho_{\mathbf{w}}})+\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}})=C\sum_{i=1}^m\Phi\left(y_i\tfrac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right)+\frac{\|\mathbf{w}\|^2}{2}.$$

This algorithm, called PBGD3, realizes a trade-off between the empirical risk (expressed by the loss Φ) and the complexity of the learned linear classifier (expressed by the regularizer $\|\mathbf{w}\|^2$). A practical drawback of PBGD3 is that the objective function is non-convex and the gradient descent implementation needs many random restarts. In fact, we made extensive empirical experiments and saw that PBGD3 performs equivalently (and at a fraction of the running time) by replacing the loss function Φ by its convex relaxation $\Phi_{\mathrm{cvx}}(a)\stackrel{\text{def}}{=}\frac{1}{2}-\frac{a}{\sqrt{2\pi}}$ if $a\leq 0$, $\Phi(a)$ otherwise.

In the following, we will see that using this approach in a DA way is a relevant strategy. To do so, we specialize the bound of Theorem 5 to linear classifiers.

5.2. Minimizing the PAC-Bayesian DA-Bound

Under the assumption that the non-estimable quantities λ_ρ and $R_{P_T}(G_{\rho_T^*})$ of Theorem 5 are negligible, we propose to design a PAC-Bayesian algorithm[7] for DA inspired by PBGD3. Therefore, given a source sample $S=\{(\mathbf{x}_i^s,y_i^s)\}_{i=1}^m$ and a target sample $T=\{(\mathbf{x}_i^t)\}_{i=1}^m$ we focus on the minimization, according to $\rho_{\mathbf{w}}$, of

$$C\,m\,R_S(G_{\rho_{\mathbf{w}}})+A\,m\ \mathrm{dis}_{\rho_{\mathbf{w}}}(S,T)+\mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}})\,,\quad(6)$$

where $\mathrm{dis}_{\rho_{\mathbf{w}}}(S,T)=\left|\mathop{\mathbf{E}}_{h,h'\sim\rho_{\mathbf{w}}^2}R_S(h,h')-\mathop{\mathbf{E}}_{h,h'\sim\rho_{\mathbf{w}}^2}R_T(h,h')\right|$ is the empirical domain disagreement between S and T specialized to a distribution $\rho_{\mathbf{w}}$ over linear classifiers. The values $A>0$, $C>0$ are hyperparameters

[7] Code available at http://graal.ift.ulaval.ca/pbda

of the algorithm. Note that the constants α and c of Theorem 5 can be recovered from any A and C. Given $\Phi_{\text{dis}}(a) \stackrel{\text{def}}{=} 2\,\Phi(a)\,\Phi(-a)$, we have for any marginal D,

$$\begin{aligned}
\mathop{\mathbf{E}}_{h,h'\sim\rho_{\mathbf{w}}^2} R_D(h,h') &= \mathop{\mathbf{E}}_{x\sim D}\ \mathop{\mathbf{E}}_{h,h'\sim\rho_{\mathbf{w}}^2} \mathbf{I}[h(\mathbf{x}) \neq h'(\mathbf{x})] \\
&= 2 \mathop{\mathbf{E}}_{x\sim D}\ \mathop{\mathbf{E}}_{h,h'\sim\rho_{\mathbf{w}}^2} \mathbf{I}[h(\mathbf{x}) = 1]\,\mathbf{I}[h'(\mathbf{x}) = -1] \\
&= 2 \mathop{\mathbf{E}}_{x\sim D}\ \mathop{\mathbf{E}}_{h\sim\rho_{\mathbf{w}}} \mathbf{I}[h(\mathbf{x}) = 1] \mathop{\mathbf{E}}_{h'\sim\rho_{\mathbf{w}}} \mathbf{I}[h'(\mathbf{x}) = -1] \\
&= 2 \mathop{\mathbf{E}}_{x\sim D} \Phi\left(\tfrac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right) \Phi\left(-\tfrac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right) = \mathop{\mathbf{E}}_{x\sim D} \Phi_{\text{dis}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right).
\end{aligned}$$

Thus, finding the optimal $\rho_{\mathbf{w}}$ in Equation (6) is equivalent to find the vector $\mathbf{w}$ that minimizes,

$$C\sum_{i=1}^{m}\Phi\left(y_i^s\,\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)+A\left|\sum_{i=1}^{m}\Phi_{\text{dis}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)-\Phi_{\text{dis}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\right|+\frac{\|\mathbf{w}\|^2}{2}.$$

The latter equation is highly non-convex. To make the optimization problem more tractable, we replace the loss function Φ by its convex relaxation Φ_{cvx} (as in Section 5.1) and minimize the resulting cost function by gradient descent. Even if this optimization task is still not convex (Φ_{dis} is quasiconcave), our empirical study shows no need to perform many restarts to find a suitable solution. We name this DA algorithm PBDA. Note that the kernel trick allows us to work with dual weight vector $\boldsymbol{\alpha} \in \mathbb{R}^{2m}$ that is a linear classifier in an augmented space. Given a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we have $h_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^m \alpha_i k(\mathbf{x}_i^s, \mathbf{x}) + \sum_{i=1}^m \alpha_{i+m} k(\mathbf{x}_i^t, \mathbf{x})$. See Supplementary Material for algorithm details.

6. Experiments

PBDA has been evaluated on a toy problem and a sentiment dataset. We compare it with two non-DA algorithms, SVM and PBGD3 (presented in Section 5.1), but also with the DA algorithm DASVM[8] (Bruzzone & Marconcini, 2010) and the DA co-training method CODA[9]. In Chen et al. (2011), CODA has showed best results on the dataset considered in our Section 6.2. Each parameters are selected with a grid search via a classical 5-folds cross-validation (^{CV}) on the source sample for PBGD3 and SVM, and via a 5-folds reverse validation (^{RCV}) on the source and the (unlabeled) target samples (see Bruzzone & Marconcini (2010); Zhong et al. (2010)) for CODA, DASVM, and PBDA.

6.1. Toy Problem: Two Inter-Twinning Moons

The source domain considered here is the classical binary problem with two inter-twinning moons, each class corresponding to one moon (Figure 1). We then consider 7 different target domains by rotating anticlockwise the source domain according to 7 angles (from 10° to 90°). The higher the angle, the more difficult the problem becomes. For each domain, we generate 300 instances (150 of each class). Moreover, to assess the generalization ability of our approach, we evaluate each algorithm on an independent test set of 1,000 target points (not provided to the algorithms). We make use of a Gaussian kernel for all the methods. Each DA problem is repeated 10 times, and we report the average error rates on Table 1. Note that since CODA decomposes features for applying co-training, it is not appropriate here (we have only 2 features). We remark that our PBDA provides the best performances except for 50° and 20°, indicating that PBDA accurately tackles DA tasks. It shows a nice adaptation ability, especially for the hardest problem, probably due to the fact that dis_ρ is tighter and seems to be a good regularizer in a DA situation. The adaptation versus risk minimization trade-off suggested by Theorem 5 appears in Figure 1. Indeed, the plot illustrates that PBDA accepts to have a lower source accuracy to maintain its performance on the target domain, at least when the source and the target domains are not so different. Note however that for large angles, PBDA prefers to "focus" on the source accuracy. We claim that this is a reasonable behavior for a DA algorithm.

