PAC-Bayesian Theorems for Domain Adaptation with Specialization to Linear Classifiers

Pascal Germain, Amaury Habrard, François Laviolette, Emilie Morvant

Journal of Machine Learning Research ? (????) ?-? Submitted 10/14; Published ??/??

PAC-Bayesian Theorems 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@UNIV-ST-ETIENNE.FR
Laboratoire Hubert Curien UMR CNRS 5516
Université Jean Monnet, 42000 St-Etienne, France

Editor: ???

Abstract

In this paper, we provide two main contributions in PAC-Bayesian theory for domain adaptation where the objective is to learn, from a source distribution, a well-performing majority vote on a different target distribution. On the one hand, we propose an improvement of the previous approach proposed by Germain et al. (2013), that relies on a novel distribution pseudodistance based on a disagreement averaging, allowing us to derive a new tighter PAC-Bayesian domain adaptation bound for the stochastic Gibbs classifier. 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. On the other hand, we generalize these results to multisource domain adaptation allowing us to take into account different source domains. This study opens the door to tackle domain adaptation tasks by making use of all the PAC-Bayesian tools.

Keywords: Domain Adaptation, PAC-Bayesian Theory, Multisource

1. Introduction

As human beings, we learn from what we saw before. Think about our education process: when a student attends to a new course, he has to make use of the knowledge he acquired during previous courses. However, in machine learning the most commun assumption is based on the fact that the learning and test data are drawn from the same probability distribution. This strong assumption may be clearly irrelevant for a lot of real tasks including those where we desire to adapt a model from one task to another one. 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 or several users 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* 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 domain adaptation is a hard and challenging task even under strong assumptions (Ben-David and Urner, 2012; Ben-David et al., 2010b; Ben-David and Urner, 2014). Note that domain adaptation with learning data coming from different source domains is referred to as multisource or multiple sources domain adaptation (Crammer et al., 2007; Mansour et al., 2009c; Ben-David et al., 2010a).

Among the existing approaches in the literature to address domain adaptation, the instance weighting-based methods allow one to deal with the covariate-shift problem (*e.g.,* Huang et al., 2006; Sugiyama et al., 2008), where source and target domains diverge only in their marginals, *i.e.*, they share the same labeling function. Another technique is to exploit self-labeling procedures, where the objective is to transfer the source labels to the target unlabeled points (*e.g.*, Bruzzone and Marconcini (2010); Habrard et al. (2013); Morvant (2014). A third solution is to learn a new common representation from the unlabeled part of source and target data. Then, a standard supervised learning algorithm can be executed on the source labeled instances (*e.g.*, Glorot et al. (2011); Chen et al. (2012)).

The work presented in this paper stands into a popular class of approaches, that relies on a distance between the source distribution and the target distribution. Such distance depends on the set $\mathcal{H}$ of hypothesis (or classifiers) considered by the learning algorithm. The intuition behind this approach is that one must look for a set $\mathcal{H}$ that minimizes the distance while preserving good performances on the source data; if the distributions are close under this measure, then generalization ability may be "easier" to quantify. In fact, defining such a measure to quantify how much the domains are related is a major issue in domain adaptation. For example, in the context of binary classification with the 0-1 loss function, Ben-David et al. (2010a); and 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 domain adaptation generalization bound based on the VC-dimension theory. The discrepancy distance proposed by 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 and Mohri (2011, 2014) have specialized the minimization of the discrepancy to regression with kernels. In these situations, domain adaptation can be viewed as a multiple trade-off between the complexity of the hypothesis class $\mathcal{H}$, the adaptation ability of $\mathcal{H}$ 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 and Bilmes, 2007) which favors classifiers closer to the best source model. However, a majority of methods prefer to perform a two step approach: *(i)* first construct a suitable representation by minimizing the divergence, then *(ii)* learn a model on the source domain in the new representation space.

1. See the surveys proposed by Jiang (2008); Quionero-Candela et al. (2009); and Margolis (2011).

The novelty of our contribution is to explore the PAC-Bayesian framework to tackle domain adaptation 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. Indeed, based on this disagreement measure, we derived in a previous work (Germain et al., 2013) a first PAC-Bayesian domain adaptation bound expressed as a ρ-averaging. In this paper, we provide a new version of this result, that does not change the philosophy supported by the previous bound, but clearly improves the theoretical result: The domain adaptation bound is now tighter and easier to interpret. Thanks to this new result, we also derive[2] three new PAC-Bayesian domain adaptation generalization bounds. Then, in contrast to the majority of methods that perform a two steps procedure, we design an algorithm tailored to linear classifiers, called PBDA, which jointly minimizes the multiple trade-offs implied by the bounds. 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. Finally, we extend our results to domain adaptation with multiple sources by considering a mixture of different source domains as done by Ben-David et al. (2010a).

The rest of the paper is structured as follows. Section 2 deals with two seminal works on domain adaptation. The PAC-Bayesian framework is then recalled in Section 3. Note that for sake of completeness, we provide for the first time the explicit deviation of the algorithm PBGD3 (Germain et al., 2009a) tailored to linear classifiers in supervised learning. Our main contribution, which consists in a domain adaptation bound suitable for PAC-Bayesian learning, is presented in Section 4. Then, we derive our new algorithm for PAC-Bayesian domain adaptation in Section 5, that we experiment in Section 6. Afterwards, we generalize this analysis to multisource domain adaptation in Section 7. Before concluding in Section 9, we discuss two important points in Section 8: *(i)* two different results for the multisource setting that imply open-questions for deriving new algorithms, *(ii)* the comparison between our new result and the one provided in (Germain et al., 2013).

2. In this paper, we was very keen to improve the readability of our proofs, particularly those provided by Germain et al. (2013) as supplementary material. The proof techniques may be of independent interest.

2. Domain Adaptation Related Works

In this section, we review the two seminal works in domain adaptation that are based on a divergence measure between the domains (Ben-David et al., 2010a; Ben-David et al., 2006; Mansour et al., 2009a).

2.1 Notations and Setting

We consider domain adaptation 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 $X \times Y$ (unknown and fixed), 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* $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ constituted of m examples drawn *i.i.d.*[3] from P_S, and an *unlabeled target sample* $T = \{\mathbf{x}_j^t\}_{j=1}^{m'}$ constituted of m' examples drawn *i.i.d.* from D_T. Note that, we denote the distribution of a m-sample by $(P_S)^m$. We suppose that $\mathcal{H}$ is a set of hypothesis functions for X to Y. The *expected source error* and the *expected target error* of $h \in \mathcal{H}$ over P_S, respectively P_T, are the probability that h errs on the entire distribution P_S, respectively P_T,

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

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 *empirical source error* $R_S(\cdot)$ on the learning sample S is

$$R_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{(\mathbf{x}^s, y^s) \in S} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}^s), y^s\big).$$

The main objective in domain adaptation is then to learn—without target label—a classifier $h \in \mathcal{H}$ leading to the lowest expected target error $R_{P_T}(h)$.

We also introduce the *expected source disagreement* $R_{D_S}(h, h')$ and the *expected target disagreement* $R_{D_T}(h, h')$ of $(h', h) \in \mathcal{H}^2$, which measure the probability that two classifiers h and h' do not agree on the respective marginal distributions, and are defined by

$$R_{D_S}(h, h') \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{x}^s \sim D_S} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}^s), h'(\mathbf{x}^s)\big) \quad \text{and} \quad R_{D_T}(h, h') \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{x}^t \sim D_T} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}^t), h'(\mathbf{x}^t)\big).$$

The *empirical source disagreement* $R_S(h, h')$ on S and the *empirical target disagreements* $R_T(h, h')$ on T are

$$R_S(h, h') \stackrel{\text{def}}{=} \frac{1}{m} \sum_{\mathbf{x}^s \in S} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}^s), h'(\mathbf{x}^s)\big) \quad \text{and} \quad R_T(h, h') \stackrel{\text{def}}{=} \frac{1}{m'} \sum_{\mathbf{x}^t \in T} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}^t), h'(\mathbf{x}^t)\big).$$

Note that, 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 $\{\mathbf{x}_i^s\}_{i=1}^m$.

3. *i.i.d.* stands for *independent and identically distributed.*

Note also that the expected error $R_P(h)$ on a distribution P can be viewed as a shortcut notation for the expected disagreement between a hypothesis h and a labeling function f_P that assigns the true label to an example description according with respect to P. We have

$$R_P(h) \;=\; R_D(h, f_P) \;=\; \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), f_P(\mathbf{x})\big)\,,$$

where D is the marginal distribution of P over X.

2.2 Necessity of a Domain Divergence

The domain adaptation objective is to find a low-error target hypothesis, even if the target labels are not available. Even under strong assumptions, this task can be impossible to solve (Ben-David and Urner, 2012; Ben-David et al., 2010b). However, for deriving generalization ability in a domain adaptation situation (with the help of a domain adaptation 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 have proposed different quantities to estimate how a domain is close to another one (C. Zhang, 2012; Ben-David et al., 2010a; Mansour et al., 2009a,b; Ben-David et al., 2006; Li and Bilmes, 2007). 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 taking 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.

2.3 Domain Adaptation Bounds for Binary Classification

We now review the first two seminal works which propose domain adaptation bounds based on a 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); and Ben-David et al. (2006) have provided the following domain adaptation bound.

Theorem 1 (Ben-David et al. (2010a); Ben-David et al. (2006)) *Let $\mathcal{H}$ be a (symmetric[4]) hypothesis class. We have*

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

where

$$\tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) \;\stackrel{\text{def}}{=}\; \sup_{(h,h')\in\mathcal{H}^2} \big|R_{D_T}(h, h') - R_{D_S}(h, h')\big|$$

is the $\mathcal{H}\Delta\mathcal{H}$-distance between the marginals D_S and D_T, and

$$\mu_{h^*} \;\stackrel{\text{def}}{=}\; R_{P_S}(h^*) + R_{P_T}(h^*)$$

4. In a symmetric hypothesis space $\mathcal{H}$, for every $h \in \mathcal{H}$, its inverse $-h$ is also in $\mathcal{H}$.

is the error of the best hypothesis overall, denoted h^*, and defined by

$$h^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{h\in\mathcal{H}} \left(R_{P_S}(h) + R_{P_T}(h)\right).$$

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 hypotheses 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 term $\mu_{h^*} = R_{P_S}(h^*) + R_{P_T}(h^*)$ is 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^* does not have a good performance on both the source and the target domain, then there is no way one can adapt from this source to this target. Hence, as pointed out by the authors, Equation (1), together with the usual VC-bound theory, express 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 $\mathcal{L} : [-1, +1]^2 \to \mathbb{R}^+$ such a loss, the discrepancy $\text{disc}_{\mathcal{L}}(D_S, D_T)$ between D_S and D_T is

$$\text{disc}_{\mathcal{L}}(D_S, D_T) \stackrel{\text{def}}{=} \sup_{(h,h')\in\mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}^t\sim D_T} \mathcal{L}(h(\mathbf{x}^t), h'(\mathbf{x}^t)) - \mathop{\mathbf{E}}_{\mathbf{x}^s\sim D_S} \mathcal{L}(h(\mathbf{x}^s), h'(\mathbf{x}^s)) \right|.$$

Note that with the 0-1 loss in binary classification, we have

$$\tfrac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) = \text{disc}_{\mathcal{L}_{0\text{-}1}}(D_S, D_T).$$

Even if these two divergences may coincide, the following domain adaptation bound of Mansour et al. (2009a) differs from Theorem 1.

Theorem 2 (Mansour et al. (2009a)) *Let $\mathcal{H}$ be a (symmetric) hypothesis class. We have*

$$\forall h \in \mathcal{H},\ R_{P_T}(h) - R_{P_T}(h_T^*) \leq R_{D_S}(h_S^*, h) + \text{disc}_{\mathcal{L}_{0\text{-}1}}(D_S, D_T) + \nu_{(h_S^*, h_T^*)}, \qquad (2)$$

where

$$\nu_{(h_S^*, h_T^*)} \stackrel{\text{def}}{=} R_{D_S}(h_S^*, h_T^*)$$

is the disagreement between the ideal hypothesis on the target and source domains defined respectively as

$$h_T^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{h\in\mathcal{H}} R_{P_T}(h), \quad \textit{and} \quad h_S^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{h\in\mathcal{H}} R_{P_S}(h).$$

In this context, Equation (2) can be tighter[5] 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

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

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 domain adaptation 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 in Section 4 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; Germain et al., 2009a; Parrado-Hernández et al., 2012).

3. PAC-Bayesian Theory in Supervised Learning

Let us now review the classical supervised binary classification framework called the PAC-Bayesian theory, first introduced by McAllester (1999). This theory succeeds to provide tight generalization guarantees on majority vote classifiers, without relying on any validation set.

Throughout this section, we adopt an algorithm design perspective: we interpret the various forms of the PAC-Bayesian theorem as a guide to derive new machine learning algorithms. Indeed, the PAC-Bayesian analysis of domain adaptation provided in the forthcoming sections is oriented by the motivation of creating a new adaptive algorithms.

