Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions

Kamiar Rahnama Rad
Baruch College
City University of New York

Wenda Zhou
Columbia University

Arian Maleki
Columbia University

Abstract

We study the problem of out-of-sample risk estimation in the high dimensional regime where both the sample size n and number of features p are large, and n/p can be less than one. Extensive empirical evidence confirms the accuracy of leave-one-out cross validation (LO) for out-of-sample risk estimation. Yet, a unifying theoretical evaluation of the accuracy of LO in high-dimensional problems has remained an open problem. This paper aims to fill this gap for penalized regression in the generalized linear family. With minor assumptions about the data generating process, and without any sparsity assumptions on the regression coefficients, our theoretical analysis obtains finite sample upper bounds on the expected squared error of LO in estimating the out-of-sample error. Our bounds show that the error goes to zero as $n, p \to \infty$, even when the dimension p of the feature vectors is comparable with or greater than the sample size n. One technical advantage of the theory is that it can be used to clarify and connect some results from the recent literature on scalable approximate LO.

Keywords: High-dimensional statistics, Regularized estimation, Out-of-sample risk estimation, Cross validation, Generalized linear models, Model selection.

1 Introduction

Balancing the sensible level of model *complexity* against model *fitness* is a fundamental challenge faced by any learning algorithm. A model that is too simple can fail to capture the essential pattern in the data, and a model that is too complex is oversensitive to the idiosyncrasies of the particular data, resulting in highly variable patterns that are mere mirages in the noise. The learning algorithm's ability to perform well on *new, previously unseen data* is typically used to set the model complexity. This performance is known as the *out-of-sample error.*

To be concrete, let $D = \{(y_1, \boldsymbol{x_1}), \ldots, (y_n, \boldsymbol{x_n})\}$ be our dataset where $\boldsymbol{x_i} \in \mathrm{R}^p$ and $y_i \in \mathrm{R}$ denote the features and response, respectively. The goal is to obtain an estimate of the response for a newly observed feature vector. We assume observations are independent and identically distributed draws from some joint unknown distribution $q(y_i, \boldsymbol{x_i})$. We model this distribution as $q(y_i, \boldsymbol{x_i}) = q_1(y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}_*) q_2(\boldsymbol{x}_i)$, and estimate $\boldsymbol{\beta}_*$ using the optimization problem

$$\hat{\boldsymbol{\beta}} \triangleq \underset{\boldsymbol{\beta} \in \mathrm{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} \ell(y_i \mid \boldsymbol{x_i}^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}) \Big\}, \tag{1}$$

where ℓ is called the loss function, and $r(\boldsymbol{\beta})$ is called the regularizer. Both the regularizer $r(\boldsymbol{\beta})$ and the regularization parameter λ have significant effects on the performance of the estimate by controlling the complexity of the model. Hence, for picking a good regularizer, r, or tuning the parameter λ one would like to estimate the *out-of-sample prediction error*, defined as

$$\mathrm{Err}_{\mathrm{out}} \triangleq \mathbb{E}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}) \mid D], \tag{2}$$

where $(y_o, \boldsymbol{x}_o)$ is a *new, previously unseen* sample from the unknown distribution $q(y, \boldsymbol{x})$ independent of $\mathcal{D}$, and ϕ is a function that measures the closeness of y_o to $\boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}$. A standard choice for ϕ is $\ell(y \mid \boldsymbol{x}^\top \boldsymbol{\beta})$.

The problem of risk estimation has been extensively studied in the past fifty years and popular estimates, such as k-fold cross validation [Stone, 1974] are used extensively in practical systems. However, the emergence of high-dimensional estimation problems in which the number of features p is comparable or even larger than

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

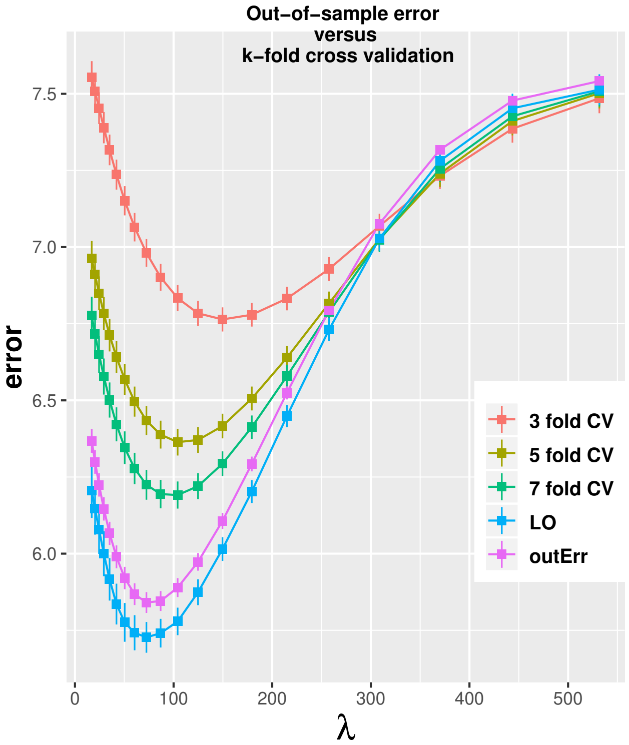

Figure 1: Comparison of K-fold cross validation (for $K = 3, 5, 7$) and leave-one-out cross validation with the true (oracle-based) out-of-sample error for the elastic-net problem where $\ell(y \mid \boldsymbol{x}^\top \boldsymbol{\beta}) = \frac{1}{2}(y - \boldsymbol{x}^\top \boldsymbol{\beta})^2$ and $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1/2 + \|\boldsymbol{\beta}\|_2^2/4$. The upward bias of K-fold CV clearly decreases as number of folds increase. $y_i \sim \mathrm{N}(\boldsymbol{x}_{\boldsymbol{i}}^\top \boldsymbol{\beta}^*, \sigma^2)$ and $\boldsymbol{x}_{\boldsymbol{i}} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$. The number of nonzero elements of the true $\boldsymbol{\beta}^*$ is set to k and their values is set to $\frac{1}{3\sqrt{2}}$. Dimensions are $(p, n, k) = \big(2000, 500, 100\big)$ and $\sigma^2 = 2$. Extra-sample test data is $y_o \sim \mathrm{N}(\boldsymbol{x}_o^\top \boldsymbol{\beta}^*, \sigma^2)$ where $\boldsymbol{x}_o \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$. The true (oracle-based) out-of-sample prediction error is $\mathrm{Err}_{\mathrm{out}} = \mathbb{E}[(y_o - \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}})^2 | D] = \sigma^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$. All depicted quantities are averages based on 100 random independent samples, and error bars depict one standard error.

the number of observations n, deemed many standard techniques in-accurate. For instance, Figure 1 compares the estimates obtained from k-fold cross validation for different values of k. As is clear in this figure, given the importance of each observation in high-dimensional settings, standard techniques, such as 5-fold suffer from a large bias.