Table 1. Average error rate results for 7 rotation angles.

	10°	20°	30°	40°	50°	70°	90°
PBGD3^{CV}	*0*	0.088	0.210	0.273	0.399	0.776	0.824
SVM^{CV}	*0*	0.104	0.24	0.312	0.4	0.764	0.828
DASVM^{RCV}	*0*	*0*	0.259	0.284	*0.334*	0.747	0.82
PBDA^{RCV}	*0*	0.094	*0.103*	*0.225*	0.412	*0.626*	*0.687*

6.2. Sentiment Analysis Dataset

We consider the popular *Amazon reviews* dataset (Blitzer et al., 2006) composed of reviews of four types of Amazon.com products (books, DVDs, electronics, kitchen appliances). Originally, the reviews corresponded to a rate between 1 and 5 stars and the feature space (of unigrams and bigrams) has on average a dimension of 100,000. We follow the simplified binary setting proposed by Chen et al. (2011). More precisely, we regroup ratings in two classes (products rated higher that 3 stars and products rated lower than 4 stars). Also, the dimensionality is reduced in the following way: we only keep the features that appear at least 10 times in a particular DA task, reducing the number of features to about 40,000). Finally, the data are pre-processed with a standard tf-idf re-weighting. One type of product is a domain, then we perform 12

[8] DASVM try to maximize iteratively a notion of margin on self-labeled target examples.

[9] CODA looks iteratively for target features related to the training set.

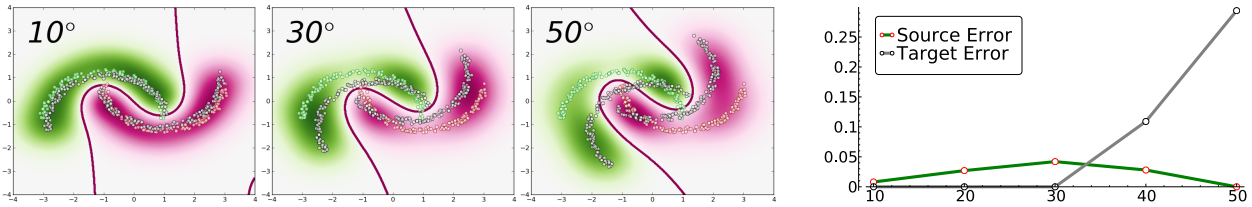

Figure 1. Illustration of the decision boundary of PBDA on 3 rotations angles for fixed parameters $A = C = 1$. The two classes of the source sample are green and pink, and target (unlabeled) sample is grey. The right plot shows corresponding source and target errors. We intentionally avoid to tune PBDA parameters to highlight its inherent adaptation behavior.

Table 2. Error rates for sentiment analysis dataset. B, D, E, K respectively denotes books, DVDs, electronics, kitchen.

	B→D	B→E	B→K	D→B	D→E	D→K	E→B	E→D	E→K	K→B	K→D	K→E	Avg.
PBGD3^{CV}	*0.174*	0.275	0.236	*0.192*	0.256	0.211	0.268	0.245	*0.127*	0.255	0.244	0.235	0.226
SVM^{CV}	0.179	0.290	0.251	0.203	0.269	0.232	0.287	0.267	0.129	0.267	0.253	0.149	0.231
DASVM^{RCV}	0.193	*0.226*	*0.179*	0.202	*0.186*	0.183	0.305	*0.214*	0.149	0.259	*0.198*	0.157	*0.204*
CODA^{RCV}	0.181	0.232	0.215	0.217	0.214	*0.181*	0.275	0.239	0.134	*0.247*	0.238	0.153	0.210
PBDA^{RCV}	0.183	0.263	0.229	0.197	0.241	0.186	*0.232*	0.221	0.141	*0.247*	0.233	*0.129*	0.208

DA tasks. For example, "books→DVDs" corresponds to the task for which books is the source domain and DVDs the target one. The algorithms use a linear kernel and consider 2,000 labeled source examples and 2,000 unlabeled target examples. We evaluate them on separate target test sets proposed by Chen et al. (2011) (between 3,000 and 6,000 examples), and we report the results on Table 2. We make the following observations. First, as expected, the DA approaches provide the best average results. Then, PBDA is on average better than CODA, but less accurate than DASVM. However, PBDA is competitive: the results are not significantly different from CODA and DASVM. Moreover, we have observed that PBDA is significantly faster than CODA and DASVM: these two algorithms are based on costly iterative procedures increasing the running time by at least a factor of 5 in comparison of PBDA. In fact, the clear advantage of PBDA is that we jointly optimize the terms of our bound in one step. PAC-Bayes appears thus relevant in the context of DA and we could imagine to improve PBDA by making use of the tools offered by the PAC-Bayesian theory.

7. Conclusion and Future Work

In this paper, we define a domain divergence pseudo-metric that is based on an average disagreement over a set of classifiers, along with consistency bounds for justifying its estimation from samples. This measure helps us to derive a first PAC-Bayesian bound for domain adaptation. Moreover, from this bound we design a well-founded and competitive algorithm (PBDA) that can directly optimize the bound for linear classifiers. We think that this PAC-Bayesian analysis opens the door to develop new domain adaptation methods by making use of the possibilities offered by the PAC-Bayesian theory, and gives rise to new interesting directions of research, among which the following ones.

PAC-Bayes allows one to deal with an *a priori* belief on what are the best classifiers; in this paper we opted for a non-informative prior that consists on a Gaussian centered at the origin of the linear classifier space. The question of finding a relevant prior in a DA situation is an exciting direction which could also be exploited when some few target labels are available.

Another promising issue is to address the problem of the hyperparameter selection. Indeed, the adaptation capability of our algorithm PBDA could be even put further with a specific PAC-Bayesian validation procedure. An idea would be to propose a kind of (reverse) validation technique that takes into account some particular prior distributions. This is also linked with model selection for domain adaptation tasks.

Besides, deriving a result similar to Equation (4) (the C-bound) for domain adaptation could be of high interest. Indeed, such an approach considers the first two moments of the margin of the weighted majority vote. This could help us to take into account both a kind of margin information over unlabeled data and the distribution disagreement (these two elements seem of crucial importance in domain adaptation).

Acknowledgments This work was supported in part by the French projects VideoSense ANR-09-CORD-026 and LAMPADA ANR-09-EMER-007-02, and in part by NSERC discovery grant 262067. Computations were performed on Compute Canada and Calcul Québec infrastructures (founded by CFI, NSERC and FRQ).

References

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

Ben-David, S. and Urner, R. On the hardness of domain adaptation and the utility of unlabeled target samples. In *ALT*, pp. 139–153, 2012.

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.W. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010a.

Ben-David, S., Lu, T., Luu, T., and Pal, D. Impossibility theorems for domain adaptation. *JMLR W&CP, AISTAT*, 9:129–136, 2010b.

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

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

C. Zhang, L. Zhang, J. Ye. Generalization bounds for domain adaptation. In *NIPS*, 2012.

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

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

Cortes, C. and Mohri, M. Domain adaptation in regression. In *ALT*, pp. 308–323, 2011.

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

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

Jiang, J. A literature survey on domain adaptation of statistical classifiers. Technical report, CS Department at Univ. of Illinois at Urbana-Champaign, 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*, 2006.

