Sparse Approximate Cross-Validation for High-Dimensional GLMs

Abstract

Leave-one-out cross validation (LOOCV) can be particularly accurate among CV variants for estimating out-of-sample error. Unfortunately, LOOCV requires re-fitting a model $N$ times for a dataset of size $N$. To avoid this prohibitive computational expense, a number of authors have proposed approximations to LOOCV. These approximations work well when the unknown parameter is of small, fixed dimension but suffer in high dimensions; they incur a running time roughly cubic in the dimension, and, in fact, we show their accuracy significantly deteriorates in high dimensions. We demonstrate that these difficulties can be surmounted in $\ell_1$-regularized generalized linear models when we assume that the unknown parameter, while high dimensional, has a small support. In particular, we show that, under interpretable conditions, the support of the recovered parameter does not change as each datapoint is left out. This result implies that the previously proposed heuristic of only approximating CV along the support of the recovered parameter has running time and error that scale with the (small) support size even when the full dimension is large. Experiments on synthetic and real data support the accuracy of our approximations.

1 Introduction

Modern complex data analyses, on large data sets and in high dimensions, pose computational challenges not just for machine learning algorithms but also for evaluation methods. For instance, practitioners widely use cross validation (CV) to assess out-of-sample error and variability in machine learning methods. But CV requires substantial computation, namely re-running a machine learning algorithm many times, especially in the case of leave-one-out CV (LOOCV). This expense has led to recent proposals to approximate LOOCV [Obuchi and Kabashima, 2016, 2018, Beirami et al., 2017, Rad and Maleki, 2019, Wang et al., 2018, Giordano et al., 2019]. Theory and empirics demonstrate that these approximations are fast and accurate – as long as the dimension $D$ of the unknown parameter in a problem is low. Unfortunately a number of issues arise in high dimensions. These LOOCV approximations use local quadratic fits in the parameter space, which incur an $O(D^3)$ time cost for inverting a $D \times D$ matrix. In fact, these matrices may not even be invertible when $D$ is large relative to $N$. Moreover, existing error bounds for LOOCV approximations either assume a fixed $D$ or suffer from poor error scaling when $D$ grows with $N$. One might wonder whether the theory could be improved, but our own experiments (see, e.g., Fig. 1) confirm that LOOCV approximations can suffer considerable error degradation in high dimensions in practice. We emphasize, though, that this high-dimensional case is exactly the case where we need LOOCV the most, rather than the computationally cheaper $K$-fold CV; the fact that $K$-fold CV trains on a smaller portion of the dataset than LOOCV makes it a more biased estimate of the out-of-sample error [Arlot and Celisse, 2010, Sec. 5.1.1], and in many cases this effect becomes more extreme when $D$ is large relative to $N$ (e.g., [Rad and Maleki, 2019, Fig. 1, Wang et al. 2018, Fig. 1]).

In this work, we demonstrate that the difficulties of approximate LOOCV in high dimensions can be surmounted by carefully exploiting sparsity. We focus on generalized linear models (GLMs) as they have a clear mechanism for engendering sparsity (namely, setting many parameter components to zero). And we focus on $\ell_1$ regularization as an effective and popular means to recover a sparse parameter support. Since $\ell_1$-regularized penalties are not twice differentiable, they cannot be used directly with existing CV approximations. And we show empirically that smooth approximations to the $\ell_1$ penalty yield slow and inaccurate CV approximations. Instead, we validate a simple and
easy-to-use proposal: to first run the full machine learning algorithm once and then use approximate LOOCV only on the recovered (low-dimensional) support. This proposal will work if exact CV rounds recover a shared support. Our major theoretical contribution is to show that, under mild and interpretable conditions, the recovered support is in fact shared across rounds of LOOCV with very high probability. Obuchi and Kabashima [2016, 2018] have considered a similar setup and shown that the effect of the change in support is asymptotically negligible for $\ell_1$-regularized linear regression; however, they do not show the support is actually shared. We show that the support is shared with high probability in the practical finite-data setting – even for the very high-dimensional case $D = o(e^N)$ – and also extend to logistic regression. Our support stability result may be of independent interest and allows us to show that, with high probability under finite data, the error and time cost of our LOOCV approximation will depend on the support size – typically much smaller than the full dimension – rather than $D$. Our experiments confirm these findings empirically on simulated and real data.

2 Overview of Approximations

Let $\theta \in \Theta \subseteq \mathbb{R}^D$ be an unknown parameter of interest. Consider a dataset of size $N$, where $n \in [N] := \{1, 2, \ldots, N\}$ indexes the data point. Then a number of problems – such as maximum likelihood, general M-estimation, and regularized loss minimization – can be expressed as solving

$$
\hat{\theta} := \arg \min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} f_n(\theta) + \lambda R(\theta),
$$

where $\lambda \geq 0$ is a constant, and $R : \Theta \to \mathbb{R}_+$ and $f_n : \Theta \to \mathbb{R}$ are functions. For instance, $f_n$ might be the loss associated with the $n$th data point, $R$ the regularizer, and $\lambda$ the amount of regularization. Consider a data set where the $n$th data point has covariates $x_n \in \mathbb{R}^D$ and response $y_n \in \mathbb{R}$. In what follows, we will be interested in taking advantage of sparsity. With this in mind, we focus on generalized linear models (GLMs), where $f_n(\theta) = f(x_n^T \theta, y_n)$, as they offer a natural framework where sparsity can be expressed by choosing many parameter dimensions to be zero.

In LOOCV, we are interested in the solutions of the same problem with the $n$th data point removed. To that end, define $\hat{\theta}_{\setminus n} := \arg \min_{\theta \in \Theta} \frac{1}{N-1} \sum_{m \neq n} f_m(\theta) + \lambda R(\theta)$. Note our choice of $1/N$ scaling here – instead of $1/(N-1)$. While we believe this choice is not of particular importance in the case of LOOCV, we observe this scaling does not seem to be a settled issue in the literature; see Appendix A for a brief discussion. In any case, computing $\hat{\theta}_{\setminus n}$ exactly across $n$ usually requires $N$ runs of an optimization procedure – a prohibitive cost. Various approximations, which we detail next, address the poor performance in high dimensions.

Two approximations. Assume that $f$ and $R$ are twice differentiable functions of $\theta$. Let $F(\theta) := (1/N) \sum_n f(x_n^T \theta, y_n)$ be the unregularized objective, and let $H(\theta) := \nabla^2 f(\theta) + \lambda \nabla^2 R(\theta)$ be the Hessian matrix of the full objective. For the moment, we assume appropriate terms in each approximation below are invertible. Beirami et al. [2017], Rad and Maleki [2019], Wang et al. [2018] approximate $\hat{\theta}_{\setminus n}$ by taking a Newton step (“NS”) on the objective $(1/N) \sum_{m \neq n} f_m + \lambda R$ starting from $\hat{\theta}$; see Appendix B.4 for details. We thus call this approximation $\tilde{NS}_{\setminus n}(R)$ for regularizer $R$:

$$
\tilde{NS}_{\setminus n}(R) := \hat{\theta} + (1/N) \left( H(\hat{\theta}) - (1/N) \nabla^2 f_n(\hat{\theta}) \right)^{-1} \nabla \theta f_n(\hat{\theta}).
$$

In the case of GLMs, Theorem 8 of Rad and Maleki [2019] gives conditions on $x_n$ and $f(\cdot, \cdot)$ that imply, for fixed $D$, the error of $\tilde{NS}_{\setminus n}(R)$ averaged over $n$ is $o(1/N)$ as $N \to \infty$. 

![Figure 1: Log percent error (Eq. (9)) of existing approximate LOOCV methods (“IJ” and “NS”) as a function of data set size $N$ for $\ell_2$ regularized logistic regression. Dashed lines show Eq. (8) (“NS”) and solid show Eq. (9) (“IJ”). Blue lines have fixed data/parameter dimension $D$, while red lines have $D = N/10$, although the true parameter has a fixed support size of $D_{\text{eff}} = 2$ (see Section 2 for a full description). Existing methods fail to capture this low “effective dimension” and suffer from substantially worse performance in high dimensions.](image-url)
The second approximation we consider is due to Giordano et al. [2019]. While their work applies much more generally than LOOCV, its application to LOOCV was originally formulated only for \( \lambda = 0 \). We show in Appendix B that their approach can be adapted to the more general \( \lambda \geq 0 \) case. As their approximation is inspired by the infinitesimal jackknife ("I") [Jaekel 1972; Efron 1982], we denote it by \( \overline{I}_\gamma \); see Appendix B.1 for details.

\[
\overline{I}_\gamma := \bar{\theta} + (1/N)H(\bar{\theta})^{-1} \nabla \theta f_n(\bar{\theta}).
\]  

(3)

When \( \lambda = 0 \), Corollary 1 of Giordano et al. [2019] shows that the accuracy of Eq. (3) is bounded by \( C/N \) in general or, in the case of bounded gradients \( \| \nabla \theta f(x_n^T \theta, y_n) \| \leq B \), by \( C' B^2 / N^2 \). The constants \( C, C' \) may depend on \( D \) but not \( N \). Our Proposition 2 in Appendix B.3 extends this result to the case of \( \lambda \geq 0 \). Still, we are left with the fact that \( C \) and \( C' \) depend on \( D \) in an unknown way.

In what follows, we consider both \( \overline{NS}_\gamma \) and \( \overline{I}_\gamma \), as the two approximations have complimentary strengths. In terms of empirics, we find that \( \overline{NS}_\gamma \) performs better in our LOOCV GLM experiments here. But \( \overline{I}_\gamma \) is computationally efficient for a wider range of losses \( f_n \) and CV variants. E.g., one can derive an NS-style approximation for general leave-k-out CV for GLMs, but every round will require inversion of a matrix that differs from \( H(\bar{\theta}) \) by a new rank-k matrix. Meanwhile, the IJ approximation still requires only a single inversion, namely of \( H(\bar{\theta}) \). In terms of theory, \( \overline{NS}_\gamma \) has a generally tighter error bound of \( o(1/N) \), but given a good bound on the gradients, the theory behind \( \overline{I}_\gamma \) may provide a tighter rate.

**Problems in high dimensions.** The first challenge for both approximations given high dimension \( D \) is computational. Since every variant of CV or approximate CV requires running the machine learning algorithm of interest at least once, we will focus on the cost of the approximations after this single run. Given \( \bar{\theta} \), both approximations require the inversion of a \( D \times D \) matrix. Calculation of \( \overline{I}_\gamma \) across \( n \in [N] \) requires a single matrix inversion and \( N \) matrix multiplications for a runtime in \( O(D^3 + ND^2) \). In general, calculating \( \overline{NS}_\gamma \) has runtime of \( O(ND^3) \) due to needing an inversion for each \( n \). In the case of GLMs, though, \( \nabla \theta f_n \) is a rank-one matrix, so standard rank-one updates give a runtime of \( O(D^3 + ND^2) \) as well.

The second challenge for both approximations is the invertibility of \( H(\bar{\theta}) \) and \( H(\bar{\theta}) - (1/N)\nabla \theta f(x_n^T \theta, y_n) \) that was assumed in defining \( \overline{NS}_\gamma \) and \( \overline{I}_\gamma \). We note that, if \( \nabla^2 R(\bar{\theta}) \) is only positive semidefinite, then invertibility of both matrices may be impossible when \( D \geq N \), presenting a major problem in high dimension. We discuss this point further in Appendix B.2.

The third and final challenge for both approximations is accuracy in high dimensions. Not only do existing error bounds behave poorly (or not exist) in high dimensions, but empirical performance degrades as well. To create Fig. 1, we generated datasets from a sparse logistic regression model with \( N \) ranging from 500 to 5,000. For the blue lines, we set \( D = 2 \), and for the red lines we set \( D = N/10 \). In both cases, we see that error is much lower when \( D \) is small and fixed.

We recall that for large \( N \) and small \( D \), training error often provides a fine estimate of the out-of-sample error (e.g., see [Vapnik 1992]). That is, CV is needed precisely in the high-dimensional regime, and this case is exactly where current approximation struggles both computationally and statistically. Thus, we wish to understand whether there are high-D cases where approximate CV is useful. In what follows, we answer this question in the affirmative.

**Approximate CV with \( \ell_1 \) Regularization.** For statistical or interpretability reasons, we often wish to recover a parameter that has some low "effective dimension" \( D_{\text{eff}} \) with \( D_{\text{eff}} \ll D \). One way to achieve low \( D_{\text{eff}} \) is sparsity: i.e., we have \( D_{\text{eff}} := |\text{supp} \bar{\theta}| \ll D \), where \( \hat{S} := \text{supp} \bar{\theta} \) collects the indices of the non-zero entries of \( \bar{\theta} \). A way to achieve sparsity is choosing \( R(\bar{\theta}) = ||\theta||_1 \). We show that proper use of \( \ell_1 \) regularization allows approximate CV to be far more accurate and fast. Note that \( \overline{NS}_\gamma \) and \( \overline{I}_\gamma \) cannot be applied directly in this case as \( ||\theta||_1 \) is not twice-differentiable. One proposal put forward by Rad and Maleki [2019], [Wang et al. 2018] is to use a smoothed approximation to \( ||\cdot||_1 \); however, as we show in Section 4, this approach is both inaccurate and slow. We instead take the intuitive approach of approximating CV only on the dimensions in \( \text{supp} \bar{\theta} \).

For notation, let \( X \in \mathbb{R}^{N \times D} \) be the covariate matrix, with rows \( x_n \). For \( S \subseteq D \), let \( X_{\cdot, S} \) be the submatrix of \( X \) with column indices in \( S \); define \( x_n^{S} \) and \( \theta^S \) similarly. Let \( D_{\text{eff}}(2) := \left[ d^2 f(z, y_n)/dz^2 \right]_{z = x_n^T \bar{\theta}} \), and define the restricted Hessian evaluated at \( \bar{\theta} \): \( H_{SS}(2) := X_{\cdot, S}^T \text{diag}(D_{\text{eff}}(2))X_{\cdot, S} \). Further define the LOO restricted Hessian, \( H_{\hat{S}\hat{S}} := H_{SS} - [\nabla \theta f(x_n^T \bar{\theta}, y_n)]_{\hat{S}\hat{S}} \). Finally, without loss of generality, assume \( \hat{S} = \{1, 2, \ldots, \hat{D}_{\text{eff}}\} \). We now define
versions of $\bar{\text{NS}}_\setminus(R)$ and $\bar{\text{IJ}}_\setminus(R)$ restricted to the entries in $\text{supp } \hat{\theta}$:

$$
\text{NS}_\setminus := \left( \hat{\theta}_S + (H^{-1}_S)^{\dagger} \begin{bmatrix} \nabla \phi f(x'_{n,S} \hat{\theta}, y_n) \end{bmatrix}_S \right), \quad \text{IJ}_\setminus := \left( \hat{\theta}_S + H^{-1}_S \begin{bmatrix} \nabla \phi f(x'_{n,S} \hat{\theta}, y_n) \end{bmatrix}_S \right).
$$

(4)

Other authors have previously considered $\text{NS}_\setminus$. Rad and Maleki [2019], Wang et al. [2018] derive $\text{NS}_\setminus$ by considering a smooth approximation to $\ell_1$ and then taking the limit of $\bar{\text{NS}}_\setminus(R)$ as the amount of smoothness goes to zero. In Appendix [C] we show a similar argument can yield $\text{IJ}_\setminus$. Also, Obuchi and Kabashima [2016, 2018] directly propose the special case of $\text{NS}_\setminus$ for linear and logistic regression without using $\bar{\text{NS}}_\setminus(R)$ as a starting point. We now show how $\text{NS}_\setminus$ and $\text{IJ}_\setminus$ avoid many of the high-dimensional challenges with $\bar{\text{NS}}_\setminus(R)$ and $\bar{\text{IJ}}_\setminus(R)$ we discussed above.

The first challenge was that compute time for $\bar{\text{NS}}_\setminus(R)$ and $\bar{\text{IJ}}_\setminus(R)$ scaled poorly with $D$. That $\text{NS}_\setminus$ and $\text{IJ}_\setminus$ do not share this issue is immediate from their definitions.