One of the existing estimates of $\mathrm{Err}_{\mathrm{out}}$ that seems to be accurate in high-dimensional settings is the leave-one-out cross validation (LO), which is defined through the following formula:

$$\mathrm{LO} \triangleq \frac{1}{n} \sum_{i=1}^{n} \phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}), \tag{3}$$

where

$$\hat{\boldsymbol{\beta}}_{/i} \triangleq \underset{\boldsymbol{\beta} \in \mathrm{R}^p}{\arg\min} \Big\{ \sum_{j \neq i} \ell(y_j \mid \boldsymbol{x}_j^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}) \Big\}, \tag{4}$$

is the leave-i-out estimate. The simulation results reported in Figure 1 and elsewhere [Rahnama Rad and Maleki, 2019, Wang et al., 2018, Stephenson and Broderick, 2019, Beirami et al., 2017, Takahashi and Kabashima, 2018] have demonstrated the good performance of LO in a wide range of high-dimensional problems. Despite the existence of extensive simulation results, the theoretical properties of LO have not been studied in the high-dimensional settings.

In this paper, we study the expected squared error of LO in estimating the out-of-sample error, in the high-dimensional setting, where both n and p are large, and n/p can be less than one. We focus on regularized regression in the generalized linear family, and we make no sparsity assumption on the vector of regression coefficients. In short, we obtain an almost sharp upper bound on the error $|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}|$. These bounds not only show that $|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}| \to 0$ as $n, p \to \infty$, but they also capture the rate of this convergence. This finally establishes what has been observed in empirical studies; LO obtains accurate out-of-sample risk estimates even in high-dimensional problems.

An important advantage of our theoretical results is that they can be used to clarify and connect some results from the recent literature on computationally efficient approximation to LO. For instance, [Rahnama Rad and Maleki, 2019] showed that in the same high dimensional regime, $|\mathrm{ALO} - \mathrm{LO}| \to 0$ as $n, p \to \infty$, where ALO stands for a computationally efficient approximation of LO we formally refer to in Section 1.2. A major consequence of our theory is that it shows that ALO is a consistent estimator of $\mathrm{Err}_{\mathrm{out}}$. We make these statements more concrete in the next sections.

1.1 Notation

We first review the notations that will be used in the rest of the paper. Let $\boldsymbol{x_i}^\top \in \mathrm{R}^{1\times p}$ stand for the ith row of $\boldsymbol{X} \in \mathrm{R}^{n\times p}$. $\boldsymbol{y}_{/i} \in \mathrm{R}^{(n-1)\times 1}$ and $\boldsymbol{X}_{/i} \in \mathrm{R}^{(n-1)\times p}$ stand for $\boldsymbol{y}$ and $\boldsymbol{X}$, excluding the ith entry y_i and the ith row $\boldsymbol{x_i}^\top$, respectively, and let $\boldsymbol{X}_{/ij}$ be defined likewise. Additionally, let $\hat{\boldsymbol{\beta}}_{/ij}$ stand for the regularized estimate in (1) when $(y_i, \boldsymbol{x}_i)$ and $(y_j, \boldsymbol{x}_j)$ are excluded. Moreover, define

$$
\begin{aligned}
\dot\phi(y,z) &\triangleq \frac{\partial \phi(y,z)}{\partial z},\\
\dot\ell(y_i \mid \boldsymbol{x_i}^\top \boldsymbol{\beta}) &\triangleq \frac{\partial \ell(y_i \mid z)}{\partial z}\Big|_{z=\boldsymbol{x}_i^\top\boldsymbol{\beta}},\\
\ddot\ell_i(\boldsymbol{\beta}) &\triangleq \frac{\partial^2 \ell(y_i \mid z)}{\partial z^2}\Big|_{z=\boldsymbol{x}_i^\top\boldsymbol{\beta}},\\
\ddot{\boldsymbol{\ell}}_{/i}(\cdot) &\triangleq [\ddot\ell_1(\cdot), \cdots, \ddot\ell_{i-1}(\cdot), \ddot\ell_{i+1}(\cdot), \ldots, \ddot\ell_n(\cdot)]^\top.
\end{aligned}
$$

Likewise, define $\ddot{\boldsymbol{\ell}}_{/ij}(\boldsymbol{\beta})$. The notation $\mathrm{poly}\log \mathrm{n}$ denotes polynomial of $\log n$ with a finite degree. Let $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ stand for the largest and smallest eigenvalues of $\boldsymbol{A}$, respectively. We state $x_n = O_p(a_n)$ when the set of values x_n/a_n is stochastically bounded.

1.2 Computational complexity of LO and its approximation

The high computational cost of repeatedly refitting models is a major hurdle in using LO in high dimensional settings. A typical approach to alleviate this problem analytically approximates the leave-i-out model based on the full-data model. A large body of work has addressed computationally efficient approximations to the leave-one-out cross validation error for ridge regularized estimation problems (and its variants) [Allen, 1974, Craven and Wahba, 1979, Golub et al., 1979, O'Sullivan et al., 1986, Burman, 1990, Cessie and Houwelingen, 1992, Opper and Winther, 2000, Cawley and Talbot, 2008, Meijer and Goeman, 2013, Vehtari et al., 2016, Mousavi et al., 2018]. Extensions to a wide array of regularizers, such as LASSO [Obuchi and Kabashima, 2018, Rahnama Rad and Maleki, 2019, Stephenson and Broderick, 2019] and nuclear norm [Wang et al., 2018] were recently studied and the validity of these approximations in estimating LO (and its variants) were theoretically studied in [Obuchi and Kabashima, 2016, Beirami et al., 2017, Rahnama Rad and Maleki, 2019, Giordano et al., 2019, Stephenson and Broderick, 2019, Xu et al., 2019].

