Annexe de
Une analyse PAC-Bayésienne de l'adaptation de domaine

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

Preuve détaillée du Théorème 3

Nous rappelons le th. 3

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 $\delta \in (0,1]$, 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,*

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

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

Démonstration. Tout d'abord, nous bornons :

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

Pour cela, nous considérons un classifieur "abstrait" $\hat{h} \stackrel{\text{def}}{=} (h,h') \in \mathcal{H}^2$ choisi selon une distribution $\hat{\rho}$, telle que $\hat{\rho}(\hat{h}) = \rho(h)\rho(h')$. Notons qu'avec $\hat{\pi}(\hat{h}) = \pi(h)\pi(h')$, on obtient $\mathrm{KL}(\hat{\rho}\|\hat{\pi}) = 2\mathrm{KL}(\rho\|\pi)$. En effet :

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

Nous définissons la fonction 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 "abstrait" de $\hat{h}$ sur la distribution jointe est défini par : $R^{(1)}_{D_{S\times T}}(\hat{h}) = \mathop{\mathbf{E}}_{\mathbf{x}^s\sim D_S}\mathop{\mathbf{E}}_{\mathbf{x}^t\sim D_T}\ell_{d^{(1)}}(\hat{h},\mathbf{x}^s,\mathbf{x}^t)$. Ainsi, le risque du classifieur de Gibbs associé à cette perte est :

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

Les versions empiriques de ces deux quantités sont :

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

et

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

Puis que $\ell_{d^{(1)}}$ retourne des valeurs entre $[0,1]$, on peut borner $R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})$ en suivant le principe de la preuve du th. 2 (avec $c=2\alpha$). Pour ce faire, nous définissons la fonction convexe

$$\mathcal{F}(p) \stackrel{\text{def}}{=} -\ln\left[1-(1-\exp(-2\alpha))\,p\right],$$

et nous considérons la variable aléatoire non négative suivante :

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

Nous appliquons ensuite l'inégalité de Markov (voir le th. 9). Pour tout $\delta\in(0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S\times T\sim(D_{S\times T})^m$, on a :

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

En prenant le logarithme de chaque côté de l'inégalité, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S\times T \sim (D_{S\times T})^m$, et pour toute distribution posterior $\hat{\rho}$, on a :

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

Le choix de $\mathcal{F}$ implique :

$$\mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m}\ \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}\ \mathrm{e}^{m\left(\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))-2\alpha R^{(1)}_{S\times T}(\hat{h})\right)}$$
$$= \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}\ \mathrm{e}^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))}\ \mathop{\mathbf{E}}_{S\times T\sim(D_{S\times T})^m}\ \mathrm{e}^{-2m\alpha R^{(1)}_{S\times T}(\hat{h})}$$
$$= \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}\ \mathrm{e}^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))}\sum_{k=0}^{m}\Pr_{S\times T\sim(D_{S\times T})^m}\left(R^{(1)}_{S\times T}(\hat{h})=\tfrac{k}{m}\right)\mathrm{e}^{-2\alpha k}$$
$$= \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}\mathrm{e}^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))}\sum_{k=0}^{m}\tbinom{m}{k}\left(R^{(1)}_{S\times T}(\hat{h})\right)^k\left(1-R^{(1)}_{S\times T}(\hat{h})\right)^{m-k}\mathrm{e}^{-2\alpha k}$$
$$= \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}\mathrm{e}^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))}\sum_{k=0}^{m}\tbinom{m}{k}\left(R^{(1)}_{S\times T}(\hat{h})e^{-2\alpha}\right)^k\left(1-R^{(1)}_{S\times T}(\hat{h})\right)^{m-k}$$
$$= \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}\mathrm{e}^{m\mathcal{F}(R^{(1)}_{D_{S\times T}}(\hat{h}))}\left[R^{(1)}_{S\times T}(\hat{h})e^{-2\alpha}+\left(1-R^{(1)}_{S\times T}(\hat{h})\right)\right]^m$$
$$= \mathop{\mathbf{E}}_{\hat{h}\sim\hat{\rho}}\ 1$$
$$= 1\,.$$

On a :

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

On insère $\mathrm{KL}(\rho\|\pi)$ dans la partie gauche de l'inégalité précédente et on trouve une borne inférieure en appliquant deux fois l'inégalité de Jensen (voir th. 10) : tout d'abord sur la fonction logarithme (concave), puis sur la fonction convexe $\mathcal{F}$. On obtient :

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

On a :

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

puis l'inégalité $1-\mathrm{e}^{-x}\leq x$, donne :

