From PAC-Bayes Bounds to Quadratic Programs for Majority Votes

François Laviolette FRANCOIS.LAVIOLETTE@IFT.ULAVAL.CA
Mario Marchand MARIO.MARCHAND@IFT.ULAVAL.CA
Jean-Francis Roy JEAN-FRANCIS.ROY.1@ULAVAL.CA
Département d'informatique et de génie logiciel, Université Laval, Québec, Canada, G1V 0A6

Abstract

We propose to construct a weighted majority vote on a set of basis functions by minimizing a risk bound (called the C-bound) that depends on the first two moments of the margin of the Q-convex combination realized on the data. This bound minimization algorithm turns out to be a quadratic program that can be efficiently solved. A first version of the algorithm is designed for the supervised inductive setting and turns out to be very competitive with AdaBoost, MDBoost and the SVM. The second version is designed for the transductive setting. It competes well against TSVM. We also propose a new PAC-Bayes theorem that bounds the difference between the "true" value of the C-bound and its empirical estimate and that, unexpectedly, contains no KL-divergence.

1. Introduction

In this paper, we propose a new algorithm, that we call MinCq, for constructing a weighted majority vote of basis functions. One version of this algorithm is designed for the supervised inductive framework and minimizes a risk bound for majority votes, known as the C-bound (Lacasse et al., 2007). A second version of MinCq minimizes the C-bound in the transductive setting. Both versions can be expressed as quadratic programs on positive semi-definite matrices.

As it is the case for boosting algorithms, (Schapire & Singer, 1999), MinCq is designed to output a Q-weighted majority vote of functions (that we call voters) which perform rather poorly individually and, consequently, are often called weak learners. Hence, the decision of each vote is based on a small majority. Moreover, minimizing the C-bound favors votes whose voters are maximally uncorrelated.

Unfortunately, minimizing the empirical value of the C-bound tends to overfit the data. To overcome this problem, MinCq uses a distribution Q of voters which is constrained to be *quasi-uniform* (i.e., close to the uniform distribution in a very specific way) and for which the first moment of the margin of the Q-convex combination realized on the training data is fixed to some precise value $\mu > 0$. This new learning strategy is justified by a new PAC-Bayes bound (Theorem 2) dedicated to quasi-uniform posteriors that, unexpectedly, contains no KL-divergence between the uniform prior and the quasi-uniform posterior. MinCq is also justified by two important properties of majority votes. First (Proposition 3), there is no generality loss for restricting ourselves to quasi-uniform distributions. Second (Proposition 4), for any margin threshold $\mu > 0$, and for any quasi-uniform distribution Q whose margin is at least μ, there is another quasi-uniform distribution Q' whose margin is exactly μ and that achieves the same majority vote and C-bound value.

We will see that to minimize the C-bound, the learner must reduce substantially the variance of the margin distribution. Many learning algorithms actually exploit this strategy in different ways. Indeed, the variance of the margin distribution is controlled by Breiman (2001) for producing random forests, by Dredze et al. (2010) in the transfer learning setting, and by Shen & Li (2010) in the boosting setting. Thus, the idea of minimizing the variance of the margin is well-known. In this paper, we propose a new theoretical justification for all these types of algorithms and propose two novel learning algorithms, called MinCq and TMinCq, that directly minimize the C-bound.

Finally, our experiments show that MinCq is very competitive with Adaboost, MDBoost and the SVM in the supervised inductive setting, and competes with TSVM in the transductive setting.

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

2. Basic Definitions and notations

We consider binary classification problems where the input space $\mathcal{X}$ consists of an arbitrary subset of $\mathbb{R}^n$ and the output space $\mathcal{Y}=\{-1,+1\}$. An example $\mathbf{z} \stackrel{\text{def}}{=} (\mathbf{x}, y)$ is an input-output pair where $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Throughout this paper, we adopt the PAC setting where each example $\mathbf{z}$ is drawn iid according to a fixed, but unknown, probability distribution D on $\mathcal{X} \times \mathcal{Y}$. We denote by $D_{\mathcal{X}}$ the $\mathcal{X}$-marginal distribution of D. The training set is denoted by $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$.

We only consider learning algorithms that construct majority votes based on a (finite) hypothesis space $\mathcal{H} = \{h_1, \ldots, h_{2n}\}$ of real value functions. These functions can be classifiers such as decision stumps or can be given by a kernel k evaluated on the examples of S such as $h_i(\cdot) = k(\mathbf{x}_i, \cdot)$. Given any $\mathbf{x} \in \mathcal{X}$, the output $B_Q(\mathbf{x})$ of a Q-weighted majority vote classifier B_Q (also called *Bayes classifier*) is given by

$$B_Q(\mathbf{x}) \;=\; \text{sgn}\left[\mathop{\mathbf{E}}_{h\sim Q} h(\mathbf{x})\right], \qquad (1)$$

where $\text{sgn}(a) = 1$ if $a > 0$ and -1 otherwise. Hence, even if the voters of $\mathcal{H}$ are not classifiers, B_Q is always a classifier. The *risk* $R_{D'}(B_Q)$ *of any Bayes classifier is defined as the probability that it misclassifies an example drawn according to a* $\mathcal{X}\times\mathcal{Y}$*-distribution* D' :

$$R_{D'}(B_Q) \;\stackrel{\text{def}}{=}\; \Pr_{(\mathbf{x},y)\sim D'}\Big(B_Q(\mathbf{x}) \neq y\Big)$$

Hence, we retrieve the usual notion of *risk* if $D' = D$, and the usual notion of *empirical risk* when $D' = \text{U}_S$, the uniform distribution on the set S. Throughout the paper, D' will generically represent either the true (and unknown) distribution D, or its empirical counterpart U_S. Moreover, for notational simplicity, we will often replace U_S by S. In this paper, we also assume that $\mathcal{H}$ is *auto-complemented*, meaning that for any $\mathbf{x} \in \mathcal{X}$ and any $i \in \{1, \ldots, n\}$,

$$h_{i+n}(\mathbf{x}) \;=\; -h_i(\mathbf{x})\,.$$

Moreover, on any auto-complemented $\mathcal{H}$, we only consider *quasi-uniform* (q-u) distributions, i.e., distributions Q such that for any $i \in \{1, \ldots, n\}$,

$$Q(h_i) + Q(h_{i+n}) \;=\; 1/n\,.$$

We will see that quasi-uniform distributions constitute a rich and interesting family in our context. Another important notion, related to a majority votes, is the Q-*margin*[1] realized on an example $(\mathbf{x}, y)$:

$$\mathcal{M}_Q(\mathbf{x}, y) \;\stackrel{\text{def}}{=}\; y \cdot \mathbf{E}_{h\sim Q}\, h(\mathbf{x})\,.$$

