Learning with $L_{q<1}$ vs L_1-Norm Regularisation with Exponentially Many Irrelevant Features

Ata Kabán and Robert J. Durrant

School of Computer Science, The University of Birmingham,
Birmingham, B15 2TT, UK
A.Kaban@cs.bham.ac.uk

Abstract. We study the use of fractional norms for regularisation in supervised learning from high dimensional data, in conditions of a large number of irrelevant features, focusing on logistic regression. We develop a variational method for parameter estimation, and show an equivalence between two approximations recently proposed in the statistics literature. Building on previous work by A.Ng, we show the fractional norm regularised logistic regression enjoys a sample complexity that grows logarithmically with the data dimensions and polynomially with the number of relevant dimensions. In addition, extensive empirical testing indicates that fractional-norm regularisation is more suitable than L1 in cases when the number of relevant features is very small, and works very well despite a large number of irrelevant features.

1 $L_{q<1}$-Regularised Logistic Regression

Consider a training set of pairs $z = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^n$ drawn i.i.d. from some unknown distribution P. $\boldsymbol{x}_j \in \mathcal{R}^m$ are m-dimensional input points and $y_j \in \{-1, 1\}$ are the associated target labels for these points. Given z, the aim in supervised learning is to learn a mapping from inputs to targets that is then able to predict the target values for previously unseen points that follow the same distribution as the training data.

We are interested in problems with large number m of input features, of which only a few $r << m$ are relevant to the target. In particular, we focus on a form of regularised logistic regression for this purpose:

$$\max_{\boldsymbol{w}} \sum_{j=1}^{n} \log p(y_j|\boldsymbol{x}_j, \boldsymbol{w}) \tag{1}$$

$$\text{subject to} ||\boldsymbol{w}||_q \leq A \tag{2}$$

or, in the Lagrangian formulation:

$$\max_{\boldsymbol{w}} \sum_{j=1}^{n} \log p(y_j|\boldsymbol{x}_j, \boldsymbol{w}) - \alpha||\boldsymbol{w}||_q^q \tag{3}$$

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part I, LNAI 5211, pp. 580–596, 2008.

where α is the Lagrange multiplier that balances between fitting the data well and having small parameter values.

In the above, the likelihood of predicting y for some input $\boldsymbol{x}$ in logistic regression is

$$p(y|\boldsymbol{w}^T\boldsymbol{x}) = 1/(1+\exp(-y\boldsymbol{w}^T\boldsymbol{x})),$$

parameterised by $\boldsymbol{w} \in \mathcal{R}^{1\times m}$. The norm that forms the regularisation term is defined as

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

Note, with $q = 2$ or $q = 1$, this is L2- or L1-regularised logistic regression respectively. The generalisation ability and sample complexity of L2- vs. L1-regularised logistic regression have been comprehensively studied in [10] — showing the impressive superiority of the latter in problems with large m and small r. Here we seek to extend their study to the case of $q \in (0, 1)$, which we refer to as $L_{q<1}$-regularisation or 'fractional norm'-regularisation.

The fractional norm is not strictly a norm in the mathematical sense, since it does not satisfy the triangle inequality. In addition it is non-differentiable at zero and non-convex, which make its use technically more challenging than that of the more common L1 or L2 norms. Nevertheless, work in a number of disjoint areas independently indicate added value to this norm, in terms of certain specific criteria. It may be interesting to note, the fractional norms were previously studied in the databases and data engineering literature [5], for mitigating the dimensionality curse. Work in statistics [3,16] have established the oracle properties of such and related [14] non-convex regularisation, despite the existence of several local optima. Good empirical results were also reported in signal reconstruction [2] and in SVM classification [8,13]. Furthermore, using fractional norm regularisation, consistently superior empirical results were reported reported on real genomic data sets [7]. Related ideas of using non-convex ('zero-norm') regularisation [13] were found useful in many application settings, though this appeared to be data dependent.

It is therefore of interest to know more exactly in what conditions fractional norm regularisation would be superior to other commonly used norms for machine learning, in terms of its generalisation ability and sample complexity. The work of [10] elucidated such issues for L1 vs. L2 regularisation, but to our best of knowledge, there is no such systematic study assessing fractional norm regularisation, and this is what we attempt in this paper.

2 Analysis

The analysis presented in this section follows standard techniques in statistical learning theory, and in particular the techniques previously employed in [10] for analysing the L_1-regularised logistic regression. We give more details to show that, their results extend straightforwardly to the $L_{q<1}$-regularised case, namely that $L_{q<1}$-regularised logistic regression enjoys a logarithmic sample complexity w.r.t. the data dimensionality, and polynomial in the number of relevant features. Thus, it can learn to generalise from data with exponentially many irrelevant features. Recall that, a logarithmic sample complexity corresponds to the best known bounds for feature selection (see e.g. [10] and references therein).

Denote by $G = \{g : g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}, \boldsymbol{x} \in \mathcal{R}^m\}$ the linear function class, and $H = \{h_g(\boldsymbol{x}, y) = -\log p(y|g(\boldsymbol{x})) : g \in G, \boldsymbol{x} \in \mathcal{R}^m, y \in \{-1, 1\}\}$ is the function class under study, i.e. the parameterised negative log likelihood of the logistic regression model.

The learning algorithm we consider is the following. Divide the available labelled set z in two disjoint sets z_1 and z_2, with sizes $|z_1| = n_1, |z_2| = n - n_1$ respectively, where z_1 (the training set) is used for learning $\boldsymbol{w}$ by optimising (1)-(2) or (3) with a fixed A (or α) – this outputs the function $L(z_1) = \min_{h \in H} \hat{er}_{z_1}(h)$ – and z_2 (the validation set) is used to select the optimal A (or equivalently α).

We start by considering the probability of error[1] at the training stage of the above learning algorithm. For a given $h \in H$, denote $er_P(h) = E_{(\boldsymbol{x},y)\sim \text{iid} P}[h(\boldsymbol{x}, y)]$ the true error of h w.r.t. the unknown data distribution P under the i.i.d. sampling assumption. Further, $\hat{er}_{z_1}(h) = \frac{1}{n_1} \sum_{i=1}^{n_1} h(\boldsymbol{x}_i, y_i)$ is the sample error achieved by h on the training set z_1, and $opt_P(H) = \inf_{h \in H} er_P(h)$ is the approximation error of our function class H. $h^* \in H$ will denote the function in H that is closest to the one where this infimum is attained.

