

SETTING AND MOTIVATION

← **Previously on Part I:** Pascal introduces the \mathcal{C} -bound and PAC-Bayesian theorems that bound the risk of the majority vote classifier.

Theorem (C-bound) For any distribution Q on a set of voters and any distribution D on $\mathcal{X} \times \{-1, 1\}$, if $\mu_1(M_Q^D) > 0$, we have

$$R_D(B_Q) \leq C_Q^D \stackrel{\text{def}}{=} 1 - \frac{(\mu_1(M_Q^D))^2}{\mu_2(M_Q^D)}.$$

Minimizing the \mathcal{C} -bound favors majority votes for which the voters are maximally uncorrelated. **We want an algorithm that minimizes the related PAC-Bayesian upper bounds** (see Part I). However, empirical experiments shown that $\text{KL}(Q||P)$ is a poor regularizer in this case, as its empirical value tends to be overweighted in comparison with the empirical value of the \mathcal{C} -bound (i.e., C_Q^S). **Let's get rid of it!**

ALIGNED DISTRIBUTIONS

Self-complemented set of voters A set of voters \mathcal{H} is said to be *self-complemented* if there exists a bijection $c : \mathcal{H} \rightarrow \mathcal{H}$ such that for any $f \in \mathcal{H}$, $c(f) = -f$. In other words, *each voter $f \in \mathcal{H}$ has a complement $-f \in \mathcal{H}$.*

Aligned posterior Given a self-complemented set \mathcal{H} , a posterior distribution Q is *aligned* on a prior distribution P if

$$Q(f) + Q(c(f)) = P(f) + P(c(f)), \quad \forall f \in \mathcal{H}.$$

Quasi-uniform distribution If P is a uniform prior distribution and Q is aligned on P , we say that Q is a *quasi-uniform* distribution.

L_∞ -norm regularization For a self-complemented set \mathcal{H} of $2n$ voters and a quasi-uniform distribution Q on \mathcal{H} , the weight of any voter is lower-bounded by 0 and upper-bounded by $\frac{1}{n}$, as for any $i \in \mathcal{H}$, $Q(f_i) + Q(f_{i+n}) = \frac{1}{n}$.

A NEW CHANGE OF MEASURE INEQUALITY

A key step of PAC-Bayesian proofs is the *change of measure inequality* (Banerjee [2006], Seldin and Tishby [2010]):

Lemma (Change of measure inequality) For any set \mathcal{H} , for any distribution P and Q on \mathcal{H} , and for any measurable function $\phi : \mathcal{H} \rightarrow \mathbb{R}$, we have

$$\mathbf{E}_{f \sim Q} \phi(f) \leq \text{KL}(Q||P) + \ln \left(\mathbf{E}_{f \sim P} e^{\phi(f)} \right).$$

Using aligned posteriors and a suitable ϕ function, the KL term does not appear.

Theorem (Change of measure inequality for aligned posteriors) For any self-complemented set \mathcal{H} , for any distribution P on \mathcal{H} , any distribution Q aligned on P , and for any measurable function $\phi : \mathcal{H} \rightarrow \mathbb{R}$ such that $\phi(f) = \phi(c(f))$ for all $f \in \mathcal{H}$, we have

$$\mathbf{E}_{f \sim Q} \phi(f) \leq \ln \left(\mathbf{E}_{f \sim P} e^{\phi(f)} \right).$$

Proof.

$$\begin{aligned} 2 \cdot \mathbf{E}_{f \sim Q} \phi(f) &= \int_{\mathcal{H}} df Q(f) \phi(f) + \int_{\mathcal{H}} df Q(c(f)) \phi(c(f)) \\ &= \int_{\mathcal{H}} df Q(f) \phi(f) + \int_{\mathcal{H}} df Q(c(f)) \phi(f) \\ &= \int_{\mathcal{H}} df (Q(f) + Q(c(f))) \phi(f) \\ &= \int_{\mathcal{H}} df (P(f) + P(c(f))) \phi(f) = \dots = 2 \cdot \mathbf{E}_{f \sim P} \phi(f). \end{aligned}$$

