Approximate Leave-One-Out for Fast Parameter Tuning in High Dimensions

Shuaiwen Wang [*1] **Wenda Zhou** [*1] **Haihao Lu** [2] **Arian Maleki** [1] **Vahab Mirrokni** [3]

Abstract

Consider the following class of leaning schemes:

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \ \sum_{j=1}^{n} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + \lambda R(\boldsymbol{\beta}), \quad (1)$$

where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ denote the i^{th} feature and response variable respectively. Let ℓ and R be the loss function and regularizer, $\boldsymbol{\beta}$ denote the unknown weights, and λ be a regularization parameter. Finding the optimal choice of λ is a challenging problem in high-dimensional regimes where both n and p are large. We propose two frameworks to obtain a computationally efficient approximation ALO of the leave-one-out cross validation (LOOCV) risk for nonsmooth losses and regularizers. Our two frameworks are based on the primal and dual formulations of (1). We prove the equivalence of the two approaches under smoothness conditions. This equivalence enables us to justify the accuracy of both methods under such conditions. We use our approaches to obtain a risk estimate for several standard problems, including generalized LASSO, nuclear norm regularization, and support vector machines. We empirically demonstrate the effectiveness of our results for non-differentiable cases.

1. Introduction

1.1. Motivation

Consider a standard prediction problem in which a dataset $\{(y_j, \boldsymbol{x}_j)\}_{j=1}^n \subset \mathbb{R} \times \mathbb{R}^p$ is employed to learn a model for inferring information about new datapoints that are yet to be observed. One of the most popular classes of learning schemes, especially in high-dimensional settings, studies the following optimization problem:

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \ \sum_{j=1}^{n} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + \lambda R(\boldsymbol{\beta}), \quad (2)$$

where $\ell : \mathbb{R}^2 \to \mathbb{R}$ is the loss function, $R : \mathbb{R}^p \to \mathbb{R}$ is the regularizer, and λ is the tuning parameter that specifies the amount of regularization. By applying an appropriate regularizer in (2), we are able to achieve better bias-variance trade-off and pursue special structures such as sparsity and low rank structure. However, the performance of such techniques hinges upon the selection of tuning parameters.

The most generally applicable tuning method is cross validation (Stone, 1974). One common choice is k-fold cross validation, which however presents potential bias issues in high-dimensional settings where n is comparable to p. For instance, the phase transition phenomena that happen in such regimes (Amelunxen et al., 2014; Donoho et al., 2009; Donoho & Tanner, 2005) indicate that any data splitting may cause dramatic effects on the solution of (2) (see Figure 1 for an example). Hence, the risk estimates obtained from k-fold cross validation may not be reliable. The bias issues of k-fold cross validation may be alleviated by choosing the number of folds k to be large. However, such schemes are computationally demanding and may not be useful for emerging high-dimensional applications. An alternative choice of cross validation is LOOCV, which is unbiased in high-dimensional problems. However, the computation of LOOCV requires training the model n times, which is unaffordable for large datasets.

The high computational complexity of LOOCV has motivated researchers to propose computationally less demanding approximations of the quantity. Early examples offered approximations for the case $R(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2$ and the loss function being smooth (Allen, 1974; O'sullivan et al., 1986; Le Cessie & Van Houwelingen, 1992; Cawley & Talbot, 2008; Meijer & Goeman, 2013; Opper & Winther, 2000). In (Beirami et al., 2017), the authors considered such approximations for smooth loss functions and smooth regularizers. In this line of work, the accuracy of the approximations was either not studied or was only studied in the n large, p fixed regime. In a recent paper, (Rad & Maleki, 2018) employed a similar approximation strategy to obtain approx-

[*]Equal contribution [1]Department of Statistics, Columbia University, New York, USA [2]Mathematics Department and Operation Research Center, Massachusetts Institute of Technology, Massachusetts, USA [3]Google Research, New York, USA. Correspondence to: Shuaiwen Wang <sw2853@columbia.edu>, Wenda Zhou <wz2335@columbia.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

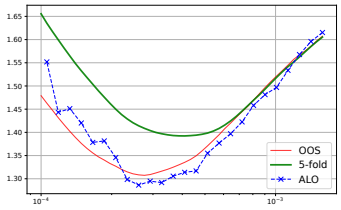

Figure 1. Risk estimates of LASSO based on 5-fold CV and ALO proposed in this paper, compared with the true out-of-sample prediction error (OOS). In this example, 5-fold CV exhibits significant bias, wherease ALO is unbiased. Here we use $n = 5000$, $p = 4000$ and iid Gaussian design.

imate leave-one-out formulas for smooth loss functions and smooth regularizers. They show that under some mild conditions, such approximations are accurate in high-dimensional settings. Unfortunately, the approximations offered in (Rad & Maleki, 2018) only cover twice differentiable loss functions and regularizers. On the other hand, numerous modern regularizers, such as generalized LASSO and nuclear norm, and also many loss functions are not smooth.

In this paper, we propose two powerful frameworks for calculating an approximate leave-one-out estimator (ALO) of the LOOCV risk that are capable of offering accurate parameter tuning even for non-differentiable losses and regularizers. Our first approach is based on the smoothing and quadratic approximation of the primal problem (2). The second approach is based on the approximation of the dual of (2). While the two approaches consider different approximations that happen in different domains, we will show that when both ℓ and r are twice differentiable, the two frameworks produce the same ALO formulas, which are also the same as the formulas proposed in (Rad & Maleki, 2018).

We use our platforms to obtain concise formulas for several popular examples including generalized LASSO, support vector machine (SVM) and nuclear norm minimization. As will be clear from our examples, despite of the equivalence of the two frameworks for smooth loss functions and regularizers, the technical aspects of the derivations involved for obtaining ALO formulas have major variations in different examples. Finally, we present extensive simulations to confirm the accuracy of our formulas on various important machine learning models. Code is available at `github.com/wendazhou/alocv-package`.

1.2. Other Related Work

The importance of parameter tuning in learning systems has encouraged many researchers to study the problem from different perspectives. In addition to cross validation, other approaches have been proposed including Stein's unbiased risk estimate (SURE), Akaike information criterion (AIC), and Mallow's C_p. While AIC is designed for smooth parametric models, SURE has been extended to emerging optimization problems, such as generalized LASSO and nuclear norm minimization (Candes et al., 2013; Dossal et al., 2013; Tibshirani et al., 2012; Vaiter et al., 2017; Zou et al., 2007).

Unlike cross validation which approximates the out-of-sample prediction error, SURE, AIC, and C_p offer estimates for in-sample prediction error (Hastie et al., 2009). This makes cross validation more appealing for many learning systems. Furthermore, unlike ALO, both SURE and C_p only work on linear models (and not generalized linear models) and their unbiasedness is only guaranteed under the Gaussian model for the errors. There has been little success in extending SURE beyond this model (Efron, 2004).

