Iterative Approximate Cross-Validation

Yuetian Luo [1]   Zhimei Ren [2]   Rina Foygel Barber [2]

Abstract

Cross-validation (CV) is one of the most popular tools for assessing and selecting predictive models. However, standard CV suffers from high computational cost when the number of folds is large. Recently, under the empirical risk minimization (ERM) framework, a line of works proposed efficient methods to approximate CV based on the solution of the ERM problem trained on the full dataset. However, in large-scale problems, it can be hard to obtain the exact solution of the ERM problem, either due to limited computational resources or due to early stopping as a way of preventing overfitting. In this paper, we propose a new paradigm to efficiently approximate CV when the ERM problem is solved via an iterative first-order algorithm, without running until convergence. Our new method extends existing guarantees for CV approximation to hold along the whole trajectory of the algorithm, including at convergence, thus generalizing existing CV approximation methods. Finally, we illustrate the accuracy and computational efficiency of our method through a range of empirical studies.

1. Introduction

In machine learning and statistics, cross-validation (CV) (Allen, 1974; Stone, 1974; Geisser, 1975) is one of the most popular methods for the tasks of assessing and selecting predictive models. It is conceptually simple and easy to implement. Among many variants of CV, one popular choice is leave-one-out CV (also called the Jackknife), which often offers the most accurate prediction for the out-of-sample risk in many practical and challenging scenarios (Arlot & Celisse, 2010; Stephenson & Broderick, 2020; Rad & Maleki, 2020). Leave-one-out CV uses $n-1$ out of n data

[1]Data Science Institute, University of Chicago, Chicago IL 60637, USA [2]Department of Statistics, University of Chicago, Chicago IL 60637, USA. Correspondence to: Yuetian Luo <yuetian@uchicago.edu>.

points for training, and the remaining one for testing, and then repeats for each of the n data points in the sample. The resulting n cross-validation errors can be used for model assessment or selection, tuning parameter selection, etc.

However, the superior performance of leave-one-out CV is accompanied by high computational cost when n is large, as the potentially complex model needs to be fitted n times. To be able to run leave-one-out CV but avoid the high computational burden, several methods have been proposed to approximate the expensive model refitting step with an inexpensive surrogate step, with much progress in the recent literature particularly in the context of empirical risk minimization (ERM) (Obuchi & Kabashima, 2016; Beirami et al., 2017; Giordano et al., 2019; Koh & Liang, 2017; Wang et al., 2018; Wilson et al., 2020; Rad & Maleki, 2020; Stephenson & Broderick, 2020).

Background: approximate CV for ERM Suppose one aims to estimate some parameter of interest θ^* via solving the following regularized ERM problem:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} F(\mathcal{Z}; \theta), \tag{1}$$

where, for a dataset $\mathcal{Z}$ of size n,

$$F(\mathcal{Z}; \theta) := \sum_{i=1}^{n} \ell(Z_i; \theta) + \lambda \pi(\theta).$$

Here $\ell(Z; \theta)$ is the loss function for data point Z and parameter θ, $\mathcal{Z} = \{Z_i\}_{i=1}^n$ is the set of observed data points, $\pi(\cdot)$ is a regularization term, and $\lambda \geq 0$ is a tuning parameter. Under this setting, the widely used leave-one-out CV loss for estimating the prediction performance of $\widehat{\theta}$ is given by

$$\mathrm{CV}(\{\widehat{\theta}_{-i}\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i; \widehat{\theta}_{-i}), \tag{2}$$

where $\widehat{\theta}_{-i} := \arg\min_{\theta \in \mathbb{R}^p} F(\mathcal{Z}_{-i}; \theta)$ is the minimizer of the leave-one-out objective $F(\mathcal{Z}_{-i}; \theta) := \sum_{j=1, j \neq i}^{n} \ell(Z_j; \theta) + \lambda \pi(\theta)$.

Since computing $\widehat{\theta}_{-i}$ for every $i \in [n] := \{1, \ldots, n\}$ is often expensive, this motivates finding an approximation to $\mathrm{CV}(\{\widehat{\theta}_{-i}\}_{i=1}^n)$ that does not require fitting n many models.

Observe that $\widehat{\theta}$ and $\widehat{\theta}_{-i}$ are the solutions to two highly similar minimization problems—they minimize $F(\mathcal{Z};\theta)$ and $F(\mathcal{Z}_{-i};\theta) = F(\mathcal{Z};\theta) - \ell(Z_i;\theta)$, respectively. One approach in the recent literature proposes approximating $\widehat{\theta}_{-i}$ by initializing at $\theta = \widehat{\theta}$ and taking a single Newton step (NS) on the objective $F(\mathcal{Z}_{-i};\theta)$ (Beirami et al., 2017; Rad & Maleki, 2020),

$$\widehat{\theta}_{-i}^{\mathrm{NS}} = \widehat{\theta} - \left(\nabla_\theta^2 F(\mathcal{Z}_{-i};\widehat{\theta})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i};\widehat{\theta}). \quad (3)$$

Another approach is based on the infinitesimal jackknife (IJ) (Jaeckel, 1972; Efron, 1982; Giordano et al., 2019):

$$\widehat{\theta}_{-i}^{\mathrm{IJ}} = \widehat{\theta} - \left(\nabla_\theta^2 F(\mathcal{Z};\widehat{\theta})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i};\widehat{\theta}). \quad (4)$$

(Here for simplicity we consider the case where F is twice differentiable; we will discuss other settings below.)

The challenge: iterative algorithms The accuracy of the estimators in (3) and (4) relies heavily on the assumption that $\widehat{\theta}$ can be computed exactly, which may not be the case in many scenarios. For example, when (1) is solved via an iterative algorithm such as gradient descent (GD), we may not be able to run the algorithm into convergence due to the limited computational resources. Other algorithms we might use have a very slow rate of convergence, such as stochastic gradient descent (SGD). Another important example is that, in some settings, we may intentionally stop training early to avoid overfitting in learning machine learning models (see e.g. (Jabbar & Khan, 2015)).

Figure 1 shows a simple simulation that illustrates the loss of accuracy suffered by the NS and IJ methods, when run with an estimate of $\widehat{\theta}$ obtained before convergence. Specifically, consider logistic regression with a ridge penalty, with the estimator obtained by running GD or SGD for t steps—that is, we would like to estimate $\mathrm{CV}(\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^n)$, where $\widehat{\theta}_{-i}^{(t)}$ is the solution after t steps of GD or SGD on the objective function $F(\mathcal{Z}_{-i};\theta)$. $\widehat{\theta}_{-i}^{(t)}$ is then approximated by running either NS (3) or IJ (4), but with $\widehat{\theta}^{(t)}$ (the tth step of GD or SGD on the full objective function $F(\mathcal{Z};\theta)$) in place of the exact minimizer $\widehat{\theta}$. We see highly inaccurate approximations of NS and IJ methods to the leave-one-out CV loss $\mathrm{CV}(\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^n)$ during the early iterations of GD, and even after 1000 iterations for SGD. Details of this simulation are given in Section 6.

Our contribution: iterative approximate CV The central question of our paper is this:

Can we develop new approximation schemes to leave-one-out CV when $\widehat{\theta}$ is not known exactly, but is estimated with an iterative algorithm that is not run to convergence?
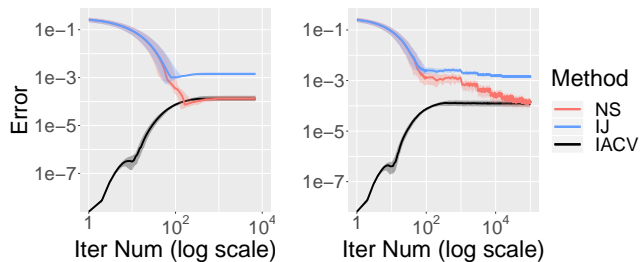

Figure 1. The red and blue lines show the error in estimating $\mathrm{CV}(\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^n)$ for the NS and IJ methods, run with $\widehat{\theta}^{(t)}$ as an approximation to the exact solution $\widehat{\theta}$, while the black line shows the error for our proposed method, iterative approximate CV (IACV). The objective function is given by logistic regression plus a ridge penalty, solved iteratively with GD (left, $n = 1000$) and SGD (right, $n = 1000$, batch-size = 400). Solid lines represent the median value over 100 repetitions and the shaded area shows the region between lower and upper quartiles.

In this work, we provide a positive answer to this question. We find that before the convergence of the iterative algorithm, efficient leave-one-out CV approximation can still be possible if we can leverage information about the iterative nature of the algorithm. In Figure 1, we see that our proposed method, Iterative Approximate CV (IACV), achieves an accurate approximation to leave-one-out CV loss $\mathrm{CV}(\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^n)$ across all iterations t, even far before convergence; at convergence, IACV yields the same results as NS. Theoretically, we are able to show that, under some regularity conditions, IACV enjoys guaranteed CV approximation along the whole trajectory of the algorithm, and moreover, at convergence, IACV recovers the NS method (3).

1.1. Additional Related Literature

In addition to the NS and IJ methods, described earlier, many other works in the literature also offer methods and theories for the problem of approximating leave-one-out CV. The closest to ours is the work by Koh & Liang (2017); Ghosh et al. (2020), which also considers the problem of approximating CV with an inexact $\widehat{\theta}$, but with results of a very different flavor—this line of work assumes that $\widehat{\theta}^{(t)}$ is already an accurate estimate of $\widehat{\theta}$ (i.e., t is large and the iterative algorithm is near convergence), and bounds the error in approximating the leave-one-out models $\widehat{\theta}_{-i}$ as a function of this convergence error $\|\widehat{\theta}^{(t)} - \widehat{\theta}\|_2$. In contrast, our work instead estimates the models $\widehat{\theta}_{-i}^{(t)}$ (i.e., at each time t rather than at convergence), and does not assume that $\|\widehat{\theta}^{(t)} - \widehat{\theta}\|_2$ is small.

Another recent line of work studies the problem of efficient

data deletion in a trained model under the ERM framework, where the aim is to provide an estimator of θ that is approximately independent of data point i, for the sake of the ith individual's privacy. These problems are often referred to as "data deletion", "machine unlearning", or "decremental learning" in the literature (Tsai et al., 2014; Cao & Yang, 2015; Bourtoule et al., 2021; Neel et al., 2021). Like our work, these methods also aim to approximate the leave-one-out model efficiently given the trained full model, but additionally require that the approximated model is "statistically indistinguishable" from a model that would have resulted from retraining without the ith individual's data (Izzo et al., 2021; Guo et al., 2020; Ginart et al., 2019). These works generally need to add noise in the approximation procedure in order to ensure this constraint is satisfied.

2. Iterative Approximate CV

In this section, we present our method for approximating leave-one-out CV in the setting of iterative optimization, where the algorithm may not be run to convergence. We find that even when $\widehat{\theta}$ can not be exactly computed, we can still approximate the leave-one-out iterates $\widehat{\theta}_{-i}^{(t)}$ accurately by leveraging the structure of the iterative update. In our theoretical results below, we will see that our estimates have guaranteed accuracy (under mild conditions) along all iterations $t \geq 1$ of the algorithm, i.e., even at iterations when the algorithm is far from convergence.

2.1. Framework

The iterative procedures that we will study in this paper—gradient descent, stochastic gradient descent, and proximal gradient descent—can all be expressed within the following general framework. Suppose our objective function can be written in the form

$$F(\mathcal{Z}; \theta) = g(\mathcal{Z}; \theta) + h(\theta)$$

where $g(\mathcal{Z}; \theta)$ is twice-differentiable in θ while $h(\theta)$ may be nondifferentiable. We consider solving the problem (1) iteratively as follows: for each step $t \geq 1$, we take a gradient step on g (possibly after subsampling the data points), and a proximal step on h. Specifically, the iterations are given by

$$\widehat{\theta}^{(t)} = \arg\min_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}$$
$$\text{where } \theta' = \widehat{\theta}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}), \quad (5)$$

where $S_t \subseteq [n]$ is a subset of indices, $\mathcal{Z}_{S_t} := \{Z_i : i \in S_t\}$ is the corresponding subset of the data, and $\alpha_t > 0$ is the learning rate.