3.1 Notations and Setting

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 $R_{P_S}(B_\rho)$ 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:

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

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

$$R_{P_S}(B_\rho) \;\leq\; 2\,R_{P_S}(G_\rho)\,.$$

Another result on the relation between $R_{P_S}(B_\rho)$ and $R_{P_S}(G_\rho)$ is the C-bound of Lacasse et al. (2006) expressed as

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

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

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

Equation (4) suggests that for a fixed numerator, *i.e.*, a fixed risk of the Gibbs classifier, 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).

Finally, we introduce the notion of *expected joint error* of a pair of classifiers (h, h') drawn according to the distribution ρ, defined as

$$e_{P_S}(G_\rho, G_\rho) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), y\big) \times \mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}), y\big)\,. \tag{6}$$

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 drawn *i.i.d.* from P_S and the Kullback-Leibler divergence $\mathrm{KL}(\rho\|\pi) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)}$. The three main PAC-Bayes theorems, that we present in the next section, have been proposed by McAllester (1999); Seeger (2002); Langford (2005); and Catoni (2007).

3.2 Three Versions of the PAC-Bayesian Theorem

First, let consider the KL-divergence $\mathrm{kl}(a\,\|\,b)$ between two Bernoulli distributions with success probability a and b, defined by

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

Seeger (2002); and Langford (2005) have derived the following PAC-Bayesian theorem in which the trade-off between the complexity and the risk is handled by $\mathrm{kl}(\cdot\|\cdot)$.

Theorem 3 (Seeger (2002); Langford (2005)) *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*

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

This version of the PAC-Bayes theorem offers a tight bound, especially for low empirical risk. However, due to the $\mathrm{kl}\left(R_S(G_\rho)\,\|\,R_{P_S}(G_\rho)\right)$ term, this bound remains difficult to interpret: the link between the empirical risk $R_S(G_\rho)$ and the "true" risk $R_{P_S}(G_\rho)$ is not given by a close form. Thus, from an algorithmic point of view, finding the distribution ρ that minimizes the bound on $R_{P_S}(G_\rho)$ given by Theorem 3 might be a difficult task.

The following version of the PAC-Bayes theorem, which was the first proposed (McAllester, 1999), appears easier to interpret since it links the terms $R_S(G_\rho)$ and $R_{P_S}(G_\rho)$ by a linear relation. Note that Theorem 4 can be straightforwardly obtained from Theorem 3 using Pinsker's inequality:

$$2(q-p)^2 \;\leq\; \mathrm{kl}(q\,\|\,p)\,. \tag{7}$$

Theorem 4 (McAllester (1999)) *For any domain P_S over $X \times Y$, any set of hypothesis $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, and 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*

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

Theorems 3 and 4 suggest that, in order to minimize the expected risk, a learning algorithm should performs a trade-off between the empirical risk minimization $R_S(G_\rho)$ and KL-divergence minimization $\mathrm{KL}(\rho \,\|\, \pi)$ (roughly speaking the complexity term).

The nature of this trade-off can be explicitly controlled in Theorem 5 below. This PAC-Bayesian result, first proposed by Catoni (2007), is defined with a hyperparameter (here named c). It appears to be a natural tool to design PAC-Bayesian algorithms. We present this result in the simplified form suggested by Germain et al. (2009b).

Theorem 5 (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*

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

The bound given by Theorem 5 has two interesting characteristics. First, choosing $c = \frac{1}{\sqrt{m}}$, the bound becomes consistent: it converges to $1 \times [R_S(G_\rho) + 0]$ as m grows. Second, as described in Section 3.3, 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). Hence, the value c allows us to control the trade-off between the empirical risk $R_S(G_\rho)$ and the complexity term $\frac{1}{m} \mathrm{KL}(\rho \| \pi)$.

3.3 Supervised PAC-Bayesian Learning of Linear Classifiers

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

$$h_{\mathbf{w}'}(\mathbf{x}) \stackrel{\text{def}}{=} \mathrm{sgn}\left(\mathbf{w}' \cdot \mathbf{x}\right),$$

where $\cdot$ denotes the dot product.

By restricting the prior and the posterior distributions over $\mathcal{H}$ to be Gaussian distributions, Langford and Shawe-Taylor (2002); Ambroladze et al. (2006); and Parrado-Hernández et al. (2012) have specialized the PAC-Bayesian theory in order to bound the expected risk of any linear classifier $h_{\mathbf{w}} \in \mathcal{H}$. 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 $h_{\mathbf{w}'} \in \mathcal{H}$, we have

$$\pi_{\mathbf{0}}(h_{\mathbf{w}'}) \stackrel{\text{def}}{=} \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left(-\frac{1}{2}\|\mathbf{w}'\|^2\right),$$

$$\text{and} \quad \rho_{\mathbf{w}}(h_{\mathbf{w}'}) \stackrel{\text{def}}{=} \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left(-\frac{1}{2}\|\mathbf{w}' - \mathbf{w}\|^2\right).$$

An interesting property of these Gaussian distributions is that the prediction of the $\rho_{\mathbf{w}}$-weighted majority vote $B_{\rho_{\mathbf{w}}}(\cdot)$ coincides with the one of the linear classifier $h_{\mathbf{w}}(\cdot)$. Indeed, we have

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

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

$$\begin{aligned}R_{P_S}(G_{\rho_{\mathbf{w}}}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \mathop{\mathbf{E}}_{h_{\mathbf{w}'}\sim\rho_{\mathbf{w}}} \mathcal{L}_{0\text{-}1}\left(h_{\mathbf{w}'}(\mathbf{x}), y\right) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \mathop{\mathbf{E}}_{h_{\mathbf{w}'}\sim\rho_{\mathbf{w}}} \mathbf{I}\left(h_{\mathbf{w}'}(\mathbf{x}) \neq y\right) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \mathop{\mathbf{E}}_{h_{\mathbf{w}'}\sim\rho_{\mathbf{w}}} \mathbf{I}\left(y\, \mathbf{w}'\cdot\mathbf{x} \leq 0\right) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \frac{1}{\sqrt{2\boldsymbol{\pi}}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\|\mathbf{w}'-\mathbf{w}\|^2\right) \mathbf{I}\left(y\, \mathbf{w}'\cdot\mathbf{x} \leq 0\right) d\,\mathbf{w}' \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \left[1 - \Pr_{t\sim\mathcal{N}(0,1)}\left(t \ \leq\ y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right)\right] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \Phi\left(y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right),\end{aligned}$$

where we defined

$$\Phi(a) \ \stackrel{\text{def}}{=}\ \frac{1}{2}\left[1 - \mathbf{Erf}\left(\frac{a}{\sqrt{2}}\right)\right],$$

with $\mathbf{Erf}(\cdot)$ is the Gauss error function defined as

$$\mathbf{Erf}\,(b) \ \stackrel{\text{def}}{=}\ \frac{2}{\sqrt{\boldsymbol{\pi}}} \int_0^b \exp\left(-t^2\right) \mathrm{d}t\,. \tag{8}$$

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

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

3.3.1 Objective Function and Gradient

Based on the specialization of the PAC-Bayesian theory to linear classifiers, Germain et al. (2009a) suggested to minimize a PAC-Bayesian bound on $R_{P_S}(G_{\rho_{\mathbf{w}}})$. For sake of completeness, we provide here more mathematical details than in the original conference paper (Germain et al., 2009a). We will build on this PAC-Bayesian learning algorithm (for supervised leaning) in our domain adaptation work.

Given a sample $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ and a hyperparameter $C > 0$, the learning algorithm performs a gradient descent in order to find an optimal weight vector $\mathbf{w}$ that minimizes

$$\begin{aligned}F(\mathbf{w}) &= CmR_S(G_{\rho_{\mathbf{w}}}) + \mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}}) \\ &= C\sum_{i=1}^m \Phi\left(y_i\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right) + \frac{1}{2}\|\mathbf{w}\|^2\,.\end{aligned} \tag{9}$$

It turns out that the optimal vector $\mathbf{w}$ corresponds to the distribution $\rho_{\mathbf{w}}$ that minimizes the value of the bound on $R_{P_S}(G_{\rho_{\mathbf{w}}})$ given by Theorem 5, with the parameter c of the theorem being the hyperparameter C of the learning algorithm. It is important to point out that PAC-Bayesian theorems bound simultaneously $R_{P_S}(G_{\rho_{\mathbf{w}}})$ *for every* $\rho_{\mathbf{w}}$ *on* $\mathcal{H}$. Therefore, one can "freely" explore the domain of objective function F to choose a posterior distribution $\rho_{\mathbf{w}}$ that gives, thanks to Theorem 5, a bound valid with probability $1-\delta$.

The minimization of Equation (9) by gradient descent corresponds to the learning algorithm called PBGD3 of Germain et al. (2009a). The gradient of $F(\mathbf{w})$ is given the vector $\nabla F(\mathbf{w})$:

$$\nabla F(\mathbf{w}) \;=\; C\sum_{i=1}^{m}\Phi'\left(y_i\frac{\mathbf{w}\cdot\mathbf{x}_i}{\|\mathbf{x}_i\|}\right)\frac{y_i\,\mathbf{x}_i}{\|\mathbf{x}_i\|}+\mathbf{w}\,,$$

where $\Phi'(a)=-\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}a^2\right)$ is the derivative of $\Phi(\cdot)$ at point a.

Similarly to the SVM, the learning algorithm PBGD3 realizes a trade-off between the empirical risk (expressed by the loss $\Phi(\cdot)$) and the complexity of the learned linear classifier (expressed by the regularizer $\|\mathbf{w}\|^2$). This similarity increases when we use a kernel function, as described next.

3.3.2 Using a kernel function

The kernel trick allows to substitute inner products by a kernel function $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ in Equation (9). If k is a Mercer kernel, it implicitly represents a function $\phi:X\to\mathbb{R}^{d'}$ that maps an example of X into an arbitrary d'-dimensional space[6], such that

$$\forall(\mathbf{x},\mathbf{x}')\in X^2,\quad k(\mathbf{x},\mathbf{x}') \;=\; \phi(\mathbf{x})\cdot\phi(\mathbf{x}')\,.$$

Then, a dual weight vector $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_m)\in\mathbb{R}^m$ encodes the linear classifier $\mathbf{w}\in\mathbb{R}^{d'}$ as a linear combination of examples of S:

$$\mathbf{w} \;=\; \sum_{i=1}^{m}\alpha_i\,\phi(\mathbf{x}_i)\,,\qquad\text{and thus}\qquad h_{\mathbf{w}}(\mathbf{x}) \;=\; \operatorname{sgn}\left[\sum_{i=1}^{m}\alpha_i k(\mathbf{x}_i,\mathbf{x})\right].$$

By the representer theorem (Schölkopf et al., 2001), the vector $\mathbf{w}$ minimizing Equation (9) can be recovered by finding the vector $\boldsymbol{\alpha}$ that minimizes

$$F(\boldsymbol{\alpha}) \;=\; C\sum_{i=1}^{m}\Phi\left(y_i\frac{\sum_{j=1}^{m}\alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right)+\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j K_{i,j}\,, \tag{10}$$

where K is the kernel matrix of size $m\times m$. That is, $K_{i,j}\stackrel{\text{def}}{=}k(\mathbf{x}_i,\mathbf{x}_j)$. The gradient of $F(\boldsymbol{\alpha})$ is simply given the vector $\nabla F(\boldsymbol{\alpha})=(\alpha'_1,\alpha'_2,\ldots\alpha'_m)$, with

$$\alpha'_{\#} \;=\; C\sum_{i=1}^{m}\Phi\left(y_i\frac{\sum_{j=1}^{m}\alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right)\frac{y_i\,K_{i,\#}}{\sqrt{K_{i,i}}}+\sum_{j=1}^{m}\alpha_i K_{i,\#}\,,\qquad\text{for }\#\in\{1,2,\ldots,m\}\,.$$

6. We consider here that the induced space is finite-dimensional.

3.3.3 Improving the Algorithm Using a Convex Objective

An annoying 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 after the ones described by Germain et al. (2009a) and saw that PBGD3 achieves an equivalent accuracy (and at a fraction of the running time) by replacing the loss function $\Phi(\cdot)$ of Equations (9) and (10) by its convex relaxation, which is

$$\Phi_{\text{cvx}}(a) \stackrel{\text{def}}{=} \max\left\{\Phi(a), \frac{1}{2} - \frac{a}{\sqrt{2\pi}}\right\} = \begin{cases} \frac{1}{2} - \frac{a}{\sqrt{2\pi}} & \text{if } a \leq 0, \\ \Phi(a) & \text{otherwise.} \end{cases}$$

The derivative of $\Phi_{\text{cvx}}(\cdot)$ at point a is then $\Phi'_{\text{cvx}}(a) = \frac{-1}{\sqrt{2\pi}}$ if $a < 0$, and $\Phi'(a)$ otherwise. Note that Figure 1 in Section 5 illustrates the functions $\Phi(\cdot)$ and $\Phi_{\text{cvx}}(\cdot)$.