We also consider the *first moment* $\mathcal{M}_Q^{D'}$ and the *second moment* $\mathcal{M}_{Q^2}^{D'}$ of the Q-margin as a random variable defined on the probability space generated by D':

$$\begin{aligned}
\mathcal{M}_Q^{D'} &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} \mathcal{M}_Q(\mathbf{x}, y) \\
&= \mathop{\mathbf{E}}_{h\sim Q} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} y\, h(\mathbf{x}) \\
\mathcal{M}_{Q^2}^{D'} &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} \left(\mathcal{M}_Q(\mathbf{x}, y)\right)^2 \\
&= \mathop{\mathbf{E}}_{(h,h')\sim Q^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} h(\mathbf{x})\, h'(\mathbf{x})\,.
\end{aligned}$$

Note that, since $y^2 = 1$, there is no label y present in the last equation. Moreover, for any $i \in \{1, \ldots, 2n\}$, we also make use of the following notation

$$\begin{aligned}
\mathcal{M}_h^{D'} &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} y\, h(\mathbf{x})\,; \\
\mathcal{M}_{(h,h')}^{D'} &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} h(\mathbf{x})\, h'(\mathbf{x})\,. \qquad (2)
\end{aligned}$$

We then have

$$\mathcal{M}_Q^{D'} \;=\; \mathop{\mathbf{E}}_{h\sim Q} \mathcal{M}_h^{D'} \quad ; \quad \mathcal{M}_{Q^2}^{D'} \;=\; \mathop{\mathbf{E}}_{(h,h')\sim Q^2} \mathcal{M}_{(h,h')}^{D'}\,.$$

3. The C-bound, an upper bound of the risk of Majority vote classifier

It is well known that minimizing $R_S(B_Q)$ is NP-hard. To recover tractability we often replace $R_S(B_Q)$ by some convex function of Q that upper bounds $R_S(B_Q)$. As examples, in boosting, $R_S(B_Q)$ is replaced by the so called empirical exponential loss $\mathbf{E}_{(\mathbf{x},y)\sim S} \frac{1}{2} \exp[-\beta\, y \mathbf{E}_{h\sim Q}\, h(\mathbf{x})]$ for some $\beta \geq 1$. In the PAC-Bayes approach (McAllester, 2003) , it is replaced by the empirical Gibbs's risk $R_S(G_Q)$ that can be defined in terms of the Q-margin as $R_S(G_Q) \stackrel{\text{def}}{=} \frac{1}{2} - \frac{1}{2}\mathbf{E}_{(\mathbf{x},y)\sim S}\, \mathbf{E}_{h\sim Q}\, yh(\mathbf{x})$ (Lacasse et al., 2007). It is well known that both functions are upper bounds of $R_S(B_Q)/2$.

In both cases, however, the bound can be very loose. Indeed, let us consider the case where the majority vote is obtained by boosting very weak learners or even the case of the Support Vector Machine[2]. In both cases, each voter is very likely to err on about

[1] In (Schapire & Singer, 1999), it is called the *margin of the Q-convex combination* realized on $(\mathbf{x}, y)$.

[2] Note that the output of an SVM is of the form: $f_{\text{SVM}}(\mathbf{x}) = \text{sgn}\Big(\sum_{i=1}^m y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b\Big)$ where $\alpha_i \geq 0$ and $b \in \mathbb{R}$. Thus, it can be viewed as a majority vote whose

half of the examples, implying that for each example $(\mathbf{x}, y)$, the Q-margin $\mathcal{M}_Q(\mathbf{x}, y)$ will be close to 0. Thus, both the empirical exponential loss and the empirical Gibbs's risk will tend to be close to 1/2. In these circumstances, we therefore obtain a non informative bound on $R(B_Q)$ saying that it is basically less than 1, even if for AdaBoost and the SVM we observe that $R(B_Q)$ is in general much closer to 0. This is due to the fact that voting can dramatically improve performance when the community of voters tends to compensate the individual errors. The following bound, that we refer to as the C-bound, is in this sense more interesting than the two others we have just discussed, and can also be stated in terms of the Q-margin.

Theorem 1. (The C-bound) *For any distribution Q over a class $\mathcal{H}$ of functions and any distribution D' over $\mathcal{X}\times\mathcal{Y}$, if $\mathcal{M}_Q^{D'} > 0$ then $R_{D'}(B_Q) \leq C_Q^{D'}$ where,*

$$C_Q^{D'} \stackrel{def}{=} \frac{\mathbf{Var}_{(\mathbf{x},y)\sim D'}\left(\mathcal{M}_Q(\mathbf{x},y)\right)}{\mathbf{E}_{(\mathbf{x},y)\sim D'}\left(\mathcal{M}_Q(\mathbf{x},y)\right)^2} = 1 - \frac{\left(\mathcal{M}_Q^{D'}\right)^2}{\mathcal{M}_{Q^2}^{D'}}\,.$$

Proof. It follows from Equations (1) that B_Q classifies correctly an example if its Q-margin is strictly positive. Hence, we have $R_{D'}(B_Q) \leq \Pr_{(\mathbf{x},y)\sim D'}\left(\mathcal{M}_Q(\mathbf{x},y) \leq 0\right)$. The result follows from the Cantelli-Chebychev's inequality (Devroye et al., 1996):

$$\Pr\left(X \leq \mathbf{E}\,X - a\right) \leq \frac{\mathbf{Var}\ X}{\mathbf{Var}\ X + a^2} \quad \text{for any} \quad a \geq 0\,,$$

replacing X by the random variable $\mathcal{M}_Q(\mathbf{x}, y)$ and a by $\mathcal{M}_Q^{D'}$. Recall that, according to our definitions, we have $\mathbf{Var}_{(\mathbf{x},y)\sim D'}\left(\mathcal{M}_Q(\mathbf{x},y)\right) = \mathcal{M}_{Q^2}^{D'} - \left(\mathcal{M}_Q^{D'}\right)^2$. □