**Proposition 1.** For general $f_n$, the time to compute $\text{NS}_\setminus$ or $\text{IJ}_\setminus$ scales with $D_{\text{eff}}$, rather than $D$. In particular, computing $\text{NS}_\setminus$ across all $n \in [N]$ takes $O(N D^3_{\text{eff}})$ time, and computing $\text{IJ}_\setminus$ across all $n \in [N]$ takes $O(D_{\text{eff}}^3 + N D_{\text{eff}}^2)$ time. Furthermore, when $f_n$ takes the form of a GLM, computing $\text{NS}_\setminus$ across all $n \in [N]$ takes $O(D_{\text{eff}}^3 + N D_{\text{eff}}^2)$ time.

Thus, when $D_{\text{eff}} \ll D$, $\text{NS}_\setminus$ and $\text{IJ}_\setminus$ will be much faster to compute than $\bar{\text{NS}}_\setminus(R)$ and $\bar{\text{IJ}}_\setminus(R)$. The second challenge was that $H$ and $H^{\text{inv}}$ may not be invertible when $D \geq N$. Notice the relevant matrices in Eq. (4) are of dimension $D_{\text{eff}} = |S|$. So, to calculate $\text{NS}_\setminus$ and $\text{IJ}_\setminus$, we need only make the much less restrictive assumption that $D_{\text{eff}} < N$, rather than $D < N$. We address the third and final challenge of accuracy in the next section.

### 3 Approximation quality in high dimensions

Recall that the accuracy of $\bar{\text{NS}}_\setminus(R)$ and $\bar{\text{IJ}}_\setminus(R)$ in general has a poor dependence on dimension $D$. We now show that the accuracy of $\text{NS}_\setminus$ and $\text{IJ}_\setminus$ depends on (the hopefully small) $D_{\text{eff}}$ rather than $D$. We start by assuming a “true” population parameter $\theta^* \in \mathbb{R}^D$ that minimizes the population-level loss, $\theta^* := \arg\min \mathbb{E}_{x,y}[f(x^T \theta, y)]$, where the expectation is over $x, y$ from some population distribution. Assume $\theta^*$ is sparse with $S := \text{supp } \theta^*$ and $D_{\text{eff}} := |S|$. Our parameter estimate would be faster and more accurate if an oracle told us $S$ in advance and we worked just over $S$:

$$
\hat{\phi} := \arg\min_{\phi \in \mathbb{R}^{D_{\text{eff}}}} \frac{1}{N} \sum_{n=1}^{N} f(x'_{n,S} \phi, y_n) + \lambda \| \phi \|_1.
$$

(5)

We define $\hat{\phi}_\setminus$ as the leave-one-out variant of $\hat{\phi}$ (as $\hat{\theta}_\setminus$ is to $\hat{\theta}$). Let $\text{RNS}_\setminus$ and $\text{RIJ}_\setminus$ be the result of applying the approximation in $\text{NS}_\setminus$ or $\text{IJ}_\setminus$ to the restricted problem in Eq. (5); note that $\text{RNS}_\setminus$ and $\text{RIJ}_\setminus$ have accuracy that scales with the (small) dimension $D_{\text{eff}}$.

Our analysis of the accuracy of $\text{NS}_\setminus$ and $\text{IJ}_\setminus$ will depend on the idea that if, for all $n$, $\text{NS}_\setminus$ and $\text{IJ}_\setminus$ run over the same $D_{\text{eff}}$-dimensional subspace, then the accuracy of $\text{NS}_\setminus$ and $\text{IJ}_\setminus$ must be identical to that of $\text{RNS}_\setminus$ and $\text{RIJ}_\setminus$. In the case of $\ell_1$ regularization, this idea specializes to the following condition, under which our main result in Theorem [I] will be immediate.

**Condition 1.** For all $n \in [N]$, we have $\text{supp } \text{IJ}_\setminus = \text{supp } \text{NS}_\setminus = \text{supp } \hat{\theta}_\setminus = S$.

**Theorem 1.** Assume Condition [I] holds. Then for all $n$, $\hat{\theta}_\setminus$ and $\text{IJ}_\setminus$ are (1) zero outside the dimensions $S$ and (2) equal to their restricted counterparts from Eq. (5):

$$
\hat{\theta}_\setminus = \left( \hat{\theta}_{\setminus, S} \right) = \left( \hat{\phi}_{\setminus} \right), \quad \text{IJ}_\setminus = \left( \text{IJ}_{\setminus, S} \right) = \left( \text{RIJ}_{\setminus} \right).
$$

(6)

It follows that the error is the same in the full problem as in the low-dimensional restricted problem: $\| \hat{\theta}_\setminus - \text{IJ}_\setminus \|_2 = \| \hat{\phi}_\setminus - \text{RIJ}_\setminus \|_2$. The same results hold for $\text{IJ}_\setminus$ and $\text{RIJ}_\setminus$ replaced by $\text{NS}_\setminus$ and $\text{RNS}_\setminus$.

---

\(^1\)This assumption may not be necessary to prove the dependence of $\text{NS}_\setminus$ and $\text{IJ}_\setminus$ on $D_{\text{eff}}$, but it allows us to invoke existing $\ell_1$ support results in our proofs.
Again, Theorem 1 is immediate if one is willing to assume Condition 1, but when does Condition 1 hold? There exist general conditions in the $\ell_1$ literature under which $\text{supp} \hat{\theta} = S$ (Lee et al., 2014, Li et al., 2015). If one assumed these conditions held for all $F^n := (1/N) \sum_{m \neq n} f_m$, then Condition 1 would directly follow. However, it is not immediate that any models of interest meet these conditions. In what follows, we give a few interpretable assumptions under which Condition 1 holds.

In fact, we show that we need just four principal assumptions in the case of linear and logistic regression; we conjecture that similar results hold for other GLMs. The first assumption arises from the intuition that, if individual data points are very extreme, the support will certainly change for some $n$. To avoid these extremes with high probability, we assume that the covariates follow a sub-Gaussian distribution:

**Definition 1.** (Vershynin, 2018) For $c_e > 0$, a random variable $V$ is $c_e$-sub-Gaussian if $\mathbb{E} \left[ \exp \left( V^2 / c_e^2 \right) \right] \leq 2$.

**Assumption 1.** Each $x_n \in \mathbb{R}^D$ has zero-mean i.i.d. $c_e$-sub-Gaussian entries with $\mathbb{E} [x_{id}^2] = 1$.

We conjecture that the unit-variance part of the assumption is unnecessary. Conditions on the distributions of the responses $y_n$ will be specific to linear and logistic regression and will be given in Assumptions 5 and 6 respectively. Our results below will hold with high probability under these distributions. Note there are reasons to expect we cannot do better than high-probability results. In particular, Xu et al. (2012) suggest that there exist worst-case training datasets for which sparsity-inducing methods like $\ell_1$ regularization are not stable as each datapoint is left out.

Our second principal assumption is a full-data high-probability incoherence condition.

**Assumption 2.** The incoherence condition holds with high probability over the data:

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

Authors in the $\ell_1$ literature often assume that incoherence holds deterministically for a given design matrix $X$—starting from the introduction of incoherence in Zhao and Yu (2006) and continuing in more recent work (Lee et al., 2014, Li et al., 2015). Similarly, we will take our high probability version in Assumption 2 as given. But we note that Assumption 2 is at least known to hold for the case of linear regression with an i.i.d. Gaussian design matrix (e.g., see Exercise 11.5 of Hastie et al., 2013). We next place some restrictions on how quickly $D$ and $D_{\text{eff}}$ grow as functions of $N$.

**Assumption 3.** As functions of $N$, $D$ and $D_{\text{eff}}$ satisfy: (1) $D = o(e^N)$, (2) $D_{\text{eff}} = o([N/\log N]^{2/5})$, and (3) $D_{\text{eff}}^{3/2} \sqrt{\log D} = o(N)$.

The constraints on $D$ here are particularly loose. While those on $D_{\text{eff}}$ are tighter, we still allow polynomial growth of $D_{\text{eff}}$ in $N$ for some lower powers of $N$. Our final assumption is on the smallest entry of $\theta^*_S$. Such conditions—typically called beta-min conditions—are frequently used in the $\ell_1$ literature to ensure $\hat{S} = S$ (Wainwright, 2009, Lee et al., 2014, Li et al., 2015).

**Assumption 4.** $\theta^*$ satisfies $\min_{s \in S} \theta^*_s > \sqrt{D_{\text{eff}} T_{\text{min}}} \lambda$, where $T_{\text{min}}$ is some constant relating to the objective function $f$; see Assumption 13 in Appendix G.1 for an exact description.

### 3.1 Linear regression

We now give the distributional assumption on the responses $y_n$ in the case of linear regression and then show that Condition 1 holds.

**Assumption 5.** $\forall n, y_n = x_n^T \theta^* + \varepsilon_n$, where the $\varepsilon_n$ are i.i.d. $c_e$-sub-Gaussian random variables.

**Theorem 2** (Linear Regression). Take Assumptions 1 to 5. Suppose the regularization parameter $\lambda$ satisfies

$$
\lambda \geq \frac{C}{\alpha - M_{\text{lin}}} \left( \frac{c_e^2 c_e^2 \log D}{N} + \frac{25c_e^2 c_e^2}{N} + \frac{4c_e c_e (\log(ND) + 26)}{N} \right),
$$

where $C > 0$ is a constant in $N, D, D_{\text{eff}}, c_e$, and $c_e$, and $M_{\text{lin}}$ is a scalar given by Eq. (35) in Appendix G.1 that satisfies, as $N \to \infty$, $M_{\text{lin}} = o(1)$. Then for $N$ sufficiently large, Condition 1 holds with probability at least $1 - 26c^{-25}$.

A full statement and proof of Theorem 2 including the exact value of $M_{\text{lin}}$, appears in Appendix G. A corollary of Theorem 1 and Theorem 2 together is that, under Assumptions 1 to 5, the LOOCV approximations $\hat{D}_n$ and $\hat{N}_n$ have accuracy that depends on (the ideally small) $D_{\text{eff}}$ rather than (the potentially large) $D$. 


It is worth considering how the allowed values of $\lambda$ in Eq. (7) compare to previous results in the $\ell_1$ literature for the support recovery of $\hat{\theta}$. We will talk about a sequence of choices of $\lambda$ scaling with $N$ denoted by $\lambda_N$. Theorem 11.3 of [Hastie et al. 2015] shows that $\lambda_N \geq c\sqrt{\log(D)/N}$ (for some constant $c$ in $D$ and $N$) is sufficient for ensuring that $\text{supp} \hat{\theta} \subseteq S$ with high probability in the case of linear regression. Thus, we ought to set $\lambda_N \geq c\sqrt{\log(D)/N}$ to ensure support recovery of $\hat{\theta}$. Compare this to the bound implied by Eq. (7). We have that $M_{\text{fin}} = o(1)$ as $N \to \infty$, so that, for large $N$, the bound in Eq. (7) becomes $\lambda_N \geq c'\sqrt{\log(D)/N}$ for some constant $c'$. We immediately see that the sequence of $\lambda_N$ satisfying Eq. (7) scales at exactly the same rate as those that ensure $\text{supp} \hat{\theta} \subseteq S$. This is important, as the error in $\hat{\theta}$, $\|\hat{\theta} - \theta^*\|_2^2$, is typically proportional to $\lambda_N$. The fact that we have not increased the asymptotic scaling of $\lambda_N$ therefore means that we can enjoy the same decay of $\|\hat{\theta} - \theta^*\|_2^2$ while ensuring $\text{supp} \hat{\theta}_{\lambda_n} = S$ for all $n$.

3.2 Logistic regression

We now give the distributional assumption on the responses $y_n$ in the case of logistic regression.

**Assumption 6.** $\forall n$, we have $y_n \in \{\pm 1\}$ with $\Pr[y_n = 1] = 1/(1 + e^{-\log(1+\eta)}).

We will also need a condition on the minimum eigenvalue of the Hessian.

**Assumption 7.** Assume for some scalar $L_{\text{min}}$ that may depend on $N, D_{\text{eff}},$ and $c_x$, we have

$$\Pr[\lambda_{\text{min}} \left(\nabla_y^2 F(\theta)_{SS}\right) \leq L_{\text{min}}] \leq e^{-25}.$$ 

Furthermore, assume the scaling of $L_{\text{min}}$ in $N$ and $D_{\text{eff}}$ is such that, under Assumption 3 and for sufficiently large $N$, $L_{\text{min}} \geq CN$ for some constant $C$ that may depend on $c_x$.

In the case of linear regression, we did not need an analogue of Assumption 7 as standard matrix concentration results tell us that its Hessian satisfies Assumption 7 with $L_{\text{min}} = N - Cc_x^2\sqrt{D_{\text{eff}}}$ (see Lemma 2 in Appendix C). The Hessian for logistic regression is significantly more complicated, and it is typical in the $\ell_1$ literature to make some kind of assumption about its eigenvalues [Bach 2011, Li et al. 2015]. Empirically, Assumption 7 is satisfied when Assumptions 1 and 6 hold; however we are unaware of any results in the literature showing this is the case.

**Theorem 3 (Logistic Regression).** Take Assumptions 7, 8, 9, 10, and 11. Suppose the regularization parameter $\lambda$ satisfies:

$$\lambda \geq \frac{C}{\alpha - M_{\loggr}} \left(\sqrt{c_x^2 + \frac{25 + \log{D}}{N}} + \frac{\sqrt{2c_x^2 \log(ND) + 50c_x^2}}{N}\right),$$

where $C, C'$ are constants in $N, D, D_{\text{eff}},$ and $c_x$, and $M_{\loggr}$ is a scalar given by Eq. (66), that, as $N \to \infty$, satisfies $M_{\loggr} = o(1)$. Then for $\lambda$ sufficiently large, Condition 7 is satisfied with probability at least $1 - 43e^{-25}$.

A restatement and proof of Theorem 3 are given as Theorem 3 in Appendix C. Similar to the remarks after Theorem 2, Theorem 3 implies that when applied to logistic regression, the LOOCV approximations $\text{iJ}_{\lambda_n}$ and $\text{nS}_{\lambda_n}$ have accuracy that depends on (the ideally small) $D_{\text{eff}}$ rather than (the potentially large) $D$, even when $D = o(\sqrt{N})$.

Theorem 3 has implications for the work of Obuchi and Kabashima [2018]. The authors of that paper conjecture that, as $N \to \infty$, the change in support of $\ell_1$ regularized logistic regression becomes negligible as each datapoint is left out and use this assumption to derive a version of $\text{nS}_{\lambda_n}$ customized to logistic regression. Our Theorem 3 confirms this conjecture by proving the stronger fact that the support is actually unchanged with high probability for finite data.

4 Experiments

**Two alternatives to the limiting approximation.** Our theoretical results imply that $\text{nS}_{\lambda_n}$ and $\text{iJ}_{\lambda_n}$ have high accuracy and low computational cost. We begin our empirical investigation of this claim by first comparing $\text{iJ}_{\lambda_n}$ to two more straightforward alternatives: (A) just use Eq. (3) with some smoothed version of $\|\theta\|_1$, or (B) approximate exact CV by exactly computing $\hat{\theta}_{\lambda_n}$ for just a few random values of $n$. To illustrate (A), we consider the smooth approximation given by Rad and Maleki [2019]: $R^\eta(\theta) := \sum_{d=1}^{D} \frac{1}{\eta}(\log(1+e^{\eta\theta_d}) + \log(1+e^{-\eta\theta_d}))$. While $\lim_{\eta \to \infty} R^\eta(\theta) = \|\theta\|_1$, we found that this approximation became numerically unstable for the purposes of optimization when $\eta$ was much larger than 100, so we set $\eta = 100$ in our experiments. To test (A) and (B), we trained logistic regression models on twenty-five random datasets in which $x_{nd} \sim N(0,1)$ with $N = 500$ and $D = 40,000$. The true $\theta^*$ was supported on its first five entries. We computed our approximation to CV as: $\text{ALOO} = \frac{1}{N} \sum_{n=1}^{N} f(x_n^T \text{iJ}_{\lambda_n}, y_n)$, and we computed
Sparse Approximate Cross-Validation for High-Dimensional GLMs

Figure 2: (Left): Error (Eq. (9)) across approximations for $\ell_1$ LOOCV (legend shared with right). The error for $\tilde{I}_J|_n$ (blue) is too small to see, but nonzero; it varies between $-0.06\%$ and $0.04\%$. (Right): Runtimes for the same experiment with exact CV (red) included for comparison. The $D \times D$ matrix inversion in the smoothed problem is so slow that even exact CV with an efficient $\ell_1$ solver is faster.