In the following we present our contributions on PAC-Bayesian domain adaptation.

4. PAC-Bayesian Theorems for Domain Adaptation

The originality of our contribution is to theoretically design a domain adaptation framework for PAC-Bayesian approach. In Section 4.1, we propose a domain comparison pseudometric suitable in this context. We then derive PAC-Bayesian domain adaptation bounds in Section 4.2, that improves the result proposed in Germain et al. (2013). Finally, note that in Section 5 we see that using the previous approach in a domain adaptation way is a relevant strategy: we specialize our result to linear classifiers.

4.1 A Domain Divergence for PAC-Bayesian Analysis

In the following, while the domain adaptation bounds presented in Section 2 focus on a single classifier, we first define a ρ-average disagreement measure to compare the marginals. Then, this leads us to derive our domain adaptation bound suitable for the PAC-Bayesian approach.

As discussed in Section 2.2, the derivation of generalization ability in domain adaptation critically needs a divergence measure between the source and target marginals.

4.1.1 Designing the Divergence

We define a *domain disagreement pseudometric*[7] 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 the C-bound presented in 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

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

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 6 *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*

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

Note that $\mathrm{dis}_\rho(\cdot,\cdot)$ is symmetric and fulfills the triangle inequality.

4.1.2 Comparison of the $\mathcal{H}\Delta\mathcal{H}$-divergence and our domain disagreement

While the $\mathcal{H}\Delta\mathcal{H}$-divergence of Theorem 1 is difficult to jointly optimize with the empirical source error, our empirical disagreement measure is easier to manipulate: we simply need to compute the ρ-average of the classifiers disagreement instead of finding the pair of classifiers that maximizes the disagreement. Indeed, $\mathrm{dis}_\rho(\cdot,\cdot)$ 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 reweighing, space representation changing or family of classifiers modification. On the contrary, $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\cdot,\cdot)$ is a supremum over all $h \in \mathcal{H}$ and hence, does not depend on the h on which the risk is considered. Moreover, $\mathrm{dis}_\rho(\cdot,\cdot)$ (the ρ-average) is lower than the $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\cdot,\cdot)$ (the worst case). Indeed, for every $\mathcal{H}$ and ρ over $\mathcal{H}$, we have

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

4.1.3 PAC-Bayesian bounds for our domain disagreement

The following theorems show 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, let first suppose that $m = m'$, *i.e.*, the size of S and T are equal. 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*

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

Proof Deferred to Appendix B. ■

Here is a "McAllester's type" PAC-Bayesian bound for our domain disagreement dis_ρ obtained straightforwardly from Theorem 7.

Corollary 8 *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*

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

Proof The result is obtained by using Pinsker's inequality (Equation (7)) on Theorem 7. ■

Here is a "Catoni's type" PAC-Bayesian bound which helps us to derive a domain adaptation algorithm in the following.

Theorem 9 *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*

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

Proof Deferred to Appendix C. ■

Similarly to the empirical risk bound of Catoni (2007) shown by Theorem 5, 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.

The last result of this section tackles the situation where $m \neq m'$, *i.e.*, the sizes of S and T are different.

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*

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

Proof Deferred to Appendix D. ■

Note that Theorem 10 is very similar to the result of Corollary 8. In fact, in the particular case $m = m'$, Theorem 10 differs from Corollary 8 only by the $4\sqrt{m}$ term inside the logarithm, instead of $2\sqrt{m}$.

4.2 PAC-Bayesian Theorems for Domain Adaptation

We now derive our main result in the following theorem: a domain adaptation bound relevant in a PAC-Bayesian setting.

4.2.1 A domain adaptation bound for the stochastic Gibbs classifier

Theorem 11 below relies on the domain disagreement of Definition 6, and also on *expected joint error* of Equation (6).

Theorem 11 *Let $\mathcal{H}$ be a hypothesis class. We have*

$$\forall \rho \text{ on } \mathcal{H},\ R_{P_T}(G_\rho) \ \leq\ R_{P_S}(G_\rho) + \frac{1}{2}\,\mathrm{dis}_\rho(D_S, D_T) + \lambda_\rho\,,$$

where λ_ρ is the deviation between the expected joint errors of G_ρ on the target and source domains:

$$\begin{aligned}\lambda_\rho \ &\stackrel{\text{def}}{=}\ \left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_T}\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}),y\big)\mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}),y\big) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S}\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}),y\big)\mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}),y\big)\right]\right| \\ &=\ \Big|\,e_{P_T}(G_\rho,G_\rho) - e_{P_S}(G_\rho,G_\rho)\,\Big|\,. \qquad (11)\end{aligned}$$

Proof First, notice that for any distribution P on $X\times Y$ (and corresponding marginal distribution D on X), we have

$$R_P(G_\rho) \ =\ \frac{1}{2}R_D(G_\rho,G_\rho) + e_P(G_\rho,G_\rho)\,, \qquad (12)$$

as

$$\begin{aligned}2\,R_P(G_\rho) \ &=\ \mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\Big[\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}),y\big) + \mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}),y\big)\Big] \\ &=\ \mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\Big[1\times\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}),h'(\mathbf{x})\big) + 2\times\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}),y\big)\,\mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}),y\big)\Big] \\ &=\ R_D(G_\rho,G_\rho) + 2\times e_P(G_\rho,G_\rho)\,.\end{aligned}$$

Therefore,

$$\begin{aligned}R_{P_T}(G_\rho) - R_{P_S}(G_\rho) \ &=\ \frac{1}{2}\Big(R_{D_T}(G_\rho,G_\rho) - R_{D_S}(G_\rho,G_\rho)\Big) + \Big(e_{P_T}(G_\rho,G_\rho) - e_{P_S}(G_\rho,G_\rho)\Big) \\ &\leq\ \frac{1}{2}\Big|R_{D_T}(G_\rho,G_\rho) - R_{D_S}(G_\rho,G_\rho)\Big| + \Big|e_{P_T}(G_\rho,G_\rho) - e_{P_S}(G_\rho,G_\rho)\Big| \\ &=\ \frac{1}{2}\,\mathrm{dis}_\rho(D_S,D_T) + \lambda_\rho\,.\end{aligned}$$

■

Our bound is, in general, incomparable with the ones of Theorems 1 and 2. It can be seen as a trade-off between different quantities. The terms $R_{P_S}(G_\rho)$ and $\mathrm{dis}_\rho(D_S,D_T)$ are similar to the first two terms of the domain adaptation 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 term λ_ρ measures the deviation between the expected joint target and source errors of G_ρ. According to this theory, a good domain adaptation is possible if this deviation is low. However, since we suppose that we do not have any label in the target sample, we cannot control or estimate it. In practice, we suppose that λ_ρ is low and we neglect it. In other words, we assume that the labeling information between the two domains is related and that considering only the marginal agreement and the source labels is sufficient to find a good majority vote. Another important point comes from the fact that this bound is not degenerated when the source and target distributions are the same or close, see Section 8.2 for a discussion on this point.

We provide in the next section three PAC-Bayesian theorems that justifies the empirical optimization of the bound of Theorem 11.

4.2.2 PAC-Bayesian theorems for domain adaptation

Finally, our Theorem 11 leads to a PAC-Bayesian bounds based on both the empirical source error of the Gibbs classifier and the empirical domain disagreement pseudometric estimated on a source and target samples.

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

Theorem 12 *For any domains P_S and P_T (respectively with marginals D_S and D_T) over $X \times Y$, any set of hypothesis $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, and 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*

$$R_{P_T}(G_\rho) \;\leq\; \sup \mathcal{R}_\rho + \tfrac{1}{2} \sup \mathcal{D}_\rho + \lambda_\rho \,,$$

where λ_ρ is defined by Equation (11), and

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

Proof The result is obtained by inserting Theorems 3 and 7 (with $\delta := \frac{\delta}{2}$) in Theorem 11. ■

The following bound is based on the Catoni's approach and corresponds to the one from which we derive—in Section 5—our algorithm for PAC-Bayesian domain adaptation.

Theorem 13 *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 $S \times T \sim (P_S \times D_T)^m$, for every posterior distribution ρ on $\mathcal{H}$, we have*

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

where λ_ρ is defined by Equation (11), and where $c' \stackrel{\text{def}}{=} \dfrac{c}{1-e^{-c}}$, *and* $\alpha' \stackrel{\text{def}}{=} \dfrac{2\alpha}{1-e^{-2\alpha}}$.

Proof In Theorem 11, we replace $R_S(G_\rho)$ and $\mathrm{dis}_\rho(S,T)$ by their upper bound, obtained from Theorem 5 and Theorem 9, with δ chosen respectively as $\frac{\delta}{3}$ and $\frac{2\delta}{3}$. In the latter case, we use

$$\begin{aligned} 2\,\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2}{2\delta/3} &= 2\,\mathrm{KL}(\rho\|\pi) + \ln\tfrac{3}{\delta} \\ &< 2\left(\mathrm{KL}(\rho\|\pi) + \ln\tfrac{3}{\delta}\right) . \end{aligned}$$

■

We now present a result based on the McAllester's bound, which allows us to easily deal with different size of samples.

Theorem 14 *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 $S_1 \sim (P_S)^{m_1}$, $S_2 \sim (D_S)^{m_2}$, and $T \sim (D_T)^{m'}$, for every ρ over $\mathcal{H}$, we have*

$$\begin{aligned} R_{P_T}(G_\rho) \leq\ & R_{S_1}(G_\rho) + \tfrac{1}{2}\,\mathrm{dis}_\rho(S_2,T) + \lambda_\rho \\ & + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m_1}}{\delta}}{2m_1}} + \sqrt{\frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{8\sqrt{m_2}}{\delta}}{8m_2}} + \sqrt{\frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{8\sqrt{m'}}{\delta}}{8m'}} . \end{aligned}$$

where λ_ρ is defined by Equation (11).

Proof We insert Theorems 4 and 10 (with $\delta := \frac{\delta}{2}$) in Theorem 11. ■

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 domain adaptation algorithm without target label is to minimize the bound of Theorem 13 by disregarding λ_ρ. Notice that a major advantage of our domain adaptation 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

In this section, we design a learning algorithm for domain adaptation inspired by the PAC-Bayesian learning algorithm of Germain et al. (2009a). That is, we adopt the specialization of the PAC-Bayesian theory to linear classifier described in Section 3.3. Note that the code of our algorithm is available on-line.[8]

5.1 Minimizing the PAC-Bayesian Domain Adaptation Bound

Let consider a prior $\pi_{\mathbf{0}}$ and a posterior $\rho_{\mathbf{w}}$ that are spherical Gaussian distributions over a space of linear classifiers, exactly as defined in Section 3.3.

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

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 of the bound given by Theorem 13. We work under the assumption that the term $\lambda_{\rho_\mathbf{w}}$ of the bound is negligible. Thus, the posterior distribution $\rho_\mathbf{w}$ that minimizes the bound on $R_T(G_{\rho_\mathbf{w}})$ is the same that minimizes

$$C\, m\, R_S(G_{\rho_\mathbf{w}}) + A\, m\ \mathrm{dis}_{\rho_\mathbf{w}}(S,T) + \mathrm{KL}(\rho_\mathbf{w} \| \pi_\mathbf{0})\,. \tag{13}$$

The values $A > 0$ and $C > 0$ are hyperparameters of the algorithm. Note that the constants α and c of Theorem 13 can be recovered from any A and C.

5.1.1 Domain Disagreement of Linear Classifiers

We know from Equation (9) how to compute the terms $R_S(G_{\rho_\mathbf{w}})$ and $\mathrm{KL}(\rho_\mathbf{w}\|\pi_\mathbf{0})$ of Equation (13). Let now derive the value of $\mathrm{dis}_{\rho_\mathbf{w}}(S,T)$, *i.e.*, the empirical domain disagreement between S and T of a distribution $\rho_\mathbf{w}$ over linear classifiers.

