Une analyse PAC-Bayésienne de l'adaptation de domaine et sa spécialisation aux classifieurs linéaires

Pascal Germain[1], Amaury Habrard[2], François Laviolette[1], et Emilie Morvant[3]

[1]Département d'informatique et de génie logiciel, Université Laval, Québec, Canada
[2]Université Jean Monnet de Saint-Étienne, Laboratoire Hubert Curien, UMR CNRS 5516
[3]Aix-Marseille Univ., LIF-QARMA, UMR CNRS 7279

27 mai 2013

Résumé

Dans cet article, nous nous intéressons au problème de l'adaptation de domaine (AD) correspondant au cas où les données d'apprentissage et de test sont issues de distributions différentes. Nous proposons une analyse PAC-Bayésienne de ce problème dans le cadre de la classification binaire sans information supervisée sur les données de test. La théorie PAC-Bayésienne permet d'obtenir des garanties théoriques sur le risque d'un vote de majorité sur un ensemble d'hypothèses. Notre contribution au cadre de l'AD repose sur une nouvelle mesure de divergence entre distributions basée une notion d'espérance de désaccords entre hypothèses. Cette mesure nous permet de dériver une première borne PAC-Bayésienne pour le classifieur stochastique de Gibbs. Cette borne a l'avantage d'être optimisable directement pour tout espace d'hypothèses et nous en donnons une illustration dans le cas de classifieurs linéaires. L'algorithme proposé dans ce contexte montre des résultats intéressants sur un problème jouet ainsi que sur une tâche courante d'analyse d'avis. Ces résultats ouvrent de nouvelles perspectives pour appréhender le problème de l'adaptation domaine grâce aux outils offerts par la théorie PAC-Bayésienne.

Mots-clef : Adaptation de domaine, Théorie PAC-Bayésienne, Classifieurs linéaires

1 Introduction

En apprentissage automatique, la majorité des méthodes d'apprentissage de classifieurs repose sur l'hypothèse selon laquelle les données d'apprentissage et de test sont générées de manière *i.i.d.* selon la même distribution de probabilité. Cependant, cette hypothèse forte n'est pas réaliste pour de nombreuses applications. Prenons l'exemple d'un système de filtrage de *spams*. Un tel système bien adapté à un utilisateur donné peut ne pas convenir à un autre utilisateur recevant des *e-mails* différents. En d'autres termes, les données d'apprentissage associées au premier utilisateur ne se sont pas nécessairement représentatives des données de test d'un autre utilisateur. Il s'avère donc nécessaire de développer des méthodes permettant d'adapter un classifieur utilisant des données d'apprentissage (sources) à des données de test (cibles) différentes. Ce type de problème fait référence au cadre de l'adaptation de domaine [1] (AD). On parle d'AD lorsque la distribution de probabilité génératrice des données cibles (appelée le domaine cible) est différente de celle générant les données sources (appelée le domaine source). Dans ce contexte, l'AD est connue pour être une tâche difficile même sous des hypothèses fortes [BDU12], parmi lesquelles le *covariate-shift* où les domaines source et cible diffèrent uniquement en leur marginale sur l'espace d'entrée (c'est-à-dire qu'ils partagent la même fonction d'étiquetage). Une problématique majeure en AD est la définition d'une mesure de divergence permettant de quantifier à quel point les domaines sont reliés entre eux. Lorsque les domaines sont similaires selon cette mesure, les garanties en généralisation sur le domaine cible peuvent être plus "simples" à obtenir. Par exemple, dans le contexte de la classification binaire (avec la fonction perte 0-1), [BDBC$^+$10] ont proposé de considérer la notion de $\mathcal{H}$-divergence entre les distributions marginales selon l'espace d'entrée. Cette quantité se définit comme le désaccord maximal entre deux classifieurs

1. Voir [Jia08, QCSSL09] pour un état de l'art.

d'une même classe d'hypothèses. Elle permet de dériver une borne en généralisation pour le problème de l'AD, principalement basée sur une analyse de type VC (Vapnik-Chervonenkis). La mesure appelée *discrepancy distance* [MMR09a] généralise la divergence précédente à des fonctions réelles ainsi qu'à d'autres fonctions de perte et est utilisée pour obtenir une borne en généralisation basée sur la complexité de Rademacher. Dans ces deux situations, une tâche d'AD peut être vue comme un compromis entre la complexité de la classe d'hypothèses $\mathcal{H}$, la capacité d'adaptation de $\mathcal{H}$ selon une divergence entre les marginales et le risque empirique sur le domaine source. D'autres mesures ont été proposées sous différentes hypothèses, comme la divergence de Rényi, pertinente pour le problème d'*importance weighting*, ou celle proposée par [ZZY12], prenant en compte à la fois l'étiquetage réel source et l'étiquetage réel cible.

La nouveauté de notre contribution est l'utilisation du cadre PAC-Bayésien pour proposer une nouvelle analyse pour l'AD dans une situation de classification binaire sans étiquettes cibles (on parle parfois d'AD non-supervisé). Étant donnée une distribution a priori (prior) sur une famille de classifieurs $\mathcal{H}$, la théorie PAC-Bayésienne (introduite par [McA99]) étudie les algorithmes construisant une distribution a posteriori (posterior) ρ sur $\mathcal{H}$ (par opposition aux analyses plus classiques qui se concentrent sur un classifieur unique $h \in \mathcal{H}$). Autrement dit, nous nous intéressons à un vote pondéré sur $\mathcal{H}$ selon ρ que nous appelons un ρ-moyennage. En suivant ce principe, nous définissons une pseudo-métrique quantifiant la divergence entre les domaines en fonction de l'espérance, selon ρ, des désaccords entre classifieurs sur les distributions marginales des deux domaines. Cette mesure de désaccords possède plusieurs avantages importants. Tout d'abord, elle est très bien adaptée au cadre PAC-Bayésien puisqu'elle s'exprime comme une ρ-moyenne sur $\mathcal{H}$. Elle est également très précise et inférieure à la $\mathcal{H}$-divergence et, au contraire de cette dernière, est spécifique au classifieur final considéré. Enfin, elle s'avère très facile à estimer à partir d'échantillons finis. Cette pseudo-métrique nous permet de dériver une première borne en généralisation PAC-Bayésienne pour le problème de l'AD. De manière pratique, l'optimisation de cette borne repose sur un compromis entre trois quantités. Les deux premières sont en fait deux termes habituels de la théorie PAC-Bayésienne : la complexité du vote de majorité mesurée par la divergence de Kullback-Leibler (KL-divergence) entre la prior et la posterior et le risque empirique mesuré par le ρ-moyennage des erreurs sur l'échantillon source. La troisième quantité correspond à notre divergence entre marginales, qui estime la capacité de la posterior à distinguer la différence structurelle entre les échantillons source et cible. Une caractéristique intéressante de notre approche est que ces quantités peuvent être optimisées simultanément. Nous l'illustrons en proposant un premier algorithme pour optimiser cette borne dans le cas des classifieurs linéaires.

