Approximate Cross Validation for Sparse Generalized Linear Models

by

William T. Stephenson

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2019

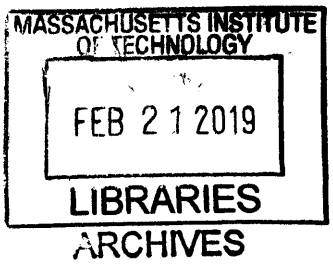
Author .. Signature redacted .....
Department of Electrical Engineering and Computer Science
January 4, 2019

Signature redacted

Certified by... .....
Tamara Broderick
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by ....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students



Approximate Cross Validation for Sparse Generalized Linear Models

by

William T. Stephenson

Submitted to the Department of Electrical Engineering and Computer Science
on January 4, 2019, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Cross validation (CV) is an effective yet computationally expensive tool for assessing the out of sample error for many methods in machine learning and statistics. Previous work has shown that methods to approximate CV can be very accurate and computationally cheap, but only for low dimensional problems. In this thesis, a modification of existing methods is developed to extend the high accuracy of these techniques to high dimensional settings.

Thesis Supervisor: Tamara Broderick
Title: Assistant Professor of Electrical Engineering and Computer Science

Contents

List of Figures

Chapter 1

Introduction

Cross validation is a useful tool for assessing the accuracy of many methods in machine learning and statistics. Although generically applicable and straightforward to use, it has the downside of requiring many re-fittings of the same model. As machine learning models often use as much computation as possible, even a single fitting can be very time consuming. To this end, practitioners typically use k-fold CV with a small k (e.g., five or ten), as this only requires k re-fittings of the model. While k-fold CV is more computationally efficient, leave-1-out CV is known to be asymptotically more accurate for assessing the out of sample error [3, 1]. Unfortunately, leave-1-out CV is hugely computationally expensive, as it requires N refittings, where N is the number of observed datapoints.

To this end, the topic of approximate leave-1-out CV has recently become an active area of research [2, 12, 17, 4]. The core methods proposed by these four works are fairly similar (see Chapter 2 for an overview), have been shown to be empirically successful, and some work has been done to show their accuracy theoretically. [2] demonstrate that, given both the parameter-space and data-space are bounded, the approximation error is $O_p(1/N^2)$, for the amount of data N growing and the dimension of the parameter θ to be estimated, $D = \dim\theta$, staying fixed. [12] make some less and some more restrictive assumptions than [2] and are able to recover an error scaling of $O_p(1/\sqrt{N})$ for the high-dimensional case of D/N converging to a constant. Finally, with the exception of assuming a fixed dimension, [4] make the least restrictive

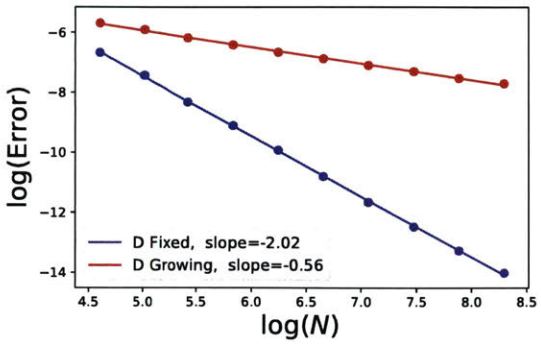


Figure 1-1: Scaling of existing methods for approximate cross validation for unregularized linear regression. When the dimension D of the regression parameters is in constant ratio with the number of observations, $D/N = 1/10$, we see that the approximation error goes down at the $O(1/\sqrt{N})$ rate described in [12]. With a fixed dimension of $D = 2$, however, we see the error goes down at the significantly faster rate of $O(1/N^2)$ as described in [2]. We describe conditions under which we can recover the same $O(1/N^2)$ for a dimension that grows as $o(e^N)$.

assumptions, and obtain an error rate of $O_p(1/N)$.

Our key takeaway from the above works is that all recently studied approaches to approximate cross-validation are very accurate with a small, fixed dimension, but have either unstudied or significantly reduced accuracy when the dimension of the parameters is high relative to the amount of data; see Fig. 1-1 for an illustration of this in the case of linear regression. We argue that this is a fairly important point, as the major cases of interest when N is large are also high dimensional. In particular, if N is significantly larger than D, the training loss is usually a good approximation to the test error, and any form of CV is relatively unnecessary; for example, well known results in empirical risk minimization imply that, for fixed dimension, the training error fairly quickly approaches the true error (see, e.g., [14]).

A common theme in high dimensional statistics and machine learning is to assume that a high dimensional problem really has a much lower "effective dimension." For example, the high dimensional data may be low rank, or the true parameter underlying the data may be sparse. Since parameter estimation can be made significantly more accurate in the presence of such low effective dimension, one might wonder if the same can be said of approximate CV. The example in Fig. 1-1 provides evidence that this

is not immediately true: the higher dimensional problem shown there (the blue line) has a sparse true parameter but still suffers from a reduction in accuracy.

Our main contribution is to demonstrate one case in which this notion of effective dimension is helpful for approximate CV – that of ℓ_1 regularized generalized linear models (GLMs) with data generated by a sparse parameter. The theoretical properties of ℓ_1 regularization and other sparsity inducing regularizers has been extensively studied [11, 10, 8, 16]; it is generally shown that, under certain conditions on the amount of regularization and the structure of the problem, the recovered parameter $\hat{\theta}$ is very accurate and/or has the correct support. Drawing on this body of work, we show that ℓ_1 regularized GLMs have special structure that is not present when using other regularizers that allows approximate CV methods to be highly accurate, even when the dimension grows with N. In particular, we propose a modification of existing approximations and prove conditions under which it successfully takes advantage of the sparse structure in ℓ_1 regularized problems to obtain an accuracy and computational expense equivalent to that of a small, fixed dimensional problem. As a major step along the way, we prove conditions under which the exact solution to the ℓ_1 regularized problem remains stable as each datapoint is held out. In experiments with synthetic and real data, we show that this increased accuracy and decreased runtime is realized in practice: the sparsity of ℓ_1 regularized problems allows our approximation to be significantly more accurate and quicker compared to existing approximate CV methods run with different regularizers on the same data.

Chapter 2

Overview of Approximation

At the heart of all previous approximate cross validation methods is a linear approximation to the solution $\hat{\theta}$ of the optimization problem:

$$\hat{\theta} \triangleq \arg\min_{\theta \in \mathbb{R}^D} F(\theta) + \lambda R(\theta), \tag{2.1}$$

where $\lambda \geq 0$ is a regularization parameter, $R : \mathbb{R}^D \to \mathbb{R}$ is some regularization function, and $F : \mathbb{R}^D \to \mathbb{R}$ is some function that decomposes into N terms: $F(\theta) = (1/N)\sum_{n=1}^{N} f_n(\theta)$. In this work, we focus on the case of generalized linear models (GLMs), in which $F(\theta) = (1/N)\sum_n f(x_n^T\theta, y_n)$, where $x_n \in \mathbb{R}^D$ and $y_n \in \mathbb{R}$ are observed data. Let $\hat{\theta}^{\backslash n}$ be the solution to the same problem with the nth datapoint held out and $H(\hat{\theta})$ the Hessian of $F + \lambda R$ evaluated at $\hat{\theta}$. Then, if we assume f and R are twice differentiable and $F(\theta) + \lambda R(\theta)$ is strongly convex at $\hat{\theta}$, the following approach from [4] gives an approximation $\tilde{\theta}^{\backslash n}$ to $\hat{\theta}^{\backslash n}$:

$$\hat{\theta}^{\backslash n} \approx \tilde{\theta}^{\backslash n} := \hat{\theta} - H(\hat{\theta})^{-1}\nabla_\theta f(x_n^T\hat{\theta}, y_n). \tag{2.2}$$

An alternative is to use the approach given by [2, 12, 17]:

$$\hat{\theta}^{\backslash n} \approx \hat{\theta} - \left(H(\hat{\theta}) - \nabla_\theta^2 f(x_n^T\hat{\theta}, y_n)\right)^{-1} \nabla_\theta f(x_n^T\hat{\theta}, y_n). \tag{2.3}$$

In general, this approach requires the inversion of a $D \times D$ matrix per each $\tilde{\theta}^{\backslash n}$ evaluated, whereas Eq. (2.2) only requires a single inversion to evaluate all $\tilde{\theta}^{\backslash n}$. In general, this is a $O(ND^3)$ computational cost to evaluate all $\tilde{\theta}^{\backslash n}$ by Eq. (2.3) versus a $O(D^3 + ND^2)$ cost for using Eq. (2.2) (one matrix inversion plus N matrix multiplications). In the special case of generalized linear models, each $\nabla^2_\theta f$ is actually a rank one matrix, so standard rank-one update formulas give that only one inversion is required; see Appendix A for a derivation and discussion of both approaches. Looking forward to more general cases, however, we prefer to study the generally computationally cheaper approach given by Eq. (2.2). We stress, though, that even a single inversion of a $D \times D$ matrix can be very expensive, as the time complexity of $O(D^3)$ and memory usage of $O(D^2)$ quickly become prohibitively expensive for D in the tens of thousands.

Even if both approximations were computationally cheap in high dimensions, there is a further issue: their accuracy is much poorer in high dimensions. As noted above, Fig. 1-1 shows that, for a fixed dimensional parameter $D = \dim \theta = 2$, the approximation error in Eq. (2.2) goes down as $O(1/N^2)$, whereas for a growing dimension, $D/N = 1/10$, the error only decays as $O(1/\sqrt{N})$. For D much larger than N, we expect the error to stay constant or even grow.

The above observations about computational expense and statistical accuracy can ruin the original point of these types of approximations. Specifically, the original hope is that, for a small fixed computational budget, using Eq. (2.2) or Eq. (2.3) will be significantly more accurate than, say, running k-fold CV or sub-sampling leave-1-out. The above discussion tells us that in high dimensions, a) computing Eq. (2.2) or Eq. (2.3) is probably impossible on a small budget, and b) has dubious usefulness anyway. This motivates our main observation in the next section: through appropriate use of ℓ_1 regularization, we can retain the $O(1/N^2)$ scaling and low computational cost present in low dimensions for nearly arbitrarily large dimension D.

2.1 Approximate CV with ℓ_1 Regularization

A common assumption in high dimensional settings is that only a small number of dimensions of each x_n are actually relevant for predicting the outcomes y_n; that is, there exists some θ^* such that $|\operatorname{supp}\theta^*| \triangleq |S|$ is much smaller than D. One of the most popular choices for R in such settings is $R(\theta) = \|\theta\|_1$ (a.k.a., "the Lasso") due to its excellent empirical and theoretical properties. Intuitively, it typically correctly recovers $\operatorname{supp}\hat{\theta} = \operatorname{supp}\theta^*$, at which point Eq. (2.1) is reduced to a $|S|$, rather than D, dimensional problem.

Unfortunately, Eq. (2.2) is not immediately applicable when $R = \|\cdot\|_1$, as this choice of R is not twice-differentiable. One suggestion put forward by [17, 12] is to use Eq. (2.3) with a smoothed approximation of $\|\theta\|_1$. A less obvious alternative, also given by [17, 12], is to observe that Eq. (2.3) has a closed form as the amount of smoothness goes to zero. While technically one can use a large amount of smoothing to achieve the same effect, we find that this is not achievable in practice due to numerical issues; see Section 4.1 for an empirical illustration. Below, we make two observations about this limiting argument: 1) the same argument immediately applies to the generally more computationally efficient Eq. (2.2), and 2) the limiting version has better statistical and computational properties than a smoothed version of Eq. (2.2).

