La théorie PAC-Bayes en apprentissage supervisé

Présentation au LRI de l'université Paris XI

François Laviolette,
Laboratoire du GRAAL,

Université Laval, Québec, Canada

14 dcembre 2010

Summary

Aujourd'hui, j'ai l'intention de

- vous présenter les mathématiques qui sous tendent la théorie PAC-Bayes
- vous présenter des algorithmes qui consistent en la minimisation d'une borne PAC-Bayes et comparer ces derniers avec des algorithmes existants.

Definitions

- Each example $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, +1\}$, is drawn acc. to D.
- The (true) risk $R(h)$ and training error $R_S(h)$ are defined as:

$$R(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(h(\mathbf{x}) \neq y) \quad ; \quad R_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} I(h(\mathbf{x}_i) \neq y_i) \, .$$

- The learner's goal is to choose a **posterior distribution** Q on a space $\mathcal{H}$ of classifiers such that the risk of the Q-weighted **majority vote** B_Q is as small as possible.

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

- B_Q is also called the *Bayes classifier*.

The Gibbs clasifier

- PAC-Bayes approach does not directly bounds the risk of B_Q
- It bounds the risk of the **Gibbs classifier** G_Q:
  - to predict the label of $\mathbf{x}$, G_Q draws h from $\mathcal{H}$ and predicts $h(\mathbf{x})$

- The risk and the training error of G_Q are thus defined as:

$$R(G_Q) = \mathop{\mathbf{E}}_{h\sim Q} R(h) \quad ; \quad R_S(G_Q) = \mathop{\mathbf{E}}_{h\sim Q} R_S(h)\,.$$

G_Q, B_Q, and $\mathrm{KL}(Q\|P)$

- If B_Q misclassifies **x**, then at least half of the classifiers (under measure Q) err on **x**.
  - Hence: $R(B_Q) \leq 2R(G_Q)$
  - **Thus, an upper bound on $R(G_Q)$ gives rise to an upper bound on $R(B_Q)$**

- PAC-Bayes makes use of a **prior distribution** P on $\mathcal{H}$.

- The risk bound depends on the **Kullback-Leibler divergence**:

$$\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)} \,.$$

A PAC-Bayes bound to rule them all !

J.R.R. Tolkien, roughly
or John Langford, less roughly.

Theorem 1 Germain et al. 2009

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\mathcal{H}$ of classifiers, for any prior distribution P of support $\mathcal{H}$, for any $\delta \in (0, 1]$, and for any convex function $\mathcal{D} : [0, 1] \times [0, 1] \to \mathbb{R}$, we have

$$\Pr_{S \sim D^m} \left(\forall\, Q \text{ on } \mathcal{H} : \ \mathcal{D}(R_S(G_Q), R(G_Q)) \leq \frac{1}{m} \left[\mathrm{KL}(Q \| P) + \ln \left(\frac{1}{\delta} \mathop{\mathbf{E}}_{S \sim D} \mathop{\mathbf{E}}_{h \sim P} e^{m \mathcal{D}(R_S(h), R(h))} \right) \right] \right) \geq 1 - \delta \,.$$

Proof of Theorem 1

- Since $\underset{h\sim P}{\mathbf{E}} e^{m\mathcal{D}(R_S(h),R(h))}$ is a non-negative r.v., Markov's inequality gives

$$\Pr_{S\sim D^m}\left(\underset{h\sim P}{\mathbf{E}} e^{m\mathcal{D}(R_S(h),R(h))} \leq \tfrac{1}{\delta}\underset{S\sim D^m}{\mathbf{E}}\underset{h\sim P}{\mathbf{E}} e^{m\mathcal{D}(R_S(h),R(h))}\right)\geq 1-\delta\,.$$

- Hence, by taking the logarithm on each side of the inequality and by transforming the expectation over P into an expectation over Q :

$$\Pr_{S\sim D^m}\left(\forall Q:\ \ln\left[\underset{h\sim Q}{\mathbf{E}}\ \tfrac{P(h)}{Q(h)}e^{m\mathcal{D}(R_S(h),R(h))}\right]\leq \ln\left[\tfrac{1}{\delta}\underset{S\sim D^m}{\mathbf{E}}\underset{h\sim P}{\mathbf{E}} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right)\geq 1-\delta\,.$$

- Then, exploiting the fact that the logarithm is a concave function, by an application of Jensen's inequality, we obtain

$$\Pr_{S\sim D^m}\left(\forall Q:\ \underset{h\sim Q}{\mathbf{E}}\ \ln\left[\tfrac{P(h)}{Q(h)}e^{m\mathcal{D}(R_S(h),R(h))}\right]\leq \ln\left[\tfrac{1}{\delta}\underset{S\sim D^m}{\mathbf{E}}\underset{h\sim P}{\mathbf{E}} e^{m\mathcal{D}(R_S(h),R(h))}\right]\right)\geq 1-\delta\,.$$

Proof of Theorem 1 (cont)

$$\Pr_{S\sim D^m}\left(\forall Q:\ \mathop{\mathbf{E}}_{h\sim Q}\ \ln\left[\frac{P(h)}{Q(h)}e^{m\mathcal{D}(R_S(h),R(h))}\right]\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{D}(R_S(h),R(h))}\right]\right)\geq 1-\delta\ .$$

- From basic logarithm properties, and from the fact that $\mathop{\mathbf{E}}_{h\sim Q}\ \ln\left[\frac{P(h)}{Q(h)}\right]\stackrel{\text{def}}{=}-\mathrm{KL}(Q\|P)$, we now have

$$\Pr_{S\sim D^m}\left(\forall Q:\ -\mathrm{KL}(Q\|P)+\mathop{\mathbf{E}}_{h\sim Q}\ m\mathcal{D}(R_S(h),R(h))\leq \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{D}(R_S(h),R(h))}\right]\right)\geq 1-\delta\ .$$

- Then, since $\mathcal{D}$ has been supposed convexe, again by the Jensen inequality, we have

$$\mathop{\mathbf{E}}_{h\sim Q}\ m\mathcal{D}(R_S(h),R(h))\ =\ m\,\mathcal{D}\left(\mathop{\mathbf{E}}_{h\sim Q}\ R_S(h),\mathop{\mathbf{E}}_{h\sim Q}\ R(h)\right),$$

which immediately implies the result. □

Applicability of Theorem 1

How can we estimate $\ln \left[\frac{1}{\delta} \underset{S \sim D^m}{\mathbf{E}} \underset{h \sim P}{\mathbf{E}} e^{m\mathcal{D}(R_S(h), R(h))} \right]$?

