PAC-Bayes and Domain Adaptation

Pascal Germain, Amaury Habrard, François Laviolette, Emilie Morvant

PAC-Bayes and Domain Adaptation

Pascal Germain PASCAL.GERMAIN@INRIA.FR
Département d'informatique de l'ENS,
École normale supérieure, CNRS, PSL Research University, 75005 Paris, France
and INRIA, Sierra Project-Team

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

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

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

Abstract

We 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, but related, target distribution. Firstly, we propose an improvement of the previous approach we proposed in Germain et al. (2013), which relies on a novel distribution pseudodistance based on a disagreement averaging, allowing us to derive a new tighter domain adaptation bound for the target risk. While this bound stands in the spirit of common domain adaptation works, we derive a second bound (recently introduced in Germain et al., 2016) that brings a new perspective on domain adaptation by deriving an upper bound on the target risk where the distributions' divergence—expressed as a ratio—controls the trade-off between a source error measure and the target voters' disagreement. We discuss and compare both results, from which we obtain PAC-Bayesian generalization bounds. Furthermore, from the PAC-Bayesian specialization to linear classifiers, we infer two learning algorithms, and we evaluate them on real data.

Keywords: Domain Adaptation, PAC-Bayesian Theory

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, the knowledge he has acquired from previous courses helps him to understand the current one. However, traditional machine learning approaches assume that the learning and test data are drawn from the same probability distribution. This assumption may be too strong for a lot of real-world tasks, in particular those where we desire to reuse 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*). Note that, it is well known that domain adaptation is a hard and challenging task even under strong assumptions (Ben-David et al., 2010b; Ben-David and Urner, 2012, 2014).

Many approaches exist in the literature to address domain adaptation, often with the same underlying idea: If we are able to apply a transformation in order to "move closer" the distributions, then we can learn a model with the available labels. This process can be performed by reweighting the importance of labeled data (Huang et al., 2006; Sugiyama et al., 2007; Cortes et al., 2010, 2015). This is one of the most popular methods when one wants to deal with the covariate-shift issue (*e.g.,* Huang et al., 2006; Sugiyama et al., 2008), where source and target domains diverge only in their marginals, *i.e.*, when 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, 2015). A third solution is to learn a new common representation space from the unlabeled part of source and target data. Then, a standard supervised learning algorithm can be run on the source labeled instances (*e.g.*, Glorot et al., 2011; Chen et al., 2012; Courty et al., 2016; Ganin et al., 2016). The work presented in this paper stands into a fourth popular class of approaches, which has been especially explored to derive generalization bounds for domain adaptation. This kind of approaches relies on the control of a measure of divergence/distance between the source distribution and target distribution (*e.g.* Ben-David et al., 2006; Ben-David et al., 2010a; Mansour et al., 2009a; Li and Bilmes, 2007; Zhang et al., 2012; Morvant et al., 2012; Cortes and Mohri, 2014). Such a distance usually depends on the set $\mathcal{H}$ of hypotheses considered by the learning algorithm. The intuition is that one must look for a set $\mathcal{H}$ that minimizes the distance between the distributions 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, for 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 source and target marginal distributions. This quantity depends on the maximal disagreement between two classifiers, and allowed 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 Zhang et al. (2012) which

1. The reader can refer to the surveys proposed by Jiang (2008); Quionero-Candela et al. (2009); and Margolis (2011) (domain adaptation in often associated with *transfer learning* (Pan and Yang, 2010)).

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.

Given the multitude of concurrent approaches for domain adaptation, and the nonexistence of a predominant one, we believe that the problem still needs to be studied from different perspectives for a global comprehension to emerge. We aim to contribute to this study from a PAC-Bayesian standpoint. One particularity of the *PAC-Bayesian theory* (first set out by McAllester, 1999) is that it focuses on algorithms that output a *posterior distribution* ρ over a classifier set $\mathcal{H}$ (*i.e.*, a ρ-average over $\mathcal{H}$) rather than just a single predictor $h \in \mathcal{H}$ (as in Ben-David et al., 2006, and other works cited above). More specifically, we tackle the *unsupervised domain adaptation* setting for binary classification, where no target labels are provided to the learner. We propose two domain adaptation analyses, both introduced separately in previous conference papers (Germain et al., 2013, 2016). We refine these results, and provide in-depth comparison, full proofs and technical details. Our analyses highlight different angles that one can adopt when studying domain adaptation.

Our first approach follows the philosophy of the seminal work of Ben-David et al. (2010a); Ben-David et al. (2006); and Mansour et al. (2009b): The risk of the target model is upper-bounded jointly by the model's risk on the source distribution, a divergence between the marginal distributions, and a non-estimable term[2] related to the ability to adapt in the current space. To obtain such a result, we define a pseudometric which is ideal for the PAC-Bayesian setting by evaluating the domains' divergence according to the ρ-average disagreement of the classifiers over the domains. Additionally, we prove that this domains' divergence is always lower than the popular $\mathcal{H}\Delta\mathcal{H}$-divergence, and is easily estimable from samples. Note that, 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. We provide here a new version of this result, that does not change the underlying philosophy supported by the previous bound, but clearly improves the theoretical result: The domain adaptation bound is now tighter and easier to interpret.

Our second analysis (introduced in Germain et al., 2016) consists in a target risk bound that brings an original way to think about domain adaptation problems. Concretely, the risk of the target model is still upper-bounded by three terms, but they differ in the information they capture. The first term is estimable from unlabeled data and relies on the disagreement of the classifiers only on the target domain. The second term depends on the expected accuracy of the classifiers on the source domain. Interestingly, this latter is weighted by a divergence between the source and the target domains that enables controlling the relationship between domains. The third term estimates the "volume" of the target domain living apart from the source one[3], which has to be small for ensuring adaptation.

Thanks to these results, we derive PAC-Bayesian generalization bounds for our two domain adaptation bounds. Then, in contrast to the majority of methods that perform a

2. More precisely, this term can only be estimated in the presence of labeled data from both the source and the target domains.
3. Here we do not focus on learning a new representation to help the adaptation: We directly aim at adapting in the current representation space.

two-step procedure, we design two algorithms tailored to linear classifiers, called PBDA and DALC, which jointly minimize the multiple trade-offs implied by the bounds. On the one hand, PBDA is inspired by our first analysis for which the first two quantities being, as usual in the PAC-Bayesian approach, the complexity of the ρ-weighted 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 domains' divergence and assesses the capacity of the posterior distribution to distinguish some structural difference between the source and target samples. On the other hand, DALC is inspired by our second analysis from which we deduce that a good adaptation strategy consists in finding a ρ-weighted majority vote leading to a suitable trade-off—controlled by the domains' divergence—between the first two terms (and the usual Kullback-Leibler divergence): Minimizing the first one corresponds to look for classifiers that disagree on the target domain, and minimizing the second one to seek accurate classifiers on the source.

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 the sake of completeness, we provide for the first time the explicit derivation of the algorithm PBGD3 (Germain et al., 2009a) tailored to linear classifiers in supervised learning. Our main contribution, which consists in two domain adaptation bounds suitable for PAC-Bayesian learning, is presented in Section 4, the associated generalization bounds are derived in Section 5. Then, we design our new algorithms for PAC-Bayesian domain adaptation in Section 6, that we experiment in Section 7. We conclude in Section 8.

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 and Mansour et al. 2009a).

2.1 Notations and Setting

We consider domain adaptation for binary classification tasks where $\mathbf{X} \subseteq \mathbb{R}^d$ is the input space of dimension d, and $Y = \{-1, +1\}$ is the output/label set. The *source domain* $\mathcal{S}$ and the *target domain* $\mathcal{T}$ are two different distributions (unknown and fixed) over $\mathbf{X} \times Y$, $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ being the respective marginal distributions over $\mathbf{X}$. We tackle the challenging task where we have no target labels, known as *unsupervised domain adaptation*. A learning algorithm is then provided with a *labeled source sample* $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m_s}$ consisting of m_s examples drawn *i.i.d.*[4] from $\mathcal{S}$, and an *unlabeled target sample* $T = \{\mathbf{x}_j\}_{j=1}^{m_t}$ consisting of m_t examples drawn *i.i.d.* from $\mathcal{T}_{\mathbf{X}}$. We denote the distribution $\mathcal{D}$ of a m-sample by $(\mathcal{D})^m$. We suppose that $\mathcal{H}$ is a set of hypothesis functions for $\mathbf{X}$ to Y. The *expected source error* and the *expected target error* of $h \in \mathcal{H}$ over $\mathcal{S}$, respectively $\mathcal{T}$, are the probability that h errs on the entire distribution $\mathcal{S}$, respectively $\mathcal{T}$,

$$\mathrm{R}_{\mathcal{S}}(h) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), y\big)\,, \quad \text{and} \quad \mathrm{R}_{\mathcal{T}}(h) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), y\big)\,,$$

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

where $\mathcal{L}_{0\text{-}1}(a,b) = \mathrm{I}[a \neq b]$ is the 0-1-loss function which returns 1 if $a \neq b$ and 0 otherwise. The *empirical source error* $\widehat{\mathrm{R}}_S(h)$ of h on the learning source sample S is

$$\widehat{\mathrm{R}}_S(h) \;=\; \frac{1}{m_s} \sum_{(\mathbf{x},y)\in S} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), y\big)\,.$$

The main objective in domain adaptation is then to learn—without target labels—a classifier $h \in \mathcal{H}$ leading to the lowest expected target error $\mathrm{R}_{\mathcal{T}}(h)$.

Given two classifiers $(h', h) \in \mathcal{H}^2$, we also introduce the notion of *expected source disagreement* $\mathrm{R}_{\mathcal{S}_\mathbf{X}}(h,h')$ and the *expected target disagreement* $\mathrm{R}_{\mathcal{T}_\mathbf{X}}(h,h')$, which measure the probability that h and h' do not agree on the respective marginal distributions, and are defined by

$$\mathrm{R}_{\mathcal{S}_\mathbf{X}}(h,h') \;=\; \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_\mathbf{X}} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), h'(\mathbf{x})\big) \quad \text{and} \quad \mathrm{R}_{\mathcal{T}_\mathbf{X}}(h,h') \;=\; \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_\mathbf{X}} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), h'(\mathbf{x})\big)\,.$$

The *empirical source disagreement* $\widehat{\mathrm{R}}_S(h,h')$ on S and the *empirical target disagreements* $\widehat{\mathrm{R}}_T(h,h')$ on T are

$$\widehat{\mathrm{R}}_S(h,h') \;=\; \frac{1}{m_s}\sum_{\mathbf{x}\in S} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), h'(\mathbf{x})\big) \quad \text{and} \quad \widehat{\mathrm{R}}_T(h,h') \;=\; \frac{1}{m_t}\sum_{\mathbf{x}\in T} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), h'(\mathbf{x})\big)\,.$$

Note that, depending on the context, S denotes either the source labeled sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m_s}$ or its unlabeled part $\{\mathbf{x}_i\}_{i=1}^{m_s}$. We can remark that the expected error $R_\mathcal{D}(h)$ on a distribution $\mathcal{D}$ can be viewed as a shortcut notation for the expected disagreement between a hypothesis h and a labeling function $f_\mathcal{D} : \mathbf{X} \to Y$ that assigns the true label to an example description with respect to $\mathcal{D}$. We have

$$\begin{aligned}\mathrm{R}_\mathcal{D}(h) \;&=\; \mathrm{R}_{\mathcal{D}_\mathbf{X}}(h, f_\mathcal{D}) \\ &=\; \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{X}} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), f_\mathcal{D}(\mathbf{x})\big)\,.\end{aligned}$$

2.2 Necessity of a Domains' 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 et al., 2010b; Ben-David and Urner, 2012, 2014). 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 (Ben-David et al., 2006; Li and Bilmes, 2007; Mansour et al., 2009a,b; Ben-David et al., 2010a; Zhang et al., 2012). Concretely, two domains $\mathcal{S}$ and $\mathcal{T}$ differ if their marginals $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$ 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 $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$, and one between the labeling. If we have some target labels, we can combine the two distances as done by Zhang et al. (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 $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ 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 divergence between the two domains.

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. (2006), and Ben-David et al. (2010a) 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[5]) hypothesis class. We have*

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

where

$$\tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = \sup_{(h,h') \in \mathcal{H}^2} \left| \mathrm{R}_{\mathcal{T}_{\mathbf{X}}}(h, h') - \mathrm{R}_{\mathcal{S}_{\mathbf{X}}}(h, h') \right|$$

is the $\mathcal{H}\Delta\mathcal{H}$-distance between the marginals $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$, and $\mu_{h^*} = \mathrm{R}_{\mathcal{S}}(h^*) + \mathrm{R}_{\mathcal{T}}(h^*)$ is the error of the best hypothesis overall $h^* = \operatorname{argmin}_{h \in \mathcal{H}} \big(\mathrm{R}_{\mathcal{S}}(h) + \mathrm{R}_{\mathcal{T}}(h)\big)$.

This bound relies on three terms. $\mathrm{R}_{\mathcal{S}}(h)$ is the classical source domain expected error. $\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ depends on $\mathcal{H}$ and corresponds to the maximum deviation between the source and target 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 $\mu_{h^*} = \mathrm{R}_{\mathcal{S}}(h^*) + \mathrm{R}_{\mathcal{T}}(h^*)$ is related to the best hypothesis h^* over the domains and acts 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) expresses a multiple trade-off between the accuracy of some particular hypothesis h, the complexity of $\mathcal{H}$ (quantified by Ben-David et al. with the usual VC-bound theory), and the "incapacity" of hypotheses 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 $\mathrm{disc}_{\mathcal{L}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ between $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$ is

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

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

$$\tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) = \mathrm{disc}_{\mathcal{L}_{0\text{-}1}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})\,.$$

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

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

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

$$\forall h \in \mathcal{H},\ \mathrm{R}_{\mathcal{T}}(h) - \mathrm{R}_{\mathcal{T}}(h^*_{\mathcal{T}}) \ \leq\ \mathrm{R}_{\mathcal{S}_{\mathbf{X}}}(h^*_{\mathcal{S}}, h) + \mathrm{disc}_{\mathcal{L}_{0\text{-}1}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \nu_{(h^*_{\mathcal{S}}, h^*_{\mathcal{T}})}\,, \tag{2}$$

where $\nu_{(h^*_{\mathcal{S}}, h^*_{\mathcal{T}})} = \mathrm{R}_{\mathcal{S}_{\mathbf{X}}}(h^*_{\mathcal{S}}, h^*_{\mathcal{T}})$ is the disagreement between the ideal hypothesis on the target and source domains: $h^*_{\mathcal{T}} = \mathrm{argmin}_{h\in\mathcal{H}} \mathrm{R}_{\mathcal{T}}(h)$, and $h^*_{\mathcal{S}} = \mathrm{argmin}_{h\in\mathcal{H}} \mathrm{R}_{\mathcal{S}}(h)$.