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

Laviolette, F., Marchand, M., and Roy, J.-F. From PAC-Bayes bounds to quadratic programs for majority votes. In *ICML*, 2011.

Li, X. and Bilmes, J. A bayesian divergence prior for classifier adaptation. In *AISTATS-2007*, 2007.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *COLT*, pp. 19–30, 2009a.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Multiple source adaptation and the rényi divergence. In *UAI*, pp. 367–374, 2009b.

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

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N.D. *Dataset Shift in Machine Learning*. MIT Press, 2009. ISBN 0262170051, 9780262170055.

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*, 2010.

Supplementary Material to A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers

Pascal Germain PASCAL.GERMAIN@IFT.ULAVAL.CA
Département d'informatique et de génie logiciel, Université Laval, Québec, Canada

Amaury Habrard AMAURY.HABRARD@UNIV-ST-ETIENNE.FR
Laboratoire Hubert Curien UMR CNRS 5516, Université Jean Monnet, 42000 St-Etienne, France

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

Emilie Morvant EMILIE.MORVANT@LIF.UNIV-MRS.FR
Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, 13013, Marseille, France

In this document, Section 1 contains some lemmas used in subsequent proofs, Section 2 presents an extended proof of the bound on the domain disagreement $\mathrm{dis}_\rho(D_S, D_T)$ (Theorem 3 of the main paper), Section 3 introduces other PAC-Bayesian bounds for $\mathrm{dis}_\rho(D_S, D_T)$ and $R_{P_T}(G_\rho)$, Section 4 shows equations and implementation details about PBDA (our proposed learning algorithm for PAC-Bayesian DA tasks).

1. Some tools

Lemma 1 (Markov's inequality). *Let Z be a random variable and $t \geq 0$, then,*

$$P(|Z| \geq t) \;\leq\; \mathbf{E}\,(|Z|)\,/\,t\,.$$

Lemma 2 (Jensen's inequality). *Let Z be an integrable real-valued random variable and $g(\cdot)$ any function.*

If $g(\cdot)$ is convex, then,

$$g(\mathbf{E}\,[Z]) \;\leq\; \mathbf{E}\,[g(Z)]\,.$$

If $g(\cdot)$ is concave, then,

$$g(\mathbf{E}\,[Z]) \;\geq\; \mathbf{E}\,[g(Z)]\,.$$

Lemma 3 (Maurer (2004)). *Let $X = (X_1, \ldots, X_m)$ be a vector of i.i.d. random variables, $0 \leq X_i \leq 1$, with $\mathbf{E}\,X_i = \mu$. Denote $X' = (X'_1, \ldots, X'_m)$, where X'_i is the unique Bernoulli ($\{0,1\}$-valued) random variable with $\mathbf{E}\,X'_i = \mu$. If $f : [0,1]^n \to \mathbb{R}$ is convex, then,*

$$\mathbf{E}\,[f(X)] \;\leq\; \mathbf{E}\,[f(X')]\,.$$

Lemma 4 (from Inequalities (1) and (2) of Maurer (2004)). *Let $m \geq 8$, and $X = (X_1, \ldots, X_m)$ be a vector of i.i.d. random variables, $0 \leq X_i \leq 1$. Then,*

$$\sqrt{m} \leq \mathbf{E}\,\exp\left(m\mathrm{kl}\left(\frac{1}{m}\sum_{i=1}^{n} X_i \,\middle\|\, \mathbf{E}\,[X_i]\right)\right) \leq 2\sqrt{m},$$

where, $\quad \mathrm{kl}(a\,\|\,b) \stackrel{\text{def}}{=} a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}\,. \qquad (7)$

2. Detailed Proof of Theorem 3

We recall the Theorem 3 of the main paper.

Theorem 3. *For any distributions D_S and D_T over X, any set of hypothesis $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta \in (0,1]$, and any real number $\alpha > 0$, with a probability at least $1-\delta$ over the choice of $S \times T \sim (D_S \times D_T)^m$, for every ρ on $\mathcal{H}$, we have,*

$$\mathrm{dis}_\rho(D_S, D_T) \leq \frac{2\alpha\Big[\,\mathrm{dis}_\rho(S,T) + \frac{2\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}}{m\times\alpha} + 1\Big] - 1}{1 - e^{-2\alpha}},$$

where $\mathrm{dis}_\rho(S,T)$ is the empirical domain disagreement.

Proof. Firstly, we propose to upper-bound,

$$d^{(1)} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} [R_{D_S}(h,h') - R_{D_T}(h,h')]\,,$$

by its empirical counterpart,

$$d^{(1)}_{S\times T} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} [R_S(h,h') - R_T(h,h')]\,.$$

and some extra terms related to the Kullback-Leibler divergence between the posterior and the prior.

To do that, we consider an "abstract" classifier $\hat{h} \stackrel{\text{def}}{=} (h, h') \in \mathcal{H}^2$ chosen according a distribution $\hat{\rho}$, with $\hat{\rho}(\hat{h}) = \rho(h)\rho(h')$. Notice that with $\hat{\pi}(\hat{h}) = \pi(h)\pi(h')$, we obtain that $\mathrm{KL}(\hat{\rho}\|\hat{\pi}) = 2\mathrm{KL}(\rho\|\pi)$,

$$\begin{aligned}\mathrm{KL}(\hat{\rho}\|\hat{\pi}) &= \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \ln \frac{\rho(h)\rho(h')}{\pi(h)\pi(h')} \\ &= \mathop{\mathbf{E}}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)} + \mathop{\mathbf{E}}_{h'\sim\rho} \ln \frac{\rho(h')}{\pi(h')} \\ &= 2 \mathop{\mathbf{E}}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)} = 2\mathrm{KL}(\rho\|\pi)\,. \qquad (8)\end{aligned}$$

Let us define the "abstract" loss of $\hat{h}$ on a pair of examples $(\mathbf{x}^s, \mathbf{x}^t) \sim D_{S\times T} = D_S \times D_T$ by,

$$\mathcal{L}_{d^{(1)}}(\hat{h}, \mathbf{x}^s, \mathbf{x}^t) \stackrel{\text{def}}{=} \frac{1 + \mathcal{L}_{0\text{-}1}(h(\mathbf{x}^s), h'(\mathbf{x}^s)) - \mathcal{L}_{0\text{-}1}(h(\mathbf{x}^t), h'(\mathbf{x}^t))}{2}.$$

Therefore, the "abstract" risk of $\hat{h}$ on the joint distribution is defined as,

$$R^{(1)}_{D_{S\times T}}(\hat{h}) = \mathop{\mathbf{E}}_{\mathbf{x}^s\sim D_S} \mathop{\mathbf{E}}_{\mathbf{x}^t\sim D_T} \mathcal{L}_{d^{(1)}}(\hat{h}, \mathbf{x}^s, \mathbf{x}^t)\,,$$

and the error of the related Gibbs classifier associated with this loss is,

$$R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}}) = \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} R^{(1)}_{D_{S\times T}}(\hat{h})\,.$$

