

# **L<sub>q</sub>-regularised sparse classifiers: A PAC-Bayes analysis**

**Ata Kabán**

School of Computer Science  
The University of Birmingham  
Birmingham B15 2TT, UK  
<http://www.cs.bham.ac.uk/~axk>

Workshop on Sparsity in Machine learning and Statistics  
Cumberland Lodge, 1-3 April, 2009.

# Introduction / Background

L1-regularisation - a workhorse in both machine learning and compressive sensing

- sparsity
- convexity
- recovers the L0 solution in some cases

Lq regularisation is non-convex - but seems to work better

- statistics (Fan & Li, '01): oracle property
- compressive sensing (Chartrand, '07)
- signal denoising (Moulin, '99)

How about learning / generalisation?

- what  $q$  to use in which case?
- when is the smaller  $q$  better?

So far:

- 0-norm SVM classification (Weston et al., '03) (results data-dependent)
- genomic data classification report better prediction using  $L_q$  (Liu et al., '07)

- we empirically found the smaller  $q$  works better when many features are irrelevant (Kaban & Durrant,'08a)
- derived a data-independent PAC-bound — somewhat informative but far too loose
- also noted that as the data dimensionality increases,  $L_q$ -regularisation with a smaller  $q$  falls pray to the curse of dimensionality at a slower rate (Kaban & Durrant,'08b)

Here we derive a data-dependent PAC-Bayes bound, to better understand the generalisation behaviour of  $L_q$ -regularised classifiers.

## Lq-regularised logistic regression

- Training set  $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$ , where  $\mathbf{x}_j \in \mathbb{R}^m$  inputs and  $y_j \in \{-1, 1\}$  their labels.
- Lq-regularised logistic regression:

$$\max_{\mathbf{w}} \sum_{i=1}^n -\log(1 + \exp(-y\mathbf{w}^T \mathbf{x}_i)) - \lambda \|\mathbf{w}\|_q^q \quad (1)$$

where  $\|\mathbf{w}\|_q = (\sum_{i=1}^m |w_i|^q)^{1/q}$ .

$q \leq 1$ : non-differentiable at zero  $\Rightarrow$  sparsity.

- To derive a PAC-Bayes bound, interpret the regularisation term as the log prior of a Generalised Gaussian Distribution (GGD), which has the following form:

$$GGD(w_j | \mu_j, \lambda, q) = \frac{q\lambda^{1/q}}{2\Gamma(1/q)} \exp\{-\lambda|w_j - \mu_j|^q\} \quad (2)$$

where  $\lambda > 0$  and we are interested in  $q \in (0, 1]$ .

## PAC-Bayes Theorem for $L_q$ -reg. logistic regression.

For all i.i.d. distributions  $D$  over the data,  $\forall P(h)$  prior over the classifiers  $h : \mathbf{X} \rightarrow Y$ , and  $\forall \delta \in (0, 1]$  risk,

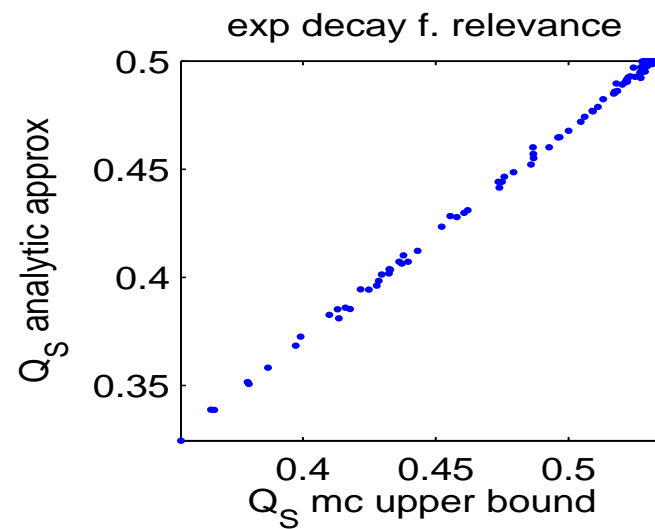
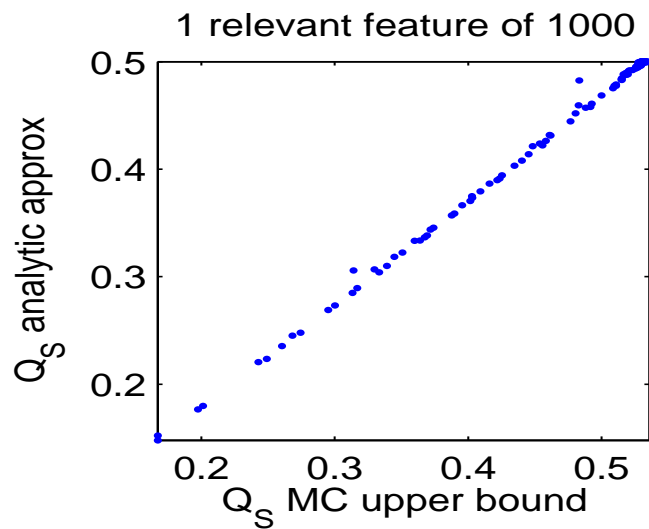
$$\Pr_{S \sim D^n} \left\{ \forall Q(h) : KL_+[ \hat{Q}_S || Q_D ] \leq \frac{1}{n} [KL(Q || P) + \ln \frac{n+1}{\delta}] \right\} \geq 1 - \delta \quad (3)$$