Theorem1. $\forall \epsilon > 0, \forall \delta > 0, \forall m, n_1 \geq 1$ and $\forall A \geq 0$ fixed, in order to ensure that

$$er_P(L(z_1)) \leq opt_P(H) + \epsilon \tag{4}$$

with probability $1 - \delta$, it is enough to have the following training set size:

$$n_1(L, \epsilon, \delta) = \frac{2048(A+1)^2}{\epsilon^2}[\log \frac{8(2m+1)}{\delta} + \frac{256A^2}{\epsilon^2} + 1] \tag{5}$$

Hence the sample complexity $n_1 = \Omega((\log m) \times poly(A, 1/\epsilon, \log(1/\delta))$ is logarithmic in the data dimensionality m and polynomial in A and other quantities of interest.

The proof is given in the Appendix.

In the above, the regularisation parameter A was fixed. Now the error from the validation procedure for selecting A should be considered. We employ the same hold-out validation scheme as [10], the implementation of which is to select A from the pre-defined set of possible values $\{0, 1, 2, 4, 8, ..., C\}$ such that the logloss is minimised on the hold-out set. If r denotes the number of relevant features, we have that $|w_{i_j}| \leq K, j = 1, ..., r$ for some constant K and for all others the entry in $\boldsymbol{w}$ is zero. So this gives us the following:

$$|w_{i_j}|^q \leq K^q \Rightarrow ||\boldsymbol{w}||_q^q \leq rK^q \Rightarrow ||\boldsymbol{w}||_q \leq r^{1/q}K \tag{6}$$

Therefore, from the above set of possible values, the optimal choice of A is $r^{1/q}K \leq A \leq 2r^{1/q}K$ (which recovers in the $q = 1$ case the relationship between

[1] Although the theory concerns the logloss error, an upper bound on the logloss also implies an upper bound on the 0-1 misclassification error [10].

A and r given in [10]). From (6), we see A grows polynomially with r, so with this optimal choice, the previous result also implies the sample complexity is polynomial in the number of relevant features r. Now, for the same standard arguments as in [10], the hold-out validation procedure ensures that with probability $1 - \delta$ the selected parameter will have performance at most ϵ worse than that with the best parameter. Adding this together with the previous result (eq (42) in Appendix), we have that:

$$er_P(L(z_1)) \leq opt_P(H) + 3\epsilon \tag{7}$$

with probability $1-2\delta$. We can replace δ with $\delta/2$ and ϵ with $\epsilon/3$, and the sample complexity remains in the same complexity class.

Comments. It may be interesting to note, from the expression of the sample complexity we can see this model is advantageous as long as the $\log m$ term is dominant – i.e. when A is small, or equivalently, when the number of relevant features is small. The sample complexity grows with the 4-th power of A. So when the number of relevant features is large enough for this to become the dominant term, we might expect to lose the benefits of a sparsity-inducing regularisation. Also, from (6) we can see that, the smaller the $q < 1$, the faster A grows in r. Hence we might expect a small exponent q to work the best in a setting where the number of relevant features is very small.

3 Implementation

The likelihood function is not convex and is also non-differentiable at zero, which makes the implementation non-trivial.

3.1 Method 1: Using a Smooth Approximation

The following smooth approximation to the regularisation term has been proposed in [7]:

$$\sum_{i=1}^{m} |w_i|^q \approx \sum_{i=1}^{m} (w_i^2 + \gamma)^{q/2} \tag{8}$$

where γ is set to a small value. The approximate log likelihood is then differentiable and any nonlinear optimisation method applies.

Although this approach seems practically convenient and easy to implement, it has several drawbacks. The main difficulty is that there is no obvious or principled way to set the smoothing parameter γ. Ideally, one would like it to be as small as it can be without causing numerical instability problems — which is not easy to determine. As one would expect, we observed in our experiments that, the smaller the q is, the more likely the danger of running into numerical instability when γ is chosen to be too small. On the other hand, a larger γ tends to over-smooth the regularisation term, specially when q is further from zero, and this causes it to loose its beneficial sparsity-inducing effect. In addition, iterative optimisers applied to this approximation are not guaranteed to produce an increase in the likelihood of the model at each iteration. The next section presents a more principled alternative that bypasses all of these limitations.

3.2 Method 2: Local Quadratic Variational Approximation

A local quadratic approximation was proposed in [3], which, as we shall see, is actually a strict lower bound on the model likelihood. This becomes evident by deriving it from convex duality [6].

With $q < 1$, the function $|w_i|^q$ is concave, so we can write:

$$f(w_i) = |w_i|^q = \min_{\lambda_i} \left\{ \lambda_i w_i^2 - f^*(\lambda_i) \right\} \tag{9}$$

$$f^*(\lambda_i) = \min_{\eta_i} \left\{ \lambda_i \eta_i^2 - f(\eta_i) \right\} \tag{10}$$

In convex analysis, the function $f^*(.)$ is termed the conjugate (or dual) function of $f(.)$. Geometrically, $f^*(\lambda_i)$ represents the amount of vertical shift applied to λw_i^2 to obtain the quadratic upper bound with precision parameter λ, that touches $f(w_i)$.

Denoting $g(\eta_i) = \lambda_i \eta_i^2 - f(\eta_i)$, the maximum occurs either at $\eta_i = 0, g(\eta_i = 0) = 0$ or at a solution of the stationary equation when $\eta_i \neq 0$:

$$g'(\eta_i) = 2\lambda_i \eta_i - f'(\eta_i) = 0 \quad \Rightarrow \lambda_i = \frac{f'(\eta_i)}{2\eta_i} \tag{11}$$

and $f'(\eta_i) = q|\eta_i|^{q-1} sign(\eta_i)$. Replacing in (9) yields the variational bound:

$$|w_i|^q \leq \frac{f'(\eta_i)}{2\eta_i}(w_i^2 - \eta_i^2) + f(\eta_i) = \frac{1}{2} \left\{ q|\eta_i|^{q-2} w_i^2 + (2-q)|\eta_i|^q \right\} \tag{12}$$

The new parameters η_i are called variational parameters, and the resulting upper bound is tangent to $|w_i|^q$ in $\eta_i = \pm|w_i|$ — see Figure 1.

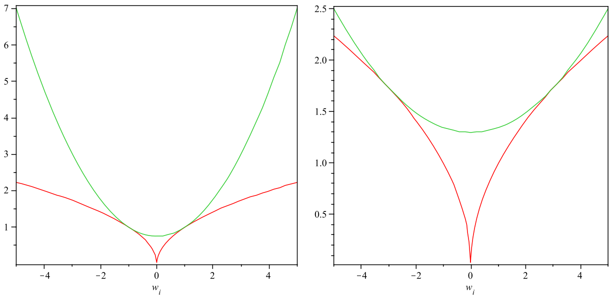

Fig. 1. Examples: Left: $q = 0.5$, $\eta_i = 1$, so the quadratic upper bound is tangent in ± 1. Right: $q = 0.5$, $\eta_i = 3$, so the quadratic upper bound is tangent in ± 3.

