Failures and Successes of Cross-Validation for Early-Stopped Gradient Descent

Pratik Patil
University of California, Berkeley

Yuchen Wu
University of Pennsylvania

Ryan J. Tibshirani
University of California, Berkeley

Abstract

We analyze the statistical properties of generalized cross-validation (GCV) and leave-one-out cross-validation (LOOCV) applied to early-stopped gradient descent (GD) in high-dimensional least squares regression. We prove that GCV is generically inconsistent as an estimator of the prediction risk of early-stopped GD, even for a well-specified linear model with isotropic features. In contrast, we show that LOOCV converges uniformly along the GD trajectory to the prediction risk. Our theory requires only mild assumptions on the data distribution and does not require the underlying regression function to be linear. Furthermore, by leveraging the individual LOOCV errors, we construct consistent estimators for the entire prediction error distribution along the GD trajectory and consistent estimators for a wide class of error functionals. This in particular enables the construction of pathwise prediction intervals based on GD iterates that have asymptotically correct nominal coverage conditional on the training data.

1 INTRODUCTION

Cross-validation (CV) is a widely used tool for assessing and selecting models in various predictive applications of statistics and machine learning. It is often used to tune the level of regularization strength in *explicitly regularized* methods, such as ridge regression and lasso. In general, CV error is based on an iterative scheme that allows each data sample to play a role in training and validation in different iterations. Minimizing CV error helps to identify a trade-off between bias and variance that favors prediction accuracy (Hastie et al., 2009).

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Meanwhile, especially in the modern era, techniques such as gradient descent (GD) and its variants are central tools for optimizing the parameters of machine learning models. Even when applied to models without explicit regularization, these algorithms are known to induce what is called *implicit regularization* in various settings (Bartlett et al., 2021; Belkin, 2021; Ji and Telgarsky, 2019; Nacson et al., 2019). For example, in the simplest case of least squares regression, GD and stochastic GD iterates bear a close connection to explicitly regularized ridge regression estimates (Suggala et al., 2018; Neu and Rosasco, 2018; Ali et al., 2019, 2020).

This naturally leads to the following question:

Can we reliably use CV to assess model performance along the trajectory of iterative algorithms?

An affirmative answer to this question would enable the use of cross-validation to determine when to stop the GD training procedure, preventing overfitting and appropriately balancing the level of implicit regularization. Motivated by this, we investigate the statistical properties of two popular CV procedures, namely generalized cross-validation (GCV) and leave-one-out cross-validation (LOOCV), along the gradient descent trajectory in high-dimensional linear regression.

Previously, it has been noted that some common variants of CV: split-sample validation and K-fold CV with small K (such as 5 or 10), can suffer from significant bias when the number of observations and features scale proportionally (Rad and Maleki, 2020; Rad et al., 2020). Although LOOCV in most cases mitigates bias issues, it is typically computationally expensive to implement. Fortunately, for estimators that are linear smoothers (linear in the response vector), GCV serves as an efficient approximation to LOOCV in classical low-dimensional problems (Golub et al., 1979; Jansen et al., 1997). Furthermore, recent work has shown that both LOOCV and GCV are consistent for estimating the out-of-sample prediction risk of ridge regression in high-dimensional settings (Patil et al., 2021, 2022b; Wei et al., 2022; Han and Xu, 2023).

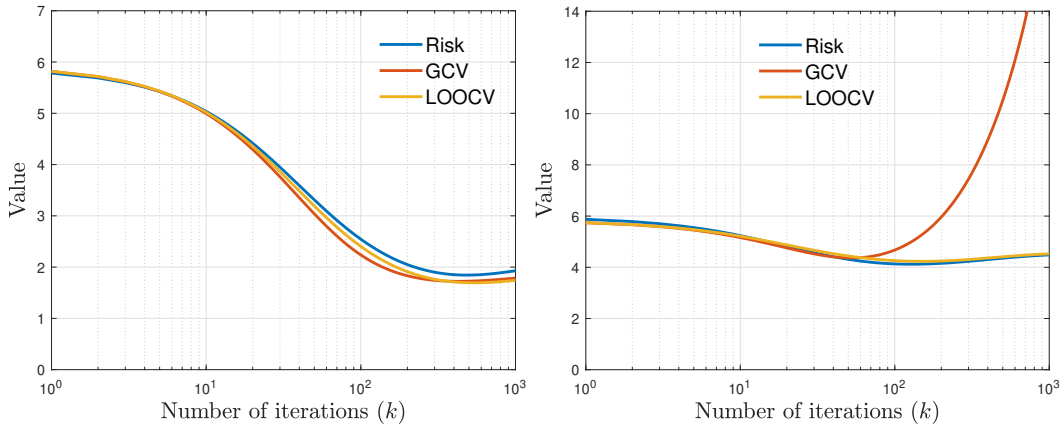

Figure 1: **GCV can perform poorly in overparameterized problems, yet LOOCV gives accurate risk estimates.** We investigate the risk of early-stopped gradient descent, applied to the least squares loss, as a function of iteration number. The *left* panel shows an underparameterized experiment with $n = 3000$, $p = 1500$, and the *right* panel an overparameterized experiment with $n = 3000$, $p = 6000$. In both cases, the data is generated from a linear model with i.i.d. standard normal features, a true signal vector with ℓ_2 norm of 5, and noise standard deviation of 1. GD uses a constant step size of 0.01. In the overparameterized case, we can see that the GCV risk estimate deviates wildly from the true risk, whereas LOOCV remains accurate throughout the entire path.

Noting that for least squares loss the GD iterates are linear smoothers, and recalling the connection between ridge regression and early-stopped GD in this problem setting (Ali et al., 2019), a natural idea would then be to use GCV to estimate the out-of-sample prediction risk of early-stopped GD iterates. To our knowledge, the performance of GCV in this setting has not yet been studied.

In this work, we derive precise theory for both GCV and LOOCV applied to the GD iterates from high-dimensional least squares. Our first and somewhat surprising result establishes that GCV is generically inconsistent for the out-of-sample prediction risk of early-stopped GD, even in the most idealized setting of a well-specified linear model with isotropic features. This inconsistency becomes particularly pronounced in the overparameterized regime, where the number of features is greater than the number of observations. In such a case, the gap between GCV and risk can be substantial, especially as the GD iteration progresses. This is, of course, problematic for model tuning, as these are precisely the scenarios in which the optimal stopping time for GD can occur at a large iteration that allows for (near) interpolation (for the analogous theory for ridge regression, see Kobak et al. (2020); Wu and Xu (2020); Richards et al. (2021)).

Our second result concerns LOOCV and establishes that it is consistent for the out-of-sample prediction risk, in a uniform sense over the GD path. For this, we make only weak assumptions on the feature distribution and do not assume a well-specified model (i.e., allowing the true regression function to be nonlinear). One interpretation this suggests is that the failure of GCV lies in its ability to approximate LOOCV, and not with LOOCV itself. Figure 1 showcases an empirical illustration of our main results, which we summarize below.

1.1 Summary of Main Results

1. **GCV inconsistency.** Under a proportional asymptotics model where the number of features p and observations n scale proportionally, and assuming a well-specified linear model and isotropic features, we show that GCV is inconsistent for estimating the prediction risk throughout basically the entire GD path (Theorem 1). We prove this result by separately deriving the asymptotic limits for the GCV estimator and the true risk of early-stopped GD, and then showing that they do not match.

2. **LOOCV consistency.** Under a proportional asymptotics model again, we show that LOOCV is consistent for estimating the prediction risk of early-stopped GD, in a uniform sense over the GD iterations (Theorem 2). Our analysis only requires the distributions of the features and noise to satisfy a T_2-inequality, which is quite weak. In particular, we do not assume any specific model for the regression function. As a consequence of uniformity, we establish that the risk of the LOOCV-tuned iterate almost surely matches that of the oracle-tuned iterate. Furthermore, we also propose an implementation of the LOOCV with lower computational complexity compared to the naive implementation (Proposition 7).

3. **Functional consistency.** Beyond prediction risk, we propose a natural extension of LOOCV to estimate general functionals of the prediction error

distribution for early-stopped GD, which is a plug-in approach based on the empirical distribution of LOOCV errors (Theorem 3). As an application, we use this to consistently estimate the quantiles of the prediction error distribution for GD iterates, allowing the construction of prediction intervals with asymptotically correct nominal coverage conditional on the training data (Theorem 4).

1.2 Related Work

GD and its variants are central tools for training modern machine learning models. These methods, especially stochastic gradient methods, can be highly scalable. But, somewhat surprisingly, overparameterized models trained with GD and variants also often generalize well, even in the absence of explicit regularizers and with noisy labels (Zhang et al., 2017). This behavior is often attributed to the fact that the GD iterates are subject to a kind of implicit regularization (Wilson et al., 2017; Gunasekar et al., 2018a,b). Implicit regularization has a rich history in machine learning and has appeared in some of the first insights into the advantages of early stopping in neural network training (Morgan and Bourlard, 1989). A parallel idea in numerical analysis is known as the Landweber iteration (Landweber, 1951; Strand, 1974). There is a rich literature on early stopping in the context of boosting (Bühlmann and Yu, 2003; Rosset et al., 2004; Zhang and Yu, 2005; Yao et al., 2007; Bauer et al., 2007; Raskutti et al., 2014; Wei et al., 2017). Furthermore, several precise correspondences between GD and ridge penalized estimators have been established by Suggala et al. (2018); Neu and Rosasco (2018); Ali et al. (2019, 2020), among others.

CV is a standard approach in statistics for parameter tuning and model selection. For classic work on CV, see, e.g., Allen (1974); Stone (1974, 1977); Geisser (1975). For practical surveys, see Arlot and Celisse (2010); Zhang and Yang (2015). More recently, there has been renewed interest in developing a modern theory for CV, with contributions from Kale et al. (2011); Kumar et al. (2013); Celisse and Guedj (2016); Austern and Zhou (2020); Bayle et al. (2020); Lei (2020); Rad et al. (2020), among others. As LOOCV is, in general, computationally expensive, there has also been recent work in designing and analyzing approximate leave-one-out methods to address the computational burden; see, e.g., Wang et al. (2018); Stephenson and Broderick (2020); Wilson et al. (2020); Rad and Maleki (2020); Auddy et al. (2023).

GCV is an approximation to LOOCV and is closely connected to what is called the "shortcut" leave-one-out formula for linear smoothers. The classic work on GCV includes Craven and Wahba (1979); Golub et al. (1979); Li (1985, 1986, 1987). Recently, GCV has garnered significant interest, as it has been found to be consistent for out-of-sample prediction risk in various high-dimensional settings; see, e.g., Hastie et al. (2022); Adlam and Pennington (2020); Patil et al. (2021, 2022b); Wei et al. (2022); Du et al. (2023); Han and Xu (2023); Patil and LeJeune (2024). While originally defined for linear smoothers, the idea of using similar degrees-of-freedom adjustments can be extended beyond this original scope to nonlinear predictors; see, e.g., Bayati and Montanari (2011); Bayati et al. (2013); Miolane and Montanari (2021); Bellec and Shen (2022); Bellec (2023).

Most of the aforementioned papers on CV have focused on estimators that are defined as solutions to empirical risk minimization problems. There has been little work that studies CV for iterates of optimization algorithms like GD, which are commonly used to find solutions (train models) in practice. Very recently, Luo et al. (2023) consider approximating LOOCV for iterative algorithms. They propose an algorithm that is more efficient than the naive LOOCV when $p \ll n$. They also show that their method approximates LOOCV well. In our work, we instead focus on analyzing LOOCV itself, along with GCV, for least squares problems, which we view as complementary to their work. Moreover, our analysis is in the proportional asymptotic regime, where $p \asymp n$.

2 PRELIMINARIES

In this section, we define the main object of study: early-stopped GD applied to the least squares loss. We then precisely define the risk metric of interest and describe the risk estimators based on LOOCV and GCV.

2.1 Early-Stopped Gradient Descent

Consider a standard regression setting, where we observe independent and identically distributed samples $\{(x_i, y_i)\} \in \mathbb{R}^{p+1} \times \mathbb{R}$ for $i \in [n]$. Here, each $x_i \in \mathbb{R}^{p+1}$ denotes a feature vector and $y_i \in \mathbb{R}$ its response value. The last entry of each x_i is set to 1 to accommodate an intercept term in the regression model. Let $X \in \mathbb{R}^{n \times (p+1)}$ denote the feature matrix whose i-th row contains $x_i^\top$, and $y \in \mathbb{R}^n$ the response vector whose i-th entry contains y_i.

We focus on the ordinary least squares problem:

$$\underset{\beta \in \mathbb{R}^{p+1}}{\text{minimize}} \ \frac{1}{2n} \|y - X\beta\|_2^2, \tag{1}$$

and we study the sequence of estimates defined by applying gradient descent (GD) to the squared loss in (1). Specifically, given step sizes $\delta = (\delta_0, \ldots, \delta_{K-1}) \in \mathbb{R}^K$, and initializing GD at the origin, $\widehat{\beta}_0 = 0$, the GD iterates are defined recursively as follows:

$$\widehat{\beta}_k = \widehat{\beta}_{k-1} + \frac{\delta_{k-1}}{n} X^\top (y - X\widehat{\beta}_{k-1}), \quad k \in [K]. \tag{2}$$

Let $(x_0, y_0) \in \mathbb{R}^{p+1} \times \mathbb{R}$ denote a test point drawn independently from the same distribution as the training data. We are interested in estimating the out-of-sample prediction risk along the GD path. More precisely, we are interested in estimating the squared prediction error $R(\widehat{\beta}_k)$ achieved by the GD iterate at each step $k \in [K]$, defined as:

$$R(\widehat{\beta}_k) = \mathbb{E}_{x_0,y_0}\big[(y_0 - x_0^\top \widehat{\beta}_k)^2 \mid X, y\big]. \tag{3}$$

Note that our notion of risk here is conditional on the training features and responses, X, y.

2.2 GCV and LOOCV

Next, we present an overview of the LOOCV and GCV estimators associated with GD iterates. First, we describe the estimators that correspond to the squared prediction risk. The exact LOOCV estimator for the squared prediction risk of $\widehat{\beta}_k$ is defined as:

$$\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \widehat{\beta}_{k,-i})^2, \tag{4}$$

where $\widehat{\beta}_{k,-i}$ denotes the GD iterate after k iterations trained on the data X_{-i}, y_{-i}, which excludes the i-th sample from the full data X, y. To be explicit, X_{-i} is the result of removing the i-th row of X, and y_{-i} is the result of removing the i-th coordinate of y.

Towards defining GCV, suppose that we have a predictor $\widehat{f}\colon \mathbb{R}^{p+1} \to \mathbb{R}$ which is a linear smoother, i.e., $\widehat{f}(x) = s_x^\top y$ for some vector $s_x \in \mathbb{R}^n$ which depends only on the feature matrix X and the test point x. The smoothing matrix associated with the predictor $\widehat{f}$ is denoted $S \in \mathbb{R}^{n\times n}$ and defined to have rows $s_{x_1}^\top, \ldots, s_{x_n}^\top$. The GCV estimator of the prediction risk of $\widehat{f}$ is defined as:

$$\widehat{R}^{\mathrm{gcv}}(\widehat{f}) = \frac{\|y - Sy\|_2^2/n}{(1 - \mathrm{tr}[S]/n)^2}.$$

The numerator here is the training error, which is of course typically biased downward, meaning that it typically underestimates the prediction error. The denominator corrects for this downward bias, often referred to as the "optimism" of the training error, with $1 - \mathrm{tr}[S]/n$ acting as a degrees-of-freedom correction, which is smaller the more complex the model (the larger the trace of S).

A short calculation shows that each GD iterate can be represented as a linear smoother, i.e., the in-sample predictions can be written as $X\widehat{\beta}_k = H_k y$, where

$$H_k = \sum_{j=0}^{k-1} \frac{\delta_j}{n} X \prod_{r=1}^{k-j-1} \big(I_{p+1} - \delta_{k-r}\widehat{\Sigma}\big)X^\top,$$

and we denote by $\widehat{\Sigma} = X^\top X/n$ the sample covariance matrix. This motivates us to estimate its prediction risk using GCV:

$$\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - x_i^\top \widehat{\beta}_k)^2}{(1 - \mathrm{tr}[H_k]/n)^2}. \tag{5}$$

Perhaps surprisingly, as we will see shortly in Section 3, GCV does not consistently estimate the prediction risk for GD iterates, even if we assume a well-specified linear model. On the other hand, we will show in Section 4 that LOOCV is uniformly consistent along the GD path. We also later propose a modified "shortcut" in Section 5 that (1) exactly tracks the LOOCV estimates, and (2) is computationally more efficient than the naive implementation of LOOCV.

3 GCV INCONSISTENCY

In this section, we prove that GCV is generically inconsistent for estimating the squared prediction risk, even under a well-specified linear model with isotropic Gaussian features. For simplicity, in this section only, we consider fixed step sizes $\delta_k = \delta$ and omit the intercept term. We impose the following assumptions on the feature and response distributions.

Assumption A (Feature distribution)**.** Each feature vector $x_i \in \mathbb{R}^p$, for $i \in [n]$, contains i.i.d. Gaussian entries with mean 0 and variance 1.

Assumption B (Response distribution)**.** Each response variable y_i, for $i \in [n]$, follows a well-specified linear model: $y_i = x_i^\top \beta_0 + \varepsilon_i$. Here, $\beta_0 \in \mathbb{R}^p$ is an unknown signal vector satisfying $\lim_{p\to\infty} \|\beta_0\|_2^2 = r^2 < \infty$, and ε_i is a noise variable, independent of x_i, drawn from a Gaussian distribution with mean 0 and variance $\sigma^2 < \infty$.

The zero-mean condition for each y_i is used only for simplicity. (Accordingly, we do not include an additional intercept term in the model, implying that $x_i \in \mathbb{R}^p$.) Although one could establish the inconsistency of GCV under more relaxed assumptions, we choose to work under Assumptions A and B to highlight that GCV fails even under favorable conditions.

We analyze the behavior of the estimator in the proportional asymptotics regime, where both the number of samples n and the number of features p tend to infinity, and their ratio p/n converges to a constant $\zeta_* \in (0, \infty)$. This regime has received considerable attention recently in high-dimensional statistics and machine learning theory.

The dynamics of GD are determined by both the step size δ and the iterate number k. We study a regime in which $\delta \to 0$ and $k \to \infty$ as $n, p \to \infty$, which effectively reduces the GD iterates to a continuous-time gradient flow, as studied in other work (Ali et al.,

2019; Celentano et al., 2021; Berthier et al., 2023). Our main negative result, on GCV, is given next.

Theorem 1 (Inconsistency of GCV). *Suppose that (x_i, y_i), $i \in [n]$ are i.i.d., and satisfy both Assumptions A and B, where either $r^2 > 0$ or $\sigma^2 > 0$. As $n, p \to \infty$, assume $p/n \to \zeta_*$, and $k \to \infty$, $\delta \to 0$ such that $k\delta \to T$, where $T, \zeta_* > 0$ are constants. Then, for every fixed $\zeta_* > 0$, it holds that for almost all $T > 0$ (i.e., all $T > 0$ except for a set of Lebesgue measure zero),*

$$\left| \widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k) - R(\widehat{\beta}_k) \right| \not\xrightarrow{\mathrm{p}} 0, \tag{6}$$

where we recall that $\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k)$ and $R(\widehat{\beta}_k)$ are as defined in (5) and (3), respectively.

In other words, the theorem says that GCV does not consistently track the true prediction risk at basically any point along the GD path (in the sense that GCV can only possibly be consistent at a Lebesgue measure zero set of times T). It is worth noting that the inconsistency here can be severe especially in the overparameterized regime, when $\zeta_* > 1$. In particular, in this regime, it is easy to show that if $k\delta \to \infty$ (rather than $k\delta \to T$ for a finite limit T), then $\lim_{k\to\infty} \widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k) \to \infty$, while $R(\widehat{\beta}_K) \to r^2 + \sigma^2$, under the assumptions of Theorem 1. This is evident in Figure 1.

4 LOOCV CONSISTENCY

Despite the inconsistency of GCV, LOOCV remains consistent for GD. This section establishes a uniform consistency result for LOOCV along the GD path.

4.1 Squared Risk

We begin by focusing on squared prediction risk. The technical crux of our analysis revolves around establishing certain concentration properties of the LOOCV estimator $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$, and to do so, we leverage Talagrand's T_2-inequality (Gozlan, 2009). Specifically, under the assumption that both the entries of the feature and noise distributions satisfy the T_2-inequality, we show that $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$ behaves approximately as a Lipschitz function of these random variables. Together, these results enable us to leverage powerful dimension-free concentration inequalities.

The inspiration for using T_2-inequality comes from the recent work of Avelin and Viitasaari (2022). They assume that the data distribution satisfies the logarithmic Sobolev inequality (LSI), which is a strictly stronger condition than what we assume here. Furthermore, they only consider fixed p and do not consider iterative algorithms. The extensions we pursue present considerable technical challenges and require us to delicately upper bound the norms of various gradients involved.

Below we give a formal definition of what it means for a distribution to satisfy the T_2-inequality.

Definition 1 (T_2-inequality). We say a distribution μ satisfies the T_2-inequality if there exists a constant $\sigma(\mu) \geq 0$, such that for every distribution ν,

$$W_2(\mu, \nu) \leq \sqrt{2\sigma^2(\mu) D_{\mathrm{KL}}(\nu \| \mu)}, \tag{7}$$

where $W_2(\cdot, \cdot)$ is the 2-Wasserstein distance, and $D_{\mathrm{KL}}(\cdot \| \cdot)$ the Kullback-Leibler divergence.

The T_2-inequality is, in some sense, a necessary and sufficient condition for dimension-free concentration. We refer interested readers to Theorem 4.31 in Van Handel (2014) for more details (see also Appendix S.5.2 for further facts related to the T_2-inequality).

One prominent example of distributions that satisfy the T_2-inequality are distributions that satisfy the log Sobolev inequality (LSI); Appendix S.5.1 gives more details. We note that all distributions that are strongly log-concave satisfy the LSI, as do many non-log-concave distributions, such as Gaussian convolutions of distributions with bounded support (Chen et al., 2021). Next, we formally state our assumptions for this section, starting with the feature distribution.

Assumption C (Feature distribution).

1. Each feature vector $x_i \in \mathbb{R}^{p+1}$, for $i \in [n]$, decomposes as $x_i^\top = ((\Sigma^{1/2} z_i)^\top, 1)$, where $z_i \in \mathbb{R}^p$ has i.i.d. entries z_{ij} drawn from μ_z.
2. The distribution μ_z has mean 0, variance 1, and satisfies the T_2-inequality with constant σ_z.
3. There covariance matrix satisfies $\|\Sigma\|_{\mathrm{op}} \leq \sigma_\Sigma$ for a constant σ_Σ.

To be clear, in the above σ_z, σ_Σ are constants that are not allowed to change with n, p. It is worth emphasizing that we do not require the smallest eigenvalue of Σ in Assumption C to be bounded away from 0. This is possible because the iterates along the GD path are implicitly regularized. This is similar to not requiring a lower bound on the smallest eigenvalue for ridge regression when $\lambda > 0$ (as opposed to ridgeless regression, where we do need such an assumption); see Dobriban and Wager (2018); Patil et al. (2021). We also impose the following assumptions on the response distribution.

Assumption D (Response distribution).

1. Each $y_i = f(x_i) + \varepsilon_i$, for $i \in [n]$,[1] where ε_i is independent of x_i and drawn from μ_ε.
2. The distribution μ_ε has mean 0 and satisfies the T_2-inequality with constant σ_ε.
3. The regression function f is L_f-Lipschitz continuous, where without loss of generality, $L_f \leq 1$.

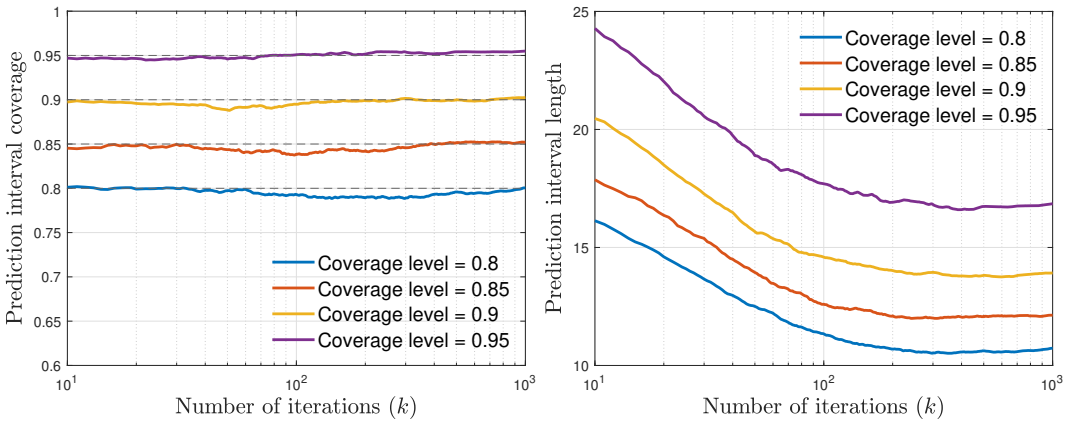

Figure 2: **LOOCV provides (asymptotically) valid prediction intervals, for various nominal coverage levels.** We investigate the empirical coverage and length of LOOCV prediction intervals along the GD path, at varying coverage levels. We consider an overparameterized regime with $n = 2500$ and $p = 5000$. The features are drawn from a Gaussian distribution with a covariance structure: $\Sigma_{ij} = \rho^{|i-j|}$ for all i, j and $\rho = 0.25$. The response is generated from a nonlinear model with heavy-tailed noise: t-distribution with 5 degrees of freedom. The linear component of $\mathbb{E}[y_i \,|\, x_i = x]$ is aligned with the top eigenvector of Σ. GD is run with a constant step size of 0.01. We can see that the prediction intervals generally have excellent finite-sample coverage along the entire path (*left*), and the smallest prediction length is typically obtained at a large iteration of GD (*right*).

4. Finally, $\mathbb{E}[y_i^8] \le m_8$, $\mathbb{E}[y_i^4] \le m_4$, and $\mathbb{E}[y_i^2] \le m_2$.

In the above $\sigma_\varepsilon, m_2, m_4, m_8$ are constants that are not allowed to change with n, p. We note that the assumptions we impose in this section are strictly weaker than those in Section 3. In particular, it is notable that we do not require $\mathbb{E}[y_i \,|\, x_i = x]$ to be linear in x. We are ready to give our first main positive result, on LOOCV for squared risk.

Theorem 2 (Squared risk consistency of LOOCV)**.** *Suppose that (x_i, y_i), $i \in [n]$ are i.i.d., and satisfy both Assumptions C and D. In addition, assume that there are constants $\Delta, B_0, \zeta_L, \zeta_U$ (independent of n, p) such that: (1) $\sum_{k=1}^{K} \delta_{k-1} \le \Delta$, (2) $\|\widehat{\beta}_0\|_2 \le B_0$, and (3) $0 < \zeta_L \le p/n \le \zeta_U < \infty$. Furthermore, let $K = o(n \cdot (\log n)^{-3/2})$. Then, as $n, p \to \infty$,*

$$\max_{k \in [K]} \left| \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k) - R(\widehat{\beta}_k) \right| \xrightarrow{\text{a.s.}} 0, \tag{8}$$

where we recall that $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$ and $R(\widehat{\beta}_k)$ are as defined in (4) and (3), respectively.

The convergence guarantee in Theorem 2 is strong in the sense that it is uniform across the entire GD path, and convergence occurs conditional on the training data. Uniformity in particular allows us to argue that tuning based on LOOCV guarantees asymptotically optimal risk. We cover this next, where we also generalize our study from squared error to general error functionals.

[1]Our result holds under a more general setting where $y_i = f(x_i, \varepsilon_i)$, with f being L_f-Lipschitz continuous. In the appendix, we provide the proof under this more general condition.

4.2 General Risk Functionals

We now extend our theory from the last subsection to cover general risk functionals, subject to only mild regularity conditions. Let $\psi \colon \mathbb{R}^2 \to \mathbb{R}$ be an error function, which takes as input the predictand (first argument) and prediction (second argument). We define a corresponding risk functional as:

$$\Psi(\widehat{\beta}_k) = \mathbb{E}_{x_0, y_0}\big[\psi(y_0, x_0^\top \widehat{\beta}_k) \mid X, y\big]. \tag{9}$$

One can naturally define an estimator for $\Psi(\widehat{\beta}_k)$ based on LOOCV using the "plug-in" principle:

$$\widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_k) = \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, x_i^\top \widehat{\beta}_{k,-i}). \tag{10}$$

Our second main positive result shows that this LOOCV plug-in estimator is uniformly consistent along the GD path.

