

Learning with $L_{q<1}$ vs L_1 -norm regularisation with exponentially many irrelevant features

Ata Kabán

Robert J. Durrant

School of Computer Science
The University of Birmingham
Birmingham B15 2TT, UK

ECML'08, 15-19 September 2008, Antwerp, Belgium

Roadmap

- Intro: $L_{q < 1}$ -norm regularised logistic regression
- Analysis: generalisation & sample complexity
- Implementation
- Experiments & results
- Conclusions

Intro: $L_{q<1}$ -regularised logistic regression

Training set $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$, with $\mathbf{x}_j \in \mathbb{R}^m$ the inputs and $y_j \in \{-1, 1\}$ their labels.

Scenario of interest: few $r \ll m$ relevant features, small sample size $n \ll m$.

Consider regularised logistic regression for concreteness:

$$\max_{\mathbf{w}} \sum_{j=1}^n \log p(y_j | \mathbf{x}_j, \mathbf{w}) \text{ subject to } \|\mathbf{w}\|_q \leq A \quad (1)$$

where $p(y | \mathbf{w}^T \mathbf{x}) = 1 / (1 + \exp(-y \mathbf{w}^T \mathbf{x}))$, and $\|\mathbf{w}\|_q = (\sum_{i=1}^m |w_i|^q)^{1/q}$.

- if $q = 2$: L2-regularised ('ridge')
- if $q = 1$: L1-regularised ('lasso')
- if $q < 1$: $L_{q<1}$ -regularised : non-convex, non-differentiable at 0

Intro: $L_{q<1}$ -regularised logistic regression

L1-regularisation - a workhorse in machine learning

- sparsity
- convexity
- sample complexity logarithmic in m (Ng, ICML'04) [L2-reg's is linear]

Non-convex norm regularisation seems to have added value

- statistics (Fan & Li, '01): oracle property
- signal reconstruction (Wipf & Rao, '05)
- compressed sensing (Chartrand, '08)
- 0-norm SVM classification (Weston et al., '03) - but results are data-dependent
- genomic data classification (Liu et al., '07) - used with data resampling

Question: When is $L_{q<1}$ -norm regularisation superior for Machine Learning?
i.e. in terms of generalisation ability & sample complexity

Analysis: Sample complexity bound

$H = \{h(\mathbf{x}, y) = -\log p(y|\mathbf{w}^T \mathbf{x}) : \mathbf{x} \in \mathcal{R}^m, y \in \{-1, 1\}\}$ the function class

$er_P(h) = E_{(\mathbf{x}, y) \sim \text{iid} P}[h(\mathbf{x}, y)]$ the true error of h

$\hat{er}_z(h) = \frac{1}{n} \sum_{j=1}^n h(\mathbf{x}_j, y_j)$ the sample error of h on training set z of size n

$opt_P(H) = \inf_{h \in H} er_P(h)$ the approximation error of H

$L(z) = \min_{h \in H} \hat{er}_z(h)$ the function returned by the learning algorithm

Theorem (extending a result of A.Ng, '04 from L_1 to $L_{q < 1}$).

$\forall \epsilon > 0, \forall \delta > 0, \forall m, n \geq 1$, in order to ensure that

$er_P(L(z)) \leq opt_P(H) + \epsilon$ with probability $1 - \delta$, it is enough to have the training set size:

$$n = \Omega((\log m) \times \text{poly}(A, r^{1/q}, 1/\epsilon, \log(1/\delta))) \quad (2)$$

- logarithmic in the data dimensionality m ;
- polynomial in #relevant features, but growing with $r^{1/q}$