Another class of parameter tuning schemes are based on approximate message passing (Bayati et al., 2013; Mousavi et al., 2017; Obuchi & Kabashima, 2016). As pointed out in (Obuchi & Kabashima, 2016), this approach is intuitively related to LOOCV. It offers consistent parameter tuning in high-dimensions (Mousavi et al., 2017), but the results strongly depend on the independence of the elements of $\boldsymbol{X}$.

1.3. Notation

Lowercase and uppercase bold letters denote vectors and matrices, respectively. For subsets $A \subset \{1, 2, \dots, n\}$ and $B \subset \{1, 2, \dots, p\}$ of indices and a matrix $\boldsymbol{X}$, let $\boldsymbol{X}_{A,\cdot}$ and $\boldsymbol{X}_{\cdot,\boldsymbol{B}}$ denote the submatrices that include only rows of $\boldsymbol{X}$ in A, and columns of $\boldsymbol{X}$ in B respectively. Let $\{a_i\}_{i\in S}$ denote the vector whose components are a_i for $i \in S$. We may omit S, in which case we consider all indices valid in the context. For a function $f : \mathbb{R} \to \mathbb{R}$, let $\dot{f}$, $\ddot{f}$ denote its 1st and 2nd derivatives. For a vector $\boldsymbol{a}$, we use $\mathrm{diag}[\boldsymbol{a}]$ to denote a diagonal matrix $\boldsymbol{A}$ with $A_{ii} = a_i$. Finally, let ∇R and $\nabla^2 R$ denote the gradient and Hessian of a function $R : \mathbb{R}^p \to \mathbb{R}$.

2. Preliminaries

2.1. Problem Description

In this paper, we study the statistical learning models in form (2). For each value of λ, we evaluate the following LOOCV risk estimate with respect to some error function d:

$$\mathrm{loo}_\lambda := \frac{1}{n}\sum_{i=1}^{n} d(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{/i}), \tag{3}$$

where $\hat{\boldsymbol{\beta}}^{/i}$ is the solution of the leave-i-out problem

$$\hat{\boldsymbol{\beta}}^{/i} := \arg\min_{\boldsymbol{\beta}} \sum_{j\neq i} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + \lambda R(\boldsymbol{\beta}). \tag{4}$$

Calculating (4) requires training the model n times, which may be time-consuming in high-dimensions. As an alternative, we propose an estimator $\tilde{\boldsymbol{\beta}}^{/i}$ to approximate $\hat{\boldsymbol{\beta}}^{/i}$ based

on the full-data estimator $\hat{\boldsymbol{\beta}}$ to reduce the computational complexity. We consider two frameworks for obtaining $\tilde{\boldsymbol{\beta}}^{/i}$, and denote the corresponding risk estimate by:

$$\mathrm{alo}_\lambda := \frac{1}{n}\sum_{i=1}^{n} d(y_i, \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{/i}). \tag{5}$$

The estimates we obtain will be called approximated leave-one-out (ALO) throughout the paper.

2.2. Primal and Dual Correspondence

The objective function of penalized regression problem with loss ℓ and regularizer R is given by:

$$P(\boldsymbol{\beta}) := \sum_{j=1}^{n} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + R(\boldsymbol{\beta}). \tag{6}$$

Here and subsequently, we absorb the value of λ into R to simplify the notation. We also consider the Lagrangian dual problem, which can be written in the form:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n} D(\boldsymbol{\theta}) := \sum_{j=1}^{n} \ell^*(-\theta_j; y_j) + R^*(\boldsymbol{X}^\top \boldsymbol{\theta}), \tag{7}$$

where ℓ^* and R^* denote the *Fenchel conjugates* [1] of ℓ and R respectively. See the derivation in Appendix A.

It is known that under mild conditions, (6) and (7) are equivalent (Boyd & Vandenberghe, 2004). In this case, we have the primal-dual correspondence relating the primal optimal $\hat{\boldsymbol{\beta}}$ and the dual optimal $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned} \hat{\boldsymbol{\beta}} \in \partial R^*(\boldsymbol{X}^\top \hat{\boldsymbol{\theta}}), \quad & \boldsymbol{X}^\top \hat{\boldsymbol{\theta}} \in \partial R(\hat{\boldsymbol{\beta}}), \\ \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} \in \partial \ell^*(-\hat{\theta}_j; y_j), \quad & -\hat{\theta}_j \in \partial \ell(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j), \end{aligned} \tag{8}$$

where ∂f denotes the set of subgradients of a function f. Below we will use both primal and dual perspectives for approximating loo_λ.

3. Approximation in the Dual Domain

3.1. The First Example: LASSO

Let us first start with a simple example that illustrates our dual method in deriving an approximate leave-one-out (ALO) formula for the standard LASSO. The LASSO estimator, first proposed in (Tibshirani, 1996), can be formulated as the penalized regression framework in (6) by setting $\ell(\mu; y) = (\mu - y)^2/2$, and $R(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$.

We recall the general formulation of the dual for penalized regression problems (7), and note that in the case of the LASSO we have:

$$\ell^*(\theta_i; y_i) = \frac{1}{2}(\theta_i - y_i)^2, \; R^*(\boldsymbol{\beta}) = \begin{cases} 0 & \text{if } \|\boldsymbol{\beta}\|_\infty \le \lambda, \\ +\infty & \text{otherwise.} \end{cases}$$

In particular, we note that the solution of the dual problem (7) can be obtained from:

$$\hat{\boldsymbol{\theta}} = \Pi_{\Delta_X}(\boldsymbol{y}). \tag{9}$$

Here Π_{Δ_X} denotes the projection onto Δ_X, where Δ_X is the polytope given by:

$$\Delta_X = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{X}^\top \boldsymbol{\theta}\|_\infty \le \lambda\}.$$

Let us now consider the leave-i-out problem. Unfortunately, the dimension of the dual problem is reduced by 1 for the leave-i-out problem, making it difficult to leverage the information from the full-data solution to approximate the leave-i-out solution. We augment the leave-i-out problem with a virtual i^{th} observation that does not affect the result of the optimization, but restores the dimensionality of the problem.