Theorem 3 (Functional consistency of LOOCV)**.** *Under the conditions of Theorem 2, suppose that $\psi : \mathbb{R}^2 \to \mathbb{R}$ is differentiable and satisfies $\|\nabla \psi(u)\|_2 \le C_\psi \|u\|_2 + \bar{C}_\psi$ for all $u \in \mathbb{R}^2$ and for constants $C_\psi, \bar{C}_\psi \ge 0$. Then, as $n, p \to \infty$,*

$$\max_{k \in [K]} \left| \widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_k) - \Psi(\widehat{\beta}_k) \right| \xrightarrow{\text{a.s.}} 0. \tag{11}$$

where we recall that $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$ and $R(\widehat{\beta}_k)$ are as defined in (10) and (9), respectively.

As consequence of (11), LOOCV can be used to tune early stopping. Specifically, if we define $k_* = \arg\min_{k \in [K]} \widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_k)$, then as $n, p \to \infty$,

$$\left| \Psi(\widehat{\beta}_{k_*}) - \min_{k \in [K]} \Psi(\widehat{\beta}_k) \right| \xrightarrow{\text{a.s.}} 0. \tag{12}$$

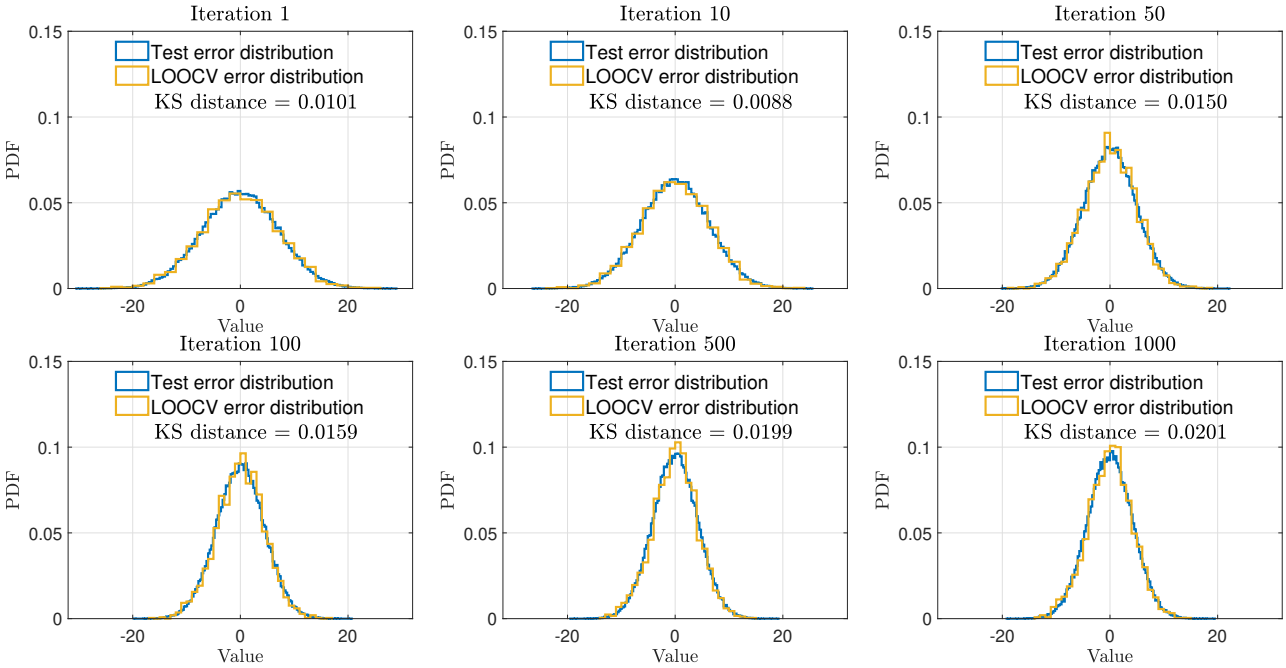

Figure 3: **Empirical distribution of LOOCV errors tracks the true test error distribution along the entire GD trajectory.** We consider the same setup as in Figure 2 with an overparameterized regime of $n = 2500$ and $p = 5000$. The blue curve in each panel represents a histogram of true prediction error errors (computed via Monte Carlo), while the yellow curve represents a histogram of the LOOCV errors. Each panel represents a given GD iteration, and we see strong agreement in the histograms throughout. Furthermore, due to the structure of the simulation setup, the test error distribution begins to exhibit lower variance as the iterations proceed.

Thanks to Theorem 3, we can consistently estimate the quantiles of the prediction error distribution using the empirical quantiles of the distribution that puts $1/n$ mass at each LOOCV residual.

Theorem 4 (Coverage guarantee)**.** *Under the conditions of Theorem 3, assume further that the distribution of the noise ε_i is continuous with density bounded by κ_{pdf}. Denote by $\widehat{\alpha}_k(q)$ the q-quantile of $\{y_i - x_i^\top \widehat{\beta}_{k,-i} : i \in [n]\}$. Then, for any quantile levels $0 \le q_1 \le q_2 \le 1$, letting $\mathcal{I}_k = [\widehat{\alpha}_k(q_1), \widehat{\alpha}_k(q_2)]$, we have as $n, p \to \infty$,*

$$\max_{k \in [K]} \mathbb{P}_{(x_0,y_0)}\big(y_0 - x_0^\top \widehat{\beta}_k \in \mathcal{I}_k \mid X, y\big) \xrightarrow{\text{a.s.}} q_2 - q_1. \quad (13)$$

Note that Theorem 4 provides *conditional* rather than *marginal* coverage guarantees for the specific data X, y that we observe. Figure 2 provides an example. Finally, we remark that the empirical distribution of the LOOCV errors can be shown to weakly converge to the true error distribution, almost surely. This is illustrated in Figure 3.

5 DISCUSSION

In the paper, we establish a significant discrepancy between LOOCV and GCV when it comes to estimating the prediction risk of early-stopped GD for least squares regression in high dimensions. While LOOCV is consistent in a strong uniform sense, GCV fails along essentially the entire path. This is especially curious considering that both LOOCV and GCV are uniformly consistent for the risk of explicitly regularized estimators such as ridge regression (Patil et al., 2021, 2022b). Therefore, this discrepancy also highlights a difference between GD and ridge regression, which is interesting in light of all of the existing work that establishes similarities between the two (Suggala et al., 2018; Neu and Rosasco, 2018; Ali et al., 2019).

Recall that GCV is generally tied to the "shortcut" formula for the leave-one-out (LOO) predictions in linear smoothers, where we adjust the training error for the i-th sample by $1 - \mathrm{tr}[S]/n$ (GCV), in place of $1 - S_{ii}/n$ (shortcut formula). A key part of the failure of GCV for GD is that its LOO predictions behave differently than those in ridge regression, as we discuss in what follows.

5.1 LOO Predictions in Ridge Regression versus Gradient Descent

For ridge regression, the LOO predictions, and hence LOOCV residuals, can be computed directly from the residuals of the full model (the model fit on the full data X, y) using a shortcut formula (Golub et al., 1979; Hastie, 2020). This is computationally important since

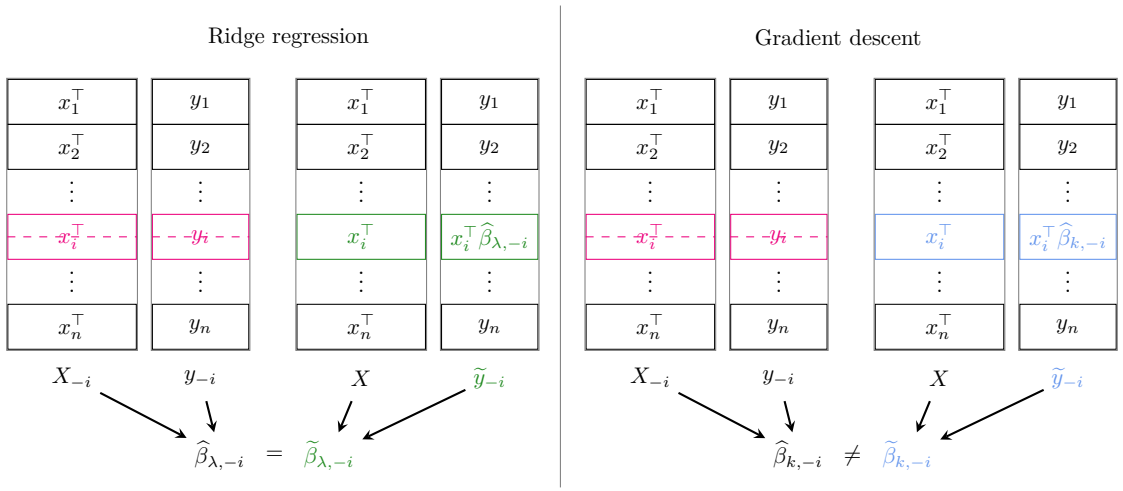

Figure 4: Illustrations of the differences between the LOO systems for ridge regression (*left*) and GD (*right*).

it means we can compute LOOCV without any refitting.

An elegant way to verify this shortcut formula involves creating an augmented system that allows us to identify the LOO prediction, which we briefly describe here. (We omit the intercept in the model, for simplicity.) For a given data point (x_i, y_i) that is to be left out, we seek to solve the problem:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|y_{-i} - X_{-i}\beta\|_2^2 + \lambda\|\beta\|_2^2. \tag{14}$$

Denoting its solution by $\widehat{\beta}_{\lambda,-i}$, the corresponding LOO prediction is therefore $x_i^\top \widehat{\beta}_{\lambda,-i}$. Let us now imagine that we "augment" the dataset X_{-i}, y_{-i} by adding the pair $(x_i, x_i^\top \widehat{\beta}_{\lambda,-i})$ in place of the i-th sample. Denote by $\widetilde{y}_{-i} \in \mathbb{R}^n$ the response vector in the augmented data set, and X the feature matrix (it is unchanged from the original data set). Denote by $\widetilde{\beta}_{\lambda,-i}$ the ridge estimator fit on the augmented data set $X, \widetilde{y}_{-i}$, which solves:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|\widetilde{y}_{-i} - X\beta\|_2^2 + \lambda\|\beta\|_2^2. \tag{15}$$

Problems (14) and (15) admit the same solution, because in the latter we have only added a single sample $x_i^\top \widehat{\beta}_{\lambda,-i}$ and this attains zero loss at the solution in the former. Thus, we have $\widetilde{\beta}_{\lambda,-i} = \widehat{\beta}_{\lambda,-i}$, and we can write the predicted value for the i-th observation as follows:

$$x_i^\top \widehat{\beta}_{\lambda,-i} = \sum_{j\neq i} [H_\lambda]_{ij} y_j + [H_\lambda]_{ii} (x_i^\top \widehat{\beta}_{\lambda,-i}),$$

where $H_\lambda \in \mathbb{R}^{n\times n}$ is the ridge smoothing matrix associated with full feature matrix X at regularization level λ. Rearranging, we have:

$$x_i^\top \widehat{\beta}_{\lambda,-i} = \frac{\sum_{j\neq i} [H_\lambda]_{ij} y_j}{1 - [H_\lambda]_{ii}},$$

or equivalently, in terms of residuals:

$$y_i - x_i^\top \widehat{\beta}_{\lambda,-i} = \frac{y_i - \sum_j [H_\lambda]_{ij} y_j}{1 - [H_\lambda]_{ii}} = \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [H_\lambda]_{ii}}. \tag{16}$$

Meanwhile, for the GD path, the analogous construction does not reproduce the LOO predictions. More precisely, let $\widehat{\beta}_{k,-i}$ be the GD iterate at step k, run on the LOO data set X_{-i}, y_{-i}. As before, imagine that we augment this data set with the pair $(x_i, x_i^\top \widehat{\beta}_{k,-i})$. Denote again by X the feature matrix and $\widetilde{y}_{-i} \in \mathbb{R}^n$ the response vector in the augmented data set, and denote by $\widetilde{\beta}_{k,-i}$ the result of running k iterations of GD on $X, \widetilde{y}_{-i}$. In general, we will have $\widehat{\beta}_{k,-i} \neq \widetilde{\beta}_{k,-i}$.

The underlying reason for this is that, even though the GD iterates can be written as a solution to a regularized least squares problem, the regularizer in this problem depends on the data (which is not true in ridge). For constant step sizes all equal to δ, the GD iterate (2) can be shown to solve:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \ \|y - X\beta\|_2^2/2n + \beta^\top Q_k \beta,$$

where $Q_k = X^\top X/n((I_p - \delta X^\top X/n)^k - I_p)^{-1}$. The regularization term is not only a function of δ and k, but also of X. This complicates the LOO predictions.

5.2 Modified Augmented System for LOO in GD

Identifying this failure of GD, as summarized in Figure 4, also helps us modify the augmentation trick so that we can recover the LOO predictions. Precisely, for $k \in [K]$ and $i, j \in [n]$, let

$$(\widetilde{y}_{k,-i})_j = \begin{cases} y_j, & j \neq i \\ x_i^\top \widehat{\beta}_{k,-i}, & j = i. \end{cases}$$

Define the vector $\widetilde{y}_{k,-i} = (\widetilde{y}_{k,-i})_{j\leq n}$, and let $\widetilde{\beta}_{k,-i}$ be the iterate obtained by running GD for k steps where at each step $\ell \leq k$, the augmented response vector $\widetilde{y}_{\ell,-i}$ is used in the gradient update. See Figure 5 for an illustration. Next, we show that this scheme recovers the LOO coefficients along the GD path.

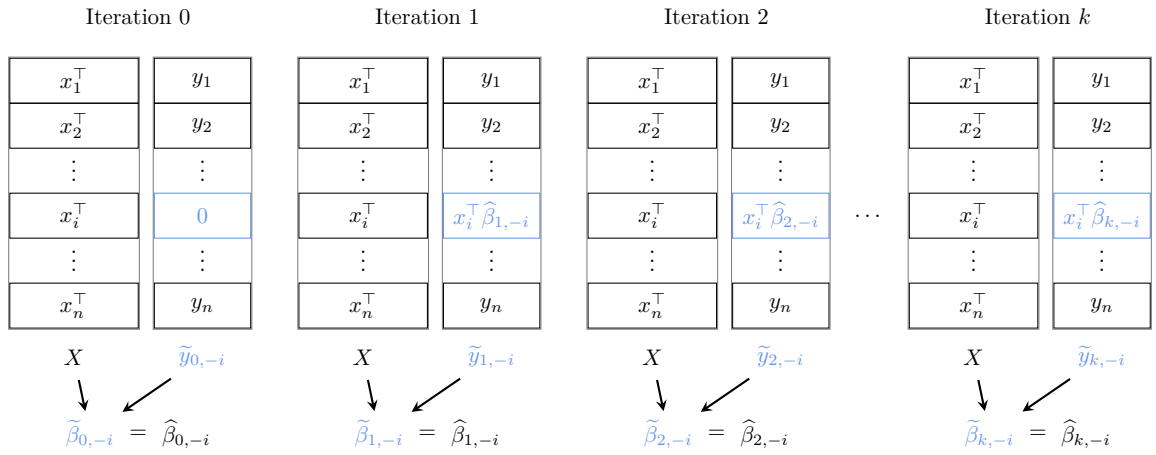

Figure 5: Illustration of the modified augmented system for LOO in GD.

Proposition 5 (Correctness of the modified augmented system)**.** *For all $k \in [K]$ and $i \in [n]$, it holds that $\widetilde{\beta}_{k,-i} = \widehat{\beta}_{k,-i}$.*

In other words, to recreate LOO coefficients from k-step GD, we must use an augmented response vector not only at step k but *at every iteration before* k as well. With this insight, we can represent the LOO predictions in a certain linear smoother form.

Proposition 6 (Smoother representation for the modified augmented system)**.** *For all $k \in [K]$ and $i \in [n]$, there is a vector $(h_{ij}^{(k)})_{j \le n}$ and scalar $b_i^{(k)}$ depending $\delta = (\delta_0, \dots \delta_{k-1})$ and X such that:*

$$x_i^\top \widehat{\beta}_{k,-i} = x_i^\top \widetilde{\beta}_{k,-i} = \sum_{j=1}^{n} h_{ij}^{(k)} y_j + b_i^{(k)}.$$

5.3 Towards Exact and Efficient LOOCV for GD?

We can unravel the relationships in LOO coefficients between iterations of the modified augmented system to arrive at explicit recursive forms for $(h_{ij}^{(k)})_{j \le n}$ and $b_i^{(k)}$ in Proposition 6, given next.

Proposition 7 (Recursive shortcut formula for LOO predictions in GD)**.** *For all $k \in [K]$ and $i \in [n]$,*

$$x_i^\top \widehat{\beta}_{k,-i} = x_i^\top \widehat{\beta}_k + A_{i,k} \|x_i\|_2^2 + \sum_{j=1}^{k-1} B_{i,k}^{(j)} x_i^\top (X^\top X)^j x_i,$$

where

$$A_{i,k+1} = A_{i,k} + \frac{2\delta_k A_{i,k} \|x_i\|_2^2}{n} + \sum_{j=1}^{k-1} \frac{2\delta_k B_{i,k}^{(j)} x_i^\top (X^\top X)^j x_i}{n} + \frac{2\delta_{k+1}(x_i^\top \widehat{\beta}_k - y_i)}{n},$$

$$B_{i,k+1}^{(1)} = B_{i,k}^{(1)} - \frac{2\delta_k A_{i,k}}{n},$$

$$B_{i,k+1}^{(j)} = B_{i,k}^{(j)} - \frac{2\delta_k B_{i,k}^{(j-1)}}{n}, \quad 2 \le j \le k,$$

and we make the convention that $B_{i,k}^{(k)} = 0$.

Using this proposition, we can estimate generic prediction risk functionals as follows. Abbreviating $\mathcal{H}_{ij} = x_i^\top (X^\top X)^j x_i$, to estimate the risk functional (9), we use:

$$\Psi^{\mathrm{loo}}(\widehat{\beta}_k) = \frac{1}{n} \sum_{i=1}^{n} \psi\Big(y_i,\, x_i^\top \widehat{\beta}_k + A_{i,k} \|x_i\|_2^2 + \sum_{j=1}^{k-1} B_{i,k}^{(j)} \mathcal{H}_{ij}\Big).$$

To be clear, this is an *exact* shortcut formula for (10).

In the $p \asymp n$ regime, the computational cost of a naive implementation of LOOCV for k-step GD is $O(n^3 k)$. (Each GD step costs $O(n^2)$, as we must compute p inner products, each of length n; then multiply this by k steps and n LOO predictions). In comparison, the shortcut formula given above can be shown to require $O(n^3 + n^2 k + n k^2)$ operations using a spectral decomposition of X. If k is large, say, itself proportional to n, then we can see that the shortcut formula is more efficient.

This is certainly not meant to be the final word on efficient LOOCV along the GD path. For one, a spectral decomposition is prohibitive for large problems (more expensive than solving the original least squares problem (1)), and there may be alternative perspectives on the shortcut formula given in Proposition 7 that lead to faster implementation. Further, if n is large enough, then stochastic variants of GD would be preferred in place of batch GD. All that said, the above analysis should be seen as a demonstration that exact shortcuts for LOO predictions in GD are *possible*, and may inspire others to develop more practical exact or approximate LOO methods.

Acknowledgements

We thank Alnur Ali, Arun Kumar Kuchibhotla, Arian Maleki, Alessandro Rinaldo, and Yuting Wei for enjoyable discussions related to this project, and for lending ears to parts of these results a while back. We also thank the anonymous reviewers for their constructive feedback, which has helped improve the clarity of the manuscript. The idea of providing proof blueprints is inspired by the `leanblueprint` plugin used in the Lean theorem prover. PP and RJT were supported by ONR grant N00014-20-1-2787.

References

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, 2020.

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Alnur Ali, Edgar Dobriban, and Ryan J. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, 2020.

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

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

Arnab Auddy, Haolin Zou, Kamiar Rahnama Rad, and Arian Maleki. Approximate leave-one-out cross validation for regression with ℓ_1 regularizers. *arXiv preprint arXiv:2310.17629*, 2023.

Morgane Austern and Wenda Zhou. Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*, 2020.

Benny Avelin and Lauri Viitasaari. Concentration inequalities for leave-one-out cross validation. *arXiv preprint arXiv:2211.02478*, 2022.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, pages 87–201, 2021.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

Mohsen Bayati, Murat A. Erdogdu, and Andrea Montanari. Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems*, 2013.

Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester Mackey. Cross-validation confidence intervals for test error. *arXiv preprint arXiv:2007.12671*, 2020.

Misha Belkin. Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, pages 203–248, 2021.

Pierre C. Bellec. Out-of-sample error estimate for robust M-estimators with convex penalty. *Information and Inference*, 12(4):2782–2817, 2023.

Pierre C. Bellec and Yiwei Shen. Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory*, 2022.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

Peter Bühlmann and Bin Yu. Boosting with the ℓ_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

Alain Celisse and Benjamin Guedj. Stability revisited: New generalisation bounds for the leave-one-out. *arXiv preprint arXiv:1608.06412*, 2016.

Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-Sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236, 2021.

Peter Craven and Grace Wahba. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics*, 46(1):247–279, 2018.

Jin-Hong Du, Pratik Patil, and Arun Kumar Kuchibhotla. Subsample ridge ensembles: Equivalences and generalized cross-validation. In *International Conference on Machine Learning*, 2023.

László Erdos and Horng-Tzer Yau. *A Dynamical Approach to Random Matrix Theory*. Courant Lecture Notes in Mathematics, 2017.

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

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Nathael Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *Annals of Probability*, 37(6):2480–2498, 2009.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, 2018a.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, 2018b.

Qiyang Han and Xiaocong Xu. The distribution of ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*, 2023.

Trevor Hastie. Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433, 2020.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. Second edition.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Maarten Jansen, Maurits Malfait, and Adhemar Bultheel. Generalized cross validation for wavelet thresholding. *Signal Processing*, 56(1):33–44, 1997.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, 2019.

Satyen Kale, Ravi Kumar, and Sergei Vassilvitskii. Cross-validation and mean-square stability. In *Proceedings of the Second Symposium on Innovations in Computer Science*, 2011.

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

Wolfram Koepf. Hypergeometric summation. *Vieweg, Braunschweig/Wiesbaden*, 5(6), 1998.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Springer, 2002.

Ravi Kumar, Daniel Lokshtanov, Sergei Vassilvitskii, and Andrea Vattani. Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, 2013.

Louis Landweber. An iteration formula for Fredholm integral equations of the first kind. *American Journal of Mathematics*, 73(3):615–624, 1951.

Jing Lei. Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532): 1978–1997, 2020.

Ker-Chau Li. From Stein's unbiased risk estimates to the method of generalized cross validation. *Annals of Statistics*, 13(4):1352–1377, 1985.

Ker-Chau Li. Asymptotic optimality of C_ℓ and generalized cross-validation in ridge regression with application to spline smoothing. *Annals of Statistics*, 14 (3):1101–1112, 1986.

Ker-Chau Li. Asymptotic optimality for C_p, C_ℓ, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15(3):958–975, 1987.

Yuetian Luo, Zhimei Ren, and Rina Foygel Barber. Iterative approximate cross-validation. *arXiv preprint arXiv:2303.02732*, 2023.

Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *Annals of Statistics*, 49 (4):2313–2335, 2021.

Nelson Morgan and Hervé Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. In *Advances in Neural Information Processing Systems*, 1989.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, 2018.

Pratik Patil and Daniel LeJeune. Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. In *International Conference on Learning Representations*, 2024.

Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan J. Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Pratik Patil, Arun Kumar Kuchibhotla, Yuting Wei, and Alessandro Rinaldo. Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*, 2022a.

Pratik Patil, Alessandro Rinaldo, and Ryan J. Tibshirani. Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, 2022b.

Kamiar Rahnama Rad and Arian Maleki. 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*, 82(4):965–996, 2020.

Kamiar Rahnama Rad, Wenda Zhou, and Arian Maleki. Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

William Stephenson and Tamara Broderick. Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, 36(2):111–133, 1974.

Mervyn Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35, 1977.

Otto Neall Strand. Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind. *SIAM Journal on Numerical Analysis*, 11(4):798–825, 1974.

Arun S. Suggala, Adarsh Prasad, and Pradeep Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, 2018.

Ramon Van Handel. Probability in high dimension. Technical report, Princeton University, 2014.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Shuaiwen Wang, Wenda Zhou, Haihao Lu, Arian Maleki, and Vahab Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. In *International Conference on Machine Learning*, 2018.

Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*, 2022.

Yuting Wei, Fanny Yang, and Martin J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. In *Advances in Neural Information Processing Systems*, 2017.

Ashia Wilson, Maximilian Kasy, and Lester Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, 2017.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, 2020.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.

Yongli Zhang and Yuhong Yang. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112, 2015.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   [Yes] Pointers below:
   - Setting and model: Sections 2.1 and 4.2.
   - Algorithms (estimators): Sections 2.2 and 5.3.
   - Assumptions: Assumptions A to D.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   [Yes] Pointers below:
   - Theoretical analyses: Theorems 1 to 4.
   - Complexities: Section 5.3.

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   [Yes] See the README file in the code folder in the attached supplementary materials.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   [Yes] Pointers below:
   - Theorem 1: Assumptions A and B.
   - Theorems 2 to 4: Assumptions C and D.

   (b) Complete proofs of all theoretical results.
   [Yes] Pointers below:
   - Theorem 1: Appendix S.3.
   - Theorems 2 and 3: Appendices S.6 and S.8.
   - Theorem 4: Appendix S.9.

   (c) Clear explanations of any assumptions.
   [Yes] Pointers below:
   - Assumptions A and B: Section 3.
   - Assumptions C and D: Section 4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   [Not applicable] For numerical illustrations in Figures 1 to 3, all the appropriate details are provided in the respective figure captions.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   [Yes] Pointers below:
   - Figures 1 and 2: Please see the caption.
   - Figure 3: The same setup as Figure 2.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   [Not applicable] All figures are for a single data realization (to illustrate almost sure statements for the given training dataset).

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   [Not applicable] No GPUs or clusters used.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets.
   [Not Applicable]

   (b) The license information of the assets, if applicable.
   [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable.
   [Not Applicable]

   (d) Information about consent from data providers/curators.
   [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
   [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   [Not Applicable]

Supplementary Materials to "Failures and Successes of Cross-Validation for Early-Stopped Gradient Descent"

Pratik Patil
University of California, Berkeley

Yuchen Wu
University of Pennsylvania

Ryan J. Tibshirani
University of California, Berkeley

This document serves as a supplement to the paper "Failures and Successes of Cross-Validation for Early-Stopped Gradient Descent." The structure of the supplement is outlined below, followed by a summary of the notation and conventions used in both the main paper and this supplement. The section and figure numbers in this supplement begin with the letter "S" and the equation numbers begin with the letter "E" to differentiate them from those appearing in the main paper.

Organization

- Appendix S.1 provides the main steps involved in the proofs of Theorem 1.

Section	Content	Purpose
Appendix S.1.1	Lemmas 8 and 9	Equivalences between gradient descent and flow for risk and GCV
Appendix S.1.2	Lemmas 10 and 11	Asymptotics of risk and GCV for gradient flow
Appendix S.1.3	Lemma 12	Mismatch of risk and GCV asymptotics for gradient flow

- Appendix S.2 contains supporting lemmas that are used in the proof of Theorem 1.

Section	Content	Purpose
Appendix S.2.1	Lemma 13	Closeness between gradient descent and flow
Appendix S.2.2	Lemmas 14 and 15	Statements of concentration results for linear and quadratic forms

- Appendix S.3 contains the proof of Theorem 1.

Section	Content	Purpose
Appendix S.3.1		Proof of Lemma 8
Appendix S.3.2		Proof of Lemma 9
Appendix S.3.3		Proof of Lemma 10
Appendix S.3.4		Proof of Lemma 11
Appendix S.3.5		Proof of Lemma 12
Appendix S.3.6		A helper lemma related to the Marchenko-Pastur law

- Appendix S.4 provides the main steps involved in the proofs of Theorem 2.