The empirical counterparts of these two quantities are,

$$R^{(1)}_{S\times T}(\hat{h}) = \mathop{\mathbf{E}}_{(\mathbf{x}^s,\mathbf{x}^t)\sim S\times T} \mathcal{L}_{d^{(1)}}(\hat{h}, \mathbf{x}^s, \mathbf{x}^t)$$

and,

$$R^{(1)}_{S\times T}(G_{\hat{\rho}}) = \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} R^{(1)}_{S\times T}(\hat{h})\,.$$

It is easy to show that,

$$d^{(1)} = 2R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}}) - 1, \qquad (9)$$

$$d^{(1)}_{S\times T} = 2R^{(1)}_{S\times T}(G_{\hat{\rho}}) - 1. \qquad (10)$$

As $\mathcal{L}_{d^{(1)}}$ lies in $[0,1]$, we can bound the true $R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})$ following the proof process of Th. 2 of the main paper (with $c=2\alpha$). To do so, we define the convex function,

$$\mathcal{F}(p) \stackrel{\text{def}}{=} -\ln[1 - (1 - e^{-2\alpha})p]\,, \qquad (11)$$

and consider the non-negative random variable,

$$\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h})) - 2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}.$$

We apply Markov's inequality (Lemma 1 of this Supp. Material). For every $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S\times T \sim (D_{S\times T})^m$, we have,

$$\begin{aligned}&\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h})) - 2\alpha R^{(1)}_{S\times T}(\hat{h})\right)} \\ &\le \frac{1}{\delta} \mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h})) - 2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}.\end{aligned}$$

By taking the logarithm on each side of the previous inequality, and transforming the expectation over $\hat{\pi}$ into an expectation over $\hat{\rho}$, we obtain that,

$$\begin{aligned}&\ln\left[\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} \frac{\hat{\pi}(\hat{h})}{\hat{\rho}(\hat{h})} e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h})) - 2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}\right] \\ &\le \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h})) - 2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}\right] \\ &= \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))} \mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} e^{-2m\alpha R^{(1)}_{S\times T}(\hat{h})}\right]. \qquad (12)\end{aligned}$$

For a classifier $\hat{h}$, let us define a random variable $X_{\hat{h}}$ that follows a binomial distribution of m trials with a probability of success $R^{(1)}_{D_{S\times T}}(\hat{h})$ denoted by $B\big(m, R^{(1)}_{D_{S\times T}}(\hat{h})\big)$. Lemma 3 gives,

$$\begin{aligned}&\mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} e^{-2m\alpha R^{(1)}_{S\times T}(\hat{h})} \\ &\quad\le \mathop{\mathbf{E}}_{X_{\hat{h}}\sim B(m, R^{(1)}_{D_{S\times T}}(\hat{h}))} e^{-2\alpha X_{\hat{h}}} \\ &\quad= \sum_{k=0}^{m} \Pr_{X_{\hat{h}}\sim B(m, R^{(1)}_{D_{S\times T}}(\hat{h}))}\Big(X_{\hat{h}} = k\Big) e^{-2\alpha k} \\ &\quad= \sum_{k=0}^{m} \binom{m}{k} \big(R^{(1)}_{S\times T}(\hat{h})\big)^k \big(1 - R^{(1)}_{S\times T}(\hat{h})\big)^{m-k} e^{-2\alpha k} \\ &\quad= \sum_{k=0}^{m} \binom{m}{k} \big(R^{(1)}_{S\times T}(\hat{h}) e^{-2\alpha}\big)^k \left(1 - R^{(1)}_{S\times T}(\hat{h})\right)^{m-k} \\ &\quad= \left[R^{(1)}_{S\times T}(\hat{h}) e^{-2\alpha} + \left(1 - R^{(1)}_{S\times T}(\hat{h})\right)\right]^m.\end{aligned}$$

The last line result, together with the choice of $\mathcal{F}$ (Eq. (11)), leads to,

$$\begin{aligned}&\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))} \mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} e^{-2m\alpha R^{(1)}_{S\times T}(\hat{h})} \\ &\le \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))} \left[R^{(1)}_{S\times T}(\hat{h}) e^{-2\alpha} + \left(1 - R^{(1)}_{S\times T}(\hat{h})\right)\right]^m \\ &= \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} 1 = 1\,.\end{aligned}$$

We can now upper bound Eq. (12) simply by,

$$\ln\left[\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} \frac{\hat{\pi}(\hat{h})}{\hat{\rho}(\hat{h})} e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h})) - 2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}\right] \le \ln\frac{1}{\delta}\,.$$

Let us insert the term $\mathrm{KL}(\rho\|\pi)$ in the left-hand side of the last inequality and find a lower bound by using Jensen's inequality (Lemma 2) twice, first on the concave logarithm function and then on the convex function $\mathcal{F}$,

$$\begin{aligned}
&\ln\left[\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}\frac{\hat{\pi}(\hat{h})}{\hat{\rho}(\hat{h})}e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))-2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}\right]\\
&= \ln\left[\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))-2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}\right]-2\mathrm{KL}(\rho\|\pi)\\
&\geq \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))-2\alpha R^{(1)}_{S\times T}(\hat{h})\right)-2\mathrm{KL}(\rho\|\pi)\\
&\geq m\mathcal{F}(\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}R^{(1)}_{D_{S\times T}}(\hat{h}))-2m\alpha\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}R^{(1)}_{S\times T}(\hat{h})-2\mathrm{KL}(\rho\|\pi)\\
&= m\mathcal{F}(R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}}))-2m\alpha R^{(1)}_{S\times T}(G_{\hat{\rho}})-2\mathrm{KL}(\rho\|\pi)\,.
\end{aligned}$$

We then have,

$$m\mathcal{F}(\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}R^{(1)}_{D_{S\times T}}(\hat{h}))-2m\alpha\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}R^{(1)}_{S\times T}(\hat{h})-2\mathrm{KL}(\rho\|\pi)\;\leq\;\ln\frac{1}{\delta}.$$

This, in turn, implies that,

$$\mathcal{F}(R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}}))\leq 2\alpha R^{(1)}_{S\times T}(G_{\hat{\rho}})+\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m}\,.$$

Now, by isolating $R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})$, we obtain,

$$R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})\leq\frac{1}{1-e^{-2\alpha}}\left[1-e^{-\left(2\alpha R^{(1)}_{S\times T}(G_{\hat{\rho}})+\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m}\right)}\right],$$

and from the inequality $1-e^{-x}\leq x$,

$$R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})\leq\frac{1}{1-e^{-2\alpha}}\left[2\alpha R^{(1)}_{S\times T}(G_{\hat{\rho}})+\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m}\right].$$

It then follows from Equations (9) and (10) that, with probability at least $1-\frac{\delta}{2}$ over the choice of $S\times T\sim(D_S\times D_T)^m$, we have,

$$\frac{d^{(1)}+1}{2}\leq\frac{2\alpha}{1-e^{-2\alpha}}\left[\frac{d^{(1)}_{S\times T}+1}{2}+\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m\times 2\alpha}\right],$$