The C-bound has first been proposed by Lacasse et al. (2007) for the restricted case where the voters are all classifiers (i.e., having outputs in $\{-1, +1\}$). In their paper, they showed that the bound can be arbitrary close to 0 even when $\mathcal{M}_Q^{D'}$ is close to 0, as long as there is a sufficiently large population of classifiers for which their errors are "sufficiently uncorrelated". Hence, the C-bound seems to take into consideration situations when the "community of voters tend to compensate the individual errors". Moreover, empirical experiments in Lacasse et al. (2007) indicate that C_Q^D is a good predictor of $R_D(B_Q)$. This, therefore, provides a motivation for an algorithm whose objective is to minimize the C-bound.

voters are the functions $h_+(\mathbf{x}) = 1$, $h_-(\mathbf{x}) = -1$ and the functions $h_i(\mathbf{x}) = y_i k(\mathbf{x}_i, \mathbf{x})$, $i = 1, .., m$. Indeed, if $b \geq 0$, the Q-weights are respectively $\frac{b}{Z}$, 0 and $\frac{\alpha_i}{Z}$ with $i = 1, ..m$, and where $Z = b + \sum_{i=1}^m \alpha_i$. Similarly, if $b < 0$, the Q-weights are 0, $\frac{-b}{Z}$ and $\frac{\alpha_i}{Z}$.

4. From the C-bound to the MinCq learning algorithm

Our first attempts to minimize the C-bound has confronted us to two problems.

Problem 1: an empirical C-bound minimization without any regularization tends to overfit the data.

Problem 2: most of the time, the distributions Q minimizing the C-bound C_Q^S are such that both $\mathcal{M}_Q^S$ and $\mathcal{M}_{Q^2}^S$ are very close to 0. Since $C_Q^S = 1 - (\mathcal{M}_Q^S)^2/\mathcal{M}_{Q^2}^S$, this gives a 0/0 numerical instability. Since $(\mathcal{M}_Q^D)^2/\mathcal{M}_{Q^2}^D$ can only be empirically estimated by $(\mathcal{M}_Q^S)^2/\mathcal{M}_{Q^2}^S$, Problem 2 amplifies Problem 1.

PAC-Bayes theorems that bound the difference between C_Q^S and C_Q^D are proposed in Lacasse et al. (2007). This opens the way to structural C-bound minimization algorithms. As for all PAC-Bayes results, the bound on C_Q^D depends on an empirical estimate of it and on the Kullback-Leibler divergence $KL(Q\|P)$ between the output distribution Q and an *a priori* defined distribution P. Our attempts to construct an algorithm regularized by such a divergence was unsuccessful. Surprisingly, the KL-divergence is a poor regularizer in this case. However, restricting ourselves to quasi-uniform distributions Q had a much better regularization effect in practice. This is also supported by the following PAC-Bayes bound on C_Q^D that contains no KL term. From the following theorem, an upper (resp. lower) bound on C_Q^D can be obtained by taking the lower (resp. upper) bound on $\mathcal{M}_Q^D$ together with the upper (resp. lower) bound on $\mathcal{M}_{Q^2}^D$. Theorem 2 is restricted to B-bounded voters (i.e., voters h such that $|h(\mathbf{x})| \leq B \ \ \forall \mathbf{x} \in \mathcal{X}$).

Theorem 2. *For any distribution D, for any $m \geq 8$, for any auto-complemented family $\mathcal{H}$ of B-bounded real value functions, and for any $\delta \in (0, 1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all } q\textit{-u distribution } Q \textit{ on } \mathcal{H}:\\ \quad |\mathcal{M}_Q^D - \mathcal{M}_Q^S| \leq \frac{2B\sqrt{\ln\frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}}\end{array}\right) \geq 1-\delta\,,$$

and

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all } q\textit{-u distribution } Q \textit{ on } \mathcal{H}:\\ \quad |\mathcal{M}_{Q^2}^D - \mathcal{M}_{Q^2}^S| \leq \frac{2B^2\sqrt{\ln\frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}}\end{array}\right) \geq 1-\delta\,.$$

Proof. Because of a lack of space, the proof of the second bound has been omitted. See Laviolette et al. (2011), for the complete proof.

Let $\mathcal{H}$ be a (possibly infinite) auto-complemented set of B-bounded functions. In the general setting, we say that $\mathcal{H}$ is *auto-complemented* if there exists a bijection $c : \mathcal{H} \to \mathcal{H}$ such that $c(h) = -h$ for any $h \in \mathcal{H}$.

Moreover, a distribution on $\mathcal{H}$ will be said *quasi-uniform* if for any $h \in \mathcal{H}$, we have $Q(h) + Q(c(h)) = P(h) + P(c(h))$, where P is the uniform distribution on $\mathcal{H}$. Note that this implies $\mathcal{M}^{D'}_{c(h)} = -\mathcal{M}^{D'}_{h}$.

Let us now consider the following Laplace transform

$$X_P \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim P} e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2} .$$

Note that the function $\mathcal{D}(q,p) \stackrel{\text{def}}{=} \frac{1}{2B^2}(q-p)^2$ used in the Laplace transform is convex, because its Hessian matrix $\nabla^2\mathcal{D}$ is positive semi-definite. Moreover, $(\mathcal{M}^S_{c(h)} - \mathcal{M}^D_{c(h)})^2 = (-\mathcal{M}^S_h - (-\mathcal{M}^D_h))^2 = (\mathcal{M}^S_h - \mathcal{M}^D_h)^2$. Hence, for any quasi-uniform distribution Q, we have

$$\begin{aligned}
2\cdot \mathop{\mathbf{E}}_{h\sim P} & e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2} \\
&= \int_{h\in\mathcal{H}} dh\ P(h)\, e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2} \\
&\qquad + \int_{h\in\mathcal{H}} dh\ P(c(h))\, e^{\frac{m}{2B^2}(\mathcal{M}^S_{c(h)} - \mathcal{M}^D_{c(h)})^2} \\
&= \int_{h\in\mathcal{H}} dh\ (P(h) + P(c(h)))\, e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2} \\
&= \int_{h\in\mathcal{H}} dh\ (Q(h) + Q(c(h)))\, e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2} \\
&\ \vdots \\
&= 2\cdot \mathop{\mathbf{E}}_{h\sim Q} e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2}
\end{aligned}$$