Analysis: Comments

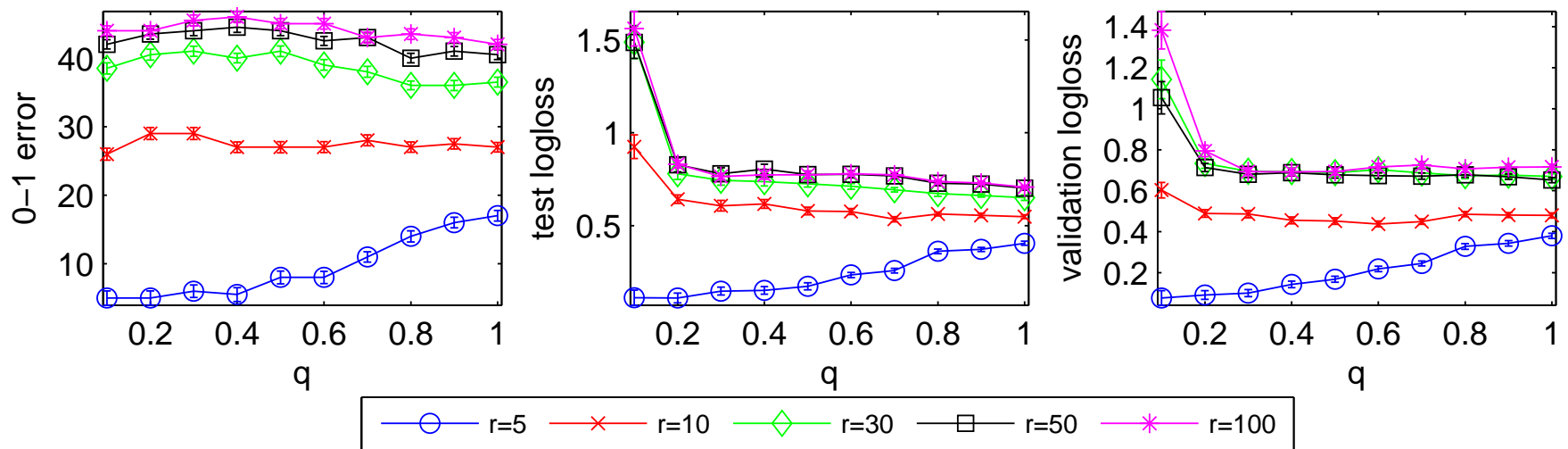
The sample complexity bound derived is an upper-bound on the logloss error, and consequently also on the 0-1 error.

Quantitatively quite loose.

However, it gives some insight into the behaviour of L_q -regularised logistic regression

- for smaller q the sample complexity grows faster in r
 - so, a small q is more advantageous for the small r case (rather than the large r case)
 - this is also intuitive: we may expect the sparsity-inducing effect is not beneficial in such a case (under-fitting)
- given the v small r case, it doesn't say if the smaller q is actually better, though experiments and some existing other kinds of analyses indicate so.

Analysis: Empirical check



Experiments on $m = 200$ dimensional data sets, varying the number of relevant features $r \in \{5, 10, 30, 50, 100\}$. The medians of 60 independent trials are shown and the error bars represent one standard error. The 0-1 errors are out of 100.

Implementation

Using the Lagrangian form of the objective. Maximise w.r.t. \mathbf{w} :

$$\mathcal{L} = - \sum_{j=1}^n \log \left\{ 1 + \exp(-y_j \mathbf{w}^T \mathbf{x}_j) \right\} - \alpha \sum_{i=1}^m |w_i|^q \quad (3)$$

Implementation is non-trivial since the objective function is not convex and not differentiable at zero. We discuss 3 methods:

- Method 1: Smoothing at zero
- Method 2: Local quadratic variational bounds
- Method 3: Local linear variational bounds

Methods 2 and 3 turn out to be equivalent in terms of finding a local optimum of the same objective.

Method 1: Using a smooth approximation

Proposed in (Liu et al., Stat App to Gen & Mol. Biol, '07)

$$\sum_{i=1}^m |w_i|^q \approx \sum_{i=1}^m (w_i^2 + \gamma)^{q/2} \quad (4)$$

where γ is set to a "small" value.

Pro: Easy to implement, any nonlinear optimisation method applies

Con: no good way to set γ

- if γ is too small, numerically unstable, specially when q is also small
- if γ is larger, over-smoothing, losing the sparsity effect

Con: No guarantee to produce an increase in the objective at each iteration

Method 2: Local quadratic variational bounds

With $q < 1$, $|w_i|^q$ is concave. Using convex duality,

$$f(w_i) = |w_i|^q = \min_{\lambda_i} \{ \lambda_i w_i^2 - f^*(\lambda_i) \} \quad (5)$$

$$f^*(\lambda_i) = \min_{\eta_i} \{ \lambda_i \eta_i^2 - f(\eta_i) \} \quad (6)$$

