Approximate Cross-Validation in High Dimensions with Guarantees

William T. Stephenson
MIT CSAIL

Tamara Broderick
MIT CSAIL

Abstract

Leave-one-out cross-validation (LOOCV) can be particularly accurate among cross-validation (CV) variants for machine learning assessment tasks – e.g., assessing methods' error or variability. But it is expensive to re-fit a model N times for a dataset of size N. Previous work has shown that approximations to LOOCV can be both fast and accurate – when the unknown parameter is of small, fixed dimension. But these approximations incur a running time roughly cubic in dimension – and we show that, besides computational issues, their accuracy dramatically deteriorates in high dimensions. Authors have suggested many potential and seemingly intuitive solutions, but these methods have not yet been systematically evaluated or compared. We find that all but one perform so poorly as to be unusable for approximating LOOCV. Crucially, though, we are able to show, both empirically and theoretically, that one approximation can perform well in high dimensions – in cases where the high-dimensional parameter exhibits sparsity. Under interpretable assumptions, our theory demonstrates that the problem can be reduced to working within an empirically recovered (small) support. This procedure is straightforward to implement, and we prove that its running time and error depend on the (small) support size even when the full parameter dimension is large.

1 Introduction

Assessing the performance of machine learning methods is an important task in medicine, genomics, and other applied fields. Experts in these areas are interested in understanding methods' error or variability and, for these purposes, often turn to cross validation (CV); see, e.g., Saeb et al. [2017], Powers et al. [2019], Carrera et al. [2009], Joshi et al. [2009], Chandrasekaran et al. [2011], Biswal et al. [2001], Roff and Preziosi [1994]. Even after decades of use [Stone, 1974, Geisser, 1975], CV remains relevant in modern high-dimensional and complex problems. In these cases, CV provides, for example, better out-of-sample error estimates than simple test error or training error [Stone, 1974]. Moreover, among variants of CV, leave-one-out CV (LOOCV) offers to most closely capture performance on the dataset size of interest. For instance, LOOCV is particularly accurate for out-of-sample error estimation [Arlot and Celisse, 2010, Sec. 5].[1]

Modern datasets, though, pose computational challenges for CV. For instance, CV requires running a machine learning algorithm many times, especially in the case of LOOCV. This expense has led to recent proposals to *approximate* LOOCV [Obuchi and Kabashima, 2016, 2018, Beirami et al., 2017, Rad and Maleki, 2020, Wang et al., 2018, Giordano et al., 2019b, Xu 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, the exact case of modern interest. First, 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. Second, even if the approximations were accurate in high dimensions, these approximations require solving a D-dimensional linear system, which incurs an $O(D^3)$ cost.

Previous authors have proposed a number of potential solutions for one or both of these problems, but these

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

[1] In the case of linear regression, LOOCV provides the least biased and lowest variance estimate of out-of-sample error among other CV methods [Burman, 1989].

methods have not yet been carefully evaluated and compared. (#1) Koh and Liang [2017] use a randomized solver [Agarwal et al., 2017] successfully for qualitative analyses similar to high-dimensional approximate CV, so it is natural to think the same technique might speed up approximate CV in high dimensions. Another option is to consider that the unknown parameter may effectively exist in some subspace with much lower dimension that D. For instance, ℓ_1 regularization offers an effective and popular means to recover a sparse parameter support.[2] Since existing approximate CV methods require twice differentiability of the regularizer, they cannot be applied directly with an ℓ_1 penalty. (#2) Thus, a second proposal – due to Rad and Maleki [2020], Wang et al. [2018] – is to apply existing approximate CV methods to a smoothed version of the ℓ_1 regularizer. (#3) A third proposal – made by, e.g., Burman [1989] – is to ignore modern approximate CV methods, and speed up CV by uniform random subsampling of LOOCV folds.

We show that all three of these methods fail to address the issues of approximate CV in high dimensions. (#4) A fourth proposal – due to Rad and Maleki [2020], Wang et al. [2018], Obuchi and Kabashima [2016, 2018], Beirami et al. [2017] – is to again consider ℓ_1 regularization for sparsity. But in this case, the plan is to fit the model once with the full dataset to find a sparse parameter subspace and then apply existing approximate CV methods to only this small subspace.

In what follows, we demonstrate with both empirics and theory that proposal #4 is the only method that is fast and accurate for assessing out-of-sample error. We emphasize, moreover, its simplicity and ease of implementation. On the theory side, we show in Section 4 that proposal #4 will work if exact LOOCV rounds recover a shared support. Our major theoretical contribution is to prove that, under mild and interpretable conditions, the recovered support is in fact shared across rounds of LOOCV with very high probability (Sections 4.1 and 4.2). Obuchi and Kabashima [2016] have considered a similar setup and shown that the effect of the change in support is asymptotically negligible for ℓ_1-regularized linear regression; however, they do not show the support is actually shared. Additionally, Beirami et al. [2017], Obuchi and Kabashima [2018] make the same approximation in the context of other GLMs but without theoretical justification. We justify such approximations by proving that the support is shared with high probability in the practical *finite-data* setting – even for the very high-dimensional case $D = o(e^N)$ – for both linear and logistic regression (Theorems 2 and 3). 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 proposal #4 will depend on the support size – typically much smaller than the full dimension – rather than D. Our experiments in Section 5 on real and simulated data confirm these theoretical results.

[2]Note that sparsity, induced by ℓ_1 regularization, is typically paired with a focus on generalized linear models (GLMs) since these models simplify when many parameters are set to zero, are tractable to analyze with theory, and typically form the building blocks for even more complex models.

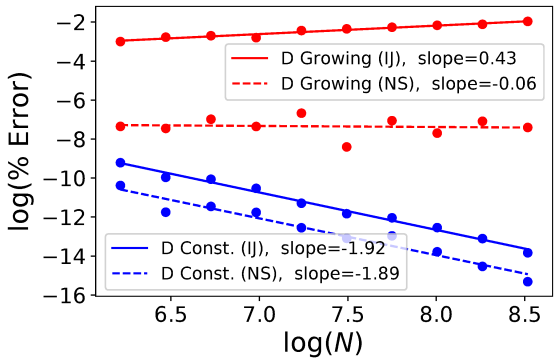

Figure 1: Log percent error (Eq. (10)) of existing approximate LOOCV methods ("IJ" and "NS") as a function of dataset size N for ℓ_2 regularized logistic regression. Dashed lines show Eq. (2) ("NS") and solid show Eq. (3) ("IJ"). Blue lines have fixed data/parameter dimension D, while red lines have $D = N/10$, although the true parameter has a fixed support size of $D_{\text{eff}} = 2$ (see Section 2 for a full description). IJ and NS fail to capture this low "effective dimension" and suffer from substantially worse performance in high dimensions.

Model assessment vs. selection. Stone [1974], Geisser [1975] distinguish at least two uses of CV: model assessment and model selection. Model assessment refers to estimating the performance of a single, fixed model. Model selection refers to choosing among a collection of competing models. We focus almost entirely on model assessment – for two principal reasons. First, as discussed above, CV is widely used for model assessment in critical applied areas – such as medicine and genetics. Before we can safely apply approximate CV for model assessment in these areas, we need to empirically and theoretically verify our methods. Second, historically, rigorous analysis of the properties of model selection even for *exact* CV has required significant additional work beyond analyzing CV for model assessment. In fact, exact CV for model selection has only recently begun to be theoretically understood for ℓ_1 regularized linear regression [Homrighausen and McDonald, 2013, 2014, Chetverikov et al., 2020]. Our

experiments in Appendix H confirm that approximate CV for model selection exhibits complex behavior. We thus expect significant further work, outside the scope of the present paper, to be necessary to develop a theoretical understanding of approximate CV for model selection. Indeed, to the best of our knowledge, all existing theory for the accuracy of approximate CV applies only to model assessment [Beirami et al., 2017, Rad and Maleki, 2020, Giordano et al., 2019b, Xu et al., 2019, Koh et al., 2019].

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

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

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 nth data point, R the regularizer, and λ the amount of regularization. Consider a dataset where the nth 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 solutions of the same problem with the nth data point removed.[3] To that end,[4] define $\hat{\theta}_{\backslash n} := \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{m:\, m \neq n} f_m(\theta) + \lambda R(\theta)$. Computing $\hat{\theta}_{\backslash n}$ exactly across n usually requires N runs of an optimization procedure – a prohibitive cost. Various approximations, detailed next, address this cost by solving Eq. (1) only once.

Two approximations. Assume that f and R are twice differentiable functions of θ. Let $F(\theta) := (1/N) \sum_n f(x_n^T\theta, y_n)$ be the unregularized objective, and let $H(\theta) := \nabla_\theta^2 F(\theta) + \lambda \nabla_\theta^2 R(\theta)$ 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 [2020], Wang et al. [2018], Koh et al. [2019] approximate $\hat{\theta}_{\backslash n}$ by taking a Newton step ("NS") on the objective $(1/N) \sum_{m:\, m \neq n} f_m + \lambda R$ starting from $\hat{\theta}$; see Appendix D.4 for details. We thus call this approximation $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ for regularizer R:

$$\widetilde{\mathrm{NS}}_{\backslash n}(R) := \hat{\theta} + \frac{1}{N} \left(H(\hat{\theta}) - \frac{1}{N} \nabla_\theta^2 f_n(\hat{\theta}) \right)^{-1} \nabla_\theta f_n(\hat{\theta}). \tag{2}$$

In the case of GLMs, Theorem 8 of Rad and Maleki [2020] gives conditions on x_n and $f(\cdot, \cdot)$ that imply, for fixed D, the error of $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ averaged over n is $o(1/N)$ as $N \to \infty$.

Koh and Liang [2017], Beirami et al. [2017], Giordano et al. [2019b], Koh et al. [2019] consider a second approximation. As their approximation is inspired by the *infinitesimal jackknife* ("IJ") [Jaeckel, 1972, Efron, 1982], we denote it by $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$; see Appendix D.1.

$$\widetilde{\mathrm{IJ}}_{\backslash n}(R) := \hat{\theta} + \frac{1}{N} H(\hat{\theta})^{-1} \nabla_\theta f_n(\hat{\theta}). \tag{3}$$

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

In what follows, we consider both $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ and $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$, as they have complimentary strengths. Empirically, we find that $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ performs better in our LOOCV GLM experiments. But $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ is computationally efficient beyond LOOCV and GLMs. E.g., for general models, computation of $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ requires inversion of a new Hessian for each n, whereas $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ needs only the inversion of $H(\hat{\theta})$ for all n. In terms of theory, $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ has a tighter error bound of $o(1/N)$ for GLMs. But the theory behind $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ applies more generally, and, given a good bound on the gradients, may provide a tighter rate.

3 Problems in high dimensions

In the above discussion, we noted that there exists encouraging theory governing the behavior of $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ and $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ when D is fixed and N grows large. We now describe issues with $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ and $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ when D is large relative to N. The first challenge for both approximations given large 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

[3] See Appendix A for a brief review of CV methods.

[4] 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, this issue does not seem to be settled in the literature; see Appendix B.

after this single run. Given $\hat{\theta}$, both approximations require the inversion of a $D \times D$ matrix. Calculation of $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ 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 $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ has runtime of $O(ND^3)$ due to needing an inversion for each n. In the case of GLMs, though, $\nabla^2_\theta f_n$ 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 $H(\hat{\theta})$ and $H(\hat{\theta}) - (1/N)\nabla^2_\theta f(x_n^T\theta, y_n)$ that was assumed in defining $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ and $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$. We note that, if $\nabla^2 R(\hat{\theta})$ is only positive semidefinite, then invertibility of both matrices may be impossible when $D \geq N$; see Appendix D.2 for more discussion.

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 approximations struggle both computationally and statistically. Thus, we wish to understand whether there are high-D cases where approximate CV is useful. In what follows, we consider a number of options for tackling one or more of these issues and show that only one method is effective in high dimensions.

Proposal #1: Use randomized solvers to reduce computation. Previously, Koh and Liang [2017] have utilized $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ for qualitative purposes, in which they are interested in its sign and relative magnitude across different n. They tackle the $O(D^3)$ scaling of $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ by using the randomized solver from Agarwal et al. [2017]. While one might hope to replicate the success of Koh and Liang [2017] in the context of approximate CV, we show in Appendix C that this randomized solver performs poorly for approximating CV: while it can be faster than exactly solving the needed linear systems, it provides an approximation to exact CV that can be an order of magnitude less accurate.

3.1 Sparsity via ℓ_1 regularization.

Intuitively, if the exact $\hat{\theta}_{\setminus n}$'s have some low "effective dimension" $D_{\text{eff}} \ll D$, we might expect approximate CV's accuracy to depend only on D_{eff}. One way to achieve low D_{eff} is sparsity: i.e., we have $\hat{D}_{\text{eff}} := |\operatorname{supp}\hat{\theta}| \ll D$, where $\hat{S} := \operatorname{supp}\hat{\theta}$ collects the indices of the non-zero entries of $\hat{\theta}$. A way to achieve sparsity is choosing $R(\theta) = \|\theta\|_1$. However, note that $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ and $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ cannot be applied directly in this case as $\|\theta\|_1$ is not twice-differentiable. **Proposal #2**: Rad and Maleki [2020], Wang et al. [2018] propose the use of a smoothed approximation to $\|\cdot\|_1$; however, as we show in Section 5, this approach is often multiple orders of magnitude more inaccurate and slower than Proposal #4 below.

Proposal #3: Subsample exact CV. Another option is to bypass all the problems of approximate CV in high-D by uniformly subsampling a small collection of LOOCV folds. This provides an unbiased estimate of exact CV and can be used with exact ℓ_1 regularization. However, our experiments (Section 5) show that, under a time budget, the results of this method are so variable that their error is often multiple orders of magnitude higher than Proposal #4 below.