Equation (2) can be tighter than Equation (1)[6] since it bounds the difference between the target error of a classifier and the one of the optimal $h^*_{\mathcal{T}}$. Based on Theorem 2 and a Rademacher complexity analysis, Mansour et al. (2009a) provide a generalization bound on the target risk, that expresses a trade-off between the disagreement (between h and the best source hypothesis $h^*_{\mathcal{S}}$), the complexity of $\mathcal{H}$, and—again—the "incapacity" of hypotheses to detect differences between the domains.

To conclude, the domain adaptation bounds of Theorems 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 two *average case* approaches 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). Our first approach (see Section 4.1) stands in the philosophy of these seminal works, and the second one (see Section 4.2) brings a different and novel point of view by taking advantages of the PAC-Bayesian framework we recall in the next section.

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—without relying on any validation set—on weighted majority votes, *i.e.*, for ensemble methods (Dietterich, 2000; Re and Valentini, 2012) where several classifiers (or voters) are assigned a specific weight.

Throughout this section, we adopt an algorithm design perspective. Indeed, the PAC-Bayesian analysis of domain adaptation provided in the forthcoming sections is oriented by the motivation of creating new adaptive algorithms.

3.1 Notations and Setting

Traditionally, PAC-Bayesian theory considers weighted majority votes over a set $\mathcal{H}$ of binary hypothesis, often called voters. Let $\mathcal{D}$ be a fixed yet unknown distribution over $\mathbf{X} \times Y$, and S be a learning set where each example are drawn *i.i.d.* from $\mathcal{D}$. Then, given a *prior distribution* π over $\mathcal{H}$ (independent from the learning set S), the "PAC-Bayesian" learner aims at finding a *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

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

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

However, minimizing the risk of B_ρ, defined as

$$\mathrm{R}_{\mathcal{D}}(B_\rho) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathcal{L}_{0\text{-}1}\big(B_\rho(\mathbf{x}), y\big)\,,$$

is known to be NP-hard. To tackle this issue, the PAC-Bayesian approach deals with the risk of the stochastic *Gibbs classifier* G_ρ associated with ρ and closely related to B_ρ. In order to predict the label of an example $\mathbf{x} \in \mathbf{X}$, the Gibbs classifier first draws a hypothesis h from $\mathcal{H}$ according to ρ, then returns $h(\mathbf{x})$ as label. Then, the error of the Gibbs classifier on a domain $\mathcal{D}$ corresponds to the expectation of the errors over ρ:

$$\mathrm{R}_{\mathcal{D}}(G_\rho) \;=\; \mathop{\mathbf{E}}_{h\sim\rho} \mathrm{R}_{\mathcal{D}}(h)\,. \tag{3}$$

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

$$\mathrm{R}_{\mathcal{D}}(B_\rho) \;\leq\; 2\,\mathrm{R}_{\mathcal{D}}(G_\rho)\,.$$

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

$$\mathrm{R}_{\mathcal{D}}(B_\rho) \;\leq\; 1 - \frac{\big(1 - 2\,\mathrm{R}_{\mathcal{D}}(G_\rho)\big)^2}{1 - 2\,\mathrm{d}_{\mathcal{D}_{\mathbf{X}}}(\rho)}\,, \tag{4}$$

where $\mathrm{d}_{\mathcal{D}_{\mathbf{X}}}(\rho)$ corresponds to the *expected disagreement* of the classifiers over ρ:

$$\mathrm{d}_{\mathcal{D}_{\mathbf{X}}}(\rho) \;=\; \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathrm{R}_{\mathcal{D}_{\mathbf{X}}}(h,h')\,. \tag{5}$$

Equation (4) suggests that for a fixed numerator, *i.e.*, a fixed risk of the Gibbs classifier, the best ρ-weighted majority vote is the one associated with the lowest denominator, *i.e.*, with the greatest disagreement between its voters (for further analysis, see Germain et al., 2015b).

We now introduce the notion of *expected joint error* of a pair of classifiers $(h,h') \in \mathcal{H}^2$ drawn according to the distribution ρ, defined as

$$\mathrm{e}_{\mathcal{D}}(\rho) \;=\; \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), y\big) \times \mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}), y\big)\,. \tag{6}$$

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

$$\mathrm{R}_{\mathcal{D}}(G_\rho) \;=\; \frac{1}{2}\,\mathrm{d}_{\mathcal{D}_{\mathbf{X}}}(\rho) + \mathrm{e}_{\mathcal{D}}(\rho)\,. \tag{7}$$

Indeed, we have

$$\begin{aligned}
2\,\mathrm{R}_{\mathcal{D}}(G_\rho) \;&=\; \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \Big[\mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), y\big) + \mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}), y\big)\Big] \\
&=\; \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \Big[1 \times \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), h'(\mathbf{x})\big) + 2 \times \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), y\big)\,\mathcal{L}_{0\text{-}1}\big(h'(\mathbf{x}), y\big)\Big] \\
&=\; \mathrm{d}_{\mathcal{D}_{\mathbf{X}}}(\rho) + 2 \times \mathrm{e}_{\mathcal{D}}(\rho)\,.
\end{aligned}$$

Lastly, PAC-Bayesian theory allows one to bound the expected error $\mathrm{R}_{\mathcal{D}}(G_\rho)$ in terms of two major quantities: The empirical error

$$\widehat{\mathrm{R}}_S(G_\rho) \;=\; \mathop{\mathbf{E}}_{h\sim\rho} \widehat{\mathrm{R}}_S(h)$$

estimated on a sample $S \sim (\mathcal{D})^m$ and the Kullback-Leibler divergence

$$\mathrm{KL}(\rho\|\pi) \;=\; \mathop{\mathbf{E}}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)}\,.$$

We present in the next section the PAC-Bayesian theorem proposed by Catoni (2007).[7]

3.2 A Usual PAC-Bayesian Theorem

Usual PAC-Bayesian theorems suggest that, in order to minimize the expected risk, a learning algorithm should perform a trade-off between the empirical risk minimization $\widehat{\mathrm{R}}_S(G_\rho)$ 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 3 below. This PAC-Bayesian result, first proposed by Catoni (2007), is defined with a hyperparameter (here named ω). 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 3 (Catoni, 2007) *For any domain $\mathcal{D}$ over $\mathbf{X}\times Y$, for any set of hypotheses $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta \in (0,1]$, and any real number $\omega > 0$, with a probability at least $1-\delta$ over the random choice of $S \sim (\mathcal{D})^m$, for every ρ on $\mathcal{H}$, we have*

$$\mathrm{R}_{\mathcal{D}}(G_\rho) \;\leq\; \frac{\omega}{1-e^{-\omega}} \left[\widehat{\mathrm{R}}_S(G_\rho) + \frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{1}{\delta}}{m \times \omega} \right]. \tag{8}$$

Similarly to McAllester and Keshet (2011), we could choose to restrict $\omega \in (0,2)$ to obtain a slightly looser but simpler bound. Using $e^{-\omega} \leq 1 - \omega - \frac{1}{2}\omega^2$ to upper-bound on the right-hand side of Equation (8), we obtain

$$\mathrm{R}_{\mathcal{D}}(G_\rho) \;\leq\; \frac{1}{1-\frac{1}{2}\omega} \left[\widehat{\mathrm{R}}_S(G_\rho) + \frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{1}{\delta}}{m \times \omega} \right]. \tag{9}$$

The bound of Theorem 3—in both forms of Equations (8) and (9)—has two appealing characteristics. First, choosing $\omega = 1/\sqrt{m}$, the bound becomes consistent: It converges to $1 \times \left[\widehat{\mathrm{R}}_S(G_\rho) + 0 \right]$ 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 ω allows us to control the trade-off between the empirical risk $\widehat{\mathrm{R}}_S(G_\rho)$ and the "complexity term" $\frac{1}{m}\mathrm{KL}(\rho\|\pi)$.

7. Two other common forms of the PAC-Bayesian theorem are the one of McAllester (1999) and the one of Seeger (2002); Langford (2005). We refer the reader to our research report (Germain et al., 2015a) for a larger variety of PAC-Bayesian theorems in a domain adaptation context.

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 $\mathbf{w}' \in \mathbb{R}^d$:

$$h_{\mathbf{w}'}(\mathbf{x}) \;=\; \operatorname{sign}\left(\mathbf{w}' \cdot \mathbf{x}\right),$$

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) 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

$$\pi_{\mathbf{0}}(h_{\mathbf{w}'}) \;=\; \left(\frac{1}{\sqrt{2\boldsymbol{\pi}}}\right)^d \exp\left(-\frac{1}{2}\|\mathbf{w}'\|^2\right),$$
$$\text{and} \quad \rho_{\mathbf{w}}(h_{\mathbf{w}'}) \;=\; \left(\frac{1}{\sqrt{2\boldsymbol{\pi}}}\right)^d \exp\left(-\frac{1}{2}\|\mathbf{w}' - \mathbf{w}\|^2\right).$$

An interesting property of these distributions—also seen as multivariate normal distributions, $\pi_{\mathbf{0}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\rho_{\mathbf{w}} = \mathcal{N}(\mathbf{w}, \mathbf{I})$—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

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

Moreover, the expected risk of the Gibbs classifier $G_{\rho_{\mathbf{w}}}$ on a domain $\mathcal{D}$ is then given by[8]

$$\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}}) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \Phi_{\mathrm{R}}\left(y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right), \tag{10}$$

where

$$\Phi_{\mathrm{R}}(x) \;=\; \tfrac{1}{2}\left[1 - \operatorname{Erf}\left(\tfrac{x}{\sqrt{2}}\right)\right], \tag{11}$$

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

$$\operatorname{Erf}(x) \;=\; \frac{2}{\sqrt{\boldsymbol{\pi}}} \int_0^x \exp\left(-t^2\right) \mathrm{d}t\,. \tag{12}$$

Here, $\Phi_{\mathrm{R}}(x)$ can be seen as a *smooth* surrogate of the 0-1-loss function $\mathrm{I}\,[x \leq 0]$ relying on $y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}$. This function Φ_{R} is sometimes called the *probit–loss* (*e.g.*, McAllester and Keshet, 2011). It is worth noting that $\|\mathbf{w}\|$ plays an important role on the value of $\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}})$, but not on $\mathrm{R}_{\mathcal{D}}(h_{\mathbf{w}})$. Indeed, $\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}})$ tends to $\mathrm{R}_{\mathcal{D}}(h_{\mathbf{w}})$ as $\|\mathbf{w}\|$ grows, which can provide *very*

8. The calculations leading to Equation (10) can be found in Langford (2005). For sake of completeness, we provide a slightly different derivation in Appendix B.

tight bounds (see the empirical analyses of Ambroladze et al., 2006; Germain et al., 2009a). Finally, the KL-divergence between $\rho_{\mathbf{w}}$ and $\pi_{\mathbf{0}}$ becomes simply

$$\begin{aligned} \mathrm{KL}\big(\rho_{\mathbf{w}} \| \pi_{\mathbf{0}}\big) &= \mathrm{KL}\big(\mathcal{N}(\mathbf{w}, \mathbf{I}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big) \\ &= \tfrac{1}{2}\|\mathbf{w}\|^2 \,, \end{aligned}$$

and turns out to be a measure of *complexity* of the learned classifier.

3.3.1 Objective Function and Gradient

Based on the specialization of the PAC-Bayesian theory to linear classifiers, Germain et al. (2009a) suggested minimizing a PAC-Bayesian bound on $\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}})$. For sake of completeness, we provide here more mathematical details than in the original conference paper (Germain et al., 2009a). In forthcoming Section 6, we will extend this supervised learning algorithm to the domain adaptation setting.

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

$$\begin{aligned} F(\mathbf{w}) &= \Omega\, m\, R_S(G_{\rho_{\mathbf{w}}}) + \mathrm{KL}(\rho_{\mathbf{w}} \| \pi_{\mathbf{0}}) \\ &= \Omega \sum_{i=1}^m \Phi_{\mathrm{R}}\left(y \frac{\mathbf{w} \cdot \mathbf{x}_i}{\|\mathbf{x}_i\|} \right) + \frac{1}{2}\|\mathbf{w}\|^2 \,. \end{aligned} \tag{13}$$

It turns out that the optimal vector $\mathbf{w}$ corresponds to the distribution $\rho_{\mathbf{w}}$ minimizing the value of the bound on $\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}})$ given by Theorem 3, with the parameter ω of the theorem being the hyperparameter Ω of the learning algorithm. It is important to point out that PAC-Bayesian theorems bound simultaneously $\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}})$ *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 3, a bound valid with probability $1 - \delta$.

The minimization of Equation (13) 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 $\nabla F(\mathbf{w})$:

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

where $\Phi'_{\mathrm{R}}(x) = -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$ is the derivative of Φ_{R} at point x.