Figure 3: We emphasize the logarithmic vertical scale in both plots. (Left): Percent accuracy (Eq. (9)) for real data experiments (legend shared with right). For each dataset, we give the accuracy of approximate CV compared to exact CV using both a smoothed approximation to $\ell_1$ and the $\tilde{I}_J|_n$, $\tilde{N}_S|_n$ approximations considered here. In the case of the bcTCGA dataset (linear regression), the nearly quadratic objective seems to be extremely well approximated by one Newton step, making $\tilde{N}_S|_n(R^\theta)$ significantly more accurate than $\tilde{I}_J|_n(R^\theta)$ (see the note at the end of Appendix B.4 about the exactness of $\tilde{N}_S|_n(R)$ on quadratic objectives). (Right): Timings for the same experiments, with exact CV included for comparison.

The exact CV estimate, denoted by LOO, similarly. We computed the accuracy of our approximation as a percent error:

$$\frac{|\hat{\theta}_{\text{LOO}} - \text{LOO}|}{\text{LOO}}.$$  \hspace{3cm} (9)

Fig. 2 compares the accuracy and run times of (A) and (B) versus $\tilde{I}_J|_n$. We chose the number of subsamples so that subsampling CV would have about the same runtime as computing $\tilde{I}_J|_n$ for all $n$. In this case, we see that its accuracy is much worse than $\tilde{I}_J|_n$ for nearly every trial. Using Eq. (3) with $R^{100}(\theta)$ as a regularizer is far worse: while subsampling exact CV is fast and unbiased (although high variance), the smoothed approximation has to form an approximation over all $D$ dimensions; the resulting approximation is orders of magnitude less accurate and slower.

The importance of correct support recovery. Our theoretical results show that each $\hat{\theta}_{\text{LOO}}$ having correct support (i.e., $\text{supp} \hat{\theta}_{\text{LOO}} = \text{supp} \theta^*$) is a sufficient condition for obtaining the fixed-dimensional error scaling shown in blue in Fig. 1. One might wonder if such a condition is necessary for approximate CV to be accurate. We offer evidence in Appendix D that this is the case. In particular, we empirically show that when $\text{supp} \theta$ grows with $N$, the error in $\tilde{I}_J|_n$ also grows with $N$.

Real Data Experiments. We next aim to study how dependent our results are on the particular random design we chose in Theorems 2 and 3. We explore this question with a number of publicly available datasets [bcTCA, 2018].

\footnote{Specifically, we computed 41 different $\hat{\theta}_{\text{LOO}}$ for each trial in order to roughly match the time cost of computing $\tilde{I}_J|_n$ for all $N = 500$ datapoints.}
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 E. For each dataset, we approximate CV for the $\ell_1$ regularized model using $I_{J_n}$ and $NS_{I_n}$. For comparison, we report the accuracy of the $I_{J_n}(R^n)$ and $NS_{I_n}(R^n)$ with $\eta = 100$. Our results, reported in Fig. 3 show that approximate CV for the $\ell_1$ regularized problem is significantly faster and more accurate.

To briefly demonstrate the scalability of our approximations, we re-ran our RCV1 experiment on a larger version of the dataset with $N = 20,242$ and $D = 30,000$. Based on the time to compute exact LOOCV for twenty datapoints, we estimate exact LOOCV would have taken over two weeks to complete, whereas computing both $NS_{I_n}$ and $I_{J_n}$ for all $n$ took three minutes.

### 5 Selection of $\lambda$ and future work

We have shown that in both theory and practice, we can expect $I_{J_n}$ and $NS_{I_n}$ to harness the low effective dimension present in $\ell_1$ regularized GLMs to give significant improvements in both computation time and approximation accuracy. However, these results, as well as those in all previous works on approximate CV, only apply when assessing the generalization error for a single fixed model (i.e., for a fixed $\lambda$). These analyses do not address one of the more common uses of CV, which is to train with many different values of $\lambda$ and select the one with the lowest CV error.

In Appendix E, we give some empirical evidence that approximate CV methods are sometimes – but not always – suitable for this purpose. Specifically, we show that the infinitesimal jackknife approximation sometimes selects $\lambda = 0$. Fortunately, this failure mode is easy to detect, suggesting the following workflow: run approximate CV to select $\lambda$, and in the event that $\lambda = 0$ is selected, run exact CV instead. Still, this suggestion is purely based on empirics. We believe that understanding the uses and limitations of approximate CV for selecting $\lambda$ is one of the most important directions for future work in this area.

### Acknowledgements

This research is supported in part by DARPA, the CSAIL-MSR Trustworthy AI Initiative, an NSF CAREER Award, an ARO YIP Award, and ONR.

### References

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

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4, 2010.

bcTCGA. *Breast cancer gene expression data*, Nov 2018. Available at http://myweb.uiowa.edu/pbrohenn/data/bcTCGA.html.

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.

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated Lasso. *arXiv Preprint*, jan 2019.

B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*, volume 38. Society for Industrial and Applied Mathematics, 1982.

R. Giordano, W. T. Stephenson, R. Liu, M. I. Jordan, and T. Broderick. A Swiss army infinitesimal jackknife. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, April 2019.

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.

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

D. Homrighausen and D. J. McDonald. The lasso, persistence, and cross-validation. In *International Conference in Machine Learning (ICML)*, 2013.

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.

L. Jaeckel. The infinitesimal jackknife, memorandum. Technical report, MM 72-1215-11, Bell Lab. Murray Hill, NJ, 1972.
J. D. Lee, Y. Sun, and J. E. Taylor. On model selection consistency of regularized m-estimators. *arXiv Preprint*, Oct 2014.

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.

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.

L. Miolane and A. Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv Preprint*, November 2018.

T. Obuchi and Y. Kabashima. Cross validation in lasso and its acceleration. *Journal of Statistical Mechanics*, May 2016.

T. Obuchi and Y. Kabashima. Accelerating cross-validation in multinomial logistic regression with l1-regularization. *Journal of Machine Learning Research*, September 2018.

K. R. Rad and A. Maleki. A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv Preprint*, February 2019.

R. van Handel. *Probability in High Dimensions*. Lecture Notes, December 2016.

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.

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

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.

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.

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.

J. Xu, A. Maleki, K. R. Rad, and D. Hsu. Consistent risk estimation in high-dimensional linear regression. *arXiv Preprint*, February 2019.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.
A Scaling of the leave-one-out objective

We defined $\hat{\theta}_{\setminus n}$ as the solution to the following optimization problem:

$$
\hat{\theta}_{\setminus n} := \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{m \neq n} f_m(\theta) + \lambda R(\theta).
$$

An alternative would be to use the objective $1/(N-1) \sum_{m \neq n} f_m + \lambda R$ in order to keep the scaling between the regularizer and the objective the same as in the full-data problem. Indeed, all existing theory that we are aware of for CV applied to $\ell_1$ regularized problems seems to follow the $1/(N-1)$ scaling [Homrighausen and McDonald 2014, 2013, Molalane and Montanari 2018, Chetverikov et al. 2019]. On the other hand, all existing approaches to approximate LOOCV for regularized problems have used the $1/N$ scaling that we have given [Beirami et al. 2017, Rad and Maleki 2019, Wang et al. 2018, Xu et al. 2019, Obuchi and Kabashima 2016, 2018]. Note that the scaling is not relevant in Giordano et al. [2019], as they do not consider the regularized case. As our work is aimed at identifying when existing approximations work well in high dimensions, we have followed the literature on approximate LOOCV. The different results from using the two scalings may be insignificant when leaving only one datapoint out. But one might expect the difference to be substantial for, e.g., $K$-fold CV. We leave an understanding of what the effect of this scaling is (if any) to future work.

B Further details of Eq. (2) and Eq. (3)

In Section 2, we briefly outlined the approximations $\tilde{\mathcal{S}}_{\setminus n}(R)$ and $\tilde{\mathcal{I}}_{\setminus n}(R)$ to $\hat{\theta}_{\setminus n}$; we give more details about these approximations and their derivations here. Recall that we defined $H(\hat{\theta}) := (1/N) \sum_{n=1}^{N} \nabla^2_{\theta} f(x_n^T \hat{\theta}, y_n) + \lambda \nabla^2 \theta R(\hat{\theta})$. We first restate the "infinitesimal jackknife" approximation from the main text, which was derived by the same approach taken by Giordano et al. [2019]:

$$
\hat{\theta}_{\setminus n} \approx \tilde{\mathcal{I}}_{\setminus n}(R) := \hat{\theta} + \frac{1}{N} H(\hat{\theta})^{-1} \nabla \theta f(x_n^T \hat{\theta}, y_n).
$$

The "Newton step" approximation, similar to the approach in Beirami et al. [2017] and identical to the approximation in Rad and Maleki [2019], Wang et al. [2018], is:

$$
\hat{\theta}_{\setminus n} \approx \tilde{\mathcal{S}}_{\setminus n}(R) := \hat{\theta} + \frac{1}{N} \left( H(\hat{\theta}) - \frac{1}{N} \nabla^2_{\theta} f(x_n^T \hat{\theta}, y_n) \right)^{-1} \nabla \theta f(x_n^T \hat{\theta}, y_n).
$$

B.1 Derivation of $\tilde{\mathcal{I}}_{\setminus n}(R)$

We will see in Appendix B.3 that, after some creative algebra, $\tilde{\mathcal{I}}_{\setminus n}(R)$ is an instance of $\hat{\theta}_{ij}$ from Definition 2 of Giordano et al. [2019]. However, this somewhat obscures the motivation for considering Eq. (10). As an alternative to jamming our problem setup into that considered by Giordano et al. [2019], we can more directly obtain the approximation in Eq. (10) by a derivation only slightly different from that in Giordano et al. [2019]. We begin by defining $\hat{\theta}^w$ as the solution to a weighted optimization problem with weights $w_n \in \mathbb{R}$:

$$
\hat{\theta}^w := \arg\min_{\theta \in \Theta} G(w, \theta) := \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} w_n f(x_n^T \theta, y_n) + \lambda R(\theta),
$$

where we assume $G$ to be twice continuously differentiable with an invertible Hessian at $\hat{\theta}^1$ (where $\hat{\theta}^1$ is the solution in Eq. (12) with all $w_n = 1$). For example, we have that $\hat{\theta}_{\setminus n} = \hat{\theta}^w$ if $w$ is the $N$-dimensional vector of all ones but with a zero in the $n$th coordinate. We will form a linear approximation to $\hat{\theta}^w$ as a function of $w$. To do so, we will need to compute the derivatives $d\hat{\theta}^w/dw_n$ for each $n$. To compute these derivatives, we begin with the first order optimality condition of Eq. (12) and take a total derivative with respect to $w_n$:

$$
\frac{\partial G}{\partial \theta} \bigg|_{w=1, \theta=\hat{\theta}^1} = 0 \quad \Rightarrow \quad \frac{d}{dw_n} \frac{\partial G}{\partial \theta} \bigg|_{w=1, \theta=\hat{\theta}^1} = \frac{\partial^2 G}{\partial \theta \partial w_n} \bigg|_{w=1, \theta=\hat{\theta}^1} dw_n + \frac{\partial^2 G}{\partial \theta^2} \bigg|_{w=1, \theta=\hat{\theta}^1} d\hat{\theta}^w dw_n \bigg|_{w=1} = 0.
$$
Re-arranging, defining $H(\hat{\theta}^1) := \nabla_{\theta}^2 \mathbb{G}(w, \hat{\theta}^1)$, and using the assumed invertibility of $H(\hat{\theta}^1)$ gives:

$$\frac{d\hat{\theta}}{dw_n} \bigg|_{w=1} = -\left( \frac{\partial^2 G}{\partial \theta^2} \bigg|_{w=1, \theta=\hat{\theta}^1} \right)^{-1} \frac{\partial^2 G}{\partial w_n \partial \theta} \bigg|_{w=1, \theta=\hat{\theta}^1} = -\frac{1}{N} H(\hat{\theta})^{-1} \nabla_{\theta} f(x_n^T \hat{\theta}, y_n). \quad (13)$$

In the final equality, we used the fact that $\hat{\theta}^1 = \hat{\theta}$. Now, by a first order Taylor expansion around $w = (1, 1, \ldots, 1)$, we can write:

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

$$= \hat{\theta} - \frac{1}{N} \sum_{n=1}^{N} H(\hat{\theta})^{-1} \nabla_{\theta} f(x_n^T \hat{\theta}, y_n)(w_n - 1). \quad (15)$$

For the special case of $w$ being the vector of all ones with a zero in the $n$th coordinate (i.e., the weighting for LOOCV), we recover Eq. (10).

### B.2 Invertibility in the definition of $\widetilde{\text{NS}}_\setminus(R)$ and $\widetilde{\text{JI}}_\setminus(R)$

In writing Eqs. (2) and (3) we have assumed the invertibility of $H(\hat{\theta})$ and $H(\hat{\theta}) - (1/N) \nabla_{\theta} f(x_n^T \theta, y_n)$. Here we note a number of common cases where this invertibility holds. First, if $\nabla^2 R$ is positive definite for all $\theta$ (as in the case of $R = \| \cdot \|_2^2$), then these matrices are always invertible. If $R$ is merely convex, then $H(\hat{\theta})$ is always invertible if $N \leq D$. If $R$ is convex, $H(\hat{\theta}) - (1/N) \nabla_{\theta} f(x_n^T \theta, y_n)$ is invertible if $\text{Span} \{ x_m \}_{m:n \neq n} = \mathbb{R}^D$. This condition on the span holds almost surely if the $x_n$ are sampled from a continuous distribution and $D < N$.

### B.3 Accuracy of $\widetilde{\text{JI}}_\setminus(R)$ for regularized problems

As noted in the main text, [Giordano et al. 2019] show that the error of $\widetilde{\text{JI}}_\setminus(R)$ is bounded by $C/N$ for some $C$ that is constant in $N$. However, their results apply only to the unregularized case (i.e., $\lambda = 0$). We show here that their results can be extended to the case of $\lambda > 0$ with mild additional assumptions; the proof of Proposition 2 appears below.

**Proposition 2.** Assume that the conditions for Corollary 1 of [Giordano et al. 2019] are satisfied by $F(\theta)$. Furthermore, assume that we are restricted to $\theta$ in some compact subset $\Theta$ of $\mathbb{R}^D$, $\lambda = O(1/\sqrt{N})$, $F + \lambda R$ is twice continuously differentiable for all $\theta$, and that $\nabla^2 R(\theta)$ is positive definite for all $\theta \in \Theta$. Then $\widetilde{\text{JI}}_\setminus(R)$ can be seen as an application of the approximation in Definition 2 of [Giordano et al. 2019]. Furthermore, the assumptions of their Corollary 1 are met, which implies:

$$\max_{n \in [N]} \| \hat{\psi}_{I,n}^n - \hat{\theta}_{\setminus n} \|_2 \leq \frac{C'}{N^2} \sup_{\theta \in \Theta} \max_{n \in [N]} \| \nabla_{\theta} f(x_n^T \theta, y_n) \|_\infty \leq \frac{C'}{N}, \quad (16)$$

where $C$ and $C'$ are problem-specific constants independent of $N$ that may depend on $D$.

Proposition 2 provides two bounds on the error $\| \widetilde{\text{JI}}_\setminus(R) - \hat{\theta}_{\setminus n} \|_2$: either $C'/N^2$ times the maximum of the gradient or just $C/N$. One bound or the other may be easier to use, depending on the specific problem. It is worth discussing the conditions of Proposition 2 before going into its proof. The first major assumption is that $\theta$ is restricted to some compact set $\Theta$. Although this assumption may not be satisfied by problems of interest, one may be willing to assume that $\theta$ lives in some bounded set in practice. In any case, such an assumption seems necessary to apply the results of [Giordano et al. 2019] to most unregularized problems, as they, for example, require $\sup_{\theta \in \Theta} F(\theta)$ to be bounded. We will require the compactness of $\Theta$ to show that $\sup_{\theta \in \Theta} F(\theta) + \lambda R(\theta)$ is bounded.