First, for any marginal D, we obtain

$$\begin{aligned}
R_D(G_{\rho_\mathbf{w}}, G_{\rho_\mathbf{w}}) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\ \mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), h'(\mathbf{x})\big) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\ \mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \mathbf{I}[h(\mathbf{x}) \neq h'(\mathbf{x})] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\ \mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \Big(\mathbf{I}[h(\mathbf{x}) = 1]\,\mathbf{I}[h'(\mathbf{x}) = -1] + \mathbf{I}[h(\mathbf{x}) = -1]\,\mathbf{I}[h'(\mathbf{x}) = 1]\Big) \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\ \mathop{\mathbf{E}}_{(h,h')\sim\rho_\mathbf{w}^2} \mathbf{I}[h(\mathbf{x}) = 1]\,\mathbf{I}[h'(\mathbf{x}) = -1] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\ \mathop{\mathbf{E}}_{h\sim\rho_\mathbf{w}} \mathbf{I}[h(\mathbf{x}) = 1] \mathop{\mathbf{E}}_{h'\sim\rho_\mathbf{w}} \mathbf{I}[h'(\mathbf{x}) = -1] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \Phi\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right) \Phi\left(-\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{dis}_{\rho_\mathbf{w}}(S,T) &= \Big| R_S(G_{\rho_\mathbf{w}}, G_{\rho_\mathbf{w}}) - R_T(G_{\rho_\mathbf{w}}, G_{\rho_\mathbf{w}}) \Big| \\
&= \left| \frac{1}{m}\sum_{i=1}^m \Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) - \frac{1}{m}\sum_{i=1}^m \Phi_{\mathrm{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right) \right|,
\end{aligned}$$

where

$$\Phi_{\mathrm{dis}}(a) \stackrel{\mathrm{def}}{=} 2\,\Phi(a)\,\Phi(-a)\,. \tag{14}$$

5.1.2 Objective Function and Gradient

From the results of Sections 3.3.1 and 5.1.1, we obtain that Equation (13) equals to

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

which is highly non-convex. To make the optimization problem more tractable, we replace the loss function $\Phi(\cdot)$ by its convex relaxation $\Phi_{\mathrm{cvx}}(\cdot)$ (as in Section 3.3.3) and minimize the

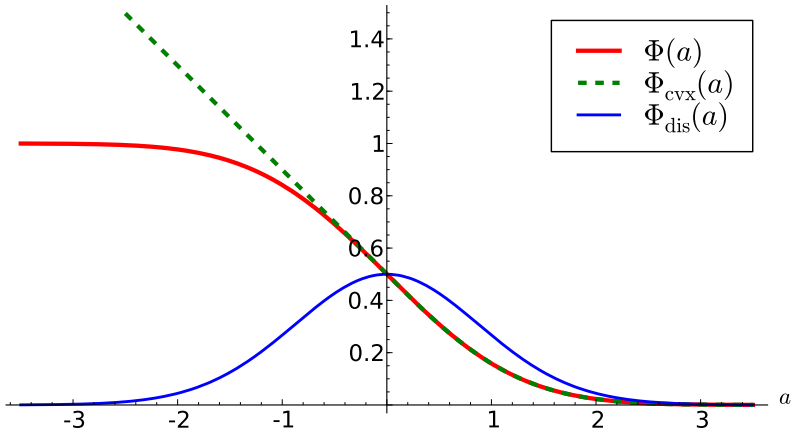

Figure 1: Behaviour of functions $\Phi(\cdot)$, $\Phi_{\text{cvx}}(\cdot)$ and $\Phi_{\text{dis}}(\cdot)$.

resulting cost function by gradient descent. Even if this optimization task is still not convex ($\Phi_{\text{dis}}(\cdot)$ is quasiconcave), our empirical study shows no need to perform many restarts to find a suitable solution.[9]

We name this domain adaptation algorithm PBDA. To sum up, 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 hyperparameters A and C, the algorithm PBDA performs gradient descent to minimize the following objective function:

$$G(\mathbf{w}) = C \sum_{i=1}^{m} \Phi_{\text{cvx}}\left(y_i^s \frac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) + A\left|\sum_{i=1}^{m}\left[\Phi_{\text{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) - \Phi_{\text{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\right]\right| + \frac{1}{2}\|\mathbf{w}\|^2 . \quad (15)$$

where

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

with $\mathbf{Erf}(\cdot)$ the Gauss error function defined in Equation (8). Figure 1 illustrates these three functions.

The gradient $\nabla G(\mathbf{w})$ of the Equation (15) is then given by

$$\begin{aligned}
\nabla G(\mathbf{w}) &= C \sum_{i=1}^{m} \Phi'_{\text{cvx}}\left(\frac{y_i^s \mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)\frac{y_i^s \mathbf{x}_i^s}{\|\mathbf{x}_i^s\|} + \mathbf{w} \\
&\quad + s \times A\left(\sum_{i=1}^{m}\left[\Phi'_{\text{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\frac{\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|} - \Phi'_{\text{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)\frac{\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right]\right),
\end{aligned}$$

9. We observe empirically that a good strategy is to first find the vector $\mathbf{w}$ minimizing the convex problem of PBGD3 described in Section 3.3.3, and then use this $\mathbf{w}$ as a starting point for the gradient descent of PBDA.

where $\Phi'_{\text{cvx}}(a)$ and $\Phi'_{\text{dis}}(a)$ are respectively the derivatives of functions $\Phi_{\text{cvx}}(\cdot)$ and $\Phi_{\text{dis}}(\cdot)$ evaluated at point a, and

$$s = \text{sgn}\left(\sum_{i=1}^{m} \left[\Phi_{\text{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|} \right) - \Phi_{\text{dis}}\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|} \right) \right] \right) .$$

We extend these equations to kernels in the following subsection.

5.1.3 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

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

Let us denote K the kernel matrix of size $2m \times 2m$ such as $K_{i,j} \stackrel{\text{def}}{=} k(\mathbf{x}_i, \mathbf{x}_j)$, 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 (15) is rewritten in term of the vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots \alpha_{2m})$ as

$$\begin{aligned} G(\boldsymbol{\alpha}) = C \sum_{i=1}^{m} \Phi_{\text{cvx}} \left(y_i^s \frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) \\ + A \left| \sum_{i=1}^{m} \left[\Phi_{\text{dis}} \left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) - \Phi_{\text{dis}} \left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i+m,j}}{\sqrt{K_{i+m,i+m}}} \right) \right] \right| + \frac{1}{2} \sum_{i=1}^{2m} \sum_{j=1}^{2m} \alpha_i \alpha_j K_{i,j} \, . \end{aligned}$$

The gradient of the latter equation is given by the vector $\nabla G(\boldsymbol{\alpha}) = (\alpha'_1, \alpha'_2, \ldots \alpha'_{2m})$, with

$$\begin{aligned} \alpha'_{\#} \;&= C \sum_{i=1}^{m} \Phi'_{\text{cvx}} \left(y_i^s \frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) \frac{y_i^s K_{i,\#}}{\sqrt{K_{i,i}}} + \sum_{j=1}^{2m} \alpha_i K_{i,\#} \\ &\quad + s \times A \left(\sum_{i=1}^{m} \left[\Phi'_{\text{dis}} \left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) \frac{K_{i,\#}}{\sqrt{K_{i,i}}} - \Phi'_{\text{dis}} \left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i+m,j}}{\sqrt{K_{i+m,i+m}}} \right) \frac{K_{i+m,\#}}{\sqrt{K_{i+m,i+m}}} \right] \right) , \end{aligned}$$

where

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

6. Experiments

6.1 General Setup

PBDA[10] has been evaluated on a toy problem and a sentiment dataset. For our experiments, we minimize the objective function using a *Broyden-Fletcher-Goldfarb-Shanno method (BFGS)* implemented in the *scipy* python library[11]. PBDA has been compared with:

- SVM learned only from the source domain, *i.e.*, without adaptation. We made use of the SVM-light library (Joachims, 1999).
- PBGD3, presented in Section 3.3, and learned only from the source domain, *i.e.*, without adaptation.
- DASVM of Bruzzone and Marconcini (2010), an iterative domain adaptation algorithm which tries to maximize iteratively a notion of margin on self-labeled target examples. We implemented DASVM with the LibSVM library (Chang and Lin, 2001).
- CODA of Chen et al. (2011), a co-training domain adaptation algorithm, which looks iteratively for target features related to the training set. We used the implementation provided by the authors. Note that in (Chen et al., 2011) has showed best results on the dataset considered in our Section 6.4.

Each parameter is 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/circular validation (^{RCV}) on the source and the (unlabeled) target samples for CODA, DASVM, and PBDA. We describe this latter point in the following section. Note that for PBDA 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.

6.2 A Note about the Reverse Validation

A crucial question in domain adaptation is the validation of the hyperparameters. One solution is to follow the principle proposed by Zhong et al. (2010) which relies on the use of a reverse validation approach. This approach is based on a so-called reverse classifier evaluated on the source domain. We propose to follow it for tuning the parameters of PBDA, DASVM and CODA. Note that Bruzzone and Marconcini (2010) have proposed a similar method, called circular validation, in the context of DASVM.

Concretely, in our setting, given k-folds on the source labeled sample ($S = S_1 \cup \ldots \cup S_k$), k-folds on the unlabeled target T sample ($T = T_1 \cup \ldots \cup T_k$) and a learning algorithm (parametrized by a fixed tuple of hyperparameters), the reverse cross validation risk on the i^{th} fold is computed as follow. Firstly, the source set $S \setminus S_i$ is used as labeled examples and the target set $T \setminus T_i$ is used as unlabeled sample for learning a classifier h'. Secondly, using the same algorithm, a reverse classifier h'^r is learned using the *self-labeled* sample $\{(\mathbf{x}, h'(\mathbf{x}))\}_{\mathbf{x} \in T \setminus T_i}$ as the source set and the unlabeled part of $S \setminus S_i$ as target sample. Finally, the reverse classifier h'^r is evaluated on S_i. We summarize this principle on Figure 2. The

10. We made our code available at the following URL: `http://graal.ift.ulaval.ca/pbda/`
11. Available at `http://www.scipy.org/`

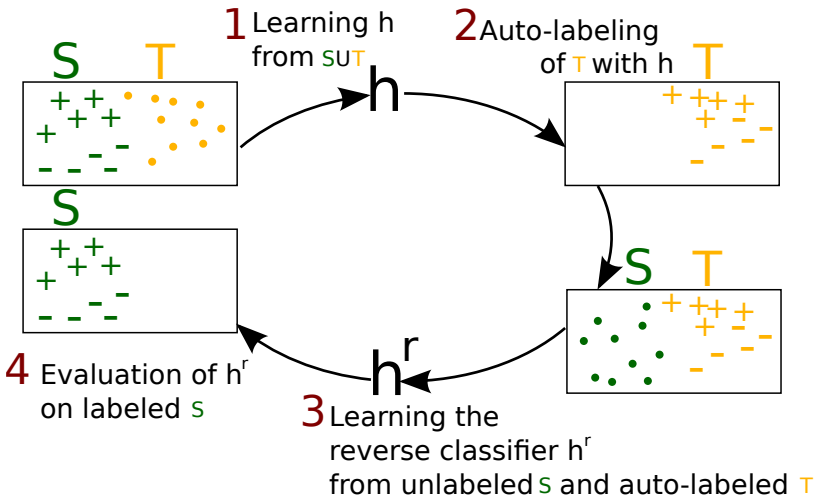

Figure 2: The principle of the reverse/circular validation in our setting.

process is repeated k times to obtain the reverse cross validation risk averaged across all folds.

6.3 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 3). We then consider seven different target domains by rotating anticlockwise the source domain according to seven 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 domain adaptation problem is repeated ten 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 two features).

We remark that our PBDA provides the best performances except for 50° and 20°, indicating that PBDA accurately tackles domain adaptation 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 domain adaptation situation. The adaptation versus risk minimization trade-off suggested by Theorem 14 appears in Figure 3. 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 domain adaptation algorithm.

6.4 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).

Table 1: Average error rate results for seven rotation angles.

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

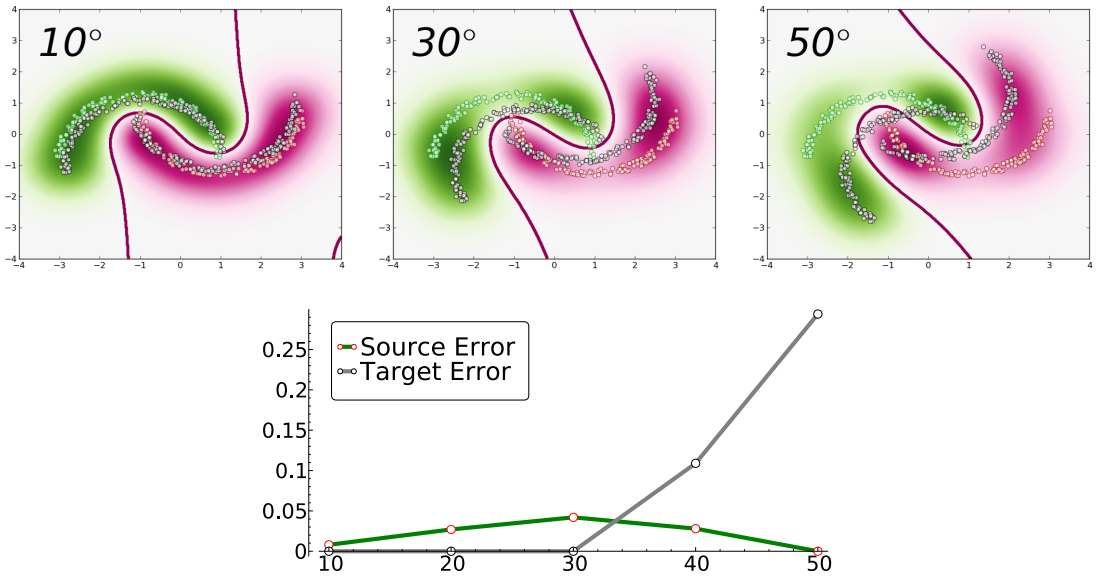

Figure 3: Illustration of the decision boundary of PBDA on three rotations angles for fixed parameters $A = C = 1$. The two classes of the source sample are green and pink, and target (unlabeled) sample is grey. The bottom plot shows corresponding source and target errors. We intentionally avoid to tune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

Originally, the reviews corresponded to a rate between one and five stars and the feature space (of unigrams and bigrams) has on average a dimension of $100,000$. For sake of simplicity and for considering a binary classification task, we propose to follow a setting similar to the one proposed by Chen et al. (2011). Then the two possible classes are: $+1$ for the products with a rank higher that 3 stars, -1 for those with a rank lower or equal to 3 stars. The dimensionality is reduced in the following way: Chen et al. (2011) only kept the features that appear at least ten times in a particular DA task (it remains about $40,000$ features), and pre-processed the data with a standard tf-idf re-weighting. One type of product is a domain, then we perform twelve domain adaptation tasks. For example, "books→DVDs" corresponds to the task for which books is the source domain and

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

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

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 domain adaptation 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 five in comparison of PBDA. In fact, the clear advantage of PBDA is that we jointly optimize the terms of our bound in one step.