This, in turn implies[3],

$$X_P = \mathop{\mathbf{E}}_{h\sim Q} e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2} .$$

Now, by Markov's inequality we have

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

By taking the logarithm on each side of the innermost inequality, we have

$$\Pr_{S\sim D^m}\left(\ln\left[\mathop{\mathbf{E}}_{h\sim Q} e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2}\right] \le \ln\left[\tfrac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m} X_P\right] \right) \ge 1-\delta .$$

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

$$\ln\left[\mathop{\mathbf{E}}_{h\sim Q} e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2}\right] \ge \mathop{\mathbf{E}}_{h\sim Q} \frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2 .$$

Again from the Jensen's inequality, applied to the convex function $m\cdot\mathcal{D}(q,p) = \frac{m}{2B^2}(q-p)^2$, we then obtain:

$$\mathop{\mathbf{E}}_{h\sim Q} \frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2 \ge \frac{m}{2B^2}(\mathcal{M}^S_Q - \mathcal{M}^D_Q)^2 .$$

[3]Note that it is because of this equality that there is no $\mathrm{KL}(Q \,\|\, P)$ term in those PAC-Bayes bounds.

Thus, from what precedes, we have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\, q\text{-}u \text{ distribution } Q \text{ on } \mathcal{H}: \\ \dfrac{m}{2B^2}\left(\mathcal{M}^S_Q - \mathcal{M}^D_Q\right)^2 \le \ln\left[\dfrac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m} X_P\right]\end{array}\right) \ge 1-\delta. \quad (3)$$

Now, let us bound the value of $\mathop{\mathbf{E}}_{S\sim D^m} X_P$:

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} X_P &= \mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S\sim D^m} e^{\frac{m}{2B^2}(\mathcal{M}^S_h - \mathcal{M}^D_h)^2} && (4)\\
&= \mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S\sim D^m} e^{m\, 2\left(\left(\frac{1}{2} - \frac{\mathcal{M}^S_h}{2B}\right) - \left(\frac{1}{2} - \frac{\mathcal{M}^D_h}{2B}\right)\right)^2} \\
&\le \mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S\sim D^m} e^{m\,\mathrm{kl}\left(\frac{1}{2} - \frac{\mathcal{M}^S_h}{2B} \,\middle\|\, \frac{1}{2} - \frac{\mathcal{M}^D_h}{2B}\right)} && (5)\\
&\le \mathop{\mathbf{E}}_{S\sim P} 2\sqrt{m} = 2\sqrt{m} && (6)
\end{aligned}$$

Line (4) follows from the fact that P has been chosen before seeing the data S. Thus, one can exchange the order of the two expectations.

Line (5) follows from the inequality $2(q-p)^2 \le \mathrm{kl}(q \,\|\, p)$ that is valid for any $p, q \in [0,1]$ provided that if $p=0$ then so is q and if $p=1$ then so is q. Indeed, we have $0 \le \frac{1}{2} - \frac{\mathcal{M}^S_h}{2B} \le 1$ and $0 \le \frac{1}{2} - \frac{\mathcal{M}^D_h}{2B} \le 1$. Moreover, since the elements of $\mathcal{H}$ are B-bounded and S is drawn iid from D, it follows from the definition of the margin that $\mathcal{M}^D_h = -B \ \Rightarrow\ \mathcal{M}^S_h = -B$ and $\mathcal{M}^D_h = B \ \Rightarrow\ \mathcal{M}^S_h = B$. We therefore have $\frac{1}{2} - \frac{\mathcal{M}^D_h}{2B} = 0 \ \Rightarrow\ \frac{1}{2} - \frac{\mathcal{M}^S_h}{2B} = 0$ and $\frac{1}{2} - \frac{\mathcal{M}^D_h}{2B} = 1 \ \Rightarrow\ \frac{1}{2} - \frac{\mathcal{M}^S_h}{2B} = 1$, as wanted.

For Line (6), first observe that $\frac{1}{2} - \frac{\mathcal{M}^S_h}{2B}$ is an arithmetic mean of m iid random variables. Thus Line (6) is obtained applying Maurer's Lemma (Maurer, 2004), with $M(X)$ replaced by $\frac{1}{2} - \frac{\mathcal{M}^S_h}{2B}$, n replaced by m, and ν replaced by $\frac{1}{2} - \frac{\mathcal{M}^D_h}{2B}$.

The first bound of the theorem then follows from Equations (3) and (6). □

Note that Theorem 2 is not valid in the sample compression case, that is when $\mathcal{H}$ consists of functions whose definition depend on the training data such as $\mathcal{H} = \{\pm k(|\mathbf{x}_i, \cdot) \mid (\mathbf{x}_i, y_i) \in S\}$ for some kernel k. However, it can be extended to this framework using the techniques proposed in Laviolette & Marchand (2007). Note also that, even if in this setting $\mathcal{H}$ is assumed to be finite, the theorem is also true if $\mathcal{H}$ is infinite.

There has already been some attempts to develop PAC-Bayes bounds that do not rely on the KL-divergence (see the localized Priors of Catoni (2007) for example). The usual idea is to bound the KL-divergence via some concentration inequality. In the

bound of Theorem 2, the KL-term simply vanishes from the bound, provided that we restrict ourselves to quasi-uniform posteriors. To our knowledge, this is new in PAC-Bayes theory. The fact that Theorem 2 contains no KL divergence between the prior P and the posterior Q indicates that the restriction to quasi-uniform distributions has some "built in" regularization action. Indeed, for such a Q, we have $0 \leq Q(h) \leq 1/n$ for all $h \in \mathcal{H}$—which is a ℓ_∞ norm regularization. However, the next proposition shows that this restriction on Q does not reduce the set of possible majority votes, and hence, the possible outcomes of an algorithm that minimizes C_Q^S.