Section	Content	Purpose
Appendix S.4.1	Lemmas 22 and 23	Concentration of the LOOCV estimator
Appendix S.4.2	Lemma 24	Concentration of the risk
Appendix S.4.3	Lemmas 25 and 26	LOOCV bias analysis

- Appendix S.5 contains supporting lemmas that are used in the proofs of Theorems 2 to 4.

Section	Content	Purpose
Appendix S.5.1	Definition 2	Technical preliminaries
Appendix S.5.2	Proposition 27	Useful property of the T_2-inequality
Appendix S.5.3	Lemma 28	Dimension-free concentration inequality
Appendix S.5.4	Lemmas 29 and 30	Upper bounds on operator norm of $\widehat{\Sigma}$ and $\|y\|_2$
Appendix S.5.5	Lemmas 31 and 32	Upper bounds on $\|\mathbb{E}[y_0 x_0]\|$ and sub-exponential of $\|\widehat{\Sigma}\|_{\text{op}}$
Appendix S.5.6	Lemma 33 and Corollary 34	Upper bounds on $\|\widehat{\beta}_k\|_2$ and $\|\widehat{\beta}_{k,-i}\|_2$
Appendix S.5.7	Lemma 35	Upper bounds on LOOCV residuals $\{\|y_i - x_i^\top \widehat{\beta}_{k,-i}\|\}_{i\in[n]}$

- Appendix S.6 contains the proof of Theorem 2.

Section	Content	Purpose
Appendix S.6.1		Proof of Lemma 23
Appendix S.6.2		Proof of Lemma 24
Appendix S.6.3		Proof of Lemma 25
Appendix S.6.4		Proof of Lemma 26

- Appendix S.7 contains the proof of Lemma 22 that forms a key component in the proof of Theorem 2.

Section	Contents	Purpose
Appendix S.7.1	Lemmas 36 to 38	Upper bounding norm of the gradient with respect to the features
Appendix S.7.2	Lemma 39	Upper bounding norm of the gradient with respect to the response

- Appendix S.8 contains the proof of Theorem 3 for general risk functionals.

Section	Contents	Purpose
Appendix S.8.1	Lemma 40	Concentration analysis for LOOCV estimator and prediction risk
Appendix S.8.2	Lemma 41	Demonstrating that projection has little effect on quantities of interest

- Appendix S.9 contains the proof of Theorem 4. The proof uses the component Lemma 42.
- Appendix S.10 provides proofs of results related to the naive and modified augmentation systems (Propositions 5 and 7 and Proposition 6) for LOOCV along the gradient path in Section 5.

Section	Content	Purpose
Appendix S.10.1		Proof of Proposition 5
Appendix S.10.2		Proof of Proposition 6
Appendix S.10.3		Proof of Proposition 7

Conventions

- Throughout, C and C' (not to be confused with derivative) denote positive absolute constants.
- If no subscript is specified for the norm $\|x\|$ of a vector x, then it is assumed to be the ℓ_2 norm.
- If a proof of a statement is separated from the statement, the statement is restated (while keeping the original numbering) along with the proof for the reader's convenience.

S.1 Proof Sketch for Theorem 1

In this section, we outline the idea behind the proof of Theorem 1. The detailed proof can be found in Appendix S.3.

S.1.1 Step 1: Closeness between Gradient Descent and Gradient Flow

This step involves establishing equivalences between gradient descent and gradient flow, specifically for the downstream analysis of risk and generalized cross-validation.

Smoothers for gradient descent and flow. We start by rearranging the terms in (2) in the form of a first-order difference equation:

$$\frac{\widehat{\beta}_k - \widehat{\beta}_{k-1}}{\delta} = \frac{1}{n} X^\top (y - X\widehat{\beta}_{k-1}). \tag{E.1}$$

(Recall we are considering a fixed step size of δ and initialization at the origin $\widehat{\beta}_0 = 0$.) To consider a continuous time analog of (E.1), we imagine the interval $(0, t)$ is divided into k pieces each of size δ. Letting $\widehat{\beta}^{\mathrm{gf}}_t = \widehat{\beta}_k$ at time $t = k\delta$ and taking the limit $\delta \to 0$, we arrive at an ordinary differential equation:

$$\frac{\partial}{\partial t}\widehat{\beta}^{\mathrm{gf}}_t = \frac{1}{n} X^\top (y - X\widehat{\beta}^{\mathrm{gf}}_t), \tag{E.2}$$

with the initial condition $\widehat{\beta}^{\mathrm{gf}}_0 = 0$. We refer to (E.2) as the gradient flow differential equation. The gradient flow (GF) estimate has a closed-form solution:

$$\widehat{\beta}^{\mathrm{gf}}_t = \widehat{\Sigma}^\dagger \big(I_p - \exp(-t\widehat{\Sigma})\big) \cdot \frac{1}{n} X^\top y, \tag{E.3}$$

where $\widehat{\Sigma}^\dagger$ stands for the Moore-Penrose generalized inverse of $\widehat{\Sigma}$, which we recall is $X^\top X/n$. Also, recall from Section 2.2 that by rolling out the iterations, the gradient descent iterate at step k can be expressed as:

$$\widehat{\beta}_k = \sum_{j=0}^{k-1} \delta \big(I_p - \delta\widehat{\Sigma}\big)^{k-j-1} \cdot \frac{1}{n} X^\top y. \tag{E.4}$$

We can define the corresponding GCV estimates for the squared risk as follows:

$$\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k) = \frac{1}{n} \frac{\|y - X\widehat{\beta}_k\|_2^2}{(1 - \mathrm{tr}(H_k)/n)^2} \quad \text{and} \quad \widehat{R}^{\mathrm{gcv}}(\widehat{\beta}^{\mathrm{gf}}_t) = \frac{1}{n} \frac{\|y - X\widehat{\beta}^{\mathrm{gf}}_t\|_2^2}{(1 - \mathrm{tr}(H^{\mathrm{gf}}_t)/n)^2},$$

where

$$H_k = \sum_{j=0}^{k-1} \frac{\delta}{n} X \big(I_p - \delta\widehat{\Sigma}\big)^{k-j-1} X^\top \quad \text{and} \quad H^{\mathrm{gf}}_t = \frac{1}{n} X (\widehat{\Sigma})^\dagger \big(I_p - \exp(-t\widehat{\Sigma})\big) X^\top. \tag{E.5}$$

We first show that under the conditions of Theorem 1, estimates obtained from GD are in some sense asymptotically equivalent to that obtained from gradient flow (GF), which we define below.

Lemma 8 (Prediction risks are asymptotically equivalent)**.** *Under the assumptions of Theorem 1, we have*

$$|R(\widehat{\beta}_k) - R(\widehat{\beta}^{\mathrm{gf}}_T)| \xrightarrow{\text{a.s.}} 0.$$

Lemma 9 (GCV risk estimates are asymptotically equivalent)**.** *Under the assumptions of Theorem 1, we have*

$$\big|\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k) - \widehat{R}^{\mathrm{gcv}}(\widehat{\beta}^{\mathrm{gf}}_T)\big| \xrightarrow{\text{a.s.}} 0.$$

The proofs of these equivalences in Lemmas 8 and 9 are provided in Appendices S.3.1 and S.3.2, respectively.

S.1.2 Step 2: Limiting Risk and GCV

This step focuses on obtaining asymptotics (limiting behaviors) for risk and GCV when using gradient flow.

According to Lemmas 8 and 9, to show that the GCV estimator is inconsistent for the GD risk, it suffices to show that it is inconsistent for the GF risk. We next separately derive the limiting expressions for $R(\widehat{\beta}_T^{\mathrm{gf}})$ and $\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_T^{\mathrm{gf}})$, respectively.

Let $F_{\zeta_*}(s)$ denote the Marchenko-Pastur law:

- *Underparameterized.* For $\zeta_* \le 1$, the density is given by:

$$\frac{\mathrm{d}F_{\zeta_*}(s)}{\mathrm{d}s} = \frac{1}{2\pi\zeta_* s}\sqrt{(b-s)(s-a)}\cdot \mathbb{1}_{[a,b]}(s). \tag{E.6}$$

  The density is supported on $[a,b]$, where $a=(1-\sqrt{\zeta_*})^2$ and $b=(1+\sqrt{\zeta_*})^2$.

- *Overparameterized.* For $\zeta_* > 1$, the law F_{ζ_*} has an additional point mass at 0 of probability $1-1/\zeta_*$. In other words,

$$\frac{\mathrm{d}F_{\zeta_*}(s)}{\mathrm{d}s} = \left(1-\frac{1}{\zeta_*}\right)\delta_0(s) + \frac{1}{2\pi\zeta_* s}\sqrt{(b-s)(s-a)}\cdot \mathbb{1}_{[a,b]}(s). \tag{E.7}$$

  Here, δ_0 is the Dirac delta function at 0.

Lemma 10 (Risk limit for gradient flow)**.** *Under the assumptions of Theorem 1,*

$$R(\widehat{\beta}_T^{\mathrm{gf}}) \xrightarrow{\text{a.s.}} r^2\int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) + \zeta_*\sigma^2\int z^{-1}(1-\exp(-Tz))^2\,\mathrm{d}F_{\zeta_*}(z) + \sigma^2.$$

Lemma 11 (GCV limit for gradient flow)**.** *Under the assumptions of Theorem 1,*

$$\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k) \xrightarrow{\text{a.s.}} \frac{r^2\int z\exp(-2Tz)\mathrm{d}F_{\zeta_*}(z) + \sigma^2(1-\zeta_*) + \sigma^2\zeta_*\int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)}{\left(1-\zeta_*\int(1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2}.$$

The proofs of these asymptotic limits in Lemmas 10 and 11 are provided in Appendices S.3.3 and S.3.4, respectively.

S.1.3 Step 3: Limits Mismatch

The final step involves showing a mismatch between the asymptotics of risk and GCV for gradient flow.

Lemma 12 (Limits mismatch)**.** *Let F_{ζ_*} be the Marchenko-Pastur law. Then, assuming either $r^2>0$ or $\sigma^2>0$, for all $T>0$, except for a set of Lebesgue measure zero,*

$$\begin{aligned} &r^2\int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) + \zeta_*\sigma^2\int z^{-1}(1-\exp(-Tz))^2\,\mathrm{d}F_{\zeta_*}(z) + \sigma^2 \\ &\neq \frac{r^2\int z\exp(-2Tz)\mathrm{d}F_{\zeta_*}(z) + \sigma^2(1-\zeta_*)_+ + \sigma^2\zeta_*\int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)}{\left(1-\zeta_*\int(1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2}. \end{aligned} \tag{E.8}$$

The proof of this asymptotic mismatch in Lemma 12 is provided in Appendix S.3.5.

S.2 Supporting Lemmas for the Proof of Theorem 1

S.2.1 Connections between Gradient Descent and Gradient Flow

We first show that under the conditions of Theorem 1, estimates obtained from gradient descent (GD) are in some sense asymptotically equivalent to that obtained from gradient flow (GF).

We next establish connections between GD and GF. This step is achieved by showing that the hat matrices as defined in Equation (E.5) when $k\to\infty$ and $k\delta\to T$ (for H_k) and when $t=T$ (for H_t) get closer under the matrix operator norm.

Observe that the two matrices in Equation (E.5) share a common set of eigenvectors, and the eigenvalues are obtained by applying separate scalar transformations to the eigenvalues of $\widehat{\Sigma}$. Hence, to show that H_k and H_T^{gf} are close in terms of operator norm, a natural first step is to show that the scalar transformations are uniformly close to each other. We characterize such closeness in Lemma 13 below.

Let $g_{\delta,k}(x) = \sum_{j=0}^{k-1} \delta x(1-\delta x)^{k-j-1}$ and $g_T(x) = 1 - \exp(-tx)$, which are exactly the scalar transformations of the hat matrices in Equation (E.5). Our next lemma says that as $k \to \infty$ and $\delta \to 0$ with $k\delta \to T$, $g_{\delta,k}$ uniformly approximates g_T on a compact interval.

Lemma 13 (Scalar uniform approximation for GD and GF smoothing functions)**.** *We assume $k \to \infty$, $\delta \to 0$, and $k\delta \to T$. Here, T is a fixed positive constant. Then it holds that*

$$\sup_{0 \le x \le \zeta_* + 2\sqrt{\zeta_*} + 2} \left| g_{\delta,k}(x) - g_T(x) \right| \to 0,$$

where we recall that ζ_* is the limit of the aspect ratio.

Proof. For notational simplicity, we let $J_{\zeta_*} = [\,0, \zeta_* + 2\sqrt{\zeta_*} + 2\,]$. We will first show that

$$\sup_{x \in J_{\zeta_*}} \left| k \log(1-\delta x) + k\delta x \right| \to 0. \tag{E.9}$$

To this end, we consider the first-order derivatives of the function inside the above absolute value sign with respect to x, which gives $-\delta k/(1-\delta x) + k\delta$. This quantity under the current conditions goes to zero uniformly for all $x \in J_{\zeta_*}$, thus proving Equation (E.9). This further implies the following uniform convergence result:

$$\sup_{x \in J_{\zeta_*}, j+1 \in [k]} \left| (k-j-1)\log(1-\delta x) + (k-j-1)\delta \right| \to 0.$$

As a direct consequence of the above equation, we obtain

$$\sup_{x \in J_{\zeta_*}, j+1 \in [k]} \left| (1-\delta x)^{k-j-1} - \exp(-\delta(k-j-1)x) \right| \to 0,$$

which further gives

$$\sup_{x \in J_{\zeta_*}} \left| \sum_{j=0}^{k-1} \delta x (1-\delta x)^{k-j-1} - \sum_{j=0}^{k-1} \delta x \exp(-\delta(k-j-1)x) \right| \to 0$$

as $\sum_{j=0}^{k-1} \delta x$ is uniformly upper bounded for all $x \in J_{\zeta_*}$.

Considering the derivative of an exponential function, it is not hard to see that

$$\sup_{j+1 \in [k]} \sup_{(k-j-1)\delta \le z \le (k-j)\delta} \left| \exp(-\delta(k-j-1)x) - \exp(-zx) \right| \to 0.$$

Therefore,

$$\sup_{x \in J_{\zeta_*}} \left| \sum_{j=0}^{k-1} \delta x \exp(-\delta(k-j-1)x) - \int_0^{k\delta} x \exp(-zx)\,\mathrm{d}z \right| \to 0.$$

Further, we note that

$$\sup_{x \in J_{\zeta_*}} \left| \int_0^{k\delta} x \exp(-zx)\,\mathrm{d}z - \int_0^{T} x \exp(-zx)\,\mathrm{d}z \right| \to 0$$

and

$$\int_0^T x \exp(-zx)\,\mathrm{d}z = 1 - \exp(-Tx).$$

This completes the proof. □

We can apply Lemma 13 to establish several useful connections between GD and GF, which we state as Lemmas 8 and 9. The proof of these two lemmas can be found in Appendices S.3.2 and S.3.1, respectively.

S.2.2 Useful Concentration Results

The following lemma provides the concentration of a linear form of a random vector with independent components. It follows from a moment bound from Lemma 7.8 of Erdos and Yau (2017), along with the Borel-Cantelli lemma and is adapted from Lemma S.8.5 of Patil et al. (2022a).

Lemma 14 (Concentration of linear form with independent components). *Let $z_p \in \mathbb{R}^p$ be a sequence of random vector with i.i.d. entries z_{pi}, $i = 1, \ldots, p$ such that for each i, $\mathbb{E}[z_{pi}] = 0$, $\mathbb{E}[z_{pi}^2] = 1$, $\mathbb{E}[|z_{pi}|^{4+\alpha}] \leq M_\alpha$ for some $\alpha > 0$ and constant $M_\alpha < \infty$. Let $a_p \in \mathbb{R}^p$ be a sequence of random vectors independent of z_p such that $\limsup_p \|a_p\|_2^2/p \leq M_0$ almost surely for a constant $M_0 < \infty$. Then $a_p^\top z_p/p \to 0$ almost surely as $p \to \infty$.*

The following lemma provides the concentration of a quadratic form of a random vector with independent components. It follows from a moment bound from Lemma B.26 of Bai and Silverstein (2010), along with the Borel-Cantelli lemma and is adapted from Lemma S.8.6 of Patil et al. (2022a).

Lemma 15 (Concentration of quadratic form with independent components). *Let $z_p \in \mathbb{R}^p$ be a sequence of random vector with i.i.d. entries z_{pi}, $i = 1, \ldots, p$ such that for each i, $\mathbb{E}[z_{pi}] = 0$, $\mathbb{E}[z_{pi}^2] = 1$, $\mathbb{E}[|z_{pi}|^{4+\alpha}] \leq M_\alpha$ for some $\alpha > 0$ and constant $M_\alpha < \infty$. Let $D_p \in \mathbb{R}^{p \times p}$ be a sequence of random matrix such that $\limsup \|D_p\|_{\text{op}} \leq M_0$ almost surely as $p \to \infty$ for some constant $M_0 < \infty$. Then $z_p^\top D_p z_p/p - \text{tr}[D_p]/p \to 0$ almost surely as $p \to \infty$.*

S.3 Proof of Theorem 1

Theorem 1 (Inconsistency of GCV). *Suppose that (x_i, y_i), $i \in [n]$ are i.i.d., and satisfy both Assumptions A and B, where either $r^2 > 0$ or $\sigma^2 > 0$. As $n, p \to \infty$, assume $p/n \to \zeta_*$, and $k \to \infty$, $\delta \to 0$ such that $k\delta \to T$, where $T, \zeta_* > 0$ are constants. Then, for every fixed $\zeta_* > 0$, it holds that for almost all $T > 0$ (i.e., all $T > 0$ except for a set of Lebesgue measure zero),*

$$\left| \widehat{R}^{\text{gcv}}(\widehat{\beta}_k) - R(\widehat{\beta}_k) \right| \not\xrightarrow{\text{p}} 0, \tag{6}$$

where we recall that $\widehat{R}^{\text{gcv}}(\widehat{\beta}_k)$ and $R(\widehat{\beta}_k)$ are as defined in (5) and (3), respectively.

S.3.1 Proof of Lemma 8

Lemma 8 (Prediction risks are asymptotically equivalent). *Under the assumptions of Theorem 1, we have*

$$|R(\widehat{\beta}_k) - R(\widehat{\beta}_T^{\text{gf}})| \xrightarrow{\text{a.s.}} 0.$$

Proof. Note that the prediction risks admit the following expressions:

$$R(\widehat{\beta}_k) = \|\beta_0 - \widehat{\beta}_k\|_2^2 + \sigma^2 \quad \text{and} \quad R(\widehat{\beta}_T^{\text{gf}}) = \|\beta_0 - \widehat{\beta}_T^{\text{gf}}\|_2^2 + \sigma^2.$$

We define $\bar{g}_{\delta,k}(x) = \sum_{j=0}^{k-1} \delta(1-\delta x)^{k-j-1}$ and $\bar{g}_T(x) = x^{-1}(1 - \exp(-Tx))$. We claim that

$$\|x^{1/2}(\bar{g}_{\delta,k}(x) - \bar{g}_T(x))\mathbb{1}_{x \in J_{\zeta_*}}\|_\infty \to 0 \tag{E.10}$$

under the asymptotics $k \to \infty$, $\delta \to 0$, and $k\delta \to T$. Proof for this claim is similar to that for Lemma 13, and we skip it for the compactness of presentation.

We note that

$$\widehat{\beta}_k - \widehat{\beta}_T^{\text{gf}} = \frac{1}{\sqrt{n}} V^\top \Big(\bar{g}_{\delta,k}(\Lambda^\top \Lambda) - \bar{g}_T(\Lambda^\top \Lambda) \Big) \Lambda^\top U^\top y, \tag{E.11}$$

where we recall that $X/\sqrt{n} = V \Lambda U$ is the spectral decomposition. It is straightforward to obtain the following upper bound:

$$\left\| \Big(\bar{g}_{\delta,k}(\Lambda^\top \Lambda) - \bar{g}_T(\Lambda^\top \Lambda) \Big) \Lambda^\top \right\|_{\text{op}} \leq \sup_{i \in [n]} \left| \lambda_i^{1/2} (\bar{g}_{\delta,k}(\lambda_i) - \bar{g}_T(\lambda_i)) \right|.$$

Recall that $\max_{i \in [n]} \lambda_i \xrightarrow{\text{a.s.}} (1 + \sqrt{\zeta_*})^2$, hence the right-hand side of the above equation converges to zero almost surely (using Equation (E.10)). By the law of large numbers, we obtain $\|y\|_2/\sqrt{n} \xrightarrow{\text{a.s.}} \mathbb{E}[y_1^2]^{1/2}$. Plugging these

results into Equation (E.11) gives $\|\widehat{\beta}_k - \widehat{\beta}_T^{\mathrm{gf}}\|_2 \xrightarrow{\text{a.s.}} 0$ as $n, p \to \infty$. Furthermore, by Equations (E.3) and (E.4), we have

$$\begin{aligned} \big\|\widehat{\beta}_T^{\mathrm{gf}}\big\|_2 &\le \max_{i\in[n]} \lambda_i^{1/2} \cdot \bar{g}_T\Big(\max_{i\in[n]} \lambda_i\Big) \cdot \frac{1}{\sqrt{n}} \|y\|_2, \\ \|\widehat{\beta}_k\|_2 &\le \max_{i\in[n]} \lambda_i^{1/2} \cdot \bar{g}_{\delta,k}\Big(\max_{i\in[n]} \lambda_i\Big) \cdot \frac{1}{\sqrt{n}} \|y\|_2. \end{aligned} \tag{E.12}$$

Standard analysis implies that $\sup_{x\in J_{\zeta_*}} \sqrt{x}\bar{g}_T(x) < \infty$ and $\limsup_{k\to\infty,\delta\to 0} \sup_{x\in J_{\zeta_*}} \sqrt{x}\bar{g}_{\delta,k}(x) < \infty$.

Finally, combining all these results we have obtained, we conclude that

$$\Big| \|\beta_0 - \widehat{\beta}_k\|_2^2 - \|\beta_0 - \widehat{\beta}_T^{\mathrm{gf}}\|_2^2 \Big| \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$. This is equivalent to saying

$$|R(\widehat{\beta}_k) - R(\widehat{\beta}_T^{\mathrm{gf}})| \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$. □

S.3.2 Proof of Lemma 9

Lemma 9 (GCV risk estimates are asymptotically equivalent)**.** *Under the assumptions of Theorem 1, we have*

$$\big|\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k) - \widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_T^{\mathrm{gf}})\big| \xrightarrow{\text{a.s.}} 0.$$

Proof. In the sequel, we will apply Lemma 13 to prove closeness between $\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k)$ and $\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_T^{\mathrm{gf}})$. This consists of proving the following three pairs of quantities are close:

(1) $(1 - \mathrm{tr}(H_k)/n)^{-2}$ and $(1 - \mathrm{tr}(H_T^{\mathrm{gf}})/n)^{-2}$.

(2) $\widehat{\beta}_k^\top \widehat{\Sigma} \widehat{\beta}_k$ and $(\widehat{\beta}_T^{\mathrm{gf}})^\top \widehat{\Sigma} \widehat{\beta}_T^{\mathrm{gf}}$.

(3) $y^\top X \widehat{\beta}_k / n$ and $y^\top X \widehat{\beta}_T^{\mathrm{gf}} / n$.

In what follows, we shall separately justify each of these closeness results.

Closeness result (1)

We denote by $\{\lambda_i\}_{i\le n}$ the top n eigenvalues of $\widehat{\Sigma}$. From Bai and Silverstein (2010, Theorem 5.8), we know that $\max_{i\in[n]} \lambda_i \xrightarrow{\text{a.s.}} (1 + \sqrt{\zeta_*})^2$. Note that

$$\frac{1}{n}\mathrm{tr}(H_k) = \frac{1}{n}\sum_{i=1}^n g_{\delta,k}(\lambda_i) \quad \text{and} \quad \frac{1}{n}\mathrm{tr}(H_T^{\mathrm{gf}}) = \frac{1}{n}\sum_{i=1}^n g_T(\lambda_i).$$

Invoking Lemma 13, we obtain that with probability one

$$\limsup_{n,p\to\infty} \frac{1}{n}\Big|\mathrm{tr}(H_k) - \mathrm{tr}(H_T^{\mathrm{gf}})\Big| \le \sup_{0\le x\le \zeta_* + 2\sqrt{\zeta_*}+2} \Big| g_{\delta,k}(x) - g_T(x)\Big|,$$

which vanishes as $n, p \to \infty$. As a result, we derive that $|\mathrm{tr}(H_k) - \mathrm{tr}(H_T^{\mathrm{gf}})|/n \xrightarrow{\text{a.s.}} 0$ as $n, p \to \infty$.

Let $F_{\zeta_*}(s)$ denote the Marchenko-Pasture law as defined in (E.6) and (E.6). Standard results in random matrix theory (Bai and Silverstein, 2010) tell us that the empirical spectral distribution of $\widehat{\Sigma}$ almost surely converges in distribution to F_{ζ_*}. Note that g_T is a bounded continuous function on $[0, \zeta_* + 2\sqrt{\zeta_*} + 2]$, thus

$$\frac{1}{n}\sum_{i=1}^n g_T(\lambda_i) \xrightarrow{\text{a.s.}} \int \big(1 - \exp(-Tz)\big)\, \mathrm{d}F_{\zeta_*}(z),$$

which one can verify is strictly smaller than 1 for all $\zeta_* \in (0, \infty)$. Putting together the above analysis, we can deduce that both $(1 - \mathrm{tr}(H_k)/n)^{-2}$ and $(1 - \mathrm{tr}(H_T^{\mathrm{gf}})/n)^{-2}$ converge almost surely to one finite constant, hence concluding the proof for this part.

Closeness result (2)
We denote by $X/\sqrt{n} = U\Lambda V$ the singular value decomposition of $X/\sqrt{n}$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices. Combining Equations (E.3) and (E.4), we arrive at the following equation:

$$\widehat{\beta}_k^\top \widehat{\Sigma}\widehat{\beta}_k - (\widehat{\beta}_T^{\mathrm{gf}})^\top \widehat{\Sigma}\widehat{\beta}_T^{\mathrm{gf}} = y^\top U^\top \cdot \left\{ g_{\delta,k}(\Lambda\Lambda^\top)^2 - g_T(\Lambda\Lambda^\top)^2 \right\} \cdot Uy/n. \tag{E.13}$$

By the strong law of large numbers, we have $\|y\|_2^2/n \xrightarrow{\text{a.s.}} \mathbb{E}[y_1^2]$. By Lemma 13 and the fact that $\max_{i\in[n]} \lambda_i \xrightarrow{\text{a.s.}} (1+\sqrt{\zeta_*})^2$, we conclude that

$$\left\| g_{\delta,k}(\Lambda\Lambda^\top)^2 - g_T(\Lambda\Lambda^\top)^2 \right\|_{\mathrm{op}} \xrightarrow{\text{a.s.}} 0.$$

Plugging these arguments into Equation (E.13), we obtain

$$\left| \widehat{\beta}_k^\top \widehat{\Sigma}\widehat{\beta}_k - (\widehat{\beta}_T^{\mathrm{gf}})^\top \widehat{\Sigma}\widehat{\beta}_T^{\mathrm{gf}} \right| \xrightarrow{\text{a.s.}} 0,$$

which concludes the proof of closeness result (2).

Closeness result (3)
Finally, we show one more closeness result (3). We note that

$$\frac{1}{n}\left(y^\top X\widehat{\beta}_k - y^\top X\widehat{\beta}_T^{\mathrm{gf}}\right) = y^\top U^\top \cdot \left\{ g_{\delta,k}(\Lambda\Lambda^\top) - g_T(\Lambda\Lambda^\top) \right\} \cdot Uy/n,$$

which by the same argument as that we used to derive result (2) almost surely converges to zero as $n, p \to \infty$.

