A PAC-Bayes Sample-compression Approach to Kernel Methods : Supplementary material

Pascal Germain, Alexandre Lacoste, François Laviolette, Mario Marchand and Sara Shanian.

Département d'informatique et de génie logiciel, Université Laval
Québec, Canada, G1V 0A6

Last Revision: July 18, 2011

This supplementary refers to the ICML 2011 paper *A PAC-Bayes Sample-compression Approach to Kernel Methods* of Germain et al. (2011).

The lastest revision of this document and related source code are available at:
`http://graal.ift.ulaval.ca/publications.php`

Contents

1 Proof of Claim 1 in the proof of Theorem 5

Claim 1 : $\mathbf{E}_{S\sim D^m} X_{\overline{P}} \;\leq\; e^{4ld}\cdot 2\sqrt{m}\,.$

The claim uses the following lemma.

Lemma 0. *(Maurer (2004)) Let $n \geq 8$, and suppose that $X = (X_1,\ldots,X_n)$ is a vector of iid random variables, $0 \leq X_i \leq 1$, $\mathbf{E}[X_i] = \nu$ and let $M(X) = \frac{1}{n}\sum_{j=1}^n X_i$ be the arithmetic mean of the random variables. Then*

$$\sqrt{n} \;\;\leq\;\; \mathbf{E}\, e^{n\,\mathrm{kl}(M(X)\|\nu)} \;\;\leq\;\; 2\sqrt{n}\,.$$

Proof of Claim 1. Let us first recall that

$$X_{\overline{P}} \;\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{\bar{h}\sim\overline{P}} e^{(m-|\mathbf{i}_{\bar{h}}|)\cdot 2(\overline{R}_S(\bar{h})-\overline{R}_D(\bar{h}))^2}.$$

Let us also introduce a new notation. Let $\widetilde{R}_S(\bar{h})$ be the *abstract empirical risk* computed on the examples of S that *are not* in the compression sequence of $\bar{h}$. More formally,

$$\widetilde{R}_S(\,\overline{h_1..h_k}\,) \;\stackrel{\text{def}}{=}\; \frac{1}{m-|\mathbf{i}_{\overline{h_1..h_k}}|}\sum_{j=1}^m I\left(\neg\veebar_{i=1}^k (h_i(x_j)\neq y_j)\right) I\left((x_j,y_j)\notin \mathbf{i}_{\overline{h_1..h_k}}\right),$$

where $\veebar$ denotes the exclusive or. Based on the definition of Equation (9) in the main paper, one can easily show that $\widetilde{R}_S(\bar{h}) = \overline{R}_U(\bar{h})$ with $U = S\backslash S_{\mathbf{i}_{\bar{h}}}$, where $\mathbf{i}_{\bar{h}}$ points to the examples belonging to the union of all compression sequences of the sc-classifiers in $\bar{h}$. Moreover, note that, contrarily to $\overline{R}_S(\bar{h})$ which may contain some bias, $\widetilde{R}_S(\bar{h})$ is an arithmetic mean of truly iid $(m-|\mathbf{i}_{\bar{h}}|)$ random variables. On another hand, these two random variables have values that are very close to each other. Indeed, since

$$0 \;\;\leq\;\; m\cdot\overline{R}_S(\bar{h}) - (m-|\mathbf{i}_{\bar{h}}|)\cdot\widetilde{R}_S(\bar{h}) \;\;\leq\;\; |\mathbf{i}_{\bar{h}}|\,,$$

we have

$$-|\mathbf{i}_{\bar{h}}| \;\leq\; -|\mathbf{i}_{\bar{h}}|\cdot\widetilde{R}_S(\bar{h}) \;\leq\; m\cdot\overline{R}_S(\bar{h}) - m\cdot\widetilde{R}_S(\bar{h}) \;\leq\; |\mathbf{i}_{\bar{h}}| - |\mathbf{i}_{\bar{h}}|\cdot\widetilde{R}_S(\bar{h}) \;\leq\; |\mathbf{i}_{\bar{h}}|\,.$$

Therefore,

$$\left|\overline{R}_S(\bar{h}) - \widetilde{R}_S(\bar{h})\right| \;\leq\; \frac{|\mathbf{i}_{\bar{h}}|}{m}\,. \tag{20}$$

Given a compression sequence $S_{\mathbf{i}}$, let $\mathbf{i}^c$ be the vector of indices of $\mathcal{I}$ that are not in the vector $\mathbf{i}$.. We now have

$$\begin{aligned}\mathbf{E}_{S\sim D^m} X_{\overline{P}} \;&\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{\bar{h}\sim\overline{P}} e^{(m-|\mathbf{i}_{\bar{h}}|)\cdot 2(\overline{R}_S(\bar{h})-\overline{R}_D(\bar{h}))^2}\\ &=\; \mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{P}_{\mathcal{I}}}\mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}}\mathop{\mathbf{E}}_{\mu\sim\overline{P}_{S_{\mathbf{i}}}}\mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-|\mathbf{i}|}} e^{(m-|\mathbf{i}|)\cdot 2\,(\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))^2},\end{aligned}$$

Hence, to prove Claim 1, it suffices to show that for all $\mathbf{i}\in\mathcal{I}$, $S_{\mathbf{i}}\in(\mathcal{X}\times\mathcal{Y})^{|\mathbf{i}|}$, and $\mu\in\mathcal{M}_{S_{\mathbf{i}}}$, we have

$$\mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-|\mathbf{i}|}} e^{(m-|\mathbf{i}|)\cdot 2\,(\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))^2} \;\;\leq\;\; e^{4ld}\cdot 2\sqrt{m}\,. \tag{21}$$

Here is the sketch of the proof of Equation (21). Justification for Line (22) to (25) follows below.

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-|\mathbf{i}|}} e^{(m-|\mathbf{i}|)\cdot 2\,(\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))^2} \\
&\quad= \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{(m-|\mathbf{i}|)}} e^{(m-|\mathbf{i}|)\cdot 2\left(\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})-\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})+\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right)^2} \\
&\quad\le \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{(m-|\mathbf{i}|)}} e^{(m-|\mathbf{i}|)\cdot 2\left(\left[\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})-\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})\right]^2+2\left|\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})-\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})\right|\cdot\left|\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right|+\left[\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right]^2\right)} \\
&\quad\le \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{(m-|\mathbf{i}|)}} e^{(m-|\mathbf{i}|)\cdot 2\left(\left[\frac{|\mathbf{i}|}{m}\right]^2+2\frac{|\mathbf{i}|}{m}\cdot 1+\left[\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right]^2\right)} \qquad (22) \\
&\quad\le \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{(m-|\mathbf{i}|)}} e^{4ld\,+\,(m-|\mathbf{i}|)\cdot 2\left[\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right]^2} \qquad (23) \\
&\quad\le \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{(m-|\mathbf{i}|)}} e^{4ld\,+\,(m-|\mathbf{i}|)\cdot \mathrm{kl}\left(\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})\|\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right)} \qquad (24) \\
&\quad= e^{4ld}\cdot \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{(m-|\mathbf{i}|)}} e^{(m-|\mathbf{i}|)\,\cdot\,\mathrm{kl}\left(\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})\|\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right)} \\
&\quad\le e^{4ld}\cdot 2\sqrt{m-|\mathbf{i}|} \qquad (25) \\
&\quad\le e^{4ld}\cdot 2\sqrt{m}
\end{aligned}
$$