We now bound $d^{(2)}\stackrel{\text{def}}{=}\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}[R_{D_T}(h,h')-R_{D_S}(h,h')]$ using exactly the same argument as for $d^{(1)}$ except that we instead consider the following "abstract" loss of $\hat{h}$ on a pair of examples $(\mathbf{x}^s,\mathbf{x}^t)\sim D_{S\times T}=D_S\times D_T$:

$$\mathcal{L}_{d^{(1)}}(\hat{h},\mathbf{x}^s,\mathbf{x}^t)\stackrel{\text{def}}{=}\frac{1+\mathcal{L}_{0\text{-}1}(h(\mathbf{x}^t),h'(\mathbf{x}^t)-\mathcal{L}_{0\text{-}1}(h(\mathbf{x}^s),h'(\mathbf{x}^s))}{2}.$$

We then obtain that, with probability at least $1-\frac{\delta}{2}$ over the choice of $S\times T\sim(D_S\times D_T)^m$,

$$\frac{d^{(2)}+1}{2}\leq\frac{2\alpha}{1-e^{-2\alpha}}\left[\frac{d^{(2)}_{S\times T}+1}{2}+\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m\times 2\alpha}\right].$$

To finish the proof, note that by definition, we have that $d^{(1)}=-d^{(2)}$, hence

$$|d^{(1)}|=|d^{(2)}|=\mathrm{dis}_\rho(D_S,D_T),$$

and,

$$|d^{(1)}_{S\times T}|=|d^{(2)}_{S\times T}|=\mathrm{dis}_\rho(S,T).$$

Then, the maximum of the bound on $d^{(1)}$ and the bound on $d^{(2)}$ gives a bound on $\mathrm{dis}_\rho(D_S,D_T)$.

Finally, by the union bound, we have that, with probability $1-\delta$ over the choice of $S\times T\sim(D_S\times D_T)^m$, we have,

$$\frac{|d^{(1)}|+1}{2}\leq\frac{\alpha}{1-e^{-2\alpha}}\left[|d^{(1)}_{S\times T}|+1+\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{2}{\delta}}{m\times\alpha}\right],$$

or, which is equivalent,

$$\mathrm{dis}_\rho(D_S,D_T)\leq\frac{2\alpha\left[\mathrm{dis}_\rho(S,T)+\frac{2\mathrm{KL}(\rho\|\pi)+\ln\frac{2}{\delta}}{m\times\alpha}+1\right]-1}{1-e^{-2\alpha}},$$

and we are done. □

3. Other PAC-Bayesian Bounds

3.1. PAC-Bayesian Bounds with the kl term

Let us recall the PAC-Bayesian bound proposed by Seeger (2002), in which the trade-off between the complexity and the risk is handled by the kl function defined by Equation (7) in this supplementary materials.

Theorem 6 (Seeger (2002))**.** *For any domain P_S over $X\times Y$, any set of hypothesis $\mathcal{H}$, and any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of $S\sim(P_S)^m$, for every ρ over $\mathcal{H}$, we have,*

$$\mathrm{kl}\Big(R_S(G_\rho)\,\Big\|\,R_{P_S}(G_\rho)\Big)\;\leq\;\frac{1}{m}\left[\mathrm{KL}(\rho\,\|\,\pi)+\ln\frac{2\sqrt{m}}{\delta}\right].$$

Here is a "Seeger's type" PAC-Bayesian bound for our domain disagreement dis_ρ.

Theorem 7. *For any distributions D_S and D_T over X, any set of hypothesis $\mathcal{H}$, and any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of $S\times T\sim(D_S\times D_T)^m$, for every ρ on $\mathcal{H}$, we have,*

$$\mathrm{kl}\Big(\tfrac{\mathrm{dis}_\rho(S,T)+1}{2}\,\Big\|\,\tfrac{\mathrm{dis}_\rho(D_S,D_T)+1}{2}\Big)\leq\tfrac{1}{m}\Big[2\mathrm{KL}(\rho\|\pi)+\ln\tfrac{2\sqrt{m}}{\delta}\Big].$$

Proof. Similarly as in the proof of Theorem 3, we will first bound,

$$d^{(1)} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[R_{D_S}(h,h') - R_{D_T}(h,h')\right],$$

by its empirical counterpart,

$$d^{(1)}_{S\times T} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[R_S(h,h') - R_T(h,h')\right],$$

and some extra terms related to the Kullback-Leibler divergence between the posterior and the prior. However, a notable difference with the proof of Theorem 3 is that the obtained bound will be simultaneously valid as an upper and a lower bound. Because of this, there will no need here to redo the all the proof to bound

$$d^{(2)} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[R_{D_T}(h,h') - R_{D_S}(h,h')\right],$$

and also, the present proof will not require the use of the union bound argument.

Again, we consider "abstract" classifiers $\hat{h} \in \mathcal{H}^2$ whose loss on a pair of examples $(\mathbf{x}^s, \mathbf{x}^t) \sim D_{S\times T}$ is defined by,

$$\mathcal{L}_{d^{(1)}}(\hat{h}, \mathbf{x}^s, \mathbf{x}^t) \stackrel{\text{def}}{=} \frac{1 + \mathcal{L}_{0\text{-}1}(h(\mathbf{x}^s), h'(\mathbf{x}^s)) - \mathcal{L}_{0\text{-}1}(h(\mathbf{x}^t), h'(\mathbf{x}^t))}{2}.$$

Note that, again, $\mathcal{L}_{d^{(1)}}$ lies in $[0,1]$, and that $R^{(1)}_{S\times T}(\hat{h})$ and $R^{(1)}_{D_{S\times T}}(\hat{h})$ are as defined in the proof of Theorem 3.

Now, let us consider the non-negative random variable,

$$\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h}) \middle\| R^{(1)}_{D_{S\times T}}(\hat{h})\right)}.$$

We apply Markov's inequality (Lemma 1). For every $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S\times T \sim (D_{S\times T})^m$, we have,

$$\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h}) \middle\| R^{(1)}_{D_{S\times T}}(\hat{h})\right)} \leq \frac{1}{\delta} \mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h}) \middle\| R^{(1)}_{D_{S\times T}}(\hat{h})\right)}.$$

By taking the logarithm on each side of the previous inequality, and transforming the expectation over $\hat{\pi}$ into an expectation over $\hat{\rho}$, we then obtain that,

$$\begin{aligned}
&\ln\left[\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} \frac{\hat{\pi}(\hat{h})}{\hat{\rho}(\hat{h})} e^{m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h}) \middle\| R^{(1)}_{D_{S\times T}}(\hat{h})\right)}\right] \qquad (13)\\
&\leq \ln\left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h}) \middle\| R^{(1)}_{D_{S\times T}}(\hat{h})\right)}\right]\\
&\leq \ln\frac{2\sqrt{m}}{\delta}.
\end{aligned}$$

The last inequality comes from the Maurer's lemma (Lemma 4).