It is interesting to note that in the case of $q = 1$, the bound (12) recovers exactly the bound proposed in [4] for deriving an exact estimation algorithm for L1-regularised logistic regression.

Using the above local bounds, the log likelihood is lower-bounded:

$$\begin{aligned}\mathcal{L} &= -\sum_{j=1}^{n} \log\left\{1+\exp(-y_j \boldsymbol{w}^T \boldsymbol{x}_j)\right\} - \alpha \sum_{i=1}^{m} |w_i|^q \\ &\geq -\sum_{j=1}^{n} \log\left\{1+\exp(-y_j \boldsymbol{w}^T \boldsymbol{x}_j)\right\} - \alpha \sum_{i=1}^{m} \frac{1}{2}\left\{q|\eta_i|^{q-2}w_i^2 + (2-q)|\eta_i|^q\right\} \\ &= \tilde{\mathcal{L}}^{quad}(\boldsymbol{w}, \boldsymbol{\eta}) \qquad (13)\end{aligned}$$

Now, maximising the lower bound to the log likelihood can be done iteratively. Each iteration will alternate between maximising w.r.t. $\boldsymbol{w}$ while keeping $\boldsymbol{\eta}$ fixed and maximising w.r.t. the variational parameters $\boldsymbol{\eta}$, i.e. tightening the variational bound while keeping $\boldsymbol{w}$ fixed. Convergence to a local optimum is guaranteed, convergence proofs for this kind of algorithms are detailed in e.g. [6] and [3], and are essentially based on the fact that the sequence of log likelihood estimates is non-decreasing and bounded from above (analogous to E-M algorithms).

Consider first the maximisation w.r.t. the variational parameters $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_n)$, with $\boldsymbol{w}$ being fixed to their current value. Solving the stationary equations w.r.t. η_i, i.e. $\frac{\partial \tilde{\mathcal{L}}^{quad}}{\partial \eta_i} = 0$ yields:

$$\boldsymbol{\eta} = |\boldsymbol{w}| \qquad (14)$$

This is indeed where we have seen the bound touches the function.

Now, the maximisation w.r.t. w, with fixed $\boldsymbol{\eta}$ is technically an L2-regularised logistic regression problem, since (13) depends quadratically on $\boldsymbol{w}$, which may be carried out using existing methods. One convenient option, which we briefly reproduce here for completeness, is to employ the local quadratic lower bound to the log likelihood term as proposed in [6], based on the fact that the logistic function is convex as a function of the square root of its argument. This is the following,

$$\begin{aligned}-\log\left\{1+\exp(-y_j \boldsymbol{w}^T \boldsymbol{x}_j)\right\} \geq -\frac{1}{4\xi_j}\tanh(\frac{\xi_j}{2})\left\{(\boldsymbol{w}^T \boldsymbol{x}_j)^2 - \xi_j^2\right\} \\ + (y_j \boldsymbol{w}^T \boldsymbol{x}_j - \xi_j)/2 - \log\left\{1+\exp(\xi_j)\right\} \qquad (15)\end{aligned}$$

where $\xi_j, j = 1, ..., n$ are new variational parameters that control the tightness of the locally quadratic bound on the log likelihood. Replacing this into (13), we obtain a lower bound expression on the log likelihood (parameterised by both $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$) which we can maximise iteratively by solving for $\boldsymbol{w}$ and $\boldsymbol{\xi}$ in turn. These optimisation problems now have closed form solutions, and we get:

$$\begin{aligned}\boldsymbol{w} &= [\boldsymbol{X}\boldsymbol{\Xi}\boldsymbol{X}^T + \boldsymbol{\Lambda}]^{-1}\boldsymbol{X}\boldsymbol{y}/2 \\ &= \left\{\boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{X}[\boldsymbol{X}^T\boldsymbol{\Lambda}\boldsymbol{X} + \boldsymbol{\Xi}]^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}\right\}\boldsymbol{X}\boldsymbol{y}/2 \qquad (16) \\ \boldsymbol{\xi} &= |\boldsymbol{X}^T\boldsymbol{w}| \qquad (17)\end{aligned}$$

where $\boldsymbol{\Lambda} = \text{diag}(\frac{|\eta_i|^{2-q}}{q\alpha})$ and $\boldsymbol{\Xi} = \text{diag}(\frac{2\xi_j}{\tanh(\xi_j/2)})$, and the form (16) is more convenient in the $m >> n$ case. This inner loop can then be interleaved with the re-estimation of $\boldsymbol{\eta}$ and so the overall algorithm consists of performing (16)-(17)-(14) in turn until convergence to a local optimum.

It should be noted that, although each step of the algorithm yields a unique solution for the parameter being re-estimated, the overall solution is not unique since the likelihood (and the bound on it) is not convex. There are multiple local optima, and so the initialisation may be important. In the reported experiments, we initialised all variational parameters to one. Then, cf. (16), $\boldsymbol{w}$ is a deterministic function of these values and the training set. While this is just one of several possible reasonable choices, it worked well, and as we shall see, even the local optima obtained is able to produce more accurate classification (and feature recovery) than the alternative convex approach of L_1-regularised logistic regression.

3.3 Method 3: Local Linear Variational Approximation

Now, convex duality [6] will be used to create a linear upper bound on the regularisation term. The idea of creating local linear (rather than quadratic) bounds for this problem was proposed in a recent statistics paper [16], and our use of convex duality is just a convenient framework for deriving variational bounds in a more systematic manner.

With $q < 1$, the function $|w_i|^q$ is concave, so we can write:

$$f(w_i) = |w_i|^q = \min_{\lambda_i} \{\lambda_i |w_i| - f^*(\lambda_i)\} \tag{18}$$

$$f^*(\lambda_i) = \min_{\eta_i} \{\lambda_i |\eta_i| - f(\eta_i)\} \tag{19}$$

Again, the function $f^*(.)$ is the conjugate (or dual) function of $f(.)$, and $f^*(\lambda_i)$ represents the amount of vertical shift to be applied to $\lambda|w_i|$ in order to obtain the linear upper bound with slope λ, that touches $f(w_i)$. For every $\pm|w_i|$, there is an optimal slope λ_i from the family of upper bounds.