Similarly to SVM, the learning algorithm PBGD3 realizes a trade-off between the empirical risk—expressed by the loss Φ_{R}—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 (13). If k is a Mercer kernel, it implicitly represents a function $\phi : X \to \mathbb{R}^{d'}$ that maps an example of $\mathbf{X}$ into an arbitrary d'-dimensional space, such that

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

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:

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

By the representer theorem (Schölkopf et al., 2001), the vector $\mathbf{w}$ minimizing Equation (13) can be recovered by finding the vector $\boldsymbol{\alpha}$ that minimizes

$$F(\boldsymbol{\alpha}) \;=\; C \sum_{i=1}^{m} \Phi_{\mathrm{R}}\left(y_i \frac{\sum_{j=1}^{m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}}\right) + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K_{i,j}\,, \tag{14}$$

where K is the kernel matrix of size $m \times m$.[9] That is, $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. The gradient of $F(\boldsymbol{\alpha})$ is simply given the vector $\nabla F(\boldsymbol{\alpha}) = (\alpha'_1, \alpha'_2, \ldots \alpha'_m)$, with

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

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 Φ_{R} of Equations (13) and (14) by its convex relaxation, which is

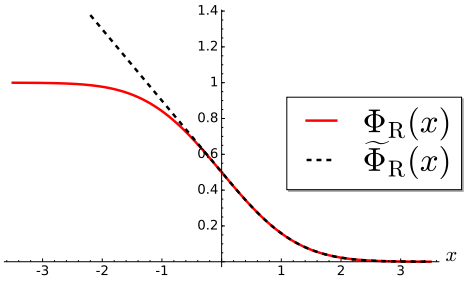
$$\begin{aligned} \widetilde{\Phi}_{\mathrm{R}}(x) \;&=\; \max\left\{\Phi_{\mathrm{R}}(x),\, \frac{1}{2} - \frac{x}{\sqrt{2\boldsymbol{\pi}}}\right\} \\ &=\; \begin{cases} \dfrac{1}{2} - \dfrac{x}{\sqrt{2\boldsymbol{\pi}}} & \text{if } x \leq 0, \\ \Phi_{\mathrm{R}}(x) & \text{otherwise.} \end{cases} \end{aligned} \tag{15}$$

The derivative of $\widetilde{\Phi}_{\mathrm{R}}$ at point x is then $\widetilde{\Phi}'_{\mathrm{R}}(x) = \Phi'_{\mathrm{R}}(\max\{0, x\})$, *i.e.*, $\widetilde{\Phi}'_{\mathrm{R}}(x) = -1/\sqrt{2\boldsymbol{\pi}}$ if $x < 0$, and $\Phi'_{\mathrm{R}}(x)$ otherwise. Figure 1a illustrates the functions Φ_{R} and $\widetilde{\Phi}_{\mathrm{R}}$. Note that the latter can be interpreted as a *smooth* version the SVM's hinge loss, $\max\{0, 1-x\}$. The toy experiment of Figure 1d (described in the next subsection) provides another empirical evidence that the minima of Φ_{R} and $\widetilde{\Phi}_{\mathrm{R}}$ tend to coincide.

3.3.4 Illustration on a Toy Dataset

To illustrate the trade-off coming into play in PBGD3 algorithm (and its convexified version), we conduct a small experiment on a two-dimensional toy dataset. That is, we generate

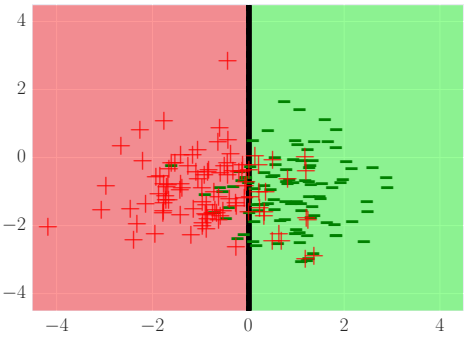
9. It is non-trivial to show that the kernel trick holds when $\pi_{\mathbf{0}}$ and $\rho_{\mathbf{w}}$ are Gaussian over infinite-dimensional feature space. As mentioned by McAllester and Keshet (2011), it is, however, the case provided we consider Gaussian processes as measure of distributions $\pi_{\mathbf{0}}$ and $\rho_{\mathbf{w}}$ over (infinite) $\mathcal{H}$. The same analysis holds for the kernelized versions of the two forthcoming domain adaptation algorithms (Section 6.3.3).



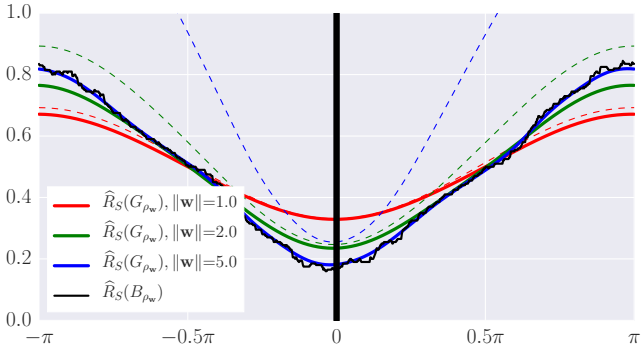
(a) Loss functions for linear classifiers.

FUNCTION	DERIVATIVE
$\Phi_{\mathrm{R}}(x) = \frac{1}{2}\big[1 - \mathrm{Erf}\big(\frac{x}{\sqrt{2}}\big)\big]$	$\Phi'_{\mathrm{R}}(x) = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$
$\widetilde{\Phi}_{\mathrm{R}}(x) = \max\big\{\Phi_{\mathrm{R}}(x), \frac{1}{2} - \frac{x}{\sqrt{2\pi}}\big\}$	$\widetilde{\Phi}'_{\mathrm{R}}(x) = \Phi'_{\mathrm{R}}\big(\max\{0, x\}\big)$

(b) Loss functions definitions and their derivatives.



(c) Toy dataset, and the decision boundary for $\theta = 0$ (matching the vertical line of Fig.(d)).



(d) Risk values according to θ, for $\|\mathbf{w}\| \in \{1, 2, 5\}$. Each dashed line shows convex counterpart of the continuous line of the same color.

Figure 1: Understanding PBGD3 supervised learning algorithm in terms of loss functions. Upper Figures (a-b) show the loss functions, and lower Figures (c-d) illustrate the behavior on a toy dataset.

100 positive examples according a Gaussian of mean $(-1, -1)$ and 100 negative examples generated by a Gaussian of mean $(-1, +1)$ (both of these Gaussian have a unit variance), as shown by Figure 1c. We then compute the risks associated with linear classifiers $h_{\mathbf{w}}$, with $\mathbf{w} = \|\mathbf{w}\|(\cos\theta, \sin\theta) \in \mathbb{R}^2$. Figure 1d shows the risks of three different classifiers for $\|\mathbf{w}\| \in \{1, 2, 5\}$, while rotating the decision boundary $\theta \in [-\pi, +\pi]$ around the origin. The 0-1-loss associated with the majority vote classifier $\widehat{\mathrm{R}}_S(B_{\rho_{\mathbf{w}}})$ does not rely on the norm $\|\mathbf{w}\|$. However, we clearly see that probit-loss of the Gibbs classifier $\widehat{\mathrm{R}}_S(B_{\rho_{\mathbf{w}}})$ converges to $\widehat{\mathrm{R}}_S(B_{\rho_{\mathbf{w}}})$ as $\|\mathbf{w}\|$ increases (the dashed lines correspond to the convex surrogate of the probit-loss given by Equation 15). Thus, thanks to the specialization of to the linear classifier, the *smoothness* of the surrogate loss is regularized by the norm $\|\mathbf{w}\|^2$.

4. Two New Domain Adaptation Bounds

The originality of our contribution is to theoretically design two domain adaptation frameworks suitable for the PAC-Bayesian approach. In Section 4.1, we first follow the spirit of the seminal works recalled in Section 2 by proving a similar trade-off for the Gibbs classifier. Then in Section 4.2, we propose a novel trade-off based on the specificities of the Gibbs classifier that come from Equation (7).

4.1 In the Spirit of the Seminal Works

In the following, while the domain adaptation bounds presented in Section 2 focus on a single classifier, we first define a ρ-average divergence measure to compare the marginals. This leads us to derive our first domain adaptation bound.

4.1.1 A Domains' Divergence for PAC-Bayesian Analysis

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. For the PAC-Bayesian setting, we propose a *domain disagreement pseudometric*[10] 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 spurred by the C-bound of Equation (4): Given a source domain $\mathcal{S}$, a target domain $\mathcal{T}$, and a posterior distribution ρ, if $\mathrm{R}_{\mathcal{S}}(G_\rho)$ and $\mathrm{R}_{\mathcal{T}}(G_\rho)$ are similar, then $\mathrm{R}_{\mathcal{S}}(B_\rho)$ and $\mathrm{R}_{\mathcal{T}}(B_\rho)$ are similar when $\mathrm{d}_{\mathcal{S}_\mathbf{X}}(\rho)$ and $\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ are also similar. Thus, the domains $\mathcal{S}$ and $\mathcal{T}$ are close according to ρ if the expected disagreement over the two domains tends to be close. We then define our pseudometric as follows.

Definition 4 *Let $\mathcal{H}$ be a hypothesis class. For any marginal distributions $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$ over $\mathbf{X}$, any distribution ρ on $\mathcal{H}$, the domain disagreement* $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ *between $\mathcal{S}_\mathbf{X}$ and $\mathcal{T}_\mathbf{X}$ is defined by*

$$\begin{aligned}\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X}) &= \left| \mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho) - \mathrm{d}_{\mathcal{S}_\mathbf{X}}(\rho) \right| \\ &= \left| \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left[\mathrm{R}_{\mathcal{T}_\mathbf{X}}(h,h') - \mathrm{R}_{\mathcal{S}_\mathbf{X}}(h,h') \right] \right| .\end{aligned}$$

Note that dis_ρ 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, $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ depends on the majority vote, which suggests that we can directly minimize it via its empirical counterpart. This can be done without instance reweighting, space representation changing or family of classifiers modification. On the contrary, $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ is a supremum over all $h \in \mathcal{H}$ and hence, does not depend on the classifier on which the risk is considered. Moreover, $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ (the ρ-average) is lower than the $\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ (the worst case). Indeed, for every $\mathcal{H}$ and

10. A pseudometric d is a metric for which the property $d(x,y) = 0 \iff x = y$ is relaxed to $d(x,y) = 0 \impliedby x = y$.

ρ over $\mathcal{H}$, we have

$$\begin{aligned}\tfrac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) &= \sup_{(h,h')\in\mathcal{H}^2} \left|\mathrm{R}_{\mathcal{T}_{\mathbf{X}}}(h,h') - \mathrm{R}_{\mathcal{S}_{\mathbf{X}}}(h,h')\right| \\ &\geq \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \left|\mathrm{R}_{\mathcal{T}_{\mathbf{X}}}(h,h') - \mathrm{R}_{\mathcal{S}_{\mathbf{X}}}(h,h')\right| \\ &\geq \mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})\,.\end{aligned}$$

4.1.3 A domain adaptation bound for the stochastic Gibbs classifier

We now derive our first main result in the following theorem: A domain adaptation bound relevant in a PAC-Bayesian setting, and that relies on the domain disagreement of Definition 4.

Theorem 5 *Let $\mathcal{H}$ be a hypothesis class. We have*

$$\forall \rho \text{ on } \mathcal{H},\ \mathrm{R}_{\mathcal{T}}(G_\rho) \leq \mathrm{R}_{\mathcal{S}}(G_\rho) + \frac{1}{2}\,\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \lambda_\rho\,,$$

where λ_ρ is the deviation between the expected joint errors (Equation 6) of G_ρ on the target and source domains:

$$\lambda_\rho = \Big|\,\mathrm{e}_{\mathcal{T}}(\rho) - \mathrm{e}_{\mathcal{S}}(\rho)\,\Big|\,. \tag{16}$$

Proof First, from Equation (7), we recall that, given a domain $\mathcal{D}$ on $\mathbf{X}\times Y$ and a distribution ρ over $\mathcal{H}$, we have:

$$\mathrm{R}_{\mathcal{D}}(G_\rho) = \frac{1}{2}\mathrm{d}_{\mathcal{D}_{\mathbf{X}}}(\rho) + \mathrm{e}_{\mathcal{D}}(\rho)\,.$$

Therefore,

$$\begin{aligned}\mathrm{R}_{\mathcal{T}}(G_\rho) - \mathrm{R}_{\mathcal{S}}(G_\rho) &= \frac{1}{2}\Big(\mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho) - \mathrm{d}_{\mathcal{S}_{\mathbf{X}}}(\rho)\Big) + \Big(\mathrm{e}_{\mathcal{T}}(\rho) - \mathrm{e}_{\mathcal{S}}(\rho)\Big) \\ &\leq \frac{1}{2}\Big|\mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho) - \mathrm{d}_{\mathcal{S}_{\mathbf{X}}}(\rho)\Big| + \Big|\mathrm{e}_{\mathcal{T}}(\rho) - \mathrm{e}_{\mathcal{S}}(\rho)\Big| \\ &= \frac{1}{2}\,\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + \lambda_\rho\,.\end{aligned}$$

■

4.1.4 Meaningful quantities

Similarly than the bounds of Theorems 1 and 2, our bound can be seen as a trade-off between different quantities. Concretely, the terms $\mathrm{R}_{\mathcal{S}}(G_\rho)$ and $\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ are akin to the first two terms of the domain adaptation bound of Theorem 1: $\mathrm{R}_{\mathcal{S}}(G_\rho)$ is the ρ-average risk over $\mathcal{H}$ on the source domain, and $\mathrm{dis}_\rho(\mathcal{T}_{\mathbf{X}}, \mathcal{S}_{\mathbf{X}})$ measures the ρ-average disagreement between the marginals but is specific to the current model depending on 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 is that the above theorem improves the one we proposed in Germain et al. (2013) in two points[11]. On the one hand, this bound is not degenerated when the source and target distributions are the same or close. On the other hand, our result contains only the half of $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X}, \mathcal{T}_\mathbf{X})$ contrary to our first bound proposed in Germain et al. (2013). Finally, due to the dependence of $\mathrm{dis}_\rho(\mathcal{T}_\mathbf{X}, \mathcal{S}_\mathbf{X})$ and λ_ρ on the learned posterior, our bound is, in general incomparable with the ones of Theorems 1 and 2. However, it brings the same underlying idea: Supposing that the two domains are sufficiently related, one must look for a model that minimizes a trade-off between its source risk and a distance between the domains' marginal.