where:

- $Q := GGD(w_j | \hat{w}_j, \tilde{\lambda}, \tilde{q})$ , with  $\tilde{\lambda}, \tilde{q}$  chosen to tighten the bound
- $\hat{Q}_S \equiv \mathbb{E}_{h \sim Q} [\frac{1}{n} \sum_{i=1}^n I(h(\mathbf{x}_i) \neq y_i)]$  (the expected training error) and  $KL(Q || P) = \mathbb{E}_{h \sim Q} \ln \frac{Q(h)}{P(h)}$  are specific to the  $L_q$  regularised logistic regression model.
- $Q_D \equiv Pr(\mathbf{x}, y) \sim D \{h(\mathbf{x}) \neq y\}$  is the true (generalisation) error, which is solved numerically from (3).

- *The  $\hat{Q}_S$  term:* MC-based upper-bound is one option; alternatively:

$$\hat{Q}_S \approx \frac{1}{n} \sum_{i=1}^n \Phi \left\{ -\frac{y_i \hat{\mathbf{w}}^T \mathbf{x}_i}{\tilde{\lambda}^{-1/\tilde{q}} \sqrt{\frac{\Gamma(3/\tilde{q})}{\Gamma(1/\tilde{q})}} \|\mathbf{x}_i\|_2} \right\} \quad (4)$$



- *The  $KL(Q||P)$  term:  $\sum_{j=1}^m KL[Q(w_j)||P(w_j)]$*   
This is intractable, but using the  $q$ -triangle inequality (valid for  $q \leq 1$ ) yields a tractable upper-bound:

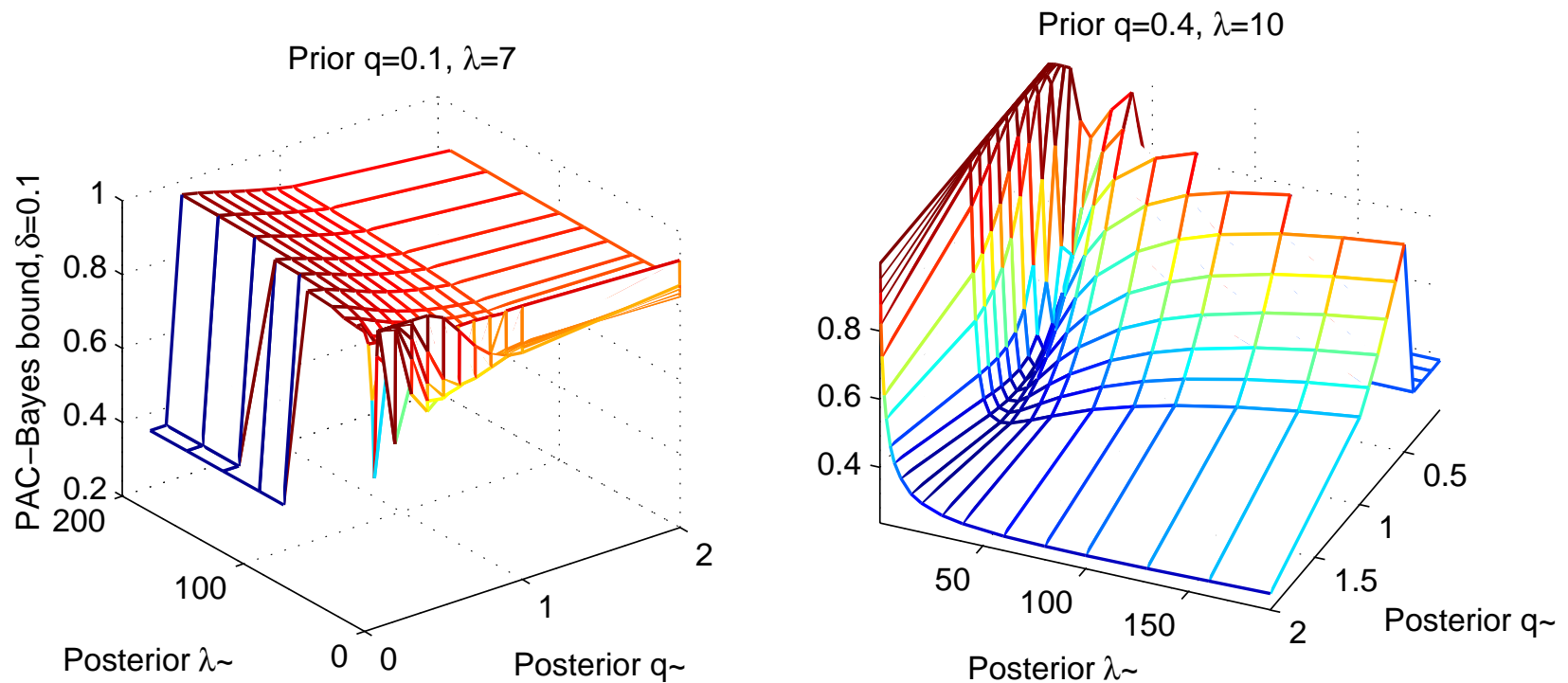
$$\begin{aligned}
 KL[Q(w_j)||P(w_j)] &\leq KL[GGD(w_j|0, \tilde{\lambda}, \tilde{q})||GGD(w_j|0, \lambda, q)] + \lambda|\hat{w}_j|^q \quad (5) \\
 &= \log \frac{\tilde{q}\tilde{\lambda}^{1/\tilde{q}}\Gamma(1/q)}{q\lambda^{1/q}\Gamma(1/\tilde{q})} + \frac{\lambda}{\tilde{\lambda}^{q/\tilde{q}}} \frac{\Gamma((q+1)/\tilde{q})}{\Gamma(1/\tilde{q})} - \frac{1}{\tilde{q}} + \lambda|\hat{w}_j|^q
 \end{aligned}$$

where '=' holds whenever  $\hat{w}_j = 0$ .