More precisely, let $\boldsymbol{y}_a$ be the same as $\boldsymbol{y}$, except that its i^{th} coordinate is replaced by $\hat{y}_i^{/i} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{/i}$, the leave-$i$-out predicted value. We note that the leave-i-out solution $\hat{\boldsymbol{\beta}}^{/i}$ is also the solution for the following augmented problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{j=1}^{n} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_{a,j}) + R(\boldsymbol{\beta}). \tag{10}$$

Let $\hat{\boldsymbol{\theta}}^{/i}$ be the corresponding dual solution of (10). Then, by (9), we know that

$$\hat{\boldsymbol{\theta}}^{/i} = \Pi_{\Delta_X}(\boldsymbol{y}_a).$$

Additionally, the primal-dual correspondence (8) gives that $\hat{\boldsymbol{\theta}}^{/i} = \boldsymbol{y}_a - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{/i}$, which is the residual in the augmented problem, and hence that $\hat{\theta}_i^{/i} = 0$. These two features allow us to characterize the leave-i-out predicted value $\hat{y}_i^{/i}$:

$$\boldsymbol{e}_i^\top \Pi_{\Delta_X}\left(\boldsymbol{y} - (y_i - \hat{y}_i^{/i})\boldsymbol{e}_i\right) = 0 \tag{11}$$

where $\boldsymbol{e}_i$ denotes the i^{th} standard vector. Solving exactly for the above equation is in general a procedure that is computationally comparable to fitting the model, which may be expensive. However, we may attempt to obtain an approximate solution of (11) by linearizing the projection operator at the full data solution $\hat{\boldsymbol{\theta}}$, or equivalently performing a single Newton step to solve the leave-i-out problem from the full data solution. The approximate leave-i-out fitted value $\tilde{y}_i^{/i}$ is thus given by:

$$\tilde{y}_i^{/i} = y_i - \frac{\hat{\theta}_i}{J_{ii}}, \tag{12}$$

[1] The Fenchel conjugate f^* of a function f is defined as $f^*(x) := \sup_y\{\langle x, y\rangle - f(y)\}$.

where $\boldsymbol{J}$ denotes the Jacobian of the projection operator Π_{Δ_X} at the full data problem $\boldsymbol{y}$. We can substitute the $\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{/i}$ in (5) with the $\tilde{y}^{/i}$ found above to obtain our ALO risk estimates. Note that Δ_X is a polytope, and thus the projection onto Δ_X is almost everywhere locally affine (Tibshirani et al., 2012). Furthermore, it is straightforward to calculate the Jacobian of Π_{Δ_X}. Let $E = \{j : |\boldsymbol{X}_j^\top \hat{\boldsymbol{\theta}}| = \lambda\}$ be the equicorrelation set (where $\boldsymbol{X}_j$ denotes the j^{th} column of $\boldsymbol{X}$). Then the projection at the full data problem $\boldsymbol{y}$ is locally given by a projection onto the orthogonal complement of the span of $\boldsymbol{X}_{\cdot,E}$, thus giving $\boldsymbol{J} = \boldsymbol{I} - \boldsymbol{X}_{\cdot,E}(\boldsymbol{X}_{\cdot,E}^\top \boldsymbol{X}_{\cdot,E})^{-1}\boldsymbol{X}_{\cdot,E}^\top$. We can then obtain $\tilde{y}^{/i}$ by plugging $\boldsymbol{J}$ in (12). Finally, by replaceing $\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{/i}$ with $\tilde{y}_i^{/i}$ in (5) we obtain an estimate of the risk.

3.2. General Case

In this section we extend the dual approach outlined in Section 3.1 to more general loss functions and regularizers.

General regularizers Let us first extend the dual approach to other regularizers, while the loss function remains $\ell(\mu, y) = (\mu - y)^2/2$. In this case the dual problem (7) has the following form:

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\sum_{j=1}^{n}(\theta_j - y_j)^2 + R^*(\boldsymbol{X}^\top \boldsymbol{\theta}). \tag{13}$$

We note that the optimal value of $\boldsymbol{\theta}$ is by definition the value of the proximal operator of $R^*(\boldsymbol{X}^\top \cdot)$ at $\boldsymbol{y}$:

$$\hat{\boldsymbol{\theta}} = \mathbf{prox}_{R^*(\boldsymbol{X}^\top \cdot)}(\boldsymbol{y}).$$

Following the argument of Section 3.1, we obtain

$$\tilde{y}_i^{/i} = y_i - \frac{\hat{\theta}_i}{J_{ii}}, \tag{14}$$

with $\boldsymbol{J}$ now denoting the Jacobian of $\mathbf{prox}_{R^*(\boldsymbol{X}^\top \cdot)}$. We note that the Jacobian matrix $\boldsymbol{J}$ exists almost everywhere, because the non-expansiveness of the proximal operator guarantees its almost-everywhere differentiability (Combettes & Pesquet, 2011). In particular, if $\boldsymbol{y}$ has distribution which is absolutely continuous with respect to the Lebesgue measure, $\boldsymbol{J}$ exists with probability 1. This approach is particularly useful when R is a norm, as its Fenchel conjugate is then the convex indicator of the unit ball of the dual norm, and the proximal operator reduces to a projection operator.

General smooth loss Let us now assume we have a convex smooth loss in (6), such as those that appear in generalized linear models. As we are arguing from a second-order perspective by considering Newton's method, we will expand the loss as a quadratic form around the full data solution. We will thus consider the approximate problem obtained by expanding ℓ^* around the dual optimal $\hat{\boldsymbol{\theta}}$:

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\sum_{j=1}^{n}\ddot{\ell}^*(-\hat{\theta}_j; y_j)\left(\theta_j - \hat{\theta}_j - \frac{\dot{\ell}^*(-\hat{\theta}_j; y_j)}{\ddot{\ell}^*(-\hat{\theta}_j; y_j)}\right)^2 + R^*(\boldsymbol{X}^\top \boldsymbol{\theta}). \tag{15}$$

The constant term has been removed from (15) for simplicity. We have reduced the problem to that of a weighted ℓ_2 loss which may be further reduced to a simple ℓ_2 problem by a change of variable and a rescaling of $\boldsymbol{X}$. Indeed, let $\boldsymbol{K}$ be the diagonal matrix such that $K_{jj} = \sqrt{\ddot{\ell}^*(-\hat{\theta}_j; y_j)}$, and note that we have: $\dot{\ell}^*(-\hat{\theta}_j; y_j) = \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} := \hat{y}_j$ by the primal-dual correspondence (8). Consider the change of variable $\boldsymbol{u} = \boldsymbol{K}\boldsymbol{\theta}$ to obtain:

$$\min_{\boldsymbol{u}} \frac{1}{2}\sum_{j=1}^{n}\left(u_j - \frac{\hat{\theta}_j \ddot{\ell}^*(-\hat{\theta}_j; y_j) + \hat{y}_j}{\sqrt{\ddot{\ell}^*(-\hat{\theta}_j; y_j)}}\right)^2 + R^*(\boldsymbol{X}^\top \boldsymbol{K}^{-1}\boldsymbol{u}).$$

We may thus reduce to the ℓ_2 loss case in (13) with a modified $\boldsymbol{X}$ and $\boldsymbol{y}$:

$$\boldsymbol{X}_u = \boldsymbol{K}^{-1}\boldsymbol{X}, \quad \boldsymbol{y}_u = \left\{\frac{\hat{\theta}_j \ddot{\ell}^*(-\hat{\theta}_j; y_j) + \hat{y}_j}{\sqrt{\ddot{\ell}^*(-\hat{\theta}_j; y_j)}}\right\}_j. \tag{16}$$

Similar to (14), the ALO formula in the case of general smooth loss can be obtained as $\tilde{y}_i^{/i} = K_{ii}\tilde{y}_{u,i}^{/i}$, with

$$\tilde{y}_{u,i}^{/i} = y_{u,i} - \frac{K_{ii}\hat{\theta}_i}{J_{ii}}, \tag{17}$$

where $\boldsymbol{J}$ is the Jacobian of $\mathbf{prox}_{R^*(\boldsymbol{X}_u^\top \cdot)}$.