The result is obtained by changing the expectation over Q to an expectation over P , and then by applying Jensen's inequality:

$$\mathbf{E}_{f \sim Q} \phi(f) = \mathbf{E}_{f \sim P} \phi(f) = \mathbf{E}_{f \sim P} \ln e^{\phi(f)} \leq \ln \left(\mathbf{E}_{f \sim P} e^{\phi(f)} \right). \quad \square$$

PAC-BAYESIAN BOUNDS REVISITED

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

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

Corollary If \mathcal{H} is a *self-complemented* set of voters, we have

$$\Pr_{S \sim D^m} \left(\text{For all posteriors } Q \text{ aligned on } P : \mu_1(M_Q^D) \geq \mu_1(M_Q^S) - \sqrt{\frac{2}{m} \left[\ln \frac{2\sqrt{m}}{\delta} \right]} \right) \geq 1 - \delta.$$

A PAC-BAYESIAN BOUND WITHOUT KL

PAC-Bound 3 For any distribution D on $\mathcal{X} \times \{-1, 1\}$, for any self-complemented set \mathcal{H} of voters $\mathcal{X} \rightarrow [-1, 1]$, for any prior distribution P on \mathcal{H} , and any $\delta \in (0, 1]$, we have, with probability $1 - \delta$ over the choice of $S \sim D^m$

$$\forall Q \text{ aligned on } P : R_D(B_Q) \leq 1 - \frac{\left(\max \left(0, \mu_1(M_Q^S) - \sqrt{\frac{2}{m} \left[\ln \frac{2\sqrt{m}}{\delta/2} \right]} \right) \right)^2}{\min \left(1, \mu_2(M_Q^S) + \sqrt{\frac{2}{m} \left[\ln \frac{2\sqrt{m}}{\delta/2} \right]} \right)}.$$

PROBLEM 1: MINIMIZING C_Q^S TENDS TO OVERFIT

Solution We restrict Q to be *quasi-uniform*. It **does not reduce the set of possible majority votes**: different distributions Q that give rise to a same majority vote have the same (real and empirical) \mathcal{C} -bound values.

Theorem Let \mathcal{H} be a self-complemented set. 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 \mathcal{C} -bound values, i.e.,

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

PROBLEM 2: 0/0 NUMERICAL INSTABILITY

Distribution Q minimizing the \mathcal{C} -bound are such that both $\mu_1(M_Q^S)$ and $\mu_2(M_Q^S)$ are very close to zero.

Solution We want to constraint $\mu_1(M_Q^S)$ to be higher to some threshold value μ , which is equivalent to **constraint the margin to be exactly equal to μ** : among all the posteriors Q that minimize the \mathcal{C} -bound, there is always one whose empirical margin $\mu_1(M_Q^S)$ is as close to 0 as we want.

Theorem Let \mathcal{H} be a self-complemented set. For all $\mu \in (0, 1]$ and for all quasi-uniform distributions Q on \mathcal{H} having an empirical margin $\mu_1(M_Q^S) \geq \mu$, there exists a quasi-uniform distribution Q' on \mathcal{H} , having an empirical margin equal to μ , such that Q and Q' induce the same majority vote and have the same empirical and true \mathcal{C} -bound values, i.e.,

$$\mu_1(M_{Q'}^S) = \mu, \quad B_{Q'} = B_Q, \quad C_{Q'}^S = C_Q^S \quad \text{and} \quad C_{Q'}^D = C_Q^D.$$

MINCQ: THE ALGORITHM MINIMIZING THE \mathcal{C} -BOUND

MinCq Algorithm Given a self-complemented set \mathcal{H} of $2n$ voters, a training set S , and a S -realizable $\mu > 0$, among all *quasi-uniform distributions Q of empirical margin $\mu_1(M_Q^S)$ exactly equal to μ* , the algorithm MinCq consists in finding one that **minimizes $\mu_2(M_Q^S)$** .