6.5 Combining PBDA and Representation Learning

As discussed in the introduction, there exist several families of approaches used to tackle the domain adaptation problem. The present work focuses on the minimization of a distance metric between the source and target distributions. Now, we ask ourselves whether it can be fruitful to combine our PBDA algorithm with another approach. To do so, we executed PBDA on top of the Marginalized Stacked Denoising Autoencoders (mSDA) introduced by Chen et al. (2012).

In brief, mSDA is an unsupervised algorithm that learn a new representation of the training samples. As a "denoising autoencoders" algorithm, it finds a representation from which one can (approximatively) reconstruct the original features of an example from its noisy counterpart. The originality of mSDA is to learn a representation that allows reconstructing

Table 3: Error rates for mSDA representations on sentiment analysis dataset.

	SVM^{CV}	PBDA^{CV+RCV}	PBDA^{RCV}	PBDA^{CV}	SVM^{TEST}	PBDA^{TEST}
B→D	*0.172*	0.174	0.181	0.174	0.171	*0.170*
B→E	0.243	*0.235*	*0.235*	0.308	0.221	*0.179*
B→K	0.189	*0.181*	*0.181*	0.185	*0.158*	*0.158*
D→B	0.179	*0.178*	*0.178*	0.189	*0.174*	0.175
D→E	*0.223*	0.233	0.233	0.327	0.195	*0.165*
D→K	*0.152*	0.155	0.155	0.163	0.152	*0.147*
E→B	*0.239*	0.246	0.246	0.251	*0.226*	0.233
E→D	0.233	0.232	*0.230*	0.232	*0.225*	0.230
E→K	0.128	*0.123*	*0.123*	0.133	0.127	*0.115*
K→B	0.229	0.230	0.230	*0.225*	0.221	*0.217*
K→D	0.209	0.216	0.311	*0.208*	0.209	*0.200*
K→E	0.138	*0.134*	0.142	*0.134*	0.138	*0.133*
Average	*0.195*	*0.195*	0.204	0.211	0.185	*0.177*

both source and target unlabeled examples. Then, one can execute any supervised learning algorithm on the new representation of source samples, for which the labels are known.

That is, given a source sample $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ and a target sample $T = \{(\mathbf{x}_i^t)\}_{i=1}^{m'}$, mSDA takes the unlabeled parts of S and T, $\{\mathbf{x}_1^s, \ldots, \mathbf{x}_m^s, \mathbf{x}_1^t, \ldots, \mathbf{x}_{m'}^t\}$, and learn a feature map $f : X \to X'$, where X' is a new input space (of real-valued vector). In (Chen et al., 2012), a linear SVM is executed using $S_f = \{(f(\mathbf{x}_i^s), y_i^s)\}_{i=1}^m$ as training data, and the hyper-parameter C is selected by standard cross-validation.

We compare the performance of SVM on mSDA representation to PBDA on the same representations. That is, we obtain a new representation of both source $S_f = \{(f(\mathbf{x}_i^s), y_i^s)\}_{i=1}^m$ and target $T_f = \{(f(\mathbf{x}_i^t))\}_{i=1}^{m'}$ data, using mSDA. Then, we execute PBDA using S_f and T_f.

This comparison is done using the *Amazon reviews* dataset. For the sake of comparison, we used the dataset pre-processed by Chen et al. (2012), which is slightly different from the one used in Section 6.4. Indeed, each domain share the same $5,000$ features, and no tf-idf re-weighting is applied. For each pair source-target, mSDA representations are generated using a *corruption probability* of 50% and a *number of layers* of 5. Then, SVM and PBDA are executed on the same representations.

The results are reported in Table 3. The PBDA algorithm, when we select the hyperparameter by reverse cross-validation (PBDA^{RCV}), is not always as good as the cross-validated SVM (SVM^{CV}). However, by looking closer at the results, we notice that there often exists hyperparameters for which PBDA is better on the testing set than the best achievable SVM (as reported by the columns PBDA^{TEST} and SVM^{TEST}). This suggests that it might be advantageous to mix mSDA and PBDA learning strategies. However, the hyperparameters selection is still a challenge in domain adaption, when we do not have any target labels, even if the reverse cross-validation method is a sound strategy. For exploratory purposes, we report on Table 3 the risk of PBDA while performing the model selection by standard cross-validation (PBDA^{CV}) and while we consider the mean of the cross-validation and the reverse cross-validation score (PBDA^{CV+RCV}). Interestingly, the latter method is a better

selection criterion than taking one or the other validation risk separately in this experiment, both being misleading in some situations.[12]

7. Generalization of the PAC-Bayesian Domain Adaptation Theorems to Multisource Domain Adaptation

In this section, we generalize our main analysis to multisource domain adaptation.

7.1 Multisource Domain Adaptation Setting

We now consider n different source domains $\{P_{S_j}\}_{j=1}^n$ over $X \times Y$ (along with $\{D_{S_j}\}_{j=1}^n$ the associated marginal distributions over X). In addition to the target m'-sample T with m' unlabeled examples drawn *i.i.d.* from the target marginal D_T, we have one *i.i.d.* source learning sample S_j per domains P_{S_j} (possibly of different size).

Similarly to Ben-David et al. (2010a), we study this issue when the relationship between the source domains and the target one is captured by a distribution v over the set of source domains $\{P_{S_j}\}_{j=1}^n$. This distribution defines a mixture of source domains that we denote by P_S^v, and its marginal over X by D_S^v, and $S^v = \{S_j\}_{j=1}^n$ corresponds to the set of source samples. On the source domains, we then consider the following v-weighted true error of the Gibbs classifier G_ρ:

$$\begin{aligned} R_{P_S^v}(G_\rho) &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{P_{S_j}\sim v} R_{P_{S_j}}(G_\rho) \\ &= \mathop{\mathbf{E}}_{P_{S_j}\sim v} \mathop{\mathbf{E}}_{h\sim\rho} R_{P_{S_j}}(h) \\ &= \sum_{j=1}^n v(P_{S_j}) \mathop{\mathbf{E}}_{h\sim\rho} R_{P_{S_j}}(h)\,. \end{aligned}$$

Its empirical counterpart is defined as

$$R_{S^v}(G_\rho) \stackrel{\text{def}}{=} \sum_{j=1}^n v(P_{S_j}) \mathop{\mathbf{E}}_{h\sim\rho} R_{S_j}(h)\,.$$

Note that another solution for tackling multisource domain adaptation in a PAC-Bayesian philosophy could be to learn different posterior distribution over $\mathcal{H}$ from different sources. Indeed, instead of learning a shared ρ on every domains (including the target one), we can learn a model for each domain, and then try to learn a good target majority vote over this set of models. In this situation, one could derive a PAC-Bayesian analysis similar to the one provided by Pentina and Lampert (2014) for life-long learning. However, this setting clearly appears to be not pertinent to extend our one-source domain analysis to multiple sources, since they treat the prior distribution as a random variable, which is not our setting.

7.2 Generalization of the ρ-Disagreement to Multiple Sources

One natural solution to generalize the ρ-disagreement of Definition 6 to the multisource setting described in above is to make use of the v-weighted sum of each ρ-disagreement

12. It is important point out that experiments on other datasets showed us that the $CV + RCV$ method not systematically outperforms the reverse cross-validation method alone.

between a source distribution and the target one $\mathbf{E}_{D_{S_j}\sim v}\,\mathrm{dis}_\rho(D_{S_j}, D_T)$, for which we can easily extend Theorem 11. However, we prefer to consider the following definition that is clearly tighter than the latter one.

Definition 15 *Let $\mathcal{H}$ be a hypothesis class. For marginal distributions $\{D_{S_j}\}_{j=1}^n$ and D_T over X, any distribution v on $\{D_{S_j}\}_{j=1}^n$, any distribution ρ on $\mathcal{H}$, the domain disagreement $\mathrm{dis}_\rho(D_S^v, D_T)$ between the mixture of source distribution D_S^v and the target distribution D_T is defined by*

$$\begin{aligned}\mathrm{dis}_\rho(D_S^v, D_T) &\stackrel{\text{def}}{=} \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[R_{D_T}(h,h') - \mathop{\mathbf{E}}_{D_{S_j}\sim v} R_{D_{S_j}}(h,h') \right] \right| \\ &= \left| R_{D_T}(G_\rho, G_\rho) - \mathop{\mathbf{E}}_{D_{S_j}\sim v} R_{D_{S_j}}(G_\rho, G_\rho) \right| .\end{aligned}$$

As noticed before, we trivially have

$$\mathrm{dis}_\rho(D_S^v, D_T) \;\le\; \mathop{\mathbf{E}}_{D_{S_j}\sim v} \mathrm{dis}_\rho(D_{S_j}, D_T)\,. \tag{16}$$

Therefore, one can use the various PAC-Bayesian bounds presented in Section 4.1.3 to obtain an empirical guarantee over $\mathrm{dis}_\rho(D_S^v, D_T)$ from a collection of observations from each domain. In particular, Corollary 16 below is directly obtained from Theorem 9.

For sake of simplicity, the results presented for the multisource setting suppose that every sample shares the same size m. We use the shortcut notation $S^v \sim (P_S^v)^m$ to denote the collection of n source samples of m examples. That is, $S^v = \{S_j\}_{j=1}^n$, where $S_j \sim (P_{S_j})^m$.

Corollary 16 *For any distributions $\{D_{S_j}\}_{j=1}^n$ and D_T over X, any set of hypothesis $\mathcal{H}$, any distribution v over $\{D_{S_j}\}_{j=1}^n$, 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 $S^v \sim (P_S^v)^m$ and $T \sim (D_T)^m$, for every ρ on $\mathcal{H}$, we have*

$$\mathrm{dis}_\rho(D_S^v, D_T) \;\le\; \frac{2\alpha}{1-e^{-2\alpha}} \left[\mathop{\mathbf{E}}_{D_{S_j}\sim v} \mathrm{dis}_\rho(S^v, T) + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta} + \ln n}{m\times\alpha} + 1 \right] - 1\,.$$

Proof We upper bound the right hand side of Equation (16) by upper-bounding each individual term of the expectation using Theorem 9. That is, we bound

$$v(P_{S_1})\,\mathrm{dis}_\rho(S_1, T),\;\; v(P_{S_2})\,\mathrm{dis}_\rho(S_2, T),\;\; \ldots,\;\; v(P_{S_n})\,\mathrm{dis}_\rho(S_n, T)\,,$$

each one with probability $1-\frac{\delta}{n}$. Thereafter, we regroup these n bounds together to obtain the final result, which stands with probability $1-\delta$. ■

The bound given by Corollary 16 can suffer from the inequality of Equation (16). A better generalization guarantee is given by Theorem 17 below, that bounds directly $\mathrm{dis}_\rho(D_S^v, D_T)$, and does not rely on a term "$\ln n$" like we have in Corollary 16.

Theorem 17 *For any distributions $\{D_{S_j}\}_{j=1}^n$ and D_T over X, any set of hypothesis $\mathcal{H}$, any distribution v over $\{D_{S_j}\}_{j=1}^n$, 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 $S^v \sim (P_S^v)^m$ and $T \sim (D_T)^m$, for every ρ on $\mathcal{H}$, we have*

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

Proof Deferred to Appendix E. ■

Note that Theorem 7, Corollary 8 and Theorem 10 can also be rewritten to bound the multisource domain disagreement following the same proof techniques than we used for Theorem 17.

7.3 Multisource Domain Adaptation Bound for the Stochastic Gibbs Classifier

Let now generalize the domain adaption bound of $R_{P_T}(G_\rho)$ presented by Theorem 11 to our multisource setting.

