A COMPLETE PROOF OF THEOREM 2

In what follows, we assume an arbitrary set $\mathcal{H}$ of classifiers and distributions $P$ and $Q$ on $\mathcal{H}$. When $\mathcal{H}$ is a discrete set, $P(h)$ and $Q(h)$ denote probability masses at $h$. When $\mathcal{H}$ is continuous, $P(h)$ and $Q(h)$ denote the probability densities at $h$ associated to $P$ and $Q$ when they exist.

Let us first recall the change of measure inequality, which is an important step in most PAC-Bayesian proofs.

**Lemma 1** (Change of measure inequality [Seldin and Tishby, 2010, McAllester, 2013]). Let $\mathcal{H}$ be a set of classifiers and let $P$ be a distribution on $\mathcal{H}$. Let $Q$ be a distribution on $\mathcal{H}$ with a support entirely contained within the support of $P$. Then for any function $\phi : \mathcal{H} \to \mathbb{R}$ measurable with respect to $P$, we have

$$\ln \left( \mathbb{E}_{h \sim P} \exp \left[ \phi(h) \right] \right) \leq \mathbb{E}_{h \sim Q} \phi(h) - \text{KL}(Q\|P).$$

**Proof.** This proof is very similar to the proofs of Seldin and Tishby [2010], McAllester [2013], but we provide it for completeness.

Given $\mathcal{H}$, let $\mathcal{H}_P \subseteq \mathcal{H}$ denote the support of $P$ and $\mathcal{H}_Q \subseteq \mathcal{H}_P$ denote the support of $Q$. In the continuous case, for any $h \in \mathcal{H}_Q$, we have that $P(h)/Q(h) = dP(h)/dQ(h)$; which is the Radon-Nykodym derivative. Hence, for any $\psi : \mathcal{H} \to \mathbb{R}$ measurable with respect to $P$ and $Q$, we have

$$\mathbb{E}_{h \sim P} \psi(h) = \int_{\mathcal{H}_P} \psi(h) dP(h) \geq \int_{\mathcal{H}_Q} \psi(h) dP(h) = \int_{\mathcal{H}_Q} \frac{dP(h)}{dQ(h)} \psi(h) dQ(h) = \int_{\mathcal{H}_Q} P(h) \psi(h) dQ(h) \triangleq \mathbb{E}_{h \sim Q} P(h) \psi(h).$$

The same result holds trivially in the discrete case. This gives us the rule of how to transform the expectation over $P$ to an expectation over $Q$. By using Jensen’s inequality and by exploiting the concavity of $\ln(\cdot)$, we then obtain

$$\ln \left( \mathbb{E}_{h \sim P} \exp \left[ \phi(h) \right] \right) \geq \ln \left( \mathbb{E}_{h \sim Q} \exp \left[ \phi(h) \frac{P(h)}{Q(h)} \right] \right)$$

$$\geq \mathbb{E}_{h \sim Q} \left( \exp \left[ \phi(h) \frac{P(h)}{Q(h)} \right] \right)$$

$$= \mathbb{E}_{h \sim Q} \phi(h) - \ln \left( \frac{Q(h)}{P(h)} \right)$$

$$= \mathbb{E}_{h \sim Q} \phi(h) - \text{KL}(Q\|P).$$

We also need the following modified version of this lemma, which takes into account pairs of voters.

**Lemma 2** (Change of measure inequality for pairs of voters [Germain et al., 2015]). For any set $\mathcal{H}$, for any distributions $P$ and $Q$ on $\mathcal{H}$, and for any measurable function $\phi : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, we have

$$\ln \left( \mathbb{E}_{(h,h') \sim P^2} \exp \left[ \phi(h,h') \right] \right) \geq \mathbb{E}_{(h,h') \sim Q^2} \phi(h,h') - 2\text{KL}(Q\|P).$$

**Proof.** This result is an application of Lemma 1, with $P = P^2$, $Q = Q^2$, together with the observation that $\text{KL}(Q^2\|P^2) = 2\text{KL}(Q\|P)$ (see the definition of the KL-divergence, Definition 2).

