Failures and Successes of Cross-Validation for Early-Stopped Gradient Descent in High-Dimensional Least Squares

Pratik Patil
University of California, Berkeley

Yuchen Wu
University of Pennsylvania

Ryan 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) iterates in high-dimensional least squares regression. Surprisingly, our results show that GCV can be inconsistent for estimating the squared prediction risk, even under a well-specified linear model with isotropic design. In contrast, we prove that LOOCV converges uniformly along the GD trajectory to the prediction risk. Our theory holds under mild assumptions on the data distribution and does not require the underlying regression function to be linear. Furthermore, by suitably extending LOOCV, we construct consistent estimators for the entire prediction error distribution along the GD trajectory and for a wide class of its functionals. This in particular enables the construction of pathwise prediction intervals for the unknown response with asymptotically correct nominal coverage conditional on the training data.

1 INTRODUCTION

Cross-validation (CV) is a widely used method for assessing and selecting models in various machine learning applications. It is particularly effective in evaluating the performance of *explicitly regularized* methods, such as ridge regression and lasso. Using multiple validation sets, CV helps to identify the optimal trade-off between bias and variance as a function of the level of regularization and enables the practitioner to choose the level of regularization that maximizes predictive accuracy (Hastie et al., 2009; Györfi et al., 2006).

On the other hand, iterative algorithms such as gradient descent (GD) and its derivatives are a popular class of algorithms for optimizing the parameters of machine learning models (Celentano et al., 2020, 2021; Montanari and Wu, 2022). These algorithms are known to induce the so-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 share a close connection to explicitly regularized ridge estimates (Ali et al., 2019, 2020). This naturally leads to the following question: *Can we reliably use CV to assess model performance along the entire trajectory of iterative algorithms?*

An affirmative answer to this question would enable the cross-validation of a model's performance along the entire algorithm trajectory. This would then allow us to determine whether and when to early stop the training procedure before convergence, 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.

Common CV choices are: split CV, K-fold CV, and LOOCV. However, it has been observed that split CV and K-fold CV with small K (such as 5 or 10) can suffer from severe biases in many high-dimensional problems (Rad and Maleki, 2020; Rad et al., 2020). Although LOOCV in most cases mitigates bias issues, it is often computationally expensive to implement. Fortunately, for estimators that are constructed based on linear smoothers, GCV serves as an efficient approximation to LOOCV (Golub et al., 1979a; Jansen et al., 1997). Both LOOCV and GCV have been shown to be consistent for estimating the out-of-sample prediction risks for ridge regression in various high-dimensional settings (Patil et al., 2021, 2022b; Wei et al., 2022). Noting that the iterates along the GD trajectory are linear smoothers, one natural choice is to apply GCV to estimate the out-of-sample prediction risk for estimators trained using GD. However, the efficacy of either of GCV and LOOCV in the iterative context is not yet well studied.

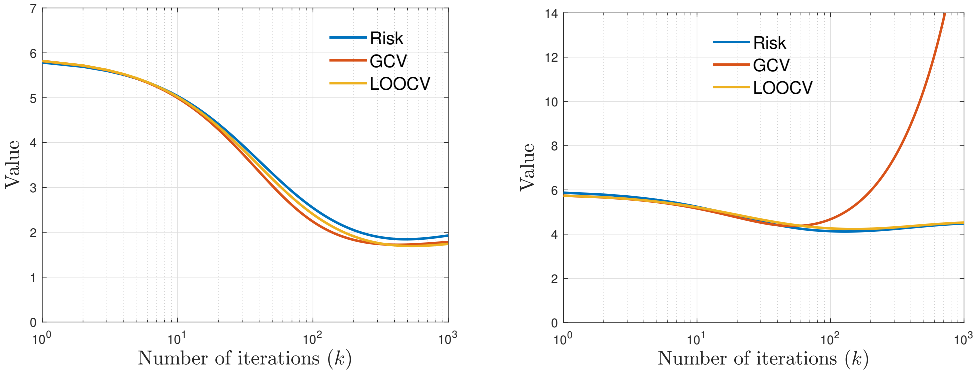

Figure 1: **GCV performs poorly in overparameterized regimes, while LOOCV provides accurate risk estimates.** We illustrate GCV and LOOCV versus squared prediction risk, as a function of the GD iteration number k. Here $n = 3000$, $p = 1500$ on the **left** (underparameterized), and $n = 3000$, $p = 6000$ on the **right** (overparameterized). We assume a linear model in both settings, where the features are sampled i.i.d. from a standard normal distribution, the linear coefficient vector has Euclidean norm 5, and the additive noise has standard deviation 1. We initialize GD iterations with a null model and use a constant step size of 0.01. The curves are obtained using a single realization of the dataset.

In this paper, we provide theoretical analyses for both GCV and LOOCV along the GD trajectory in high-dimensional linear regression. The first, and somewhat surprising, result of this paper shows that GCV is inconsistent for the out-of-sample prediction error, even in the most basic context of a correctly specified linear response model with isotropic design. This inconsistency becomes particularly pronounced in overparameterized regimes, where the number of features is greater than the number of observations. In such cases, the (multiplicative) gap between GCV and risk can be substantial, especially as the iteration proceeds. This is problematic for model assessment, as these are precisely the scenarios in which the optimal stopping time can occur at a far iteration that allows (near) interpolation (Kobak et al., 2020; Wu and Xu, 2020; Richards et al., 2020).

Partially towards understanding the failure of GCV, in the second part of this paper, we analyze the theoretical properties of LOOCV along the entire GD path. We demonstrate that the failure of the GCV lies in its approximation and not with the LOOCV itself. Our main result for this part shows that LOOCV is uniformly consistent along the GD path for both square and general risk functionals. Our analysis accommodates the regimes where the number of features p can scale proportionally with the number of observations n, and in particular includes situations where p is much higher than n, where the iterates along the GD path can (nearly) interpolate training data as training progresses. 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 standard regression setting, we show that GCV is almost everywhere inconsistent for estimating the squared prediction risk throughout the GD trajectory (Theorem 1). We prove this result by separately deriving the asymptotic limits of the GCV estimator and the true risk and demonstrating that they do not match.

2. **LOOCV consistency.** In contrast, we show that LOOCV is uniformly consistent for estimating the prediction risk, not only for the squared loss (Theorem 2) but also for general pseudo-Lipschitz losses (Theorem 3). Our analysis allows the number of GD steps to increase with the size of the input and only requires the distributions of the feature and the noise to satisfy a T_2-inequality. In particular, we do not assume any specific model for the response. As a consequence of uniformity, we show that the risk of the LOOCV-tuned iterate almost surely matches that of the oracle-tuned iterate. Furthermore, we also propose a faster implementation of the LOOCV that has lower computational complexity compared to the naive implementation (Proposition 5).

3. **Pathwise predictive inference.** We propose a natural extension of LOOCV to estimate the distribution of out-of-sample prediction error along the GD trajectory based on empirical distributions of LOOCV errors. As an application, we demonstrate that this distribution can be used to consistently estimate quantiles of the prediction error distribution uniformly along the GD path. This further enables the construction of uniformly consistent prediction intervals that have the 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. From an optimization point of view, they are efficient in terms of scalability. 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, 2021). This behavior is attributed to the implicit regularization effect of GD (Gunasekar et al., 2018a,b). Implicit regularization has a rich history in machine learning, such as early insights into the advantages of early stopping in neural network training (Morgan and Bourlard, 1989). A parallel idea in numerical analysis is known as Landweber iteration (Landweber, 1951). There is also a rich literature on early stopping in the context of boosting (Bühlmann and Yu, 2003; Rosset et al., 2004; Zhang and Yu, 2005). Explicit connections between squared regularization and the iterates generated by GD are made by Friedman and Popescu (2004). Recent works have further shown favorable performance of early stopped GD over squared regularization in various nonparametric models (Yao et al., 2007; Bauer et al., 2007; Raskutti et al., 2014; Wei et al., 2017). Furthermore, several precise pathwise estimator couplings and risk correspondences between GD and its variants and the ridge regression are established by Ali et al. (2019, 2020), among others.

CV is a classical approach for parameter tuning and model selection (Allen, 1974; Stone, 1974, 1977; Geisser, 1975). For reviews of CV variants used in practice, we refer the readers to Arlot and Celisse (2010); Zhang and Yang (2015). The simplest version of CV is the sample-split CV (Hastie et al., 2009), which holds a specific portion of the data to evaluate the performance of the model under different parameters. By repeatedly fitting candidate models on multiple subsets of the data, the K-fold CV extends the idea of sample splitting and further reduces the variance of the procedure. In a high-dimensional regime where the number of variables is comparable to the number of observations, the commonly used small values of K, such as 5 or 10, suffer from severe bias issues (Rad and Maleki, 2020). LOOCV, that is the case where $K = n$, alleviates such bias issues. The theoretical properties of LOOCV have been analyzed in recent years by Kale et al. (2011); Kumar et al. (2013); Rad et al. (2020). However, LOOCV, as discussed above, in general is computationally expensive, and there has also been work designing and analyzing approximate LOOCV to address this computational problem; see, for example, Wang et al. (2018); Stephenson and Broderick (2020); Wilson et al. (2020); Xu et al. (2019); Rad et al. (2020); Auddy et al. (2023).

GCV serves as an approximation to the so-called "shortcut" leave-one-out formula. GCV was initially studied in the context of fixed-X design settings for linear smoothers (Golub et al., 1979b; Craven and Wahba, 1979). The consistency of GCV in these scenarios has been investigated by Li (1985, 1986, 1987). More recently, the focus has shifted towards the random-X setting, where GCV has garnered significant interest. Specifically, its consistency for ridge regression has been confirmed under various data settings (Hastie et al., 2022; Patil et al., 2021, 2022b; Wei et al., 2022). Beyond linear smoothers, a risk estimator like GCV can be defined in terms of degrees-of-freedom adjustment. See Bayati and Montanari (2011); Bayati et al. (2013); Miolane and Montanari (2021) for lasso, Bellec (2020); Bellec and Zhang (2021); Bellec and Shen (2022) for general M-estimators, among others. Our paper extends this existing literature by examining the behavior of GCV in the context of iterates along the GD path for high-dimensional linear regression problems. Unlike ridge regression, we show in this paper that GCV in fact turns out to be inconsistent for estimating the prediction risk along the GD path almost everywhere.

Most of the aforementioned CV papers have focused on full solutions to empirical risk minimization problems. While iterative algorithms are frequently applied to train machine learning models, there has been little work that studies LOOCV or approximate CV for iterative algorithms. Very recently, Luo et al. (2023) considers approximating LOOCV for iterative algorithms. They propose an algorithm that is more efficient than naive LOOCV when $p \ll n$. They also show that their method approximates LOOCV well. In our work, we instead focus more on analyzing LOOCV itself (along with GCV), thus complementing their work. Moreover, we also allow overparameterized regimes where $p \asymp n$.

Finally, another thread of CV research is on the inferential properties of CV; see, for example, Wager et al. (2014); Lei (2020); Bates et al. (2021). Asymptotic distributions of suitably normalized K-fold CV are obtained in Austern and Zhou (2020), under certain stability conditions on the predictors. Central limit theorems for CV error and a consistent estimator of its variance are derived in Bayle et al. (2020), which assumes stability conditions similar to Kumar et al. (2013); Celisse and Guedj (2016). Their results yield asymptotically valid confidence intervals for the prediction error and can be applied to both K-fold CV and LOOCV. In a related direction, a body of work has considered predictive inference, in the form of prediction intervals; see, e.g., Patil et al. (2022b) for results related to ridge regression. Apart from studying consistency properties of CV, in this paper, we also consider predictive inference along the iterative GD trajectory.

2 PRELIMINARIES

In this section, we set our notation by formally defining the risk functionals of the least squares GD iterates and present estimators based on LOOCV and GCV.

2.1 Gradient Descent and Squared Risk

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

We focus on the ordinary least squares problem:

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

In the sequel, we consider minimizing the square loss in (1) using GD. Specifically, given step sizes $\boldsymbol{\delta} = (\delta_0, \dots, \delta_{K-1}) \in \mathbb{R}^K$ and an arbitrary initialization vector $\widehat{\boldsymbol{\beta}}_0 \in \mathbb{R}^{p+1}$, GD is defined recursively as follows:

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

Let $(\boldsymbol{x}_0, y_0) \in \mathbb{R}^{p+1} \times \mathbb{R}$ be 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, the squared prediction risk achieved by the GD iterate at step $k \in [K]$, denoted by $R(\widehat{\boldsymbol{\beta}}_k)$, is defined as:

$$R(\widehat{\boldsymbol{\beta}}_k) := \mathbb{E}_{\boldsymbol{x}_0, y_0}\left[(y_0 - \boldsymbol{x}_0^\top\widehat{\boldsymbol{\beta}}_k)^2 \mid \boldsymbol{X}, \boldsymbol{y}\right]. \tag{3}$$

2.2 GCV and LOOCV

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

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

where $\widehat{\boldsymbol{\beta}}_{k,-i}$ denotes the GD estimate with k iterations trained on dataset $(\boldsymbol{X}_{-i}, \boldsymbol{y}_{-i})$ that excludes the i-th data point from the full data $(\boldsymbol{X}, \boldsymbol{y})$.

A short calculation shows that the GD iterates can be represented as linear smoothers, i.e., the in-sample predictions can be expressed as a linear transformation of the response vector $\boldsymbol{X}\widehat{\boldsymbol{\beta}}_k = \boldsymbol{H}_k\boldsymbol{y} + \widetilde{\boldsymbol{H}}_k\widehat{\boldsymbol{\beta}}_0$, where $\boldsymbol{H}_k, \widetilde{\boldsymbol{H}}_k$ are smoothing matrices:

$$\boldsymbol{H}_k = \sum_{j=0}^{k-1}\frac{\delta_j}{n}\boldsymbol{X}\prod_{r=1}^{k-j-1}\left(\boldsymbol{I}_{p+1} - \delta_{k-r}\widehat{\boldsymbol{\Sigma}}\right)\boldsymbol{X}^\top, \tag{5}$$

$$\widetilde{\boldsymbol{H}}_k = \boldsymbol{X}\prod_{r=1}^{k}\left(\boldsymbol{I}_{p+1} - \delta_{k-r}\widehat{\boldsymbol{\Sigma}}\right). \tag{6}$$

In displays (5) and (6) above, we denote by $\widehat{\boldsymbol{\Sigma}} := \boldsymbol{X}^\top\boldsymbol{X}/n$, the sample covariance matrix.

Towards defining GCV, suppose that we have a predictor $\widehat{f} : \mathbb{R}^{p+1} \to \mathbb{R}$ that is a linear smoother. In other words, $\widehat{f}(\mathbf{x}) = \mathbf{s}_{\mathbf{x}}^\top\mathbf{y}$ for some vector $\mathbf{s}_{\mathbf{x}} \in \mathbb{R}^n$ that depends only on the design matrix $\mathbf{X}$ and the test point $\mathbf{x}$. The smoothing matrix associated with the prediction $\widehat{f}$ is then the matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$ with rows $\mathbf{s}_{\mathbf{x}_1}^\top, \dots, \mathbf{s}_{\mathbf{x}_n}^\top$. For any linear smoother, the GCV estimator (Craven and Wahba, 1978; Golub et al., 1979b) of the prediction risk of the predictor $\widehat{f}$ is defined as:

$$\widehat{R}^{\mathrm{gcv}}(\widehat{f}) := \frac{\|\mathbf{y} - \mathbf{S}\mathbf{y}\|_2^2/n}{(1 - \mathrm{tr}[\mathbf{S}]/n)^2}. \tag{7}$$

The numerator of the GCV estimator represents the training error. This error is typically biased downward, meaning that it often underestimates the true error. The denominator of the GCV estimator corrects for this downward bias, often referred to as the training "optimism" of the predictor. In particular, the term $1 - n^{-1}\mathrm{tr}(\mathbf{S})$ acts as a degrees-of-freedom correction, adjusting the estimate to more accurately reflect the complexity of the model (Bellec, 2020).

This smoother representation motivates us to estimate the prediction risk following the idea of GCV:

$$\widehat{R}^{\mathrm{gcv}}(\widehat{\boldsymbol{\beta}}_k) := \frac{1}{n}\sum_{i=1}^n\frac{(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_k)^2}{(1 - \mathrm{tr}[\boldsymbol{H}_k]/n)^2}. \tag{8}$$

When defining GCV above, we implicitly assume $\widehat{\boldsymbol{\beta}}_0 = \mathbf{0}_{p+1}$. We refer the reader to Golub et al. (1979a); Jansen et al. (1997) for more examples of GCV. There is a general way of understanding this estimator and also of extending the definition of GCV for estimators that are not necessarily linear smoothers, by employing the idea of degrees-of-freedom adjustments; see, for example, Bayati and Montanari (2011); Bellec (2020); Bellec and Shen (2022).

We shall shortly see in Section 3 that GCV does not consistently estimate the prediction risk even under 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 demonstrate that GCV is pointwise 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 $\boldsymbol{\delta} = \delta \mathbf{1}_K$, and omit the intercept term. We initialize GD at the origin: $\widehat{\boldsymbol{\beta}}_0 = \mathbf{0}_p$. For the specific result presented in this section, we will impose the following structural assumptions on feature and response distributions.

Assumption A (Feature distribution)**.** Each feature vector $\boldsymbol{x}_i$, 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: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i$. Here, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown signal vector satisfying $\lim_{p\to\infty} \|\boldsymbol{\beta}_0\|_2^2 = r^2 < \infty$, and ε_i is an unobserved error variable independent of $\boldsymbol{x}_i$ following a Gaussian distribution with mean 0 and variance $\sigma^2 > 0$.

The condition for each y_i to be centered is assumed only for simplicity. Accordingly, we do not include an additional intercept term in the feature vector, implying that $\boldsymbol{x}_i \in \mathbb{R}^p$. Although one could establish the inconsistency of GCV under more relaxed assumptions, we choose to work under these conditions 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, with their ratio p/n converging to a constant $\zeta_* \in (0, \infty)$. This regime has received considerable attention recently in high-dimensional statistics and machine learning theory; see, e.g., Hastie et al. (2022); Bartlett et al. (2021), among others.

The dynamics of GD are determined by both the step size δ and the number of iterations K. In many settings, $\delta \to 0$ and $K \to \infty$ as $n, p \to \infty$, reducing the GD iterations to a continuous-time gradient flow (Ali et al., 2019; Celentano et al., 2021; Berthier et al., 2023). For the scope of our results on inconsistency, we assume that the product of K and δ approaches a constant $T > 0$, that is, $K\delta \to T$. Note that the limit is taken simultaneously for all parameters K, δ, n, and p.

Theorem 1 (Pointwise inconsistency of GCV)**.** *Suppose Assumptions A and B hold. Let $p/n \to \zeta_*$, $K \to \infty$, and $\delta \to 0$ simultaneously such that $K\delta \to T$, where $T, \zeta_* > 0$ are constants that are independent of (n, p). Then, for every fixed $\zeta_* > 0$, it holds that for all $T > 0$, except for a set of Lebesgue measure zero,*

$$|\widehat{R}^{\mathrm{gcv}}(\widehat{\boldsymbol{\beta}}_K) - R(\widehat{\boldsymbol{\beta}}_K)| \overset{\mathrm{p}}{\not\to} 0. \tag{9}$$

In other words, the theorem says that GCV does not consistently track the true risk along GD, even for well-specified linear models with isotropic features. Furthermore, the inconsistency occurs almost surely along the GD path for all signal-to-noise ratios, and both under the underparameterized regime (when $\zeta_* < 1$) and the overparameterized regime (when $\zeta_* > 1$).