Line (22) follows from Equation (20) and the fact that the exponential function is increasing. For Line (23), since $|\mathbf{i}| \le ld$, we simply have to show that $(m-|\mathbf{i}|)\cdot 2\left(\frac{|\mathbf{i}|}{m^2}+\frac{2}{m}\right) \le 4$, which follows from direct calculations. Line (24) follows directly from the property : $2(q-p)^2 \le \mathrm{kl}(q\,\|\,p)$, which can also be proved via straightforward calculations. Finally, for Line (25), first observe that $\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})$ is an arithmetic mean of $(m-|\mathbf{i}|)$ iid random variables. Thus Line (25) is simply an application of Lemma 0 with $M(X)$ replaced by $\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})$, n replaced by $m-|\mathbf{i}|$, and ν replaced by $\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})$. Thus, Claim 1 is proved. □

2 Proof of Theorem 1: The PAC-Bayes bound with $KL(Q\|P)$

Theorem 1. *For any D, any family $(\mathcal{H}^S)_{S\in D^m}$ of sets of sc-classifiers of size at most l, any prior $\mathcal{P}$, any $\delta \in (0,1]$, any positive real number C_1, and any margin loss function ζ such that $l\cdot \deg(\zeta) < m$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall Q \text{ on } \mathcal{H}^S\text{:}\\ \zeta_D^Q \leq \zeta(1)[C'-1] + C'\cdot\left(\zeta_S^Q + \frac{2}{m\cdot C_1}[\zeta'(1)\cdot \mathrm{KL}(Q\|P) + \zeta(1)\cdot \ln\frac{1}{\delta}]\right)\end{array}\right) \geq 1-\delta$$

where $\mathrm{KL}(\cdot\|\cdot)$ *is the Kullback-Leibler divergence, and where* $C' = \frac{C_1\cdot\frac{m}{m-l\cdot\deg\zeta}}{1-e^{-C_1\cdot\frac{m-l\cdot\deg\zeta}{m}}}$.

Proof. Let S be any training sequence, $d \stackrel{\text{def}}{=} \deg\zeta$. Let $\mathcal{F}$ be a convex function to be defined later, and $\mathcal{D}(q,p) \stackrel{\text{def}}{=} \mathcal{F}(p) - C_1\cdot q$.

For each $k\in\{0,..,d\}$ and any k-tupple $(h_1,..,h_k)$, let us define $\bar{h} = \overline{h_1..h_k}$ as an "abstract" sc-classifier whose "abstract" true and empirical risks are respectively defined as

$$\begin{array}{rcl}\overline{R}_D(\,\overline{h_1..h_k}\,) &\stackrel{\text{def}}{=}& \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(\neg\underline{\vee}_{i=1}^k(h_i(x)\neq y))\,,\\ \overline{R}_S(\,\overline{h_1..h_k}\,) &\stackrel{\text{def}}{=}& \mathop{\mathbf{E}}_{(x,y)\sim S} I(\neg\underline{\vee}_{i=1}^k(h_i(x)\neq y)) \;=\; \frac{1}{m}\sum_{j=1}^m I(\neg\underline{\vee}_{i=1}^k(h_i(x_j)\neq y_j))\,,\end{array}$$

where $\underline{\vee}$ denotes the exclusive or.

Considering the above definitions, we also have

$$\overline{R}_D(\,\overline{h_1..h_k}\,) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(\neg\underline{\vee}_{i=1}^k(h_i(x)\neq y)) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\frac{1}{2}\left[1+\prod_{i=1}^k -yh_i(x)\right].$$

Since the compression sequence size of each h_i's is at most l, we have $|\mathbf{i}_{\bar{h}}| \leq l\cdot k$ for any $\bar{h} = \overline{h_1..h_k}$. Moreover, we have $\overline{R}_D(\overline{h_1..h_0}) = 1$ when $k=0$ because the exclusive or over an empty sequence outputs *false*. For each S, let $\overline{\mathcal{H}^S}$ be the set of all such sc-classifiers, and for each distribution Q on $\mathcal{H}^S$, denote by $\overline{Q}$ the following distribution on $\overline{\mathcal{H}^S}$:

$$\overline{Q}(\,\overline{h_1..h_k}\,) \;\stackrel{\text{def}}{=}\; \frac{a_k}{\zeta(1)}\,Q(h_1)\cdot\ldots\cdot Q(h_k) = \frac{a_k}{\zeta(1)}\prod_{i=1}^k Q(h_i)\,.$$

Since $\zeta(1) = \sum_{k=0}^d a_k$, $\overline{Q}$ is a probability distribution. We also denote by $\overline{P}$ the following distribution on $\overline{\mathcal{H}^S}$:

$$\overline{P}(\overline{h_1..h_k}) \stackrel{\text{def}}{=} \frac{a_k}{\zeta(1)}\cdot P(h1)\cdot\ldots\cdot P(h_k) \;=\; \frac{a_k}{\zeta(1)}\prod_{i=1}^k P(h_i)\,.$$

Moreover, for $U=D$ and $U=S$, we have

$$\begin{aligned} R_U(G_{\overline{Q}}) &= \mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} \overline{R}_U(\bar{h}) \\ &= \sum_{k=0}^{\deg(\zeta)} \frac{a_k}{\zeta(1)} \mathop{\mathbf{E}}_{h_1\sim Q} \cdots \mathop{\mathbf{E}}_{h_k\sim Q} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \frac{1}{2}\left[1+\prod_{i=1}^{k} -yh_i(x)\right] \\ &= \sum_{k=0}^{\deg(\zeta)} \frac{a_k}{\zeta(1)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \frac{1}{2}\left[1+\prod_{i=1}^{k} \mathop{\mathbf{E}}_{h_i\sim Q} -yh_i(x)\right] \\ &= \frac{1}{2}\left[1+\frac{1}{\zeta(1)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \sum_{k=0}^{\deg(\zeta)} a_k \big(\mathop{\mathbf{E}}_{h\sim Q} -yh(x)\big)^k\right] \\ &= \frac{1}{2}\left[1+\frac{1}{\zeta(1)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \sum_{k=0}^{\deg(\zeta)} a_k (-M_Q(x,y))^k\right] \\ &= \frac{1}{2}\left[1+\frac{1}{\zeta(1)}\ \zeta_U^Q\right]. \end{aligned} \tag{26}$$