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

Ainsi, puisque $d^{(1)}=2R^{(1)}_{D_{S\times T}}(G_{\hat{\rho}})-1$, on obtient une borne sur $d^{(1)}$ à partir de son estimation empirique (notée $d^{(1)}_{S\times T}$), Nous obtenons donc avec une probabilité d'au moins $1-\frac{\delta}{2}$ sur le choix de $S\times T\sim(D_S\times D_T)^m$,

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

Ensuite, nous bornons $d^{(2)}\stackrel{\text{def}}{=}\mathop{\mathbf{E}}_{h,h'\sim\rho^2}[R_{D_T}(h,h')-R_{D_S}(h,h')]$ en utilisant la même méthode :

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

Notons que $|d^{(1)}|=|d^{(2)}|=\mathrm{dis}_\rho(D_S,D_T)$ et $|d^{(1)}_{S\times T}|=|d^{(2)}_{S\times T}|=\mathrm{dis}_\rho(S,T)$. Ainsi, la borne sur $\mathrm{dis}_\rho(D_S,D_T)$ est obtenue en prenant le maximum de la borne sur $d^{(1)}$ et de la borne $d^{(2)}$. Finalement en appliquant la borne de l'union, avec une probabilité d'au moins $1-\delta$ sur le choix de $S\times T\sim(D_S\times D_T)^m$, on a

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

□

Bornes PAC-Bayésienne si $m\neq m'$

Nous rappelons tout d'abord la borne PAC-Bayésienne proposée par [McA03], énoncée sans terme de contrôle du compromis complexité/risque.

Théorème 6 ([McA03]). *Pour tout P_S sur $X\times Y$, pour tout ensemble $\mathcal{H}$, et pour toute distribution prior π sur $\mathcal{H}$, pour tout $\delta\in(0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S\sim(P_S)^m$, pour toute distribution posterior ρ sure $\mathcal{H}$, on a*

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

Nous pouvons maintenant prouver la borne de consistance suivante pour $\mathrm{dis}_\rho(D_S,D_T)$, pour le cas $m \neq m'$.

Théorème 7. *Pour toute distribution marginale D_S et D_T sur X, pour tout ensemble $\mathcal{H}$, pour toute distribution prior π sur $\mathcal{H}$, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (D_S)^m$ et $T \sim (D_T)^{m'}$, pour toute distribution posterior ρ sur $\mathcal{H}$, on a :*

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

Démonstration. Considérons la variable aléatoire non négative :

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

On applique l'inégalité de Markov (voir th. 9), pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S \sim (D_S)^m$, on a :

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

En prenant le logarithme de chaque côté de cette inégalité, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix $S \sim (D_S)^m$ et pour toute distribution ρ, on a :

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

Puis que $\ln(\cdot)$ est une fonction concave, on applique l'inégalité de Jensen (voir th. 10). Ainsi, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix $S \sim (D_S)^m$ et pour toute distribution ρ, on a :

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

D'après les propriétés de la KL-divergence (éq. (8)), on a :

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

Ainsi :

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

Pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix $S \sim (D_S)^m$ et pour toute distribution ρ, on a :

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

Puisque $2(a-b)^2$ est une fonction convexe, on applique encore une fois l'inégalité de Jensen :

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

Donc pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix $S \sim (D_S)^m$ et pour toute distribution ρ, on a :

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

Il ne reste plus qu'à borner :

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

On a :

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim(D_S)^m}\mathop{\mathbf{E}}_{h,h'\sim\pi^2} \mathrm{e}^{\frac{m}{2}(R_{D_S}(h,h') - R_S(h,h'))^2} \\
&= \mathop{\mathbf{E}}_{h,h'\sim\pi^2}\mathop{\mathbf{E}}_{S\sim(D_S)^m} \mathrm{e}^{\frac{m}{2}(R_{D_S}(h,h') - R_S(h,h'))^2} && (9)\\
&\leq \mathop{\mathbf{E}}_{h,h'\sim\pi^2}\mathop{\mathbf{E}}_{S\sim(D_S)^m} \mathrm{e}^{\frac{m}{4}\mathrm{kl}(R_{D_S}(h,h')\|R_S(h,h'))} && (10)\\
&\leq \sqrt{m}. && (11)
\end{aligned}$$

