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*

$$\ln\left(\mathop{\mathbf{E}}_{h\sim P} \exp\left[\phi(h)\right]\right) \;\geq\; \mathop{\mathbf{E}}_{h\sim Q} \phi(h) - \mathrm{KL}(Q\|P)\,.$$

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

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

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

$$\begin{aligned}
\ln\left(\mathop{\mathbf{E}}_{h\sim P} \exp\left[\phi(h)\right]\right) \;&\geq\; \ln\left(\mathop{\mathbf{E}}_{h\sim Q} \exp\left[\phi(h)\right] \frac{P(h)}{Q(h)}\right)\\
&\geq\; \mathop{\mathbf{E}}_{h\sim Q} \ln\left(\exp\left[\phi(h)\right] \frac{P(h)}{Q(h)}\right)\\
&=\; \mathop{\mathbf{E}}_{h\sim Q} \left[\phi(h) - \ln\left(\frac{Q(h)}{P(h)}\right)\right]\\
&=\; \mathop{\mathbf{E}}_{h\sim Q} \phi(h) - \mathrm{KL}(Q\|P)\,.
\end{aligned}$$

□

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*

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

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

$$\mathrm{kl}(q\|p) \;\triangleq\; q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}\,.$$

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*

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

Proof. Given a voter $h : \mathcal{X} \to [-1, 1]$ and a distribution D' on $\mathcal{X} \times \mathcal{Y}$, let $M_h^{D'} \triangleq \mathbf{E}_{(x,y)\sim D'}\, y \cdot h(x)$.

First, note that $\mathbf{E}_{h\sim P} \exp\left[\frac{m}{2}\left(M_h^S - M_h^D\right)^2\right]$ 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

$$\mathop{\mathbf{E}}_{h\sim P} \exp\left[\frac{m}{2}\left(M_h^S - M_h^D\right)^2\right] \;\leq\; \frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} \exp\left[\frac{m}{2}\left(M_h^S - M_h^D\right)^2\right]. \tag{7}$$

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

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} \exp\left[\frac{m}{2}\left(M_h^S - M_h^D\right)^2\right] &= \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \exp\left[\frac{m}{2}\left(M_h^S - M_h^D\right)^2\right] && (8)\\
&= \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \exp\left[m \cdot 2\left(\tfrac{1}{2}\left(1-M_h^S\right) - \tfrac{1}{2}\left(1-M_h^D\right)\right)^2\right] \\
&\leq \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S\sim D^m} \exp\left[m \cdot \mathrm{kl}\left(\tfrac{1}{2}\left(1-M_h^S\right) \middle\| \tfrac{1}{2}\left(1-M_h^D\right)\right)\right] && (9)\\
&\leq \mathop{\mathbf{E}}_{h\sim P} 2\sqrt{m} \;=\; 2\sqrt{m}\,, && (10)
\end{aligned}$$

where Line (8) comes from the fact that P is independent of S, Line (9) is an application of Pinsker's inequality $2(q - p)^2 \;\leq\; \mathrm{kl}(q\|p)$, 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

$$\ln\left(\mathop{\mathbf{E}}_{h\sim P} \exp\left[\frac{m}{2}\left(M_h^S - M_h^D\right)^2\right]\right) \;\leq\; \ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

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

$$\begin{aligned}
\ln\left(\mathop{\mathbf{E}}_{h\sim P} \exp\left[\frac{m}{2}\left(M_h^S - M_h^D\right)^2\right]\right) &\geq \mathop{\mathbf{E}}_{h\sim Q} \frac{m}{2}\left(M_h^S - M_h^D\right)^2 - \mathrm{KL}(Q\|P)\\
&\geq \frac{m}{2}\left(\mathop{\mathbf{E}}_{h\sim Q} M_h^S - \mathop{\mathbf{E}}_{h\sim Q} M_h^D\right)^2 - \mathrm{KL}(Q\|P)\\
&= \frac{m}{2}\left(\mu_1(M_Q^S) - \mu_1(M_Q^D)\right)^2 - \mathrm{KL}(Q\|P)
\end{aligned}$$

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

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

The result immediately follows. $\square$

The second result provides an upper bound on the second moment of the margin from its empirical estimate. It requires techniques provided in Lacasse et al. [2006], Laviolette et al. [2011], Germain et al. [2011] which are less common in the PAC-Bayesian literature as they make use of random variables considering pairs of voters.