Proposal #4: Use the sparsity from $\hat{\theta}$. Instead, in what follows, we take the intuitive approach of approximating CV only on the dimensions in $\operatorname{supp}\hat{\theta}$. Unlike all previously discussed options, we show that this approximation is fast and accurate in high dimensions in both theory and practice. For notation, let $X \in \mathbb{R}^{N\times D}$ be the covariate matrix, with rows x_n. For $S \subset [D]$, let $X_{\cdot,S}$ be the submatrix of X with column indices in S; define x_{nS} and θ_S similarly. Let $\hat{D}_n^{(2)} := \left[d^2 f(z, y_n)/dz^2\right]_{z=x_n^T\hat{\theta}}$, and define the restricted Hessian evaluated at $\hat{\theta}$: $H_{\hat{S}\hat{S}} := X_{\cdot,\hat{S}}^T \operatorname{diag}\{\hat{D}_n^{(2)}\} X_{\cdot,\hat{S}}$. Further define the LOO restricted Hessian, $H^{\setminus n}_{\hat{S}\hat{S}} := H_{\hat{S}\hat{S}} - [\nabla^2_\theta f(x_n^T\hat{\theta}, y_n)]_{\hat{S}\hat{S}}$. Finally, without loss of generality, assume $\hat{S} = \{1, 2, \ldots, \hat{D}_{\text{eff}}\}$. We now define versions of $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ and $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ restricted to the entries in $\operatorname{supp}\hat{\theta}$:

$$\mathrm{NS}_{\setminus n} := \begin{pmatrix} \hat{\theta}_{\hat{S}} + (H^{\setminus n}_{\hat{S}\hat{S}})^{-1}\left[\nabla_\theta f(x_n^T\hat{\theta}, y_n)\right]_{\hat{S}} \\ 0 \end{pmatrix} \tag{4}$$

$$\mathrm{IJ}_{\setminus n} := \begin{pmatrix} \hat{\theta}_{\hat{S}} + H^{-1}_{\hat{S}\hat{S}}\left[\nabla_\theta f(x_n^T\hat{\theta}, y_n)\right]_{\hat{S}} \\ 0 \end{pmatrix}. \tag{5}$$

Other authors have previously considered $\mathrm{NS}_{\setminus n}$. Rad and Maleki [2020], Wang et al. [2018] derive $\mathrm{NS}_{\setminus n}$ by considering a smooth approximation to ℓ_1 and then taking the limit of $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ as the amount of smoothness goes to zero. In Appendix E, we show a similar argument can yield $\mathrm{IJ}_{\setminus n}$. Also, Obuchi and Kabashima [2016, 2018], Beirami et al. [2017] directly propose $\mathrm{NS}_{\setminus n}$ without using $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ as a starting point. We now show how $\mathrm{NS}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$ avoid the

three major high-dimensional challenges with $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ and $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ we discussed above.

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

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

The second high-dimensional challenge was that H and $H^{\setminus n}$ may not be invertible when $D \geq N$. Notice the relevant matrices in $\mathrm{NS}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$ are of dimension $\hat{D}_{\text{eff}} = |\hat{S}|$. So we need only make the much less restrictive assumption that $\hat{D}_{\text{eff}} < N$, rather than $D < N$. We address the third and final challenge of accuracy in the next section.

4 Approximation quality in high dimensions

Recall that the accuracy of $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ and $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ in general has a poor dependence on dimension D. We now show that the accuracy of $\mathrm{NS}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$ depends on (the hopefully small) $\hat{D}_{\text{eff}}$ rather than D. We start by assuming a "true" population parameter[5] $\theta^* \in \mathbb{R}^D$ 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 θ^* is sparse with $S := \operatorname{supp}\theta^*$ 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:

$$\hat{\phi} := \underset{\phi \in \mathbb{R}^{D_{\text{eff}}}}{\arg\min} \frac{1}{N} \sum_{n=1}^{N} f(x_{nS}^T \phi, y_n) + \lambda \|\phi\|_1 . \quad (6)$$

We define $\hat{\phi}_{\setminus n}$ as the leave-one-out variant of $\hat{\phi}$ (as $\hat{\theta}_{\setminus n}$ is to $\hat{\theta}$). Let $\mathrm{RNS}_{\setminus n}$ and $\mathrm{RIJ}_{\setminus n}$ be the result of applying the approximation in $\mathrm{NS}_{\setminus n}$ or $\mathrm{IJ}_{\setminus n}$ to the restricted problem in Eq. (6); note that $\mathrm{RNS}_{\setminus n}$ and $\mathrm{RIJ}_{\setminus n}$ have accuracy that scales with the (small) dimension D_{eff}.

Our analysis of the accuracy of $\mathrm{NS}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$ will depend on the idea that if, for all n, $\mathrm{NS}_{\setminus n}$, $\mathrm{IJ}_{\setminus n}$, and $\hat{\theta}_{\setminus n}$ run over the same D_{eff}-dimensional subspace, then the accuracy of $\mathrm{NS}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$ must be identical to that of $\mathrm{RNS}_{\setminus n}$ and $\mathrm{RIJ}_{\setminus n}$. In the case of ℓ_1 regularization, this idea specializes to the following condition, under which our main result in Theorem 1 will be immediate.

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

Theorem 1. *Assume Condition 1 holds. Then for all n, $\hat{\theta}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$ are (1) zero outside the dimensions S and (2) equal to their restricted counterparts from Eq. (6):*

$$\hat{\theta}_{\setminus n} = \begin{pmatrix} \hat{\theta}_{\setminus n,S} \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{\phi}_{\setminus n} \\ 0 \end{pmatrix},$$
$$\mathrm{IJ}_{\setminus n} = \begin{pmatrix} \mathrm{IJ}_{\setminus n,S} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathrm{RIJ}_{\setminus n} \\ 0 \end{pmatrix}. \quad (7)$$

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

Taking Condition 1 as a given, Theorem 1 tells us that for ℓ_1 regularized problems, $\mathrm{IJ}_{\setminus n}$ and $\mathrm{NS}_{\setminus n}$ inherit the fixed-dimensional accuracy of $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ and $\widetilde{\mathrm{NS}}_{\setminus n}(R)$ shown empirically in Fig. 1 and described theoretically in the references from Section 1. Taking a step further, one could show that $\mathrm{IJ}_{\setminus n}$ and $\mathrm{NS}_{\setminus n}$ are accurate for model assessment tasks by using results on the accuracy of exact CV for assessment (e.g., [Abou-Moustafa and Szepesvári, 2018, Steinberger and Leeb, 2018, Barber et al., 2019]).

Again, Theorem 1 is immediate if one is willing to assume Condition 1, but when does Condition 1 hold? There exist assumptions in the ℓ_1 literature under which $\operatorname{supp}\hat{\theta} = S$ [Lee et al., 2014, Li et al., 2015]. If one took these assumptions to hold for all $F^{\setminus n} := (1/N)\sum_{m:\, m\neq n} f_m$, then Condition 1 would directly follow. However, it is not immediate that any models of interest meet such assumptions. Rather than taking such uninterpretable assumptions or just taking Condition 1 as an assumption directly, we will give a set of more interpretable assumptions under which Condition 1 holds.

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

[5] This assumption may not be necessary to prove the dependence of $\mathrm{NS}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$ on $\hat{D}_{\text{eff}}$, but it allows us to invoke existing ℓ_1 support results in our proofs.

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

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] show that there exist worst-case training datasets for which sparsity-inducing methods like ℓ_1 regularization are not stable as each datapoint is left out.

Our second principal assumption is an *incoherence* condition.

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

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

Authors in the ℓ_1 literature often assume that incoherence holds deterministically for a given design matrix X – starting from the introduction of incoherence by 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. [2015]). We next place some restrictions on how quickly D and D_{eff} grow as functions of N.

Assumption 3. *As functions of N, D and D_{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_{eff} are tighter, we still allow polynomial growth of D_{eff} in N for some lower powers of N. Our final assumption is on the smallest entry of θ^*_S. Such conditions – typically called *beta-min conditions* – are frequently used in the ℓ_1 literature to ensure $\hat{S} = S$ [Wainwright, 2009, Lee et al., 2014, Li et al., 2015].

Assumption 4. *θ^* satisfies $\min_{s\in S}|\theta^*_s| > \sqrt{D_{\text{eff}}}T_{min}\lambda$, where T_{min} is some constant relating to the objective function f; see Assumption 15 in Appendix I.1 for an exact description.*

4.1 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 ε_n are i.i.d. c_ε-sub-Gaussian random variables.*

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

$$\lambda \geq \frac{C}{\alpha - M_{\text{lin}}}\left(\sqrt{\frac{c_x^2c_\varepsilon^2\log D}{N} + \frac{25c_x^2c_\varepsilon^2}{N}} + \frac{4c_xc_\varepsilon(\log(ND)+26)}{N}\right), \quad (8)$$

where $C > 0$ is a constant in $N, D, D_{\text{eff}}, c_x$, and c_ε, and M_{lin} is a scalar given by Eq. (36) in Appendix I that satisfies, as $N \to \infty$, $M_{\text{lin}} = o(1)$. Then for N sufficiently large, Condition 1 holds with probability at least $1 - 26e^{-25}$.

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

It is worth considering how the allowed values of λ in Eq. (8) compare to previous results in the ℓ_1 literature for the support recovery of $\hat{\theta}$. We will talk about a sequence of choices of λ scaling with N denoted by λ_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 $\operatorname{supp}\hat{\theta} \subseteq S$ 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 constraint on λ_N to the constraint implied by Eq. (8). We have that $M_{\text{lin}} = o(1)$ as $N \to \infty$, so that, for large N, the bound in Eq. (8) becomes $\lambda_N \geq c'\sqrt{\log(D)/N}$ for some constant c'. Thus, the sequence of λ_N satisfying Eq. (8) scales at exactly the same rate as those that ensure $\operatorname{supp}\hat{\theta} \subseteq S$. The scaling of λ_N is important, as the error in $\hat{\theta}$, $\|\hat{\theta} - \theta^*\|_2^2$, is typically proportional to λ_N. The fact that we have not increased the asymptotic scaling of λ_N therefore means that we can enjoy the same decay of $\|\hat{\theta} - \theta^*\|_2^2$ while ensuring $\operatorname{supp}\hat{\theta}_{\setminus n} = S$ for all n.

4.2 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 $\Pr[y_n = 1] = 1/(1 + e^{-x_n^T\theta^*})$.*

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

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

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

Furthermore, assume the scaling of L_{min} in N and D_{eff} is such that, under Assumption 3 and for sufficiently large N, $L_{min} \geq CN$ 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 $L_{min} = N - Cc_x^2\sqrt{ND_{\text{eff}}}$ (see Lemma 2 in Appendix I). The Hessian for logistic regression is significantly more complicated, and it is typical in the ℓ_1 literature to make some kind of assumption about its eigenvalues [Bach, 2010, 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 1 to 4, 6 and 7. Suppose the regularization parameter λ satisfies:*

$$\lambda \geq \frac{C}{\alpha - M_{\text{logr}}}\left(\sqrt{c_x^2\frac{25+\log D}{N}} + \frac{\sqrt{2c_x^2\log(ND)}+\sqrt{50c_x^2}}{N}\right), \quad (9)$$

where C, C' are constants in N, D, D_{eff}, and c_x, and M_{logr} is a scalar given by Eq. (67), that, as $N \to \infty$, satisfies $M_{\text{logr}} = o(1)$. Then for N sufficiently large, Condition 1 is satisfied with probability at least $1 - 43e^{-25}$.

A restatement and proof of Theorem 3 are given as Theorem 5 in Appendix I. Similar to the remarks after Theorem 2, Theorem 3 implies that when applied to logistic regression, $\text{IJ}_{\setminus n}$ and $\text{NS}_{\setminus n}$ have accuracy that depends on (the ideally small) D_{eff} rather than (the potentially large) D, even when $D = o(e^N)$.

Theorem 3 has implications for the work of Obuchi and Kabashima [2018], who conjecture that, as $N \to \infty$, the change in support of ℓ_1 regularized logistic regression becomes negligible as each datapoint is left out; this assumption is used to derive a version of $\text{NS}_{\setminus n}$ for logistic regression. Our Theorem 3 confirms this conjecture by proving the stronger fact that the support is unchanged with high probability for finite data.

5 Experiments

We now empirically verify the good behavior of $\text{NS}_{\setminus n}$ and $\text{IJ}_{\setminus n}$ (i.e. proposal #4) and show that it far outperforms #2 (smoothing ℓ_1) and #3 (subsampling) in our high-dimensional regime of interest. All the code to run our experiments here is available online.[6] We

[6] `https://bitbucket.org/wtstephe/sparse_appx_cv/`

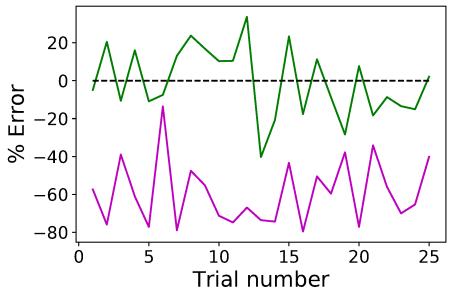

Figure 2: Error (Eq. (10)) across approximations for ℓ_1 LOOCV (legend shared with Fig. 3). The error for $\text{IJ}_{\setminus n}$ (black dashed) is too small to see, but nonzero; it varies between -0.06% and 0.04%.