It is worth mentioning that the inconsistency can be severe. In particular, it is easy to show that $\lim_{K\to\infty} \widehat{R}^{\mathrm{gcv}}(\widehat{\boldsymbol{\beta}}_K) \to \infty$, while $R(\widehat{\boldsymbol{\beta}}_K) \to r^2 + \sigma^2$ under the assumptions of Theorem 1. This is evident in Figure 1. This follows from the proof of Theorem 1. Appendix C in the supplement provides a proof outline and Appendix E provides a detailed proof. Finally, we note that the conclusions of Theorem 1 can also be extended to more relaxed assumptions, but our main goal is simply to confirm the inconsistency. We employ relaxed assumptions for our LOOCV analysis next.

4 LOOCV CONSISTENCY

Despite the inconsistency of GCV, LOOCV remains consistent for GD. In this section, we establish a uniform consistency result for LOOCV along the GD path.

4.1 Squared Risk Consistency

We begin by focusing on the squared risk. The technical crux of our analysis revolves around establishing the concentration of the estimator $\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$. To do this, we leverage Talagrand's T_2-inequality (Gozlan, 2009). Specifically, under the assumption that both the entries of the feature vector and noise distributions satisfy the T_2-inequality (see Assumption C for details), we demonstrate that $\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$ behaves approximately as a Lipschitz function of these random variables. Together, these results enable us to apply a powerful concentration inequality leading to the desired result.

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

We will shortly provide a formal definition (in Definition 1) of distributions that satisfy the T_2-inequality. But first we build some high-level intuition and utility. Roughly speaking, there is a dimension-free concentration inequality that gives concentration bounds for Lipschitz functions of random variables, each satisfying the T_2-inequality. In fact, satisfying 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 Proposition 23 in the supplement and Appendix G.1 for further properties of the T_2-inequality). In the following, we formally describe what

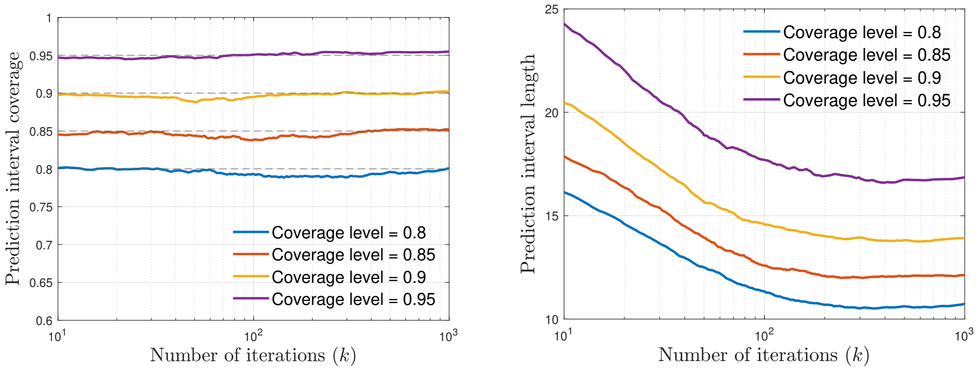

Figure 2: **LOOCV provides coverage consistent prediction intervals across different nominal coverage levels.** We illustrate the empirical coverage and length of LOOCV prediction intervals at varying coverage levels along the gradient descent iterations. We consider an overparameterized regime with $n = 2500$ and $p = 5000$. The features are drawn randomly with a covariance structure that satisfies $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$ for all i, j and $\rho = 0.25$. The response is generated from a nonlinear model with a nonrandom signal vector β_0 and t-distributed noise with 5 degrees-of-freedom. We align the signal vector with the top eigenvector of $\mathbf{\Sigma}$, similar to that of Kobak et al. (2020). The gradient descent is run with step sizes of 0.01. See Appendix L for further details on the experimental setup. Throughout the GD path, the prediction intervals have excellent finite-sample coverage (**left**), and the smallest prediction length is obtained at a far enough iteration (**right**).

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 and only if there exists a constant $\sigma(\mu) \geq 0$, such that for every distribution ν,

$$W_2(\mu, \nu) \leq \sqrt{2\sigma^2(\mu) D_{\texttt{KL}}(\nu \| \mu)}, \qquad (10)$$

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

One prominent example of probability distributions that satisfy the T_2-inequality is the family of distributions that satisfy the logarithmic Sobolev inequality (LSI). We provide more details in **??**. Exercise 4.13 in Van Handel (2014) tells us that if a distribution satisfies LSI, then it also satisfies the T_2-inequality. We note that all distributions that are strongly log-concave satisfy LSI. In addition, many non-log-concave distributions, such as Gaussian convolutions of distributions with bounded support, also satisfy this condition (Chen et al., 2021). A modified LSI inequality is also intimately related to the convergence of Markov chains (see Chapter 3 of Van Handel (2014) for more details).

Next we formally state our assumptions for this section.

Assumption C (Feature distribution)**.** In the following, we require the constants $(\sigma_\Sigma, \sigma_z)$ to be independent of (n, p).

1. We assume each feature $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ for $i \in [n]$ decomposes into $\boldsymbol{x}_i^\top = ((\mathbf{\Sigma}^{1/2}\boldsymbol{z}_i)^\top, 1)$, where $\boldsymbol{z}_i \in \mathbb{R}^p$ contains i.i.d. entries $z_{ij} \overset{\text{i.i.d.}}{\sim} \mu_z$. We also assume that μ_z has mean 0, variance 1, and satisfies the T_2-inequality with constant σ_z.
2. There exists a constant σ_Σ, such that $\|\mathbf{\Sigma}\|_{\text{op}} \leq \sigma_\Sigma$.

It is worth mentioning that to show uniform consistency, we do not assume that the smallest eigenvalue of $\mathbf{\Sigma}$ in Assumption C is bounded away from 0. This is possible because the iterates along the GD path are implicitly regularized. This is similar to not having to assume a lower bound on the smallest eigenvalue for ridge regression when $\lambda > 0$ (as opposed to ridgeless regression when $\lambda \to 0^+$, where we do need such an assumption) (Dobriban and Wager, 2018; Hastie, 2020).

We also make the following assumptions about the response distribution.

Assumption D (Response distribution)**.** In the following, we require the constants $(\sigma_\varepsilon, m_2, m_4, m_8)$ to be independent of (n, p).

1. We assume that $(\boldsymbol{x}_i, y_i)$ are sampled i.i.d., satisfying $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$.[1]
2. We assume $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mu_\varepsilon$ for $i \in [n]$, where 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. In addition, we assume $\mathbb{E}[y_1^8] \leq m_8$, $\mathbb{E}[y_1^4] \leq m_4$ and $\mathbb{E}[y_1^2] \leq m_2$. Without loss, we let $L_f \geq 1$.

The assumptions we impose in this section are strictly weaker than those in Section 3. Under these assumptions, we are able to establish our main theorem:

Theorem 2 (Uniform squared consistency of LOOCV)**.** *Suppose Assumptions C and D hold. In addition, assume that there exist constants $\Delta, B_0, \zeta_L, \zeta_U$ that are independent of (n, p), such that the following conditions are satisfied: (1) $\sum_{k=1}^{K} \delta_{k-1} \leq \Delta$, (2) $\|\widehat{\boldsymbol{\beta}}_0\|_2 \leq B_0$, (3)*

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

$0 < \zeta_L \le \zeta \le \zeta_U < \infty$, where $\zeta := p/n$. Furthermore, let $K = o(n \cdot (\log n)^{-3/2})$. Then, as $n, p \to \infty$,

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

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

We comment that Theorem 2 does not require an explicit response model and can accommodate general test error functionals as we show next. Furthermore, we do not require that the aspect ratio p/n converges to a constant. Finally, convergence is strong in the sense that it is uniform along the whole GD path, and convergence occurs conditional on the training data.

Finally, for a fixed $T < \infty$, it is possible to carry out a similar analysis as that in Theorem 2 for gradient flow in the limit as $K \to \infty$ and $\delta \to 0$ while $K\delta \to T$. This is in a similar spirit to that for Theorem 1. We leave the details to the interested readers.

4.2 Extension to General Risk Functionals

We now demonstrate that Theorem 2 can be extended to work with general loss functionals subject to mild regularity conditions. First, we define the general prediction risk functional that we wish to estimate. More precisely, let $\psi : \mathbb{R}^2 \to \mathbb{R}$ be an error function. We define the corresponding prediction risk functional as:

$$\Psi(\widehat{\boldsymbol{\beta}}_k) = \mathbb{E}_{\boldsymbol{x}_0, y_0}\left[\psi(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k) \mid X, y\right]. \tag{12}$$

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

$$\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) := \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i}). \tag{13}$$

Our next theorem establishes that $\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$ is uniformly consistent for $\Psi(\widehat{\boldsymbol{\beta}}_k)$ along the GD path.

Theorem 3 (Functional consistency of LOOCV)**.** *Suppose assumptions of Theorem 2 hold. Moreover, assume that $\psi : \mathbb{R}^2 \to \mathbb{R}$ is differentiable and satisfies $\|\nabla \psi(x)\|_2 \le C_\psi \|x\|_2 + \bar{C}_\psi$ for all $x \in \mathbb{R}^2$ with some nonnegative constants $C_\psi, \bar{C}_\psi$. Then, as $n, p \to \infty$,*

$$\sup_{k \in \{0\} \cup [K]} \left| \widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) - \Psi(\widehat{\boldsymbol{\beta}}_k) \right| \xrightarrow{\text{a.s.}} 0. \tag{14}$$

As a direct consequence of the convergence (14) above, LOOCV can be used to tune the early stopping. Specifically, let $k_* = \arg\min_{k \in \{0\} \cup [K]} \widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$ be the estimated early stopping iteration. Then, as $n, p \to \infty$,

$$\left| \Psi(\widehat{\boldsymbol{\beta}}_{k_*}) - \min_{k \in \{0\} \cup [K]} \Psi(\widehat{\boldsymbol{\beta}}_k) \right| \xrightarrow{\text{a.s.}} 0. \tag{15}$$

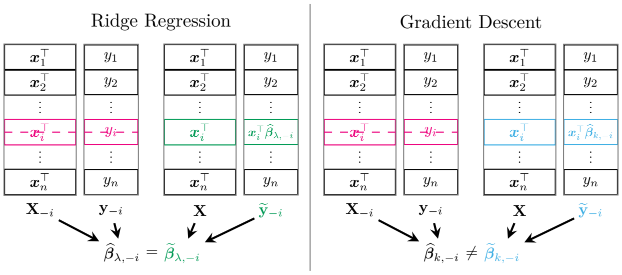
Ridge Regression
Gradient Descent
$\boldsymbol{x}_1^\top$
y_1
$\boldsymbol{x}_2^\top$
y_2
$\boldsymbol{x}_i^\top$
y_i
$\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i}$
$\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i}$
$\boldsymbol{x}_n^\top$
y_n
$\mathbf{X}_{-i}$
$\mathbf{y}_{-i}$
$\mathbf{X}$
$\widetilde{\mathbf{y}}_{-i}$
$\widehat{\boldsymbol{\beta}}_{\lambda,-i} = \widetilde{\boldsymbol{\beta}}_{\lambda,-i}$
$\widehat{\boldsymbol{\beta}}_{k,-i} \neq \widetilde{\boldsymbol{\beta}}_{k,-i}$

Figure 3: Illustration of the differences between LOO augmentations for ridge regression versus gradient descent.

Thanks to Theorem 3, among other things, 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 of the LOOCV residuals $\{y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i}\}_{i \in [n]}$. This allows the construction of prediction intervals of the form $\mathcal{I}_k := [(\widehat{\alpha}_k(q_1), \widehat{\alpha}_k(q_2)]$ that have the right coverage asymptotically almost surely conditional on the training data, as shown next:

Theorem 4 (Coverage guarantee)**.** *Conditioning on $(\boldsymbol{X}, \boldsymbol{y})$, denote by $\widehat{\alpha}_k(q)$ the q-quantile of the uniform distribution over n points $\{y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i} : i \in [n]\}$. Assume the conditions of Theorem 3. Further, assume that the noise ε_i in Assumption D has bounded p.d.f. with an upper bound κ_{pdf} that is independent of (n, p). Then, for any $0 \le q_1 \le q_2 \le 1$, as $n, p \to \infty$,*

$$\sup_{k \in \{0\} \cup [K]} \mathbb{P}_{(\boldsymbol{x}_0, y_0)}\left(y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k \in \mathcal{I}_k \mid \boldsymbol{X}, \boldsymbol{y}\right) \xrightarrow{\text{a.s.}} q_2 - q_1.$$

Note that Theorem 4 provides *conditional*, rather than *marginal*, coverage guarantees for the specific dataset that we observe, which is useful in practice for model selection and tuning. See Figure 2 for an illustration.

5 DISCUSSION

One of the key takeaways from this paper is the discrepancy between GCV and LOOCV along the GD path. While LOOCV is consistent in a strong uniform sense, GCV fails along the entire path, even though both behave very similarly for explicitly regularized estimators like ridge regression. This inconsistency highlights an interesting disconnect between gradient descent and ridge regression, although both have very close relationships in their regularization effects (Ali et al., 2019, 2020). We next discuss the root cause for this disconnect in Section 5.1, which also helps build an efficient exact LOOCV implementation in Section 5.2.

5.1 Understanding GCV Failure

A comparison with ridge regression offers insights into the GCV inconsistency observed in gradient descent. For ridge regression, the LOOCV residuals (the difference between the observed and predicted responses for the left-out observation) can be computed directly from the residuals of the full model (the model fitted

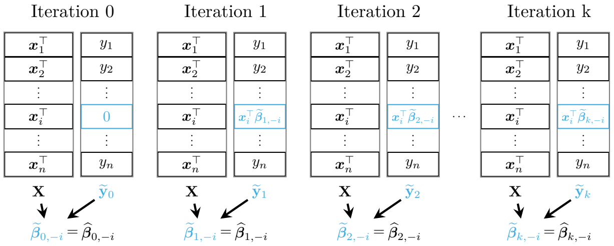
Iteration 0
Iteration 1
Iteration 2
Iteration k
$x_1^\top$
$x_2^\top$
$x_i^\top$
$x_n^\top$
y_1
y_2
y_n
0
$x_i^\top \widetilde{\beta}_{1,-i}$
$x_i^\top \widetilde{\beta}_{2,-i}$
$x_i^\top \widetilde{\beta}_{k,-i}$
X
$\widetilde{\mathbf{y}}_0$
$\widetilde{\mathbf{y}}_1$
$\widetilde{\mathbf{y}}_2$
$\widetilde{\mathbf{y}}_k$
$\widetilde{\beta}_{0,-i} = \widehat{\beta}_{0,-i}$
$\widetilde{\beta}_{1,-i} = \widehat{\beta}_{1,-i}$
$\widetilde{\beta}_{2,-i} = \widehat{\beta}_{2,-i}$
$\widetilde{\beta}_{k,-i} = \widehat{\beta}_{k,-i}$

Figure 4: Illustration of the modified augmentation system.

on all observations). This is established using an elegant "augmentation trick" (Golub et al., 1979b; Hastie, 2020). The augmentation trick involves creating an augmented dataset $(\boldsymbol{X}, \widetilde{\boldsymbol{y}}_{-i})$, where the left-out observation $\boldsymbol{x}_i$ is added back into the i-th leave-one-out dataset $(\boldsymbol{X}_{-i}, \boldsymbol{y}_{-i})$, but with its corresponding response replaced by the leave-one-out prediction $y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i}$ on the i-th observation $\boldsymbol{x}_i$. Let $\widetilde{\boldsymbol{\beta}}_{\lambda,-i}$ denote the ridge estimator fit on the augmented dataset. It is easy to show that the augmentation does not change the solution and we have $\widetilde{\boldsymbol{\beta}}_{\lambda,-i} = \widehat{\boldsymbol{\beta}}_{\lambda,-i}$. On the other hand, for gradient descent, if we construct a similar "augmented" dataset, then the resulting iterates turn out not to be the same as the leave-one-out iterates; i.e., if $\widetilde{\boldsymbol{\beta}}_{k,-i}$ denotes the gradient descent iterate at iteration k on the augmented dataset $(\boldsymbol{X}, \widetilde{\boldsymbol{y}}_{-i})$, we have $\widehat{\boldsymbol{\beta}}_{k,-i} \neq \widetilde{\boldsymbol{\beta}}_{k,-i}$ in general. See Figure 3 for an illustration of these standard augmented systems for ridge and GD.

The underlying reason for this failure is that while the GD iterates are closely related to the solution of the ridge regularized least squares, unlike the ridge, the regularizer now depends on the data also! Identifying this failure also helps us modify the "augmentation" so that the solution to the system at every iteration recovers the leave-one-out estimator (see Proposition 6) and the leave-one-out predictions (see Lemma 7). We showcase the modified augmented system in Figure 4. We leave further details on modified augmentation and why it works to Appendix A due to space constraints.

5.2 Towards Exact Efficient LOOCV

Naively computing LOOCV requires running GD n times, which is computationally expensive. Capitalizing on the modified augmented system, we propose an efficient way to compute exact leave-one-out residuals, which does not require refitting the model n times.

Proposition 5. *The equality below holds for $i \in [n]$:*

$$\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{K,-i} = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_K + A_{i,K} \|\boldsymbol{x}_i\|^2 + \sum_{j=1}^{K-1} B_{i,K}^{(j)} \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^j \boldsymbol{x}_i.$$

Here we define the quantities $\{A_{i,k} : i \in [n], k \in [K]\}$ and $\{B_{i,k}^{(j)} : i \in [n], j \in [k-1], 2 \le k \le K\}$ as follows:

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

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

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

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

Table 1: Computational complexity of LOOCV when $p \asymp n$.

Method	Time	Space
Naive	$O(n^3 K)$	$O(n^2)$
Proposed	$O(n^3 + n^2 K + nK^2)$	$O(n^2 + nK^2)$

By updating $\{A_{i,k} : i \in [n], k \in \{0\} \cup [K]\}$ and $\{B_{i,k}^{(j)} : i \in [n], 1 \le j \le k-1, 2 \le k \le K\}$, we are able to compute the full LOOCV efficiently. More precisely, we propose to estimate the out-of-sample prediction risk functionals via the following formula. Denoting by $\mathcal{H}_i = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^j \boldsymbol{x}_i$ for $i \in [n]$, the estimator (12) for general risk functionals can be computed using:

$$\Psi^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) = \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, x_i^\top \widehat{\boldsymbol{\beta}}_K + A_{i,K} \|\boldsymbol{x}_i\|^2 + \sum_{j=1}^{k-1} B_{i,K}^{(j)} \mathcal{H}_i).$$

Choosing $\psi : (y, z) \mapsto (y - z)^2$ yields the estimator (4).