Now, let us consider the following Laplace transform

$$X_{\overline{P}} \stackrel{\text{def}}{=} \mathbf{E}_{\bar{h}\sim\overline{P}}\, e^{(m-|\mathbf{i}_{\bar{h}}|)\mathcal{D}(\overline{R}_S(\bar{h}),\overline{R}_D(\bar{h}))}\,.$$

It then follows from Markov's inequality that

$$\Pr_{S\sim D^m}\left(X_{\overline{P}} \le \frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} X_{\overline{P}}\right) \ge 1-\delta\,.$$

By taking the logarithm on each side of the innermost inequality and by transforming the expectation over $\overline{P}$ into an expectation over $\overline{Q}$, we obtain

$$\Pr_{S\sim D^m}\left(\forall Q\colon \ln\left[\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} \frac{\overline{P}(\bar{h})}{\overline{Q}(\bar{h})} e^{(m-|\mathbf{i}_{\bar{h}}|)\mathcal{D}(\overline{R}_S(\bar{h}),\overline{R}_D(\bar{h}))}\right] \le \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{\bar{h}\sim\overline{P}} e^{(m-|\mathbf{i}_{\bar{h}}|)\mathcal{D}(\overline{R}_S(\bar{h}),\overline{R}_D(\bar{h}))}\right]\right) \ge 1-\delta\,. \tag{27}$$

Since $\zeta'(1)=\sum_{k=1}^{d} k\cdot a_k\,,$ by straightforward calculation, one can show that

$$\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} \ln\frac{\overline{P}(\bar{h})}{\overline{Q}(\bar{h})} = -\frac{\zeta'(1)}{\zeta(1)}\cdot KL(Q\|P)\,.$$

This, together with Jensen's inequality applied to the concave $\ln(x)$, gives

$$\ln\left[\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} \frac{\overline{P}(\bar{h})}{\overline{Q}(\bar{h})}\, e^{(m-|\mathbf{i}_{\bar{h}}|)\mathcal{D}(\overline{R}_S(\bar{h}),\overline{R}_D(\bar{h}))}\right] \ge -\frac{\zeta'(1)}{\zeta(1)}\cdot \mathrm{KL}(Q\|P) + \mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} (m-|\mathbf{i}_{\bar{h}}|)\,\mathcal{D}(\overline{R}_S(\bar{h}),\overline{R}_D(\bar{h}))\,. \tag{28}$$

Again from the Jensen's inequality applied to the convex function $\mathcal{F}$, together with Equation (26), and the fact that $m-l\cdot d \le (m-|\mathbf{i}_{\bar{h}}|) \le m\,,$ we obtain

$$\begin{aligned} \mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} (m-|\mathbf{i}_{\bar{h}}|)\ \mathcal{D}(\overline{R}_S(\bar{h}),\overline{R}_D(\bar{h})) &\ge (m-ld) \mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} \mathcal{F}(\overline{R}_D(\bar{h})) - m \mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} C_1\cdot\overline{R}_S(\bar{h})) \\ &\ge (m-ld)\,\mathcal{F}\left(\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\ \zeta_D^Q\right]\right) - m\,C_1\cdot\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\ \zeta_S^Q\right]. \end{aligned} \tag{29}$$

Let us analyze the value of $\mathbf{E}_{S\sim D^m} X_{\overline{P}}$, appearing in the right-hand side of the innermost inequality of Equation (27). First, let us define $\mathbf{i}^c$ as the vector of indices of $\mathcal{I}$ that are not in the vector $\mathbf{i}$. Thus, $|\mathbf{i}^c| = m - |\mathbf{i}|$. Now, note that

$$\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{\bar{h}\sim\overline{P}} e^{(m-|\mathbf{i}_{\bar{h}}|)\,\mathcal{D}(\overline{R}_S(\bar{h}),\overline{R}_D(\bar{h}))} = \mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{P}_{\mathcal{I}}} \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}} \mathop{\mathbf{E}}_{\mu\sim\overline{P}_{S_{\mathbf{i}}}} \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-|\mathbf{i}|}} e^{|\mathbf{i}^c|\,\mathcal{D}(\overline{R}_S(\overline{h^\mu_{\mathbf{i}}}),\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))}\,.$$

Now, for each $\overline{h^\mu_{\mathbf{i}}} \in \overline{\mathcal{H}^S}$, define $a^\mu_{S_{\mathbf{i}}} \stackrel{\text{def}}{=} \sum_{(x,y)\in S_{\mathbf{i}}} I(\overline{h^\mu_{\mathbf{i}}}(x) \neq y)$. Observe that $m\cdot\overline{R}_S(\overline{h^\mu_{\mathbf{i}}}) - a^\mu_{S_{\mathbf{i}}}$ is the number of errors made by $\overline{h^\mu_{\mathbf{i}}}$ on $S_{\mathbf{i}^c}$. Since the later is iid and disjoint from the compression sequence of $\overline{h^\mu_{\mathbf{i}}}$, we have that $m\cdot\overline{R}_S(\overline{h^\mu_{\mathbf{i}}}) - a^\mu_{S_{\mathbf{i}}}$ is a random variable following a binomial law of parameters $(|\mathbf{i}^c|, \overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))$. Since, $(m-ld)\cdot\frac{k}{m} \leq |\mathbf{i}^c|\cdot\frac{k+a^\mu_{S_{\mathbf{i}}}}{m}$ for any $k\in\{0,..,|\mathbf{i}^c|\}$, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-|\mathbf{i}|}} e^{|\mathbf{i}^c|\mathcal{D}(\overline{R}_S(\overline{h^\mu_{\mathbf{i}}}),\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))} &= \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-|\mathbf{i}|}} e^{|\mathbf{i}^c|\mathcal{F}(\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))-C_1|\mathbf{i}^c|\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})} \\
&= e^{|\mathbf{i}^c|\mathcal{F}(\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))\cdot\sum_{k=0}^{|\mathbf{i}^c|}\Pr_{S\sim D^m}\left(m\cdot\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})-a^\mu_{S_{\mathbf{i}}}=k\right)\,e^{-C_1|\mathbf{i}^c|\cdot\frac{k+a^\mu_{S_{\mathbf{i}}}}{m}}} \\
&\leq e^{|\mathbf{i}^c|\mathcal{F}(\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))\cdot\sum_{k=0}^{|\mathbf{i}^c|}\Pr_{S\sim D^m}\left(m\cdot\overline{R}_S(\overline{h^\mu_{\mathbf{i}}})-a^\mu_{S_{\mathbf{i}}}=k\right)\,e^{-C_1\cdot(m-ld)\cdot\frac{k}{m}}} \\
&= e^{|\mathbf{i}^c|\mathcal{F}(\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))\cdot\sum_{k=0}^{|\mathbf{i}^c|}\binom{|\mathbf{i}^c|}{k}\left(\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right)^k\left(1-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\right)^{|\mathbf{i}^c|-k}\,e^{-C_1\cdot\frac{m-ld}{m}\cdot k}} \\
&= e^{|\mathbf{i}^c|\mathcal{F}(\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))}\left(1-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\,[1-e^{-C_1\cdot\frac{m-ld}{m}}]\right)^{|\mathbf{i}^c|}\,.
\end{aligned} \tag{30}$$

The last equation being obtained from the Newton binomial $\sum_{i=0}^k\binom{m}{k}x^k y^{m-k} = (x+y)^k$.

