Risk Bounds for the Majority Vote (Part I)

From a PAC-Bayesian Analysis to a Learning Algorithm

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, Jean-Francis Roy

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

This poster presents a recently accepted JMLR paper. Go to http://graal.ift.ulaval.ca/majorityvote/ for a preliminary version, source code, and more!

Setting and Motivation

Training set We draw m examples *i.i.d.* from a distribution D on $\mathcal{X}\times\{-1,+1\}$:

$$S = \{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\} \sim D^m\,.$$

Majority vote classifier Given a set $\mathcal{H}$ of *voters* $f:\mathcal{X}\to[-1,+1]$ and a *posterior* distribution Q on $\mathcal{H}$, we study the Q-weighted majority vote classifier:

$$B_Q(x) \stackrel{\text{def}}{=} \operatorname{sgn}\left[\mathop{\mathbf{E}}_{f\sim Q} f(x)\right].$$

Majority votes are everywhere! Many learning algorithms output ensemble of voters: Bagging, Boosting, Random Forests... Even kernel method classifiers (such SVM) and neurons of Neural Networks can be expressed as majority votes.

Bayes Risk versus Gibbs Risk

Bayes risk Risk of majority vote classifier B_Q on data-generating distribution D:

$$R_D(B_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} I\Big[B_Q(x)\neq y\Big] = \mathop{\mathbf{E}}_{(x,y)\sim D} I\Big[\mathop{\mathbf{E}}_{f\sim Q} y\cdot f(x)\le 0\Big],$$

where $I(a)=1$ if predicate a is true and 0 otherwise.

Gibbs risk The *Gibbs* classifier $G_Q(x)$ draws f according to Q and returns $f(x)$. Its risk is defined as

$$R_D(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D}\mathop{\mathbf{E}}_{f\sim Q}\mathcal{L}(f(x),y) = \frac{1}{2}\left(1-\mathop{\mathbf{E}}_{(x,y)\sim D}\mathop{\mathbf{E}}_{f\sim Q} y\cdot f(x)\right),$$

where $\mathcal{L}(y,y') \stackrel{\text{def}}{=} \frac{1}{2}(1-y\cdot y')$ is the *linear loss*.

The Margin of the Majority Vote

Margin on an example $M_Q(x,y) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{f\sim Q} y\cdot f(x)$

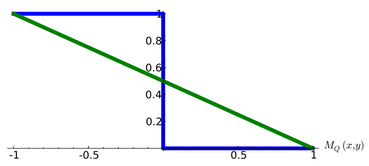
Margin on a distribution Let M_Q^D be the random variable that, given any example (x,y) drawn according to D, outputs $M_Q(x,y)$.

Thus, we rewrite the **Bayes risk** and the **Gibbs risk**,

$$R_D(B_Q) = \Pr_{(x,y)\sim D}\Big(M_Q(x,y)\le 0\Big) \quad\text{and}\quad R_D(G_Q) = \frac{1}{2}\Big(1-\mu_1(M_Q^D)\Big),$$

and we recover the *factor two* bound by Markov's inequality ($\Pr(X\ge a)\le \frac{\mathbf{E}\,X}{a}$) :

$$\begin{aligned} R_D(B_Q) &= \Pr_{(x,y)\sim D}\big(1-M_Q(x,y)\ge 1\big)\\ &\le \mathop{\mathbf{E}}_{(x,y)\sim D}\big(1-M_Q(x,y)\big) = 2\,R_D(G_Q)\,. \end{aligned}$$