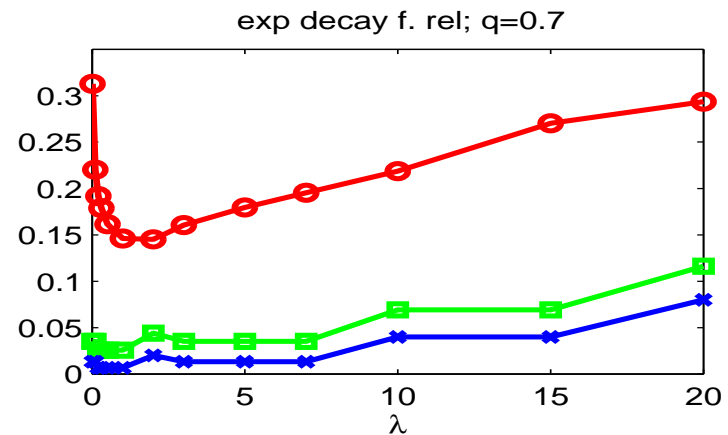
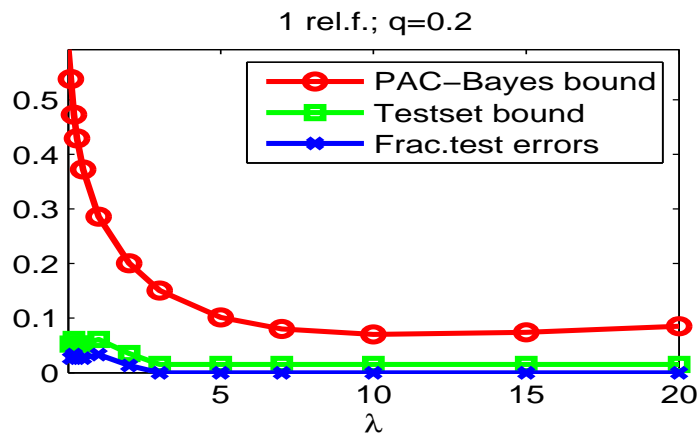
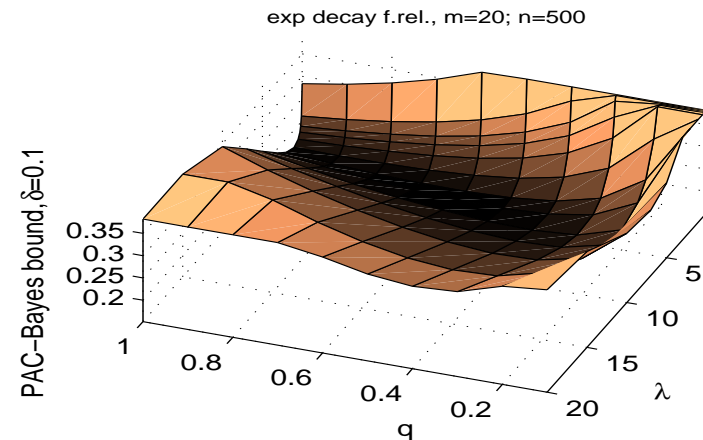
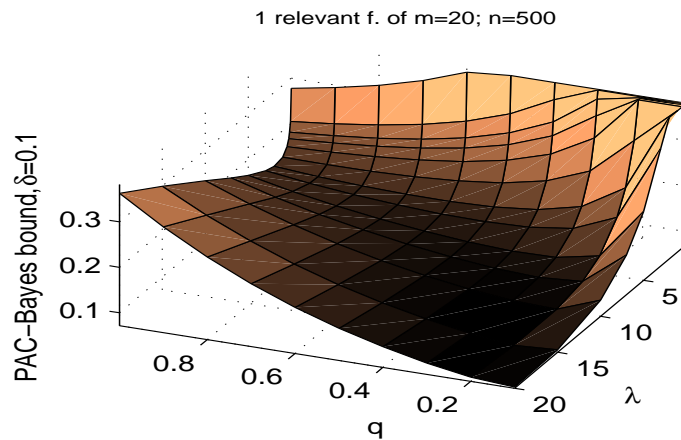
Interesting to observe that for small  $\hat{Q}_S$ , we have:

$$\min_{\tilde{\lambda}, \tilde{q}} KL[Q||P] \leq \lambda ||\hat{\mathbf{w}}||_q^q \quad (6)$$

# The PAC-Bayes bounds at work

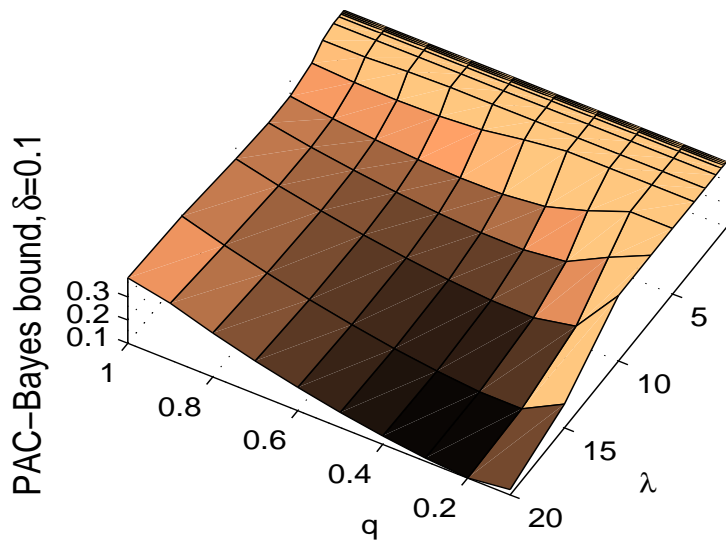


Examples of searching for the parameters of  $Q$  to optimise the generalisation bound. Right: 1 relevant feature of 200; Left: exp decay feature relevance.  $n = 500$ .