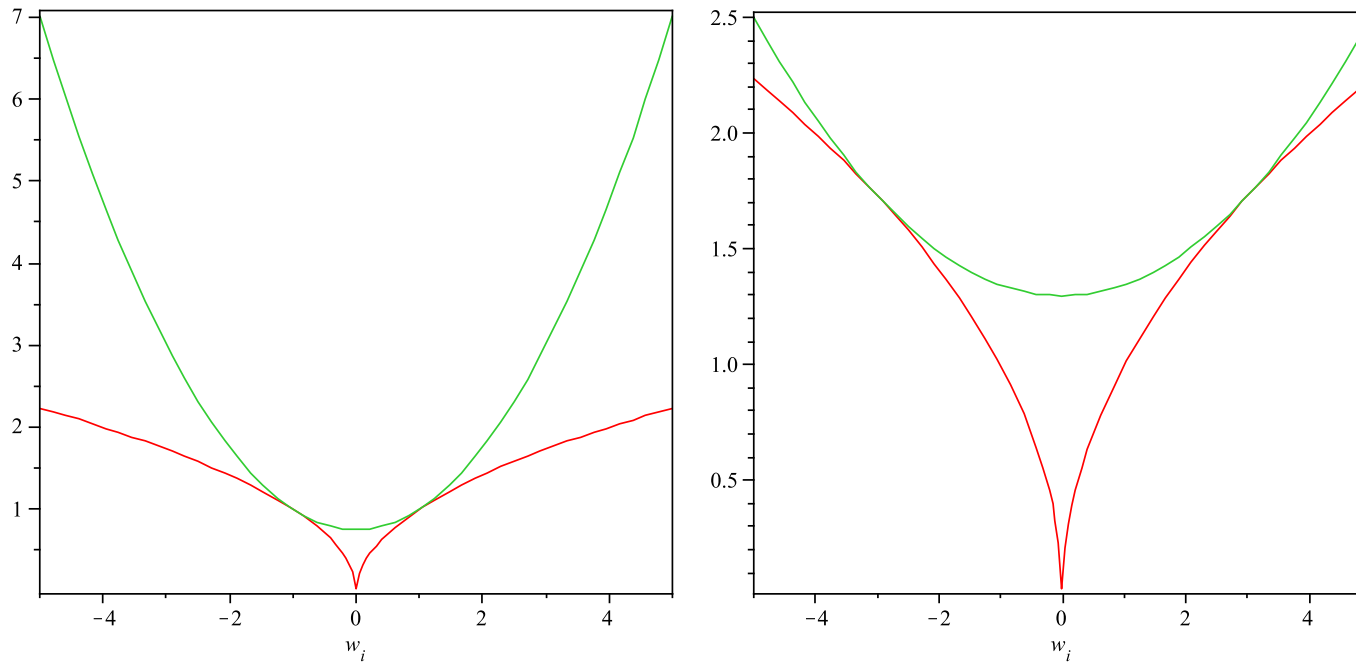
we get:

$$|w_i|^q \leq \frac{f'(\eta_i)}{2\eta_i} (w_i^2 - \eta_i^2) + f(\eta_i) = \frac{1}{2} \left\{ q|\eta_i|^{q-2} w_i^2 + (2-q)|\eta_i|^q \right\} \quad (7)$$

with equality when $\eta_i = \pm|w_i|$, and η_i are variational parameters.

- quadratic in w .
- we note in the case of $q = 1$, the bound (7) recovers exactly the bound proposed in (Krishnapuram et al., PAMI, '05).
- for $q < 1$ (in linear regression), the expression on the r.h.s. of (7) was proposed as an approximation in (Fan & Li, JASA '01). Now we see it is a rigorous upper bound (in fact an 'auxiliary function').

$$|w_i|^q \leq \frac{1}{2} \{q|\eta_i|^{q-2}w_i^2 + (2-q)|\eta_i|^q\}$$



Examples: Left: $q = 0.5$, $\eta_i = 1$, so the quadratic upper bound is tangent in ± 1 . Right: $q = 0.5$, $\eta_i = 3$, so the quadratic upper bound is tangent in ± 3 .