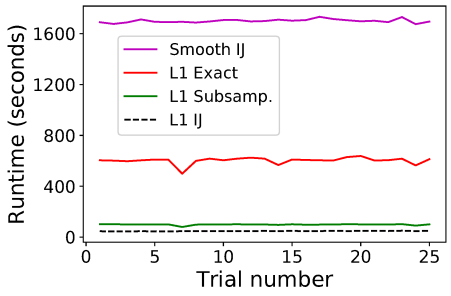

Figure 3: Runtimes for the experiments in Fig. 2 with exact CV (red) included for comparison. The $D \times D$ matrix inversion in the smoothed problem is so slow that even exact CV with an efficient ℓ_1 solver is faster.

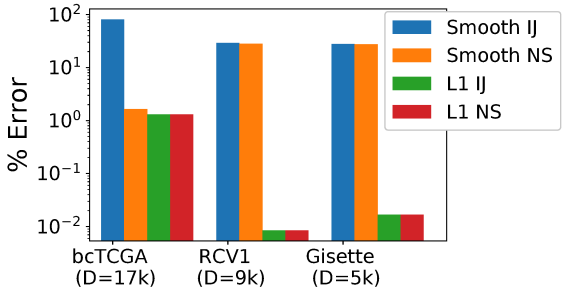

Figure 4: Log percent accuracy (Eq. (10)) for real data experiments. For each dataset, we give the accuracy of approximate CV compared to exact CV using both a smoothed approximation to ℓ_1 and the $\text{IJ}_{\setminus n}$, $\text{NS}_{\setminus n}$ approximations. For the bcTCGA dataset (linear regression), the nearly quadratic objective seems to be extremely well approximated by one Newton step, making $\widetilde{\text{NS}}_{\setminus n}(R^\eta)$ significantly more accurate than $\widetilde{\text{IJ}}_{\setminus n}(R^\eta)$; see the note at the end of Appendix D.4 about the exactness of $\widetilde{\text{NS}}_{\setminus n}(R)$ on quadratic objectives.

focus comparisons in this section on proposals #2–#4, as they all directly address ℓ_1-regularized problems. For an illustration of the failings of proposal #1, see Appendix C. To illustrate #2, we consider the smooth approximation given by Rad and Maleki [2020]: $R^\eta(\theta) := \sum_{d=1}^D \frac{1}{\eta}\big(\log(1 + e^{\eta\theta_d}) + \log(1 + e^{-\eta\theta_d})\big)$. While $\lim_{\eta\to\infty} R^\eta(\theta) = \|\theta\|_1$, we found that this approximation became numerically unstable for optimization when η was much larger than 100, so we set $\eta = 100$ in our experiments.

Simulated experiments. First, we trained logistic regression models on twenty-five random datasets in which $x_{nd} \overset{i.i.d.}{\sim} N(0,1)$ with $N = 500$ and $D = 40{,}000$. We set $\lambda = 1.5\sqrt{\log(D)/N}$ to mimic our condition in Eq. (9). The true θ^* was supported on its first five entries. We evaluate our approximations by comparing the CV estimate of out-of-sample error ("LOO") to the approximation ALOO $:= \frac{1}{N}\sum_{n=1}^N f(x_n^T \mathrm{IJ}_{\setminus n}, y_n)$. We report percent error:

$$|\mathrm{ALOO} - \mathrm{LOO}|/\mathrm{LOO}. \tag{10}$$

Fig. 2 compares the accuracy and run times of proposals #2 and #3 versus $\mathrm{IJ}_{\setminus n}$. We chose the number of subsamples so that subsampling CV would have about the same runtime as computing $\mathrm{IJ}_{\setminus n}$ for all n.[7] We see that subsampling usually has much worse accuracy than $\mathrm{IJ}_{\setminus n}$. Using $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ with $R^{100}(\theta)$ as a regularizer is even worse, as we approximate over all D dimensions; the resulting approximation is slower and less accurate – by multiple orders of magnitude – across all trials.

The importance of setting λ. Our theoretical results heavily depend on particular settings of λ to obtain the fixed-dimensional error scaling shown in blue in Fig. 1. One might wonder if such a condition on λ is necessary for approximate CV to be accurate. We offer evidence in Appendix F that this scaling is necessary by empirically showing that when λ violates our condition, the error in $\mathrm{IJ}_{\setminus n}$ grows with N.

Real data experiments. We next study how dependent our results are on the particular distributional assumptions in Theorems 2 and 3. We explore this question with a number of publicly available datasets [bcTCGA, 2018, Lewis et al., 2004, Guyon et al., 2004]. We chose these datasets because they have a high enough dimension to observe the effect of our results, yet are not so large that running exact CV for comparison is prohibitively expensive; see Appendix G for details (including our settings of λ). For each dataset, we approximate CV for the ℓ_1 regularized model using $\mathrm{IJ}_{\setminus n}$ and $\mathrm{NS}_{\setminus n}$. For comparison, we report the accuracy of $\widetilde{\mathrm{IJ}}_{\setminus n}(R^\eta)$ and $\widetilde{\mathrm{NS}}_{\setminus n}(R^\eta)$ with $\eta = 100$. Our results in Fig. 4 show that $\mathrm{IJ}_{\setminus n}$ is significantly faster and more accurate than exact CV or smoothing.

[7] Specifically, we computed 41 different $\hat{\theta}_{\setminus n}$ for each trial in order to roughly match the time cost of computing $\mathrm{IJ}_{\setminus n}$ for all $N = 500$ datapoints.

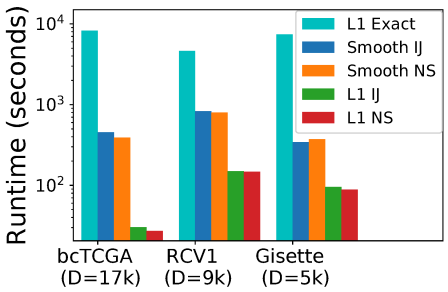

Figure 5: Log runtimes for experiments in Fig. 4, with exact CV included for comparison.

To 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 $\mathrm{NS}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$ for *all* n took three minutes.

6 Conclusions and future work

We have provided the first analysis of when CV can be approximated quickly *and* accurately in high dimensions with guarantees on quality. We have seen that, out of a number of proposals in the literature, running approximate CV on the recovered support (i.e., $\mathrm{NS}_{\setminus n}$ and $\mathrm{IJ}_{\setminus n}$) forms the only proposal that reaches these goals both theoretically and empirically. We hope this analysis will serve as a starting point for further understanding of when approximate CV methods work for high-dimensional problems.

We see three interesting directions for future work. First, this work has focused entirely on approximate CV for model assessment. In Appendix H, we show that approximate CV for model *selection* can have unexpected and undesirable behavior; we believe understanding this behavior is one of the most important future directions in this area. Second, one could extend our results to results to the higher order infinitesimal jackknife presented in Giordano et al. [2019a]. Finally, it would be interesting to consider our approximations as a starting point for subsampling estimators, as proposed in Magnusson et al. [2019].

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

K. T. Abou-Moustafa and C. Szepesvári. An exponential tail bound for Lq stable learning rules. Application to k-folds cross-validation. In *ISAIM*, 2018.

N. Agarwal, B. Bullins, and E. Hazan. Second-order stochastic optimization in linear time. *Journal of Machine Learning Research*, 2017.

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.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *arXiv Preprint*, December 2019.

bcTCGA. *Breast cancer gene expression data*, Nov 2018. Available at http://myweb.uiowa.edu/pbreheny/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 (NeurIPS)*, pages 3458–3468, 2017.

B. B. Biswal, P. A. Taylor, and J. L. Ulmer. Use of jackknife resampling techniques to estimate the confidence intervals of fmri parameters. *Journal of Computer Assisted Tomography*, 25, 2001.

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

J. Carrera, G. Rodrigo, and A. Jaramillo. Model-based redesign of global transcription regulation. *Nucleic Acids Research*, 39(5), 2009.

S. Chandrasekaran, S. Ament, J. Eddy, S. Rodriguez-Zas, B. Schatz, N. Price, and G. Robinson. Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proceedings of the National Academy of Sciences of the United States of America*, 108(44), 2011.

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated Lasso in high dimensions. *arXiv Preprint*, February 2020.

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

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1):1–22, 2009.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, June 1975.

R. Giordano, T. Broderick, and M. I. Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

R. Giordano, M. I. Jordan, and T. Broderick. A higher-order Swiss army infinitesimal jackknife. *arXiv Preprint*, July 2019a.

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 2019b.

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 (NeurIPS)*, 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.

A. Joshi, R. De Smet, K. Marchal, Y. Van de Peer, and T. Michoel. Module networks revisited: Computational assessment and prioritization of model predictions. *Bioinformatics*, 25(4), 2009.

P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference in Machine Learning (ICML)*, 2017.

P. W. Koh, K. S. Ang, H. Teo, and P. Liang. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

J. D. Lee, Y. Sun, and J. E. Taylor. On model selection consistency of regularized M-estimators. *arXiv Preprint*, October 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.

M. Magnusson, M. R. Andersen, J. Jonasson, and A. Vehtari. Bayesian leave-one-out cross-validation for large data. In *International Conference in Machine Learning (ICML)*, 2019.

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.

A. Powers, M. Pinto, O. Tang, J. Chen, C. Doberstein, and W. Asaad. Predicting mortality in traumatic intracranial hemorrhage. *Journal of Neurosurgery*, To Appear 2019.

K. R. Rad and A. Maleki. A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv Preprint*, January 2020.

D. A. Roff and R. Preziosi. The estimation of the genetic correlation: the use of the jackknife. *Heredity*, 73, 1994.

S. Saeb, L. Lonini, A. Jayaraman, D. Mohr, and K. Kording. The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5), 2017.

L. Steinberger and H. Leeb. Conditional predictive inference for high-dimensional stable algorithms. *arXiv Preprint*, sep 2018.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the American Statistical Association*, 36(2):111–147, 1974.

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

V. Vapnik. Principles of risk minimization for learning theory. In *Neural Information Processing Systems (NeurIPS)*, 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 Cross-validation methods

In this appendix, we review standard cross-validation (CV) for optimization problems of the form:

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

where $\Theta \subseteq \mathbb{R}^D$. By leave-one-out cross-validation (LOOCV), we mean the process of repeatedly computing:

$$\hat{\theta}_{\backslash n} := \sum_{m:\, m\neq n} f_m(\theta) + \lambda R(\theta).$$

The parameter estimates $\{\hat{\theta}_{\backslash n}\}_{n=1}^N$ might then be used to produce an estimate of variability or out-of-sample error; e.g., to estimate the out-of-sample error, one computes $(1/N)\sum_n f_n(\hat{\theta}_{\backslash n})$. By K-fold cross-validation, we mean the process of splitting up the dataset into K disjoint folds, $S_1, \ldots, S_K$ with $S_1 \cup \cdots \cup S_K = [N]$. One then estimates the parameters:

$$\hat{\theta}_{\backslash S_k} := \arg\min_{\theta\in\Theta} \sum_{n:\, n\notin S_k} f_n(\theta) + \lambda R(\theta).$$

The parameter estimates $\{\hat{\theta}_{\backslash S_k}\}_{k=1}^K$ might then be used to produce an estimate of variability or out-of-sample error.

B Scaling of the leave-one-out objective

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

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

An alternative would be to use the objective $1/(N-1)\sum_{m:\, m\neq n} f_m + \lambda R$ 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 ℓ_1 regularized problems seems to follow the $1/(N-1)$ scaling [Homrighausen and McDonald, 2014, 2013, Miolane and Montanari, 2018, Chetverikov et al., 2020]. 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, 2020, Wang et al., 2018, Xu et al., 2019, Obuchi and Kabashima, 2016, 2018]. Note that the scaling is not relevant in Giordano et al. [2019b], 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 $1/N$ choice from 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.

C Approximately solving $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ and $\widetilde{\mathrm{NS}}_{\backslash n}(R)$

We have seen $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ and $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ are in general not accurate for high-dimensional problems. Even worse, they can become prohibitively costly to compute due to the $O(D^3)$ cost required to solve the needed linear systems. One idea to at least alleviate this computational burden, proposed by Koh and Liang [2017] in a slightly different context, is to use a stochastic inverse Hessian-vector-product from Agarwal et al. [2017] to approximately compute $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ and $\widetilde{\mathrm{NS}}_{\backslash n}(R)$. Although this method works well for the purposes of Koh and Liang [2017], we will see that in the context of approximate CV, it adds a large amount of extra error on top of the already inaccurate $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ and $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$.