Let us now define $\mathcal{F}$ such that $1 = e^{|\mathbf{i}^c|\mathcal{F}(\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))}\left(1-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})\,[1-e^{-C_1\cdot\frac{m-ld}{m}}]\right)^{|\mathbf{i}^c|}$. Equivalently, let

$$\mathcal{F}(\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))) \stackrel{\text{def}}{=} -\ln\left(1-\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))\,\left[1-e^{-C_1\frac{m-ld}{m}}\right]\right)\,. \tag{31}$$

With this choice, we have $\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{\overline{h^\mu_{\mathbf{i}}}\sim\overline{P}} e^{|\mathbf{i}^c|\mathcal{D}(\overline{R}_S(\overline{h^\mu_{\mathbf{i}}}),\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))} = 1\,.$

To finish the proof, let us combine Equations (28), (29), (30) and (31), in order to rewrite the innermost inequality of Equation (27) as follows

$$\begin{aligned}
&(m-ld)\cdot\mathcal{F}\left(\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\,\zeta_D^Q\right]\right) - mC_1\cdot\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\,\zeta_S^Q\right] - \frac{\zeta'(1)}{\zeta(1)}\cdot\mathrm{KL}(Q\|P) \leq \ln\frac{1}{\delta} \\
&(m-ld)\left\{-\ln\left(1-\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\,\zeta_D^Q\right]\left[1-e^{-C_1\frac{m-ld}{m}}\right]\right)\right\} \leq mC_1\cdot\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\,\zeta_S^Q\right] + \frac{\zeta'(1)}{\zeta(1)}\cdot\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} \\
&\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\,\zeta_D^Q\right]\left[1-e^{-C_1\frac{m-ld}{m}}\right] \leq 1-\exp\left\{-\left(\frac{1}{m-ld}\right)\left(mC_1\cdot\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\,\zeta_S^Q\right] + \frac{\zeta'(1)}{\zeta(1)}\cdot\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right)\right\} \\
&\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\,\zeta_D^Q\right]\left[1-e^{-C_1\frac{m-ld}{m}}\right] \leq \left(\frac{1}{m-ld}\right)\left(mC_1\cdot\frac{1}{2}\left[1+\frac{1}{\zeta(1)}\,\zeta_S^Q\right] + \frac{\zeta'(1)}{\zeta(1)}\cdot\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right)\,,
\end{aligned}$$

where the last transformation is an application of the inequality $1 - e^{-x} \leq x$. We are now able to isolate ζ_D^Q to obtain

$$\begin{aligned}
\zeta_D^Q &\leq \left(\frac{2 \cdot \zeta(1)}{1 - e^{-C_1 \frac{m-ld}{m}}}\right)\left(\frac{1}{m-ld}\right)\left(mC_1 \cdot \frac{1}{2}\left[1 + \frac{1}{\zeta(1)}\ \zeta_S^Q\right] + \frac{\zeta'(1)}{\zeta(1)} \cdot \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right) - \zeta(1) \\
&= \left(\frac{C_1 \frac{m}{m-ld}}{1 - e^{-C_1 \frac{m-ld}{m}}}\right)\left(\zeta(1) + \zeta_S^Q + \frac{2}{mC_1}[\zeta'(1) \cdot \mathrm{KL}(Q\|P) + \zeta(1) \cdot \ln\frac{1}{\delta}]\right) - \zeta(1) \\
&= \zeta(1)[C' - 1] + C' \cdot \left(\zeta_S^Q + \frac{2}{mC_1}[\zeta'(1) \cdot \mathrm{KL}(Q\|P) + \zeta(1) \cdot \ln\frac{1}{\delta}]\right),
\end{aligned}$$

where

$$C' = \frac{C_1 \cdot \frac{m}{m - l \cdot \deg\zeta}}{1 - e^{-C_1 \cdot \frac{m - l \cdot \deg\zeta}{m}}}.$$

□

3 Details related to the PBSC algorithms

In this section, we present the theoretical development leading the two PBSC learning algorithms and the optimization procedures. Both algorithms build a majority vote of sc-classifier of compression sequence size of at most one as defined in Section 2.2 of the main paper. The two algorithms that we present minimize a bound on the quadratic loss. Given a fixed parameter q, the loss an example having margin $(-\alpha)$ is given by:

$$\zeta(\alpha) \;=\; \left(1+\tfrac{1}{q}\,\alpha\right)^2 .$$

The first algorithm, named *PBSC-A*, works with a strongly aligned posterior Q. The second algorithm, named *PBSC-N*, works with a non-aligned posterior Q.

3.1 PBSC-A: The aligned case

As seen in Section 2.1 of the paper, the strongly aligned posterior Q is totally defined by a vector $\mathbf{w} \stackrel{\text{def}}{=} (w_0, w_1, \ldots, w_m)$. For Q to remain a valid distribution, each component of the vector $\mathbf{w}$ must remain in the interval $\left[-\frac{1}{m+1}, +\frac{1}{m+1}\right]$.