The Seeger's bound (2002)

Seeger Bound

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

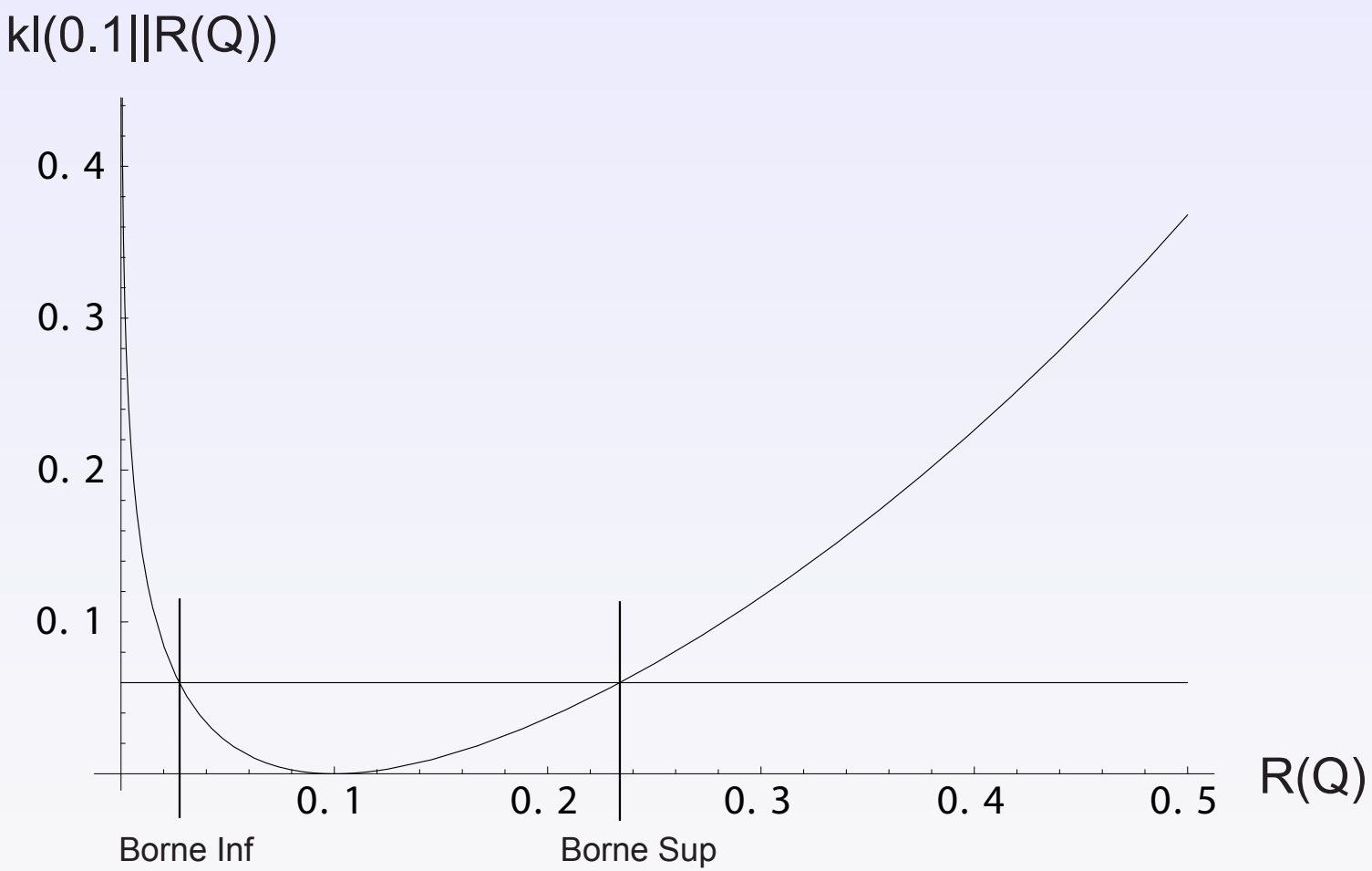
$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}:\ \mathrm{kl}(R_S(G_Q), R(G_Q)) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\xi(m)}{\delta}\right]\right) \geq 1-\delta\,,$$

where $\mathrm{kl}(q,p) \stackrel{\text{def}}{=} q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$,

and where $\xi(m) \stackrel{\text{def}}{=} \sum_{k=0}^{m}\binom{m}{k}(k/m)^k(1-k/m)^{m-k}$.

- Note: $\xi(m) \leq 2\sqrt{m}$

Graphical illustration of the Seeger bound



Proof of the Seeger bound

Follows immediately from Theorem 1 by choosing $\mathcal{D}(q,p) = \mathrm{kl}(q,p)$.

- Indeed, in that case we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{D}(R_S(h),R(h))} &= \mathop{\mathbf{E}}_{h\sim P}\ \mathop{\mathbf{E}}_{S\sim D^m}\ \left(\frac{R_S(h)}{R(h)}\right)^{mR_S(h)}\left(\frac{1-R_S(h)}{1-R(h)}\right)^{m(1-R_S(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P}\ \sum_{k=0}^{m}\ \Pr_{S\sim D^m}\left(R_S(h)=\tfrac{k}{m}\right)\left(\frac{\frac{k}{m}}{R(h)}\right)^{k}\left(\frac{1-\frac{k}{m}}{1-R(h)}\right)^{m-k} \\
&= \sum_{k=0}^{m}\binom{m}{k}(k/m)^k(1-k/m)^{m-k}\,, \qquad (1) \\
&\leq 2\sqrt{m}\,.
\end{aligned}$$

□

- Note that, in Line (1) of the proof, $\Pr_{S\sim D^m}\left(R_S(h) = \frac{k}{m}\right)$ is replaced by the probability mass function of the binomial.
- This is **only true if** the examples of S are drawn iid. (i.e., $S \sim D^m$)
- So this result is no longuer valid in the non iid case, even if Theorem 1 is.

The McAllester's bound (1998)

Put $\mathcal{D}(q,p) = \frac{1}{2}(q-p)^2$, Theorem 1 then gives

McAllester Bound

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, we have

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H}:\ \frac{1}{2}(R_S(G_Q), R(G_Q))^2 \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\xi(m)}{\delta}\right]\right) \geq 1-\delta\,,$$

where $\mathrm{kl}(q,p) \stackrel{\text{def}}{=} q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$,

and where $\xi(m) \stackrel{\text{def}}{=} \sum_{k=0}^{m}\binom{m}{k}(k/m)^k(1-k/m)^{m-k}$.

The Catoni's bound (2004)

In Theorem 1, let $\mathcal{D}(q,p) = \mathcal{F}(p) - C \cdot q.$, then

Catoni's bound

For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, and any positive real number C, we have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\, Q \text{ on } \mathcal{H}: \\ R(G_Q) \leq \frac{1}{1-e^{-C}}\left\{1-\exp\left[-\left(C\cdot R_S(G_Q) \right.\right.\right. \\ \qquad\qquad \left.\left.\left. +\frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right]\right)\right]\right\}\end{array}\right) \quad \geq \quad 1 - \delta.$$

- Because,

$$\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{h\sim P}\ e^{m\,\mathcal{D}(R_S(h),R(h))} \quad = \quad \mathop{\mathbf{E}}_{h\sim P}\ e^{m\mathcal{F}(R(h))}\Big(R(h)e^{-C}+(1-R(h))\Big)^m.$$

Bounding $\mathop{\mathbf{E}}_{S\sim D}\ \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R(h))}$: other ways

- via concentration inequality
  - used in the original proof of Seeger (and in the one due to Langford).
  - used by Higgs (2009) to generalized the Seeger's bound the the transductive case
  - used by Ralaivola et al. (2008) for the non iid case.
- via martingales
  - used by Lever et al (2010) to generalized PAC-Bayes bound to U-statistics of order > 1.