Now, let us first define the Kullback-Leibler divergence between two Bernoulli distributions, which will be used in the proof of Theorems 3 and 4, below.

**Definition 3.** The Kullback-Leibler divergence between two Bernoulli distributions with probability of success $q$ and probability of success $p$ is given by

$$\text{kl}(q\|p) \triangleq q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}.$$
To prove Theorem 2 that relies on an upper bound on the first moment of the margin and a lower bound on the second moment, we will first prove these two bounds independently. The first provides a lower bound on the first moment of the margin from its empirical estimate, and is very similar to the classical PAC-Bayesian bounds on the risk of the stochastic Gibbs classifier, which can be recovered with a linear transformation of the first moment of the margin: \( R_{D'}(G_Q) = \frac{1}{2} \left( 1 - \mu_1(M_{Q}^{D'}) \right) \).

**Theorem 3.** For any distribution \( D \) on \( \mathcal{X} \times \mathcal{Y} \), for any set \( \mathcal{H} \) of real-valued voters \( h : \mathcal{X} \to [-1, 1] \), for any prior distribution \( P \) on \( \mathcal{H} \), and any \( \delta \in (0, 1] \), we have

\[
\Pr_{S \sim D^m} \left( \forall Q \text{ on } \mathcal{H}, \mu_1(M_Q^S) \geq \mu_1(M_Q^D) - \frac{2}{m} \left[ \text{KL}(Q \| P) + \ln \left( \frac{2\sqrt{m}}{\delta} \right) \right] \right) \geq 1 - \delta.
\]

**Proof.** Given a voter \( h : \mathcal{X} \to [-1, 1] \) and a distribution \( D' \) on \( \mathcal{X} \times \mathcal{Y} \), let \( M_{D'}^P = E_{(x,y) \sim D'} y \cdot h(x). \)

First, note that \( E_{h \sim P} \exp \left[ \frac{m}{2} (M_h^S - M_h^D)^2 \right] \) is a non-negative random variable. By applying Markov’s inequality, with probability at least \( 1 - \delta \) over the choice of \( S \sim D^m \), we have

\[
E_{h \sim P} \exp \left[ \frac{m}{2} (M_h^S - M_h^D)^2 \right] \leq \frac{1}{\delta} E_{S \sim D^m} E_{h \sim P} \exp \left[ \frac{m}{2} (M_h^S - M_h^D)^2 \right]. \tag{7}
\]

Let us now upper-bound the right-hand side of the inequality:

\[
E_{S \sim D^m} E_{h \sim P} \exp \left[ \frac{m}{2} (M_h^S - M_h^D)^2 \right] = E_{h \sim P} E_{S \sim D^m} \exp \left[ \frac{m}{2} (M_h^S - M_h^D)^2 \right] \tag{8}
\]

\[
= E_{h \sim P} \exp \left[ m \cdot \frac{1}{2} (M_h^S)^2 - \frac{1}{2} (M_h^P)^2 \right] \tag{9}
\]

\[
\leq E_{h \sim P} 2\sqrt{m} = 2\sqrt{m}, \tag{10}
\]

where Line (8) comes from the fact that \( P \) is independent of \( S \), Line (9) is an application of Pinsker’s inequality \( 2(q - p)^2 \leq \text{kl}(q \| p) \), and Line (10) is an application of the main result of Maurer [2004], which is valid for arbitrary random variables which lie within \([0, 1] \).

