

A norm-concentration argument for non-convex regularisation

Ata Kabán

Robert J. Durrant

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

ICML/UAI/COLT Workshop on Sparse Optimization and Variable Selection
Helsinki, 9 July 2008.

Introduction

L1-regularisation - a workhorse in machine learning

- sparsity
- convexity
- logarithmic sample complexity

Non-convex norm regularisation - seems to have added value

- statistics (Fan & Li, '01): oracle property
- signal proc. (Chartland, '07), signal reconstruction (Wipf & Rao, '05)
- 0-norm SVM classification (Weston et al., '03) (results data-dependent)
- genomic data classification (Liu et al., '07)

Regularised regression in high dimensions

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

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

Consider regularised logistic regression for concreteness:

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

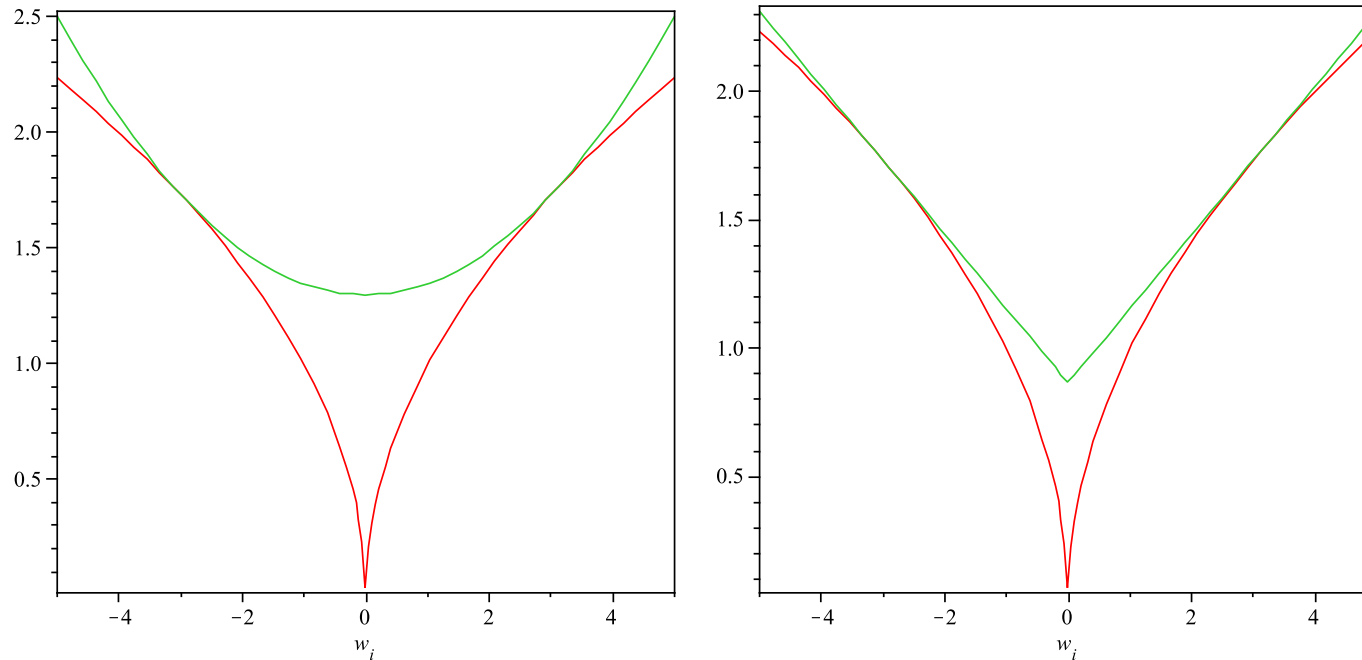
where $\mathbf{w} \in \mathbb{R}^{1 \times m}$ are unknown parameters, $p(y | \mathbf{w}^T \mathbf{x}) = 1 / (1 + \exp(-y \mathbf{w}^T \mathbf{x}))$, and $\|\mathbf{w}\|_q = (\sum_{i=1}^m |w_i|^q)^{1/q}$.

If $q = 2$: L2-regularised ('ridge') logistic regression.

If $q = 1$: L1-regularised ('lasso') logistic regression.