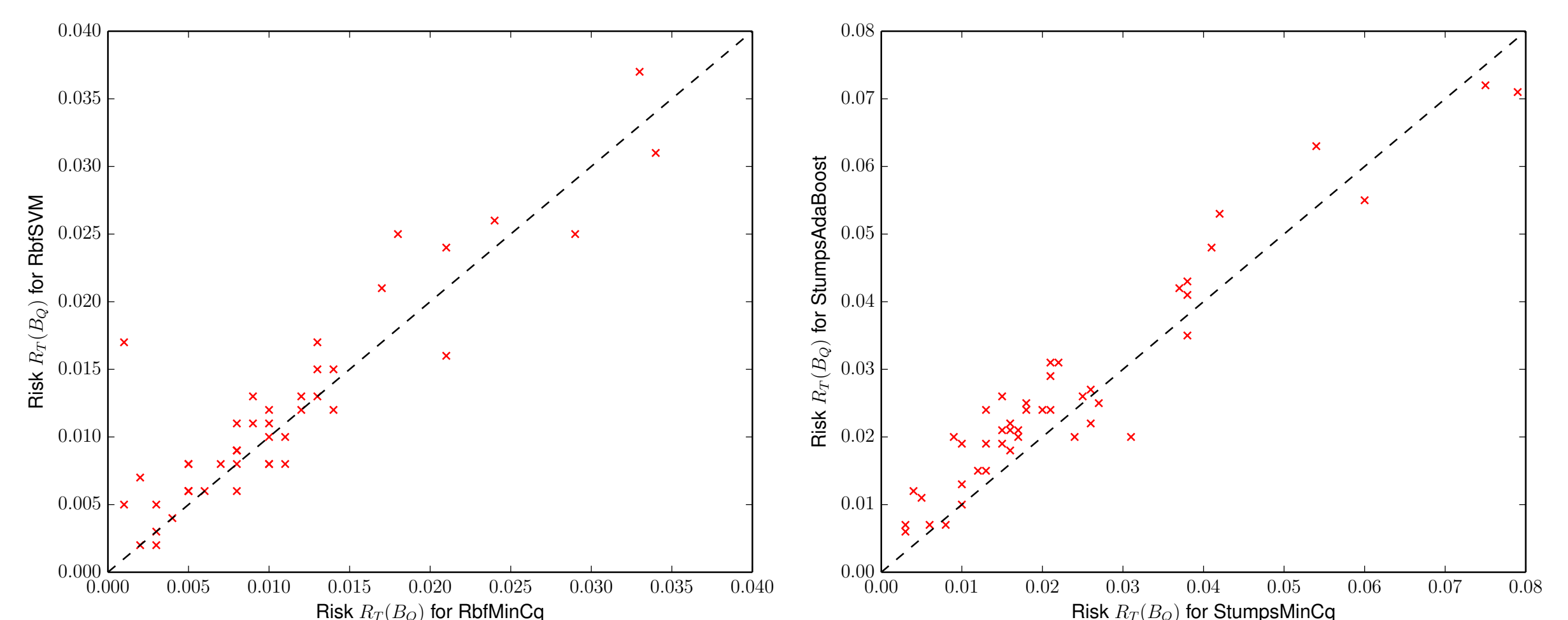
Can be translated as a simple quadratic program (QP) that has only n variables.

EMPIRICAL RESULTS

We compare MinCq using decision stumps as voters against AdaBoost, and using *kernel voters* against SVM.

A **kernel voter** is defined as $f_i(\cdot) = y_i k(x_i, \cdot)$ for some example (x_i, y_i) and kernel function k . One can interpret its output as a confidence level.

Handwritten Digits Recognition (MNIST)



Sentiment Analysis (Amazon Reviews)

Dataset information			Risk $R_T(B_Q)$ for each algorithm	
Name	S	T	LinearMinCq	LinearSVM
Books	2000	4465	0.158	0.158
DVD	2000	3586	0.162	0.163
Kitchen	2000	5945	0.130	0.131
Electronics	2000	5681	0.116	0.118

Statistical comparison on MNIST, Amazon and 31 UCI datasets.

	MNIST		Amazon	UCI	
	RBF Kernel	Stumps	Linear Kernel	RBF Kernel	Stumps
Poisson binomial test*	88%	99%	68%	54%	48%
Sign test (p-value)*	.01	.00	0.31	0.36	0.35

* For the Poisson binomial test, a higher value is better. For the sign test, a lower value is better.