Now, by applying Line 10 in Inequality (7) and by taking the logarithm on each side, with probability at least \( 1 - \delta \) over the choice of \( S \sim D^m \), we have

\[
\ln \left( E_{h \sim P} \exp \left[ \frac{m}{2} (M_h^S - M_h^D)^2 \right] \right) \leq \ln \left( \frac{2\sqrt{m}}{\delta} \right).
\]

By applying the change of measure inequality of Lemma 1 on the left-hand side of the inequality with \( \phi(h) = \frac{m}{2} (M_h^S - M_h^D)^2 \), and by using Jensen’s inequality exploiting the convexity of \( \frac{m}{2} (M_h^S - M_h^D)^2 \), we obtain that for all distributions \( Q \) on \( \mathcal{H} \),

\[
\ln \left( E_{h \sim P} \exp \left[ \frac{m}{2} (M_h^S - M_h^D)^2 \right] \right) \geq E_{h \sim Q} m \frac{1}{2} (M_h^S - M_h^D)^2 - \text{KL}(Q \| P)
\]

\[
\geq m \frac{1}{2} \left( E_{h \sim Q} M_h^S - E_{h \sim Q} M_h^D \right)^2 - \text{KL}(Q \| P)
\]

\[
= m \frac{1}{2} \left( \mu_1(M_Q^S) - \mu_1(M_Q^D) \right)^2 - \text{KL}(Q \| P).
\]

We then have that with probability at least \( 1 - \delta \) over the choice of \( S \sim D^m \), for all \( Q \) on \( \mathcal{H} \),

\[
\frac{m}{2} \left( \mu_1(M_Q^S) - \mu_1(M_Q^D) \right)^2 - \text{KL}(Q \| P) \leq \ln \left( \frac{2\sqrt{m}}{\delta} \right).
\]

The result immediately follows. \( \square \)
Let us now upper-bound the right-hand side of the last inequality:

\[ \Pr_{S \sim D^m} \left( \forall Q \text{ on } \mathcal{H}, \mu_2(M_Q^S) \leq \mu_2(M_Q^D) + \sqrt{\frac{2}{m} \left[ 2 \KL(Q \| P) + \ln \left( \frac{2\sqrt{m}}{\delta} \right) \right]} \right) \geq 1 - \delta. \]

Proof. Given a voter \( h : \mathcal{X} \to [-1, 1] \) and a distribution \( D' \) on \( \mathcal{X} \times \mathcal{Y} \), let \( M_{h,h'}^{D'} \equiv E_{(x,y) \sim D'} h(x) h'(x) \).

First, note that \( E_{(h,h')} \sim P^2 \exp \left[ \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \right] \) is a non-negative random variable. By applying Markov’s inequality, with probability at least \( 1 - \delta \) over the draws of \( S \sim D^m \), we have

\[
\frac{E}{(h,h')} \sim P^2 \ E \exp \left[ \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \right] \leq \frac{1}{\delta} \frac{E}{S \sim D^m (h,h')} \sim P^2 \ E \exp \left[ \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \right].
\]

Let us now upper-bound the right-hand side of the last inequality:

\[
E_{S \sim D^m} \ E_{(h,h') \sim P^2} \exp \left[ \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \right] = E_{(h,h') \sim P^2} \ E_{S \sim D^m} \exp \left[ \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \right]
\]

\[
= \frac{m \cdot 2 \left( 1 - M_{h,h'}^S \right) - 2 \left( 1 - M_{h,h'}^D \right) \right)^2}{S_{h,h'} \sim D^m} \ E_{(h,h') \sim P^2} \ E_{S \sim D^m} \exp \left[ m \cdot \text{kl} \left( \frac{1}{2} \left( 1 - M_{h,h'}^S \right) \right) \left( 1 - M_{h,h'}^D \right) \right]
\]

\[
\leq \frac{m}{S_{h,h'} \sim D^m} \ E_{(h,h') \sim P^2} \ E \\left( 2\sqrt{m} = 2\sqrt{m}, \right)
\]

where Line (12) comes from the fact that distribution \( P \) is independent of \( S \), Line (13) is an application of Pinsker’s inequality \( 2(q - p)^2 \leq \text{kl}(q \| p) \), and Line (14) is an application of the main result of Maurer [2004], which is valid for arbitrary random variables which lie within \([0, 1]\).