Observations about Catoni's bound

- G_Q is minimizing the Catoni's bound iff it minimizes the following cost function (linear in $R_S(G_Q)$):

$$C\, m\, R_S(G_Q) + \mathrm{KL}(Q\|P)$$

- We have a **hyperparameter** C to tune (in contrast with the Seeger' bound).
- Seeger' bound gives a bound which is always tighter except for a narrow range of C values.
  - In fact, if we would replace $\xi(m)$ by one, LS-bound would always be a tighter.

Observations about Catoni's bound (cont)

- Given any prior P, the posterior Q^* minimizing the Catoni's bound is given by the Boltzman distribution:

$$Q^*(h) \;=\; \frac{1}{Z} P(h) e^{-C \cdot m R_S(h)} \,.$$

- We could sample Q^* by Markov Chain Monté Carlo.
  - But the mixing time being unknown, we have few control over the precision of the approximation.

- To avoid MCMC, let us analyse the case where Q is chosen from a **parameterized set of distributions** over the (continuous) space of **linear classifiers**.

The problem of bounding $R(G_Q)$ instead of $R(B_Q)$

The main problem PAC-Bayes theory is the fact that it allows us to bound the Gibbs risk but, most of the time, it is the Bayes risk we are interested in. For this problem I will discuss here two possible answers:

- Answer#1: if a non too small "part" of the classifier of $\mathcal{H}$ are strong, then one can obtained a quiet tight bound (exemple: if $\mathcal{H}$ is the set of all linear classifiers in a high-dimensional feature vectors space, like in SVM)
- Answer#2: otherwise, extend the PAC-Bayes bound to something else than the Gibbs's Risk

Specialization to Linear classifiers

- Each $\mathbf{x}$ is mapped to a high-dimensional feature vector $\boldsymbol{\phi}(\mathbf{x})$:

$$\boldsymbol{\phi}(\mathbf{x}) \stackrel{\text{def}}{=} (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x})) \,.$$

- $\boldsymbol{\phi}$ is often implicitly given by a Mercer kernel

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{x}') \,.$$

- The output $h_{\mathbf{v}}(\mathbf{x})$ of linear classifier $h_{\mathbf{v}}$ with weight vector $\mathbf{v}$ is given by

$$h_{\mathbf{v}}(\mathbf{x}) \;=\; \text{sgn}\left(\mathbf{v} \cdot \boldsymbol{\phi}(\mathbf{x})\right) \,.$$

- Let us moreover suppose that each posterior $Q_{\mathbf{w}}$ is an isotropic Gaussian centered on $\mathbf{w}$:

$$Q_{\mathbf{w}}(\mathbf{v}) \;=\; \left(\tfrac{1}{\sqrt{2\pi}}\right)^N \exp\left(-\tfrac{1}{2}\|\mathbf{v}-\mathbf{w}\|^2\right)$$

Bayes-equivalent classifiers

- With this choice for $Q_{\mathbf{w}}$, the majority vote $B_{Q_{\mathbf{w}}}$ is the same classifier as $h_{\mathbf{w}}$ since:

$$B_{Q_{\mathbf{w}}}(\mathbf{x}) \;=\; \operatorname{sgn}\left(\mathop{\mathbf{E}}_{\mathbf{v}\sim Q_{\mathbf{w}}} \operatorname{sgn}\left(\mathbf{v}\cdot\boldsymbol{\phi}(\mathbf{x})\right)\right) \;=\; \operatorname{sgn}\left(\mathbf{w}\cdot\boldsymbol{\phi}(\mathbf{x})\right) \;=\; h_{\mathbf{w}}(\mathbf{x})\,.$$

- Thus $R(h_{\mathbf{w}}) = R(B_{Q_{\mathbf{w}}}) \leq 2R(G_{Q_{\mathbf{w}}})$: an upper bound on $R(G_{Q_{\mathbf{w}}})$ also provides an upper bound on $R(h_{\mathbf{w}})$.

- The prior $P_{\mathbf{w}_p}$ is also an isotropic Gaussian centered on $\mathbf{w}_p$. Consequently:

$$\mathrm{KL}(Q_{\mathbf{w}} \| P_{\mathbf{w}_p}) \;=\; \frac{1}{2}\|\mathbf{w}-\mathbf{w}_p\|^2\,.$$

Gibbs' risk

We need to compute Gibb's risk $R_{(\mathbf{x},y)}(G_{Q_\mathbf{w}})$ on $(\mathbf{x}, y)$ since:

$$R_{(\mathbf{x},y)}(G_{Q_\mathbf{w}}) \stackrel{\text{def}}{=} \int_{\mathbb{R}^N} Q_\mathbf{w}(\mathbf{v})\, I(y\mathbf{v}\cdot\boldsymbol{\phi}(\mathbf{x})<0)\, d\mathbf{v}$$

we have:

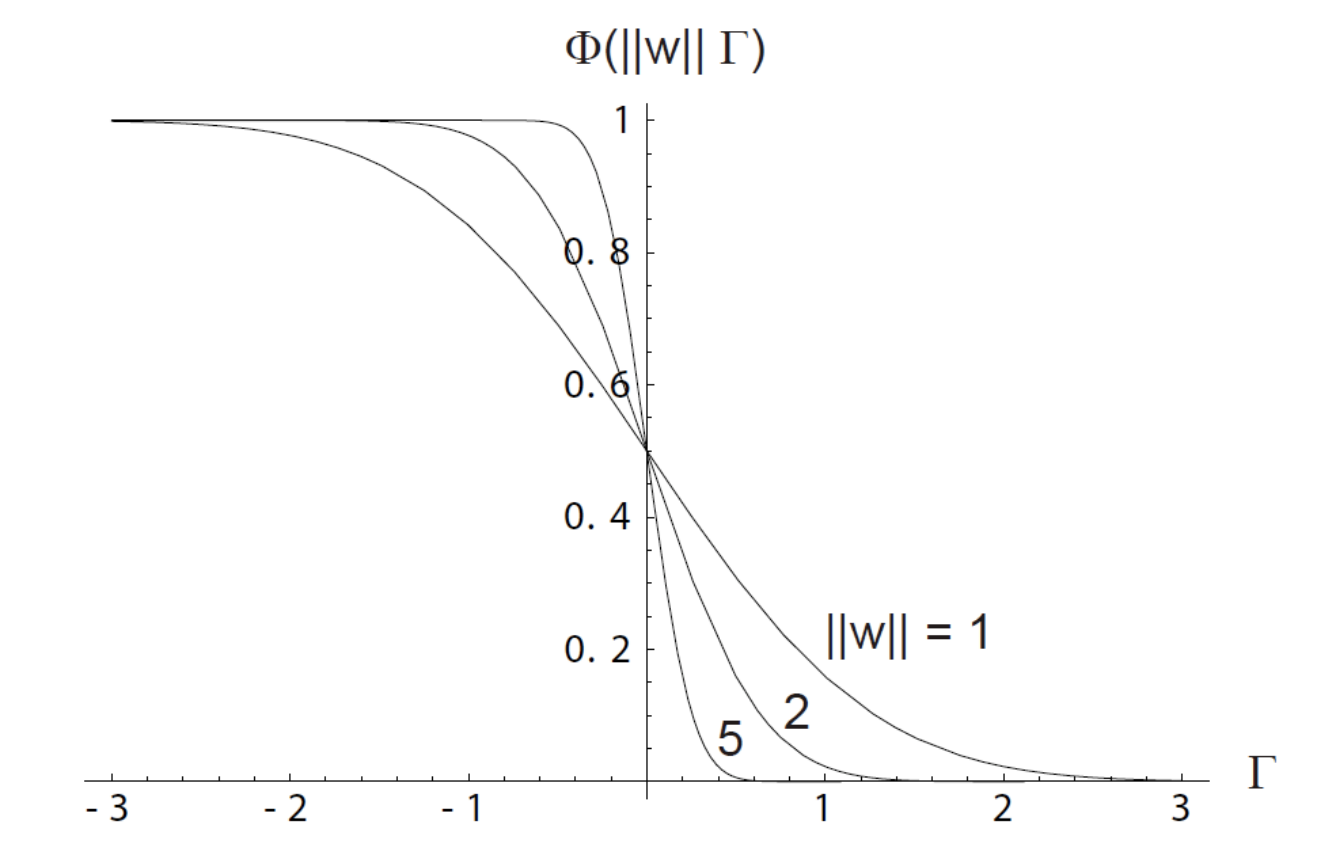
$$R(G_{Q_\mathbf{w}}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} R_{(\mathbf{x},y)}(G_{Q_\mathbf{w}}) \quad \text{and} \quad R_S(G_{Q_\mathbf{w}}) = \frac{1}{m}\sum_{i=1}^m R_{(\mathbf{x}_i,y_i)}(G_{Q_\mathbf{w}})\,.$$

Moreover, as in Langford (2005), the Gaussian integral gives:

$$R_{(\mathbf{x},y)}(G_{Q_\mathbf{w}}) = \Phi\Big(\|\mathbf{w}\|\,\Gamma_\mathbf{w}(\mathbf{x},y)\Big)$$

where: $\Gamma_\mathbf{w}(\mathbf{x},y) \stackrel{\text{def}}{=} \frac{y\mathbf{w}\cdot\boldsymbol{\phi}(\mathbf{x})}{\|\mathbf{w}\|\,\|\boldsymbol{\phi}(\mathbf{x})\|}$ and $\Phi(a) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}}\int_a^\infty \exp\Big(-\frac{1}{2}x^2\Big)\,dx\,.$

Probit loss



Objective function from Catoni's bound

Recall that, to minimize the Catoni's bound, for fixed C and $\mathbf{w}_p$, we need to find $\mathbf{w}$ that minimizes:

$$C\, m\, R_S(G_{Q_\mathbf{w}}) + \mathrm{KL}(Q_\mathbf{w} \| P_{\mathbf{w}_p})$$

Which, according to preceding slides, corresponds of minimizing

$$C \sum_{i=1}^{m} \Phi\left(\frac{y_i \mathbf{w} \cdot \phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_i)\|}\right) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_p\|^2$$

Objective function from Catoni's bound

So PAC-Bayes tells us to minimize

$$C \sum_{i=1}^{m} \Phi\left(\frac{y_i \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)}{\|\boldsymbol{\phi}(\mathbf{x}_i)\|}\right) + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_p\|^2$$

Note that, when $\mathbf{w}_p = \mathbf{0}$ (absence of prior knowledge), this is very similar to SVM . Indeed, SVM minimizes:

$$C \sum_{i=1}^{m} \max\Big(0, 1 - y_i \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)\Big) + \frac{1}{2}\|\mathbf{w}\|^2 ,$$

- The probit loss is simply replaced by the convex hinge loss.
- Up to convexe relaxation, PAC-Bayes theory has rediscover SVM !!!

Numerical result [ICML09]

Dataset				(s) SVM		(1) PBGD1			(2) PBGD2			(3) PBGD3	
Name	$\|S\|$	$\|T\|$	n	$R_T(\mathbf{w})$	*Bnd*	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	*Bnd*	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	*Bnd*	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$
Usvotes	235	200	16	0.055	0.370	0.080	0.117	0.244	**0.050**	0.050	0.153	0.075	0.085
Credit-A	353	300	15	0.183	0.591	**0.150**	0.196	0.341	**0.150**	0.152	0.248	0.160	0.267
Glass	107	107	9	0.178	0.571	**0.168**	0.349	0.539	0.215	0.232	0.430	**0.168**	0.316
Haberman	144	150	3	0.280	0.423	0.280	0.285	0.417	0.327	0.323	0.444	**0.253**	0.250
Heart	150	147	13	0.197	0.513	0.190	0.236	0.441	**0.184**	0.190	0.400	0.197	0.246
Sonar	104	104	60	0.163	0.599	0.250	0.379	0.560	0.173	0.231	0.477	**0.144**	0.243
BreastCancer	343	340	9	**0.038**	0.146	0.044	0.056	0.132	0.041	0.046	0.101	0.047	0.051
Tic-tac-toe	479	479	9	0.081	0.555	0.365	0.369	0.426	0.173	0.193	0.287	**0.077**	0.107
Ionosphere	176	175	34	0.097	0.531	0.114	0.242	0.395	0.103	0.151	0.376	**0.091**	0.165
Wdbc	285	284	30	0.074	0.400	0.074	0.204	0.366	**0.067**	0.119	0.298	0.074	0.210
MNIST:0vs8	500	1916	784	**0.003**	0.257	0.009	0.053	0.202	0.007	0.015	0.058	0.004	0.011
MNIST:1vs7	500	1922	784	0.011	0.216	0.014	0.045	0.161	**0.009**	0.015	0.052	0.010	0.012
MNIST:1vs8	500	1936	784	0.011	0.306	0.014	0.066	0.204	0.011	0.019	0.060	**0.010**	0.024
MNIST:2vs3	500	1905	784	**0.020**	0.348	0.038	0.112	0.265	0.028	0.043	0.096	0.023	0.036
Letter:AvsB	500	1055	16	**0.001**	0.491	0.005	0.043	0.170	0.003	0.009	0.064	**0.001**	0.408
Letter:DvsO	500	1058	16	0.014	0.395	0.017	0.095	0.267	0.024	0.030	0.086	**0.013**	0.031
Letter:OvsQ	500	1036	16	0.015	0.332	0.029	0.130	0.299	0.019	0.032	0.078	**0.014**	0.045
Adult	1809	10000	14	**0.159**	0.535	0.173	0.198	0.274	0.180	0.181	0.224	0.164	0.174
Mushroom	4062	4062	22	**0.000**	0.213	0.007	0.032	0.119	0.001	0.003	0.011	**0.000**	0.001

Majority vote of weak classifiers

- The classical PAC-Bayes theory bounds the risk of the majority vote $R(B_Q)$, trought twice the Gibbs's risk $2R(G_Q)$
- In the case of linear classifiers, there exists Q s.t. $R(G_Q)$ is relatively small, it seems to be a good idea,
- but what if the set $\mathcal{H}$ of voters is only composed of weak voters ? (Like in Boosting)
  - In that case, the Gibbs's risk cannot be a good predictor for the Bayes's risk.
  - Indeed, it is well-known that voting can dramatically improve performance when the "community" of classifiers tend to compensate the individual errors.
- So what can we do in this case ?