In order to state this more formally, let $\hat{S} \triangleq \operatorname{supp}\hat{\theta}$ be the support recovered by the full ℓ_1 regularized optimization problem, $X \in \mathbb{R}^{N\times D}$ the data matrix with rows x_n, $X_{\cdot,\hat{S}}$ its submatrix formed by taking only the columns in $\hat{S}$, and finally define

$$\hat{D}_n^{(2)} \triangleq \left.\frac{d^2 f(z, y_n)}{dz^2}\right|_{z=x_n^T\hat{\theta}}. \tag{2.4}$$

Theorem 1. *Define the restricted Hessian of F as $H_{\hat{S}\hat{S}} \triangleq X_{\cdot,\hat{S}}^T \operatorname{diag}(\hat{D}_n^{(2)}) X_{\cdot,\hat{S}}$, and assume without loss of generality that $\hat{S} = \{1, 2, \ldots, |\hat{S}|\}$. If $H_{\hat{S}\hat{S}}$ has strictly positive eigenvalues and one considers any "reasonable" smooth approximation to $\|\theta\|_1$ and*

takes the limit as the amount of smoothness goes to zero, the limit of Eq. (2.2) *is:*

$$\begin{pmatrix} \hat{\theta}^{\setminus n}_{\hat{S}} \\ \hat{\theta}^{\setminus n}_{\hat{S}^c} \end{pmatrix} \approx \begin{pmatrix} \hat{\theta}_{\hat{S}} - H^{-1}_{\hat{S}\hat{S}} \left[\nabla_\theta f(x_n^T \hat{\theta}, y_n) \right]_{\hat{S}} \\ 0 \end{pmatrix}. \tag{2.5}$$

Furthermore, this limiting approximation has the following properties:

1. *It has the same computational expense as an $|\hat{S}|$-dimensional version of Eq.* (2.2).
2. *If the conditions discussed Proposition 2 hold for the full data problem, and those discussed in Proposition 1 hold for each of the leave-1-out problems, then the error in Eq.* (2.5) *behaves as if the underlying problem were $|\hat{S}|$ dimensional.*

Proof. That Eq. (2.5) is the limit of a smoothed version of Eq. (2.2) follows directly from the arguments in either of [17, 12], who derive the limit of a smoothed approximation to Eq. (2.3) (and also give precise meaning to a "reasonable" smoothing of $\|\theta\|_1$). The point about computation follows by noting Eq. (2.5) only acts along $|\hat{S}|$ dimensions, which tells us that we only need to invert and perform multiplications with a $|\hat{S}| \times |\hat{S}|$, rather than $D \times D$, matrix. The point about accuracy follows from Proposition 1 and Proposition 2, which together imply that 1) all leave-1-out problems are really optimization problems restricted to the dimensions $\hat{S}$ (i.e., $\operatorname{supp} \hat{\theta}^{\setminus n} \subseteq \hat{S}$), and 2) the approximation in Eq. (2.5) runs over only these dimensions. □

The most important step in proving this theorem is giving conditions under which Eq. (2.1) and each of the leave-1-out problems are actually $|\hat{S}|$-dimensional optimization problems and our approximation acts only along these dimensions. After having shown this, the conclusions 1) and 2) in its statement are immediate: the accuracy and computational expense behave as if we were dealing with a $|\hat{S}|$-dimensional, rather than D-dimensional, problem because we *actually are* dealing with a $|\hat{S}|$-dimensional problem.

Of course, Theorem 1 is not particularly useful if its conditions are not satisfied by practical examples. The most difficult to check are the conditions of Proposition 1; in Section 3.3 and Section 3.4 we will see that, with appropriate random data

(X, Y), these conditions are satisfied for linear and logistic regression with very high probability.

Chapter 3

Bounds on Approximation Quality

One of the properties that has made the use of ℓ_1 regularization so popular is that, given a true parameter $\theta^* \in \mathbb{R}^D$ with $\operatorname{supp}\theta^* = S$ such that $|S| \ll D$, the optimal solution $\hat{\theta}$ to Eq. (2.1) often has $\operatorname{supp}\hat{\theta} = S$, even when $N \ll D$. In this section we will show an important piece of the proof of Theorem 1: conditions under which not only $\operatorname{supp}\hat{\theta} = S$, but also $\operatorname{supp}\hat{\theta}^{\backslash n} = \operatorname{supp}\hat{\theta} = S$. We start by reviewing a common technique for proving the support recovery properties of ℓ_1 regularized problems. Notably, these proofs proceed by showing that solving ℓ_1 regularized problems is actually the same as solving a $|S|$-dimensional problem, which exactly fits with the discussion right after Theorem 1 above.

3.1 Primal dual witnesses and support recovery

At the heart of many proofs for showing that ℓ_1 regularized problems recover the correct support $S = \operatorname{supp}\theta^*$ is a proof showing that the solution to a lower dimensional optimization problem is the unique solution to Eq. (2.5) [16, 8, 10]. The idea is to consider the "oracle estimator," which for S^c being the complement of S, sets $\hat{\theta}_{S^c} = 0$, and $\hat{\theta}_S$ as the solution to the restricted version of Eq. (2.1)[1]:

[1]The "oracle" here is needed to tell us the unknown set S.

$$\hat{\theta}_S = \underset{\theta_S \in \mathbb{R}^{|S|}}{\arg\min} \sum_{n=1}^{N} f(x_{nS}^T \theta_S, y_n) + \lambda \|\theta_S\|_1 . \tag{3.1}$$

If this problem is strongly convex, this $\hat{\theta}_S$ is unique. If $\hat{\theta} = (\hat{\theta}_S, \mathbf{0})$ satisfies the first order optimality conditions for the un-restricted problem:

$$\frac{1}{N} \sum_n \nabla_\theta f(x_n^T \hat{\theta}, y_n) + \lambda \hat{z} = 0, \tag{3.2}$$

where $\hat{z}$ is some element of the subdifferential $\partial \|\hat{\theta}\|_1$, then this $\hat{\theta}$ is also an optimal solution to Eq. (2.1). The main difficulty comes in showing that this $\hat{\theta}$ is the *unique* optimal solution. One condition for this is given by a lemma from [6]:

Lemma 1 (Lemma 11.2 from [6]). *If the $\hat{z}$ satisfying the first order optimality conditions Eq.* (3.2) *also satisfies $\|\hat{z}_{S^c}\|_\infty < 1$, then the oracle-estimated $\hat{\theta}$ is the unique optimal solution to Eq.* (2.1). *That is, we have* $\operatorname{supp} \hat{\theta} \subseteq \operatorname{supp} \theta^*$.

This has an immediate corollary relevant to our problem:

Corollary 1. *Let $\hat{z}^{\backslash n}$ satisfy the first order optimality conditions Eq.* (3.2) *for the nth leave-1-out problem. If $\max_n \|\hat{z}_{S^c}^{\backslash n}\|_\infty < 1$, then all the leave-1-out problems are really optimization problems in the same $|S|$ dimensional space.*

Even for just the original problem, it is not obvious that $\|\hat{z}_{S^c}\|_\infty < 1$ will hold in many situations; a large amount of work has gone into identifying conditions for which this holds high probability for various M-estimators. After reviewing a recent version of these conditions, we describe our main insight: that the conditions under which it holds for all leave-1-out problems are almost identical to those for the full data problem.

3.2 Conditions under which $\|\hat{z}_{S^c}^{\backslash n}\|_\infty < 1$

Conditions under which $\|z_{S^c}\|_\infty < 1$ holds vary throughout the literature. Studying the behavior of $\|z_{S^c}\|_\infty$ was introduced in [16], where conditions for its success are

specialized to ℓ_1 regularized linear regression. [8] generalizes this argument to other M-estimators, as well as other types of regularization. Most recently, [10] gave conditions specific to ℓ_1 regularization, but that allow for a wide class of M-estimators. We find the conditions in [10] to be the easiest to check as each datapoint is held out and so choose to work with them here. The main result from [10] is:

Proposition 1 (Theorem 5.1 from [10]). *For $\hat{z}$ defined in Eq. (3.2), $\|\hat{z}_{S^c}\|_\infty < 1$ if the following conditions hold:*

1. *(LSSC) F satisfies the (θ^*, N_{θ^*}) locally structured smoothness condition with constant K. This condition, proposed by [10], is not required to understand our results, so we defer its definition to Appendix C.1.*

2. *(Strong convexity) The restricted problem Eq. (3.1) is strongly convex:*

$$\lambda_{min}\left(\nabla^2 F(\theta^*)_{SS}\right) \geq \lambda_{min} > 0 \tag{3.3}$$

   for some λ_{min}.

3. *(Incoherence) A less interpretable, but common in the literature, condition is the* incoherence *condition:*

$$\left\|\nabla F(\theta^*)_{S^c,S}\left(\nabla^2 F(\theta^*)_{SS}\right)^{-1}\right\|_\infty < 1 - \gamma \tag{3.4}$$

   for some $\gamma > 0$.

4. *(Bounded gradient) The gradient of F evaluated at the true parameters θ^* is small relative to the amount of regularization:*

$$\|\nabla F(\theta^*)\|_\infty \leq \frac{\gamma}{4}\lambda \tag{3.5}$$

5. *(λ is sufficiently small) The above are satisfiable with a regularization parameter that is not too large:*

$$\lambda < \frac{\lambda_{min}^2}{4(\gamma+4)^2}\frac{\gamma}{|S|K}, \tag{3.6}$$

with the understanding that puts no constraint on λ if $K = 0$.

An immediate consequence is that, if each of these conditions hold for each $F_{\backslash n}(\theta)$, then we have $\|z_{S^c}^{\backslash n}\|_\infty < 1$ for each datapoint n. The proofs of Theorem 2 and Theorem 3 take the obvious approach of checking that the conditions in Proposition 1 are true for each of the leave-1-out problems. This will require conditions on the data matrix X; intuitively, one might imagine that if one of its rows x_n were particularly "extreme," the conditions of Proposition 1 will not be satisfied for $F_{\backslash n}$. While all our results could be given explicitly in terms of the entries of X, they are not especially interpretable in this form. Instead, we will assume a particular random form for X and study what happens as D and N grow. In particular, we will assume throughout that the data matrix X is comprised of i.i.d. sub-Gaussian entries:

Definition 1. *[Sub-Gaussian random variable [15]] A random variable X is sub-Gaussian with parameter $c_x > 0$ if:*

$$E\left[\exp\left[\frac{X^2}{c_x^2}\right]\right] \leq 2. \tag{3.7}$$

Our results will be stated both in terms of the sub-Gaussian parameters of the data as well as various global constants all of which will be denoted by $C > 0$; these constants are related to various relationships between sub-Gaussian random variables and are completely independent of the problem. We note that high probability results for random data are in some sense the best sort of result one might hope to get about the stability of ℓ_1 regularization under leave-1-out. Specifically, the results of [18] imply that there exist worst-case training datasets (X, Y) for which sparsity inducing methods like ℓ_1 regularization are not stable as each datapoint is left out. In this sense, our Theorem 2 and Theorem 3 can be interpreted as showing that ℓ_1 regularizaion is stable with high probability. In any case, in order to make such an analysis we will need one major assumption:

Assumption 1. *Assume that the incoherence condition Eq. (3.4) holds with high*

probability for the full data problem:

$$\Pr\left[\left\|\nabla F(\theta^*)_{S^c,S}\left(\nabla^2 F(\theta^*)_{SS}\right)^{-1}\right\|_\infty < 1-\gamma\right] \leq e^{-25},$$

for some fixed $\gamma > 0$, where the probability is taken over the random data X, Y.