We first describe this stochastic inverse Hessian-vector-product technique and argue that it is not suitable for approximating cross-validation. The main idea from Agarwal et al. [2017] is to use the series:

$$H^{-1} = \sum_{k=0}^{\infty} (I - H)^k,$$

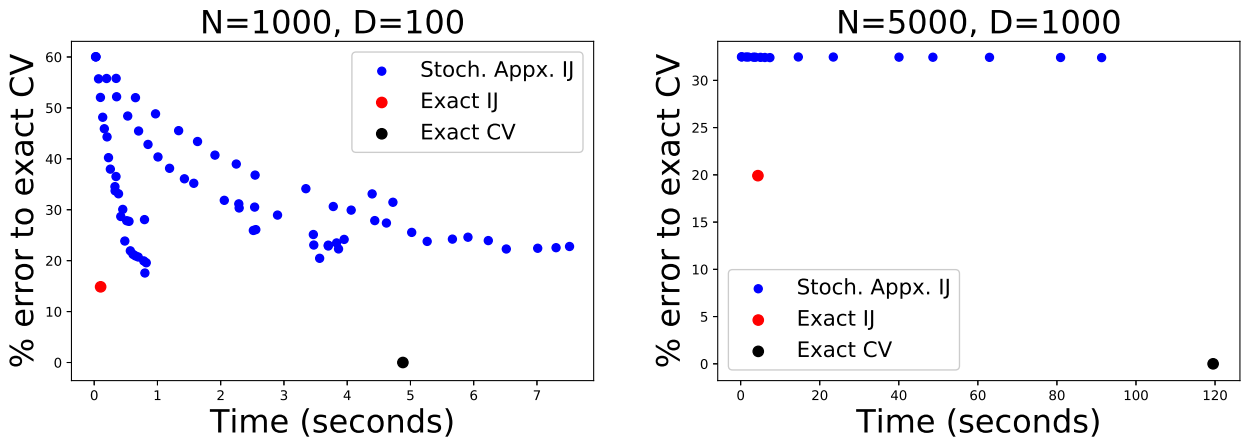

Figure 6: Stochastic Hessian experiments from Appendix C. We show percent error of approximation versus compute time for two different dataset sizes. We show three techniques for computing CV: exactly computing CV (black dot, which naturally has 0% error), $\widetilde{\text{IJ}}_{\setminus n}(R)$ with exactly computing the needed linear systems (red dot), and $\widetilde{\text{IJ}}_{\setminus n}(R)$ with the stochastic solves described in Appendix C (blue dots, one for each setting of the parameters K and M). Values of M and K used are described in Appendix C; we use an extended range of settings for the smaller dataset to more extensively illustrate the approximation's behavior. Settings of K and M for which the stochastic solves are roughly as fast as exactly computing $\widetilde{\text{IJ}}_{\setminus n}(R)$ result in a significantly less accurate approximation of CV.

which holds for any positive definite H with $\|H\|_{op} \leq 1$. Now, we can both truncate this series at some level K and write it recursively as:

$$H^{-1} \approx H_K^{-1} := I + (I - H)H_{K-1}^{-1},$$

where $H_0^{-1} = I$. Next, to avoid computing H explicitly, we can note that if A_k is some random variable with $E[A_k] = H$, we can instead just sample a new A_k at each iteration to define:

$$\bar{H}_k^{-1} := I + (I - A_k)\bar{H}_{K-1}^{-1}.$$

In our case, we pick a random $n_k \in [N]$ and set $A_k = \nabla_\theta^2 f(x_{n_k}^T \theta, y_{n_k}) + (1/N)\lambda\nabla^2 R(\theta)$. Finally, Agarwal et al. [2017] suggest taking M samples of $\bar{H}_K^{-1}$ and averaging the results to lower the variance of the estimator. This leaves us with two parameters to tune: M and K. Increasing either will make the estimate more accurate and more expensive to compute. Koh and Liang [2017] use this approximation to compute $\widetilde{\text{IJ}}_{\setminus n}(R)$ for high dimensional models such as neural networks; however, we remark that their interest lies in the qualitative properties of $\widetilde{\text{IJ}}_{\setminus n}(R)$, such as signs and relative magnitudes across various values of n. It remains to be seen whether this stochastic solver can be successfully used to approximate CV.

To test the application of this approximation to approximate CV, we generated a synthetic logistic regression dataset with covariates $x_{nd} \overset{i.i.d.}{\sim} N(0, 1)$. We use $R(\theta) = \|\theta\|_2^2$. In Fig. 6, we show that for two settings of N and D there are no settings of M and K for which using $\bar{H}_K^{-1}$ to compute $\widetilde{\text{IJ}}_{\setminus n}(R)$ provides a both fast and accurate approximation to CV. Specifically, we range $K \in \{1, 20, 30, 50, 60, 80, 100, 120\}$ and $M \in \{2, 25\}$ and see that when the stochastic approximation is faster, it provides only a marginal speedup while providing a significantly worse approximation error.

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

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

$$\hat{\theta}_{\backslash n} \approx \widetilde{\mathrm{IJ}}_{\backslash n}(R) := \hat{\theta} + \frac{1}{N} H(\hat{\theta})^{-1} \nabla_\theta f(x_n^T \hat{\theta}, y_n). \tag{11}$$

The "Newton step" approximation, similar to the approach in Beirami et al. [2017] and identical to the approximation in Rad and Maleki [2020], Wang et al. [2018], is:

$$\hat{\theta}_{\backslash n} \approx \widetilde{\mathrm{NS}}_{\backslash n}(R) := \hat{\theta} + \frac{1}{N} \left(H(\hat{\theta}) - \frac{1}{N} \nabla_\theta^2 f(x_n^T \hat{\theta}, y_n) \right)^{-1} \nabla_\theta f(x_n^T \hat{\theta}, y_n). \tag{12}$$

D.1 Derivation of $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$

We will see in Appendix D.3 that, after some creative algebra, $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$ is an instance of $\hat{\theta}_{IJ}$ from Definition 2 of Giordano et al. [2019b]. However, this somewhat obscures the motivation for considering Eq. (11). As an alternative to jamming our problem setup into that considered by Giordano et al. [2019b], we can more directly obtain the approximation in Eq. (11) by a derivation only slightly different from that in Giordano et al. [2019b]. We begin by defining $\hat{\theta}^w$ as the solution to a weighted optimization problem with weights $w_n \in \mathbb{R}$:

$$\hat{\theta}^w := \arg\min_{\theta \in \Theta} G(w, \theta) := \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^N w_n f(x_n^T \theta, y_n) + \lambda R(\theta), \tag{13}$$

where we assume G to be twice continuously differentiable with an invertible Hessian at $\hat{\theta}^{\mathbf{1}}$ (where $\hat{\theta}^{\mathbf{1}}$ is the solution in Eq. (13) with all $w_n = 1$). For example, we have that $\hat{\theta}_{\backslash n} = \hat{\theta}^w$ if w is the N-dimensional vector of all ones but with a zero in the nth 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. (13) and take a total derivative with respect to w_n:

$$\begin{aligned}
&\frac{\partial G}{\partial \theta}\Big|_{w=1, \theta=\hat{\theta}^{\mathbf{1}}} = 0 \\
&\implies \frac{d}{dw_n} \frac{\partial G}{\partial \theta}\Big|_{w=1, \theta=\hat{\theta}^{\mathbf{1}}} = \frac{\partial^2 G}{\partial \theta \partial w_n}\Big|_{w=1, \theta=\hat{\theta}^{\mathbf{1}}} \frac{dw_n}{dw_n} + \frac{\partial^2 G}{\partial \theta^2}\Big|_{w=1, \theta=\hat{\theta}^{\mathbf{1}}} \frac{d\hat{\theta}^w}{dw_n}\Big|_{w=1} = 0.
\end{aligned}$$

Re-arranging, defining $H(\hat{\theta}^{\mathbf{1}}) := \nabla_\theta^2 G(w, \hat{\theta}^{\mathbf{1}})$, and using the assumed invertibility of $H(\hat{\theta}^{\mathbf{1}})$ gives:

$$\frac{d\hat{\theta}}{dw_n}\Big|_{w=1} = -\left(\frac{\partial^2 G}{\partial \theta^2}\Big|_{w=1, \theta=\hat{\theta}^{\mathbf{1}}} \right)^{-1} \frac{\partial^2 G}{\partial w_n \partial \theta}\Big|_{w=1, \theta=\hat{\theta}^{\mathbf{1}}} = -\frac{1}{N} H(\hat{\theta})^{-1} \nabla_\theta f(x_n^T \hat{\theta}, y_n). \tag{14}$$

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

$$\hat{\theta}^w \approx \hat{\theta} + \sum_{n=1}^N \frac{d\hat{\theta}}{dw_n}\Big|_{w=1} (w_n - 1) \tag{15}$$

$$= \hat{\theta} - \frac{1}{N} \sum_{n=1}^N H(\hat{\theta})^{-1} \nabla_\theta f(x_n^T \hat{\theta}, y_n)(w_n - 1). \tag{16}$$

For the special case of w being the vector of all ones with a zero in the nth coordinate (i.e., the weighting for LOOCV), we recover Eq. (11).

D.2 Invertibility in the definition of $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ and $\widetilde{\mathrm{IJ}}_{\backslash n}(R)$

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

D.3 Accuracy of $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ for regularized problems

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

$$\max_{n\in[N]} \|\tilde{\psi}_{IJ}^{\setminus n} - \hat{\theta}_{\setminus n}\|_2 \leq \frac{C'}{N^2} \sup_{\theta\in\Theta} \max_{n\in[N]} \left\|\nabla_\theta f(x_n^T\theta, y_n)\right\|_\infty \leq \frac{C}{N}, \tag{17}$$

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

Proposition 2 provides two bounds on the error $\|\widetilde{\mathrm{IJ}}_{\setminus n}(R) - \hat{\theta}_{\setminus n}\|_2$: 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 θ is restricted to some compact set Θ. Although this assumption may not be satisfied by problems of interest, one may be willing to assume that θ lives in some bounded set in practice. In any case, such an assumption seems necessary to apply the results of Giordano et al. [2019b] to most unregularized problems, as they, for example, require $\sup_{\theta\in\Theta} F(\theta)$ to be bounded. We will require the compactness of Θ 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 ℓ_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.

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. [2019b] and then showing that the re-written problem satisfies the assumptions of their Corollary 1. First, the framework of Giordano et al. [2019b] applies to weighted optimization problems of the form:

$$\hat{\theta}^w := \theta \in \Theta \; s.t. \; \frac{1}{N}\sum_{n=1}^{N} w_n g_n(\theta) = 0. \tag{18}$$

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 $R(\theta)$:

$$\frac{1}{N+1} w_0 (N+1)\lambda \nabla R(\theta) + \frac{1}{N+1}\sum_{n=1}^{N} w_n \frac{N+1}{N} \nabla f(x_n^T\theta, y_n). \tag{19}$$

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 $\hat{\theta}^w = \hat{\theta}_{\setminus n}$. Finally, then, we can apply Definition 2 of Giordano et al. [2019b] to find the approximation $\theta_{IJ}(w)$ for the w that corresponds to leaving out n. We see that $\theta_{IJ}(w)$ in this case is exactly equal to $\widetilde{\mathrm{IJ}}_{\setminus n}(R)$ in our notation here.

Now that we know our approximation is actually an instance of $\theta_{IJ}(w)$, we need to check that Eq. (19) meets the assumptions of Corollary 1 of Giordano et al. [2019b] to apply their theoretical analysis to our problem. We check these below, first stating the assumption from Giordano et al. [2019b] and then covering why it holds for our problem.

1. (*Assumption 1*): *for all $\theta \in \Theta$, each g_n is continuously differentiable in θ.*
   For our problem, by assumption, $R(\theta)$ and $f(x_n^T\theta, y_n)$ are twice continuously differentiable functions of θ, so this assumption holds.

2. (*Assumption 2*): *for all $\theta \in \Theta$, the Hessian matrix, $H(\theta, 1) := (1/N)\sum_n \partial g_n(\theta)/\partial\theta^T$ is invertible and satisfies $\sup_{\theta\in\Theta} \|H^{-1}(\theta,1)\|_{op} \leq C_{op} < \infty$ for some constant C_{op}, where $\|\cdot\|_{op}$ denotes the operator norm on matrices with respect to the ℓ_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 θ, which implies:

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

   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 dth eigenvalue of a matrix with $\lambda_1 = \lambda_{min}$. Then:

$$\lambda_d(A+B) = \min_{\substack{E\subseteq\mathbb{R}^D \\ \dim E = d}} \max_{\substack{x\in E \\ \|x\|_2=1}} x^T(A+B)x \geq \min_{\substack{E\subseteq\mathbb{R}^D \\ \dim E = d}} \max_{\substack{x\in E \\ \|x\|_2=1}} x^T A x = \lambda_d(A),$$

   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 $(N+1) \times D \times D$ stack of Hessians, respectively. That is, $g(\theta)_{nd} := (\nabla_\theta f(x_n^T\theta, y_n))_d$ for $n = 1, \ldots, N$ and $g(\theta)_{N+1,d} := (\nabla_\theta R(\theta))_d$, with h defined similarly. Let $\|g(\theta)\|_2$ be the ℓ_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:*

$$\sup_{\theta\in\Theta} \frac{1}{\sqrt{N+1}} \|g(\theta)\|_2 \leq C_g < \infty$$
$$\sup_{\theta\in\Theta} \frac{1}{\sqrt{N+1}} \|h(\theta)\|_2 \leq C_h < \infty$$

   To see that this holds for our problem, we have that:

$$\begin{aligned}\|g(\theta)\|_2 &:= \left[\sum_{d=1}^{D}\left((\lambda(N+1)\nabla R(\theta)_d)^2 + \sum_{n=1}^{N}\left(\frac{N+1}{N}\right)^2 (\nabla f(x_n^T\theta, y_n)_d)^2\right)\right]^{1/2} \\ &\leq \lambda(N+1)\|\nabla R(\theta)\|_2 + \frac{N+1}{N}\left[\sum_{d=1}^{D}\sum_{n=1}^{N}(\nabla f(x_n^T\theta, y_n)_d)^2\right]^{1/2}.\end{aligned}$$

   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 λ is $O(1/\sqrt{N})$, the first term is equal to $O(\sqrt{N})\|\nabla R(\theta)\|_2$. The compactness of Θ 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 $\frac{O(\sqrt{N})}{\sqrt{N+1}}\|\nabla R(\theta)\|_2 \leq C_g^{(2)}$ for some constant $C_g^{(2)}$. Thus, we have that the assumption on $\|g(\theta)\|_2$ holds with $C_g = \frac{(N+1)}{N}C_g^{(1)} + C_g^{(2)}$. That the condition on $\|h(\theta)\|_2$ holds follows by the same reasoning.