If $q < 1$: $L_{q < 1}$ -regularised logistic regression: non-convex, non-differentiable at 0

A word on some recent estimation algorithms



Local quadratic (Fan & Li,'01) vs. local linear (Zou & Li,'08) bound, tangent at ± 3 . [Despite the latter appears to be a closer approximation, framing the iterative estimation into the E-M methodology framework, it turns out they are in fact equivalent (Kaban & Durrant, ECML'08)]

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

$\hat{er}_z(h) = \frac{1}{n} \sum_{i=1}^n h(\mathbf{x}_i, y_i)$ 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

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

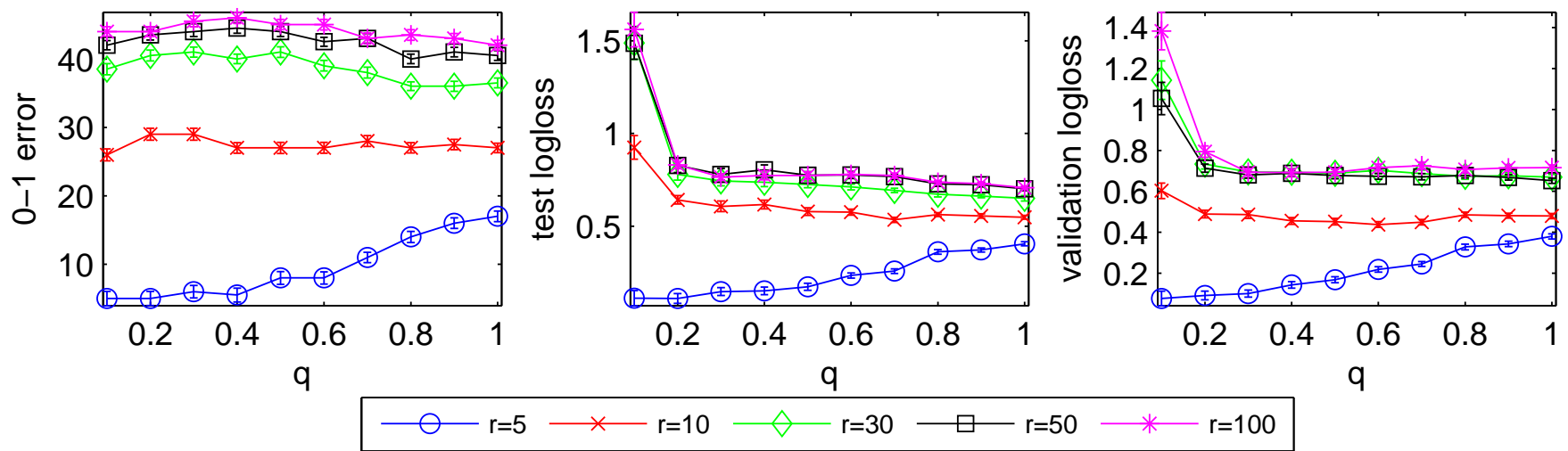
Theorem (A.Ng, '04, extended from L_1 to $L_{q < 1}$).

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

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

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

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



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.

A norm concentration view

Consider the un-regularised version of the problem. Because $n \ll m$, the system is under-determined, and so, $m - n$ components of \boldsymbol{w} can be set arbitrarily.

We can model the arbitrary components of \boldsymbol{w} as being i.i.d. Uniform: $w_i \sim \text{Unif}[-a, a], \forall i \in \{n + 1, \dots, m\}$ with some large a .

The regularisation term is meant to constrain the problem to make it well-posed.

However, in very high dimensions, a counter-intuitive phenomenon known as the concentration of distances and norms comes into play. The regularisation term becomes essentially the same for all the infinitely many possible maximisers of the likelihood term.

Distance concentration

Distance concentration is the counter-intuitive phenomenon that, as the data dimensionality increases without bounds, all pairwise distances between points become identical.

This phenomenon affects every area, where high-dimensional data processing is required — e.g. database indexing & retrieval, data analysis, statistical machine learning.