In general, there is not much understanding of when Eq. (3.4) holds, let alone when it holds with high probability as in Assumption 1. It seems to be standard to assume that Eq. (3.4) holds (starting from its introduction in [19] and continuing in more recent work [8, 10]); we make the somewhat stronger assumption that the condition holds with high probability under a random sub-Gaussian design matrix. While we will not attempt to prove Assumption 1, it is worth noting that it is known to hold for the simplified case of linear regression with an i.i.d. Gaussian design matrix (e.g., see exercise 11.5 of [6]).

3.3 Linear Regression

Under such a random design, we can show that ℓ_1 regularized linear regression recovers the same support as each datapoint is left out with high probability; we do so by checking that the conditions stated in Proposition 1 hold for each leave-1-out problem. Specifically, assume a linear regression model $y_n = x_n^T\theta^* + w_n$, where $x_n \in \mathbb{R}^D$ has i.i.d. c_x-sub-Gaussian components with $E[x_{nd}^2] = 1$ and w_n is c_w-sub-Gaussian. We then have the following:

Theorem 2. *Assume that Assumption 1 holds and that D, as a function of N, grows as $o(e^N)$. Consider the linear regression model above, and set the regularization parameter as:*

$$\lambda = \frac{1}{\alpha - M_J}\sqrt{\frac{c_x^2 c_w^2 \log D}{NC} + \frac{25 c_x^2 c_w^2}{NC}} + \frac{4 c_x c_w (\log(ND) + 26)}{N(\alpha - M_J)} \tag{3.8}$$

where $C > 0$ is a global constant, and M_J is defined as:

$$M_J = O\left(\frac{|S|\sqrt{2c_x \log(D-|S|)}\left(\sqrt{|S|} + \sqrt{2Cc_x \log N}\right)}{N}\right)$$

(for a non-big-O statement, see Appendix C). Then each of the leave-1-out problems has $\operatorname{supp}\hat{\theta}^{\backslash n} \subseteq S$ *with a high, fixed probability. That is,*

$$\Pr\left[\max_n \left\|\hat{z}^{\backslash n}_{S^c}\right\|_\infty = 1\right] \leq 22e^{-25} \tag{3.9}$$

Proof. We check that, for the value of λ given in the theorem statement, the conditions of Proposition 1 are all satisfied with probability at least $1 - 22e^{-25}$. See Theorem 4 in Appendix C for details. □

It is worth noting how the λ given in Eq. (3.8) compares to that typically given for successful support recovery for the original $\hat{\theta}$. Theorem 11.3 of [6] gives the commonly stated condition $\lambda \geq O(\sqrt{\log(D)/N})$ as sufficient for ensuring that $\operatorname{supp}\hat{\theta} \subseteq S$ with high probability in the case of linear regression. Although this is true for any scaling of D and N, the error in $\hat{\theta}$, $\|\hat{\theta} - \theta^*\|_2$, is usually proportional to λ, so to have asymptotically decaying error in $\hat{\theta}$, one needs $D = o(e^N)$. This same scaling is relevant in Theorem 2: if $D = o(e^N)$ and we additionally assume that $|S|$ is a constant, we get $M_J = o(1)$, so that the λ we require is also $O(\sqrt{\log(D)/N})$. In this sense, if some λ is large enough to guarantee support recovery and good squared error error in the original problem, using $\lambda + o(1)$ gives support recovery for all of the leave-1-out problems.

3.4 Logistic Regression

We can next state a very similar result for logistic regression. Assume a logistic regression model such that the data $y_n \in \{-1, 1\}$ with $\Pr[y_n = 1] = 1/(1 + e^{-x_n^T\theta^*})$, where $x_n \in \mathbb{R}^D$ has i.i.d. c_x-sub-Gaussian components.

Theorem 3. *Assume that Assumption 1 holds and that D, as a function of N, grows as $o(e^N)$. Consider the logistic regression model above, and set the regularization parameter as:*

$$\lambda = \frac{1}{\alpha - M_J}\sqrt{\frac{25 + \log D}{NC}} + \frac{\sqrt{2c_x \log(ND)} + \sqrt{50c_x}}{N(\alpha - M_J)} \tag{3.10}$$

where C is a global constant relating to relationships between sub-Gaussian random variables, and M_J is defined as in Theorem 2. Then each of the leave-1-out problems has support $\operatorname{supp}\hat{\theta}^{\backslash n} \subseteq S$ *with a high, fixed probability. That is,*

$$\Pr\left[\max_n \left\|\hat{z}^{\backslash n}_{S^c}\right\|_\infty = 1\right] \leq 28e^{-25} \tag{3.11}$$

Proof. The proof runs very similarly to that of Theorem 2; see Theorem 5 in Appendix C for details. □

A similar analysis of the λ required by Theorem 3 applies here, as [10] show that $\lambda \geq O(\sqrt{\log(D)/N})$ is sufficient to ensure good squared error and accurate support recovery of $\hat{\theta}$.

3.5 Conditions under which $\operatorname{supp}\hat{\theta} = \operatorname{supp}\theta^*$

We have just seen that, in the case of linear and logistic regression, if a particular λ is sufficient for ensuring $\|\hat{z}_{S^c}\|_\infty < 1$, a slightly increased λ is sufficient for ensuring $\max_n \|z^{\backslash n}_{S^c}\|_\infty < 1$. According to Lemma 1, this condition ensures that $\operatorname{supp}\hat{\theta}^{\backslash n} \subseteq \operatorname{supp}\theta^*$ for all n; however, this is not enough to establish Theorem 1. The issue is that our approximation in Eq. (2.5) only runs over the dimensions in $\hat{S}$. If $\hat{S}$ is a strict subset of the true S, it is possible that, for some n, we will have $\hat{S} \subset \operatorname{supp}\hat{\theta}^{\backslash n}$ so that our approximation does not cover the extra dimensions in $\hat{\theta}^{\backslash n}$. However, a commonly stated condition ensures $\hat{S} = S$:

Proposition 2. *Assume that the true θ^* has non-zero entries that are not too small;*

that is

$$\min_{s \in S} |\theta_s^*| > \frac{\sqrt{|S|}(\gamma + 4)}{\lambda_{min}} \lambda \tag{3.12}$$

where γ *and* λ_{min} *are as defined in Proposition 1 for the full data problem, and also assume* $\|\hat{z}_{S^c}\|_\infty < 1$. *Then* $\operatorname{supp} \hat{\theta} = \operatorname{supp} \theta^*$.

Proof. This is part of Theorem 5.1 in [10]. □

We note that this is not an extra condition beyond that required for the success of the original estimate $\hat{\theta}$; conditions on the minimum entries of θ_S^* are typically used in the ℓ_1 literature to ensure $\hat{S} = S$ (e.g., in [8, 10, 16]), and Eq. (3.12) is an exact duplicate of the condition in [10].

Assuming the conditions of Proposition 2 hold, Theorem 2 and Theorem 3 now tell us that we can expect our approximation in Eq. (2.5) to be highly accurate and computationally cheap, even when the dimension D grows as $o(e^N)$.

Chapter 4

Experiments

4.1 Two alternatives to Eq. (2.5)

Our theoretical results imply that Eq. (2.5), derived by using a smooth approximation to the ℓ_1 norm with Eq. (2.2) and then taking the amount of smoothness to zero, has high accuracy and low computational cost. We begin our empirical investigation of this claim by first considering two more straightforward alternatives: 1) ignore the limiting argument and just use Eq. (2.2) with some smoothed version of $\|\theta\|_1$, or 2) approximate exact CV by exactly computing $\hat{\theta}^{\backslash n}$ for just a few random values of n. For the former point, we consider the smooth approximation $R^\eta(\theta)$ for $\eta > 0$ suggested by [12]:

$$R^\eta(\theta) := \sum_{d=1}^{D} \frac{1}{\eta}\Big(\log(1 + e^{\eta\theta_d}) + \log(1 + e^{-\eta\theta_d}) \Big).$$

While $\lim_{\eta\to\infty} R^\eta(\theta) = \|\theta\|_1$, we found this approximation to become numerically unstable for the purposes of optimization when η was much larger than 100, so we set $\eta = 100$ in our experiments. To test this along with subsampling exact CV, we trained logistic regression models on twenty five high dimensional random datasets in which $x_{nd} \overset{i.i.d.}{\sim} N(0,1)$ with $N = 500$ and $D = 40,000$, and the true θ^* was supported on its first five entries. We compute the accuracy of our various approximations to

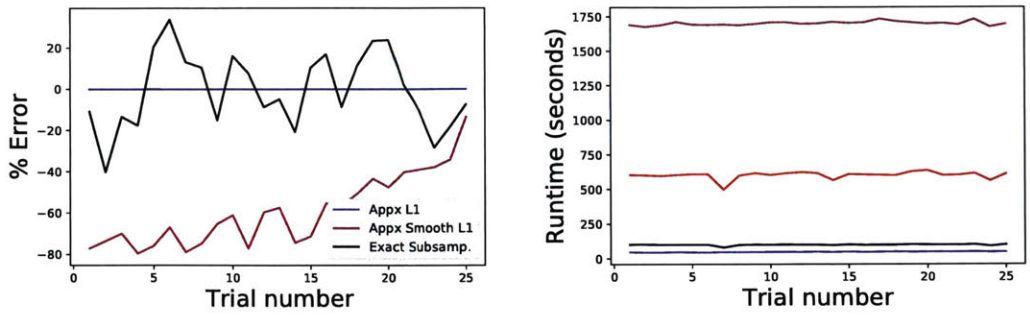


Figure 4-1: (*Left:*) Accuracy of experiments from Section 4.1. Percent error is computed as in Eq. (4.1). Note the error in Eq. (2.5) (red curve) is not noticeable but is nonzero: it varies between -0.06% and 0.04%. (*Right:*) Timings of results from Section 4.1. The legend is the same as on the left, with the addition of blue showing the runtime of exact CV for the ℓ_1 regularized model (the $D \times D$ matrix inversion needed for approximating CV in the smoothed problem is so slow that even exact CV with an efficient ℓ_1 solver is faster).

full exact CV as the percent error of exact CV:

$$\frac{|\text{approximation} - \text{exactCV}|}{\text{exactCV}} \tag{4.1}$$

Fig. 4-1 shows the accuracy of these two alternatives to Eq. (2.5), and Fig. 4-1 compares their timings. By design, subsampling exact CV has almost exactly the same runtime as using Eq. (2.5)[1]; however, we see that its accuracy is significantly reduced for nearly every trial. Using Eq. (2.2) with $R^{100}(\theta)$ as a regularizer is far worse: while subsampling exact CV is fast and unbiased (although very high variance), the smoothed approximation has to deal with a the full $D \times D$ matrix and an approximation over all D dimensions, resulting in an approximation that is orders of magnitude less accurate and slower.

4.2 The importance of correct support recovery

The discussion in Chapter 3 revolved around the point that each $\hat{\theta}^{\backslash n}$ having correct support (i.e. $\operatorname{supp} \hat{\theta}^{\backslash n} = \operatorname{supp} \theta^*$) was sufficient for obtaining the fixed-dimensional

[1]Specifically, we computed 41 different $\hat{\theta}^{\backslash n}$ for each trial in order to roughly match the computational cost of computing Eq. (2.5) for all $N = 500$ datapoints.