4. Approximation in the Primal Domain

4.1. Smooth Loss and Regularizer

To obtain loo_λ we need to solve

$$\hat{\boldsymbol{\beta}}^{/i} := \arg\min_{\boldsymbol{\beta}} \sum_{j \neq i} \ell(\boldsymbol{x}_j^\top \boldsymbol{\beta}; y_j) + R(\boldsymbol{\beta}). \tag{18}$$

Assuming $\hat{\boldsymbol{\beta}}^{/i}$ is close to $\hat{\boldsymbol{\beta}}$, we can take a *Newton step* from $\hat{\boldsymbol{\beta}}$ towards $\hat{\boldsymbol{\beta}}^{/i}$ to obtain its approximation $\tilde{\boldsymbol{\beta}}^{/i}$ as:

$$\tilde{\boldsymbol{\beta}}^{/i} = \hat{\boldsymbol{\beta}} + \Big[\sum_{j \neq i} \boldsymbol{x}_j \boldsymbol{x}_j^\top \ddot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j) + \nabla^2 R(\hat{\boldsymbol{\beta}})\Big]^{-1} \boldsymbol{x}_i \dot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i). \tag{19}$$

We have by the matrix inversion lemma (Hager, 1989):

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{/i} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii}}{1 - H_{ii}\ddot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i)} \dot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i), \tag{20}$$

$$\boldsymbol{H} = \boldsymbol{X}[\boldsymbol{X}^\top \text{diag}[\{\ddot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i)\}_i]\boldsymbol{X} + \nabla^2 R(\hat{\boldsymbol{\beta}})]^{-1}\boldsymbol{X}^\top. \tag{21}$$

This is the formula reported in (Rad & Maleki, 2018). By calculating $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{H}$ in advance, we can cheaply approximate the leave-i-out prediction for all i and efficiently evaluate the LOOCV risk. On the other hand, in order to use the above strategy, twice differentiability of both the loss and the regularizer is necessary in a neighborhood of $\hat{\boldsymbol{\beta}}$. However, this assumption is violated for many machine learning models including LASSO, Nuclear norm, and SVM. In the next two sections, we introduce a smoothing technique which lifts the scope of the above primal approach to nondifferentiable losses and regularizers.

4.2. Nonsmooth Loss and Smooth Regularizer

In this section we study the piecewise smooth loss functions and twice differentiable regularizers. Such problems arise in SVM (Cortes & Vapnik, 1995) and robust regression (Huber, 1973). Before proceeding further, we clarify our assumptions on the loss function.

Definition 4.1. *A singular point of a function is called q^{th} order, if at this point the function is q times differentiable, but its $(q+1)^{th}$ order derivative does not exist.*

Below we assume the loss ℓ is piecewise twice differentiable with k zero-order singularities $v_1, \dots, v_k \in \mathbb{R}$. The existence of singularities prohibits us from directly applying strategies in (19) and (20), where twice differentiability of ℓ and R is necessary. A natural solution is to first smooth the loss ℓ, then apply the framework in Section 4.1 to the smoothed version and finally reduce the smoothness to recover the ALO formula for the original nonsmooth problem.

As the first step, consider the following smoothing idea:

$$\ell_h(\mu; y) =: \frac{1}{h}\int \ell(u; y)\phi((\mu - u)/h)du,$$

where $h > 0$ is fixed and ϕ is a smooth symmetric function with the following properties:

Normalization: $\int \phi(w)dw = 1$, $\phi(w) \geq 0$, $\phi(0) > 0$;
Compact support: $\mathrm{supp}(\phi) = [-C, C]$ for some $C > 0$.

Now plug in this smooth version ℓ_h into (18) to obtain the following formula from (19):

$$\begin{aligned} \boldsymbol{G}_h &:= \sum_{j\neq i} \boldsymbol{x}_j \boldsymbol{x}_j^\top \ddot{\ell}_h(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h; y_j) + \nabla^2 R(\hat{\boldsymbol{\beta}}_h), \\ \tilde{\boldsymbol{\beta}}_h^{/i} &:= \hat{\boldsymbol{\beta}}_h + \boldsymbol{G}_h^{-1}\boldsymbol{x}_i \dot{\ell}_h(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_h; y_i). \end{aligned} \tag{22}$$

where $\hat{\boldsymbol{\beta}}_h$ is the minimizer on the full data from loss ℓ_h and R. $\tilde{\boldsymbol{\beta}}_h^{/i}$ is a good approximation to the leave-i-out estimator $\hat{\boldsymbol{\beta}}_h^{/i}$ based on smoothed loss ℓ_h. Setting $h \to 0$, we have $\ell_h(\mu, y)$ converge to $\ell(\mu, y)$ uniformly in the region of interest (see Appendix C.1 for the proof), implying that $\lim_{h\to 0} \tilde{\boldsymbol{\beta}}_h^{/i}$ serves as a good estimator of $\lim_{h\to 0} \hat{\boldsymbol{\beta}}_h^{/i}$, which is heuristically close to the true leave-i-out $\hat{\boldsymbol{\beta}}^{/i}$. Equation (22) can be simplified in the limit $h \to 0$. We define the sets of indices V and S for the samples at singularities and smooth parts respectively:

$$\begin{aligned} V &:= \{j : \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} = v_t \text{ for some } t \in \{1, \dots, k\}\}, \\ S &:= \{1, \dots, n\} \setminus V. \end{aligned}$$

We characterize the limit of $\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}_h^{/i}$ below.

Theorem 4.1. *Under some mild conditions, as $h \to 0$,*

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}_h^{/i} \to \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + a_i g_{\ell,i}$$

where

$$\begin{aligned} a_i &= \begin{cases} \frac{W_{ii}}{1 - W_{ii}\ddot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i)} & \text{if } i \in S, \\ \frac{1}{[(\boldsymbol{X}_{V\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V\cdot}^\top)^{-1}]_{ii}} & \text{if } i \in V, \end{cases} \\ \boldsymbol{Y} &= \nabla^2 R(\hat{\boldsymbol{\beta}}) + \boldsymbol{X}_{S\cdot}^\top \mathrm{diag}[\{\ddot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}})\}_{j\in S}]\boldsymbol{X}_{S\cdot}, \\ W_{ii} &= \boldsymbol{x}_i^\top \boldsymbol{Y}^{-1}\boldsymbol{x}_i - \boldsymbol{x}_i^\top \boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top (\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}\boldsymbol{Y}^{-1}\boldsymbol{x}_i. \end{aligned}$$

For $i \in S$, $g_{\ell,i} = \dot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i)$, and for $i \in V$, we have:

$$\boldsymbol{g}_{\ell,V} = (\boldsymbol{X}_{V,\cdot}\boldsymbol{X}_{V,\cdot}^\top)^{-1}\boldsymbol{X}_{V,\cdot}[\nabla R(\hat{\boldsymbol{\beta}}) - \sum_{j\in S} \boldsymbol{x}_j \dot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}; y_j)].$$

We can obtain the ALO estimate of prediction error by plugging $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + a_i g_{\ell,i}$ instead of $\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{/i}$ in (5). The conditions and proof of Theorem 4.1 can be found in Appendix C.3.