Concentration of the L2-norm (Demartinez,'94) Let $\mathbf{x} \in \mathbb{R}^m$ a random vector with i.i.d. components of any distribution. Then,

$$\lim_{m \rightarrow \infty} \frac{\mathbb{E}[\|\mathbf{x}\|_2]}{m^{1/2}} = \text{const.}; \quad \lim_{m \rightarrow \infty} \text{Var}[\|\mathbf{x}\|_2] = \text{const.} \quad (3)$$

Concentration of arbitrary dissimilarity functions in arbitrary multivariate distributions (Beyer et al., 99).

Let $F_m, m = 1, 2, \dots$ be an infinite sequence of data distributions and $\mathbf{x}_1^{(m)}, \dots, \mathbf{x}_n^{(m)}$ a random sample of n independent data vectors distributed as F_m .

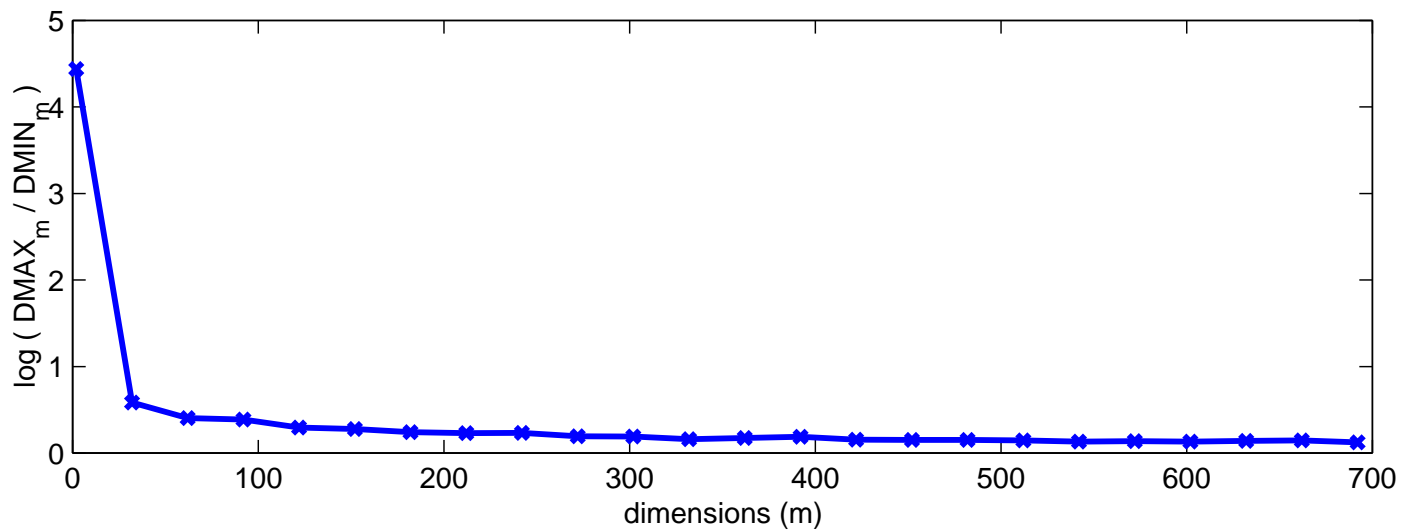
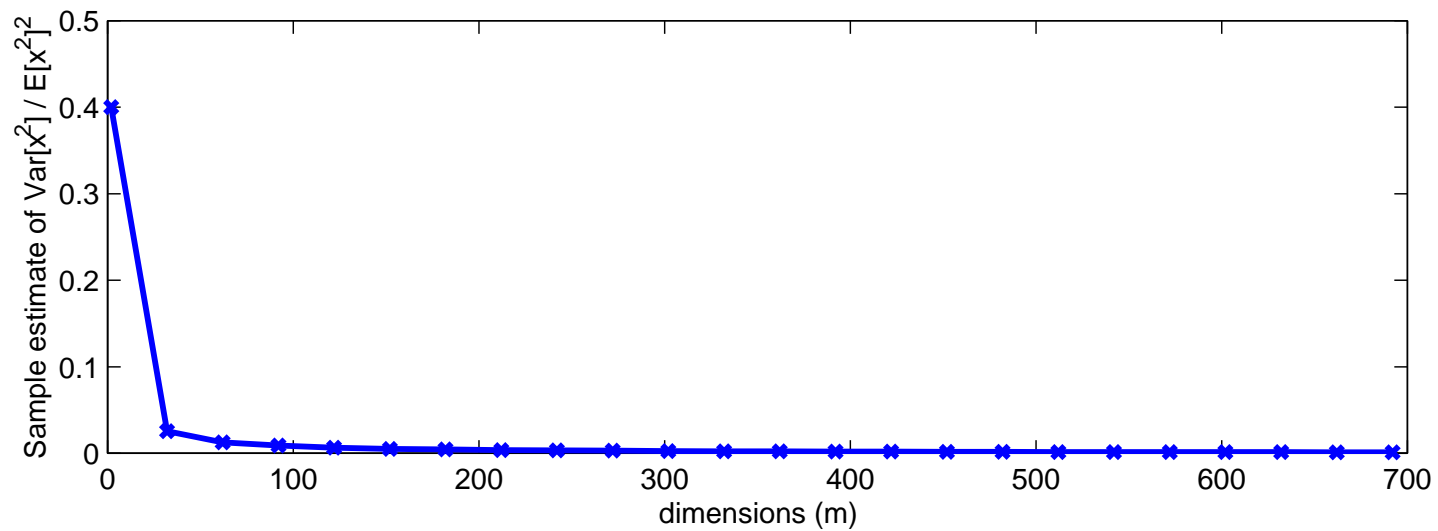
For each m , let $\|\cdot\| : \text{dom}(F_m) \rightarrow \mathbb{R}^+$ be a function that takes a point from the domain of F_m and returns a positive real value.