Denoting $g(\eta_i) = \lambda_i|\eta_i| - f(\eta_i)$, the maximum occurs either at $\eta_i = 0, g(\eta_i = 0) = 0$ or at a solution of the stationary equation when $\eta_i \neq 0$:

$$g'(\eta_i) = \lambda_i \text{sign}(\eta_i) - f'(\eta_i) = 0 \quad \Rightarrow \tag{20}$$

$$\lambda_i = \frac{f'(\eta_i)}{\text{sign}(\eta_i)} = q|\eta_i|^{q-1} \quad \text{and} \tag{21}$$

$$f^*(\lambda_i) \leq (q-1)|\eta_i|^q \tag{22}$$

Replacing in (18) yields the variational bound:

$$|w_i|^q \leq q|\eta_i|^{q-1}|w_i| + (1-q)|\eta_i|^q \tag{23}$$

The new parameters η_i are called variational parameters, and the resulting upper bound is tangent to $|w_i|^q$ in $\eta_i = \pm|w_i|$ — see Figure 2.

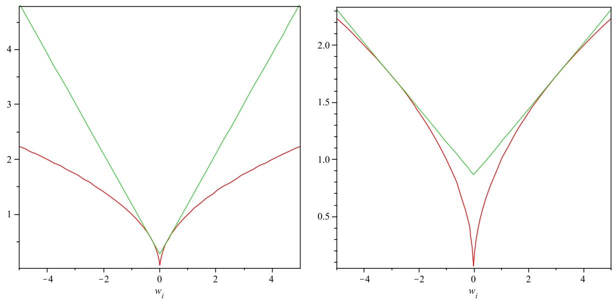

Fig. 2. Examples - Left: $q = 0.5$, $\eta_i = 0.3$, so the linear upper bound is tangent in ± 0.3. Right: $q = 0.5$, $\eta_i = 3$, so the linear upper bound is tangent in ± 3.

This variational approximation casts back the initial problem of fractional regularisation to solving a number of L1-regularised problems instead.

$$\begin{aligned}\mathcal{L} &= -\sum_{j=1}^{n} \log\left\{1+\exp(-y_j \boldsymbol{w}^T \boldsymbol{x}_j)\right\} - \alpha \sum_{i=1}^{m} |w_i|^q \\ &\geq -\sum_{j=1}^{n} \log\left\{1+\exp(-y_j \boldsymbol{w}^T \boldsymbol{x}_j)\right\} - \alpha \sum_{i=1}^{m} \left\{q|\eta_i|^{q-1}|w_i| + (1-q)|\eta_i|^q\right\} \\ &= \tilde{\mathcal{L}}^{lin}(\boldsymbol{w}, \boldsymbol{\eta}) \end{aligned} \tag{24}$$

As before, maximising the lower bound to the log likelihood can be done iteratively. Each iteration will alternate between maximising w.r.t. $\boldsymbol{w}$ — which is now an L1-regularised version of the problem — and maximising w.r.t. the variational parameters $\boldsymbol{\eta}$, i.e. tightening the bound. Convergence to a local optimum is guaranteed, proofs are detailed in e.g. [6] and [16].

Maximising w.r.t. the variational parameters is straightforward, and after some algebra we get the stationary equation,

$$\frac{\partial \tilde{\mathcal{L}}^{lin}}{\partial \eta_i} = q(q-1)|\eta_i|^{q-2} sign(\eta_i)(|w_i| - |\eta_i|) = 0 \tag{25}$$

the solution of which is, as one would expect, $|\eta_i| = |w_i|$. Due to symmetry, since the objective is a function of $|\eta_i|$, it is enough to take:

$$\boldsymbol{\eta} = |\boldsymbol{w}| \tag{26}$$

For maximising w.r.t. $\boldsymbol{w}$, we can use any of the several existing efficient methods for solving an L_1-regularised logistic regression. The algorithm is then to iterate between these two steps till convergence, and we see this requires us to solve an L1-regularised regression problem in each iteration. However, in the sequel we shall show instead, that — perhaps surprisingly — the method described in this section is actually equivalent to the previous one.

Equivalence of Local Linear and Local Quadratic Approximation. Of course, the local linear approximation appears to be a tighter bound to the $L_{q<1}$-term than the local quadratic one, which has been the main motivation for [16] proposing it. However, we will show in fact they are both finding a local optimum of the same objective. To see this, we develop a generalised E-M [9] estimation algorithm for the local linear bound, which will turn out to be identical to the iterative estimation algorithm we have developed for the local quadratic bound.

Let us start by rewriting the $|w_i|$ term as follows:

$$|w_i| = -\log \int_0^\infty \frac{1}{\sqrt{2\pi\tau_i}} \exp\left\{-\frac{w_i^2 + \tau_i^2}{2\tau_i}\right\} d\tau_i \tag{27}$$

$$= -\log\left\{2\int_0^\infty N(w_i|0,\tau_i) Exp(\tau_i) d\tau_i\right\} \tag{28}$$

where $N(w_i|0,\tau_i) = \frac{1}{\sqrt{2\pi\tau_i}}\exp(-w_i^2/(2\tau_i))$ is the Gaussian density with zero men and variance τ_i and $Exp(\tau_i) = \frac{1}{2}\exp(-\tau_i/2)$ is the exponential density with parameter 1.

In this rewriting, τ may be seen as a latent variable, so we could use the E-M methodology for iterative estimation of $\boldsymbol{w}$ from (24) (in an inner-loop) while keeping $\boldsymbol{\eta}$ fixed at its currently estimated value (from the outer loop). We should emphasise, this is not a proposal for a practical algorithm, but serves to show the equivalence relationship with the local quadratic approximation approach — which is obviously more convenient to implement.

It is well known from the theory of E-M [9] that the expectation of the log complete likelihood forms a so-called auxiliary function, meaning that an increase in this function corresponds to an increase in the initial objective (in our case $\tilde{\mathcal{L}}^{lin}(\boldsymbol{w},\boldsymbol{\eta})$), and a local optimum of the auxiliary function is also a local optimum of the initial objective. For (24), the expected log complete likelihood is the following:

$$Q(\boldsymbol{w}, p(\boldsymbol{\tau}|\boldsymbol{w}^{old}), \boldsymbol{\eta}) = -\sum_{j=1}^{n}\log\{1+\exp(-y_j\boldsymbol{w}^T\boldsymbol{x}_j)\} + \alpha q\sum_{i=1}^{m}|\eta_i|^{q-1}\int_0^\infty p(\boldsymbol{\tau}|\boldsymbol{w}^{old}) \times$$
$$\times \{\log N(w_i|0,\tau_i) + \log Exp(\tau_i)\} + const_{\boldsymbol{w},\boldsymbol{\tau}} \tag{29}$$

where $const_{\boldsymbol{w},\boldsymbol{\tau}}$ is independent of both $\boldsymbol{w}$ and $\boldsymbol{\tau}$.