4.3. Nonsmooth Regularizer and Smooth Loss

The smoothing technique proposed in the last section can also handle many nonsmooth regularizers. In this section we focus on separable regularizers R, defined as $R(\boldsymbol{\beta}) = \sum_{l=1}^p r(\beta_l)$, where $r : \mathbb{R} \to \mathbb{R}$ is piecewise twice differentiable with finite number of zero-order singularities in $v_1, \dots, v_k \in \mathbb{R}$. (Examples on non-separable regularizers are studied in Section 6.) We further assume the loss function ℓ to be twice differentiable and denote by $A = \{l : \hat{\beta}_l \neq v_t, \text{ for any } t \in \{1, \dots, k\}\}$ the active set.

For the coordinates of $\hat{\boldsymbol{\beta}}$ that lie in A, our objective function, constrained to these coordinates, is locally twice differentiable. Hence we expect $\hat{\boldsymbol{\beta}}_A^{/i}$ to be well approximated by the ALO formula using $\hat{\boldsymbol{\beta}}_A$. On the other hand, components not in A are trapped at singularities. As long as they are not on the boundary of being in or out of the singularities, we expect these locations of $\hat{\boldsymbol{\beta}}^{/i}$ to stay at the same values.

Technically, consider a similar smoothing scheme for r:

$$r_h(w) = \frac{1}{h}\int r(u)\phi((w - u)/h)du,$$

and let $R_h(\boldsymbol{\beta}) = \sum_{l=1}^{p} r_h(\beta_l)$. We then consider the ALO formula of Model (18) with regularizer R_h:

$$\begin{aligned} \boldsymbol{G}_h &:= \sum_{j \neq i} \boldsymbol{x}_j \boldsymbol{x}_j^\top \ddot{\ell}(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_h; y_j) + \nabla^2 R_h(\hat{\boldsymbol{\beta}}_h), \\ \tilde{\boldsymbol{\beta}}_h^{/i} &:= \hat{\boldsymbol{\beta}}_h + \boldsymbol{G}_h^{-1} \boldsymbol{x}_i \dot{\ell}_h(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i). \end{aligned} \tag{23}$$

Setting $h \to 0$, (23) reduces to a simplified formula which heuristically serves as a good approximation to the true leave-i-out estimator $\hat{\boldsymbol{\beta}}^{/i}$, stated as the following theorem:

Theorem 4.2. *Under some mild conditions, as $h \to 0$,*

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}_h^{/i} \to \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{H_{ii} \dot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i)}{1 - H_{ii} \ddot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i)},$$

with

$$\boldsymbol{H} = \boldsymbol{X}_{\cdot,A} [\boldsymbol{X}_{\cdot,A}^\top \mathrm{diag}[\{\ddot{\ell}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}; y_i)\}_i] \boldsymbol{X}_{\cdot,A} + \nabla^2 R(\hat{\boldsymbol{\beta}}_A)]^{-1} \boldsymbol{X}_{\cdot,A}^\top.$$

The conditions and proof of Theorem 4.2 can be found in the Appendix C.2.

Remark 4.1. *For nonsmooth problems, higher order singularities do not cause issues: the set of tuning values which cause $\hat{\beta}_l$ (for regularizer) or $\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}$ (for loss) to fall at those higher order singularities has measure zero.*

Remark 4.2. *For both nonsmooth losses and regularizers, we need to invert some matrices in the ALO formula. Although the invertibility does not seem guaranteed in the general formula, as we apply ALO to specific models, the structures of the loss and/or the regularizer ensures this invertibility. For example, for LASSO, we have that the size of the active set $|E| \leq \min(n, p)$.*

Remark 4.3. *We note that the dual approach is typically powerful for models with smooth losses and norm-type regularizers, such as the SLOPE norm and the generalized LASSO. On the other hand, the primal approach is valuable for models with nonsmooth loss or when the Hessian of the regularizer is feasible to calculate. Such regularizers often exhibit some type of separability or symmetry, such as in the case of SVM or nuclear norm.*

5. Equivalence Between Primal and Dual Methods

Although the primal and dual methods may be harder or easier to carry out depending on the specific problem at hand, one may wonder if they always obtain the same result. In this section, we outline a unifying view for both methods, and state an equivalence theorem.

As both the primal and dual methods are based on a first-order approximation strategy, we will study them not as approximate solutions to the leave-i-out problem, but will instead show that they are exact solutions to a surrogate leave-i-out problem. Indeed, recall that the leave-i-out problem is given by (4), which cannot be solved in closed form. However, we note that the solution does exist in closed form in the case where both ℓ and R are quadratic functions.

We may thus consider the approximate leave-i-out problem, where both ℓ and R have been replaced in the leave-i-out problem (4) by their quadratic expansion at $\hat{\boldsymbol{\beta}}$:

$$\min_{\boldsymbol{\beta}^{/i}} \sum_{j \neq i} \tilde{\ell}(\boldsymbol{x}_j^\top \boldsymbol{\beta}^{/i}; y_j) + \tilde{R}(\boldsymbol{\beta}^{/i}). \tag{24}$$

When both ℓ and R are twice differentiable at the full data solution, $\tilde{\ell}$ and $\tilde{R}$ correspond to their respective second order Taylor expansions at $\hat{\boldsymbol{\beta}}$. When ℓ or R is not twice differentiable at the full data solution, we have seen that it is still possible to obtain an ALO estimator through the proximal map (in the case of the dual) or through smoothing arguments (in the case of the primal). The corresponding quadratic surrogates may then be formulated as partial quadratic functions, i.e. convex quadratic functions restricted to an affine subspace. However, due to space limitations we only focus on twice differentiable losses and regularizers here.

The way we obtain $\tilde{\boldsymbol{\beta}}^{/i}$ in (19) indicates that the primal formula in (20) and (21) are the exact leave-i-out solution of the surrogate primal problem (24). On the other hand, we may also wish to consider the surrogate dual problem, by replacing ℓ^* and R^* by their quadratic expansion at full data dual solution $\hat{\boldsymbol{\theta}}$ in the dual problem (7). One may possibly worry that the surrogate dual problem is then different from the dual of the surrogate primal problem (24). This does not happen, and we have the following theorem.