In the setting where the objective function is twice-differentiable, we can take $g(\mathcal{Z}; \theta) = F(\mathcal{Z}; \theta)$ and $h \equiv 0$.

For example, in the regularized ERM setting, if the loss function ℓ and the penalty function $\pi(\theta)$ are twice-differentiable, we can simply take $g(\mathcal{Z}; \theta) = \sum_{i=1}^n \ell(Z_i; \theta) + \lambda\pi(\theta)$. Then (5) yields gradient descent if we choose the full dataset $S_t = [n]$ at each iteration,

$$\text{GD}: \quad \widehat{\theta}^{(t)} = \widehat{\theta}^{(t-1)} - \alpha_t \nabla_\theta F(\mathcal{Z}; \widehat{\theta}^{(t-1)}), \quad (6)$$

or stochastic gradient descent if at each iteration we sample a random batch $S_t \subseteq [n]$,

$$\text{SGD}: \quad \widehat{\theta}^{(t)} = \widehat{\theta}^{(t-1)} - \alpha_t \nabla_\theta F(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}). \quad (7)$$

In other settings, we may have a nondifferentiable term $h(\theta)$, which has an inexpensive proximal map (i.e., the solution to (5) can be computed efficiently for each iteration). For instance, in regularized ERM with a nonsmooth penalty function π such as the ℓ_1 norm, we might take $g(\mathcal{Z}; \theta) = \sum_{i=1}^n \ell(Z_i; \theta)$ and $h(\theta) = \lambda\pi(\theta)$. Then the general iterative scheme (5), run again with the full dataset $S_t = [n]$ at each iteration, reduces to proximal gradient descent,

$$\text{ProxGD}: \widehat{\theta}^{(t)} = \arg\min_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}$$
$$\text{where } \theta' = \widehat{\theta}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}; \widehat{\theta}^{(t-1)}). \quad (8)$$

2.2. Estimation Procedure: General Case

We will now provide an algorithm for Iterative Approximate CV (IACV) for general iterative procedures of the form (5). After deriving the general formulation, we will then show how it specializes to each of the three settings, GD, SGD, and ProxGD.

The targets $\widehat{\theta}_{-i}^{(t)}$ for $i \in [n]$ are obtained by running the same iterative solver as in (5) except with i-th data point being left out: for each $t \geq 1$,

$$\widehat{\theta}_{-i}^{(t)} = \arg\min_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}$$
$$\text{where } \theta' = \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}_{-i}^{(t-1)}), \quad (9)$$

In many settings, running the iterative algorithm for each $i \in [n]$ is computationally too expensive, since at each iteration t, it requires computing gradients for g at n different parameter vectors $\{\widehat{\theta}_{-i}^{(t-1)}\}_{i=1}^n$. Thus, we aim to find computationally inexpensive surrogates for these gradients.

We now define our procedure, which produces approximations $\widetilde{\theta}_{-i}^{(t)} \approx \widehat{\theta}_{-i}^{(t)}$, at each iteration $t \geq 1$ and for each $i \in [n]$. At iteration t, if we are not running the exact leave-one-out procedure, then computing θ' in (9) above involves two unknowns: the previous iterate, $\widehat{\theta}_{-i}^{(t-1)}$, which we simply approximate with our previous iteration's estimate $\widetilde{\theta}_{-i}^{(t-1)}$,

and its gradient, $\nabla_\theta g(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}_{-i}^{(t-1)})$, which we now address. If $\theta \mapsto g(\cdot; \theta)$ is twice-differentiable, a natural idea is to approximate $\nabla_\theta g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}_{-i}^{(t-1)})$ by its Taylor expansion at $\widehat{\theta}^{(t-1)}$. The reason we choose $\widehat{\theta}^{(t-1)}$ as the base point is that it can be shared across the n problems of approximating $\nabla_\theta g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}_{-i}^{(t-1)})$ for each $i \in [n]$. We thus take the approximation

$$\nabla_\theta g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}_{-i}^{(t-1)}) \approx \nabla_\theta g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)})$$
$$+ \nabla_\theta^2 g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)})[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}], \quad (10)$$

and then again plug in $\widetilde{\theta}_{-i}^{(t-1)}$ for the unknown $\widehat{\theta}_{-i}^{(t-1)}$ in the last term.

Thus, for each $i \in [n]$ each $t \geq 1$, the IACV approximation is given by

$$\text{IACV:} \quad \widetilde{\theta}_{-i}^{(t)} = \arg\min_\theta \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}$$
$$\text{where } \theta' = \widetilde{\theta}_{-i}^{(t-1)} - \alpha_t \big(\nabla_\theta g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)})$$
$$+ \nabla_\theta^2 g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]\big). \quad (11)$$

2.3. Estimation Procedure for GD, SGD, and ProxGD

Next, we give the details for implementing this general procedure in the specific settings of GD, SGD, and ProxGD.

Gradient descent (GD)   For GD, we have $h(\theta) \equiv 0$ and $S_t \equiv [n]$, so the steps of IACV reduce to

$$\widetilde{\theta}_{-i}^{(t)} = \widetilde{\theta}_{-i}^{(t-1)} - \alpha_t \big(\nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})$$
$$+ \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]\big). \quad (12)$$

Stochastic gradient descent (SGD)   For SGD, we again have $h(\theta) \equiv 0$, but the sets S_t consist of small batches of data. The steps of IACV reduce to

$$\widetilde{\theta}_{-i}^{(t)} = \widetilde{\theta}_{-i}^{(t-1)} - \alpha_t \big(\nabla_\theta F(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)})$$
$$+ \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]\big). \quad (13)$$

Proximal gradient descent (ProxGD)   For ProxGD, we again take $S_t \equiv [n]$, but the function $h(\theta)$ is nontrivial (e.g., a nonsmooth regularizer). The steps of IACV reduce to

$$\widetilde{\theta}_{-i}^{(t)} = \arg\min_\theta \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}$$
$$\text{where } \theta' = \widetilde{\theta}_{-i}^{(t-1)} - \alpha_t \big(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})$$
$$+ \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]\big). \quad (14)$$

	Computational Cost	
	IACV	Exact LOO CV
GD	$n(A_p + B_p) + np^2$	$n^2 A_p + np$
SGD	$K(A_p + B_p) + np^2$	$nK A_p + np$
ProxGD	$n(A_p + B_p + D_p) + np^2$	$n^2 A_p + n D_p + np$

Table 1. The order of per-iteration computation complexity of iterative approximate CV (IACV) as compared to exact leave-one-out (LOO) CV. See Section 3 for details.

3. Computation and Memory Complexity

In this section, we compare the computation and memory complexity of the proposed leave-one-out CV approximation method (11) with the exact leave-one-out CV (9). Note that running IACV (11) assumes that we have access to the iterates $\widehat{\theta}^{(t)}$, obtained by running the iterative procedure (5) on the full dataset. In addition, at each time t, we also need to calculate $\nabla_\theta g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)})$ and $\nabla_\theta^2 g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)})$ in order to compute the estimator. In other words, at each time t we need to compute the gradient and Hessian of g for datasets $\mathcal{Z}_{S_t \setminus i}$ that *vary* with i but at a parameter vector $\widehat{\theta}^{(t-1)}$ that is *constant* over i—in general, this can be done efficiently due to the separability of the data in the ERM problem. This is different from the computational cost of exact leave-one-out CV, which instead requires computing $\nabla_\theta g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}_{-i}^{(t-1)})$ for each i at each iteration t, i.e., the gradient of g needs to be computed at n *different* parameter vectors $\{\widehat{\theta}_{-i}^{(t-1)}\}_{i=1}^n$.

For simplicity, let us consider the case where $g(\mathcal{Z}; \theta) = \sum_{i=1}^n \ell(Z_i; \theta)$. We denote the computational cost of one call to $\nabla_\theta \ell(Z_i, \cdot)$ as A_p, and one call to $\nabla_\theta^2 \ell(Z_i, \cdot)$ as B_p, which depend on the dimension p of the parameter θ. We also assume that the cost of computing $\nabla_\theta g(\mathcal{Z}; \cdot) = \sum_{i=1}^n \nabla_\theta \ell(Z_i; \cdot)$ is equal to $n A_p$, and similarly $n B_p$ for $\nabla_\theta^2 g(\mathcal{Z}; \cdot)$; this assumption is mild since, in most typical settings, the gradient or Hessian of the loss can only be computed via evaluation on each data point one-by-one. Finally, let D_p denotes the computational cost of one call to the proximal operator $\theta' \mapsto \arg\min_\theta \{\frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta)\}$.

In Table 1, we provide the per-iteration computation complexity of computing $\{\widetilde{\theta}_{-i}^{(t)}\}_{i=1}^n$ for a single iteration t in our proposed method IACV, as compared to computing $\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^n$ in the exact leave-one-out CV, for the three iterative algorithms GD, SGD (with batch size K), and ProxGD. Details for these calculations are given in Appendix B.

For example, for GD, we can see that IACV is more efficient as long as $B_p + p^2 \ll n A_p$. Since typically, A_p and B_p are of order p and p^2, respectively, we can see that condition $B_p + p^2 \ll n A_p$ is typically satisfied when $n \gg p$ (which is exactly the regime considered in the literature for analyzing the existing CV approximation methods (Beirami et al.,

2017; Wilson et al., 2020)), and thus IACV is computationally more efficient in this regime.

One limitation of IACV is an increased memory cost—the proposed method needs to store the data, gradient, and Hessian, which costs $O(np + p^2)$ space, while the exact leave-one-out CV costs $O(np)$ as it only needs to store the data and gradient. In the $n \gg p$ regime, however, the memory cost of the two methods is on the same order.

4. Accuracy Guarantees for IACV

In this section, we provide theoretical guarantees for IACV in the GD and SGD settings. (Guarantees for ProxGD are given in Appendix A.) For both GD and SGD, we will consider the objective function $F(\mathcal{Z}; \theta) = \sum_{i=1}^{n} \ell(Z_i; \theta) + \lambda \pi(\theta)$, where π is a twice-differentiable regularizer, as before.

We will study two notions of accuracy: the *approximation error*,

$$\text{Err}_{\text{approx}}^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2, \quad (15)$$

measuring our accuracy in approximating the exact leave-one-out estimators $\widehat{\theta}_{-i}^{(t)}$, and the *CV error*,

$$\text{Err}_{\text{CV}}^{(t)} = |\text{CV}(\{\widetilde{\theta}_{-i}^{(t)}\}_{i=1}^{n}) - \text{CV}(\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^{n})|, \quad (16)$$

measuring our accuracy in estimating the leave-one-out CV loss for $\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^{n}$. (This latter notion of error is the quantity that was plotted in Figure 1.)

4.1. Guarantees for GD

We first state our assumptions for the GD setting.

Assumption 4.1. For all $t \geq 1$, $\min_{i \in [n]} \lambda_{\min}(\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})) \geq n\lambda_0$ and $\max_{i \in [n]} \lambda_{\max}(\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})) \leq n\lambda_1$ for some positive constants $\lambda_0, \lambda_1 > 0$.

This first assumption says the Hessian of F is well-conditioned along all iterates. Note that in the literature for analyzing the NS (3) or IJ (4) estimators, the Hessian is often assumed to be well-conditioned at $\widehat{\theta}$, the minimizer of (1) (Beirami et al., 2017; Wilson et al., 2020; Giordano et al., 2019), which is sufficient since $\widetilde{\theta}_{-i}^{\text{NS}}$ and $\widetilde{\theta}_{-i}^{\text{IJ}}$ in (3) and (4) are based on $\widehat{\theta}$ only. Here we require a stronger assumption because our estimator depends on $\widehat{\theta}^{(t)}$ for all t.

The second assumption is about the gradient.

Assumption 4.2. For all $t \geq 1$, $i \in [n]$, we have $\|\nabla_\theta \ell(Z_i; \widehat{\theta}^{(t-1)})\|_2 \leq \eta_i$ for some $\eta_i > 0$.

The final assumption requires the Hessian to be Lipschitz.