For example, a single Newton step around $\hat{\boldsymbol{\beta}}$ was used in [Wang et al., 2018, Rahnama Rad and Maleki, 2019] to approximate $\hat{\boldsymbol{\beta}}_{/i}$ by

$$
\tilde{\boldsymbol{\beta}}_{/i} \triangleq \hat{\boldsymbol{\beta}} + \Big(\sum_{j\neq i} \boldsymbol{x_j}\boldsymbol{x_j}^\top \ddot\ell(y_j \mid \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}) + \lambda \boldsymbol{\nabla^2 r}(\hat{\boldsymbol{\beta}})\Big)^{-1} \cdot \boldsymbol{x_i}\dot\ell(y_i \mid \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})
$$

and using the Woodburry lemma, the following scalable approximate LO (ALO) formula was obtained:

$$
\begin{aligned}
\mathrm{ALO} &\triangleq \frac{1}{n}\sum_{i=1}^n \phi(y_i, \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}_{/i})\\
&= \frac{1}{n}\sum_{i=1}^n \phi\Big(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \big(\frac{H_{ii}}{1-H_{ii}}\big)\frac{\dot\ell(y_i \mid \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}{\ddot\ell(y_i \mid \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})}\Big) \qquad (5)
\end{aligned}
$$

where

$$
\boldsymbol{H} \triangleq \boldsymbol{X}(\boldsymbol{X}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X} + \lambda \boldsymbol{\nabla^2 r}(\hat{\boldsymbol{\beta}}))^{-1}\boldsymbol{X}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]
$$

This result was extended to nonsmooth regularizers. For example, [Wang et al., 2018, Rahnama Rad and Maleki, 2019, Stephenson and Broderick, 2019] showed that for $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$, the same ALO formula is a valid approximation of LO if the following $\boldsymbol{H}$ matrix is used:

$$
\boldsymbol{H} = \boldsymbol{X}_S\left(\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]\boldsymbol{X}_S\right)^{-1}\boldsymbol{X}_S^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})]
$$

where S is the active set of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_S$ is the matrix $\boldsymbol{X}$ restricted to columns indexed by S. With minor assumptions about the data generating process and without any sparsity assumption on the vector of regression coefficients, [Rahnama Rad and Maleki, 2019] (Theorem 3 and Corollary 1) proved that for various regularizers and regression methods $|\mathrm{ALO} - \mathrm{LO}| = O_p(\frac{\mathrm{poly}\log \mathrm{n}}{n})$ in the high dimensional setting where $n/p = \delta$ is constant while $n, p \to \infty$.

Our finite sample bounds in the next section show that with similar (easy to check) regularity conditions, for various regularizers and regression methods, $|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}| \to 0$ estimate go to zero as $n, p \to \infty$ but $n/p = \delta$ is a fixed number. As a byproduct of this result, we show that in this high dimensional regime $|\mathrm{ALO} - \mathrm{Err}_{\mathrm{out}}| \to 0$ as $n, p \to \infty$. We will more formally state these claims in Section 3.

2 Main results

2.1 Our assumptions

Our goal is to evaluate the accuracy of LO in estimating the out-of-sample prediction error in the high-dimensional regime. Our results are valid for finite values of n and p. Later, in order to make asymptotic conclusions, we suppose that $n/p = \delta$ is constant while $n, p \to \infty$.

We now state our assumptions for theorem 1. For simplicity of exposition, we start by stating a *strong* version of our assumptions, which often requires uniform bounds. Weaker analogues are discussed in 2.3. As the assumptions may appear somewhat opaque and technical, we will discuss them in the context of usual assumptions and concrete examples of standard generalized linear models.

Assumption 1. *The vectors $\boldsymbol{x}_i$ are independent zero mean vectors with covariance $\boldsymbol{\Sigma} \in \mathrm{R}^{p\times p}$ such that $\sigma_{\max}(\boldsymbol{\Sigma}) \leq \rho/p$ for a nonnegative constant ρ.*

Assumption 1 characterizes the different distributions obtained for each n and p. The rows $\boldsymbol{x}_i^\top$ are scaled in a way that ensures $\mathbb{E}\,\|\boldsymbol{x}_i\|_2^2 = O(1)$ and $\mathrm{Var}(\boldsymbol{x}_i^\top\boldsymbol{\beta}) = \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} = O(1)$, assuming that β_i (for $i = 1, \cdots, p$) is $O(1)$, e.g. $\|\boldsymbol{\beta}\|_2^2 = O(p)$. For instance, under the linear model $y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \epsilon_i$, this scaling ensures that the signal-to-noise ratio in each observation remains fixed as n, p grow (when the noise variance is a non-zero constant). Unless we make explicit assumptions about the sparsity of $\boldsymbol{\beta}$, without the $1/p$ scaling, the Hessian of the optimization problem (1) is dominated by the data, making the regularizer, and in turn λ, irrelevant. In this paper, we make *no sparsity assumption* on the vector of regression coefficients. For similar finite signal-to-noise ratio scalings in the high-dimensional asymptotic analysis see [El Karoui, 2017, El Karoui et al., 2013, Bean et al., 2013, Donoho and Montanari, 2016, Donoho et al., 2011, Bayati and Montanari, 2012, Nevo and Ritov, 2016, Su et al., 2017, Dobriban and Wager, 2018, Rahnama Rad and Maleki, 2019, Xu et al., 2019]. Under this scaling, the optimal value of λ will be $O_p(1)$ [Mousavi et al., 2018].

Assumption 2. *We assume the functions $\ell(y \mid z)$ and $\phi(y,z)$ are twice differentiable in z. We also assume that $\ell(y \mid z)$ and $r(\boldsymbol{\beta})$ are convex in z and $\boldsymbol{\beta}$, respectively. Let $(y_o, \boldsymbol{x}_o)$ be a sample from the unknown distribution $q(y, \boldsymbol{x})$ independent of $D = \{(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)\}$. We assume there exists constants c_0 and c_1, such that, for all i, j, uniformly:*

$$c_0 \geq \max\left(|\dot{\ell}(y_i \mid \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})|, |\dot{\ell}(y_o \mid \boldsymbol{x}_o^\top\hat{\boldsymbol{\beta}}_{/i})|\right),$$

$$c_1 \geq \sup_{t\in[0,1]} \sqrt{\mathbb{E}\big[\dot{\phi}(y_o, t\boldsymbol{x}_{\boldsymbol{o}}^\top\hat{\boldsymbol{\beta}}_{/i} + \bar{t}\boldsymbol{x}_{\boldsymbol{o}}^\top\hat{\boldsymbol{\beta}}_{/ij})^2 \mid D_{/i}\big]},$$

$$c_1 \geq \sup_{t\in[0,1]} \sqrt{\mathbb{E}\,\dot{\phi}(y_o, t\boldsymbol{x}_{\boldsymbol{o}}^\top\hat{\boldsymbol{\beta}}_{/1} + \bar{t}\boldsymbol{x}_{\boldsymbol{o}}^\top\hat{\boldsymbol{\beta}}_{/1,2})^2},$$

$$c_1 \geq \sup_{t\in[0,1]} \sqrt{\mathbb{E}\big[\dot{\phi}(y_o, t\boldsymbol{x}_o^\top\hat{\boldsymbol{\beta}} + \bar{t}\boldsymbol{x}_o^\top\hat{\boldsymbol{\beta}}_{/i})^2 \mid D\big]},$$

where $D_{/i} \triangleq D \setminus \{(y_i, \boldsymbol{x}_i)\}$, and $\bar{t} = 1 - t$.