$p > 0$ an arbitrary positive constant

Assume that $E[\|\mathbf{x}^{(m)}\|^p]$ and $\text{Var}[\|\mathbf{x}^{(m)}\|^p]$ are finite and $E[\|\mathbf{x}^{(m)}\|^p] \neq 0$.

If $\lim_{m \rightarrow \infty} \frac{\text{Var}[(\|\mathbf{x}^{(m)}\|)^p]}{E[(\|\mathbf{x}^{(m)}\|)^p]^2} = 0$, then,

$$\forall \epsilon > 0, \lim_{m \rightarrow \infty} P \left[\max_{1 \leq j \leq n} \|\mathbf{x}_j^{(m)}\| \leq (1 + \epsilon) \min_{1 \leq j \leq n} \|\mathbf{x}_j^{(m)}\| \right] = 1.$$



Applying this to our problem. Denote $RV_m^{(p)}(\|\mathbf{x}\|_q) = \frac{\text{Var}[(\|\mathbf{x}\|_q)^p]}{\mathbb{E}[(\|\mathbf{x}\|_q)^p]^2}$

Using the independence of w_{n+1}, \dots, w_m , we get:

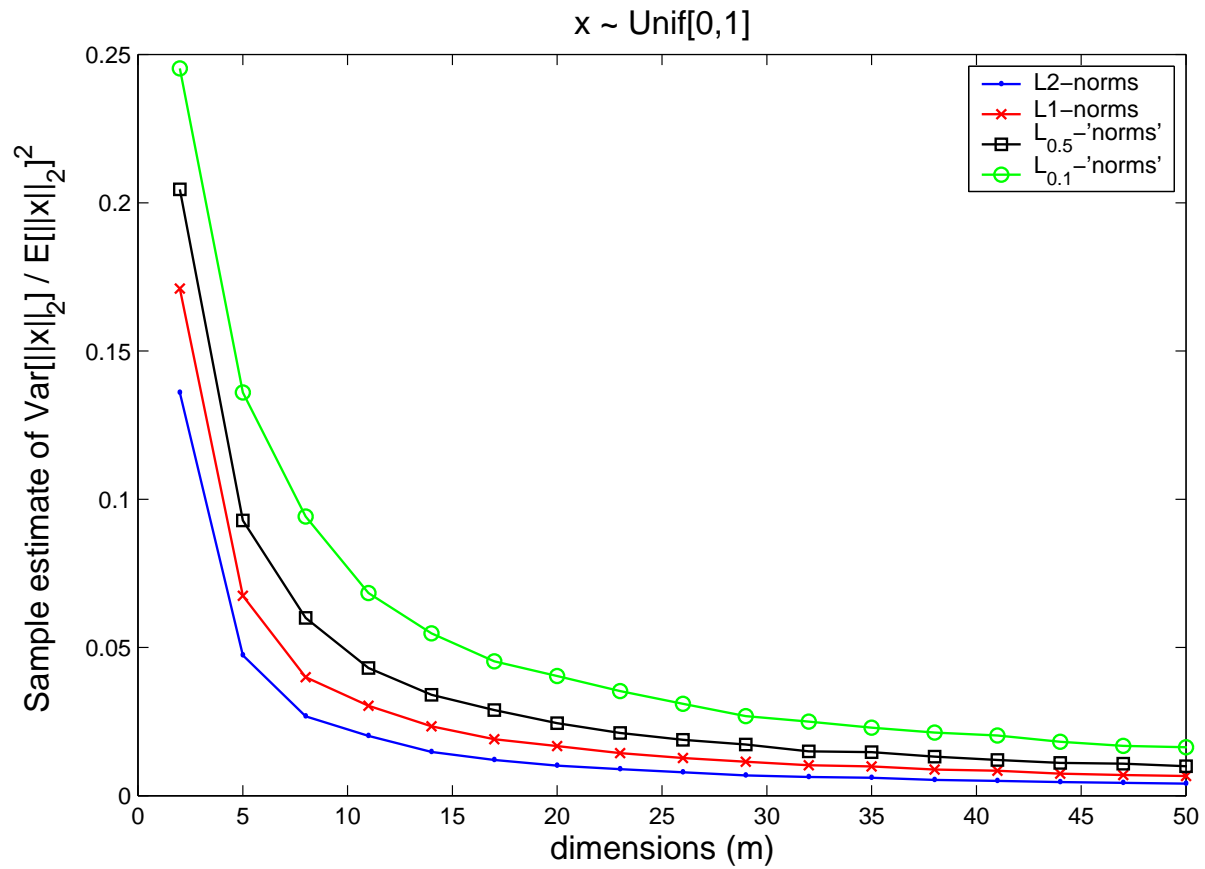
$$RV_m^{(q)}(\|\mathbf{w}\|_q) = \frac{\sum_{i=1}^n \sum_{j=1}^n \text{Cov}[|w_i|^q, |w_j|^q] + \sum_{i=n+1}^m \text{Var}[|w_i|^q]}{\sum_{i=1}^m \sum_{j=1}^m \mathbb{E}[|w_i|^q] \mathbb{E}[|w_j|^q]}$$

which converges to 0 as $m \rightarrow \infty$.

Hence, the problem remains ill-posed despite the regularisation.

The effect of q

Fortunately, not all norms concentrate at the same rate.



Theorem (Francois et al.'07, extended). If $\mathbf{w} \in \mathbb{R}^m$ is a random vector with no more than $n < m$ non-iid components, where n is finite, and all the other components being i.i.d, then

$$\lim_{m \rightarrow \infty} m \frac{\text{Var}[\|\mathbf{w}\|_q]}{\mathbb{E}[\|\mathbf{w}\|_q]^2} = \frac{1}{q^2} \frac{\sigma^2}{\mu^2} \quad (4)$$

where $\mu = \mathbb{E}[w_{n+1}]$, $\sigma^2 = \text{Var}[w_{n+1}]$ and $n + 1$ is one of the i.i.d. dimensions of \mathbf{w} .

Applying this to \mathbf{w} , we can use (4) to approximate

$$\frac{\text{Var}[\|\mathbf{w}\|_q]}{\mathbb{E}[\|\mathbf{w}\|_q]^2} \approx \frac{1}{m} \frac{1}{q^2} \frac{\sigma^2}{\mu^2} = \dots \quad (5)$$

for some large m .