Now, by applying Line (14) in Inequality (11) and by taking the logarithm on each side, with probability at least \( 1 - \delta \) over the draws of \( S \sim D^m \), we have

\[
\ln \left( \frac{E}{(h,h') \sim P^2} \exp \left[ \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \right] \right) \leq \ln \left( 2\sqrt{m} \delta \right).
\]

We now apply the change of measure inequality of Lemma 2 on the left-hand side of the inequality, with \( \phi(h, h') = \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \). We then use Jensen’s inequality exploiting the convexity of \( \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \).

We obtain that for all distributions \( Q \) on \( \mathcal{H} \),

\[
\ln \left( \frac{E}{(h,h') \sim P^2} \exp \left[ \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 \right] \right) \geq \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 - 2 \KL(Q \| P)
\]

\[
\geq \frac{m}{2} \left( M_{h,h'}^S - M_{h,h'}^D \right)^2 - 2 \KL(Q \| P)
\]

\[
= \frac{m}{2} (\mu_2(M_Q^S) - \mu_2(M_Q^D))^2 \leq \ln \left( \frac{2\sqrt{m}}{\delta} \right).
\]

The result then immediately follows. □
B DETAILED CALCULATIONS OF THE LAGRANGIAN DUALITY

Partial derivative for getting from Lagrangian (4) to first optimality constraint (5). The result is obtained by making the last line equal to 0 and by isolating \(-\xi + \nu 1\).

\[
\frac{\partial}{\partial q^*} \Lambda(q^*, \gamma^*, \alpha, \beta, \xi, \nu) = \frac{\partial}{\partial q^*} \left[ \frac{1}{m} \gamma^T \gamma^* + \alpha^T (\gamma^* - \text{diag}(y)Hq^*) + \beta \left( \frac{1}{m} 1^T \gamma^* - \mu \right) - \xi^T q^* + \nu 1^T q^* - 1 \right]
\]

Straightforward calculations details for substituting Equation (5) in Lagrangian (4).

\[
\Lambda(q^*, \gamma^*, \alpha, \beta, \xi, \nu) = \frac{1}{m} \gamma^T \gamma^* + \alpha^T (\gamma^* - \text{diag}(y)Hq^*) + \beta \left( \frac{1}{m} 1^T \gamma^* - \mu \right) - \xi^T q^* + \nu 1^T q^* - 1
\]

First substitution using Eq. (5)

\[
= \frac{1}{m} \gamma^T \gamma^* + \alpha^T \gamma^* - \text{diag}(y)Hq^* + \beta \frac{1}{m} 1^T \gamma^* - \beta \mu - \xi^T q^* + \nu 1^T q^* - \nu
\]

Simplification

\[
\alpha^T \gamma^* + \beta \frac{1}{m} 1^T \gamma^* - \beta \mu - \nu
\]

Second substitution using Eq. (5)

\[
\alpha^T \gamma^* + \beta \frac{1}{m} 1^T \gamma^* - \beta \mu - \nu
\]

Partial derivative for getting from Lagrangian (4) to second optimality constraint (5). The result is obtained by making the last line equal to 0 and by isolating \(\gamma^*\).

\[
\frac{\partial}{\partial \gamma^*} \Lambda(q^*, \gamma^*, \alpha, \beta, \xi, \nu) = \frac{\partial}{\partial \gamma^*} \left[ \frac{1}{m} \gamma^T \gamma^* + \alpha^T (\gamma^* - \text{diag}(y)Hq^*) + \beta \left( \frac{1}{m} 1^T \gamma^* - \mu \right) - \xi^T q^* + \nu 1^T q^* - 1 \right]
\]

\[
= \frac{1}{m} \gamma^T \gamma^* + \alpha^T \gamma^* - \text{diag}(y)Hq^* + \beta \frac{1}{m} 1^T \gamma^* - \beta \mu - \xi^T q^* + \nu 1^T q^* - \nu
\]

\[
= \frac{1}{m} \gamma^T \gamma^* + \alpha^T \gamma^* + \beta \frac{1}{m} 1^T \gamma^*
\]

\[
= \frac{2}{m} \gamma^* + \alpha + \beta \frac{1}{m} 1
\]
C RESULTS USING RBF KERNELS AS VOTERS