The E-step of this inner loop estimation procedure would then be to compute the posterior $p(\tau_i|w_i^{old})$ and the M-step would maximise Q with respect to $\boldsymbol{w}$. Observe, however, that the only posterior statistic required for the estimation of $\boldsymbol{w}$ is the expectation $E[1/\tau_i|w_i^{old}]$ w.r.t. $p(\tau_i|w_i^{old})$. Thus, computing this completes the E-step:

$$E[1/\tau_i|w_i^{old}] = \frac{\int_0^\infty 1/\tau_i N(w_i^{old}|0,\tau_i)Exp(\tau_i)d\tau_i}{\int_0^\infty N(w_i^{old}|0,\tau_i)Exp(\tau_i)d\tau_i} = \frac{1}{|w_i^{old}|} \tag{30}$$

Hence we have, in more compact notation:

$$E[\boldsymbol{\tau}^{-1}|\boldsymbol{w}^{old}] = |\boldsymbol{w}^{old}|^{-1} \tag{31}$$

which, after one E-step, is identical to the inverse of the estimate of $\boldsymbol{\eta}$ that we obtained previously (26).

The M-step is to compute:

$$\begin{aligned} \boldsymbol{w} &= \underset{\boldsymbol{w}}{\operatorname{argmax}}\, Q(\boldsymbol{w}, E[\boldsymbol{\tau}^{-1}|\boldsymbol{w}^{old}], \boldsymbol{\eta}) && (32)\\ &= \underset{\boldsymbol{w}}{\operatorname{argmax}} - \sum_{j=1}^{n} \log\left\{1 + \exp(-y_j \boldsymbol{w}^T \boldsymbol{x}_j)\right\} - \alpha q \sum_{i=1}^{m} \frac{1}{2} |\eta_i|^{q-1} \mathrm{E}[1/\tau_i | w_i^{old}] w_i^2 && (33) \end{aligned}$$

where we replaced (31) into the terms of Q that depend on $\boldsymbol{w}$. Observe, in the first M-step this is exactly the same as $\underset{\boldsymbol{w}}{\operatorname{argmax}}\, \tilde{\mathcal{L}}^{quad}(\boldsymbol{w}, \boldsymbol{\eta})$.

Now, rather than *maximising* $\tilde{\mathcal{L}}^{lin}(\boldsymbol{w}, \boldsymbol{\eta})$ w.r.t. $\boldsymbol{w}$ with $\boldsymbol{\eta}$ fixed, by iterating the E and M steps to convergence in an inner-loop (and reestimate $\boldsymbol{\eta}$ in the outer loop), it is sufficient, cf. the generalised E-M [9], to make an *increase* in $\tilde{\mathcal{L}}^{lin}(\boldsymbol{w}, \boldsymbol{\eta})$ w.r.t. $\boldsymbol{w}$ before reestimating $\boldsymbol{\eta}$. This leads to merging the inner and outer loops in a single loop, while still having the guarantee of a monotonic convergence to a local optimum of $\tilde{\mathcal{L}}^{lin}(\boldsymbol{w}, \boldsymbol{\eta})$ w.r.t. $\boldsymbol{w}$. In particular, we have the following convergent sequence, where t is the iteration index:

$$\begin{aligned} \tilde{\mathcal{L}}^{lin}(\boldsymbol{w}^{t-1}, \boldsymbol{\eta}^{t-1}) &\leq \tilde{\mathcal{L}}^{lin}(\boldsymbol{w}^{t-1}, \boldsymbol{\eta}^{t}) = Q(\boldsymbol{w}^{t-1}, \mathrm{E}[\boldsymbol{\tau}^{-1}|\boldsymbol{w}^{t-1}], \boldsymbol{\eta}^{t})\\ &\leq Q(\boldsymbol{w}^{t}, \mathrm{E}[\boldsymbol{\tau}^{-1}|\boldsymbol{w}^{t-1}], \boldsymbol{\eta}^{t}) = \tilde{\mathcal{L}}^{quad}(\boldsymbol{w}^{t}, \boldsymbol{\eta}^{t})\\ &\leq Q(\boldsymbol{w}^{t}, \mathrm{E}[\boldsymbol{\tau}^{-1}|\boldsymbol{w}^{t}], \boldsymbol{\eta}^{t})\\ &= \tilde{\mathcal{L}}^{lin}(\boldsymbol{w}^{t}, \boldsymbol{\eta}^{t}) \leq \tilde{\mathcal{L}}^{lin}(\boldsymbol{w}^{t}, \boldsymbol{\eta}^{t+1}) = Q(\boldsymbol{w}^{t}, \mathrm{E}[\boldsymbol{\tau}^{-1}|\boldsymbol{w}^{t}], \boldsymbol{\eta}^{t+1})\\ &\leq Q(\boldsymbol{w}^{t+1}, \mathrm{E}[\boldsymbol{\tau}^{-1}|\boldsymbol{w}^{t}], \boldsymbol{\eta}^{t+1}) = \tilde{\mathcal{L}}^{quad}(\boldsymbol{w}^{t+1}, \boldsymbol{\eta}^{t+1}) \leq \ldots \end{aligned}$$

Now, observe, this is in fact identical to the monotonic sequence produced when optimising the local quadratic bound. Hence there is no advantage to the extra sophistication brought by the local linear approximation. Because of this, we implemented and used the former in the reported experiments.

4 Experiments

From the theoretical analysis we found they both enjoy a logarithmic sample complexity w.r.t. the data dimension and so they can both learn with exponentially many irrelevant features. We may also expect in the light of that analysis that a smaller q should work best in a setting where the number of relevant features is very small. However, to find out how $L_{q<1}$-regularisation compares with L_1 regularisation in such a setting, we conduct an empirical comparison in this section.