Theorem 4. *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*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall Q \text{ on } \mathcal{H},\\ \mu_2(M_Q^D)\ \le\ \mu_2(M_Q^S)+\sqrt{\dfrac{2}{m}\left[2\mathrm{KL}(Q\|P)+\ln\left(\dfrac{2\sqrt{m}}{\delta}\right)\right]}\end{array}\right)\ \ge\ 1-\delta\,.$$

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

First, note that $\mathbf{E}_{(h,h')\sim P^2}\exp\left[\frac{m}{2}\left(M_{h,h'}^S-M_{h,h'}^D\right)^2\right]$ 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

$$\mathop{\mathbf{E}}_{(h,h')\sim P^2}\exp\left[\frac{m}{2}\left(M_{h,h'}^S-M_{h,h'}^D\right)^2\right]\ \le\ \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{(h,h')\sim P^2}\exp\left[\frac{m}{2}\left(M_{h,h'}^S-M_{h,h'}^D\right)^2\right].\tag{11}$$

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

$$\begin{aligned}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{(h,h')\sim P^2}\exp\left[\frac{m}{2}\left(M_{h,h'}^S-M_{h,h'}^D\right)^2\right] &= \mathop{\mathbf{E}}_{(h,h')\sim P^2}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[\frac{m}{2}\left(M_{h,h'}^S-M_{h,h'}^D\right)^2\right] &(12)\\ &= \mathop{\mathbf{E}}_{(h,h')\sim P^2}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[m\cdot 2\left(\tfrac{1}{2}\left(1-M_{h,h'}^S\right)-\tfrac{1}{2}\left(1-M_{h,h'}^D\right)\right)^2\right]\\ &\le \mathop{\mathbf{E}}_{(h,h')\sim P^2}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[m\cdot\mathrm{kl}\left(\tfrac{1}{2}\left(1-M_{h,h'}^S\right)\middle\|\tfrac{1}{2}\left(1-M_{h,h'}^D\right)\right)\right] &(13)\\ &\le \mathop{\mathbf{E}}_{(h,h')\sim P^2}2\sqrt{m}\ =\ 2\sqrt{m}\,, &(14)\end{aligned}$$

where Line (12) comes from the fact that distribution P is independent of S, Line (13) is an application of Pinsker's inequality $2(q-p)^2\le\mathrm{kl}(q\|p)$, 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

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

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 $\mathcal{H}$,

$$\begin{aligned}\ln\left(\mathop{\mathbf{E}}_{(h,h')\sim P^2}\exp\left[\frac{m}{2}\left(M_{h,h}^S-M_{h'h'}^D\right)^2\right]\right) &\ge \mathop{\mathbf{E}}_{(h,h')\sim Q^2}\frac{m}{2}\left(M_{h,h'}^S-M_{h,h'}^D\right)^2-2\,\mathrm{KL}(Q\|P)\\ &\ge \frac{m}{2}\left(\mathop{\mathbf{E}}_{(h,h')\sim Q^2}M_{h,h'}^S-\mathop{\mathbf{E}}_{(h,h')\sim Q^2}M_{h,h'}^D\right)^2-2\,\mathrm{KL}(Q\|P)\\ &= \frac{m}{2}\left(\mu_2(M_Q^S)-\mu_2(M_Q^D)\right)^2-2\,\mathrm{KL}(Q\|P)\,.\end{aligned}$$