Answer # 1

- Suppose $\mathcal{H} = \{h_1, .., h_n, h_{n+1}, .., h_{2n}\}$ with $h_{i+n} = -h_i$,
- and consider instead, the set of *all the majority votes* over $\mathcal{H}$

$$\mathcal{H}^{MV} \stackrel{\text{def}}{=} \{\text{sgn}\,(\mathbf{v} \cdot \boldsymbol{\phi}(\mathbf{x})) : \mathbf{v} \in \mathbb{R}^{|\mathcal{H}|}\}$$

where $\boldsymbol{\phi}(\mathbf{x}) \stackrel{\text{def}}{=} (h_1(\mathbf{x}), \ldots, h_{2n}(\mathbf{x}))$.

- Then we are back to the linear classifier specialization.

Numerical result [ICML09], with decision stumps as weak learners

Dataset				(a) AdaBoost		(1) PBGD1			(2) PBGD2			(3) PBGD3		
Name	$\|S\|$	$\|T\|$	n	$R_T(\mathbf{w})$	Bnd	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	Bnd	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	Bnd	$R_T(\mathbf{w})$	$G_T(\mathbf{w})$	Bnd
Usvotes	235	200	16	**0.055**	0.346	0.085	0.103	0.207	0.060	0.058	0.165	0.060	0.057	0.261
Credit-A	353	300	15	0.170	0.504	0.177	0.243	0.375	0.187	0.191	0.272	**0.143**	0.159	0.420
Glass	107	107	9	0.178	0.636	0.196	0.346	0.562	0.168	0.176	0.395	**0.150**	0.226	0.581
Haberman	144	150	3	**0.260**	0.590	0.273	0.283	0.422	0.267	0.287	0.465	0.273	0.386	0.424
Heart	150	147	13	0.259	0.569	**0.170**	0.250	0.461	0.190	0.205	0.379	0.184	0.214	0.473
Sonar	104	104	60	0.231	0.644	0.269	0.376	0.579	0.173	0.168	0.547	**0.125**	0.209	0.622
BreastCancer	343	340	9	0.053	0.295	**0.041**	0.058	0.129	0.047	0.054	0.104	0.044	0.048	0.190
Tic-tac-toe	479	479	9	0.357	0.483	0.294	0.384	0.462	**0.207**	0.208	0.302	**0.207**	0.217	0.474
Ionosphere	176	175	34	0.120	0.602	0.120	0.223	0.425	0.109	0.129	0.347	**0.103**	0.125	0.557
Wdbc	285	284	30	0.049	0.447	0.042	0.099	0.272	0.049	0.048	0.147	**0.035**	0.051	0.319
MNIST:0vs8	500	1916	784	0.008	0.528	0.015	0.052	0.191	0.011	0.016	0.062	**0.006**	0.011	0.262
MNIST:1vs7	500	1922	784	**0.013**	0.541	0.020	0.055	0.184	0.015	0.016	0.050	0.016	0.017	0.233
MNIST:1vs8	500	1936	784	0.025	0.552	0.037	0.097	0.247	0.027	0.030	0.087	**0.018**	0.037	0.305
MNIST:2vs3	500	1905	784	0.047	0.558	0.046	0.118	0.264	0.040	0.044	0.105	**0.034**	0.048	0.356
Letter:AvsB	500	1055	16	0.010	0.254	0.009	0.050	0.180	**0.007**	0.011	0.065	**0.007**	0.044	0.180
Letter:DvsO	500	1058	16	0.036	0.378	0.043	0.124	0.314	0.033	0.039	0.090	**0.024**	0.038	0.360
Letter:OvsQ	500	1036	16	**0.038**	0.431	0.061	0.170	0.357	0.053	0.053	0.106	0.042	0.049	0.454
Adult	1809	10000	14	**0.149**	0.394	0.168	0.196	0.270	0.169	0.169	0.209	0.159	0.160	0.364
Mushroom	4062	4062	22	**0.000**	0.200	0.046	0.065	0.130	0.016	0.017	0.030	0.002	0.004	0.150

Answer # 2: generalize the PAC-Bayes theorem to something else than the Gibbs's risk !

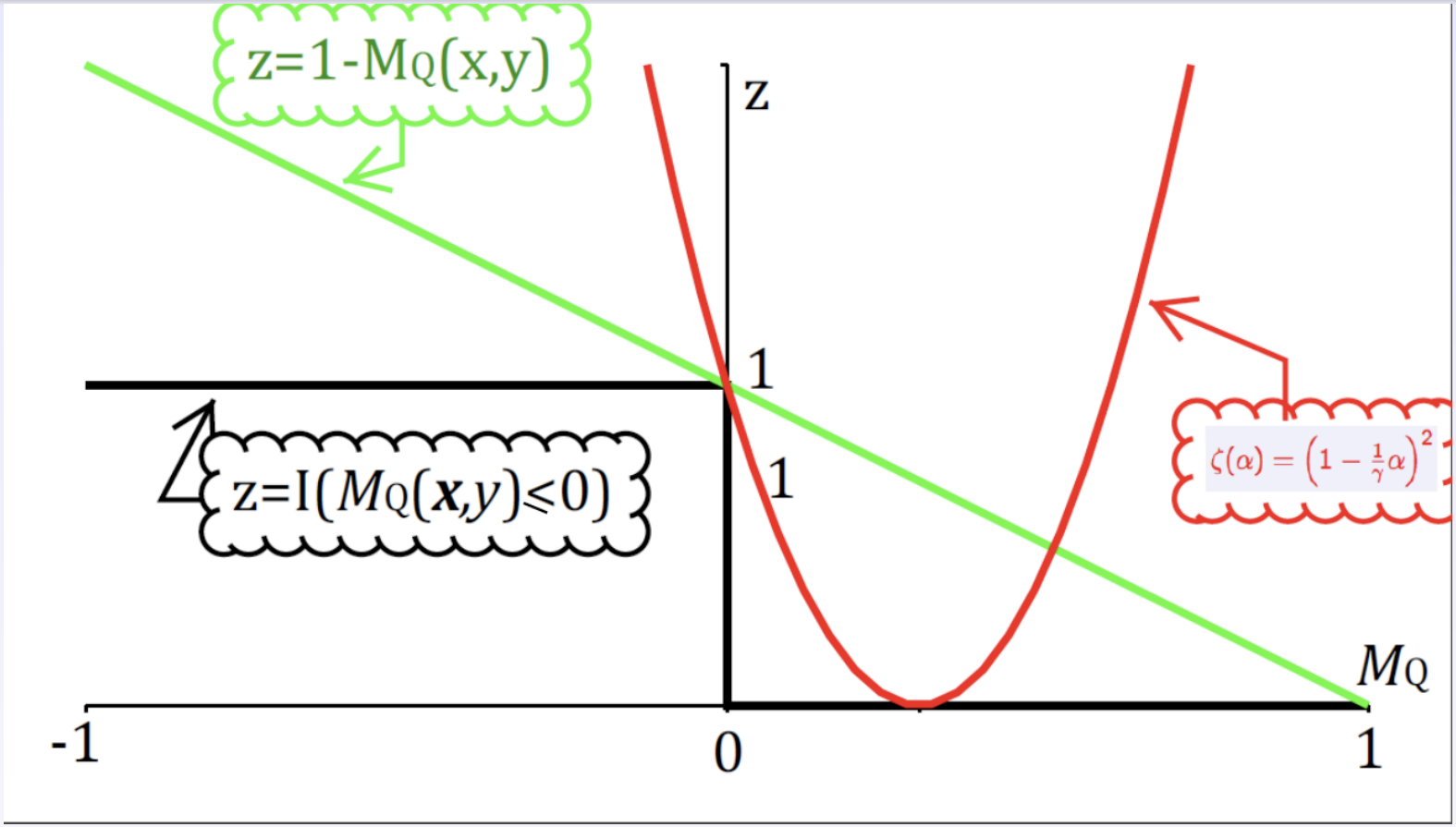
- Consider the margin on an example: $M_Q(\mathbf{x}, y) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim Q} y h(\mathbf{x})$
- and any convex margin loss function $\zeta_Q(\alpha)$ that can be expanded in a Taylor series around $M_Q(\mathbf{x}, y) = 0$:

$$\zeta_Q(M_Q(\mathbf{x}, y)) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k \left(M_Q(\mathbf{x}, y)\right)^k$$

and that upper bounds the risk of the majority vote B_Q, *i.e.*,

$$\zeta_Q(M_Q(\mathbf{x},y)) \geq I(M_Q(\mathbf{x},y)\leq 0) \quad \forall Q,\mathbf{x},y\,.$$

- Conclusion: if we can obtain a PAC-Bayes bound on $\zeta_Q(\mathbf{x}, y)$, we will then have a "new" bound on $R(B_Q)$



Note: $1 - M_Q(\mathbf{x}, y) = 2R(G_Q)$

Thus the green and the black curves illustrate: $R(B_Q) \leq 2R(G_Q)$

Catoni's bound for a general loss

If we define

$$\begin{aligned} \zeta_Q &\stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \zeta_Q(M_Q(\mathbf{x},y)) \\ \widehat{\zeta_Q} &\stackrel{\text{def}}{=} \tfrac{1}{m}\sum_{i=1}^{m} \zeta_Q(M_Q(\mathbf{x}_i,y_i)) \\ c_a &\stackrel{\text{def}}{=} \zeta(1) \\ \bar{k} &= \zeta'(1) \end{aligned}$$

Catoni's bound become :

Theorem 3.2. *For any D, any $\mathcal{H}$, any P of support $\mathcal{H}$, any $\delta \in (0,1]$, any positive real number C', any loss function $\zeta_Q(\mathbf{x},y)$ defined above, we have*

$$\Pr_{S\sim D^m}\left(\forall Q \text{ on } \mathcal{H}\colon \zeta_Q \le g(c_a,C') + \frac{C'}{1-e^{-C'}}\left[\widehat{\zeta_Q} + \frac{2c_a}{mC'}\left[\bar{k}\cdot \mathrm{KL}(Q\|P) + \ln\frac{1}{\delta}\right]\right]\right) \ge 1-\delta\,,$$

where $g(c_a,C') \stackrel{\text{def}}{=} 1 - c_a + \frac{C'}{1-e^{-C'}}\cdot(c_a-1)$.

Answer # 2 (cont)

The trick !

- $\zeta_Q(\mathbf{x}, y)$ can be expressed in terms of the risk on example $(\mathbf{x}, y)$ of a Gibbs classifier described by a *transformed* posterior $\overline{Q}$ on $\mathbb{N} \times \mathcal{H}^\infty$

$$\zeta_Q(M_Q(\mathbf{x}, y)) \;=\; c_a \left[M_{\overline{Q}}(\mathbf{x}, y) \right] ,$$

where $c_a \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} a_k$ and where

$$R_{\{(\mathbf{x},y)\}}(G_{\overline{Q}}) \stackrel{\text{def}}{=} \frac{1}{c_a} \sum_{k=1}^{\infty} |a_k| \mathop{\mathbf{E}}_{h_1 \sim Q} \cdots \mathop{\mathbf{E}}_{h_k \sim Q} I\Big((-y)^k h_1(\mathbf{x}) \ldots h_k(\mathbf{x}) \;=\; -\mathrm{sgn}(a_k) \Big) .$$

- Since $R_{\{(\mathbf{x},y)\}}(G_{\overline{Q}})$ is the expectation of boolean random variable, the Catoni's bound holds if we replace (P, Q) by $(\overline{P}, \overline{Q})$

Minimizing Catoni's bound for a general loss

Minimizing this version of the Catoni's bound is equivalent to finding Q that minimizes

$$f(Q) \stackrel{\text{def}}{=} C \sum_{i=1}^{m} \zeta_Q(\mathbf{x}_i, y_i) + \mathrm{KL}(Q\|P)\,,$$

here: $C \stackrel{\text{def}}{=} C'/(2c_a\overline{k})$.

Minimizing Catoni's bound for a general loss

- To compare the proposed learning algorithms with AdaBoost, we will consider, for $\zeta_Q(\mathbf{x}, y)$, the *exponential loss* given by

$$\exp\Big(-\tfrac{1}{\gamma}\, y \sum_{h\in\mathcal{H}} Q(h)h(\mathbf{x})\Big) \;=\; \exp\Big(\tfrac{1}{\gamma}\,[M_Q(\mathbf{x},y)]\Big)\,.$$

- Because of its simplicity, let us also consider, for $\zeta_Q(\mathbf{x}, y)$, the *quadratic loss* given by

$$\Big(\tfrac{1}{\gamma}\, y \sum_{h\in\mathcal{H}} Q(h)h(\mathbf{x}) - 1\Big)^2 \;=\; \Big(\tfrac{1}{\gamma}\, M_Q(\mathbf{x},y) - 1\Big)^2\,.$$

Empirical results (Nips[09])