Theorem 5.1. *Let ℓ and R be twice differentiable convex functions. Let $\tilde{\ell}$ and $\tilde{R}$ denote the quadratic surrogates of the loss and regularizer at $\hat{\boldsymbol{\beta}}$, and let $\tilde{\ell}_D^*$ and $\tilde{R}_D^*$ denote the quadratic surrogates of the conjugate loss and regularizer at the dual full data solution $\hat{\boldsymbol{\theta}}$. We have that the following problems are equivalent (have the same minimizer):*

$$\min_{\boldsymbol{\theta}} \textstyle\sum_{j=1}^n \tilde{\ell}^*(-\theta_j; y_j) + \tilde{R}^*(\boldsymbol{X}^\top \boldsymbol{\theta}), \tag{25}$$

$$\min_{\boldsymbol{\theta}} \textstyle\sum_{j=1}^n \tilde{\ell}_D^*(-\theta_j; y_j) + \tilde{R}_D^*(\boldsymbol{X}^\top \boldsymbol{\theta}). \tag{26}$$

Additionally, we note that the dual method described in Section 3 solves the surrogate dual problem (26).

Theorem 5.2. *Let $\boldsymbol{X}_u$, $\boldsymbol{y}_u$ be as in (16), and let $\tilde{y}_{u,i}^{/i}$ be the transformed ALO obtained in (17). Let $\tilde{\boldsymbol{y}}_a$ be the same as $\boldsymbol{y}_u$ except $\tilde{y}_{a,i} = \tilde{y}_{u,i}^{/i}$. Then $\tilde{\boldsymbol{y}}_a$ satisfies*

$$[\mathbf{prox}_{\tilde{g}}(\tilde{\boldsymbol{y}}_a)]_i = 0, \tag{27}$$

where $\tilde{g}(\boldsymbol{u}) = \tilde{R}^*(\boldsymbol{X}_u^\top \boldsymbol{u})$ and $\tilde{R}$ denotes the quadratic surrogate of the regularizer. In particular, $\tilde{y}_i^{/i} = K_{ii} \tilde{y}_{u,i}^{/i}$ is the exact leave-i-out predicted value for the surrogate problem in Theorem 5.1.

We refer the reader to Appendix B for the proofs. These two theorems imply that for twice differentiable losses and regularizers, the frameworks we laid out in Sections 3 and 4 lead to exactly the same ALO formulas.

6. Applications

6.1. Generalized LASSO

The generalized LASSO (Tibshirani & Taylor, 2011) is a generalization of the LASSO problem which captures many applications such as the fused LASSO (Tibshirani et al., 2005), ℓ_1 trend filtering (Kim et al., 2009) and wavelet smoothing in a unified framework. The generalized LASSO problem solves the following penalized regression problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\sum_{j=1}^{n}(y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1. \tag{28}$$

where the regularizer is parameterized by a fixed matrix $\boldsymbol{D} \in \mathbb{R}^{m\times p}$ which captures the desired structure in the data. We note that the regularizer is a semi-norm. Hence we can formulate the dual problem as a projection. In fact, a dual formulation of (28) can be obtained as (see Appendix D):

$$\min_{\boldsymbol{\theta},\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{y}\|_2^2 \quad \text{s.t. } \|\boldsymbol{u}\|_\infty \le \lambda \text{ and } \boldsymbol{X}^\top\boldsymbol{\theta} = \boldsymbol{D}^\top\boldsymbol{u}.$$

The dual optimal solution satisfies $\hat{\boldsymbol{\theta}} = \Pi_{\Delta_X}(\boldsymbol{y})$, where Δ_X is the polytope given by:

$$\Delta_X = \{\boldsymbol{\theta} \in \mathbb{R}^n : \exists \boldsymbol{u}, \|u\|_\infty \le \lambda \text{ and } \boldsymbol{X}^\top\boldsymbol{\theta} = \boldsymbol{D}^\top u\}.$$

The projection onto the polytope $C = \{\boldsymbol{D}^\top\boldsymbol{u} : \|\boldsymbol{u}\|_\infty \le \lambda\}$ is given in (Tibshirani & Taylor, 2011) as locally being the projection onto the affine space orthogonal to the nullspace of $\boldsymbol{D}_{\cdot,-E}$, where $E = \{i : |\hat{u}_i| = \lambda\}$ and $-E = \{1,\dots,p\} \setminus E$. Since $\Delta_X = [\boldsymbol{X}^\top]^{-1}C$ is the inverse image of C under the linear map given by $\boldsymbol{X}^\top$, the projection onto Δ_X is given locally by the projection onto the affine space normal to the space spanned by the columns of $[\boldsymbol{X}^\top]^+ \text{null}\,\boldsymbol{D}_{\cdot,-E}$, provided $\boldsymbol{X}$ has full column rank. Here, $[\boldsymbol{X}^\top]^+$ denotes the Moore-Penrose pseudoinverse of $\boldsymbol{X}^\top$. To obtain a spanning set of this space, we consider $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{B}$, where $\boldsymbol{B}$ is a set of vectors spanning the nullspace of $\boldsymbol{D}_{\cdot,-E}$. This allows us to compute $\boldsymbol{H} = \boldsymbol{A}\boldsymbol{A}^+$, the projection onto the normal space required to compute the ALO.

6.2. Nuclear Norm

Consider the following matrix sensing problem

$$\hat{\boldsymbol{B}} := \arg\min_{\boldsymbol{B}} \frac{1}{2}\sum_{j=1}^{n}(y_j - \langle \boldsymbol{X}_j, \boldsymbol{B}\rangle)^2 + \lambda\|\boldsymbol{B}\|_* \tag{29}$$

with $\boldsymbol{B}, \boldsymbol{X}_j \in \mathbb{R}^{p_1\times p_2}$. $\langle \boldsymbol{X}, \boldsymbol{B}\rangle = \text{trace}(\boldsymbol{X}^\top\boldsymbol{B})$ denotes the inner product. We use $\|\cdot\|_*$ for nuclear norm, which is defined as the sum of the singular values of a matrix. The nuclear norm is a unitarily invariant function of the matrix (Lewis, 1995). Such functions are only indirectly related to the components of the matrix, making their analysis difficult even when they are smooth, and exacerbating the difficulties when they are non-smooth such as in the case of the nuclear norm. In particular, the smoothing framework described in Section 4.3 cannot be applied directly.

We are nonetheless able to leverage the specific structure of such functions and apply the smoothing trick to the singular values to obtain the following theorem. For more details on the derivation, please refer to Appendix E.3.