Proposition 3. *For all distributions Q on $\mathcal{H}$, there exists a quasi-uniform distribution Q' on $\mathcal{H}$ that gives the same majority vote as Q, and that has the same empirical and true C-bound values, i.e.,*

$$B_{Q'} = B_Q \quad , \quad C_{Q'}^S = C_Q^S \quad \text{and} \quad C_{Q'}^D = C_Q^D \, .$$

Proof. Let Q be a distribution on $\mathcal{H}$, let $M \stackrel{\text{def}}{=} \max_{i \in \{1,..,n\}} |Q(h_{i+n}) - Q(h_i)|$, and let Q' be defined as $Q'(h_i) \stackrel{\text{def}}{=} \frac{1}{2n} + \frac{Q(h_i) - Q(h_{i+n})}{2nM}$ where the indices of h are defined modulo $2n$ (i.e., $h_{(i+n)+n} = h_i$). Then it is easy to show that Q' is a quasi-uniform distribution. Moreover, for any example $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{h \sim Q'} h(\mathbf{x}) &\stackrel{\text{def}}{=} \sum_{i=1}^{2n} Q'(h_i)\, h_i(\mathbf{x}) \\
&= \sum_{i=1}^{n} (Q'(h_i) - Q'(h_{i+n}))\, h_i(\mathbf{x}) \\
&= \sum_{i=1}^{n} \frac{2Q(h_i) - 2Q(h_{i+n})}{2nM}\, h_i(\mathbf{x}) \\
&= \frac{1}{nM} \sum_{i=1}^{2n} Q(h_i)\, h_i(\mathbf{x}) \;=\; \frac{1}{nM} \mathop{\mathbf{E}}_{h \sim Q'} h(\mathbf{x}) \, .
\end{aligned}$$

This implies that $B_{Q'}(\mathbf{x}) = B_Q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. It also shows that $\mathcal{M}_{Q'}(\mathbf{x}, y) = \frac{1}{nM} \mathcal{M}_Q(\mathbf{x}, y)$, which implies that $\left(\mathcal{M}_{Q'}^{D'}\right)^2 = \left(\frac{1}{nM} \mathcal{M}_Q^{D'}\right)^2$ and $\mathcal{M}_{(Q')^2}^{D'} = \left(\frac{1}{nM}\right)^2 \mathcal{M}_{Q^2}^{D'}$ for both $D' = D$ and $D' = S$. The result then follows from the definition of the C-bound. □

Proposition 3 points out a nice property of the C-bound: different distributions Q that give rise to a same majority vote have the same (real and empirical) C-bound values. Since the C-bound is a bound on majority votes, this is a suitable property. Moreover, Theorem 2 and Proposition 3 indicate that restricting ourselves to quasi-uniform distributions is a natural solution to the problem of overfitting (see Problem 1). Unfortunately, Problem 2 remains present in a strong way since a consequence of the next proposition is that, among all the distributions that minimize (or ϵ-minimize) the C-bound, there is always one whose empirical margin $\mathcal{M}_Q^S$ is as close to 0 as we want.

Proposition 4. *For all $\mu \in]0, 1]$ and for all quasi-uniform distribution Q on $\mathcal{H}$ having an empirical margin $\mathcal{M}_Q^S \geq \mu$, there exists a quasi-uniform distribution Q' on $\mathcal{H}$, having an empirical margin equal to μ, such that Q and Q' induce same majority vote and have the same empirical and true C-bound values, i.e.,*

$$\mathcal{M}_{Q'}^S = \mu \, , \quad B_{Q'} = B_Q \, , \quad C_{Q'}^S = C_Q^S \quad \text{and} \quad C_{Q'}^D = C_Q^D \, .$$

Proof. Let Q be a quasi-uniform distribution on $\mathcal{H}$ such that $\mathcal{M}_Q^S \geq \mu$ and define Q' as

$$Q'(h_i) \stackrel{\text{def}}{=} \frac{\mu}{\mathcal{M}_Q^S} \cdot Q(h_i) + (1 - \frac{\mu}{\mathcal{M}_Q^S}) \cdot 1/2n \, , \quad i \in \{1, .., 2n\} \, .$$

Clearly, Q' is a quasi-uniform distribution since it is a convex combination of a quasi-uniform distribution and the uniform one. Then, similarly as in the proof of Proposition 3, one can easily show that $\mathbf{E}_{h \sim Q'} h(\mathbf{x}) = \frac{\mu}{\mathcal{M}_Q^S} \mathbf{E}_{h \sim Q} h(\mathbf{x})$, which implies the result. □

One way to overcome the instability identified in Problem 2 is to restrict ourselves to quasi-uniform distributions whose empirical margins are greater or equal than some threshold μ. By Proposition 4, this is equivalent at restricting ourselves to distributions having empirical margin *exactly equal to* μ. From Theorem 1 and Proposition 4, it then follows that *minimizing the C-bound, under the constraint $\mathcal{M}_Q^S \geq \mu$, is equivalent at minimizing $\mathcal{M}_{Q^2}^S$, under the constraint $\mathcal{M}_Q^S = \mu$*, which is the simple quadratic program described by Program 1, below.

Training set bounds (as VC-bounds for example) are known to degrade when the capacity of classification increases. As shown by Proposition 4 for the majority vote setting, this capacity increases as μ decreases to 0. Thus, we expect that any training set bound degrades for small μ. This is clearly not the case for the C-bound itself, but the C-bound is not a training set bound. To obtain a training set bound, we have to relate the empirical value C_Q^S to the true one C_Q^D. This is done via the PAC-Bayes bounds of Theorem 2. In the resulting bound, there is indeed a degradation as μ decreases because the true C-bound is of the form $1 - (\mathcal{M}_Q^D)^2 / \mathcal{M}_{Q^2}^D$. Since $\mu = \mathcal{M}_Q^S$, and because a small $\mathcal{M}_Q^S$ tends to produce small $\mathcal{M}_{Q^2}^S$, the bound on C_Q^D given C_Q^S that we obtain from Theorem 2 is much looser for small μ because of the 0/0 instability.