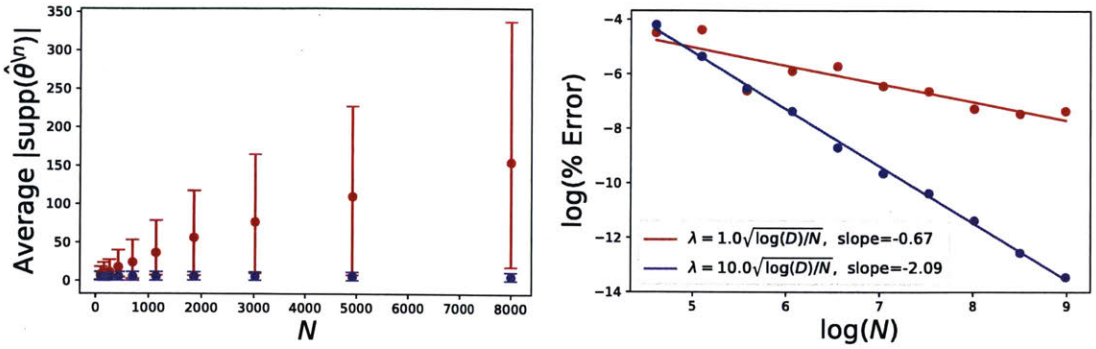


Figure 4-2: Illustration of the role of support recovery in the accuracy of the approximation in Eq. (2.5) in the case of linear regression. On the left, we show the average $|\text{supp}\,\hat{\theta}^{\backslash n}|$, with the average taken over a few random values of n and error bars showing the min and max $|\text{supp}\,\hat{\theta}^{\backslash n}|$. For $\lambda = 10.0\sqrt{\log(D)/N}$, the mean recovered support remains constant with N. On the other hand, for $\lambda = 1.0\sqrt{\log(D)/N}$, $|\text{supp}\,\hat{\theta}^{\backslash n}|$ is growing with N, as well as hugely varying for different values of n. As a result, we are forced to approximate the behavior of a much higher dimensional optimization problem. The right plot shows the resulting reduced accuracy in terms of Eq. (4.1) as D scales with N: when the support recovery is constant, we recover an error scaling of $O(1/N^2)$, whereas a growing support results in a much slower decay.

error scaling shown in Fig. 1-1. Here, we give some brief empirical evidence that this is actually necessary, at least in the case of linear regression. For values of N ranging from 1,000 to 8,000, we set $D = \lceil N/10 \rceil$ and generate a design matrix with i.i.d. $N(0,1)$ entries. The true θ^* is supported on its first five entries, with the rest set to zero. We then generate observations $y_n = x_n^T\theta^* + w_n$, for $w_n \overset{i.i.d.}{\sim} N(0,1)$. To examine what happens when the recovered supports are and are not correct, we use slightly different values of the regularization parameter λ. Specifically, the results of [16] (especially Theorem 1) tell us that the support recovery of ℓ_1 regularized linear regression will change sharply around $\lambda \approx 4\sqrt{\log(D)/N}$, where lower values of λ will fail to correctly recover the support.

With this in mind, we choose two settings of λ: $1.0\sqrt{\log(D)/N}$ and $10.0\sqrt{\log(D)/N}$. As expected, the righthand side of Fig. 4-2 shows that the quality of the approximation in Eq. (2.5) is drastically different in these two situations. The lefthand plot of Fig. 4-2 offers an explanation for this observation: the support of $\text{supp}\,\hat{\theta}^{\backslash n}$ grows with N under the former value of λ, whereas the latter value of λ ensures that

$|\text{supp}\,\hat{\theta}^{\backslash n}| = |\text{supp}\,\theta^*| = \text{const}$. Empirically, this confirms the idea that accurate support recovery of each $\hat{\theta}^{\backslash n}$ is also necessary to recover the "low-dimensional" error scaling described in Fig. 1-1.

That the approximation quality relies so heavily on the exact setting of λ is somewhat concerning. However, we emphasize that this is just as much a criticism of ℓ_1 regularization in general; as previously noted, [16] demonstrated similarly drastic behavior of $\text{supp}\,\hat{\theta}$ in the same exact linear regression setup that we use here. On the other hand, [7] do show that, despite this sensitive behavior, using exact leave-1-out CV to select λ for ℓ_1 regularized linear regression does give reasonable results. In Section 4.4, we empirically show that this is sometimes, but not always, the case for our and other approximate CV methods.

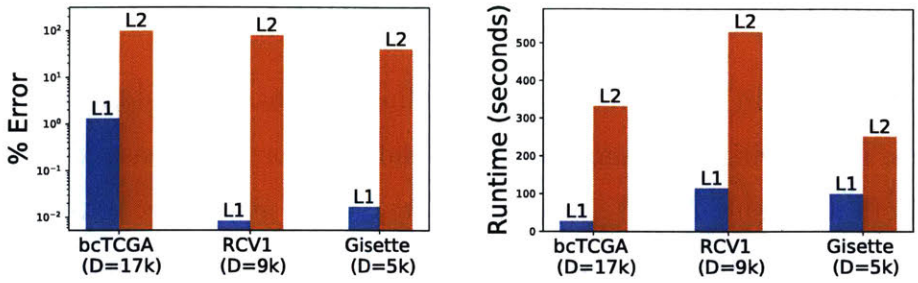


Figure 4-3: (*Left:*) Accuracy for real data experiments in Section 4.3. For each dataset, we give the accuracy of approximate CV compared to exact CV for both ℓ_2 regularized models using the existing Eq. (2.2) and ℓ_1 regularized models using our proposed Eq. (2.5). We compute "% Error" as in Eq. (4.1). (*Right:*) Timings for the same experiments.

4.3 Real data experiments

While we have shown increased accuracy of Eq. (2.5) on synthetic data, it is important to understand how dependent our results are on the particular random design we chose. We explore this question by running on a number of publicly available datasets. We chose the particular datasets shown here for having a high enough dimension to observe the effect of our results, yet not so high in dimension nor number of datapoints that running exact CV for comparison was prohibitively expensive. For a description of each dataset as well as our exact experimental setup, see Appendix B. For each dataset, we approximate CV for the ℓ_1 regularized model using Eq. (2.5). To understand how Eq. (2.5)'s accuracy is improved by the special structure present in ℓ_1 regularized problems, we compare to the accuracy of the approximation in Eq. (2.2) on the same problem, but with ℓ_2 regularization, in which there is no sparsity, and the approximation runs over all D dimensions. The accuracy we report on the left of Fig. 4-3 is the percent error compared to exact CV as in Eq. (4.1). Our results, reported in Fig. 4-3 show that approximate CV on the ℓ_1 regularized problem is significantly more accurate. Additionally, the timings on the right of show in Fig. 4-3, show that the approximations to the ℓ_1 regularized problems have significantly reduced runtimes.

4.4 Selection of λ and future work

Most previous work in approximate CV, including ours, has focused on showing that with a fixed model with increasing amounts of data, N, (or, in the case of our approximation, both increasing N and D), approximate CV will give an accurate assessment of the out of sample error. However, this does not address one of the more common uses of CV, which is to train many models with varying values of the regularization parameter λ and select the one with the lowest CV error. We consider this issue here.

We generate a synthetic ℓ_1 regularized logistic regression problem with $N = 300$ observations and $D = 150$ dimensions. The data matrix X has $N(0, 1)$ entries, and the true θ^* is supported on only its first five entries. As a measure of the "true" out of sample error, we construct a test set with ten thousand observations. For a range of values of λ, we solve Eq. (2.1), and measure the train, test, exact leave-1-out, and approximate leave-1-out errors; the results are plotted in Fig. 4-4. Not only does our approximate CV select a λ that gives a significantly worse test error than exact CV, it selects the obviously incorrect value of $\lambda = 0.0$. This issue is somewhat suggested by our theory above: that is, for an appropriately large λ, we will recover a small, correct support $\hat{S}$ and approximate CV will be highly accurate, whereas small λ will cause a larger support $\hat{S}$, which in turn causes a degredation in approximation quality. While the results in Fig. 4-4 come from using our Eq. (2.5) to approximate CV for an ℓ_1 regularized problem, we note that this issue is not specific to the current work; we observed similar behavior when using ℓ_2 regularization with both Eq. (2.2) and the computationally slower Eq. (2.3).

Still, all is not lost for approximate CV: the righthand side Fig. 4-4 shows that for the same problem setup with $D = 75$ dimensions, the error vs λ curve constructed by approximate CV is significantly different. In particular, it is convex and has its minimum very close to that of exact CV. We believe a further understanding of this issue is the most pressing direction for future work. In the meantime, this failure mode of approximate CV is at least easy to spot, assuming one believes that the true out of sample error is a convex function of λ.

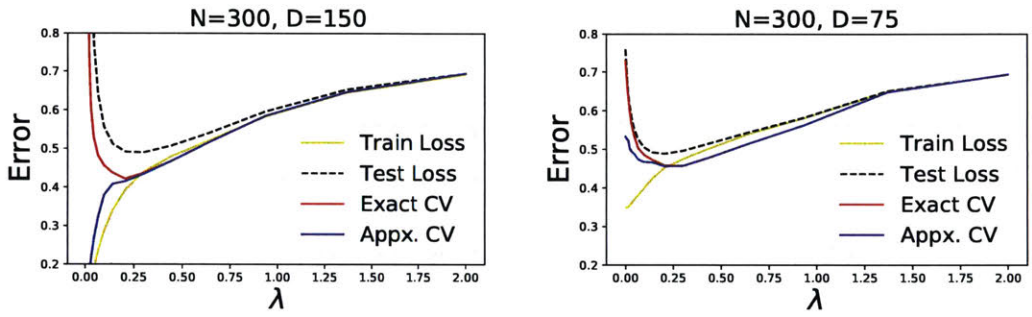


Figure 4-4: Experiment for selecting λ from Section 4.4. (*Left:*) Despite being very accurate for higher values of λ, approximate CV's degredation in accuracy for lower values of λ (which corresponds to a larger $\hat{S}$) causes the selection of a λ that is far from optimal in terms of test loss. (*Right:*) For a lower dimensional problem, the curve constructed by approximate CV much more closely mirrors that of exact CV for all values of λ.

Appendix A

Differences between Eq. (2.2) and Eq. (2.3)

In Chapter 2, we briefly outlined the differences between Eq. (2.2) and Eq. (2.3); we examine the differences in more detail here. Recall that we defined $H(\hat{\theta}) \triangleq (1/N)\sum_{n=1}^{N} \nabla_\theta^2 f(x_n^T\hat{\theta}, y_n)$. We restate the "simple" approximation given by [4] as:

$$\hat{\theta}^{\backslash n} \approx \hat{\theta} - H(\hat{\theta})^{-1}\nabla_\theta f(x_n^T\hat{\theta}, y_n), \tag{A.1}$$

whereas the "modified" approximation given by [2, 12, 17] is:

$$\hat{\theta}^{\backslash n} \approx \hat{\theta} - \left(H(\hat{\theta}) - \nabla_\theta^2 f(x_n^T\hat{\theta}, y_n)\right)^{-1} \nabla_\theta f(x_n^T\hat{\theta}, y_n). \tag{A.2}$$

A.1 Derivation of "simple" approximation

[4] derive Eq. (A.1) by appealing to the implicit function theorem. Specifically, they define $\hat{\theta^w}$ as the solution to a weighted optimization problem:

$$\hat{\theta^w} \triangleq \underset{\theta\in\mathbb{R}^D}{\arg\min} \frac{1}{N}\sum_{n=1}^{N} w_n f(x_n^T\theta, y_n) + R(\theta). \tag{A.3}$$

For example, Leave-1-out CV with the first datapoint left out corresponds to solving Eq. (A.3) with $w = (0, 1, 1, \ldots, 1)$. [4] note that the derivatives of $\hat{\theta}$ (i.e. the solution to Eq. (A.3) with $w = \mathbf{1}$) with respect to the weights w_n can be computed using the implicit function theorem to get:

$$\frac{d\hat{\theta}}{dw_n}\Big|_{w_n=1} = -H(\hat{\theta})^{-1}\nabla_\theta f(x_n^T\theta, y_n). \tag{A.4}$$

By a first order Taylor expansion around $w = (1, 1, \ldots, 1)$, we can write:

$$\hat{\theta^w} \approx \hat{\theta} + \sum_{n=1}^{N} \frac{d\hat{\theta}}{dw_n}\Big|_{w_n=1}(w_n - 1) \tag{A.5}$$

$$= \hat{\theta} - \sum_{n=1}^{N} H(\hat{\theta})^{-1}\nabla_\theta f(x_n^T\theta, y_n)(w_n - 1). \tag{A.6}$$

For the special case of w being the vector of all 1's with a zero in one coordinate (i.e. the weighting for leave-1-out CV), we recover Eq. (A.1).