Theorem 5 suggests to minimize the bound on ζ_D^Q given by the following expression:

$$\zeta_D^Q \;\leq\; \zeta_S^Q + \frac{\zeta(1)}{\sqrt{\frac{1}{2}(m - l\deg\zeta)}}\sqrt{4\,l\deg\zeta + \ln\frac{2\sqrt{m}}{\delta}}\;.$$

To do so, we only need to minimize the empirical risk ζ_S^Q because the last term of the right hand side is constant. The empirical risk, ζ_S^Q, is given by

$$\zeta_S^Q \quad = \quad \frac{1}{mq^2}\sum_{j=1}^{m}\left(q - y_j\left[w_0 + \sum_{i=1}^{m} w_i k(x_i, x_j)\right]\right)^2 .$$

Lets define a matrix $\mathbf{G}$ of size $m+1 \times m$ as

$$G_{i,j} = \begin{cases} 1 & \text{for } i = 0, \\ k(x_i, x_j) & \text{for } 1 \leq i, j \leq m\;. \end{cases}$$

With this notation, the optimization problem of *PBSC-A* can be written as

$$\begin{array}{lll} \text{Minimize:} & f_A(\mathbf{w}) & = \displaystyle\sum_{j=1}^{m}\left(q - y_j\sum_{i=0}^{m} w_i G_{i,j}\right)^2 \\ \text{subject to:} & & |w_i| \leq \frac{1}{m+1} \text{ for } i = 0, 1, \ldots, m\,. \end{array}$$

We propose to solve this optimization problem by minimizing f_A coordinate-wise, similarly as it is done for AdaBoost (Schapire et al. (1998)), with the difference that we will have to ensure that Q remains an aligned distribution at each step of the algorithm. Starting from the uniform distribution P (i.e., $\mathbf{w} = \mathbf{0}$), the learning algorithm iteratively chooses (at random) $k \in \{0, \ldots, m\}$, and updates $w_k \leftarrow w_k + \delta$ (without updating the other weights) according to some optimally chosen value of δ. Let $\mathbf{w}_\delta$ be the new weight vector obtained with such an update. After an update, the objective function becomes:

$$f_A(\mathbf{w}_\delta) \;=\; \sum_{j=1}^{m}\left[q - y_j\left(\sum_{i=0}^{m} w_i G_{i,j} + \delta G_{k,j}\right)\right]^2 .$$

The optimal value for δ is obtained when $\frac{df_A(\mathbf{w}_\delta)}{d\delta} = 0$, provided that $w_k + \delta \in [\frac{-1}{m+1}, \frac{1}{m+1}]$. The derivative of f_A with respect to the δ is given by

$$\begin{aligned}
\frac{\partial f_A(\mathbf{w}_\delta)}{\partial \delta} &= \sum_{j=1}^{m} 2\left(q - y_j \sum_{i=0}^{m} w_i G_{i,j} - y_j \delta G_{k,j}\right)(-y_j G_{k,j}) \\
&= 2\sum_{j=1}^{m}\left[\delta G_{k,j}^2 + y_j G_{k,j}\left(y_j \sum_{i=0}^{m} w_i G_{i,j} - q\right)\right] \\
&= 2\left[\delta \sum_{j=1}^{m} G_{k,j}^2 + \sum_{j=1}^{m} G_{k,j}\left(\sum_{i=0}^{m} w_i G_{i,j} - q y_j\right)\right] \\
&= 2\left[\delta \sum_{j=1}^{m} G_{k,j}^2 + \sum_{j=1}^{m} G_{k,j} D_{\mathbf{w}}(j)\right], \qquad (32)
\end{aligned}$$

where $D_{\mathbf{w}}(j) \stackrel{\text{def}}{=} \sum_{i=0}^{m} w_i G_{i,j} - q y_j$.

Equation (32) implies that the optimal value for δ is given by

$$\delta = -\frac{\sum_{j=1}^{m} G_{k,j}\, D_{\mathbf{w}}(j)}{\sum_{j=1}^{m} G_{k,j}^2}\,. \qquad (33)$$

Algorithm 1 presents the complete optimization procedure that we have used.

Algorithm 1 : PBSC-A optimization procedure

1: **Initialize:** $w_i = 0 \;\; \forall i \in \{0, \ldots, m\}$ and $D_{\mathbf{w}}(j) = -q\, y_j \;\; \forall j \in \{1, \ldots, m\}$.
2: **repeat**
3: Choose at random $k \in \{0, .., m\}$.
4: Compute δ given by Equation (33) .
5: If $[w_k + \delta > \frac{1}{m+1}]$ then $\delta \leftarrow \frac{1}{m+1} - w_k$.
6: If $[w_k + \delta < \frac{-1}{m+1}]$ then $\delta \leftarrow \frac{-1}{m+1} - w_k$.
7: $w_k \leftarrow w_k + \delta$.
8: Update $D_{\mathbf{w}}(j) \leftarrow D_{\mathbf{w}}(j) + \delta\, G_{k,j} \;\; \forall j \in \{1, \ldots, m\}$.
9: **until** Convergence

3.2 PBSC-N: The non-aligned case

We consider the non-aligned scenario where the posterior Q is defined by a vector $\mathbf{v} \stackrel{\text{def}}{=} (v_+, v_1, \ldots, v_{2m}, v_-)$:

$$\begin{aligned}
Q(h_{S_{\langle\rangle}}^{(\varepsilon,+)}) = v_+, \quad Q(h_{S_{\langle i\rangle}}^{(\sigma,+)}) = v_i \tfrac{1}{|\mathcal{M}^1|} I(\sigma \in \mathcal{M}^1), \\
Q(h_{S_{\langle\rangle}}^{(\varepsilon,-)}) = v_-, \quad Q(h_{S_{\langle i\rangle}}^{(\sigma,-)}) = v_{m+i} \tfrac{1}{|\mathcal{M}^1|} I(\sigma \in \mathcal{M}^1),
\end{aligned}$$

under the constraints $v \geq 0$ for all $v \in \mathbf{v}$ and $\sum_{v \in \mathbf{v}} v = 1$.

Lets compute the Kullback-Leibler divergence $\mathrm{KL}(Q\|P)$ between this posterior Q and the uniform prior P. We find that

$$\begin{aligned}
\mathrm{KL}(Q\|P) &= \mathop{\mathbf{E}}_{h\sim Q} \ln \tfrac{Q(h)}{P(h)} \\
&= \sum_{i=1}^{m} \sum_{s\in\{-,+\}} \int_{\mathcal{M}^1} d\sigma Q(h^{(\sigma,s)}_{S_{\langle i\rangle}}) \ln \left[\frac{Q(h^{(\sigma,s)}_{S_{\langle i\rangle}})}{P(h^{(\sigma,s)}_{S_{\langle i\rangle}})}\right] + \sum_{s\in\{-,+\}} Q(h^{(\varepsilon,s)}_{S_{\langle\rangle}}) \ln \left[\frac{Q(h^{(\varepsilon,s)}_{S_{\langle\rangle}})}{P(h^{(\varepsilon,s)}_{S_{\langle\rangle}})}\right] \\
&= \sum_{i=1}^{2m} \int_{\mathcal{M}^1} d\sigma \frac{v_i}{|\mathcal{M}^1|} \ln \left[\frac{\frac{v_i}{|\mathcal{M}^1|}}{\frac{1}{2|\mathcal{M}^1|(m+1)}}\right] + v_+ \ln \frac{v_+}{\frac{1}{2(m+1)}} + v_- \ln \frac{v_-}{\frac{1}{2(m+1)}} \\
&= \sum_{v\in\mathbf{v}} v \ln \left[\frac{v}{\frac{1}{2(m+1)}}\right] \\
&= \ln(2m+2) + \sum_{v\in\mathbf{v}} v \ln [v] \ .
\end{aligned}$$

Moreover, we show in the main paper that the empirical risk ζ_S^Q is given by

$$\zeta_S^Q = \frac{1}{mq^2} \sum_{j=1}^{m} \left(q - y_j \left[v_+ - v_- + \sum_{i=1}^{m} (v_i - v_{i+m}) k(x_i, x_j) \right] \right)^2 .$$

Minimizing the bound of Theorem 1 amounts at finding $\mathbf{v}$ that minimizes

$$C \cdot \zeta_S^Q + \mathrm{KL}(Q\|P) \ ,$$

for some constant $C > 0$. However, in Theorem 1, we have $C = \frac{m\cdot C_1}{2\,\zeta'(1)}$.

