
A norm-concentration argument for non-convex regularisation

Ata Kabán & Robert J. Durrant

A.KABAN@CS.BHAM.AC.UK

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

1. Introduction

L1-regularisation has become a workhorse in statistical machine learning, because of its sparsity-inducing property and convenient convexity. In addition, detailed theoretical and empirical analysis (Ng, 2004) has shown its ability to learn with exponentially many irrelevant features, in the context of L1-regularised logistic regression.

However, independent results in several areas indicate added value to non-convex norm regularisation, despite the existence of local optima. Work in statistics (Fan & Li, 2001) and signal reconstruction (Wipf & Rao, 2005) have established the oracle properties of non-convex regularisers. Good empirical results were also reported in signal processing (Chartland, 2007) and SVM classification (Weston et.al, 2003). Furthermore, using a family of non-convex norms that we shall refer to as fractional-norms in the rest of the paper, turned out to consistently outperform the L_1 regulariser in real high-dimensional genomic data classification (Liu et.al, 2007), both in terms of error rates and interpretability. Related ideas, termed as 'zero-norm' regularisation (Weston et.al, 2003) were also found useful in many other applications, though their success appeared to be data dependent. It is therefore of interest to gain a better understanding of the potential advantages of non-convex norm regularisers, which is our purpose.

2. Regularised regression in high dimensions

Given a training set of input-target pairs $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$, where $\mathbf{x}_j \in \mathbb{R}^m$ are m -dimensional input points and $y_j \in \{-1, 1\}$ are their labels. We are interested in high-dimensional problems, with few $r \ll m$ relevant features and 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 parameters, and the norm in the regularisation term is defined as $\|\mathbf{w}\|_q =$

$(\sum_{i=1}^m |w_i|^q)^{1/q}$. Note, if $q = 2$ or $q = 1$, this is L2- or L1-regularised regression respectively. However, if $q \in (0, 1)$, we have a non-convex regularisation term, which we will refer to as $L_{q < 1}$ -regularisation or 'fractional norm'-regularisation. This is not strictly a norm in the mathematical sense, since it does not satisfy the triangle inequality. Also, parameter estimation becomes more difficult than with the more common L1 or L2 norms, because the $L_{q < 1}$ -norm is non-differentiable at zero and non-convex. Some recent algorithms were developed (Fan & Li, 2001; Kabán & Durrant, 2008) that we use in the reported numerical simulations.

Sample complexity. Noticing that $\|\mathbf{w}\|_{q < 1} \geq \|\mathbf{w}\|_1, \forall \mathbf{w}$, and extending the result of (Ng, 2004) obtained for L_1 -regularised logistic regression, it can be shown (Kabán & Durrant, 2008) that $L_{q < 1}$ -norm regularised logistic regression also enjoys a sample complexity that is logarithmic in the data dimensionality m and polynomial in the number of relevant features r and other quantities of interest, $n = \Omega((\log m) \times \text{poly}(A, r^{1/q}, 1/\epsilon, \log(1/\delta)))$. Logarithmic bounds are the best known bounds for feature selection.

However, we are also interested to know whether, and in which cases there is any advantage in using $q < 1$ rather than $q = 1$. Fractional norms were previously studied in the databases and data engineering literature (Aggarwal et.al, 2001; François et.al, 2007), for mitigating the dimensionality curse. In the sequel, we use some of their results to better understand the effects of q in the regularisation term.

2.1. A norm-concentration view

Consider the un-regularised version of the problem. Because $n \ll m$, the system is under-determined and so the set of solutions is infinite. For the analysis that follows, we would like to capture the distribution of the set of solutions to the unregularised model. To ease notations and without loss of generality, we consider the $m - n$ free variables of \mathbf{w} , that can be set arbitrarily, start from component $n + 1$. We can model the distribution of these arbitrary components as being i.i.d. uniform. So we will have

$w_i \sim \text{Unif}[-a, a], \forall i \in \{n+1, \dots, m\}$ with some large a — in fact, the result of our analysis will turn out not to depend on a .

It is well known that in such ill-conditioned problems, the regularisation term is meant to constrain the problem and make it well-posed. This is indeed so, as long as m is not too large. However, in very high dimensions, the rather counter-intuitive phenomenon known as the concentration of distances and norms comes into play, which is overlooked in previous analyses of regularisation. That is, as the dimensionality grows, the norm that appears in the regularisation constraint becomes essentially the same for all minimisers of the likelihood term. Therefore the problem remains ill-conditioned despite of the regularisation. To see this, we use a result due to (Beyer et al, 1999).