Assumption 4.3. For all $i \in [n]$, $\nabla_\theta^2 F(\mathcal{Z}_{-i}; \cdot)$ is $n\gamma$-Lipschitz, i.e., for any $\theta_1, \theta_2 \in \mathbb{R}^p$,

$$\|\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta_1) - \nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta_1)\| \leq n\gamma\|\theta_1 - \theta_2\|_2.$$

We are now ready to state our first main result about IACV.

Theorem 4.4 (Approximation Error of IACV for GD). *Suppose $\widehat{\theta}^{(0)} = \widehat{\theta}_{-i}^{(0)} = \widetilde{\theta}_{-i}^{(0)}$ for all $i \in [n]$, $\alpha_t \leq 1/(n\lambda_1)$ for $t \geq 1$, and Assumptions 4.1–4.3 are satisfied with $\gamma < 2\lambda_0$ and $n \geq \frac{4\|\eta\|_\infty}{2\lambda_0 - \gamma}$. Then for all $t \geq 1$ and $i \in [n]$, we have $\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \leq \frac{2\eta_i}{(2\lambda_0 - \gamma)n}$ and $\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \leq \frac{4\gamma\eta_i^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2}$.*

In Theorem 4.4, we give upper bounds that are independent of t, for clarity of the result; tighter, iteration-dependent upper bounds are discussed in Appendix D. As an immediate consequence of the last bound, the approximation error (15) is bounded as

$$\text{Err}_{\text{approx}}^{(t)} \leq \frac{4\gamma\|\eta\|_\infty^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2} \text{ for all } t \geq 1.$$

In Theorem 4.4, we see that the error bound for $\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ is of a smaller order than the one for $\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$. This is exactly what we need—the goal of leave-one-out CV is to determine how the leave-one-out estimators $\widehat{\theta}_{-i}^{(t)}$ behave differently from the estimator $\widehat{\theta}^{(t)}$ computed on the full dataset, and thus the approximations $\widetilde{\theta}_{-i}^{(t)}$ are useful only if they can improve over the "baseline" accuracy of $\widehat{\theta}^{(t)}$ itself.

In addition, it has been shown in Beirami et al. (2017); Wilson et al. (2020) that the NS (3) and IJ (4) estimators achieve an approximation error of order $1/n^2$, which is the same as our result above. However, NS and IJ only achieve this error when run with the exact value of $\widehat{\theta}$ (as we have seen in Figure 1, this assumption is crucial for empirical accuracy), while IACV achieves the $O(1/n^2)$ approximation error bound along all iterations of the algorithm rather than only at convergence.

Based on the approximation error bounds established in Theorem 4.4, we are also able to provide guarantees for IACV's ability to approximate the leave-one-out CV loss $\text{CV}(\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^{n})$, with error on the order of $1/n^2$. We will need one additional assumption on the gradient of the loss, which can be viewed as a stronger version of Assumption 4.2:

Assumption 4.5. Suppose for any $i \in [n]$ and all $t \geq 1$, there exists $\eta_i' > 0$ such that $\sup_{a \in [0,1]} \|\nabla \ell(Z_i; a\widetilde{\theta}_{-i}^{(t)} + (1-a)\widehat{\theta}_{-i}^{(t)})\|_2 \leq \eta_i'$.

For example, this assumption will be satisfied if $\ell(Z_i; \cdot)$ is η_i'-Lipschitz for each i.

Theorem 4.6 (CV Error of IACV for GD). *If the assumptions in Theorem 4.4 are satisfied, and Assumption 4.5 holds, then for all $t \geq 1$, $\text{Err}_{\text{CV}}^{(t)} \leq \frac{1}{n} \sum_{i=1}^{n} \frac{4\gamma \eta_i' \eta_i^2}{\lambda_0 (2\lambda_0 - \gamma)^2 n^2}$.*

Theorem 4.6 suggests IACV might be useful when we perform early stopping based on CV loss as it has guaranteed accurate estimates of $\text{CV}(\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^{n})$ along the entire trajectory, including times t that are well before convergence.

When are these assumptions satisfied? Next, we provide an example of a setting where Assumptions 4.1–4.5 are likely satisfied. Suppose the data is generated from the generalized linear model with a canonic link function (McCullagh & Nelder, 1989). Specifically, suppose Y_i has density $\exp(\eta_i Y_i - \phi(\eta_i)) h(Y_i)$, where $\phi : \mathbb{R} \mapsto \mathbb{R}$ is a given function and η_i is linked with X_i via $\eta_i = X_i^{\top} \theta^*$. Then given i.i.d. data $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1}^{n}$, the gradient, and the Hessian of the negative log-likelihood function with ridge regularizer are given as

$$\nabla_\theta F(\mathcal{Z}; \theta) = -\sum_{i=1}^{n} \left(Y_i - \phi'(\theta^{\top} X_i) \right) X_i + \lambda \theta,$$

$$\nabla_\theta^2 F(\mathcal{Z}; \theta) = \sum_{i=1}^{n} \phi''(\theta^{\top} X_i) X_i X_i^{\top} + \lambda \mathbf{I}_p.$$

Since $\mathbb{E}(Y_i) = \phi'(\theta^{*\top} X_i)$ and $\text{Var}(Y_i) = \phi''(\theta^{*\top} X_i) > 0$, we expect the first and second derivatives of the objective will be well-controlled when θ is close to θ^*, the X_is are bounded, and $n \geq p$. For many non-convex problems of interest, if we could have a warm-start initialization, then the problem at local typically satisfies certain restricted strong convexity and smoothness properties, and the Hessians and gradients along the iterates are often well-controlled as well (Loh & Wainwright, 2013; Chi et al., 2019).

4.2. Guarantees for SGD

In the SGD setting, we assume the batches $\{S_t\}_{t \geq 1}$ are drawn i.i.d. such that each data point is included in S_t with probability K/n. To establish the approximation guarantee for IACV in this setting, we need the following three assumptions, which are stochastic analogues of Assumptions 4.1–4.3.

Assumption 4.7. Suppose $F(\cdot; \theta)$ is twice differentiable in θ and there exists $\lambda_0, \lambda_1 > 0$ such that for all $t \geq 1$, $i \in [n]$, and $\alpha_t \leq \frac{1}{K\lambda_1}$, we have

$$\mathbb{E}(\|[\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})][\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}]\|_2)$$
$$\leq (1 - \alpha_t K \lambda_0) \mathbb{E}\left[\|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2\right],$$

where the expectation is taken over randomly drawn batches $S_1, \ldots, S_t$.

For example, this assumption would be satisfied under the stronger assumption that for all $a, \theta \in \mathbb{R}^p$, all $i \in [n]$, and for a randomly drawn batch S,

$$\mathbb{E}_S(\|(\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{S \setminus \{i\}}; \theta)) v\|_2) \leq (1 - \alpha_t K \lambda_0) \|a\|_2.$$

Essentially, we can interpret $K\lambda_0$ and $K\lambda_1$ as bounding the smallest and largest eigenvalues of $\nabla_\theta^2 F(\mathcal{Z}_{S \setminus \{i\}}; \theta)$.

Our next assumptions bound the gradient and the Lipschitz constant of the Hessian.

Assumption 4.8. For all $t \geq 1$ and $i \in [n]$, we have $\mathbb{E}_{S_1, \ldots, S_{t-1}} \|\nabla_\theta \ell(Z_i; \widehat{\theta}^{(t-1)})\|_2 \leq \eta_i$ for some $\eta_i > 0$.

Assumption 4.9. There exists $\gamma > 0$ such that for any $i \in [n]$ and $t \geq 1$,

$$\mathbb{E} \sup_{\theta'} \left\| (\nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})) \right.$$
$$\left. [\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \right\|_2$$
$$\leq \gamma K \mathbb{E} \|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2^2,$$

where the supremum is taken over $\theta' = a\widehat{\theta}_{-i}^{(t-1)} + (1 - a)\widehat{\theta}^{(t-1)}$ for any $a \in [0, 1]$.

For example, this assumption would be satisfied if given any $v, \theta_1, \theta_2 \in \mathbb{R}^p$,

$$\mathbb{E}_S \|(\nabla_\theta^2 F(\mathcal{Z}_{S \setminus \{i\}}; \theta_1) - \nabla_\theta^2 F(\mathcal{Z}_{S \setminus \{i\}}; \theta_2) v\|_2$$
$$\leq \gamma K \|v\|_2 \|\theta_1 - \theta_2\|_2.$$

In addition, we need an extra assumption as follows.

Assumption 4.10. There exists $\beta \geq 1$ such that for all $t \geq 1$, we have $\mathbb{E}_{S_1, \ldots, S_{t-1}} \|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2^2 \leq \beta \frac{n}{K} \left(\mathbb{E}_{S_1, \ldots, S_{t-1}} \|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2 \right)^2$.

Assumption 4.10 can be viewed as a reverse Jensen's inequality, but there is an inflation factor $\beta \frac{n}{K}$ on the right-hand side. The order $\frac{n}{K}$ of this inflation factor is expected. A simple way to reveal this is to examine the inequality when $t = 2$. If $\widehat{\theta}_{-i}^{(0)} = \widehat{\theta}^{(0)}$, then $\widehat{\theta}_{-i}^{(1)} - \widehat{\theta}^{(1)} = 0$ with probability $(1 - \frac{K}{n})$ and $\widehat{\theta}_{-i}^{(1)} - \widehat{\theta}^{(1)} = \alpha_1 \nabla_\theta \ell(Z_i; \widehat{\theta}^{(0)})$ with probability $\frac{K}{n}$ (depending on whether data point i is excluded or included in the first batch S_1 at time $t = 1$). Thus, we have $\mathbb{E}_{S_1} \|\widehat{\theta}_{-i}^{(1)} - \widehat{\theta}^{(1)}\|_2^2 = \frac{n}{K} (\mathbb{E}_{S_1} \|\widehat{\theta}_{-i}^{(1)} - \widehat{\theta}^{(1)}\|)^2$.

Now we are ready to present a guarantee on the approximation error of IACV in the SGD setting.

Theorem 4.11 (Approximation Error of IACV for SGD). *Suppose $\widehat{\theta}^{(0)} = \widehat{\theta}_{-i}^{(0)} = \widetilde{\theta}_{-i}^{(0)}$ for all $i \in [n]$, $\alpha_t \leq 1/(K\lambda_0)$ for all $t \geq 1$ and Assumptions 4.7–4.10 are satisfied with $\gamma\beta < 2\lambda_0$ and $K \geq \frac{4\|\eta\|_\infty}{2\lambda_0 - \gamma\beta}$. Then for all $t \geq 1$, $i \in [n]$, we*

have $\mathbb{E}\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \leq \frac{2\eta_i}{(2\lambda_0 - \gamma\beta)n}$ and $\mathbb{E}\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \leq \frac{4\gamma\beta\eta_i^2}{\lambda_0(2\lambda_0 - \gamma\beta)^2 nK}$, where the expectation is taken over the randomly drawn batches $S_1, \ldots, S_t$.

By Theorem 4.11, the expected approximation error in the SGD case is bounded as

$$\mathbb{E}\left(\mathrm{Err}_{\mathrm{approx}}^{(t)}\right) \leq \frac{4\gamma\beta\|\eta\|_\infty^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2} \text{ for all } t \geq 1.$$

Interestingly, we find that the "baseline" approximation error $\mathbb{E}\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ does not depend on K and is at the same order $\frac{1}{n}$ as the one in Theorem 4.4. In contrast, the IACV error $\mathbb{E}\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ scales at the order of $\frac{1}{nK}$, as compared to order $\frac{1}{n^2}$ in the previous setting. This factor comes from the inflation factor $\frac{n}{K}$ in Assumption 4.10; as discussed immediately below this assumption, the factor $\frac{n}{K}$ cannot be removed if we expect the assumption to hold in practice for all $t \geq 1$, but a remaining open question is whether the factor $\frac{n}{K}$ can be removed if we only require it to hold for all $t \geq T_0$ for some large T_0. If this is the case, then (for sufficiently large t) the IACV error in Theorem 4.11 will scale as $\frac{1}{n^2}$ rather than $\frac{1}{nK}$.