Let us now re-write a part of the equation as $\mathrm{KL}(\rho\|\pi)$ and let us then find a lower bound by using twice the Jensen's inequality (Lemma 2), first on the concave logarithm function, and then on the convex function kl,

$$\begin{aligned}
&\ln\left[\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} \frac{\hat{\pi}(\hat{h})}{\hat{\rho}(\hat{h})} e^{m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h}) \middle\| R^{(1)}_{D_{S\times T}}(\hat{h})\right)}\right]\\
&= \ln\left[\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} e^{m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h}) \middle\| R^{(1)}_{D_{S\times T}}(\hat{h})\right)}\right] - 2\mathrm{KL}(\rho\|\pi)\\
&\geq \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h}) \,\middle\|\, R^{(1)}_{D_{S\times T}}(\hat{h})\right) - 2\mathrm{KL}(\rho\|\pi)\\
&\geq m\,\mathrm{kl}\left(\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} R^{(1)}_{S\times T}(\hat{h}) \,\middle\|\, \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} R^{(1)}_{D_{S\times T}}(\hat{h})\right) - 2\mathrm{KL}(\rho\|\pi)\\
&\geq m\,\mathrm{kl}\left(R^{(1)}_{S\times T}(G_{\hat{\rho}}) \,\middle\|\, R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})\right) - 2\mathrm{KL}(\rho\|\pi).
\end{aligned}$$

This implies that,

$$\mathrm{kl}\left(R^{(1)}_{S\times T}(G_{\hat{\rho}}) \middle\| R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})\right) \leq \frac{1}{m}\left[2\mathrm{KL}(\rho\,\|\,\pi) + \ln\frac{2\sqrt{m}}{\delta}\right].$$

Since, as in the proof of Theorem 3 for $d^{(1)}$, we have: $d^{(1)} = 2R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}}) - 1$ and $d^{(1)}_{S\times T} = 2R^{(1)}_{S\times T}(G_{\hat{\rho}}) - 1$, the previous line directly implies a bound on $d^{(1)}$ from its empirical counterpart $d^{(1)}_{S\times T}$. Hence, with probability at least $1-\delta$ over the choice of $S\times T \sim (D_S \times D_T)^m$, we have,

$$\mathrm{kl}\left(\frac{d^{(1)}_{S\times T}+1}{2} \middle\| \frac{d^{(1)}+1}{2}\right) \leq \frac{1}{m}\left[2\mathrm{KL}(\rho\,\|\,\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]. \quad (14)$$

We claim that we also have,

$$\mathrm{kl}\left(\frac{|d^{(1)}_{S\times T}|+1}{2} \middle\| \frac{|d^{(1)}|+1}{2}\right) \leq \frac{1}{m}\left[2\mathrm{KL}(\rho\,\|\,\pi) + \ln\frac{2\sqrt{m}}{\delta}\right], \quad (15)$$

which, since

$$|d^{(1)}| = \mathrm{dis}_\rho(D_S, D_T) \;\text{ and }\; |d^{(1)}_{S\times T}| = \mathrm{dis}_\rho(S,T),$$

implies the result. Hence to finish the proof, let us prove the claim of Equation (15). There are four cases to consider.

Case 1: $d^{(1)}_{S\times T} \geq 0$ and $d^{(1)} \geq 0$. There is nothing to prove since in that case, Equations (14) and (15) coincide.

Case 2: $d^{(1)}_{S\times T} \leq 0$ and $d^{(1)} \leq 0$. This case reduces to Case 1 because of the following property of $\mathrm{kl}(\cdot\|\cdot)$:

$$\mathrm{kl}\left(\tfrac{a+1}{2} \middle\| \tfrac{b+1}{2}\right) = \mathrm{kl}\left(\tfrac{-a+1}{2} \middle\| \tfrac{-b+1}{2}\right). \quad (16)$$

Case 3: $d^{(1)}_{S\times T} \leq 0$ and $d^{(1)} \geq 0$. From straightforward calculations, one can show that,

$$\begin{aligned}
&\mathrm{kl}\left(\tfrac{|d^{(1)}_{S\times T}|+1}{2}\middle\|\tfrac{|d^{(1)}|+1}{2}\right) - \mathrm{kl}\left(\tfrac{d^{(1)}_{S\times T}+1}{2}\middle\|\tfrac{d^{(1)}+1}{2}\right)\\
&= \mathrm{kl}\left(\tfrac{-d^{(1)}_{S\times T}+1}{2}\middle\|\tfrac{d^{(1)}+1}{2}\right) - \mathrm{kl}\left(\tfrac{d^{(1)}_{S\times T}+1}{2}\middle\|\tfrac{d^{(1)}+1}{2}\right)\\
&= \left(\tfrac{-d^{(1)}_{S\times T}+1}{2} - \tfrac{d^{(1)}_{S\times T}+1}{2}\right)\ln\left(\frac{1}{\frac{d^{(1)}+1}{2}}\right)\\
&\quad + \left(\left(1-\tfrac{-d^{(1)}_{S\times T}+1}{2}\right) - \left(1-\tfrac{d^{(1)}_{S\times T}+1}{2}\right)\right)\ln\left(\frac{1}{1-\frac{d^{(1)}+1}{2}}\right)\\
&= \left(-d^{(1)}_{S\times T}\right)\ln\left(\frac{1}{\frac{d^{(1)}+1}{2}}\right) + \left(d^{(1)}_{S\times T}\right)\ln\left(\frac{1}{1-\frac{d^{(1)}+1}{2}}\right)\\
&= \left(-d^{(1)}_{S\times T}\right)\ln\left(\frac{1}{\frac{d^{(1)}+1}{2}}\right) + \left(d^{(1)}_{S\times T}\right)\ln\left(\frac{1}{\frac{-d^{(1)}+1}{2}}\right)\\
&= d^{(1)}_{S\times T}\ \ln\left(\tfrac{d^{(1)}+1}{-d^{(1)}+1}\right)\\
&\leq 0\,. \qquad (17)
\end{aligned}$$

The last inequality follows from the fact that we have $d^{(1)}_{S\times T} \leq 0$ and $d^{(1)} \geq 0$.

Hence, from Equations (17) and (14), we have,

$$\begin{aligned}
\mathrm{kl}\left(\tfrac{|d^{(1)}_{S\times T}|+1}{2}\middle\|\tfrac{|d^{(1)}|+1}{2}\right) &\leq \mathrm{kl}\left(\tfrac{d^{(1)}_{S\times T}+1}{2}\middle\|\tfrac{d^{(1)}+1}{2}\right)\\
&\leq \frac{1}{m}\left[2\mathrm{KL}(\rho\,\|\,\pi)+\ln\frac{2\sqrt{m}}{\delta}\right],
\end{aligned}$$

as wanted.

Case 4: $d^{(1)}_{S\times T} \geq 0$ and $d^{(1)} \leq 0$. Again because of Equation (16), this case reduces to Case 3, and we are done. □

From the preceding "Seeger's type" results, one can then obtain the following PAC-Bayesian DA-bound.

Theorem 8. *For any domains P_S and P_T (respectively with marginals D_S and D_T) over $X\times Y$, any set of hypothesis $\mathcal{H}$, and any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of $S\times T\sim(P_S\times D_T)^m$, we have,*

$$R_{P_T}(G_\rho) - R_{P_T}(G_{\rho^*_T}) \ \leq\ \sup\mathcal{R}_\rho + \sup\mathcal{D}_\rho + \lambda_\rho\,,$$

where $\lambda_\rho \stackrel{\text{def}}{=} R_{D_T}(G_\rho, G_{\rho^*_T}) + R_{D_S}(G_\rho, G_{\rho^*_T})$ *and,*

$$\mathcal{R}_\rho \stackrel{\text{def}}{=} \left\{r : \mathrm{kl}\big(R_S(G_\rho)\|r\big) \leq \tfrac{1}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\tfrac{4\sqrt{m}}{\delta}\right]\right\},$$
$$\mathcal{D}_\rho \stackrel{\text{def}}{=} \left\{d : \mathrm{kl}\big(\tfrac{\mathrm{dis}_\rho(S,T)+1}{2}\big\|\tfrac{d+1}{2}\big) \leq \tfrac{1}{m}\left[2\mathrm{KL}(\rho\|\pi)+\ln\tfrac{4\sqrt{m}}{\delta}\right]\right\}.$$

Proof. The result is obtained by inserting Ths. 6 and 7 (with $\delta := \frac{\delta}{2}$) in Th. 4 of the main paper. □

3.2. PAC-Bayesian Bounds when $\mathbf{m\neq m'}$

In the main paper, for the sake of simplicity, we restrict to the case where m (the size of the source set S) and m' (the size of the target set T) are equal. All the results generalize to the $m \neq m'$ case. In this subsection, we will show how it can be done from a "McAllester's type" of bound (Similar results can be achieved for "Catoni's type" or "Seeger's type").