Theorem 18 *Let $\mathcal{H}$ be a hypothesis class. We have*

$$\forall\rho \text{ on } \mathcal{H},\ \forall v \text{ on } \{P_{S_j}\}_{j=1}^n,\quad R_{P_T}(G_\rho) \;\leq\; R_{P_S^v}(G_\rho) + \frac{1}{2}\,\mathrm{dis}_\rho(D_S^v, D_T) + \lambda_\rho^v\,,$$

where λ_ρ^v is the deviation between the expected joint error of G_ρ on the source domains and the target one:

$$\begin{aligned}\lambda_\rho^v &\stackrel{\text{def}}{=} \left|\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_T}\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}),y\big)\mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}),y\big) - \mathop{\mathbf{E}}_{P_{S_j}\sim v}\ \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_{S_j}}\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}),y\big)\mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}),y\big)\right]\right| \\ &= \left|e_{P_T}(G_\rho,G_\rho) - \mathop{\mathbf{E}}_{P_{S_j}\sim v} e_{P_{S_j}}(G_\rho,G_\rho)\right|. \qquad (17)\end{aligned}$$

See Equation (6) for the definition of $e_{P_{S_j}}(G_\rho, G_\rho)$.

Proof We follow the same steps than in the proof of Theorem 11. Indeed, from Equation (12), we have

$$\begin{aligned}&R_{P_T}(G_\rho) - R_{P_S^v}(G_\rho) \\ &\quad= \frac{1}{2}\Big(R_{D_T}(G_\rho,G_\rho) - \mathop{\mathbf{E}}_{P_{S_j}\sim v} R_{D_{S_j}}(G_\rho,G_\rho)\Big) + \Big(e_{P_T}(G_\rho,G_\rho) - \mathop{\mathbf{E}}_{P_{S_j}\sim v} e_{P_{S_j}}(G_\rho,G_\rho)\Big) \\ &\quad\leq \frac{1}{2}\Big|R_{D_T}(G_\rho,G_\rho) - \mathop{\mathbf{E}}_{P_{S_j}\sim v} R_{D_{S_j}}(G_\rho,G_\rho)\Big| + \Big|e_{P_T}(G_\rho,G_\rho) - \mathop{\mathbf{E}}_{P_{S_j}\sim v} e_{P_{S_j}}(G_\rho,G_\rho)\Big| \\ &\quad= \frac{1}{2}\,\mathrm{dis}_\rho(D_S^v, D_T) + \lambda_\rho^v\,.\end{aligned}$$

■

7.4 PAC-Bayesian Theorem for Multisource Domain Adaptation

Building on Theorems 17 and 18, we now present a PAC-Bayesian theorem for multisource domain adaptation.

Theorem 19 *For any domains $\{P_{S_j}\}_{j=1}^n$ and P_T (respectively with marginals $\{D_S\}_{j=1}^n$ and D_T) over $X \times Y$, any distribution v over $\{P_{S_j}\}_{j=1}^n$, 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 $S^v \sim (P_S^v)^m$ and $T \sim (D_T)^m$, for every ρ over $\mathcal{H}$, we have*

$$R_{P_T}(G_\rho) \;\leq\; c'\, R_{S^v}(G_\rho) + \alpha' \tfrac{1}{2} \operatorname{dis}_\rho(S^v, T) + \left(\frac{c'}{c} + \frac{\alpha'}{\alpha}\right) \frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{3}{\delta}}{m} + \lambda_\rho^v + \tfrac{1}{2}(\alpha' - 1)\,,$$

where λ_ρ^v is defined by Equation (17), and where $c' \stackrel{\text{def}}{=} \dfrac{c}{1-e^{-c}}$ and $\alpha' \stackrel{\text{def}}{=} \dfrac{2\alpha}{1-e^{-2\alpha}}$.

Proof In Theorem 18, replace $R_{S^v}(G_\rho)$ and $\operatorname{dis}_\rho(D_S^v, D_T)$ by their upper bound, obtained from Theorem 5 and Theorem 17, with δ chosen respectively as $\frac{\delta}{3}$ and $\frac{2\delta}{3}$. ■

Theorem 19 above is a generalization of Theorem 13. It is straightforward to generalize Theorems 12 and 14 as well to the multisource setting.

It is important to point out that the above theorem, which naturally generalizes our one-source domain analysis, supposes that the distribution v over P_S^v is fixed (or known). However, we can prove generalization bounds that involve v given a prior distribution u over P_S^v. On the one hand, it is possible to derive a result for a distribution ρ on $\mathcal{H}$ fixed. On the other hand, such a result can be also derive on v and ρ at the same time. These two results can be helpful to derive another kind of approaches, and we detail and discuss these bounds in the in Section 8.1.

7.5 PBDA for Multisource Domain Adaptation

Regarding the results of Section 7, optimizing the PAC-Bayesian multisource domain adaptation bounds of Theorem 19 is equivalent to minimize the following trade-off

$$C\, m\, R_{S^v}(G_{\rho_{\mathbf{w}}}) + A\, m\, \operatorname{dis}_{\rho_{\mathbf{w}}}(S^v, T) + \mathrm{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}})\,,$$

where

$$\operatorname{dis}_{\rho_{\mathbf{w}}}(S^v, T) \;=\; \Big|\, R_{S^v}(G_{\rho_{\mathbf{w}}}, G_{\rho_{\mathbf{w}}}) - R_T(G_{\rho_{\mathbf{w}}}, G_{\rho_{\mathbf{w}}}) \,\Big|,$$

and $S^v = \{S_j\}_{j=1}^n = \Big\{\{(\mathbf{x}_{ij}^s, y_{ij}^s\}_{i=1}^m\Big\}_{j=1}$ are the n source samples coming from the mixture of source domains P_S^v, and $T = \{(\mathbf{x}_i^t)\}_{i=1}^m$ is the target sample. Given the vectors of weights $\mathbf{v} = \{v(P_{S_j})\}_{j=1}^n$ over the source domains, finding the optimal $\rho_{\mathbf{w}}$ is then equivalent to find the vector $\mathbf{w}$ that minimizes

$$C \sum_{j=1}^n \sum_{i=1}^m v(P_{S_j})\, \Phi\!\left(y_{ij}^s \frac{\mathbf{w}\cdot\mathbf{x}_{ij}^s}{\|\mathbf{x}_{ij}^s\|}\right) + A \left| \sum_{i=1}^m \left[\sum_{j=1}^n v(P_{S_j})\, \Phi_{\text{dis}}\!\left(\frac{\mathbf{w}\cdot\mathbf{x}_{ij}^s}{\|\mathbf{x}_{ij}^s\|}\right) - \Phi_{\text{dis}}\!\left(\frac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\right] \right| + \frac{\|\mathbf{w}\|^2}{2}.$$

Note that if v is an uniform distribution, *i.e.*, every source domain is equally probable, one can solve above optimization problem using the learning algorithm PBDA of Section 5, with $S := \bigcup_{j=1}^{n} S_j$ as the source sample. In Section 8.1, we discuss the possibility of creating other kind of learning algorithm, namely by learning v, the weights of source distributions.

8. Discussions

In this section, we discuss two points related to this paper. Firstly, we present two others results in multisource domain adaptation that lead to open-questions related to the deviation of new multisource algorithms. Secondly, we point out the differences between our new version of the PAC-Bayesian domain adaptation bound (Theorem 11) and the version proposed in Germain et al. (2013).

8.1 Other Results for Multiple Source Domain Adaptation

In Section 7, we studied multisource domain adaptation when we suppose that we know the distribution v over P_S^v. However, this ideal situation cannot be always verified. Then either one can fix v as the uniform distribution, or one can learn v given a prior distribution u on P_S^v. This latter point can be justified by the two following theorems.

Firstly, we can prove a bound similar to Theorem 19, but applied on the distribution v on the source domains instead of the distribution ρ on $\mathcal{H}$.

Theorem 20 *For any domains $\{P_{S_j}\}_{j=1}^n$ and P_T (respectively with marginals $\{D_S\}_{j=1}^n$ and D_T) over $X \times Y$, any prior distribution u over $\{P_{S_j}\}_{j=1}^n$, and for any set $\mathcal{H}$ of hypothesis, for any fixed distribution[13] π over $\mathcal{H}$, any $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of $S^v \sim (P_S^v)^m$ and $T \sim (D_T)^m$, for every v over $\{P_{S_j}\}_{j=1}^n$, we have*

$$R_{P_T}(G_\pi) \;\leq\; c'\, R_{S^v}(G_\pi) + \alpha' \tfrac{1}{2} \operatorname{dis}_\rho(S^v, T) + \left(\frac{c'}{c} + \frac{\alpha'}{2\,\alpha}\right) \frac{\mathrm{KL}(v\|u) + \ln \frac{3}{\delta}}{m} + \lambda_\rho^v + \tfrac{1}{2}(\alpha' - 1)\,,$$

where λ_ρ^v is defined by Equation (17), and where $c' \stackrel{\text{def}}{=} \dfrac{c}{1-e^{-c}}$ and $\alpha' \stackrel{\text{def}}{=} \dfrac{2\alpha}{1-e^{-\alpha}}$.

Proof Deferred to Appendix F. ■

Secondly, it is possible to prove the same kind of generalization bounds for the distribution v over the source domains and the distribution ρ over $\mathcal{H}$ at the same time. This result is stated in the next theorem.

Theorem 21 *For any domains $\{P_{S_j}\}_{j=1}^n$ and P_T (respectively with marginals $\{D_S\}_{j=1}^n$ and D_T) over $X \times Y$, any prior distribution u over $\{P_{S_j}\}_{j=1}^n$, 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 $S^v \sim (P_S^v)^m$ and $T \sim (D_T)^m$, for every v over $\{P_{S_j}\}_{j=1}^n$, and every ρ over $\mathcal{H}$, we*

13. To avoid confusion with ρ that we usually want to learn, we denote this fixed distribution π.

have

$$R_{P_T}(G_\rho) \;\leq\; c'\, R_{S^v}(G_\rho) + \alpha'\, \tfrac{1}{2}\, \mathrm{dis}_\rho(S^v, T) + \left(\frac{c'}{c} + \frac{\alpha'}{\alpha}\right) \frac{\mathrm{KL}(\rho\|\pi) + \mathrm{KL}(v\|u) + \ln \frac{3}{\delta}}{m} + \lambda_\rho^v + \tfrac{1}{2}(\alpha' - 1)\,,$$

Proof Deferred to Appendix G. ■

These two theorems open the door to the conception of two different algorithms for PAC-Bayesian multisource domain adaptation when we desire to learn both the distributions v on P_S^v and ρ on $\mathcal{H}$. On the one hand, Theorem 20 suggests that one could derive a two steps algorithm for PAC-Bayesian multisource domain adaptation, according the following principle:

(i) Given a fixed distribution π over $\mathcal{H}$, we can learn v by minimizing a trade-off between $R_{S^v}(G_\pi)$, $\mathrm{dis}_\rho(S^v, T)$ and $\mathrm{KL}(v\|u)$.

(ii) Then, for learning ρ, we simply have to optimize PBDA given this learned v.

On the other hand, Theorem 21 implies that we can jointly learn v and ρ by optimizing the trade-off between $R_{S^v}(G_\rho)$, $\mathrm{dis}_\rho(S^v, T)$, $\mathrm{KL}(v\|u)$ and $\mathrm{KL}(\rho\|\pi)$. This leads to exiting research directions.

8.2 Comparison with the first PAC-Bayesian domain adaptation bound

As said in Section 4, our PAC-Bayesian domain adaptation bound (of Theorem 11) improves the one provided in Germain et al. (2013). We recall that our bound is expressed as follows. For every distribution ρ on $\mathcal{H}$, we have

$$R_{P_T}(G_\rho) \leq R_{P_S}(G_\rho) + \frac{1}{2}\,\mathrm{dis}_\rho(D_S, D_T) + \underbrace{\left|\, e_{P_T}(G_\rho, G_\rho) - e_{P_S}(G_\rho, G_\rho) \,\right|}_{\lambda_\rho}. \tag{18}$$

Germain et al. (2013) proved the next result.[14] For every distribution ρ on $\mathcal{H}$, we have

$$R_{P_T}(G_\rho) \leq R_{P_S}(G_\rho) + \mathrm{dis}_\rho(D_S, D_T) + \underbrace{R_{P_T}(G_{\rho_T^*}) + R_{D_T}(G_\rho, G_{\rho_T^*}) + R_{D_S}(G_\rho, G_{\rho_T^*})}_{\lambda_{\rho,\rho_T^*}}, \tag{19}$$

where $\rho_T^* = \mathrm{argmin}_\rho\, R_{P_T}(G_\rho)$ is the best distribution on the target domain.