Theorem 1 (Beyer et al, 1999). Let $\mathbf{w}_1, \dots, \mathbf{w}_n$ be a random sample of size n drawn from the multivariate distribution of \mathbf{w} as defined by the likelihood term, and $p > 0$ arbitrary. If $\lim_{m \rightarrow \infty} \frac{\text{Var}[(\|\mathbf{w}\|_q)^p]}{\text{E}[(\|\mathbf{w}\|_q)^p]^2} = 0$, then

$$\forall \epsilon > 0 \lim_{m \rightarrow \infty} P \left[\max_{1 \leq j \leq n} \|\mathbf{w}_j\|_q \leq (1 + \epsilon) \min_{1 \leq j \leq n} \|\mathbf{w}_j\|_q \right] = 1, \text{ where } \text{E}[\cdot] \text{ and } \text{Var}[\cdot] \text{ are the theoretical expectation and variance, and the probability on the r.h.s. is over an arbitrary random sample of size } n.$$

Applying this to our case, denoting $RV_m^{(p)} = \frac{\text{Var}[(\|\mathbf{w}\|_q)^p]}{\text{E}[(\|\mathbf{w}\|_q)^p]^2}$, choosing $p = q$ to ease the computations, and using the independence of w_{n+1}, \dots, w_m ,

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

which converges to 0 as $m \rightarrow \infty$. Hence, in very high dimensions, the regularisation term of any of the possible solutions become essentially indistinguishable. In fact, simulations in (Beyer et al, 1999) indicate the problem may become of concern already at 10-20 dimensions.

2.1.1. THE EFFECT OF q

Fortunately, not all norms concentrate equally fast as the dimensionality increases. In particular, the family of norms used in our regularisation term was studied in this respect by (Aggarwal et.al, 2001; François et.al, 2007). Here we propose a straightforward extension of a recent result by (François et.al, 2007) to give us insight into the effect of q and guide our choice of q for the type of problems considered. For this part of the analysis, $p = 1$.

Theorem 2 (Francois (2007), 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, then

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

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

Proof (sketch). To allow for a finite number of possibly non-iid marginal distributions we write $\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m |w_i|^q = \lim_{m \rightarrow \infty} \frac{1}{m} \left\{ \sum_{i=1}^n |w_i|^q + \frac{\sum_{i=n+1}^m |w_i|^q}{m-n} (m-n) \right\} = \lim_{m \rightarrow \infty} \frac{\sum_{i=n+1}^m |w_i|^q}{m-n}$ and $\lim_{m \rightarrow \infty} \frac{\text{Var}[\|\mathbf{w}\|_q]}{m} = \lim_{m \rightarrow \infty} \frac{1}{m} \left\{ \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(|w_i|^q, |w_j|^q) + \sum_{i=n+1}^m \text{Var}(|w_i|^q) \right\} = \text{Var}[w_{n+1}^q]$. Using these, the rest of the proof follows the same steps as the original theorem: We can show that $\lim_{m \rightarrow \infty} \frac{\text{E}[\|\mathbf{w}\|_q]}{m^{1/q}} = \mu^{1/q}$, and $\lim_{m \rightarrow \infty} \frac{\text{Var}[\|\mathbf{w}\|_q]}{m^{2/q-1}} = \frac{\sigma^2}{q^2 \mu^{2(q-1)/q}}$ — which put together gives us the required result. Q.E.D.

As in (François et.al, 2007), we can then approximate $RV_m^{(1)}$ for some large m using (2), so we can read off the optimal q by computing the maximiser of this expression. For $w_{n+1} \sim \text{Unif}[-a, a]$, we get:

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

so we have

$$\frac{\text{Var}[\|\mathbf{w}\|_q]}{\text{E}[\|\mathbf{w}\|_q]^2} \approx \frac{1}{m} \frac{1}{2q+1} \quad (3)$$

which rather conveniently turns out to be independent of a . Most importantly, we see that (3) is a monotonically decreasing function of q . In other words, in small sample size problems and with increasing input dimensions, the q -norm regularisation term will concentrate the slowest if we choose the smallest q . Therefore, in such settings, from this analysis we can conclude that, the 0-norm regulariser represents the best choice, from the point of keeping the problem from becoming ill-conditioned till fairly high dimensions.

3. Results

We generated synthetic data sets similarly to (Ng, 2004), having $r = 1$ and $r = 3$. We keep r small to suppress the effect of q on the sample complexity w.r.t r , and follow differences attributable to the norm concentration effects. In each experiment, the training and validation set size (to select the regularisation parameter) was 70 and 30 respectively and the performance was measured on an independent test set of size 100.