4. (*Assumption 4*): *There exists some $\Delta_\theta > 0$ and $L_h < \infty$ such that if $\left\|\theta - \hat{\theta}\right\|_2 \leq \Delta_\theta$, then $\frac{1}{\sqrt{N+1}}\left\|h(\theta) - h(\hat{\theta})\right\|_2 \leq L_h\left\|\theta - \hat{\theta}\right\|_2$.*

We can show this holds for our problem by:

$$\left\|h(\theta) - h(\hat{\theta})\right\|_2 := \left\|\frac{N+1}{N}\nabla^2 F(\theta) + \lambda(N+1)\nabla^2 R(\theta) - \frac{N+1}{N}\nabla^2 F(\hat{\theta}) - \lambda(N+1)\nabla^2 R(\hat{\theta})\right\|_2 \leq (N+1)\lambda\left\|\nabla^2 R(\theta) - \nabla^2 R(\hat{\theta})\right\|_2 + \frac{N+1}{N}\left\|\nabla^2 F(\theta) - \nabla^2 F(\hat{\theta})\right\|_2,$$

where we have abused notation to denote $\left\|\nabla^2 F(\theta)\right\|_2 := \sqrt{\sum_{i,j=1}^{D}\sum_{n=1}^{N}\nabla^2_\theta(f(x_n^T\theta, y_n)_{ij})^2}$. 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 $\frac{N+1}{(\sqrt{N+1})(N)}\left\|\nabla^2 F(\theta) - \nabla^2 F(\hat{\theta})\right\|_2 \leq L_h^{(1)}\|\theta - \hat{\theta}\|_2$ 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 θ's, we know that $\nabla^2 R(\theta)$ is Lipschitz over this domain. Using this along with the assumption that λ is $O(1/\sqrt{N})$, we have that:

$$\frac{\lambda(N+1)}{\sqrt{N+1}}\left\|\nabla^2 R(\theta) - \nabla^2 R(\hat{\theta})\right\|_2 = O(1)\left\|\nabla^2 R(\theta) - \nabla^2 R(\hat{\theta})\right\|_2 \leq L_h^{(2)}\left\|\theta - \hat{\theta}\right\|_2,$$

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 $\frac{1}{\sqrt{N+1}}\|w\|_2 \leq C_w$ for some constant C_w.* This is immediately true for our definition of W, which, for all $w \in W$, has $\|w\|_2 = \sqrt{N}$.

□

D.4 Derivation of $\widetilde{\mathrm{NS}}_{\backslash n}(R)$

Wang et al. [2018] and Rad and Maleki [2020] derive $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ in Eq. (12) by taking a single Newton step on the objective $F^{\backslash n} + \lambda R$ starting at the point $\hat{\theta}$. For completeness, we include a derivation here. Recall that the objective with one datapoint left out is:

$$F^{\backslash n}(\theta) + \lambda R(\theta) := \frac{1}{N}\sum_{m=1}^{N} f(x_m^T\theta, y_m) - \frac{1}{N}f(x_n^T\theta, y_n) + \lambda R(\theta), \tag{20}$$

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

$$\hat{\theta}_{\backslash n} \approx \hat{\theta} - \left(H(\hat{\theta}) - \frac{1}{N}\nabla^2_\theta f(x_n^T\hat{\theta}, y_n)\right)^{-1}\left(\frac{1}{N}\sum_{m=1}^{N}\nabla_\theta f(x_m^T\hat{\theta}, y_m) - \frac{1}{N}\nabla_\theta f(x_n^T\hat{\theta}, y_n) + \lambda\nabla R(\hat{\theta})\right). \tag{21}$$

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

$$\hat{\theta}_{\backslash n} \approx \hat{\theta} + \frac{1}{N}\left(H(\hat{\theta}) - \frac{1}{N}\nabla^2_\theta f(x_n^T\hat{\theta}, y_n)\right)^{-1}\nabla_\theta f(x_n^T\hat{\theta}, y_n), \tag{22}$$

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

As $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ can be interpreted as a single Newton step on the objective $F^{\backslash n} + \lambda R$, it follows that $\widetilde{\mathrm{NS}}_{\backslash n}(R)$ 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, ℓ_2 regularized linear regression has $\widetilde{\mathrm{NS}}_{\backslash n}(R) = \hat{\theta}_{\backslash n}$ for all n. We further note that somewhat similar behavior can hold for ℓ_1 regularized linear regression. Specifically, when $\mathrm{sign}\hat{\theta} = \mathrm{sign}\hat{\theta}_{\backslash n}$, we have that the objective $F^{\backslash n} + \lambda\|\cdot\|_1$ is a quadratic when restricted to the dimensions in $\hat{S}$. In this case, $\mathrm{NS}_{\backslash n}$ can be interpreted as taking a Newton step on $F^{\backslash n} + \lambda\|\cdot\|_1$ restricted to the dimensions in $\hat{S}$. It follows that $\mathrm{NS}_{\backslash n} = \hat{\theta}_{\backslash n}$ when $\mathrm{sign}\hat{\theta} = \mathrm{sign}\hat{\theta}_{\backslash n}$ for ℓ_1 regularized linear regression.

D.5 Computation time of approximations

There is a major computational difference between Eq. (12) and Eq. (11): 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 [2020], Wang et al. [2018], Eq. (12) is much cheaper when considering the special case of generalized linear models. In this case, $\nabla^2_\theta f_n$ 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. (12) given only H^{-1}; this is how Equation 8 in Rad and Maleki [2020] and Equation 21 in Wang et al. [2018] are derived. Even though we only consider GLMs in this work, we still study Eq. (11) with the hope of retaining scalability in more general problems.

E Derivation of $\mathrm{IJ}_{\backslash n}$ and $\mathrm{NS}_{\backslash n}$ via smoothed approximations

As noted in Section 2, Rad and Maleki [2020], Wang et al. [2018] derive the $\mathrm{NS}_{\backslash n}$ approximation by considering $\widetilde{\mathrm{NS}}_{\backslash n}(R^\eta)$ with R^η being some smoothed approximation to the ℓ_1 norm, and then taking the limit of $\widetilde{\mathrm{NS}}_{\backslash n}(R^\eta)$ as the amount of smoothness goes to zero. We review this approach and then state our Proposition 4, which says that the same technique can be used to derive $\mathrm{IJ}_{\backslash n}$.

We first give two possible ways to smooth the ℓ_1 norm. The first is given by Rad and Maleki [2020]:

$$\|\theta\|_1 \approx R^\eta(\theta) := \sum_{d=1}^{D} \frac{1}{\eta}\Big(\log(1 + e^{\eta\theta_d}) + \log(1 + e^{-\eta\theta_d}) \Big), \tag{23}$$

The second option is to use the more general smoothing framework described by Wang et al. [2018]. They allow selection of a function $q : \mathbb{R} \to \mathbb{R}$ 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:

$$R^\eta(\theta) := \eta \sum_{d=1}^{D} \int_{-\infty}^{\infty} |u| q\big(\eta(\theta_d - u)\big) du, \tag{24}$$

In both Eqs. (23) and (24), we have $\lim_{\eta\to\infty} = \|\theta\|_1$. Notice that either choice of R^η is twice differentiable for any $\eta < \infty$, so one can consider the approximations $\widetilde{\mathrm{NS}}_{\backslash n}(R^\eta), \widetilde{\mathrm{IJ}}_{\backslash n}(R^\eta)$. We now state two assumptions, both of which are given by Rad and Maleki [2020], Wang et al. [2018], under which one can show the limits of these approximations as $\eta \to \infty$ are equal to $\mathrm{NS}_{\backslash n}$ and $\mathrm{IJ}_{\backslash n}$.

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 $\left\|\hat{z}_{\hat{S}^c}\right\|_\infty < 1$.*

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 [2020]; Theorem 4.2 of Wang et al. [2018])**.** *Take Assumptions 8 and 9. Suppose $H_{\hat{S}\hat{S}}$ has strictly positive eigenvalues. Let $H^{\backslash n}_{\hat{S}\hat{S}} := H_{\hat{S}\hat{S}} - [\nabla^2_\theta f(x_n^T\hat{\theta}, y_n)]_{\hat{S}\hat{S}}$, and suppose that, for all n, $H^{\backslash n}_{\hat{S}\hat{S}}$ is invertible. Then, for R^η as in Eq. (23) or Eq. (24),*

$$\mathrm{NS}_{\backslash n} := \lim_{\eta\to\infty} \widetilde{\mathrm{NS}}_{\backslash n}(R^\eta) = \left(\hat{\theta}_{\hat{S}} + (H^{\backslash n}_{\hat{S}\hat{S}})^{-1} \left[\nabla_\theta f(x_n^T\hat{\theta}, y_n)\right]_{\hat{S}}\right). \tag{25}$$

As noted in the main text, we show that a very similar result holds for the limit of $\widetilde{\mathrm{IJ}}_{\backslash n}(R^\eta)$:

Proposition 4. *Take Assumptions 8 and 9. Suppose $H_{\hat{S}\hat{S}}$ is invertible. Then for R^η as in Eq. (23) or Eq. (24):*

$$\mathrm{IJ}_{\backslash n} := \lim_{\eta\to\infty} \widetilde{\mathrm{IJ}}_{\backslash n}(R^\eta) = \begin{pmatrix} \hat{\theta}_{\hat{S}} + H^{-1}_{\hat{S}\hat{S}} \left[\nabla_\theta f(x_n^T\hat{\theta}, y_n)\right]_{\hat{S}} \\ 0 \end{pmatrix}. \tag{26}$$

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^η in the next two subsections.

E.1 Proof of Proposition 4 using Eq. (23)

This proof is almost identical to the proof of Theorem 1 from Rad and Maleki [2020]. First we will need some notation. Let $\hat{\theta}^\eta$ be the solution to Eq. (1) using R^η from Eq. (23) as the regularizer. Let $\hat{S}_\eta := \left\{ i \,:\, |\hat{\theta}^\eta| > c/\eta \right\}$ for some constant c. We know from the arguments in Appendix A.2 of Rad and Maleki [2020] that for an appropriately chosen c and $\eta > C$ for some large constant $C > 0$, we have $S^\eta = \hat{S} =: \operatorname{supp} \hat{\theta}$. Next, define the scalars $\hat{D}_n^{(1,\eta)}$ and $\hat{D}_n^{(2,\eta)}$ as the derivatives of f evaluated at $\hat{\theta}^\eta$:

$$\hat{D}_n^{(1,\eta)} := \frac{df(z, y_n)}{dz}\Big|_{z=x_n^T\hat{\theta}^\eta}, \quad \hat{D}_n^{(2,\eta)} := \frac{d^2 f(z, y_n)}{dz^2}\Big|_{z=x_n^T\hat{\theta}^\eta}. \tag{27}$$

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

$$A := X^T_{\cdot,\hat{S}_\eta} \operatorname{diag}\left\{\hat{D}_n^{(2,\eta)}\right\} X_{\cdot,\hat{S}_\eta} + \lambda \nabla^2 R^\eta(\hat{\theta}^\eta), \qquad B := X^T_{\cdot,\hat{S}_\eta^c} \operatorname{diag}\left\{\hat{D}_n^{(2,\eta)}\right\} X_{\cdot,\hat{S}_\eta} + \lambda \nabla^2 R^\eta(\hat{\theta}^\eta)$$
$$C := X^T_{\cdot,\hat{S}_\eta^c} \operatorname{diag}\left\{\hat{D}_n^{(2,\eta)}\right\} X_{\cdot,\hat{S}_\eta^c} + \lambda \nabla^2 R^\eta(\hat{\theta}^\eta), \qquad D := (A - BC^{-1}B^T)^{-1}$$

We can then compute the block inverse of the Hessian of the smoothed problem, H_η^{-1} as:

$$H_\eta^{-1} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} = \begin{pmatrix} D & -DBC^{-1} \\ -C^{-1}B^T D & A^{-1} + A^{-1}BDB^T A^{-1} \end{pmatrix}. \tag{28}$$

Rad and Maleki [2020] show that all blocks of H_η^{-1} converge to zero as $\eta \to \infty$ except for the upper left, which has $D \to X^T_{\cdot,\hat{S}} \operatorname{diag}\left\{\hat{D}_n^{(2)}\right\} X_{\cdot,\hat{S}}$. So, we have that the limit of $\widetilde{\mathrm{IJ}}_{\setminus n}(R^\eta)$ is:

$$\lim_{\eta\to\infty} \widetilde{\mathrm{IJ}}_{\setminus n}(R^\eta) = \lim_{\eta\to\infty} H_\eta^{-1} \hat{D}_n^{(1,\eta)} x_n = \hat{D}_n^{(1)} \begin{pmatrix} (X^T_{\cdot,\hat{S}} \operatorname{diag}\left\{\hat{D}_n^{(2)}\right\} X_{\cdot,\hat{S}})^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_{n\hat{S}} \\ x_{n\hat{S}^c} \end{pmatrix}, \tag{29}$$

where we used that $\hat{\theta}_\eta \to \hat{\theta}$ by Lemma 15 of Rad and Maleki [2020], 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 $\hat{D}_n^{(1)} x_{n\hat{S}} = [\nabla_\theta f(x_n^T\hat{\theta}, y_n)]_{\hat{S}}$.

E.2 Proposition 4 using Eq. (24)

This proof proceeds along the exact same direction as when using Eq. (23). In their proof of their Theorem 4.2, Wang et al. [2018] provide essentially all the same ingredients that Rad and Maleki [2020] do, except for the general class of smoothed approximations given by Eq. (24). 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. (29).

