A New PAC-Bayesian Perspective on Domain Adaptation – Supplementary Material

Pascal Germain, Amaury Habrard, François Laviolette, Emilie Morvant

1 Proof of Theorem 4

Proof. We use the following shorthand notation:

$$\mathcal{L}_{\mathcal{D}}(h) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} \ell(h,\mathbf{x},y) \quad \text{and} \quad \mathcal{L}_S(h) = \frac{1}{m}\sum_{(\mathbf{x},y)\in S} \ell(h,\mathbf{x},y)\,.$$

Consider any convex function $\Delta : [0,1]\times[0,1] \to \mathbb{R}$. Applying consecutively Jensen's Inequality and the *change of measure inequality* (see Seldin & Tishby (2010, Lemma 4) and McAllester (2013, Equation (20))), we obtain

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

with

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

Then, Markov's Inequality gives

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

and

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

where the last inequality is due to Maurer (2004, Lemma 3) (we have an equality when the output of ℓ is in $\{0,1\}$). As shown in Germain et al. (2009, Corollary 2.2), by fixing

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

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

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

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

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

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

2 Using DALC with a kernel function

Let $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^{m_s}$, $T=\{\mathbf{x}'_i\}_{i=1}^{m_t}$ and $\mathrm{M}=m_s+m_t$. We will denote

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

The kernel trick allows us to work with dual weight vector $\boldsymbol{\alpha}\in\mathbb{R}^{\mathrm{M}}$ 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

$$h_{\mathbf{w}}(\cdot) \;=\; \operatorname{sign}\left[\sum_{i=1}^{\mathrm{M}}\alpha_i k(\mathbf{x}_i,\cdot)\right].$$

Let us denote K the kernel matrix of size $\mathrm{M}\times\mathrm{M}$ such as $K_{i,j}=k(\mathbf{x}_i,\mathbf{x}_j)$. In that case, the objective function—Equation (13) of the main paper—can be rewritten in term of the vector

$$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots\alpha_{\mathrm{M}})$$

as

$$C\times\sum_{i=m_s}^{\mathrm{M}}\Phi\left(\frac{\sum_{j=1}^{\mathrm{M}}\alpha_jK_{i,j}}{\sqrt{K_{i,i}}}\right)\Phi\left(-\frac{\sum_{j=1}^{\mathrm{M}}\alpha_jK_{i,j}}{\sqrt{K_{i,i}}}\right)+B\times\sum_{i=1}^{m_s}\left[\Phi\left(y_i\frac{\sum_{j=1}^{\mathrm{M}}\alpha_jK_{i,j}}{\sqrt{K_{i,i}}}\right)\right]^2+\sum_{i=1}^{\mathrm{M}}\sum_{j=1}^{\mathrm{M}}\alpha_i\alpha_jK_{i,j}\,.$$

For our experiments, we minimize this objective function using a *Broyden-Fletcher-Goldfarb-Shanno method (BFGS)* implemented in the *scipy* python library Jones et al. (2001–).

We initialize the optimization procedure at $\alpha_i=\frac{1}{\mathrm{M}}$ for all $i\in\{1,\ldots,\mathrm{M}\}$.

3 Experimental Protocol

For obtaining the DALC^{RCV} results of Table 1, the reverse validation procedure searches on a 20×20 parameter grid for a C between 0.01 and 10^6 and a parameter B between 1.0 and 10^8, both on a logarithm scale. The results of the other algorithms are reported from Germain et al. (2013).

References

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *ICML*, pp. 353–360, 2009.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *ICML*, pp. 738–746, 2013.

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

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

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

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