In what follows, μ represents such a restriction on the margin. Moreover, we say that a value μ is D'-*realizable* if there exists some quasi-uniform distribution Q such that $\mathcal{M}_Q^{D'} = \mu$. The proposed algorithm, called MinCq, is then defined as follows.

Definition 5. – the MinCq algorithm. *Given a set $\mathcal{H}$ of voters, a training set S, and a S-realizable $\mu > 0$, among all quasi-uniform distributions Q of empirical margin $\mathcal{M}_Q^S$ exactly equal to μ, the MinCq algorithm consists in finding one that minimizes $\mathcal{M}_{Q^2}^S$.*

Because of the quasi-uniformity assumption, we only need to consider the first n values of Q, and only $\mathcal{M}_{h_i}^S$ and $\mathcal{M}_{(h_i,h_j)}^S$ for i and j in $\{1,\ldots,n\}$ (defined in Equation (2) for $D' = S$). Consequently, let $\mathbf{Q} \stackrel{\text{def}}{=} (Q(h_1),\ldots,Q(h_n))^T$, let $\mathbf{M}_S$ be the $n \times n$ matrix formed by $\mathcal{M}_{(h_i,h_j)}^S$ for i and $j \in \{1,\ldots,n\}$. Also let

$$\mathbf{m}_S \stackrel{\text{def}}{=} \left(\mathcal{M}_{h_1}^S,\ldots,\mathcal{M}_{h_n}^S\right)^T, \quad \text{and}$$
$$\mathbf{A}_S \stackrel{\text{def}}{=} \left(\frac{1}{n}\sum_{j=1}^{n}\mathcal{M}_{(h_1,h_j)}^S,\ldots,\frac{1}{n}\sum_{j=1}^{h_n}\mathcal{M}_{(h_n,h_j)}^S\right)^T.$$

From these definitions, it follows from tedious straightforward calculations (see Laviolette et al. (2011), for the details) that

$$\frac{\mathcal{M}_Q^S}{2} = \mathbf{m}_S^T\mathbf{Q} - \frac{1}{2n}\sum_{i=1}^{n}\mathcal{M}_{h_i}^S, \quad \text{and}$$
$$\mathcal{M}_{Q^2}^S = 4\left[\mathbf{Q}^T\mathbf{M}_S\mathbf{Q} - \mathbf{A}_S^T\mathbf{Q}\right] + \frac{1}{n^2}\sum_{i,j=1}^{n}\mathcal{M}_{(h_i,h_j)}^S.$$

Hence, given any S-realizable μ, up to the multiplicative constant 4 and the additive constant $\frac{1}{n^2}\sum_{i,j=1}^{n}\mathcal{M}_{(h_i,h_j)}^S$, the MinCq algorithm solves the optimization problem described by Program 1.

Program 1 : **MinCq**
a quadratic program for classification

1: **Solve** $\quad \text{argmin}_{\mathbf{Q}} \quad \mathbf{Q}^T\,\mathbf{M}_S\,\mathbf{Q} \;-\; \mathbf{A}_S^T\,\mathbf{Q}$
2: **under constraints:** $\mathbf{m}_S^T\,\mathbf{Q} = \frac{\mu}{2} + \frac{1}{2n}\sum_{i=1}^{n}\mathcal{M}_{h_i}^S$
3: **and:** $0 \le Q_i \le \frac{1}{n} \quad \forall i \in \{1,\ldots,n\}$

To prove that Program 1 is a quadratic program, it suffices to show that $\mathbf{M}_S$ is a positive semi-definite matrix. This is a direct consequence of the fact that each $\mathcal{M}_{(h_i,h_j)}^S$ can be viewed as a scalar product since $\mathcal{M}_{(h_i,h_j)}^S = \left\langle \left(\sqrt{\frac{1}{|S|}}\,h_i(\mathbf{x})\right)_{\mathbf{x}\in S_\mathcal{X}}, \left(\sqrt{\frac{1}{|S|}}\,h_j(\mathbf{x})\right)_{\mathbf{x}\in S_\mathcal{X}}\right\rangle$, where $S_\mathcal{X} \stackrel{\text{def}}{=} \{\mathbf{x}\colon (\mathbf{x},y) \in S\}$.

5. TMinCq, a transductive extension

An interesting property of MinCq is that the labels only appear in the constraints—they are not involved in the function to be optimized. This opens the way to many natural extensions to the transductive setting. In this section, we explore one such extension.

Given access to a set $S = \{(\mathbf{x}_1,y_1)\ldots(\mathbf{x}_{|S|},y_{|S|})\}$ of labeled examples and a set $U = \{\mathbf{x}_{|S|+1}\ldots\mathbf{x}_{|S|+|U|}\}$ of unlabeled examples, the task of the transductive learner is to label as accurately as possible the data from U. To give us insight on what we should optimize on S and U, we developped a PAC-Bayes bound that relates the C-bound on the training set S to the C-bound on $V^l \stackrel{\text{def}}{=} \{(\mathbf{x}_1,y_1),\ldots(\mathbf{x}_{|S|+|U|},y_{|S|+|U|})\}$, where $y_{|S|+1}$,.., $y_{|S|+|U|}$ are the correct labels associated to $\mathbf{x}_{|S|+1}$,.., $\mathbf{x}_{|S|+|U|}$.

More formally, in the transductive setting, we assume that a set $S_\mathcal{X}$ is obtained by selecting uniformly at random (without replacement) the examples in some given set V of unlabeled examples. The training set S is then obtained by adding the correct labels to the examples in $S_\mathcal{X}$ and $U \stackrel{\text{def}}{=} V\backslash S_\mathcal{X}$. Thus, in this setting, for any voter h, the random variable $m \cdot R_S(h)$ follows an hypergeometric law of parameters $|V^l|$, $|S|$, and $R_{V^l}(h)$. Recall that, in the inductive setting, it follows a binomial law of parameters $|S|$ and $R(h)$.