We then have that with probability at least $1-\delta$ over the draws of $S\sim D^m$,

$$\forall Q \text{ on } \mathcal{H},\qquad \frac{m}{2}\left(\mu_2(M_Q^S)-\mu_2(M_Q^D)\right)^2-2\,\mathrm{KL}(Q\|P)\ \le\ \ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

The result then immediately follows. $\square$

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 $\mathbf{0}$ and by isolating $-\boldsymbol{\xi}+\nu\mathbf{1}$.

$$
\begin{aligned}
&\frac{\partial}{\partial \mathbf{q}^\star}\Lambda(\mathbf{q}^\star,\boldsymbol{\gamma}^\star,\boldsymbol{\alpha},\beta,\boldsymbol{\xi},\nu)\\
&\quad= \frac{\partial}{\partial \mathbf{q}^\star}\left[\frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\left(\boldsymbol{\gamma}^\star-\mathrm{diag}(\mathbf{y})\mathbf{H}\,\mathbf{q}^\star\right)+\beta\left(\frac{1}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star-\mu\right)-\boldsymbol{\xi}^\top\mathbf{q}^\star+\nu\left(\mathbf{1}^\top\mathbf{q}^\star-1\right)\right]\\
&\quad= \frac{\partial}{\partial \mathbf{q}^\star}\left[\boldsymbol{\alpha}^\top\left(\boldsymbol{\gamma}^\star-\mathrm{diag}(\mathbf{y})\mathbf{H}\,\mathbf{q}^\star\right)-\boldsymbol{\xi}^\top\mathbf{q}^\star+\nu\mathbf{1}^\top\mathbf{q}^\star-\nu\right]\\
&\quad= \frac{\partial}{\partial \mathbf{q}^\star}\left[\boldsymbol{\alpha}^\top\boldsymbol{\gamma}^\star-\frac{1}{m}\boldsymbol{\alpha}^\top\mathrm{diag}(\mathbf{y})\mathbf{H}\,\mathbf{q}^\star-\boldsymbol{\xi}^\top\mathbf{q}^\star+\nu\mathbf{1}^\top\mathbf{q}^\star\right]\\
&\quad= \frac{\partial}{\partial \mathbf{q}^\star}\left[-\boldsymbol{\alpha}^\top\mathrm{diag}(\mathbf{y})\mathbf{H}\,\mathbf{q}^\star-\boldsymbol{\xi}^\top\mathbf{q}^\star+\nu\mathbf{1}^\top\mathbf{q}^\star\right]\\
&\quad= -\mathbf{H}^\top\mathrm{diag}(\mathbf{y})\,\boldsymbol{\alpha}-\boldsymbol{\xi}+\nu\mathbf{1}
\end{aligned}
$$

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

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\gamma}^\star}\Lambda(\mathbf{q}^\star,\boldsymbol{\gamma}^\star,\boldsymbol{\alpha},\beta,\boldsymbol{\xi},\nu)\\
&\quad= \frac{\partial}{\partial \boldsymbol{\gamma}^\star}\left[\frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\left(\boldsymbol{\gamma}^\star-\mathrm{diag}(\mathbf{y})\mathbf{H}\,\mathbf{q}^\star\right)+\beta\left(\frac{1}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star-\mu\right)-\boldsymbol{\xi}^\top\mathbf{q}^\star+\nu\left(\mathbf{1}^\top\mathbf{q}^\star-1\right)\right]\\
&\quad= \frac{\partial}{\partial \boldsymbol{\gamma}^\star}\left[\frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\boldsymbol{\gamma}^\star-\boldsymbol{\alpha}^\top\mathrm{diag}(\mathbf{y})\mathbf{H}\,\mathbf{q}^\star+\frac{\beta}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star-\beta\mu-\boldsymbol{\xi}^\top\mathbf{q}^\star+\nu\mathbf{1}^\top\mathbf{q}^\star-\nu\right]\\
&\quad= \frac{\partial}{\partial \boldsymbol{\gamma}^\star}\left[\frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\boldsymbol{\gamma}^\star+\frac{\beta}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star\right]\\
&\quad= \frac{2}{m}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}+\frac{\beta}{m}\mathbf{1}
\end{aligned}
$$