Indroducing the $\mathcal{C}$-bound

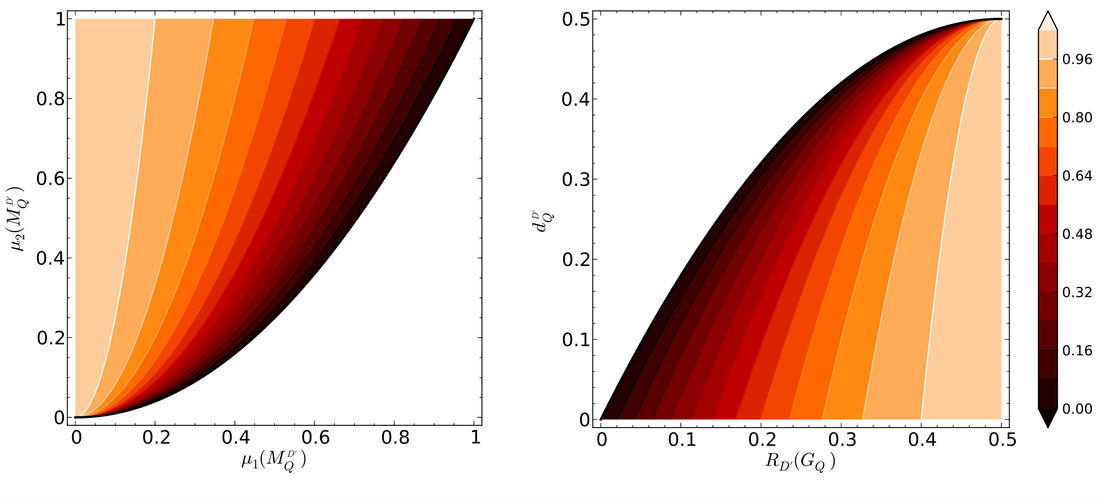
Markov's uses only the first moment $\mu_1(M_Q^D)$ of the margin. We exploit the second moment $\mu_2(M_Q^D)$ by Chebychev's inequality ($\Pr(X-\mathbf{E}\,X\ge a)\le \frac{\mathbf{Var}\,X}{a^2+\mathbf{Var}\,X}$).

Theorem ($\mathcal{C}$-bound) *For any distribution Q on a set of voters and any distribution D on $\mathcal{X}\times\{-1,1\}$, if $\mu_1(M_Q^{D'})>0$ (i.e., $R_D(G_Q)<1/2$), we have*

$$R_D(B_Q) \le \mathcal{C}_Q^D \stackrel{\text{def}}{=} 1-\frac{\big(\mu_1(M_Q^D)\big)^2}{\mu_2(M_Q^D)} = 1-\frac{\big(1-2\cdot R_D(G_Q)\big)^2}{1-2\cdot d_Q^D},$$

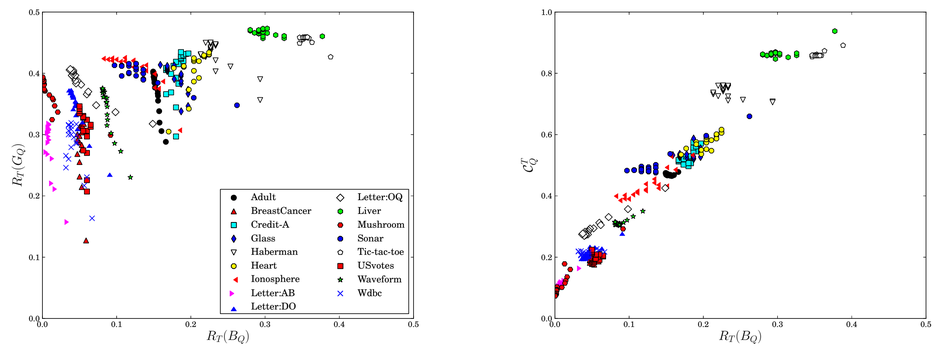
where d_Q^D is the ***expected disagreement*** *between voters:*

$$d_Q^D \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,\cdot)\sim D}\mathop{\mathbf{E}}_{f_1\sim Q}\mathop{\mathbf{E}}_{f_2\sim Q}\mathcal{L}\Big(f_1(x),f_2(x)\Big) = \frac{1}{2}\Big(1-\mu_2(M_Q^D)\Big).$$



Predictive power of the $\mathcal{C}$-bound

We run AdaBoost on UCI datasets to obtain a collection of majority vote classifiers, and we compare the factor two and the $\mathcal{C}$-bound on test sets.



PAC-Bayesian Bound for the Gibbs Risk

Theorem *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

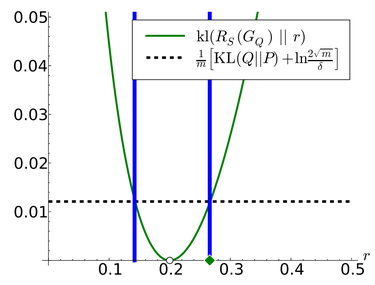
$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors } Q \text{ on } \mathcal{H}:\\ \mathrm{kl}\big(R_S(G_Q)\big\|R_D(G_Q)\big)\le \dfrac{1}{m}\left[\mathrm{KL}(Q\|P)+\ln\dfrac{2\sqrt{m}}{\delta}\right]\end{array}\right)\ge 1-\delta\,.$$