Putting together (1), (2), and (3), we conclude the proof of the lemma. □

S.3.3 Proof of Lemma 10

Lemma 10 (Risk limit for gradient flow)**.** *Under the assumptions of Theorem 1,*

$$R(\widehat{\beta}_T^{\mathrm{gf}}) \xrightarrow{\text{a.s.}} r^2 \int \exp(-2Tz)\, \mathrm{d}F_{\zeta_*}(z) + \zeta_*\sigma^2 \int z^{-1}(1-\exp(-Tz))^2\, \mathrm{d}F_{\zeta_*}(z) + \sigma^2.$$

Proof. Applying Equation (E.3) and the risk decomposition formula, we obtain

$$\begin{aligned} R(\widehat{\beta}_T^{\mathrm{gf}}) =& \beta_0^\top \exp(-2T\widehat{\Sigma})\beta_0 - \frac{2}{n}\beta_0^\top \exp(-T\widehat{\Sigma})\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\varepsilon \\ &+ \frac{1}{n^2}\varepsilon^\top X(I_p - \exp(-T\widehat{\Sigma}))(\widehat{\Sigma}^\dagger)^2(I_p - \exp(-T\widehat{\Sigma}))X^\top\varepsilon + \sigma^2. \end{aligned}$$

Note that

$$\frac{2}{\sqrt{n}}\left\|\beta_0^\top \exp(-T\widehat{\Sigma})\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\right\|_2 \le 2\|\beta_0\|_2 \cdot \sup_{i\in[n]} \frac{\exp(-T\lambda_i)(1-\exp(-T\lambda_i))}{\lambda_i^{1/2}},$$

where it is understood that $\lambda^{-1/2}e^{-T\lambda}(1-e^{-T\lambda})\,|_{\lambda=0} = 0$. Recall that $\max_i \lambda_i \xrightarrow{\text{a.s.}} (1+\sqrt{\zeta_*})^2$ and $\|\beta_0\|_2^2 \to r^2$. Hence, there exists a constant M_0 such that almost surely

$$\limsup_{n,p\to\infty} \|\beta_0^\top \exp(-T\widehat{\Sigma})\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\|_2^2/n \le M_0.$$

Therefore, we can apply Lemma 14 and deduce that

$$\frac{2}{n}\beta_0^\top \exp(-T\widehat{\Sigma})\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\varepsilon \xrightarrow{\text{a.s.}} 0. \tag{E.14}$$

By Lemma 15, we have

$$\left| n^{-2}\varepsilon^\top X(I_p - \exp(-T\widehat{\Sigma}))(\widehat{\Sigma}^\dagger)^2(I_p - \exp(-T\widehat{\Sigma}))X^\top\varepsilon \right.$$

$$-n^{-2}\sigma^2 \operatorname{tr}(X(I_p - \exp(-T\widehat{\Sigma}))(\widehat{\Sigma}^\dagger)^2(I_p - \exp(-T\widehat{\Sigma}))X^\top)\Big| \xrightarrow{\text{a.s.}} 0.$$

Standard random matrix theory result implies that almost surely the empirical spectral distribution of $\widehat{\Sigma}$ converges in distribution to F_{ζ_*}, which is the Marchenko-Pastur law defined in (E.6) and (E.7). Furthermore, $\|\widehat{\Sigma}\|_{\text{op}} \xrightarrow{\text{a.s.}} (1+\sqrt{\zeta_*})^2$. Therefore, we conclude that

$$\begin{aligned} &n^{-2}\sigma^2 \operatorname{tr}(X(I_p - \exp(-T\widehat{\Sigma}))(\widehat{\Sigma}^\dagger)^2(I_p - \exp(-T\widehat{\Sigma}))X^\top) \\ &\xrightarrow{\text{a.s.}} \zeta_*\sigma^2 \int z^{-1}(1-\exp(-Tz))^2 \,\mathrm{d}F_{\zeta_*}(z). \end{aligned} \tag{E.15}$$

Finally, we study the limit of $\beta_0^\top \exp(-2T\widehat{\Sigma})\beta_0$. Let $\Omega \in \mathbb{R}^{p\times p}$ be a uniformly distributed orthogonal matrix that is independent of anything else. Since by assumption $\|\beta_0\|_2 \to r$, we can then couple $\Omega\beta_0$ with $g \sim \mathcal{N}(0, I_p)$, so that (1) g is independent of $\widehat{\Sigma}$, and (2) $\|\Omega\beta_0 - rg/\sqrt{p}\|_2 \xrightarrow{\text{a.s.}} 0$. Note that all eigenvalues of $\exp(-2T\widehat{\Sigma})$ are between 0 and 1, hence

$$\left|\beta_0^\top \exp(-2T\widehat{\Sigma})\beta_0 - \frac{r^2}{p} g^\top \exp(-2T\widehat{\Sigma})g\right| \xrightarrow{\text{a.s.}} 0.$$

Leveraging Lemma 15, we obtain

$$r^2 g^\top \exp(-2T\widehat{\Sigma})g/p \xrightarrow{\text{a.s.}} r^2 \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z).$$

Combining this with (E.14) and (E.15), we finish the proof. □

S.3.4 Proof of Lemma 11

Lemma 11 (GCV limit for gradient flow)**.** *Under the assumptions of Theorem 1,*

$$\widehat{R}^{\text{gcv}}(\widehat{\beta}_k) \xrightarrow{\text{a.s.}} \frac{r^2 \int z\exp(-2Tz)\mathrm{d}F_{\zeta_*}(z) + \sigma^2(1-\zeta_*) + \sigma^2\zeta_* \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)}{\left(1-\zeta_*\int(1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2}.$$

Proof. We separately discuss the numerator and the denominator. We start with the denominator. Recall that the empirical spectral distribution of $\widehat{\Sigma}$ almost surely converges to F_{ζ_*} and $\|\widehat{\Sigma}\|_{\text{op}} \xrightarrow{\text{a.s.}} (1+\sqrt{\zeta_*})^2$. Hence,

$$(1-\operatorname{tr}(H_T^{\text{gf}})/n)^{-2} \xrightarrow{\text{a.s.}} \left(1-\zeta_*\int(1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^{-2}. \tag{E.16}$$

Next, we consider the numerator. Straightforward computation implies that

$$\begin{aligned} \frac{1}{n}\|y - X\widehat{\beta}_T^{\text{gf}}\|_2^2 =& \beta_0^\top \exp(-T\widehat{\Sigma})\widehat{\Sigma}\exp(-T\widehat{\Sigma})\beta_0 + \frac{1}{n}\left\|\left(I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\right)\varepsilon\right\|_2^2 \\ &+ \frac{2}{n}\left\langle \beta_0, \exp(-T\widehat{\Sigma})X^\top\left(I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\right)\varepsilon\right\rangle. \end{aligned}$$

Since $\|\widehat{\Sigma}\|_{\text{op}} \xrightarrow{\text{a.s.}} (1+\sqrt{\zeta_*})^2$, we then obtain almost surely

$$\limsup_{n,p\to\infty} \left\|\exp(-T\widehat{\Sigma})X^\top\left(I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\right)\right\|_{\text{op}} \le G(\zeta_*) < \infty,$$

where $G(\zeta_*)$ is a function of ζ_*. Therefore, by Lemma 14, we obtain

$$\frac{2}{n}\left\langle \beta_0, \exp(-T\widehat{\Sigma})X^\top\left(I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\right)\varepsilon\right\rangle \xrightarrow{\text{a.s.}} 0. \tag{E.17}$$

Using the same argument that we used to compute the limiting expression of $\beta_0^\top \exp(-T\widehat{\Sigma})\beta_0$, we conclude that

$$\beta_0^\top \exp(-T\widehat{\Sigma})\widehat{\Sigma}\exp(-T\widehat{\Sigma})\beta_0 \xrightarrow{\text{a.s.}} r^2 \int z\exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z). \tag{E.18}$$

In addition, by Lemma 15, we have

$$\frac{1}{n}\big\|\big(I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top\big)\varepsilon\big\|_2^2 \xrightarrow{\text{a.s.}} \sigma^2(1-\zeta_*) + \sigma^2\zeta_* \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z). \tag{E.19}$$

To see the limit in (E.19), we expand the matrix of the quadratic form as follows:

$$\begin{aligned}
&(I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top)(I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top) \\
&= (I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top) - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma})X^\top)(I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top) \\
&= (I_n - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))X^\top) - \frac{1}{n}X\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma}))(I_p - \widehat{\Sigma}\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma})))X^\top.
\end{aligned}$$

The normalized (by n) trace of the matrix above is

$$\begin{aligned}
&1 - \zeta_*\,\mathrm{tr}[(I_p - \exp(-T\widehat{\Sigma}))]/p - \zeta_*\,\mathrm{tr}[(I_p - \exp(-T\widehat{\Sigma}))\exp(-T\widehat{\Sigma})]/p \\
&= 1 - \zeta_* + \zeta_*\,\mathrm{tr}[\exp(-2T\widehat{\Sigma})]/p.
\end{aligned}$$

In the above simplification, we used the fact that

$$\widehat{\Sigma}\widehat{\Sigma}^\dagger(I_p - \exp(-T\widehat{\Sigma})) = (I_p - \exp(-T\widehat{\Sigma})).$$

This fact follows because $\widehat{\Sigma}^\dagger\widehat{\Sigma}$ is the projection onto the row space of X. But the image of $I_p - \exp(-t\widehat{\Sigma})$ is already in the row space. The limit for (E.19) therefore is

$$\sigma^2(1-\zeta_*) + \sigma^2\zeta_* \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z).$$

We can do quick sanity checks for this limit:

- When $T = 0$, we should get σ^2 irrespective of ζ_* because we start with a null model.
- When $T = \infty$, we should get the training error of the least squares or ridgeless estimator due to noise. There are two cases:
  - When $\zeta_* < 1$: this is the variance component of the residual of least squares. This should be $\sigma^2(1-\zeta_*)$.
  - When $\zeta_* > 1$: this is the variance component of the training error of the ridgeless interpolator, which should be zero.

To check the last point, it is worth noting that

$$\lim_{T\to\infty} \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) = \begin{cases} 0 & \zeta_* < 1 \\ 1 - \dfrac{1}{\zeta_*} & \zeta_* > 1. \end{cases}$$

Now, Equations (E.16) to (E.19) together imply the stated result. □

S.3.5 Proof of Lemma 12

Lemma 12 (Limits mismatch)**.** *Let F_{ζ_*} be the Marchenko-Pastur law. Then, assuming either $r^2 > 0$ or $\sigma^2 > 0$, for all $T > 0$, except for a set of Lebesgue measure zero,*

$$\begin{aligned}
&r^2 \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) + \zeta_*\sigma^2 \int z^{-1}(1-\exp(-Tz))^2\,\mathrm{d}F_{\zeta_*}(z) + \sigma^2 \\
&\neq \frac{r^2 \int z\exp(-2Tz)\mathrm{d}F_{\zeta_*}(z) + \sigma^2(1-\zeta_*)_+ + \sigma^2\zeta_* \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)}{\left(1 - \zeta_* \int (1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2}.
\end{aligned} \tag{E.8}$$

Proof. Recall the asymptotics of the risk from Lemma 10:

$$R(\widehat{\beta}_T^{\mathrm{gf}}) \tag{E.20}$$

$$\xrightarrow{\text{a.s.}} r^2 \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) + \zeta_*\sigma^2 \int z^{-1}(1-\exp(-Tz))^2\,\mathrm{d}F_{\zeta_*}(z) + \sigma^2$$

$$= r^2 \left\{ \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) \right\} + \sigma^2 \left\{ 1 + \zeta_* \int z^{-1}(1-\exp(-Tz))^2\,\mathrm{d}F_{\zeta_*}(z) \right\}. \tag{E.21}$$

Recall also the asymptotics of GCV from Lemma 11:

$$\widehat{R}^{\mathrm{gcv}}(\widehat{\beta}_k) \tag{E.22}$$

$$\xrightarrow{\text{a.s.}} \frac{r^2 \int z\exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) + \sigma^2(1-\zeta_*) + \sigma^2\zeta_* \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)}{\left(1-\zeta_* \int (1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2}$$

$$= r^2 \frac{\int z\exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)}{\left(1-\zeta_* \int (1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2} + \sigma^2 \frac{(1-\zeta_*) + \zeta_* \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)}{\left(1-\zeta_* \int (1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2}. \tag{E.23}$$

Here, F_{ζ_*} is the Marchenko-Pasture law, as defined in (E.6) and (E.7). Observe that both functions (E.21) and (E.23) are analytic (i.e., they can be represented by a convergent power series in a neighborhood of every point in their domain). From the identity theorem for analytic functions (see, e.g., Chapter 1 of Krantz and Parks (2002)), it suffices to show that the functions do not agree in a neighborhood of a point inside the domain. We will do this in the neighborhood of $t = 0$. The function value and the derivatives match, but the second derivatives mismatch. This is shown in Appendices S.3.5.2 and S.3.5.3. This supplies us with the desired function disagreement and concludes the proof. □

A remark on the proof of Lemma 12: Observe that both the risk and the GCV asymptotics in (E.21) and (E.23) split into bias or bias-like and variance or variance-like components, respectively. The bias or bias-like component is scaled by the signal energy, and the variance or variance-like component is scaled by the noise energy. We can also show that except for a set of Lebesgue measure 0, we have

$$\int \exp(-2Ts)\,\mathrm{d}F_{\zeta_*}(s) \neq \frac{\int s\exp(-2Ts)\,\mathrm{d}F_{\zeta_*}(s)}{\left(1-\zeta_* \int (1-\exp(-Ts))\,\mathrm{d}F_{\zeta_*}(s)\right)^2}, \tag{E.24}$$

$$1 + \zeta_* \int \frac{(1-\exp(-Ts))^2}{s}\,\mathrm{d}F_{\zeta_*}(s) \neq \frac{(1-\zeta_*) + \zeta_* \int \exp(-2Ts)\,\mathrm{d}F_{\zeta_*}(s)}{\left(1-\zeta_* \int (1-\exp(-Ts))\,\mathrm{d}F_{\zeta_*}(s)\right)^2}. \tag{E.25}$$

In the following, we will refer to (E.24) as the signal component mismatch and (E.25) as the noise component mismatch. The functions on both sides of (E.24) and (E.25) are again analytic in T. The mismatch of the second derivatives for the sum above in fact is a consequence of mismatches for the individual signal and noise component. This is shown in Appendices S.3.5.2 and S.3.5.3.

S.3.5.1 Combined Sum Mismatch

Our goal is to show that the two limiting functions (of T) in (E.21) and (E.23) differ on a neighborhood of $T = 0$. Since the common denominator in the two terms in Equation (E.23) are away from 0, it suffices to show that in the neighborhood around $T = 0$, the following function is not identically zero:

$$\mathcal{D}(T) = r^2 \left\{ \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) \left(1-\zeta_* \int (1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2 - \int z\exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z) \right\}$$

$$+\sigma^2\left\{\left(1+\zeta_*\int z^{-1}(1-\exp(-Tz))^2\,\mathrm{d}F_{\zeta_*}(z)\right)\left(1-\zeta_*\int(1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^2\right.$$
$$\left.-(1-\zeta_*)-\zeta_*\int\exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)\right\}. \tag{E.26}$$

As argued in the proof of Lemma 12, the function $\mathcal{D}$ is analytic and it suffices to examine the Taylor coefficients. Both $\mathcal{D}(0)$ and $\mathcal{D}'(0)$ are 0 but it turns out that $\mathcal{D}''(0)\neq 0$ for $\zeta_*>0$. Thus, our subsequent goal will be to compute $\mathcal{D}''(T)$ and evaluate it at $T=0$. We will make use of double derivative calculations in Appendices S.3.5.2 and S.3.5.3 for this purpose, as summarized below.

Claim 16 (Second derivatives mismatch for combined sum)**.** For the function $\mathcal{D}$ as defined in (E.26), we have $\mathcal{D}''(T)=-2\zeta_*(2r^2+\sigma^2)$. Thus, when $\zeta_*>0$ and either $r^2>0$ or $\sigma^2>0$, we have $\mathcal{D}''(0)\neq 0$.

Proof. The calculation follows from Claims 17 and 19. Specifically, using the notation defined in these claims, we have

$$\mathcal{D}''(T)=r^2(\mathcal{B}_\ell''(T)-\mathcal{B}_r''(T))-\sigma^2(\mathcal{V}_\ell''(T)-\mathcal{V}_r''(T)). \tag{E.27}$$

Evaluating (E.27) at $T=0$ yields

$$\mathcal{D}''(0)=r^2(4+12\zeta_*+4\zeta_*^2)-r^2(4+14\zeta_*+4\zeta_*^2)-\sigma^2(4\zeta_*^2)+\sigma^2(4\zeta_*+4\zeta_*^2). \tag{E.28}$$

Simplifying (E.28), we obtain the desired conclusion. □

S.3.5.2 Signal Component Mismatch

Claim 17 (Second derivatives mismatch for signal component)**.** Let F_{ζ_*} be the Marchenko-Pasture law as defined in (E.6) and (E.7). Let $\mathcal{B}_\ell$ and $\mathcal{B}_r$ be two functions defined as follows:

$$\mathcal{B}_\ell(T)=\left(1-\zeta_*\int(1-\exp(-Ts))\,\mathrm{d}F_{\zeta_*}(s)\right)^2\int\exp(-2Ts)\,\mathrm{d}F_{\zeta_*}(s),$$
$$\mathcal{B}_r(T)=\int s\exp(-2Ts)\,\mathrm{d}F_{\zeta_*}(s).$$

We have $\mathcal{B}_\ell''(0)=4+12\zeta_*+4\zeta_*^2$ and $\mathcal{B}_r''(0)=4+14\zeta_*+4\zeta_*^2$, and hence $B_\ell''(0)\neq B_r''(0)$.

Proof. For ease of notation, define the functions w, v, and u as follows:

$$w(T)=\left(1-\zeta_*\int(1-\exp(-Ts))\,\mathrm{d}F_{\zeta_*}(s)\right)^2,$$
$$v(T)=\int\exp(-2Ts)\,\mathrm{d}F_{\zeta_*}(s),$$
$$u(T)=\int s\exp(-2Ts)\,\mathrm{d}F_{\zeta_*}(s).$$

Then we have $\mathcal{B}_\ell(T)=w(T)v(T)$ and $\mathcal{B}_r(T)=u(T)$. The first-order derivatives are $\mathcal{B}_\ell'(T)=w'(T)v(T)+w(T)v'(T)$ and $\mathcal{B}_r'(T)=u'(T)$. The second-order derivatives are $\mathcal{B}_\ell''(T)=w''(T)+2w'(T)v'(T)+v''(T)$ and $\mathcal{B}_r''(T)=u''(T)$. From Claim 18, we obtain

$$\mathcal{B}_\ell''(0)=2\zeta_*(1+2\zeta_*)+8\zeta_*+4(1+\zeta_*)=4+14\zeta_*+4\zeta_*^2.$$

On the other hand, from Claim 18 again, we have

$$\mathcal{B}_r'(0)=4(1+3\zeta_*+\zeta_*^2)=4+12\zeta_*+4\zeta_*^2.$$

Thus, for any $\zeta_*>0$, we have that $\mathcal{B}_\ell''(0)\neq\mathcal{B}_r''(0)$, as desired. □

Claim 18 (Second derivatives of various parts signal component)**.** Let w, v, and u be functions defined in the proof of Claim 17. Then the following claims hold.

- $w(0) = 1$, $w'(0) = -2\zeta_*$, and $w''(0) = 2\zeta_*(1 + 2\zeta_*)$.
- $v(0) = 1$, $v'(0) = -2$, and $v''(0) = 4(1 + \zeta_*)$.
- $u(0) = 1$, $u'(0) = -2(1 + \zeta_*)$, and $u''(0) = 4(1 + 3\zeta_* + \zeta_*^2)$.

Proof. The functional evaluations are straightforward. We will split the first- and second-order derivative calculations into separate parts below. For $k \geq 0$, let $M_k = \int s^k \, \mathrm{d}F_{\zeta_*}(s)$ be the k-th moment of the Marchenko-Pastur law.

Part 1. Denote the inner integral by

$$I(T) = \zeta_* \int (1 - \exp(-Ts)) \, \mathrm{d}F_{\zeta_*}(s).$$

Then, $w(T) = (1 - I(T))^2$. The first derivative of $w(T)$ is

$$w'(T) = -2(1 - I(T)) \cdot I'(T) \quad \text{with} \quad I'(T) = \zeta_* \int s \exp(-Ts) \, \mathrm{d}F_{\zeta_*}(s).$$

The second derivative of $w(T)$ is

$$w''(T) = 2(I'(T))^2 - 2(1 - I(T)) \cdot I''(T) \quad \text{with} \quad I''(T) = -\zeta_* \int s^2 \exp(-Ts) \, \mathrm{d}F_{\zeta_*}(s).$$

From (E.29), note that $I(0) = 0$, $I'(0) = \zeta_* M_1 = \zeta_*$, and $I''(0) = -\zeta_* M_2 = -\zeta_* - \zeta_*^2$. Thus, we have $w'(0) = -2\zeta_* M_1 = -2\zeta_*$ and $w''(0) = 2\zeta_*^2 + 2\zeta_* + 2\zeta_*^2 = 2\zeta_* + 4\zeta_*^2$.

Part 2. For $v(T)$, the derivatives are straightforward. The first derivative is

$$v'(T) = -2 \int s \exp(-2Ts) \, \mathrm{d}F_{\zeta_*}(s).$$

The second derivative is

$$v''(T) = 4 \int s^2 \exp(-2Ts) \, \mathrm{d}F_{\zeta_*}(s).$$

Hence, from (E.29), we then get $v'(0) = -2M_1 = -2$ and $v''(0) = 4M_2 = 4 + 4\zeta_*$.

Part 3. For $u(T)$, the derivatives are similarly straightforward. The first derivative is

$$u'(T) = -2 \int s^2 \exp(-2Ts) \, \mathrm{d}F_{\zeta_*}(s).$$

The second derivative is

$$u''(T) = 4 \int s^3 \exp(-2Ts) \, \mathrm{d}F_{\zeta_*}(s).$$

From Equation (E.29) again, we obtain that $u'(0) = -2M_2 = -2(1 + \zeta_*)$ and $u''(0) = 4M_3 = 4(1 + 3\zeta_* + \zeta_*)$. □

S.3.5.3 Noise Component Mismatch

Claim 19 (Second derivatives mismatch for noise component)**.** Let F_{ζ_*} be the Marchenko-Pasture law as defined in (E.6) and (E.7). Let $\mathcal{V}_\ell$ and $\mathcal{V}_r$ be two functions defined as follows:

$$\mathcal{V}_\ell(T) = \left(1 + \zeta_* \int \frac{(1 - \exp(-Ts))^2}{s} \, \mathrm{d}F_{\zeta_*}(s)\right) \left(1 - \zeta_* \int (1 - \exp(-Ts)) \, \mathrm{d}F_{\zeta_*}(s)\right)^2$$

$$\mathcal{V}_r(T) = (1 - \zeta_*) + \zeta_* \int \exp(-2Ts) \, \mathrm{d}F_{\zeta_*}(s).$$

We have $\mathcal{V}_\ell''(0) = 4\zeta_*^2$ and $\mathcal{V}_r''(0) = 4\zeta_* + 4\zeta_*^2$, and hence $\mathcal{V}_\ell''(0) \neq \mathcal{V}_r''(0)$ for $\zeta_* > 0$.

Proof. For ease of notation, define the functions w, $\widetilde{v}$, and $\widetilde{u}$ such that

$$w(T) = \left(1 - \zeta_* \int (1 - \exp(-Ts))\, \mathrm{d}F_{\zeta_*}(s)\right)^2,$$
$$\widetilde{v}(T) = 1 + \zeta_* \int \frac{(1 - \exp(-Ts))^2}{s}\, \mathrm{d}F_{\zeta_*}(s),$$
$$\widetilde{u}(T) = (1 - \zeta_*) + \zeta_* \int \exp(-2Ts)\, \mathrm{d}F_{\zeta_*}(s).$$

(Note that the function w is the same function as defined in Claim 18.) Then we have $\mathcal{V}_\ell(T) = w(T)\widetilde{v}(T)$ and $\mathcal{V}_r(T) = \widetilde{u}(T)$. The first-order derivatives are $\mathcal{V}_\ell'(T) = w'(T)\widetilde{v}(T) + w(T)\widetilde{v}'(T)$ and $\mathcal{V}_r'(T) = \widetilde{u}'(T)$. The second-order derivatives are $\mathcal{V}_\ell''(T) = w''(T) + 2w'(T)\widetilde{v}'(T) + \widetilde{v}''(T)$ and $\mathcal{V}_r''(T) = \widetilde{u}''(T)$. From Claim 20, we obtain

$$\mathcal{V}_\ell''(0) = 2\zeta_*(1 + 2\zeta_*) - 2\zeta_* = 4\zeta_*^2.$$

On the other hand, from Claim 20 again, we have

$$\mathcal{V}_r''(0) = 4\zeta_* + 4\zeta_*^2.$$

Thus, we have that $\mathcal{V}_\ell''(0) \neq \mathcal{V}_r''(0)$ for any $\zeta_* > 0$. This concludes the proof. □

Claim 20 (Second derivatives of various parts of noise component)**.** Let w, $\widetilde{v}$, and $\widetilde{u}$ be functions defined in the proof of Claim 19. Then the following claims hold.

- $w(0) = 1$, $w'(0) = -2\zeta_*$, and $w''(0) = 2\zeta_*(1 + 2\zeta_*)$.
- $\widetilde{v}(0) = 1$, $\widetilde{v}'(0) = 0$, and $\widetilde{v}''(0) = -2\zeta_*$.
- $\widetilde{u}(0) = 1$, $\widetilde{u}'(0) = -2$, and $\widetilde{u}''(0) = 4(1 + \zeta_*)$.

Proof. The functional evaluations are straightforward. We will split the first- and second-order derivative calculations into separate parts below. Recall that for $k \geq 0$, we denote by $M_k = \int s^k\, \mathrm{d}F_{\zeta_*}(s)$, the k-th moment of the Marchenko-Pastur law.

Part 1. This part is the same as Part 1 of Claim 18.

Part 2. We start by computing the derivative of the integrand.

$$\frac{\partial}{\partial T}\left(\frac{(1 - \exp(-Ts))^2}{s}\right) = 2(1 - \exp(-Ts)) \cdot (-\exp(-Ts)) \cdot (-s) \cdot \frac{1}{s} = 2(1 - \exp(-Ts))\exp(-Ts)$$

For the second derivative, note that

$$\frac{\partial^2}{\partial T^2}\left(\frac{(1 - \exp(-Ts))^2}{s}\right) = \frac{\partial}{\partial T}(2(1 - \exp(-Ts))\exp(-Ts)) = -2s\exp(-Ts) + 4s\exp(-2Ts).$$

Therefore, we have

$$\widetilde{v}'(T) = 2\zeta_* \int (1 - \exp(-Ts))\exp(-Ts)\, \mathrm{d}F_{\zeta_*}(s)$$

and

$$\widetilde{v}''(T) = 2\zeta_* \int s\exp(-Ts)(1 - 2\exp(-Ts))\, \mathrm{d}F_{\zeta_*}(s).$$

Thus, $\widetilde{v}'(0) = 0$ and $\widetilde{v}''(0) = -2\zeta_* M_1 = -2\zeta_*$.

Part 3. For $\widetilde{u}(T)$, the derivatives are straightforward. The first derivative is

$$\widetilde{u}'(T) = -2\zeta_* \int s\exp(-2Ts)\, \mathrm{d}F_{\zeta_*}(s).$$

The second derivative is

$$\widetilde{u}''(T) = 4\zeta_* \int s^2\exp(-2Ts)\, \mathrm{d}F_{\zeta_*}(s).$$

From Equation (E.29) again, we obtain that $\widetilde{u}'(0) = -2\zeta_* M_1 = -2\zeta_*$ and $\widetilde{u}''(0) = 4\zeta_* M_2 = 4\zeta_*(1 + \zeta_*)$. □

S.3.6 A Helper Lemma related to the Marchenko-Pastur Law

Lemma 21 (Moments of the Marchenko-Pasture distribution). *Let F_{ζ_*} be the Marchenko-Pasture law as defined in (E.6) and (E.7). For $k \geq 1$, we have*

$$\int s^k \, \mathrm{d}F_{\zeta_*}(s) = \sum_{i=0}^{k-1} \frac{1}{i+1} \binom{k}{i} \binom{k-1}{i} \zeta_*^i.$$

The explicit moment formula in Lemma 21 is well-known. See, for example, Lemma 3.1 of Bai and Silverstein (2010). It can be derived using the Chu-Vandermonde identity, also known as Vandermonde's convolutional formula for binomial coefficients (Koepf, 1998, Chapter 3).