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

$$
\begin{aligned}
&\Lambda(\mathbf{q}^\star,\boldsymbol{\gamma}^\star,\boldsymbol{\alpha},\beta,\boldsymbol{\xi},\nu)\\
&\quad= \frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\left(\boldsymbol{\gamma}^\star-\mathrm{diag}(\mathbf{y})\mathbf{H}\,\mathbf{q}^\star\right)+\beta\left(\frac{1}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star-\mu\right)-\boldsymbol{\xi}^\top\mathbf{q}^\star+\nu\left(\mathbf{1}^\top\mathbf{q}^\star-1\right)\\
&\quad= \frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\boldsymbol{\gamma}^\star-\boldsymbol{\alpha}^\top\mathrm{diag}(\mathbf{y})\mathbf{H}\,\mathbf{q}^\star+\frac{\beta}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star-\beta\mu-\boldsymbol{\xi}^\top\mathbf{q}^\star+\nu\mathbf{1}^\top\mathbf{q}^\star-\nu\\
&\quad= \frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\boldsymbol{\gamma}^\star-\left(\mathbf{H}^\top\mathrm{diag}(\mathbf{y})\,\boldsymbol{\alpha}\right)^\top\mathbf{q}^\star+\frac{\beta}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star-\beta\mu-(\boldsymbol{\xi}+\nu\mathbf{1})^\top\mathbf{q}^\star-\nu\\
&\quad= \frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\boldsymbol{\gamma}^\star+(\boldsymbol{\xi}+\nu\mathbf{1})^\top\mathbf{q}^\star+\frac{\beta}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star-\beta\mu-(\boldsymbol{\xi}+\nu\mathbf{1})^\top\mathbf{q}^\star-\nu && \langle\text{First substitution using Eq. (5)}\rangle\\
&\quad= \frac{1}{m}\boldsymbol{\gamma}^{\star\top}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}^\top\boldsymbol{\gamma}^\star+\frac{\beta}{m}\mathbf{1}^\top\boldsymbol{\gamma}^\star-\beta\mu-\nu && \langle\text{Simplification}\rangle\\
&\quad= \left(\frac{1}{m}\boldsymbol{\gamma}^\star+\boldsymbol{\alpha}+\frac{\beta}{m}\mathbf{1}\right)^\top\boldsymbol{\gamma}^\star-\beta\mu-\nu\\
&\quad= \left(\frac{1}{m}\left(-\frac{m}{2}\boldsymbol{\alpha}-\frac{\beta}{2}\mathbf{1}\right)+\boldsymbol{\alpha}+\frac{\beta}{m}\mathbf{1}\right)^\top\left(-\frac{m}{2}\boldsymbol{\alpha}-\frac{\beta}{2}\mathbf{1}\right)-\beta\mu-\nu && \langle\text{Second substitution using Eq. (5)}\rangle\\
&\quad= \left(-\frac{1}{2}\boldsymbol{\alpha}-\frac{\beta}{2m}\mathbf{1}+\boldsymbol{\alpha}+\frac{\beta}{m}\mathbf{1}\right)^\top\left(-\frac{m}{2}\boldsymbol{\alpha}-\frac{\beta}{2}\mathbf{1}\right)-\beta\mu-\nu\\
&\quad= \left(\frac{1}{2}\boldsymbol{\alpha}+\frac{\beta}{2m}\mathbf{1}\right)^\top\left(-\frac{m}{2}\boldsymbol{\alpha}-\frac{\beta}{2}\mathbf{1}\right)-\beta\mu-\nu\\
&\quad= -\frac{m}{4}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}-\frac{\beta}{4}\boldsymbol{\alpha}^\top\mathbf{1}-\frac{\beta}{4}\mathbf{1}^\top\boldsymbol{\alpha}-\frac{\beta^2}{4m}\mathbf{1}^\top\mathbf{1}-\beta\mu-\nu\\
&\quad= -\frac{m}{4}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}-\frac{\beta}{2}\mathbf{1}^\top\boldsymbol{\alpha}-\frac{\beta^2}{4}-\beta\mu-\nu
\end{aligned}
$$