Assumption 2 characterizes the smoothness of the GLM problem (and its associated leave-one-out versions). As we will show below there are many examples, such as logistic and robust regression, in which we can find c_0 and c_1. However, in some other popular examples, such as linear or Poisson regression, $|\dot{\ell}(y_i \mid \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})|$ is a random quantity and we cannot find an absolute constant to dominate it everywhere. As will be discussed later in Section 2.3, we can weaken Assumption 2 at the expense of a slightly stronger moment condition on the feature vector $\boldsymbol{x}_i$.

Example 1. In the generalized linear model family, for the negative logistic regression log-likelihood $\ell(y \mid \boldsymbol{x}^\top\boldsymbol{\beta}) = -y\boldsymbol{x}^\top\boldsymbol{\beta} + \log(1 + e^{\boldsymbol{x}^\top\boldsymbol{\beta}})$, where $y \in \{0, 1\}$, for $\phi(y, z) = \ell(y \mid z)$ it is easy to show that $\dot{\ell}(y \mid z) \leq 2$ for any y and z, leading to $c_0 = c_1 = 2$.

Example 2. Our next example is about a smooth approximation of the Huber loss used in robust estimation, known as the pseudo-Huber loss:

$$f_H(z) = \gamma^2\left(\sqrt{1 + \frac{z^2}{\gamma^2}} - 1\right), \tag{6}$$

where $\gamma > 0$ is a fixed number. If we use this loss for the linear regression problem, and set $\ell(y \mid \boldsymbol{x}^\top\boldsymbol{\beta}) = \phi(y, \boldsymbol{x}^\top\boldsymbol{\beta}) = f_H(y - \boldsymbol{x}^\top\boldsymbol{\beta})$. It is easy to show that $\dot{\ell}(y \mid z) \leq \gamma$ for any y and z, leading to $c_0 = c_1 = \gamma$.

Our next example is concerned with another popular loss function in linear regression, namely the absolute deviation. However, since we would like our loss functions to be differentiable, we use the following smooth approximation of the absolute deviation loss, $\ell(y \mid z) = |y - z|$, introduced in [Schmidt et al., 2007]:

$$\ell_\gamma(y \mid z) \triangleq \frac{1}{\gamma}\Big(\log(1 + e^{\gamma(y-z)}) + \log(1 + e^{-\gamma(y-z)})\Big),$$

where $\gamma > 0$ is fixed.[1]

Example 3. For $\ell(y \mid \boldsymbol{x}^\top\boldsymbol{\beta}) = \phi(y, \boldsymbol{x}^\top\boldsymbol{\beta}) = \ell_\gamma(y \mid z)$, we have $c_0 = c_1 = 1$. In fact, it is straightforward to show that $\dot{\ell}_\gamma(y \mid z) \leq 1$ for any y and z.

Assumption 3. *For $t \in [0, 1]$ define the two matrices*

$$\begin{aligned}\boldsymbol{A}_{t,/i} &\triangleq \boldsymbol{X}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}})]\boldsymbol{X}_{/i} \\ &\quad + \lambda\boldsymbol{\nabla}^{\boldsymbol{2}}\boldsymbol{r}(t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}}), \\ \boldsymbol{A}_{t,/i,j} &\triangleq \boldsymbol{X}_{/ij}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/ij}(t\hat{\boldsymbol{\beta}}_{/ij} + (1-t)\hat{\boldsymbol{\beta}}_{/i})]\boldsymbol{X}_{/ij} \\ &\quad + \lambda\boldsymbol{\nabla}^{\boldsymbol{2}}\boldsymbol{r}(t\hat{\boldsymbol{\beta}}_{/ij} + (1-t)\hat{\boldsymbol{\beta}}_{/i}).\end{aligned} \tag{7}$$

We assume that there exists a fixed number ν, such that

$$\nu \leq \min\Big(\min_{1\leq i\leq n}\inf_{t\in[0,1]}\sigma_{\min}(\boldsymbol{A}_{t,/i}), \min_{1\leq i\leq n}\inf_{t\in[0,1]}\sigma_{\min}(\boldsymbol{A}_{t,/i,j})\Big).$$

[1] Note that $\lim_{\gamma\to\infty}\sup_{y,z}\big||y - z| - \ell_\gamma(y \mid z)\big| = 0$.

Assumption 3 characterizes the curvature of the GLM problem (and its associated leave-one-out versions). In some examples, such as the ones that have ridge or smoothed elastic-net as the regularizer, it is straightforward to confirm this assumption. For instance, for the ridge regularization, $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2/2$, we have $\nu > \lambda$. In Section 2.3, we explain how this assumption can be relaxed (at the expense of requiring more stringent moment conditions on $\boldsymbol{x}_i$) to cover more examples.

Having stated our assumptions, we now move on to stating our main result before proposing a number of examples to demonstrate how this result can be applied in common GLM cases.

2.2 Main theorem

Based on these assumptions we can now evaluate the accuracy of LO in estimating $\mathrm{Err}_{\mathrm{out}}$. The following theorem proves that the expected square error of LO in estimating $\mathrm{Err}_{\mathrm{out}}$ is small even in high-dimensional asymptotic settings.

Theorem 1. *Let $\delta \triangleq n/p$. If Assumptions 1, 2 and 3 hold, then*

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}) - \mathbb{E}\big[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}) \mid D\big]\right)^2 \leq \frac{C_v}{n},$$

where the outer expectation is taken with respect to the data D and:

$$C_v = \mathbb{E}\,\mathrm{Var}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}_{/1}) \mid D_{/1}] + 2C_b + 2C_b^{1/2}\sqrt{\mathbb{E}\,\mathrm{Var}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}_{/1}) \mid D_{/1}] + C_b},$$

and $C_b = \left(\frac{c_0 c_1 \rho \delta^{1/2}}{\nu}\right)^2$.

The proof can be found in Appendix C.

The only term that is not explicitly computed in terms of the constants in our assumptions is $\mathbb{E}\,\mathrm{Var}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}_{/i}) \mid D_{/i}]$. Hence, to obtain an explicit quantitative bound for a specific GLM problem requires computing this quantity. We present two examples below.