4.2 A Novel Perspective on Domain Adaptation

In this section, we introduce an original approach to upper-bound the non-estimable risk of a ρ-weighted majority vote on a target domain $\mathcal{T}$ thanks to a term depending on its marginal distribution $\mathcal{T}_\mathbf{X}$, another one on a related source domain $\mathcal{S}$, and a term capturing the "volume" of the source distribution uninformative for the target task. We base our bound on Equation (7) (recalled below) that decomposes the Gibbs classifier into the trade-off between the half of the expected disagreement $\mathrm{d}_{\mathcal{D}_\mathbf{X}}(\rho)$ of Equation (5) and the expected joint error $\mathrm{e}_\mathcal{D}(\rho)$ of Equation (6):

$$\mathrm{R}_\mathcal{D}(G_\rho) \;=\; \tfrac{1}{2}\,\mathrm{d}_{\mathcal{D}_\mathbf{X}}(\rho) + \mathrm{e}_\mathcal{D}(\rho)\,. \tag{7}$$

A key observation is that the *voters' disagreement does not rely on labels*; we can compute $\mathrm{d}_{\mathcal{D}_\mathbf{X}}(\rho)$ using the marginal distribution $\mathcal{D}_\mathbf{X}$. Thus, in the present domain adaptation context, we have access to $\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ even if the target labels are unknown. However, the expected joint error can only be computed on the labeled source domain, that is what we kept in mind to define our new domains divergence.

4.2.1 Another Domains Divergence for the PAC-Bayesian Approach

We design a domains' divergence that allows us to link the target joint error $\mathrm{e}_\mathcal{T}(\rho)$ with the source one $\mathrm{e}_\mathcal{S}(\rho)$ by reweighting the latter. This new divergence is called the β_q-divergence and is parametrized by a real value $q > 0$:

$$\beta_q(\mathcal{T}\|\mathcal{S}) \;=\; \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right)^q\right]^{\frac{1}{q}}. \tag{17}$$

It is worth noting that considering some q values allow us to recover well-known divergences. For instance, choosing $q=2$ relates our result to the χ^2-distance, between the domains as $\beta_2(\mathcal{T}\|\mathcal{S}) = \sqrt{\chi^2(\mathcal{T}\|\mathcal{S})+1}$. Moreover, we can link $\beta_q(\mathcal{T}\|\mathcal{S})$ to the Rényi divergence[12], which

11. More details are given in our research report (Germain et al., 2015a).

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

has led to generalization bounds in the specific context of importance weighting (Cortes et al., 2010). We denote the limit case $q \to \infty$ by

$$\beta_\infty(\mathcal{T}\|\mathcal{S}) \;=\; \sup_{(\mathbf{x},y)\in \textsc{supp}(\mathcal{S})} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right), \tag{18}$$

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

$$\eta_{\mathcal{T}\backslash\mathcal{S}} = \Pr_{(\mathbf{x},y)\sim\mathcal{T}}\Big((\mathbf{x},y) \notin \textsc{supp}(\mathcal{S})\Big) \sup_{h\in\mathcal{H}} \mathrm{R}_{\mathcal{T}\backslash\mathcal{S}}(h)\,. \tag{19}$$

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

4.2.2 A Novel Domain Adaptation Bound

Let us state the result underlying the novel domain adaptation perspective of this paper.

Theorem 6 *Let $\mathcal{H}$ be a hypothesis space, let $\mathcal{S}$ and $\mathcal{T}$ respectively be the source and the target domains on $\mathbf{X}\times Y$. Let $q>0$ be a constant. We have, for all ρ on $\mathcal{H}$,*

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

where $\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho)$, $\mathrm{e}_{\mathcal{S}}(\rho)$, $\beta_q(\mathcal{T}\|\mathcal{S})$ and $\eta_{\mathcal{T}\backslash\mathcal{S}}$ are respectively defined by Equations (5), (6), (17) *and* (19)*.*

Proof From Equation (7), we know that $\mathrm{R}_{\mathcal{T}}(G_\rho) = \frac{1}{2}\,\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho) + \mathrm{e}_{\mathcal{T}}(\rho)$. Let us split $\mathrm{e}_{\mathcal{T}}(\rho)$ in two parts:

$$\begin{aligned}
\mathrm{e}_{\mathcal{T}}(\rho) \;&=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}),y\right) \mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}),y\right) \\
&=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}),y\right) \mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}),y\right) \qquad (20)\\
&\quad + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}} \mathrm{I}\left[(\mathbf{x},y) \notin \textsc{supp}(\mathcal{S})\right] \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}),y\right) \mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}),y\right). \qquad (21)
\end{aligned}$$

(i) On the one hand, we upper-bound the first part (Line 20) using the Hölder's inequality (see Lemma 15 in Appendix A), with p such that $\frac{1}{p}=1-\frac{1}{q}$:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)} \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}),y\right) \mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}),y\right) \\
&\quad\le\; \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left(\frac{\mathcal{T}(\mathbf{x},y)}{\mathcal{S}(\mathbf{x},y)}\right)^q\right]^{\frac{1}{q}} \left[\mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{S}} \left[\mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}),y\right) \mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}),y\right)\right]^p\right]^{\frac{1}{p}} \\
&\quad= \beta_q(\mathcal{T}\|\mathcal{S}) \left[\mathrm{e}_{\mathcal{S}}(\rho)\right]^{\frac{1}{p}},
\end{aligned}$$

where we have removed the exponent from expression $[\mathcal{L}_{0\text{-}1}(h(\mathbf{x}), y)\mathcal{L}_{0\text{-}1}(h'(\mathbf{x}), y)]^p$ without affecting its value, which is either 1 or 0.

(ii) On the other hand, we upper-bound the second part (Line 21) by the term $\eta_{\mathcal{T}\setminus\mathcal{S}}$:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}}\left(\mathrm{I}\left[(\mathbf{x},y)\notin \textsc{supp}(\mathcal{S})\right]\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}),y\right)\mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}),y\right)\right)\\
&\quad= \left(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}}\mathrm{I}[(\mathbf{x},y)\notin\textsc{supp}(\mathcal{S})]\right)\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}\setminus\mathcal{S}}\mathop{\mathbf{E}}_{(h,h')\sim\rho^2}\mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}),y\right)\,\mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}),y\right)\\
&\quad= \left(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}}\mathrm{I}[(\mathbf{x},y)\notin\textsc{supp}(\mathcal{S})]\right)\mathrm{e}_{\mathcal{T}\setminus\mathcal{S}}(\rho)\\
&\quad= \left(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}}\mathrm{I}[(\mathbf{x},y)\notin\textsc{supp}(\mathcal{S})]\right)\left(\mathrm{R}_{\mathcal{T}\setminus\mathcal{S}}(G_\rho)-\tfrac{1}{2}\mathrm{d}_{\mathcal{T}\setminus\mathcal{S}}(\rho)\right)\\
&\quad\le \left(\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{T}}\mathrm{I}[(\mathbf{x},y)\notin\textsc{supp}(\mathcal{S})]\right)\sup_{h\in\mathcal{H}}\mathrm{R}_{\mathcal{T}\setminus\mathcal{S}}(h)\;=\;\eta_{\mathcal{T}\setminus\mathcal{S}}\,.
\end{aligned}$$

■

Note that the bound of Theorem 6 is reached whenever the domains are equal ($\mathcal{S}=\mathcal{T}$). Thus, when adaptation is not necessary, our analysis is still sound and non-degenerated:

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

4.2.3 Meaningful Quantities and Connection with Some Domain Adaptation Assumptions

Similarly to the previous results recalled in Section 2, our domain adaptation theorem bounds the target risk by a sum of three terms. However, our approach breaks the problem into *atypical* quantities:

(i) The expected disagreement $\mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho)$ captures *second degree* information about the target domain (without any label).

(ii) The β_q-divergence $\beta_q(\mathcal{T}\|\mathcal{S})$ is not an additional term: it weighs the influence of the expected joint error $\mathrm{e}_{\mathcal{S}}(\rho)$ of the source domain; the parameter q allows us to consider different relationships between $\beta_q(\mathcal{T}\|\mathcal{S})$ and $\mathrm{e}_{\mathcal{S}}(\rho)$.

(iii) The term $\eta_{\mathcal{T}\setminus\mathcal{S}}$ quantifies the worst feasible target error on the regions where the source domain is uninformative for the target one. In the current work, we assume that this area is small.

We now establish some connections with existing common domain adaptation assumptions in the literature. Recall that in order to characterize which domain adaptation task may be *learnable*, Ben-David et al. (2012) presented three *assumptions that can help domain adaptation*. Our Theorem 6 does not rely on these assumptions, and remains valid in the absence of these assumptions, but they can be interpreted in our framework as discussed below.

On the covariate shift. A domain adaptation task fulfills the *covariate shift* assumption (Shimodaira, 2000) if the source and target domains only differ in their marginals according to the input space, *i.e.,* $\mathcal{T}_{Y|\mathbf{x}}(y) = \mathcal{S}_{Y|\mathbf{x}}(y)$. In this scenario, one may estimate the values of $\beta_q(\mathcal{T}_{\mathbf{X}} \| \mathcal{S}_{\mathbf{X}})$, and even $\eta_{\mathcal{T}\setminus\mathcal{S}}$, by using unsupervised density estimation methods. Interestingly, with the additional assumption that the domains share the same support, we have $\eta_{\mathcal{T}\setminus\mathcal{S}} = 0$. Then from Line (20) we obtain

$$\mathrm{R}_{\mathcal{T}}(G_\rho) \;=\; \tfrac{1}{2}\mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho) + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}} \frac{\mathcal{T}_{\mathbf{X}}(\mathbf{x})}{\mathcal{S}_{\mathbf{X}}(\mathbf{x})} \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{h'\sim\rho} \mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}), y\right) \mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}), y\right),$$

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

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

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

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

On the cluster assumption. A target domain fulfills the *cluster assumption* when examples of the same label belong to a common "area" of the input space, and the differently labeled "areas" are well separated by *low-density regions* (formalized by the *probabilistic Lipschitzness* of Urner et al., 2011). Once specialized to linear classifiers, $\mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho)$ behaves nicely in this context (see Section 6).

On representation learning. The main assumption underlying our domain adaptation algorithm exhibited in Section 6 is that the support of the target domain is mostly included in the support of the source domain, *i.e.*, the value of the term $\eta_{\mathcal{T}\setminus\mathcal{S}}$ is small. In situations when $\mathcal{T}\setminus\mathcal{S}$ is sufficiently large to prevent proper adaptation, one could try to reduce its volume while taking care to preserve a good compromise between $\mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho)$ and $\mathrm{e}_{\mathcal{S}}(\rho)$, using a *representation learning* approach, *i.e.*, by projecting source and target examples into a new common input space, as done for example by Chen et al. (2012); Ganin et al. (2016).

4.3 Comparison of the Two Domain Adaptation Bounds

Since they rely on different approximations, the gap between the bounds of Theorems 5 and 6 varies according to the context. As presented in Sections 4.1.4 and 4.2.3, the main difference between our two bounds lies in the estimable terms, from which we will derive algorithms in Section 6. In Theorem 6, the non-estimable terms are the domains' divergence $\beta_q(\mathcal{T}\|\mathcal{S})$ and the term $\eta_{\mathcal{T}\setminus\mathcal{S}}$. Contrary to the non-controllable term λ_ρ of Theorem 5, these terms do not depend on the *learned* posterior distribution ρ: For every ρ on $\mathcal{H}$, $\beta_q(\mathcal{T}\|\mathcal{S})$

and $\eta_{\mathcal{T}\setminus\mathcal{S}}$ are constant values measuring the relation between the domains for the considered task. Moreover, the fact that the β_q-divergence is not an additive term but a multiplicative one (as opposed to $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X})+\lambda_\rho$ in Theorem 5) is a contribution of our new perspective. Consequently, $\beta_q(\mathcal{T}\|\mathcal{S})$ can be viewed as a hyperparameter allowing us to tune the trade-off between the target voters' disagreement and the source joint error. Experiments of Section 7 confirm that this hyperparameter can be successfully selected.

Note that we can upper-bound the term λ_ρ of Theorem 5 by using the same trick as in Theorem 6 proof. This leads to

$$\begin{aligned}\lambda_\rho \;=\; |\mathrm{e}_\mathcal{T}(\rho)-\mathrm{e}_\mathcal{S}(\rho)| &\le \left|\beta_q(\mathcal{T}\|\mathcal{S})\times\left[\mathrm{e}_\mathcal{S}(\rho)\right]^{1-\frac{1}{q}}+\eta_{\mathcal{T}\setminus\mathcal{S}}-\mathrm{e}_\mathcal{S}(\rho)\right| \\ &\le \left|\beta_q(\mathcal{T}\|\mathcal{S})\times\left[\mathrm{e}_\mathcal{S}(\rho)\right]^{1-\frac{1}{q}}-\mathrm{e}_\mathcal{S}(\rho)\right|+\eta_{\mathcal{T}\setminus\mathcal{S}}\,.\end{aligned}$$

Thus, we can rewrite Theorem 5 statement as

$$\forall\rho \text{ on } \mathcal{H},\;\; \mathrm{R}_\mathcal{T}(G_\rho) \;\le\; \mathrm{R}_\mathcal{S}(G_\rho)+\frac{1}{2}\,\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X})+\left|\beta_q(\mathcal{T}\|\mathcal{S})\times\left[\mathrm{e}_\mathcal{S}(\rho)\right]^{1-\frac{1}{q}}-\mathrm{e}_\mathcal{S}(\rho)\right|+\eta_{\mathcal{T}\setminus\mathcal{S}}\,.$$

It turns out that, if $\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho)\ge\mathrm{d}_{\mathcal{S}_\mathbf{X}}(\rho)$ and $\beta_q(\mathcal{T}\|\mathcal{S})\times\left[\mathrm{e}_\mathcal{S}(\rho)\right]^{1-\frac{1}{q}}\ge\mathrm{e}_\mathcal{S}(\rho)$, the above statement reduces to the one of Theorem 6. In words, this occurs in the very particular case where the target disagreement is greater than the source one and when the density of the target instances is superior to source ones on the source support which may be interpreted as a rather favorable situation. However, Theorem 6 is tighter in all other cases. This highlights that introducing absolute values in Theorem 5 proof leads to a crude approximation. Remember that we have first followed this path to stay aligned with classical domain adaptation analysis, but our second approach leads to a more suitable analysis in a PAC-Bayesian context. Our experiments of Subsection 6.4 illustrate this empirically, once the domain adaptation bounds are converted into PAC-Bayesian generalization guarantees for linear classifiers.