A.2 Derivation of "modified" approximation

Eq. (A.2) is derived by taking a single Newton step on the objective $F_{\backslash n}$ starting at the point $\hat{\theta}$. Specifically, recall that the objective with one datapoint left out is:

$$F_{\backslash n}(\theta) = \frac{1}{N}\sum_{n=1}^{N} f(x_n^T\theta, y_n) - \frac{1}{N} f(x_n^T\theta, y_n), \tag{A.7}$$

which has $\nabla^2_\theta F_{\backslash n}(\theta) = H(\theta) - \nabla^2_\theta f(x_n^T\theta, y_n)$ as its Hessian. This gives a single Newton step from the point $\hat{\theta}$ as exactly Eq. (A.2).

A.3 Comparison of approximations

There is a major computational difference between Eq. (A.2) and Eq. (A.1): the former requires the inversion of a $D \times D$ matrix for *each* $\hat{\theta}^{\backslash n}$ approximated, while the latter requires a single $D \times D$ matrix inversion for *all* $\hat{\theta}^{\backslash n}$ inverted, which incurs a cost

of $O(N + ND^3)$ versus a cost of $O(N + D^3)$. Even for small D, this is a significant additional expense.

However, as noted by [12, 17], Eq. (A.2) is much cheaper when considering the special case of generalized linear models. In this case, $\nabla^2_\theta f$ is some scalar times $x_n x_n^T$ – a rank one matrix. The Sherman-Morrison formula then allows us to cheaply compute the needed inverse in Eq. (A.2) given only H^{-1}; this is how Equation 8 in [12] and Equation 21 in [17] are derived. Even though we only consider GLMs in this work, we, as noted above, still prefer to study Eq. (A.1) with the hope of retaining scalibility in more general problems.

Appendix B

Details of real experiments

We use three publicly available datasets for our experiments in Section 4.3:

1. The "Gisette" dataset [5] is available from the UCI repository at `https://archive.ics.uci.edu/ml/datasets/Gisette`. The dataset is constructed from the MNIST handwritten digits dataset. Specifically, the task is to differentiate between handwritten images of either "4" or "9." There are $N = 6,000$ training examples, each of which has $D = 5,000$ features, some of which are junk "distractor features" added to make the problem more difficult.

2. The "bcTCGA" is a dataset of breast cancer samples from The Cancer Genome Atlas, which we downloaded from `http://myweb.uiowa.edu/pbreheny/data/bcTCGA.html`. The dataset consists of $N = 536$ samples of tumors, each of which has the real-valued expression levels of $D = 17,322$ genes measured. The task is to predict the real-valued expression level of the BRCA1 gene, which is known to correlate with breast cancer.

3. The "RCV1" dataset [9] is a dataset of Reuters' news articles given one of four categorical labels according to their subject: "Corporate/Industrial," "Economics," "Government/Social," and "Markets." We downloaded a pre-processed binarized version from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`, which combines the first two categories into a "positive" label and the later two into a "negative" label. The full dataset contains

$N = 20,242$ articles, each of which has $D = 47,236$ features. Running exact CV on this dataset would have been prohibitively slow, so we created a smaller dataset. First, the data matrix X is extremely sparse (i.e., is mostly equal to zero), so we selected the top 10,000 most common features and threw away the rest. We then randomly chose 5,000 documents to keep as our training set. After throwing away any of the 10,000 features that were now not observed in this subset, we were left with a dataset of size $N = 5,000$ and $D = 9,836$.

In order to run ℓ_1 regularized regression on each of these datasets, we first needed to select a value of λ. Since all of these datasets are fairly high dimensional, Section 4.4 suggests our approximation should will likely be innacurate for values of λ that are "too small." In an attempt to get the order of magnitude for λ correct, we used the theoretically motivated value of $\lambda = C\sqrt{\log(D)/N}$ for some constant C (e.g., [10] shows this scaling of λ will recover the correct support for both linear and logistic regression). Section 4.2 suggests that the constant C can be very important for the accuracy of our approximation, and our experiments there suggest that this is caused by too large a recovered support size $|\text{supp}\,\hat{\theta}|$. For these experiments, we guessed a reasonable sounding value for C, solved for $\hat{\theta}$, and confirmed that $|\text{supp}\,\hat{\theta}|$ was not too large. For all of the datasets here, we used $C = 1.50$ to get the results reported in Section 4.3.

Appendix C

Proofs from Chapter 3

We first state a few existing results about the maxima of sub-Gaussian and sub-Exponential random variables that will be useful in our proofs.

Lemma 2 (Lemma 5.2 from [13]). *Suppose that we have real valued random variables* $Z_1, \ldots, Z_N$ *that satisfy* $\log E[e^{\lambda x_n}] \leq \psi(\lambda)$ *for all* n *and all* $\lambda \geq 0$ *for some convex* ψ *with* $\psi(0) = \psi'(0) = 0$. *Then for any* $u \geq 0$ *:*

$$\Pr\left[\max_{n=1,\ldots,N} Z_n \geq \psi^{*-1}(\log N + u)\right] \leq e^{-u}.$$

where ψ^{*-1} *is the inverse of the Legendre dual of* ψ.

Remembering the definition of a sub-Gaussian random variable from Definition 1, Lemma 2 can be used to show the following:

Corollary 2. *Let* $Z_1, \ldots, Z_N$ *be i.i.d. sub-Gaussian random variables with parameter* c_x. *Then:*

$$\Pr\left[\max_{n=1,\ldots,N} Z_n \geq E[Z_n] + \sqrt{2c_x \log N} + u\right] \leq e^{-\frac{u^2}{2c_x}} \tag{C.1}$$

$$\Pr\left[\max_{n=1,\ldots,N} Z_n^2 \geq E[Z_n^2] + c_x^2(\log N + 1 + u)\right] \leq e^{-u} \tag{C.2}$$

Proof. For the first inequality, the definition of a sub-Gaussian random variable is that $\log E e^{\lambda Z_n} \leq \lambda^2 c_x/2 \triangleq \psi(\lambda)$, which has $\psi^{*-1}(x) = \sqrt{2c_x x}$. For the second

inequality, use the fact that Z_n^2 is sub-Exponential with parameter c_x^2 so that it satisfies $\log Ee^{\lambda Z_n^2} \leq \psi(\lambda)$, where:

$$\psi(\lambda) \triangleq \begin{cases} \lambda c_x^2, & 0 \leq t \leq 1/c_x^2 \\ \infty, & \text{o.w.} \end{cases}$$

which, for $x \geq 0$, has inverse dual $\psi^{*-1}(x) = c_x^2(x+1)$. □

Proposition 3. *Let $x_1, \ldots, x_N \in \mathbb{R}^D$ be random vectors with i.i.d. sub-Gaussian components with parameter c_x and $E[x_{nd}^2] = \sigma^2$. Then:*

$$\Pr\left[\max_{n=1,\ldots,N} \|x_n\|_2 \geq E[\|x_n\|_2] + \sqrt{2C\sigma c_x^2 \log N} + u\right] \leq e^{-\frac{u^2}{2C\sigma c_x^2}}, \tag{C.3}$$

where $C > 0$ is some global constant, independent of c_x, D, and N.

Proof. From Theorem 3.1.1 of [15], we have that $\|x_n\|_2 - \sigma\sqrt{D}$ is sub-Gaussian with parameter $C\sigma c_x^2$, where C is some constant. Using the first part of Corollary 2 gives the result. □

C.1 Local structured smoothness condition (LSSC)

The local structured smoothness condition (LSSC) was first introduced in [10] for the purpose of extending proof techniques for the support recovery of ℓ_1 regularized linear regression to more general ℓ_1 regularized M-estimators. Essentially, it provides a condition on the smoothness of the third derivatives of the objective $F(\theta)$ around the true sparse θ^*. One can then analyze a second order Taylor expansion of the loss and use the LSSC to show that the remainder in this expansion is not too large. To formalize the LSSC, we need to define the third order Fréchet derivative of F evaluated along a direction $u \in \mathbb{R}^D$:

$$D^3F(\theta)[u] := \lim_{t \to 0} \frac{\nabla^2 F(\theta + tu) - \nabla^2 F(\theta)}{t}.$$

In the cases considered in this paper, this is just a $D \times D$ matrix. We can then naturally define the scalar $D^3F(\theta)[u, v, w]$ as:

$$D^3[u, v, w] := v^T D^3 F(\theta)[u]w$$

We can now define the LSSC:

Definition 2 (LSSC). *Let $F : \mathbb{R}^D \to \mathbb{R}$ be a continuously three-times differentiable function. For $\theta^* \in \mathbb{R}^D$ and $N_{\theta^*} \subseteq \mathbb{R}^D$, the function F the (θ^*, N_{θ^*}) LSSC with constant $K \geq 0$ if for any $u \in \mathbb{R}^D$:*

$$|D^3 f(\theta^* + \delta)[u, u, e_j]| \leq K \|u\|_2^2, \tag{C.4}$$

where $e_j \in \mathbb{R}^D$ is the jth coordinate vector, and $\delta \in \mathbb{R}^D$ is any vector such that $\theta^* + \delta \in N_{\theta^*}$.

We note that this definition is actually equivalent to the original definition given in [10], who prove the two to be equivalent in their Proposition 3.1. [10] goes on to prove bounds on the LSSC constants for linear and logistic regression, which we state as Proposition 7 and Proposition 9 below.

C.2 Linear Regression

Assume a linear regression model $y_n = x_n^T \theta^* + w_n$, where $x_n \in \mathbb{R}^D$ has i.i.d. c_x-sub-Gaussian components with $E[x_{nd}^2] = 1$ and w_n is c_w-sub-Gaussian. The probability that we're interested in is:

$$\Pr\left[\max_n \left\|\hat{z}_{S^c}^{/n}\right\|_\infty \geq 1\right],$$

where the probability is taken over the random y_n and x_n. Using Proposition 1 from the main text, we can upper bound this by the probability that any of the conditions in Proposition 1 are violated by any of the leave-1-out problems. Throughout, $C > 0$ will be a global constant in that it is independent of any characteristic of the problem (e.g. N, D, amounts of noise, $|S|$, etc.) We will frequently use X_S to denote the

$N \times |S|$ matrix formed by taking the columns of X that are in S, x_{nS} to denote the coordinates of the nth datapoint x_n that are in the set S, and $X_{S\setminus n}$ to denote the matrix X_S with the nth row removed. We will show the following theorem, stated slightly more concisely as Theorem 2 in the main text:

Theorem 4. *Consider the linear regression model above, and set the regularization parameter as:*

$$\lambda = \frac{1}{\alpha - M_J}\sqrt{\frac{c_x^2 c_w^2 \log D}{NC} + \frac{25c_x^2 c_w^2}{NC}} + \frac{4c_x c_w(\log(ND) + 26)}{N(\alpha - M_J)} \tag{C.5}$$

where C is a global constant relating to relationships between sub-Gaussian random variables, and M_J is defined as:

$$M_J = \frac{4|S|\left(EX_{nd} + \sqrt{50c_x} + \sqrt{2c_x \log(N(D-|S|))}\right)\left(E\left\|X_{ns}\right\|_2 + \sqrt{50Cc_x} + \sqrt{2Cc_x \log N}\right)}{N - 3Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)} + \frac{16|S|\left(|S|c_x^2(\log N + 26)\right)\left(E\left\|X_{\cdot,S}\right\|_2 + \sqrt{50Cc_x^2} + \sqrt{2Cc_x^2 \log(D-|S|)}\right)\left(\sqrt{N|S|} + 5Cc_x\right)}{\left(N - 3Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)\right)^2} \tag{C.6}$$