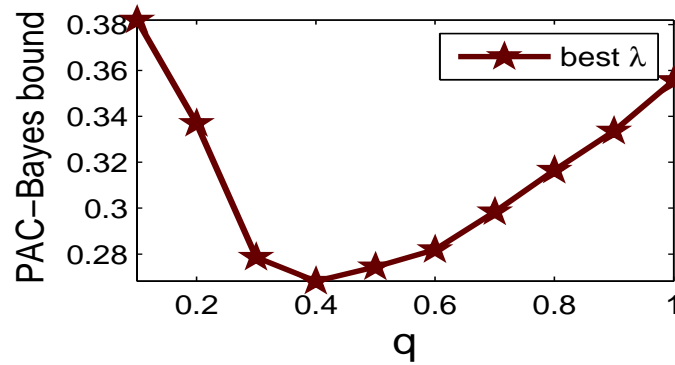
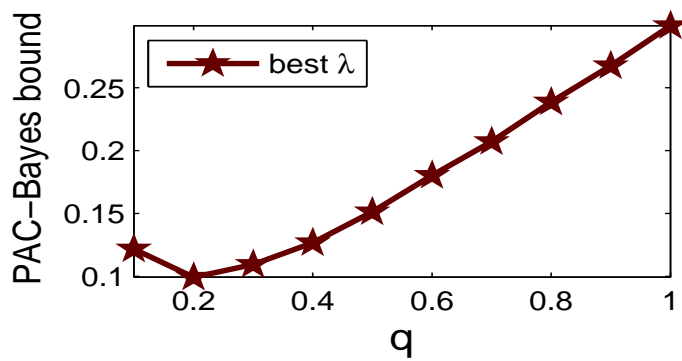
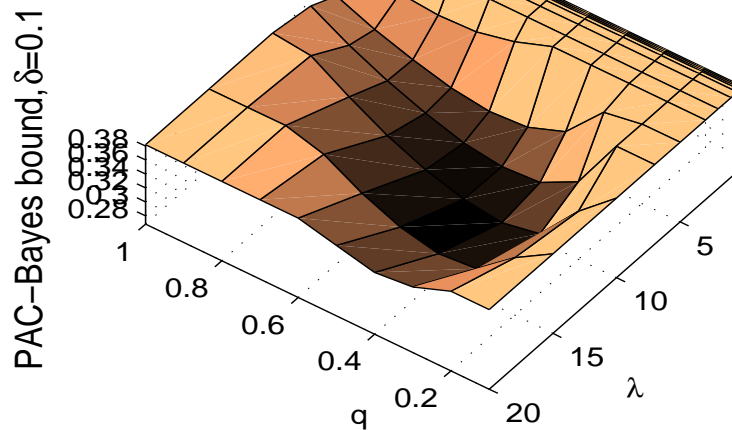


Toy-example with  $m = 20$  features and  $m = 500$  samples. Best  $q$  compared against a test set bound and the actual fraction of test error counts. For the latter, 30% of the data was held out for testing.