The improvement of Equation (18) over Equation (19) relies on two main points. On the one hand, our new result contains only the half of $\mathrm{dis}_\rho(D_S, D_T)$. On the other hand, contrary to λ_{ρ,ρ_T^*} of Equation (19), the term λ_ρ of Equation (18) does not depend anymore on the best ρ_T^* on the target domain. This implies that our new bound is not degenerated

14. The proof of Equation (19) relies on several triangle inequalities and on an artificial introduction of a source error term $R_{P_S}(G_\rho)$ (see Germain et al. (2013) for more technical details). Therefore, the proof of Equation (18) seems simpler as it is only based the rewriting of the risk introduced by Equation (12).

when the two distributions P_S and P_T are equal (or very close). Conversely, when $P_S = P_T$, the bound of Equation (19) gives

$$R_{P_T}(G_\rho) \leq R_{P_T}(G_\rho) + R_{P_T}(G_{\rho_T^*}) + 2R_{D_T}(G_\rho, G_{\rho_T^*}) \,,$$

which is at least $2R_{P_T}(G_{\rho_T^*})$. Moreover, the term $2R_{D_T}(G_\rho, G_{\rho_T^*})$ is greater than zero for any ρ when the support of ρ and ρ_T^* in $\mathcal{H}$ is constituted of at least two different classifiers.

Finally, note that these improvements do not change the form and the philosophy of the PAC-Bayesian theorems of Section 4.2.2, and then of the algorithm PBDA of Section 5. Indeed, the only differences stand in $\frac{1}{2}\,\mathrm{dis}_\rho(D_S, D_T)$ and in the value of λ_ρ.

9. Conclusion and Future Work

In this paper, we define a domain divergence pseudometric 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 jointly optimize the multiple trade-offs implied by th bound for linear classifiers. In addition, we generalize our analysis to multisource domain adaptation, allowing us to take into account informations from different source domains according to their relations to the target one.

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.

Firstly, the PAC-Bayesian approach 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 domain adaptation situation is an exciting direction which could also be exploited when some few target labels are available. Moreover, as pointed out by Pentina and Lampert (2014), this notion of prior distribution could modelize an information learned from previous tasks. This suggests that we can extend our multisource analysis to issues related to lifelong learning where the objective is to perform well on future tasks, for which so far no data has been observed (Thrun and Mitchell, 1995).

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. Another possible solution could be to explicitly control the neglected term in the domain adaptation bound. 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, and by the European Research Concil under the European Unions Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036. Computations were performed on Compute Canada and Calcul Québec infrastructures (founded by CFI, NSERC and FRQ). We thank Christoph Lampert and Anastasia Pentina for helpful discussions. A part of the work of this paper was carried out while E. Morvant was affiliated with IST Austria.

Appendix A. Some Tools

Lemma 22 (Markov's inequality) *Let Z be a random variable and $t \geq 0$, then*

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

Lemma 23 (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 24 (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 25 (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*

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

Lemma 26 (Change of measure inequality) *For any set $\mathcal{H}$, for any distributions π and ρ on $\mathcal{H}$, and for any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, we have*

$$\mathop{\mathbf{E}}_{f\sim\rho} \phi(f) \;\leq\; \mathrm{KL}(\rho\|\pi) + \ln\left(\mathop{\mathbf{E}}_{f\sim\pi} e^{\phi(f)}\right)\,.$$

Lemma 27 *Given any set $\mathcal{H}$, and any distributions π and ρ on $\mathcal{H}$, let $\hat{\rho}$ and $\hat{\pi}$ two distributions over $\mathcal{H}^2$ such that $\hat{\rho}(h,h') \stackrel{\text{def}}{=} \rho(h)\rho(h')$ and $\hat{\pi}(h,h') \stackrel{\text{def}}{=} \pi(h)\pi(h')$. Then*

$$\mathrm{KL}(\hat{\rho}\|\hat{\pi}) \;=\; 2\,\mathrm{KL}(\rho\|\pi)\,.$$

Proof

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

■

Appendix B. Proof of Theorem 7

Proof Firstly, we propose to upper-bound

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

by its empirical counterpart

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

To achieve this, we consider an "abstract" classifier $\hat{h} \stackrel{\text{def}}{=} (h,h') \in \mathcal{H}^2$ chosen according a distribution $\hat{\rho}$, with $\hat{\rho}(\hat{h}) = \rho(h)\rho(h')$. 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

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

which empirical counterpart is

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

The error of the related Gibbs classifier of these two quantities are

$$R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}}) = \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} R^{(1)}_{D_{S\times T}}(\hat{h}) \quad \text{and} \quad R^{(1)}_{S\times T}(G_{\hat{\rho}}) = \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} R^{(1)}_{S\times T}(\hat{h})\,. \tag{20}$$

It is easy to show that

$$d^{(1)} = 2R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}}) - 1 \quad \text{and} \quad d^{(1)}_{S\times T} = 2R^{(1)}_{S\times T}(G_{\hat{\rho}}) - 1\,. \tag{21}$$

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 22). 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

$$\begin{aligned}
\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h})\middle\|R^{(1)}_{D_{S\times T}}(\hat{h})\right)} &\le \frac{1}{\delta}\mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m}\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h})\middle\|R^{(1)}_{D_{S\times T}}(\hat{h})\right)}\\
&= \frac{1}{\delta}\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}}\mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} e^{m\mathrm{kl}\left(R^{(1)}_{S\times T}(\hat{h})\middle\|R^{(1)}_{D_{S\times T}}(\hat{h})\right)}\\
&\le \frac{1}{\delta}\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} 2\sqrt{m}\,,
\end{aligned}$$

where the last inequality comes from the Maurer's lemma (Lemma 25).
By taking the logarithm of each outermost sides of the previous equation, we then obtain

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

Let us now find a lower bound of the left side of the last equation by using the change of measure inequality (Lemma 26) and the Jensen's inequality (Lemma 23) on the convex function $\mathrm{kl}(\cdot\|\cdot)$. We have

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

Note that the last equality is obtained from Equation (20) and Lemma 27.

We finally obtain

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

With Equation (21), the previous line gives us 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 $S\times T\sim(D_S\times D_T)^m$,

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

Lemma 28 (stated below) gives

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

which, since $|d^{(1)}| = \mathrm{dis}_\rho(D_S, D_T)$ and $|d^{(1)}_{S\times T}| = \mathrm{dis}_\rho(S,T)$, implies the result. ■

Lemma 28 *For $a, b \in [-1, +1]$, we have*

$$\mathrm{kl}\left(\tfrac{1+|a|}{2} \,\middle\|\, \tfrac{1+|b|}{2}\right) \;\leq\; \mathrm{kl}\left(\tfrac{1+a}{2} \,\middle\|\, \tfrac{1+b}{2}\right) .$$

Proof There are four cases to consider.

Case 1: Let $a \geq 0$ and $b \geq 0$.
This first case is trivial, since $|a| = a$ and $|b| = b$.

Case 2: Let $a \leq 0$ and $b \leq 0$.
This case reduces to **Case 1** because $\mathrm{kl}(q\|p) = \mathrm{kl}(1-q\|1-p)$ for all $(q, p) \in [0, 1]^2$.
Then

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

Case 3: Let $a \leq 0$ and $b \geq 0$.
From straightforward calculations, we show that

$$\begin{aligned}
&\mathrm{kl}\left(\tfrac{1+|a|}{2}\middle\|\tfrac{1+|b|}{2}\right) - \mathrm{kl}\left(\tfrac{a+1}{2}\middle\|\tfrac{1+b}{2}\right) \\
&\qquad = \;\; \mathrm{kl}\left(\tfrac{1-a}{2}\middle\|\tfrac{1+b}{2}\right) - \mathrm{kl}\left(\tfrac{1+a}{2}\middle\|\tfrac{1+b}{2}\right) \\
&\qquad = \;\; \left(\tfrac{1-a}{2} - \tfrac{1+a}{2}\right) \ln\left(\frac{1}{\frac{1+b}{2}}\right) + \left(\left(1 - \tfrac{1-a}{2}\right) - \left(1 - \tfrac{1+a}{2}\right)\right) \ln\left(\frac{1}{1 - \frac{1+b}{2}}\right) \\
&\qquad = \;\; -a \ln\left(\frac{1}{\frac{1+b}{2}}\right) \;+\; a \ln\left(\frac{1}{1 - \frac{1+b}{2}}\right) \;=\; -a \ln\left(\frac{1}{\frac{1+b}{2}}\right) \;+\; a \ln\left(\frac{1}{\frac{1-b}{2}}\right) \\
&\qquad = \;\; a \;\ln\left(\tfrac{1+b}{1-b}\right) \\
&\qquad \leq \;\; 0 \,.
\end{aligned}$$

Case 4: Let $a \geq 0$ and $b \leq 0$.
This case reduces to **Case 3**, since $\mathrm{kl}(q\|p) = \mathrm{kl}(1-q\|1-p)$ for all $(q, p) \in [0, 1]^2$.
Hence,

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

■

Appendix C. Detailed Proof of Theorem 9

Proof Similarly as in the proof of Theorem 7 (see Appendix B), we will first bound

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

by its empirical counterpart.
Refer to the proof of Theorem 7 for the definitions of $R^{(1)}_{D_{S\times T}}(\hat{h})$ and $R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})$, as well as their empirical counterparts $R^{(1)}_{S\times T}(\hat{h})$ and $R^{(1)}_{S\times T}(G_{\hat{\rho}})$.

As $\mathcal{L}_{d^{(1)}}$ lies in $[0,1]$, we can bound $R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})$ following the proof process of Theorem 5 (with $c = 2\alpha$). To do so, we define the convex function,

$$\mathcal{F}(x) \stackrel{\text{def}}{=} -\ln\big[\,1-(1-e^{-2\alpha})\,x\,\big]\,, \tag{22}$$

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 22). For every $\delta\in(0,1]$, with a probability at least $1-\frac{\delta}{2}$ over the choice of $S\times T\sim(D_{S\times T})^m$, we have

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

By taking the logarithm on each side of the previous inequality, we obtain

$$\ln\left[\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))-2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}\right] = \ln\left[\frac{2}{\delta}\mathop{\mathbf{E}}_{\hat{h}\sim\hat{\pi}} e^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))} \mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m} e^{-2m\alpha R^{(1)}_{S\times T}(\hat{h})}\right]. \tag{23}$$

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 24 gives

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

The last line result, together with the choice of $\mathcal{F}$ (Equation (22)), leads to

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

We can now upper bound Equation (23) simply by

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

Let us now find a lower bound of the left side of the last equation by using the change of measure inequality (Lemma 26) and the Jensen's inequality (Lemma 23) on the convex function $\mathcal{F}$:

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

The last equality is obtained from Equation (20) and Lemma 27. This, in turn, implies

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

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

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

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

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

It then follows from Equation (21) 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

$$\frac{d^{(1)}+1}{2} \leq \frac{2\alpha}{1-e^{-2\alpha}}\left[\frac{d^{(1)}_{S\times T}+1}{2} + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}}{m\times 2\alpha}\right]. \tag{24}$$

We now bound

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

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 $(\mathbf{x}^s, \mathbf{x}^t) \sim D_{S\times T} = D_S \times D_T$:

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

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

$$\frac{d^{(2)} + 1}{2} \leq \frac{2\alpha}{1 - e^{-2\alpha}} \left[\frac{d^{(2)}_{S\times T} + 1}{2} + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}}{m \times 2\alpha} \right]. \tag{25}$$

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

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

Then, the maximum of the bound on $d^{(1)}$ (Equation (24)) and the bound on $d^{(2)}$ (Equation (25)) gives a bound on $\mathrm{dis}_\rho(D_S, D_T)$. By the union bound, 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 to

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

and we are done. ■

Appendix D. Proof of Theorem 10

Proof Let us consider the non-negative random variable $\underset{(h,h')\sim\pi^2}{\mathbf{E}} e^{2m(R_{D_S}(h,h') - R_S(h,h'))^2}$.