Then each of the leave-one-out problems recovers the support with a high, fixed probability. That is,

$$\Pr\left[\max_n \left\|\hat{z}_{S^c}^{\setminus n}\right\|_\infty \geq 1\right] \leq 22e^{-25} \tag{C.7}$$

Proof. For a fixed regularization parameter λ and random data X and Y, we are interested in $\Pr[\max_n \|\hat{z}_{S^c}^{\setminus n}\|_\infty \geq 1]$. As noted above, Proposition 1 allows us to upper bound this by the probability that any of the conditions in Proposition 1 are violated. For convenience, define $J_{nd} \in \mathbb{R}^D$, for $d \in S^c$, as:

$$J_{nd} := \left(X_{S\setminus n}^T X_{S\setminus n}\right)^{-1} X_{S\setminus n}^T X_{\cdot,d}.$$

It is easiest to show the conditions in Proposition 1 hold with high probability sepa-

rately, rather than all together, so we apply a union bound to get:

$$\begin{aligned}
&\Pr\left[\max_n \left\|\hat{z}^{/n}_{S^c}\right\|_\infty \geq 1\right] \leq \\
&\quad \Pr\left[\min_n \lambda_{min}(X^T_{S\backslash n} X_{S\backslash n}) = 0\right] \\
&\quad + \Pr\left[\max_n \max_{d\in S^c} \|J_{nd}\|_1 \geq 1\right] \\
&\quad + \Pr\left[\max_n \left\|\nabla F_{\backslash n}\right\|_\infty > \frac{\lambda \max_n \max_{d\in S^c} \|J_{nd}\|_1}{4}\right] \\
&\quad + \Pr\left[\frac{\min_n \lambda^2_{min}(X^T_{S\backslash n} X_{S\backslash n})}{4\left(\max_n \max_{d\in S^c} \|J_{nd}\|_1 + 4\right)^2} \frac{\max_n \max_{d\in S^c} \|J_{nd}\|_1}{|S|K} \leq \lambda\right]
\end{aligned}$$

Using Lemma 3, Lemma 5, and Lemma 6, the first three terms are bounded by $22e^{-25}$. As noted in Proposition 7, we have $\Pr[K = 0] = 1$, so the final probablity is equal to zero (as the event reduces to the probability that $\infty < \lambda$). □

What remains is to prove the lemmas needed to show that the above probabilities are small.

C.3 Linear regression: minimum eigenvalue

All we want to bound right now is the probability that the minimum eigenvalue is actually equal to zero; however, we will later want to show that it is, in fact roughly $O(N)$. The lemma we prove in this section shows exactly this. We will start with two propositions.

Proposition 4. *If X_S is an $N \times |S|$ matrix with independent c_x-sub-Gaussian entries with unit variance, then:*

$$\Pr\left[\lambda_{min}(X_S^T X_S) \leq N - 2Cc_x^2\sqrt{N}(\sqrt{|S|} + 5)\right] \leq 2e^{-25}, \tag{C.8}$$

where $C > 0$ is a global constant.

Proof. Theorem 4.6.1 of [15] gives a concentration inequality for the minimum singular

value, $s_{min}(X_S)$, of X_S:

$$\Pr\left[s_{min}(X_S) \leq \sqrt{N} - Cc_x^2(\sqrt{|S|}+t)\right] \leq 2e^{-t^2}. \tag{C.9}$$

Using the fact that the minimum eigenvalue of $X_S^T X_S$ is the square of the minimum singular value of X_S and putting in $t = 5$:

$$\Pr\left[\lambda_{min}(X_S^T X_S) \leq N - 2Cc_x^2\sqrt{N}(\sqrt{|S|}+5) + C^2c_x^4(\sqrt{|S|}+5)^2\right] \leq 2e^{-25}.$$

Dropping the $C^2c_x^4(\sqrt{|S|}+5)^2$ gives the result.

the needed concentration inequality on the minimum singular value of X_S. □

Proposition 5. *If $X_{S\setminus n}$ is the $N-1 \times |S|$ matrix formed by removing the nth row from X_S, we have:*

$$\lambda_{min}(X_{S\setminus n}^T X_{S\setminus n}) \geq \lambda_{min}(X_S^T X_S) - \|x_{nS}\|_2^2, \tag{C.10}$$

where x_n is the nth row of X_S.

Proof. Looking at the variational characterization of the minimum eigenvalue:

$$\begin{aligned}
\lambda_{min}(X_{S\setminus n}^T X_{S\setminus n}) &= \min_{z\in\mathbb{R}^{|S|}\,:\,\|z\|_2=1}\left[z^T X_S^T X_S z - z^T x_{nS} x_{nS}^T z\right] \\
&\geq \min_z z^T X_S^T X_S z - \max_z z^T x_{nS} x_{nS}^T z \\
&= \lambda_{min}(X_S^T X_S) - \|x_{nS}\|_2^2.
\end{aligned}$$

□

The above two propositions now allow us to prove the bound we want on $\min_n \lambda_{min}(X_{S\setminus n}^T X_{S\setminus n})$.

Lemma 3. *If X_S is a $N \times |S|$ matrix with independent c_x-sub-Gaussian entries and N is at least moderately large compared to $|S|$, then*

$$\Pr\left[\min_{n=1,\ldots N} \lambda_{min}(X_{S\setminus n}^T X_{S\setminus n}) \leq N - 3Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)\right] \leq 3e^{-25} \tag{C.11}$$

Proof. By using the generic inequality, for any events A and B:

$$\begin{aligned}\Pr[A] &= \Pr[A \mid B]\Pr[B] + \Pr[A \mid B^c]\Pr[B^c] \\ &\leq \Pr[A \mid B] + \Pr[B^c],\end{aligned} \tag{C.12}$$

Calling the probability on the left hand side of Eq. (C.11) P, we can break P down as, for some constant λ_{min}:

$$\begin{aligned}P \leq & \\ &\Pr\left[\min_{n=1,\ldots N} \lambda_{min}(X_{S\backslash n}^T X_{S\backslash n}) \leq N - 3Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right) \;\middle|\; \lambda_{min}(X_S^T X_S) \geq \lambda_{min}\right] \\ &\qquad + \Pr\left[\lambda_{min}(X_S^T X_S) \leq \lambda_{min}\right] \\ &\leq \Pr\left[\min_{n=1,\ldots N} \lambda_{min} - \|x_{nS}\|_2^2 \leq N - 3Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)\right] + \Pr\left[\lambda_{min}(X_S^T X_S) \leq \lambda_{min}\right] \\ &\leq \Pr\left[\max_n \|x_{nS}\|_2^2 \geq \lambda_{min} - N + 3Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)\right] + \Pr\left[\lambda_{min}(X_S^T X_S) \leq \lambda_{min}\right]\end{aligned}$$

Picking $\lambda_{min} = N - 2Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)$, we have that the second probability at most $2e^{-25}$ by Proposition 5. For this choice of λ_{min}, choose $u = 25$ in the second statement of Corollary 2; this tells us that the first probability is at most e^{-25} if $E[\|x_{nS}\|_2^2] + c_x^2(\log N + 26) = |S| + c_x^2(\log N + 26)$ is less than $Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)$, which the requirement that N be "moderately large" compared to $|S|$ in the statement of the lemma. □

C.4 Linear regression: incoherence

The following lemma will be useful:

Lemma 4. *Let $z \in \mathbb{R}^N$ be any vector and $z_{\backslash n} \in \mathbb{R}^{N-1}$ the same vector with the nth coordinate removed. Also let $X_S \in \mathbb{R}^{N\times|S|}$ be some matrix with $X_{S\backslash n}$ the same matrix with the nth row removed. Define, for any vector $z \in R^N$:*

$$J_{nz} \triangleq \left(X_{S\backslash n}^T X_{S\backslash n}\right)^{-1} X_{S\backslash n}^T z_{\backslash n}, \tag{C.13}$$

and J_z the same but with no row removed. Then:

$$\|J_{nz} - J_z\|_1 \leq |S| \frac{|z_n| \, \|x_{nS}\|_2}{\lambda_{min}(X_{S\setminus n}^T X_{S\setminus n})} + |S| \frac{\|x_{nS}\|_2^2}{\lambda_{min}^2(X_{S\setminus n}^T X_{S\setminus n})} \|z\|_2 \, \|X_S\|_2 ,$$

where $\|X_S\|_2 \triangleq \sqrt{\sum_{n=1}^N \sum_{s \in S} X_{ns}^2}$.

Proof. We can rewrite $J_z = (X_S^T X_S)^{-1} X_S^T z$ by noting that $X_S^T X_S$ and $X_{S\setminus n}^T X_{S\setminus n}$ differ by a rank one update and then applying the Sherman-Morrison formula:

$$J_z = (X_S^T X_S)^{-1} X_S^T z \tag{C.14}$$

$$= \left((X_{S\setminus n}^T X_{S\setminus n})^{-1} - \frac{(X_{S\setminus n}^T X_{S\setminus n})^{-1} X_{Sn} X_{Sn}^T (X_{S\setminus n}^T X_{S\setminus n})^{-1}}{1 + X_{Sn}^T (X_{S\setminus n}^T X_{S\setminus n})^{-1} X_{Sn}} \right) X_S^T z \tag{C.15}$$

$$= \left(J_{nz} + (X_{S\setminus n}^T X_{S\setminus n})^{-1} x_{nS} z_n \right) - \frac{(X_{S\setminus n}^T X_{S\setminus n})^{-1} X_{Sn} X_{Sn}^T (X_{S\setminus n}^T X_{S\setminus n})^{-1}}{1 + x_n^T (X_{S\setminus n}^T X_{S\setminus n})^{-1} X_{Sn}} X_S^T z \tag{C.16}$$

To cleanup notation a bit, let $B := X_{S\setminus n}^T X_{S\setminus n}$. We can continue to rewrite the above as:

$$= (J_{nz} + B x_{nS} z_n) - \frac{B^{-1} x_{nS}}{1 + x_{nS}^T B^{-1} x_{nS}} \sum_{m=1}^N z_m x_{nS}^T B^{-1} x_{mS} \tag{C.17}$$

Now, we're interested in $\|J_{nz} - J_z\|_1$, which we will bound by subtracting J_{nz} from both sides of the above equation, and then examining each coordinate by multiplying by the ith unit vector e_i:

$$|e_i^T (J_{nz} - J_z)| \leq |e_i^T B^{-1} x_{nS}| |z_n| + \frac{|e_i^T B^{-1} x_{nS}|}{1 + x_{nS}^T B^{-1} x_{nS}} \sum_{m=1}^N |z_m| |x_{nS}^T B^{-1} x_{mS}| \tag{C.18}$$

$$\leq |z_n| \lambda_{max}(B^{-1}) \, \|x_{nS}\|_2 + \frac{\lambda_{max}^2(B^{-1}) \, \|x_{nS}\|_2^2}{1 + \lambda_{min}(B^{-1}) \, \|x_{nS}\|_2^2} \sum_{m=1}^N |z_m| \, \|x_{mS}\|_2 \tag{C.19}$$

Dropping quantities that are making the denominators larger, and using the fact that, for the positive semidefinite matrix B we have:

1. $\lambda_{min}(B^{-1}) = 1/\lambda_{max}(B)$

2. $\lambda_{max}(B^{-1}) = 1/\lambda_{min}(B)$

we get:

$$|e_i^T(J_{nz} - J_z)| \leq \frac{|z_n| \, \|x_{nS}\|_2}{\lambda_{min}(B)} + \frac{\|x_{nS}\|_2^2}{\lambda_{min}^2(B)} \sum_{m=1}^{N} |z_m| \, \|x_{mS}\|_2 \,. \tag{C.20}$$

Finally, use Cauchy-Schwarz to get $\sum_{m=1}^{N} |z_m| \, \|x_{mS}\|_2 \leq \|z\|_2 \, \|X_S\|_2$. Notice that our upper bound is now independent of the index i; we then have a bound on any coordinate i of $|(J_{nz} - J_z)|$, so multiplying this bound by $|S|$ upper bounds $\|J_{nz} - J_z\|_1$, which gives the result. □

To get a high probability upper bound on $\|J_{nd}\|_1$, the idea will be to use $\|J_{nd}\|_1 \leq \|J_d\|_1 + \|J_{nd} - J_d\|_1$, and then put high probability bounds on the bound given by Lemma 4.