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 $K(x, x') \triangleq \exp\left(-\|x - x'\|^2/2\sigma^2\right)$, where σ 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 μ 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 10^6. 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 ϵ of all column generation algorithms has been set to 10^{-8}.

	CqBoost		MDBoost		LPBoost		CG-Boost		AdaBoost		MinCq		SVM	
Dataset	Risk	Cols.	Risk	Cols.	Risk	Cols.	Risk	Cols.	Risk	Cols.	Risk	Cols.	Risk	Cols.
australian	0.142	**31***	0.151	62	0.145	71	**0.136**	345	0.157	46	0.128*	690	0.133	218
balance	0.054	25	0.038	89	**0.029***	**23***	0.032	313	0.032	23*	0.058	624	0.035	37
breast	**0.040**	35	**0.040**	33	**0.040**	**4***	**0.040**	350	0.040	10	0.037*	700	0.040	51
bupa	**0.272***	30	0.277	**23***	0.295	39	0.283	174	0.283	37	0.295	344	0.272*	110
car	0.094	**32***	0.054	169	**0.034***	87	0.197	504	0.268	74	0.302	1000	0.034*	97
cmc	0.317	**28***	**0.312**	39	0.323	30	0.322	501	0.312	50	0.316	1000	0.306*	323
credit	0.133	**21***	**0.130***	137	0.139	73	0.133	345	0.145	62	0.133	690	0.130*	118
cylinder	0.307	36	**0.296**	144	0.359	**17***	0.363	270	0.300	41	0.315	540	0.267*	152
ecoli	**0.060***	25	0.065	48	0.113	**12***	0.113	169	0.095	39	0.095	336	0.101	42
glass	0.187	38	0.187	43	**0.159***	**29***	0.290	110	0.234	37	0.243	214	0.187	64
heart	0.156	17	**0.148***	27	**0.148***	**14**	0.170	135	0.148*	12*	0.156	270	0.156	87
hepatitis	**0.156***	**12***	0.182	65	0.182	18	0.195	78	0.182	14	0.208	156	0.182	33
horse	0.158	**31***	0.163	32	**0.136***	33	0.196	184	0.179	34	0.185	368	0.201	85
ionosphere	0.131	**31***	0.154	71	**0.097***	45	0.120	176	0.126	37	0.120	352	0.097*	43
letter:ab	0.016	26	**0.008***	104	0.012	**22**	0.016	500	0.018	16*	0.019	1000	0.014	67
monks	**0.245**	**18***	**0.245**	61	**0.245**	50	0.329	216	0.287	47	0.347	432	0.208*	96
optdigits	0.090	**25***	**0.066***	147	0.088	77	0.098	500	0.087	58	0.142	1000	0.096	77
pima	0.263	32	0.258	36	**0.247***	**15***	0.250	384	0.253	17	0.263	768	0.260	254
titanic	**0.220***	**13***	**0.220***	15	0.227	49	0.222	500	0.220*	16	0.220*	1000	0.227	234
vote	**0.051***	**33***	0.055	110	0.055	37	0.055	218	0.055	41	0.060	436	0.051*	54
wine	**0.034**	27	**0.034**	29	0.045	**16***	0.045	89	0.045	19	0.022*	178	0.056	30
yeast	0.279	**33***	**0.277***	65	0.288	88	0.278	502	0.282	80	0.299	1000	0.278	337
zoo	0.059	24	0.059	27	**0.000***	**18**	0.098	50	0.000*	23	0.039	100	0.137	12*

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.