Figure 1 gives the results in terms of both the number of 0-1 errors and the logloss, for different values of q , and varying the data dimensionality. On these plots, each point represents the median of 140 independent trials. We see the logarithmic increase of errors with the data dimensionality, as predicted from the theory, is well supported. More interestingly, smaller values of q do systematically achieve significant improvements. Also the relevant features are more correctly recovered, which clearly favours interpretability. L1 regularisation in turn tends to retain too many features in high dimensions.

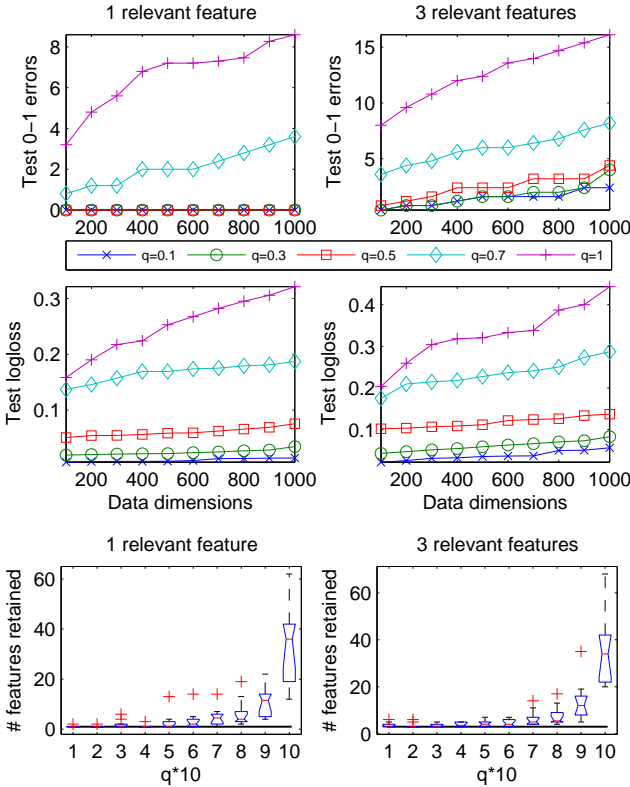


Figure 1. Comparative results (see text for details). The 0-1 errors are out of 100.

In a subsequent experiment we generated 5000-dimensional data with only 1 relevant feature, and even smaller training set sizes. This is shown on Figure 2. $L_{q<1}$ -regularisation still has excellent performance (the median of the 0-1 errors is still 0) and the improvement to L1-regularisation becomes even larger.

To summarise, in this work we considered a special problem setting, which nevertheless is quite often relevant in gene expression arrays for example, and the superior empirical results of (Liu et.al, 2007) was in

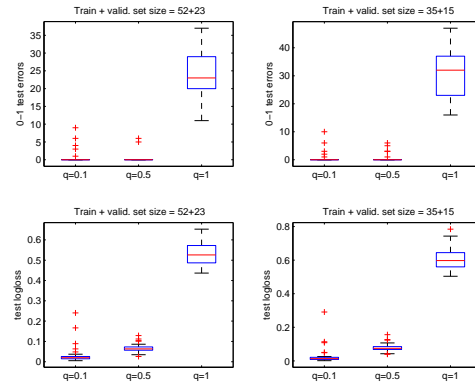


Figure 2. Results on 5000-dimensional synthetic data with only one relevant feature and small sample size. Each box-plot summarises 30 independent trials. The 0-1 errors are out of 100.

fact a motivating factor for our study. Our analysis gave some new insights that complement other analysis frameworks and our numerical simulations are in agreement with the theory.

ACKNOWLEDGEMENT

RJD was funded by an EPSRC CTA studentship.

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.

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

R Chartrand. Exact reconstructions of sparse signals via non-convex minimization. IEEE Signal Process. Lett., vol. 14, pp. 707-710, 2007.

J Fan & R Li. Variable Selection via Non-concave Penalized Likelihood and its Oracle Properties. J. Amer. Stat. Assoc, Dec 2001, Vol. 96, No. 456.

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.

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.

J Weston, A Elisseeff, B Schölkopf, & M Tipping. Use of the Zero-Norm with Linear Models and Kernel Methods. J. Machine Learning Research 3, pp. 1439-1461, 2003.

D.P. Wipf & B.D. Rao, “ ℓ_0 -Norm Minimization for Basis Selection”, NIPS 17, MIT Press, 2005.