Corollary 1. *(Ridge regularized logistic regression) Consider the negative logistic regression log-likelihood $\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta}) = -y\boldsymbol{x}^\top\boldsymbol{\beta} + \log(1+e^{\boldsymbol{x}^\top\boldsymbol{\beta}})$, and the regularizer $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2/2$, where $y \in \{0,1\}$. Furtherassume that $\boldsymbol{x}_i$ is iid $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\sigma_{\max}(\boldsymbol{\Sigma}) \leq \frac{\rho}{p}$. If $\phi(y,z) = \ell(y|z)$, then there exists a constant C_v such that*

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}) - \mathbb{E}\big[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}) \mid D\big]\right)^2 \leq \frac{C_v}{n},$$

where

$$\begin{aligned} C_v &= 6 + \frac{5\rho\delta}{\lambda} + 2\left(\frac{4\rho\delta^{1/2}}{\lambda}\right)^2 \\ &+ 2\left(\frac{4\rho\delta^{1/2}}{\lambda}\right)\sqrt{6 + \frac{5\rho\delta}{\lambda} + \left(\frac{4\rho\delta^{1/2}}{\lambda}\right)^2}. \quad (8)\end{aligned}$$

The proof of this corollary can be found in Section D of the supplementary material.

Corollary 2. *(Pseudo-Huber loss with strongly convex regularizer) We consider again the pseudo-Huber loss defined in (6) with parameter γ. As this loss is typically used in regression settings, we consider a linear regression model $y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}^* + \epsilon_i$, where ϵ_i denotes i.i.d. zero-mean noise, and $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\sigma_{\max}(\boldsymbol{\Sigma}) \leq \frac{\rho}{p}$. We additionally assume that the regularizer is strongly convex with parameter ν_r,[2] $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2$, and $\frac{1}{p}\boldsymbol{\beta}^{*\top}\boldsymbol{\beta}^* \leq b$. Under these conditions, there exists a fixed number C_v (depending on γ, σ_ϵ, b, ρ, δ and ν_r) such that*

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}) - \mathbb{E}\big[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}) \mid D\big]\right)^2 \leq \frac{C_v}{n}.$$

The proof of this corollary can be found in Section E of the supplementary material.

To summarize, the examples presented in Corollary 1 and 2 satisfy the assumption needed for Theorem 1.

2.3 Extensions

As we discussed in Section 2.1, we can weaken the assumptions without a major change in our proofs or the main conclusions of our result. In this section, we aim to present one such weaker set of assumptions that enables our analyses to cover several other popular examples, such as the Poisson and linear regression.

Assumption 1′. *We assume that $\boldsymbol{x}_i$ are i.i.d. zero mean vectors with covariance $\boldsymbol{\Sigma} \in \mathrm{R}^{p\times p}$ such that $\sigma_{\max}(\boldsymbol{\Sigma}) \leq \rho/p$ for a non-negative constant ρ. Furthermore, there exists a fixed number c_4, such that $\mathbb{E}(\|\boldsymbol{x}_i\|_2^4) \leq c_4$.*

Note that this assumption is more stringent than Assumption 1. However, in essence the only extra requirement of this assumption is a bound on the fourth moments. Hence, it holds for a wide range of random features including sub-Gaussian and sub-exponential features. Thanks to this slightly stronger moment assumption we can weaken the other

[2] Note that this is a fairly benign assumption in practice: it is common to introduce a slight ridge penalty which automatically satisfies this assumption.

assumptions.

Assumption 2′. *We assume the functions $\ell(y \mid z)$ and $\phi(y, z)$ are twice differentiable in z. Moreover, assume $\ell(y \mid z)$ and $r(\boldsymbol{\beta})$ are convex in z and $\boldsymbol{\beta}$, respectively. Let $(y_o, \boldsymbol{x}_o)$ be a sample from the unknown distribution $q(y, \boldsymbol{x})$ independent of $D = \{(y_1, \boldsymbol{x}_1), \cdots, (y_n, \boldsymbol{x}_n)\}$. We assume that there exist constants $\tilde{c}_0$ and $\tilde{c}_1$, such that for all i, j, uniformly*

$$\tilde{c}_0 \geq \mathbb{E}|\dot{\ell}(y_1 \mid \boldsymbol{x}_1^\top \hat{\boldsymbol{\beta}})|^8,$$

$$\tilde{c}_0 \geq \mathbb{E}|\dot{\ell}(y_o \mid \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}_{/1})|^8,$$

$$\tilde{c}_1 \geq \sup_{t \in [0,1]} \sqrt{\mathbb{E}\left[\dot{\phi}(y_o, t\boldsymbol{x_o}^\top \hat{\boldsymbol{\beta}}_{/i} + \bar{t}\boldsymbol{x_o}^\top \hat{\boldsymbol{\beta}}_{/ij})^2 \mid D_{/i}\right]},$$

$$\tilde{c}_1 \geq \sup_{t \in [0,1]} \sqrt{\mathbb{E}\, \dot{\phi}(y_o, t\boldsymbol{x_o}^\top \hat{\boldsymbol{\beta}}_{/1} + \bar{t}\boldsymbol{x_o}^\top \hat{\boldsymbol{\beta}}_{/1,2})^2},$$

$$\tilde{c}_1 \geq \sup_{t \in [0,1]} \sqrt{\mathbb{E}[\dot{\phi}(y_o, t\boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}} + \bar{t}\boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}_{/i})^2 \mid D]},$$

where $D_{/i} \triangleq D \setminus \{(y_i, \boldsymbol{x}_i)\}$, and $\bar{t} = 1 - t$.

Compared to Assumption 2 that requires $|\dot{\ell}(y_i \mid \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})|$ to be bounded everywhere, this assumption requires the 8th moment of $|\dot{\ell}(y_i \mid \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})|$ to be bounded. This simple modification enables our theoretical results to be applied to a much broader set of regression techniques, including Poisson, linear, and negative binomial regression. These three examples will be studied later in this section.

Assumption 3′. *Let $\boldsymbol{A}_{t,/1}$ and $\boldsymbol{A}_{t,/1,2}$ be as defined in Assumption 3. We assume that there exists a fixed number $\tilde{\nu}$, such that*

$$\tilde{\nu} \geq \mathbb{E}\left(\inf_{t \in [0,1]} \sigma_{\min}\left(\boldsymbol{A}_{t,/1}\right)\right)^{-8},$$

$$\tilde{\nu} \geq \mathbb{E}\left(\inf_{t \in [0,1]} \sigma_{\min}\left(\boldsymbol{A}_{t,/1,2}\right)\right)^{-8}.$$

Again, compared to Assumption 3, this assumption only bounds the moments of the minimum eigenvalue of the matrix. The following example shows an example in which it is impossible to find a positive lower bound for the minimum eigenvalue, but still the moments of the inverse of the minimum eigenvalue are bounded.