The second major assumption of Proposition 2 is that $\lambda = O(1/\sqrt{N})$. We need this assumption to ensure that the term $\lambda R(\theta)$ is sufficiently well behaved. In practice this assumption may be somewhat limiting; however, we note that for fixed $D$, such a scaling is usually assumed—and in some situations is necessary— to obtain standard theoretical results for $\ell_1$ regularization (e.g., [Wainwright 2009] gives the standard scaling for linear regression, $\lambda = \Omega(\sqrt{\log(D)/N})$). Our Theorems 2 and 3 also satisfy such a scaling when $D$ is fixed. In any case, we stress that this assumption—as well as the assumption on compactness—are needed only to prove Proposition 2 and not any of our other results. We prove Proposition 2 to demonstrate the baseline results that exist in the literature so that we can then show how our results build on these baselines.
We will also need a set of weight vectors \( W \).

**Proof.** We proceed by showing that the regularized optimization problem in our Eq. (1) can be written in the framework of Eq. (1) of [Giordano et al. 2019] and then showing that the re-written problem satisfies the assumptions of their Corollary 1. First, the framework of [Giordano et al. 2019] applies to weighted optimization problems of the form:

\[
\hat{\theta}^w := \theta \in \Theta \text{ s.t. } \frac{1}{N} \sum_{n=1}^{N} w_n g_n(\theta) = 0. \tag{17}
\]

In order to match this form, we will rewrite the gradient of the objective in Eq. (1) as a weighted sum with \( N + 1 \) terms, where the first term, with weight \( w_0 = 1 \), will correspond to \( R(\theta) \):

\[
\frac{1}{N+1} w_0 (N+1) \lambda \nabla R(\theta) + \frac{1}{N+1} \sum_{n=1}^{N} w_n \frac{N+1}{N} \nabla f(x_n^T \theta, y_n). \tag{18}
\]

We will also need a set of weight vectors \( W \subseteq \mathbb{R}^{N+1} \) for which we are interested in evaluating \( \hat{\theta}^w \). We choose this set as follows. In the set, we include each weight vector that is equal to one everywhere except \( w_n = 0 \) for exactly one of \( n \in \{1, \ldots, N\} \). Thus, for each \( n \), there is a \( w \in W \) such that \( \hat{\theta}^w = \theta_{\backslash n} \). Finally, then, we can apply Definition 2 of [Giordano et al. 2019] to find the approximation \( \theta_{IJn}(w) \) for the \( w \) that corresponds to leaving out \( n \). We see that \( \theta_{IJn}(w) \) in this case is exactly equal to \( \text{I\textbackslash}_n(R) \) in our notation here.

Now that we know our approximation is actually an instance of \( \theta_{IJn}(w) \), we need to check that Eq. (18) meets the assumptions of Corollary 1 of [Giordano et al. 2019] to apply their theoretical analysis to our problem. We check these below, first stating the assumption from [Giordano et al. 2019] and then covering why it holds for our problem.

1. **(Assumption 1):** for all \( \theta \in \Theta \), each \( g_n \) is continuously differentiable in \( \theta 
\)

For our problem, by assumption, \( R(\theta) \) and \( f(x_n^T \theta, y_n) \) are twice continuously differentiable functions of \( \theta \), so this assumption holds.

2. **(Assumption 2):** for all \( \theta \in \Theta \), the Hessian matrix, \( H(\theta, 1) := (1/N) \sum \partial^2 g_n(\theta)/\partial \theta^T \) is invertible and satisfies \( \| H^{-1}(\theta, 1) \|_{op} \leq C_{op} < \infty \) for some constant \( C_{op} \), where \( \| \cdot \|_{op} \) denotes the operator norm on matrices with respect to the \( \ell_2 \) norm (i.e., the maximum eigenvalue of the matrix).

For our problem, by assumption, the inverse matrix \( (\nabla^2 F(\theta))^{-1} \) exists and has bounded maximum eigenvalue for all \( \theta \in \Theta \). Also by assumption, \( R \) has a positive semidefinite Hessian for all \( \theta \), which implies:

\[
\sup_{\theta \in \Theta} \| H^{-1}(\theta, 1) \|_{op} = \sup_{\theta \in \Theta} \left\| (\nabla^2 F(\theta) + \lambda \nabla^2 R(\theta))^{-1} \right\|_{op} \leq \sup_{\theta \in \Theta} \left\| (\nabla^2 F(\theta))^{-1} \right\|_{op}.
\]

To see that the inequality holds, first note that for a positive semi-definite (PSD) matrix \( A, \| A^{-1} \|_{op} = 1/\lambda_{\min}(A) \). The inequality would then follow if \( \lambda_{\min}(\nabla^2 F(\theta) + \lambda \nabla^2 R(\theta)) \geq \lambda_{\min}(\nabla^2 F(\theta)) \). To see that this holds, take any two \( D \times D \) PSD matrices \( A \) and \( B \). Let \( \lambda_d(\cdot) \) be the \( d \)th eigenvalue of a matrix with \( \lambda_1 = \lambda_{\min} \). Then:

\[
\lambda_d(A + B) = \min_{E \subseteq \mathbb{R}^D} \max_{x \in E} \frac{x^T (A + B) x}{\|x\|_2^2} = \min_{E \subseteq \mathbb{R}^D} \max_{x \in E} \frac{x^T A x}{\|x\|_2^2} = \lambda_d(A),
\]

where the inequality holds because \( B \) is PSD. So, \( \lambda_{\min}(A + B) \geq \lambda_{\min}(A) \), which finishes the proof. We have thus showed that the operator norm of \( H^{-1}(\theta, 1) \) is bounded by that of \( (\nabla^2 F(\theta))^{-1} \) for all \( \theta \in \Theta \).

3. **(Assumption 3):** Let \( g(\theta) \) and \( h(\theta) \) be the \((N+1) \times D \) stack of gradients and \((N+1) \times D \) stack of Hessians, respectively. That is, \( g(\theta)_{nd} := (\nabla g(x_n^T \theta, y_n))_d \) for \( n = 1, \ldots, N \) and \( g(\theta)_{N+1,d} := (\nabla g(R(\theta)))_d \) with \( h \) defined similarly. Let \( \| g(\theta) \|_2 \) be the \( \ell_2 \) norm of \( g \) flattened into a vector with \( \| h(\theta) \|_2 \) defined similarly. Then assume that there exist constants \( C_g \) and \( C_h \) such that:

\[
\sup_{\theta \in \Theta} \frac{1}{N+1} \| g(\theta) \|_2 \leq C_g < \infty
\]

\[
\sup_{\theta \in \Theta} \frac{1}{N+1} \| h(\theta) \|_2 \leq C_h < \infty
\]
To see that this holds for our problem, we have that:
\[
\|g(\theta)\|_2 := \left[ \sum_{d=1}^{D} \left( (\lambda(N+1)\nabla R(\theta))_d + \frac{N+1}{N} \sum_{n=1}^{N} \left( \nabla f(x_n^T \theta, y_n)_d \right)^2 \right) \right]^{1/2} 
\leq \lambda(N+1)\|\nabla R(\theta)\|_2 + \frac{N+1}{N} \left[ \sum_{d=1}^{D} \sum_{n=1}^{N} \left( \nabla f(x_n^T \theta, y_n)_d \right)^2 \right]^{1/2}.
\]

We need to show this is bounded by \(\sqrt{N+1}C_g\) for some constant \(C_g\). By assumption in the statement of Proposition 2 we have \(\frac{1}{\sqrt{N+1}} \|\nabla F(\theta)\|_2 \leq \frac{1}{\sqrt{N}} \|\nabla F(\theta)\|_2 \leq C_g^{(1)}\) for some constant \(C_g^{(1)}\). Because \(\lambda\) is \(O(1/\sqrt{N})\), the first term is equal to \(O(\sqrt{N})\|\nabla R(\theta)\|_2\). The compactness of \(\Theta\) and the continuity of \(\nabla R(\theta)\) imply that \(\|\nabla R(\theta)\|_2\) is bounded by a constant for all \(\theta \in \Theta\). So, we know that \(\frac{1}{\sqrt{N}} \|\nabla R(\theta)\|_2 \leq C_g^{(2)}\) for some constant \(C_g^{(2)}\). Thus, we have that the assumption on \(\|g(\theta)\|_2\) holds with \(C_g = (N+1)C_g^{(1)} + C_g^{(2)}\). That the condition on \(\|h(\theta)\|_2\) follows by the same reasoning.

4. (Assumption 4): There exists some \(\Delta_\theta > 0\) and \(L_h < \infty\) such that if \(\|\theta - \hat{\theta}\|_2 \leq \Delta_\theta\), then
\[
\frac{1}{\sqrt{N+1}} \|h(\theta) - h(\hat{\theta})\|_2 \leq L_h \|\theta - \hat{\theta}\|_2.
\]
We can show this holds for our problem by:
\[
\|h(\theta) - h(\hat{\theta})\|_2 := \left\| \frac{N+1}{N} \nabla^2 F(\theta) + \lambda(N+1)\nabla^2 R(\theta) - \frac{N+1}{N} \nabla^2 F(\hat{\theta}) - \lambda(N+1)\nabla^2 R(\hat{\theta}) \right\|_2 
\leq (N+1)\lambda \left\| \nabla^2 R(\theta) - \nabla^2 R(\hat{\theta}) \right\|_2 + \frac{N+1}{N} \left\| \nabla^2 F(\theta) - \nabla^2 F(\hat{\theta}) \right\|_2,
\]
where we have abused notation to denote \(\|\nabla^2 F(\theta)\|_2 := \sqrt{\sum_{d,j=1}^{D,N} \sum_{n=1}^{N} \nabla^2_2 f(x_n^T \theta, y_n)_{ij}^2}\). Now, we want to show that this quantity divided by \(\sqrt{N+1}\) is bounded by \(L_h\|\theta - \hat{\theta}\|_2\) for some constant \(L_h\). By assumption in the statement of Proposition 2 we have that Assumption 4 holds for \(F\); this implies that \(\frac{1}{\sqrt{N+1}} \left\| \nabla^2 F(\theta) - \nabla^2 F(\hat{\theta}) \right\|_2 \leq L_h^{(1)} \|\theta - \hat{\theta}\|_2\) for some constant \(L_h^{(1)}\). As \(R\) is twice continuously differentiable and the condition of Assumption 4 needs only to hold over a compact set of \(\theta\)'s, we know that \(\nabla^2 R(\theta)\) is Lipschitz over this domain. Using this along with the assumption that \(\lambda\) is \(O(1/\sqrt{N})\), we have that:
\[
\frac{\lambda(N+1)}{\sqrt{N+1}} \left\| \nabla^2 R(\theta) - \nabla^2 R(\hat{\theta}) \right\|_2 = O(1) \left\| \nabla^2 R(\theta) - \nabla^2 R(\hat{\theta}) \right\|_2 
\leq L_h^{(2)} \|\theta - \hat{\theta}\|_2,
\]
for some constant \(L_h^{(2)}\). So, Assumption 4 holds with constant \(L_h = L_h^{(1)} + L_h^{(2)}\).

5. (Assumption 5): For all \(w \in W\), we have \(\frac{1}{\sqrt{N+w}} \|w\|_2 \leq C_w\) for some constant \(C_w\). This is immediately true for our definition of \(W\), which, for all \(w \in W\), has \(\|w\|_2 = \sqrt{N}\).

\[\Box\]

B.4 Derivation of \(\widetilde{NS}_{\setminus n}(R)\)

Wang et al. 2018 and Rad and Maleki 2019 derive \(\widetilde{NS}_{\setminus n}(R)\) in Eq. (11) by taking a single Newton step on the objective \(F^{\setminus n} + \lambda R\) starting at the point \(\hat{\theta}\). For completeness, we include a derivation here. Recall that the objective with one datapoint left out is:

\[
F^{\setminus n}(\theta) + \lambda R(\theta) := \frac{1}{N} \sum_{m=1}^{N} f(x_m^T \theta, y_m) - \frac{1}{N} f(x_n^T \theta, y_n) + \lambda R(\theta),
\]

(19)
which has $H(\theta) - (1/N)\nabla^2_\theta f(x_n^T \hat{\theta}, y_n)$ as its Hessian. Now consider approximating $\hat{\theta}_{\backslash n}$ by performing a single Newton step on $F^{\backslash n}$ starting from $\hat{\theta}$:

$$\hat{\theta}_{\backslash n} \approx \hat{\theta} - \left( H(\hat{\theta}) - \frac{1}{N} \nabla^2_\theta f(x_n^T \hat{\theta}, y_n) \right)^{-1} \left( \frac{1}{N} \sum_{n=1}^{N} \nabla f(x_n^T \hat{\theta}, y_n) - \frac{1}{N} \nabla f(x_n^T \hat{\theta}, y_n) + \lambda \nabla R(\hat{\theta}) \right). \tag{20}$$

Using the fact that, by definition of $\hat{\theta}$, $(1/N) \sum_{n=1}^{N} \nabla f(x_n^T \hat{\theta}, y_n) + \lambda \nabla R(\hat{\theta}) = 0$, we have that this simplifies to:

$$\hat{\theta}_{\backslash n} \approx \hat{\theta} + \frac{1}{N} \left( H(\hat{\theta}) - \frac{1}{N} \nabla^2_\theta f(x_n^T \hat{\theta}, y_n) \right)^{-1} \nabla f(x_n^T \hat{\theta}, y_n), \tag{21}$$

which is exactly $\tilde{NS}_{\backslash n}(R)$.

As $\tilde{NS}_{\backslash n}(R)$ can be interpreted as a single Newton step on the objective $F^{\backslash n} + \lambda R$, it follows that $\tilde{NS}_{\backslash n}(R)$ is exactly equal to $\hat{\theta}_{\backslash n}$ in the case that $F^{\backslash n} + \lambda R$ is a quadratic, as noted by [Beirami et al. 2017]. For example, $\ell_2$ regularized linear regression has $\tilde{NS}_{\backslash n}(R) = \hat{\theta}_{\backslash n}$ for all $n$. We further note that somewhat similar behavior holds for $\ell_1$ regularized linear regression. Specifically, when $\text{sign} \hat{\theta} = \text{sign} \hat{\theta}_{\backslash n}$, we have that the objective $F^{\backslash n} + \lambda \| \theta \|_1$ is a quadratic when restricted to the dimensions in $\tilde{S}$. In this case, $NS_{\backslash n}$ can be interpreted as taking a Newton step on $F^{\backslash n} + \lambda \| \cdot \|_1$ restricted to the dimensions in $\tilde{S}$. It follows that $NS_{\backslash n} = \hat{\theta}_{\backslash n}$ when $\text{sign} \hat{\theta} = \text{sign} \hat{\theta}_{\backslash n}$ for $\ell_1$ regularized linear regression.

### B.5 Computation time of approximations

There is a major computational difference between Eq. (11) and Eq. (10): 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 $O(ND^3)$ versus a cost of $O(D^3)$. Even for small $D$, this is a significant additional expense.

However, as noted by [Rad and Maleki 2019, Wang et al. 2018], Eq. (11) is much cheaper when considering the special case of generalized linear models. In this case, $\nabla^2_\theta f_{\backslash n}$ 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. (11) given only $H^{-1}$; this is how Equation 8 in [Rad and Maleki 2019] and Equation 21 in [Wang et al. 2018] are derived. Even though we only consider GLMs in this work, we still study Eq. (10) with the hope of retaining scalability in more general problems.