L'article s'organise comme suit. La section 2 présente les notations et les deux résultats de référence en AD. Le cadre de la théorie PAC-Bayésienne est rappelé en section 3. Notre contribution principale, la borne d'AD pour un apprentissage PAC-Bayésien, est présentée en section 4. Nous dérivons ensuite notre nouvel algorithme en section 5. Avant de conclure en section 7, nous expérimentons notre approche en section 6.

2 Notations et théorie de l'AD

Nous considérons les tâches d'AD pour la classification binaire où $X \subseteq \mathbb{R}^d$ est l'espace d'entrée et $Y = \{-1, +1\}$ est l'espace de sortie. Le domaine source P_S et le domaine cible P_T sont deux distributions différentes sur $X \times Y$, D_S et D_T étant les distributions marginales respectives sur X. Nous nous intéressons à la tâche pour laquelle nous ne disposons d'aucune étiquette cible. Un algorithme d'apprentissage prend donc en entrée un échantillon source étiqueté $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ tiré *i.i.d.* depuis P_S, et un cible non étiqueté $T = \{\mathbf{x}_j^t\}_{j=1}^{m'}$ tiré *i.i.d.* selon D_T. Soit $h : X \to Y$ une hypothèse. L'erreur source réelle de h sur P_S correspond à la probabilité que h commette une erreur,

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

où $\ell_{0\text{-}1}(a, b) \stackrel{\text{def}}{=} \mathbf{I}[a \neq b]$ est la fonction perte 0-1 retournant 1 si $a \neq b$ et 0 sinon. L'erreur cible réelle $R_{P_T}(\cdot)$ sur P_T est définie de la même manière. $R_S(\cdot)$ est l'erreur source empirique mesurée sur l'échantillon S. L'objectif principal en AD est d'apprendre — sans étiquette cible — un classifieur ayant l'erreur cible réelle $R_{P_T}(h)$ la plus faible possible.

En outre, nous introduisons la notion de désaccord source réel entre h' et h, mesurant la probabilité que les deux classifieurs h et h' soient en désaccord sur D_S,

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

Le désaccord cible réel $R_{D_T}(\cdot, \cdot)$ sur D_T est défini de la même manière. $R_S(\cdot, \cdot)$ et $R_T(\cdot, \cdot)$ sont les désaccords source et cible empiriques mesurés respectivement sur S et T. En fonction du contexte, S peut

désigner soit l'échantillon source étiqueté $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$, soit l'échantillon non étiqueté associé $\{\mathbf{x}_i^s\}_{i=1}^m$.

2.1 Nécessité d'une divergence entre les domaines

Le but de l'AD est de trouver une hypothèse ayant une erreur cible faible même lorsqu'aucune information sur les étiquettes cibles n'est disponible. Ce problème peut clairement s'avérer difficile à résoudre même sous des hypothèses fortes [BDU12]. Pour dériver des bornes en généralisation en AD (nous appelons un tel résultat une borne d'AD), il est essentiel de faire appel à une mesure de divergence entre les domaines source et cible : plus les domaines sont similaires, plus l'adaptation est aisée. Dans la littérature plusieurs approches ont été proposées [ZZY12, BDBC+10, MMR09a, MMR09b] pour estimer à quel point le domaine source est proche du domaine cible. Concrètement, les deux domaines P_S et P_T diffèrent si leurs marginales D_S et D_T sont différentes, ou si la fonction d'étiquetage source diffère de la fonction cible, ou si les deux cas précédents se produisent. Ceci suggère de prendre en compte deux divergences : une entre les marginales D_S et D_T et l'autre entre les étiquetages. Lorsque des étiquettes cibles sont disponibles, les deux mesures peuvent être combinées (comme dans [ZZY12]). Dans le cas contraire, il est préférable de séparer les deux mesures puisque dans une telle situation il sera impossible d'estimer le meilleur étiquetage cible. Habituellement, on admet qu'il existe un lien entre l'étiquetage source et l'étiquetage cible. Une solution consiste alors à chercher une représentation pour laquelle les marginales D_S et D_T sont proches tout en gardant de bonnes performances sur le domaine source.

2.2 Bornes d'AD pour la classification

Nous rappelons maintenant les deux premiers résultats de référence proposant une borne d'AD basée sur une divergence entre les marginales.

Tout d'abord [BDBC+10] ont prouvé la borne d'AD suivante qui s'avère précise lorsqu'il existe un classifieur (dans $\mathcal{H}$) à la fois performant sur le domaine source et sur le domaine cible.

Théorème 1 ([BDBC+10]). *Soit $\mathcal{H}$ une classe d'hypothèses, alors pour tout $h \in \mathcal{H}$, on a,*

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

où $\frac{1}{2} d_{\mathcal{H}}(D_S, D_T) \stackrel{\text{def}}{=} \sup_{(h,h') \in \mathcal{H}^2} \left| R_{D_T}(h,h') - R_{D_S}(h,h') \right|$ est la $\mathcal{H}\Delta\mathcal{H}$-divergence entre les marginales D_S et D_T et $h^* \stackrel{\text{def}}{=} \operatorname{argmin}_{h \in \mathcal{H}} \big(R_{P_S}(h) + R_{P_T}(h)\big)$ est l'hypothèse jointe idéale.

Cette borne dépend de quatre termes. $R_{P_S}(h)$ est l'erreur classique sur le domaine source. La divergence $\frac{1}{2} d_{\mathcal{H}}(D_S, D_T)$ dépend de $\mathcal{H}$ et correspond au désaccord maximal entre deux hypothèses de $\mathcal{H}$. Autrement dit, elle quantifie à quel point les hypothèses de $\mathcal{H}$ peuvent "détecter" les différences entre les marginales. Les garanties en généralisation seront meilleures si cette capacité de détection est faible. Les derniers termes $R_{P_S}(h^*)$ et $R_{P_T}(h^*)$ sont reliés à la meilleure hypothèse h^* sur les deux domaines et mesurent la qualité de $\mathcal{H}$ en terme d'étiquetage. Si h^* est mauvaise, il sera difficile de trouver une hypothèse performante sur le domaine cible. Enfin, l'équation (1) combinée avec la théorie VC expriment un compromis entre la performance source d'une hypothèse h, la complexité de $\mathcal{H}$ et "l'incapacité" des hypothèses de $\mathcal{H}$ à détecter les différences entre les domaines.