We generated synthetic data sets as in [10] and followed the same experimental protocol in the first instance. In each experiment, 30% of the data was used as a validation set to select the regularisation parameter. The training + validation set size was 70+30 and the performance was measured on an independent test set of size 100. The number of features was varied between 100 and 1000, out

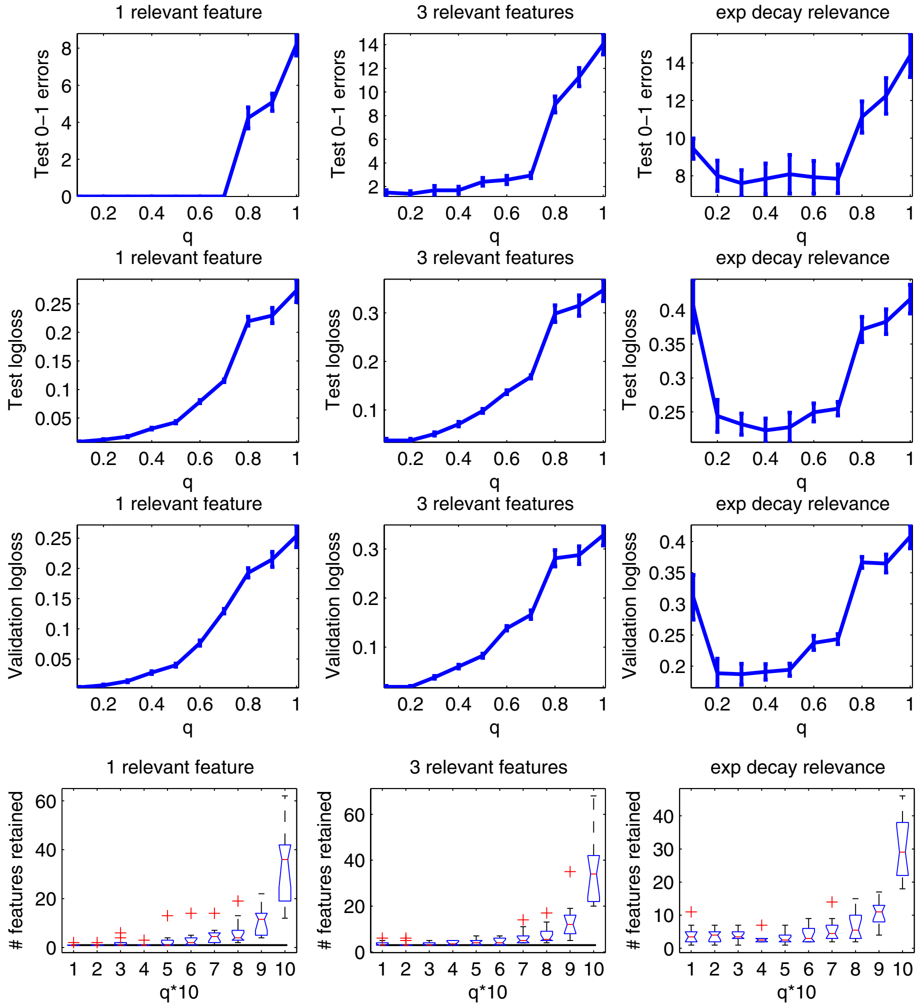

Fig. 3. Results on synthetic data sets when varying q. The training set size was $n_1 = 70$, the validation set size=30, and the out-of-sample test set size=100. The statistics are over 10 independent runs with overall feature size (dimensionality) ranging from 100 to 1000. The upper figures represent average ± standard error.

of which just a few are relevant, as follows. We created and studied the same 3 types of data sets as in [10]: (1) a single feature is relevant; (2) 3 features are relevant; and (3) exponentially decaying relevance of the features. See [10] for more details on the data generation procedure. Figure 3 summarises the results, measured by three different criteria: the number of 0-1 errors on the test set (out of 100), the logloss achieved on the test set and the number of features retained vs. the true number of relevant features. We also plotted the loglosses on validation set, to ascertain these are in good agreement with the

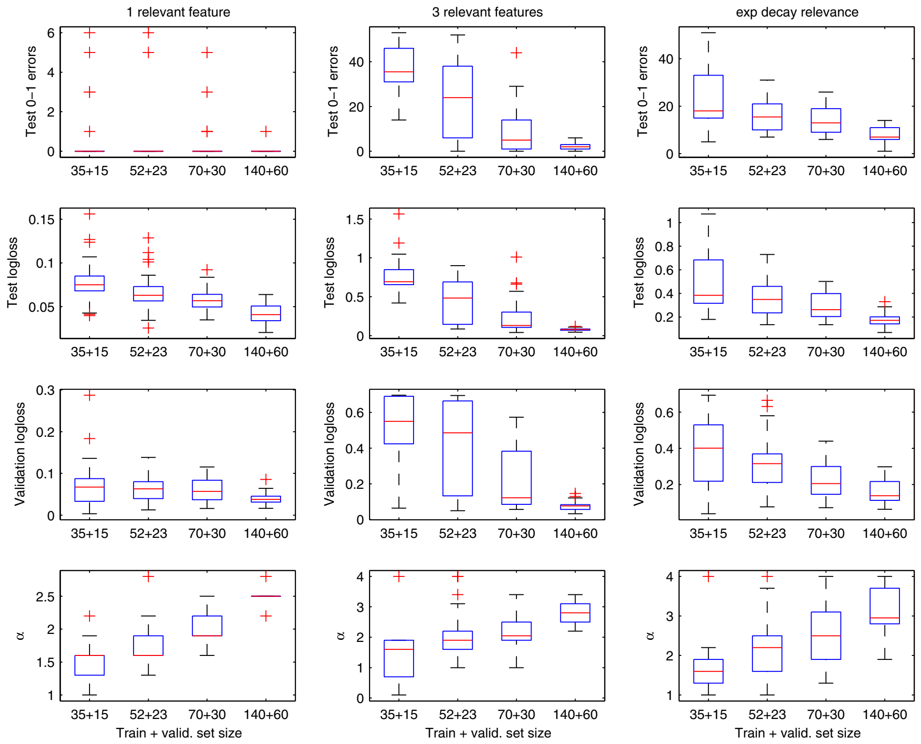

Fig. 4. Varying the training + validation set size. q=0.5, m=3000. The box-plots summarise 20 independent trials for each experiment.

other measures. For each q, the error bars represent averages and standard errors of results obtained for all data dimensions tested. As we can see, the variation w.r.t. this factor turns out to be much smaller than that w.r.t. varying q. It is clear from the figure that small values of q work extremely well when only one feature is relevant (column 1), despite of the large number of relevant dimensions. The improvement achieved over the L_1-regularised logistic regression is both statistically and practically significant, w.r.t. all measures. The picture is very similar when 3 features are relevant (column 2). Moreover, we notice that for certain values of q, the fractional-norm regularisation is also able to improve over L_1 in the case of exponentially decaying feature relevance (column 3). However, unsurprisingly, small values are not the best in this latter case — since in this case all features have some degree of information about the target, a relatively small but not too small number of them is needed for good prediction.