First we recall the PAC-Bayesian bound proposed by McAllester (2003), which is stated without a term allowing to control the trade-off between the complexity and the risk.

Theorem 9 (McAllester (2003))**.** *For any domain P_S over $X\times Y$, any set of hypothesis $\mathcal{H}$, and any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of $S\sim(P_S)^m$, for every ρ over $\mathcal{H}$, we have,*

$$\left|R_{P_S}(G_\rho) - R_S(G_\rho)\right| \ \leq\ \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\,\|\,\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]}.$$

Now we can prove the following consistency bound for $\mathrm{dis}_\rho(D_S,D_T)$, when $m \neq m'$.

Theorem 10. *For any marginal distributions D_S and D_T over X, any set of hypothesis $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of $S\sim(D_S)^m$ and $T\sim(D_T)^{m'}$, for every ρ over $\mathcal{H}$, we have,*

$$\begin{aligned}
\left|\,\mathrm{dis}_\rho(D_S,D_T) - \mathrm{dis}_\rho(S,T)\right| \leq &\sqrt{\frac{1}{2m}\left[2\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]}\\
&+\sqrt{\frac{1}{2m'}\left[2\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m'}}{\delta}\right]}.
\end{aligned}$$

Proof. Let us consider the non-negative random variable,

$$\mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2}.$$

We apply Markov's inequality (Lemma 1). For every $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of $S\sim(D_S)^m$, we have,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2}\\
&\leq \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(D_S)^m}\ \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2}.
\end{aligned}$$

By taking the logarithm on each side of the previous inequality and transforming the expectation over π^2 into an expectation over ρ^2, we obtain that for every $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S \sim (D_S)^m$, and for every posterior distribution ρ, we have,

$$\ln \left[\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \frac{\pi(h)\pi(h')}{\rho(h)\rho(h')} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \right] \leq \ln \left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(D_S)^m} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \right].$$

Since $\ln(\cdot)$ is a concave function, we can apply the Jensen's inequality (Lemma 2). Then, for every $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S \sim (D_S)^m$, and for every posterior distribution ρ, we have,

$$\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \ln \left[\frac{\pi(h)\pi(h')}{\rho(h)\rho(h')} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \right] \leq \ln \left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(D_S)^m} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{(2m(R_{D_S}(h,h')-R_S(h,h'))^2} \right].$$

By the Equation (8),

$$\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \ln \left[\frac{\pi(h)\pi(h')}{\rho(h)\rho(h')} \right] = -2\mathrm{KL}(\rho\|\pi).$$

For every $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S \sim (D_S)^m$, and for every posterior distribution ρ, we have,

$$-2\mathrm{KL}(\rho\|\pi) + \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} m\, 2(R_{D_S}(h,h') - R_S(h,h'))^2 \leq \ln \left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(D_S)^m} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \right].$$

Since $2(a-b)^2$ is a convex function, we again apply Jensen inequality,

$$\left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} (R_{D_S}(h,h') - R_S(h,h')) \right)^2 \leq \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} (R_{D_S}(h,h') - R_S(h,h'))^2.$$

Thus, for every $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S \sim (D_S)^m$, and for every posterior distribution ρ, we have,

$$2m \left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_S}(h,h') - \mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_S(h,h') \right)^2 \leq 2\mathrm{KL}(\rho\|\pi) + \ln \left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(D_S)^m} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \right].$$

Let us now bound,

$$\ln \left[\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(D_S)^m} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \right].$$

To do so, we have,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim(D_S)^m} \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \\
&= \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \mathop{\mathbf{E}}_{S\sim(D_S)^m} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} && (18)\\
&\leq \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} \mathop{\mathbf{E}}_{S\sim(D_S)^m} e^{\mathrm{kl}(R_S(h,h')\|R_{D_S}(h,h'))} && (19)\\
&\leq 2\sqrt{m}. && (20)
\end{aligned}$$

Line (18) comes from the independence between D_S and π^2. The Pinsker's inequality,

$$2(q-p)^2 \leq \mathrm{kl}(q\|p) \quad \text{for any } p,q \in [0,1],$$

gives Line (19). The last Line (20) comes from the Maurer's lemma (Lemma 4).

Thus for every $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S \sim (D_S)^m$, and for every posterior distribution ρ, we obtain,

$$\begin{aligned}
&2m \left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_S}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_S(h,h') \right)^2 \leq 2\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \\
\Leftrightarrow \quad &\left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_S}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_S(h,h') \right)^2 \leq \frac{1}{2m} \left[2\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right] \\
\Leftrightarrow \quad &\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_S}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_S(h,h') \right| \leq \sqrt{\frac{1}{2m} \left[2\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m}}{\delta} \right]}. \qquad (21)
\end{aligned}$$

Following the same proof process for bounding $\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_T}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_T(h,h') \right|$, we obtain the following result.

For every $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $T \sim (D_T)^{m'}$, and for every posterior distribution ρ,

$$\left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_T}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_T(h,h') \right| \leq \sqrt{\frac{1}{2m'} \left[\mathrm{KL}(\rho\|\pi) + \ln \frac{2\sqrt{m'}}{\delta} \right]}. \qquad (22)$$

Finally, let us substitute δ by $\frac{\delta}{2}$ in Inequalities (21) and (22). This, together with the union bound that assure that both results hold simultaneously, gives the

result because,

$$\left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\left[R_{D_T}(h,h')-R_{D_S}(h,h')\right]\right| = \mathrm{dis}_\rho(D_S,D_T),$$

$$\left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\left[R_{T}(h,h')-R_{S}(h,h')\right]\right| = \mathrm{dis}_\rho(S,T),$$

and because if $|a_1-b_1|\leq c_1$ and $|a_2-b_2|\leq c_2$, then $|(a_1-a_2)-(b_1-b_2)|\leq c_1'+c_2'$. □

Then we can obtain the following PAC-Bayesian DA-bound.