Theorem 6.1. *Consider the nuclear-norm penalized matrix regression problem (29), and let $\hat{\boldsymbol{B}} = \hat{\boldsymbol{U}}\text{diag}[\hat{\boldsymbol{\sigma}}]\hat{\boldsymbol{V}}^\top$ be the SVD of the full data estimator $\hat{\boldsymbol{B}}$, with $\hat{\boldsymbol{U}} \in \mathbb{R}^{p_1\times p_1}$, $\hat{\boldsymbol{V}} \in \mathbb{R}^{p_2\times p_2}$. Let $m = \text{rank}(\hat{\boldsymbol{B}})$ be the number of nonzero $\hat{\sigma}_j$'s for $\hat{\boldsymbol{B}}$. Let $\tilde{\boldsymbol{B}}_h^{/i}$ denote the approximate of $\hat{\boldsymbol{B}}^{/i}$ obtained from the smoothed problem. Then, as $h \to 0$*

$$\langle \boldsymbol{X}_i, \tilde{\boldsymbol{B}}_h^{/i}\rangle \to \langle \boldsymbol{X}_i, \hat{\boldsymbol{B}}\rangle + \frac{H_{ii}}{1-H_{ii}}(\langle \boldsymbol{X}_i, \hat{\boldsymbol{B}}\rangle - y_i),$$
$$\boldsymbol{H} = \boldsymbol{\mathcal{X}}_{\cdot,E}[\boldsymbol{\mathcal{X}}_{\cdot,E}^\top\boldsymbol{\mathcal{X}}_{\cdot,E} + \lambda\boldsymbol{\mathcal{G}}]^{-1}\boldsymbol{\mathcal{X}}_{\cdot,E}^\top,$$

with $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n\times p_1p_2}$ and $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{m(p_1+p_2-m)\times m(p_1+p_2-m)}$ a symmetric matrix given by:

$$\boldsymbol{\mathcal{X}}_{j,kl} = \hat{\boldsymbol{U}}_{\cdot,k}^\top\boldsymbol{X}_j\hat{\boldsymbol{V}}_{\cdot,l},$$

$$\boldsymbol{\mathcal{G}}_{kl,st} = \begin{cases} 0 & s=t=k=l\le m \\ \frac{1}{\hat{\sigma}_s+\hat{\sigma}_t} & 1\le s\ne t\le m, (k,l)=(s,t) \\ \frac{1}{\hat{\sigma}_s} & 1\le s\le m<t\le p_2, (k,l)=(s,t) \\ \frac{1}{\hat{\sigma}_t} & 1\le t\le m<s\le p_1, (k,l)=(s,t) \\ -\frac{1}{\hat{\sigma}_s+\hat{\sigma}_t} & 1\le s\ne t\le m, (k,l)=(t,s) \\ -\frac{g_r[\hat{\sigma}_t]}{\hat{\sigma}_s} & 1\le s\le m<t\le p_2, (k,l)=(t,s) \\ -\frac{g_r[\hat{\sigma}_s]}{\hat{\sigma}_t} & 1\le t\le m<s\le p_2, (k,l)=(t,s) \\ 0 & \textit{otherwise.} \end{cases} \tag{30}$$

where for $t > m$, $\hat{\sigma}_t = 0$ and $g_r[\hat{\sigma}_t]$ is the corresponding subgradient at this singular value, which can be obtained through the SVD of $\frac{1}{\lambda}\sum_{j=1}^{n}(y_j - \langle \boldsymbol{X}_j, \hat{\boldsymbol{B}}\rangle)\boldsymbol{X}_j$. The set E is then defined as:

$$E = \{(k,l) : k \le m \text{ or } l \le m\}.$$

Note that the indices of $\boldsymbol{\mathcal{G}}$ and the index set E are consistent.

6.3. Linear SVM

The linear SVM optimization can be written as

$$\arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{n}(1 - y_j\boldsymbol{x}_j^\top\boldsymbol{\beta})_+ + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2,$$

with $y_j \in \{-1, 1\}$ and $(\cdot)_+ = \max\{\cdot, 0\}$. Note that this is a special case of the problem we studied in Section 4.2. Here, $\ell(u; y_j) = (1 - y_j u)_+$ has only one zero order singularity at y_j. Using Theorem 4.1 and simplifying the expressions, we obtain the following ALO formula for SVM:

$$\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}^{/i} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + a_i g_{\ell,i},$$

where

$$a_i = \begin{cases} \frac{1}{\lambda} \boldsymbol{x}_i^\top (\boldsymbol{I}_p - \boldsymbol{X}_{V,\cdot}^\top (\boldsymbol{X}_{V,\cdot} \boldsymbol{X}_{V,\cdot}^\top)^{-1} \boldsymbol{X}_{V,\cdot}) \boldsymbol{x}_i & i \in S, \\ \left(\lambda [(\boldsymbol{X}_{V,\cdot} \boldsymbol{X}_{V,\cdot}^\top)^{-1}]_{ii}\right)^{-1} & i \in V, \end{cases}$$

and for $i \in S$, $g_{\ell,i} = -y_i$ if $y_i \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} < 1$, $g_{\ell,i} = 0$ if $y_i \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} > 1$, and for $i \in V$

$$\boldsymbol{g}_{\ell,V} = (\boldsymbol{X}_{V,\cdot} \boldsymbol{X}_{V,\cdot}^\top)^{-1} \boldsymbol{X}_{V,\cdot} [\lambda \hat{\boldsymbol{\beta}} + \textstyle\sum_{j: y_j \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} < 1} y_j \boldsymbol{x}_j].$$

Recall that $V = \{j : \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} = y_j\}$ and $S = [1, \ldots, n] \backslash V$.

7. Numerical Experiments

We illustrate the performance of ALO through three experiments. The first two compare the ALO risk estimate with that of LOOCV. The third experiment compares the computational complexity of ALO with that of LOOCV. We have also evaluated the performance of ALO on real-world datasets. Due to lack of space, these results are presented in Appendix F.2. For the first experiment (Figure 2a), we run ALO and LOOCV for the three models studied in Section 6 (using fused LASSO (Tibshirani et al., 2005) as a special case of generalized LASSO) and compare their risk estimates under the settings $n > p$ and $n < p$ respectively. The full details of the experiments are provided in Appendix F.

For the second experiment (Figure 2b), we consider the risk estimates for LASSO from ALO and LOOCV under settings with model mis-specification, heavy-tail noise and correlated design. For all three cases, ALO approximates LOOCV well.

Table 1. Timing (in *sec*) of one single fit, ALO and LOOCV. In the upper and lower tables, we fix $n = 800$ and $p = 800$ respectively.

p	200	400	1600
single fit	0.035 ± 0.001	0.13 ± 0.01	0.60 ± 0.01
ALO	0.060 ± 0.001	0.21 ± 0.01	0.89 ± 0.01
LOOCV	27.52 ± 0.03	107.4 ± 0.5	479 ± 2
n	200	400	1600
single fit	0.055 ± 0.002	0.19 ± 0.01	0.76 ± 0.02
ALO	0.065 ± 0.001	0.24 ± 0.01	1.20 ± 0.01
LOOCV	11.44 ± 0.049	74.7 ± 0.5	1249 ± 3

In general, we observe that the estimates given by ALO are close to LOOCV, although the performance may deteriorate for very small values of λ, as is clear in the fused-LASSO ($n < p$) example. These values of λ correspond to "dense" solutions, and are far from the optimal choice. Hence, such inaccuracies do not harm the parameter tuning algorithm.