Next, noting that the theoretical guarantees assume a large number of examples, and although the previous set of experiments considered a small but fixed sample size throughout, we conducted an additional set of experiments designed to test the variability of L_q-regularised logistic regression when varying the number of samples in both directions. Here $q = 0.5$ is fixed, as from the previous experiments we observe this value tends to win over on average. Figure 4 presents these results for all three types of data sets previously considered. The overall dimensionality is 3000 this time, so these data sets must be harder than the previous ones. The size of the independent test set is 100 in all cases, so the misclassifications reported are again out of 100.We see the results remain reasonably good as long as the ratio of sample size vs. relevant dimensions is not too small. The results with a single relevant feature in 3000 dimensions (column 1) are in fact excellent — even with very small sample sizes the median of the 0-1 missclassification error rate is still zero. In turn, as the number of relevant features increases to 3 and the training + validation size is as small as 35+15, we see in the leftmost box-plot (middle column) this is where the method reaches its limits and ceases to work.

5 Conclusions

We studied fractional-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 local quadratic and local linear approximations to the regularisation term. Based on our results, fractional-norm regularisation is more suitable than L1 in cases when the number of relevant features is very small, and works very well despite a large number of irrelevant features.

A Word about a Bayesian Interpretation. $L_{q<1}$ regularisation may be interpreted as the MAP estimate of a Generalised Laplacian Distribution prior. Our variational approach makes it tractable to compute variational posterior distributions and hence to obtain uncertainty estimates. Whether those will keep having such favourable sample complexity properties or not, remains to be elucidated in the future.

Acknowledgments. RJD was supported by an EPSRC CTA studentship for MSc in Natural Computation.

References

1. Anthony, M., Bartlett, P.L.: Neural Network Learning: Theoretical Foundations. Cambridge University Press, Cambridge (2001)
2. Chartrand, R.: Exact reconstructions of sparse signals via non-convex minimization. IEEE Signal Process. Lett. 14, 707–710 (2007)
3. Fan, J., Li, R.: Variable Selection via Non-concave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association, Theory and Methods 96(456) (December 2001)
4. Krishnapuram, B., Carin, L., Figueiredo, M., Hartemink, A.: Learning sparse Bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6), 957–968 (2005)

5. François, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE Trans. on Knowledge and Data Engineering 19(7) (July 2007)
6. Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: An Introduction to Variational Methods for Graphical models. In: Jordan, M. (ed.) Learning in Graphical Models. The MIT Press, Cambridge (1998)
7. Liu, Z., Jiang, F., Tian, G., Wang, S., Sato, F., Meltzer, S.J., Tan, M.: Sparse Logistic Regression with Lp Penalty for Biomarker Identification. Statistical Applications in Genetics and molecular Biology 6(1) (2007)
8. Bradley, P.S., Mangasarian, O.L.: Feature Selection via Concave Minimization and Support Vector Machines. In: ICML 1998, pp. 82–90 (1998)
9. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. JohnWiley and Sons, New York (1997)
10. Ng, A.: Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In: ICML 2004 (2004)
11. Pollard, D.: Empirical Processes:Theory and Applications. Springer, Heidelberg (1984)
12. Tipping, M.: Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research 1, 211–244
13. Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M.: Use of the Zero-Norm with Linear Models and Kernel Methods. Journal of Machine Learning Research 3, 1439–1461 (2003)
14. Wipf, D.P., Rao, B.D.: ℓ_0-Norm Minimization for Basis Selection. In: Saul, L., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 17. MIT Press, Cambridge (2005)
15. Zhang, T.: Covering number bounds of certain regularized linear function classes. Journal of Machine Learning Research 2, 527–550 (2002)
16. Zou, H., Li, R.: One-step sparse estimates in non-concave penalized likelihood models. The Annals of Statistics (2008)

Appendix

Proof of Theorem 1

First, it is useful to notice that

$$\forall \boldsymbol{w}, ||\boldsymbol{w}||_{q<1} \geq ||\boldsymbol{w}||_1 \tag{34}$$

The plan is then the following. We show that $h \in H$ is bounded, so we can apply standard results to bound its error probability by a uniform covering number. This will then be bounded further using (34) and an existing result of [15] for regularised function classes, in the same manner as previously employed in [10]. Finally, the sufficient sample complexity is computed by requiring the error bound to be smaller or equal than the user-defined confidence parameter δ.

To see (34), note first that the inequality $||\boldsymbol{w}||_q \geq ||\boldsymbol{w}||_p, \forall \boldsymbol{w}$ is well known for $1 < q < p$ from measure theory. Extending it to $0 < q < p < 1$ is straightforward, rewrite the required inequality as follows:

$$\left(\sum_{i=1}^{m}|w_i|^q\right)^{1/q} \geq \left(\sum_{i=1}^{m}|w_i|^p\right)^{1/p}$$

$$\left(\sum_{i=1}^{m}|\frac{w_i}{(\sum_{i=1}^{m}|w_i|^p)^{1/p}}|^q\right)^{1/q} \geq \left(\sum_{i=1}^{m}|\frac{w_i}{(\sum_{i=1}^{m}|w_i|^p)^{1/p}}|^p\right)^{1/p} = 1$$

$$\left(\sum_{i=1}^{m}|u_i|^q\right)^{1/q} \geq \left(\sum_{i=1}^{m}|u_i|^p\right)^{1/p} = 1$$

where we denoted $|u_i| \equiv |\frac{w_i}{(\sum_{i=1}^{m}|w_i|^p)^{1/p}}|$. Now, since $\sum_{i=1}^{m}|u_i|^p = 1 \Rightarrow |u_i|^p \leq 1 \Rightarrow |u_i| \leq 1$. In consequence, and since $q \leq p$, we have $|u_i|^q \geq |u_i|^p$. Summing both sides w.r.t. i yields $\sum_{i=1}^{m}|u_i|^q \geq \sum_{i=1}^{m}|u_i|^p$. But we know the r.h.s. is one, so $\sum_{i=1}^{m}|u_i|^q \geq 1$. Finally, raising both sides to $1/q$, the required result follows (for any $q > 0$), i.e. $(\sum_{i=1}^{m}|u_i|^q)^{1/q} \geq 1$.