Finally, we provide a CV error guarantee under the SGD setting, with one additional assumption:

Assumption 4.12. Suppose for any $i \in [n]$ and all $t \geq 1$, there exists $\eta_i' > 0$ such that

$$\mathbb{E}\left\{ \sup_{a \in [0,1]} |\nabla_\theta \ell(Z_i; a\widetilde{\theta}_{-i}^{(t)} + (1-a)\widehat{\theta}_{-i}^{(t)})^\top (\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)})| \right\}$$
$$\leq \eta_i' \mathbb{E}\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2.$$

For example, this assumption will be satisfied if $\ell(Z_i, \cdot)$ is η_i'-Lipschitz for each i.

Theorem 4.13 (CV Error of IACV for SGD). *If the assumptions in Theorem 4.11 and Assumption 4.12 are satisfied, then for all $t \geq 1$:*

$$\mathbb{E}\left(\mathrm{Err}_{\mathrm{CV}}^{(t)}\right) \leq \frac{1}{n}\sum_{i=1}^n \frac{4\gamma\beta\eta_i'\eta_i^2}{\lambda_0(2\lambda_0 - \gamma\beta)^2 nK},$$

where the expectation is taken over the randomly drawn batches $S_1, \ldots, S_t$.

As before, this result is analogous to Theorem 4.6 for the GD setting, but with rate $1/nK$ in place of $1/n^2$.

5. Limiting Behavior of IACV

As we have seen in Section 4, IACV achieves accurate approximation along all iterations of the algorithm under some regularity conditions. Thus, in the setting when $\widehat{\theta}^{(t)}$ indeed converges to $\widehat{\theta}$, IACV also has the same approximation properties in the limit, and its approximation error bound is comparable to the guarantee of the one-step Newton (NS) estimator in (3) as we have mentioned in Section 4.1. In this section, we show this is not a coincidence—in fact, there is a close connection of the proposed estimator $\widetilde{\theta}_{-i}^{(t)}$ to the NS estimator when $t \to \infty$ and the algorithm converges to $\widehat{\theta}$.

Theorem 5.1. *Suppose $\widehat{\theta}^{(t)}$ converges to $\widehat{\theta}$, Assumption 4.1 is satisfied along all iterations, $\alpha_t = \alpha < 1/(n\lambda_1)$ for all $t \geq 1$, and $\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta)$, $\nabla_\theta F(\mathcal{Z}_{-i}; \theta)$ are continuous in θ for all $i \in [n]$. Then $\widetilde{\theta}_{-i}^{(t)}$ defined in (12) converges to $\widetilde{\theta}_{-i}^{\mathrm{NS}}$ in (3), as $t \to \infty$.*

Moreover, in Theorem A.3, we will show similar guarantees continue to hold in the ProxGD setting. (In the SGD setting, however, the techniques for proving Theorem 5.1 fail—we will give an intuition for why this is the case, in Appendix E.2.)

In view of the results in Theorem 5.1, we can regard IACV as an extension of $\widetilde{\theta}_{-i}^{\mathrm{NS}}$ in (3) to iterative solvers with provable per-iteration guarantees. This provides a safe and efficient way to approximate CV in practice where it is often agnostic whether the algorithm will converge to the solution of ERM or not.

6. Simulation Studies

In this section, we conduct numerical studies to investigate the empirical performance of the proposed IACV method and to verify our theoretical findings for GD, SGD, and ProxGD.[1] The data is generated from a logistic regression model, with $Z_i = (X_i, Y_i) \in \mathbb{R}^p \times \{0, 1\}$, with dimension $p = 20$ and with X_i drawn with i.i.d. $N(0, 1)$ entries while

$$Y_i \sim \mathrm{Bernoulli}(\exp(X_i^\top \theta^*)/(1 + \exp(X_i^\top \theta^*)))$$

for true parameter vector θ^* which has 5 randomly chosen nonzero entries drawn as $N(0, 1)$, and all other entries zero. Our objective function is given by regularized negative log-likelihood, $F(\mathcal{Z}; \theta) = \sum_{i=1}^n \ell(Z_i; \theta) + \lambda\pi(\theta)$, where

$$\ell(Z_i; \theta) = -Y_i \cdot X_i^\top \theta + \log(1 + \exp(X_i^\top \theta)).$$

For GD and SGD, we use ridge regularization, with $\pi(\theta) = \|\theta\|_2^2$ and penalty parameter $\lambda = 10^{-6} \cdot n$. For ProxGD we instead use the logistic Lasso, with $\pi(\theta) = \|\theta\|_1$ and $\lambda = 10^{-6} \cdot n$. We initialize the algorithm at the origin. Each simulation study is repeated for 100 independent trials.

Our proposed method is given by the IACV estimator $\widetilde{\theta}_{-i}^{(t)}$, defined in (12) for GD and in (13) for SGD. For comparison,

---

[1]Code to reproduce all experiments is available at https://github.com/yuetianluo/IACV.

we also implement the one-step Newton (NS) and infinitesimal jackknife (IJ) estimators along the optimization path, i.e., we use the tth iteration $\widehat{\theta}^{(t)}$ in place of the true minimizer $\widehat{\theta}$ in the definition of the NS (3) or IJ (4) estimators, leading to the approximate NS estimator

$$\widetilde{\theta}_{-i}^{\mathrm{NS}(t)} = \widehat{\theta}^{(t)} - \left(\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t)})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t)})$$

and similarly the approximate IJ estimator

$$\widetilde{\theta}_{-i}^{\mathrm{IJ}(t)} = \widehat{\theta}^{(t)} - \left(\nabla_\theta^2 F(\mathcal{Z}; \widehat{\theta}^{(t)})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t)}).$$

Finally, we also compare to a "baseline" estimator where we simply approximate the leave-one-out iterate $\widehat{\theta}_{-i}^{(t)}$ with the full-data iterate $\widehat{\theta}^{(t)}$: $\widehat{\theta}_{-i}^{\mathrm{baseline}(t)} = \widehat{\theta}^{(t)}$. Of course, this baseline is not useful in practice (since data point i has not actually been removed from the estimator), and thus any proposed method is only meaningful if it can perform substantially better than this baseline. We measure the accuracy of the four methods (baseline, NS, IJ, and IACV) in terms of the averaged approximation error $\mathrm{Err}_{\mathrm{approx}}^{(t)}$ defined in (15), and the relative CV error, $\mathrm{RelErr}_{\mathrm{CV}}^{(t)} = \mathrm{Err}_{\mathrm{CV}}^{(t)}/\mathrm{CV}(\{\widehat{\theta}_{-i}^{(t)}\}_{i=1}^n)$, where $\mathrm{Err}_{\mathrm{CV}}^{(t)}$ is defined in (16).

6.1. Gradient Descent (GD)

For the gradient descent simulation, we consider sample sizes $n = 250$ and $n = 1000$, and $\alpha_t = 0.5/n$. In the top panels of Figure 2, we can see that along the iterations of the algorithm, the approximation error of IACV is always better than NS and IJ before the convergence of the algorithm. As a result of that, IACV also achieves better CV error as we illustrate in the middle panels of Figure 2. In fact, during early iterations, the error of both NS and IJ is higher than the (noninformative) baseline estimator $\widehat{\theta}_{-i}^{\mathrm{baseline}(t)}$, while IACV's error is substantially lower, and also shows a smaller variance in estimation as it has narrower shaded areas. When the algorithm converges, we find that the performance of IACV and the NS estimator are almost the same, which matches the theoretical prediction in Theorem 5.1 (where we see that, under mild assumptions, IACV will converge to the NS estimator, i.e., the black line and the red line will meet in the limit), while the IJ method shows higher error even at convergence.

In addition, we observe that as sample size n increases from 250 to 1000, the limiting approximation error of IACV decreases from 1.5×10^{-3} to 6.8×10^{-5}. This roughly matches what we have shown in Theorem 4.4 that the approximation error decreases quadratically with respect to the sample size.

Finally, in the bottom panels of Figure 2, we report the runtime of IACV as compared to the exact leave-one-out
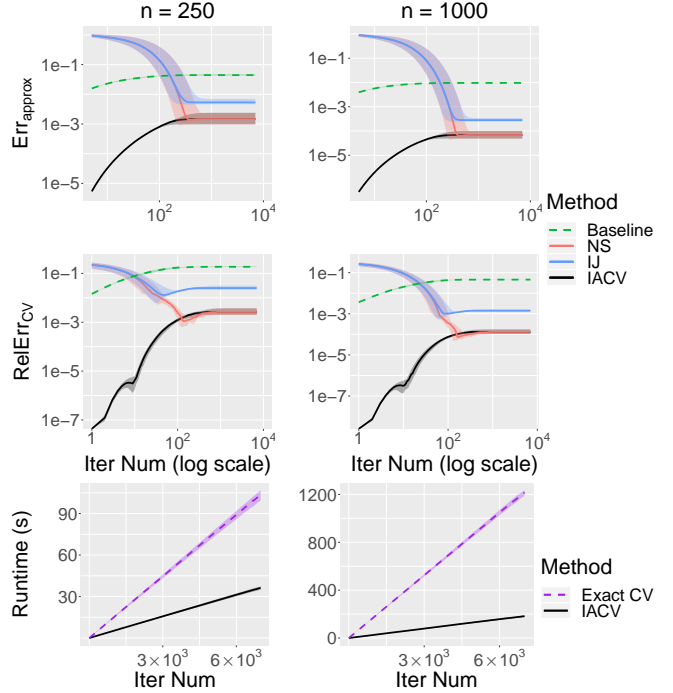

Figure 2. Comparison of the baseline, NS, IJ, and IACV methods in gradient descent for logistic regression with ridge regularizer. Top: approximation error comparison; middle: relative CV error comparison; bottom: runtime comparison of exact leave-one-out CV and IACV. Solid lines represent the median value over 100 experiments and shaded areas denote the region between lower and upper quartiles.

CV. We can see that IACV is much faster than the exact leave-one-out CV method; in particular when $n = 1000$, IACV shows approximately 6-7 times speed-up. A larger scale simulation for GD is provided in Appendix C.

6.2. Stochastic Gradient Descent (SGD)

For the stochastic gradient descent simulation, we take sample size $n = 1000$, and test batch size $K = 100$ and $K = 400$. We choose α_t based on a common strategy called "epoch doubling" in the literature, where we run $T_0 = 1000$ steps with step size $\alpha = 0.5/K$, then run $2T_0$ steps with step size $\alpha/2$, and so on.

We plot the accuracy of the different methods in Figure 3, in the top (approximation error) and middle (CV error) panels. We can see that IACV has a clear advantage over the other methods. In contrast to the simulation results in Figure 2 for GD, here the red line (for NS) does not reach the black line (for IACV) even after $T = 10^5$ iterations, in terms of approximation error. This may be due to the slow convergence of SGD; nonetheless, it is still unclear whether the NS estimator $\widetilde{\theta}_{-i}^{\mathrm{NS}(t)}$ and our estimator $\widetilde{\theta}_{-i}^{(t)}$ will converge to the same limit or not, since we do not know whether a
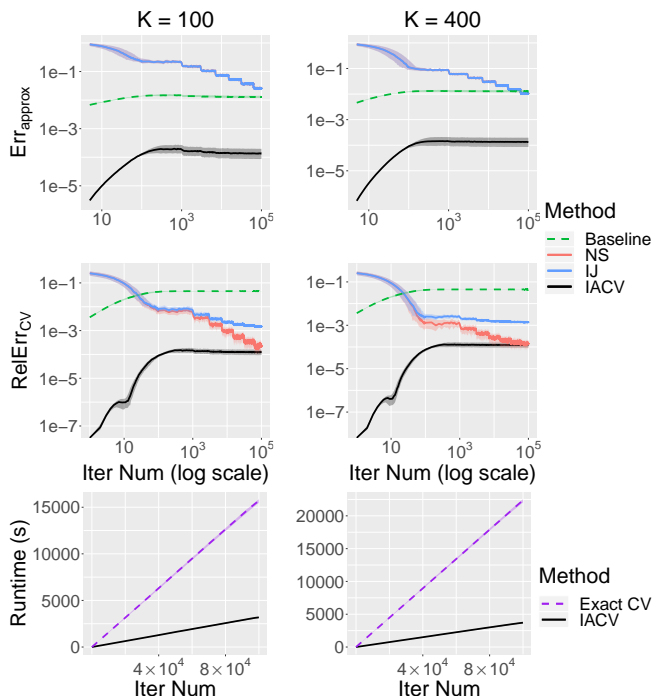
Figure 3. Comparison of the baseline, NS, IJ, and IACV methods in stochastic gradient descent for logistic regression with ridge regularizer; details of the panels are the same as in Figure 2. Note that in the top panels, the results for NS and IJ (red and blue) are nearly perfectly overlapping and thus difficult to distinguish.