Moreover, as pointed out in Section 2, $\mathcal{M}_{Q^2}^{V^l}$ can be computed in this setting because no labels are needed to obtain its value. However, we do need labels to obtain $\mathcal{M}_Q^{V^l}$. Thus, in the transductive version of the algorithm that minimizes the C-bound, we consider $\mathcal{M}_{Q^2}^{V^l}$, but we replace $\mathcal{M}_Q^{V^l}$ by its empirical counterpart $\mathcal{M}_Q^S$. The PAC-Bayes bound theoratically corroborating this idea can be found in Laviolette et al. (2011).

Definition 6. – the TMinCq algorithm. *Given a set $\mathcal{H}$ of voters, a set S of labeled examples, a set U of unlabeled data, and a S-realizable $\mu > 0$, among all q-u distributions Q of margin $\mathcal{M}_Q^S = \mu$, the TMinCq algorithm consists in finding one that minimizes $\mathcal{M}_{Q^2}^{V^l}$.*

Following the definitions of the matrices for Program 1, replacing S by V^l, TMinCq solves the quadratic program described by Program 2.

Program 2 : **TMinCq**
a transductive quadratic program for classification

1: **Solve** $\quad \text{argmin}_{\mathbf{Q}} \quad \mathbf{Q}^T\,\mathbf{M}_{V^l}\,\mathbf{Q} \;-\; \mathbf{A}_{V^l}^T\,\mathbf{Q}$
2: **under constraints:** $\mathbf{m}_S^T\,\mathbf{Q} = \frac{\mu}{2} + \frac{1}{2n}\sum_{i=1}^{n}\mathcal{M}_{h_i}^S$
3: **and:** $0 \le Q_i \le \frac{1}{n} \quad \forall i \in \{1,\ldots,n\}$

6. Experiments

For all experiments, the QPs MinCq and TMinCq were solved using CVXOPT (Dahl & Vandenberghe, 2007), an off-the-shelf convex optimization solver. The first two experiments were performed in the inductive supervised framework. Except for MNIST, all datasets were taken from the UCI repository. Each dataset was randomly split into a training set S of $|S|$ examples and a testing set T of $|T|$ examples. We also specify the number of features of each dataset. For all algorithms, $R_T(B_Q)$ refers to the frequency of errors, measured on the testing set, of the resulting majority vote.

We first compared MinCq using decision stumps as voters (refered in Table 1 as MinCq-stumps), to AdaBoost (Schapire & Singer, 1999) and MDBoost (Shen & Li, 2010). For all algorithms, we used 10 decision stumps per feature. The fixed margin parameter μ of MinCq was selected using 10-fold cross-validation (CV) among 9 values between 0.0001 and 0.05. The number of iterations of AdaBoost was fixed to 200. The paremeter D of MDBoost was selected using 10-fold CV among the 14 values proposed in Shen & Li (2010). The results are summarized in Table 1.

In the second experiment, we compared MinCq using RBF kernel functions[4] as voters (refered in Table 1 as MinCq-RBF) to the SVM. For both algorithms, the kernel parameter γ was chosen by 10-fold CV among the set of 15 values proposed in Ambroladze et al. (2007). For the SVM, the soft-margin parameter C was chosen by 10-fold CV among a set of 15 values proposed in Ambroladze et al. (2007). For MinCq, parameter μ was selected as in the first experiment.

These results show that MinCq has an edge over AdaBoost (11 wins and 6 losses), MDBoost (12 wins and 6 losses) and the SVM (11 wins and 6 losses). Using the sign test methodology (Mendenhall, 1983), we obtain a p-value (for the null hypothesis: "there is no difference") of 0.17 against AdaBoost, 0.12 against MDBoost and 0.03 against SVM, implying that MinCq is better than SVM, with a confidence of 97%. Also note that the performance of MinCq can vary significantly by changing the nature of the voters (see Ionosphere and Tic-tac-toe).

The last experiment was performed in the transductive setting by using the benchmark framework of Chapelle et al. (2006). It provides natural and artificial datasets that can be used to compare transductive and semi-supervised learning algorithms. For all datasets, 12 random splits between labeled and unlabeled examples are provided for $|S| \in \{10, 100\}$. The small number of labeled examples makes it difficult to perform model selection. Indeed, many authors that published results in Chapelle et al. (2006) had to fix the values of their algorithms' hyperparameters to some value that was "experimentally known to perform well".

Within this framework, we compared TMinCq with TSVM on 6 datasets provided in the benchmark. Both algorithms were using the RBF kernel. The TSVM experiment follows the method proposed in Chapelle et al. (2006), i.e. the hyperparameter γ of the RBF kernel had its value set to the median of the pairwise distances between the training examples; and the parameter C was fixed to 100. For TMinCq, RBF kernel functions over the labeled examples were chosen. The parameters μ and γ were selected by 5-fold CV among the same values used in the second experiment. Note that for each CV-fold (and the final training), all available unlabeled examples are used. Table 2 shows the mean test error frequency of the 12 splits.

Even if this experiment shows that TSVM has generally an edge over TMinCq, it also shows that TMinCq has the potential to perform very well in this setting.

7. Conclusion

We have proposed two new bound-minimization learning algorithms which reduce to a quadratic program. This was made possible, firstly, by using quasi-uniform posteriors which do not limit the expressiveness of weighted majority votes and, secondly, by providing new PAC-Bayes bounds free of any KL-divergence (which also explains why the proposed learning algorithms avoid overfitting).

The proposed algorithms are always quadratic programs regardless of the choice for the set $\mathcal{H}$ of voters—which can be classifiers or similarity measures k evaluated on the examples. In contrast to the SVM, MinCq remains a quadratic program even if indefinite similarity measures are used.

Empirically, MinCq performs very well when compared with AdaBoost, MDBoost and the SVM. In the transductive setting, TMinCq competes with TSVM even if TSVM seems to perform better overall.

References

Ambroladze, Amiran, Parrado-Hernández, Emilio, and Shawe-Taylor, John. Tighter PAC-Bayes bounds. In *Proc. of the 2006 conference on Neural Information Processing Systems (NIPS-06)*, 2007.

Breiman, Leo. Random Forests. *Machine Learning*, 45

[4] For a RBF kernel k, $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}||\mathbf{x} - \mathbf{x}'||^2/\gamma^2)$.