5. PAC-Bayesian Generalization Guarantees

To compute our domain adaptation bounds, one needs to know the distributions $\mathcal{S}$ and $\mathcal{T}_\mathbf{X}$, which is never the case in real life tasks. PAC-Bayesian theory provides tools to convert the bounds of Theorems 5 and 6 into generalization bounds on the target risk computable from a pair of source-target samples $(S,T)\sim(\mathcal{S})^{m_s}\times(\mathcal{T}_\mathbf{X})^{m_t}$. To achieve this goal, we first provide generalization guarantees for the terms involved in our domain adaptation bounds: $\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho)$, $\mathrm{e}_\mathcal{S}(\rho)$, and $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X})$. These results are presented as corollaries of Theorem 7 below, that generalizes the PAC-Bayesian of Catoni (2007) (see Theorem 3 in Section 3.2) to arbitrary loss functions. Indeed, Theorem 7, with $\ell(h,\mathbf{x},y)=\mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}),y\right)$ and Equation (3), gives the usual bound on the Gibbs risk.

Note that the proofs of Theorem 7 (deferred in Appendix C) and Corollary 8 (below) reuse techniques from related results presented in Germain et al. (2015b). Indeed, PAC-Bayesian bounds on $\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and $\mathrm{e}_\mathcal{S}(\rho)$ appeared in the latter, but under different forms.

Theorem 7 *For any domain $\mathcal{D}$ over $\mathbf{X} \times Y$, for any set of hypotheses $\mathcal{H}$, any prior π over $\mathcal{H}$, any loss $\ell : \mathcal{H}\times\mathbf{X}\times Y \to [0,1]$, any real number $\alpha > 0$, with a probability at least $1-\delta$ over the random choice of $\{(\mathbf{x}_i, y_i)\}_{i=1}^m \sim (\mathcal{D})^m$, we have, for all ρ on $\mathcal{H}$,*

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

We now exploit Theorem 7 to obtain generalization guarantees on the expected disagreement, the expected joint error, and the domain disagreement. In Corollary 8 below, we are especially interested in the possibility of controlling the trade-off—between the empirical estimate computed on the samples and the complexity term $\mathrm{KL}(\rho\|\pi)$—with the help of parameters a, b and c.

Corollary 8 *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X} \times Y$, any set of voters $\mathcal{H}$, any prior π over $\mathcal{H}$, any $\delta \in (0,1]$, any real numbers $a > 0$, $b > 0$ and $c > 0$, we have:*
— with a probability at least $1-\delta$ over $T \sim (\mathcal{T}_\mathbf{X})^{m_t}$,

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

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

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

— with a probability at least $1-\delta$ over $S\times T \sim (\mathcal{S}_\mathbf{X}\times\mathcal{T}_\mathbf{X})^m$,

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

where $\widehat{\mathrm{d}}_T(\rho)$, $\widehat{\mathrm{e}}_S(\rho)$, and $\widehat{\mathrm{dis}}_\rho(S,T)$ are the empirical estimations of the target voters' disagreement, the source joint error, and the domain disagreement.

Proof Given π and ρ over $\mathcal{H}$, we consider a new prior π^2 and a new posterior ρ^2, both over $\mathcal{H}^2$, such that: $\forall\, h_{ij} = (h_i, h_j) \in \mathcal{H}^2$, $\pi^2(h_{ij}) = \pi(h_i)\pi(h_j)$,and $\rho^2(h_{ij}) = \rho(h_i)\rho(h_j)$. Thus, $\mathrm{KL}(\rho^2\|\pi^2) = 2\,\mathrm{KL}(\rho^2\|\pi^2)$ (see Lemma 21 in Appendix A). Let us define four new loss functions for a "paired voter" $h_{ij} \in \mathcal{H}^2$:

$$\begin{aligned}
\ell_d(h_{ij},\mathbf{x},y) &= \mathcal{L}_{0\text{-}1}\left(h_i(\mathbf{x}), h_j(\mathbf{x})\right),\\
\ell_e(h_{ij},\mathbf{x},y) &= \mathcal{L}_{0\text{-}1}\left(h_i(\mathbf{x}), y\right) \times \mathcal{L}_{0\text{-}1}\left(h_j(\mathbf{x}), y\right),\\
\ell_{d^{(1)}}(h_{ij},(\mathbf{x}^s,\mathbf{x}^t),\cdot) &= \frac{1 + \mathcal{L}_{0\text{-}1}\left(h_i(\mathbf{x}^s), h_j(\mathbf{x}^s)\right) - \mathcal{L}_{0\text{-}1}\left(h_i(\mathbf{x}^t), h_j(\mathbf{x}^t)\right)}{2},\\
\ell_{d^{(2)}}(h_{ij},(\mathbf{x}^s,\mathbf{x}^t),\cdot) &= \frac{1 + \mathcal{L}_{0\text{-}1}\left(h_i(\mathbf{x}^t), h_j(\mathbf{x}^t)\right) - \mathcal{L}_{0\text{-}1}\left(h_i(\mathbf{x}^s), h_j(\mathbf{x}^s)\right)}{2}.
\end{aligned}$$

Thus, from Theorem 7:

- The bound on $\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ is obtained with $\ell \coloneqq \ell_d$, and Equation (5);

- The bound on $\mathrm{e}_{\mathcal{S}}(\rho)$ is similarly obtained with $\ell \coloneqq \ell_e$, and Equation (6);
- The bound on $\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ is obtained with $\ell \coloneqq \ell_{d^{(1)}}$, by upper-bounding

$$d^{(1)} = \mathrm{d}_{\mathcal{S}_{\mathbf{X}}}(\rho) - \mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho) = 2 \mathop{\mathbf{E}}_{h_{ij}\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x}^s,\mathbf{x}^t)\sim\mathcal{S}_{\mathbf{X}}\times\mathcal{T}_{\mathbf{X}}} \ell_{d^{(1)}}(h_{ij}, (\mathbf{x}^s, \mathbf{x}^t), \cdot) - 1\,,$$

from its empirical counterpart

$$\widehat{d}^{(1)} = \widehat{\mathrm{d}}_S(\rho) - \widehat{\mathrm{d}}_T(\rho) = \frac{2}{m} \mathop{\mathbf{E}}_{h_{ij}\sim\rho^2} \sum_{k=1}^{m} \ell_{d^{(1)}}(h_{ij}, (\mathbf{x}^s_k, \mathbf{x}^t_k), \cdot) - 1\,.$$

We then have, with probability $1 - \frac{\delta}{2}$ over the choice of $S \times T \sim (\mathcal{S}_{\mathbf{X}} \times \mathcal{T}_{\mathbf{X}})^m$,

$$\frac{|d^{(1)}| + 1}{2} \;\leq\; \frac{a}{1 - e^{-2a}} \left[|\widehat{d}^{(1)}| + 1 + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}}{m \times a} \right].$$

In turn, with $\ell \coloneqq \ell_{d^{(2)}}$ we bound $d^{(2)} = \mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho) - \mathrm{d}_{\mathcal{S}_{\mathbf{X}}}(\rho)$ by its empirical counterpart $\widehat{d}^{(2)} = \widehat{\mathrm{d}}_T(\rho) - \widehat{\mathrm{d}}_S(\rho)$, with probability $1 - \frac{\delta}{2}$ over the choice of $S \times T \sim (\mathcal{S}_{\mathbf{X}} \times \mathcal{T}_{\mathbf{X}})^m$,

$$\frac{|d^{(2)}| + 1}{2} \;\leq\; \frac{a}{1 - e^{-2a}} \left[|\widehat{d}^{(2)}| + 1 + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}}{m \times a} \right].$$

Finally, by the union bound, with probability $1 - \delta$, we have

$$\mathrm{dis}_\rho(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) \;=\; \max\left\{ d^{(1)}, d^{(2)} \right\} \leq \frac{2a}{1 - e^{-2a}} \left[\widehat{\mathrm{dis}}_\rho(S,T) + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{\delta}}{m \times a} + 1 \right] - 1\,,$$

and we are done. ■

The following bound is based on the above Catoni's approach for our domain adaptation bound of Theorem 5 and corresponds to the one from which we derive—in Section 6—PBDA our first algorithm for PAC-Bayesian domain adaptation.

Theorem 9 *For any domains $\mathcal{S}$ and $\mathcal{T}$ (resp. with marginals $\mathcal{S}_{\mathbf{X}}$ and $\mathcal{T}_{\mathbf{X}}$) over $\mathbf{X} \times Y$, any set of hypotheses $\mathcal{H}$, any prior distribution π over $\mathcal{H}$, any $\delta \in (0, 1]$, any real numbers $\omega > 0$ and $a > 0$, with a probability at least $1 - \delta$ over the choice of $S \times T \sim (\mathcal{S} \times \mathcal{T}_{\mathbf{X}})^m$, for every posterior distribution ρ on $\mathcal{H}$, we have*

$$\mathrm{R}_{\mathcal{T}}(G_\rho) \;\leq\; \omega' \widehat{\mathrm{R}}_S(G_\rho) + a' \tfrac{1}{2} \widehat{\mathrm{dis}}_\rho(S,T) + \left(\frac{\omega'}{\omega} + \frac{a'}{a} \right) \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{\delta}}{m} + \lambda_\rho + \tfrac{1}{2}(a' - 1)\,,$$

where $\widehat{\mathrm{R}}_S(G_\rho)$ and $\widehat{\mathrm{dis}}_\rho(S,T)$ are the empirical estimates of the target risk and the domain disagreement; λ_ρ is defined by Equation (16); $\omega' = \frac{\omega}{1 - e^{-\omega}}$ and $a' = \frac{2a}{1 - e^{-2a}}$.

Proof In Theorem 5, we replace $\widehat{\mathrm{R}}_S(G_\rho)$ and $\widehat{\mathrm{dis}}_\rho(S,T)$ by their upper bound, obtained from Theorem 3 and Corollary 8, with δ chosen respectively as $\frac{\delta}{3}$ and $\frac{2\delta}{3}$. In the latter case, we use $2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{2}{2\delta/3} = 2\,\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{\delta} < 2\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{3}{\delta}\right)$. ■

We now derive a PAC-Bayesian generalization bound for our second domain adaptation bound of Theorem 6 from which we derive—in Section 6—our second algorithm for PAC-Bayesian domain adaptation DALC. For algorithmic simplicity, we deal with Theorem 6 when $q\to\infty$. Thanks to Corollary 8, we obtain the following generalization bound defined with respect to the empirical estimates of the target disagreement and the source joint error.

Theorem 10 *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}\times Y$, any set of voters $\mathcal{H}$, any prior π over $\mathcal{H}$, any $\delta\in(0,1]$, any real numbers $b>0$ and $c>0$, with a probability at least $1-\delta$ over the choices of $S\sim(\mathcal{S})^{m_s}$ and $T\sim(\mathcal{T}_\mathbf{X})^{m_t}$, for every posterior distribution ρ on $\mathcal{H}$, we have*

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

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

Proof We bound separately $\mathrm{d}_{\mathcal{T}_\mathbf{X}}(\rho)$ and $\mathrm{e}_\mathcal{S}(\rho)$ using Corollary 8 (with probability $1-\frac{\delta}{2}$ each), and then combine the two upper bounds according to Theorem 6. ■

From an optimization perspective, the problem suggested by the bound of Theorem 10 is much more convenient to minimize than the PAC-Bayesian bound derived in Theorem 9. The former is *smoother* than the latter: The absolute value related to the domain disagreement $\mathrm{dis}_\rho(\mathcal{S}_\mathbf{X},\mathcal{T}_\mathbf{X})$ disappears in benefit of the domain divergence $\beta_\infty(\mathcal{T}\|\mathcal{S})$, which is constant and can be considered as an hyperparameter of the algorithm. Additionally, Theorem 9 requires equal source and target sample sizes while Theorem 10 allows $m_s\neq m_t$. Moreover, for algorithmic purposes, we ignore the ρ-dependent non-constant term λ_ρ of Theorem 9. In our second analysis, such compromise is not mandatory in order to apply the theoretical result to real problems, since the non-estimable term $\eta_{\mathcal{T}\setminus\mathcal{S}}$ is constant and does not depend on the learned ρ. Hence, we can neglect $\eta_{\mathcal{T}\setminus\mathcal{S}}$ without any impact on the optimization problem described in the next section. Besides, it is realistic to consider $\eta_{\mathcal{T}\setminus\mathcal{S}}$ as a small quantity in situations where the source and target supports are similar.

6. PAC-Bayesian Domain Adaptation Learning of Linear Classifiers

In this section, we design two learning algorithms for domain adaptation[13] 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 classifiers described in Section 3.3. The taken approach is the one privileged in numerous PAC-Bayesian works (*e.g.*, Langford

13. The code of our algorithms are available on-line. See `http://graal.ift.ulaval.ca/pbda` and `https://github.com/GRAAL-Research/domain_adaptation_of_linear_classifiers`.

and Shawe-Taylor, 2002; Ambroladze et al., 2006; McAllester and Keshet, 2011; Parrado-Hernández et al., 2012; Germain et al., 2009a, 2013), as it makes the risk of the linear classifier $h_{\mathbf{w}}$ and the risk of a (properly parametrized) majority vote coincide, while in the same time promoting large margin classifiers.

6.1 Domain disagreement, Expected Disagreement, Joint Error of Linear Classifiers

Let us 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. We seek to express the *domain disagreement* $\mathrm{dis}_{\rho_{\mathbf{w}}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$, *expected disagreement* $\mathrm{d}_{\mathcal{D}_{\mathbf{X}}}(\rho_{\mathbf{w}})$ and the *expected joint error* $\mathrm{e}_{\mathcal{D}}(\rho_{\mathbf{w}})$.