Table 2 below shows the results of the experiments considering RBF kernels as base voters. In this setting, for each training example \((x, y)\), we consider the voters \(h(\cdot) = \pm K(x, \cdot)\), where \(K(x, x') = \exp(-||x - x'||^2/2\sigma^2)\), where \(\sigma\) is the width parameter of the kernel and is set to the mean squared distance between pairs of training examples.

Again, the hyperparameter value of each algorithm has been selected by 5-folds cross-validation on the training set, among 15 values on a logarithmic scale. The value of hyperparameter \(\mu\) of CqBoost and MinCq is selected among values between \(10^{-5}\) and \(10^{-2}\). The value of hyperparameter \(D\) of MDBoost is chosen between \(10^2\) and \(10^5\). The value of hyperparameter \(C\) of LPBoost and CG-Boost is selected among values between \(10^{-3}\) and \(10^3\). The number of iterations of AdaBoost is selected among values between \(10^3\) and \(10^7\). The value of hyperparameter \(C\) of SVM has been chosen between \(10^{-4}\) and \(10^4\). The stopping criterion additive constant \(\epsilon\) of all column generation algorithms has been set to \(10^{-8}\).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CqBoost</th>
<th>MDBoost</th>
<th>LPBoost</th>
<th>CG-Boost</th>
<th>AdaBoost</th>
<th>MinCq</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>australian</td>
<td>0.142</td>
<td>31*</td>
<td>0.151</td>
<td>62</td>
<td>0.145</td>
<td>71</td>
<td>0.136</td>
</tr>
<tr>
<td>balance</td>
<td>0.054</td>
<td>25</td>
<td>0.038</td>
<td>89</td>
<td>0.029*</td>
<td>23*</td>
<td>0.032</td>
</tr>
<tr>
<td>breast</td>
<td>0.040</td>
<td>35</td>
<td>0.040</td>
<td>33</td>
<td>0.040</td>
<td>4*</td>
<td>0.040</td>
</tr>
<tr>
<td>bupa</td>
<td>0.272*</td>
<td>30</td>
<td>0.277</td>
<td>23*</td>
<td>0.265</td>
<td>39</td>
<td>0.283</td>
</tr>
<tr>
<td>car</td>
<td>0.094</td>
<td>32*</td>
<td>0.054</td>
<td>169</td>
<td>0.034*</td>
<td>87</td>
<td>0.197</td>
</tr>
<tr>
<td>cmc</td>
<td>0.317</td>
<td>28*</td>
<td>0.312</td>
<td>39</td>
<td>0.323</td>
<td>30</td>
<td>0.322</td>
</tr>
<tr>
<td>credit</td>
<td>0.133</td>
<td>21*</td>
<td>0.130*</td>
<td>137</td>
<td>0.139</td>
<td>73</td>
<td>0.133</td>
</tr>
<tr>
<td>cylinder</td>
<td>0.307</td>
<td>36</td>
<td>0.296</td>
<td>144</td>
<td>0.359</td>
<td>17*</td>
<td>0.363</td>
</tr>
<tr>
<td>ecoli</td>
<td>0.060*</td>
<td>25</td>
<td>0.065</td>
<td>48</td>
<td>0.113</td>
<td>12*</td>
<td>0.113</td>
</tr>
<tr>
<td>glass</td>
<td>0.187</td>
<td>38</td>
<td>0.187</td>
<td>43</td>
<td>0.159*</td>
<td>29*</td>
<td>0.290</td>
</tr>
<tr>
<td>heart</td>
<td>0.156</td>