Table 1. Results for AdaBoost and MDBoost vs MinCq on decision stumps, and SVM vs MinCq on RBF kernel.

Dataset				AdaBoost	MDBoost		MinCq-stumps		SVM			MinCq-RBF		
Name	$\|S\|$	$\|T\|$	#feat	$R_T(B_Q)$	$R_T(B_Q)$	D	$R_T(B_Q)$	μ	$R_T(B_Q)$	C	γ	$R_T(B_Q)$	μ	γ
Adult	1809	10000	14	**0.149**	0.150	30	0.152	0.04	0.159	100	0.0357	**0.157**	0.001	0.1429
BreastW	343	340	9	0.053	**0.047**	20	0.050	0.01	**0.038**	0.5	0.0035	0.044	0.01	0.0011
Credit-A	353	300	15	0.170	**0.143**	8	0.157	0.04	0.183	500	0.0083	**0.143**	0.02	0.3000
Glass	107	107	9	0.178	0.178	10	**0.168**	0.04	0.178	2	0.5000	**0.168**	0.02	2.0000
Haberman	144	150	3	0.260	0.260	5	**0.253**	0.02	**0.280**	0.02	0.0034	**0.280**	0.02	0.0417
Heart	150	147	13	0.259	**0.197**	5	0.224	0.05	**0.197**	1	0.1539	**0.197**	0.01	0.1539
Ionosphere	176	175	34	0.120	**0.097**	70	0.143	0.01	0.097	10	0.1324	**0.029**	0.0005	0.2353
Letter:AB	500	1055	16	0.010	0.009	30	**0.002**	0.05	**0.001**	0.02	0.2813	0.002	0.0005	0.1250
Letter:DO	500	1058	16	0.036	0.031	50	**0.023**	0.05	0.014	20	0.0078	**0.009**	0.0005	0.0313
Letter:OQ	500	1036	16	**0.038**	0.054	40	0.043	0.04	0.015	4	0.0313	**0.012**	0.001	0.0313
Liver	170	175	6	**0.320**	0.331	15	0.331	0.01	**0.314**	5	0.0013	**0.314**	0.01	0.0023
MNIST:08	500	1916	784	**0.008**	0.031	30	0.016	0.0001	**0.003**	2	0.0159	**0.003**	0.0001	0.0313
MNIST:17	500	1922	784	0.013	0.053	20	**0.012**	0.05	0.011	5	0.0102	**0.007**	0.0005	0.0057
MNIST:18	500	1936	784	**0.025**	0.071	12	**0.025**	0.03	**0.011**	1	0.0408	**0.011**	0.0005	0.0313
MNIST:23	500	1905	784	0.047	0.132	2	**0.033**	0.04	0.020	5	0.0230	**0.016**	0.0005	0.0159
Mushroom	4062	4062	22	**0.000**	0.001	90	**0.000**	0.02	**0.000**	10	0.0227	**0.000**	0.0001	0.0909
Sonar	104	104	60	0.231	0.298	90	**0.144**	0.05	0.163	2	0.4083	**0.135**	0.0001	0.4083
Tic-tac-toe	479	479	9	0.357	0.349	2	**0.344**	0.05	0.081	10	0.2222	**0.017**	0.0001	0.2222
Usvotes	235	200	16	**0.055**	**0.055**	5	**0.055**	0.02	**0.055**	5	0.0313	0.065	0.02	0.0313
Wdbc	285	284	30	**0.049**	**0.049**	90	0.053	0.04	0.074	0.5	0.0003	**0.067**	0.02	0.0003

Table 2. Summary of mean risks for 12 provided labeled/unlabeled splits, comparing TSVM vs TMinCq on RBF kernel.

Dataset				TSVM	TMinCq-RBF	Dataset		TSVM	TMinCq-RBF
Name	#feat.	$\|S\|$	$\|U\|$	$R_U(B_Q)$	$R_U(B_Q)$	$\|S\|$	$\|U\|$	$R_U(B_Q)$	$R_U(B_Q)$
BCI	241	10	1490	**0.487**	0.488	100	1400	**0.330**	0.359
COIL2	241	10	1490	0.431	**0.428**	100	1400	0.162	**0.153**
Digit1	241	10	1490	**0.197**	0.283	100	1400	0.064	**0.059**
g241c	241	10	1490	**0.247**	0.429	100	1400	**0.186**	0.267
g241n	241	10	1490	0.468	**0.445**	100	1400	**0.219**	0.263
USPS	241	10	1490	0.284	**0.278**	100	1400	**0.102**	0.129

(1):5–32, October 2001.

Catoni, Olivier. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Monograph series of the Institute of Mathematical Statistics, http://arxiv.org/abs/0712.0248, 2007.

Chapelle, Olivier, Schölkopf, Bernhard, and Zien, Alexander (eds.). *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

Dahl, Joachim and Vandenberghe, Lieven. CVXOPT, 2007. http://mloss.org/software/view/34/.

Devroye, Luc, Györfi, László, and Lugosi, Gábor. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, New York, NY, 1996.

Dredze, Mark, Kulesza, Alex, and Crammer, Koby. Multi-domain learning by confidence-weighted parameter combination. *Mach. Learn.*, 79(1-2):123–149, 2010.

Lacasse, Alexandre, Laviolette, François, Marchand, Mario, Germain, Pascal, and Usunier, Nicolas. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Proceedings of the 2006 conference on Neural Information Processing Systems (NIPS-06)*, 2007.

Laviolette, François and Marchand, Mario. PAC-Bayes risk bounds for stochastic averages and majority votes of sample-compressed classifiers. *Journal of Machine Learning Research*, 8:1461–1487, 2007.

Laviolette, François, Marchand, Mario, and Roy, Jean-Francis. From PAC-Bayes bounds to quadratic programs for majority votes (extended version), 2011. http://graal.ift.ulaval.ca/publications.php.

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

McAllester, David. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

Mendenhall, W. Nonparametric statistics. *Introduction to Probability and Statistics*, 604, 1983.

Schapire, Robert E. and Singer, Yoram. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.

Shen, Chunhua and Li, Hanxi. Boosting through optimization of margin distributions. *IEEE Transactions on Neural Networks*, 21(4):659–666, 2010.