Ensuite, [MMR09a] ont étendu la $\mathcal{H}\Delta\mathcal{H}$-divergence à la *discrepancy divergence* disc_ℓ pour les problèmes de régression et pour toute fonction perte symétrique ℓ vérifiant l'inégalité triangulaire. Soit $\ell : [-1,+1]^2 \mapsto \mathbb{R}^+$ une telle fonction, la disc_ℓ entre D_S et D_T est : $\text{disc}_\ell(D_S, D_T) \stackrel{\text{def}}{=} \sup_{(h,h') \in \mathcal{H}^2} \left| \mathop{\mathbf{E}}_{\mathbf{x}^t \sim D_T} \ell(h(\mathbf{x}^t), h'(\mathbf{x}^t)) - \mathop{\mathbf{E}}_{\mathbf{x}^s \sim D_S} \ell(h(\mathbf{x}^s), h'(\mathbf{x}^s)) \right|$. Dans le contexte de la classification binaire avec la perte 0-1, on a : $\frac{1}{2} d_{\mathcal{H}}(D_S, D_T) = \text{disc}_{\ell_{0\text{-}1}}(D_S, D_T)$. Même si ces deux divergences coïncident, la borne d'AD de [MMR09a] diffère du théorème 1 :

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

où $h_T^* \stackrel{\text{def}}{=} \operatorname{argmin}_{h \in \mathcal{H}} R_{P_T}(h)$ et $h_S^* \stackrel{\text{def}}{=} \operatorname{argmin}_{h \in \mathcal{H}} R_{P_S}(h)$ sont respectivement les hypothèses idéales sur le domaine cible et le domaine source. Dans cette situation, l'équation (2) borne directement la différence entre l'erreur cible d'un classifieur et celle de la meilleure hypothèse cible ($R_{P_T}(h_T^*)$). Cette borne s'exprime comme un compromis entre le désaccord entre h et l'hypothèse source idéale h_S^*, la complexité de la classe $\mathcal{H}$ (exprimée par la complexité de Rademacher dans [MMR09a]), et — encore une fois — "l'incapacité" d'une hypothèse de $\mathcal{H}$ à détecter les différences entre les deux domaines.

Les bornes des équations (1) et (2) suggèrent que si la divergence entre les domaines est petite, un classifieur avec une erreur faible sur le domaine source peut

être performant sur le domaine cible. La divergence peut être vue comme un désaccord dans le "pire cas". Nous proposons dans la suite une analyse en moyenne grâce à l'essence de la théorie PAC-Bayésienne, connue pour offrir des bornes en généralisations plus précises [McA99, APHST06].

3 La théorie PAC-Bayésienne

Nous rappelons dans cette section le cadre PAC-Bayésien classique pour la classification supervisée binaire, introduit par [McA99]. Traditionnellement, on considère un vote de majorité pondéré sur un ensemble $\mathcal{H}$ d'hypothèses. Étant donnés une distribution prior π sur $\mathcal{H}$ et un échantillon d'apprentissage S, le but de l'apprenant est de trouver la distribution posterior ρ sur $\mathcal{H}$ correspondant au vote de majorité B_ρ pondéré par ρ (parfois appelé le classifieur de Bayes) avec les meilleures garanties en généralisation. B_ρ est défini par,

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

Il est connu que la minimisation de l'erreur de B_ρ est NP-difficile. Dans l'approche PAC-Bayésienne, nous substituons l'erreur de B_ρ par l'erreur du classifieur stochastique de Gibbs G_ρ, qui prédit l'étiquette d'un exemple $\mathbf{x}$ en tirant aléatoirement selon ρ une hypothèse h dans $\mathcal{H}$, puis en retournant $h(\mathbf{x})$. Remarquons que l'erreur du classifieur de Gibbs sur un domaine P_S correspond à l'espérance des erreurs selon ρ,

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

Dans ce contexte, si B_ρ commet une erreur sur un exemple $\mathbf{x}$, alors au moins la moitié des classifieurs (selon ρ) commettent une erreur sur $\mathbf{x}$. On a donc trivialement : $R_{P_S}(B_\rho) \leq 2R_{P_S}(G_\rho)$. Un autre résultat sur $R_{P_S}(B_\rho)$ est la C-borne (introduite par [LLM+06]) qui relie $R_{P_S}(B_\rho)$ à la moyenne et la variance du classifieur de Gibbs et est définie par,

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

où $R_{D_S}(G_\rho, G_\rho)$ correspond au désaccord moyen des classifieurs selon ρ défini par,

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

Comme suggéré dans [LMR11], pour un numérateur fixé, le meilleur vote de majorité est celui associé au plus faible dénominateur, *c.-à-d.* amenant à des classifieurs en grand désaccord. La C-borne est connue pour être une bonne approximation de $R_{P_S}(B_\rho)$.

La théorie PAC-Bayésienne permet de borner l'espérance des erreurs $R_{P_S}(G_\rho)$ en fonction de deux quantités principales : l'erreur empirique $R_S(G_\rho) = \mathbf{E}_{h\sim\rho} R_S(h)$, estimée sur un échantillon S *i.i.d.* selon P_S, et la KL-divergence entre ρ et π : $\text{KL}(\rho\|\pi) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim\rho} \ln \frac{\rho(h)}{\pi(h)}$. Dans cet article, nous utilisons la borne PAC-Bayésienne proposée par Catoni [Cat07] dans sa forme simplifiée suggérée dans [GLL+09].

Théorème 2 ([Cat07]). *Pour tout domaine P_S sur $X \times Y$, pour toute classe d'hypothèses $\mathcal{H}$, pour toute distribution π sur $\mathcal{H}$, pour tout $\delta \in (0, 1]$, pour tout réel $c > 0$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (P_S)^m$, pour toute distribution ρ sur $\mathcal{H}$, on a,*

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

Cette borne possède deux avantages. D'une part, lorsque ρ est une gaussienne isotropique sur l'espace des classifieurs linéaires, sa minimisation est étroitement liée à celle du problème associé aux SVM [GLLM09]. D'autre part, elle dépend d'un paramètre c, qui nous permet de contrôler le compromis entre l'erreur empirique $R_S(G_\rho)$ et le terme de complexité $\frac{\text{KL}(\rho\|\pi)}{m}$. Si $c = \frac{1}{\sqrt{m}}$, cette borne devient consistante : lorsque m tend vers $+\infty$, la borne tend vers $1 \times [R_S(G_\rho) + 0]$.

