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.

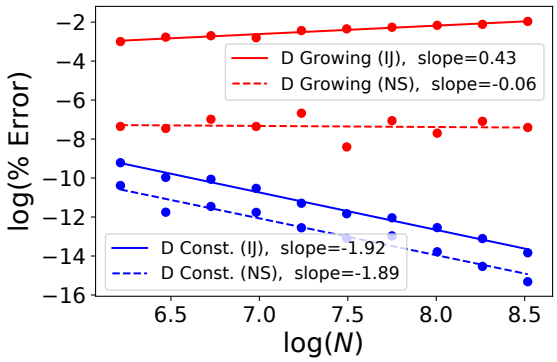

Figure 1: Log percent error (Eq. (10)) of existing approximate LOOCV methods ("IJ" and "NS") as a function of dataset size N for ℓ_2 regularized logistic regression. Dashed lines show Eq. (2) ("NS") and solid show Eq. (3) ("IJ"). Blue lines have fixed data/parameter dimension D, while red lines have $D = N/10$, 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/

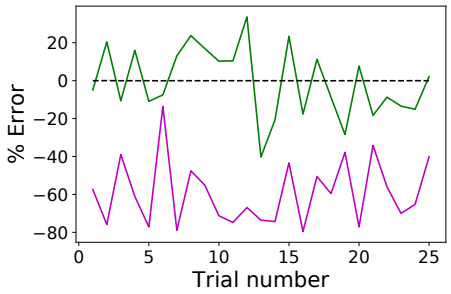

Figure 2: Error (Eq. (10)) across approximations for ℓ_1 LOOCV (legend shared with Fig. 3). The error for $\text{IJ}_{\setminus n}$ (black dashed) is too small to see, but nonzero; it varies between -0.06% and 0.04%.

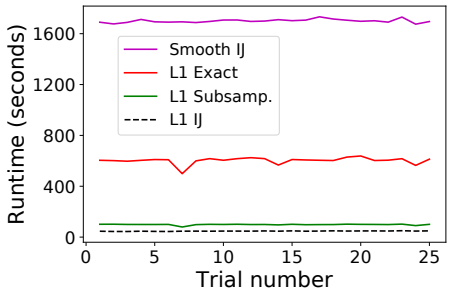

Figure 3: Runtimes for the experiments in Fig. 2 with exact CV (red) included for comparison. The $D \times D$ matrix inversion in the smoothed problem is so slow that even exact CV with an efficient ℓ_1 solver is faster.

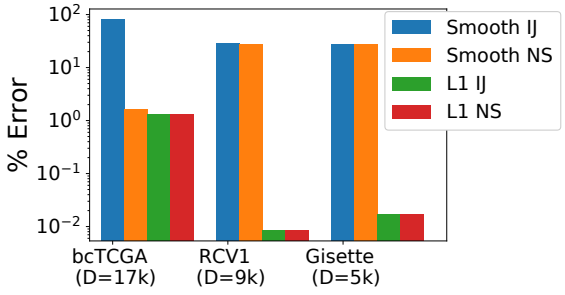

Figure 4: Log percent accuracy (Eq. (10)) for real data experiments. For each dataset, we give the accuracy of approximate CV compared to exact CV using both a smoothed approximation to ℓ_1 and the $\text{IJ}_{\setminus n}$, $\text{NS}_{\setminus n}$ approximations. For the bcTCGA dataset (linear regression), 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.

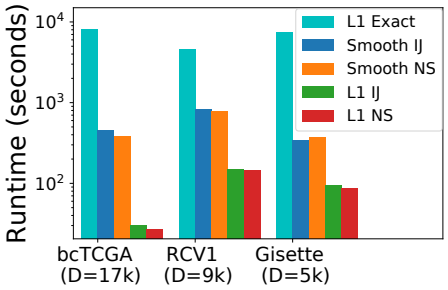

Figure 5: Log runtimes for experiments in Fig. 4, with exact CV included for comparison.

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.