First, for any marginal $\mathcal{D}_{\mathbf{X}}$, the expected disagreement for linear classifiers is:

$$\begin{aligned}
\mathrm{d}_{\mathcal{D}_{\mathbf{X}}}(\rho_{\mathbf{w}}) &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}}^2} \mathcal{L}_{0\text{-}1}\big(h(\mathbf{x}), h'(\mathbf{x})\big) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}}^2} \mathrm{I}[h(\mathbf{x}) \neq h'(\mathbf{x})] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}}^2} \Big(\mathrm{I}[h(\mathbf{x}) = 1]\,\mathrm{I}[h'(\mathbf{x}) = -1] + \mathrm{I}[h(\mathbf{x}) = -1]\,\mathrm{I}[h'(\mathbf{x}) = 1]\Big) \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \mathop{\mathbf{E}}_{(h,h')\sim\rho_{\mathbf{w}}^2} \mathrm{I}[h(\mathbf{x}) = 1]\,\mathrm{I}[h'(\mathbf{x}) = -1] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \mathop{\mathbf{E}}_{h\sim\rho_{\mathbf{w}}} \mathrm{I}[h(\mathbf{x}) = 1] \mathop{\mathbf{E}}_{h'\sim\rho_{\mathbf{w}}} \mathrm{I}[h'(\mathbf{x}) = -1] \\
&= 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \Phi_{\mathrm{R}}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right) \Phi_{\mathrm{R}}\left(-\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{X}}} \Phi_{\mathrm{d}}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right), \qquad (22)
\end{aligned}$$

where

$$\Phi_{\mathrm{d}}(x) \;=\; 2\,\Phi_{\mathrm{R}}(x)\,\Phi_{\mathrm{R}}(-x)\,. \qquad (23)$$

Thus, the domain disagreement for linear classifiers is

$$\begin{aligned}
\mathrm{dis}_{\rho_{\mathbf{w}}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) &= \Big|\, \mathrm{d}_{\mathcal{S}_{\mathbf{X}}}(\rho_{\mathbf{w}}) - \mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho_{\mathbf{w}}) \,\Big| \\
&= \left| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}_{\mathbf{X}}} \Phi_{\mathrm{d}}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right) - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}_{\mathbf{X}}} \Phi_{\mathrm{d}}\left(\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right) \right|. \qquad (24)
\end{aligned}$$

Following a similar approach, the expected joint error is, for all $\mathbf{w} \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{e}_{\mathcal{D}}(\rho_{\mathbf{w}}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho_{\mathbf{w}}} \mathop{\mathbf{E}}_{h'\sim\rho_{\mathbf{w}}} \mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}), y\right) \times \mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}), y\right) \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \mathop{\mathbf{E}}_{h\sim\rho_{\mathbf{w}}} \mathcal{L}_{0\text{-}1}\left(h(\mathbf{x}), y\right) \mathop{\mathbf{E}}_{h'\sim\rho_{\mathbf{w}}} \mathcal{L}_{0\text{-}1}\left(h'(\mathbf{x}), y\right) \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \Phi_{\mathrm{e}}\left(y\,\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right), \qquad (25)
\end{aligned}$$

with

$$\Phi_{\mathrm{e}}(x) \;=\; \big[\Phi_{\mathrm{R}}(x)\big]^2\,. \qquad (26)$$

Functions Φ_{e} and Φ_{d} defined above can be interpreted as loss functions for linear classifiers (illustrated by Figure 2a).

6.2 Domain Adaptation Bounds

Theorem 5 and 6 (with $q{\to}\infty$) specialized to linear classifiers give the two following corollaries. Note that, as mentioned before, $\mathrm{R}_{\mathcal{T}}(h_{\mathbf{w}}) = \mathrm{R}_{\mathcal{T}}(B_{\rho_{\mathbf{w}}}) \leq 2\,\mathrm{R}_{\mathcal{T}}(G_{\rho_{\mathbf{w}}})$.

Corollary 11 *Let $\mathcal{S}$ and $\mathcal{T}$ respectively be the source and the target domains on $\mathbf{X} \times Y$. For all $\mathbf{w} \in \mathbb{R}$, we have*

$$\mathrm{R}_{\mathcal{T}}(h_{\mathbf{w}}) \;\leq\; 2\,\mathrm{R}_{\mathcal{S}}(h_{\mathbf{w}}) + \mathrm{dis}_{\rho_{\mathbf{w}}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}}) + 2\lambda_{\rho_{\mathbf{w}}}\,,$$

where $\mathrm{dis}_{\rho_{\mathbf{w}}}(\mathcal{S}_{\mathbf{X}}, \mathcal{T}_{\mathbf{X}})$ and $\lambda_{\rho_{\mathbf{w}}}$ are respectively defined by Equations (24) *and* (16)*.*

Corollary 12 *Let $\mathcal{S}$ and $\mathcal{T}$ respectively be the source and the target domains on $\mathbf{X} \times Y$. For all $\mathbf{w} \in \mathbb{R}$, we have*

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

where $\mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho_{\mathbf{w}})$, $\mathrm{e}_{\mathcal{S}}(\rho_{\mathbf{w}})$, $\beta_\infty(\mathcal{T}\|\mathcal{S})$ and $\eta_{\mathcal{T}\setminus\mathcal{S}}$ are respectively defined by Equations (22), (25), (18) *and* (19)*.*

For fixed values of $\beta_\infty(\mathcal{T}\|\mathcal{S})$ and $\eta_{\mathcal{T}\setminus\mathcal{S}}$, the target risk $\mathrm{R}_{\mathcal{T}}(h_{\mathbf{w}})$ is upper-bounded by a (β_∞-weighted) sum of two losses. The expected Φ_{e}-loss (*i.e.*, the joint error) is computed on the (labeled) source domain; it aims to label the source examples correctly, but is more permissive on the required margin than the Φ-loss (*i.e.*, the Gibbs risk). The expected Φ_{d}-loss (*i.e.*, the disagreement) is computed on the target (unlabeled) domain; it promotes large *unsigned* target margins. Thus, if a target domain fulfills the *cluster assumption* (described in Section 4.2.3), $\mathrm{d}_{\mathcal{T}_{\mathbf{X}}}(\rho_{\mathbf{w}})$ will be low when the decision boundary crosses a low-density region between the homogeneous labeled clusters. Hence, Corollary 12 reflects that some source errors may be allowed if, doing so, the separation of the target domain is improved. Figure 2a leads to an insightful geometric interpretation of the two domain adaptation trade-off promoted by Corollaries 11 and 12.

6.3 Generalization Bounds and Learning Algorithms

6.3.1 First Domain Adaptation Learning Algorithm (pbda).

Theorem 9 specialized to linear classifiers gives the following.

Corollary 13 *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}{\times}Y$, any $\delta{\in}(0,1]$, any $\omega{>}0$ and $a{>}0$, with a probability at least $1{-}\delta$ over the choices of $S{\sim}(\mathcal{S})^m$ and $T{\sim}(\mathcal{T}_{\mathbf{X}})^m$, we have, for all $\mathbf{w} \in \mathbb{R}$,*

$$\mathrm{R}_{\mathcal{T}}(h_{\mathbf{w}}) \leq\; 2\,\omega'\,\widehat{\mathrm{R}}_S(h_{\mathbf{w}}) + a'\,\widehat{\mathrm{dis}}_{\rho_{\mathbf{w}}}(S,T) + 2\lambda_{\rho_{\mathbf{w}}} + 2\left(\frac{\omega'}{\omega} + \frac{a'}{a}\right)\frac{\|\mathbf{w}\|^2 + \ln\frac{3}{\delta}}{m} + (a'-1)\,,$$

where $\widehat{\mathrm{R}}_S(G_{\rho_{\mathbf{w}}})$ and $\widehat{\mathrm{dis}}_{\rho_{\mathbf{w}}}(S,T)$, are the empirical estimates of the target risk and the domain disagreement; $\lambda_{\rho_{\mathbf{w}}}$ is obtained using Equation (16)*; $\omega' = \frac{\omega}{1-e^{-\omega}}$, and $a' = \frac{2a}{1-e^{-2a}}$.*

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 $\widehat{\mathrm{R}}_T(G_{\rho_\mathbf{w}})$ is the same that minimizes

$$\begin{aligned}
&\Omega\, m\, \widehat{\mathrm{R}}_S(G_{\rho_\mathbf{w}}) + A\, m\, \widehat{\mathrm{dis}}_{\rho_\mathbf{w}}(S,T) + \mathrm{KL}(\rho_\mathbf{w}\|\pi_\mathbf{0}) \\
&\quad = \Omega \sum_{i=1}^m \Phi_\mathrm{R}\left(y_i^s \tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) + A \left| \sum_{i=1}^m \left[\Phi_\mathrm{d}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) - \Phi_\mathrm{d}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\right]\right| + \tfrac{1}{2}\|\mathbf{w}\|^2 . \qquad (27)
\end{aligned}$$

The values $\Omega > 0$ and $A > 0$ are hyperparameters of the algorithm. Note that the constants ω and a of Theorem 9 can be recovered from any Ω and A.

Equation (27) is difficult to minimize by gradient descent, as it contains an absolute value and it is highly non-convex. To make the optimization problem more tractable, we replace the loss function Φ_R by its convex relaxation $\widetilde{\Phi}_\mathrm{R}$ (as in Section 3.3.3). Even if this optimization task is still not convex (Φ_d is quasiconcave), our empirical study shows no need to perform many restarts while performing gradient descent to find a suitable solution.[14] 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:

$$G(\mathbf{w}) = \Omega \sum_{i=1}^m \widetilde{\Phi}_\mathrm{R}\left(y_i^s \tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) + A \left| \sum_{i=1}^m \left[\Phi_\mathrm{d}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) - \Phi_\mathrm{d}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\right]\right| + \tfrac{1}{2}\|\mathbf{w}\|^2 , \qquad (28)$$

where $\widetilde{\Phi}_\mathrm{R}(x) = \max\left\{\Phi_\mathrm{R}(x),\, \frac{1}{2} - \frac{x}{\sqrt{2\boldsymbol{\pi}}}\right\}$ and $\Phi_\mathrm{d}(x) = 2\,\Phi_\mathrm{R}(x)\,\Phi_\mathrm{R}(-x)$ have been defined by Equations (15) and (23). Figure 2a illustrates these three functions.

The gradient $\nabla G(\mathbf{w})$ of the Equation (28) is then given by

$$\nabla G(\mathbf{w}) = \Omega \sum_{i=1}^m \widetilde{\Phi}'_\mathrm{R}\left(\tfrac{y_i^s \mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) \tfrac{y_i^s \mathbf{x}_i^s}{\|\mathbf{x}_i^s\|} + s \times A \left(\sum_{i=1}^m \left[\Phi'_\mathrm{d}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right) \tfrac{\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|} - \Phi'_\mathrm{d}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) \tfrac{\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right]\right) + \mathbf{w} ,$$

where $\widetilde{\Phi}'_\mathrm{R}(x)$ and $\Phi'_\mathrm{d}(x)$ are respectively the derivatives of functions $\widetilde{\Phi}_\mathrm{R}$ and Φ_d evaluated at point x, and

$$s = \mathrm{sign}\left(\sum_{i=1}^m \left[\Phi_\mathrm{d}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) - \Phi_\mathrm{d}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\right]\right) .$$

We extend these equations to kernels in the following subsection.

6.3.2 Second Domain Adaptation Learning Algorithm (dalc).

Now, Theorem 10 specialized to linear classifiers gives the following.

14. 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.

Corollary 14 *For any domains $\mathcal{S}$ and $\mathcal{T}$ over $\mathbf{X}\times Y$, any $\delta\in(0,1]$, any $b>0$ and $c>0$, with a probability at least $1-\delta$ over the choices of $S\sim(\mathcal{S})^{m_s}$ and $T\sim(\mathcal{T}_{\mathbf{X}})^{m_t}$, we have, for all $\mathbf{w}\in\mathbb{R}$,*

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

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

For a source $S=\{(\mathbf{x}_i^s,y_i^s)\}_{i=1}^{m_s}$ and a target $T=\{(\mathbf{x}_i^t)\}_{i=1}^{m_t}$ samples of potentially *different size*, and some hyperparameters $B>0$, $C>0$, minimizing the next objective function *w.r.t* $\mathbf{w}\in\mathbb{R}$ is equivalent to minimize the above bound.

$$C\,\widehat{\mathrm{d}}_T(\rho_{\mathbf{w}}) + B\,\widehat{\mathrm{e}}_S(\rho_{\mathbf{w}}) + \|\mathbf{w}\|^2 \;=\; C\sum_{i=1}^{m_t}\Phi_{\mathrm{d}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right) + B\sum_{i=1}^{m_s}\Phi_{\mathrm{e}}\left(y_i^s\,\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right) + \|\mathbf{w}\|^2\,. \qquad (29)$$

We call the optimization of Equation (29) by gradient descent the DALC algorithm, for Domain Adaptation of Linear Classifiers. The gradient of Equation (29) is

$$C\sum_{i=1}^{m_t}\Phi'_{\mathrm{d}}\left(\tfrac{\mathbf{w}\cdot\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|}\right)\tfrac{\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|} + B\sum_{i=1}^{m_s}\Phi'_{\mathrm{e}}\left(y_i^s\,\tfrac{\mathbf{w}\cdot\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|}\right)\tfrac{y_i^s\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|} + \tfrac{1}{2}\,\mathbf{w}\,.$$

Contrary to the algorithm PBDA described above, our empirical study shows that there is no need to convexify any component of Equation (29): Starting the gradient descent from a uniform vector ($w_i=\frac{1}{d}$ for $i\in\{1,\ldots,d\}$), we obtain as good prediction accuracies than performing multiple random restarts. Even if the objective function is not convex, the gradient descent is easy to perform. Indeed, Φ_{d} is smooth and its derivative is continuous, in contrast with the absolute value of $\widehat{\mathrm{dis}}_{\rho_{\mathbf{w}}}(S,T)$ in Equation (27) (see also the forthcoming toy experiment of Figure 2d). Thus, the actual optimization problem of DALC is closer to the theoretical analysis than the PBDA one.