Dataset				(1) AdB	(2) RR		(3) KL-EL			(4) KL-QL		
Name	$\|S\|$	$\|T\|$	a	R_T	R_T	C	R_T	C	γ	R_T	C	γ
BreastCancer	343	340	9	0.053	0.050	10	**0.047**	0.1	0.1	**0.047**	0.02	0.4
Liver	170	175	6	0.320	0.309	5	0.360	0.5	0.02	**0.286**	0.02	0.3
Credit-A	353	300	15	0.170	**0.157**	2	0.227	0.1	0.2	0.183	0.02	0.05
Glass	107	107	9	**0.178**	0.206	5	0.187	500	0.01	0.196	0.02	0.01
Haberman	144	150	3	0.260	0.273	100	**0.253**	500	0.2	0.260	0.02	0.5
Heart	150	147	13	0.252	0.197	1	0.211	0.2	0.1	**0.177**	0.05	0.2
Ionosphere	176	175	34	0.120	0.131	0.05	0.120	20	0.0001	**0.097**	0.2	0.1
Letter:AB	500	1055	16	0.010	**0.004**	0.5	0.006	0.1	0.02	0.006	1000	0.1
Letter:DO	500	1058	16	0.036	0.026	0.05	**0.019**	500	0.01	0.020	0.02	0.05
Letter:OQ	500	1036	16	**0.038**	0.045	0.5	0.043	10	0.0001	0.047	0.1	0.05
MNIST:0vs8	500	1916	784	0.008	0.015	0.05	**0.006**	500	0.001	0.015	0.2	0.02
MNIST:1vs7	500	1922	784	0.013	**0.012**	1	0.014	500	0.02	0.014	1000	0.1
MNIST:1vs8	500	1936	784	0.025	0.024	0.2	**0.016**	0.2	0.001	0.031	1	0.02
MNIST:2vs3	500	1905	784	0.047	0.033	0.2	0.035	500	0.0001	**0.029**	0.02	0.05
Mushroom	4062	4062	22	**0.000**	0.001	0.5	**0.000**	10	0.001	**0.000**	1000	0.02
Ringnorm	3700	3700	20	0.043	0.037	0.05	**0.025**	500	0.01	0.039	0.05	0.05
Sonar	104	104	60	0.231	0.192	0.05	0.135	500	0.05	**0.115**	1000	0.1
Usvotes	235	200	16	**0.055**	0.060	2	0.060	0.5	0.1	**0.055**	1000	0.05
Waveform	4000	4000	21	0.085	**0.079**	0.02	0.080	0.2	0.05	0.080	0.02	0.05
Wdbc	285	284	30	0.049	0.049	0.2	**0.039**	500	0.02	0.046	1000	0.1

From $\mathrm{KL}(Q\|P)$ to ℓ_2 regularization

We can recover ℓ_2 regularization if we upper-bound $\mathrm{KL}(Q\|P)$ by a quadratic function.

PAC-Bayes vs Boosting and Ridge regression (cont)

- With this approximation, the objective function to minimize becomes

$$f_{\ell_2}(\mathbf{w}) \;=\; C'' \sum_{i=1}^{m} \zeta\left(\frac{1}{\gamma}\, y_i \mathbf{w}\cdot\mathbf{h}(\mathbf{x}_i)\right) + \|\mathbf{w}\|_2^2\,,$$

  subject to the ℓ_∞ constraint $|w_j| \leq 1/n \;\; \forall j \in \{1,\ldots,n\}$.

- Here $\|\mathbf{w}\|_2$ denotes the Euclidean norm of $\mathbf{w}$ and $\zeta(x) = (x-1)^2$ for the quadratic loss and e^{-x} for the exponential loss.

- If, instead, we minimize f_{ℓ_2} for $\mathbf{v} \stackrel{\text{def}}{=} \mathbf{w}/\gamma$ and remove the ℓ_∞ constraint, we recover *exactly*
  - ridge regression for the quadratic loss case !
  - ℓ_2-regularized boosting for the exponential loss case !!

Answer#2 and kernel methods

- Note that in contrast to the approach Answer#1, the approach Answer#2 can not, as it is presently stated, construct kernel based algorithm.
- For that we need to extend the PAC-Bayes theorem to the sample compression setting (to be submitted to ICML).

Dataset				Rbf kernel				Sigmoid kernel		
Name	$\|T\|$	$\|S\|$	n	SVM	PBSC1	PBSC2	LINEAR	SVM		PBSC2
Adult	10000	1809	14	0.160	**0.157**	**0.157**	0.193	**0.157**		0.158
BreastCancer	340	343	9	0.038	**0.035**	**0.035**	0.144	0.376	(76% n/a)	**0.032**
Credit-A	300	353	15	0.187	0.170	**0.150**	0.200	0.183		**0.143**
Letter:AB	1055	500	16	**0.001**	**0.001**	**0.001**	0.038	0.498	(4% n/a)	**0.130**
Letter:DO	1058	500	16	0.013	**0.009**	**0.009**	0.069	0.490	(19% n/a)	**0.210**
Letter:OQ	1036	500	16	**0.014**	0.017	0.017	0.123	0.488	(10% n/a)	**0.157**
MNIST:0vs8	1916	500	784	0.005	**0.004**	**0.004**	0.031	**0.007**	(1% n/a)	0.011
MNIST:1vs7	1922	500	784	0.012	**0.008**	0.010	0.161	0.015		**0.010**
MNIST:1vs8	1936	500	784	0.013	**0.011**	**0.011**	0.292	**0.024**		0.038
MNIST:2vs3	1905	500	784	0.023	**0.016**	0.018	0.114	**0.029**		0.032
Mushroom	4062	4062	22	**0.000**	**0.000**	**0.000**	0.022	0.280	(66% n/a)	**0.007**
Ringnorm	3700	3700	20	0.015	0.016	**0.012**	0.103	0.056	(26% n/a)	**0.023**
Tic-tac-toe	479	479	9	0.023	**0.015**	**0.015**	0.365	0.365	(61% n/a)	**0.042**
Waveform	4000	4000	21	**0.068**	0.069	**0.068**	0.143	0.108	(54% n/a)	**0.069**
Wdbc	284	285	30	0.070	0.092	**0.067**	0.180	**0.366**		**0.366**

MinCq, another bound minimization algorithm

Definition

Recall that the *Q-margin* realized on an example $(\mathbf{x}, y)$ is :

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

- Now, 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' (D' being either D or S):

$$\mathcal{M}_Q^{D'} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} \mathcal{M}_Q(\mathbf{x}, y) = \mathop{\mathbf{E}}_{h\sim Q} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} y\, h(\mathbf{x})$$

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

- Note that, since $y^2 = 1$, there is no label y in the last equation.

MinCq is based on the following theorem

Theorem

(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 $\mathcal{M}_Q^{D'} \geq 0$ we then have*

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

Proof.

Since $B_Q(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname{sgn}\left[\mathop{\mathbf{E}}_{h\sim Q} h(\mathbf{x})\right]$, B_Q misclassifies an example if its Q-margin is strictly negative and that B_Q classifies it correctly if its Q-margin is strictly positive. Hence, we have $R_{D'}(B_Q) \leq \Pr_{(\mathbf{x},y)\sim D'}(\mathcal{M}_Q(\mathbf{x},y) \leq 0)$. The result follows from the Cantelli-Chebychev's inequality. □

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 training 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 $C_Q^S = 1 - \mathcal{M}_Q^S / \mathcal{M}_{Q^2}^S$, this gives a $0/0$ numerical instability.
- Moreover, since $\mathcal{M}_Q^D / \mathcal{M}_{Q^2}^D$ can only be empirically estimated by $\mathcal{M}_Q^S / \mathcal{M}_{Q^2}^S$, Problem 2, therefore, amplifies Problem 1.