Alors que les bornes d'AD mentionnées à la section 2 se concentrent sur un classifieur unique, nous croyons qu'une borne appropriée au contexte PAC-Bayesien doit considérer un désaccord entre classifieurs défini selon un ρ-moyennage (combiné avec une borne PAC-Bayésienne classique). Cette intuition nous amène à définir un désaccord moyenné selon ρ comparant les distributions marginales source et cible, pour dériver ensuite notre borne d'AD.

4 Borne d'AD pour le classifieur de Gibbs

L'originalité de notre contribution réside en la définition d'un cadre théorique d'AD pour l'approche PAC-Bayésienne. La section 4.1 présente notre pseudo-métrique permettant de comparer les domaines dans ce contexte. Ensuite nous dérivons en section 4.2 la première borne d'AD PAC-Bayésienne.

4.1 Une divergence entre domaines pour une analyse PAC-Bayésienne

Comme discuté en section 2.1, pour dériver une borne d'AD il est crucial de faire appel à une mesure de divergence entre les domaines source et cible. Dans notre situation où nous ne disposons pas d'étiquette cible, nous définissons une divergence entre les distributions marginales associées. Cette divergence a l'avantage d'être facilement estimable à partir d'échantillons.

Une divergence pour le PAC-Bayes. Nous définissons une pseudo-métrique [2] portant sur l'écart entre les désaccords mesurés sur les domaines. Elle quantifie la différence structurelle entre les marginales en terme de posterior ρ sur $\mathcal{H}$. Puisqu'en "PAC-Bayes" le but est d'apprendre un vote majorité B_ρ amenant à de bonnes garanties en généralisation, nous proposons de suivre l'idée portée par l'équation (4) (la C-borne). Concrètement, étant donnés P_S et P_T et une distribution ρ sur $\mathcal{H}$, si $R_{P_S}(G_\rho)$ et $R_{P_T}(G_\rho)$ sont similaires, alors $R_{P_S}(B_\rho)$ et $R_{P_T}(B_\rho)$ sont similaires lorsque $\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_S}(h,h')$ et $\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_T}(h,h')$ sont proches. Ainsi, les domaines P_S et P_T sont proches selon ρ si l'écart entre $\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_S}(h,h')$ et $\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_T}(h,h')$ est faible. Notre divergence est alors définie comme suit.

Définition 1. *Soit $\mathcal{H}$ une classe d'hypothèses. Pour toutes distributions marginales D_S et D_T sur X, et pour toute distribution ρ sur $\mathcal{H}$, le désaccord* $\mathrm{dis}_\rho(D_S,D_T)$ *entre D_S et D_T est défini par,*

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

Notons qu'il est trivial de démontrer que $\mathrm{dis}_\rho(\cdot,\cdot)$ est symétrique et vérifie l'inégalité triangulaire. Le théorème suivant garantit la consistance de cette mesure en majorant le désaccord $\mathrm{dis}_\rho(D_S,D_T)$ par les quantités classiques en PAC-Bayes : le désaccord empirique $\mathrm{dis}_\rho(S,T)$ estimé à partir d'un échantillon source et d'un échantillon cible (non étiquetés) et $\mathrm{KL}(\rho\|\pi)$. Pour simplifier [3], nous supposons $m=m'$.

Théorème 3. *Pour toutes marginales D_S et D_T sur X, pour tout espace d'hypothèses $\mathcal{H}$, pour toute distribution prior π sur $\mathcal{H}$, pour tout $\delta\in(0,1]$, pour tout réel $\alpha>0$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S\times T\sim(D_S\times D_T)^m$, pour toute distribution ρ sur $\mathcal{H}$, on a,*

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

où $\mathrm{dis}_\rho(S,T)$ *est le désaccord empirique.*

D'une manière similaire à la borne de Catoni [Cat07] (*c.f.* théorème 2 en section 3), le résultat précédent est consistent lorsque $\alpha=\frac{1}{2\sqrt{m}}$. En fait, la borne tend vers $1\times[\mathrm{dis}_\rho(S,T)+0+1]-1$ lorsque m tend vers $+\infty$. Dans l'annexe, nous fournissons une autre borne où la consistance apparaît plus directement, mais qui est moins adaptée pour dériver un algorithme.

Démonstration du Th.3. (détails en annexe) Nous majorons $d^{(1)} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h,h'\sim\rho^2}[R_{D_S}(h,h')-R_{D_T}(h,h')]$. Soit le classifieur "abstrait" $\hat{h} \stackrel{\text{def}}{=} (h,h') \in \mathcal{H}^2$ choisi selon une distribution $\hat{\rho}$, avec $\hat{\rho}(\hat{h})=\rho(h)\rho(h')$. Avec $\hat{\pi}(\hat{h})=\pi(h)\pi(h')$, on a $\mathrm{KL}(\hat{\rho}\|\hat{\pi})=2\mathrm{KL}(\rho\|\pi)$. Nous définissons la perte "abstraite" de $\hat{h}$ sur une paire d'exemples $(\mathbf{x}^s,\mathbf{x}^t)\sim D_S\times D_T$ par,
$\ell_{d^{(1)}}(\hat{h},\mathbf{x}^s,\mathbf{x}^t) \stackrel{\text{def}}{=} \frac{1+\ell_{0\text{-}1}(h(\mathbf{x}^s),h'(\mathbf{x}^s))-\ell_{0\text{-}1}(h(\mathbf{x}^t),h'(\mathbf{x}^t))}{2}$.
Le risque du classifieur de Gibbs suivant cette perte est $R^{(1)}_{D_S\times D_T}(G_{\hat{\rho}}) = \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}} \mathop{\mathbf{E}}_{\mathbf{x}^s\sim D_S} \mathop{\mathbf{E}}_{\mathbf{x}^t\sim D_T} \ell_{d^{(1)}}(\hat{h},\mathbf{x}^s,\mathbf{x}^t)$. Comme $\ell_{d^{(1)}}$ renvoie des valeurs dans $[0,1]$, nous suivons le principe de la preuve du Th. 2 (avec $c=2\alpha$). On borne alors la valeur réelle de $R^{(1)}_{D_S\times D_T}(G_{\hat{\rho}})$ (voir en annexe). Puis, nous majorons $d^{(1)}$ par sa valeur empirique ($d^{(1)}_{S\times T}$), car $d^{(1)}=2R^{(1)}_{D_S\times D_T}(G_{\hat{\rho}})-1$. Enfin, avec une *proba.* d'au moins $1-\frac{\delta}{2}$ sur le choix de $S\times T\sim(D_S\times D_T)^m$, on a,