Computing μ and σ^2 for $w_{n+1} \sim \text{Unif}[-a, a]$:

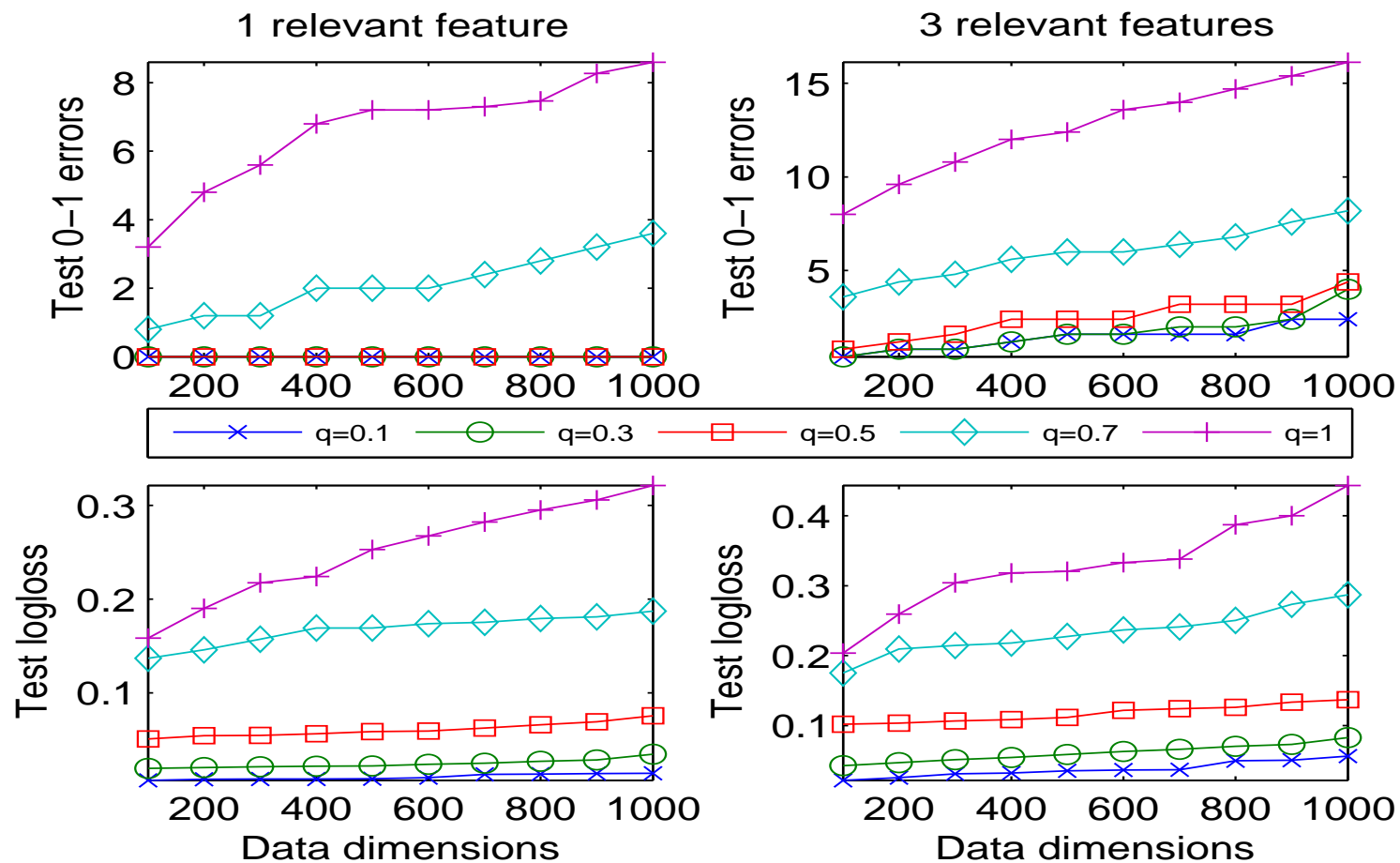
$$\begin{aligned}\mu &= \mathbb{E}[|w_{n+1}|^q] = \int_{-a}^a |w_{n+1}|^q \frac{1}{2a} = \frac{a^q}{q+1} \\ \sigma^2 &= \mathbb{E}[|w_{n+1}|^{2q}] - \mathbb{E}[|w_{n+1}|^q]^2 = \frac{a^{2q}q^2}{(2q+1)(q+1)^2}\end{aligned}$$