Iterative estimation algorithm

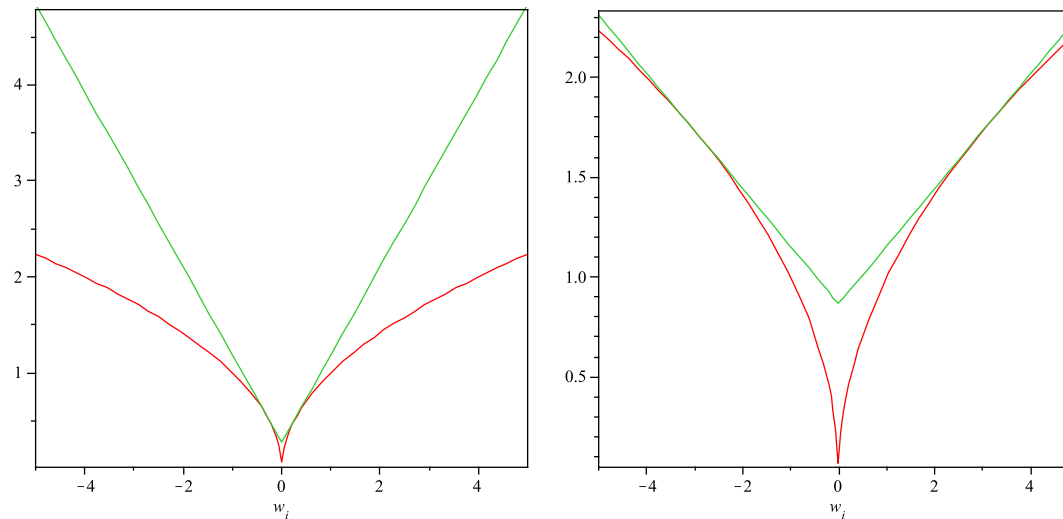
- Initialisation
- Loop until convergence to a local optimum:
 - solve for the var parameters η , i.e. tighten the bound
 - solve for w , i.e. an **L2-regularised logistic regression problem**
- End Loop

[Details in the paper.]

- we initialised η to ones, which worked well. Other schemes could be investigated.
- we used a variational bound to the logistic term due to Jaakkola & Jordan, which then conveniently yielded close form updates throughout. Other existing methods could also be used instead.

Method 3: Local linear variational bound

Recently proposed by (Zou & Li, The Annals of Stats, '08), this appears to be a closer approximation.



Examples - Left: $q = 0.5$, $\eta_i = 0.3$, so the linear upper bound is tangent in ± 0.3 . Right: $q = 0.5$, $\eta_i = 3$, so the linear upper bound is tangent in ± 3

We can derive it from convex duality in the a similar way as before (details in paper). We get:

$$|w_i|^q \leq q|\eta_i|^{q-1}|w_i| + (1 - q)|\eta_i|^q \quad (8)$$

with equality when $\eta_i = \pm|w_i|$.

We can derive it from convex duality in the a similar way as before (details in paper). We get:

$$|w_i|^q \leq q|\eta_i|^{q-1}|w_i| + (1 - q)|\eta_i|^q \quad (9)$$

with equality when $\eta_i = \pm|w_i|$.

The estimation algorithm becomes:

- Initialisation
- Loop until convergence to a local optimum:
 - solve for the var parameters $\boldsymbol{\eta}$, i.e. tighten the bound
 - solve for \boldsymbol{w} , i.e. an **L1-regularised logistic regression problem**
- End Loop

We can derive it from convex duality in the a similar way as before (details in paper). We get:

$$|w_i|^q \leq q|\eta_i|^{q-1}|w_i| + (1 - q)|\eta_i|^q \quad (10)$$

with equality when $\eta_i = \pm|w_i|$.

However, the L1-regularised problem can be estimated exactly by the use of a local quadratic bound approximation.

Consequently, deriving a generalised EM algorithm turns out to be the same as the algorithm we had before with local quadratic bounding method! [details in paper]

Therefore we decided to implement Method 2 for the experiments reported.