Our next goal is to assess the computational complexity of the proposed algorithm under the proportional asymptotic regime $p \asymp n$. In fact, we can evaluate $\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^j \boldsymbol{x}_i$ for all $i \in [n]$ and $j + 1 \in [K]$ with computational complexity $O(n^2 K + n^3)$. To be specific, denote by $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{\Omega} \boldsymbol{\Lambda} \boldsymbol{\Omega}^\top$ the spectral decomposition of matrix $\boldsymbol{X}^\top \boldsymbol{X}$ and let $\boldsymbol{Q} = \boldsymbol{X} \boldsymbol{\Omega} \in \mathbb{R}^{n \times (p+1)}$. Implementing the spectral decomposition and computing the matrix $\boldsymbol{Q}$ requires $O(n^3)$ time complexity. We further denote by $\boldsymbol{q}_i \in \mathbb{R}^{p+1}$ the i-th row of $\boldsymbol{Q}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\{\lambda_i\}_{i \le p+1})$. Then we can write $\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^j \boldsymbol{x}_i = \sum_{r=1}^{p+1} q_{ir}^2 \lambda_r^j$, which takes only $O(n)$ time to evaluate for each pair (i, j). Note that the computational complexity of $O(n^3)$ would also be the cost required to implement GCV.

We compare the computational complexity (time and space) of our method with the naive LOOCV in Table 1. When K, for instance, is of order n, the proposed implementation offers a significant time reduction. However, note that the improvement comes at the cost of additional space complexity. It is an interesting open question to construct an exact LOOCV alternative that simultaneously improves on time and space complexity.

In concluding the paper, we briefly highlight two other directions for future work. First, our work focused on GCV and exact LOOCV from both statistical and computational viewpoints. Exploring other approximate LOOCV techniques (Luo et al., 2023; Rad and Maleki, 2020; Stephenson and Broderick, 2020; Xu et al., 2019) and possibly correcting GCV to achieve consistency, especially under overparameterized proportional regimes, is an interesting direction. Second, our results are specifically tailored to squared training loss. Whether our analytical approach can encompass more general loss functions, such as the MLE losses used in generalized linear models, is an open question.

References

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 (extended version). *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. *arXiv preprint arXiv:2103.09177*, 2021.

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: What does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*, 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. *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.

Mikhail Belkin. Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Pierre C. Bellec. Out-of-sample error estimate for robust M-estimators with convex penalty. *arXiv preprint arXiv:2008.11840*, 2020.

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.

Pierre C. Bellec and Cun-Hui Zhang. Second-order Stein: SURE for SURE and other applications in high-dimensional inference. *The Annals of Statistics*, 49(4):1864–1903, 2021.

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, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory*, 2020.

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. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31 (4):377–403, 1978.

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. *The Annals of Statistics*, 46(1):247–279, 2018.

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

Jerome Friedman and Bogdan E. Popescu. Gradient directed regularization. *Unpublished manuscript, http://www-stat. stanford. edu/˜ jhf/ftp/pathlite. pdf*, 2004.

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, 1979a.

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, 1979b.

Nathael Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *The Annals of Probability*, pages 2480–2498, 2009.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 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.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer, 2006.

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. *The 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.

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. *The Annals of Statistics*, pages 1352–1377, 1985.

Ker-Chau Li. Asymptotic optimality of C_ℓ and generalized cross-validation in ridge regression with application to spline smoothing. *The 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. *The 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.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

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

Andrea Montanari and Yuchen Wu. Statistically optimal first order algorithms: A proof via orthogonalization. *arXiv preprint arXiv:2201.05101*, 2022.

Nelson Morgan and Hervé Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. *Advances in neural information processing systems*, 2:630–637, 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.

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 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 (Statistical Methodology)*, 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. *arXiv preprint arXiv:2003.01770*, 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. *arXiv preprint arXiv:2006.06386*, 2020.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5(Aug):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.

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.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

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. *arXiv preprint arXiv:2203.06176*, 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.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.

Ji Xu, Arian Maleki, and Kamiar Rahnama Rad. Consistent risk estimation in high-dimensional linear regression. *arXiv preprint arXiv:1902.01753*, 2019.

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.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The 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.

Supplementary Materials

This document serves as a supplement to the paper “Failures and Successes of Cross-Validation for Early-Stopped Gradient Descent in High-Dimensional Least Squares.” The structure of the supplement is outlined below, followed by a summary of notation used in both the main paper and this supplement.

Organization

- Appendix A provides additional details and proofs of results related to the naive and modified augmentation systems (Propositions 5 and 6 and Lemma 7) for LOOCV along the gradient path, expanding on Section 5.

Section	Content	Purpose
Appendix A.1		Leave-one-out augmentation system for ridge and gradient descent
Appendix A.2	Proposition 6, Lemma 7	Modified leave-one-out augmentation system for gradient descent
Appendix B		Proof of Proposition 5

- Appendix C provides the main steps involved in the proofs of Theorems 1 and 2.

Section	Content	Purpose
Appendix C.1	Lemmas 8 to 12	Proof outline for Theorem 1
Appendix C.2	Lemmas 13 to 17	Proof outline for Theorem 2

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

Section	Content	Purpose
Appendix D.1	Lemma 18	Closeness between gradient descent and flow
Appendix D.2	Lemmas 19 and 20	Statements of concentration results for linear and quadratic forms

- Appendix E contains proof of Theorem 1.

Section	Content	Purpose
Appendix E.1		Proof schematic
Appendix E.2	Lemma 8	Equivalences between gradient descent and flow for risk
Appendix E.3	Lemma 9	Equivalences between gradient descent and flow for GCV
Appendix E.4	Lemma 10	Asymptotics of risk for gradient flow
Appendix E.3	Lemma 11	Asymptotics of GCV for gradient flow
Appendix E.6	Lemma 12	Mismatch of risk and GCV asymptotics for gradient flow

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

Section	Content	Purpose
Appendix G.1	Proposition 23	Useful property of the T_2-inequality
Appendix G.2	Lemma 24	Dimension-free concentration inequality
Appendix G.3	Lemmas 25 and 26	Upper bounds on operator norm of $\widehat{\mathbf{\Sigma}}$ and $\|\boldsymbol{y}\|_2$
Appendix G.4	Lemmas 27 and 28	Upper bounds on $\|\mathbb{E}[y_0 \boldsymbol{x}_0]\|$ and sub-exponential of $\|\widehat{\mathbf{\Sigma}}\|_{\mathrm{op}}$
Appendix G.5	Lemma 29 and Corollary 30	Upper bounds on $\|\widehat{\boldsymbol{\beta}}_k\|_2$ and $\|\widehat{\boldsymbol{\beta}}_{k,-i}\|_2$
Appendix G.6	Lemma 31	Upper bounds on LOOCV residuals $\{\lvert y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i}\rvert\}_{i \in [n]}$

- Appendix H contains proof of Theorem 2.

Section	Content	Purpose
Appendix H.1		Proof schematic
Appendix H.2	Lemma 14	Studying the concentration of LOOCV estimate $\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$
Appendix H.3	Lemma 15	Studying the concentration of LOOCV estimator risk $R(\widehat{\boldsymbol{\beta}}_k)$
Appendix H.4	Lemma 16	Understanding the effect of projection onto Ω on the prediction risk
Appendix H.5	Lemma 17	Showing the prediction risk is stable to the sample size

- Appendix I contains proof of Lemma 13 that forms a key component in the proof of Theorem 2.

Section	Contents	Purpose
Appendix I.1		Proof schematic
Appendix I.2	Lemmas 32 to 34	Upper bounding norm of the gradient with respect to the features
Appendix I.3	Lemma 35	Upper bounding norm of the gradient with respect to the response

- Appendix J contains proof of Theorem 3 for general risk functionals.

Section	Contents	Purpose
Appendix J.1		Proof schematic
Appendix J.2	Lemma 36	Concentration analysis for LOOCV estimator and prediction risk
Appendix J.3	Lemma 37	Demonstrating that projection has little effect on quantities of interest

- Appendix K contains proof of Theorem 4. The proof uses component Lemma 38.
- Appendix L provides an additional numerical illustration and details of the setups for Figures 1 and 2.

Section	Content	Purpose
??		Setup details for Figure 1
??		Setup details for Figure 2
??	Figure 13	Additional illustration for prediction intervals with intermediate optimal stopping

Notation

- **General notation.** We denote scalars in non-bold lower or upper case (e.g., x, X), vectors in bold lower case (e.g., $\boldsymbol{x}$), and matrices in bold upper case (e.g., $\boldsymbol{X}$). We use blackboard letters to denote some special sets: $\mathbb{N}$ denotes the set of positive integers and $\mathbb{R}$ denotes the set of real numbers. For a positive integer n, we use the shorthand $[n]$ to denote the set $\{1, \dots, n\}$. For a pair of real numbers x and y, we use $x \wedge y$ to denote $\min\{x, y\}$, and $x \vee y$ to denote $\max\{x, y\}$. For an event or set A, $\mathbb{1}_A$ denotes the indicator random variable associated with A.
- **Vector and matrix notation.** For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_2$ denotes its ℓ_2 norm. (If no subscript is specified for norm $\|\boldsymbol{x}\|$ of a vector $\boldsymbol{x}$, then it is assumed to be the ℓ_2 norm of $\boldsymbol{x}$.) For $\boldsymbol{v} \in \mathbb{R}^n$ and $k \in \mathbb{N}_+$, we let $\boldsymbol{v}_{1:k} \in \mathbb{R}^k$ be the vector that contains the first k coordinates of $\boldsymbol{v}$. For a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{X}^\top \in \mathbb{R}^{p \times n}$ denotes its transpose, and $\boldsymbol{X}^\dagger \in \mathbb{R}^{p \times n}$ denotes its Moore-Penrose inverse. For a square matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, $\mathrm{tr}[\boldsymbol{A}]$ denotes its trace, and $\boldsymbol{A}^{-1} \in \mathbb{R}^{p \times p}$ denotes its inverse, provided that it is invertible. For a positive semidefinite matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{1/2}$ denotes its principal square root. A $p \times p$ identity matrix is denoted $\boldsymbol{I}_p$, or simply by $\boldsymbol{I}$, when it is clear from the context. For a matrix $\boldsymbol{X}$, we denote its operator norm with respect to ℓ_2 vector norm by $\|\boldsymbol{X}\|_{\mathrm{op}}$ and its Frobenius norm by $\|\boldsymbol{X}\|_F$. For a matrix $\boldsymbol{M}$, $\|\boldsymbol{X}\|_{\mathrm{tr}}$ denotes the trace norm of $\boldsymbol{M}$, which is the sum of all its singular values.
- **Asymptotics notation.** For a nonnegative quantity Y, we use $X = O_\alpha(Y)$ to denote the deterministic big-O notation that indicates the bound $|X| \le C_\alpha Y$, where C_α is some numerical constant that can depend on the ambient parameter α but otherwise does not depend on other parameters in the context. We denote the probabilistic big-O notation by O_p. We denote convergence in probability by "$\xrightarrow{\mathrm{p}}$" and almost sure convergence by "$\xrightarrow{\mathrm{a.s.}}$".

A Additional Details for Section 5

It is instructive to compare the inconsistency results with that for ridge regression. Gradient descent iterates are known to be very closely related to ridge regression (Ali et al., 2019). However, for ridge regression, GCV is uniformly consistent over the regularization parameter (Patil et al., 2021, 2022b). The discrepancy between gradient descent iterates and ridge regression is curious in this regard. Understanding this discrepancy helps shed light on the inconsistency of GCV in Theorem 1.

A.1 Leave-one-out augmentation systems for ridge and gradient descent

For ridge regression, the LOOCV residuals (the difference between the observed and predicted responses for the left-out observation) can be computed directly from the residuals of the full model (the model fit on all observations). This is done using an elegant "augmentation trick" (Golub et al., 1979b; Hastie, 2020).

The augmentation trick involves creating an augmented system where the left-out observation is added back into the leave-one-out system but with its response replaced by the unknown leave-one-out error on that observation. This allows us to express each leave-one-out error in terms of the full smoothing matrix (a matrix that transforms the observed responses into the predicted responses) and the response. In other words, the leave-one-out predictions are linear in the response, just like the predictions of the full estimator. This means that we can compute the LOOCV residuals directly from the residuals of the full model, without having to fit the model n times. We briefly describe the trick below.

For each data point $(\boldsymbol{x}_i, y_i)$ that is left out in the leave-one-out system, we need to solve the problem:

$$\widehat{\boldsymbol{\beta}}_{\lambda,-i} := \arg\min \|\boldsymbol{y}_{-i} - \boldsymbol{X}_{-i}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2. \tag{16}$$

The predicted value of leave-one-out for the i-th observation is $\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i}$. Let us now imagine that we "augment" the data point $(\boldsymbol{x}_i, \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i})$ into the leave-one-out dataset $(\boldsymbol{X}_{-i}, \boldsymbol{y}_{-i})$ (which has size $n-1$). The augmented dataset has feature matrix $\boldsymbol{X}$. Let $\widetilde{\boldsymbol{y}}_{-i} \in \mathbb{R}^n$ denote the augmented response vector. Let $\widetilde{\boldsymbol{\beta}}_{\lambda,-i}$ denote the ridge estimator fit on the augmented dataset $(\boldsymbol{X}, \widetilde{\boldsymbol{y}}_{-i})$. In other words,

$$\widetilde{\boldsymbol{\beta}}_{\lambda,-i} := \arg\min \|\widetilde{\boldsymbol{y}}_{-i} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\beta\|_2^2. \tag{17}$$

It is easy to show that the augmentation does not change the solution to (16) and we have $\widetilde{\boldsymbol{\beta}}_{\lambda,-i} = \widehat{\boldsymbol{\beta}}_{\lambda,-i}$. Briefly, this is because the newly added point $(\boldsymbol{x}_i, \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i})$ yields zero loss on the objective function of (16). Now, let us consider the augmented dataset (of size n) again. Since ridge regression is a linear smoother, we can write the predicted value on the i-th observation as follows:

$$\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i} = \sum_{j \neq i} [\boldsymbol{H}_\lambda]_{ij} y_j + [\boldsymbol{H}_\lambda]_{ii} \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i} \tag{18}$$

Here, $\boldsymbol{H}_\lambda \in \mathbb{R}^{n \times n}$ is the smoothing matrix associated with ridge regression at regularization level λ. Rearranging, we have

$$\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i} = \frac{\sum_{j \neq i} [\boldsymbol{H}_\lambda]_{ij} y_j}{1 - [\boldsymbol{H}_\lambda]_{ii}}. \tag{19}$$

One can equivalently write this in terms of residuals as follows:

$$y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i} = \frac{y_i - \sum_j [\boldsymbol{H}_\lambda]_{ij} y_j}{1 - [\boldsymbol{H}_\lambda]_{ii}} = \frac{y_i - \widehat{y}_i}{1 - [\boldsymbol{H}_\lambda]_{ii}}. \tag{20}$$

From (20), we can see that the difference between the actual response y_i and the predicted response $\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\lambda,-i}$ is equal to the difference between the actual response y_i and the predicted response $\widehat{y}_i$ (from the full model), divided by $1 - [\boldsymbol{H}_\lambda]_{ii}$. Here, $\widehat{y}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_\lambda$ is the predicted response from the full model.

For gradient descent iterates, if we construct a similar "augmented" system by replacing the response for the i-th feature, the resulting estimator is actually not the same as the leave-one-out estimator. More precisely, let $\widehat{\boldsymbol{\beta}}_{k,-i}$ be the gradient descent iterate at step k run on the leave-one-out dataset $(\boldsymbol{X}_{-i}, \boldsymbol{y}_{-i})$. As before, imagine that we "augment" the data point $(\boldsymbol{x}_i, \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i})$ into the leave-one-out dataset $(\boldsymbol{X}_{-i}, \boldsymbol{y}_{-i})$. The augmented dataset again has the feature matrix $\boldsymbol{X}$. and let $\widetilde{\boldsymbol{y}}_{-i} \in \mathbb{R}^n$ denote the augmented response vector. Let $\widetilde{\boldsymbol{\beta}}_{k,-i}$ denote the ridge estimator fit on the augmented dataset $(\boldsymbol{X}, \widetilde{\boldsymbol{y}}_{-i})$. In general, we have $\widehat{\boldsymbol{\beta}}_{k,-i} \neq \widetilde{\boldsymbol{\beta}}_{k,-i}$. The underlying reason for this is that, even though the iterates (2) can be written as a solution to the regularized least-squares problem, unlike ridge regression, the regularizer depends on the data. For instance, assuming equal step sizes δ, the k-th iterate is a solution to the problem:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \ \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2/2n + \boldsymbol{\beta}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}/n((\boldsymbol{I}_p - \delta \boldsymbol{X}^\top \boldsymbol{X}/n)^k - \boldsymbol{I}_p)^{-1} \right) \boldsymbol{\beta}.$$

Note that the regularization term is a function of both δ and k (tuning parameters), as well as the feature matrix $\boldsymbol{X}$.

A.2 Modified augmentation system for gradient descent

Identifying the failure of the LOOCV augmentation system for gradient descent in Appendix A.1 also helps us to modify the "augmentation" so that the solution to the system at every iteration recovers the leave-one-out estimator.

For $k \in \{0\} \cup [K]$ and $i, j \in [n]$, we let

$$\widetilde{y}_{ij}^{(k)} = \begin{cases} y_j, & j \neq i, \\ \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i}, & j = i, \end{cases}$$

and define $\widetilde{\boldsymbol{y}}_i^{(k)} = (\widetilde{y}_{ij}^{(k)})_{j \le n}$. Let $\widetilde{\boldsymbol{\beta}}_{k,-i}$ be the iterate obtained by running the gradient descent on $(\boldsymbol{X}, \widetilde{\boldsymbol{y}}_i^{(k)})$. The following proposition shows that $\widetilde{\boldsymbol{\beta}}_{k,-i} = \widehat{\boldsymbol{\beta}}_{k,-i}$.

Proposition 6. *For all $k \in [K]$ and $i \in [n]$, we have $\widetilde{\boldsymbol{\beta}}_{k,-i} = \widehat{\boldsymbol{\beta}}_{k,-i}$.*

Proof of Proposition 6. We prove the lemma through induction over k. For $k = 0$, by definition $\widetilde{\boldsymbol{\beta}}_{0,-i} = \widehat{\boldsymbol{\beta}}_{0,-i} = \boldsymbol{\beta}_0$ for all $i \in [n]$. Suppose that we have $\widetilde{\boldsymbol{\beta}}_{k,-i} = \widehat{\boldsymbol{\beta}}_{k,-i}$ 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

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_{k+1,-i} &= \widetilde{\boldsymbol{\beta}}_{k,-i} - \frac{2\delta_{k+1}}{n} \boldsymbol{X}^\top \boldsymbol{X} \, \widetilde{\boldsymbol{\beta}}_{k,-i} + \frac{2\delta_{k+1}}{n} \boldsymbol{X}^\top \widetilde{\boldsymbol{y}}_i^{(k)} \\
&= \widetilde{\boldsymbol{\beta}}_{k,-i} - \frac{2\delta_{k+1}}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{X}_{-i} \, \widetilde{\boldsymbol{\beta}}_{k,-i} + \frac{2\delta_{k+1}}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{y}_{-i} - \frac{2\delta_{k+1}}{n} \boldsymbol{x}_i \big(\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_{k,-i} - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i} \big)
\end{aligned}$$

$$
\begin{aligned}
&= \widehat{\boldsymbol{\beta}}_{k,-i} - \frac{2\delta_{k+1}}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{X}_{-i} \widehat{\boldsymbol{\beta}}_{k,-i} + \frac{2\delta_{k+1}}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{y}_{-i} \\
&= \widehat{\boldsymbol{\beta}}_{k+1,-i},
\end{aligned}
$$

thus completing the proof of the lemma by induction. □

In other words, what Proposition 6 shows is that, to obtain the i-th leave-one-out prediction at time step k, one needs to augment the system so that at every time step before k in the trajectory the i-th response is the "correct" leave-one-out prediction. See Figure 4 for an illustration. This is in contrast to setting the i-th response to the leave-one-out prediction step k naively for each step before, as illustrated in Appendix A.1. Our new shortcut for exactly computing LOOCV, presented in Section 5, builds on this idea.