We will use Lemma 21 to obtain the following moments explicitly. Let M_k denote the k-th moment $\int s^k \, \mathrm{d}F_{\zeta_*}(s)$ of the Marchenko-Pastur distribution. We have

$$M_0 = 1, \quad M_1 = 1, \quad M_2 = 1 + \zeta_*, \quad M_3 = 1 + 3\zeta_* + \zeta_*^2, \quad M_4 = 1 + 6\zeta_* + 6\zeta_*^2 + \zeta_*^3. \tag{E.29}$$

S.4 Proof Sketch for Theorem 2

In this section, we outline the proof idea of Theorem 2. The extension to general test functionals can be found in Appendix S.8. We will prove both the theorems for a general starting estimator $\widehat{\beta}_0$.

S.4.1 Step 1: LOO Concentration

The most challenging part of our proof is establishing concentration for $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$. This is achieved by upper bounding the norm of the gradient of the mapping $(w_1, \cdots, w_n) \mapsto \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$, where $w_i = (x_i, y_i)$. Although this mapping is not exactly Lipschitz, it is approximately Lipschitz in the sense that its gradient is bounded on a set that occurs with high probability.

For $k \in \{0\} \cup [K]$, we define $f_k : \mathbb{R}^{n(p+2)} \mapsto \mathbb{R}$ as $f_k(w_1, \cdots, w_n) = \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$. Our goal is to upper bound $\|\nabla f_k\|_2$. It will become clear that f_k is Lipschitz continuous on a closed convex set Ω. We define Ω as follows:

$$\Omega = \left\{ \|\widehat{\Sigma}\|_{\mathrm{op}} \leq C_{\Sigma,\zeta}, \ \|y\|_2^2 \leq n(m + \log n) \right\}, \tag{E.30}$$

where $C_{\Sigma,\zeta} = 2C_0\sigma_\Sigma(1+\zeta)+1$, $m = m_2$, and $C_0 > 0$ is a numerical constant. It can be verified that Ω is a convex set of the data. Standard concentration results (see Lemma 29 and Lemma 30) imply that with an appropriately selected C_0, we have $\mathbb{P}(\Omega) \geq 1 - 2(n+p)^{-4} - n^{-1}m_4 \log^{-2} n$. In other words, for large (n, p), the input samples will fall inside Ω with high probability.

In the following, we establish the Lipschitz continuity of f_k when restricted to Ω, which is a closed convex set. This can be equivalently stated as the Lipschitz continuity of the composition of the projection onto Ω and f_k. To prove this, we upper bound the Euclidean norm of the gradient, as detailed in Lemma 22. The proof of Lemma 22 can be found in Appendix S.7.

Lemma 22 (Gradient upper bound). *There exists a constant $\xi(C_{\Sigma,\zeta}, \Delta, m, B_0) > 0$ that depends only on $(C_{\Sigma,\zeta}, \Delta, m, B_0)$, such that on the set Ω, it holds that*

$$\|\nabla_W f_k(W)\|_F \leq \frac{K\xi(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \log n}{\sqrt{n}}$$

for all $k \in \{0\} \cup [K]$. In the above display, we define $W = (w_1, \cdots, w_n)$ and K, we recall, is the total number of GD iterations.

We define $h : \mathbb{R}^{n(p+2)} \mapsto \mathbb{R}^{n(p+2)}$ as the projection that projects its inputs onto Ω. Define $\widetilde{f}_k = f_k \circ h$. Lemma 22 implies that $\widetilde{f}_k$ is a Lipschitz continuous mapping with a Lipschitz constant as stated in Lemma 22. By assumption, the input data distribution satisfies a T_2-inequality, allowing us to apply a powerful concentration inequality stated in Proposition 27 to obtain the desired concentration result. We state this result as Lemma 23 below, and its proof can be found in Appendix S.6.1.

Lemma 23 (LOO concentration). *We assume the assumptions of Theorem 2. Then with probability at least* $1-2(n+p)^{-4}-(n\log^2 n)^{-1}m_4-2(K+1)C_{\mathsf{T}_2}n^{-2}$, *it holds that for all* $k\in\{0\}\cup[K]$

$$\left|\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)-\mathbb{E}[\widetilde{f}_k(w_1,\cdots,w_n)]\right|\le\frac{2\sigma_{\mathsf{T}_2}LK\xi(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot(\log n)^{3/2}}{\sqrt{n}},$$

where we let $L=(L_f^2\sigma_\Sigma+L_f^2+\sigma_\Sigma)^{1/2}$, $\sigma_{\mathsf{T}_2}^2=\sigma_z^2\vee\sigma_\varepsilon^2$, *and* C_{T_2} *is a positive numerical constant that appears in Proposition 27.*

S.4.2 Step 2: Risk Concentration

In the second part, we provide concentration bounds for the prediction risk $R(\widehat{\beta}_k)$. We follow a similar approach as in Step 1, establishing that $R(\widehat{\beta}_k)$ is a Lipschitz function of the input data with high probability. Leveraging the assumption of a T_2-inequality in the data distribution, we apply Proposition 27 to derive a concentration result. The proof of this result is presented in Appendix S.6.2. We state the concentration result as Lemma 24.

Lemma 24 (Risk concentration). *We write* $R(\widehat{\beta}_k)=r_k(w_1,\cdots,w_n)$ *and define* $\widetilde{r}_k(w_1,\cdots,w_n)=r_k(h(w_1,\cdots,w_n))$. *Then under the assumptions of Theorem 2, with probability at least* $1-2(n+p)^{-4}-(n\log^2 n)^{-1}m_4-2(K+1)C_{\mathsf{T}_2}n^{-2}$, *for all* $k\in\{0\}\cup[K]$ *we have*

$$\left|R(\widehat{\beta}_k)-\mathbb{E}[\widetilde{r}_k(w_1,\cdots,w_n)]\right|\le\frac{2\sigma_{\mathsf{T}_2}L\bar{\xi}(C_{\Sigma,\zeta},\Delta,m,B_0)(\log n)^{3/2}}{\sqrt{n}},$$

where $\bar{\xi}(C_{\Sigma,\zeta},\Delta,m,B_0)>0$ *depends uniquely on* $(C_{\Sigma,\zeta},\Delta,m,B_0)$.

S.4.3 Step 3: LOO Bias Analysis

In Steps 1 and 2, we have proven concentration results for both $R(\widehat{\beta}_k)$ and $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$. Specifically, we have shown that $R(\widehat{\beta}_k)$ concentrates around $\mathbb{E}[\widetilde{r}_k(w_1,\cdots,w_n)]$ and $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$ concentrates around $\mathbb{E}[\widetilde{f}_k(w_1,\cdots,w_n)]$. These expectations represent the target functionals composed with the projection h.

Next, we demonstrate that incorporating the projection h into the expectation does not significantly alter the quantities of interest. This result is presented as Lemma 25 below.

Lemma 25 (Projection effects). *Under the assumptions of Theorem 2, it holds that*

$$\begin{aligned}\sup_{k\in\{0\}\cup[K]}|\mathbb{E}[\widetilde{r}_k(w_1,\cdots,w_n)]-\mathbb{E}[r_k(w_1,\cdots,w_n)]|&=o_n(1),\\ \sup_{k\in\{0\}\cup[K]}\left|\mathbb{E}[\widetilde{f}_k(w_1,\cdots,w_n)]-\mathbb{E}[f_k(w_1,\cdots,w_n)]\right|&=o_n(1).\end{aligned}\tag{E.31}$$

Finally, we aim to establish a result showing that the prediction risk is stable with respect to the sample size. Specifically, we seek to demonstrate that $\mathbb{E}[R(\widehat{\beta}_k)]$ is approximately equal to $\mathbb{E}[R(\widehat{\beta}_{k,-1})]$, which is equivalent to $\mathbb{E}[r_k(w_1,\cdots,w_n)]\approx\mathbb{E}[f_k(w_1,\cdots,w_n)]$.

Formally speaking, we prove the following lemma.

Lemma 26 (LOO bias). *Under the assumptions of Theorem 2, it holds that*

$$\sup_{k\in\{0\}\cup[K]}\left|\mathbb{E}[R(\widehat{\beta}_k)]-\mathbb{E}[R(\widehat{\beta}_{k,-1})]\right|=o_n(1).$$

This is equivalently saying

$$\sup_{k\in\{0\}\cup[K]}\left|\mathbb{E}[r_k(w_1,\cdots,w_n)]-\mathbb{E}[f_k(w_1,\cdots,w_n)]\right|=o_n(1).$$

We defer the proofs of Lemma 25 and Lemma 26 to Sections S.6.3 and S.6.4, respectively.

Theorem 2 then follows from these three steps. To be precise, by putting together Lemmas 23 to 26, we obtain that with probability at least $1-4(n+p)^{-4}-2(n\log^2 n)^{-1}m_4-4(K+1)C_{\mathsf{T}_2}n^{-2}$, for all $k\in\{0\}\cup[K]$, we have

$$\sup_{k\in\{0\}\cup[K]}\left|R(\widehat{\beta}_k)-\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)\right|$$

$$\leq \frac{2\sigma_{\mathsf{T}_2} LK\xi(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot (\log n)^{3/2} + 2\sigma_{\mathsf{T}_2} L\bar{\xi}(C_{\Sigma,\zeta}, \Delta, m, B_0)(\log n)^{3/2}}{\sqrt{n}}. \tag{E.32}$$

Since $\zeta = p/n$ is both lower and upper bounded, thus we can conclude that

$$\sum_{n=1}^{\infty} \left\{4(n+p)^{-4} + 2(n\log^2 n)^{-1} m_4 + 4(K+1)C_{\mathsf{T}_2} n^{-2}\right\} < \infty.$$

Hence, Theorem 2 follows immediately by applying the first Borel–Cantelli lemma. More precisely, we prove that almost surely the event depicted in (E.32) occurs only finitely many times.

S.5 Supporting Lemmas for the Proofs of Theorems 2 to 4

We present in this section several supporting lemmas that are useful for the analysis presented in Appendix S.6 and Appendix S.7. Without any loss of generality, in this section, we always assume $n \geq 3$, thus $\log n \geq 1$.

S.5.1 Technical Preliminaries

We define below what it means for a distribution to satisfy log Sobolev inequality (LSI).

Definition 2 (LSI)**.** We say a distribution μ satisfies LSI if there exists a constant $\sigma(\mu) \geq 0$ such that for all smooth function f, it holds that

$$\mathrm{Ent}_{w\sim\mu}[f(w)^2] \leq 2\sigma^2(\mu)\mathbb{E}_{w\sim\mu}\left[\|\nabla f(w)\|_2^2\right], \tag{E.33}$$

where the entropy of a non-negative random variable Z is defined as

$$\mathrm{Ent}[Z] = \mathbb{E}[Z\log Z] - \mathbb{E}[Z]\log\mathbb{E}[Z].$$

S.5.2 Concentration Based on T_2-inequality

In this section, we discuss useful properties of the T_2-inequality. An important result that will be applied multiple times throughout the proof is Theorem 4.31 of Van Handel (2014), which we include below for the convenience of the readers. See also Gozlan (2009).

Proposition 27 (Equivalent characterizations of T_2-inequality)**.** *Let μ be a probability measure on a Polish space $(\mathcal{X}, d)$, and let $\{X_i\}_{i\leq n}$ be i.i.d. $\sim \mu$. Denote by $d_n(x,y) = [\sum_{i=1}^n d(x_i,y_i)^2]^{1/2}$. Then the following are equivalent:*

1. *μ satisfies the T_2-inequality:*

$$W_2(\mu,\nu) \leq \sqrt{2\sigma^2\mathcal{D}_{\mathrm{KL}}(\nu\,||\,\mu)} \quad \text{for all } \nu.$$

2. *$\mu^{\otimes n}$ satisfies the T_1-inequality for every $n \geq 1$:*

$$W_1(\mu^{\otimes n},\nu) \leq \sqrt{2\sigma^2\mathcal{D}_{\mathrm{KL}}(\nu\,||\,\mu^{\otimes n})} \quad \text{for all } \nu \text{ and } n \geq 1.$$

3. *There is an absolute constant C_{T_2}, such that*

$$\mathbb{P}\left(f(X_1,\cdots,X_n) - \mathbb{E}[f(X_1,\cdots,X_n)] \geq t\right) \leq C_{\mathsf{T}_2} e^{-t^2/2\sigma^2} \tag{E.34}$$

for every $n \geq 1$, $t \geq 0$ and 1-Lipschitz function f.

S.5.3 Dimension-Free Concentration

Define $w_i = (x_i, y_i)$. The following lemma is a straightforward consequence of the assumptions and the T_2-inequality.

Lemma 28 (Dimension-free concentration)**.** *We let $\sigma_{\mathsf{T}_2}^2 = \sigma_z^2 \vee \sigma_\varepsilon^2$, and $L = (L_f^2\sigma_\Sigma + L_f^2 + \sigma_\Sigma)^{1/2}$. Then for any $n \geq 1$, $t \geq 0$, and 1-Lipschitz function f, it holds that*

$$\mathbb{P}\left(f(w_1,\cdots,w_n) - \mathbb{E}[f(w_1,\cdots,w_n)] \geq Lt\right) \leq C_{\mathsf{T}_2} e^{-t^2/2\sigma_{\mathsf{T}_2}^2},$$

where we recall that $C_{\mathsf{T}_2} > 0$ is an absolute constant introduced in Proposition 27.

Proof. Since f is 1-Lipschitz, for any $w_i, \widetilde{w}_i \in \mathbb{R}^p$

$$\begin{aligned}|f(w_1, \cdots, w_n) - f(\widetilde{w}_1, \cdots, \widetilde{w}_n)| \leq & \sqrt{\sum_{i=1}^n \|w_i - \widetilde{w}_i\|_2^2} \\ = & \sqrt{\sum_{i=1}^n \|x_i - \widetilde{x}_i\|_2^2 + \sum_{i=1}^n |y_i - \widetilde{y}_i|^2} \\ \leq & \sqrt{\sum_{i=1}^n \sigma_\Sigma (L_f^2 + 1)\|z_i - \widetilde{z}_i\|_2^2 + \sum_{i=1}^n L_f^2 |\varepsilon_i - \widetilde{\varepsilon}_i|^2} \\ \leq & L\sqrt{\sum_{i=1}^n \|z_i - \widetilde{z}_i\|_2^2 + \sum_{i=1}^n |\varepsilon_i - \widetilde{\varepsilon}_i|^2}.\end{aligned}$$

Invoking Corollary 4.16 of Van Handel (2014), we obtain that

$$W_1(\mu_z^{\otimes n} \otimes \mu_\varepsilon^{\otimes n}, \nu) \leq \sqrt{2\sigma_{\mathsf{T}_2}^2 \mathcal{D}_{\mathtt{KL}}(\nu \,||\, \mu_z^{\otimes n} \otimes \mu_\varepsilon^{\otimes n})}$$

for all ν. We then see that the desired concentration inequality is a straightforward consequence of Proposition 27. □

S.5.4 Upper Bounding Operator Norms and Response Energy

We then state several technical lemmas required for our analysis. Recall that $\widehat{\Sigma} = X^\top X/n$. Our first lemma upper bounds the operator norm of $\widehat{\Sigma}$.

Lemma 29. *We assume the assumptions of Theorem 2. Then there exists a numerical constant $C_0 > 0$, such that with probability at least $1 - (n+p)^{-4}$*

$$\|\widehat{\Sigma}\|_{\mathrm{op}} \leq 2C_0\sigma_\Sigma(1 + \zeta) + 1.$$

Proof. Note that the operator norm of $\widehat{\Sigma}$ is equal to the operator norm of $Z\Sigma Z^\top/n + 1_{n\times n}/n \in \mathbb{R}^{n\times n}$.

To proceed, we will utilize a canonical concentration inequality that bounds the operator norm of random matrices with sub-Gaussian entries. This further requires the introduction of several related concepts.

To be specific, we say a random variable R is *sub-Gaussian* if and only if there exists $K_R > 0$ such that $\|R\|_{L^d} \leq K_R\sqrt{d}$ for all $d \geq 1$. Proposition 2.5.2 of Vershynin (2018) tells us that when such upper bound is satisfied, the sub-Gaussian norm of this random variable $\|Z\|_{\Psi_2}$ is no larger than $4K_R$.

By Assumption C and Proposition 27, it holds that

$$\mathbb{P}\left(|z_{11}| \geq t\right) \leq 2C_{\mathsf{T}_2} e^{-t^2/2\sigma_z^2}.$$

Leveraging the above upper bound and applying an appropriate integral inequality, we can conclude that for all $d \geq 1$,

$$\mathbb{E}[|z_{11}|^d] \leq C_{\mathsf{T}_2} d(d/2)^{d/2},$$

hence $\|z_{11}\|_{\Psi_2} \leq 8 + 8C_{\mathsf{T}_2}$. By Theorem 4.4.5 of Vershynin (2010), we see that for all $t \geq 0$, with probability at least $1 - 2\exp(-t^2)$

$$\|Z\|_{\mathrm{op}} \leq C'(8 + 8C_{\mathsf{T}_2})(\sqrt{n} + \sqrt{p} + t), \tag{E.35}$$

where $C' > 0$ is a numerical constant. Taking $t = 2\sqrt{\log(p+n)}$, we conclude that $\|Z\|_{\mathrm{op}} \leq C'(8 + 8C_{\mathsf{T}_2})(\sqrt{n} + \sqrt{p} + 2\sqrt{\log(n+p)})$ with probability at least $1 - 2(p+n)^{-4}$. When this occurs, a straightforward consequence is that

$$n\|\widehat{\Sigma}\|_{\mathrm{op}} \leq \|Z\|_{\mathrm{op}}^2 \|\Sigma\|_{\mathrm{op}} + n \leq C_0\sigma_\Sigma(n + p + \log(n+p)) + n$$

for some positive numerical constant C_0, thus completing the proof of the lemma. □

Our next lemma upper bounds $\|y\|_2^2/n$. This lemma is a direct consequence of Chebyshev's inequality, and we skip the proof for the compactness of the presentation.

Lemma 30. *We assume the assumptions of Theorem 2. Then with probability at least* $1 - n^{-1}m_4 \log^{-2} n$, *we have* $\|y\|_2^2/n \le m_2 + \log n$.

S.5.5 Other Useful Norm Bounds

Our next lemma upper bounds the Euclidean norm of $\theta = \mathbb{E}[y_0 x_0] \in \mathbb{R}^{p+1}$, where we recall that $(x_0, y_0) \stackrel{d}{=} (x_1, y_1)$.

Lemma 31. *Under the assumptions of Theorem 2, we have* $\|\theta\|_2 \le (\sigma_\Sigma^{1/2} + 1) m_2^{1/2}$.

Proof. We notice that

$$\theta = \begin{pmatrix} \Sigma^{1/2}\mathbb{E}[y_0 z_0] \\ \mathbb{E}[y_0] \end{pmatrix}.$$

We let $x_0^\top = (z_0^\top \Sigma^{1/2}, 1)$. By assumption, z_0 is isotropic. Hence, y_0 admits the following decomposition:

$$y_0 = \sum_{i=1}^{p} \mathbb{E}[y_0 z_{0,i}] z_{0,i} + \omega, \qquad \mathbb{E}[\omega z_{0,i}] = 0 \ \text{ for all } i \in [n].$$

In addition, $\mathbb{E}[y_0^2] = \mathbb{E}[w^2] + \sum_{i\in[p]} \mathbb{E}[y_0 z_{0,i}]^2$. As a result, we are able to deduce that $\|\mathbb{E}[y_0 z_0]\|_2 \le m_2^{1/2}$, where we recall that $m_2 = \mathbb{E}[y_0^2]$. This further tells us $\|\theta\|_2 \le \|\Sigma\|_{\rm op}^{1/2} \times \|\mathbb{E}[y_0 z_0]\|_2 + m_2^{1/2} \le (\sigma_\Sigma^{1/2} + 1) m_2^{1/2}$, thus completing the proof of the lemma. □

We next prove that $\|\widehat{\Sigma}\|_{\rm op}$ is sub-exponential.

Lemma 32. *We define* $\widetilde{C}_0 = C' \sigma_\Sigma (8 + 8C_{\mathsf{T}_2})$, *where we recall that* C' *is a positive numerical constant that appears in Equation (E.35). Under the assumptions of Theorem 2, for all* $\lambda \ge 0$ *and* $n \ge \lambda \widetilde{C}_0^2 + 1$, *there exists a constant* $\mathcal{E}(\widetilde{C}_0, \zeta, \lambda) > 0$ *that depends only on* $(\widetilde{C}_0, \zeta, \lambda)$, *such that*

$$\mathbb{E}[\exp(\lambda \|\widehat{\Sigma}\|_{\rm op})] \le \mathcal{E}(\widetilde{C}_0, \zeta, \lambda).$$

Proof. By Equation (E.35), for all $t \ge 0$, with probability at least $1 - 2\exp(-nt^2)$

$$\|\widehat{\Sigma}\|_{\rm op}^{1/2} = n^{-1/2}\|X\|_{\rm op} \le \widetilde{C}_0(1 + \zeta^{1/2} + t).$$

As a result, for all $\lambda \ge 0$,

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda \|\widehat{\Sigma}\|_{\rm op})] \\
&\le 1 + \int_0^\infty 2\lambda s e^{\lambda s^2} \mathbb{P}\Big(\|\widehat{\Sigma}\|_{\rm op}^{1/2} \ge s\Big) \mathrm{d}s \\
&\le 1 + 2\lambda \widetilde{C}_0^2 (1 + \zeta^{1/2})^2 e^{\lambda \widetilde{C}_0^2 (1+\zeta^{1/2})^2} + \int_{\widetilde{C}_0(1+\zeta^{1/2})} 2\lambda s e^{\lambda s^2} \mathbb{P}\Big(\|\widehat{\Sigma}\|_{\rm op}^{1/2} \ge s\Big) \mathrm{d}s \\
&\le 1 + 2\lambda \widetilde{C}_0^2 (1 + \zeta^{1/2})^2 e^{\lambda \widetilde{C}_0^2 (1+\zeta^{1/2})^2} + \int_0^\infty 4\lambda \widetilde{C}_0^2 (1 + \zeta^{1/2} + t) e^{\lambda \widetilde{C}_0^2 (1+\zeta^{1/2}+t)^2 - nt^2} \mathrm{d}t \le \mathcal{E}(\widetilde{C}_0, \zeta, \lambda),
\end{aligned}$$

thus completing the proof of the lemma. □

S.5.6 Upper Bounding $\|\widehat{\beta}_k\|_2$ and $\|\widehat{\beta}_{k,-i}\|_2$

We then prove that on Ω, the Euclidean norm of the coefficient estimates $\{\widehat{\beta}_k, \widehat{\beta}_{k,-i} : k \in [K], i \in [n]\}$ are uniformly upper bounded. In addition, apart from a logarithmic factor, this upper bound depends only on the constants from our assumptions and in particular is independent of (n, p).

Lemma 33. *For the sake of simplicity, we let*

$$B_* = (B_0 + \Delta C_{\Sigma,\zeta}^{1/2}\sqrt{m+1}) \cdot e^{C_{\Sigma,\zeta}\Delta}. \tag{E.36}$$

Then on the set Ω, for all $k \in \{0\} \cup [K]$ and $i \in [n]$, it holds that

$$\|\widehat{\beta}_k\|_2 \le B_*\sqrt{\log n}, \qquad \|\widehat{\beta}_{k,i}\|_2 \le B_*\sqrt{\log n}.$$

Proof. By definition,

$$\begin{aligned}\widehat{\beta}_{k+1} =& \widehat{\beta}_k + \frac{\delta_k}{n}\sum_{i=1}^{n}(y_i - x_i^\top \widehat{\beta}_k)x_i \\ =& \widehat{\beta}_k - \delta_k \widehat{\Sigma}\widehat{\beta}_k + \frac{\delta_k}{n}X^\top y.\end{aligned}$$

Applying the triangle inequality, we obtain the following upper bound:

$$\begin{aligned}\|\widehat{\beta}_{k+1}\|_2 \le& \|\widehat{\beta}_k\|_2 + \delta_k\|\widehat{\Sigma}\|_{\mathrm{op}} \cdot \|\widehat{\beta}_k\|_2 + \delta_k \cdot \|\widehat{\Sigma}\|_{\mathrm{op}}^{1/2} \cdot \|y/\sqrt{n}\|_2 \\ \le& (1 + \delta_k C_{\Sigma,\zeta}) \cdot \|\widehat{\beta}_k\|_2 + \delta_k C_{\Sigma,\zeta}^{1/2}\sqrt{m + \log n}.\end{aligned}$$

By induction, we see that on Ω

$$\|\widehat{\beta}_k\|_2 \le \left(B_0 + \Delta C_{\Sigma,\zeta}^{1/2}\sqrt{m+\log n}\right) \cdot e^{C_{\Sigma,\zeta}\Delta}$$

for all $k \in [K]$. The upper bound for $\|\widehat{\beta}_{k,-i}\|_2$ follows using exactly the same argument. We complete the proof of the lemma as $\log n \ge 1$. □

The following corollary is a straightforward consequence of Lemma 33 and the Cauchy-Schwartz inequality.

Corollary 34. *On the set Ω, it holds that*

$$\begin{aligned}\frac{1}{n}\|y - X\widehat{\beta}_{k,-i}\|_2^2 &\le \left(2m + 2 + 2C_{\Sigma,\zeta}B_*^2\right) \cdot \log n, \\ \frac{1}{n}\|y - X\widehat{\beta}_k\|_2^2 &\le \left(2m + 2 + 2C_{\Sigma,\zeta}B_*^2\right) \cdot \log n\end{aligned}$$

for all $k \in \{0\} \cup [K]$ and $i \in [n]$.

For the compactness of future presentation, we define

$$\bar{B}_* = (2m + 2 + 2C_{\Sigma,\zeta}B_*^2)^{1/2} \tag{E.37}$$

We comment that both B_* and $\bar{B}_*$ depend only on $(C_{\Sigma,\zeta}, \Delta, m, B_0)$.

S.5.7 Upper Bounding $|y_i - x_i^\top \beta_{k,-i}|$

We next upper bound $|y_i - x_i^\top \widehat{\beta}_{k,-i}|$ on Ω. More precisely, we shall upper bound collectively the Frobenius norms of

$$\begin{aligned}a_k &= (y_i - x_i^\top \widehat{\beta}_{k,-i})_{i=1}^n \in \mathbb{R}^n \qquad \text{and} \\ E_k &= \left[X(\widehat{\beta}_k - \widehat{\beta}_{k,-1}) \mid \cdots \mid X(\widehat{\beta}_k - \widehat{\beta}_{k,-n})\right] \in \mathbb{R}^{n\times n}\end{aligned}$$

respectively and recursively. For the base case $k = 0$, we have

$$\|a_0\|_2^2 \le \bar{B}_*^2 n \log n, \qquad \|E_0\|_F^2 = 0,$$

where the first upper bound follows from Corollary 34.