Lemma 5. *Under Assumption 1, for the linear regression setup given above and the number M_J given in Theorem 2, we have:*

$$\Pr\left[\max_{n=1,\ldots,N} \max_{d \in S^c} \|J_{nd}\|_1 \geq 1 - \alpha + M_J\right] \leq 8e^{-25}, \tag{C.21}$$

where J_{nd} is shorthand for $J_{nX_{\cdot,d}}$ defined in Lemma 4.

Proof. First, for any n and d, we have $\|J_{nd}\|_1 \leq \|J_d\|_1 + \|J_{nd} - J_d\|_1$. We can upper bound $\|J_{nd} - J_d\|_1$ using Lemma 4 and then apply a high probability upper bound. Arguing by the same conditioning trick as in the proof of Lemma 3, we can condition

on the events, the complement of each of which has a small constant probability:

$$\left\{\min_n \lambda_{min}(X_{S\backslash n}^T X_{S\backslash n}) \geq N - 3Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)\right\} \tag{C.22}$$

$$\left\{\|X_{\cdot,S}\|_2 \leq \sqrt{N|S|} + 5Cc_x\right\} \tag{C.23}$$

$$\left\{\max_n \|X_{nS}\|_2 \leq E[\|X_{nS}\|_2] + \sqrt{50Cc_x^2} + \sqrt{2Cc_x^2 \log N}\right\} \tag{C.24}$$

$$\left\{\max_{d\in S^c} \|X_{\cdot,d}\|_2 \leq E[\|X_{\cdot,d}\|_2] + \sqrt{50Cc_x^2} + \sqrt{2Cc_x^2 \log(D-|S|)}\right\} \tag{C.25}$$

$$\left\{\max_n \max_{d\in S^c} |X_{n,d}| \leq E[X_{nd}] + \sqrt{50c_x} + \sqrt{2c_x \log(N(D-|S|))}\right\} \tag{C.26}$$

The probability of the complement of the first event is $\leq 2e^{-25}$ by Lemma 3, the second is $\leq e^{-25}$ by noting that $\|X_{\cdot,S}\|_2$ is a Cc_x^2-sub-Gaussian random variable and applying a standard sub-Gaussian bound, the third is $\leq e^{-25}$ by applying Proposition 3, the fourth is $\leq e^{-25}$ by the same reasoning as the third, and the fifth is $\leq 2e^{-25}$ by Corollary 2. All in all, these probabilities sum up to $7e^{-25}$. Conditioned on all these events, we can upper bound the upper bound on $\|J_{nd} - J_d\|_1$ given by Lemma 4 to get:

$$\|J_{nd} - J_d\|_1 \leq$$
$$\frac{4|S|\left(EX_{nd} + \sqrt{50c_x} + \sqrt{2c_x \log(N(D-|S|))}\right)\left(E\|X_{ns}\|_2 + \sqrt{50Cc_x^2} + \sqrt{2Cc_x^2 \log N}\right)}{N - 3CC_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)} +$$
$$\frac{16|S|\left(|S|c_x^2(\log N + 26)\right)\left(E\|X_{\cdot,S}\|_2 + \sqrt{50Cc_x^2} + \sqrt{2Cc_x^2 \log(D-|S|)}\right)\left(\sqrt{N|S|} + 5Cc_x\right)}{\left(N - 3Cc_x^2\sqrt{N}\left(\sqrt{|S|}+5\right)\right)^2}$$

Call the quantity on the right-hand side of this inequality M_J and the union of the above four events the event F. Then by conditioning on F and manipulating

probabilities as in Eq. (C.12), we get:

$$\Pr\left[\max_{n=1,\ldots,N}\max_{d\in S^c}\|J_{nd}\|_1 \geq 1-\alpha+M_J\right] \leq \tag{C.27}$$

$$\Pr\left[\max_{n}\max_{d\in S^c}\|J_{nd}-J_n\|_1 \geq M_J \mid F\right] + \Pr\left[F^c\right] + \Pr\left[\max_{d\in S^c}\|J_d\|_1 \geq 1-\alpha\right] \tag{C.28}$$

By the definition of M_J above, we know that the first probability is zero, the probability of the second is $\leq 7e^{-25}$, and the third is $\leq e^{-25}$ by Assumption 1. □

C.5 Linear regression: bounded gradient

We need to bound the probability

$$\Pr\left[\max_{n=1,\ldots,N}\left\|\nabla F_{\setminus n}(\theta^*)\right\|_\infty \geq \frac{\lambda \max_n \max_{d\in S^c}\|J_{nd}\|_1}{4}\right]$$

$$\leq \Pr\left[\max_{n}\left(\|\nabla F(\theta^*)\|_\infty + \left\|\nabla f(x_n^T\theta^*, y_n)\right\|_\infty\right) \geq \frac{\lambda \max_n \max_{d\in S^c}\|J_{nd}\|_1}{4}\right]$$

Conditioning on the event that $\|\nabla F(\theta^*)\|_\infty \leq B_G$ for some number B_G and the event that $\max_n \max_{d\in S^c}\|J_{nd}\|_1 \leq 1-\alpha+M_J$, we get that this probability is less than or equal to:

$$\leq \Pr\left[\max_{n=1,\ldots,N}\left\|\nabla f(x_n^T\theta^*, y_n)\right\|_\infty \geq \frac{\lambda(1-\alpha+M_J)}{4} - B_G\right]$$

$$+ \Pr\left[\|\nabla F(\theta^*)\|_\infty \geq B_G\right] + \Pr\left[\max_{n}\max_{d\in S^c}\|J_{nd}\|_1 \geq 1-\alpha+M_J\right] \tag{C.29}$$

The following proposition gives a reasonable value for B_G:

Proposition 6. *In the above setup for linear regression,*

$$\Pr\left[\|\nabla F(\theta^*)\|_\infty \geq \left[\frac{c_x^2 c_w^2 \log D}{NC} + \frac{25 c_x^2 c_w^2}{NC}\right]^{1/2}\right] \leq e^{-25} \tag{C.30}$$

Proof. The dth coordinate of the gradient is $(\nabla F(\theta^*))_d = 1/N \sum_n w_n x_{nd}$. First, we

have that $1/N \sum_n w_n x_{nd}$ is a $c_x c_w$-sub-Exponential random variable. By Bernstein's inequality (see Theorem 2.8.1 from [15]), we have:

$$\Pr\left[\frac{1}{N}|\sum_{n=1}^{N} w_n x_{nd}| \geq \left[\frac{c_x^2 c_w^2 \log D}{NC} + \frac{25 c_x^2 c_w^2}{NC}\right]^{1/2}\right] \leq e^{-25-\log D}$$

If we union bound over the D dimensions of $\nabla F(\theta^*)$, we get that the probability in the proposition's statement is $\leq De^{-25-\log D} = e^{-25}$, as claimed. □

Now we can prove the bound we need on the probability that any $\left\|\nabla F_{\backslash n}(\theta^*)\right\|_\infty$ is large:

Lemma 6. *For the above setup for linear regression and the λ given in Theorem 2, we have:*

$$\Pr\left[\max_{n=1,\ldots,N} \left\|\nabla F_{\backslash n}(\theta^*)\right\|_\infty \geq \frac{\lambda \max_n \max_{d \in S^c} \|J_{nd}\|_1}{4}\right] \leq 11e^{-25} \tag{C.31}$$

Proof. We can first apply the bound worked out in Eq. (C.29). Picking B_G to be the value given in Proposition 6, the second probability is $\leq e^{-25}$ by Proposition 6, and the third is $\leq 8e^{-25}$ by Lemma 5. To analyze the first probability, note that we can write the event as:

$$\Pr\left[\frac{1}{N} \max_n \max_d |w_n x_{nd}| \geq \frac{\lambda(\alpha - M_J)}{4} - B_G\right].$$

Looking at the form of λ given in Theorem 2, we get that this is equal to:

$$= \Pr\left[\frac{1}{N}|\max_n \max_d w_n x_{nd}| \geq 4c_x c_w(\log(ND) + 26)\right].$$

The event we're considering is just the absolute value of the max of ND sub-Exponential variables with parameter $c_x c_w$. Plugging into Corollary 2 gives that this probability is $\leq 2e^{-25}$. □

C.6 Linear regression: λ small enough

To check Eq. (3.6), we need to know the LSSC constant K for linear regression:

Proposition 7. *For the linear regression setup in Theorem 4, the loss $F(\theta)$ satisfies the (θ^*, N_{θ^*}) LSSC with constant $K = 0$ for any θ^*, N_{θ^*}, and any data X, Y.*

Proof. This follows from the fact that $F(\theta) = \|X\theta - Y\|_2^2$ has zero third derivatives, implying that $D^3F(\theta)[u, u, e_j] = 0$ for any $\theta, u \in \mathbb{R}^D$ and coordinate vector $e_j \in R^D$. □

As linear regression has a LSSC constant K that is deterministically equal to zero, the only constraint implied by Eq. (3.6) is that $\lambda < \infty$, which is always satisfied by the value of λ given in Theorem 4.