### C Derivation of $IJ_{\backslash n}$ and $NS_{\backslash n}$ via smoothed approximations

As noted in Section 2 [Rad and Maleki 2019, Wang et al. 2018] derive the $NS_{\backslash n}$ approximation by considering $\tilde{NS}_{\backslash n}(R^\eta)$ with $R^\eta$ being some smoothed approximation to the $\ell_1$ norm, and then taking the limit of $\tilde{NS}_{\backslash n}(R^\eta)$ as the amount of smoothness goes to zero. We review this approach and then state our Proposition 3 which says that the same technique can be used to derive $IJ_{\backslash n}$.

We first give two possible ways to smooth the $\ell_1$ norm. The first is given by [Rad and Maleki 2019]:

$$\| \theta \|_1 \approx R^\eta(\theta) := \sum_{d=1}^{D} \frac{1}{\eta} \left( \log(1 + e^{\eta \theta_d}) + \log(1 + e^{-\eta \theta_d}) \right), \tag{22}$$

The second option is to use the more general smoothing framework described by [Wang et al. 2018]. They allow selection of a function $q: \mathbb{R} \rightarrow \mathbb{R}$ satisfying: (1) $q$ has compact support, (2) $\int q(u) \, du = 1$, $q(0) > 0$, and $q \geq 0$, and (3) $q$ is symmetric around 0 and twice continuously differentiable on its domain, and then define a smoothed approximation:

$$R^\eta(\theta) := \eta \sum_{d=1}^{D} \int_{-\infty}^{\infty} |u| q(\eta(\theta_d - u)) \, du, \tag{23}$$

In both Eqs. (22) and (23), we have $\lim_{\eta \rightarrow \infty} = \| \theta \|_1$. Notice that either choice of $R^\eta$ is twice differentiable for any $\eta < \infty$, so one can consider the approximations $\tilde{NS}_{\backslash n}(R^\eta), IJ_{\backslash n}(R^\eta)$. We now state two assumptions, both of which are given by [Rad and Maleki 2019, Wang et al. 2018], under which one can show the limits of these approximations as $\eta \rightarrow \infty$ are equal to $NS_{\backslash n}$ and $IJ_{\backslash n}$.
Assumption 8. For any element \( \hat{z} \in \mathbb{R}^D \) of the subdifferential \( \partial \|\theta\|_1 \) evaluated at \( \hat{\theta} \) such that \( \nabla F(\hat{\theta}) + \lambda \hat{z} = 0 \), we have \( \|\hat{z}\|_{\infty} < 1 \).

Assumption 9. For any \( y_n \in \mathbb{R} \), \( f(z,y_n) \) is a twice continuously differentiable function as a function of \( z \in \mathbb{R} \).

Proposition 3 (Theorem 1 of Rad and Maleki [2019]; Theorem 4.2 of Wang et al. [2018]). Take Assumptions 8 and 9. Suppose \( H_{SS}^{\infty} \) has strictly positive eigenvalues. Let \( H_{SS}^{\infty} = H_{SS} - \|\nabla_2^2 f(x_n^T \hat{\theta}, y_n)\|_{SS} \) and suppose that, for all \( n \), \( H_{SS}^{\infty} \) is invertible. Then, for \( R^0 \) as in Eq. (22) or Eq. (23),

\[
\text{NS}_{\lambda} := \lim_{n \to \infty} \text{NS}_{\lambda}(R^0) = \left( \hat{\theta}_S + (H_{SS}^{\infty})^{-1} \nabla_2 f(x_n^T \hat{\theta}, y_n) \right)_S.
\]  

(24)

As noted in the main text, we show that a very similar result holds for the limit of \( \text{I}_I\lambda_n(R^0) \):

Proposition 4. Take Assumptions 8 and 9. Suppose \( H_{SS}^{\infty} \) is invertible. Then for \( R^0 \) as in Eq. (22) or Eq. (23),

\[
\text{I}_I\lambda_n := \lim_{n \to \infty} \text{I}_I\lambda_n(R^0) = \left( \hat{\theta}_S + H_{SS}^{\infty} \nabla_2 f(x_n^T \hat{\theta}, y_n) \right)_S.
\]  

(25)

The proof of Proposition 4 is a straightforward adaptation of the proof of Proposition 3. We prove it separately for the two different forms of \( R^0 \) in the next two subsections.

C.1 Proof of Proposition 4 Using Eq. (22)

This proof is almost identical to the proof of Theorem 1 from Rad and Maleki [2019]. First we will need some notation. Let \( \hat{\theta}^0 \) be the solution to Eq. (1) using \( R^0 \) from Eq. (22) as the regularizer. Let \( \hat{S}_n := \{ i : |\hat{\theta}_n| > c/\eta \} \) for some constant \( c \). We now use the arguments in Appendix A.1 of Rad and Maleki [2019] that for an appropriately chosen \( c \) and \( \eta > C \) for some large constant \( C > 0 \), we have \( S^\eta = \hat{S} = \supp \hat{\theta} \). Next, define the scalars \( \hat{D}_n^{(1,\eta)} \) and \( \hat{D}_n^{(2,\eta)} \) as the derivatives of \( \hat{D}_n^{(1,\eta)} \) and \( \hat{D}_n^{(2,\eta)} \) as the derivatives of \( f \) evaluated at \( \hat{\theta} \):

\[
\hat{D}_n^{(1,\eta)} := \frac{d f(z,y_n)}{dz} \bigg|_{z=x_n^T \hat{\theta}^0}, \quad \hat{D}_n^{(2,\eta)} := \frac{d^2 f(z,y_n)}{dz^2} \bigg|_{z=x_n^T \hat{\theta}^0}.
\]  

(26)

Finally, divide the Hessian of the smooth problem up into blocks by defining:

\[
A := X_{\cdot,\hat{S}_n}^T \diag \{ \hat{D}_n^{(2,\eta)} \} X_{\cdot,\hat{S}_n} + \lambda \nabla^2 \hat{R}^\eta(\hat{\theta}), \quad B := X_{\cdot,\hat{S}_n}^T \diag \{ \hat{D}_n^{(2,\eta)} \} X_{\cdot,\hat{S}_n} + \lambda \nabla^2 \hat{R}^\eta(\hat{\theta})
\]

\[
C := X_{\cdot,\hat{S}_n}^T \diag \{ \hat{D}_n^{(2,\eta)} \} X_{\cdot,\hat{S}_n} + \lambda \nabla^2 \hat{R}^\eta(\hat{\theta}), \quad D := (A - BC^{-1} B^T)^{-1}
\]

We can then compute the block inverse of the Hessian of the smooth problem, \( H_n^{-1} \) as:

\[
H_n^{-1} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} = \begin{pmatrix} D & -DBC^{-1} \\ -C^{-1} B^T D & A^{-1} + A^{-1} BDB \end{pmatrix}
\]  

(27)

Rad and Maleki [2019] show that all blocks of \( H_n^{-1} \) converge to zero as \( \eta \to \infty \) except for the upper left, which has \( D \to X_{\cdot,\hat{S}_n}^T \diag \{ \hat{D}_n^{(2,\eta)} \} X_{\cdot,\hat{S}_n} \). So we have that the limit of \( \text{I}_I\lambda_n(R^0) \) is:

\[
\lim_{\eta \to \infty} \text{I}_I\lambda_n(R^0) = \lim_{\eta \to \infty} H_n^{-1} \hat{D}_n^{(1,\eta)} x_n = \hat{D}_n^{(1)} \left( X_{\cdot,\hat{S}_n}^T \diag \{ \hat{D}_n^{(2)} \} X_{\cdot,\hat{S}_n} \right)^{-1} \begin{pmatrix} x_n \hat{S} \\ 0 \end{pmatrix},
\]  

(28)

where we used that \( \hat{\theta}_n \to \hat{\theta} \) by Lemma 3 of Rad and Maleki [2019], which gives that \( \hat{D}_n^{(1,\eta)} \to \hat{D}_n^{(1)} \) by Assumption 9. The resulting approximation is exactly that given in the statement of Proposition 4 by noting that \( \hat{D}_n^{(1)} x_n \hat{S} = [\nabla f(x_n^T \hat{\theta}, y_n)]_\hat{S} \).

C.2 Proposition 4 Using Eq. (23)

This proof proceeds along the exact same direction as when using Eq. (22). In their proof of their Theorem 4.2, Wang et al. [2018] provide essentially all the same ingredients that Rad and Maleki [2019] do, except for the general class of smoothed approximations given by Eq. (23). This allows the same argument of taking the limit of each block of the Hessian individually and finishing by taking the limit as in Eq. (28).
D The importance of correct support recovery

Theorem 1 shows that each \( \hat{\theta}_n \) having correct support (i.e., \( \text{supp} \hat{\theta}_n = \text{supp} \theta^* \)) is a sufficient condition for obtaining the fixed-dimensional error scaling shown in blue in Fig. 1. Here, we give some brief empirical evidence that this condition is necessary in the case of linear regression when using \( \text{IJ}_n \), as an approximation. For values of \( N \) ranging from 1,000 to 8,000, we set \( D = N/10 \) and generate a design matrix with i.i.d. \( N(0,1) \) entries. The true \( \theta^* \) is supported on its first five entries, with the rest set to zero. We then generate observations \( y_n = x_n^T \theta^* + \varepsilon_n \), for \( \varepsilon_n \sim N(0,1) \).

To examine what happens when the recovered supports are and are not correct, we use slightly different values of the regularization parameter \( \lambda \). Specifically, the results of Wainwright [2009] (especially their Theorem 1) tell us that the support recovery of \( \ell_1 \) regularized linear regression will change sharply around \( \lambda \approx 4\sqrt{\log(D)/N} \), where lower values of \( \lambda \) will fail to correctly recover the support. With this in mind, we choose two settings of \( \lambda \): \( 1.0\sqrt{\log(D)/N} \) and \( 10.0\sqrt{\log(D)/N} \). As expected, the right-hand side of Fig. 4 shows that the accuracy of \( \text{IJ}_n \) is drastically different in these two situations. The lefthand plot of Fig. 4 offers an explanation for this observation: the support of \( \hat{\theta}_n \) grows with \( N \) under the lower value of \( \lambda \), whereas the larger value of \( \lambda \) ensures that \( |\text{supp} \hat{\theta}_n| = |\text{supp} \theta^*| = \text{const} \). Empirically, these results suggest that, for high-dimensional problems, approximate CV methods are accurate estimates of exact CV only when taking advantage of some kind of low “effective dimensional” structure.

Figure 4: Illustration of the role of support recovery in the accuracy of \( \text{IJ}_n \) in the case of linear regression. Left: Points show the average of \( |\text{supp} \hat{\theta}_n| \) over random values of \( n \). Error bars show the min and max \( |\text{supp} \hat{\theta}_n| \) over these \( n \). For \( \lambda = 1.0\sqrt{\log(D)/N} \) (blue), the mean recovered support is constant with \( N \). For \( \lambda = 10.0\sqrt{\log(D)/N} \) (red), \( |\text{supp} \hat{\theta}_n| \) grows with \( N \), and varies dramatically for different values of \( n \). Right: Percent error (Eq. (9)) as \( D \) scales with \( N \). When the support recovery is constant, we recover an error scaling of roughly \( 1/N^2 \), whereas a growing support results in a much slower decay.

That the approximation quality relies so heavily on the exact setting of \( \lambda \) is somewhat concerning. However, we emphasize that sensitivity exists for \( \ell_1 \) regularization in general; as previously noted, Wainwright [2009] demonstrated similarly drastic behavior of \( \hat{\theta} \) in the same exact linear regression setup that we use here. On the other hand, Homrigausen and McDonald [2014] do show that using exact LOOCV to select \( \lambda \) for \( \ell_1 \) regularized linear regression gives reasonable results. In Appendix E we empirically show this is sometimes, but not always, the case for our and other approximate CV methods.

E Details of real experiments

We use three publicly available datasets for our real-data experiments in Section 4.

1. The “Gisette” dataset [Guyon et al. 2004] is available from the UCI repository at [https://archive.ics.uci.edu/ml/datasets/Gisette](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” [bcTCGA 2018] is a dataset of breast cancer samples from The Cancer Genome Atlas, which we downloaded from [http://myweb.uiowa.edu/pbreheny/data/bcTCGA.html](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. 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 [Lewis et al. 2004] 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 use 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 latter 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 covariate matrix \( X \) is extremely sparse (i.e., most entries are 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 \( \ell_1 \) regularized regression on each of these datasets, we first needed to select a value of \( \lambda \). Since all of these datasets are fairly high dimensional, our experiments in Appendix B suggests our approximation will be inaccurate for values of \( \lambda \) that are “too small.” In an attempt to get the order of magnitude for \( \lambda \) correct, we used the theoretically motivated value of \( \lambda = C \sqrt{\log(D)/N} \) for some constant \( C \) (e.g., [Li et al. 2015] shows this scaling of \( \lambda \) will recover the correct support for both linear and logistic regression). Section 4 suggests that the constant \( C \) can be very important for the accuracy of our approximation, and our experiments there suggest that inaccuracy is caused by too large a recovered support size \(|\hat{\theta}|\). For the RCV1 and Gisette datasets, both run with logistic regression, we guessed a value of \( C = 1.5 \), as this sits roughly in the range of values that give support recovery for logistic regression on synthetic datasets. After confirming that \(|\hat{\theta}|\) was not too large (i.e., of size ten or twenty), we proceeded with these experiments. Although we found linear regression on synthetic data typically needed a larger value of \( C \) than logistic regression on synthetic data, we found that \( C = 1.5 \) also produced reasonable results for the bcTCGA dataset.

F Selection of \( \lambda \)

We generate two synthetic \( \ell_1 \) regularized logistic regression problems with \( N = 300 \) observations and \( D = \{75, 150\} \) dimensions. The matrix of covariates \( X \) has i.i.d. \( N(0, 1) \) entries, and the true \( \theta^* \) has its first five entries drawn i.i.d. as \( N(0, 1) \) with the rest set to zero. 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 \( \lambda \), we find \( \hat{\theta} \), and measure the train, test, exact LOOCV, and approximate LOOCV errors via both \( \text{NS}_{\lambda,n} \) and \( \text{IJ}_{\lambda,n} \); the results are plotted in Fig. 5: \( \text{NS}_{\lambda,n} \) (magenta curve) is an extremely close approximation to exact CV (red curve) in both datasets and selects a \( \lambda \) that gives a test error very close to the \( \lambda \) selected by exact CV. On the other hand, \( \text{IJ}_{\lambda,n} \) (blue curve) performs very differently on the two datasets. For \( D = 75 \), it selects a somewhat reasonable value for \( \lambda \); however, for \( D = 150 \), \( \text{IJ}_{\lambda,n} \) goes disastrously wrong by selecting the obviously incorrect value of \( \lambda = 0 \). While the results in Fig. 5 come from using our \( \text{IJ}_{\lambda,n} \) to approximate CV for an \( \ell_1 \) regularized problem, we note that this issue is not specific to the current work; we observed similar behavior when using \( \ell_2 \) regularization and the pre-existing \( \text{IJ}_{\lambda,n}(\ell_2) \).

While \( \text{NS}_{\lambda,n} \) performs far better than \( \text{IJ}_{\lambda,n} \) in the experiments here, it too has a limitation when \( D > N \). In particular, when \( \lambda \) is small enough, we will eventually recover \(|\hat{S}| = N \). At this point, the matrix we need to invert in the definition of \( \text{NS}_{\lambda,n} \) in Eq. 4 will be a \( N \times N \) matrix that is the sum of \( N - 1 \) rank-one matrices. As such, it will not be invertible, meaning that we cannot compute \( \text{NS}_{\lambda,n} \) for small \( \lambda \) when \( D > N \). Even when \( D \) is less than – but still close to – \( N \), we have observed numerical issues in computing \( \text{NS}_{\lambda,n} \) when \( \lambda \) is sufficiently small; typically, these issues show up as enormously large values for ALOO for small values of \( \lambda \).

Given the above discussion and the extremely common use of CV to select \( \lambda \), we believe that an understanding of the behavior of \( \text{IJ}_{\lambda,n} \) and \( \text{NS}_{\lambda,n} \) for the purposes of hyperparameter tuning is a very important direction for future work.