Solution: restricting to quasi-uniform distributions

Definition

Assume that $\mathcal{H}$ is finite and *auto-complemented*, meaning that

$$h_{i+n}(\mathbf{x}) \;=\; -h_i(\mathbf{x}) \qquad \text{for any } \mathbf{x} \in \mathcal{X} \text{ and any } i\,.$$

A distribution Q is *quasi-uniform* if

$$Q(h_i) + Q(h_{i+n}) \;=\; 1/n \qquad \text{for any } i \in \{1, \ldots, n\}\,.$$

Quasi-uniform distributions is a rich family

Proposition

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.,

$$B_{Q'}(\mathbf{x}) = B_Q(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{X} \quad , \quad C^S_{Q'} = C^S_Q \quad \text{and} \quad C^D_{Q'} = C^D_Q \, .$$

Proposition

For all $\mu \in]0, 1]$ and for all quasi-uniform distribution Q on $\mathcal{H}$ having an empirical margin $\mathcal{M}^S_Q \geq \mu$, there exists a quasi-uniform distribution Q' on $\mathcal{H}$, having an empirical margin equal to μ,

$$\mathcal{M}^S_{Q'} = \mu \ , \quad B_{Q'}(\mathbf{x}) = B_Q(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{X} \ , \quad C^S_{Q'} = C^S_Q \quad \text{and} \quad C^D_{Q'} = C^D_Q \, .$$

Quasi-uniform distributions have a nice PAC-Bayes property ... they have no $\mathrm{KL}(Q||P)$!!!

Theorem

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

$$\Pr_{S\sim D^m}\left(\begin{array}{l} \text{For all quasi-uniform distribution } Q \text{ on } \mathcal{H}, \text{ we have :} \\ \mathcal{M}_Q^S - \frac{2B\sqrt{\ln\frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}} \leq \mathcal{M}_Q^D \leq \mathcal{M}_Q^S + \frac{2B\sqrt{\ln\frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}} \\ \text{and } \mathcal{M}_{Q^2}^S - \frac{2B^2\sqrt{\ln\frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}} \leq \mathcal{M}_{Q^2}^D \leq \mathcal{M}_{Q^2}^S + \frac{2B^2\sqrt{\ln\frac{2\sqrt{m}}{\delta}}}{\sqrt{2m}} \end{array}\right) \geq 1-\delta$$

The algorithm MinCq

Definition

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 $\mathcal{M}_{Q^2}^S$.

- MinCq is a quadratic program

Empirical results

Dataset				AdaBoost	minCq-stumps		SVM			minCq-RBF		
Name	$\|S\|$	$\|T\|$	#feat.	$R_T(B_Q)$	$R_T(B_Q)$	Margin	$R_T(B_Q)$	C	γ	$R_T(B_Q)$	Margin	γ
Adult	1809	10000	14	**0.149**	0.152	0.04	0.159	100	0.03571	**0.157**	0.001	0.14286
BreastCancer	343	340	9	0.053	**0.050**	0.01	**0.038**	0.5	0.00347	0.044	0.01	0.00113
Credit-A	353	300	15	0.170	**0.157**	0.04	0.183	500	0.00833	**0.143**	0.02	0.30000
Glass	107	107	9	0.178	**0.168**	0.04	0.178	2	0.50000	**0.168**	0.02	2.00000
Haberman	144	150	3	0.260	**0.253**	0.02	**0.280**	0.02	0.00340	**0.280**	0.02	0.04166
Heart	150	147	13	0.259	**0.224**	0.05	**0.197**	1	0.15385	**0.197**	0.01	0.15385
Ionosphere	176	175	34	**0.120**	0.143	0.01	0.097	10	0.13235	**0.029**	0.0005	0.23529
Letter:AB	500	1055	16	0.010	**0.002**	0.05	**0.001**	0.02	0.28125	0.002	0.0005	0.12500
Letter:DO	500	1058	16	0.036	**0.023**	0.05	0.014	20	0.00781	**0.009**	0.0005	0.03125
Letter:OQ	500	1036	16	**0.038**	0.043	0.04	0.015	4	0.03125	**0.012**	0.001	0.03125
Liver	170	175	6	**0.320**	0.331	0.01	**0.314**	5	0.00130	**0.314**	0.01	0.00232
MNIST:08	500	1916	784	0.008	0.016	0.0001	**0.003**	2	0.01594	**0.003**	0.0001	0.03125
MNIST:17	500	1922	784	0.013	**0.012**	0.05	0.011	5	0.01020	**0.007**	0.0005	0.00574
MNIST:18	500	1936	784	0.025	**0.025**	0.03	**0.011**	1	0.04082	**0.011**	0.0005	0.03125
MNIST:23	500	1905	784	0.047	**0.033**	0.04	0.020	5	0.02296	**0.016**	0.0005	0.01594
Mushroom	4062	4062	22	**0.000**	**0.000**	0.02	**0.000**	10	0.02273	**0.000**	0.0001	0.09091
Sonar	104	104	60	0.231	**0.144**	0.05	0.163	2	0.40833	**0.135**	0.0001	0.40833
Tic-tac-toe	479	479	9	0.357	**0.344**	0.05	0.081	10	0.22222	**0.017**	0.0001	0.22222
Usvotes	235	200	16	**0.055**	**0.055**	0.02	**0.055**	5	0.03125	0.065	0.02	0.03125
Wdbc	285	284	30	**0.049**	0.053	0.04	0.074	0.5	0.00026	**0.067**	0.02	0.00026

Conclusion

- Theorem 1, being relatively simple, represents a good starting point for an introduction to PAC-Bayes theory
- Again because of its simplicity, it represents an interesting tool for developping new PAC-Bayes bounds (not necessary in binary classification under the iid assumption).
- Up to some convex relaxation PAC-Bayes rediscovers existing algorithms,
  - this is nice
  - and should be interesting for other paradigms than iid supervised learning, where our knowledge is not as "extended".

Conclusion

- Minimizing PAC-Bayes bounds seems to produce performing algorithms !!!
- but these algorithms nevertheless need to have some parameter to be tune via cross-validation in order to perform as well as the state of the art
  - Why this is so ?
  - Possibly because the loss of those bounds are only based on the margin
  - The U-statistic involved here is therefore of order one,
    - what if we consider higher order ?
    - Note: PAC-Bayes bound of U-statistic of high orders will be in a non iid setting

QUESTIONS ?

Suggestion de lectures

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. A pac-bayes risk bound for general loss functions. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 449–456. MIT Press, Cambridge, MA, 2007.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Sara Shanian. From PAC-Bayes bounds to KL regularization. In J. Lafferty and C. Williams, editors, *Advances in Neural Information Processing Systems 22 (accepted)*, page accepted, Cambridge, MA, 2009. MIT Press.

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. 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.