Let $v_0 \stackrel{\text{def}}{=} v_+$ and $v_{2m+1} \stackrel{\text{def}}{=} v_-$. Let us define a matrix $\mathbf{G}$ of size $2m+2 \times m$ as

$$G_{i,j} = \begin{cases} 1 & \text{if } i = 0, \\ k(\mathbf{x}_i, \mathbf{x}_j) & \text{if } 1 \le i, j \le m, \\ -k(\mathbf{x}_{i-m}, \mathbf{x}_j) & \text{if } m+1 \le i \le 2m \text{ (and } 1 \le j \le m), \\ -1 & \text{if } i = 2m+1 \, . \end{cases}$$

With this notation, the optimization problem for *PBSC-N* can be written as

$$\begin{aligned}
\text{Minimize:} \quad f_N(\mathbf{v}) &= \frac{C}{mq^2} \sum_{j=1}^{m} \left(q - y_j \sum_{i=0}^{2m+1} v_i G_{i,j} \right)^2 + \sum_{i=0}^{2m+1} v_i \ln v_i \\
\text{subject to:} \quad & v_i \ge 0 \quad \text{for } i = 0, 1, \ldots, 2m+1 \, , \\
& \sum_{i=0}^{2m+1} v_i = 1 \, .
\end{aligned}$$

We propose to solve this optimization problem by minimizing f_N with a coordinate-pair descent algorithm that works iteratively by exchanging weights between two components of $\mathbf{v}$. Starting from the uniform distribution P (i.e., $v_i = \frac{1}{2m+2}$ for $i = 0, 1, \ldots, 2m+1$), the learning algorithm iteratively chooses (at random) $k, l \in \{0, \ldots, 2m+1\}$ (with $k \neq l$), and updates $v_k \leftarrow v_k + \delta$ and $v_l \leftarrow v_l - \delta$ (without updating the other weights) according to some optimally chosen value of δ. Let $\mathbf{v}_\delta$ be the new weight vector obtained

with such an update. After an update, the objective function becomes

$$\begin{aligned} f_N(\mathbf{v}_\delta) &= \frac{C}{mq^2}\sum_{j=1}^{m}\left[q - y_j\left(\sum_{i=0}^{2m+1} v_i G_{i,j} + \delta G_{k,j} - \delta G_{l,j}\right)\right]^2 \\ &\quad + \sum_{i=0}^{2m+1} I(i \notin \{k,l\}) \cdot v_i \ln v_i + (v_k+\delta)\ln(v_k+\delta) + (v_l-\delta)\ln(v_l-\delta) \end{aligned}$$

The optimal value for δ is obtained when $\frac{df_N(\mathbf{v}_\delta)}{d\delta} = 0$, provided that $v_k+\delta \in [0, v_k+v_l]$ and $v_l-\delta \in [0, v_k+v_l]$. The derivative of f_N with respect to the δ is given by

$$\begin{aligned} \frac{\partial f_N(\mathbf{v}_\delta)}{\partial \delta} &= \frac{C}{mq^2}\sum_{j=1}^{m} 2\left(q - y_j\sum_{i=0}^{2m+1} v_i G_{i,j} - y_j\delta\left(G_{k,j}-G_{l,j}\right)\right)\left(-y_j\left(G_{k,j}-G_{l,j}\right)\right) + \ln\frac{v_k+\delta}{v_l-\delta} \\ &= \frac{2C}{mq^2}\sum_{j=1}^{m}\left[\delta\left(G_{k,j}-G_{l,j}\right)^2 + y_j\left(G_{k,j}-G_{l,j}\right)\left(y_j\sum_{i=0}^{2m+1} v_i G_{i,j} - q\right)\right] + \ln\frac{v_k+\delta}{v_l-\delta} \\ &= \frac{2C}{mq^2}\left[\delta\sum_{j=1}^{m}\left(G_{k,j}-G_{l,j}\right)^2 + \sum_{j=1}^{m}\left(G_{k,j}-G_{l,j}\right)\left(\sum_{i=0}^{2m+1} v_i G_{i,j} - q y_j\right)\right] + \ln\frac{v_k+\delta}{v_l-\delta} \\ &= \frac{2C}{mq^2}\left[\delta\sum_{j=1}^{m}\left(G_{k,j}-G_{l,j}\right)^2 + \sum_{j=1}^{m}\left(G_{k,j}-G_{l,j}\right) D_{\mathbf{v}}(j)\right] + \ln\frac{v_k+\delta}{v_l-\delta}\,, \end{aligned} \tag{34}$$

where $D_{\mathbf{v}}(j) = \sum_{i=0}^{2m+1} v_i G_{i,j} - q y_j$.

We find the optimal value for δ with the help of a root finding method. Algorithm 2 presents the complete optimization procedure that we have used.

Algorithm 2 : PBSC-N optimization procedure

1: **Initialize:** $v_i = \frac{1}{2m+2} \quad \forall i \in \{0,\ldots,m\}$ and $D_{\mathbf{v}}(j) = -q\,y_j \quad \forall j \in \{1,\ldots,m\}$.
2: **repeat**
3: Choose at random $k, l \in \{0,\ldots,2m+1\}$ (with $k \neq l$).
4: Find δ given by the root of Equation (34).
5: If $\delta > v_l$ then $\delta \leftarrow v_l$.
6: If $\delta < -v_k$ then $\delta \leftarrow -v_k$.
7: $v_k \leftarrow v_k + \delta$ and $v_l \leftarrow v_l - \delta$.
8: Update $D_{\mathbf{v}}(j) \leftarrow D_{\mathbf{v}}(j) + \delta\left(G_{k,j}-G_{l,j}\right) \quad \forall j \in \{1,\ldots,m\}$.
9: **until** Convergence

4 Empirical Results

As mentioned in Section 2.4 of the main paper, this section presents all the empirical results that we have obtained in our experiments.

Table 1 of the main paper only presents the results for a subset of the data. Due to a lack of space, we have removed from Table 1 of the main paper the data sets having the smallest number of examples because they were less likely to show significative differences between the different algorithms. Here is the complete table of the results of our experiments.