G Proofs from Section 3

As mentioned in the main text, there exist somewhat general assumptions in the \( \ell_1 \) literature under which \( \text{supp} \hat{\theta} = \hat{S} \) [Lee et al. 2014, Li et al. 2015]. By taking these assumptions for all leave-one-out problems, we immediately get that \( \text{supp} \hat{\theta}_{\lambda,n} = S \) for all \( n \). Our method for proving Theorems 2 and 3 will be to show that the assumptions of those theorems imply those from the \( \ell_1 \) literature for all leave-one-out problems.
G.1 Assumptions from Li et al. [2015]

We choose to use the conditions from Li et al. [2015], as we find them easier to work with for our problem. Li et al. [2015] gives conditions on $F = (1/N) \sum f(x, \theta; y)$ under which $\text{supp } \hat{\theta} = S$. We are interested in $\text{supp } \hat{\theta} \setminus n$, so we state versions of these conditions for $F_{\setminus n} = (1/N) \sum_{m \neq n} f(x, \theta; y)$. 

Assumption 10 (LSSC). $\forall n, F_{\setminus n}$ satisfies the $(\theta^*, R^D)$ locally structured smoothness condition (LSSC) with constant $K$. We recall this condition, due to Li et al. [2015], in Appendix G.2.

Assumption 11 (Strong convexity). For a matrix $A$, let $\lambda_{\text{min}}(A)$ be the smallest eigenvalue of $A$. Then, $\forall n$ and for some constant $L_{\text{min}}$, the Hessian of $F_{\setminus n}$ is positive definite at $\theta^*$ when restricted to the dimensions in $S$: $\lambda_{\text{min}}(\nabla^2 F_{\setminus n}(\theta^*)_{SS}) \geq L_{\text{min}} > 0$.

Assumption 12 (Incoherence). $\forall n$ and for some $\gamma > 0$,

$$\left\| \nabla^2 F_{\setminus n}(\theta^*)_{SS}^{-1} \nabla^2 F_{\setminus n}(\theta^*)_{SS} \right\|_{\infty} < 1 - \gamma. \quad (29)$$

Assumption 13 (Bounded gradient). For $\gamma$ from Assumption 12, $\forall n$, the gradient of $F_{\setminus n}$ evaluated at the true parameters $\theta^*$ is small relative to the amount of regularization: $\left\| \nabla F_{\setminus n}(\theta^*) \right\|_{\infty} \leq (\gamma/4)\lambda$.

Assumption 14 (λ sufficiently small). For $K$, $L_{\text{min}}$ and $\gamma$ as in Assumptions 10 to 12, the regularization parameter is sufficiently small: $\lambda < L_{\text{min}}^{-1} \gamma/(4(\gamma + 4)^2 D_{\text{eff}} K)$, where there is no constraint on $\lambda$ if $K = 0$.

We see in Appendix G.3 that a minor adaptation of Theorem 5.1 from Li et al. [2015] tells us that Assumptions 10 to 14 imply $\forall n, \text{supp } \hat{\theta} \setminus n \subseteq S$. To prove the accuracy of $\text{NS}_{\setminus n}$ and $\text{IJ}_{\setminus n}$, though, we further need that $\text{supp } \hat{\theta} \setminus n \subseteq \hat{S}$ so that all LOOCV problems run over the same low-dimensional space as the full-data problem. It will be easier to state conditions for a stronger result, that $\text{supp } \hat{\theta} \setminus n = \hat{S} = S$. This will follow from an assumption on the smallest entry of $\theta^*_S$, which we stated as Assumption 4 in the main text. We stated Assumption 4 using the quantity $T_{\text{min}}$ to avoid stating Assumptions 11 and 12 in the main text. We can now state its full version.

Assumption 15 (full version of Assumption 4). For $L_{\text{min}}$ and $\gamma$ from Assumptions 11 and 12, $\min_{s \in S} |\theta^*_s| > (\sqrt{D_{\text{eff}}(\gamma + 4)}/L_{\text{min}})\lambda$.

Proposition 5. If Assumptions 10 to 15 hold, then $\forall n, \text{supp } \hat{\theta} \setminus n = \hat{S} = S$.

Proof. This is immediate from Theorem 5.1 of Li et al. [2015].

---

3 Readers familiar with the LSSC may see choosing the neighborhood of $\theta^*$ as $\mathbb{R}^D$ to be too restrictive. This choice is not necessary for our results; we state Assumption 10 this way only for simplicity. See Appendix G.2 for an explanation.
We now define the local structured smoothness condition (LSSC). The LSSC was introduced by Li et al. [2015] for the purpose of extending proof techniques for the support recovery of $\ell_1$ regularized linear regression to more general $\ell_p$ regularized $M$-estimators. Essentially, it provides a condition on the smoothness of the third derivatives of the objective $F(\theta)$ near the true sparse $\theta^*$. 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 derivative of $F$ evaluated along a direction $u \in \mathbb{R}^D$:

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

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

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

**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 $F$ satisfies the $(\theta^*, N_{\theta^*})$ LSSC with constant $K \geq 0$ if for any $u \in \mathbb{R}^D$:

$$|D^3 f(\theta^* + \delta)[u, u, e_j]| \leq K \|u\|_2^2,$$

(30)

where $e_j \in \mathbb{R}^D$ is the $j$th 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 different from the original definition given in Li et al. [2015], who prove the two to be equivalent in their Proposition 3.1. Li et al. [2015] go on to prove bounds on the LSSC constants for linear and logistic regression, which we state as Proposition 11 and Proposition 13 below.

Note that Assumption 10 in the main text states that the LSSC holds with $N_{\theta^*} = \mathbb{R}^D$. We state Assumption 10 in this form purely for conciseness; we will only consider checking Assumption 10 for linear and logistic regression, both of which satisfy the LSSC with $N_{\theta^*} = \mathbb{R}^D$. Going beyond these cases, it is easily possible to state a version of our results with $N_{\theta^*} \neq \mathbb{R}^D$; however, this will require an extra assumption along the lines of Condition 7 of Theorem 5.1 in Li et al. [2015], which is trivially satisfied when $N_{\theta^*} = \mathbb{R}^D$. In order to avoid stating an extra assumption that is trivially satisfied in the cases we consider, we chose to simply state the LSSC with $N_{\theta^*} = \mathbb{R}^D$.

**G.3 Assumptions 10 to 14 imply supp $\hat{\theta}_{\setminus n} \subseteq S$ for all n**

Theorem 5.1 of Li et al. [2015] gives conditions on $F$ under which supp $\hat{\theta} = S$. So, if these conditions hold for all $F \setminus n$, then we have supp $\hat{\theta}_{\setminus n} = S$ for all $n$. Their Theorem 5.1 actually has two extra assumptions beyond Assumptions 10 to 14. The first is their Assumption 7; however, this is immediately implied by the fact that we assume the LSSC holds with $N_{\theta^*} = \mathbb{R}^D$. The second is their analogue of our Assumption 15; however, they use this condition to imply that $\hat{\theta} = S$ after having shown that $\hat{\theta} \subseteq S$.

**G.4 Useful results for proving Theorems 2 and 3**

Before going on to Theorems 2 and 3, we will give a few useful results. We first define a sub-Exponential random variable:

**Definition** [Vershynin 2018]. A random variable $V$ is $c_x$-sub-Exponential if $E[\exp(V/c_x)] \leq 2$.

We will frequently use the fact that if $X$ is $c_x$-sub-Gaussian, then $X^2$ is $c_x^2$-sub-Exponential. Now we state a few existing results about the maxima of sub-Gaussian and sub-Exponential random variables that will be useful in our proofs.

**Lemma 1** (Lemma 5.2 from van Handel [2016]). Suppose that we have real valued random variables $Z_1, \ldots, Z_N$ that satisfy $\log E[\exp(\lambda Z_n)] \leq \psi(\lambda)$ for all $n = 1, \ldots, N$ and all $\lambda \geq 0$ for some convex function $\psi : \mathbb{R} \to \mathbb{R}$ 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 $\psi^{-1}$ is the inverse of the Legendre dual of $\psi$. 

19
Remembering the definition of a sub-Gaussian random variable from Definition 1, Lemma 1 can be used to show the following:

**Corollary 1.** Let \( Z_1, \ldots, Z_N \) be i.i.d. sub-Gaussian random variables with parameter \( c_{x} \). Then:

\[
\Pr \left[ \max_{n=1,\ldots,N} Z_n \geq E[Z_n] + \sqrt{2C_{x}^2 \log N} + u \right] \leq e^{-C_{x}^2 u^2 / 2} \tag{31}
\]

\[
\Pr \left[ \max_{n=1,\ldots,N} Z_n^2 \geq E[Z_n^2] + C_{x}^2 (\log N + 1 + u) \right] \leq e^{-u} \tag{32}
\]

**Proof.** For the first inequality, the definition of a sub-Gaussian random variable is that \( \log E e^{\lambda Z_n} \leq \lambda^2 c_{x}^2 / 2 =: \psi(\lambda) \), which has \( \psi^{*}(y) = y^2 / (2C_{x}^2) \) and \( \psi^{-1}(x) = \sqrt{2C_{x}^2 x} \). We use the upper bound:

\[
\psi^{-1}(\log N + u) = \sqrt{2C_{x}^2 (\log N + u)} \leq \sqrt{2C_{x}^2 \log N} + \sqrt{2C_{x}^2 u},
\]

Using this upper bound with Lemma 1 and changing variables \( u \rightarrow u^2 / (2C_{x}^2) \) gives the first inequality.

For the second inequality, use the fact that \( Z_n^2 \) is sub-Exponential with parameter \( c_{x}^2 \) so that it satisfies \( \log E e^{\lambda Z_n^2} \leq \psi(\lambda) \), where:

\[
\psi(\lambda) := \begin{cases} \lambda C^2_{x}, & 0 \leq t \leq 1/c_{x}^2 \\ \infty, & \text{otherwise.} \end{cases}
\]

For \( x \geq 0 \), this \( \psi^{-1} \) has inverse Legendre dual \( \psi^{-1}(x) = C_{x}^2 (x + 1) \). Plugging into Lemma 1 gives the result.

**Proposition 6.** Let \( x_1, \ldots, x_N \) be random vectors in \( \mathbb{R}^D \) with i.i.d. \( c_{x} \)-sub-Gaussian components and \( E[x^2_{n,l}] = 1 \). Then:

\[
\Pr \left[ \max_{n=1,\ldots,N} \|x_n\|_2 \geq \sqrt{D} + \sqrt{2C_{x}^4 \log N} + u \right] \leq e^{-C_{x}^2 u^2 / 2}, \tag{33}
\]

where \( C > 0 \) is some global constant, independent of \( c_{x}, D, \) and \( N \).

**Proof.** From Theorem 3.1.1 of Vershynin [2018], we have that \( \|x_n\|_2 - \sqrt{D} \) is sub-Gaussian with parameter \( Cc_{x}^2 \), where \( C \) is some constant. Using the first part of Corollary 1 gives the result.

**G.5 Proof of Theorem 2 (Linear Regression)**

Recall Assumptions 1 and 5, we assume a linear regression model \( y_n = x_n^T \theta^* + \varepsilon_n \), where \( x_n \in \mathbb{R}^D \) has i.i.d. \( c_{x} \)-sub-Gaussian components with \( E[x^2_{n,l}] = 1 \) and \( \varepsilon_n \) is \( c_{x} \)-sub-Gaussian. For notation throughout this section, we will let \( \lambda \) denote an absolute constant independent of any aspect of the problem \( (N, D, D_{\text{eff}}, c_{x}, \text{or } c_{z}) \) that will change from line to line (e.g. we may write \( 5C_{x}^2 = C \)). We will frequently use \( \lambda Z \) to denote the \( N \times D_{\text{eff}} \) matrix formed by taking the columns of \( X \) that are in \( S \), \( x_{n,s} \) to denote the coordinates of the \( n \)th vector of covariates \( x_n \) that are in the set \( S \), and \( X_{n,s} \) to denote the matrix \( X_{s,s} \) with the \( n \)th row removed. We will show the following theorem, stated more concisely as Theorem 2 in the main text:

**Theorem 4 (Restated version of Theorem 2 from main text).** Take Assumptions 7 to 13, 5, and 7. Suppose the regularization parameter \( \lambda \) satisfies:

\[
\lambda \geq \frac{1}{\alpha - M_{\text{lin}}} \sqrt{\frac{c_{x}^2 c_{z}^2 \log D}{NC} + \frac{25c_{x}^2 c_{z}^3}{NC} + \frac{4c_{x} c_{z} \log(ND) + 26}{N(\alpha - M_{\text{lin}})}}, \tag{34}
\]

where \( C \) is a constant in \( N, D, D_{\text{eff}}, c_{x} \) and \( c_{z} \), and \( M_{\text{lin}} \) is defined as:

\[
M_{\text{lin}} = \frac{C D_{\text{eff}} \left( \sqrt{50c_{x}^2} + \sqrt{2c_{x}^2 \log(N(D - D_{\text{eff}}))} \right) \left( \sqrt{D_{\text{eff}}} + \sqrt{50c_{x}^3} + \sqrt{2c_{x}^4 \log N} \right)}{N - 3c_{x}^2 \sqrt{N} (\sqrt{D_{\text{eff}}} + 5)} + \frac{C D_{\text{eff}} (D_{\text{eff}} + D_{\text{eff}} c_{x}^2 (\log N + 26)) \left( \sqrt{N} + \sqrt{50c_{x}^4} + \sqrt{2c_{x}^4 \log(D - D_{\text{eff}})} \right) \left( \sqrt{ND_{\text{eff}}} + \sqrt{50c_{x}^4} \right)}{(N - 3c_{x}^2 \sqrt{N} (\sqrt{D_{\text{eff}}} + 5))^2} \tag{35}
\]

Then for \( N \) sufficiently large, Condition 1 holds with probability at least \( 1 - 26e^{-25} \), where the probability is over the random data \( \{(x_n, y_n)\}_{n=1}^{N} \).
Proof. For a fixed regularization parameter $\lambda$ and random data $\{x_n, y_n\}_{n=1}^N$, we are interested in the probability that any of Assumptions [10] to [14] are violated, as Proposition 5 then proves the result. For convenience in writing the incoherence condition, define $J_{nd} \in \mathbb{R}^D$, for $d \in S^c$, as:

$$J_{nd} := \left(X_{\setminus n,S}^T X_{\setminus n,S} \right)^{-1} X_{\setminus n,S}^T X_{\setminus n,d}. \quad (36)$$

It is easiest to show that each of Assumptions [10] to [14] hold with high probability separately, rather than all together, so we apply a union bound to get:

\[
\Pr[\text{any assumption violated}] \leq \Pr[\min_n \lambda_{\text{min}}(X_{\setminus n,S}^T X_{\setminus n,S}) = 0] + \Pr[\max_n \max_{d \in S^c} \|J_{nd}\|_1 \geq 1] + \Pr[\max_n \|\nabla F_n\|_\infty \geq \frac{\lambda(1 - \max_n \max_{d \in S^c} \|J_{nd}\|_1)}{4}] + \Pr\left[\frac{\min_n \lambda_{\text{min}}^2(X_{\setminus n,S}^T X_{\setminus n,S})}{\left(1 - \max_n \max_{d \in S^c} \|J_{nd}\|_1\right) + 4}\frac{\left(1 - \max_n \max_{d \in S^c} \|J_{nd}\|_1\right)}{D_{\text{eff}}} \leq \lambda\right].
\]

We will bound each term by appealing to the Lemmas and Propositions proved below. Using Lemma [2] and Lemma [4], the first and third terms are bounded by $16^{-25}$. As noted in Proposition [11], we have $\Pr[K = 0] = 1$, so the final probability is equal to zero (as the event reduces to $\alpha < \lambda$). To bound the second probability, we have that Lemma [5] says that:

$$\Pr[\max_n \max_{d \in S^c} \|J_{nd}\|_1 \geq 1 - \alpha + M_{\text{lin}}] \leq 9e^{-25}.$$ 

As $\alpha > 0$, if $M_{\text{lin}} = o(1)$ as $N \to \infty$, we will have that $1 - \alpha + M_{\text{lin}} < 1$ for large enough $N$. This would imply the third probability is $\leq 9e^{-25}$ for $N$ large enough. Under our conditions on the growth of $D_{\text{eff}}$ and $D$, we can show that $M_{\text{lin}} = o(1)$. We have, hiding constants and lower order terms in $N, D,$ and $D_{\text{eff}}$:

$$M_{\text{lin}} = O\left(\frac{D_{\text{eff}} \sqrt{\log(N) + \log(D)}}{N - \sqrt{ND_{\text{eff}}}} + \frac{D_{\text{eff}}^{5/2} \log(N) \left(\sqrt{N} + \sqrt{\log(D)}\right) \sqrt{N}}{(N - \sqrt{ND_{\text{eff}}})^2}\right) + \frac{D_{\text{eff}} \left(\sqrt{\log(N) + \log(D)} + \sqrt{\log(N) \log(D)}\right)}{N - \sqrt{ND_{\text{eff}}}} + \frac{D_{\text{eff}}^{5/2} \log(N) \log(N)}{(N - \sqrt{ND_{\text{eff}}})^2}. \quad (37)$$

where the second statement follows from using $\sqrt{\log(N) + \log(D)} \leq \sqrt{\log(N)} + \sqrt{\log(D)}$ and $D = o(e^N)$. Now, given that $D_{\text{eff}} = o([N/ \log(N)]^{2/5})$, the second term in Eq. [37] is $o(1)$. The first term is also $o(1)$ by combining $D_{\text{eff}} = o([N/ \log(N)]^{2/5})$ with $D_{\text{eff}}^{5/2} \sqrt{\log(D)} = o(N)$. Thus, $M_{\text{lin}} = o(1)$, which completes the proof.