La ligne (9) provient de l'indépendance entre D_S et ρ. L'inégalité $2(q-p)^2 \leq kl(q\|p)$ pour tout $p,q \in [0,1]$ implique la ligne (10). La dernière ligne (11) est due au lemme de Maurer (voir lem. 1).
Ainsi, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix $S \sim (D_S)^m$ et pour toute

distribution ρ, on a :

$$\begin{aligned}
&\frac{m}{2}\left(\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_S}(h,h') - \mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_S(h,h')\right)^2 \\
&\quad\leq 2\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta} \\
\Leftrightarrow &\left(\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_S}(h,h') - \mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_S(h,h')\right)^2 \\
&\quad\leq \frac{2}{m}\left[2\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right] \\
\Leftrightarrow &\left|\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_S}(h,h') - \mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_S(h,h')\right| \\
&\quad\leq \sqrt{\frac{2}{m}\left[2\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]}. \qquad (12)
\end{aligned}$$

En suivant le même principe de preuve, on borne

$$\left|\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_T}(h,h') - \mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_T(h,h')\right|.$$

Pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix $S \sim (D_S)^m$ et pour toute distribution ρ, on a :

$$\begin{aligned}
&\left|\mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_{D_T}(h,h') - \mathop{\mathbf{E}}_{h,h'\sim\rho^2} R_T(h,h')\right| \\
&\leq \sqrt{\frac{2}{m'}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m'}}{\delta}\right]}. \qquad (13)
\end{aligned}$$

On remplace δ par $\frac{\delta}{2}$ dans les Inégalités (12) et (13). Puis en appliquant la borne de l'union, les deux résultats sont vrais simultanément avec une probabilité d'au moins $1-\delta$, et nous les réunissons [1] pour borner : $\left|\mathop{\mathbf{E}}_{h,h'\sim\rho^2}[R_{D_T}(h,h') - R_{D_S}(h,h')]\right| = \mathrm{dis}_\rho(D_S,D_T)$. □

On obtient alors la borne suivante.

Théorème 8. *Pour tout domaine P_S et P_T (de marginale respective D_S et D_T) sur $X \times Y$, et pour tout ensemble $\mathcal{H}$, pour toute distribution π sur $\mathcal{H}$, pour tout $\delta \in (0,1]$, avec une probabilité d'au moins $1-\delta$ sur le choix de $S_1 \sim (D_S)^{m_1}$, $S_2 \sim (D_S)^{m_2}$, et $T \sim (D_T)^{m'}$, pour toute distribution posterior ρ sure $\mathcal{H}$, on a :*

$$\begin{aligned}
R_{P_T}(G_\rho) - R_{P_T}(G_{\rho^*_T}) \leq R_S(G_\rho) + \mathrm{dis}_\rho(S,T) + \lambda_\rho \\
+\sqrt{\frac{2}{m_1}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{8\sqrt{m_1}}{\delta}\right]} \\
+\sqrt{\frac{2}{m_2}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{8\sqrt{m_2}}{\delta}\right]} \\
+\sqrt{\frac{2}{m'}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{4\sqrt{m'}}{\delta}\right]}.
\end{aligned}$$

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

Démonstration. Le résultat est directement obtenu en insérant les ths. 6, 7 (avec $\delta := \frac{\delta}{2}$) dans Th. 4. □

Quelques outils

Théorème 9 (Inégalité de Markov)**.** *Soit Z une variable aléatoire et $\epsilon \geq 0$, alors :*

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

Théorème 10 (Inégalité de Jensen)**.** *Soit X une variable aléatoire réelle intégrable et $g(\cdot)$ convexe, alors :*

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

Lemme 1 (des inégalités (1) et (2) de [Mau04])**.** *Soit $m \geq 8$, soit $X = (X_1, \ldots, X_m)$ un vecteur i.i.d. de variables aléatoires, $0 \leq X_i \leq 1$. Alors :*

$$\sqrt{m} \leq \mathbf{E}\ \exp(m\mathrm{kl}(\tfrac{1}{m}\textstyle\sum_{i=1}^n X_i \big\| \mathbf{E}\ [X_i])) \leq 2\sqrt{m},$$

où $\mathrm{kl}(a\|b) = a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$.

Références

[Mau04] A. Maurer. A note on the pac bayesian theorem. *CoRR*, cs.LG/0411099, 2004.

[McA03] D. McAllester. PAC-Bayesian stochastic model selection. *Mach. Learn.*, 51 :5–21, 2003.

1. Si $-c'_1 \leq a_1 - b_1 \leq c_1$ et $-c'_2 \leq a_2 - b_2 \leq c_2$, alors $-(c'_1 + c'_2) \leq (a_1 - a_2) - (b_1 - b_2) \leq c'_1 + c'_2$.