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

En suivant le même principe, on majore $d^{(2)} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h,h'\sim\rho^2}[R_{D_T}(h,h')-R_{D_S}(h,h')]$ par sa valeur empirique $d^{(2)}_{S\times T}$. Comme $|d^{(1)}|=|d^{(2)}|=\mathrm{dis}_\rho(D_S,D_T)$ et $|d^{(1)}_{S\times T}|=|d^{(2)}_{S\times T}|=\mathrm{dis}_\rho(S,T)$, le maximum entre la borne sur $d^{(1)}$ et la borne sur $d^{(2)}$ donne une borne sur $\mathrm{dis}_\rho(D_S,D_T)$. En appliquant la borne de l'union, on a avec une *proba.* d'au moins $1-\delta$ sur le choix de $S\times T\sim(D_S\times D_T)^m$,

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

□

Avant de dériver en section 4.2, une borne d'AD pour le classifieur de Gibbs, nous comparons notre désaccord entre domaines avec la $\mathcal{H}$-divergence.

2. Contrairement à une métrique, deux points ne sont pas nécessairement discrenable par la pseudo-métrique : on peut avoir $d(x,x')=0$ pour des valeurs distinctes $x\neq x'$.

3. Notons qu'il est possible de dériver une borne PAC-Bayésienne lorsque $m\neq m'$, voir en annexe.

Comparaison de $\frac{1}{2}d_{\mathcal{H}}$ et dis_ρ. Estimer la $\mathcal{H}$-divergence est connue pour être NP-difficile et revient à trouver la paire de classifieurs maximisant le désaccord. Notre mesure de désaccord empirique apparaît donc plus simple à calculer puisqu'il suffit de calculer un ρ-moyennage des désaccords entre classifieurs. En effet, $\mathrm{dis}_\rho(D_S, D_T)$ dépend de la posterior ρ considérée : son optimisation peut se faire directement en minimisant son estimation empirique $\mathrm{dis}_\rho(S,T)$ et la KL-divergence sans avoir à modifier la représentation (repondération des instances, projection, ...). De plus, la valeur de la $\mathcal{H}$-divergence est la même pour toutes les hypothèses de $\mathcal{H}$ et ne dépend pas du classifieur considéré, alors que notre désaccord dis_ρ est adapté au classifieur de Gibbs considéré. En outre, dis_ρ est inférieur ou égale à $\frac{1}{2}d_{\mathcal{H}}$ rendant la borne plus précise : pour tout $\mathcal{H}$ et pour toute ρ sur $\mathcal{H}$, on a,

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

4.2 La borne d'AD PAC-Bayésienne

Le prochain théorème énonce notre résultat principal. Par souci de lisibilité, nous préférons utiliser les notations $R_P(G_\rho)$ et $R_D(G_\rho, \cdot)$, correspondant aux ρ-moyennages respectifs de $\mathop{\mathbf{E}}_{h\sim\rho} R_P(h)$ et $\mathop{\mathbf{E}}_{h\sim\rho} R_D(h,\cdot)$.

Théorème 4. *Soit $\mathcal{H}$ un espace d'hypothèses. Pour toute distribution ρ sur $\mathcal{H}$, on a,*

$$\begin{aligned}R_{P_T}(G_\rho) - R_{P_T}(G_{\rho_T^*}) \leq R_{P_S}(G_\rho) + \mathrm{dis}_\rho(D_S, D_T) \\ + R_{D_T}(G_\rho, G_{\rho_T^*}) + R_{D_S}(G_\rho, G_{\rho_T^*}), \quad (6)\end{aligned}$$

avec $\rho_T^* = \mathrm{argmin}_\rho R_{P_T}(G_\rho)$ le posterior optimal sur le domaine cible et $R_D(G_\rho, G_{\rho_T^*}) = \mathop{\mathbf{E}}_{h\sim\rho}\mathop{\mathbf{E}}_{h'\sim\rho_T^*} R_D(h,h')$.

Démonstration. Soit $\mathcal{H}$ une classe d'hypothèses, ρ sur $\mathcal{H}$, $\rho_T^* = \mathrm{argmin}_\rho R_{P_T}(G_\rho)$ la posterior optimale pour P_T. D'après l'inégalité triangulaire et puisque pour tout h et pour toute marginale D : $R_D(G_\rho, h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\mathbf{x}\sim D} \mathbf{I}[G_\rho(\mathbf{x}) \neq h(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim D}\mathop{\mathbf{E}}_{h'\sim\rho} \mathbf{I}[h'(\mathbf{x}) \neq h(\mathbf{x})]$, on a,

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

En général, cette borne n'est pas comparable à celles présentées en section 2. Cependant, comme la borne d'AD de l'équation (2) [MMR09a], nous majorons directement l'écart entre le ρ-moyennage des erreurs et le moyennage optimal. Notre borne s'exprime comme un compromis entre différentes quantités. $R_{P_S}(G_\rho)$ et $\mathrm{dis}_\rho(D_S, D_T)$ sont semblables aux deux premiers termes de la borne d'AD de l'équation (1) [BDBC+10] : $R_{P_S}(G_\rho)$ est le ρ-moyennage sur $\mathcal{H}$ des erreurs sources et $\mathrm{dis}_\rho(D_T, DS)$ quantifie le ρ-moyennage des désaccords mais est spécifique à la distribution ρ considérée. Les autres termes $R_{D_T}(G_\rho, G_{\rho_T^*})$ et $R_{D_S}(G_\rho, G_{\rho_T^*})$ mesurent à quel point la distribution ρ est proche (en terme de désaccords) du classifieur de Gibbs optimal, à la fois sur le domaine cible et le domaine source. D'après cette théorie, l'adaptation est possible si la distribution optimale ρ_T^* admet une erreur cible faible (hypothèse classique). De plus, la quantité $R_{D_T}(G_\rho, G_{\rho_T^*}) + R_{D_S}(G_\rho, G_{\rho_T^*})$, vue comme une mesure de la capacité d'adaptation en terme d'étiquetage, doit être faible : G_ρ doit être en accord (sur les deux domaines) avec la solution optimale. Finalement, le théorème 5 énonce notre borne PAC-Bayésienne basée à la fois sur l'erreur source empirique du classifieur de Gibbs et la pseudo-métrique de désaccords estimée sur des échantillons source et cible.