Our lemma for this part can be formally stated as follows:

Lemma 35. *We define*

$$\mathcal{G}_1(C_{\Sigma,\zeta},\Delta,m,B_0)=\bar{B}_*\sqrt{e^{3\Delta C_{\Sigma,\zeta}+2\Delta^2C_{\Sigma,\zeta}^2}(\Delta C_{\Sigma,\zeta}+2\Delta^2C_{\Sigma,\zeta}^2)}, \tag{E.38}$$

$$\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)=\bar{B}_*+\Delta C_{\Sigma,\zeta}\sqrt{8\bar{B}_*^2+2\mathcal{G}_1(C_{\Sigma,\zeta},\Delta,m,B_0)^2}. \tag{E.39}$$

Then on the set Ω, for all $k\in\{0\}\cup[K]$ we have

$$\frac{1}{\sqrt{n}}\|E_k\|_F\le\mathcal{G}_1(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\sqrt{\log n}, \tag{E.40}$$

$$\frac{1}{\sqrt{n}}\|a_k\|_2\le\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\sqrt{\log n}. \tag{E.41}$$

Proof. We first prove Equation (E.40). We denote by $X_{-i}\in\mathbb{R}^{(n-1)\times(p+1)}$ the matrix obtained by deleting the i-th row from X. By definition,

$$\begin{aligned}X(\widehat{\beta}_{k+1}-\widehat{\beta}_{k+1,-i})=&X(\widehat{\beta}_k-\widehat{\beta}_{k,-i})+\frac{\delta_k(y_i-x_i^\top\widehat{\beta}_k)}{n}Xx_i-\frac{\delta_k}{n}X\sum_{j\neq i}x_jx_j^\top(\widehat{\beta}_k-\widehat{\beta}_{k,-i})\\=&X(\widehat{\beta}_k-\widehat{\beta}_{k,-i})+\frac{\delta_k(y_i-x_i^\top\widehat{\beta}_k)}{n}Xx_i-\frac{\delta_k}{n}XX_{-i}^\top X_{-i}(\widehat{\beta}_k-\widehat{\beta}_{k,-i}),\end{aligned}$$

which further implies

$$\begin{aligned}&\|X(\widehat{\beta}_{k+1}-\widehat{\beta}_{k+1,-i})\|_2^2\\&\le(1+\delta_kC_{\Sigma,\zeta})^2\|X(\widehat{\beta}_k-\widehat{\beta}_{k,-i})\|_2^2+\frac{\delta_k^2(y_i-x_i^\top\widehat{\beta}_k)^2}{n^2}\|Xx_i\|_2^2\\&\quad+\frac{2\delta_k(1+\delta_kC_{\Sigma,\zeta})\cdot|y_i-x_i^\top\widehat{\beta}_k|}{n}\|X(\widehat{\beta}_k-\widehat{\beta}_{k,-i})\|_2\cdot\|Xx_i\|_2\\&\le(1+\delta_kC_{\Sigma,\zeta})^2\|X(\widehat{\beta}_k-\widehat{\beta}_{k,-i})\|_2^2+\delta_k^2C_{\Sigma,\zeta}^2(y_i-x_i^\top\widehat{\beta}_k)^2\\&\quad+\delta_kC_{\Sigma,\zeta}(1+\delta_kC_{\Sigma,\zeta})\cdot\big\{(y_i-x_i^\top\widehat{\beta}_k)^2+\|X(\widehat{\beta}_k-\widehat{\beta}_{k,-i})\|_2^2\big\}\\&\le\big(1+3\delta_kC_{\Sigma,\zeta}+2\delta_k^2C_{\Sigma,\zeta}^2\big)\cdot\|X(\widehat{\beta}_k-\widehat{\beta}_{k,-i})\|_2^2+\big(\delta_kC_{\Sigma,\zeta}+2\delta_k^2C_{\Sigma,\zeta}^2\big)\cdot(y_i-x_i^\top\widehat{\beta}_k)^2,\end{aligned}$$

where we make use of the fact that $\|Xx_i\|_2/n\le C_{\Sigma,\zeta}$ on Ω. Putting together the above upper bound and Corollary 34, then summing over $i\in[n]$, we obtain the following inequality:

$$\|E_{k+1}\|_F^2\le\big(1+3\delta_kC_{\Sigma,\zeta}+2\delta_k^2C_{\Sigma,\zeta}^2\big)\cdot\|E_k\|_F^2+\big(\delta_kC_{\Sigma,\zeta}+2\delta_k^2C_{\Sigma,\zeta}^2\big)\cdot\bar{B}_*^2\log n.$$

Employing the standard induction argument, we can conclude that

$$\frac{1}{n}\|E_k\|_F^2\le e^{3\Delta C_{\Sigma,\zeta}+2\Delta^2C_{\Sigma,\zeta}^2}(\Delta C_{\Sigma,\zeta}+2\Delta^2C_{\Sigma,\zeta}^2)\cdot\bar{B}_*^2\log n=\mathcal{G}_1(C_{\Sigma,\zeta},\Delta,m,B_0)^2\log n$$

for all $k\in\{0\}\cup[K]$. This completes the proof of Equation (E.40).

Next, we prove Equation (E.41). By definition,

$$\begin{aligned}y_i-x_i^\top\widehat{\beta}_{k+1,-i}=&\,y_i-x_i^\top\widehat{\beta}_{k,-i}-\frac{\delta_k}{n}\sum_{j\neq i}(y_j-x_j^\top\widehat{\beta}_{k,-i})x_i^\top x_j\\=&\,y_i-x_i^\top\widehat{\beta}_{k,-i}-\frac{\delta_k}{n}\sum_{j\neq i}(y_j-x_j^\top\widehat{\beta}_k)x_i^\top x_j-\frac{\delta_k}{n}\sum_{j\neq i}x_i^\top x_jx_j^\top(\widehat{\beta}_k-\widehat{\beta}_{k,-i}).\end{aligned}$$

We let $D=\mathrm{diag}\{(\|x_i\|_2^2/n)_{i=1}^n\}\in\mathbb{R}^{n\times n}$. We denote by $a_{k,i}$ the i-th entry of a_k. From the above equality, we can deduce that

$$(a_{k+1,i}-a_{k,i})^2\le\frac{2\delta_k^2}{n^2}\Big(\sum_{j\neq i}(y_j-x_j^\top\widehat{\beta}_k)x_i^\top x_j\Big)^2+\frac{2\delta_k^2}{n^2}\Big(\sum_{j\neq i}x_i^\top x_jx_j^\top(\widehat{\beta}_k-\widehat{\beta}_{k,-i})\Big)^2.$$

Summing over $i \in [n]$, we obtain

$$\begin{aligned}
&\|a_{k+1} - a_k\|_2^2 \\
&\le \frac{2\delta_k^2}{n^2}\big\|(XX^\top - nD)(y - X\widehat{\beta}_k)\big\|_2^2 + \frac{2\delta_k^2}{n^2}\sum_{i=1}^n \Big(\sum_{j\neq i}(x_i^\top x_j)^2\Big)\cdot\Big(\sum_{j\neq i}\big(x_j^\top(\widehat{\beta}_k - \widehat{\beta}_{k,-i})\big)^2\Big) \\
&\overset{(i)}{\le} 8n\delta_k^2 C_{\Sigma,\zeta}^2\cdot\bar{B}_*^2\log n + 2\delta_k^2 C_{\Sigma,\zeta}^2\sum_{i=1}^n\|X(\widehat{\beta}_k - \widehat{\beta}_{k,-i})\|_2^2 \\
&\overset{(ii)}{\le} 8n\delta_k^2 C_{\Sigma,\zeta}^2\cdot\bar{B}_*^2\log n + 2n\delta_k^2 C_{\Sigma,\zeta}^2\mathcal{G}_1(C_{\Sigma,\zeta},\Delta,m,B_0)^2\cdot\log n \\
&= n\delta_k^2\cdot\mathcal{G}'(C_{\Sigma,\zeta},\Delta,m,B_0)^2\cdot\log n,
\end{aligned}$$

where to derive (i), we employ the following established results: (1) On Ω we have $\|nD\|_{\mathrm{op}} \le nC_{\Sigma,\zeta}$ and $\|XX^\top\|_{\mathrm{op}} \le nC_{\Sigma,\zeta}$. (2) By Corollary 34, on Ω we have $\|y - X\widehat{\beta}_k\|_2^2/n \le \bar{B}_*^2\cdot\log n$. To derive (ii), we simply apply Equation (E.40), which we have already proved. Therefore, by triangle inequality

$$\frac{1}{\sqrt{n}}\|a_{k+1}\|_2 \le \frac{1}{\sqrt{n}}\|a_k\|_2 + \frac{1}{\sqrt{n}}\|a_{k+1} - a_k\|_2 \le \frac{1}{\sqrt{n}}\|a_k\|_2 + \delta_k\mathcal{G}'(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\sqrt{\log n}.$$

By standard induction argument, we see that for all $k \in \{0\}\cup[K]$,

$$\frac{1}{\sqrt{n}}\|a_k\|_2 \le \bar{B}_*\sqrt{\log n} + \Delta\mathcal{G}'(C_{\Sigma,\zeta},\Delta,m,B_0)\sqrt{\log n} = \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\sqrt{\log n},$$

which concludes the proof of Equation (E.41). □

S.6 Proof of Theorem 2

Theorem 2 (Squared risk consistency of LOOCV)**.** *Suppose that (x_i, y_i), $i \in [n]$ are i.i.d., and satisfy both Assumptions C and D. In addition, assume that there are constants $\Delta, B_0, \zeta_L, \zeta_U$ (independent of n, p) such that: (1) $\sum_{k=1}^K \delta_{k-1} \le \Delta$, (2) $\|\widehat{\beta}_0\|_2 \le B_0$, and (3) $0 < \zeta_L \le p/n \le \zeta_U < \infty$. Furthermore, let $K = o(n\cdot(\log n)^{-3/2})$. Then, as $n, p \to \infty$,*

$$\max_{k\in[K]}\Big|\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k) - R(\widehat{\beta}_k)\Big| \xrightarrow{\text{a.s.}} 0, \tag{8}$$

where we recall that $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$ and $R(\widehat{\beta}_k)$ are as defined in (4) and (3), respectively.

To better present our proof idea, we consider in this section the quadratic functional $\psi(y,u) = (y-u)^2$. A compact version of proof for general functional estimation can be found in Appendix S.8.

S.6.1 Proof of Lemma 23

Lemma 23 (LOO concentration)**.** *We assume the assumptions of Theorem 2. Then with probability at least $1 - 2(n+p)^{-4} - (n\log^2 n)^{-1}m_4 - 2(K+1)C_{\mathsf{T}_2}n^{-2}$, it holds that for all $k \in \{0\}\cup[K]$*

$$\Big|\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k) - \mathbb{E}[\widetilde{f}_k(w_1,\cdots,w_n)]\Big| \le \frac{2\sigma_{\mathsf{T}_2}LK\xi(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot(\log n)^{3/2}}{\sqrt{n}},$$

where we let $L = (L_f^2\sigma_\Sigma + L_f^2 + \sigma_\Sigma)^{1/2}$, $\sigma_{\mathsf{T}_2}^2 = \sigma_z^2 \vee \sigma_\varepsilon^2$, and C_{T_2} is a positive numerical constant that appears in Proposition 27.

Proof. We claim that Lemma 22 can be used to show $\widetilde{f}_k$ is Lipschitz continuous. More precisely, for $W, W' \in \mathbb{R}^{n(p+2)}$, it holds that

$$\begin{aligned}
\Big|\widetilde{f}_k(W) - \widetilde{f}_k(W')\Big| &= |f_k(h(W)) - f_k(h(W'))| \\
&\le \frac{K\xi(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n}{\sqrt{n}}\cdot\|h(W) - h(W')\|_F \\
&\le \frac{K\xi(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n}{\sqrt{n}}\cdot\|W - W'\|_F.
\end{aligned}$$

Namely, $\widetilde{f}_k$ is $n^{-1/2}K\xi(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n$-Lipschitz continuous for all $k\in\{0\}\cup[K]$. Applying Lemma 28, we conclude that

$$\mathbb{P}\left(\left|\widetilde{f}_k(w_1,\cdots,w_n)-\mathbb{E}[\widetilde{f}_k(w_1,\cdots,w_n)]\right|\geq\frac{2\sigma_{\mathsf{T}_2}LK\xi(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot(\log n)^{3/2}}{\sqrt{n}}\right)\leq 2C_{\mathsf{T}_2}n^{-2}.$$

Note that on the set Ω we have $\widetilde{f}_k(w_1,\cdots,w_n)=f_k(w_1,\cdots,w_n)=\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$ for all $k\in\{0\}\cup[K]$. This completes the proof of the lemma. □

S.6.2 Proof of Lemma 24

Lemma 24 (Risk concentration)**.** *We write $R(\widehat{\beta}_k)=r_k(w_1,\cdots,w_n)$ and define $\widetilde{r}_k(w_1,\cdots,w_n)=r_k(h(w_1,\cdots,w_n))$. Then under the assumptions of Theorem 2, with probability at least $1-2(n+p)^{-4}-(n\log^2 n)^{-1}m_4-2(K+1)C_{\mathsf{T}_2}n^{-2}$, for all $k\in\{0\}\cup[K]$ we have*

$$\left|R(\widehat{\beta}_k)-\mathbb{E}[\widetilde{r}_k(w_1,\cdots,w_n)]\right|\leq\frac{2\sigma_{\mathsf{T}_2}L\bar{\xi}(C_{\Sigma,\zeta},\Delta,m,B_0)(\log n)^{3/2}}{\sqrt{n}},$$

where $\bar{\xi}(C_{\Sigma,\zeta},\Delta,m,B_0)>0$ depends uniquely on $(C_{\Sigma,\zeta},\Delta,m,B_0)$.

Proof. For $s\in[n]$, direct computation gives

$$\begin{aligned}\nabla_{x_s}R(\widehat{\beta}_k)&=2\widehat{\beta}_k^\top\widetilde{\Sigma}\nabla_{x_s}\widehat{\beta}_k-2\widehat{\theta}^\top\nabla_{x_s}\widehat{\beta}_k,\\ \frac{\partial}{\partial_{y_s}}R(\widehat{\beta}_k)&=2\widehat{\beta}_k^\top\widetilde{\Sigma}\frac{\partial}{\partial_{y_s}}\widehat{\beta}_k-2\theta^\top\frac{\partial}{\partial_{y_s}}\widehat{\beta}_k,\end{aligned}$$

where

$$\theta=\mathbb{E}[y_0x_0]\in\mathbb{R}^{p+1},\qquad\widetilde{\Sigma}=\begin{bmatrix}\Sigma & 0_p\\ 0_p^\top & 1\end{bmatrix}\in\mathbb{R}^{(p+1)\times(p+1)}.$$

By definition,

$$\begin{aligned}\nabla_{x_s}\widehat{\beta}_{k+1}&=\nabla_{x_s}\widehat{\beta}_k-\delta_k\widehat{\Sigma}\cdot\nabla_{x_s}\widehat{\beta}_k+\frac{\delta_k}{n}(y_s-x_s^\top\widehat{\beta}_k)I_{p+1}-\frac{\delta_k}{n}x_s\widehat{\beta}_k^\top,\\ \frac{\partial}{\partial_{y_s}}\widehat{\beta}_{k+1}&=\frac{\partial}{\partial_{y_s}}\widehat{\beta}_k-\delta_k\widehat{\Sigma}\frac{\partial}{\partial_{y_s}}\widehat{\beta}_k+\frac{\delta_k}{n}x_s.\end{aligned}$$

Standard induction argument leads to the following decomposition:

$$\begin{aligned}\nabla_{x_s}\widehat{\beta}_{k+1}&=\sum_{k'=1}^{k}H_{k',k}\cdot\left(\frac{\delta_{k'}}{n}(y_s-x_s^\top\widehat{\beta}_{k'})I_{p+1}-\frac{\delta_{k'}}{n}x_s\widehat{\beta}_{k'}^\top\right),\\ \frac{\partial}{\partial_{y_s}}\widehat{\beta}_{k+1}&=\sum_{k'=1}^{k}H_{k',k}\cdot\frac{\delta_{k'}}{n}x_s,\end{aligned}$$

where $H_{k',r}=\prod_{j=k'+1}^{r}M_{k'+1+r-j}$ and $M_j=I_{p+1}-\delta_j\widehat{\Sigma}$ are defined in Lemma 37. Combining all these arguments, we arrive at the following equations:

$$\begin{aligned}v^\top\nabla_{x_s}\widehat{\beta}_{k+1}&=\sum_{k'=1}^{k}v^\top H_{k',k}\cdot\frac{\delta_{k'}}{n}(y_s-x_s^\top\widehat{\beta}_{k'})-\sum_{k'=1}^{k}\frac{\delta_{k'}}{n}x_s^\top H_{k',k}v\widehat{\beta}_{k'}^\top,\\ v^\top\frac{\partial}{\partial_{y_s}}\widehat{\beta}_{k+1}&=\sum_{k'=1}^{k}\frac{\delta_{k'}}{n}x_s^\top H_{k',k}v.\end{aligned}$$

The above equations hold for all $v\in\{\theta,\widetilde{\Sigma}\widehat{\beta}_{k+1}\}$. Recall that $\theta=\mathbb{E}[y_0x_0]$. This further implies that

$$\nabla_X R(\widehat{\beta}_{k+1})$$

$$= \sum_{k'=1}^{k} \frac{2\delta_{k'}}{n} \cdot \Big\{ (y - X\widehat{\beta}_{k'})\widehat{\beta}_{k+1}^\top \widetilde{\Sigma} H_{k',k} - X H_{k,k'} \widetilde{\Sigma} \widehat{\beta}_{k+1} \widehat{\beta}_{k'}^\top - (y - X\widehat{\beta}_{k'})\theta^\top H_{k',k} + X H_{k,k'} \theta \widehat{\beta}_{k'}^\top \Big\},$$

$$\nabla_y \mathcal{R}(\widehat{\beta}_{k+1}) = \sum_{k'=1}^{k} \frac{2\delta_{k'}}{n} \cdot \Big\{ X H_{k,k'} \widetilde{\Sigma} \widehat{\beta}_{k+1} - X H_{k,k'} \theta \Big\}.$$

Recall that B_* is defined in Equation (E.36). Invoking triangle inequality, we obtain that on Ω,

$$\begin{aligned}
\|\nabla_X R(\widehat{\beta}_{k+1})\|_F \le & \sum_{k'=1}^{k} \frac{2\delta_{k'}}{n} \cdot \Big\{ \|y - X\widehat{\beta}_{k'}\|_2 \cdot (\|\widehat{\beta}_{k+1}\|_2 \|\widetilde{\Sigma}\|_{\mathrm{op}} + \|\theta\|_2) \cdot \|H_{k',k}\|_{\mathrm{op}} \\
& + \|X\|_{\mathrm{op}} \cdot \|H_{k,k'}\|_{\mathrm{op}} \cdot (\|\widehat{\beta}_{k+1}\|_2 \|\widetilde{\Sigma}\|_{\mathrm{op}} + \|\theta\|_2) \cdot \|\widehat{\beta}_{k'}\|_2 \Big\} \\
\le & \frac{2\Delta e^{\Delta C_{\Sigma,\zeta}} \cdot \sqrt{\log n}}{\sqrt{n}} \cdot \Big(\bar{B}_* + C_{\Sigma,\zeta}^{1/2} B_* \Big) \Big(B_*(\sigma_\Sigma + 1) \cdot \sqrt{\log n} + (\sigma_\Sigma^{1/2} + 1) m_2^{1/2} \Big),
\end{aligned}$$

where the inequality follows by invoking Lemma 31 to upper bound $\|\theta\|_2$. Also, by Lemma 37 we know that $\|H_{k',r}\|_{\mathrm{op}} \le e^{\Delta C_{\Sigma,\zeta}}$. Similarly, we obtain

$$\begin{aligned}
\|\nabla_y R(\widehat{\beta}_{k+1})\|_2 \le & \sum_{k'=1}^{k} \frac{2\delta_{k'}}{n} \cdot \Big\{ \|X\|_{\mathrm{op}} \cdot \|H_{k,k'}\|_{\mathrm{op}} \cdot \|\widetilde{\Sigma}\|_{\mathrm{op}} \cdot \|\beta_{k+1}\| + \|X\|_{\mathrm{op}} \cdot \|H_{k,k'}\|_{\mathrm{op}} \cdot \|\theta\|_2 \Big\} \\
\le & \frac{2\Delta e^{\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}^{1/2}}{\sqrt{n}} \cdot \Big(B_*(\sigma_\Sigma + 1) + (\sigma_\Sigma^{1/2} + 1) m_2^{1/2} \Big) \cdot \sqrt{\log n}.
\end{aligned}$$

The above inequalities give an upper bound for $\|\nabla_W R(\widehat{\beta}_{k+1})\|_2$ on Ω. The rest parts of the proof are similar to the proof of Lemma 23 given Lemma 22. □

S.6.3 Proof of Lemma 25

Lemma 25 (Projection effects)**.** *Under the assumptions of Theorem 2, it holds that*

$$\begin{aligned}
\sup_{k \in \{0\} \cup [K]} & |\mathbb{E}[\widetilde{r}_k(w_1, \cdots, w_n)] - \mathbb{E}[r_k(w_1, \cdots, w_n)]| = o_n(1), \\
\sup_{k \in \{0\} \cup [K]} & \Big|\mathbb{E}[\widetilde{f}_k(w_1, \cdots, w_n)] - \mathbb{E}[f_k(w_1, \cdots, w_n)]\Big| = o_n(1).
\end{aligned} \tag{E.31}$$

Proof. We shall first upper bound the fourth moments $\mathbb{E}[(y_0 - x_0^\top \widehat{\beta}_{k,-1})^4]$ and $\mathbb{E}[(y_0 - x_0^\top \widehat{\beta}_k)^4]$. By standard induction, it is not hard to see that for all $0 \le k \le K$ and $i \in [n]$,

$$\|\widehat{\beta}_k\|_2 \le \exp(\Delta \|\widehat{\Sigma}\|_{\mathrm{op}}) \cdot \Big(B_0 + \Delta n^{-1} \|X\|_{\mathrm{op}} \cdot \|y\|_2 \Big), \tag{E.42}$$

$$\|\widehat{\beta}_{k,-i}\|_2 \le \exp(\Delta \|\widehat{\Sigma}\|_{\mathrm{op}}) \cdot \Big(B_0 + \Delta n^{-1} \|X\|_{\mathrm{op}} \cdot \|y\|_2 \Big). \tag{E.43}$$

For technical reasons that will become clear soon, we need to upper bound the expectations of $\|\widehat{\beta}_k\|_2$ and $\|\widehat{\beta}_{k,-i}\|_2$. To this end, we find it useful to show $\|\widehat{\Sigma}\|_{\mathrm{op}}^{1/2}$ is sub-Gaussian. Next, we will employ Lemma 32 to upper bound $\mathbb{E}[(y_0 - x_0^\top \beta_{k,-1})^4]$ and $\mathbb{E}[(y_0 - x_0^\top \beta_k)^4]$. Invoking the Cauchy-Schwartz inequality and triangle inequality, we obtain that for $n \ge N(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)$,

$$\begin{aligned}
\mathbb{E}[(y_0 - x_0^\top \widehat{\beta}_{k,-1})^4] & \le \mathbb{E}[\|(y_0, x_0^\top \widehat{\beta}_{k,-1})\|_2^4] \\
& \le 8\mathbb{E}[y_1^4] + 8\mathbb{E}[(x_1^\top \widehat{\beta}_{k,-1})^4] = 8m_4 + 8\mathbb{E}[((z_1^\top, 1)\widetilde{\Sigma}^{1/2} \widehat{\beta}_{k,-1})^4] \\
& \overset{(i)}{\le} 8m_4 + C_z \mathbb{E}[\|\widetilde{\Sigma}^{1/2} \widehat{\beta}_{k,-1}\|_2^4] \\
& \overset{(ii)}{\le} \mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)^2
\end{aligned}$$

where $C_z > 0$ is a constant that depends only on μ_z, $\mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta) \in \mathbb{R}_+$ and $N(\sigma_\Sigma, \zeta, B_0, m_8, \Delta) \in \mathbb{N}_+$ depend only on $(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)$. To derive *(i)* we use the following facts: (1) μ_z has zero expectation; (2) z_1 is independent of $\widetilde{\Sigma}^{1/2}\beta_{k,-1}$. To derive *(ii)* we apply Equation (E.43) and Lemma 32. Similarly, we can show that for $n \geq N(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)$,

$$\mathbb{E}[(y_0 - x_0^\top \beta_k)^4] \leq \mathbb{E}[\|(y_0, x_0^\top \beta_k)\|_2^4] \leq \mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)^2. \tag{E.44}$$

Finally, we are ready to establish Equation (E.31). By the Cauchy-Schwartz inequality,

$$\begin{aligned}
\left|\mathbb{E}[r_k(w_1, \cdots, w_n)] - \mathbb{E}[\widetilde{r}_k(w_1, \cdots, w_n)]\right| &\leq \mathbb{P}(\Omega^c)^{1/2}\mathbb{E}[(y_0 - x_0^\top \widehat{\beta}_k)^4]^{1/2}, \\
\left|\mathbb{E}[f_k(w_1, \cdots, w_n)] - \mathbb{E}[\widetilde{f}_k(w_1, \cdots, w_n)]\right| &\leq \mathbb{P}(\Omega^c)^{1/2}\mathbb{E}[(y_1 - x_1^\top \widehat{\beta}_{k,-1})^4]^{1/2},
\end{aligned}$$

which for $n \geq N(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)$ are upper bounded by

$$\left(2(n+p)^{-1} + n^{-1}m_4 + 2C_{\mathsf{T}_2}n^{-2}\right)^{1/2} \mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta).$$

The above upper bound goes to zero as $n, p \to \infty$, thus completing the proof of the lemma. □

S.6.4 Proof of Lemma 26

Lemma 26 (LOO bias)**.** *Under the assumptions of Theorem 2, it holds that*

$$\sup_{k \in \{0\} \cup [K]} \left|\mathbb{E}[R(\widehat{\beta}_k)] - \mathbb{E}[R(\widehat{\beta}_{k,-1})]\right| = o_n(1).$$

This is equivalently saying

$$\sup_{k \in \{0\} \cup [K]} \left|\mathbb{E}[r_k(w_1, \cdots, w_n)] - \mathbb{E}[f_k(w_1, \cdots, w_n)]\right| = o_n(1).$$

Proof. By Equation (E.42), Equation (E.43), and Lemma 32, we know that there exists a constant C'' that depends only on $(\sigma_\Sigma, \zeta, \Delta, B_0, m_2)$, such that

$$\max\left\{\mathbb{E}[\|\beta_k\|_2^2]^{1/2}, \mathbb{E}[\|\beta_{k,-i}\|_2^2]^{1/2}\right\} \leq C''. \tag{E.45}$$

To show this result, we first prove that $\widehat{\beta}_k \approx \widehat{\beta}_{k,-i}$. By definition,

$$\widehat{\beta}_{k+1} - \widehat{\beta}_{k+1,-i} = \left(I_p - \delta_k \widehat{\Sigma}\right) \cdot \left(\widehat{\beta}_k - \widehat{\beta}_{k,-i}\right) + \frac{\delta_k}{n} y_i x_i - \frac{\delta_k}{n} x_i x_i^\top \widehat{\beta}_{k,-i}.$$

Invoking the triangle and Cauchy-Schwartz inequalities, we conclude that

$$\begin{aligned}
&\|\widehat{\beta}_{k+1} - \widehat{\beta}_{k+1,-i}\|_2^2 \\
&\leq (1 + \delta_k \|\widehat{\Sigma}\|_{\mathrm{op}})^2 \|\widehat{\beta}_k - \widehat{\beta}_{k,-i}\|_2^2 + \frac{\delta_k^2}{n^2}(y_i - x_i^\top \widehat{\beta}_{k,-i})^2 \cdot \|x_i\|_2^2 \\
&\quad + \frac{2\delta_k(1 + \delta_k \|\widehat{\Sigma}\|_{\mathrm{op}})}{n} \cdot \|\widehat{\beta}_k - \widehat{\beta}_{k,-i}\|_2 \cdot |y_i - x_i^\top \widehat{\beta}_{k,-i}| \cdot \|x_i\|_2 \\
&\leq (1 + \delta_k \|\widehat{\Sigma}\|_{\mathrm{op}})(1 + 2\delta_k \|\widehat{\Sigma}\|_{\mathrm{op}}) \|\widehat{\beta}_k - \widehat{\beta}_{k,-i}\|_2^2 + \frac{\delta_k(1 + \delta_k + \delta_k \|\widehat{\Sigma}\|_{\mathrm{op}})}{n^2}(y_i - x_i^\top \widehat{\beta}_{k,-i})^2 \cdot \|x_i\|_2^2.
\end{aligned}$$

By induction,

$$\|\widehat{\beta}_{k+1} - \widehat{\beta}_{k+1,-i}\|_2^2 \leq \sum_{j=1}^{k} \frac{\delta_j \exp(3\Delta \|\widehat{\Sigma}\|_{\mathrm{op}} + \Delta)}{n^2} \cdot (y_i - x_i^\top \widehat{\beta}_{k,-i})^2 \cdot \|x_i\|_2^2.$$

By the Hölder's inequality and Lemma 32, we see that for $n \geq 12\Delta\widetilde{C}_0^2 + 1$

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{\beta}_{k+1} - \widehat{\beta}_{k+1,-i}\|_2^2\right] \\
&\leq \sum_{j=1}^{k} \frac{\delta_j}{n^2} \cdot \mathbb{E}[(y_i - x_i^\top \widehat{\beta}_{k,-i})^4]^{1/2} \cdot \mathbb{E}[\|x_i\|_2^8]^{1/4} \cdot \mathbb{E}[\exp(12\Delta\|\widehat{\Sigma}\|_{\mathrm{op}} + 4\Delta)]^{1/4} \\
&\leq \frac{\Delta e^{\Delta} \sigma_\Sigma \mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)}{n} \cdot \mathcal{E}(\widetilde{C}_0, \zeta, 12\Delta)^{1/4}.
\end{aligned} \tag{E.46}$$

In addition, direct computation gives

$$\begin{aligned}
\mathbb{E}[r_k(w_1, \cdots, w_n)] &= m_2 + \mathbb{E}[\widehat{\beta}_k^\top \Sigma \widehat{\beta}_k] + 2\langle \mathbb{E}[\widehat{\beta}_k], \theta \rangle, \\
\mathbb{E}[f_k(w_1, \cdots, w_n)] &= m_2 + \mathbb{E}[\widehat{\beta}_{k,-i}^\top \Sigma \widehat{\beta}_{k,-i}] + 2\langle \mathbb{E}[\widehat{\beta}_{k,-i}], \theta \rangle,
\end{aligned}$$

where we recall that $\theta = \mathbb{E}[y_0 x_0]$. By Lemma 31 we know that $\|\theta\|_2 \leq (\sigma_\Sigma^{1/2} + 1) m_2^{1/2}$. Therefore,

$$\begin{aligned}
&|\mathbb{E}[r_k(w_1, \cdots, w_n)] - \mathbb{E}[f_k(w_1, \cdots, w_n)]| \\
&\leq 2\|\theta\|_2 \cdot \mathbb{E}\left[\|\widehat{\beta}_{k+1} - \widehat{\beta}_{k+1,-i}\|_2^2\right]^{1/2} + \sigma_\Sigma \mathbb{E}\left[\|\widehat{\beta}_k - \widehat{\beta}_{k,-i}\|_2^2\right]^{1/2} \cdot \left(\mathbb{E}\left[\|\widehat{\beta}_k\|_2^2\right]^{1/2} + \mathbb{E}\left[\|\widehat{\beta}_{k,-i}\|_2^2\right]^{1/2}\right),
\end{aligned}$$

which by Equations (E.45) and (E.46) goes to zero as $n, p \to \infty$. Furthermore, the convergence is uniform for all $k \in \{0\} \cup [K]$. This completes the proof of the lemma. □

S.7 Proof of Lemma 22

Lemma 22 (Gradient upper bound)**.** *There exists a constant $\xi(C_{\Sigma,\zeta}, \Delta, m, B_0) > 0$ that depends only on $(C_{\Sigma,\zeta}, \Delta, m, B_0)$, such that on the set Ω, it holds that*

$$\|\nabla_W f_k(W)\|_F \leq \frac{K\xi(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \log n}{\sqrt{n}}$$

for all $k \in \{0\} \cup [K]$. In the above display, we define $W = (w_1, \cdots, w_n)$ and K, we recall, is the total number of GD iterations.