Lemma 7. *For all $k \in [K]$, there exist $(h_{ij}^{(k)})_{i,j \le n}$ and $(b_i^{(k)})_{i \le n}$ that depend uniquely on $(\boldsymbol{\beta}_0, \boldsymbol{\delta}_{1:t}, \boldsymbol{X}, k)$, such that for all $i \in [n]$, the following holds:*

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

Proof of Lemma 7. We prove this lemma by induction over k. For the base case $k = 0$, the requirement of the lemma can be satisfied by setting

$$
h_{ij}^{(0)} = 0, \qquad b_i^{(0)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0, \qquad i, j \in [n].
$$

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 $\boldsymbol{H}^{(k)} \in \mathbb{R}^{n \times n}$, $\boldsymbol{b}^{(k)} \in \mathbb{R}^n$, such that $\boldsymbol{H}_{ij}^{(k)} = h_{ij}^{(k)}$ and $\boldsymbol{b}_i^{(k)} = b_i^{(k)}$. Using induction hypothesis and Proposition 6, we have

$$
\begin{aligned}
&\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i} \\
&= \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_{k,-i} \\
&= \boldsymbol{x}_i^\top \left(\widetilde{\boldsymbol{\beta}}_{k,-i} - \frac{2\delta_{k+1}}{n} \boldsymbol{X}^\top \boldsymbol{X} \widetilde{\boldsymbol{\beta}}_{k,-i} + \frac{2\delta_{k+1}}{n} \boldsymbol{X}^\top \widetilde{\boldsymbol{y}}_i^{(k)} \right) \\
&= \boldsymbol{x}_i^\top \left(\widehat{\boldsymbol{\beta}}_{k,-i} - \frac{2\delta_{k+1}}{n} \boldsymbol{X}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}}_{k,-i} + \frac{2\delta_{k+1}}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{y}_{-i} + \frac{2\delta_{k+1}}{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i} \right) \\
&= \sum_{j=1}^n h_{ij}^{(k)} y_j + b_i^{(k)} - \frac{2\delta_{k+1}}{n} \boldsymbol{x}_i^\top \boldsymbol{X}^\top (\boldsymbol{H}^{(k)} \boldsymbol{y} + \boldsymbol{b}^{(k)}) + \frac{2\delta_{k+1}}{n} \boldsymbol{x}_i^\top \boldsymbol{X}_{-i}^\top \boldsymbol{y}_{-i} + \frac{2\delta_{k+1}}{n} \|\boldsymbol{x}_i\|_2^2 \left(\sum_{j=1}^n h_{ij}^{(k)} y_j + b_i^{(k)} \right).
\end{aligned}
$$

Note that right-hand of the display above is affine in $\boldsymbol{y}$, which completes the proof for iteration $k+1$. This completes our induction proof. □

With the new augmentation, the leave-one-out predictions are linear in the response. Thus, what Lemma 7 shows is that there exist a smoothing matrix that allows one to express the exact leave-one-out score. This perspective does not immediately suggest a computationally efficient way to compute the LOO residuals. However, we can directly unroll the LOO residuals and construct an estimator that is faster than the naive implementation of the LOO estimator. This is done in Proposition 5, whose proof we provide next.

B Proof of Proposition 5

By definition, $\widehat{\boldsymbol{\beta}}_{0,-i} = \widehat{\boldsymbol{\beta}}_0$ for all $i \in [n]$. After implementing the first step of gradient descent, we have

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_{1,-i} =& \widehat{\boldsymbol{\beta}}_{0,-i} - \frac{2\delta_1}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{X}_{-i} \widehat{\boldsymbol{\beta}}_{0,-i} + \frac{2\delta_1}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{y}_{-i} \\
=& \widehat{\boldsymbol{\beta}}_1 + \frac{2\delta_1}{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_0 - \frac{2\delta_1}{n} y_i \boldsymbol{x}_i.
\end{aligned}
$$

We define $A_{i,1} = 2\delta_1(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_0 - y_i)/n$, then $\widehat{\boldsymbol{\beta}}_{1,-i} = \widehat{\boldsymbol{\beta}}_1 + A_{i,1}\boldsymbol{x}_i$. Now suppose $\widehat{\boldsymbol{\beta}}_{k,-i}$ admits the decomposition

$$\widehat{\boldsymbol{\beta}}_{k,-i} = \widehat{\boldsymbol{\beta}}_k + A_{i,k}\boldsymbol{x}_i + \sum_{j=1}^{k-1} B_{i,k}^{(j)}(\boldsymbol{X}^\top\boldsymbol{X})^j\boldsymbol{x}_i$$

for some $A_{i,k}, B_{i,k}^{(j)} \in \mathbb{R}$. Then, in the next step of gradient descent, by definition we have

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{k+1,-i} =& \widehat{\boldsymbol{\beta}}_{k,-i} - \frac{2\delta_{k+1}}{n}\boldsymbol{X}_{-i}^\top\boldsymbol{X}_{-i}\widehat{\boldsymbol{\beta}}_{k,-i} + \frac{2\delta_{k+1}}{n}\boldsymbol{X}_{-i}^\top\boldsymbol{y}_{-i} \\ =& \widehat{\boldsymbol{\beta}}_{k+1} + A_{i,k}\boldsymbol{x}_i + \sum_{j=1}^{k-1} B_{i,k}^{(j)}(\boldsymbol{X}^\top\boldsymbol{X})^j\boldsymbol{x}_i - \frac{2\delta_{k+1}A_{i,k}}{n}\boldsymbol{X}^\top\boldsymbol{X}\,\boldsymbol{x}_i - \sum_{j=1}^{k-1}\frac{2\delta_{k+1}B_{i,k}^{(j)}}{n}(\boldsymbol{X}^\top\boldsymbol{X})^{j+1}\boldsymbol{x}_i \\ &+ \frac{2\delta_{k+1}A_{i,k}\|\boldsymbol{x}_i\|_2^2}{n}\boldsymbol{x}_i + \sum_{j=1}^{k-1}\frac{2\delta_{k+1}B_{i,k}^{(j)}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^j\boldsymbol{x}_i}{n}\boldsymbol{x}_i + \frac{2\delta_{k+1}(\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_k - y_i)}{n}\boldsymbol{x}_i.\end{aligned}$$

As a result, we obtain the following update equations

$$\begin{aligned} A_{i,k+1} &= A_{i,k} + \frac{2\delta_{k+1}A_{i,k}\|\boldsymbol{x}_i\|_2^2}{n} + \sum_{j=1}^{k-1}\frac{2\delta_{k+1}B_{i,k}^{(j)}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^j\boldsymbol{x}_i}{n} + \frac{2\delta_{k+1}(\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_k - y_i)}{n}, \\ B_{i,k+1}^{(1)} &= B_{i,k}^{(1)} - \frac{2\delta_{k+1}A_{i,k}}{n}, \\ B_{i,k+1}^{(j)} &= B_{i,k}^{(j)} - \frac{2\delta_{k+1}B_{i,k}^{(j-1)}}{n}, \quad 2 \le j \le k, \end{aligned}$$

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

C Proof Sketches for Theorems 1 and 2

This section contains proof sketches for Theorems 1 and 2.

C.1 Proof Sketch for Theorem 1

In this section, we outline the proof idea of Theorem 1. A detailed proof of the theorem can be found in Section E.

Step 1: Closeness between gradient descent and gradient flow. This step involves demonstrating certain equivalences between gradient descent and gradient flow, particularly in terms of risk and GCV. We can rearrange the terms in (2) to write:

$$\frac{\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k-1}}{\delta} = \frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{k-1}). \tag{21}$$

A continuous-time analogue of Equation (21) is stated as follows:

$$\frac{\partial}{\partial t}\widehat{\boldsymbol{\beta}}_t^{\mathrm{gf}} = \frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_t^{\mathrm{gf}}), \qquad \widehat{\boldsymbol{\beta}}_0^{\mathrm{gf}} = 0. \tag{22}$$

In our case, the GF estimate has a closed-form solution:

$$\widehat{\boldsymbol{\beta}}_t^{\mathrm{gf}} = \widehat{\boldsymbol{\Sigma}}^\dagger\left(\boldsymbol{I}_p - \exp(-t\widehat{\boldsymbol{\Sigma}})\right)\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{y}, \tag{23}$$

where $\widehat{\boldsymbol{\Sigma}}^\dagger$ stands for the Moore-Penrose generalized inverse of $\widehat{\boldsymbol{\Sigma}}$. Also, by unrolling the iterations, the gradient descent iterate at step k can be expressed as:

$$\widehat{\boldsymbol{\beta}}_k = \sum_{j=0}^{k-1}\delta\left(\boldsymbol{I}_p - \delta\widehat{\boldsymbol{\Sigma}}\right)^{k-j-1}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{y}. \tag{24}$$

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

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

where

$$\boldsymbol{H}_k = \sum_{j=0}^{k-1} \frac{\delta}{n} \boldsymbol{X} \left(\boldsymbol{I}_p - \delta\widehat{\boldsymbol{\Sigma}}\right)^{k-j-1} \boldsymbol{X}^\top, \qquad \boldsymbol{H}_t^{\mathrm{gf}} = \frac{1}{n} \boldsymbol{X}(\widehat{\boldsymbol{\Sigma}})^\dagger \cdot \left(\boldsymbol{I}_p - \exp(-t\widehat{\boldsymbol{\Sigma}})\right) \boldsymbol{X}^\top. \tag{25}$$

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*

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

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

$$\left|\widehat{R}^{\mathrm{gcv}}(\widehat{\boldsymbol{\beta}}_K) - \widehat{R}^{\mathrm{gcv}}(\widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}})\right| \xrightarrow{\mathrm{a.s.}} 0.$$

The proofs of these equivalences in Lemmas 8 and 9 are provided in Appendices E.2 and E.3, respectively.

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, in order 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 $R(\widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}})$ and $\widehat{R}^{\mathrm{gcv}}(\widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}})$, respectively.

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

- For $\zeta_* \leq 1$, the density is given by:

$$\frac{\mathrm{d}F_{\zeta_*}(s)}{\mathrm{d}s} = \frac{1}{2\pi\zeta_* s} \sqrt{(b-s)(s-a)} \cdot \mathbb{1}_{[a,b]}(s). \tag{26}$$

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

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

$$\frac{\mathrm{d}F_{\zeta_*}(s)}{\mathrm{d}s} = \left(1 - \frac{1}{\zeta_*}\right)\delta_0(s) + \frac{1}{2\pi\zeta_* s}\sqrt{(b-s)(s-a)} \cdot \mathbb{1}_{[a,b]}(s). \tag{27}$$

  Here, δ_0 is the Dirac delta function at 0.

Lemma 10. *Under the assumptions of Theorem 1,*

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

Lemma 11. *Under the assumptions of Theorem 1,*

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

The proofs of these asymptotic limits in Lemmas 10 and 11 are provided in Appendix E.4 and **??**, respectively.

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,*

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

The proof of this asymptotics mismatch in Lemma 12 is provided in Appendix E.6.

C.2 Proof Sketch for Theorem 2

In this section, we outline the proof idea of Theorem 2. Extension to general test functionals can be found in Section J.

Step 1: LOO concentration. The most challenging part of our proof amounts to establishing concentration for $\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$. This is achieved by upper bounding the norm of the gradient of the mapping $(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n) \mapsto \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$, where $\boldsymbol{w}_i := (\boldsymbol{x}_i, y_i)$. As we will see, this mapping is not exactly a Lipschitz one, but is only 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 $f_k : \mathbb{R}^{n(p+2)} \mapsto \mathbb{R}$ such that $f_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n) = \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$. Namely, we are interested in upper bounding $\|\nabla f_k\|_2$. As will become clear soon, f_k is Lipschitz continuous on a closed convex set Ω. We give definition for Ω below:

$$\Omega = \left\{ \|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}} \le C_{\Sigma,\zeta},\ \|\boldsymbol{y}\|_2^2 \le n(m + \log n) \right\}, \tag{28}$$

where $C_{\Sigma,\zeta} := 2C_0\sigma_\Sigma(1+\zeta) + 1$, $m := m_2$ and $C_0 > 0$ is a numerical constant. One can verify that Ω is a convex set of the data. Standard concentration results (see Lemmas 25 and 26) imply that with an appropriately selected C_0, it holds that $\mathbb{P}(\Omega) \ge 1 - 2(n+p)^{-4} - n^{-1}m_4 \log^{-2} n$. Namely, for large (n,p), with high probability the input samples will fall inside Ω.

In the sequel, we prove that f_k when restricted to Ω is Lipschitz continuous with respect to the input data. Since projection onto closed convex set is 1-Lipschitz, this is equivalently saying that if we first project the data onto Ω then apply f_k, then the whole procedure consists of a Lipschitz continuous mapping. As mentioned before, we prove such result by upper bounding the Euclidean norm of the gradient, which is detailed by lemma 13 below. We defer the proof of Lemma 13 to Section I.

Lemma 13. *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*

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

for all $k \in \{0\} \cup [K]$. In the above display, $\boldsymbol{W} := (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)$ 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 13 implies that $\widetilde{f}_k$ is a Lipschitz continuous mapping with Lipschitz constant as stated in Lemma 13. By assumption, the input data distribution satisfies a T_2-inequality, hence we can apply a powerful concentration inequality as stated in Proposition 23 in the appendix to obtain the desired concentration result. We state such result as Lemma 14 below, the proof of which can be found in Section H.2.

Lemma 14. *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 $k \in \{0\} \cup [K]$*

$$\left| \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) - \mathbb{E}[\widetilde{f}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] \right| \le \frac{2\sigma_{\mathsf{T}_2} LK\xi(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot (\log n)^{3/2}}{\sqrt{n}},$$

where $L := (L_f^2\sigma_\Sigma + L_f^2 + \sigma_\Sigma)^{1/2}$, $\sigma_{\mathsf{T}_2}^2 := \sigma_z^2 \vee \sigma_\varepsilon^2$, and C_{T_2} is a positive numerical constant that appears in proposition 23.

Step 2: Risk concentration. In our second part, we shall provide concentration bounds for the prediction risk $R(\widehat{\boldsymbol{\beta}}_k)$. We establish this step following a similar route as in Step 1. More precisely, we first prove that $R(\widehat{\boldsymbol{\beta}}_k)$ is with high probability a Lipschitz function of the input data. Having established this result, we shall once again use the assumption that the data distribution satisfies a T_2-inequality and apply Proposition 23 to derive a concentration result. We state this result as Lemma 15. The proof of the lemma can be found in Section H.3.

Lemma 15. *We write $R(\widehat{\boldsymbol{\beta}}_k) = r_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)$ and define $\widetilde{r}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n) = r_k(h(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n))$. 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*

$$\Big|R(\widehat{\boldsymbol{\beta}}_k) - \mathbb{E}[\widetilde{r}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]\Big| \le \frac{2\sigma_{\mathsf{T}_2}L\bar{\xi}(C_{\Sigma,\zeta}, \Delta, m, B_0)(\log n)^{3/2}}{\sqrt{n}},$$

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

Step 3: LOO bias analysis. In Steps 1 and 2, we have proved concentration results for both $R(\widehat{\boldsymbol{\beta}}_k)$ and $\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$. To be precise, we have obtained that $R(\widehat{\boldsymbol{\beta}}_k)$ concentrates around $\mathbb{E}[\widetilde{r}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]$ and $\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$ concentrates around $\mathbb{E}[\widetilde{f}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]$, both are expectations of the target functionals composite the projection h.

Next, we show that involving the projection h in the expectation does not change too much the quantities that we are interested in. We present this result as Lemma 16 below.

Lemma 16. *Under the assumptions of Theorem 2, it holds that*

$$\begin{aligned} \sup_{k\in\{0\}\cup[K]} |\mathbb{E}[\widetilde{r}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] - \mathbb{E}[r_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]| = o_n(1), \\ \sup_{k\in\{0\}\cup[K]} \Big|\mathbb{E}[\widetilde{f}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] - \mathbb{E}[f_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]\Big| = o_n(1). \end{aligned} \tag{29}$$

Finally, we show that the prediction risk is stable to sample size. To be specific, in this part we aim to establish a result of the form $\mathbb{E}[R(\widehat{\boldsymbol{\beta}}_k)] \approx \mathbb{E}[R(\widehat{\boldsymbol{\beta}}_{k,-1})]$. This is equivalently saying $\mathbb{E}[r_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] \approx \mathbb{E}[f_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]$.

Formally speaking, we show the following lemma:

Lemma 17. *Under the assumptions of Theorem 2, it holds that*

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

This is equivalently saying

$$\sup_{k\in\{0\}\cup[K]} |\mathbb{E}[r_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] - \mathbb{E}[f_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]| = o_n(1).$$

We defer the proofs of lemma 16 and lemma 17 to Sections H.4 and H.5, respectively.

Theorem 2 then follows from these three steps. To be precise, by putting together Lemmas 14 to 17, 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

$$\sup_{k\in\{0\}\cup[K]} \Big|R(\widehat{\boldsymbol{\beta}}_k) - \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)\Big| \le \frac{2\sigma_{\mathsf{T}_2}LK\xi(C_{\Sigma,\zeta}, \Delta, m, B_0)\cdot(\log n)^{3/2} + 2\sigma_{\mathsf{T}_2}L\bar{\xi}(C_{\Sigma,\zeta}, \Delta, m, B_0)(\log n)^{3/2}}{\sqrt{n}}. \tag{30}$$

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 (30) occurs only finitely many times.

D Supporting lemmas for the proof of Theorem 1

D.1 Connections between gradient descent and gradient flow

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).

We next establish connections between GD and GF. This step is achieved by showing that the hat matrices as defined in Equation (25) when $k = K$ and $t = T$ are close under matrix operator norm.

Observe that the two matrices in Equation (25) share a common set of eigenvectors, and the eigenvalues are obtained by applying separate scalar transformations to the eigenvalues of $\widehat{\mathbf{\Sigma}}$. Hence, to show that $\boldsymbol{H}_K$ and $\boldsymbol{H}_T^{\mathrm{gf}}$ 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 18 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 (25). 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 18. *We assume $K \to \infty$, $\delta \to 0$, and $K\delta \to T$. Here, T is a fixed positive constant. Then it holds that*

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

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

We can apply Lemma 18 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 E.3 and E.2, respectively.