F The importance of correct support recovery

Theorem 1 shows that each $\hat{\theta}_{\setminus n}$ having correct support (i.e., $\operatorname{supp} \hat{\theta}_{\setminus n} = \operatorname{supp} \theta^*$) 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 $\mathrm{IJ}_{\setminus n}$ 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 θ^* 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 $\varepsilon_n \overset{i.i.d.}{\sim} N(0,1)$.

To examine what happens when the recovered supports are and are not correct, we use slightly different values of the regularization parameter λ. Specifically, the results of Wainwright [2009] (especially their Theorem 1) tell us that the support recovery of ℓ_1 regularized linear regression will change sharply around $\lambda \approx 4\sqrt{\log(D)/N}$, where lower values of λ will fail to correctly recover the support. With this in mind, we choose two settings of λ: $1.0\sqrt{\log(D)/N}$ and $10.0\sqrt{\log(D)/N}$. As expected, the righthand side of Fig. 7 shows that the accuracy of $\mathrm{IJ}_{\setminus n}$ is drastically different in these two situations. The lefthand plot of Fig. 7 offers an explanation for this observation: the support of $\operatorname{supp} \hat{\theta}_{\setminus n}$ grows with N under the lower value of λ, whereas the larger value

of λ ensures that $|\operatorname{supp}\hat{\theta}_{\setminus n}| = |\operatorname{supp}\theta^*| = \text{const}$. 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.

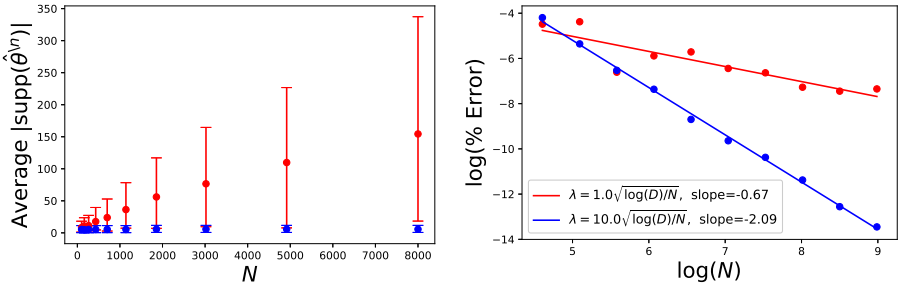

Figure 7: Illustration of the role of support recovery in the accuracy of $\text{IJ}_{\setminus n}$ in the case of linear regression. *Left*: Points show the average of $|\operatorname{supp}\hat{\theta}_{\setminus n}|$ over random values of n. Error bars show the min and max $|\operatorname{supp}\hat{\theta}_{\setminus n}|$ over these n. For $\lambda = 10.0\sqrt{\log(D)/N}$ (blue), the mean recovered support is constant with N. For $\lambda = 1.0\sqrt{\log(D)/N}$ (red), $|\operatorname{supp}\hat{\theta}_{\setminus n}|$ grows with N, and varies dramatically for different values of n. *Right*: Percent error (Eq. (10)) 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 λ is somewhat concerning. However, we emphasize that sensitivity exists for ℓ_1 regularization in general; as previously noted, Wainwright [2009] demonstrated similarly drastic behavior of $\operatorname{supp}\hat{\theta}$ in the same exact linear regression setup that we use here. On the other hand, Homrighausen and McDonald [2014] do show that using exact LOOCV to select λ for ℓ_1 regularized linear regression gives reasonable results. In Appendix H, we empirically show this is sometimes, but not always, the case for our and other approximate CV methods.

Accuracy of approximate CV by optimization error. In early experiments, we used the Python bindings for the glmnet package [Friedman et al., 2009] to solve our ℓ_1 regularized problems. However, we found that both $\text{IJ}_{\setminus n}$ and $\text{NS}_{\setminus n}$ failed to recover the roughly $1/N^2$ scaling present in fixed-dimensional problems (e.g. as shown in Fig. 1 of Section 1) that we would expect given our theoretical results. We found that this was due to the relatively loose convergence tolerance with which glmnet is implemented (e.g. parameter changes of $\leq 1 \times 10^{-4}$ between iterations), which seems to be an issue for approximate CV methods and related approximations [Giordano et al., 2019b, 2015]. We implemented our own ℓ_1 solver in Python using many of the speed-ups proposed in Friedman et al. [2009] and set a convergence theshold of 1×10^{-10} for the initial fit of $\hat{\theta}$. This solver was used to produce all of our results, including Fig. 7, which shows the expected roughly $1/N^2$ accuracy of $\text{IJ}_{\setminus n}$ in blue.

G Details of real experiments

We use three publicly available datasets for our real-data experiments in Section 5:

1. The "Gisette" dataset Guyon et al. [2004] is available from the UCI repository at `https://archive.ics.uci.edu/ml/datasets/Gisette`. The dataset is constructed from the MNIST handwritten digits dataset. Specifically, the task is to differentiate between handwritten images of either "4" or "9." There are $N = 6{,}000$ training examples, each of which has $D = 5{,}000$ features, some of which are junk "distractor features" added to make the problem more difficult.

2. The "bcTCGA" 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`. 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 ℓ_1 regularized regression on each of these datasets, we first needed to select a value of λ. Since all of these datasets are fairly high dimensional, our experiments in Appendix H suggests our approximation will be inaccurate for values of λ that are "too small." In an attempt to get the order of magnitude for λ correct, we used the theoretically motivated value of $\lambda = C\sqrt{\log(D)/N}$ for some constant C (e.g., Li et al. [2015] shows this scaling of λ will recover the correct support for both linear and logistic regression). Section 5 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 $|\operatorname{supp}\hat{\theta}|$. 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 $|\operatorname{supp}\hat{\theta}|$ 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.

H Selection of λ

Our work in this paper is almost exclusively focused on approximating CV for model assessment. However, this is not the only use-case of CV. CV is also commonly used for model selection, which, as a special case, contains hyperparameter tuning. Previous authors have used approximate CV methods for hyperparameter tuning in the way one might expect: for various values of λ, compute $\hat{\theta}$ and then use approximate CV to compute the out-of-sample error of each $\hat{\theta}$; the λ leading to the lowest out-of-sample error is then selected [Obuchi and Kabashima, 2016, 2018, Beirami et al., 2017, Rad and Maleki, 2020, Wang et al., 2018, Giordano et al., 2019b]. While many of these authors theoretically study the accuracy of approximate CV, we note that they only do so in the context of model *assessment* and only empirically study approximate CV for hyperparameter tuning. In this appendix, we add to these experiments by showing that approximate CV can exhibit previously undemonstrated complex behavior when used for hyperparameter tuning.

We generate two synthetic ℓ_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 θ^* 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 λ, we find $\hat{\theta}$, and measure the train, test, exact LOOCV, and approximate LOOCV errors via both $\mathrm{NS}_{\backslash n}$ and $\mathrm{IJ}_{\backslash n}$; the results are plotted in Fig. 8. $\mathrm{NS}_{\backslash n}$ (blue dashed curve) is an extremely close approximation to exact CV (red curve) in both datasets and selects a λ that gives a test error very close to the λ selected by exact CV. On the other hand, $\mathrm{IJ}_{\backslash n}$ (solid blue curve) performs very differently on the two datasets. For $D = 75$, it selects a somewhat reasonable value for λ; however, for $D = 150$, $\mathrm{IJ}_{\backslash n}$ goes disastrously wrong by selecting the obviously incorrect value of $\lambda = 0$. While the results in Fig. 8 come from using our $\mathrm{IJ}_{\backslash n}$ to approximate CV for an ℓ_1 regularized problem, we note that this issue is not specific to the current work; we observed similar behavior when using ℓ_2 regularization and the pre-existing $\widetilde{\mathrm{IJ}}_{\backslash n}(\ell_2)$.

While $\mathrm{NS}_{\backslash n}$ performs far better than $\mathrm{IJ}_{\backslash n}$ in the experiments here, it too has a limitation when $D > N$. In particular, when λ is small enough, we will eventually recover $|\hat{S}| = N$. At this point, the matrix we need to invert in the definition of $\mathrm{NS}_{\backslash n}$ in Eq. (5) 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 $\mathrm{NS}_{\backslash n}$ for small λ when $D > N$. Even when D is less than – but still close to – N, we have observed numerical issues in computing $\mathrm{NS}_{\backslash n}$ when λ is sufficiently small; typically, these issues show up as enormously large values for ALOO for small values of λ.

Given the above discussion, we believe that an understanding of the behavior of $\text{IJ}_{\setminus n}$ and $\text{NS}_{\setminus n}$ for the purposes of hyperparameter tuning is a very important direction for future work.

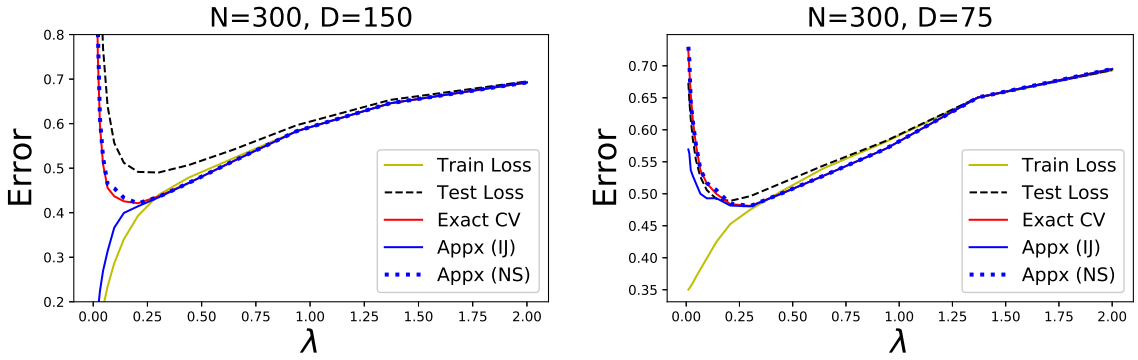

Figure 8: Experiment for selecting λ from Appendix H. (*Top:*) Despite being very accurate for higher values of λ, the degredation of the accuracy of $\text{IJ}_{\setminus n}$ for lower values of λ (which corresponds to a larger $\hat{S}$) causes the selection of a λ that is far from optimal in terms of test loss. (*Bottom:*) For a lower dimensional problem, the curve constructed by $\text{IJ}_{\setminus n}$ much more closely mirrors that of exact CV for all values of λ. In both cases, $\text{NS}_{\setminus n}$ performs well.

I Proofs from Section 4

As mentioned in the main text, there exist somewhat general assumptions in the ℓ_1 literature under which $\operatorname{supp}\hat{\theta} = S$ [Lee et al., 2014, Li et al., 2015]. By taking these assumptions for all leave-one-out problems, we immediately get that $\operatorname{supp}\hat{\theta}_{\setminus n} = S$ 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 ℓ_1 literature for all leave-one-out problems.

I.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 $F = (1/N)\sum_n f(x_n^T\theta, y_n)$ under which $\operatorname{supp}\hat{\theta} = S$. We are interested in $\operatorname{supp}\hat{\theta}_{\setminus n}$, so we state versions of these conditions for $F^{\setminus n} := (1/N)\sum_{m:\, m\neq n} f(x_m^T\theta, y_n)$.

Assumption 10 (LSSC). *$\forall n$, $F^{\setminus n}$ satisfies the $(\theta^*, \mathbb{R}^D)$ locally structured smoothness condition (LSSC)[8] with constant K. We recall this condition, due to Li et al. [2015], in Appendix I.2.*

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

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

$$\left\| \nabla^2 F^{\setminus n}(\theta^*)_{S^c,S} \left(\nabla^2 F^{\setminus n}(\theta^*)_{SS}\right)^{-1} \right\|_\infty < 1 - \gamma. \tag{30}$$

Assumption 13 (Bounded gradient). *For γ from Assumption 12, $\forall n$, the gradient of $F^{\setminus n}$ evaluated at the true parameters θ^* 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_{min} and γ as in Assumptions 10 to 12, the regularization parameter is sufficiently small: $\lambda < L_{min}^2\gamma/(4(\gamma+4)^2 D_{\text{eff}} K)$, where there is no constraint on λ if $K = 0$.*

We see in Appendix I.3 that a minor adaptation of Theorem 5.1 from Li et al. [2015] tells us that Assumptions 10 to 14 imply $\forall n, \operatorname{supp}\hat{\theta}_{\setminus n} \subseteq S$. To prove the accuracy of $\text{NS}_{\setminus n}$ and $\text{IJ}_{\setminus n}$, though, we further need that $\operatorname{supp}\hat{\theta}_{\setminus n} \subseteq \hat{S}$

[8] Readers familiar with the LSSC may see choosing the neighborhood of θ^* 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 I.2 for an explanation.

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 $\operatorname{supp}\hat{\theta}_{\backslash n} = \hat{S} = S$. This will follow from an assumption on the smallest entry of θ^*_S, which we stated as Assumption 4 in the main text. We stated Assumption 4 using the quantity T_{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_{min} and γ from Assumptions 11 and 12,* $\min_{s\in S}|\theta^*_s| > (\sqrt{D_{\text{eff}}}(\gamma+4)/L_{min})\lambda$.

Proposition 5. *If Assumptions 10 to 15 hold, then* $\forall n, \operatorname{supp}\hat{\theta}_{\backslash n} = \hat{S} = S$.

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

I.2 Local structured smoothness condition (LSSC)

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 ℓ_1 regularized linear regression to more general ℓ_1 regularized M-estimators. Essentially, it provides a condition on the smoothness of the third derivatives of the objective $F(\theta)$ near the true sparse θ^*. One can then analyze a second order Taylor expansion of the loss and use the LSSC to show that the remainder in this expansion is not too large. To formalize the LSSC, we need to define the third order derivative of F evaluated along a direction $u \in \mathbb{R}^D$:

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

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

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

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

$$|D^3 f(\theta^* + \delta)[u,u,e_j]| \leq K \left\|u\right\|_2^2, \tag{31}$$

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

We note that this definition is actually 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 $N_{\theta^*} = \mathbb{R}^D$.