Table 1: Empirical risk measured on the testing set T for the five different algorithms.

Dataset			**Rbf kernel**					**Sigmoid kernel**	
Name	$\|T\|$	$\|S\|$	**SVM**	**RLSC**	**PBSC-A**	**PBSC-N**	**LINEAR**	**SVM**	**PBSC-A**
Adult	10000	1809	0.158	0.157	**0.156**	0.160	0.193	0.163	**0.157**
BreastC	340	343	**0.038**	**0.038**	0.044	**0.038**	0.144	**0.038**	**0.038**
Credit-A	300	353	0.190	0.160	**0.140**	0.173	0.200	0.190	**0.170**
Glass	107	107	**0.150**	**0.150**	**0.150**	0.168	0.187	**0.355**	0.411
Haberman	150	144	**0.267**	0.327	0.280	**0.267**	**0.267**	**0.273**	**0.273**
Heart	147	150	**0.197**	0.211	0.204	0.218	0.238	**0.184**	0.197
Ionosphere	175	176	0.057	**0.023**	0.040	0.040	0.326	0.126	**0.091**
Letter:AB	1055	500	**0.001**	0.002	**0.001**	**0.001**	0.038	0.009	**0.005**
Letter:DO	1058	500	0.014	0.015	**0.011**	0.012	0.069	**0.022**	0.028
Letter:OQ	1036	500	0.016	**0.011**	0.016	0.014	0.123	**0.018**	0.039
Liver	175	170	0.286	0.291	**0.280**	0.286	0.349	**0.400**	**0.400**
Mnist:0vs8	1916	500	**0.003**	0.009	0.004	0.004	0.031	0.007	**0.003**
Mnist:1vs7	1922	500	0.014	0.008	0.008	**0.007**	0.161	0.012	**0.007**
Mnist:1vs8	1936	500	0.011	**0.010**	**0.010**	0.011	0.292	**0.014**	0.015
Mnist:2vs3	1905	500	0.020	0.022	**0.019**	0.020	0.114	**0.025**	0.031
Mushroom	4062	4062	**0.000**	**0.000**	**0.000**	**0.000**	0.022	**0.000**	0.010
Ringnorm	3700	3700	0.015	0.017	**0.013**	**0.013**	0.103	**0.020**	0.035
sonar	104	104	0.154	0.250	**0.125**	0.192	0.490	0.250	**0.183**
Tic-tac-toe	479	479	**0.015**	0.019	0.019	0.052	0.365	**0.023**	0.159
Usvotes	200	235	0.075	**0.065**	**0.065**	**0.065**	0.140	0.070	**0.065**
Waveform	4000	4000	0.068	0.067	0.068	**0.066**	0.143	**0.067**	**0.067**
Wdbc	284	285	**0.042**	0.067	0.049	0.074	0.180	**0.366**	**0.366**

Note that Table 2 below is exactly the same as in the main paper.

Table 2: Mean and standard deviation (in parentheses) of the empirical risk across 20 partitions.

Dataset	**Linear SVM**		**k-NN**		**PBSC-A**	
Aural Sonar	**0.1425**	(0.694)	0.1825	(0.597)	0.1500	(0.827)
Voting	0.0534	(0.193)	0.0546	(0.174)	**0.0529**	(0.184)
Yeast-5-7	**0.2688**	(0.622)	0.3063	(0.580)	0.2975	(0.668)
Yeast-5-12	**0.1075**	(0.482)	0.1275	(0.439)	0.1088	(0.598)

5 Another PAC-Bayes bound without $KL(Q\|P)$ (not stated in main the paper)

The following theorem can be viewed as a generalization of PAC-Bayes bound of Seeger (2002) to our setting. As for Theorem 5, the following version is free of any Kullback-Leibler divergence. This bound is tighter than the one in Theorem 5. We have, nevertheless, decided to state Theorem 5 in the main paper because it has a simpler statement and because we have $l \deg \zeta = 2$ for the proposed learning algorithms—a case where the bound of Theorem 5 is already quite tight.

Theorem 2. *For any D, for any family $(\mathcal{H}^S)_{S\in D^m}$ of sets of sc-classifiers of size at most l, for any prior $\mathcal{P}$, for any margin loss function ζ such that $l \cdot \deg(\zeta) < m$, and for for any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall Q \text{ aligned on } P\ :\\ \quad \mathrm{kl}^+\left(\frac{m}{m-l\cdot\deg\zeta}\left[\frac{1}{2}\left(1+\frac{1}{\zeta(1)}\ \zeta_S^Q\right)+\frac{ld}{m}\right]\ \middle\|\ \frac{1}{2}\left[1+\frac{1}{\zeta(1)}\ \zeta_D^Q\right]\right)\ \leq\ \frac{\ln\frac{2\sqrt{m}}{\delta}}{m-l\cdot\deg\zeta}\end{array}\right)\geq 1-\delta$$

where $\mathrm{kl}(q\|p) \stackrel{def}{=} q\ln\frac{q}{p}+(1-q)\ln\frac{1-q}{1-p}$, *and where* $\mathrm{kl}^+(q\|p) = \mathrm{kl}(q\|p)$ *if* $q \geq p$ *and* 0 *otherwise. Moreover, if* $l = 0$, $\mathrm{kl}^+(,)$ *can be replaced by* $\mathrm{kl}(,)$ *in the above statement—thus giving rise to both a lower and an upper bound for* ζ_D^Q.

Proof. The first part of the proof is very similar to the one of Theorem 5. We thus use here the same definitions for d, $\bar{h} = \overline{h_1..h_k}$ (with $k\in\{0,..,d\}$), $\overline{R}_D(\bar{h})$, $\overline{R}_S(\bar{h})$, $\overline{\mathcal{H}^S}$, $\overline{P}$, $\overline{Q}$, ζ_D^Q, and ζ_S^Q. However, we will instead consider the following (and quite different) Laplace transform

$$X_{\overline{P}} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\bar{h}\sim\overline{P}} e^{(m-|\mathbf{i}_{\bar{h}}|)\mathrm{kl}(\widetilde{R}_S(\bar{h}),\overline{R}_D(\bar{h}))}\,, \tag{35}$$

where $\widetilde{R}_S(\bar{h})$ is the abstract empirical risk computed on the examples of S that are not in the compression sequence of $\bar{h}$. More formally,

$$\widetilde{R}_S(\,\overline{h_1..h_k}\,)\ \stackrel{\text{def}}{=}\ \frac{1}{m-|\mathbf{i}_{\overline{h_1..h_k}}|}\sum_{j=1}^{m} I\left(\neg\bigvee_{i=1}^{k}(h_i(x_j)\neq y_j)\right) I\left((x_j,y_j)\notin\mathbf{i}_{\overline{h_1..h_k}}\right).$$