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

$$\sup_{x \in J_{\zeta_*}} \Big| K \log(1 - \delta x) + K\delta x \Big| \to 0. \tag{31}$$

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 (31). This further implies the following uniform convergence result:

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

As a direct consequence of the above equation, we obtain

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

which further gives

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

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

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

Therefore,

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

Further, we note that $\sup_{x \in J_{\zeta_*}} \big| \int_0^{K\delta} x\exp(-zx)\,\mathrm{d}z - \int_0^T x\exp(-zx)\,\mathrm{d}z \big| \to 0$ and $\int_0^T x\exp(-zx)\,\mathrm{d}z = 1 - \exp(-Tx)$. Hence, we have completed the proof of the lemma. □

D.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 19 (Concentration of linear form with independent components)**.** *Let $\boldsymbol{z}_p \in \mathbb{R}^p$ 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 $\boldsymbol{a}_p \in \mathbb{R}^p$ be a sequence of random vectors independent of $\boldsymbol{z}_p$ such that $\limsup_p \|\boldsymbol{a}_p\|^2/p \leq M_0$ almost surely for a constant $M_0 < \infty$. Then $\boldsymbol{a}_p^\top \boldsymbol{z}_p/p \to 0$ almost surely as $p \to \infty$.*

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

Lemma 20 (Concentration of quadratic form with independent components)**.** *Let $\boldsymbol{z}_p \in \mathbb{R}^p$ 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 $\mathbf{D}_p \in \mathbb{R}^{p \times p}$ be a sequence of random matrix such that* $\limsup \|\mathbf{D}_p\|_{\mathrm{op}} \leq M_0$ *almost surely as $p \to \infty$ for some constant $M_0 < \infty$. Then $\boldsymbol{z}_p^\top \mathbf{D}_p \boldsymbol{z}_p/p - \mathrm{tr}[\mathbf{D}_p]/p \to 0$ almost surely as $p \to \infty$.*

E Proof of Theorem 1

E.1 Schematic of the proof

A visual schematic for the proof of Theorem 1 is provided in Figure 5. The lemmas appear in the figure shall be introduced in later parts of this section.

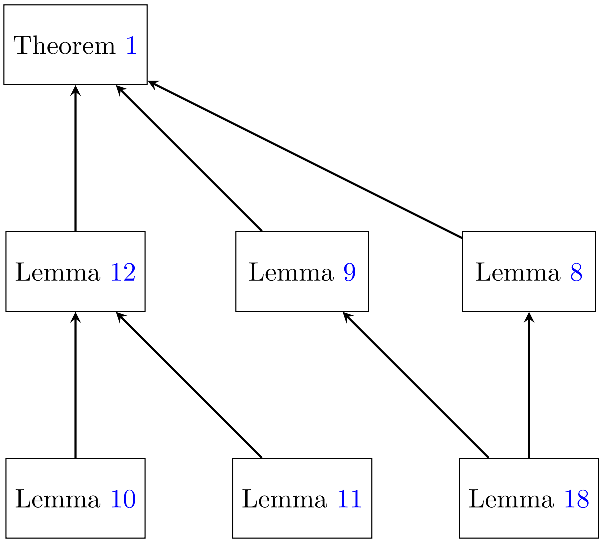

Figure 5: Schematic for the proof of Theorem 1

E.2 Proof of Lemma 8

Note that the prediction risks admit the following expressions:

$$R(\widehat{\boldsymbol{\beta}}_K) = \|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_K\|_2^2 + \sigma^2, \qquad R(\widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}}) = \|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}}\|_2^2 + \sigma^2.$$

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

$$\|x^{1/2}(\bar{g}_{\delta,K}(x) - \bar{g}_T(x))\mathbb{1}_{x \in J_{\zeta_*}}\|_\infty \to 0 \tag{32}$$

under the asymptotics $K \to \infty$, $\delta \to 0$, and $K\delta \to T$. Proof for this claim is similar to that for Lemma 18, and we skip it for the compactness of presentation.

We note that

$$\widehat{\boldsymbol{\beta}}_K - \widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}} = \frac{1}{\sqrt{n}} \boldsymbol{V}^\top \Big(\bar{g}_{\delta,K}(\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}) - \bar{g}_T(\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda})\Big) \boldsymbol{\Lambda}^\top \boldsymbol{U}^\top \boldsymbol{y}, \tag{33}$$

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

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

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 (32)). By the law of large numbers, we obtain $\|\boldsymbol{y}\|_2/\sqrt{n} \xrightarrow{\text{a.s.}} \mathbb{E}[y_1^2]^{1/2}$. Plugging these results into Equation (33) gives $\|\widehat{\boldsymbol{\beta}}_K - \widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}}\|_2 \xrightarrow{\text{a.s.}} 0$ as $n, p \to \infty$. Furthermore, by Equations (23) and (24) we have

$$\begin{aligned} \|\widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}}\|_2 &\le \max_{i\in[n]} \lambda_i^{1/2} \cdot \bar{g}_T\Big(\max_{i\in[n]} \lambda_i\Big) \cdot \frac{1}{\sqrt{n}} \|\boldsymbol{y}\|_2, \\ \|\widehat{\boldsymbol{\beta}}_K\|_2 &\le \max_{i\in[n]} \lambda_i^{1/2} \cdot \bar{g}_{\delta,K}\Big(\max_{i\in[n]} \lambda_i\Big) \cdot \frac{1}{\sqrt{n}} \|\boldsymbol{y}\|_2. \end{aligned} \tag{34}$$

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

$$\Big| \|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_K\|_2^2 - \|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}}\|_2^2 \Big| \xrightarrow{\text{a.s.}} 0$$

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

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

E.3 Proof of Lemma 9

In the sequel, we will apply lemma 18 to prove closeness between $\widehat{R}^{\mathrm{gcv}}(\widehat{\boldsymbol{\beta}}_K)$ and $\widehat{R}^{\mathrm{gcv}}(\widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}})$. This consists of proving the following three pairs of quantities are close: (1) $(1-\mathrm{tr}(\boldsymbol{H}_K)/n)^{-2}$ and $(1-\mathrm{tr}(\boldsymbol{H}_T^{\mathrm{gf}})/n)^{-2}$, (2) $\widehat{\boldsymbol{\beta}}_K^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_K$ and $(\widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}})^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}}$, (3) $\boldsymbol{y}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}}_K/n$ and $\boldsymbol{y}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}}/n$. 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 $\widehat{\boldsymbol{\Sigma}}$. By (**?**, Theorem 5.8), we know that $\max_{i\in[n]} \lambda_i \xrightarrow{\text{a.s.}} (1+\sqrt{\zeta_*})^2$. Note that

$$\frac{1}{n}\mathrm{tr}(\boldsymbol{H}_K) = \frac{1}{n}\sum_{i=1}^n g_{\delta,K}(\lambda_i), \qquad \frac{1}{n}\mathrm{tr}(\boldsymbol{H}_T^{\mathrm{gf}}) = \frac{1}{n}\sum_{i=1}^n g_T(\lambda_i).$$

Invoking Lemma 18, we obtain that with probability one

$$\limsup_{n,p\to\infty} \frac{1}{n} \left| \mathrm{tr}(\boldsymbol{H}_K) - \mathrm{tr}(\boldsymbol{H}_T^{\mathrm{gf}}) \right| \le \sup_{0\le x\le \zeta_* + 2\sqrt{\zeta_*}+2} \left| g_{\delta,K}(x) - g_T(x) \right|,$$

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

Let $F_{\zeta_*}(s)$ denote the Marchenko-Pasture law such that

- for $\zeta_* \le 1$:

  $$\frac{dF_{\zeta_*}(s)}{ds} = \frac{1}{2\pi\zeta_* s}\sqrt{(b-s)(s-a)}, \tag{35}$$

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

- for $\zeta_* > 1$:

  the density of F_{ζ_*} has an additional point mass at 0 of probability $1 - 1/\zeta_*$.

Standard result in random matrix theory Bai and Silverstein (2010) tells us that the empirical spectral distribution (ESD) of $\widehat{\boldsymbol{\Sigma}}$ 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 are able to deduce that both $(1 - \operatorname{tr}(\boldsymbol{H}_K)/n)^{-2}$ and $(1 - \operatorname{tr}(\boldsymbol{H}_T^{\text{gf}})/n)^{-2}$ converge almost surely to one finite constant,hence concluding the proof for this part.

Closeness result (2)

We denote by $\boldsymbol{X}/\sqrt{n} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}$ the singular value decomposition of $\boldsymbol{X}/\sqrt{n}$, where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ are orthogonal matrices. Combining Equations (23) and (24), we arrive at the following equation:

$$\widehat{\boldsymbol{\beta}}_K^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_K - (\widehat{\boldsymbol{\beta}}_T^{\text{gf}})^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_T^{\text{gf}} = \boldsymbol{y}^\top \boldsymbol{U}^\top \cdot \big\{ g_{\delta,K}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^2 - g_T(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^2 \big\} \cdot \boldsymbol{U}\boldsymbol{y}/n. \tag{36}$$

By the strong law of large numbers we have $\|\boldsymbol{y}\|_2^2/n \xrightarrow{\text{a.s.}} \mathbb{E}[y_1^2]$. By Lemma 18 and the fact that $\max_{i\in[n]} \lambda_i \xrightarrow{\text{a.s.}} (1 + \sqrt{\zeta_*})^2$, we conclude that

$$\big\| g_{\delta,K}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^2 - g_T(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^2 \big\|_{\text{op}} \xrightarrow{\text{a.s.}} 0.$$

Plugging these arguments into Equation (36), we obtain

$$\Big| \widehat{\boldsymbol{\beta}}_K^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_K - (\widehat{\boldsymbol{\beta}}_T^{\text{gf}})^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_T^{\text{gf}} \Big| \xrightarrow{\text{a.s.}} 0,$$

which concludes the proof of closeness result (2).

Closeness result (3)

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

$$\frac{1}{n}\big(\boldsymbol{y}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}}_K - \boldsymbol{y}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}}_T^{\text{gf}} \big) = \boldsymbol{y}^\top \boldsymbol{U}^\top \cdot \big\{ g_{\delta,K}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top) - g_T(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top) \big\} \cdot \boldsymbol{U}\boldsymbol{y}/n,$$

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.

E.4 Proof of Lemma 10

Applying Equation (23) and the risk decomposition formula, we obtain

$$\begin{aligned} R(\widehat{\boldsymbol{\beta}}_T^{\text{gf}}) =& \boldsymbol{\beta}_0^\top \exp(-2T\widehat{\boldsymbol{\Sigma}})\boldsymbol{\beta}_0 - \frac{2}{n}\boldsymbol{\beta}_0^\top \exp(-T\widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}^\dagger (\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top \boldsymbol{\varepsilon} \\ &+ \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))(\widehat{\boldsymbol{\Sigma}}^\dagger)^2(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top \boldsymbol{\varepsilon} + \sigma^2. \end{aligned}$$

Note that

$$\frac{2}{\sqrt{n}} \big\| \boldsymbol{\beta}_0^\top \exp(-T\widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}^\dagger (\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top \big\|_2 \le 2\|\boldsymbol{\beta}_0\|_2 \cdot \sup_{i\in[n]} \frac{\exp(-T\lambda_i)(1 - \exp(-T\lambda_i))}{\lambda_i^{1/2}},$$

where it is understood that $\lambda^{-1/2} e^{-T\lambda}(1 - e^{-T\lambda}) \mid_{\lambda=0} = 0$. Recall that $\max_i \lambda_i \xrightarrow{\text{a.s.}} (1 + \sqrt{\zeta_*})^2$ and $\|\boldsymbol{\beta}_0\|_2^2 \to r^2$. Hence, there exists a constant M_0 such that almost surely $\limsup_{n,p\to\infty} \|\boldsymbol{\beta}_0^\top \exp(-T\widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}^\dagger (\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top\|_2^2/n \le M_0$. Therefore, we can apply Lemma 19 and deduce that

$$\frac{2}{n}\boldsymbol{\beta}_0^\top \exp(-T\widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}^\dagger (\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top \boldsymbol{\varepsilon} \xrightarrow{\text{a.s.}} 0. \tag{37}$$

By Lemma 20, we have

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

Standard random matrix theory result implies that almost surely the empirical spectral distribution (ESD) of $\widehat{\boldsymbol{\Sigma}}$ converges in distribution to F_{ζ_*}, which is the Marchenko-Pastur law defined in Equation (35). Furthermore, $\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}} \xrightarrow{\text{a.s.}} (1+\sqrt{\zeta_*})^2$. Therefore, we conclude that

$$n^{-2}\sigma^2 \operatorname{tr}(\boldsymbol{X}(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))(\widehat{\boldsymbol{\Sigma}}^\dagger)^2(\boldsymbol{I}_p - \exp(-T\widehat{\Sigma}))\boldsymbol{X}^\top) \xrightarrow{\text{a.s.}} \zeta_*\sigma^2 \int z^{-1}(1-\exp(-Tz))^2 \,\mathrm{d}F_{\zeta_*}(z). \tag{38}$$

Finally, we study the limit of $\boldsymbol{\beta}_0^\top \exp(-2T\widehat{\boldsymbol{\Sigma}})\boldsymbol{\beta}_0$. Let $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$ be a uniformly distributed orthogonal matrix that is independent of anything else. Since by assumption $\|\boldsymbol{\beta}_0\|_2 \to r$, we can then couple $\boldsymbol{\Omega}\boldsymbol{\beta}_0$ with $\boldsymbol{g} \sim \mathsf{N}(\boldsymbol{0}, \mathbf{I}_p)$, such that (1) $\boldsymbol{g}$ is independent of $\widehat{\boldsymbol{\Sigma}}$, and (2) $\|\boldsymbol{\Omega}\boldsymbol{\beta}_0 - r\boldsymbol{g}/\sqrt{p}\|_2 \xrightarrow{\text{a.s.}} 0$. Note that all eigenvalues of $\exp(-2T\widehat{\boldsymbol{\Sigma}})$ are between 0 and 1, hence

$$\left|\boldsymbol{\beta}_0^\top \exp(-2T\widehat{\boldsymbol{\Sigma}})\boldsymbol{\beta}_0 - \frac{r^2}{p}\boldsymbol{g}^\top \exp(-2T\widehat{\boldsymbol{\Sigma}})\boldsymbol{g}\right| \xrightarrow{\text{a.s.}} 0.$$

Leveraging lemma 20, we obtain that $r^2\boldsymbol{g}^\top \exp(-2T\widehat{\boldsymbol{\Sigma}})\boldsymbol{g}/p \xrightarrow{\text{a.s.}} r^2 \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z)$. Combining this with Equations (37) and (38), we arrive at the following lemma.

E.5 Proof of Lemma 11

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

$$(1-\operatorname{tr}(\boldsymbol{H}_T^{\mathrm{gf}})/n)^{-2} \xrightarrow{\text{a.s.}} \left(1 - \zeta_* \int (1-\exp(-Tz))\,\mathrm{d}F_{\zeta_*}(z)\right)^{-2}. \tag{39}$$

Next, we consider the numerator. Straightforward computation implies that

$$\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_T^{\mathrm{gf}}\|_2^2 = \boldsymbol{\beta}_0^\top \exp(-T\widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}\exp(-T\widehat{\boldsymbol{\Sigma}})\boldsymbol{\beta}_0 + \frac{1}{n}\Big\|\big(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top\big)\boldsymbol{\varepsilon}\Big\|_2^2 \\ + \frac{2}{n}\big\langle \boldsymbol{\beta}_0, \exp(-T\widehat{\boldsymbol{\Sigma}})\boldsymbol{X}^\top\big(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top\big)\boldsymbol{\varepsilon}\big\rangle.$$

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

$$\limsup_{n,p\to\infty} \big\|\exp(-T\widehat{\boldsymbol{\Sigma}})\boldsymbol{X}^\top\big(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top\big)\big\|_{\mathrm{op}} \le G(\zeta_*) < \infty,$$

where $G(\zeta_*)$ is a function of ζ_*. Therefore, by Lemma 19, we obtain that

$$\frac{2}{n}\big\langle \boldsymbol{\beta}_0, \exp(-T\widehat{\boldsymbol{\Sigma}})\boldsymbol{X}^\top\big(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top\big)\boldsymbol{\varepsilon}\big\rangle \xrightarrow{\text{a.s.}} 0. \tag{40}$$

Using the same argument as that we used to compute the limiting expression of $\boldsymbol{\beta}_0^\top \exp(-T\widehat{\boldsymbol{\Sigma}})\boldsymbol{\beta}_0$, we conclude that

$$\boldsymbol{\beta}_0^\top \exp(-T\widehat{\boldsymbol{\Sigma}})\widehat{\boldsymbol{\Sigma}}\exp(-T\widehat{\boldsymbol{\Sigma}})\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} r^2 \int z\exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z). \tag{41}$$

In addition, by Lemma 20, we have

$$\frac{1}{n}\Big\|\big(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top\big)\boldsymbol{\varepsilon}\Big\|_2^2 \xrightarrow{\text{a.s.}} \sigma^2(1-\zeta_*)_+ + \sigma^2\zeta_* \int \exp(-2Tz)\,\mathrm{d}F_{\zeta_*}(z). \tag{42}$$

Equations (39) to (42) together imply the following lemma:

Expanding the matrix of the quadratic form, we get

$$
\begin{aligned}
&(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top)(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top) \\
&= (\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top) - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}})\boldsymbol{X}^\top(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top) \\
&= (\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\boldsymbol{X}^\top) - \frac{1}{n}\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))(\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}})))\boldsymbol{X}^\top
\end{aligned}
$$

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

$$
\begin{aligned}
&1 - \zeta_* \operatorname{tr}[(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))]/p - \zeta_* \operatorname{tr}[(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))\exp(-T\widehat{\boldsymbol{\Sigma}})]/p \\
&= 1 - \zeta_* + \zeta_* \operatorname{tr}[\exp(-2T\widehat{\boldsymbol{\Sigma}})]/p.
\end{aligned}
$$

In the simplification above, we used the fact that

$$\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Sigma}}^\dagger(\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}})) = (\boldsymbol{I}_p - \exp(-T\widehat{\boldsymbol{\Sigma}}))$$

This fact follows because $\widehat{\boldsymbol{\Sigma}}^\dagger\widehat{\boldsymbol{\Sigma}}$ is the projection onto the row space of $\boldsymbol{X}$. But the image of $\boldsymbol{I}_p - \exp(-t\widehat{\boldsymbol{\Sigma}})$ is already in the row space.

The limit for (42) 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 variance component of the training error of the ridgeless interpolator, which should be zero.

To check the last point, it is worth noting that

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

Finally, we shall combine Lemmas 10 and 11 to complete the proof.

E.6 Proof of Lemma 12

Finally, we shall combine Lemmas 10 and 11 to complete the proof.

Lemma 21 (Risk and GCV asymptotics mismatch for gradient flow)**.** *For ..., we have*

$$\left(1 - \gamma \int (1 - \exp(-ts))\, \mathrm{d}F_\gamma(s)\right)^2 \int \exp(-2ts)\, \mathrm{d}F_\gamma(s) \neq \int s\exp(-2ts)\, \mathrm{d}F_\gamma(s)$$

$$\left(1 - \gamma \int (1 - \exp(-ts))\, \mathrm{d}F_\gamma(s)\right)^2 \left\{1 + \gamma \int \frac{(1 - \exp(-ts))^2}{s}\, \mathrm{d}F_\gamma(s)\right\} \neq (1 - \gamma) + \gamma \int \exp(-2Ts)\, \mathrm{d}F_\gamma(s).$$

Proof. We will argue that both equations can not be simultaneously equal. For the data models for which $r^2 > 0$ and $\sigma^2 > 0$, this suffices to argue that GCV can not be consistent for all t.

The key for the contradiction is to notice that both the risk and the GCV numerator splits into a bias or bias-like and variance or variance-like components, respectively.