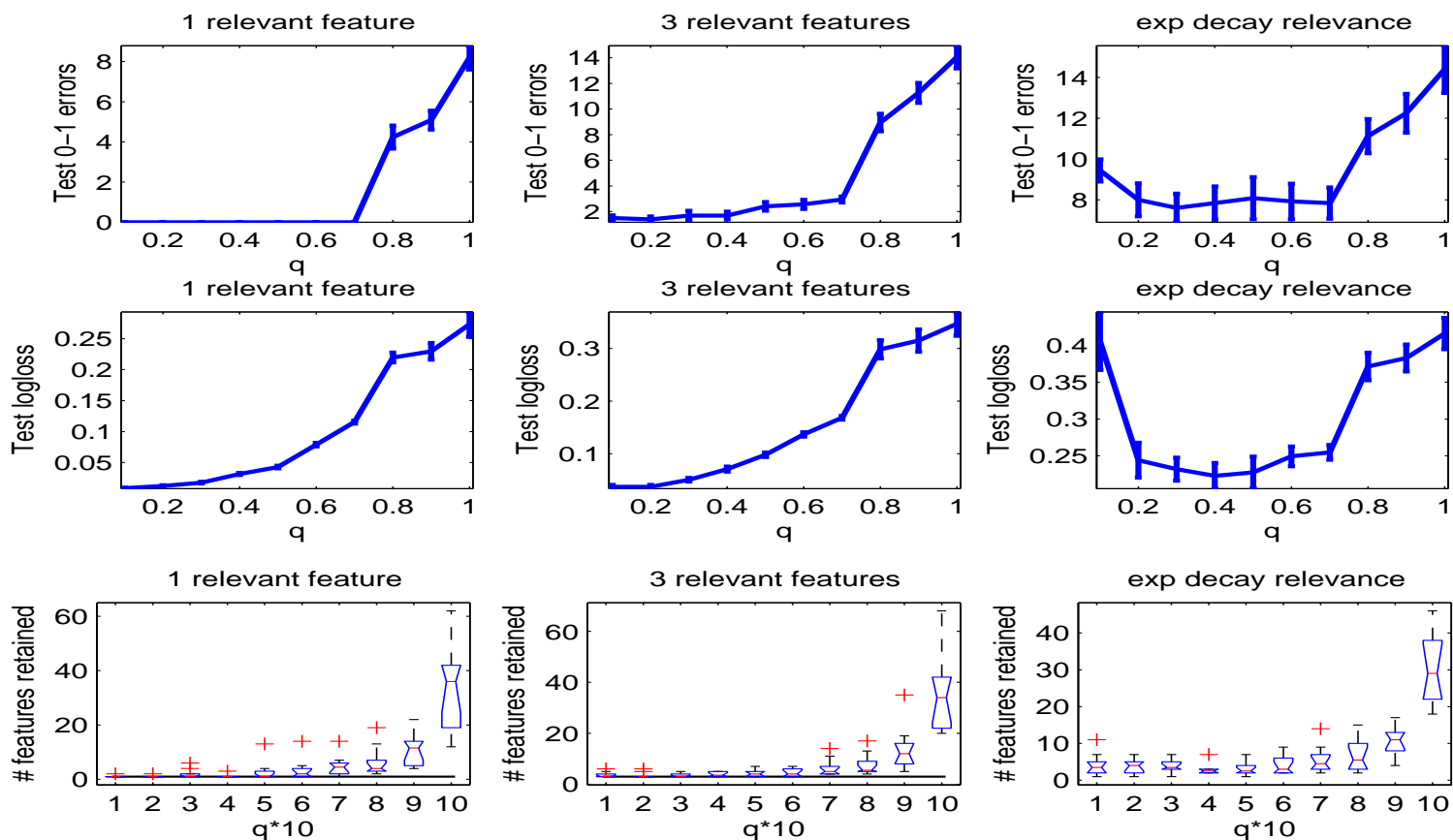
Experiments & Results

Synthetic data sets following and experimental protocol previously used in (Ng, ICML'04) in the first instance: 1000-dimensional data, of which:

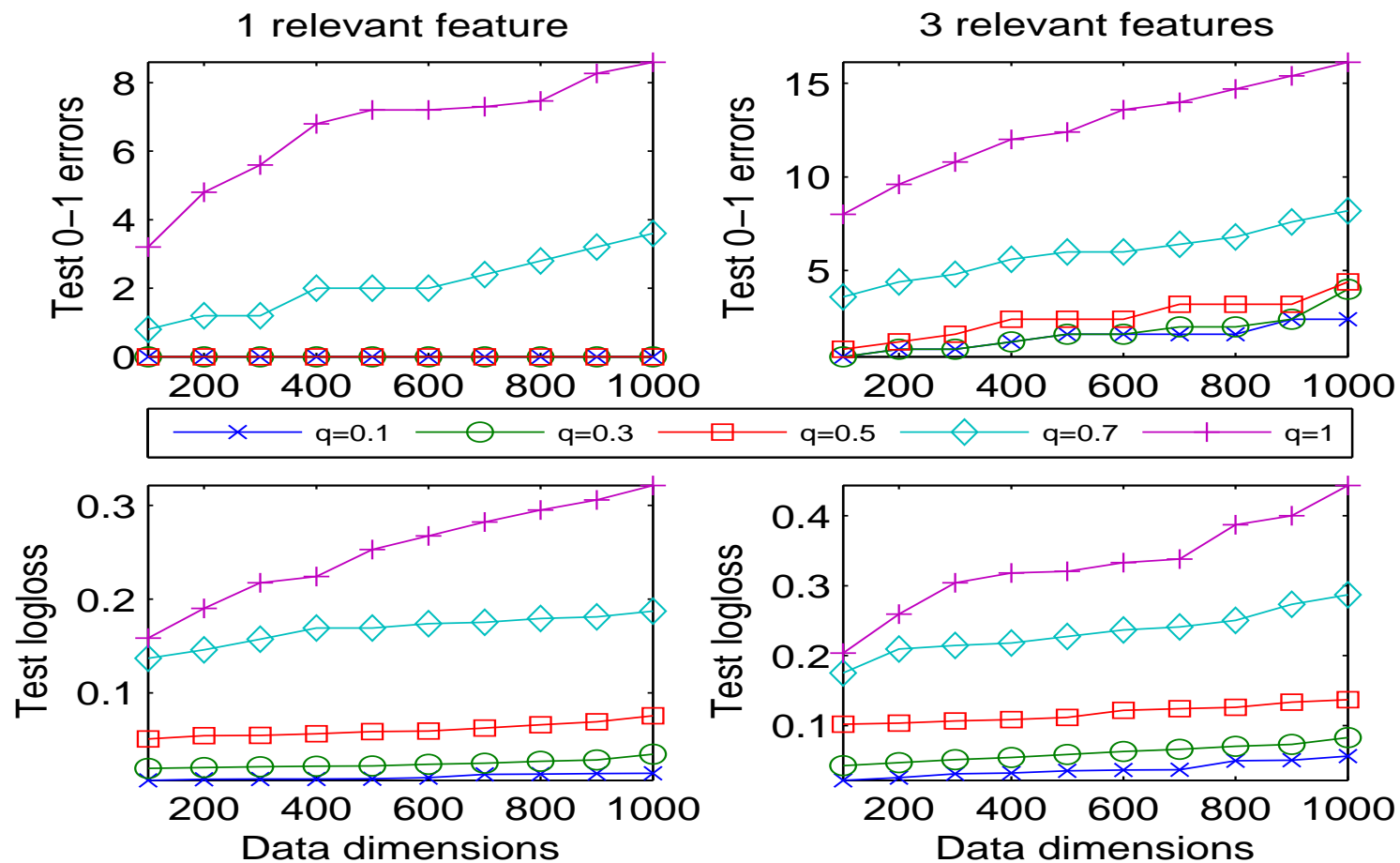
- 1 relevant feature
- 3 relevant features
- exponential decay of feature relevance

Training + validation set (to determine α) size: 70 + 30 points.