<td>17</td>
<td>0.148*</td>
<td>27</td>
<td>0.148*</td>
<td>14</td>
<td>0.170</td>
</tr>
<tr>
<td>hepatitis</td>
<td>0.156*</td>
<td>12*</td>
<td>0.182</td>
<td>65</td>
<td>0.182</td>
<td>18</td>
<td>0.195</td>
</tr>
<tr>
<td>horse</td>
<td>0.158</td>
<td>31*</td>
<td>0.163</td>
<td>32</td>
<td>0.136*</td>
<td>33</td>
<td>0.196</td>
</tr>
<tr>
<td>ionosphere</td>
<td>0.131</td>
<td>31*</td>
<td>0.154</td>
<td>71</td>
<td>0.097*</td>
<td>45</td>
<td>0.120</td>
</tr>
<tr>
<td>letter:ab</td>
<td>0.016</td>
<td>26</td>
<td>0.008*</td>
<td>104</td>
<td>0.012</td>
<td>22</td>
<td>0.016</td>
</tr>
<tr>
<td>monks</td>
<td>0.245</td>
<td>18*</td>
<td>0.245</td>
<td>61</td>
<td>0.245</td>
<td>50</td>
<td>0.329</td>
</tr>
<tr>
<td>optdigits</td>
<td>0.090</td>
<td>25*</td>
<td>0.066*</td>
<td>147</td>
<td>0.088</td>
<td>77</td>
<td>0.098</td>
</tr>
<tr>
<td>pima</td>
<td>0.263</td>
<td>32</td>
<td>0.258</td>
<td>36</td>
<td>0.247*</td>
<td>15*</td>
<td>0.250</td>
</tr>
<tr>
<td>titanic</td>
<td>0.220*</td>
<td>13*</td>
<td>0.220*</td>
<td>15</td>
<td>0.227</td>
<td>49</td>
<td>0.222</td>
</tr>
<tr>
<td>vote</td>
<td>0.051*</td>
<td>33*</td>
<td>0.055</td>
<td>110</td>
<td>0.055</td>
<td>37</td>
<td>0.055</td>
</tr>
<tr>
<td>wine</td>
<td>0.034</td>
<td>27</td>
<td>0.034</td>
<td>29</td>
<td>0.045</td>
<td>16*</td>
<td>0.045</td>
</tr>
<tr>
<td>yeast</td>
<td>0.279</td>
<td>33*</td>
<td>0.277*</td>
<td>65</td>
<td>0.288</td>
<td>88</td>
<td>0.278</td>
</tr>
<tr>
<td>zoo</td>
<td>0.059</td>
<td>24</td>
<td>0.059</td>
<td>27</td>
<td>0.000*</td>
<td>18</td>
<td>0.098</td>
</tr>
</tbody>
</table>

Table 2: Performance and sparsity comparison of CqBoost, MDBoost, LPBoost, CG-Boost, AdaBoost, MinCq and SVM, using RBF kernel functions as weak classifiers. A bold value indicates that the risk (or number of chosen columns) is the lowest among the column generation algorithms. A star indicates that the risk is the lowest among all seven algorithms.

In this setting, we observe that CqBoost, MDBoost and LPBoost show a very similar performance. We also notice that MDBoost slightly outperforms CqBoost with 10 wins and 7 losses, but with a sign test \(p\)-value of only 0.31, which is not statistically significant.

In terms of sparsity, we observe that CqBoost still reaches its goal of outputting significantly sparser solutions than MinCq, while keeping a similar performance. Using RBF kernels as voters, as opposed to the results using decision stumps, CqBoost produces slightly sparser solutions than LPBoost, even if the latter has a \(L_1\)-norm regularization term on the weight vector that directly penalizes dense solutions.