We apply Markov's inequality (Lemma 22). For every $\delta \in (0, 1]$, with a probability at least $1 - \frac{\delta}{2}$ over the choice of $S \sim (D_S)^m$, we have

$$\begin{aligned}
\underset{(h,h')\sim\pi^2}{\mathbf{E}} e^{2m(R_{D_S}(h,h') - R_S(h,h'))^2} &\leq \frac{2}{\delta} \underset{S\sim(D_S)^m}{\mathbf{E}} \underset{(h,h')\sim\pi^2}{\mathbf{E}} e^{2m(R_{D_S}(h,h') - R_S(h,h'))^2} \\
&= \frac{2}{\delta} \underset{(h,h')\sim\pi^2}{\mathbf{E}} \underset{S\sim(D_S)^m}{\mathbf{E}} e^{2m(R_{D_S}(h,h') - R_S(h,h'))^2} \\
&\leq \frac{2}{\delta} \underset{(h,h')\sim\pi^2}{\mathbf{E}} \underset{S\sim(D_S)^m}{\mathbf{E}} e^{\mathrm{kl}(R_S(h,h')\,\|\,R_{D_S}(h,h'))} \qquad (26) \\
&\leq \frac{2}{\delta} \underset{(h,h')\sim\pi^2}{\mathbf{E}} 2\sqrt{m}\,. \qquad (27)
\end{aligned}$$

Line (26) comes from Pinsker's inequality, and Line (27) comes from the Maurer's lemma (Lemma 25). By taking the logarithm on each outermost side of the previous inequality, we obtain

$$\ln \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \;\leq\; \ln \frac{4\sqrt{m}}{\delta}\,. \tag{28}$$

Let us now find a lower bound of the left side of the last equation by using the change of measure inequality (Lemma 26) and the Jensen's inequality (Lemma 23).

$$\begin{aligned}
&\ln \mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m(R_{D_S}(h,h')-R_S(h,h'))^2} \\
&\qquad\geq \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} 2m(R_{D_S}(h,h')-R_S(h,h'))^2 - \mathrm{KL}(\rho^2\|\pi^2) \\
&\qquad\geq 2m\left(\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_{D_S}(h,h') - \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} R_S(h,h')\right)^2 - \mathrm{KL}(\rho^2\|\pi^2) \\
&\qquad= 2m\Big(R_{D_S}(G_\rho,G_\rho) - R_S(G_\rho,G_\rho)\Big)^2 - 2\,\mathrm{KL}(\rho\|\pi)\,.
\end{aligned}$$

The last equality is obtained from Equation (20) and Lemma 27. We finally obtain

$$2m\Big(R_{D_S}(G_\rho,G_\rho) - R_S(G_\rho,G_\rho)\Big)^2 \;\leq\; 2\,\mathrm{KL}(\rho\,\|\,\pi) + \ln\frac{4\sqrt{m}}{\delta}\,,$$

and we conclude, with a probability at least $1-\frac{\delta}{2}$ over the choice of $S\sim(D_S)^m$,

$$\left|R_{D_S}(G_\rho,G_\rho) - R_S(G_\rho,G_\rho)\right| \;\leq\; \sqrt{\frac{1}{2m}\left[2\,\mathrm{KL}(\rho\,\|\,\pi) + \ln\frac{4\sqrt{m}}{\delta}\right]}. \tag{29}$$

Following the exact same proof process with the random variable $\mathop{\mathbf{E}}_{(h,h')\sim\pi^2} e^{2m'(R_{D_T}(h,h')-R_T(h,h'))^2}$, we obtain, with a probability at least $1-\frac{\delta}{2}$ over the choice of $T\sim(D_T)^{m'}$,

$$\left|R_{D_T}(G_\rho,G_\rho) - R_T(G_\rho,G_\rho)\right| \;\leq\; \sqrt{\frac{1}{2m'}\left[2\,\mathrm{KL}(\rho\,\|\,\pi) + \ln\frac{4\sqrt{m'}}{\delta}\right]}. \tag{30}$$

Joining Inequalities (29) and (30) with the union bound (that assure that both results hold simultaneously with probability $1-\delta$), gives the result because

$$\begin{aligned}
\left|R_{D_S}(G_\rho,G_\rho) - R_{D_T}(G_\rho,G_\rho)\right| &= \mathrm{dis}_\rho(D_S,D_T)\,,\\
\left|R_S(G_\rho,G_\rho) - R_T(G_\rho,G_\rho)\right| &= \mathrm{dis}_\rho(S,T)\,,
\end{aligned}$$

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

Appendix E. Proof of Theorem 17

Proof The proof follow all the steps of the proof of Theorem 9 (see Appendix C). The only difference is that, in order to obtain a guarantee over $\mathrm{dis}_\rho(D_S^v, D_T)$, we bound

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

by its empirical counterpart

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

To do so, we define the "abstract" loss of $\hat{h} \stackrel{\text{def}}{=} (h,h') \in \mathcal{H}^2$ on a tuple of $n+1$ examples $(\mathbf{x}^{s_1}, \ldots, \mathbf{x}^{s_n}, \mathbf{x}^t) \sim D_{S_1} \times \ldots \times D_{S_n} \times D_T$ by

$$\mathcal{L}_{\widehat{d^{(1)}}}(\hat{h}, \mathbf{x}^{s_1}, \ldots, \mathbf{x}^{s_n}, \mathbf{x}^t) \stackrel{\text{def}}{=} \frac{1}{2}\left[1 + \mathop{\mathbf{E}}_{D_{S_j}\sim v} \mathcal{L}_{0\text{-}1}(h(\mathbf{x}^{s_j}), h'(\mathbf{x}^{s_j})) - \mathcal{L}_{0\text{-}1}(h(\mathbf{x}^t), h'(\mathbf{x}^t)) \right] .$$

Again, we obtain the result by following the proof of Theorem 9. ■

Appendix F. Proof of Theorem 20

We first need the following result.

Theorem 29 *For any distributions $\{D_{S_j}\}_{j=1}^n$ and D_T over X, any set of hypothesis $\mathcal{H}$, for any prior distribution u over $\{D_{S_j}\}_{j=1}^n$, any 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 $S^v \sim D_S^v$, and $T \sim (D_T)^m$, for every distribution v over $\{D_{S_j}\}_{j=1}^n$, we have*

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

Proof The proof follows a process similar to the proof of Theorem 9 in Appendix C: we separately bound

$$R_{D_T}(G_\rho, G_\rho) - \mathop{\mathbf{E}}_{D_{S_j}\sim v} R_{D_{S_j}}(G_\rho, G_\rho) \quad \text{and} \quad \mathop{\mathbf{E}}_{D_{S_j}\sim v} R_{D_{S_j}}(G_\rho, G_\rho) - R_{D_T}(G_\rho, G_\rho) ,$$

by rescaling their value into $[0,1]$. ■

Then, we easily obtain the result of Theorem 20.

Proof of Theorem 20 In Theorem 18, replace $R_{S^v}(G_\rho)$ and $\mathrm{dis}_\rho(D_S^v, D_T)$ by their upper bound, obtained from Theorem 5 applied on $R_{P_S^v}(G_\pi) = \mathbf{E}_{P_{S_j}\sim v} R_{P_{S_j}}(G_\pi)$ (instead of $R_{P_S}(G_\rho)$) and Theorem 29, with δ chosen respectively as $\frac{\delta}{3}$ and $\frac{2\delta}{3}$. ■

Appendix G. Proof of Theorem 21

Proof Consider the data distribution $\mathcal{P} \stackrel{\text{def}}{=} P_{S_1} \times P_{S_2} \times \ldots \times P_{S_n}$. The loss of a classifier $h \in \mathcal{H}$ on a tuple of examples $\big((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\big) \sim \mathcal{P}$ is defined as the mean of the zero-loss $\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}_j), y_j\big)$ on each example of the tuple (*i.e.*, $j \in \{1, \ldots, n\}$).

Thanks to this convention, and by a slight abuse of notation, we can write the *expected risk* on $\mathcal{P}$ of a classifier $h \in \mathcal{H}$ as

$$\begin{aligned} R_{\mathcal{P}}(h) &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n))\sim\mathcal{P}} \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}_j), y_j\big) \\ &= \frac{1}{n}\sum_{j=1}^{n} R_{P_{S_j}}(h)\,, \end{aligned}$$

and the *expected disagreement* of a pair of classifiers $(h, h') \in \mathcal{H}^2$ on the corresponding marginal distribution $\mathcal{D} \stackrel{\text{def}}{=} D_{S_1} \times D_{S_2} \times \ldots \times D_{S_n}$ as

$$\begin{aligned} R_{\mathcal{D}}(h, h') &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x}_1,\ldots,\mathbf{x}_n)\sim\mathcal{D}} \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}_j), h'(\mathbf{x}_j)\big) \\ &= \frac{1}{n}\sum_{j=1}^{n} R_{D_{S_j}}(h, h')\,. \end{aligned}$$

Let now define new posterior ρ_v and prior π_u on $\mathcal{H}$:

$$\rho_v(h) \;=\; \rho(h)\sum_{j=1}^{n} v(P_{S_j}) \quad \text{and} \quad \pi_u(h) \;=\; \pi(h)\sum_{j=1}^{n} u(P_{S_j})\,.$$

From above definitions, one can easily show

$$R_{P_S^v}(G_\rho) = R_{\mathcal{P}}(G_{\rho_v})\,, \quad \text{and} \quad \mathrm{dis}_\rho(D_S^v, D_T) = \mathrm{dis}_{\rho_v}(\mathcal{D}, D_T)\,.$$

Moreover, we have

$$\begin{aligned} \mathrm{KL}(\rho_v\|\pi_u) &= \mathop{\mathbf{E}}_{h\sim\rho_v} \ln\frac{\rho_v(h)}{\pi_u(h)} \\ &= \mathop{\mathbf{E}}_{h\sim\rho} \sum_{j=1}^{n} v(P_{S_j}) \left[\ln\frac{\rho(h)}{\pi(h)} + \ln\frac{v(P_{S_j})}{u(P_{S_j})}\right] \\ &= \mathop{\mathbf{E}}_{h\sim\rho} \ln\frac{\rho(h)}{\pi(h)} + \sum_{j=1}^{n} v(P_{S_j}) \ln\frac{v(P_{S_j})}{u(P_{S_j})} \\ &= \mathrm{KL}(\rho\|\pi) + \mathrm{KL}(v\|u)\,. \end{aligned}$$

From Theorem 13, with a probability at least $1-\delta$ over the choice of $\mathcal{S}\times T \sim (\mathcal{P}\times D_T)^m$, for every posterior distribution ρ_v on $\mathcal{H}$, we have

$$R_{P_T}(G_\rho) \;\le\; c'\, R_{\mathcal{S}}(G_{\rho_v}) + \alpha' \tfrac{1}{2}\mathrm{dis}_{\rho_v}(\mathcal{S}, T) + \left(\frac{c'}{c} + \frac{\alpha'}{\alpha}\right)\frac{\mathrm{KL}(\rho_v\|\pi_u) + \ln\frac{3}{\delta}}{m} + \lambda_{\rho_v} + \tfrac{1}{2}(\alpha'-1)\,,$$

and we obtain the final result by the substitution of $R_{\mathcal{S}}(G_{\rho_v})$, $\mathrm{dis}_{\rho_v}(\mathcal{S}, T)$, and $\mathrm{KL}(\rho_v\|\pi_u)$ with their equivalent expression. ■

References

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems*, pages 9–16, 2006.

S. Ben-David and R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of Algorithmic Learning Theory*, pages 139–153, 2012.

S. Ben-David and R. Urner. Domain adaptation-can quantity compensate for quality? *Ann. Math. Artif. Intell.*, 70(3):185–202, 2014.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 137–144, 2006.

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

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

J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, pages 120–128. Association for Computational Linguistics, 2006.

L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *Transaction Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

J. Ye C. Zhang, L. Zhang. Generalization bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, 2012.

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

C.-C. Chang and C.-J. Lin. *LibSVM: a library for support vector machines*, 2001. www.csie.ntu.edu.tw/~cjlin/libsvm.

M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 2456–2464, 2011.

M. Chen, Z. E. Xu, K. Q. Weinberger, and F. Sha. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the International Conference on Machine Learning*, 2012.

C. Cortes and M. Mohri. Domain adaptation in regression. In *Algorithmic Learning Theory*, pages 308–323. Springer, 2011.

C. Cortes and M. Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Advances in Neural Information Processing Systems*, 19:321, 2007.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning*, 2009a.

P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and S. Shanian. From PAC-Bayes bounds to KL regularization. In *Advances in Neural Information Processing Systems*, pages 603–610, 2009b.

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *International Conference on Machine Learning*, pages 738–746, 2013.

X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the International Conference on Machine Learning*, pages 513–520, 2011.

A. Habrard, J.-P. Peyrache, and M. Sebban. Iterative self-labeling domain adaptation for linear structured image classification. *International Journal on Artificial Intelligence Tools*, 22(05), 2013.

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, pages 601–608, 2006.

J. Jiang. A literature survey on domain adaptation of statistical classifiers. Technical report, CS Department at Univ. of Illinois at Urbana-Champaign, 2008.

T. Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning*, pages 200–209, 1999.

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 *Advances in Neural Information Processing Systems*, 2006.

J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2002.

F. Laviolette, M. Marchand, and J.-F. Roy. From PAC-Bayes bounds to quadratic programs for majority votes. In *International Conference on Machine Learning*, 2011.

X. Li and J. Bilmes. A Bayesian divergence prior for classifier adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 275–282, 2007.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, pages 19–30, 2009a.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the Rényi divergence. In *Conference on Uncertainty in Artificial Intelligence*, pages 367–374. AUAI Press, 2009b.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, pages 1041–1048, 2009c.

A. Margolis. A literature review of domain adaptation with unlabeled data. Technical report, University of Washington, 2011.

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

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

E. Morvant. Domain adaptation of weighted majority votes via perturbed variation-based self-labeling. *Pattern Recognition Letters*, To be published, 2014.

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

A. Pentina and C. Lampert. A PAC-Bayesian bound for lifelong learning. In *JMLR W&CP, Proceedings of International Conference on Machine Learning*, volume 32, pages 991–999, 2014.

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

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Annual Conference on Computational Learning Theory, and European Conference on Computational Learning Theory*, pages 416–426, 2001.

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

M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.

S. Thrun and T. M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15(1-2):25–46, 1995.

E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *LNCS*, pages 547–562. Springer, 2010.