The functions on both sides of the equation are composed of integrals of functions that are products of analytic functions.

The function $1-\exp(-ts)$, which appears in the integrand on the left-hand side, is also analytic as it is a difference of two analytic functions.

Therefore, the functions inside the integrals on both sides are analytic.

The integral of an analytic function over a finite interval is also analytic.

Therefore, the functions on both sides of the equation are analytic in t.

If the functions on the left-hand side and the right-hand side are both analytic (i.e., they can be represented by a convergent power series in a neighborhood of every point in their domain), then they can only be equal on an interval if they are identically equal. This is a consequence of the identity theorem for analytic functions (see, e.g., Chapter 1 of Krantz and Parks (2002)). The identity theorem states that if two analytic functions agree on a set of points that has a limit point inside the domain of both functions, then the two functions are identically equal on the entire domain.

By choosing a specific value for t, we can evaluate the functions on the left-hand side and the right-hand side of the equation at this value and show that the functions are indeed not equal.

This shows that the functions on the left-hand side and the right-hand side of the original equation are not equal for $t = ...$ Therefore, they are not identically equal, and they must differ on a set of points that has positive Lebesgue measure.

□

E.6.1 Signal limits mismatch

Function calculations. Given the equation

$$\left(1-\gamma\int(1-\exp(-ts))\,\mathrm{d}F_\gamma(s)\right)^2\int\exp(-2ts)\,\mathrm{d}F_\gamma(s)\neq\int s\exp(-2ts)\,\mathrm{d}F_\gamma(s)$$

we want to show that the functions on the left-hand side and the right-hand side are not identically equal.

For $t=0$, the equation simplifies to:

$$\left(1-\gamma\int(1-1)\,\mathrm{d}F_\gamma(s)\right)^2\int 1\,\mathrm{d}F_\gamma(s)\neq\int s\,\mathrm{d}F_\gamma(s)$$

The left-hand side simplifies to:

$$(1-\gamma\cdot 0)^2\cdot 1=1$$

The right-hand side is the expected value of the distribution $F_\gamma(s)$:

$$\int s\,\mathrm{d}F_\gamma(s)=1$$

Unfortunately, the function evaluation does not supply us an immediate inequality on the level of function evaluations. However, we can take the derivative and analytically show that the derivatives do not match at $t=0$.

Derivative calculations. Let

$$w(t)=\left(1-\gamma\int(1-\exp(-ts))\,\mathrm{d}F_\gamma(s)\right)^2\quad\text{and}\quad v(t)=\int\exp(-2ts)\,\mathrm{d}F_\gamma(s)$$

Note that

$$\frac{dw(t)}{dt}=2\left(1-\gamma\int(1-\exp(-ts))\,\mathrm{d}F_\gamma(s)\right)\left(-\gamma\int s\exp(-ts)\,\mathrm{d}F_\gamma(s)\right)$$

$$= -2\gamma \left(1 - \gamma \int (1 - \exp(-ts))\,\mathrm{d}F_\gamma(s)\right) \int s \exp(-ts)\,\mathrm{d}F_\gamma(s)$$

$$\frac{dv(t)}{dt} = -2 \int s \exp(-2ts)\,\mathrm{d}F_\gamma(s)$$

Let

$$u(t) = \int s \exp(-2ts)\,\mathrm{d}F_\gamma(s)$$

Note that

$$\frac{du(t)}{dt} = \frac{d}{dt} \int s \exp(-2ts)\,\mathrm{d}F_\gamma(s) = -2 \int s^2 \exp(-2ts)\,\mathrm{d}F_\gamma(s)$$

$$\begin{aligned}
\frac{d\text{LHS}}{dt} &= \frac{dw(t)}{dt} \cdot v(t) + w(t) \cdot \frac{dv(t)}{dt} \\
&= -2\left(1 - \gamma \int (1 - \exp(-ts))\,\mathrm{d}F_\gamma(s)\right) \gamma \int s \exp(-ts)\,\mathrm{d}F_\gamma(s) \int s \exp(-ts)\,\mathrm{d}F_\gamma(s) \\
&\quad - \left(1 - \gamma \int (1 - \exp(-ts))\,\mathrm{d}F_\gamma(s)\right)^2 \int s \exp(-ts)\,\mathrm{d}F_\gamma(s) \\
&= -2\gamma \left(1 - \gamma \int (1 - \exp(-ts))\,\mathrm{d}F_\gamma(s)\right) \int s \exp(-ts)\,\mathrm{d}F_\gamma(s) \cdot \int \exp(-2ts)\,\mathrm{d}F_\gamma(s) \\
&\quad - 2\left(1 - \gamma \int (1 - \exp(-ts))\,\mathrm{d}F_\gamma(s)\right)^2 \cdot \int s \exp(-ts)\,\mathrm{d}F_\gamma(s)
\end{aligned}$$

Let

$$M_0 = \int \mathrm{d}F_\gamma(s) \quad \text{and} \quad M_1 = \int s\,\mathrm{d}F_\gamma(s) \quad \text{and} \quad M_2 = \int s^2\,\mathrm{d}F_\gamma(s)$$

We will evaluate the derivative at $t = 0$.

$$\left.\frac{d\text{LHS}}{dt}\right|_{t=0} = -2\gamma M_1 - 2M_1.$$

$$\left.\frac{d\text{RHS}}{dt}\right|_{t=0} = -2M_2$$

From Lemma 22, we have that $M_0 = 1$, $M_1 = 1$, and $M_2 = 1 + \gamma$

Numerical calculations. We will numerically verify that the functions are indeed different.

Figure 6

A side remark: The signal multiplier for both the under- and overparameterized regimes is 1 at $t = 0$. This is because the estimator is simply the null estimator at $t = 0$, which has bias of r^2.

E.6.2 Variance limits mismatch

The terms in this setting are:

$$w(t) = \left(1 - \gamma \int (1 - \exp(-ts))\,\mathrm{d}F_\gamma(s)\right)^2 \quad \text{and} \quad v(t) = 1 + \gamma \int \frac{(1 - \exp(-ts))^2}{s}\,\mathrm{d}F_\gamma(s)$$

$$u(t) = (1 - \gamma) + \gamma \int \exp(-2Ts)\,\mathrm{d}F_\gamma(s)$$

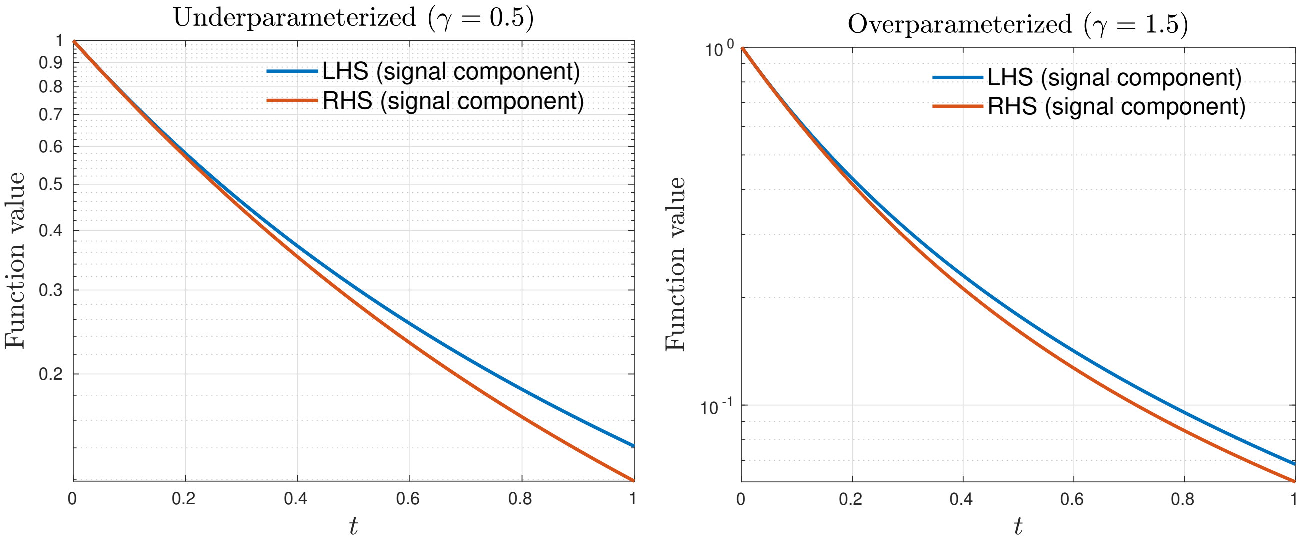

Figure 6: Comparison of the LHS and RHS

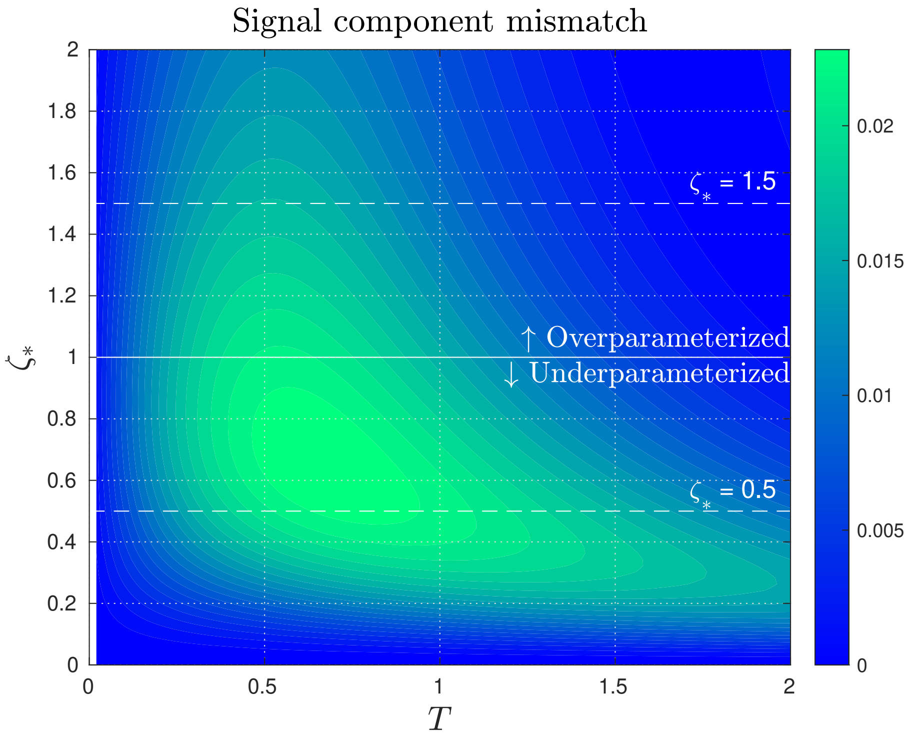

Figure 7: Contour plot of the absolute value of the difference between LHS and RHS for normalized the signal terms

E.7 Useful lemmas

Lemma 22 (Moments of the Marchenko-Pasture law)**.** *Let F_γ be the Marchenko-Pasture law as defined in ... For $k \geq 1$, we have*

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

Proof. This is a known formula that is derived using Vendermonde's identity. □

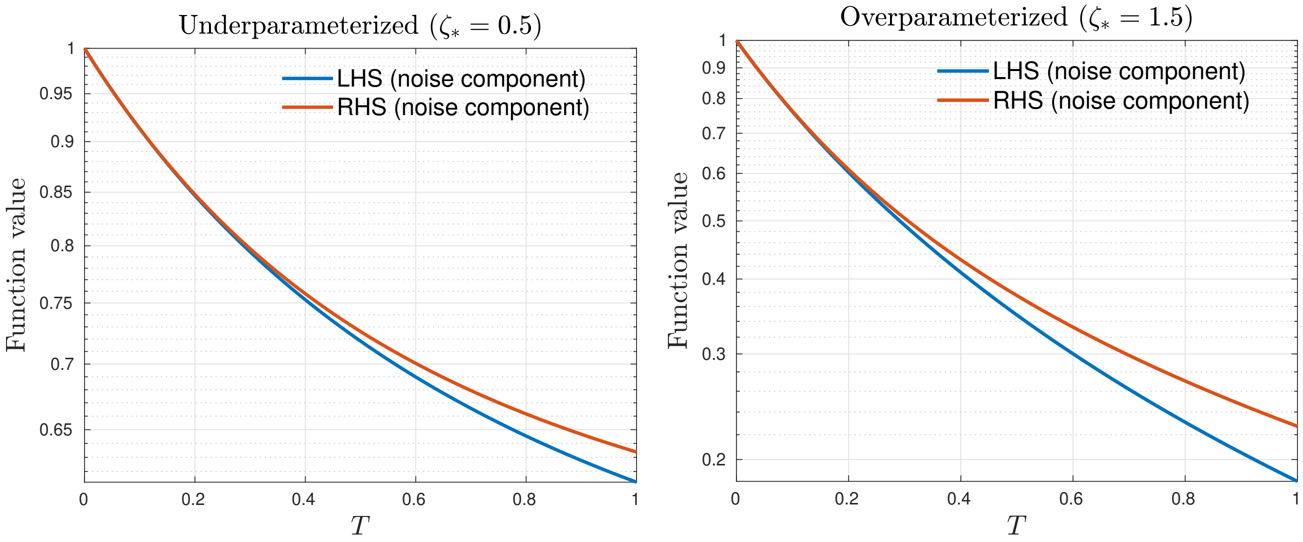

Figure 8: Comparison of the LHS and RHS

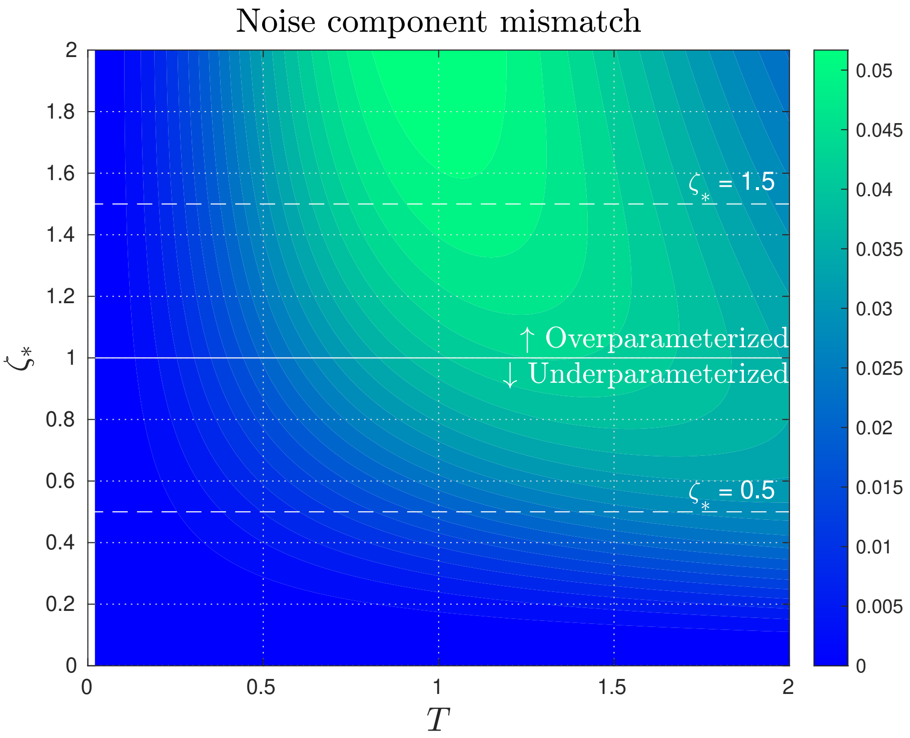

Figure 9: Contour plot of the absolute value of the difference between LHS and RHS for normalized the noise components

F Additional proofs in Section 3

F.1 Proof of Proposition 6

We prove the lemma via induction over t. For $t = 0$, by definition $\widetilde{\boldsymbol{\beta}}_{0,-i} = \boldsymbol{\beta}_{0,-i} = \boldsymbol{\beta}_0$ for all $i \in [n]$. Suppose we have $\widetilde{\boldsymbol{\beta}}_{t,-i} = \boldsymbol{\beta}_{t,-i}$ for $t = k$ and all $i \in [n]$, we then prove that it also holds for $t = k + 1$ via induction. Using its definition, we see that

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{k+1,-i} =& \widetilde{\boldsymbol{\beta}}_{k,-i} - \frac{2\delta_{k+1}}{n}\boldsymbol{X}^\top \boldsymbol{X}\,\widetilde{\boldsymbol{\beta}}_{k,-i} + \frac{2\delta_{k+1}}{n}\boldsymbol{X}^\top \widetilde{\boldsymbol{y}}_i^{(k)} \\ =& \widetilde{\boldsymbol{\beta}}_{k,-i} - \frac{2\delta_{k+1}}{n}\boldsymbol{X}_{-i}^\top \boldsymbol{X}_{-i}\,\widetilde{\boldsymbol{\beta}}_{k,-i} + \frac{2\delta_{k+1}}{n}\boldsymbol{X}_{-i}^\top \boldsymbol{y}_{-i} - \frac{2\delta_{k+1}}{n}\boldsymbol{x}_i\big(\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_{k,-i} - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{k,-i}\big)\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\beta}_{k,-i} - \frac{2\delta_{k+1}}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{X}_{-i} \boldsymbol{\beta}_{k,-i} + \frac{2\delta_{k+1}}{n} \boldsymbol{X}_{-i}^\top \boldsymbol{y}_{-i} \\
&= \boldsymbol{\beta}_{k+1,-i},
\end{aligned}
$$

thus completing the proof of the lemma by induction.

F.2 Proof of Lemma 7

We prove this lemma by induction over t. For the base case $t = 0$, the requirement of the lemma can be satisfied by setting

$$
h_{ij}^{(0)} = 0, \qquad b_i^{(0)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0, \qquad i, j \in [n].
$$

Suppose we can find $(h_{ij}^{(t)})_{i,j \le n}$ and $(b_i^{(t)})_{i \le n}$ for $t = k$, we next show that the counterpart quantities also exist for $t = k+1$. We define $\boldsymbol{H}^{(k)} \in \mathbb{R}^{n \times n}$, $\boldsymbol{b}^{(k)} \in \mathbb{R}^n$, such that $\boldsymbol{H}_{ij}^{(k)} = h_{ij}^{(k)}$ and $\boldsymbol{b}_i^{(k)} = b_i^{(k)}$. Using induction hypothesis, we obtain that

$$
\begin{aligned}
\boldsymbol{x}_i^\top \mathsf{GD}(\boldsymbol{\beta}_0, \boldsymbol{\delta}_{1:(k+1)}; \boldsymbol{X}, \boldsymbol{y}) &= \boldsymbol{x}_i^\top \left(\mathsf{GD}(\boldsymbol{\beta}_0, \boldsymbol{\delta}_{1:k}; \boldsymbol{X}, \boldsymbol{y}) - \frac{2\delta_{k+1}}{n} \boldsymbol{X}^\top \boldsymbol{X} \mathsf{GD}(\boldsymbol{\beta}_0, \boldsymbol{\delta}_{1:k}; \boldsymbol{X}, \boldsymbol{y}) + \frac{2\delta_{k+1}}{n} \boldsymbol{X}^\top \boldsymbol{y} \right) \\
&= \sum_{j=1}^n h_{ij}^{(k)} y_j + b_i^{(k)} - \frac{2\delta_{k+1}}{n} \boldsymbol{x}_i^\top \boldsymbol{X}^\top (\boldsymbol{H}^{(k)} \boldsymbol{y} + \boldsymbol{b}^{(k)}) + \frac{2\delta_{k+1}}{n} \boldsymbol{x}_i^\top \boldsymbol{X}^\top \boldsymbol{y},
\end{aligned}
$$

which completes the proof for $t = k+1$. We thus proves the lemma by induction.