PAC-Bound 0

With probability $1-\delta$ over the choice of $S\sim D^m$,

$$\forall Q \text{ on } \mathcal{H}\ :\ R_D(B_Q)\le 2\cdot\sup\mathcal{R}_{Q,S}^{\delta}\,,$$

$$\mathcal{R}_{Q,S}^{\delta} \stackrel{\text{def}}{=} \left\{r:\ \mathrm{kl}\big(R_S(G_Q)\,\|\,r\big)\le\frac{1}{m}\left[\mathrm{KL}(Q\|P)+\ln\frac{2\sqrt{m}}{\delta}\right]\right\}.$$



PAC-Bayesian Bound for the Disagreement

Theorem *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

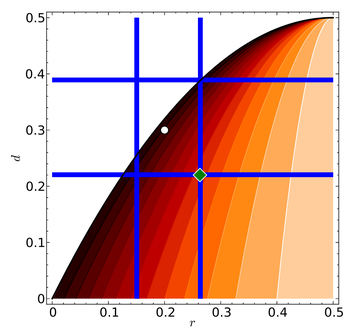
$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors } Q \text{ on } \mathcal{H}:\\ \mathrm{kl}\big(d_Q^S\big\|d_Q^D\big)\le \dfrac{1}{m}\left[2\cdot\mathrm{KL}(Q\|P)+\ln\dfrac{2\sqrt{m}}{\delta}\right]\end{array}\right)\ge 1-\delta\,.$$

PAC-Bound 1

With probability $1-\delta$ over the choice of $S\sim D^m$,

$$\forall Q \text{ on } \mathcal{H}\ :\ R_D(B_Q)\le 1-\frac{\left(1-2\cdot\sup\mathcal{R}_{Q,S}^{\delta/2}\right)^2}{1-2\cdot\inf\mathcal{D}_{Q,S}^{\delta/2}}\,,$$

$$\mathcal{D}_{Q,S}^{\delta} \stackrel{\text{def}}{=} \left\{d:\ \mathrm{kl}(d_Q^S\,\|\,d)\le\frac{1}{m}\left[2\cdot\mathrm{KL}(Q\|P)+\ln\frac{2\sqrt{m}}{\delta}\right]\right\}.$$



Another PAC-Bayesian Bound

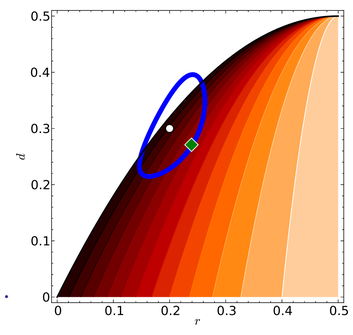
Idea Let's bound the Gibbs risk $R_D(G_Q)$ and the disagreement d_Q^D simultaneously.

PAC-Bound 2

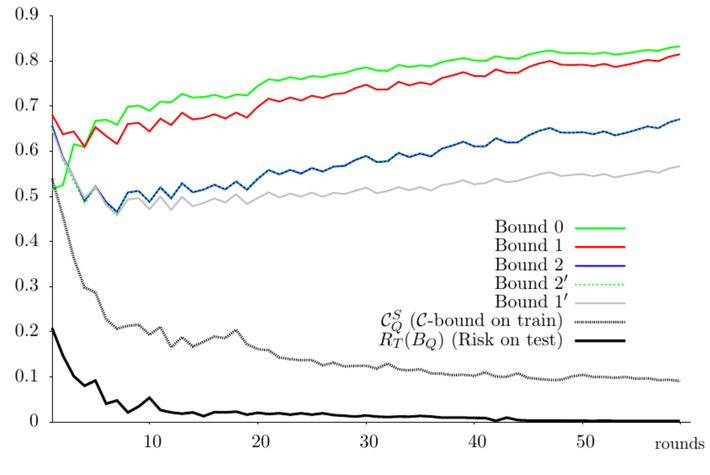
With probability $1-\delta$ over the choice of $S\sim D^m$,

$$\forall Q \text{ on } \mathcal{H}\ :\ R_D(B_Q)\le \sup_{(r,d)\in\mathcal{A}_{Q,S}^{\delta}}\left[1-\frac{(1-2\,r)^2}{1-2\,d}\right],$$

$$\mathcal{A}_{Q,S}^{\delta} \stackrel{\text{def}}{=} \left\{(r,d):\widetilde{\mathrm{kl}}(R_S(G_Q),d_Q^S\|r,d)\le\frac{1}{m}\left[2\cdot\mathrm{KL}(Q\|P)+\ln\frac{2\sqrt{m}+m}{\delta}\right]\right\}.$$



Empirical Comparison (Adaboost on Mushroom)



⇒ **Next on Part II:** Jean-Francis designs an algorithm to minimize the $\mathcal{C}$-bound.