Théorème 5. *Pour tous domaines P_S et P_T sur $X\times Y$ (resp. de marginales D_S et D_T), pour toute classe $\mathcal{H}$ d'hypothèses, pour toute distribution π sur $\mathcal{H}$, pour tout $\delta\in(0,1]$, avec une probabilité de $1-\delta$ sur le choix de $S\times T\sim(P_S\times D_T)^m$, pour toute distribution ρ sur $\mathcal{H}$, on a,*

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

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

Démonstration. Dans le Th. 3, on remplace $R_S(G_\rho)$ et $\mathrm{dis}_\rho(S,T)$ par leur borne supérieure obtenue dans les Ths. 2 et 4 avec δ choisi *resp.* comme $\frac{\delta}{3}$ et $\frac{2\delta}{3}$ (dans le dernier cas, on utilise $\ln\frac{2}{2\delta/3} = \ln\frac{3}{\delta} < 2\ln\frac{3}{\delta}$). $\square$

Sous l'hypothèse qu'il existe un lien entre les domaines en termes d'accord d'étiquetage sur les domaines source et cible, autrement dit qu'un désaccord

$\text{dis}_\rho(D_S, D_T)$ faible implique un λ_ρ négligeable, une solution naturelle pour réaliser une AD PAC-Bayésienne (sans étiquette cible) revient à minimiser la borne du théorème 5 en négligeant [4] λ_ρ. Un avantage majeur de ce résultat est qu'il justifie théoriquement la minimisation simultanée de l'erreur source et de la divergence.

5 Algorithme d'AD PAC-Bayésien pour classifieurs linéaires

Définissons maintenant $\mathcal{H}$ comme un ensemble de classifieurs linéaires $h_{\mathbf{v}}(\mathbf{x}) \stackrel{\text{def}}{=} \text{sign}\,(\mathbf{v}\cdot\mathbf{x})$ tel que $\mathbf{v}\in\mathbb{R}^d$ est un vecteur de poids. En restreignant les distributions prior et posterior à des distributions gaussiennes, [LST02, APHST06] ont spécialisé la théorie PAC-Bayésienne afin de borner l'erreur réelle d'un ρ-moyennage de classifieurs linéaires $h_{\mathbf{w}}\in\mathcal{H}$ chacun identifié par un vecteur de poids $\mathbf{w}$. Plus précisément, pour un prior $\pi_{\mathbf{0}}$ et un posterior $\rho_{\mathbf{w}}$ définis par une gaussienne sphérique de matrice de covariance égale à l'identité centrée respectivement sur les vecteurs $\mathbf{0}$ et $\mathbf{w}$, *c.-à-d.* pour tout $h_{\mathbf{v}}\in\mathcal{H}$ on a,

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

L'erreur réelle du classifieur de Gibbs $G_{\rho_{\mathbf{w}}}$ sur un domaine P_S est alors donnée par,

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

où $\Phi(a) \stackrel{\text{def}}{=} \frac{1}{2}\left[1-\mathbf{Erf}\left(\frac{a}{\sqrt{2}}\right)\right]$, et $\mathbf{Erf}(\cdot)$ est la fonction d'erreur de Gauss. Dans cette situation, la KL-divergence entre $\rho_{\mathbf{w}}$ et $\pi_{\mathbf{0}}$ devient $\text{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}}) = \frac{1}{2}\|\mathbf{w}\|^2$.

Apprentissage supervisé PAC-Bayésien. La théorie PAC-Bayésienne spécifique aux classifieurs linéaires dans le cadre supervisé classique [GLLM09] suggère de minimiser la borne sur $R_{P_S}(G_{\rho_{\mathbf{w}}})$ du théorème 2. Étant donné un échantillon $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ et un hyperparamètre $C>0$, l'algorithme d'apprentissage effectue une descente de gradient pour trouver un vecteur de poids optimal $\mathbf{w}$ qui minimise,

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

Cet algorithme, nommé PBGD3, cherche un compromis entre le taux d'erreur empirique (exprimé en fonction de la perte Φ) et la complexité du moyennage appris (mesurée par le régulariseur $\|\mathbf{w}\|^2$). L'inconvénient (en pratique) de PBGD3 est que la fonction objectif n'est pas convexe. L'implémentation de la descente de gradient nécessite donc de nombreux redémarrages. En fait, nous avons réalisé une étude empirique approfondie et montré que PBGD3 était aussi performant (tout en étant plus rapide) en remplaçant la fonction perte Φ par sa relaxation convexe $\Phi_{\text{cvx}}(a) \stackrel{\text{def}}{=} \frac{1}{2} - \frac{a}{\sqrt{2\pi}}$ si $a\leq 0$, $\Phi(a)$ sinon. Nous proposons maintenant une version de cet algorithme adaptée à la borne du théorème 5 pour des classifieurs linéaires.

Minimisation de la borne d'AD. Sous l'hypothèse que les quantités non estimables du théorème 5, λ_ρ et $R_{P_T}(G_{\rho_T^*})$, sont négligeables nous définissons un algorithme PAC-Bayésien pour l'AD inspiré de PBGD3. Étant donnés un échantillon source $S = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^m$ *i.i.d.* selon P_S et un échantillon cible $T = \{(\mathbf{x}_i^t)\}_{i=1}^m$ *i.i.d.* selon D_T, on a,

$$C\,m\,R_S(G_{\rho_{\mathbf{w}}}) + A\,m\ \text{dis}_{\rho_{\mathbf{w}}}(S,T) + \text{KL}(\rho_{\mathbf{w}}\|\pi_{\mathbf{0}}), \quad (7)$$

où $\text{dis}_{\rho_{\mathbf{w}}}(S,T) = \left|\mathop{\mathbf{E}}_{h,h'\sim\rho_{\mathbf{w}}^2} R_S(h,h') - \mathop{\mathbf{E}}_{h,h'\sim\rho_{\mathbf{w}}^2} R_T(h,h')\right|$ est le désaccord empirique entre S et T spécialisé à une distribution $\rho_{\mathbf{w}}$ sur l'espace des classifieurs linéaires considéré. Les valeurs $A>0$ et $C>0$ sont des hyperparamètres de l'algorithme. Notons que les constantes α et c du théorème 5 peuvent être retrouvées à partir de n'importe quel A et C. Étant donné $\Phi_{\text{dis}}(a) \stackrel{\text{def}}{=} 2\,\Phi(a)\,\Phi(-a)$, pour toute marginale D, on a,