G Supporting lemmas for the proof of ??

We present in this section several supporting lemmas that are useful for analysis presented in Appendix H and Appendix I. Without any loss, in this section we always assume $n \ge 3$, thus $\log n \ge 1$.

G.1 Concentration based on T_2-inequality

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

Proposition 23. *Let μ be a probability measure on a Polish space $(\mathbb{X}, d)$, and let $\{X_i\}$ 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:*

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

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

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

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

$$
\mathbb{P}\left(f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)] \ge t\right) \le C_{\mathsf{T}_2} e^{-t^2/2\sigma^2} \tag{43}
$$

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

G.2 Dimension free concentration

We define $\boldsymbol{w}_i = (\boldsymbol{x}_i, y_i)$. The following lemma is a straightforward consequence of the assumptions and the T_2-inequality.

Lemma 24. *We let $\sigma_{\mathsf{T}_2}^2 = \sigma_z^2 \vee \sigma_\varepsilon^2$, and $L = (L_\rho^2 \sigma_\Sigma + L_\rho^2 + \sigma_\Sigma)^{1/2}$. Then for any $n \ge 1$, $t \ge 0$, and 1-Lipschitz function f, it holds that*

$$
\mathbb{P}\left(f(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n) - \mathbb{E}[f(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] \ge Lt\right) \le C_{\mathsf{T}_2} e^{-t^2/2\sigma_{\mathsf{T}_2}^2},
$$

where we recall that $C_{\mathsf{T}_2} > 0$ is an absolute constant introduced in proposition 23.

Proof of Lemma 24. Since f is 1-Lipschitz, for any $\boldsymbol{w}_i, \widetilde{\boldsymbol{w}}_i \in \mathbb{R}^p$

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

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

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

for all ν. We then see that the desired concentration inequality is merely a straightforward consequence of Proposition 23. □

G.3 Upper bounding operator norms and response energy

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

Lemma 25. *We assume the assumptions of* **??**. *Then there exists a numerical constant $C_0 > 0$, such that with probability at least $1 - (n+p)^{-4}$*

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

Proof of Lemma 25. Note that the operator norm of $\widehat{\boldsymbol{\Sigma}}$ is equal to the operator norm of $\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^\top/n + \mathbf{1}_{n\times n}/n \in \mathbb{R}^{n\times n}$.

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 23, 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 are able to 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 $1 - 2\exp(-t^2)$

$$\|\boldsymbol{Z}\|_{\mathrm{op}} \leq C'(8 + 8C_{\mathsf{T}_2})(\sqrt{n} + \sqrt{p} + t), \tag{44}$$

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

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

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

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

Lemma 26. *We assume the assumptions of* **??**. *Then with probability at least* $1 - n^{-1} m_4 \log^{-2} n$, *we have* $\|\boldsymbol{y}\|_2^2/n \le m_2 + \log n$.

G.4 Other useful norm bounds

Our next lemma upper bounds the Euclidian norm of $\boldsymbol{\theta} := \mathbb{E}[y_0 \boldsymbol{x}_0] \in \mathbb{R}^{p+1}$, where we recall that $(\boldsymbol{x}_0, y_0) \overset{d}{=} (\boldsymbol{x}_1, y_1)$.

Lemma 27. *Under the assumptions of* **??**, *we have* $\|\boldsymbol{\theta}\|_2 \le (\sigma_\Sigma^{1/2} + 1) m_2^{1/2}$.

Proof of Lemma 27. We notice that

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\Sigma}^{1/2} \mathbb{E}[y_0 \boldsymbol{z}_0] \\ \mathbb{E}[y_0] \end{pmatrix}.$$

We let $\boldsymbol{x}_0^\top = (\boldsymbol{z}_0^\top \boldsymbol{\Sigma}^{1/2}, 1)$. By assumption, $\boldsymbol{z}_0$ 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 $\|\mathbb{E}[y_0 \boldsymbol{z}_0]\|_2 \le m_2^{1/2}$, where we recall that $m_2 = \mathbb{E}[y_0^2]$. This further tells us $\|\boldsymbol{\theta}\|_2 \le \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^{1/2} \times \|\mathbb{E}[y_0 \boldsymbol{z}_0]\|_2 + m_2^{1/2} \le (\sigma_\Sigma^{1/2} + 1) m_2^{1/2}$, thus completing the proof of the lemma. □

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

Lemma 28. *We define* $\widetilde{C}_0 = C' \sigma_\Sigma (8 + 8 C_{\mathsf{T}_2})$, *where we recall that* C' *is a positive numerical constant that appears in Equation (44). Under the assumptions of* **??**, *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*

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

Proof of Lemma 28. By Equation (44), for all $t \ge 0$, with probability at least $1 - 2\exp(-nt^2)$

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

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

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

thus completing the proof of the lemma. □

G.5 Upper bounding $\|\widehat{\boldsymbol{\beta}}_k\|_2$ and $\|\widehat{\boldsymbol{\beta}}_{k,-i}\|_2$

We then prove that on Ω, the Euclidean norm of the coefficient estimates $\{\widehat{\boldsymbol{\beta}}_k, \widehat{\boldsymbol{\beta}}_{k,-i} : k \in [K], i \in [n]\}$ 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 are independent of (n, p).

Lemma 29. *For the sake of simplicity, we let*

$$B_* = (B_0 + \Delta C_{\Sigma,\zeta}^{1/2}\sqrt{m+1}) \cdot e^{C_{\Sigma,\zeta}\Delta}. \tag{45}$$

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

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

Proof of Lemma 29. By definition,

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

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

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

By induction, we see that on Ω

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

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

The following corollary is a straightforward consequence of Lemma 29 and Cauchy-Schwartz inequality.

Corollary 30. *On the set Ω, it holds that*

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

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

For the compactness of future presentation, we define

$$\bar{B}_* = (2m + 2 + 2C_{\Sigma,\zeta}B_*^2)^{1/2} \tag{46}$$

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

G.6 Upper bounding $|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_{k,-i}|$

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

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

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

$$\|\boldsymbol{a}_0\|_2^2 \leq \bar{B}_*^2 n \log n, \qquad \|\boldsymbol{E}_0\|_F^2 = 0,$$

where the first upper bound follows from Corollary 30.

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

Lemma 31. *We define*

$$\mathcal{G}_1(C_{\Sigma,\zeta}, \Delta, m, B_0) = \bar{B}_* \sqrt{e^{3\Delta C_{\Sigma,\zeta} + 2\Delta^2 C_{\Sigma,\zeta}^2}(\Delta C_{\Sigma,\zeta} + 2\Delta^2 C_{\Sigma,\zeta}^2)}, \tag{47}$$

$$\mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) = \bar{B}_* + \Delta C_{\Sigma,\zeta} \sqrt{8\bar{B}_*^2 + 2\mathcal{G}_1(C_{\Sigma,\zeta}, \Delta, m, B_0)^2}. \tag{48}$$

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

$$\frac{1}{\sqrt{n}} \|\boldsymbol{E}_k\|_F \le \mathcal{G}_1(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \sqrt{\log n}, \tag{49}$$

$$\frac{1}{\sqrt{n}} \|\boldsymbol{a}_k\|_2 \le \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \sqrt{\log n}. \tag{50}$$

Proof of Lemma 31. We first prove Equation (49). We denote by $\boldsymbol{X}_{-i} \in \mathbb{R}^{(n-1)\times(p+1)}$ the matrix obtained by deleting the i-th row from $\boldsymbol{X}$. By definition,

$$\begin{aligned}\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{k+1} - \widehat{\boldsymbol{\beta}}_{k+1,-i}) =& \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i}) + \frac{\delta_k (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_k)}{n} \boldsymbol{X}\boldsymbol{x}_i - \frac{\delta_k}{n} \boldsymbol{X} \sum_{j \ne i} \boldsymbol{x}_j \boldsymbol{x}_j^\top (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i}) \\ =& \boldsymbol{X}(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i}) + \frac{\delta_k (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_k)}{n} \boldsymbol{X}\boldsymbol{x}_i - \frac{\delta_k}{n} \boldsymbol{X}\boldsymbol{X}_{-i}^\top \boldsymbol{X}_{-i} (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i}),\end{aligned}$$

which further implies

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

where we make use of the fact that $\|\boldsymbol{X}\boldsymbol{x}_i\|_2/n \le C_{\Sigma,\zeta}$ on Ω. Putting together the above upper bound and Corollary 30, then summing over $i \in [n]$, we obtain the following inequality:

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

Employing standard induction argument, we are able to conclude that

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

for all $k \in \{0\} \cup [K]$. This completes the proof of Equation (49).

Next, we prove Equation (50). By definition,

$$\begin{aligned} y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i} =& y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i} - \frac{\delta_k}{n} \sum_{j \ne i} (y_j - \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_{k,-i}) \boldsymbol{x}_i^\top \boldsymbol{x}_j \\ =& y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i} - \frac{\delta_k}{n} \sum_{j \ne i} (y_j - \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_k) \boldsymbol{x}_i^\top \boldsymbol{x}_j - \frac{\delta_k}{n} \sum_{j \ne i} \boldsymbol{x}_i^\top \boldsymbol{x}_j \boldsymbol{x}_j^\top (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i}).\end{aligned}$$

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

$$(a_{k+1,i} - a_{k,i})^2 \le \frac{2\delta_k^2}{n^2} \Big(\sum_{j \ne i} (y_j - \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_k) \boldsymbol{x}_i^\top \boldsymbol{x}_j \Big)^2 + \frac{2\delta_k^2}{n^2} \Big(\sum_{j \ne i} \boldsymbol{x}_i^\top \boldsymbol{x}_j \boldsymbol{x}_j^\top (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i}) \Big)^2.$$

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

$$
\begin{aligned}
&\|\boldsymbol{a}_{k+1} - \boldsymbol{a}_k\|_2^2 \\
&\leq \frac{2\delta_k^2}{n^2}\left\|(\boldsymbol{X}\boldsymbol{X}^\top - n\mathbf{D})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k)\right\|_2^2 + \frac{2\delta_k^2}{n^2}\sum_{i=1}^n \Big(\sum_{j\neq i}(\boldsymbol{x}_i^\top \boldsymbol{x}_j)^2\Big) \cdot \Big(\sum_{j \neq i}\big(\boldsymbol{x}_j^\top(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i})\big)^2\Big) \\
&\overset{(i)}{\leq} 8n\delta_k^2 C_{\Sigma,\zeta}^2 \cdot \bar{B}_*^2 \log n + 2\delta_k^2 C_{\Sigma,\zeta}^2 \sum_{i=1}^n \|\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i})\|_2^2 \\
&\overset{(ii)}{\leq} 8n\delta_k^2 C_{\Sigma,\zeta}^2 \cdot \bar{B}_*^2 \log n + 2n\delta_k^2 C_{\Sigma,\zeta}^2 \mathcal{G}_1(C_{\Sigma,\zeta}, \Delta, m, B_0)^2 \cdot \log n \\
&:= n\delta_k^2 \cdot \mathcal{G}'(C_{\Sigma,\zeta}, \Delta, m, B_0)^2 \cdot \log n,
\end{aligned}
$$

where to derive (i), we employ the following established results: (1) On Ω we have $\|n\mathbf{D}\|_{\mathrm{op}} \leq nC_{\Sigma,\zeta}$ and $\|\boldsymbol{X}\boldsymbol{X}^\top\|_{\mathrm{op}} \leq nC_{\Sigma,\zeta}$. (2) By Corollary 30, on Ω we have $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k\|_2^2/n \leq \bar{B}_*^2 \cdot \log n$. To derive (ii), we simply apply Equation (49), which we have already proved. Therefore, by triangle inequality

$$
\frac{1}{\sqrt{n}}\|\boldsymbol{a}_{k+1}\|_2 \leq \frac{1}{\sqrt{n}}\|\boldsymbol{a}_k\|_2 + \frac{1}{\sqrt{n}}\|\boldsymbol{a}_{k+1} - \boldsymbol{a}_k\|_2 \leq \frac{1}{\sqrt{n}}\|\boldsymbol{a}_k\|_2 + \delta_k \mathcal{G}'(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \sqrt{\log n}.
$$

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

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

which concludes the proof of Equation (50). □

H Proof of ?? (for squared risk)

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 J.

H.1 Outline of the proof

A visual schematic for the proof of **??** is provided in Figure 10.

H.2 Proof of Lemma 14

We prove Lemma 13 in Appendix I. We claim that Lemma 13 can be used to show $\widetilde{f}_k$ is Lipchitz continuous. More precisely, for $\boldsymbol{W}, \boldsymbol{W}' \in \mathbb{R}^{n(p+2)}$, it holds that

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

Namely, $\widetilde{f}_k$ is $n^{-1/2}K\xi(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot \log n$-Lipchitz continuous for all $k \in \{0\} \cup [K]$. Applying Lemma 24, we conclude that

$$
\mathbb{P}\left(\left|\widetilde{f}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n) - \mathbb{E}[\widetilde{f}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]\right| \geq \frac{2\sigma_{\mathsf{T}_2} LK\xi(C_{\Sigma,\zeta}, \Delta, m, B_0) \cdot (\log n)^{3/2}}{\sqrt{n}}\right) \leq 2C_{\mathsf{T}_2} n^{-2}.
$$

Note that on the set Ω we have $\widetilde{f}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n) = f_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n) = \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$ for all $k \in \{0\} \cup [K]$. This completes the proof of the lemma.

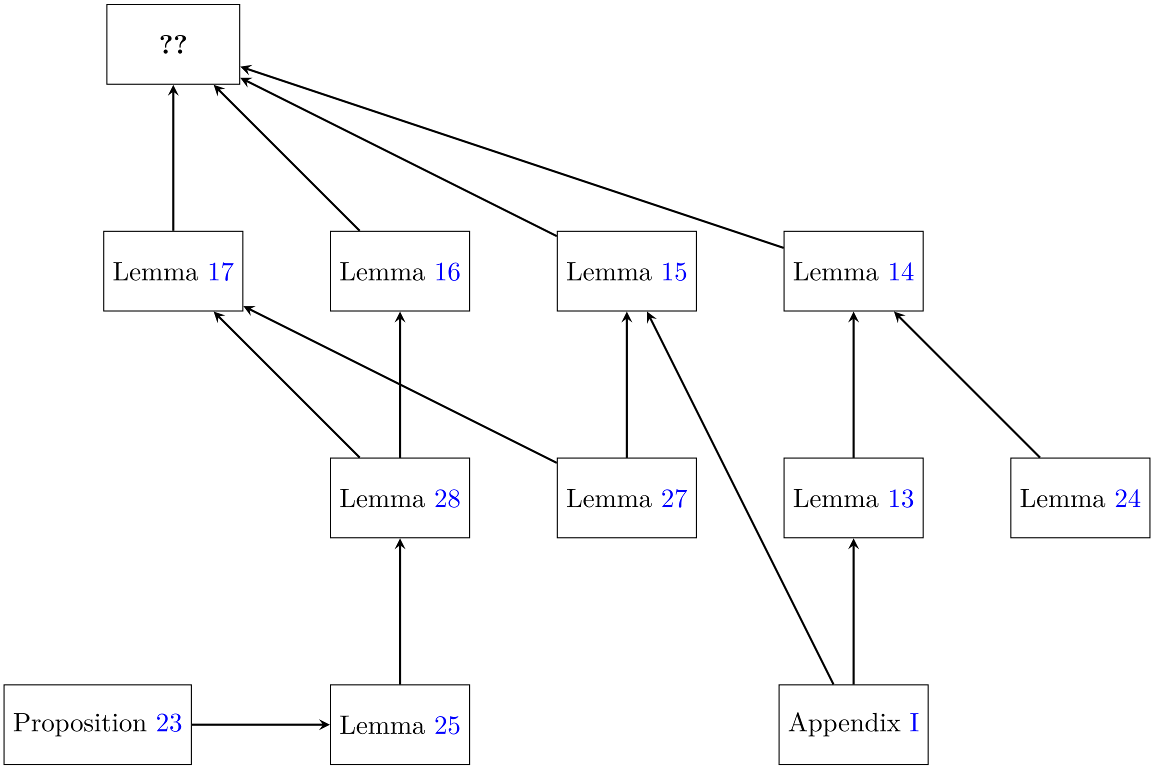

Figure 10: Schematic for the proof of **??**

H.3 Proof of Lemma 15

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

$$\nabla_{\boldsymbol{x}_s} R(\widehat{\boldsymbol{\beta}}_k) = 2\widehat{\boldsymbol{\beta}}_k^\top \widetilde{\boldsymbol{\Sigma}} \nabla_{\boldsymbol{x}_s} \widehat{\boldsymbol{\beta}}_k - 2\widehat{\boldsymbol{\theta}}^\top \nabla_{\boldsymbol{x}_s} \widehat{\boldsymbol{\beta}}_k,$$
$$\frac{\partial}{\partial_{y_s}} R(\widehat{\boldsymbol{\beta}}_k) = 2\widehat{\boldsymbol{\beta}}_k^\top \widetilde{\boldsymbol{\Sigma}} \frac{\partial}{\partial_{y_s}} \widehat{\boldsymbol{\beta}}_k - 2\boldsymbol{\theta}^\top \frac{\partial}{\partial_{y_s}} \widehat{\boldsymbol{\beta}}_k,$$

where

$$\boldsymbol{\theta} = \mathbb{E}[y_0 \boldsymbol{x}_0] \in \mathbb{R}^{p+1}, \qquad \widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0}_p \\ \boldsymbol{0}_p^\top & 1 \end{bmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}.$$

By definition,

$$\nabla_{\boldsymbol{x}_s} \widehat{\boldsymbol{\beta}}_{k+1} = \nabla_{\boldsymbol{x}_s} \widehat{\boldsymbol{\beta}}_k - \delta_k \widehat{\boldsymbol{\Sigma}} \cdot \nabla_{\boldsymbol{x}_s} \widehat{\boldsymbol{\beta}}_k + \frac{\delta_k}{n}(y_s - \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_k) \boldsymbol{I}_{p+1} - \frac{\delta_k}{n} \boldsymbol{x}_s \widehat{\boldsymbol{\beta}}_k^\top,$$
$$\frac{\partial}{\partial_{y_s}} \widehat{\boldsymbol{\beta}}_{k+1} = \frac{\partial}{\partial_{y_s}} \widehat{\boldsymbol{\beta}}_k - \delta_k \widehat{\boldsymbol{\Sigma}} \frac{\partial}{\partial_{y_s}} \widehat{\boldsymbol{\beta}}_k + \frac{\delta_k}{n} \boldsymbol{x}_s.$$