Example 1. *Suppose that $\delta = n/p > 1$ and that the loss function is strongly convex with parameter c, and the regularizer is convex. Finally, suppose that $\boldsymbol{x}_i \sim N(0, \boldsymbol{\Sigma})$, with $\sigma_{\min}(\boldsymbol{\Sigma}) = \frac{\rho}{p}$. Then, there exists a fixed number $\tilde{\nu}$ that satisfies Assumption 3′ for large enough values of n and p.*

The proof can be found in Section F of the supplementary material.

As we discussed before one can prove the accuracy of LO under Assumptions 1′, 2′, and 3′. The following theorem formalizes this claim.

Theorem 2. *Let $\delta \triangleq n/p$. If Assumptions 1′, 2′ and 3′, then*

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}) - \mathbb{E}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}) \mid D]\right)^2 \leq \frac{\tilde{C}_v}{n},$$

where the outer expectation is taken with respect to the data D and:

$$\tilde{C}_v = \mathbb{E}\,\mathrm{Var}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}_{/1}) \mid D_{/1}] + 2\tilde{C}_b + 2\tilde{C}_b^{1/2}\sqrt{\mathbb{E}\,\mathrm{Var}\big[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}_{/1}) \mid D_{/1}\big] + \tilde{C}_b}.$$

and $\tilde{C}_b = c_1^2 \rho \delta_0 \tilde{c}_0 \tilde{v} c_4$.

The proof can be found in Section G of the supplementary material. As we described before, this theorem can cover several generalized linear models, that could not be covered by Theorem 1. We mention three important examples below.

Corollary 3. *(Square loss with elastic-net penalty) Consider the data generating mechanism $y_i = \boldsymbol{x}_i \boldsymbol{\beta}^* + \epsilon_i$, where $\boldsymbol{x}_i^\top \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, and $\frac{1}{p}\|\boldsymbol{\beta}^*\|_2^2 \leq b$. Suppose that we use the smoothed elastic-net optimization*

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{n} \frac{(y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2}{2} + \lambda \sum_{j=1}^{p} r(\beta_i),$$

where for $\gamma > 0$, $r(\beta) = \gamma\beta^2 + (1-\gamma) r^\alpha(\beta)$, and $r^\alpha(\beta) = \frac{1}{\alpha}\Big(\log(1 + e^{\alpha\beta}) + \log(1 + e^{-\alpha\beta})\Big)$ is a smooth approximation of the ℓ_1-norm. Then, there exists a fixed number, $\tilde{C}_v$, such that

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}) - \mathbb{E}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}) \mid D]\right)^2 \leq \frac{\tilde{C}_v}{n}.$$

Since the proof of this claim is long, we defer it to Section H of the supplementary material.

Corollary 4. *[Poisson regression with soft-rectifying link] Consider the data-generating mechanism $y_i \sim \mathrm{Poisson}(f(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*))$, where $f(z) = \log(1 + e^z)$ denotes the soft-rectifying link, $\boldsymbol{x}_i \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\frac{1}{p}\boldsymbol{\beta}^{*\top}\boldsymbol{\beta}^* \leq b$. Finally, assume that r denotes the smoothed elastic-net regularizer introduced in Corollary 3. Under these assumptions, there exists a fixed number, $\tilde{C}_v$, such that:*

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}) - \mathbb{E}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}) \mid D]\right)^2 \leq \frac{\tilde{C}_v}{n}.$$

The proof can be found in Section I of the supplementary file.

Remark 1. *We have assumed here that $\boldsymbol{x}_i$ is multivariate Gaussian. As might be clear to the reader from the proof, this normality assumption on $\boldsymbol{x}$ may be relaxed to an 8th moment assumption at the cost of a slightly more complicated proof.*

Corollary 5 (Negative-Binomial Regression)**.** *We consider the problem of negative binomial regression with fixed shape parameter $\alpha > 0$ and exponential link. Here, the negative log-likelihood is given by:*

$$\ell(y \mid z) = (y + \alpha^{-1}) \log(1 + \alpha e^z) - yz + C(\alpha, y),$$

where $C(\alpha, y)$ denotes a constant which only depends on α and y. Assume the data generating process is such that $\mathbb{E}[y^8] \le \kappa$, and that $\phi(y, z) = \ell(y|z)$. Finally, similar to Corollary, 3 we use the smoothed elastic-net as the regularizer. Under these assumptions, there exists a fixed number, $\tilde{C}_v$, such that

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i}) - \mathbb{E}[\phi(y_o, \boldsymbol{x}_o^\top \hat{\boldsymbol{\beta}}) \mid D]\right)^2 \le \frac{\tilde{C}_v}{n}.$$

The proof can be found in Section J of the supplementary material.

3 Connection of ALO and $\mathrm{Err}_{\mathrm{out}}$

We mentioned in Section 1.2 that different approximations of LO have been proposed in the literature to reduce the computational complexity of LO. Among such approximations, the ALO formula introduced in (5), is analyzed in [Rahnama Rad and Maleki, 2019] under a similar asymptotic framework as the one discussed in our paper:

Theorem 3. *[Rahnama Rad and Maleki, 2019] Suppose that $n/p = \delta$ is constant while $n, p \to \infty$. Under the assumption $\boldsymbol{x}_i \sim N(0, \boldsymbol{\Sigma})$, for the regression problems discussed in Corollaries 1, 2, 3, and 4 we have*

$$|\mathrm{ALO} - \mathrm{LO}| = O_p\left(\frac{\text{poly} \log \text{n}}{\sqrt{n}}\right).$$

Note that the ultimate goal of ALO is to use it as an estimate of $\mathrm{Err}_{\mathrm{out}}$. Hence, while Theorem 3 confirms the accuracy of ALO in approximating LO it does not explain whether the estimates obtained by ALO or LO can be trusted in high-dimensional settings. However, we can combine this result with Theorems 1 and 2 to prove the accuracy of ALO in estimating $\mathrm{Err}_{\mathrm{out}}$. Toward this goal we first prove the following claim.