Theorem 11. *For any domains P_S and P_T (respectively with marginals D_S and D_T) over $X\times Y$, and for any set $\mathcal{H}$ of hypothesis, for any prior distribution π over $\mathcal{H}$, any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of $S_1\sim(D_S)^m$, $S_2\sim(D_S)^{m'}$, and $T\sim(D_T)^{m'}$, for every ρ over $\mathcal{H}$, we have,*

$$R_{P_T}(G_\rho)-R_{P_T}(G_{\rho_T^*})\leq R_S(G_\rho)+\mathrm{dis}_\rho(S,T)+\lambda_\rho + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{4\sqrt{m}}{\delta}\right]} + \sqrt{\frac{1}{2m}\left[2\mathrm{KL}(\rho\|\pi)+\ln\frac{8\sqrt{m}}{\delta}\right]} + \sqrt{\frac{1}{2m'}\left[2\mathrm{KL}(\rho\|\pi)+\ln\frac{8\sqrt{m'}}{\delta}\right]}.$$

where $\lambda_\rho \stackrel{\text{def}}{=} R_{D_T}(G_\rho,G_{\rho_T^*})+R_{D_S}(G_\rho,G_{\rho_T^*})$.

Proof. The result is obtained by inserting Ths. 9 and 10 (with $\delta:=\frac{\delta}{2}$) in Th. 4 of the main paper. □

4. PBDA Algorithm Details

4.1. Objective function and gradient

Given a source sample $S=\{(\mathbf{x}_i^s,y_i^s)\}_{i=1}^m$, a target sample $T=\{(\mathbf{x}_i^t)\}_{i=1}^m$, and fixed parameters $A>0$ and $C>0$, the learning algorithm PBDA consists in finding the weight vector $\mathbf{w}$ minimizing,

$$\frac{\|\mathbf{w}\|^2}{2}+C\sum_{i=1}^m\Phi_{\mathrm{cvx}}\left(y_i^s\frac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) + A\left|\sum_{i=1}^m\Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)-\Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\right|, \quad (23)$$

where, **Erf** being the Gauss error function,

$$\begin{aligned}\Phi(a) &\stackrel{\text{def}}{=} \tfrac{1}{2}\left[1-\mathbf{Erf}\left(\tfrac{a}{\sqrt{2}}\right)\right],\\ \Phi_{\mathrm{cvx}}(a) &\stackrel{\text{def}}{=} \max\left[\Phi(a),\ \tfrac{1}{2}-\tfrac{a}{\sqrt{2\pi}}\right],\\ \Phi_{\mathrm{dis}}(a) &\stackrel{\text{def}}{=} 2\times\Phi(a)\times\Phi(-a).\end{aligned}$$

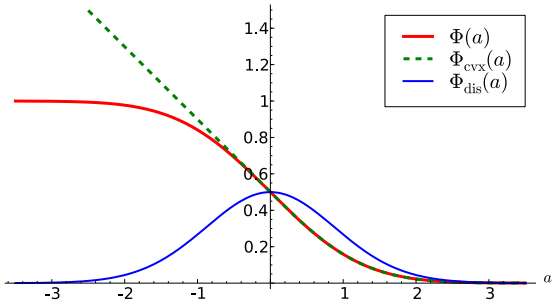

Figure 1. Behaviour of functions $\Phi(\cdot)$, $\Phi_{\mathrm{cvx}}(\cdot)$ and $\Phi_{\mathrm{dis}}(\cdot)$.

Figure 1 illustrates these three functions.

The gradient of the Equation (23) is given by,

$$\mathbf{w}+C\sum_{i=1}^m\Phi'_{\mathrm{cvx}}\left(\tfrac{y_i^s\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)\tfrac{y_i^s\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|} + s\times A\left[\sum_{i=1}^m\Phi'_{\mathrm{dis}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\tfrac{\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}-\Phi'_{\mathrm{dis}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)\tfrac{\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right],$$

where $\Phi'_{\mathrm{cvx}}(a)$ and $\Phi'_{\mathrm{dis}}(a)$ are respectively the derivatives of functions Φ_{cvx} and Φ_{dis} evaluated at point a, and $s=\mathrm{sgn}\left[\sum_{i=1}^m\Phi_{\mathrm{dis}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)-\Phi_{\mathrm{dis}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\right]$.

4.2. Using a kernel function

The kernel trick allows us to work with dual weight vector $\boldsymbol{\alpha}\in\mathbb{R}^{2m}$ that is a linear classifier in an augmented space. Given a kernel $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$, we have,

$$h_{\mathbf{w}}(\mathbf{x}) \;=\; \sum_{i=1}^m\alpha_i k(\mathbf{x}_i^s,\mathbf{x})+\sum_{i=1}^m\alpha_{i+m}k(\mathbf{x}_i^t,\mathbf{x}).$$

Let us denote K the kernel matrix of size $2m\times 2m$ such as,

$$K_{i,j}=k(\mathbf{x}_i,\mathbf{x}_j)$$

where,

$$\mathbf{x}_\# = \begin{cases}\mathbf{x}_\#^s & \text{if } \#\leq m\\ \mathbf{x}_{\#-m}^t & \text{otherwise.}\end{cases}$$

In that case, the objective function of Equation (23) is rewritten in term of the vector $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\dots\alpha_{2m})$ as,

$$\frac{1}{2}\sum_{i=1}^{2m}\sum_{j=1}^{2m}\alpha_i\alpha_jK_{i,j}+C\sum_{i=1}^m\Phi_{\mathrm{cvx}}\left(y_i^s\frac{\sum_{j=1}^{2m}\alpha_jK_{i,j}}{\sqrt{K_{i,i}}}\right) + A\left|\sum_{i=1}^m\Phi_{\mathrm{dis}}\left(\frac{\sum_{j=1}^{2m}\alpha_jK_{i,j}}{\sqrt{K_{i,i}}}\right)-\Phi_{\mathrm{dis}}\left(\frac{\sum_{j=1}^{2m}\alpha_jK_{i+m,j}}{\sqrt{K_{i+m,i+m}}}\right)\right|.$$

The gradient of the latter equation is given by the vector $\boldsymbol{\alpha}' = (\alpha'_1, \alpha'_2, \dots \alpha'_{2m})$, with $\alpha'_{\#}$ equals to,

$$\sum_{j=1}^{2m} \alpha_i K_{i,\#} + C \sum_{i=1}^{m} \Phi_{\text{cvx}}\left(y_i^s \frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right) \frac{y_i^s \, K_{i,\#}}{\sqrt{K_{i,i}}} + s \times A \left[\sum_{i=1}^{m} \Phi_{\text{dis}}\left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right) \frac{K_{i,\#}}{\sqrt{K_{i,i}}} - \Phi_{\text{dis}}\left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i+m,j}}{\sqrt{K_{i+m,i+m}}}\right) \frac{K_{i+m,\#}}{\sqrt{K_{i+m,i+m}}} \right],$$

where,

$$s = \text{sgn}\left[\sum_{i=1}^{m} \Phi_{\text{dis}}\left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right) - \Phi_{\text{dis}}\left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i+m,j}}{\sqrt{K_{i+m,i+m}}}\right) \right].$$

4.3. Implementation details

For our experiments, we minimize the objective function using a *Broyden-Fletcher-Goldfarb-Shanno method (BFGS)* implemented in the *scipy* python library[1]. We made our code available at the following URL:

http://graal.ift.ulaval.ca/pbda/

When selecting hyperparameters by reverse cross-validation, we search on a 20×20 parameter grid for a A between 0.01 and 10^6 and a parameter C between 1.0 and 10^8, both on a logarithm scale.

References

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

McAllester, D. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

Seeger, M. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.

[1] Available at http://www.scipy.org/