I.3 Assumptions 10 to 14 imply $\operatorname{supp}\hat{\theta}_{\backslash n} \subseteq S$ for all n

Theorem 5.1 of Li et al. [2015] gives conditions on F under which $\operatorname{supp}\hat{\theta} = S$. So, if these conditions hold for all $F^{\backslash n}$, then we have $\operatorname{supp}\hat{\theta}_{\backslash n} = S$ 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 $\hat{\theta} \subseteq S$.

I.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 3 (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 $\log E[e^{\lambda Z_n}] \leq \psi(\lambda)$ 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 ψ^{*-1} is the inverse of the Legendre dual of ψ.

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:*

$$\Pr\left[\max_{n=1,\ldots,N} Z_n \geq E[Z_n] + \sqrt{2Cc_x^2 \log N} + u\right] \leq e^{-C\frac{u^2}{2c_x^2}} \tag{32}$$

$$\Pr\left[\max_{n=1,\ldots,N} Z_n^2 \geq E[Z_n^2] + Cc_x^2(\log N + 1 + u)\right] \leq e^{-u} \tag{33}$$

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

$$\psi^{*-1}(\log N + u) = \sqrt{2Cc_x^2(\log N + u)} \leq \sqrt{2Cc_x^2 \log N} + \sqrt{2Cc_x^2 u},$$

Using this upper bound with Lemma 1 and changing variables $u \mapsto u^2/(2Cc_x^2)$ 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 $\log Ee^{\lambda Z_n^2} \leq \psi(\lambda)$, where:

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

For $x \geq 0$, this ψ has inverse Legendre dual $\psi^{*-1}(x) = Cc_x^2(x+1)$. 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 $E[x_{nd}^2] = 1$. Then:*

$$\Pr\left[\max_{n=1,\ldots,N} \|x_n\|_2 \geq \sqrt{D} + \sqrt{2Cc_x^4 \log N} + u\right] \leq e^{-C\frac{u^2}{2c_x^4}}, \tag{34}$$

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. □

I.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 $E[x_{nd}^2] = 1$ and ε_n is c_ε-sub-Gaussian. For notation throughout this section, we will let C denote an absolute constant independent of any aspect of the problem ($N, D, D_{\text{eff}}, c_x$, or c_ε) that will change from line to line (e.g. we may write $5C^2 = C$). We will frequently use $X_{\cdot,S}$ to denote the $N \times D_{\text{eff}}$ matrix formed by taking the columns of X that are in S, x_{nS} to denote the coordinates of the nth vector of covariates x_n that are in the set S, and $X_{\setminus n,S}$ to denote the matrix $X_{\cdot,S}$ with the nth 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 1 to 3, 5 and 15. Suppose the regularization parameter λ satisfies:*

$$\lambda \geq \frac{1}{\alpha - M_{\text{lin}}} \sqrt{\frac{c_x^2 c_\varepsilon^2 \log D}{NC} + \frac{25 c_x^2 c_\varepsilon^2}{NC}} + \frac{4 c_x c_\varepsilon (\log(ND) + 26)}{N(\alpha - M_{\text{lin}})}, \tag{35}$$

where C is a constant in $N, D, D_{\text{eff}}, c_x$ and c_ε, and M_{lin} is defined as:

$$\begin{aligned} M_{\text{lin}} = & \\ & \frac{C D_{\text{eff}} \left(\sqrt{50 c_x^2} + \sqrt{2 c_x^2 \log(N(D - D_{\text{eff}}))}\right) \left(\sqrt{D_{\text{eff}}} + \sqrt{50 c_x^4} + \sqrt{2 c_x^4 \log N}\right)}{N - 3 c_x^2 \sqrt{N} \left(\sqrt{D_{\text{eff}}} + 5\right)} + \\ & \frac{C D_{\text{eff}} \left(D_{\text{eff}} + D_{\text{eff}} c_x^2 (\log N + 26)\right) \left(\sqrt{N} + \sqrt{50 c_x^4} + \sqrt{2 c_x^4 \log(D - D_{\text{eff}})}\right) \left(\sqrt{N D_{\text{eff}}} + \sqrt{50 c_x^4}\right)}{\left(N - 3 c_x^2 \sqrt{N} \left(\sqrt{D_{\text{eff}}} + 5\right)\right)^2} \end{aligned} \tag{36}$$

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 λ 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:

$$J_{nd} := \left(X_{\setminus n,S}^T X_{\setminus n,S}\right)^{-1} X_{\setminus n,S}^T X_{\setminus n,d}. \tag{37}$$

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:

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

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 probablity is equal to zero (as the event reduces to $\infty < \lambda$). To bound the second probability, we have that Lemma 3 says that:

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

As $\alpha > 0$, if $M_{\text{lin}} = o(1)$ as $N \to \infty$, we will we 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_{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_{eff}:

$$\begin{aligned} M_{\text{lin}} &= O\left(\frac{D_{\text{eff}} \sqrt{\log(N) + \log(D)} \left(\sqrt{D_{\text{eff}}} + \sqrt{\log(N)}\right)}{N - \sqrt{N D_{\text{eff}}}} + \frac{D_{\text{eff}}^{5/2} \log(N) \left(\sqrt{N} + \sqrt{\log(D)}\right) \sqrt{N}}{(N - \sqrt{N D_{\text{eff}}})^2}\right) \\ &= O\left(\frac{D_{\text{eff}} \left(\sqrt{D_{\text{eff}} \log(N)} + \sqrt{D_{\text{eff}} \log(D)} + \log(N) + \sqrt{\log(N) \log(D)}\right)}{N - \sqrt{N D_{\text{eff}}}} + \frac{D_{\text{eff}}^{5/2} N \log(N)}{(N - \sqrt{N D_{\text{eff}}})^2}\right), \end{aligned} \tag{38}$$

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. (38) is $o(1)$. The first term is also $o(1)$ by combining $D_{\text{eff}} = o([N/\log(N)]^{2/5})$ with $D_{\text{eff}}^{3/2}\sqrt{\log(D)} = o(N)$. 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.

I.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 $X_{\cdot,S}$ is an $N \times D_{\text{eff}}$ matrix with independent c_x-sub-Gaussian entries with unit second moments, then:*

$$\Pr\left[\lambda_{min}(X_{\cdot,S}^T X_{\cdot,S}) \leq N - 2Cc_x^2\sqrt{N}(\sqrt{D_{\text{eff}}}+5)\right] \leq 2e^{-25}, \tag{39}$$

where $C > 0$ is a global constant.

Proof. Theorem 4.6.1 of Vershynin [2018] gives a concentration inequality for the minimum singular value, $s_{min}(X_{\cdot,S})$, of $X_{\cdot,S}$:

$$\Pr\left[s_{min}(X_{\cdot,S}) \leq \sqrt{N} - Cc_x^2(\sqrt{D_{\text{eff}}}+t)\right] \leq 2e^{-t^2}. \tag{40}$$

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

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

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

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

$$\lambda_{min}(X_{\setminus n,S}^T X_{\setminus n,S}) \geq \lambda_{min}(X_{\cdot,S}^T X_{\cdot,S}) - \|x_{nS}\|_2^2, \tag{41}$$

where x_n is the nth row of $X_{\cdot,S}$.

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

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

□

The above two propositions now allow us to prove the bound we want on $\min_n \lambda_{min}(X_{\setminus n,S}^T X_{\setminus n,S})$. 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 $D_{\text{eff}} = o([N/\log(N)]^{2/5})$ in Assumption 3, the current result can be stated with the less restrictive requirement of $o(N/\log(N))$.

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

$$\Pr\left[\min_{n=1,\ldots N}\lambda_{min}(X_{\setminus n,S}^T X_{\setminus n,S}) \leq N - 3Cc_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}}+5\right)\right] \leq 3e^{-25} \tag{42}$$

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:

$$\begin{aligned}\Pr[A] &= \Pr[A \mid B]\Pr[B] + \Pr[A \mid B^c]\Pr[B^c] \\ &\leq \Pr[A \mid B]Pr[B] + \Pr[B^c],\end{aligned} \tag{43}$$

Calling the probability on the left hand side of Eq. (42) P, we can break P down as, for some constant L_{min}:

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

Picking $L_{min} = N - 2Cc_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}}+5\right)$, 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 $D_{\text{eff}}c_x^2$-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 $E[\|x_{nS}\|_2^2] + Cc_x^2(\log N + 26) = D_{\text{eff}} + Cc_x^2(\log N + 26)$ is less than $Cc_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}}+5\right)$, which, for D_{eff} being $o(N/\log(N))$, is satisfied for N large enough. □

I.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 $z_{\setminus n} \in \mathbb{R}^{N-1}$ the same vector with the nth coordinate removed. Also let $X_{\cdot,S} \in \mathbb{R}^{N \times D_{\text{eff}}}$ be some matrix with $X_{\setminus n,S}$ the same matrix with the nth row removed. Define, for any vector $z \in R^N$:*

$$J_{nz} := \left(X_{\setminus n,S}^T X_{\setminus n,S}\right)^{-1} X_{\setminus n,S}^T z_{\setminus n}, \tag{44}$$

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

$$\begin{aligned}\|J_{nz} - J_z\|_1 \leq& D_{\text{eff}} \frac{|z_n|\,\|x_{nS}\|_2}{\lambda_{min}\left(X_{\setminus n,S}^T X_{\setminus n,S}\right)} \\ &+ D_{\text{eff}} \frac{\|x_{nS}\|_2^2}{\lambda_{min}^2\left(X_{\setminus n,S}^T X_{\setminus n,S}\right)} \|z\|_2 \|X_{\cdot,S}\|_2,\end{aligned}$$

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

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

$$J_z = (X_{\cdot,S}^T X_{\cdot,S})^{-1} X_{\cdot,S}^T z \tag{45}$$

$$= \left((X_{\setminus n,S}^T X_{\setminus n,S})^{-1} - \frac{(X_{\setminus n,S}^T X_{\setminus n,S})^{-1} x_{nS} x_{nS}^T (X_{\setminus n,S}^T X_{\setminus n,S})^{-1}}{1 + x_{nS}^T (X_{\setminus n,S}^T X_{\setminus n,S})^{-1} x_{nS}}\right) X_{\cdot,S}^T z \tag{46}$$

$$= \left(J_{nz} + (X_{\setminus n,S}^T X_{\setminus n,S})^{-1} x_{nS} z_n\right) - \frac{(X_{\setminus n,S}^T X_{\setminus n,S})^{-1} x_{nS} x_{nS}^T (X_{\setminus n,S}^T X_{\setminus n,S})^{-1}}{1 + x_{nS}^T (X_{\setminus n,S}^T X_{\setminus n,S})^{-1} x_{nS}} X_{\cdot,S}^T z \tag{47}$$

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

$$= \left(J_{nz} + B^{-1} x_{nS} z_n\right) - \frac{B^{-1} x_{nS}}{1 + x_{nS}^T B^{-1} x_{nS}} \sum_{m=1}^N z_m x_{nS}^T B^{-1} x_{mS} \tag{48}$$

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 ith unit vector e_i:

$$|e_i^T(J_{nz} - J_z)| \leq |e_i^T B^{-1} x_{nS}||z_n| + \frac{|e_i^T B^{-1} x_{nS}|}{1 + x_{nS}^T B^{-1} x_{nS}} \sum_{m=1}^{N} |z_m||x_{nS}^T B^{-1} x_{mS}| \tag{49}$$

$$\leq |z_n|\lambda_{max}(B^{-1}) \|x_{nS}\|_2 + \frac{\lambda_{max}^2(B^{-1}) \|x_{nS}\|_2^2}{1 + \lambda_{min}(B^{-1}) \|x_{nS}\|_2^2} \sum_{m=1}^{N} |z_m| \|x_{mS}\|_2 \tag{50}$$

The $\lambda_{min}(B^{-1}) \|x_{nS}\|_2^2$ 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 $\lambda_{min}(B^{-1}) = 1/\lambda_{max}(B)$ and $\lambda_{max}(B^{-1}) = 1/\lambda_{min}(B)$, we get:

$$|e_i^T(J_{nz} - J_z)| \leq \frac{|z_n| \|x_{nS}\|_2}{\lambda_{min}(B)} + \frac{\|x_{nS}\|_2^2}{\lambda_{min}^2(B)} \sum_{m=1}^{N} |z_m| \|x_{mS}\|_2 . \tag{51}$$

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

$$\Pr\left[\max_{n=1,\ldots,N} \max_{d \in S^c} \|J_{nd}\|_1 \geq 1 - \alpha + M_{\text{lin}}\right] \leq 10e^{-25}, \tag{52}$$

where J_{nd} is defined in Eq. (37) 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:

$$\left\{\min_n \lambda_{min}(X_{\setminus n,S}^T X_{\setminus n,S}) \geq N - 3Cc_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}} + 5\right)\right\} \tag{53}$$

$$\left\{\|X_{\cdot,S}\|_2 \leq \sqrt{ND_{\text{eff}}} + \sqrt{50Cc_x^4}\right\} \tag{54}$$

$$\left\{\max_n \|x_{nS}\|_2 \leq \sqrt{D_{\text{eff}}} + \sqrt{50Cc_x^4} + \sqrt{2Cc_x^4 \log N}\right\} \tag{55}$$

$$\left\{\max_{d \in S^c} \|X_{\cdot,d}\|_2 \leq \sqrt{N} + \sqrt{50Cc_x^4} + \sqrt{2Cc_x^4 \log(D - D_{\text{eff}})}\right\} \tag{56}$$

$$\left\{\max_n \max_{d \in S^c} |x_{n,d}| \leq \sqrt{50Cc_x^2} + \sqrt{2c_x^2 \log(N(D - D_{\text{eff}}))}\right\} \tag{57}$$

$$\left\{\max_n \|x_{nS}\|_2^2 \leq D_{\text{eff}} + c_x^2 D_{\text{eff}}(\log N + 26)\right\} \tag{58}$$

