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

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

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

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

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

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

## 1. Some tools

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

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

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

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

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

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

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

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

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

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

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

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

## 2. Detailed Proof of Theorem 3

We recall the Theorem 3 of the main paper.

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

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

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

*Proof.* Firstly, we propose to upper-bound,

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

by its empirical counterpart,

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

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

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

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

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

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

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

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

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

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

The empirical counterparts of these two quantities are,

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

and,

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

It is easy to show that,

$$\begin{aligned}
d^{(1)} &= 2R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}}) - 1, \qquad (9) \\
d^{(1)}_{S\times T} &= 2R^{(1)}_{S\times T}(G_{\hat{\rho}}) - 1. \qquad (10)
\end{aligned}$$

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

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

and consider the non-negative random variable,

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

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

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

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

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

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

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

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

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

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

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

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

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

We then have,

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

This, in turn, implies that,

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

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

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

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

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

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

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

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

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

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

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

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

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

and,

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

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

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

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

or, which is equivalent,

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

and we are done. □

## 3. Other PAC-Bayesian Bounds

### 3.1. PAC-Bayesian Bounds with the kl term

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

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

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

Here is a "Seeger's type" PAC-Bayesian bound for our domain disagreement $\mathrm{dis}_\rho$.

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

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

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

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

by its empirical counterpart,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This implies that,

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

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

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

We claim that we also have,

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

which, since

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

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

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

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

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

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

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

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

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

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

as wanted.

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

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

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

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

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

$$\mathcal{R}_\rho \overset{\text{def}}{=} \left\{ r : \mathrm{kl}\big(R_S(G_\rho)\|r\big) \le \tfrac{1}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\tfrac{4\sqrt{m}}{\delta}\right]\right\},$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

By the Equation (8),

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

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

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

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

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

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

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

Let us now bound,

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

To do so, we have,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# 4. PBDA Algorithm Details

## 4.1. Objective function and gradient

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

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

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

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

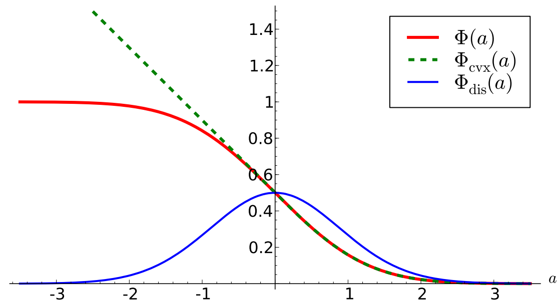


*Figure 1.* Behaviour of functions $\Phi(\cdot)$, $\Phi_{\mathrm{cvx}}(\cdot)$ and $\Phi_{\mathrm{dis}}(\cdot)$.

Figure 1 illustrates these three functions.

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

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

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

## 4.2. Using a kernel function

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

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

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

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

where,

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

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

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

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

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

where,

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

### 4.3. Implementation details

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

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

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

## References

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

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

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

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