What remains is to prove Lemmas [2] to [4] and Proposition [11] needed to prove Theorem [4]. We do this in the following four subsections.

### G.6 Linear regression: minimum eigenvalue

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

#### Proposition 7. If $X_{\setminus n} \in \mathbb{R}^{N \times D_{\text{eff}}}$ is an $N \times D_{\text{eff}}$ matrix with independent $c_{\text{sub}}$-sub-Gaussian entries with unit second moments, then:

$$\Pr[\lambda_{\text{min}}(X_{\setminus n}^T X_{\setminus n}) \leq N - 2Cc_{\text{sub}}^2 \sqrt{N(\sqrt{D_{\text{eff}}} + 5)}] \leq 2e^{-25}, \quad (38)$$

where $C > 0$ is a global constant.
Proof. Theorem 4.6.1 of [Vershynin 2018] gives a concentration inequality for the minimum singular value, $s_{\text{min}}(X_{n,S})$, of $X_{n,S}$:

$$\Pr \left[ s_{\text{min}}(X_{n,S}) \leq \sqrt{N - Cc^2 \left(\sqrt{D_{\text{eff}}} + t\right)} \right] \leq 2e^{-t^2}. \tag{39}$$

Using the fact that the minimum eigenvalue of $X_{n,S}^T X_{n,S}$ is the square of the minimum singular value of $X_{n,S}$ and putting in $t = 5$:

$$\Pr \left[ \lambda_{\text{min}}(X_{n,S}^T X_{n,S}) \leq N - 2Cc^2 \sqrt{N(\sqrt{D_{\text{eff}}} + 5)} + C^2c_4^4(\sqrt{D_{\text{eff}}} + 5)^2 \right] \leq 2e^{-25}.$$

Dropping the $C^4c_4^4(\sqrt{D_{\text{eff}}} + 5)^2$ gives the result.

**Proposition 8.** If $X_{n,S}$ is the $N - 1 \times D_{\text{eff}}$ matrix formed by removing the $n$th row from $X_{n,S}$, we have:

$$\lambda_{\text{min}}(X_{n,S}^T X_{n,S}) \geq \lambda_{\text{min}}(X_{n,S}^T X_{n,S}) - \|x_nS\|_2^2,$$ \hspace{1cm} (40)

where $x_n$ is the $n$th row of $X_{n,S}$.

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

$$\lambda_{\text{min}}(X_{n,S}^T X_{n,S}) = \min_{z \in \mathbb{R}^{D_{\text{eff}}}} \|z\|_2 \geq \left[ z^T X_{n,S}^T X_{n,S} z - z^T x_nS x_nS^T z \right] \geq 0.$$

The above two propositions now allow us to prove the bound we want on $\lambda_{\text{min}}(X_{n,S}^T X_{n,S})$. In the following lemma, we will assume that $D_{\text{eff}} = o(N/\log(N))$. While we ultimately will have the more restrictive requirement that $D_{\text{eff}} = o(\sqrt{N/\log(N)})$ in Assumption 3, the current result can be stated with the less restrictive requirement of $o(N/\log(N))$.

**Lemma 2.** Suppose $X_{n,S}$ is a $N \times D_{\text{eff}}$ matrix with independent $c_x$-sub-Gaussian entries and $D_{\text{eff}} = o(N/\log(N))$ as function of $N$. Then we have for $N$ sufficiently large:

$$\Pr \left[ \min_{n=1,...,N} \lambda_{\text{min}}(X_{n,S}^T X_{n,S}) \leq N - 3Cc^2 \sqrt{N(\sqrt{D_{\text{eff}}} + 5)} \right] \leq 3e^{-25} \tag{41}$$

**Proof.** In what follows, and repeatedly throughout the rest of our proofs, we will make use of the following generic inequality for any events $A$ and $B$:

$$\Pr[A] = \Pr[A \mid B] \Pr[B] + \Pr[A \mid B^c] \Pr[B^c] \leq \Pr[A \mid B] \Pr[B] + \Pr[B^c], \tag{42}$$

Calling the probability on the left hand side of Eq. $41$ $P$, we can break $P$ down as, for some constant $L_{\text{min}}$:

$$P \leq \Pr \left[ \min_{n=1,...,N} \lambda_{\text{min}}(X_{n,S}^T X_{n,S}) \leq N - 3Cc^2 \sqrt{N(\sqrt{D_{\text{eff}}} + 5)} \mid \lambda_{\text{min}}(X_{n,S}^T X_{n,S}) \geq L_{\text{min}} \right] \Pr \left[ \lambda_{\text{min}}(X_{n,S}^T X_{n,S}) \geq L_{\text{min}} \right]$$

$$+ \Pr \left[ \lambda_{\text{min}}(X_{n,S}^T X_{n,S}) \leq L_{\text{min}} \right].$$

Picking $L_{\text{min}} = N - 2Cc^2 \sqrt{N(\sqrt{D_{\text{eff}}} + 5)}$, we have that the second probability at most $2e^{-25}$ by Proposition 8.

Now to control the $\max_n \|x_nS\|_2^2$, note that $\|x_nS\|_2^2$ is $D_{\text{eff}}c_2^2$-sub-Exponential, and choose $u = 25$ in the second statement of Corollary 1. This tells us that the first probability is at most $e^{-25}$ if $E[\|x_nS\|_2^2] + Cc_2^2(\log N + 26) = D_{\text{eff}} + Cc_2^2(\log N + 26)$ is less than $Cc_2^2 \sqrt{N(\sqrt{D_{\text{eff}}} + 5)}$, which, for $D_{\text{eff}}$ being $o(N/\log(N))$, is satisfied for $N$ large enough.
G.7 Linear regression: incoherence

The following proposition will be useful in proving Lemma 3 below:

Proposition 9. Let $z \in \mathbb{R}^N$ be any vector and $z_n \in \mathbb{R}^{N-1}$ the same vector with the $n$th coordinate removed. Also let $X_{:S} \in \mathbb{R}^{N \times D_{\text{eff}}}$ be some matrix with $X_{:n,S}$ the same matrix with the $n$th row removed. Define, for any vector $z \in \mathbb{R}^N$:

$$J_{nz} := \left( X_{:n,S}^T X_{:n,S} \right)^{-1} X_{:n,S}^T z_n,$$

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

$$\| J_{nz} - J_z \|_1 \leq D_{\text{eff}} \frac{|z_n| \| x_{nS} \|_2}{\lambda_{\text{min}}\left( X_{:n,S}^T X_{:n,S} \right)}$$

$$+ D_{\text{eff}} \frac{x_n^2}{\lambda_{\text{min}}\left( X_{:n,S}^T X_{:n,S} \right)} \| z \|_2 \| x_{n,S} \|_2,$$

where $\| x_{n,S} \|_2 := \sqrt{\sum_{n=1}^{N} \sum_{s \in S} x_{ns}^2}$.

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

$$J_z = (X_{:n,S}^T X_{:n,S})^{-1} X_{:n,S}^T z$$

$$= \left( (X_{:n,S}^T X_{:n,S})^{-1} - \frac{(X_{:n,S}^T X_{:n,S})^{-1} x_{nS} x_{nS}^T (X_{:n,S}^T X_{:n,S})^{-1}}{1 + x_{nS}^T (X_{:n,S}^T X_{:n,S})^{-1} x_{nS}} \right) X_{:n,S}^T z$$

$$= (J_{nz} + (X_{:n,S}^T X_{:n,S})^{-1} x_{nS} z_n) - \frac{X_{:n,S}^T z_n}{1 + x_{nS}^T (X_{:n,S}^T X_{:n,S})^{-1} x_{nS}} X_{:n,S}^T z$$

(46)

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

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

(47)

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

$$| e_i^T (J_{nz} - J_z) | \leq | e_i^T B^{-1} x_{nS} | z_n | + \frac{| e_i^T B^{-1} x_{nS} |}{1 + x_{nS} B^{-1} x_{nS}} \sum_{m=1}^{N} | z_m | \| x_{nS} B^{-1} x_{mS} \|_2$$

(48)

$$\leq | z_n | \lambda_{\text{max}}(B^{-1}) \| x_{nS} \|_2 + \frac{\lambda_{\text{max}}(B^{-1}) \| x_{nS} \|_2^2}{1 + \lambda_{\text{max}}(B^{-1}) \| x_{nS} \|_2^2} \sum_{m=1}^{N} | z_m | \| x_{mS} \|_2$$

(49)

The $\lambda_{\text{min}}(B^{-1}) \| x_{nS} \|_2^2$ is strictly positive, so we can drop it from the denominator for a further upper bound. Using the fact that, for the positive semidefinite matrix $B$ we have $\lambda_{\text{min}}(B^{-1}) = 1/\lambda_{\text{max}}(B)$ and $\lambda_{\text{max}}(B^{-1}) = 1/\lambda_{\text{min}}(B)$, we get:

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

(50)

Finally, use Cauchy-Schwarz to get $\sum_{m=1}^{N} | z_m | \| x_{mS} \|_2 \leq \| z \|_2 \| X_{:,S} \|_2$, where $\| X_{:,S} \|_2 := \left( \sum_{m=1}^{N} \sum_{s \in S} x_{ms}^2 \right)^{1/2}$. Notice that our upper bound is now independent of the index $i$; this means we have a bound on any coordinate $i$ of $\| (J_{nz} - J_z) \|_1$. So, multiplying this bound by $D_{\text{eff}}$ 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 Proposition 9.
Lemma 3. Take Assumptions 2 and 5 Then, for the scalar \( M_{\text{lin}} \) defined in Theorem 2 we have:

\[
\Pr \left[ \max_{n=1}^{N} \max_{d \in S^c} \| J_{nd} \|_1 \geq 1 - \alpha + M_{\text{lin}} \right] \leq 10e^{-25},
\]

where \( J_{nd} \) is defined in Eq. (36) above.

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 Proposition 9 and then apply a high probability upper bound. Following the same idea of conditioning and peeling off terms as in the proof of Lemma 2, we can condition on the following events, the complement of each of which has a small constant probability:

\[
\min_n \lambda_{\text{min}}(X_{\gamma n, S}^TX_{\gamma n, S}) \geq N - 3Cc_2^2 \sqrt{N} \left( \sqrt{D_{\text{eff}}} + 5 \right)
\]

\[
\| X_{n, S} \|_2 \leq \sqrt{ND_{\text{eff}}} + \sqrt{50Cc_2^2}
\]

\[
\max_n \| x_{n, S} \|_2 \leq \sqrt{D_{\text{eff}}} + \sqrt{50Cc_2^4} + 2Cc_2^4 \log N
\]

\[
\max_{d \in S^c} \| X_{n, d} \|_2 \leq \sqrt{N} + \sqrt{50Cc_2^4} + 2Cc_2^4 \log(N(D - D_{\text{eff}}))
\]

\[
\max_{d \in S^c} \| x_{n, d} \| \leq \sqrt{50Cc_2^4} + 2Cc_2^4 \log(N(D - D_{\text{eff}}))
\]

\[
\max_n \| x_{n, S} \|_2 \leq \sqrt{D_{\text{eff}}} + c_2^2 \log N + 26
\]

The probability of the complement of the first event is \( \leq 3e^{-25} \) by Lemma 2, the second is \( \leq e^{-25} \) by noting that \( \| X_{n, S} \|_2 - \sqrt{ND_{\text{eff}}} \) is a \( C_2^2 \)-sub-Gaussian random variable and applying a standard sub-Gaussian bound, the third is \( \leq e^{-25} \) by applying Proposition 5 the fourth is \( \leq e^{-25} \) by the same reasoning as the third, and the fifth is \( \leq 2e^{-25} \) by the first part of Corollary 1. Finally, the sixth is \( \leq e^{-25} \) by noting that \( \| x_{n, S} \|_2 \) is a \( C_2^2 \)-sub-Exponential random variable, to which we can apply Corollary 1. All in all, these probabilities sum up to \( 9e^{-25} \). Conditioned on all these events, we can upper bound the upper bound on \( \| J_{nd} - J_d \|_1 \) given by Proposition 9 to get:

\[
\| J_{nd} - J_d \|_1 \leq \frac{CD_{\text{eff}} \left( \sqrt{50Cc_2^2} + \sqrt{2Cc_2^4 \log(N(D - D_{\text{eff}}))} \right) \left( \sqrt{D_{\text{eff}}} + \sqrt{50Cc_2^4} + 2Cc_2^4 \log N \right) + \sqrt{ND_{\text{eff}}} + 50Cc_2^4}{(N - 3Cc_2^2 \sqrt{N} \left( \sqrt{D_{\text{eff}}} + 5 \right))^2}
\]

Call the entire quantity on the right-hand side of this inequality \( M_{\text{lin}} \), and call the union of the above six events the event \( F \). Then by conditioning on \( F \) and the event \( \{ \max_{d \in S^c} \| J_d \|_1 < 1 - \alpha \} \), we get:

\[
\Pr \left[ \max_{n=1}^{N} \max_{d \in S^c} \| J_{nd} \|_1 \geq 1 - \alpha + M_{\text{lin}} \right] \leq \Pr \left[ \max_{n=1}^{N} \max_{d \in S^c} \| J_{nd} - J_n \|_1 \geq M_{\text{lin}} | F \right] + \Pr [F^c] + \Pr \left[ \max_{d \in S^c} \| J_d \|_1 \geq 1 - \alpha \right]
\]

By the definition of \( M_{\text{lin}} \) above, we know that the first probability is zero, by the argument above and a union bound we know \( \Pr [F^c] \leq 9e^{-25} \), and the third is \( \leq e^{-25} \) by Assumption 2.