6.3.3 Using a Kernel Function

Like the algorithm PBGD3 in Subsection 3.3.2, the kernel trick applies to PBDA and DALC. Given a kernel $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$, one can express a linear classifier in a *RKHS* by a dual weight vector $\boldsymbol{\alpha}\in\mathbb{R}^{m_s+m_t}$

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

Let $S=\{(\mathbf{x}_i^s,y_i^s)\}_{i=1}^{m_s}$, $T=\{\mathbf{x}_i^t\}_{i=1}^{m_t}$ and $\mathrm{M}=m_s+m_t$. We denote K the kernel matrix of size $\mathrm{M}\times\mathrm{M}$ such as $K_{i,j}=k(\mathbf{x}_i,\mathbf{x}_j)$, where

$$\mathbf{x}_{\#} \;=\; \begin{cases}\mathbf{x}_i^s & \text{if } \#\leq m_s \quad \text{(source examples)}\\ \mathbf{x}_{\#-m_s}^t & \text{otherwise.} \quad \text{(target examples)}\end{cases}$$

On the one hand, in that case, with $m_s = m_t = m$, the objective function of PBDA (Equation 28) is rewritten in terms of the vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots \alpha_{2m})$ as

$$G(\boldsymbol{\alpha}) = \Omega \sum_{i=1}^{m} \widetilde{\Phi}_{\mathrm{R}} \left(y_i \frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) + A \left| \sum_{i=1}^{m} \left[\Phi_{\mathrm{d}} \left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) - \Phi_{\mathrm{d}} \left(\frac{\sum_{j=1}^{2m} \alpha_j K_{i+m,j}}{\sqrt{K_{i+m,i+m}}} \right) \right] \right| + \tfrac{1}{2} \sum_{i=1}^{2m} \sum_{j=1}^{2m} \alpha_i \alpha_j K_{i,j} \, .$$

On the other hand, the objective function of DALC (Equation 29) can be rewritten in terms of the vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots \alpha_{\mathrm{M}})$ as

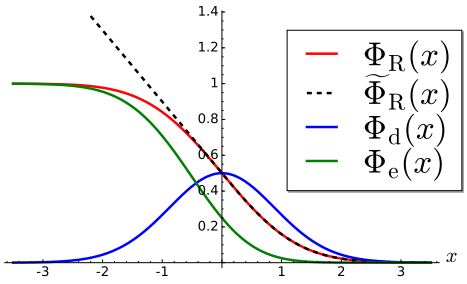
$$C \sum_{i=m_s+1}^{\mathrm{M}} \Phi_{\mathrm{d}} \left(\frac{\sum_{j=1}^{\mathrm{M}} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) + B \sum_{i=1}^{m_s} \Phi_{\mathrm{e}} \left(y_i \frac{\sum_{j=1}^{\mathrm{M}} \alpha_j K_{i,j}}{\sqrt{K_{i,i}}} \right) + \sum_{i=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} \alpha_i \alpha_j K_{i,j} \, .$$

To perform the gradient descent in terms of dual weights $\boldsymbol{\alpha}$, we start the gradient descent from the point $\alpha_i = \frac{y_i}{\mathrm{M}}$ for $i \in \{1, \dots, m_s\}$, and $\alpha_i = \frac{1}{\mathrm{M}}$ for $i \in \{m_s + 1, \dots, \mathrm{M}\}$.

6.4 Illustration on a Toy Dataset

To illustrate and compare the trade-offs of both algorithms PBDA and DALC, we extend the toy experiment of Subsection 3.3.4 (see Figure 1). To obtain the two-dimensional dataset illustrated by Figure 2c, we use, as the source sample, the 200 examples of the supervised experiment—generated by Gaussians of mean $(-1, -1)$ for the positives and $(-1, 1)$ for the negatives (see Figure 1c). Then, we generate 100 positive target examples according to a Gaussian of mean $(-1, -1)$ and 100 negative target examples according to a Gaussian of mean $(1, 1)$. All Gaussian distributions have unit variance. Note that positive source and target examples are generated by the same distribution.

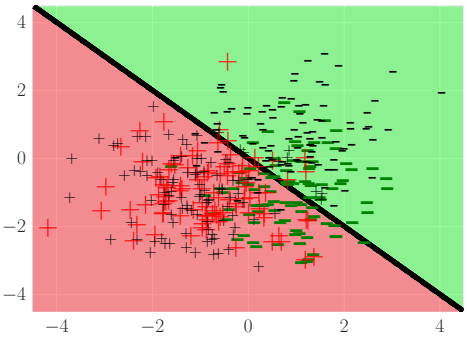
We study linear classifiers $h_{\mathbf{w}}$, with $\mathbf{w} = 2\,(\cos\theta, \sin\theta) \in \mathbb{R}^2$. That is, we fix the norm value $\|\mathbf{w}\| = 2$. Figure 2d shows the quantities varying in our two domain adaptation approaches while rotating the decision boundary $\theta \in [-\pi, \pi]$ around the origin. On the one hand, PBDA algorithm minimizes a trade-off between the domain disagreement $\widehat{\mathrm{dis}}_{\rho_{\mathbf{w}}}(S, T)$ and the source Gibbs risk $\widehat{\mathrm{R}}_S(G_{\rho_{\mathbf{w}}})$ convex surrogate given by Equation 15. On the other hand, DALC minimizes a trade-off between the target disagreement $\widehat{\mathrm{d}}_T(\rho_{\mathbf{w}})$ and source joint error $\widehat{\mathrm{e}}_S(\rho_{\mathbf{w}})$. From both Figures 2a and 2d, we see that the Gibbs risk, its convex surrogate, and the joint error behave similarly; they are following the linear classifier accuracy on the source sample. However, the domains' divergence and the target joint error values notably differ for the experiment of Figure 2d : When the target accuracy is optimal (*i.e.*, $\theta \approx \frac{\pi}{4}$) the target disagreement is close to its lowest value, while it is the opposite for the domain divergence. Thus, provided that the hyperparameters handling the trade-off between $\widehat{\mathrm{d}}_T(\rho_{\mathbf{w}})$ and $\widehat{\mathrm{e}}_S(\rho_{\mathbf{w}})$ are well chosen, DALC minimization procedure is able to find a solution *close to* the one minimizing the target risk. On the contrary, for all hyperparameters, PBDA will prefer the solution that minimizes the source risk ($\theta \approx 0$), as it minimizes $\widehat{\mathrm{dis}}_{\rho_{\mathbf{w}}}(S, T)$ and $\widehat{\mathrm{R}}_S(G_{\rho_{\mathbf{w}}})$ simultaneously.



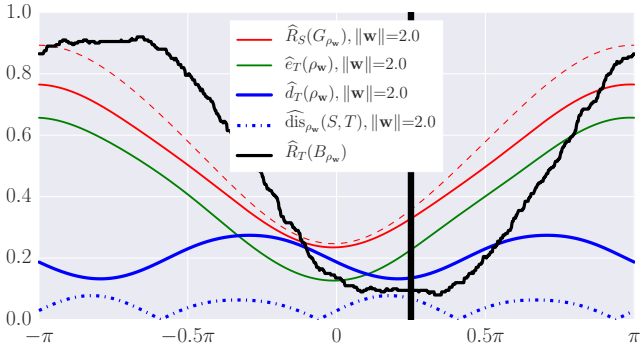
(a) Loss functions given by the specialization to linear classifiers.

FUNCTION	DERIVATIVE
$\Phi_{\mathrm{R}}(x) = \frac{1}{2}\left[1 - \mathrm{Erf}\left(\frac{x}{\sqrt{2}}\right)\right]$	$\Phi'_{\mathrm{R}}(x) = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$
$\widetilde{\Phi}_{\mathrm{R}}(x) = \max\left\{\Phi_{\mathrm{R}}(x), \frac{1}{2} - \frac{x}{\sqrt{2\pi}}\right\}$	$\Phi'_{\mathrm{R}}\left(\max\{0, x\}\right)$
$\Phi_{\mathrm{d}}(x) = 2\,\Phi_{\mathrm{R}}(x)\,\Phi_{\mathrm{R}}(-x)$	$-\sqrt{\frac{2}{\pi}}\mathrm{Erf}\left(\frac{x}{\sqrt{2}}\right) e^{-\frac{1}{2}x^2}$
$\Phi_{\mathrm{e}}(x) = \left[\Phi_{\mathrm{R}}(x)\right]^2$	$2\,\Phi_{\mathrm{R}}(x)\,\Phi'_{\mathrm{R}}(x)$

(b) Loss functions definitions and their derivatives.



(c) Toy dataset, and the decision boundary for $\theta = \frac{\pi}{4}$ (matching the vertical line of Fig.(d)). Source points are red and green, target points are black.



(d) PBDA and DALC loss values, according to θ. PBDA algorithm a trade-off between $\widehat{\mathrm{dis}}_{\rho_{\mathbf{w}}}(S,T)$ and $\widehat{\mathrm{R}}_S(G_{\rho_{\mathbf{w}}})$ (or the convex surrogate given by the red dashed line); DALC minimizes a trade-off between $\widehat{\mathrm{d}}_T(\rho_{\mathbf{w}})$ and $\widehat{\mathrm{e}}_S(\rho_{\mathbf{w}})$

Figure 2: Understanding PBDA and DALC domain adaptation learning algorithms in terms of loss functions. Upper Figures (a-b) show the loss functions, and lower Figures (c-d) illustrate the behavior on a toy dataset.

7. Experiments

Our domain adaptation algorithms PBDA[15] and DALC[16] have been evaluated on a toy problem and a sentiment dataset. In both cases, we minimize the objective function using a *Broyden-Fletcher-Goldfarb-Shanno method (BFGS)* implemented in the *scipy* python library[17] (Jones et al., 2001).

7.1 Toy Problem: Two Inter-Twinning Moons

Figure 3 illustrates the behavior of the decision boundary of our algorithms PBDA and DALC on an intertwining moons toy problem[18], where each moon corresponds to a label.

15. PBDA's code is available at the following URL: `http://graal.ift.ulaval.ca/pbda/`
16. DALC's code is available at the following URL: `https://github.com/GRAAL-Research/domain_adaptation_of_linear_classifiers`
17. The *scipy* library is available at `http://www.scipy.org/`
18. We generate each pair of moons with the `make_moons` function provided in `scikit-learn` (Pedregosa et al., 2011).

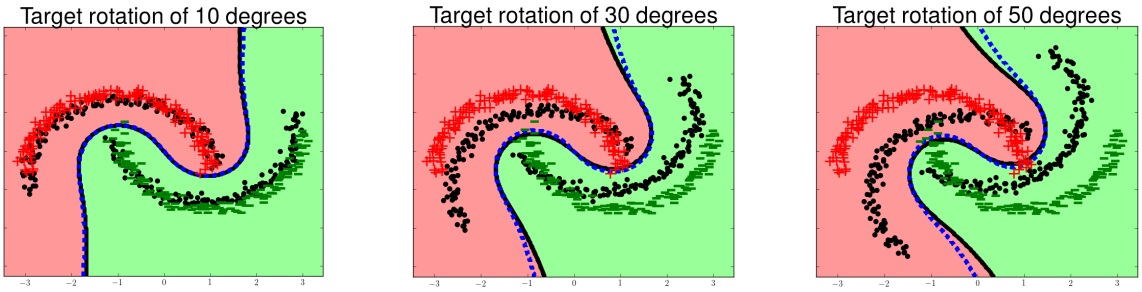


Figure 3: Decision boundaries of PBDA (in blue dashed) and DALC (in black) on the *intertwining moons* toy problem, for fixed parameters $\alpha = A = 1$ and $B{=}C{=}1$, and a RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2)$. The target points are black. The positive, *resp.* negative, source points are red, *resp.* green.

The target domain, for which we have no label, is a rotation of the source one. The figure shows clearly that PBDA and DALC succeed to adapt to the target domain, even for a rotation angle of 50°. We see that our algorithms do not rely on the restrictive *covariate shift* assumption, as some source examples are misclassified. This behavior illustrates the PBDA and DALC trade-off in action, that concede some errors on the source sample to lower the disagreement on the target sample.

7.2 Sentiment Analysis Dataset

We consider the popular *Amazon reviews* dataset (Blitzer et al., 2006) composed of reviews of four types of *Amazon.com*© products (books, DVDs, electronics, kitchen appliances). Originally, the reviews corresponded to a rate between 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 than 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 domain adaptation task (it remains about $40,000$ features), and pre-processed the data with a standard tf-idf reweighting. 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 DVDs the target one. The learning algorithms are trained $2,000$ labeled source examples and $2,000$ unlabeled target examples, and ee evaluate them on separate target test sets proposed by Chen et al. (2011) (between $3,000$ and $6,000$ examples).

7.2.1 Experiment Details

PBDA and DALC with a linear kernel have been compared with:

- SVM learned only from the source domain 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 without adaptation.

- DASVM of Bruzzone and Marconcini (2010), an iterative domain adaptation algorithm which aims 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 Chen et al. (2011) have shown best results on the dataset considered in Section 7.2.

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, PBDA, and DALC. We describe this latter method in the following subsection. For PBDA, respectively DALC, we search on a 20×20 parameter grid for a Ω, respectively C, between 0.01 and 10^6 and a parameter A, respectively B, between 1.0 and 10^8, both on a logarithm scale.

7.2.2 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 DALC, 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 follows. Firstly, the source set $S \backslash S_i$ is used as a labeled sample and the target set $T \setminus T_i$ is used as an unlabeled sample for learning a classifier h'. Secondly, using the same algorithm, a reverse classifier h'^r is learned using the *self-labeled* sample $\{(\mathbf{x}, h'(\mathbf{x}))\}_{\mathbf{x} \in T \backslash T_i}$ as the source set and the unlabeled part of $S \backslash S_i$ as target sample. Finally, the reverse classifier h'^r is evaluated on S_i. We summarize this principle on Figure 4. The process is repeated k times to obtain the reverse cross-validation risk averaged across all folds.

7.2.3 Empirical Results

Table 1 contains the test accuracies on the sentiment analysis dataset. We make the following observations. Above all, the domain adaptation approaches provide the best average results, implying that tackling this problem with a domain adaptation method is reasonable. Then, our method DALC based on the novel domain adaptation analysis is the best algorithm overall on this task. Except for the two adaptive tasks between "electronics" and "DVDs", DALC is either the best one (five times), or the second one (five times). Moreover, according to a Wilcoxon signed rank test with a 5% significance level, we obtain a probability of 89.5% that DALC is better than PBDA. This test tends to confirm that the analysis with the new perspective improves the analysis based on a domains' divergence point of view. Moreover, PBDA is on average better than CODA, but less accurate than DASVM.

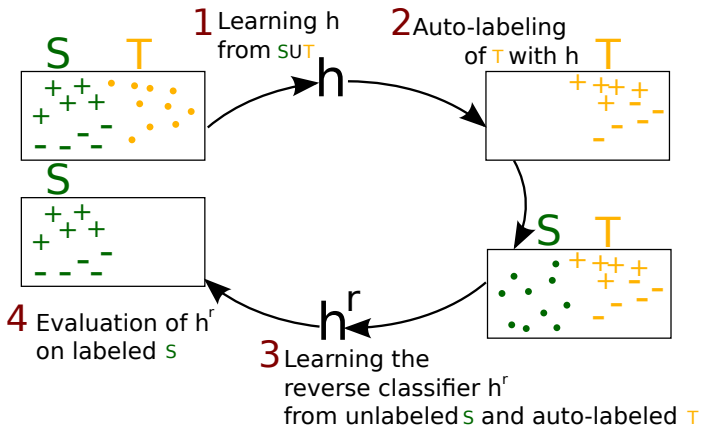


Figure 4: The principle of the reverse/circular validation in our setting.