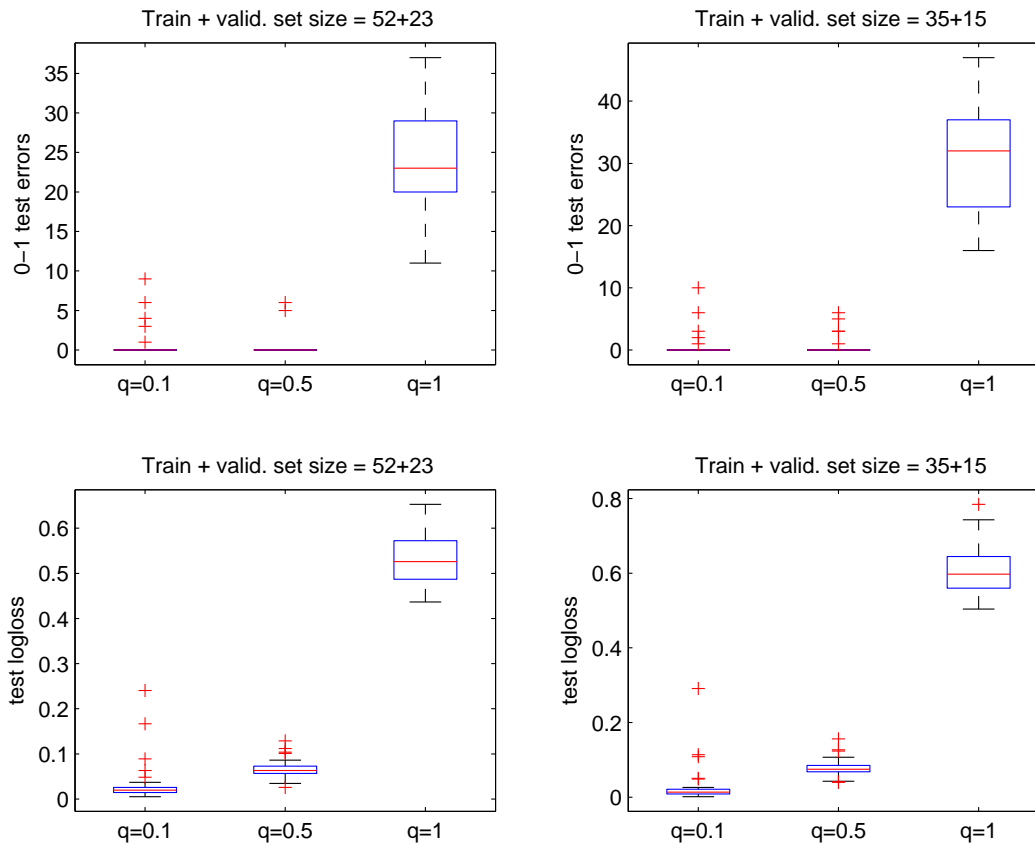
Test set of 100 points.



Results on synthetic data from (A.Ng,'04). Training set size $n_1 = 70$, validation set size=30, and the out-of-sample test set size=100. The statistics are over 10 independent runs with dimensionality ranging from 100 to 1000.



Comparative results on 1000-dimensional synthetic data from (Ng,'04). Each point is the median of > 100 indep. trials. The 0-1 errors are out of 100.



Results on 5000-dimensional synthetic data with only one relevant feature and even smaller sample size. The improvement over L1 becomes larger. (The 0-1 errors are out of 100.)

Conclusions

- We studied $L_{q<1}$ -norm regularisation for logistic regression both theoretically and empirically, for high dimensional data with many irrelevant features.
- We developed a variational method for parameter estimation, and have shown an equivalence between the use of local quadratic and local linear variational bounds to the regularisation term.
- We found that $L_{q<1}$ -norm regularisation is more suitable in cases when the number of relevant features is very small, and works very well despite a very large number of irrelevant features being present.
- In Bayesian terms, $L_{q<1}$ -norm regularisation may be interpreted as a MAP-estimation with a Generalised Laplace Prior. Future work will assess the sample complexity of a full Bayesian estimate.

Selected References

J Fan and R Li. Variable Selection via Non-concave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, Dec 2001, Vol. 96, No. 456, Theory and Methods.

Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink. Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957-968, 2005.

Z Liu, F Jiang, G Tian, S Wang, F Sato, S.J Meltzer, M Tan. Sparse Logistic Regression with L_p Penalty for Biomarker Identification. *Statistical Applications in Genetics and Molecular Biology*. Vol.6, Issue 1, 2007.

A.Y. Ng. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. *Proc. ICML 2004*.

Hui Zou and Runze Li: One-step sparse estimates in non-concave penalized likelihood models. *The Annals of Statistics*, 2008.