Figure 4. Comparison of the baseline, NS, IJ, and IACV methods in proximal gradient descent for logistic regression with ℓ_1 regularizer (the logistic Lasso); details of the panels are the same as in Figure 2.

result analogous to Theorem 5.1 holds in the setting of SGD. As for GD, we see that IACV offers error far lower than both NS and IJ during early iterations, when NS and IJ show error higher even than the baseline estimator.

Finally, we show the runtime of IACV with exact leave-one-out CV in the bottom panel in Figure 3. We can see IACV still has clear computational advantages in this setting.

6.3. Proximal Gradient Descent (ProxGD)

Finally, for the proximal gradient descent simulation, we take sample sizes $n = 250$ and $n = 1000$, and $\alpha_t = 0.5/n$, as for the GD simulation. The regularizer is now $\pi(\theta) = \|\theta\|_1$, a nonsmooth function (see Appendix A for the definition of the NS and IJ methods in this nonsmooth setting).

The results for this setting are qualitatively very similar to the GD setting; again, we see that IACV shows good accuracy in terms of both approximation error and CV error even at early iterations, while NS and IJ show error higher than the baseline during early iterations. Finally, when the algorithm converges, the performance of our estimator is similar to the NS estimator as predicted in Theorem A.3 (which is the analogue of Theorem 5.1, for the ProxGD setting).
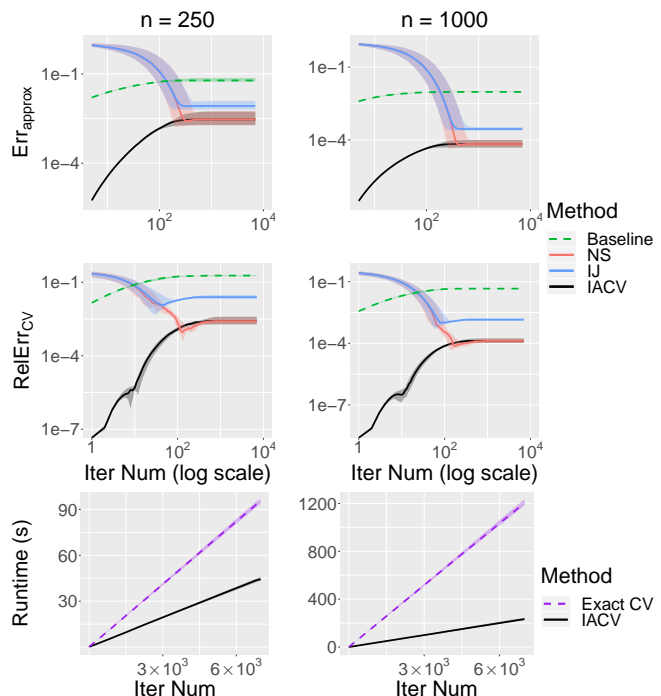
7. Conclusion and Discussions

In this paper, we provide a new method, iterative approximate cross-validation (IACV), to efficiently approximate the computationally expensive leave-one-out CV under the ERM framework when the problem is solved by common iterative algorithms. This work suggests several interesting directions for further exploration. For example, Theorem 5.1 establishes that, for gradient descent, the IACV and NS estimators coincide at convergence; we do not yet know whether an analogous result holds for SGD. We can also consider the performance of IACV in a high-dimensional setting. For instance, in regression with $p > n$, the NS estimator defined in (3) fails due to high dimensionality, but Stephenson & Broderick (2020) propose running ℓ_1-regularized regression and then running the NS estimator on the selected active set of $\hat{\theta}$. It would be interesting to see whether we can adapt the IACV method to that setting and provide theoretical guarantees for the sparse high-dimensional regime.

Acknowledgements

References

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

Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

Beirami, A., Razaviyayn, M., Shahrampour, S., and Tarokh, V. On optimal generalizability in parametric learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015.

Chi, Y., Lu, Y. M., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Efron, B. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

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

Ghosh, S., Stephenson, W., Nguyen, T. D., Deshpande, S., and Broderick, T. Approximate cross-validation for structured models. *Advances in Neural Information Processing Systems*, 33:8741–8752, 2020.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1139–1147. PMLR, 2019.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3832–3842, 2020.

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Jabbar, H. and Khan, R. Z. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, 70:163–172, 2015.

Jaeckel, L. The infinitesimal jackknife. memorandum. Technical report, MM 72-1215-11, Bell Lab. Murray Hill, NJ, 1972.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Loh, P.-L. and Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Advances in Neural Information Processing Systems*, 26, 2013.

McCullagh, P. and Nelder, J. Generalized linear models, 1989.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

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

Rad, K. R. and Maleki, A. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996, 2020.

Stephenson, W. and Broderick, T. Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2424–2434. PMLR, 2020.

Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

Tsai, C.-H., Lin, C.-Y., and Lin, C.-J. Incremental and decremental training for linear classification. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 343–352, 2014.

Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. Approximate leave-one-out for fast parameter tuning in high dimensions. In *International Conference on Machine Learning*, pp. 5228–5237. PMLR, 2018.

Wilson, A., Kasy, M., and Mackey, L. Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pp. 4530–4540. PMLR, 2020.

Wright, S. J. and Recht, B. *Optimization for data analysis*. Cambridge University Press, 2022.

A. Guarantees for IACV in the Proximal GD Setting

We now provide the analogues of Theorems 4.4 and 4.6 in the ProxGD setting, working with the objective function $F(\mathcal{Z}; \theta) = g(\mathcal{Z}; \theta) + h(\theta)$, where as before, $g(\mathcal{Z}; \theta) = \sum_{i=1}^{n} \ell(Z_i; \theta)$ is the empirical risk and $h(\theta) = \lambda\pi(\theta)$ is the nonsmooth regularization term.

Theorem A.1. *Suppose $\widehat{\theta}^{(0)} = \widehat{\theta}_{-i}^{(0)} = \widetilde{\theta}_{-i}^{(0)}$ for all $i \in [n]$, $\alpha_t \leq 1/(n\lambda_1)$ for $t \geq 1$, and Assumptions 4.1–4.3 are satisfied with $F(\cdot; \theta)$ replaced by $g(\cdot; \theta)$ and with $\gamma < 2\lambda_0$ and $n \geq \frac{4\|\eta\|_\infty}{2\lambda_0 - \gamma}$. Assume also that h is convex. Then for all $t \geq 1$ and $i \in [n]$, we have*

$$\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \leq \frac{2\eta_i}{(2\lambda_0 - \gamma)n}$$

and

$$\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \leq \frac{4\gamma\eta_i^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2}.$$

Note that these bounds are identical to the results obtained in Theorem 4.4, and thus as before, we obtain the approximation error bound

$$\mathrm{Err}_{\mathrm{approx}}^{(t)} \leq \frac{4\gamma\|\eta\|_\infty^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2} \text{ for all } t \geq 1.$$

The proof of Theorem A.1 is provided in Appendix E.1.

Next we bound the CV error for IACV in the ProxGD setting.

Theorem A.2. *Suppose the assumptions in Theorem A.1 and Assumption 4.5 are satisfied. Then for all $t \geq 1$,*

$$\mathrm{Err}_{\mathrm{CV}}^{(t)} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{4\gamma\eta_i'\eta_i^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2}.$$

Again, this bound is the same as the one obtained in Theorem 4.6 for GD. The proof of Theorem A.2 is essentially the same as the proof of Theorem 4.6 and for simplicity, we omit it here.

Next, we provide convergence theory for IACV in the ProxGD setting, to obtain a result analogous to Theorem 5.1 for the GD case comparing the IACV and NS estimators.

First, we need to define the NS estimator in this setting where $F(\mathcal{Z}; \theta)$ is nonsmooth. The main idea is to replace the one Newton step in (3) with one proximal Newton step (Wilson et al., 2020):

$$\widehat{\theta}_{-i}^{\mathrm{NS}} = \mathrm{prox}_h^{\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})}\left(\widehat{\theta} - \left(\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})\right)^{-1}\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta})\right), \tag{17}$$

where given any positive definite matrix $H \in \mathbb{R}^{p\times p}$ and convex function $h : \mathbb{R}^p \to \mathbb{R}$, $\mathrm{prox}_h^H(x)$ is defined as follows:

$$\mathrm{prox}_h^H(x) = \arg\min_{z\in\mathbb{R}^p} \frac{1}{2}(x - z)^\top H(x - z) + h(z). \tag{18}$$

(We can similarly define the IJ estimator in this setting as

$$\widetilde{\theta}_{-i}^{\mathrm{IJ}} = \mathrm{prox}_h^{\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})}\left(\widehat{\theta} - \left(\nabla_\theta^2 g(\mathcal{Z}; \widehat{\theta})\right)^{-1}\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta})\right). \tag{19}$$

In our simulations in Section 6.3, we simply replace $\widehat{\theta}$ with $\widehat{\theta}^{(t)}$ in equations (17) and (19) to obtain our approximate NS and IJ iterations.)

Next, we show $\widetilde{\theta}_{-i}^{(t)}$ in (14) will converge to $\widetilde{\theta}_{-i}^{\mathrm{NS}}$ in (17), under proper assumptions.

Theorem A.3. *Suppose $\widehat{\theta}^{(t)}$ converges to $\widehat{\theta}$, Assumption 4.1 is satisfied along all iterations with $g(\cdot; \theta)$ in place of $F(\cdot; \theta)$, $\alpha_t = \alpha < 1/(n\lambda_1)$ for all $t \geq 1$, and $\nabla_\theta^2 g(\mathcal{Z}_{-i}; \theta)$, $\nabla_\theta g(\mathcal{Z}_{-i}; \theta)$ are continuous in θ for all $i \in [n]$. In addition, we assume $h(\theta)$ is convex. Then it holds that $\widetilde{\theta}_{-i}^{(t)}$ in (14) converges to $\widetilde{\theta}_{-i}^{\mathrm{NS}}$ in (17), as $t \to \infty$.*

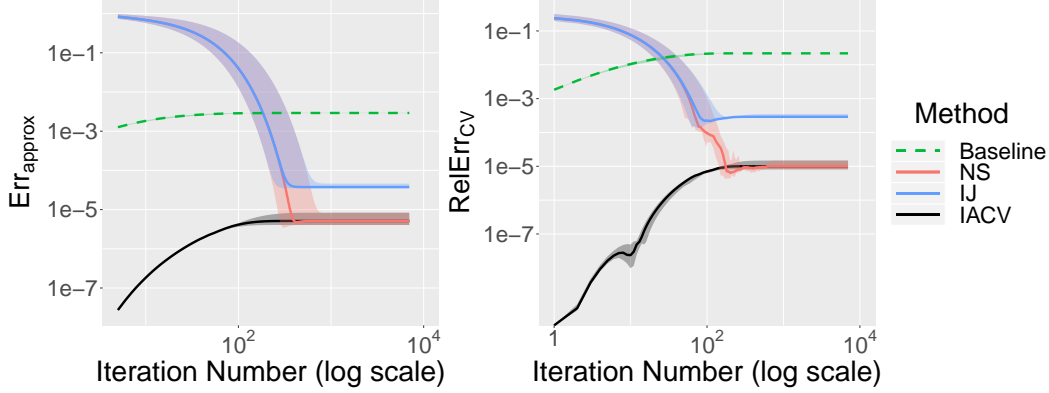The proof of this theorem is provided in Appendix E.2.