Theorem 4. *Suppose that $n/p = \delta$ is constant while $n, p \to \infty$. Under the assumption $\boldsymbol{x}_i \sim N(0, \boldsymbol{\Sigma})$, for the regression problems discussed in Corollaries 1, 2, 3, and 4 we have*

$$|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. For a fixed number M

$$\begin{aligned} &\mathrm{P}\Big(\big|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}\big| > \frac{M}{\sqrt{n}}\Big) \\ &\quad = \mathrm{P}\Big(\big|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}\big|^2 > \frac{M^2}{n}\Big) \\ &\quad \le \frac{n}{M^2}\,\mathbb{E}\,|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}|^2 \\ &\quad \le \frac{n}{M^2}\,\frac{\min(C_\nu, \tilde{C}_\nu)}{n} = \frac{\min(C_\nu, \tilde{C}_\nu)}{M^2}. \qquad (9) \end{aligned}$$

The first inequality in the above equations is due to Markov inequality, and the second inequality is a result of Theorems 1 and 2. As we discussed in Corollaries 1, 2, 3, and 4 either C_ν or $\tilde{C}_\nu$ are finite numbers. Hence, as M increases, the final probability can be reduced to the desired level. □

Before we proceed to establish the accuracy of ALO we have to clarify Theorem 4. Note that even under the idealized (but incorrect) assumption that the individual estimates $\phi(y_i, \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{/i})$ are independent and $\hat{\boldsymbol{\beta}}_{/i}$s are the same as $\hat{\boldsymbol{\beta}}$, the central limit theorem indicates that $|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}| \sim \frac{1}{\sqrt{n}}$.[3] Hence, we should not expect the error of LO to be $o_p(\frac{1}{\sqrt{n}})$. Therefore, the above theorem seems to offer the sharpest result that is possible for LO. Note that the sharpness is with regard to the rate of convergence and not the constants.

Combining the results of Theorem 2 and Theorem 4 we can finally quantify the accuracy of ALO in estimating $\mathrm{Err}_{\mathrm{out}}$.

Corollary 6. *Suppose that $n/p = \delta$ is constant while $n, p \to \infty$. Under the assumption $\boldsymbol{x}_i \sim N(0, \boldsymbol{\Sigma})$, for the regression problems discussed in Corollaries 1, 2, 3, and 4 we have*

$$|\mathrm{ALO} - \mathrm{Err}_{\mathrm{out}}| = O_p\left(\frac{\text{poly} \log \text{n}}{\sqrt{n}}\right).$$

The proof of this corollary is straightforward, and is hence skipped. Note that this corollary finally establishes the fact that ALO obtains accurate estimates of $\mathrm{Err}_{\mathrm{out}}$. While we have established this result for only four popular examples in this paper, Theorems 1, 2 and Theorem 3 of [Rahnama Rad and Maleki, 2019] can be applied to a much broader class of regression problems. Hence, a similar result is expected for such scenarios as well. Finally, we should emphasize that by comparing Theorems 4 and 6 one may notice that the accuracy of ALO might be worse than LO by a logarithmic factor. At this stage, it is not clear whether this difference is an artifact of the proof of [Rahnama Rad and Maleki,

[3] The notation $|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}| \sim \frac{1}{\sqrt{n}}$ means that we have both $|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}| \sim O_p(\frac{1}{\sqrt{n}})$ and $\frac{1}{\sqrt{n}} = O_p(|\mathrm{LO} - \mathrm{Err}_{\mathrm{out}}|)$.

n	p	MSE (SE)
40	400	0.0156 (0.0021)
80	800	0.0064 (0.0008)
120	1200	0.0039 (0.0006)
160	1600	0.0038 (0.0006)
200	2000	0.0028 (0.0004)

Table 1: Square loss with elastic-net penalty: MSE$\triangleq$ $\mathbb{E}(\mathrm{Err}_{\mathrm{out}} - \mathrm{LO})^2$ (and standard errors).

2019] or it is a real extra error that has been introduced by the approximation of LO.

4 Numerical Experiments

In this section, we present two numerical experiments to show that the $O(\frac{1}{n})$ bound given in Theorem 1 and 2 is sharp but not tight. Specifically, we generate synthetic data, and compare $\mathrm{Err}_{\mathrm{out}}$ and LO for elastic-net linear regression and ridge logistic regression. In all the examples in this section, the rows of $\boldsymbol{X}$ are $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Here we let $\boldsymbol{\Sigma} = \boldsymbol{I}/n$ and $\phi(y, z) = \ell(y \mid z)$. The codes for the Figure 1 and Table 1,2 are available at `https://github.com/RahnamaRad/LO`.

Square loss with elastic-net penalty. We set $\ell(y \mid \boldsymbol{x}^\top\boldsymbol{\beta}) = \frac{1}{2}(y - \boldsymbol{x}^\top\boldsymbol{\beta})^2$, $r(\boldsymbol{\beta}) = \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1$ and $\alpha = 0.5$. The true unknown parameter vector $\boldsymbol{\beta}^* \in \mathrm{R}^p$ is sparse with $k = 0.1n$ non-zero elements independently drawn from a zero mean unit variance Laplace distribution, leading to $\mathrm{Var}(\boldsymbol{x}^\top\boldsymbol{\beta}^*) = 0.1$ (regardless of the values of n and p). To generate data, we sample $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}^*, \sigma^2\boldsymbol{I})$. Here the out-of-sample error is:

$$\mathrm{Err}_{\mathrm{out}} = \mathbb{E}\,\ell(y_o \mid \boldsymbol{x}_o^\top\boldsymbol{\beta}) = \sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2.$$

As we increase n and p, we keep the ratio $\delta = n/p = 0.1$ constant. We numerically calculate MSE$\triangleq$ $\mathbb{E}(\mathrm{Err}_{\mathrm{out}} - \mathrm{LO})^2$ as a function of n (and $p = 10n$) based on 100 synthetic data samples, for each n, p and $\lambda = 5$. We fitted a line to model $\log(\mathrm{MSE}) \sim \log(n)$ and obtained a slope of -1.03 (SE= 0.04) and intercept of -0.46 (SE= 0.54) with an Adjusted R-squared of 0.95. The slope of -1.03 ($SE = 0.04$) shows that the bound is sharp because it confirms the $1/n$ scaling of our theory. Table 1 shows the numerical MSE as a function of n and p.

Logistic regression with ridge penalty. We set $\ell(y \mid \boldsymbol{x}^\top\boldsymbol{\beta}) = -y\boldsymbol{x}^\top\boldsymbol{\beta} + \log(1 + e^{\boldsymbol{x}^\top\boldsymbol{\beta}})$ (the negative logistic log-likelihood) and $r(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2$. To generate data, we sample $y_i \sim \mathit{Binomial}\left(\frac{e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}^*}}{1+e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}^*}}\right)$. Here the out-of-sample error

$$\begin{aligned}\mathrm{Err}_{\mathrm{out}} &= \mathbb{E}\,\ell(y_o|\boldsymbol{x}_o^\top\hat{\boldsymbol{\beta}}) \\ &= -\frac{\hat{\boldsymbol{\beta}}^\top\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|_2^2}\,\mathbb{E}\left[\frac{Ze^Z}{1+e^Z}\right] + \mathbb{E}\log(1+e^W)\end{aligned}$$

where $Z \sim \mathrm{N}(0, \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*\|_2^2)$ and $W \sim \mathrm{N}(0, \|\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\beta}}\|_2^2)$.

n	p	MSE (SE)	Bound
100	100	0.0136 (0.0019)	63.12
300	300	0.0037 (0.0005)	21.04
500	500	0.0026 (0.0005)	12.62
700	700	0.0017 (0.0002)	9.02
900	900	0.0015 (0.0002)	7.01
1100	1100	0.0012 (0.0002)	5.74

Table 2: Logistic regression with ridge penalty: MSE$\triangleq$ $\mathbb{E}(\mathrm{Err}_{\mathrm{out}} - \mathrm{LO})^2$ (and standard errors) and the upper bound based on 8 in Corollary 1 of Theorem 1.