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 $\|X_{\cdot,S}\|_2 - \sqrt{ND_{\text{eff}}}$ is a Cc_x^2-sub-Gaussian random variable and applying a standard sub-Gaussian bound, the third is $\leq e^{-25}$ by applying Proposition 6, 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 $\|x_{nS}\|_2^2$ is a $c_x^2 D_{\text{eff}}$-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:

$$\begin{aligned}
&\|J_{nd} - J_d\|_1 \leq \\
&\frac{CD_{\text{eff}}\left(\sqrt{50c_x^2} + \sqrt{2c_x^2\log(N(D-D_{\text{eff}}))}\right)\left(\sqrt{D_{\text{eff}}} + \sqrt{50c_x^4} + \sqrt{2c_x^4\log N}\right)}{N - 3c_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}}+5\right)} + \\
&\frac{CD_{\text{eff}}\left(D_{\text{eff}} + D_{\text{eff}}c_x^2(\log N + 26)\right)\left(\sqrt{N} + \sqrt{50c_x^4} + \sqrt{2c_x^4\log(D-D_{\text{eff}})}\right)\left(\sqrt{ND_{\text{eff}}} + \sqrt{50c_x^4}\right)}{\left(N - 3c_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}}+5\right)\right)^2}
\end{aligned}$$

Call the entire quantity on the right-hand side of this inequality M_{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:

$$\Pr\left[\max_{n\in[N]}\max_{d\in S^c}\|J_{nd}\|_1 \geq 1-\alpha+M_{\text{lin}}\right] \leq \tag{59}$$

$$\Pr\left[\max_{n}\max_{d\in S^c}\|J_{nd}-J_n\|_1 \geq M_{\text{lin}} \mid F\right] + \Pr\left[F^c\right] + \Pr\left[\max_{d\in S^c}\|J_d\|_1 \geq 1-\alpha\right] \tag{60}$$

By the definition of M_{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. □

I.8 Linear regression: bounded gradient

We need to bound the probability

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

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:

$$\begin{aligned}
\leq &\Pr\left[\max_{n=1,\ldots,N}\left\|\frac{1}{N}\nabla f(x_n^T\theta^*, y_n)\right\|_\infty \geq \frac{\lambda(\alpha - M_{\text{lin}})}{4} - B_G\right] \\
&+ \Pr\left[\|\nabla F(\theta^*)\|_\infty \geq B_G\right] + \Pr\left[\max_n\max_{d\in S^c}\|J_{nd}\|_1 \geq 1-\alpha+M_{\text{lin}}\right]
\end{aligned} \tag{61}$$

The following proposition gives a reasonable value for B_G:

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

$$\Pr\left[\|\nabla F(\theta^*)\|_\infty \geq \left[\frac{c_x^2c_\varepsilon^2\log D}{NC} + \frac{25c_x^2c_\varepsilon^2}{NC}\right]^{1/2}\right] \leq e^{-25} \tag{62}$$

Proof. The dth coordinate of the gradient is $(\nabla F(\theta^*))_d = 1/N\sum_n \varepsilon_n x_{nd}$. First, we have that $1/N\sum_n \varepsilon_n x_{nd}$ is a c_xc_ε-sub-Exponential random variable. By Bernstein's inequality (see Theorem 2.8.1 from Vershynin [2018]), we have:

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

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

Now we can prove the lemma we need, which bounds the probability that any $\left\|\nabla F^{\setminus n}(\theta^*)\right\|_\infty$ is large:

Lemma 4. *For the above setup for linear regression and the λ given in Theorem 2, we have:*

$$\Pr\left[\max_{n=1,\ldots,N}\left\|\nabla F^{\setminus n}(\theta^*)\right\|_\infty \geq \frac{\lambda\left(1-\max_n \max_{d\in S^c}\|J_{nd}\|_1\right)}{4}\right] \leq 13e^{-25} \tag{63}$$

Proof. We can first apply the bound worked out in Eq. (61). 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 3. To analyze the first probability, note that we can write the event as:

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

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

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

The event we're considering is just the absolute value of the max of ND sub-Exponential variables with parameter $c_x c_\varepsilon$. Plugging into Corollary 1 gives that this probability is $\leq 2e^{-25}$. □

I.9 Linear regression: λ 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 4, the loss $F(\theta)$ satisfies the (θ^*, N_{θ^*}) LSSC with constant $K = 0$ for any θ^*, N_{θ^*}, and any data X, Y.*

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

As linear regression has a LSSC constant K that is deterministically equal to zero, the only constraint implied by the bound in Assumption 14 is that $\lambda < \infty$, which is always satisfied by the value of λ given in Theorem 4.

I.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:

$$D_n^{(1)} := \frac{-y_n}{1 + e^{y_n x_n^T\theta^*}}, \quad D_n^{(2)} := \frac{e^{x_n^T\theta^*}}{(1 + e^{x_n^T\theta^*})^2}, \tag{64}$$

the derivatives of F are:

$$\nabla_\theta F(\theta^*) = \frac{1}{N}\sum_{n=1}^N D_n^{(1)} x_n, \quad \nabla_\theta^2 F(\theta^*) = \frac{1}{N}\sum_{n=1}^N D_n^{(2)} x_n x_n^T. \tag{65}$$

For comparison, things were easier for linear regression because $D_n^{(2)} = 1$ and $D_n^{(1)} = \varepsilon_n$ for some sub-Gaussian noise ε_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 ≤ 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:*

$$\lambda \geq \frac{C}{\alpha - M_{\text{logr}}}\left(\sqrt{c_x^2\frac{25 + \log D}{N}} + \frac{\sqrt{2c_x^2\log(ND)} + \sqrt{50c_x^2}}{N}\right), \tag{66}$$

where C is a constant in N, D, and c_x, and M_{logr} is defined similarly to M_{lin} from Theorem 4, but with different denominators:

$$M_{\text{logr}} = \frac{CD_{\text{eff}}\left(\sqrt{50c_x^2} + \sqrt{2c_x^2\log(N(D-D_{\text{eff}}))}\right)\left(\sqrt{D_{\text{eff}}} + \sqrt{50c_x^4} + \sqrt{2c_x^4\log N}\right)}{L_{min} - c_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}}+5\right)} + \frac{CD_{\text{eff}}\left(D_{\text{eff}} + D_{\text{eff}}c_x^2(\log N + 26)\right)\left(\sqrt{N} + \sqrt{50c_x^4} + \sqrt{2c_x^4\log(D-D_{\text{eff}})}\right)\left(\sqrt{ND_{\text{eff}}} + \sqrt{50c_x^4}\right)}{\left(L_{min} - c_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}}+5\right)\right)^2} \tag{67}$$

Then for N sufficiently large, Condition 1 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 4 – 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 $M_{\text{logr}} = o(1)$ by Assumptions 3 and 7. □

I.11 Logistic regression: lambda min

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

$$\Pr\left[\min_{n=1,\ldots,N} \lambda_{min}(\nabla_\theta^2 F^{\setminus n}(\theta^*)_{SS}) \leq L_{min} - Cc_x^2\sqrt{N}\left(\sqrt{D_{\text{eff}}}+5\right)\right] \leq 3e^{-25}, \tag{68}$$

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

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

$$\begin{aligned}\lambda_{min}(\nabla_\theta^2 F^{\setminus n}(\theta^*)_{SS}) &\geq \lambda_{min}(\nabla_\theta^2 F(\theta^*)_{SS}) - \|x_{nS}\|_2^2 |D_n^{(2)}| \\ &\geq \lambda_{min}(\nabla_\theta^2 F(\theta^*)_{SS}) - \|x_{nS}\|_2^2 .\end{aligned}$$

The rest of the proof is now exactly the same as that of Lemma 2. □

I.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^TX; 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_{logr} defined in Theorem 3, we have:*

$$\Pr\left[\max_{n=1,\ldots,N}\max_{d\in S^c}\|J_{nd}\|_1 \geq 1 - \alpha + M_{\text{logr}}\right] \leq 10e^{-25}, \tag{69}$$

where J_{nd} is defined in Eq. (37).

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 $\|J_d\|_1$ 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. (53) by:

$$\left\{\min_{n=1,\ldots,N}\lambda_{min}\left(\nabla_\theta^2 F_{SS}^{\setminus n}\right) \geq L_{min} - Cc_x^2\sqrt{N}(\sqrt{D_{\text{eff}}}+5)\right\}. \tag{70}$$

By Lemma 5, the complement of this event has probability at most $3e^{-25}$. We condition on the rest of the events in the proof of Lemma 3 and finish the proof along the same lines. □

I.13 Logistic regression: bounded gradient

Again, we are interested in bounding:

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

The same reasoning that led to Eq. (61) gives us the same bound:

$$\begin{aligned}\leq \Pr\left[\max_{n=1,\ldots,N}\left\|\frac{1}{N}\nabla_\theta f(x_n^T\theta^*, y_n)\right\|_\infty \geq \frac{\lambda(\alpha - M_{\text{logr}})}{4} - B_G\right] \\ + \Pr\left[\|\nabla F(\theta^*)\|_\infty \geq B_G\right] + \Pr\left[\max_n \max_{d\in S^c}\|J_{nd}\|_1 \geq 1-\alpha+M_{\text{logr}}\right]\end{aligned} \tag{71}$$

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:*

$$\Pr\left[\|\nabla F(\theta^*)\|_\infty \geq c_x\sqrt{\frac{25+\log D}{CN}}\right] \leq 2e^{-25}. \tag{72}$$

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

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

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

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

where the final inequality comes from noting that $|x_{nd}|$ is also c_x-sub-Gaussian and using Hoeffding's inequality (Theorem 2.6.2 from Vershynin [2018]). Union bounding over all D dimensions of $\nabla F(\theta^*)$ gives the result. □

Lemma 7. *For the above setup for logistic regression and the λ given in Theorem 3, we have:*

$$\Pr\left[\max_{n\in[N]}\left\|\nabla F^{\setminus n}(\theta^*)\right\|_\infty \geq \frac{\lambda\left(1-\max_n \max_{d\in S^c}\|J_{nd}\|_1\right)}{4}\right] \leq 14e^{-25} \tag{73}$$

Proof. Just as in the proof of Lemma 4, we will apply the upper bound in Eq. (71) and then bound each term. The second probability in Eq. (71) 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:

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

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

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

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

I.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 (θ^*, N_{θ^*}) LSSC for any θ^* and $N_{\theta^*} = \mathbb{R}^D$ 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:*

$$\Pr\left[K \geq \frac{1}{4}\left(\sqrt{2c_x^2\log(ND)} + \sqrt{50c_x^2}\right)\left(D_{\text{eff}} + c_x^2 D_{\text{eff}}(\log N + 26)\right)\right] \leq 3e^{-25} \tag{74}$$

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

$$\leq \Pr\left[\max_n \|x_{nS}\|_2^2 \geq D_{\text{eff}} + c_x^2 D_{\text{eff}}(\log N + 26)\right] + 2e^{-25}. \tag{75}$$

Noting that $\|x_{nS}\|_2^2$ is the sum of D_{eff} c_x^2-sub-Exponential random variables, $\|x_{nS}\|_2^2$ is a $D_{\text{eff}}c_x^2$-sub-Exponential random variable. Corollary 1 then gives us that Eq. (75) 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:*

$$\Pr\left[\lambda \geq \frac{\min_n \lambda_{min}^2(\nabla_\theta^2 F^{\backslash n}(\theta^*)_{SS})}{4\left(\left(1 - \max_n \max_{d\in S^c}\|J_{nd}\|_1\right) + 4\right)^2}\frac{4\left(1 - \max_{d\in S^c}\|J_{nd}\|_1\right)}{K}\right] \leq 16e^{-25} \tag{76}$$

Proof. Using Lemma 5, Lemma 6, and Proposition 14, the desired probability is $\leq 16e^{-25}$ if the following deterministic inequality holds:

$$\lambda \leq \frac{4(\alpha - M_{\text{logr}})}{4(\alpha - M_{\text{logr}} + 4)^2}\frac{(L_{min} - Cc_x^2\sqrt{D_{\text{eff}}N})^2}{\left(\sqrt{2c_x^2\log(ND)} + \sqrt{50c_x^2}\right)(D_{\text{eff}} + c_x^2 D_{\text{eff}}(\log N + 26))} \tag{77}$$

We will lower bound the right hand side and show that λ 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, $M_{\text{logr}} = o(1)$ as $N \to \infty$, so that for large enough N, we have $(\alpha - M_{\text{logr}})/(\alpha - M_{\text{logr}} + 4)^2 \geq (\alpha/2)/(\alpha/2 + 4)^2$. 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:

$$\lambda \leq C\frac{\alpha/2}{(\alpha/2+4)^2}\frac{N^2}{D_{\text{eff}}\log N\sqrt{\log(ND)}}. \tag{78}$$

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

$$\lambda \leq C\frac{\alpha/2}{(\alpha/2+4)^2}\frac{\log^{2/5}(N)N^2}{N^{2/5}\log N\sqrt{\log(N)+N}}. \tag{79}$$

Now, the right hand side goes to infinity as N gets large, while the λ given in Theorem 3 goes to 0 as N gets large. Thus, for sufficiently large N, Eq. (77) holds. □