Figure 5. Approximation error and relative CV error comparison of the baseline, NS, IJ, and IACV methods in gradient descent for logistic regression with ridge regularizer. Solid lines represent the median value over 100 experiments and shaded areas denote the region between lower and upper quartiles.

B. Details for deriving computational complexity

To derive the calculations in Table 1, for example, for GD, the main per-iteration cost of our method comes from computing $\{\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\}_{i=1}^n$ and $\{\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\}_{i=1}^n$, and performing the Hessian-gradient product. As we assume above, due to the relation $\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) = \sum_{j \in [n], j \neq i} \nabla_\theta \ell(Z_j; \widehat{\theta}^{(t-1)})$, the cost of computing $\{\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\}_{i=1}^n$ and $\{\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\}_{i=1}^n$ is of order $n(A_p + B_p)$. Moreover, the cost of the Hessian-gradient product is of order np^2. On the other hand, the main per-iteration computational cost of the exact leave-one-out CV (i.e., running an iteration of gradient descent on the dataset $\mathcal{Z}_{-i}$, for each i), comes from evaluating $\nabla_\theta g(\mathcal{Z}_{-i}, \widehat{\theta}_{-i}^{(t-1)})$ and then subtracting this gradient from the current estimate, for each i. These two steps have cost of order nA_p and p, respectively, for each i, and therefore the total cost of one iteration has order $n^2 A_p + np$. Similar calculations can be performed to compute the order of computational complexity for SGD and ProxGD as well.

Note here that we do not intend to compare the runtime of NS and IJ with IACV and exact CV because these methods are not comparable in terms of their target problem. NS and IJ are one-step methods, where given a single solution (i.e., a single $\widehat{\theta}^{(T)}$ approximating $\widehat{\theta}$), we run the method once to approximate the leave-one-out models. On the other hand, exact iterative CV as well as our IACV methods are both performed in an online fashion, with steps carried out for each $t = 1, \ldots, T$. The cost of NS and IJ will not scale with T, while naturally cost of CV and IACV must scale with T; this apparent computational benefit of NS and IJ is simply due to the fact that NS and IJ ignore the iterative nature of the algorithm and thus are completely invalid (i.e., errors are higher than the noninformative "baseline") for early, pre-convergence times T.

C. A Larger Scale Simulation for IACV in GD

In Figures 5 and 6, we provide simulation results for comparing IACV and other methods when $n = 5000$, $p = 50$. We observe a similar pattern as in the existing plots: the approximation error of IACV is always better than NS and IJ before the convergence of the algorithm and IACV is much faster than the exact leave-one-out CV. If we further grow p and n, we find it is too expensive to run the whole program as the exact leave-one-out CV takes too much time to run for many independent trials, but we nonetheless would expect to see similar performance.

D. A Iteration-Dependent Upper Bound for $\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$

Notice that in Theorem 4.4, we only present a version of the upper bound which is independent of t for convenience. It is simple and clearly illustrates the idea that the error bound holds along the whole trajectory of the learning process. But based on our proof, a t-dependent iterative bound on the approximation error can be obtained and it will typically be sharper at the early stage of training. Specifically, in our proof of Theorem 4.4, (25) shows

$$\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \leq (1 - \alpha_t n \lambda_0)\|\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 + \alpha_t n \gamma \|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2^2, \quad \forall t \geq 1.$$
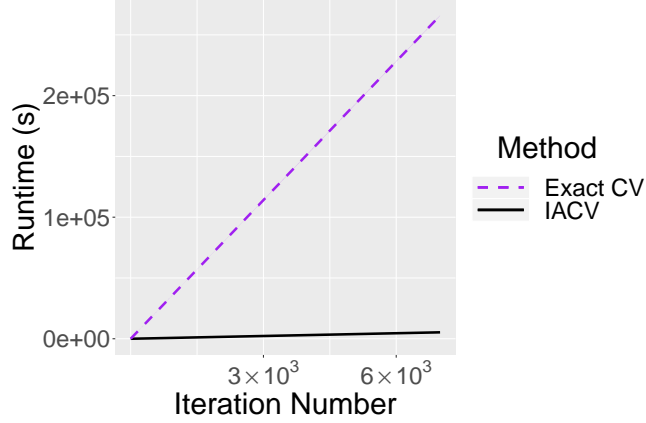
13

Figure 6. Runtime comparison of the baseline, NS, IJ, and IACV methods in gradient descent for logistic regression with ridge regularizer.

The error bound of $\widetilde{\theta}_{-i}^{(t)}$ depends on error bound of $\widetilde{\theta}_{-i}^{(t-1)}$ and $\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2$. Furthermore, $\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2$ also satisfies the following iterative upper bound by (23):

$$\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2 \leq (1 - \alpha_{t-1} n \lambda_0)\|\widehat{\theta}^{(t-2)} - \widehat{\theta}_{-i}^{(t-2)}\|_2 + \alpha_{t-1}\eta_i + \alpha_{t-1} n \gamma \|\widehat{\theta}^{(t-2)} - \widehat{\theta}_{-i}^{(t-2)}\|_2^2.$$

If we initialize $\widehat{\theta}^{(0)} = \widehat{\theta}_{-i}^{(0)} = \widetilde{\theta}_{-i}^{(0)} = 0$, we can see the error bound for $\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2$ will slightly increase from 0 over the iteration t and stabilize when the error contraction $(-\alpha_{t-1} n \lambda_0 \|\widehat{\theta}^{(t-2)} - \widehat{\theta}_{-i}^{(t-2)}\|_2)$ cancels out the per-iteration approximation error accumulation $(\alpha_{t-1}\eta_i + \alpha_{t-1} n \gamma \|\widehat{\theta}^{(t-2)} - \widehat{\theta}_{-i}^{(t-2)}\|_2^2)$. By a similar argument, we also have the error bound for $\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ will increase first and then stabilize.

E. Proofs

E.1. Proofs of error bound results

In this section we prove the approximation error bounds and CV error bounds for GD, SGD, and ProxGD.

Proof of Theorem 4.4. **Step 1.** In this step, we prove the bound on $\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ by induction. By the update rule of $\widehat{\theta}^{(t)}$ and $\widehat{\theta}_{-i}^{(t)}$, we have

$$
\begin{aligned}
\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)} &= \widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)} + \alpha_t \left(\nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}_{-i}^{(t-1)}) - \nabla_\theta F(\mathcal{Z}; \widehat{\theta}^{(t-1)}) \right) \\
&= \widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla \ell(Z_i; \widehat{\theta}^{(t-1)}) + \alpha_t \left(\nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}_{-i}^{(t-1)}) - \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) \right).
\end{aligned}
\tag{20}
$$

Given any $u \in \mathbb{R}^p$, by the Taylor expansion for $g_u(\theta) = \langle u, \nabla_\theta F(\mathcal{Z}_{-i}; \theta) \rangle$, we have

$$
\begin{aligned}
g_u(\widehat{\theta}_{-i}^{(t-1)}) &= g_u(\widehat{\theta}^{(t-1)}) + \nabla_\theta g_u(\theta') \\
\iff \langle u, \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}_{-i}^{(t-1)}) \rangle &= \langle u, \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) \rangle + \langle u, \nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta')[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \rangle \\
&= \langle u, \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) \rangle + \langle u, \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \rangle \\
&\quad + \langle u, (\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \rangle
\end{aligned}
\tag{21}
$$

where $\theta' = a\widehat{\theta}^{(t-1)} + (1 - a)\widehat{\theta}_{-i}^{(t-1)}$ for some $a \in [0, 1]$.

14

Plugging (21) into (20), we have for any $u \in \mathbb{R}^p$,

$$
\begin{aligned}
\langle \widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}, u \rangle &= \langle \widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \ell(Z_i; \widehat{\theta}^{(t-1)}) + \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}], u \rangle \\
&\quad + \alpha_t \langle u, (\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \rangle \\
&= \left\langle \left(\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) \right) [\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}] - \alpha_t \ell(Z_i; \widehat{\theta}^{(t-1)}), u \right\rangle \\
&\quad + \alpha_t \langle u, (\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \rangle.
\end{aligned}
\tag{22}
$$

Since $\widehat{\theta}^{(0)} = \widehat{\theta}_{-i}^{(0)}$, the error bound of $\widehat{\theta}^{(t)}$ to $\widehat{\theta}_{-i}^{(t)}$ holds at $t = 0$. Suppose now $\|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 \leq \frac{2\eta_i}{(2\lambda_0 - \gamma)n}$ holds for some $t \geq 1$, then by (22) and the fact that $\|\widehat{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{(t)}\|_2 = \sup_{u: \|u\|_2 = 1} \langle \widehat{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{(t)}, u \rangle$, we have

$$
\begin{aligned}
\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 &\overset{(a)}{\leq} \left\| \mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) \right\|_2 \|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 + \alpha_t \eta_i + \alpha_t n\gamma \|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2^2 \\
&\overset{(b)}{\leq} (1 - \alpha_t n\lambda_0) \|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 + \alpha_t \eta_i + \alpha_t n\gamma \|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2^2 \\
&\overset{(c)}{\leq} (1 - \alpha_t n\lambda_0) \frac{2\eta_i}{(2\lambda_0 - \gamma)n} + \alpha_t \eta_i + \alpha_t n\gamma \cdot \left(\frac{2\eta_i}{(2\lambda_0 - \gamma)n} \right)^2 \\
&\overset{(d)}{\leq} (1 - \alpha_t n\lambda_0) \frac{2\eta_i}{(2\lambda_0 - \gamma)n} + \alpha_t \eta_i + \alpha_t n\gamma \frac{\eta_i}{(2\lambda_0 - \gamma)n} = \frac{2\eta_i}{(2\lambda_0 - \gamma)n},
\end{aligned}
\tag{23}
$$

where (a) is by triangle inequality and Assumptions 4.2 and 4.3; (b) is by Assumption 4.1; (c) is by the induction assumption; (d) is because $n \geq \frac{4\|\eta\|_\infty}{2\lambda_0 - \gamma}$. This proves the desired bound on $\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$.

Step 2. We now prove the bound on $\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ by induction. First by the update rule of $\widehat{\theta}_{-i}^{(t)}$ and the Taylor expansion in (21), we have for any $u \in \mathbb{R}^p$,

$$
\begin{aligned}
\langle \widehat{\theta}_{-i}^{(t)}, u \rangle &= \left\langle \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \left(\nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \right), u \right\rangle \\
&\quad - \alpha_t \langle u, (\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \rangle.
\end{aligned}
$$

Combining with (12), we have

$$
\begin{aligned}
\langle \widehat{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{(t)}, u \rangle &= \left\langle \widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)}], u \right\rangle \\
&\quad - \alpha_t \langle u, (\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \rangle \\
&= \left\langle \left(\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) \right) [\widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)}], u \right\rangle \\
&\quad - \alpha_t \langle u, (\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \rangle.
\end{aligned}
\tag{24}
$$

Since $\widehat{\theta}_{-i}^{(0)} = \widetilde{\theta}_{-i}^{(0)}$, the error bound holds at $t = 0$. Suppose now $\|\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 \leq \frac{4\gamma\eta_i^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2}$ holds, then at time t, based on (24), we have

$$
\begin{aligned}
\|\widehat{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{(t)}\|_2 &\overset{(a)}{\leq} \left\| \mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) \right\|_2 \|\widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)}\|_2 + \alpha_t n\gamma \|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2^2 \\
&\overset{(b)}{\leq} (1 - \alpha_t n\lambda_0) \|\widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)}\|_2 + \alpha_t n\gamma \|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2^2 \\
&\overset{(c)}{\leq} (1 - \alpha_t n\lambda_0) \frac{4\gamma\eta_i^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2} + \alpha_t n\gamma \left(\frac{2\eta_i}{(2\lambda_0 - \gamma)n} \right)^2 \\
&= \frac{4\gamma\eta_i^2}{\lambda_0(2\lambda_0 - \gamma)^2 n^2}.
\end{aligned}
\tag{25}
$$

Here (a) is by the triangle inequality and Assumption 4.3; (b) is by Assumption 4.1; (c) is by the induction assumption and the error bound we have proved for $\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2$ in Step 1. This proves the desired bound on $\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$, and thuscompletes the proof of this theorem. $\square$

Proof of Theorem 4.6. First, for every $i \in [n]$, by the Taylor expansion, we have

$$\ell(Z_i; \widetilde{\theta}_{-i}^{(t)}) = \ell(Z_i; \widehat{\theta}_{-i}^{(t)}) + \nabla_\theta \ell(Z_i; \theta')^\top [\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}]$$

where $\theta' = a\widetilde{\theta}_{-i}^{(t)} + (1-a)\widehat{\theta}_{-i}^{(t)}$ for some $0 \le a \le 1$. Thus

$$|\ell(Z_i; \widetilde{\theta}_{-i}^{(t)}) - \ell(Z_i; \widehat{\theta}_{-i}^{(t)})| \le \|\nabla_\theta \ell(Z_i; \theta')\|_2 \cdot \|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \le \eta_i' \|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2. \tag{26}$$

Here the last inequality is by Assumption 4.5. This completes the proof of this theorem, once we plug in the bound on $\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ obtained in Theorem 4.4. $\qquad\square$

Proof of Theorem 4.11. **Step 1.** In this step, we prove the bound on $\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ by induction. Similar to (21), given any $u \in \mathbb{R}^p$, by the Taylor expansion for $\langle \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}_{-i}^{(t-1)}), u \rangle$, we have

$$\begin{aligned}
\langle \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}_{-i}^{(t-1)}), u \rangle = & \left\langle \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}], u \right\rangle \\
& + \langle (\nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}], u \rangle,
\end{aligned} \tag{27}$$

where $\theta' = a\widehat{\theta}^{(t-1)} + (1-a)\widehat{\theta}_{-i}^{(t-1)}$ with $0 \le a \le 1$.