Table 1: Error rates for sentiment analysis dataset. B, D, E, K respectively denotes Books, DVDs, Electronics, Kitchen. In **bold** are highlighted the best results and in *italic* the second ones.

	PBGD3 ^{CV}	SVM ^{CV}	DASVM ^{RCV}	CODA ^{RCV}	PBDA ^{RCV}	DALC ^{RCV}
B→D	**0.174**	0.179	0.193	0.181	0.183	*0.178*
B→E	0.275	0.290	*0.226*	0.232	0.263	**0.212**
B→K	0.236	0.251	**0.179**	0.215	0.229	*0.194*
D→B	*0.192*	0.203	0.202	0.217	0.197	**0.186**
D→E	0.256	0.269	**0.186**	*0.214*	0.241	0.245
D→K	0.211	0.232	0.183	*0.181*	0.186	**0.175**
E→B	0.268	0.287	0.305	0.275	**0.232**	*0.240*
E→D	0.245	0.267	**0.214**	0.239	*0.221*	0.256
E→K	*0.127*	0.129	0.149	0.134	0.141	**0.123**
K→B	0.255	0.267	0.259	*0.247*	*0.247*	**0.236**
K→D	0.244	0.253	**0.198**	0.238	0.233	*0.225*
K→E	0.235	0.149	0.157	0.153	**0.129**	*0.131*
Average	0.226	0.231	*0.204*	0.210	0.208	**0.200**

However, PBDA is competitive: the results are not significantly different from CODA and DASVM. It is important to notice that DALC and PBDA are 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 DALC and PBDA. In fact, the clear advantage of the PAC-Bayesian approach is that we jointly optimize the terms of our bounds in one step.

8. Conclusion and Future Work

In this paper, we present two domain adaptation analyses for the PAC-Bayesian framework that focuses on models that takes the form of a majority vote over a set of classifiers: The first one is based on a common principle in domain adaptation, and the second one brings a novel perspective on domain adaptation.

To begin, we follow the underlying philosophy of the seminal works of Ben-David et al. (2006); Ben-David et al. (2010a) and Mansour et al. (2009a), in other words, we derive an upper bound on the target risk (of the Gibbs classifier) thanks to a domains' divergence measure suitable for the PAC-Bayesian setting. We define this divergence as the average deviation between the disagreement over a set of classifiers on the source and target domains. This leads to a bound that takes the form of a trade-off between the source risk, the domains' divergence and a term that captures the ability to adapt for the current task. Then, we propose another domain adaptation bound while taking advantage of the inherent behavior of the target risk in the PAC-Bayesian setting. We obtain a different upper bound that is expressed as a trade-off between the disagreement only on the target domain, the joint errors of the classifiers only on the source domain, and a term reflecting the worst case error in regions where the source domain is non-informative. To the best of our knowledge, a crucial novelty of this contribution is that the trade-off is controlled by a domains' divergence: Contrary to our first bound, the divergence is not an additive term (as in many domain adaptation bounds) but is a factor weighing the importance of the source information.

Our analyses, combined with PAC-Bayesian generalization bounds, lead to two new domain adaptation algorithms for linear classifiers: PBDA associated with the previous works philosophy, and DALC associated with the novel perspective. The empirical experiments show that the two algorithms are competitive with other approaches, and that DALC outperforms significantly PBDA. Consequently, we believe that our PAC-Bayesian analyses open the door to develop new domain adaptation methods by making use of the possibilities offered by the PAC-Bayesian theory, and give 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 about the classifiers accuracy; in this paper we opted for a non-informative prior that is 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 model information learned from previous tasks. This suggests that we can extend our analyses to multisource domain adaptation (Crammer et al., 2007; Mansour et al., 2009c; Ben-David et al., 2010a) and lifelong learning where the objective is to perform well on future tasks, for which no data has been observed so far (Thrun and Mitchell, 1995).

Another promising issue is to address the problem of the hyperparameter selection. Indeed, the adaptation capability of our algorithms PBDA and DALC could be even put further with a specific PAC-Bayesian validation procedure. An idea would be to propose a (reverse) validation technique that takes into account some particular prior distributions. Another solution could be to explicitly control the neglected terms 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 to take into account both margin information over unlabeled data and the distribution disagreement (these two elements seem of crucial importance in domain adaptation).

Concerning DALC, we would like to investigate the case where the domains' divergence can be estimated, *i.e.*, when the covariate shift assumption holds or when some target labels are available. In this scenario, the domains' divergence might not be considered as a hyperparameter to tune. Last but not least, the non-estimable term of DALC—suggesting that the domains should live in the same regions—can be dealt with representation learning approach. This could be an incentive to combine DALC with existing representation learning techniques.

Acknowledgments

This work was supported in part by the French projects LIVES ANR-15-CE23-0026-03, VideoSense ANR-09-CORD-026, 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, and while P. Germain was affiliated with Département d'informatique et de génie logiciel, Université Laval, Québec, Canada.

Appendix A. Some Tools

Lemma 15 (Hölder's Inequality) *Let S be a measure space and let $(p, q) \in [1, \infty]^2$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then, for all measurable real-valued functions f and g on S,*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Lemma 16 (Markov's inequality) *Let Z be a random variable and $t \geq 0$, then*

$$\Pr\left(|Z| \geq t\right) \;\leq\; \frac{\mathbf{E}(|Z|)}{t}\,.$$

Lemma 17 (Jensen's inequality) *Let Z be an integrable real-valued random variable and $g(\cdot)$ any function. If $g(\cdot)$ is convex, then*

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

Lemma 18 (from Lemma 3 of 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 19 (from Inequalities 1 and 2 of Maurer, 2004) *Let $X = (X_1, \ldots, X_m)$ 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 20 (Change of measure inequality[19]) *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 21 (from Theorem 25 of Germain et al., 2015b) *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 $\hat{\rho}(h,h') = \rho(h)\rho(h')$ and $\hat{\pi}(h,h') = \pi(h)\pi(h')$. Then*

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

Appendix B. Proof of Equation (10)

Given $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$ and $\mathbf{w} \in \mathbb{R}^d$, we consider—without loss of generality—a vector basis where $y\frac{\mathbf{x}}{\|\mathbf{x}\|}$ is the first coordinate. Thus, the first component of any vector $\mathbf{w}' \in \mathbb{R}^d$ is given by $w'_1 = y\frac{\mathbf{w}'\cdot\mathbf{x}}{\|\mathbf{x}\|}$, which leads to

$$\begin{aligned}
\mathop{\mathbf{E}}_{h_{\mathbf{w}'}\sim\rho_{\mathbf{w}}} \mathcal{L}_{0\text{-}1}\big(h_{\mathbf{w}'}(\mathbf{x}), y\big) &= \mathop{\mathbf{E}}_{h_{\mathbf{w}'}\sim\rho_{\mathbf{w}}} \mathrm{I}\left[h_{\mathbf{w}'}(\mathbf{x}) \neq y\right] \\
&= \mathop{\mathbf{E}}_{h_{\mathbf{w}'}\sim\rho_{\mathbf{w}}} \mathrm{I}\left[y\,\mathbf{w}'\cdot\mathbf{x} \leq 0\right] \\
&= \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\boldsymbol{\pi})^n}} \exp\left(-\tfrac{1}{2}\|\mathbf{w}' - \mathbf{w}\|^2\right) \mathrm{I}\left[y\,\mathbf{w}'\cdot\mathbf{x} \leq 0\right] d\,\mathbf{w}' \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\boldsymbol{\pi}}} \exp\left(-\tfrac{1}{2}(w'_1 - w_1)^2\right) \mathrm{I}\left[w'_1 \leq 0\right] dw'_1 \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\boldsymbol{\pi}}} \exp\left(-\tfrac{1}{2}\,t^2\right) \mathrm{I}\left[t \leq -w_1\right] dt \\
&= \int_{-\infty}^{-w_1} \frac{1}{\sqrt{2\boldsymbol{\pi}}} \exp\left(-\tfrac{1}{2}\,t^2\right) dt \\
&= 1 - \Pr_{t\sim\mathcal{N}(0,1)}\left(t \;\leq\; y\,\tfrac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right) \\
&= \Phi_{\mathrm{R}}\left(y\,\tfrac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right),
\end{aligned}$$

where we used $y\,\mathbf{w}'\cdot\mathbf{x} = w'_1\|\mathbf{x}\|$, $t \coloneqq w'_1 - w_1$, $w_1 = y\frac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}$, and the definition of Φ_{R} given by Equation (11). We then have

$$\mathrm{R}_{\mathcal{D}}(G_{\rho_{\mathbf{w}}}) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \mathop{\mathbf{E}}_{h_{\mathbf{w}'}\sim\rho_{\mathbf{w}}} \mathcal{L}_{0\text{-}1}\big(h_{\mathbf{w}'}(\mathbf{x}), y\big) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P_S} \Phi_{\mathrm{R}}\left(y\,\tfrac{\mathbf{w}\cdot\mathbf{x}}{\|\mathbf{x}\|}\right).$$

19. See Seldin and Tishby (2010, Lemma 4), McAllester (2013, Equation 20), or Germain et al. (2015b, Lemma 17).

Appendix C. Proof of Theorem 7

Proof We use the shorthand notation $\mathcal{L}_{\mathcal{D}}(h) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(h,\mathbf{x},y)$ and $\mathcal{L}_S(h) = \frac{1}{m}\sum_{(\mathbf{x},y)\in S} \ell(h,\mathbf{x},y)$.

Consider any convex function $\Delta : [0,1]\times[0,1] \to \mathbb{R}$. Applying consecutively Jensen's Inequality (Lemma 17) and the *change of measure inequality* (Lemma 20), we obtain

$$\begin{aligned}\forall \rho \text{ on } \mathcal{H}, \quad m\times\Delta\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_S(h), \mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_{\mathcal{D}}(h)\right) &\le \mathop{\mathbf{E}}_{h\sim\rho} m\times\Delta\left(\mathcal{L}_S(h),\mathcal{L}_{\mathcal{D}}(h)\right)\\ &\le \mathrm{KL}(\rho\|\pi) + \ln\Big[X_\pi(S)\Big],\end{aligned}$$

with

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

Then, Markov's Inequality (Lemma 16) gives

$$\Pr_{S\sim\mathcal{D}^m}\left(X_\pi(S) \le \tfrac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m} X_\pi(S')\right) \ge 1-\delta,$$

and

$$\begin{aligned}\mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m} X_\pi(S') &= \mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m}\mathop{\mathbf{E}}_{h\sim\pi} e^{m\times\Delta(\mathcal{L}_{S'}(h),\mathcal{L}_{\mathcal{D}}(h))}\\ &= \mathop{\mathbf{E}}_{h\sim\pi}\mathop{\mathbf{E}}_{S'\sim\mathcal{D}^m} e^{m\times\Delta(\mathcal{L}_{S'}(h),\mathcal{L}_{\mathcal{D}}(h))}\\ &\le \mathop{\mathbf{E}}_{h\sim\pi}\sum_{k=0}^{m}\binom{k}{m}(\mathcal{L}_{\mathcal{D}}(h))^k(1-\mathcal{L}_{\mathcal{D}}(h))^{m-k}e^{m\times\Delta\left(\frac{k}{m},\mathcal{L}_{\mathcal{D}}(h)\right)}, \qquad (30)\end{aligned}$$

where the last inequality is given by Lemma 19 (we have an equality when the output of ℓ is in $\{0,1\}$). As shown in Germain et al. (2009a, Corollary 2.2), by fixing

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

Line (30) becomes equal to 1, and then $\mathbf{E}_{S'\sim\mathcal{D}^m} X_\pi(S') \le 1$. Hence,

$$\Pr_{S\sim\mathcal{D}^m}\left(\forall\rho \text{ on } \mathcal{H},\ -c\mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_S(h) - \ln[1-\mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_{\mathcal{D}}(h)\,(1-e^{-c})] \le \frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m}\right) \ge 1-\delta.$$

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

$$\forall\rho \text{ on } \mathcal{H},\ \mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_{\mathcal{D}}(h) \le \frac{1}{1-e^{-c}}\left[1-\exp\left(-c\mathop{\mathbf{E}}_{h\sim\rho}\mathcal{L}_S(h) - \frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{1}{\delta}}{m}\right)\right].$$

The final result is obtained by using the inequality $1-\exp(-z)\le z$. ■

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.

S. Ben-David, S. Shalev-Shwartz, and R. Urner. Domain adaptation—can quantity compensate for quality? In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2012.

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.

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.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pages 442–450, 2010.

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

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

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

T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

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

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.

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. PAC-Bayesian theorems for domain adaptation with specialization to linear classifiers. *Research Report. arXiv preprint arXiv:1503.06944*, 2015a.

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

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A new PAC-Bayesian perspective on domain adaptation. In *International Conference on Machine Learning*, volume 48, pages 859–868, 2016.

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.

E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL `http://www.scipy.org/`.

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*, pages 769–776, 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.

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. McAllester. A PAC-Bayesian tutorial with a dropout bound. *CoRR*, abs/1307.2118, 2013.

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

D. A. McAllester and J. Keshet. Generalization bounds and consistency for latent structural probit and ramp loss. In *Advances in Neural Information Processing System*, pages 2205–2212, 2011.

E. Morvant. Domain adaptation of weighted majority votes via perturbed variation-based self-labeling. *Pattern Recognition Letters*, 51:37–43, 2015.

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

S.J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

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.

M. Re and G. Valentini. Ensemble methods: a review. *Advances in machine learning and data mining for astronomy*, pages 563–582, 2012.

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.

Y. Seldin and N. Tishby. PAC-Bayesian analysis of co-clustering and beyond. *JMLR*, 11: 3595–3646, 2010.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference*, 90(2):227–244, 2000.

M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2007.

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.

R. Urner, S. Shalev-Shwartz, and S. Ben-David. Access to unlabeled data can speed up prediction time. In *International Conference on Machine Learning*, pages 641–648, 2011.

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

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.