We divide the proof of the lemma into two parts: upper bounding $\|\nabla_X \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)\|_F$ and $\|\nabla_X \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)\|_F$.

S.7.1 Upper Bounding $\|\nabla_X \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)\|_F$

We start with the most challenging part, namely, upper bounding $\|\nabla_X \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)\|_F$. We will show the following:

Lemma 36 (Bounding norm of gradient with respect to features)**.** *On the set Ω, for all $k \in \{0\} \cup [K]$,*

$$\begin{aligned}
\|\nabla_X \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)\|_F \leq{}& \frac{2B_* \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) \log n}{\sqrt{n}} + \frac{2\Delta K e^{2\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta} B_* \log n}{\sqrt{n}} \cdot \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) \\
&+ \frac{2\Delta K e^{2\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}^{1/2} \bar{B}_* \log n}{\sqrt{n}} \cdot \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0).
\end{aligned}$$

In the above equation, we recall that B_* is defined in Equation (E.36), $\bar{B}_*$ is defined in Equation (E.37), and $\mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0)$ is defined in Equation (E.39).

Proof. We prove Lemma 36 in the remainder of this section. For $s \in [n]$ and $k \in [K]$, we can compute $\nabla_{x_s} \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$, which takes the following form:

$$\nabla_{x_s} \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k) = -\frac{2}{n}(y_s - x_s^\top \widehat{\beta}_{k,-s}) \widehat{\beta}_{k,-s}^\top - \frac{2}{n} \sum_{i=1}^{n} (y_i - x_i^\top \widehat{\beta}_{k,-i}) x_i^\top \nabla_{x_s} \widehat{\beta}_{k,-i}. \tag{E.47}$$

The above formula suggests that we should analyze the Jacobian matrix $\nabla_{x_s}\widehat{\beta}_{k,-i}$, which can be done recursively. More precisely, the following update rule is a direct consequence of the gradient descent update rule:

$$\begin{aligned}
&\nabla_{x_s}\widehat{\beta}_{k+1,-i} \\
&= \nabla_{x_s}\widehat{\beta}_{k,-i} + \frac{\delta_k \mathbb{1}\{i \neq s\}}{n}(y_s - x_s^\top \widehat{\beta}_{k,-i})I_{p+1} - \frac{\delta_k \mathbb{1}\{i \neq s\}}{n} x_s \widehat{\beta}_{k,-i}^\top - \frac{\delta_k}{n}\sum_{j\neq i} x_j x_j^\top \nabla_{x_s}\widehat{\beta}_{k,-i} \\
&= \Big(I_{p+1} - \delta_k \widehat{\Sigma}\Big)\cdot \nabla_{x_s}\widehat{\beta}_{k,-i} + \frac{\delta_k}{n} x_i x_i^\top \nabla_{x_s}\widehat{\beta}_{k,-i} + \mathbb{1}\{i \neq s\}\cdot\left\{\frac{\delta_k}{n}(y_s - x_s^\top \widehat{\beta}_{k,-i})I_{p+1} - \frac{\delta_k}{n} x_s \widehat{\beta}_{k,-i}^\top\right\}.
\end{aligned}$$

Note that the above process is initialized at $\nabla_{x_s}\widehat{\beta}_{0,-i} = 0_{(p+1)\times(p+1)}$. Clearly when $i = s$, the Jacobian $\nabla_{x_s}\widehat{\beta}_{k,-i}$ remains zero for all k that is concerned, and we automatically get an upper bound for $\|\nabla_{x_s}\widehat{\beta}_{k,-i}\|_2$.

In what follows, we focus on the non-trivial case $i \neq s$. For this part, we will mostly fix i and s, and ignore the dependency on (i, s) when there is no confusion. Note that we can reformulate the Jacobian update rule as follows:

$$\nabla_{x_s}\widehat{\beta}_{k+1,-i} = M_k \nabla_{x_s}\widehat{\beta}_{k,-i} + M_{k,i}\nabla_{x_s}\widehat{\beta}_{k,-i} + \frac{\delta_k}{n}(y_s - x_s^\top \widehat{\beta}_{k,-i})I_{p+1} - \frac{\delta_k}{n} x_s \widehat{\beta}_{k,-i}^\top, \tag{E.48}$$

where $M_k = I_{p+1} - \delta_k \widehat{\Sigma}$ and $M_{k,i} = \delta_k x_i x_i^\top / n$. By induction, it is not hard to see that for all $0 \le k \le K-1$, the matrix $\nabla_{x_s}\widehat{\beta}_{k+1,-i} - R_0^{(k)}$ can be expressed as the sum of terms that take the form

$$\left(\prod_{j=1}^{k-k'} R_{k+1-j}\right) R_0^{(k')},$$

where $k' \in \{0\} \cup [k-1]$, $R_0^{(k')} = \frac{\delta_{k'}}{n}(y_s - x_s^\top \widehat{\beta}_{k',-i})I_{p+1} - \frac{\delta_{k'}}{n} x_s \widehat{\beta}_{k',-i}^\top$, and R_j is either M_j or $M_{j,i}$.

To put it formally, we summarize this result as the following lemma:

Lemma 37. *For $i, s \in [n]$ with $i \neq s$ and all $k \in \{0\} \cup [K-1]$, it holds that*

$$x_i^\top \nabla_{x_s}\widehat{\beta}_{k+1,-i} = \sum_{k'=0}^{k}\sum_{r=k'}^{k} c_{i,k,k',r}\, x_i^\top H_{k',r}\cdot\left(\frac{\delta_{k'}}{n}(y_s - x_s^\top \widehat{\beta}_{k',-i})I_{p+1} - \frac{\delta_{k'}}{n} x_s \widehat{\beta}_{k',-i}^\top\right),$$

where $c_{i,k,k',r} \in \mathbb{R}$ and $H_{k',r} = \prod_{j=k'+1}^{r} M_{k'+1+r-j}$. We adopt the convention that $H_{k',k'} = I_{p+1}$. Furthermore, on the set Ω, it holds that

$$\|H_{k',r}\|_{\mathrm{op}} \le e^{\Delta C_{\Sigma,\zeta}}, \qquad \|c_{i,k,k',r} H_{k',r}\|_{\mathrm{op}} \le e^{2\Delta C_{\Sigma,\zeta}}. \tag{E.49}$$

Proof of Lemma 37. To derive the first inequality in Equation (E.49), we simply notice that

$$\|H_{k',r}\|_{\mathrm{op}} \le \prod_{j=k'+1}^{r} \|M_{k'+1+r-j}\|_{\mathrm{op}} \le \prod_{j=k'+1}^{r}(1 + \delta_{k'+1+r-j} C_{\Sigma,\zeta}) \le e^{\Delta C_{\Sigma,\zeta}}.$$

We next prove the second inequality in Equation (E.49). As discussed before, $x_i^\top \nabla_{x_s}\beta_{k+1,-i} - x_i^\top R_0^{(k)}$ can be expressed as the sum of terms that take the form

$$x_i^\top \left(\prod_{j=1}^{k-k'} R_{k+1-j}\right) R_0^{(k')},$$

with k' ranging from 0 to $k-1$. The subtracting $x_i^\top R_0^{(k)}$ part implies that we should set $c_{i,k,k,k} = 1$ and $H_{k,k} = I_{p+1}$.

We then study $c_{i,k,k',r}$ in general. For this purpose, we analyze each summand. Without loss, we let R_{j_*} be the last matrix in the sequence $(R_{k+1-j})_{j=1}^{k-k'}$ that takes the form $M_{j_*,i}$. Then

$$x_i^\top \left(\prod_{j=1}^{k-k'} R_{k+1-j}\right) R_0^{(k')} = x_i^\top \left(\prod_{j=1}^{k-j_*} R_{k+1-j}\right) \cdot \frac{\delta_{j_*}}{n} x_i x_i^\top \cdot \left(\prod_{j=k-j_*+2}^{k-k'} R_{k+1-j}\right) R_0^{(k')}$$
$$= \frac{\delta_{j_*}}{n} x_i^\top \left(\prod_{j=1}^{k-j_*} R_{k+1-j}\right) x_i x_i^\top H_{k',j_*-1} R_0^{(k')}.$$

This implies that

$$c_{i,k,k',j_*-1} = \sum_{R_{k+1-j}\in\{M_{k+1-j}, M_{k+1-j,i}\}, 1\le j\le k-j_*} \frac{\delta_{j_*}}{n} x_i^\top \left(\prod_{j=1}^{k-j_*} R_{k+1-j}\right) x_i,$$

which further tells us

$$\begin{aligned}
&\|c_{i,k,k',j_*-1} H_{k',j_*-1}\|_{\mathrm{op}} \\
&= \left\| \sum_{R_{k+1-j}\in\{M_{k+1-j}, M_{k+1-j,i}\}, 1\le j\le k-j_*} \frac{\delta_{j_*}}{n} x_i^\top \left(\prod_{j=1}^{k-j_*} R_{k+1-j}\right) x_i \cdot H_{k',j_*-1} \right\|_{\mathrm{op}} \\
&\le \prod_{k=0}^{K-1} (1 + \|M_k\|_{\mathrm{op}} + \|M_{k,i}\|_{\mathrm{op}}) \\
&\le \prod_{k=0}^{K-1} (1 + \delta_k C_{\Sigma,\zeta} + \delta_k C_{\Sigma,\zeta}) \le e^{2\Delta C_{\Sigma,\zeta}}.
\end{aligned}$$

This completes the proof. □

As a consequence of Lemma 37, we can write

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^n (y_i - x_i^\top \widehat{\beta}_{k+1,-i}) x_i^\top \nabla_{x_s} \widehat{\beta}_{k+1,-i} \\
&= \frac{2}{n}\sum_{i=1}^n \sum_{k'=0}^k \sum_{r=k'}^k c_{i,k,k',r} (y_i - x_i^\top \widehat{\beta}_{k+1,-i}) x_i^\top H_{k',r} \cdot \left(\frac{\delta_{k'}}{n}(y_s - x_s^\top \widehat{\beta}_{k',-i}) I_{p+1} - \frac{\delta_{k'}}{n} x_s \widehat{\beta}_{k',-i}^\top\right) \\
&= \sum_{k'=0}^k \sum_{r=k'}^k \left(g_{k,k',r,s} + \bar{g}_{k,k',r,s}\right),
\end{aligned}$$

where we define

$$g_{k,k',r,s} = \frac{2\delta_{k'}}{n^2} \sum_{i=1}^n c_{i,k,k',r} (y_i - x_i^\top \widehat{\beta}_{k+1,-i})(y_s - x_s^\top \widehat{\beta}_{k',-i}) x_i^\top H_{k',r},$$
$$\bar{g}_{k,k',r,s} = -\frac{2\delta_{k'}}{n^2} \sum_{i=1}^n c_{i,k,k',r} (y_i - x_i^\top \widehat{\beta}_{k+1,-i}) x_i^\top H_{k',r} x_s \widehat{\beta}_{k',-i}^\top.$$

We define $V_{k,k',r}, \bar{V}_{k,k',r} \in \mathbb{R}^{(p+1)\times n}$ such that the s-th columns correspond to $g_{k,k',r,s}^\top$ and $\bar{g}_{k,k',r,s}^\top$, respectively. We also define $\widetilde{V}_k \in \mathbb{R}^{(p+1)\times n}$ such that the s-th column of this matrix corresponds to $2(y_s - x_s^\top \widehat{\beta}_{k+1,-s})\widehat{\beta}_{k+1,-s}/n$. Inspecting Equation (E.47), we see that to upper bound the Frobenius norm of $\nabla_X \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_{k+1})$, it suffices to upper bound the Frobenius norms of matrices $V_{k,k',r}, \bar{V}_{k,k',r}$, and $\widetilde{V}_k$, which we analyze in the lemma below.

Lemma 38. *On the set Ω, we have*

$$\|V_{k,k',r}\|_F^2 \le \frac{4\delta_{k'}^2 e^{4\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}}{n} \cdot \bar{B}_*^2 \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)^2 \cdot (\log n)^2, \tag{E.50}$$

$$\|\bar{V}_{k,k',r}\|_F^2 \le \frac{4\delta_{k'}^2 e^{4\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}^2 B_*^2}{n} \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)^2 \cdot (\log n)^2, \tag{E.51}$$

$$\|\widetilde{V}_k\|_F^2 \le \frac{4B_*^2}{n} \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)^2 \cdot (\log n)^2. \tag{E.52}$$

Proof of Lemma 38. We observe that

$$V_{k,k',r} = \frac{2\delta_{k'}}{n^2} H_{k',r} X^\top A_{k,k',r},$$

where $A_{k,k',r} \in \mathbb{R}^{n\times n}$, and $(A_{k,k',r})_{is} = c_{i,k,k',r}(y_i - x_i^\top \widehat{\beta}_{k+1,-i})(y_s - x_s^\top \widehat{\beta}_{k',-i})$. Note that on Ω, by Corollary 34 and Lemma 35, we have

$$\frac{1}{n}\|y - X\widehat{\beta}_{k',-i}\|_2^2 \le \bar{B}_*^2 \cdot \log n,$$
$$\frac{1}{n}\|a_{k+1}\|_2^2 \le \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)^2 \cdot \log n.$$

This further implies that

$$\|A_{k,k',r}\|_F^2 \le n^2 \sup_{i\in[n]} |c_{i,k,k',r}|^2 \cdot \bar{B}_*^2 \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)^2 \cdot (\log n)^2.$$

As a result,

$$\begin{aligned}\|V_{k,k',r}\|_F^2 \le& \frac{4\delta_{k'}^2}{n^4} \cdot \|H_{k',r}\|_{\mathrm{op}}^2 \cdot \|X\|_{\mathrm{op}}^2 \cdot \|A_{k,k',r}\|_F^2 \\ \le& \frac{4\delta_{k'}^2 e^{4\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}}{n} \cdot \bar{B}_*^2 \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)^2 \cdot (\log n)^2,\end{aligned}$$

which concludes the proof for the first inequality.

We then consider upper bounding $\|\bar{V}_{k,k',r}\|_F$. Note that

$$\bar{V}_{k,k',r} = -\frac{2\delta_{k'}}{n^2} Q_{k,k',r} X H_{k',r} X^\top,$$
$$Q_{k,k',r} = \left[\widehat{\beta}_{k,-1} \mid \cdots \mid \widehat{\beta}_{k,-n}\right] \cdot \mathrm{diag}\{(c_{i,k,k',r}(y_i - x_i^\top \widehat{\beta}_{k+1,-i}))_{i=1}^n\} \in \mathbb{R}^{(p+1)\times n}.$$

Therefore,

$$\begin{aligned}\|\bar{V}_{k,k',r}\|_F^2 \le& \frac{4\delta_{k'}^2}{n^4} \cdot \|Q_{k,k',r}\|_F^2 \cdot \|XX^\top\|_{\mathrm{op}}^2 \cdot \|H_{k',r}\|_{\mathrm{op}}^2 \\ \le& \frac{4\delta_{k'}^2 e^{4\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}^2 B_*^2}{n} \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)^2 \cdot (\log n)^2.\end{aligned}$$

This completes the proof of Equation (E.51). Finally, we prove Equation (E.52). By Lemma 33 and Lemma 35, we obtain

$$\|\widetilde{V}_k\|_F^2 \le \frac{4B_*^2 (\log n)^2}{n} \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)^2 \cdot (\log n)^2.$$

This is exactly what we aim to prove. □

By triangle inequality,

$$\|\nabla_X \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)\|_F \le \|\widetilde{V}_k\|_F + \sum_{k'=0}^{k}\sum_{r=k'}^{k} \left(\|\bar{V}_{k,k',r}\|_F + \|V_{k,k',r}\|_F\right).$$

The proof of Lemma 36 now follows by putting together the above upper bound and Lemma 38. □

S.7.2 Upper Bounding $\nabla_y \widehat{R}^{\text{loo}}(\widehat{\beta}_k)$

Next, we upper bound the Euclidean norm of $\nabla_y \widehat{R}^{\text{loo}}(\widehat{\beta}_k)$. This part is in spirit similar to the upper bounding of the Euclidean norm of $\nabla_X \widehat{R}^{\text{loo}}(\widehat{\beta}_k)$ that we discussed in the previous section.

More precisely, we will show the following:

Lemma 39 (Bounding norm of gradient with respect to response)**.** *On the set* Ω*,*

$$\begin{aligned}&\|\nabla_y \widehat{R}^{\text{loo}}(\widehat{\beta}_k)\|_2 \\ &\leq \frac{2\mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0)}{\sqrt{n}} \cdot \sqrt{\log n} + \frac{2\Delta K C_{\Sigma,\zeta} e^{2\Delta C_{\Sigma,\zeta}}}{\sqrt{n}} \cdot \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \sqrt{\log n}.\end{aligned} \tag{E.53}$$

Proof. For $s \in [n]$, we note that

$$\frac{\partial}{\partial y_s} \widehat{R}^{\text{loo}}(\widehat{\beta}_k) = \frac{2}{n}(y_s - x_s^\top \widehat{\beta}_{k,-s}) - \frac{2}{n}\sum_{i=1}^n (y_i - x_i^\top \widehat{\beta}_{k,-i}) x_i^\top \frac{\partial}{\partial y_s}\widehat{\beta}_{k,-i}. \tag{E.54}$$

If $i = s$, then $\frac{\partial}{\partial y_s}\widehat{\beta}_{k,-i} = 0$ for all $k \in \{0\} \cup [K]$. Moving forward, we focus on the more interesting case $i \neq s$. We also have

$$\begin{aligned}\frac{\partial}{\partial y_s}\widehat{\beta}_{k+1,-i} =& \frac{\partial}{\partial y_s}\widehat{\beta}_{k,-i} + \frac{\delta_k}{n} x_s - \frac{\delta_k}{n}\sum_{j\neq i} x_j x_j^\top \frac{\partial}{\partial y_s}\widehat{\beta}_{k,-i} \\ =& M_k \frac{\partial}{\partial y_s}\widehat{\beta}_{k,-i} + M_{k,i}\frac{\partial}{\partial y_s}\widehat{\beta}_{k,-i} + \frac{\delta_k}{n} x_s,\end{aligned}$$

where we recall that $M_k = (I_{p+1} - \delta_k \widehat{\Sigma})$ and $M_{k,i} = \delta_k x_i x_i^\top / n$. Invoking the same argument that we employed to derive Lemma 37, we can conclude that

$$\frac{\partial}{\partial y_s} x_i^\top \widehat{\beta}_{k+1,-i} = \sum_{k'=0}^k \sum_{r=k'}^k c_{i,k,k',r} x_i^\top H_{k',r} \cdot \frac{\delta_{k'}}{n} x_s.$$

Plugging this into Equation (E.54) leads to the following equality:

$$\frac{\partial}{\partial y_s}\widehat{R}^{\text{loo}}(\widehat{\beta}_{k+1}) = \frac{2}{n}(y_s - x_s^\top \widehat{\beta}_{k+1,-s}) - \sum_{k'=0}^k \sum_{r=k'}^k \eta_{k,k',r,s},$$

where

$$\eta_{k,k',r,s} = \frac{2}{n}\sum_{i=1}^n (y_i - x_i^\top \widehat{\beta}_{k+1,-i}) c_{i,k,k',r} x_i^\top H_{k',r} \cdot \frac{\delta_{k'}}{n} x_s.$$

We define $\eta_{k,k',r} = (\eta_{k,k',r,s})_{s=1}^n \in \mathbb{R}^n$. It then holds that

$$\begin{aligned}\eta_{k,k',r} &= \frac{2\delta_{k'}}{n^2} X H_{k',r} X^\top q_{k,k',r}, \\ q_{k,k',r} &= \left(c_{i,k,k',r}(y_i - x_i^\top \widehat{\beta}_{k+1,-i})\right)_{i=1}^n \in \mathbb{R}^n.\end{aligned}$$

We can upper bound the Euclidean norm of $\eta_{k,k',r}$ using Lemma 35 and 37. More precisely,

$$\|\eta_{k,k',r}\|_2 \leq \frac{2\delta_{k'}}{n^2}\|X^\top X\|_{\text{op}} \cdot \|H_{k',r}\|_{\text{op}} \cdot \|q_{k,k',r}\|_2 \leq \frac{2\delta_{k'} C_{\Sigma,\zeta} e^{2\Delta C_{\Sigma,\zeta}}}{\sqrt{n}} \cdot \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \sqrt{\log n}.$$

Note that

$$\nabla_y \widehat{R}^{\text{loo}}(\widehat{\beta}_k) = \frac{2}{n} a_k - \sum_{k'=0}^k \sum_{r=k'}^k \eta_{k,k',r}.$$

Invoking triangle inequality and Lemma 35, we obtain

$$\begin{aligned}&\|\nabla_y \widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)\|_2 \\ \leq &\frac{2\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)}{\sqrt{n}}\cdot\sqrt{\log n}+\frac{2\Delta K C_{\Sigma,\zeta}e^{2\Delta C_{\Sigma,\zeta}}}{\sqrt{n}}\cdot\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\sqrt{\log n}.\end{aligned} \tag{E.55}$$

This completes the proof. □

S.8 Proof of Theorem 3

Theorem 3 (Functional consistency of LOOCV). *Under the conditions of Theorem 2, suppose that $\psi:\mathbb{R}^2\to\mathbb{R}$ is differentiable and satisfies $\|\nabla\psi(u)\|_2\leq C_\psi\|u\|_2+\bar{C}_\psi$ for all $u\in\mathbb{R}^2$ and for constants $C_\psi,\bar{C}_\psi\geq 0$. Then, as $n,p\to\infty$,*

$$\max_{k\in[K]}\left|\widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_k)-\Psi(\widehat{\beta}_k)\right|\xrightarrow{\text{a.s.}}0. \tag{11}$$

where we recall that $\widehat{R}^{\mathrm{loo}}(\widehat{\beta}_k)$ and $R(\widehat{\beta}_k)$ are as defined in (10) and (9), respectively.

As consequence of (11), LOOCV can be used to tune early stopping. Specifically, if we define $k_*=\arg\min_{k\in[K]}\widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_k)$, then as $n,p\to\infty$,

$$\left|\Psi(\widehat{\beta}_{k_*})-\min_{k\in[K]}\Psi(\widehat{\beta}_k)\right|\xrightarrow{\text{a.s.}}0. \tag{12}$$

Once again, we will work on the set Ω, which we recall is defined in Equation (E.30). The proof idea is similar to that for the squared loss. More precisely, if we can prove Equations (E.59) to (E.61) listed below, then once again can add up the probabilities and show that the sum is finite. Next, we just apply the first Borel–Cantelli lemma, which leads to the following uniform consistency result:

$$\sup_{k\in\{0\}\cup[K]}\left|\widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_k)-\Psi(\widehat{\beta}_k)\right|\xrightarrow{\text{a.s.}}0.$$

S.8.1 Concentration Analysis

As before, we will first prove that both $\widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_{k+1})$ and $\Psi(\widehat{\beta}_{k+1})$ concentrate. To this end, we shall again analyze the gradients and show that they are Lipschitz functions of the input data. The proof for this part is similar to the proof of Lemmas 23 and 24.

We define

- $f^\psi_{k+1}(w_1,\cdots,w_n)=\widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_{k+1})$
- $\widetilde{f}^\psi_{k+1}=f^\psi_{k+1}\circ h$
- $r^\psi_k(w_1,\cdots,w_n)=\Psi(\widehat{\beta}_k)$
- $\widetilde{r}^\psi_k=r_k\circ h$

Our formal statement then is as follows.

Lemma 40 (LOO and risk concentration analysis). *Under the assumptions of Theorem 3, with probability at least $1-2(n+p)^{-4}-(n\log^2 n)^{-1}m_4-2(K+1)C_{\mathsf{T}_2}n^{-2}$, for all $k\in\{0\}\cup[K]$*

$$\begin{aligned}\left|\widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_k)-\mathbb{E}[\widetilde{f}^\psi_k(w_1,\cdots,w_n)]\right|&\leq\frac{2\sigma_{\mathsf{T}_2}LK\xi^\psi(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot(\log n)^{3/2}}{\sqrt{n}},\\ \left|\Psi(\widehat{\beta}_k)-\mathbb{E}[\widetilde{r}^\psi_k(w_1,\cdots,w_n)]\right|&\leq\frac{2\sigma_{\mathsf{T}_2}L\bar{\xi}^\psi(C_{\Sigma,\zeta},\Delta,m,B_0)(\log n)^{3/2}}{\sqrt{n}}.\end{aligned}$$

In the above display, $\xi^\psi(C_{\Sigma,\zeta},\Delta,m,B_0)$ and $\bar{\xi}^\psi(C_{\Sigma,\zeta},\Delta,m,B_0)$ are positive constants that depend only on $(C_{\Sigma,\zeta},\Delta,m,B_0)$.