So we have

$$\begin{aligned}
\langle \widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}, u \rangle = & \langle \widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta F(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}) + \alpha_t \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}_{-i}^{(t-1)}), u \rangle \\
\overset{(27)}{=} & \langle [\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})][\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}], u \rangle \\
& - \langle \alpha_t \nabla_\theta F(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}) - \alpha_t \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}), u \rangle \\
& + \alpha_t \langle (\nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}], u \rangle.
\end{aligned}$$

Then we have

$$\begin{aligned}
\mathbb{E}\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 \le & \underbrace{\mathbb{E}\left(\|[\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})][\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}]\|_2 \right)}_{(A)} \\
& + \alpha_t \underbrace{\mathbb{E}\|\nabla_\theta F(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}) - \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})\|_2}_{(B)} \\
& + \alpha_t \underbrace{\mathbb{E}\|(\nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]\|_2}_{(C)}.
\end{aligned} \tag{28}$$

Next, we bound the (A), (B), (C) terms separately. First for term (A),

$$\begin{aligned}
& \mathbb{E}\left(\|[\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})][\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}]\|_2 \right) \\
= & \mathbb{E}_{S_1, \dots, S_{t-1}} \left[\mathbb{E}_{S_t} (\|[\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})][\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}]\|_2) \right] \\
\overset{\text{Assumption (4.7)}}{\le} & (1 - \alpha_t K \lambda_0) \mathbb{E}_{S_1, \dots, S_{t-1}} \left[\|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 \right].
\end{aligned} \tag{29}$$

Next, for term (B),

$$\begin{aligned}
& \mathbb{E}\|\nabla_\theta F(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}) - \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})\|_2 \\
= & \mathbb{E}_{S_1, \dots, S_{t-1}} \left[\mathbb{E}_{S_t} \|\nabla_\theta F(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}) - \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})\|_2 \right] \\
= & \frac{K}{n} \mathbb{E}_{S_1, \dots, S_{t-1}} \|\nabla_\theta \ell(Z_i; \widehat{\theta}^{(t-1)})\|_2 \overset{\text{Assumption (4.8)}}{\le} \frac{K \eta_i}{n}.
\end{aligned} \tag{30}$$

Here the next-to-last step holds because each data point is included independently in S_t with probability K/n; thus with probability $1 - \frac{K}{n}$, we have $\nabla_\theta F(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}) - \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}) = 0$ and with probability $\frac{K}{n}$, $i \in S_t$ and in that case we have $\nabla_\theta F(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)}) - \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}) = \nabla_\theta \ell(Z_i; \widehat{\theta}^{(t-1)})$.

And, for term (C),

$$
\begin{aligned}
&\mathbb{E}\|(\nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]\|_2 \\
&\leq \mathbb{E} \sup_{b \in [0,1]} \|(\nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; b\widehat{\theta}^{(t-1)} + (1-b)\widehat{\theta}_{-i}^{(t-1)}) - \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]\|_2 \\
&\overset{\text{Assumption } 4.9}{\leq} \gamma K \mathbb{E}_{S_1, \ldots, S_{t-1}} \left[\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2^2 \right] \\
&\overset{\text{Assumption } 4.10}{\leq} \gamma K \beta \frac{n}{K} \left(\mathbb{E}\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2 \right)^2 = \gamma \beta n \left(\mathbb{E}\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2 \right)^2.
\end{aligned}
\tag{31}
$$

Since $\widehat{\theta}^{(0)} = \widehat{\theta}_{-i}^{(0)}$, the error bound of $\widehat{\theta}^{(t)}$ to $\widehat{\theta}_{-i}^{(t)}$ holds at $t = 0$. Suppose now $\mathbb{E}\|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 \leq \frac{2\eta_i}{(2\lambda_0 - \gamma\beta)n}$ holds for some $t \geq 1$, then by plugging (29), (30) and (31) into (28), we have

$$
\begin{aligned}
\mathbb{E}\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 &\leq (1 - \alpha_t K \lambda_0) \mathbb{E}\|\widehat{\theta}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 + \frac{\alpha_t K \eta_i}{n} + \alpha_t \gamma \beta n \left(\mathbb{E}\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2 \right)^2 \\
&\leq (1 - \alpha_t K \lambda_0) \frac{2\eta_i}{(2\lambda_0 - \gamma\beta)n} + \frac{\alpha_t K \eta_i}{n} + \alpha_t K \gamma \beta \frac{\eta_i}{(2\lambda_0 - \gamma\beta)n} \frac{4\eta_i}{(2\lambda_0 - \gamma\beta)K} \\
&\leq (1 - \alpha_t K \lambda_0) \frac{2\eta_i}{(2\lambda_0 - \gamma\beta)n} + \frac{\alpha_t K \eta_i}{n} + \alpha_t K \gamma \beta \frac{\eta_i}{(2\lambda_0 - \gamma\beta)n} = \frac{2\eta_i}{(2\lambda_0 - \gamma\beta)n},
\end{aligned}
\tag{32}
$$

where the next-to-last step holds since we have assumed $K \geq \frac{4\|\eta\|_\infty}{2\lambda_0 - \gamma\beta}$ by the assumption on K. This finishes the proof for the first part.

Step 2. We now prove the bound on $\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$ by induction. Based on the Taylor expansion in (27) and the update rule of $\widetilde{\theta}_{-i}^{(t)}$ in (13), we have for any $u \in \mathbb{R}^p$,

$$
\begin{aligned}
\langle \widehat{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{(t)}, u \rangle &= \langle \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}_{-i}^{(t-1)}) - \widetilde{\theta}_{-i}^{(t-1)}, u \rangle \\
&\quad + \left\langle \alpha_t \left(\nabla_\theta F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \right), u \right\rangle \\
&\overset{(27)}{=} \left\langle \left(\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}) \right) [\widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)}], u \right\rangle \\
&\quad - \alpha_t \langle (\nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}], u \rangle.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
\mathbb{E}\|\widehat{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{(t)}\|_2 &\leq \mathbb{E} \left[\left(\|\mathbf{I}_p - \alpha_t \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}) \right) [\widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)}]\|_2 \right] \\
&\quad + \alpha_t \mathbb{E} \left(\|(\nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \theta') - \nabla_\theta^2 F(\mathcal{Z}_{S_t \setminus \{i\}}; \widehat{\theta}^{(t-1)}))[\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]\|_2^2 \right) \\
&\leq (1 - \alpha_t K \lambda_0) \mathbb{E}\|\widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)}\|_2 + \alpha_t \gamma \beta n \left(\mathbb{E}\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2 \right)^2 \\
&\leq (1 - \alpha_t K \lambda_0) \mathbb{E}\|\widehat{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{(t-1)}\|_2 + \alpha_t \gamma \beta n \left(\frac{2\eta_i}{(2\lambda_0 - \gamma\beta)n} \right)^2.
\end{aligned}
\tag{33}
$$

Here the next-to-last step holds by a similar argument as in (29) and (31), while the last step holds by the error bound we have proved for $\mathbb{E}\|\widehat{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}\|_2$ in Step 1.

Since $\widehat{\theta}_{-i}^{(0)} = \widetilde{\theta}_{-i}^{(0)}$, the error bound holds at $t = 0$. Suppose now $\mathbb{E}\|\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}_{-i}^{(t-1)}\|_2 \leq \frac{4\gamma\beta\eta_i^2}{\lambda_0(2\lambda_0 - \gamma\beta)^2 nK}$ holds, then at time t, based on (33), we have

$$
\mathbb{E}\|\widehat{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{(t)}\|_2 \leq (1 - \alpha_t K \lambda_0) \frac{4\gamma\beta\eta_i^2}{\lambda_0(2\lambda_0 - \gamma\beta)^2 nK} + \alpha_t \gamma n \beta \left(\frac{2\eta_i}{(2\lambda_0 - \gamma\beta)n} \right)^2 = \frac{4\gamma\beta\eta_i^2}{\lambda_0(2\lambda_0 - \gamma\beta)^2 nK},
\tag{34}
$$

which proves the desired bound and thus completes the proof of this theorem.

$\square$

Proof of Theorem 4.13. First, for every $i \in [n]$, by the Taylor expansion, we have

$$\ell(Z_i; \widetilde{\theta}_{-i}^{(t)}) = \ell(Z_i; \widehat{\theta}_{-i}^{(t)}) + \nabla_\theta \ell(Z_i; \theta')^\top [\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}],$$

where $\theta' = a\widetilde{\theta}_{-i}^{(t)} + (1-a)\widehat{\theta}_{-i}^{(t)}$ for some $0 \le a \le 1$.

Next

$$\begin{aligned}
\mathbb{E}(\text{Err}_{\text{CV}}^{(t)}) &\le \frac{1}{n} \sum_{i=1}^n \mathbb{E}(|\ell(Z_i; \widetilde{\theta}_{-i}^{(t)}) - \ell(Z_i; \widehat{\theta}_{-i}^{(t)})|) \\
&\le \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left\{ \sup_{a \in [0,1]} |\nabla_\theta \ell(Z_i; a\widetilde{\theta}_{-i}^{(t)} + (1-a)\widehat{\theta}_{-i}^{(t)})^\top [\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}]| \right\} \\
&\le \frac{1}{n} \sum_{i=1}^n \frac{4\eta_i' \eta_i^2 \gamma \beta}{\lambda_0 (2\lambda_0 - \gamma\beta)^2 nK},
\end{aligned}$$

where the last inequality is by Assumption 4.12 and Theorem 4.11. $\square$

Proof of Theorem A.1. The proof strategy of this theorem is similar to the proof of Theorem 4.4. For convenience, we will define

$$\text{prox}_h^{\alpha_t}(\theta') := \arg\min_\theta \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}. \tag{35}$$

A key property we use here is that the proximal operator $\text{prox}_h^{\alpha_t}$ is nonexpansive for convex h (Wright & Recht, 2022, Proposition 8.19), i.e.,

$$\|\text{prox}_h^{\alpha_t}(x) - \text{prox}_h^{\alpha_t}(x')\|_2 \le \|x - x'\|_2.$$

Step 1. In this step, we derive the error bound for $\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2$. By the updating rules of $\widehat{\theta}^{(t)}$ and $\widehat{\theta}_{-i}^{(t)}$, we have

$$\begin{aligned}
\|\widehat{\theta}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 &= \left\| \text{prox}_h^{\alpha_t} \left(\widehat{\theta}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}; \widehat{\theta}^{(t-1)}) \right) - \text{prox}_h^{\alpha_t} \left(\widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}_{-i}^{(t-1)}) \right) \right\|_2 \\
&\le \left\| \widehat{\theta}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}; \widehat{\theta}^{(t-1)}) - (\widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}_{-i}^{(t-1)})) \right\|_2.
\end{aligned}$$

Here the inequality holds by the nonexpansiveness property of the proximal operator. The rest of the proof follows exactly as in Step 1 of Theorem 4.4 by replacing $F(\cdot; \theta)$ with $g(\cdot; \theta)$.