So,

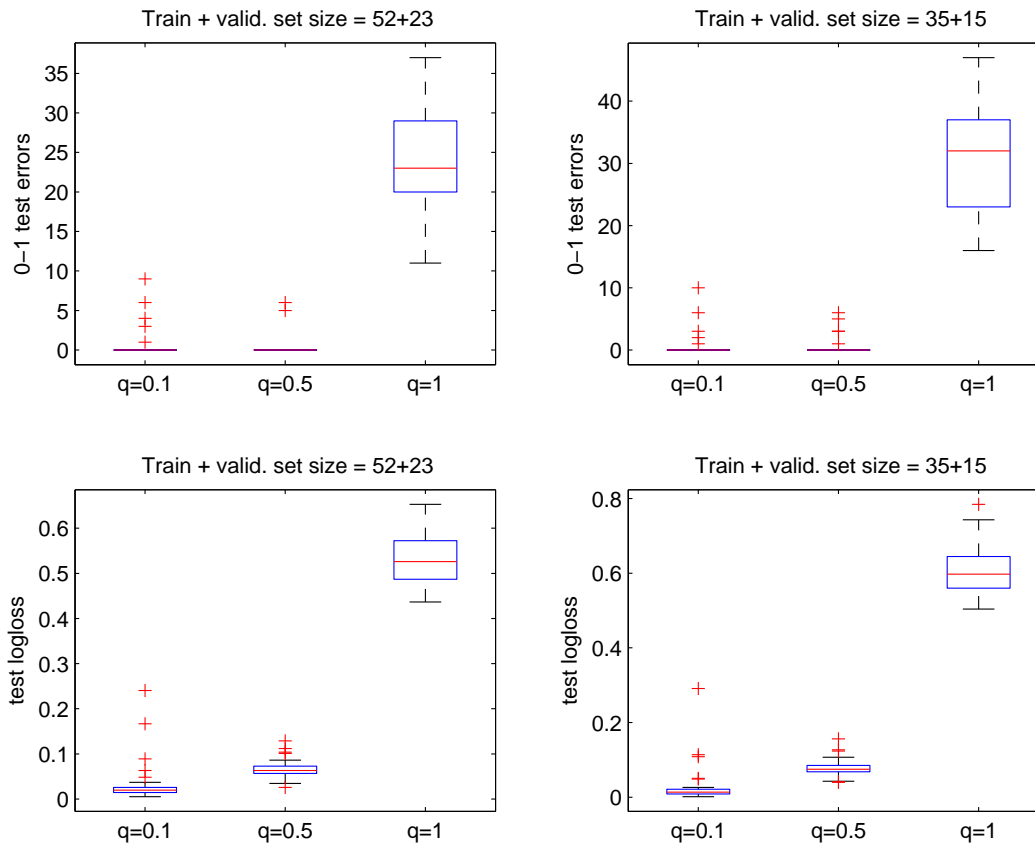
$$\frac{\text{Var}[||\mathbf{w}||_q]}{\mathbb{E}[||\mathbf{w}||_q]^2} \approx \frac{1}{m} \frac{1}{2q+1} \quad (6)$$

(Conveniently, a cancels out in this computation.)