G.8 Linear regression: bounded gradient

We need to bound the probability

\[
\Pr \left[ \max_{n \in [N]} \left\| \nabla F\backslash_n (\theta^*) \right\|_\infty \geq \frac{\lambda(1 - \max_n \max_{d \in S^c} \| J_{nd} \|_1)}{4} \right] \leq \Pr \left[ \max_{n \in [N]} \left( \| \nabla F(\theta^*) \|_\infty + \frac{1}{N} \| \nabla f(x_n^T \theta^*, y_n) \|_\infty \right) \geq \frac{\lambda(1 - \max_n \max_{d \in S^c} \| J_{nd} \|_1)}{4} \right]
\]
Conditioning on the event that \( \|\nabla F(\theta^*)\|_\infty \leq B_G \) for some number \( B_G \) and the event that \( \max_n \max_{d \in S^c} \|J_{nd}\|_1 \leq 1 - \alpha + M_{\text{lin}} \), we get that this probability is less than or equal to:

\[
\Pr \left[ \max_{n=1,\ldots,N} \left\| \frac{1}{N} \nabla f(x_n^T \theta^*, y_n) \right\|_\infty \geq \frac{\lambda(\alpha - M_{\text{lin}})}{4} - B_G \right] + \Pr \left[ \|\nabla F(\theta^*)\|_\infty \geq B_G \right] + \Pr \left[ \max_n \max_{d \in S^c} \|J_{nd}\|_1 \geq 1 - \alpha + M_{\text{lin}} \right]
\]

(60)

The following proposition gives a reasonable value for \( B_G \):

**Proposition 10.** In the above setup for linear regression,

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

(61)

**Proof.** The \( d \)-th coordinate of the gradient is \( (\nabla F(\theta^*))_d = 1/N \sum_n \epsilon_n x_{nd} \). First, we have that \( 1/N \sum_n \epsilon_n x_{nd} \) is a \( c_x c_s \)-sub-Exponential random variable. By Bernstein’s inequality (see Theorem 2.8.1 from Vershynin 2018), we have:

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

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 lemma we need, which bounds the probability that any \( \|\nabla F^{\backslash n}(\theta^*)\|_\infty \) is large:

**Lemma 4.** For the above setup for linear regression and the \( \lambda \) given in Theorem 2 we have:

\[
\Pr \left[ \max_{n=1,\ldots,N} \left\| \nabla F^{\backslash n}(\theta^*) \right\|_\infty \geq \frac{\lambda(1 - \max_n \max_{d \in S^c} \|J_{nd}\|_1)}{4} \right] \leq 13e^{-25}
\]

(62)

**Proof.** We can first apply the bound worked out in Eq. (60). Picking \( B_G \) to be the value given in Proposition 10 the second probability is \( \leq e^{-25} \) by Proposition 10 and the third is \( \leq 10e^{-25} \) by Lemma 5. To analyze the first probability, note that we can write the event as:

\[
\Pr \left[ \frac{1}{N} \max_n \max_d |\epsilon_n x_{nd}| \geq \frac{\lambda(\alpha - M_{\text{lin}})}{4} - B_G \right].
\]

Looking at the form of \( \lambda \) given in Theorem 2 we get that this is equal to:

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

The event we’re considering is just the absolute value of the max of \( ND \) sub-Exponential variables with parameter \( c_x c_s \). Plugging into Corollary 11 gives that this probability is \( \leq 2e^{-25} \).

**G.9 Linear regression: \( \lambda \) small enough**

To check the bound in Assumption 14 we need to know the LSSC constant \( K \) for linear regression:

**Proposition 11 (Li et al. 2015).** For the linear regression setup in Theorem 2 the loss \( F(\theta) \) satisfies the \( (\theta^*, N_{\theta^*}) \) LSSC with constant \( K = 0 \) for any \( \theta^*, N_{\theta^*} \) and any data \( X, Y \).

**Proof.** This follows from the fact that \( F(\theta) = \frac{1}{2} \|X \theta - Y\|_2^2 \) has zero third derivatives, implying that \( D^3 F(\theta)[u, u, e_j] = 0 \) for any \( \theta, u \in \mathbb{R}^D \) and coordinate vector \( e_j \in \mathbb{R}^D \).
G.10 Proof of Theorem 3 (Logistic Regression)

Recall Assumptions 1 and 6 we assume a logistic regression model such that the responses \( y_n \in \{-1, 1\} \) 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:

\[
D_n^{(1)} := \frac{-y_n}{1 + e^{y_n x_n^T \theta^*}}, \quad D_n^{(2)} := \frac{e^{y_n x_n^T \theta^*}}{(1 + e^{y_n x_n^T \theta^*})^2},
\]

the derivatives of \( F \) are:

\[
\nabla_{\theta} F(\theta^*) = \frac{1}{N} \sum_{n=1}^{N} D_n^{(1)} x_n, \quad \nabla_{\theta}^2 F(\theta^*) = \frac{1}{N} \sum_{n=1}^{N} D_n^{(2)} x_n x_n^T.
\]

For comparison, things were easier for linear regression because \( D_n^{(2)} = 1 \) and \( D_n^{(1)} = \varepsilon_n \) for some sub-Gaussian noise \( \varepsilon_n \). Still, we will be able to extend basically all our proof techniques for linear regression by using the fact that \(|D_n^{(2)}|\) and \(|D_n^{(1)}|\) are both \( \leq 1 \), allowing us to drop them in many of our upper bounds. This will allow us to prove a very similar result to Theorem 2. Again, we will let \( C \) denote an absolute constant independent of any aspect of the problem (\( N, D, D_\text{eff}, c_x \)) that will change from line to line (e.g. we may write \( 5C^2 = C \)).

**Theorem 5.** Take Assumptions 1 to 3, 6, 7 and 15. Suppose the regularization parameter is set as:

\[
\lambda \geq \frac{C}{\alpha - M_{\text{log}}} \left( \sqrt{\frac{c_x^2}{2}} \frac{25 + \log D}{N} + \frac{\sqrt{2c_x^2} \log(ND) + 50c_x^2}{N} \right),
\]

where \( C \) is a constant in \( N, D, c_x \), and \( M_{\text{log}} \) is defined similarly to \( M_{\text{lin}} \) from Theorem 2 but with different denominators:

\[
M_{\text{log}} = \frac{CD_{\text{eff}} \left( \sqrt{50c_x^2} + \sqrt{2c_x^2} \log(N(D - D_{\text{eff}})) \right)}{L_{\min} - c_x^2 \sqrt{N} (\sqrt{D_{\text{eff}} + 5})} + \frac{CD_{\text{eff}} \left( D_{\text{eff}} + D_{\text{eff}} c_x^2 (\log N + 26) \right)}{(L_{\min} - c_x^2 \sqrt{N} (\sqrt{D_{\text{eff}} + 5}))^2} \left( \sqrt{N} + \frac{50c_x^2}{\sqrt{D_{\text{eff}} + 5}} + \frac{2c_x^2 \log(D - D_{\text{eff}})}{\sqrt{D_{\text{eff}} + 5}} \right).
\]

Then for \( N \) sufficiently large, Condition 7 holds with probability at least \( 1 - 43e^{-25} \), where the probability is over the random data \( \{(x_n, y_n)\}_{n=1}^{N} \).

**Proof.** The proof is exactly the same as that of Theorem 2 – we bound the probability that any of Assumptions 10 to 14 are violated by a union bound – except that we use Lemmas 5 to 7 and Proposition 13 below to bound each term. Note that we have \( M_{\text{log}} = o(1) \) by Assumptions 3 and 7. \( \square \)

G.11 Logistic regression: lambda min

**Lemma 5.** Take Assumption 7. Further suppose that \( D_{\text{eff}} \) grows as \( o(N/\log(N)) \). Then for \( N \) sufficiently large:

\[
\Pr \left[ \min_{n=1,\ldots,N} \lambda_{\min}(\nabla_{\theta}^2 F^{\land n}(\theta^*))_{SS} \leq L_{\min} - Cc_x^2 \sqrt{N} (\sqrt{D_{\text{eff}} + 5}) \right] \leq 3e^{-25},
\]

where \( L_{\min} \) is the constant from Assumption 7.

**Proof.** We have by Proposition 8 and the fact that \(|D_n^{(2)}| \leq 1\):

\[
\lambda_{\min}(\nabla_{\theta}^2 F^{\land n}(\theta^*))_{SS} \geq \lambda_{\min}(\nabla_{\theta}^2 F(\theta^*))_{SS} - ||x_n s||_2^2 |D_n^{(2)}| \geq \lambda_{\min}(\nabla_{\theta}^2 F(\theta^*))_{SS} - ||x_n s||_2^2.
\]

The rest of the proof is now exactly the same as that of Lemma 2. \( \square \)
G.12 Logistic regression: incoherence

We can get exactly the same bound as in Lemma 3. To do so, we first note that Proposition 9 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 Proposition 9 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 3.

Lemma 6. Take Assumptions 1, 2, 6 and 7. Then for the scalar $M_{\text{logr}}$ defined in Theorem 3 we have:

$$\Pr \left[ \max_{n=1, \ldots, N} \max_{d \in S^c} \| J_{nd} \|_1 \geq 1 - \alpha + M_{\text{logr}} \right] \leq 10 e^{-25},$$

where $J_{nd}$ is defined in Eq. (36).

Proof. The proof is very similar to that of Lemma 3. To prove Lemma 3 we wrote $\| J_{nd} \|_1 \leq \| J_d \|_1 + \| J_{nd} - J_d \|_1$. To bound $\| J_{nd} \|_1$ with high probability, we applied Assumption 2. To bound $\| J_{nd} - J_d \|_1$, we used the bound from Proposition 9 and then conditioned on a number of high-probability events to give an overall bound. We can condition on all of the same events, except we replace the event in Eq. (52) by:

$$\left\{ \min_{n=1, \ldots, N} \lambda_{\min} \left( \nabla^2_{\theta} F_{SS}^{(n)} \right) \geq L_{\min} - C c_2^2 \sqrt{N (\sqrt{D_{\text{eff}}} + 5)} \right\}.$$ (69)

By Lemma 3 the complement of this event has probability at most $3 e^{-25}$. We condition on the rest of the events in the proof of Lemma 3 and finish the proof along the same lines.

G.13 Logistic regression: bounded gradient

Again, we are interested in bounding:

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

The same reasoning that led to Eq. (60) gives us the same bound:

$$\leq \Pr \left[ \max_{n=1, \ldots, N} \left\| \frac{\nabla f(x_n^T \theta^*, y_n)}{\sqrt{N}} \right\|_{\infty} \geq \frac{\lambda (\alpha - M_{\text{logr}})}{4} - B_G \right] + \Pr \left[ \| \nabla F(\theta^*) \|_{\infty} \geq B_G \right] + \Pr \left[ \max_{n=1, \ldots, N} \max_{d \in S^c} \| J_{nd} \|_1 \geq 1 - \alpha + M_{\text{logr}} \right]$$

(70)

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

Proposition 12. For the logistic regression setup above, we have:

$$\Pr \left[ \| \nabla F(\theta^*) \|_{\infty} \geq c_\theta \sqrt{\frac{25 + \log D}{CN}} \right] \leq 2 e^{-25}.$$ (71)

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

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

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

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

where the final inequality comes from noting that $|x_{nd}|$ is also $c_\theta$-sub-Gaussian and using Hoeffding’s inequality (Theorem 2.6.2 from Vershynin [2018]). Union bounding over all $D$ dimensions of $\nabla F(\theta^*)$ gives the result. $\square$
Lemma 7. For the above setup for logistic regression and the λ given in Theorem 3 we have:

\[
\Pr \left[ \max_{n \in [N]} \left\| \nabla F^\dagger_n(\theta^*) \right\|_\infty \geq \frac{\lambda(1 - \max_n \max_{d \in S^c} \|J_{nd}\|_1)}{4} \right] \leq 14e^{-25} \tag{72}
\]

Proof. Just as in the proof of Lemma 4, we will apply the upper bound in Eq. 70 and then bound each term. The second probability in Eq. (70) is \(\leq 10e^{-25}\) by Lemma 6. The second term is \(\leq 2e^{-25}\) by Proposition 12. We now just need to analyze the first term:

\[
\Pr \left[ \max_{n \in [N]} \left\| \frac{1}{N} \nabla f(x_n^T \theta^*, y_n) \right\|_\infty \geq \frac{\lambda(\alpha - M_{\logr})}{4} - B_G \right].
\]

Plugging in the λ given in Theorem 5 and using \(\|\nabla f(x_n^T \theta^*, y_n)\|_\infty \leq \|x_n\|_\infty\), we can further upper bound this probability:

\[
\leq \Pr \left[ \max_{n \in [N]} \|x_n\|_\infty \geq \left(2\sqrt{2c_if_2} \log N + \sqrt{50c_i^2}\right) \right].
\]

By part 1 of Corollary 1, this probability is \(\leq 2e^{-25}\).

G.14 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 the bound in Assumption 14; this is not the case for logistic regression. Li et al. [2015] prove that the LSSC holds here:

 Proposition 13 ([Li et al., 2015]). The logistic regression model given above satisfies the \((\theta^*, N_{\theta^*})\) LSSC for any \(\theta^*\) and \(N_{\theta^*} \in \mathbb{R}^D\) with a data-dependent constant \(K = 1/4(\max_n \|x_n\|_\infty)(\max_n \|x_nS\|_2^2)\).

Proof. This is proved in Section 6.2 of [Li et al., 2015].

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

 Proposition 14. 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:

\[
\Pr \left[ K \geq 4 \left( \sqrt{2c_i^2 \log(ND)} + \sqrt{50c_i^2} \right)(D_{\text{eff}} + c_i^2 D_{\text{eff}}(\log N + 26)) \right] \leq 3e^{-25} \tag{73}
\]

Proof. First, Corollary 1 implies that \(\max_n \|x_n\|_\infty \geq \sqrt{2c_i^2 \log(ND)} + \sqrt{50c_i^2}\) with probability at most \(2e^{-25}\), so the probability we are interested in is bounded by:

\[
\leq \Pr \left[ \max_n \|x_nS\|_2^2 \geq D_{\text{eff}} + c_i^2 D_{\text{eff}}(\log N + 26) \right] + 2e^{-25}. \tag{74}
\]

Noting that \(\|x_nS\|_2^2\) is the sum of \(D_{\text{eff}}c_i^2\)-sub-Exponential random variables, \(\|x_nS\|_2^2\) is a \(D_{\text{eff}}c_i^2\)-sub-Exponential random variable. Corollary 1 then gives us that Eq. 74 is bounded above by \(3e^{-25}\).

We can now prove the result we need, which is that λ satisfies the upper bound in Assumption 14 with high probability.

Lemma 8. Take Assumptions 1 to 3 and 7. Then, for the logistic regression setup in Assumption 6 and λ as given in Theorem 3 and large enough N, we have:

\[
\Pr \left[ \lambda \geq \frac{\min_n \lambda_{min}^2(\nabla F^\dagger_n(\theta^*)_{SS})}{4 \left( (1 - \max_n \max_{d \in S^c} \|J_{nd}\|_1) + 4 \right)^2} \frac{1 - \max_d \|J_{nd}\|_1}{K} \right] \leq 16e^{-25} \tag{75}
\]

Proof. Using Lemma 5, Lemma 6, and Proposition 12, the desired probability is \(\leq 16e^{-25}\) if the following deterministic inequality holds:

\[
\lambda \leq \frac{4(\alpha - M_{\logr})}{4(\alpha - M_{\logr} + 4)^2} \left( \sqrt{2c_i^2 \log(ND)} + \sqrt{50c_i^2} \right)(D_{\text{eff}} + c_i^2 D_{\text{eff}}(\log N + 26)) \tag{76}
\]
We will lower bound the right hand side and show that $\lambda$ is less than this lower bound. Throughout, $C$ will be a generic constant that changes from line-to-line. First, as noted in the proof of Theorem 3, $M_{\log r} = o(1)$ as $N \to \infty$, so that for large enough $N$, we have $(\alpha - M_{\log r})/(\alpha - M_{\log r} + 4)^2 \geq (\alpha/2)/(\alpha/2 + 4)^2$. Next, for large enough $N$, Lemma 5 implies the denominator is greater than $CN$. Also for large enough $N$, the denominator is less than $CD_{\text{eff}} \log N \sqrt{\log(ND)}$. We are left with checking the condition:

$$\lambda \leq C \frac{\alpha/2}{(\alpha/2 + 4)^2} \frac{N^2}{D_{\text{eff}} \log N \sqrt{\log(ND)}}.$$  \hspace{1cm} (77)

Under Assumption 3, we can upper bound the denominator to get a further lower bound on the right hand side:

$$\lambda \leq C \frac{\alpha/2}{(\alpha/2 + 4)^2} \frac{\log^{2/5}(N)N^2}{N^{2/5} \log N \sqrt{\log(N)} + N}.$$  \hspace{1cm} (78)

Now, the right hand side goes to infinity as $N$ gets large, while the $\lambda$ given in Theorem 3 goes to 0 as $N$ gets large. Thus, for sufficiently large $N$, Eq. (76) holds. \hfill \square