C.7 Logistic Regression

Assume a logistic regression model such that the data $y_n \in \{-1, 1\}$ with $\Pr[y_n = 1] = 1/(1 + e^{-x_n^T\theta^*})$. The derivatives are slightly more complicated here than in the case of linear regression. In particular, defining:

$$D_n^{(1)} := \frac{-y_n}{1 + e^{y_n x_n^T \theta^*}}, \quad D_n^{(2)} := \frac{e^{x_n^T\theta^*}}{(1 + e^{x_n^T\theta^*})^2}, \tag{C.32}$$

the derivatives of F are:

$$\nabla_\theta F(\theta^*) = \frac{1}{N}\sum_{n=1}^N D_n^{(1)} x_n, \quad \nabla_\theta^2 F(\theta^*) = \frac{1}{N}\sum_{n=1}^N D_n^{(2)} x_n x_n^T. \tag{C.33}$$

For comparison, things were easier for linear regression because $D_n^{(2)} = 1$ and $D_n^{(1)} = w_n$ for some sub-Gaussian noise w_n. Still, we will be able to apply basically all of the above reasoning just by using the fact that $|D_n^{(2)}|$ and $|D_n^{(1)}|$ are both ≤ 1, allowing them to drop them in many of our upper bounds. This will allow us to prove a very similar result to Theorem 2:

Theorem 5. *Consider the logistic regression model above, and assume that for some scalar λ_{min}, we have:*

$$\Pr\left[\lambda_{min}\left(\nabla^2_\theta F(\theta^*)_{SS}\right) \leq \lambda_{min}\right] \leq e^{-25} \tag{C.34}$$

Then, given the regularization parameter is set as:

$$\lambda = \frac{1}{\alpha - M_J}\sqrt{\frac{25 + \log D}{NC}} + \frac{\sqrt{2c_x \log(ND)} + \sqrt{50c_x}}{N(\alpha - M_J)} \tag{C.35}$$

where C is a global constant relating to relationships between sub-Gaussian random variables, and M_J is defined the same as in Theorem 4. Then each of the leave-1-out problems has $\operatorname{supp}\hat{\theta}^{\backslash n} \subseteq S$ *with a high, fixed probability. That is:*

$$\Pr\left[\max_n \left\|\hat{z}^{\backslash n}_{S^c}\right\|_\infty \geq 1\right] \leq 28e^{-25} \tag{C.36}$$

C.8 Logistic regression: lambda min

Lemma 7. *Assume that for some scalar λ_{min}, we have:*

$$\Pr\left[\lambda_{min}\left(\nabla^2_\theta F(\theta^*)_{SS}\right) \leq \lambda_{min}\right] \leq e^{-25}. \tag{C.37}$$

Then:

$$\Pr\left[\lambda_{min}(\nabla^2_\theta F_{\backslash n}(\theta^*)_{SS}) \leq \lambda_{min} - Cc_x^2\sqrt{N}\left(\sqrt{|S|} + 5\right)\right] \leq 3e^{-25} \tag{C.38}$$

Proof. We have by Proposition 5 and the fact that $|D_n^{(2)}| \leq 1$:

$$\begin{aligned}\lambda_{min}(\nabla^2_\theta F_{\backslash n}(\theta^*)_{SS}) &\geq \lambda_{min}(\nabla^2_\theta F_{\backslash n}(\theta^*)_{SS}) - \|x_{nS}\|_2^2\,|D_n^{(2)}| \\ &\geq \lambda_{min}(\nabla^2_\theta F_{\backslash n}(\theta^*)_{SS}) - \|x_{nS}\|_2^2\,.\end{aligned}$$

The rest of the proof is now exactly the same as that of Lemma 3. □

C.9 Logistic regression: incoherence

We can get exactly the same bound as in Lemma 5. To do so, we first note that Lemma 4 is only written to deal with Hessians of the form $X^T X$; however, if we rewrite our data as $\bar{x}_n := \sqrt{D_n^{(2)}} x_n$, the Hessian for logistic regression is equal to $\bar{X}^T \bar{X}$. We can further upper bound the upper bound in Lemma 4 by noting that $|D_n^{(2)}| \leq 1 \implies \|\bar{x}_n\|_2 \leq \|x_n\|_2$. Applying this reasoning, we get an identical lemma to Lemma 5

Lemma 8. *Under Assumption 1, for the logistic regression setup given above and the number M_J given in Theorem 3, we have:*

$$\Pr\left[\max_{n=1,\ldots,N} \max_{d \in S^c} \|J_{nd}\|_1 \geq 1 - \alpha + M_J\right] \leq 8e^{-25}, \tag{C.39}$$

where J_{nd} is shorthand for $J_{nX_{\cdot,d}}$ defined in Lemma 4.

C.10 Logistic regression: bounded gradient

Again, we are interested in bounding:

$$\Pr\left[\max_{n=1,\ldots,N} \left\|\nabla F_{\backslash n}(\theta^*)\right\|_\infty \geq \frac{\lambda \max_n \max_{d \in S^c} \|J_{nd}\|_1}{4}\right]$$

The same reasoning that led to Eq. (C.29) gives us the same bound:

$$\begin{aligned} \leq \Pr\left[\max_{n=1,\ldots,N} \left\|\nabla f(x_n^T \theta^*, y_n)\right\|_\infty \geq \frac{\lambda(1-\alpha+M_J)}{4} - B_G\right] \\ + \Pr\left[\|\nabla F(\theta^*)\|_\infty \geq B_G\right] + \Pr\left[\max_n \max_{d \in S^c} \|J_{nd}\|_1 \geq 1 - \alpha + M_J\right] \end{aligned} \tag{C.40}$$

Just as in the case of linear regression, we can first pick a reasonable value for B_G:

Proposition 8. *For the logistic regression setup above, we have:*

$$\Pr\left[\|\nabla F(\theta^*)\|_\infty \geq c_x \sqrt{\frac{25 + \log D}{CN}}\right] \leq 2e^{-25}. \tag{C.41}$$

Proof. The dth coordinate of the gradient is $(\nabla F(\theta^*))_d = 1/N \sum_n D_n^{(1)} x_{nd}$, where

$$D_n^{(1)} = \frac{-y_n}{1 + e^{y_n x_n^T \theta^*}}.$$

Noting that this satisfies $|D_n^{(1)}| \leq 1$:

$$\begin{aligned}
&\Pr\left[|\frac{1}{N}\sum_{n=1}^{N} D_n^{(1)} x_{nd}| \geq c_x \sqrt{\frac{25 + \log D}{CN}}\right] \\
&\leq \Pr\left[\sum_{n=1}^{N} |x_{nd}| \geq c_x \sqrt{N \frac{25 + \log D}{C}}\right] \\
&\leq 2e^{-25 - \log D},
\end{aligned}$$

where the final inequality comes from noting that $|x_{nd}|$ is also c_x-sub-Gaussian and using Hoeffding's inequality (Theorem 2.6.2 from [15]). Union bounding over all D dimensions of $\nabla F(\theta^*)$ gives the result. □

Lemma 9. *For the above setup for logistic regression and the λ given in Theorem 3, we have:*

$$\Pr\left[\max_{n=1,\ldots,N} \left\|\nabla F_{\backslash n}(\theta^*)\right\|_\infty \geq \frac{\lambda \max_n \max_{d \in S^c} \|J_{nd}\|_1}{4}\right] \leq 11e^{-25} \tag{C.42}$$

Proof. This follows from the same reasoning as in the proof of Lemma 6. □

C.11 Logistic regression: λ small enough

In the case of linear regression, the LSSC held with $K = 0$, so there was no work to be done in checking Eq. (3.6); this is not the case for logistic regression. [10] prove that the LSSC holds here:

Proposition 9. *The logistic regression model given above satisfies the (θ^*, N_{θ^*}) LSSC for any θ^* and N_{θ^*} with a data-dependent constant $K = 1/4(\max_n \|x_n\|_\infty)(\max_n \|x_{nS}\|_2^2)$.*

Proof. See the derivation in Section 6.2 of [10]. □

We first show that this random K is not too large with high probability under our random design:

Proposition 10. *For $x_n \in \mathbb{R}^D$ comprised of i.i.d. c_x-sub-Gaussian random variables, the random variable $K = 1/4(\max_n \|x_n\|_\infty)(\max_n \|x_{nS}\|_2^2)$ satisfies:*

$$\Pr\left[K \geq \frac{1}{4}\left(\sqrt{2c_x \log(ND)} + \sqrt{50c_x}\right)\left(c_x^2|S|(\log N + 26)\right)\right] \leq 3e^{-25} \tag{C.43}$$

Proof. First, Corollary 2 implies that $\max_n \|x_n\|_\infty \geq \sqrt{2c_x \log(ND)} + \sqrt{50c_x}$ with probability at most $2e^{-25}$, so the probability we are interested in is bounded by:

$$\leq \Pr\left[\max_n \|x_{nS}\|_2^2 \geq c_x^2|S|(\log N + 26)\right] + 2e^{-25}. \tag{C.44}$$

Noting that, as the sum of $|S|$ c_x^2-sub-Exponential random variables, each $\|x_{nS}\|_2^2$ is a $|S|c_x^2$-sub-Exponential random variable. Corollary 2 then gives us that Eq. (C.44) is bounded above by $3e^{-25}$. □

We can now prove the result we need, which is that λ satisfies the upper bound in Eq. (3.6) with high probability:

Lemma 10. *Assume that Eq.* (C.34) *holds with λ_{min} being $O(N)$ and that D is $o(e^N)$. Then, for the logistic regression setup above and λ as given in Theorem 3 and large enough N, we have:*

$$\Pr\left[\lambda \geq \frac{\min_n \lambda_{min}^2(\nabla_\theta^2 F_{\backslash n}(\theta^*))}{4\max_{n=1,\ldots,N}(\max_{d\in S^c}\|J_{nd}\|_1 + 4)^2}\frac{4\max_{d\in S^c}\|J_{nd}\|_1}{K}\right] \leq 13e^{-25} \tag{C.45}$$

Proof. Using Lemma 7, Lemma 8, and Proposition 10, the desired probability is $\leq 13e^{-25}$ if the following deterministic inequality holds:

$$\lambda \leq \frac{4(1-\alpha+M_J)}{4(1-\alpha+M_J+4)^2}\frac{(\lambda_{min} - Cc_x^2\sqrt{|S|N})^2}{\left(\sqrt{2c_x\log(ND)} + \sqrt{50c_x}\right)\left(c_x^2|S|(\log N + 26)\right)}$$

Plugging in the form of λ given in Theorem 3, rearranging, and noting that $(1-\alpha+$

$M_J)^2/(1-\alpha+M_J+4)^2 \leq 1$, we can equivalently check:

$$\frac{\sqrt{2c_x \log(ND)}}{N} + c_x\sqrt{\frac{25+\log D}{NC}} + \frac{\sqrt{50c_x}}{N} \leq \frac{(\lambda_{min} - Cc_x^2\sqrt{|S|N})^2}{\left(\sqrt{2c_x \log(ND)} + \sqrt{50c_x}\right)(c_x^2|S|(\log N + 26))} \tag{C.46}$$

For any scaling of N and D, the second term on the left hand side is dominant so that the LHS is $O(\sqrt{\log(D)/N})$. By assumption, the numerator of the right hand side is $O(N^2)$, whereas the demonimator is $O(\log^{3/2} N + \sqrt{\log N \log D})$. So, as long as N is large enough and D is $o(e^N)$, Eq. (C.46) is true[1]. □

[1]Looking at Eq. (C.46), a larger D would be fine; however, we state our result this way since other results require D to be $o(e^N)$.

Bibliography

[1] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 2010.

[2] A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh. On optimal generalizability in parametric learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3458–3468, 2017.

[3] P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76, September 1989.

[4] R. Giordano, W. Stephenson, R. Liu, M. I. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. *arXiv Preprint*, October 2018.

[5] I. Guyon, S. R. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[6] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* Chapman and Hall / CRC, 2015.

[7] D. Homrighausen and D. J. Mcdonald. Leave-one-out cross-validation is risk consistent for lasso. *Machine Learning*, 97(1-2):65–78, October 2014.

[8] J. D. Lee, Y. Sun, and J. E. Taylor. On model selection consistency of regularized m-estimators. *arXiv Preprint*, October 2014.

[9] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5, 2004.

[10] Y. Li, J. Scarlett, P. Ravikumar, and V. Cevher. Sparsistency of l1-regularized m-estimators. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[11] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[12] K. R. Rad and A. Maleki. A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv Preprint*, January 2018.

[13] R. van Handel. *Probability in High Dimensions.* Lecture Notes, December 2016. Accessible at https://web.math.princeton.edu/ rvan/APC550.pdf.

[14] V. Vapnik. Principles of risk minimization for learning theory. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 831–838. 1992.

[15] R. Vershynin. *High-dimensional probability: an introduction with applications in data science.* Cambridge University Press, August 2018.

[16] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5), 05 2009.

[17] S. Wang, W. Zhou, H. Lu, A. Maleki, and V. Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. In *International Conference in Machine Learning (ICML)*, 2018.

[18] H. Xu, C. Caramanis, and S. Mannor. Sparse algorithms are not stable: a no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1), 2012.

[19] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.