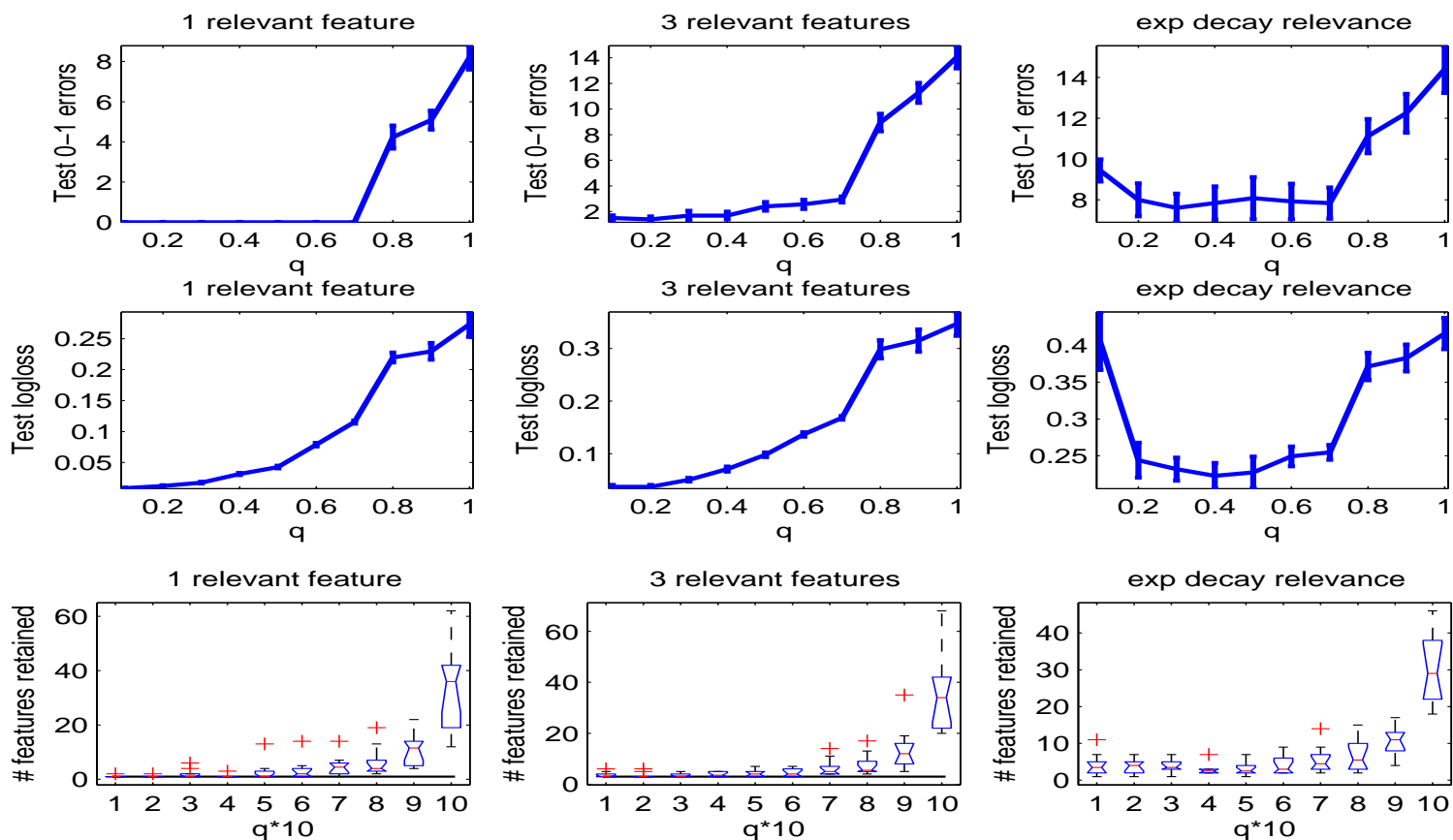
Observe this is a decreasing function of q . Thus, the smaller the q the better, from the point of countering the concentration of the norm in regularisation.



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.)



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.

Discussion & further work

The learning-theoretic sample complexity bound for generalization is an (loose) upper-bound only.

Our analysis based on norm-concentration so far only used that $n \ll m$. Further work should examine of the effect of $r \ll m$ from this perspective.

The phenomenon of concentration of norms and distances in very high dimensions impacts all high dimensional problems. Its implications for learning and generalization (and for other areas) is and open question.

References

C.C. Aggarwal, A. Hinneburg, & D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. Proc. Int. Conf. Database Theory, 2001, pp. 420-434.

K. Beyer, J. Goldstein, R. Ramakrishnan, & U. Shaft. When is nearest neighbor meaningful? Proc. Int. Conf. Database Theory, pp. 217-235, 1999.

D François, V Wertz, & M Verleysen. The concentration of fractional distances. IEEE Trans. on Knowledge and Data Engineering, vol 19, no 7, July 2007

A Kabán and R.J Durrant. Learning with $L_{q<1}$ vs. L_1 -norm regularization with exponentially many irrelevant features. Proc. ECML 2008, to appear.

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

A.Y. Ng. Feature selection, L1 vs. L2 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.