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:

\[ \mathcal{L}_{D}(h) = \mathbb{E}_{(x,y) \sim D} \ell(h, x, y) \quad \text{and} \quad \mathcal{L}_{S}(h) = \frac{1}{m} \sum_{(x,y) \in S} \ell(h, x, y). \]

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

\[ \forall \rho \text{ on } \mathcal{H} : \quad m \times \Delta \left( \mathbb{E}_{h \sim \rho} \mathcal{L}_{S}(h), \mathbb{E}_{h \sim \rho} \mathcal{L}_{D}(h) \right) \leq \mathbb{E}_{h \sim \rho} m \times \Delta (\mathcal{L}_{S}(h), \mathcal{L}_{D}(h)) \]

\[ \leq \text{KL}(\rho \| \pi) + \ln \left[ X_\pi(S) \right], \]

with

\[ X_\pi(S) = \mathbb{E}_{h \sim \pi} e^{m \times \Delta (\mathcal{L}_{S}(h), \mathcal{L}_{D}(h))}. \]

Then, Markov’s Inequality gives

\[ \Pr_{S \sim D^m} \left( X_\pi(S) \leq \frac{1}{\delta} \mathbb{E}_{S' \sim D^m} X_\pi(S') \right) \geq 1 - \delta, \]

and

\[ \mathbb{E}_{S' \sim D^m} X_\pi(S') = \mathbb{E}_{S' \sim D^m} \mathbb{E}_{h \sim \pi} e^{m \times \Delta (\mathcal{L}_{S'}(h), \mathcal{L}_{D}(h))} \]

\[ = \mathbb{E}_{h \sim \pi} \mathbb{E}_{S' \sim D^m} e^{m \times \Delta (\mathcal{L}_{S'}(h), \mathcal{L}_{D}(h))} \]

\[ \leq \mathbb{E}_{h \sim \pi} \sum_{k=0}^{m} \frac{k}{m} \left( \mathcal{L}_{D}(h) \right)^{k} \left( 1 - \mathcal{L}_{D}(h) \right)^{m-k} e^{m \times \Delta (\mathcal{L}_{S'}(h), \mathcal{L}_{D}(h))}, \]

(1)

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

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

Line 1 becomes equal to 1, and then

\[ \mathbb{E}_{S' \sim D^m} X_\pi(S') \leq 1. \]

Hence,

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

By reorganizing the terms, we have, with probability \( 1 - \delta \) over the choice of \( S \in D^m \),

\[ \forall \rho \text{ on } \mathcal{H} : \quad \mathbb{E}_{h \sim \rho} \mathcal{L}_{D}(h) \leq \frac{1}{1 - e^{-c}} \left[ 1 - \exp \left( -c \times \mathbb{E}_{h \sim \rho} \mathcal{L}_{S}(h) - \frac{\text{KL}(\rho \| \pi) + \ln \frac{1}{\delta}}{m} \right) \right]. \]

The final result is obtained by using the inequality \( 1 - \exp(-z) \leq z \).
2 Using DALC with a kernel function

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

$$x_# = \begin{cases} x_i & \text{if } # \leq m_s \\
 x'_i - m_s & \text{otherwise.} \end{cases}$$

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

$$h_w(\cdot) = \text{sign} \left[ \sum_{i=1}^{m} \alpha_i k(x_i, \cdot) \right].$$

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

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$$

as

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

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 $\alpha_i = \frac{1}{m}$ for all $i \in \{1, \ldots, M\}$.

3 Experimental Protocol

For obtaining the $\text{DALC}^{RCV}$ results of Table 1, the reverse validation procedure searches on a $20 \times 20$ 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