As in the proof of Theorem 5, we can show the following claim.
Claim : for any posterior Q aligned on P, we have

$$X_{\overline{P}}\ =\ \mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} e^{(m-|\mathbf{i}_{\bar{h}}|)\mathrm{kl}(\widetilde{R}_S(\bar{h}),\overline{R}_D(\bar{h}))}. \tag{36}$$

Now again, as in the proof of Theorem 5, by Markov's inequality, we have

$$\Pr_{S\sim D^m}\left(X_{\overline{P}} \leq \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m} X_{\overline{P}}\right)\geq 1-\delta\,.$$

Thus, by applying the claim and by taking the logarithm on each side of the innermost inequality, we obtain

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\, Q \textit{ aligned on } P:\\ \ln\left[\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} e^{(m-|\mathbf{i}_{\bar{h}}|)\mathrm{kl}(\widetilde{R}_S(\bar{h})\|\overline{R}_D(\bar{h}))}\right]\leq\ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{\bar{h}\sim\overline{P}} e^{(m-|\mathbf{i}_{\bar{h}}|)\mathrm{kl}(\widetilde{R}_S(\bar{h})\|\overline{R}_D(\bar{h}))}\right]\end{array}\right)\geq 1-\delta\,. \tag{37}$$

Jensen's inequality applied to the concave $\ln(x)$ gives

$$\ln\left[\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} e^{(m-|\mathbf{i}_{\bar{h}}|)\mathrm{kl}(\widetilde{R}_S(\bar{h})\|\overline{R}_D(\bar{h}))}\right] \geq \mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} (m-|\mathbf{i}_{\bar{h}}|)\,\mathrm{kl}(\widetilde{R}_S(\bar{h})\|\overline{R}_D(\bar{h})\,. \tag{38}$$

Again from the Jensen's inequality, applied to the convex function $\mathrm{kl}(\cdot\,\|\,\cdot)$, together with the definitions of ζ_D^Q and ζ_S^Q (see Equation (26)) and the fact that $m-|\mathbf{i}_{\bar{h}}| \geq m-l\cdot d$, we obtain

$$\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}} (m-|\mathbf{i}_{\bar{h}}|)\,\mathrm{kl}\left(\widetilde{R}_S(\bar{h})\,\|\,\overline{R}_D(\bar{h})\right) \quad\geq\quad (m-ld)\;\;\mathrm{kl}\left(\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\widetilde{R}_S(\bar{h})\,\|\,\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_D(\bar{h})\right) \tag{39}$$

Let us now analyse the value of $\mathbf{E}_{S\sim D^m} X_{\overline{P}}$. Let $\mathbf{i}^c$ be the vector of indices of $\mathcal{I}$ that are not in the vector $\mathbf{i}$, and note that

$$\mathop{\mathbf{E}}_{S\sim D^m}\;\mathop{\mathbf{E}}_{\bar{h}\sim\overline{P}} e^{(m-|\mathbf{i}_{\bar{h}}|)\,\mathrm{kl}(\widetilde{R}_S(\bar{h}),\overline{R}_D(\bar{h}))} \quad=\quad \mathop{\mathbf{E}}_{\mathbf{i}\sim\overline{P}_{\mathcal{I}}}\;\mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^{|\mathbf{i}|}}\;\mathop{\mathbf{E}}_{\mu\sim\overline{P}_{S_{\mathbf{i}}}}\;\mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-|\mathbf{i}|}} e^{|\mathbf{i}^c|\,\mathrm{kl}(\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}}),\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))}\,.$$

Since $\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})$ is an arithmetic mean of iid random variables, one can apply Lemma 0 with $M(X)$ replaced by $\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}})$, n replaced by $m-|\mathbf{i}|$, and ν replaced by $\overline{R}_D(\overline{h^\mu_{\mathbf{i}}})$ to obtain

$$\mathbf{E}_{S_{\mathbf{i}^c}\sim D^{m-|\mathbf{i}|}}\; e^{(m-|\mathbf{i}|)\,\mathrm{kl}(\widetilde{R}_S(\overline{h^\mu_{\mathbf{i}}}),\overline{R}_D(\overline{h^\mu_{\mathbf{i}}}))} \quad\leq\quad 2\sqrt{m-|\mathbf{i}|} \quad\leq\quad 2\sqrt{m}\,. \tag{40}$$

By rearranging Equation (37), and by using Equations (40), (38) and (39), we have

$$\mathop{\mathbf{E}}_{S\sim D^m} (m-ld)\;\;\mathrm{kl}\left(\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\widetilde{R}_S(\bar{h})\,\|\,\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_D(\bar{h})\right) \quad\leq\quad \ln\frac{2\sqrt{m}}{\delta}\,. \tag{41}$$

Finally, observe that for any classifier $\bar{h}\in\overline{\mathcal{H}^S}$, we have

$$\begin{aligned}\widetilde{R}_S(\bar{h}) \quad&\leq\quad \left(\overline{R}_S(\bar{h})+\frac{ld}{m}\right)\;\frac{m}{m-|\mathbf{i}|}\\ &\leq\quad \left(\overline{R}_S(\bar{h})+\frac{ld}{m}\right)\;\frac{m}{m-ld}\,.\end{aligned} \tag{42}$$

Consider the following two cases.

case 1 : $l=0$. In that case we have $\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\widetilde{R}_S(\bar{h}) = \mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_S(\bar{h})$. Hence we have

$$\mathrm{kl}\left(\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\widetilde{R}_S(\bar{h})\,\|\,\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_D(\bar{h})\right) \quad=\quad \mathrm{kl}\left(\frac{m}{m-ld}\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_S(\bar{h})+\frac{ld}{m}\,\|\,\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_D(\bar{h})\right)$$

case 2 : $l>0$. In that case, following Equation (**??**), we can show that

$$\mathrm{kl}\left(\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\widetilde{R}_S(\bar{h})\,\|\,\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_D(\bar{h})\right) \quad\geq\quad \mathrm{kl}^+\left(\frac{m}{m-ld}\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_S(\bar{h})+\frac{ld}{m}\,\|\,\mathop{\mathbf{E}}_{\bar{h}\sim\overline{Q}}\overline{R}_D(\bar{h})\right)$$

In each case, the result then follows from Equation (26), Equation (41) and straightforward calculations. □

References

Germain, Pascal, Lacoste, Alexandre, Laviolette, François, Marchand, Mario, and Shanian, Sara. A PAC-Bayes Sample Compression Approach to Kernel Methods. In *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, June 2011. URL http://www.icml-2011.org/papers.php.

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

Schapire, Robert E., Freund, Yoav, Bartlett, Peter, and Lee, Wee Sun. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26:1651–1686, 1998.

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