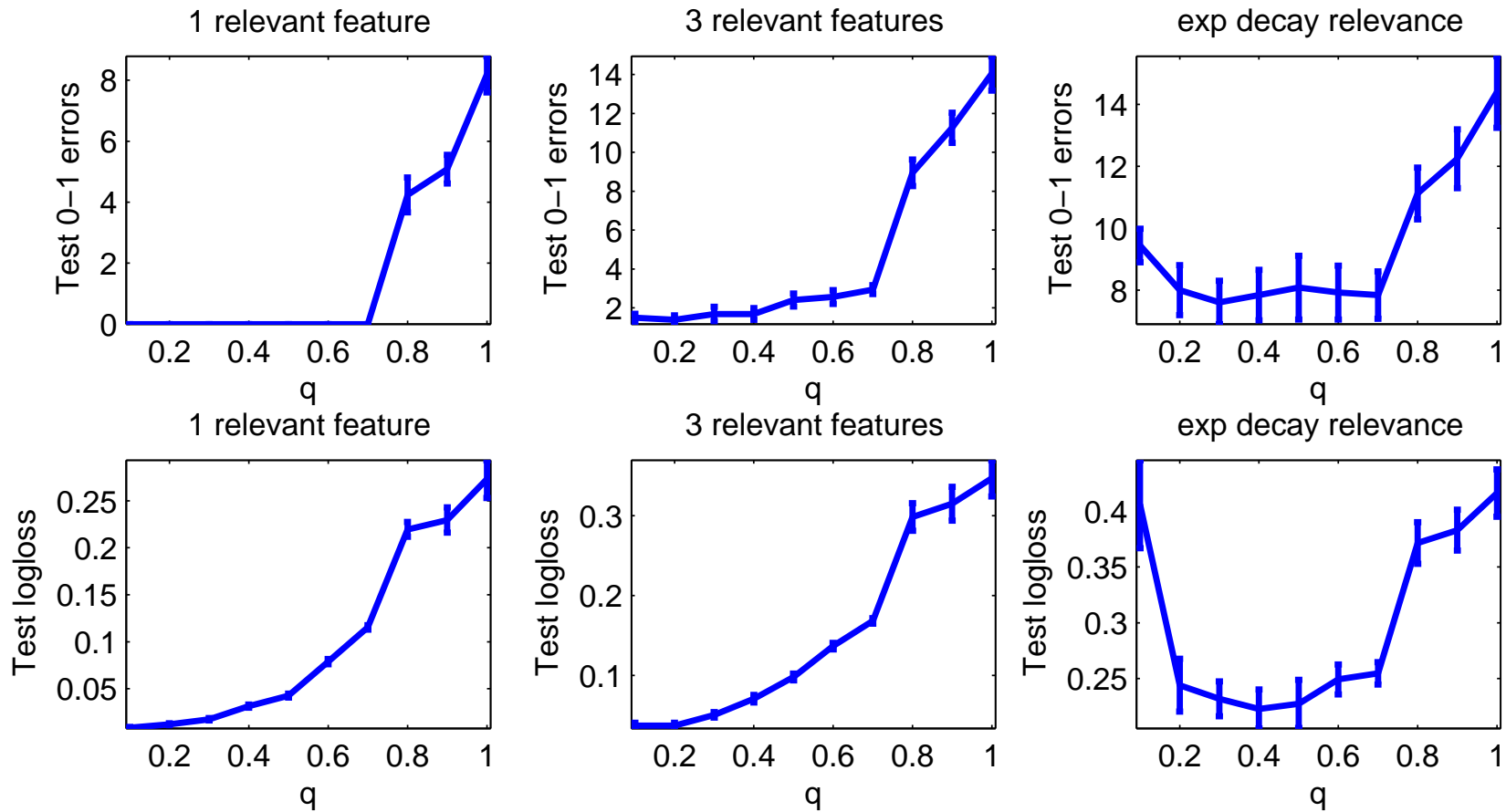
1 relevant f. of  $m=200$ ;  $n=500$



exp decay f. rel,  $m=200$ ;  $n=500$



PAC-Bayes analysis on synthetic data with  $m = 200$  dimensions,  $n = 500$  samples.



Empirical behaviour, selecting  $\lambda$  by cross-validation. Test errors refer to counts out of 100. The error bars are over dimensionality ranging from 10 to 1000.

# Conclusions

- Derived PAC-Bayes bound for studying the generalisation ability of the family of sparse classifiers expressed by  $L_{q \leq 1}$ -regularised logistic regression.
- $L_{q < 1}$  improves over  $L_1$ -regularisation when many features are irrelevant (and i.i.d.). This is when more sparsity helps.
- The PAC-Bayes bound seems well aligned with the empirical behaviour of  $L_q$ -regularised sparse classifiers, when the sample size is not excessively small.
- The optimisation involved in tuning the parameters of  $Q$  appears to become more difficult as  $q$  gets smaller. Here we used a simple grid search, but a more sophisticated optimiser may bring further improvements in the tightness of the bounds.
- Further work to experiment with correlated features and real-world data sets.

## References

- B. Krishnapuram, L. Carin, M. Figueiredo, A. Hartemink. Learning sparse Bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds. IEEE PAMI, vol. 27, no. 6, pp. 957-968, 2005.
- A. Kabán and R.J. Durrant. Learning with  $L_{q<1}$  vs.  $L_1$  regularization in exponentially many irrelevant features. Proc. ECML'08.
- A. Kabán and R.J. Durrant. A norm-concentration argument for non-convex regularization. ICML/UAI/COLT Workshop on Sparse Optimization and Variable Selection, 9 July, 2008, Helsinki, Finland.
- J. Langford. Tutorial on practical prediction theory for classification. JMLR, 2005.
- D. McAllester. PAC-Bayesian theorems. Machine Learning, 37(3):355-363, 1999.
- M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian Process classification. JMLR, 2003.