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

Ainsi, trouver la solution optimale de l'équation (7) revient à chercher le vecteur $\mathbf{w}$ qui minimise,

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

L'équation précédente est fortement non convexe. Afin de rendre le problème d'optimisation plus facile à résoudre, nous remplaçons la fonction de perte Φ par sa relaxation convexe Φ_{cvx} (similairement à PBGD3). L'optimisation se fait ensuite par une descente de gra-

4. Avec quelques étiquettes cibles, nous pourrions estimer λ_ρ.

dient. Le gradient de l'équation précédente est,

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

où $\Phi'_{\text{cvx}}(a)$, *resp.* $\Phi'_{\text{dis}}(a)$, est la valeur de la dérivée de $\Phi_{\text{cvx}}(\cdot)$, *resp.* $\Phi_{\text{dis}}(\cdot)$, évaluée au point A, et,

$$s' = \text{sign} \left[\sum_{i=1}^{m} \Phi_{\text{dis}} \left(\frac{\mathbf{w} \cdot \mathbf{x}_i^s}{\|\mathbf{x}_i^s\|} \right) - \Phi_{\text{dis}} \left(\frac{\mathbf{w} \cdot \mathbf{x}_i^t}{\|\mathbf{x}_i^t\|} \right) \right].$$

Cette nouvelle tâche d'optimisation demeure toutefois non convexe ($\Phi_{\text{dis}}(\cdot)$ est quasi-concave). Cependant, notre étude empirique montre qu'il n'est pas nécessaire d'effectuer plusieurs redémarrages pour trouver une solution convenable. Nous nommons cet algorithme d'adaptation de domaine PBDA. Notons qu'en appliquant l'astuce du noyau à PBDA, nous pouvons travailler avec le vecteur de poids dual $\boldsymbol{\alpha} \in \mathbb{R}^{2m}$. Étant donné un noyau $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, on a :

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

6 Expérimentations

Protocole expérimental. Nous avons évalué PBDA sur un jeu de données synthétiques dît des lunes jumelles, ainsi que sur un jeu de données d'analyse d'avis communément utilisé en adaptation de domaine. Nous avons comparé les résultats obtenus avec ceux de méthodes non adaptatives : la version de PBGD3 [GLLM09] convexifiée (section 5) et un SVM ; et de méthodes adaptatives : DASVM [5] [BM10], la méthode d'AD de co-apprentissage CODA [6] ayant montrée les meilleurs résultats dans [CWB11] sur le jeu de données considéré en section 6.2. La fonction objectif de PBDA est minimisée par l'algorithme *BFGS* implémenté dans la librairie python *scipy* [7]. Nous avons utilisé la bibliothèque SVM-light [Joa99] pour SVM, DASVM est implémenté avec la bibliothèque LibSVM [CL01] et nous avons utilisé l'implémentation [8] proposée par [CWB11] pour CODA. Les paramètres sont sélectionnés, à l'aide d'une grille de recherche, par une validation croisée avec 5-*folds* (^{CV}) sur l'échantillon source pour PBGD3 et SVM, et par une validation circulaire/inverse [BM10, ZFY+10] avec 5-*folds* sur l'échantillon cible (non étiqueté) pour CODA, DASVM et PBDA.

5. DASVM maximise itérativement une notion de marge sur des exemples cibles auto étiquetés.

6. CODA cherche itérativement des attributs cibles reliés à l'ensemble d'apprentissage.

7. *scipy* est disponible ici : `http://www.scipy.org/`

8. L'implémentation de CODA est disponible ici : `http://www1.cse.wustl.edu/~mchen/code/coda.tar`.

TABLE 1 – Taux d'erreur moyens pour les 7 angles.

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

6.1 Problème jouet synthétique

Nous considérons comme domaine source un problème classique de classification binaire appelé les lunes jumelles. Chacune des classes correspond à une lune distincte (cf figure 1). Nous étudions 7 domaines cibles différents en fonction de 7 rotations de la source initiale. Pour chaque domaine, nous générons 300 instances (150 de chaque classe). La capacité en généralisation des algorithmes est évaluée sur un échantillon de test composé de 1500 exemples cibles. Tous les algorithmes utilisent un noyau gaussien. Chacun des problèmes d'adaptation de domaine est répété 10 fois et nous reportons les résultats moyens dans la table 1. Notons que CODA n'est pas approprié pour cette expérimentation, puisqu'il décompose les attributs afin d'appliquer un co-apprentissage. PBDA obtient les meilleures performances, à l'exception des angles de 20° et 50°. Il montre une bonne capacité d'adaptation, en particulier pour les problèmes les plus difficiles. Ce comportement vient probablement du fait que notre divergence entre les domaines dis_ρ est plus précise et apparaît comme un bon (co-)régulariseur dans une situation d'AD. Ceci est confirmé par la figure 1. On y voit l'évolution de la frontière de décision de l'algorithme pour chacun des angles, et de la capacité d'adaptation en fonction de la minimisation du risque suggéré par le théorème 5. En effet, le graphique illustre que PBDA accepte de perdre en performance sur le domaine source pour maintenir sa performance sur le domaine cible (au moins quand les domaines ne sont pas trop différents).

6.2 Analyse d'avis

Nous considérons le jeu de données *Amazon reviews* [BMP06] constitué d'avis sur quatre types de produits issus de *Amazon.com*© (*books, DVDs, electronics, kitchen appliances*). À l'origine, les avis s'expriment à

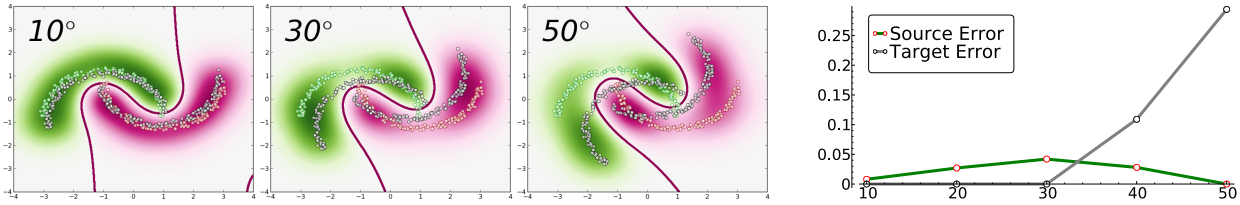

FIGURE 1 – Illustration de la frontière de décision de PBDA à paramètres fixés ($A = C = 1$). Le vert et le rose correspondent à l'échantillon source, le gris à l'échantillon cible. Le graphique de droite correspond au compromis erreur source / erreur cible.