Step 2. By our estimator in (14), we have

$$\begin{aligned}
\|\widetilde{\theta}_{-i}^{(t)} - \widehat{\theta}_{-i}^{(t)}\|_2 &= \left\| \text{prox}_h^{\alpha_t} \left(\widetilde{\theta}_{-j}^{(t-1)} - \alpha_t (\nabla_\theta g(\mathcal{Z}_{-j}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 g(\mathcal{Z}_{-j}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-j}^{(t-1)} - \widehat{\theta}^{(t-1)}]) \right) \right. \\
&\quad \left. - \text{prox}_h^{\alpha_t} \left(\widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}_{-i}^{(t-1)}) \right) \right\|_2 \\
&\le \| \widetilde{\theta}_{-j}^{(t-1)} - \alpha_t (\nabla_\theta g(\mathcal{Z}_{-j}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 g(\mathcal{Z}_{-j}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-j}^{(t-1)} - \widehat{\theta}^{(t-1)}]) - \widehat{\theta}_{-i}^{(t-1)} \\
&\quad + \alpha_t \nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}_{-i}^{(t-1)}) \|_2,
\end{aligned}$$

where the inequality is again due to the nonexpansiveness property of the proximal operator. The remainder of the proof again follows the same argument as in Step 2 of Theorem 4.4 by replacing $F(\cdot; \theta)$ with $g(\cdot; \theta)$. $\square$

E.2. Proofs of convergence results

In this section, we prove Theorems 5.1 and A.3, which establish convergence of the IACV estimator to the NS estimator in the GD and ProxGD setting, respectively.

Proof of Theorem 5.1. Since we assume $\widehat{\theta}^{(t)}$ converges to $\widehat{\theta}$ and Assumption 4.1 is satisfied at all iterations, we therefore see that $\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta})$ is invertible by the continuity of $\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta)$. Let us denote

$$b_i^t = \widehat{\theta}^{(t-1)} - (\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))^{-1} \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}).$$

First, based on the update rule for $\widetilde{\theta}_{-i}^{(t)}$ in (12), we have:

$$\widetilde{\theta}_{-i}^{(t)} - b_i^t = [\mathbf{I}_p - \alpha \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})](\widetilde{\theta}_{-i}^{(t-1)} - b_i^t).$$

This implies

$$\begin{aligned}
\widetilde{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{NS} &= [\mathbf{I}_p - \alpha \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})](\widetilde{\theta}_{-i}^{(t-1)} - b_i^t) + b_i^t - \widetilde{\theta}_{-i}^{NS} \\
&= [\mathbf{I}_p - \alpha \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})](\widetilde{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{NS}) - \alpha \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})(\widetilde{\theta}_{-i}^{NS} - b_i^t),
\end{aligned}$$

and thus

$$\begin{aligned}
\|\widetilde{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{NS}\|_2 &\leq \|\mathbf{I}_p - \alpha \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\| \cdot \|\widetilde{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{NS}\|_2 + \alpha \|\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\| \cdot \|\widetilde{\theta}_{-i}^{NS} - b_i^t\|_2 \\
&\leq (1 - \alpha n \lambda_0) \cdot \|\widetilde{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{NS}\|_2 + \|\widetilde{\theta}_{-i}^{NS} - b_i^t\|_2,
\end{aligned}$$

where the last step holds by Assumption 4.1 together with the choice of α.

Note that since $\nabla_\theta^2 F(\mathcal{Z}_{-i}; \theta)$ and $\nabla_\theta F(\mathcal{Z}_{-i}; \theta)$ are continuous in θ, we then have $b_i^t \to \widetilde{\theta}_{-i}^{NS}$ as $\widehat{\theta}^{(t)}$ converges to $\widehat{\theta}$, i.e.,

$$\lim_{t \to \infty} \|\widetilde{\theta}_{-i}^{NS} - b_i^t\|_2 = 0.$$

Combined with the above, this is sufficient to prove that

$$\lim_{t \to \infty} \|\widetilde{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{NS}\|_2 = 0,$$

which completes the proof of the theorem. $\qquad\square$

Proof of Theorem A.3. Let us first prove the claim that $\widetilde{\theta}_{-i}^{NS}$ in (17) is a fixed point of

$$a^{(t)} = \text{prox}_h^\alpha \left(a^{(t-1)} - \alpha \left(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})[a^{(t-1)} - \widehat{\theta}] \right) \right), \tag{36}$$

where prox is defined in (35). To show this, it is equivalent to show $\widetilde{\theta}_{-i}^{NS} = \text{prox}_h^\alpha \left(\widetilde{\theta}_{-i}^{NS} - \alpha \left(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})[\widetilde{\theta}_{-i}^{NS} - \widehat{\theta}] \right) \right)$.

Since h is convex, by the definition of prox in (18) and the definition of $\widetilde{\theta}_{-i}^{NS}$ in (17), the first-order optimality condition tells us that $\widetilde{\theta}_{-i}^{NS}$ must satisfy

$$0 \in -(\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}) \left(\widehat{\theta} - \widetilde{\theta}_{-i}^{NS} \right) - \nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta})) + \partial h(\widetilde{\theta}_{-i}^{NS}). \tag{37}$$

Also, the proximal operator prox_h^α has a unique solution as h is convex. By the definition of the proximal operator prox_h^α, the output of $\text{prox}_h^\alpha \left(\widetilde{\theta}_{-i}^{NS} - \alpha \left(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})[\widetilde{\theta}_{-i}^{NS} - \widehat{\theta}] \right) \right)$, say x^*, is the unique vector that satisfies

$$0 \in -\frac{\widetilde{\theta}_{-i}^{NS} - \alpha \left(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})[\widetilde{\theta}_{-i}^{NS} - \widehat{\theta}] \right) - x^*}{\alpha} + \partial h(x^*). \tag{38}$$

Notice that $x^* = \widetilde{\theta}_{-i}^{\mathrm{NS}}$ satisfies (38) because of (37). Thus, $\widetilde{\theta}_{-i}^{\mathrm{NS}}$ is a fixed point of (36) as desired.

Next, we show $\widetilde{\theta}_{-i}^{(t)}$ in (14) converges to $\widetilde{\theta}_{-i}^{\mathrm{NS}}$ under the assumptions given in the statement of the theorem. By (14) together with the calculations above, we have

$$\widetilde{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{\mathrm{NS}} = \mathrm{prox}_h^\alpha \left(\widetilde{\theta}_{-i}^{(t-1)} - \alpha(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]) \right)$$
$$- \mathrm{prox}_h^\alpha \left(\widetilde{\theta}_{-i}^{\mathrm{NS}} - \alpha \left(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})[\widetilde{\theta}_{-i}^{\mathrm{NS}} - \widehat{\theta}] \right) \right).$$

So we have

$$\|\widetilde{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{\mathrm{NS}}\|_2 = \left\| \mathrm{prox}_h^\alpha \left(\widetilde{\theta}_{-i}^{(t-1)} - \alpha(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]) \right) \right.$$
$$\left. - \mathrm{prox}_h^\alpha \left(\widetilde{\theta}_{-i}^{\mathrm{NS}} - \alpha \left(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})[\widetilde{\theta}_{-i}^{\mathrm{NS}} - \widehat{\theta}] \right) \right) \right\|_2$$
$$\leq \left\| \left(\widetilde{\theta}_{-i}^{(t-1)} - \alpha(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})[\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}]) \right) \right.$$
$$\left. - \left(\widetilde{\theta}_{-i}^{\mathrm{NS}} - \alpha \left(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}) + \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})[\widetilde{\theta}_{-i}^{\mathrm{NS}} - \widehat{\theta}] \right) \right) \right\|_2$$
$$= \left\| [\mathbf{I} - \alpha \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})](\widetilde{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{\mathrm{NS}}) + \alpha(\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}) - \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))\widetilde{\theta}_{-i}^{\mathrm{NS}} \right.$$
$$\left. - \alpha(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) - \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\widehat{\theta}^{(t-1)}) + \alpha(\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}) - \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})\widehat{\theta}) \right\|_2 \quad (39)$$
$$\leq \left\| [\mathbf{I} - \alpha \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})](\widetilde{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{\mathrm{NS}}) \right\|_2 + \alpha \Big[\|(\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}) - \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))\widetilde{\theta}_{-i}^{\mathrm{NS}}\|_2$$
$$+ \|\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) - \nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta})\|_2 + \|\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\widehat{\theta}^{(t-1)} - \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})\widehat{\theta}\|_2 \Big]$$
$$\leq (1 - \alpha n \lambda_0) \cdot \|\widetilde{\theta}_{-i}^{(t-1)} - \widetilde{\theta}_{-i}^{\mathrm{NS}}\|_2 + \alpha \Big[\|(\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}) - \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}))\widetilde{\theta}_{-i}^{\mathrm{NS}}\|_2$$
$$+ \|\nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)}) - \nabla_\theta g(\mathcal{Z}_{-i}; \widehat{\theta})\|_2 + \|\nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta}^{(t-1)})\widehat{\theta}^{(t-1)} - \nabla_\theta^2 g(\mathcal{Z}_{-i}; \widehat{\theta})\widehat{\theta}\|_2 \Big].$$

where the second step holds because prox_h^α is nonexpansive for convex h (Wright & Recht, 2022, Proposition 8.19), and the last step applies Assumption 4.1 (with g in place of F, as assumed in the Theorem). Since $\nabla_\theta^2 g(\mathcal{Z}_{-i}; \theta)$, $\nabla_\theta g(\mathcal{Z}_{-i}; \theta)$ are continuous in θ, and $\widehat{\theta}^{(t)}$ is assumed to converge to $\widehat{\theta}$, the quantity in square brackets converges to zero. As in the proof of Theorem 5.1, this therefore implies

$$\lim_{t \to \infty} \|\widetilde{\theta}_{-i}^{(t)} - \widetilde{\theta}_{-i}^{\mathrm{NS}}\|_2 \to 0.$$

$\square$

Can Theorem 5.1 be extended to SGD? A key intuition behind the proof of Theorem 5.1 is that $\widetilde{\theta}_{-i}^{\mathrm{NS}}$ in (3) is a fixed point of the following update equation

$$x^t = x^{t-1} - \alpha_t \Big(\nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}) + \nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta})[x^{t-1} - \widehat{\theta}] \Big).$$

In addition, we use the fact that given any $\epsilon > 0$, a sequence of scalars $\{x^t\}$ that satisfies $x^t = (1 - \alpha_t)x^{t-1} + \epsilon$ must have $x^t \to 0$ when we take $\alpha_t \equiv \alpha > 0$. Finally, we note in the proof of Theorem 5.1, the condition "$\widehat{\theta}$ is the solution of (1)" is actually not used, so the conclusions there continue to hold if we replace $\widehat{\theta}$ with any $\mathring{\theta}$ as long as: (1) $\widehat{\theta}^{(t)}$ converges to $\mathring{\theta}$ and (2) Assumption 4.1 is satisfied along the iterates. A similar intuition holds for the proof of Theorem A.3.

In the SGD setting, on the other hand, the techniques for proving Theorem 5.1 fail as α_t there are required to decay to zero eventually for convergence. Thus, if we have a sequence of values $\{x^t\}$ satisfying $x^t = (1 - \alpha_t)x^{t-1} + \epsilon$ for any $\epsilon > 0$, this does not necessarily imply $x^t \to 0$. Whether a result analogous to Theorem 5.1 might be provable for SGD via a different technique remains an open question.