As we increase n and p, we keep the ratio $n/p = 1$ constant. We numerically calculate MSE$\triangleq$ $\mathbb{E}(\mathrm{Err}_{\mathrm{out}} - \mathrm{LO})^2$ as a function of n (and $p = n$) based on 100 synthetic data samples, for each n, p and $\lambda = 0.1$. We fitted a line to model $\log(\mathrm{MSE}) \sim \log(n)$ and obtained a slope of -1.00 (SE= 0.04) and intercept of 0.34 (SE= 0.27) with an Adjusted R-squared of 0.99. The slope of -1.00 shows that the bound is sharp because it confirms the $1/n$ scaling of our theory. Table 2 compares the numerical MSE and the theoretical bound from Theorem 1 and Corollary 1. The theoretical upper bound was computed using 8 in Corollary 1 where in this example, we have $\lambda = 0.1$, $\rho = 1$, and $\delta = 1$, leading to $C_v = 6311.52$. The significant difference between the bound and the MSE shows that the bound is not tight.

5 Conclusion

Leave-one-out estimators (and their approximate versions) have seen renewed interest recently in the context of big data and high-dimensional problems. We show that, in general, leave-one-out risk estimators have desirable statistical behaviours in the high-dimensional setting. Although the leave-out-risk estimator itself is generally computationally intractable, this result also implies consistency for a (growing) number of approximate leave-one-out estimators, and demonstrate that such estimators offer a potentially good direction for building risk estimators for high-dimensional problems.

Acknowledgement

K.R. was supported by the NSF DMS grant 1810888, and Eugene M. Lang Junior Faculty Research Fellowship. A.M. was supported by the NSF DMS grant 1810888.

References

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

[Bayati and Montanari, 2012] Bayati, M. and Montanari, A. (2012). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

[Bean et al., 2013] Bean, D., Bickel, P. J., El Karoui, N., and Yu, B. (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568.

[Beirami et al., 2017] Beirami, A., Razaviyayn, M., Shahrampour, S., and Tarokh, V. (2017). On optimal generalizability in parametric learning. *NIPS*, pages 3455–3465.

[Burman, 1990] Burman, P. (1990). Estimation of generalized additive models. *Journal of Multivariate Analysis*, 32:230–255.

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

[Cessie and Houwelingen, 1992] Cessie, S. and Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

[Craven and Wahba, 1979] Craven, P. and Wahba, G. (1979). Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403.

[Dobriban and Wager, 2018] Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Stat.*, 46(1):247–279.

[Donoho et al., 2011] Donoho, D., Maleki, A., and Montanari, A. (2011). Noise sensitivity phase transition. *IEEE Trans. Inform. Theory*, 57(10).

[Donoho and Montanari, 2016] Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probab Theory Relat Fields*, 166(3-4):935–969.

[El Karoui, 2017] El Karoui, N. (2017). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory Related Fields*, 170(1-2):95–175.

[El Karoui et al., 2013] El Karoui, N., Bean, D., Bickel, P., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *PNAS*, 110(36):14557–14562.

[Giordano et al., 2019] Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. (2019). A swiss army infinitesimal jackknife. *JMLR*, 89(1139-1147).

[Golub et al., 1979] Golub, G., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

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

[Mousavi et al., 2018] Mousavi, A., Maleki, A., and Baraniuk, R. (2018). Consistent parameter estimation for lasso and approximate message passing. *The Annals of Statistics*, 46(1):119–148.

[Nevo and Ritov, 2016] Nevo, D. and Ritov, Y. (2016). On Bayesian robust regression with diverging number of predictors. *Electron. J. Statist.*, 10(2):3045–3062.

[Obuchi and Kabashima, 2016] Obuchi, T. and Kabashima, Y. (2016). Cross validation in lasso and its acceleration. *J. Stat. Mech. Theor. Exp.*, 53(304):1–36.

[Obuchi and Kabashima, 2018] Obuchi, T. and Kabashima, Y. (2018). Accelerating Cross-Validation in Multinomial Logistic Regression with ell1-Regularization. *Journal of Machine Learning Research*, 19:1–30.

[Opper and Winther, 2000] Opper, M. and Winther, O. (2000). Gaussian processes and SVM: Mean field results and leave-one-out. In Smola, A., Bartlett, P., Scholkopf, B., and Schuurmans, D., editors, *Advances Large Margin Classifiers*, pages 43–56. MIT Press, Cambridge, MA.

[O'Sullivan et al., 1986] O'Sullivan, F., Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *JASA*, 81(393):96–103.

[Rahnama Rad and Maleki, 2019] Rahnama Rad, K. and Maleki, A. (2019). A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv:1801.10243v3*.

[Schmidt et al., 2007] Schmidt, M., Fung, G., and Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *ECML*, pages 286–297. Springer.

[Stephenson and Broderick, 2019] Stephenson, W. and Broderick, T. (2019). Sparse Approximate Cross-Validation for High-Dimensional GLMs. *arXiv preprint arXiv:1905.13657*.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assesment of statistical predictions. *J R Stat Soc Series B*, 36(2):111–147.

[Su et al., 2017] Su, W., Bogdan, M., and Candes, E. (2017). False discoveries occur early on the Lasso path. *Ann. Stat.*, 45(5):2133–2150.

[Takahashi and Kabashima, 2018] Takahashi, T. and Kabashima, Y. (2018). A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment*, 7(073405).

[Vehtari et al., 2016] Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research*, 17(1):3581–3618.

[Wang et al., 2018] Wang, S., Zhou, W., Maleki, A., Lu, H., and Mirrokni, V. (2018). Approximate Leave-One-Out for High-Dimensional Non-Differentiable Learning Problems. *International Conference on Machine Learning*.

[Xu et al., 2019] Xu, J., Maleki, A., Rahnama Rad, K., and Hsu, D. (2019). Consistent risk estimation in high-dimensional linear regression. *arXiv:1902.01753*.