Proof. We start by writing down the gradient. For all $s \in [n]$, note that

$$\nabla_{x_s} \widehat{\Psi}^{\text{loo}}(\widehat{\beta}_{k+1}) = -\frac{1}{n}\partial_2\psi(y_s, x_s^\top \widehat{\beta}_{k+1,-s})\widehat{\beta}_{k+1,-s}^\top - \frac{1}{n}\sum_{i=1}^n \partial_2\psi(y_i, x_i^\top \widehat{\beta}_{k+1,-i}) x_i^\top \nabla_{x_s}\widehat{\beta}_{k+1,-i},$$

where ∂_i stands for taking the partial derivative with respect to the i-th input. Here, $i \in \{1, 2\}$. By Lemma 37, on Ω we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \partial_2\psi(y_i, x_i^\top \widehat{\beta}_{k+1,-i}) x_i^\top \nabla_{x_s}\widehat{\beta}_{k+1,-i} \\
&= \sum_{k'=0}^k \sum_{r=k'}^k c_{i,k,k',r} \frac{1}{n}\sum_{i=1}^n \partial_2\psi(y_i, x_i^\top \widehat{\beta}_{k+1,-i}) x_i^\top H_{k',r} \cdot \left(\frac{\delta_{k'}}{n}(y_s - x_s^\top \widehat{\beta}_{k',-i}) I_{p+1} - \frac{\delta_{k'}}{n} x_s \widehat{\beta}_{k',-i}^\top\right) \\
&= \sum_{k'=0}^k \sum_{r=k'}^k \left(g^\psi_{k,k',r,s} + \bar{g}^\psi_{k,k',r,s}\right),
\end{aligned}$$

where

$$\begin{aligned}
g^\psi_{k,k',r,s} &= \frac{\delta_{k'}}{n^2}\sum_{i=1}^n c_{i,k,k',r}\partial_2\psi(y_i, x_i^\top \widehat{\beta}_{k+1,-i})(y_s - x_s^\top \widehat{\beta}_{k',-i}) x_i^\top H_{k',r}, \\
\bar{g}^\psi_{k,k',r,s} &= -\frac{\delta_{k'}}{n^2}\sum_{i=1}^n c_{i,k,k',r}\partial_2\psi(y_i, x_i^\top \widehat{\beta}_{k+1,-i}) x_i^\top H_{k',r} x_s \widehat{\beta}_{k',-i}^\top.
\end{aligned}$$

We let $V^\psi_{k,k',r}, \bar{V}^\psi_{k,k',r} \in \mathbb{R}^{(p+1)\times n}$, such that the s-th columns are set to be $(g^\psi_{k,k',r,s})^\top$ and $(\bar{g}^\psi_{k,k',r,s})^\top$, respectively. We also define $\widetilde{V}^\psi_k \in \mathbb{R}^{(p+1)\times n}$ such that the s-th column of this matrix corresponds to $\partial_2\psi(y_s, x_s^\top \widehat{\beta}_{k+1,-s})\widehat{\beta}_{k+1,-s}/n$. Using triangle inequality, we immediately obtain that

$$\|\nabla_X \widehat{\Psi}^{\text{loo}}(\widehat{\beta}_{k+1})\|_F \le \|\widetilde{V}^\psi_k\|_F + \sum_{k'=0}^k \sum_{r=k'}^k \left\{\|V^\psi_{k,k',r}\|_F + \|\bar{V}^\psi_{k,k',r}\|_F\right\}. \tag{E.56}$$

Next, we upper bound $\|V^\psi_{k,k',r}\|_F$, $\|\bar{V}^\psi_{k,k',r}\|_F$, and $\|\widetilde{V}^\psi_k\|_F$. We observe that

$$V^\psi_{k,k',r} = \frac{\delta_{k'}}{n^2} H_{k',r} X^\top A^\psi_{k,k',r}, \qquad \bar{V}^\psi_{k,k',r} = -\frac{\delta_{k'}}{n^2} Q^\psi_{k,k',r} X H_{k',r} X^\top,$$

where

$$\begin{aligned}
Q^\psi_{k,k',r} &= [\beta_{k,-1} \mid \cdots \mid \beta_{k,-n}] \cdot \text{diag}\{(c_{i,k,k',r}\partial_2\psi(y_i, x_i^\top \widehat{\beta}_{k+1,-i}))_{i=1}^n\} \in \mathbb{R}^{(p+1)\times n}, \\
(A^\psi_{k,k',r})_{is} &= c_{i,k,k',r}\partial_2\psi(y_i, x_i^\top \widehat{\beta}_{k+1,-i})(y_s - x_s^\top \widehat{\beta}_{k',-i}).
\end{aligned}$$

We let $a^\psi_{k+1} = (\partial_2\psi(y_i, x_i^\top \widehat{\beta}_{k+1,-i}))_{i=1}^n$. Recall that $a_{k+1} = (y_i - x_i^\top \beta_{k+1,-i})_{i=1}^n$. Using triangle inequality, we obtain that

$$\|a^\psi_{k+1}\|_2 \le 3C_\psi(\|a_{k+1}\|_2 + \|y\|_2) + \sqrt{2n}\bar{C}_\psi.$$

Invoking Lemma 35, we know that on Ω, $\|a_{k+1}\|_2 \le \sqrt{n}\mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \sqrt{\log n}$. Furthermore, by definition we know that on Ω, $\|y\|_2 \le \sqrt{n(m + \log n)}$. By Corollary 34 we see that $\|y - X\beta_{k,-i}\|_2 \le \sqrt{n}\bar{B}_* \cdot \sqrt{\log n}$. By Lemma 37, we have $\|c_{i,k,k',r}H_{k',r}\|_{\text{op}} \le e^{2\Delta C_{\Sigma,\zeta}}$. Putting together all these results, we conclude that

$$\begin{aligned}
\|V^\psi_{k,k',r}\|_F &\le \frac{\delta_{k'}}{n^2} \cdot \|H_{k',r}\|_{\text{op}} \cdot \|X\|_{\text{op}} \cdot \|A^\psi_{k,k',r}\|_F \\
&\le \frac{\delta_{k'} e^{2\Delta C_{\Sigma,\zeta}} C^{1/2}_{\Sigma,\zeta} \bar{B}_* \cdot (3C_\psi \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) + 3C_\psi\sqrt{m} + \sqrt{2}\bar{C}_\psi) \cdot \log n}{\sqrt{n}}.
\end{aligned} \tag{E.57}$$

Applying Lemma 33, we deduce that

$$\|Q^{\psi}_{k,k',r}\|_F \le \sqrt{n}B_* \sup_{i\in[n]} |c_{i,k,k',r}| \cdot (3C_\psi \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0) + 3C_\psi m^{1/2} + \sqrt{2}\bar{C}_\psi)\cdot \log n.$$

Therefore,

$$\begin{aligned}
\|\bar{V}^{\psi}_{k,k',r}\|_F &\le \frac{\delta_{k'}}{n^2}\cdot \|Q^{\psi}_{k,k',r}\| \cdot \|X\|^2_{\mathrm{op}} \cdot \|H_{k',r}\|_{\mathrm{op}} \\
&\le \frac{\delta_{k'} B_* e^{2\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}\cdot (3C_\psi \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0) + 3C_\psi\sqrt{m} + \sqrt{2}\bar{C}_\psi)\cdot \log n}{\sqrt{n}}, \\
\|\widetilde{V}^{\psi}_{k}\|_F &\le \frac{1}{n}\|a^{\psi}_{k+1}\|_2 \cdot \|\widehat{\beta}_{k+1,-s}\|_2 \le \frac{B_*(3C_\psi \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0) + 3C_\psi\sqrt{m} + \sqrt{2}\bar{C}_\psi)\cdot \log n}{\sqrt{n}}.
\end{aligned} \tag{E.58}$$

Combining Equations (E.56) to (E.58), we see that there exists a constant $\xi^{\psi}_1(C_{\Sigma,\zeta},\Delta,m,B_0)$ that depends only on $(C_{\Sigma,\zeta},\Delta,m,B_0)$, such that on Ω, for all $k\in\{0\}\cup[K]$ we have

$$\|\nabla_X \widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_{k+1})\|_F \le \frac{K\xi^{\psi}_1(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n}{\sqrt{n}}.$$

Analogously, we can conclude the existence of a non-negative constant $\xi^{\psi}_2(C_{\Sigma,\zeta},\Delta,m,B_0)$, such that on Ω, it holds that

$$\|\nabla_y \widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_{k+1})\|_F \le \frac{K\xi^{\psi}_2(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n}{\sqrt{n}}.$$

Hence, we know that

$$\|\nabla_W \widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_{k+1})\|_F \le K\xi^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n$$

if we set $\xi^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0) = \xi^{\psi}_1(C_{\Sigma,\zeta},\Delta,m,B_0) + \xi^{\psi}_2(C_{\Sigma,\zeta},\Delta,m,B_0)$. Following the same steps that we used to derive Lemma 23, we deduce that with probability at least $1-2(n+p)^{-4} - (n\log^2 n)^{-1}m_4 - 2(K+1)C_{\mathsf{T}_2}n^{-2}$,

$$\left|\widehat{\Psi}^{\mathrm{loo}}(\widehat{\beta}_k) - \mathbb{E}[\widetilde{f}^{\psi}_k(w_1,\cdots,w_n)]\right| \le \frac{2\sigma LK\xi^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot(\log n)^{3/2}}{\sqrt{n}}. \tag{E.59}$$

Similarly, we can prove that with probability at least $1-2(n+p)^{-4} - (n\log^2 n)^{-1}m_4 - 2(K+1)C_{\mathsf{T}_2}n^{-2}$, for all $k\in\{0\}\cup[K]$,

$$\left|\Psi(\widehat{\beta}_k) - \mathbb{E}[\widetilde{r}^{\psi}_k(w_1,\cdots,w_n)]\right| \le \frac{2\sigma L\bar{\xi}^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)(\log n)^{3/2}}{\sqrt{n}}. \tag{E.60}$$

for some constant $\bar{\xi}^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)$ that depends only on $(C_{\Sigma,\zeta},\Delta,m,B_0)$. □

S.8.2 Uniform Consistency

Next, we shall prove that projection has little effect on the expected risk.

Lemma 41 (LOO and risk bias analysis)**.** *On the set Ω, it holds that*

$$\begin{aligned}
\sup_{k\in\{0\}\cup[K]} \left|\mathbb{E}[\widetilde{r}^{\psi}_k(w_1,\cdots,w_n)] - \mathbb{E}[r^{\psi}_k(w_1,\cdots,w_n)]\right| &= o_n(1), \\
\sup_{k\in\{0\}\cup[K]} \left|\mathbb{E}[\widetilde{f}^{\psi}_k(w_1,\cdots,w_n)] - \mathbb{E}[f^{\psi}_k(w_1,\cdots,w_n)]\right| &= o_n(1).
\end{aligned} \tag{E.61}$$

Proof. Using the Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
\left|\mathbb{E}[r^{\psi}_k(w_1,\cdots,w_n)] - \mathbb{E}[\widetilde{r}^{\psi}_k(w_1,\cdots,w_n)]\right| &\le \mathbb{P}(\Omega^c)^{1/2}\mathbb{E}[\psi(y_0, x_0^\top\widehat{\beta}_k)^2]^{1/2}, \\
\left|\mathbb{E}[f^{\psi}_k(w_1,\cdots,w_n)] - \mathbb{E}[\widetilde{f}^{\psi}_k(w_1,\cdots,w_n)]\right| &\le \mathbb{P}(\Omega^c)^{1/2}\mathbb{E}[\psi(y_1, x_1^\top\widehat{\beta}_{k,-1})^2]^{1/2}.
\end{aligned} \tag{E.62}$$

Since $\|\nabla\psi(x)\|_2 \le C_\psi\|x\|_2 + \bar{C}_\psi$, we are able to conclude that there exist constants $\phi_\psi, \bar{\phi}_\psi$ that depend only on $\psi(\cdot)$, such that $\|\psi(x)\|_2^2 \le \phi_\psi\|x\|_2^4 + \bar{\phi}_\psi$ for all $x \in \mathbb{R}^2$. Putting this and Equation (E.44) together, we obtain that

$$\mathbb{E}[\psi(y_1, x_1^\top \widehat{\beta}_{k,-1})^2] \le \phi_\psi \mathbb{E}[\|(y_1, x_1^\top \widehat{\beta}_{k,-1})\|_2^4] + \bar{\phi}_\psi \le \phi_\psi \mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)^2 + \bar{\phi}_\psi. \tag{E.63}$$

Recall that $\mathbb{P}(\Omega^c) \le 2(n+p)^{-4} + n^{-1}m_4$. Combining this, Equations (E.62) and (E.63), we can establish Equation (E.61).

To derive uniform consistency, we also need to show that the expected prediction risk is robust to the sample size. Namely, we will prove $\mathbb{E}[\Psi(\widehat{\beta}_k)] \approx \mathbb{E}[\Psi(\widehat{\beta}_{k,-1})]$.

Since $\|\nabla\psi(x)\|_2 \le C_\psi\|x\|_2 + \bar{C}_\psi$, we see that there exist constants $\varphi_\psi \in \mathbb{R}$, such that for all $x, y \in \mathbb{R}^2$,

$$|\psi(x) - \psi(y)| \le \varphi_\psi \|x - y\|_2 \cdot (1 + \|x\|_2^2 + \|y\|_2^2).$$

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}[\Psi(\widehat{\beta}_k)] - \mathbb{E}[\Psi(\widehat{\beta}_{k,-1})]\right| \\
&= \left|\mathbb{E}[r_k^\psi(w_1, \cdots, w_n)] - \mathbb{E}[f_k^\psi(w_1, \cdots, w_n)]\right| \\
&= \left|\mathbb{E}[\psi(y_0, x_0^\top \widehat{\beta}_k)] - \mathbb{E}[\psi(y_0, x_0^\top \widehat{\beta}_{k,-1})]\right| \\
&\le \varphi_\psi \mathbb{E}\left[\left(1 + \|(y_0, x_0^\top \widehat{\beta}_k)\|_2^2 + \|(y_0, x_0^\top \widehat{\beta}_{k,-1})\|_2^2\right) \cdot |x_0^\top(\widehat{\beta}_k - \widehat{\beta}_{k,-1})|\right] \\
&\le 3\varphi_\psi \mathbb{E}\left[(x_0^\top(\widehat{\beta}_k - \widehat{\beta}_{k,-1}))^2\right]^{1/2} \cdot \mathbb{E}\left[1 + \|(y_0, x_0^\top \widehat{\beta}_k)\|_2^4 + \|(y_0, x_0^\top \widehat{\beta}_{k,-1})\|_2^4\right] \\
&\le 3\varphi_\psi(\sigma_\Sigma + 1)^{1/2} \mathbb{E}\left[\|\widehat{\beta}_k - \widehat{\beta}_{k,-1}\|_2^2\right]^{1/2} \cdot \mathbb{E}\left[1 + \|(y_0, x_0^\top \widehat{\beta}_k)\|_2^4 + \|(y_0, x_0^\top \widehat{\beta}_{k,-1})\|_2^4\right],
\end{aligned}$$

which by Equations (E.44) and (E.46) goes to zero as $n, p \to \infty$. □

S.9 Proof of Theorem 4

Theorem 4 (Coverage guarantee)**.** *Under the conditions of Theorem 3, assume further that the distribution of the noise ε_i is continuous with density bounded by κ_{pdf}. Denote by $\widehat{\alpha}_k(q)$ the q-quantile of $\{y_i - x_i^\top \widehat{\beta}_{k,-i} : i \in [n]\}$. Then, for any quantile levels $0 \le q_1 \le q_2 \le 1$, letting $\mathcal{I}_k = [\widehat{\alpha}_k(q_1), \widehat{\alpha}_k(q_2)]$, we have as $n, p \to \infty$,*

$$\max_{k \in [K]} \mathbb{P}_{(x_0, y_0)}\left(y_0 - x_0^\top \widehat{\beta}_k \in \mathcal{I}_k \mid X, y\right) \xrightarrow{\text{a.s.}} q_2 - q_1. \tag{13}$$

Proof. For $z \in \mathbb{R}$, we define $\mathsf{l}_z(y, u) = \mathbb{1}\{y - u \le z\}$. We first prove that if we replace $\psi(y, u)$ by $\mathsf{l}_z(y, u)$ in Theorem 3, then as $n, p \to \infty$ we still have

$$\sup_{k \in \{0\} \cup [K]} |\widehat{\Psi}^{\text{loo}}(\widehat{\beta}_k) - \Psi(\widehat{\beta}_k)| \xrightarrow{\text{a.s.}} 0. \tag{E.64}$$

This step is achieved via uniformly approximating l_z using Lipschitz functions. To be specific, we let $\{g_j\}_{j \in \mathbb{N}_+}$ be a sequence of Lipschitz functions satisfying $\|g_j - \mathsf{l}_z\|_\infty \le 2^{-j}$. We define

$$\widehat{\Psi}_j^{\text{loo}}(\widehat{\beta}_k) = \frac{1}{n}\sum_{i=1}^n g_j(y_i - x_i^\top \widehat{\beta}_{k,-i}) \quad \text{and} \quad \Psi_j(\widehat{\beta}_k) = \mathbb{E}[g_j(y_0 - x_0^\top \widehat{\beta}_k) \mid X, y].$$

By Theorem 3, we know that for every j,

$$\sup_{k \in \{0\} \cup [K]} |\widehat{\Psi}_j^{\text{loo}}(\widehat{\beta}_k) - \Psi_j(\widehat{\beta}_k)| \xrightarrow{\text{a.s.}} 0.$$

Furthermore, notice that

$$|\widehat{\Psi}_j^{\text{loo}}(\widehat{\beta}_k) - \widehat{\Psi}^{\text{loo}}(\widehat{\beta}_k)| \le 2^{-j} \quad \text{and} \quad |\Psi_j(\widehat{\beta}_k) - \Psi(\widehat{\beta}_k)| \le 2^{-j},$$

and j is arbitrary, thus completing the proof of Equation (E.64).

We denote by $\widehat{F}_k$ the cumulative distribution function (CDF) of the uniform distribution over $\{y_i - x_i^\top \widehat{\beta}_{k,-i} : i \in [n]\}$, and denote by F_k the CDF of $y_0 - x_0^\top \widehat{\beta}_k$ conditioning on (X, y). We emphasize that both F_k and $\widehat{F}_k$ are random distributions that depend on (X, y). Next, we prove that F_k is Lipschitz continuous.

Lemma 42. *Under the conditions of Theorem 4, F_k is κ_{pdf}-Lipschitz continuous.*

Proof of Lemma 42. Note that $y_0 - x_0^\top \widehat{\beta}_k = f(x_0) - x_0^\top \widehat{\beta}_k + \varepsilon_0$, where ε_0 is independent of $f(x_0) - x_0^\top \widehat{\beta}_k$. Since ε_0 has a probability density function (PDF), we see that $y_0 - x_0^\top \widehat{\beta}_k$ also has a PDF, and we denote it by h. We denote by h_ε the PDF of ε_0 and denote by G the CDF of $f(x_0) - x_0^\top \widehat{\beta}_k$, then we have

$$h(x) = \int h_\varepsilon(x - z) dG(z),$$

which is uniformly upper bounded by κ_{pdf} for all $x \in \mathbb{R}$. □

As a consequence of Lemma 42 and the fact that $y_0 - x_0^\top \widehat{\beta}_k$ has bounded fourth moment (see Equation (E.44) for derivation), we immediately obtain that $\sup_{k \in \{0\} \cup [K]} \|\widehat{F}_k - F_k\|_\infty \xrightarrow{\text{a.s.}} 0$ as $n, p \to \infty$.

In addition, it is not hard to see that

$$\left|\widehat{F}_k(\widehat{\alpha}_k(q_i)) - q_i\right| \le n^{-1}$$

for all $i \in \{1, 2\}$ and $k \in \{0\} \cup [K]$. Therefore,

$$\sup_{k \in \{0\} \cup [K]} |F_k(\widehat{\alpha}_k(q_i)) - q_i| \le \sup_{k \in \{0\} \cup [K]} \left|\widehat{F}_k(\widehat{\alpha}_k(q_i)) - q_i\right| + \sup_{k \in \{0\} \cup [K]} \|\widehat{F}_k - F_k\|_\infty \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$, thus completing the proof of the theorem. □

S.10 Proofs in Section 5

S.10.1 Proof of Proposition 5

Proposition 5 (Correctness of the modified augmented system)**.** *For all $k \in [K]$ and $i \in [n]$, it holds that $\widetilde{\beta}_{k,-i} = \widehat{\beta}_{k,-i}$.*

Proof. We prove the lemma through induction on k. For $k = 0$, by definition $\widetilde{\beta}_{0,-i} = \widehat{\beta}_{0,-i} = \beta_0$ for all $i \in [n]$. Suppose that we have $\widetilde{\beta}_{k,-i} = \widehat{\beta}_{k,-i}$ iteration k and all $i \in [n]$, we then prove that it also holds for iteration $k+1$ via induction. Using its definition, we see that

$$\begin{aligned}
\widetilde{\beta}_{k+1,-i} &= \widetilde{\beta}_{k,-i} - \frac{2\delta_k}{n} X^\top X \widetilde{\beta}_{k,-i} + \frac{2\delta_k}{n} X^\top \widetilde{y}_{k,-i} \\
&= \widetilde{\beta}_{k,-i} - \frac{2\delta_k}{n} X_{-i}^\top X_{-i} \widetilde{\beta}_{k,-i} + \frac{2\delta_k}{n} X_{-i}^\top y_{-i} - \frac{2\delta_k}{n} x_i \big(x_i^\top \widetilde{\beta}_{k,-i} - x_i^\top \widehat{\beta}_{k,-i}\big) \\
&= \widehat{\beta}_{k,-i} - \frac{2\delta_k}{n} X_{-i}^\top X_{-i} \widehat{\beta}_{k,-i} + \frac{2\delta_k}{n} X_{-i}^\top y_{-i} \\
&= \widehat{\beta}_{k+1,-i},
\end{aligned}$$

thus completing the proof of the lemma by induction. □

S.10.2 Proof of Proposition 6

Proposition 6 (Smoother representation for the modified augmented system)**.** *For all $k \in [K]$ and $i \in [n]$, there is a vector $(h_{ij}^{(k)})_{j \le n}$ and scalar $b_i^{(k)}$ depending $\delta = (\delta_0, \dots \delta_{k-1})$ and X such that:*

$$x_i^\top \widehat{\beta}_{k,-i} = x_i^\top \widetilde{\beta}_{k,-i} = \sum_{j=1}^n h_{ij}^{(k)} y_j + b_i^{(k)}.$$

Proof. We prove this lemma by induction over k. For the base case $k = 0$, the requirement of the lemma can be satisfied by setting

$$h_{ij}^{(0)} = 0, \qquad b_i^{(0)} = x_i^\top \beta_0, \qquad i, j \in [n].$$

Suppose we can find $(h_{ij}^{(k)})_{i,j \le n}$ and $(b_i^{(k)})_{i \le n}$ for iteration k, we next show that the counterpart quantities also exist for iteration $k+1$. We define $H^{(k)} \in \mathbb{R}^{n \times n}$, $b^{(k)} \in \mathbb{R}^n$, such that $H_{ij}^{(k)} = h_{ij}^{(k)}$ and $b_i^{(k)} = b_i^{(k)}$. Using induction hypothesis and Proposition 5, we have

$$\begin{aligned}
&x_i^\top \widehat{\beta}_{k+1,-i} \\
&= x_i^\top \widetilde{\beta}_{k,-i} \\
&= x_i^\top \left(\widetilde{\beta}_{k,-i} - \frac{2\delta_k}{n} X^\top X \widetilde{\beta}_{k,-i} + \frac{2\delta_k}{n} X^\top \widetilde{y}_{k,-i} \right) \\
&= x_i^\top \left(\widehat{\beta}_{k,-i} - \frac{2\delta_k}{n} X^\top X \widehat{\beta}_{k,-i} + \frac{2\delta_k}{n} X_{-i}^\top y_{-i} + \frac{2\delta_{k+1}}{n} x_i x_i^\top \widehat{\beta}_{k,-i} \right) \\
&= \sum_{j=1}^n h_{ij}^{(k)} y_j + b_i^{(k)} - \frac{2\delta_k}{n} x_i^\top X^\top (H^{(k)} y + b^{(k)}) + \frac{2\delta_k}{n} x_i^\top X_{-i}^\top y_{-i} + \frac{2\delta_{k+1}}{n} \|x_i\|_2^2 \left(\sum_{j=1}^n h_{ij}^{(k)} y_j + b_i^{(k)} \right).
\end{aligned}$$

Note that the right-hand of the display above is affine in y, which completes the proof for iteration $k+1$. This completes our induction proof. □

S.10.3 Proof of Proposition 7

Proposition 7 (Recursive shortcut formula for LOO predictions in GD)**.** *For all $k \in [K]$ and $i \in [n]$,*

$$x_i^\top \widehat{\beta}_{k,-i} = x_i^\top \widehat{\beta}_k + A_{i,k} \|x_i\|_2^2 + \sum_{j=1}^{k-1} B_{i,k}^{(j)} x_i^\top (X^\top X)^j x_i,$$

where

$$\begin{aligned}
A_{i,k+1} &= A_{i,k} + \frac{2\delta_k A_{i,k} \|x_i\|_2^2}{n} \\
&\quad + \sum_{j=1}^{k-1} \frac{2\delta_k B_{i,k}^{(j)} x_i^\top (X^\top X)^j x_i}{n} + \frac{2\delta_{k+1}(x_i^\top \widehat{\beta}_k - y_i)}{n}, \\
B_{i,k+1}^{(1)} &= B_{i,k}^{(1)} - \frac{2\delta_k A_{i,k}}{n}, \\
B_{i,k+1}^{(j)} &= B_{i,k}^{(j)} - \frac{2\delta_k B_{i,k}^{(j-1)}}{n}, \quad 2 \le j \le k,
\end{aligned}$$

and we make the convention that $B_{i,k}^{(k)} = 0$.

Proof. By definition, $\widehat{\beta}_{0,-i} = \widehat{\beta}_0$ for all $i \in [n]$. After implementing the first step of gradient descent, we have

$$\widehat{\beta}_{1,-i} = \widehat{\beta}_{0,-i} - \frac{2\delta_1}{n} X_{-i}^\top X_{-i} \widehat{\beta}_{0,-i} + \frac{2\delta_1}{n} X_{-i}^\top y_{-i}$$

$$= \widehat{\beta}_1 + \frac{2\delta_1}{n} x_i x_i^\top \widehat{\beta}_0 - \frac{2\delta_1}{n} y_i x_i.$$

We define $A_{i,1} = 2\delta_1(x_i^\top \widehat{\beta}_0 - y_i)/n$, then $\widehat{\beta}_{1,-i} = \widehat{\beta}_1 + A_{i,1} x_i$. Now suppose $\widehat{\beta}_{k,-i}$ admits the decomposition

$$\widehat{\beta}_{k,-i} = \widehat{\beta}_k + A_{i,k} x_i + \sum_{j=1}^{k-1} B_{i,k}^{(j)} (X^\top X)^j x_i$$

for some $A_{i,k}, B_{i,k}^{(j)} \in \mathbb{R}$. Then, in the next step of gradient descent, by definition we have

$$\begin{aligned}
\widehat{\beta}_{k+1,-i} &= \widehat{\beta}_{k,-i} - \frac{2\delta_k}{n} X_{-i}^\top X_{-i} \widehat{\beta}_{k,-i} + \frac{2\delta_k}{n} X_{-i}^\top y_{-i} \\
&= \widehat{\beta}_{k+1} + A_{i,k} x_i + \sum_{j=1}^{k-1} B_{i,k}^{(j)} (X^\top X)^j x_i - \frac{2\delta_k A_{i,k}}{n} X^\top X\, x_i - \sum_{j=1}^{k-1} \frac{2\delta_k B_{i,k}^{(j)}}{n} (X^\top X)^{j+1} x_i \\
&\quad + \frac{2\delta_k A_{i,k} \|x_i\|_2^2}{n} x_i + \sum_{j=1}^{k-1} \frac{2\delta_k B_{i,k}^{(j)} x_i^\top (X^\top X)^j x_i}{n} x_i + \frac{2\delta_{k+1}(x_i^\top \widehat{\beta}_k - y_i)}{n} x_i.
\end{aligned}$$

As a result, we obtain the following update equations:

$$\begin{aligned}
A_{i,k+1} &= A_{i,k} + \frac{2\delta_k A_{i,k} \|x_i\|_2^2}{n} + \sum_{j=1}^{k-1} \frac{2\delta_k B_{i,k}^{(j)} x_i^\top (X^\top X)^j x_i}{n} + \frac{2\delta_{k+1}(x_i^\top \widehat{\beta}_k - y_i)}{n}, \\
B_{i,k+1}^{(1)} &= B_{i,k}^{(1)} - \frac{2\delta_k A_{i,k}}{n}, \\
B_{i,k+1}^{(j)} &= B_{i,k}^{(j)} - \frac{2\delta_k B_{i,k}^{(j-1)}}{n}, \quad 2 \le j \le k,
\end{aligned}$$

where we make the convention that $B_{i,k}^{(k)} = 0$. □