To see that $h \in H$ is bounded, we write:

$$M = |-\log p(y|\boldsymbol{w}^T\boldsymbol{x})| = |\log(1+\exp(-y\boldsymbol{w}^T\boldsymbol{x}))| \leq 1 + |\boldsymbol{w}^T\boldsymbol{x}| \tag{35}$$

$$\leq 1 + ||\boldsymbol{w}||_1||\boldsymbol{x}||_\infty \quad \text{(by Hölder's inequality)} \tag{36}$$

$$\leq 1 + ||\boldsymbol{w}||_{q<1}||\boldsymbol{x}||_\infty \quad \text{cf. (34)} \tag{37}$$

$$\leq 1 + A||\boldsymbol{x}||_\infty \tag{38}$$

Therefore, the classical result due to Pollard [11] (pp. 492), combined with standard results [10,1] applies, and we have the error probability bounded by a uniform covering number of the linear function class G in the L2-norm:

$$P\{\exists h \in H : |er_P(h) - \hat{er}_{z_1}(h)| > \epsilon\} \leq 8\mathcal{N}_2(G, \frac{\epsilon}{8L}, n_1)\exp\left\{-\frac{n_1\epsilon^2}{512M^2}\right\} \tag{39}$$

where M is the output bound that we previously computed and L is the Lipschitz constant of $h \in H$ as a function of $g \in G$. Since h is continuous on $[0,1]$ and differentiable on $(0,1)$, and $|h'(t)| = |\frac{d}{dt}\log\frac{1}{1+e^{-t}}| = |\frac{e^{-t}}{1+e^{-t}}| \leq 1$, therefore h satisfies the first order Lipschitz condition with Lipschitz constant L=1.

Now, in the above, it remains to approximate the uniform covering number[2] $\mathcal{N}_2(G, \epsilon/8, n_1)$. This is a combinatorial quantity that expresses the complexity of a function class at the given scale, and as such, it is affected by the regularisation constraints imposed. Similarly to the approach in [10], the following result by

[2] The definition of uniform covering numbers is as follows (see e.g. [1]). B is called an ε-cover for a real valued function class F of infinite cardinality (e.g. a parameterised function class) w.r.t. a particular training set z of size n and a distance d, if B is a finite set of functions, and $\forall f \in F, \exists b \in B$, such that $d(f(\boldsymbol{x}) - b(\boldsymbol{x})) < \varepsilon$. The size of the smallest such cover set B is the covering number of F w.r.t. z. The uniform covering number is then the maximum of these w.r.t. all training sets of size n and is denoted $\mathcal{N}_d(F, \varepsilon, n)$.

Zhang [15], developed for regularised linear function classes, can be used to bound this uniform covering number.

Lemma [15]. In a linear function class $G = \{g : g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}, \boldsymbol{x} \in \mathcal{R}^m\}$, if $||\boldsymbol{x}||_p \leq b$ and $||\boldsymbol{w}||_q \leq a$, where $1/p + 1/q = 1$ and $p \geq 2$, then

$$\log_2 \mathcal{N}(G, \epsilon, n) \leq \text{ceil}(\frac{a^2b^2}{\epsilon^2}) \log_2(2m+1)$$

where n is the sample size and m is the data dimension.

Clearly, the lemma does not apply directly to our case, since $1/q > 1$ already when $q < 1$. However using again (34), it follows by transitivity that $||\boldsymbol{w}||_1 \leq A$. Now applying the Lemma, we get:

$$\log_2 \mathcal{N}(G^{(q<1)}, \epsilon, n) \leq \text{ceil}(\frac{A^2b^2}{\epsilon^2}) \log_2(2m+1)$$

where the $L_{q<1}$-regularised linear function class is denoted by $G^{(q<1)}$, and $b = \max_i ||\boldsymbol{x}_{i\geq n}||_\infty$ and $||\boldsymbol{x}||_\infty \equiv \max_{j\geq m} |x_j|$.

Replacing these results into the generalisation bound (39) and assuming the data is normalised such that $||\boldsymbol{x}_i||_\infty \leq 1$, we get

$$P\{\exists h \in H : |er_P(h) - \hat{er}_{z_1}(h)| > \epsilon\} \leq 8 \times 2^{64A^2/\epsilon^2+1}(2m+1)\exp\left\{-\frac{n_1\epsilon^2}{512(1+A)^2}\right\} \tag{40}$$

Since we want this to be small, we set the r.h.s. of (40) to be smaller or equal to a desired (user-prescribed) confidence parameter δ, we seek to ensure that h is uniformly ϵ-good with probability at least $1-\delta$. Then, with high probability, i.e. with probability $1-\delta$, we have, from (40), that:

$$\forall h \in H,\ |er_P(h) - \hat{er}_{z_1}(h)| \leq \epsilon \tag{41}$$

Now, it is a standard result to show that from (41) it follows that $er_P(L(z_1))$ is close to $opt_P(H)$, which follows below for completeness. Indeed, if (41) holds then it must hold also for our learning algorithm $h = L(z_1)$. Hence, applying (41), the definition of $L(z_1)$, and the definition of $opt_P(H)$, we get:

$$\begin{aligned} er_P(L(z_1)) &\leq \hat{er}(L(z_1)) + \epsilon = \min_{h\in H} \hat{er}_{z_1}(h) + \epsilon \\ &\leq \hat{er}_{z1}(h^*) + \epsilon \leq er_P(h^*) + 2\epsilon = \inf_{h\in H} er_P(h) + 2\epsilon = opt_P(H) + 2\epsilon \end{aligned}$$

Thus,

$$er_P(L(z_1)) \leq opt_P(H) + 2\epsilon \tag{42}$$

which is indeed of the form (4) if we replace ϵ by $\epsilon/2$.

Finally, solving for n_1 the r.h.s. of (40)= δ when ϵ is replaced by $\epsilon/2$, yields the sample complexity of $L(z_1)$, i.e. the amount of data required for good generalisation:

$$n_1(L, \epsilon, \delta) = \frac{2048(A+1)^2}{\epsilon^2}[\log\frac{8(2m+1)}{\delta} + \frac{256A^2}{\epsilon^2} + 1] \tag{43}$$

We can now conclude that the sample complexity of fractional norm regularised logistic regression with a given regularisation parameter A is $n_1 = \Omega(\log(m) \times \text{poly}(A, 1/\epsilon, \log(1/\delta)))$. Moreover, following the argument in [10] to express n_1 as a linear function of n, i.e. $n_1 = (1-\nu)n$ where $\nu < 1$ is a constant, the sample complexity result obtained for n_1 also extends to n, i.e. we have $n = \Omega(\log(m) \times \text{poly}(A, 1/\epsilon, \log(1/\delta)))$. Q.E.D.