l'aide d'étoiles (de 1 à 5) et la dimension de la description des données (des uni-grammes et des bi-grammes) est en moyenne de 100 000. Pour simplifier la tâche de classification, nous suivons un protocole similaire à celui proposé par [CWB11]. Les deux classes possibles sont : $+1$ pour les produits ayant au moins 4 étoiles, -1 pour ceux ayant au plus 3 étoiles. Un domaine correspond à un type de produit, ce qui implique 12 tâches d'AD. Par exemple, "*books*→*DVDs*" correspond à la tâche pour laquelle *books* est la source et *DVDs* la cible. La dimension des données est réduite de la manière suivante : étant donnée une tâche d'AD, [CWB11] ont uniquement gardé les attributs qui apparaissent au moins 10 fois dans les deux domaines (il reste environ 40 000 attributs), puis ont appliqué une pondération de type tf-idf. Les algorithmes font appel à un noyau linéaire et considèrent 2 000 exemples sources étiquetés et 2 000 cibles non étiquetés. Nous les évaluons sur les ensembles de test cibles proposés par [CWB11] (entre 3 000 et 6 000 exemples), puis nous reportons les résultats sur la table 2. Tout d'abord, nous pouvons remarquer que comme espéré les approches adaptatives obtiennent les meilleurs résultats en moyenne. Ensuite, PBDA est en moyenne plus performant que CODA, mais moins performant que DASVM. Cependant, PBDA reste compétitif : les résultats ne sont pas significativement différents de ceux obtenus pour CODA et DASVM. De plus, PBDA est significativement plus rapide que CODA et DASVM : ces deux algorithmes utilisent une méthode itérative coûteuse augmentant le temps d'exécution d'au moins un facteur 5 par rapport à PBDA. En fait, un avantage certain de PBDA est de pouvoir optimiser conjointement les termes de notre borne en une seule étape.

7 Conclusion

Dans cet article, nous avons défini une divergence entre distributions basée sur une notion désaccord moyen entre hypothèses, accompagnée de bornes de consistance justifiant de son estimation consistante à partir d'échantillons. Cette mesure nous a permis de dériver une borne PAC-Bayésienne pour l'AD. De plus, nous avons proposé, à partir de cette borne, un algorithme compétitif et fondé théoriquement (PBDA) qui minimise directement la borne dans le cas des classifieurs linéaires. Nous pensons que cette première analyse PAC-Bayésienne ouvre la porte à de nouvelles méthodes d'adaptation en faisant appel aux possibilités offertes par la théorie PAC-Bayésienne, ce qui donne lieu à de nombreuses questions intéressantes.
Un des intérêts du "PAC-Bayes" est de considérer une connaissance *a priori* sur la performance des classifieurs et dans ce papier nous avons opté pour un prior non informatif (une gaussienne centrée à l'origine dans l'espace des classifieurs linéaires). La définition d'un prior pertinent pour l'adaptation de domaine est à étudier lorsque quelques étiquettes cibles sont accessibles, ou lorsque différents domaines sources sont disponibles. Une autre piste prometteuse concerne la sélection des paramètres. En effet, l'adaptabilité de notre méthode pourrait être améliorée via une procédure de validation spécifique au cadre PAC-Bayésien. Une idée serait de considérer une technique de validation inverse tirant parti des distributions prior et posterior.
En outre, la dérivation d'un résultat similaire à l'équation (4) (la C-borne) pour l'adaptation de domaine semble être d'un grand intérêt. En effet, cette approche prend en compte les deux premiers moments de la marge du vote de majorité. Ceci pourrait nous aider à utiliser à la fois une information sur la marge sur les données non étiquetées et sur le désaccord entre hypothèses (ces deux éléments semblent être d'une importance cruciale en adaptation de domaine).

Remerciements. Travail financé par les projets VideoSense ANR-09-CORD-026 et LAMPADA ANR-09-EMER-007-02, et par la subvention à la découverte 262067 du CRSNG. Les calculs ont été effectués sur les infrastructures de Calcul Québec et de Calcul Canada (financées par la FCI, le CRSNG et le FRQ).

TABLE 2 – Taux d'erreur sur *Amazon reviews*. B, D, E, K correspondent à *books, DVDs, electronics, kitchen*.

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

Références

[APHST06] A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *NIPS*, pages 9–16, 2006.

[BDBC+10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2) :151–175, 2010.

[BDU12] S. Ben-David and R. Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *ALT*, pages 139–153, 2012.

[BM10] L. Bruzzone and M. Marconcini. Domain adaptation problems : A DASVM classification technique and a circular validation strategy. *Trans. Pattern Anal. Mach. Intell.*, 32(5) :770–787, 2010.

[BMP06] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.

[Cat07] O. Catoni. *PAC-Bayesian supervised classification : the thermodynamics of statistical learning*, volume 56. Inst of Mathematical Statistic, 2007.

[CL01] C.-C. Chang and C.-J. Lin. *LIBSVM : a library for support vector machines*, 2001. `www.csie.ntu.edu.tw/~cjlin/libsvm`.

[CWB11] M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *NIPS*, pages 2456–2464, 2011.

[GLL+09] Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Sara Shanian. From PAC-Bayes bounds to kl regularization. In *NIPS*, pages 603–610. 2009.

[GLLM09] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian Learning of Linear Classifiers. In *ICML*, 2009.

[Jia08] J. Jiang. A literature survey on domain adaptation of statistical classifiers. Technical report, CS Department at Univ. of Illinois at Urbana-Champaign, 2008.

[Joa99] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[LLM+06] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. PAC-bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *NIPS*, 2006.

[LMR11] F. Laviolette, M. Marchand, and J.-F. Roy. From PAC-Bayes Bounds to Quadratic Programs for Majority Votes. In *ICML*, 2011.

[LST02] J. Langford and J. Shawe-Taylor. PAC-bayes & margins. In *NIPS*, pages 439–446, 2002.

[McA99] D. A. McAllester. Some PAC-bayesian theorems. *Mach. Learn.*, 37 :355–363, 1999.

[MMR09a] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation : Learning bounds and algorithms. In *COLT*, pages 19–30, 2009.

[MMR09b] Y. Mansour, M. Mohri, and A. Rostamizadeh. Multiple source adaptation and the Rényi divergence. In *UAI*, pages 367–374, 2009.

[QCSSL09] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

[ZFY+10] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *ECML-PKDD*, 2010.

[ZZY12] C. Zhang, L. Zhang, and J. Ye. Generalization bounds for domain adaptation. In *NIPS*, 2012.