Our last experiment compares the computational complexity of ALO with that of LOOCV. In Table 1, we provide the timing of LASSO for different values of n and p. The time required by ALO, which involves a single fit and a matrix inversion (in the construction of $\boldsymbol{H}$ matrix), is in all experiments no more than twice that of a single fit. We refer the reader to Appendix F for the details of this experiment.

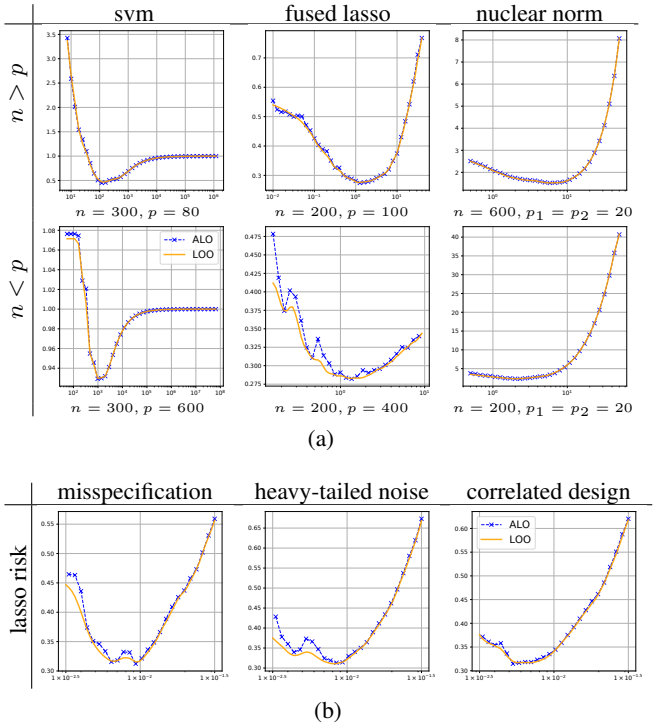

Figure 2. Risk estimates from ALO versus LOOCV. The x-axis is the tuning parameter value on log-scale, the y-axis is the risk estimate. In part (a), the comparison is based on SVM, fused LASSO and nuclear norm. For nuclear norm, p_1, p_2 are dimensions of a matrix. In part (b), we consider the risk estimates of LASSO under model mis-specification, heavy-tailed noise and correlated design scenarios.

8. Discussion

ALO offers a highly efficient approach for parameter tuning and risk estimation for a large class of statistical machine learning models. We focus on nonsmooth models and propose two general frameworks for calculating ALO. One is from the primal perspective, the other from the dual.

By approximating LOOCV, ALO inherits desirable properties of LOOCV in high-dimensional settings where n and p are comparable. In particular, ALO can overcome the bias issues that k-fold cross validation displays in these settings.

Acknowledgements

We acknowledge computing resources from Columbia University's Shared Research Computing Facility project, which is supported by NIH Research Facility Improvement Grant 1G20RR030893-01, and associated funds from the New York State Empire State Development, Division of Science Technology and Innovation (NYSTAR) Contract C090171, both awarded April 15, 2010.

References

Allen, D. M. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12): 6745–6750, 1999.

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

Bayati, M., Erdogdu, M. A., and Montanari, A. Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*, pp. 944–952, 2013.

Becker, S. R., Candès, E. J., and Grant, M. C. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3(3):165–218, 2011.

Beirami, A., Razaviyayn, M., Shahrampour, S., and Tarokh, V. On optimal generalizability in parametric learning. In *Advances in Neural Information Processing Systems*, pp. 3458–3468, 2017.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, Cambridge, 2004. ISBN 0-521-83378-7.

Candes, E. J., Sing-Long, C. A., and Trzasko, J. D. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE transactions on signal processing*, 61 (19):4643–4657, 2013.

Cawley, G. C. and Talbot, N. L. Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71(2-3):243–264, 2008.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

Combettes, P. L. and Pesquet, J.-C. *Proximal Splitting Methods in Signal Processing*, pp. 185–212. Springer New York, New York, NY, 2011. ISBN 978-1-4419-9569-8.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Donoho, D. L. and Tanner, J. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452–9457, 2005.

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Dossal, C., Kachour, M., Fadili, M., Peyré, G., and Chesneau, C. The degrees of freedom of the lasso for general design matrix. *Statistica Sinica*, pp. 809–828, 2013.

Efron, B. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

Grant, M. and Boyd, S. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. Result analysis of the nips 2003 feature selection challenge. In *Advances in neural information processing systems*, pp. 545–552, 2005.

Hager, W. W. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989.

Hastie, T., Tishirani, R., and Friedman, J. *Elements of Statistical Learning*, chapter Model Assessment and Selection. Springer-Verlag New York, 2 edition, 2009. ISBN 978-0-387-84857-0.

Huber, P. J. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, pp. 799–821, 1973.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. l_1 trend filtering. *SIAM Rev.*, 51(2):339–360, 2009.

Lan, G., Lu, Z., and Monteiro, R. D. C. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, Jan 2011.

Le Cessie, S. and Van Houwelingen, J. C. Ridge estimators in logistic regression. *Applied statistics*, pp. 191–201, 1992.

Lewis, A. S. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

Meijer, R. J. and Goeman, J. J. Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155, 2013.

Mirsky, L. Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11:50–59, 1960.

Mousavi, A., Maleki, A., Baraniuk, R. G., et al. Consistent parameter estimation for lasso and approximate message passing. *The Annals of Statistics*, 45(6):2427–2454, 2017.

Obuchi, T. and Kabashima, Y. Cross validation in lasso and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(5):053304, 2016.

Opper, M. and Winther, O. Gaussian processes and svm: Mean field results and leave-one-out. 2000.

O'sullivan, F., Yandell, B. S., and Raynor Jr, W. J. Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, 81(393):96–103, 1986.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Qian, J., Hastie, T., Friedman, J., Tibshirani, R., and Simon, N. Glmnet for matlab 2013. *URL http://www. stanford. edu/˜ hastie/glmnet_matlab*, 2013.

Rad, K. and Maleki, A. A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv preprint arXiv:1801.10243*, 2018.

Rockafellar, R. T. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pp. 111–147, 1974.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

Tibshirani, R. J. and Taylor, J. The solution path of the generalized lasso. *Ann. Statist.*, 39(3):1335–1371, 2011.

Tibshirani, R. J., Taylor, J., et al. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

Vaiter, S., Deledalle, C., Fadili, J., Peyré, G., and Dossal, C. The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, 69(4): 791–832, 2017.

Weyl, L. Das asymptotische verteilungsgestez der eigenwert linearer partieller differentialgleichungen (mit einer anwendung auf der theorie der hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912.

Woodbury, M. A. Inverting modified matrices. *Memorandum report*, 42(106):336, 1950.

Zou, H., Hastie, T., Tibshirani, R., et al. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5): 2173–2192, 2007.