Standard induction argument leads to the following decomposition:

$$\nabla_{\boldsymbol{x}_s} \widehat{\boldsymbol{\beta}}_{k+1} = \sum_{k'=1}^{k} \boldsymbol{H}_{k',k} \cdot \Big(\frac{\delta_{k'}}{n}(y_s - \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_{k'}) \boldsymbol{I}_{p+1} - \frac{\delta_{k'}}{n} \boldsymbol{x}_s \widehat{\boldsymbol{\beta}}_{k'}^\top \Big),$$
$$\frac{\partial}{\partial_{y_s}} \widehat{\boldsymbol{\beta}}_{k+1} = \sum_{k'=1}^{k} \boldsymbol{H}_{k',k} \cdot \frac{\delta_{k'}}{n} \boldsymbol{x}_s,$$

where $\boldsymbol{H}_{k',r} = \prod_{j=k'+1}^{r} \boldsymbol{M}_{k'+1+r-j}$ and $\boldsymbol{M}_j = \boldsymbol{I}_{p+1} - \delta_j \widehat{\boldsymbol{\Sigma}}$ are defined in Lemma 33. Combining all these arguments, we arrive at the following equations:

$$\boldsymbol{v}^\top \nabla_{\boldsymbol{x}_s} \widehat{\boldsymbol{\beta}}_{k+1} = \sum_{k'=1}^{k} \boldsymbol{v}^\top \boldsymbol{H}_{k',k} \cdot \frac{\delta_{k'}}{n}(y_s - \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_{k'}) - \sum_{k'=1}^{k} \frac{\delta_{k'}}{n} \boldsymbol{x}_s^\top \boldsymbol{H}_{k',k} \boldsymbol{v} \widehat{\boldsymbol{\beta}}_{k'}^\top,$$

$$\boldsymbol{v}^\top \frac{\partial}{\partial_{y_s}} \widehat{\boldsymbol{\beta}}_{k+1} = \sum_{k'=1}^{k} \frac{\delta_{k'}}{n} \boldsymbol{x}_s^\top \boldsymbol{H}_{k',k} \boldsymbol{v}.$$

The above equations hold for all $\boldsymbol{v} \in \{\boldsymbol{\theta}, \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_{k+1}\}$. Recall that $\boldsymbol{\theta} := \mathbb{E}[y_0 \boldsymbol{x}_0]$. This further implies that

$$\begin{aligned}
&\nabla_{\boldsymbol{X}} R(\widehat{\boldsymbol{\beta}}_{k+1}) \\
&= \sum_{k'=1}^{k} \frac{2\delta_{k'}}{n} \cdot \Big\{ (\boldsymbol{y} - \boldsymbol{X} \widehat{\boldsymbol{\beta}}_{k'}) \widehat{\boldsymbol{\beta}}_{k+1}^\top \widetilde{\boldsymbol{\Sigma}} \boldsymbol{H}_{k',k} - \boldsymbol{X} \boldsymbol{H}_{k,k'} \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_{k+1} \widehat{\boldsymbol{\beta}}_{k'}^\top - (\boldsymbol{y} - \boldsymbol{X} \widehat{\boldsymbol{\beta}}_{k'}) \boldsymbol{\theta}^\top \boldsymbol{H}_{k',k} + \boldsymbol{X} \boldsymbol{H}_{k,k'} \boldsymbol{\theta} \widehat{\boldsymbol{\beta}}_{k'}^\top \Big\} \\
&\nabla_{\boldsymbol{y}} \mathcal{R}(\widehat{\boldsymbol{\beta}}_{k+1}) = \sum_{k'=1}^{k} \frac{2\delta_{k'}}{n} \cdot \Big\{ \boldsymbol{X} \boldsymbol{H}_{k,k'} \widetilde{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\beta}}_{k+1} - \boldsymbol{X} \boldsymbol{H}_{k,k'} \boldsymbol{\theta} \Big\}.
\end{aligned}$$

Recall that B_* is defined in Equation (45). Invoking triangle inequality, we obtain that on Ω,

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

where the inequality follows by invoking Lemma 27 to upper bound $\|\boldsymbol{\theta}\|_2$. Also, by Lemma 33 we know that $\|\boldsymbol{H}_{k',r}\|_{\mathrm{op}} \leq e^{\Delta C_{\Sigma,\zeta}}$. Similarly, we obtain

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

The above inequalities give an upper bound for $\|\nabla_{\boldsymbol{W}} R(\widehat{\boldsymbol{\beta}}_{k+1})\|_2$ on Ω. The rest parts of the proof is similar to the proof of Lemma 14 given Lemma 13.

H.4 Proof of Lemma 16

We shall first upper bound the fourth moments $\mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_{k,-1})^4]$ and $\mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k)^4]$. By standard induction, it is not hard to see that for all $0 \leq k \leq K$ and $i \in [n]$,

$$\|\widehat{\boldsymbol{\beta}}_k\|_2 \leq \exp(\Delta \|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}) \cdot \Big(B_0 + \Delta n^{-1} \|\boldsymbol{X}\|_{\mathrm{op}} \cdot \|\boldsymbol{y}\|_2 \Big), \tag{51}$$

$$\|\widehat{\boldsymbol{\beta}}_{k,-i}\|_2 \leq \exp(\Delta \|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}) \cdot \Big(B_0 + \Delta n^{-1} \|\boldsymbol{X}\|_{\mathrm{op}} \cdot \|\boldsymbol{y}\|_2 \Big). \tag{52}$$

For technical reasons that will become clear soon, we need to upper bound the expectations of $\|\widehat{\boldsymbol{\beta}}_k\|_2$ and $\|\widehat{\boldsymbol{\beta}}_{k,-i}\|_2$. To this end, we find it useful to show $\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}^{1/2}$ is sub-Gaussian. Next, we will employ Lemma 28 to upper bound $\mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_{k,-1})^4]$ and $\mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_k)^4]$. Invoking Cauchy-Schwartz inequality and triangle inequality, we obtain that for $n \geq N(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)$,

$$\begin{aligned}
\mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_{k,-1})^4] \leq & \mathbb{E}[\|(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_{k,-1})\|_2^4] \\
\leq & 8\mathbb{E}[y_1^4] + 8\mathbb{E}[(\boldsymbol{x}_1^\top \widehat{\boldsymbol{\beta}}_{k,-1})^4] = 8m_4 + 8\mathbb{E}[((\boldsymbol{z}_1^\top, 1) \widetilde{\boldsymbol{\Sigma}}^{1/2} \widehat{\boldsymbol{\beta}}_{k,-1})^4] \\
\overset{(i)}{\leq} & 8m_4 + C_z \mathbb{E}[\|\widetilde{\boldsymbol{\Sigma}}^{1/2} \widehat{\boldsymbol{\beta}}_{k,-1}\|_2^4] \\
\overset{(ii)}{\leq} & \mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)^2
\end{aligned}$$

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) $\boldsymbol{z}_1$ is independent of $\widetilde{\mathbf{\Sigma}}^{1/2}\boldsymbol{\beta}_{k,-1}$. To derive *(ii)* we apply Equation (52) and Lemma 28. Similarly, we can show that for $n \geq N(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)$,

$$\mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_k)^4] \leq \mathbb{E}[\|(y_0, \boldsymbol{x}_0^\top \boldsymbol{\beta}_k)\|_2^4] \leq \mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)^2. \tag{53}$$

Finally, we are ready to establish Equation (29). By Cauchy-Schwartz inequality,

$$\begin{aligned}
\Big|\mathbb{E}[r_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] - \mathbb{E}[\widetilde{r}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]\Big| &\leq \mathbb{P}(\Omega^c)^{1/2}\mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k)^4]^{1/2}, \\
\Big|\mathbb{E}[f_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] - \mathbb{E}[\widetilde{f}_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]\Big| &\leq \mathbb{P}(\Omega^c)^{1/2}\mathbb{E}[(y_1 - \boldsymbol{x}_1^\top \widehat{\boldsymbol{\beta}}_{k,-1})^4]^{1/2},
\end{aligned}$$

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.

H.5 Proof of Lemma 17

By Equation (51), Equation (52), and Lemma 28, we know that there exists a constant C'' that depends only on $(\sigma_\Sigma, \zeta, \Delta, B_0, m_2)$, such that

$$\max\left\{\mathbb{E}[\|\boldsymbol{\beta}_k\|_2^2]^{1/2}, \mathbb{E}[\|\boldsymbol{\beta}_{k,-i}\|_2^2]^{1/2}\right\} \leq C''. \tag{54}$$

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

$$\widehat{\boldsymbol{\beta}}_{k+1} - \widehat{\boldsymbol{\beta}}_{k+1,-i} = \big(\boldsymbol{I}_p - \delta_k\widehat{\mathbf{\Sigma}}\big) \cdot \big(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i}\big) + \frac{\delta_k}{n}y_i\boldsymbol{x}_i - \frac{\delta_k}{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{k,-i}.$$

Invoking triangle inequality and Cauchy-Schwartz inequality, we conclude that

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

By induction,

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

By Hölder's inequality and Lemma 28, we see that for $n \geq 12\Delta\widetilde{C}_0^2 + 1$

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{k+1} - \widehat{\boldsymbol{\beta}}_{k+1,-i}\|_2^2\right] \\
&\leq \sum_{j=1}^{k}\frac{\delta_j}{n^2} \cdot \mathbb{E}[(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{k,-i})^4]^{1/2} \cdot \mathbb{E}[\|\boldsymbol{x}_i\|_2^8]^{1/4} \cdot \mathbb{E}[\exp(12\Delta\|\widehat{\mathbf{\Sigma}}\|_{\mathrm{op}} + 4\Delta)]^{1/4} \\
&\leq \frac{\Delta e^{\Delta}\sigma_\Sigma\mathcal{H}(\sigma_\Sigma, \zeta, B_0, m_8, \Delta)}{n} \cdot \mathcal{E}(\widetilde{C}_0, \zeta, 12\Delta)^{1/4}.
\end{aligned} \tag{55}$$

In addition, direct computation gives

$$\mathbb{E}[r_k(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] = m_2 + \mathbb{E}[\widehat{\boldsymbol{\beta}}_k^\top\mathbf{\Sigma}\widehat{\boldsymbol{\beta}}_k] + 2\langle\mathbb{E}[\widehat{\boldsymbol{\beta}}_k], \boldsymbol{\theta}\rangle,$$

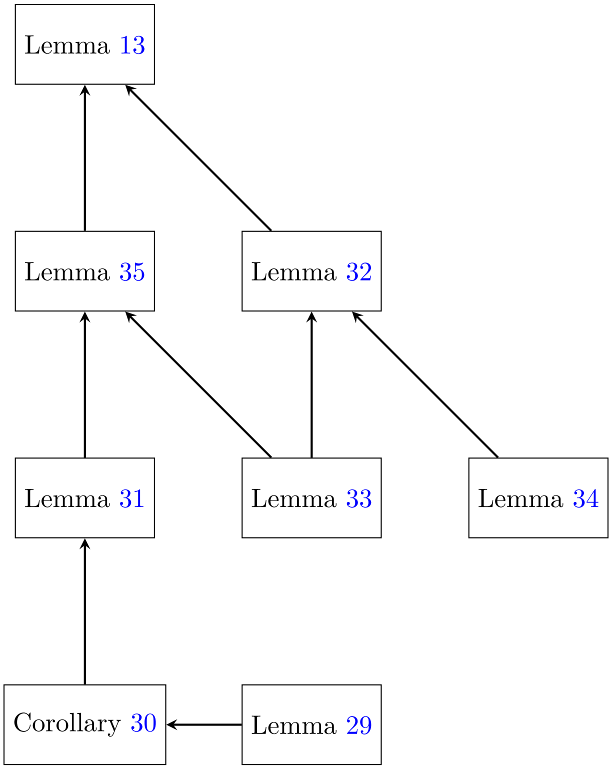

Figure 11: Schematic for the proof of Lemma 13

$$\mathbb{E}[f_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)] = m_2 + \mathbb{E}[\widehat{\boldsymbol{\beta}}_{k,-i}^\top \boldsymbol{\Sigma} \widehat{\boldsymbol{\beta}}_{k,-i}] + 2\langle \mathbb{E}[\widehat{\boldsymbol{\beta}}_{k,-i}], \boldsymbol{\theta}\rangle,$$

where we recall that $\boldsymbol{\theta} = \mathbb{E}[y_0 \boldsymbol{x}_0]$. By Lemma 27 we know that $\|\boldsymbol{\theta}\|_2 \le (\sigma_\Sigma^{1/2} + 1) m_2^{1/2}$. Therefore,

$$\begin{aligned}
&|\mathbb{E}[r_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)] - \mathbb{E}[f_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]| \\
&\le 2\|\boldsymbol{\theta}\|_2 \cdot \mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_{k+1} - \widehat{\boldsymbol{\beta}}_{k+1,-i}\|_2^2\right]^{1/2} + \sigma_\Sigma \mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-i}\|_2^2\right]^{1/2} \cdot \Big(\mathbb{E}\big[\|\widehat{\boldsymbol{\beta}}_k\|_2^2\big]^{1/2} + \mathbb{E}\big[\|\widehat{\boldsymbol{\beta}}_{k,-i}\|_2^2\big]^{1/2}\Big),
\end{aligned}$$

which by Equations (54) and (55) goes to zero as $n, p \to \infty$. Furthermore, the convergence is uniform for all $k \in \{0\} \cup [K]$. We have completed the proof of the lemma.

I Proof of Lemma 13

I.1 Outline of the proof

We divide the proof of the lemma into two parts: upper bounding $\|\nabla_{\boldsymbol{X}} \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)\|_F$ and $\|\nabla_{\boldsymbol{X}} \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)\|_F$. A visual schematic for the proof of Lemma 13 is provided in Figure 11.

I.2 Upper bounding $\|\nabla_{\boldsymbol{X}} \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)\|_F$

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

Lemma 32 (Bounding norm of gradient with respect to features)**.** *On the set Ω, for all $k \in \{0\} \cup [K]$,*

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

In the above equation, we recall that B_* is defined in Equation (45), $\bar{B}_*$ is defined in Equation (46), and $\mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0)$ is defined in Equation (48).

Proof of Lemma 32

We prove Lemma 32 in the remainder of this section. For $s \in [n]$ and $k \in [K]$, we can compute $\nabla_{\boldsymbol{x}_s}\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$, which takes the following form:

$$\nabla_{\boldsymbol{x}_s}\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) = -\frac{2}{n}(y_s - \boldsymbol{x}_s^\top\widehat{\boldsymbol{\beta}}_{k,-s})\widehat{\boldsymbol{\beta}}_{k,-s}^\top - \frac{2}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{k,-i})\boldsymbol{x}_i^\top\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k,-i}. \tag{56}$$

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

$$\begin{aligned}
&\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k+1,-i}\\
&= \nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k,-i} + \frac{\delta_k\mathbb{1}\{i\neq s\}}{n}(y_s - \boldsymbol{x}_s^\top\widehat{\boldsymbol{\beta}}_{k,-i})\boldsymbol{I}_{p+1} - \frac{\delta_k\mathbb{1}\{i\neq s\}}{n}\boldsymbol{x}_s\widehat{\boldsymbol{\beta}}_{k,-i}^\top - \frac{\delta_k}{n}\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j^\top\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k,-i}\\
&= \Big(\boldsymbol{I}_{p+1} - \delta_k\widehat{\boldsymbol{\Sigma}}\Big)\cdot\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k,-i} + \frac{\delta_k}{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k,-i} + \mathbb{1}\{i\neq s\}\cdot\left\{\frac{\delta_k}{n}(y_s - \boldsymbol{x}_s^\top\widehat{\boldsymbol{\beta}}_{k,-i})\boldsymbol{I}_{p+1} - \frac{\delta_k}{n}\boldsymbol{x}_s\widehat{\boldsymbol{\beta}}_{k,-i}^\top\right\}.
\end{aligned}$$

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

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:

$$\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k+1,-i} = \boldsymbol{M}_k\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k,-i} + \boldsymbol{M}_{k,i}\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k,-i} + \frac{\delta_k}{n}(y_s - \boldsymbol{x}_s^\top\widehat{\boldsymbol{\beta}}_{k,-i})\boldsymbol{I}_{p+1} - \frac{\delta_k}{n}\boldsymbol{x}_s\widehat{\boldsymbol{\beta}}_{k,-i}^\top, \tag{57}$$

where $\boldsymbol{M}_k = \boldsymbol{I}_{p+1} - \delta_k\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{M}_{k,i} = \delta_k\boldsymbol{x}_i\boldsymbol{x}_i^\top/n$. By induction, it is not hard to see that for all $0 \le k \le K-1$, the matrix $\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k+1,-i} - \boldsymbol{R}_0^{(k)}$ can be expressed as the sum of terms that take the form

$$\left(\prod_{j=1}^{k-k'}\boldsymbol{R}_{k+1-j}\right)\boldsymbol{R}_0^{(k')},$$

where $k' \in \{0\}\cup[k-1]$, $\boldsymbol{R}_0^{(k')} = \frac{\delta_{k'}}{n}(y_s - \boldsymbol{x}_s^\top\widehat{\boldsymbol{\beta}}_{k',-i})\boldsymbol{I}_{p+1} - \frac{\delta_{k'}}{n}\boldsymbol{x}_s\widehat{\boldsymbol{\beta}}_{k',-i}^\top$, and $\boldsymbol{R}_j$ is either $\boldsymbol{M}_j$ or $\boldsymbol{M}_{j,i}$.

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

Lemma 33. *For $i, s \in [n]$ with $i \neq s$ and all $k \in \{0\}\cup[K-1]$, it holds that*

$$\boldsymbol{x}_i^\top\nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k+1,-i} = \sum_{k'=0}^{k}\sum_{r=k'}^{k}c_{i,k,k',r}\boldsymbol{x}_i^\top\boldsymbol{H}_{k',r}\cdot\left(\frac{\delta_{k'}}{n}(y_s - \boldsymbol{x}_s^\top\widehat{\boldsymbol{\beta}}_{k',-i})\boldsymbol{I}_{p+1} - \frac{\delta_{k'}}{n}\boldsymbol{x}_s\widehat{\boldsymbol{\beta}}_{k',-i}^\top\right),$$

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

$$\|\boldsymbol{H}_{k',r}\|_{\mathrm{op}} \le e^{\Delta C_{\Sigma,\zeta}}, \qquad \|c_{i,k,k',r}\boldsymbol{H}_{k',r}\|_{\mathrm{op}} \le e^{2\Delta C_{\Sigma,\zeta}}. \tag{58}$$

Proof of Lemma 33. To derive the first inequality in Equation (58), we simply notice that

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

We next prove the second inequality in Equation (58). As discussed before, $\boldsymbol{x}_i^\top\nabla_{\boldsymbol{x}_s}\boldsymbol{\beta}_{k+1,-i} - \boldsymbol{x}_i^\top\boldsymbol{R}_0^{(k)}$ can be expressed as the sum of terms that take the form

$$\boldsymbol{x}_i^\top\left(\prod_{j=1}^{k-k'}\boldsymbol{R}_{k+1-j}\right)\boldsymbol{R}_0^{(k')},$$

with k' ranging from 0 to $k-1$. The subtracting $\boldsymbol{x}_i^\top \boldsymbol{R}_0^{(k)}$ part implies that we should set $c_{i,k,k,k} = 1$ and $\boldsymbol{H}_{k,k} = \boldsymbol{I}_{p+1}$.

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

$$\begin{aligned}\boldsymbol{x}_i^\top \left(\prod_{j=1}^{k-k'} \boldsymbol{R}_{k+1-j}\right) \boldsymbol{R}_0^{(k')} =& \boldsymbol{x}_i^\top \left(\prod_{j=1}^{k-j_*} \boldsymbol{R}_{k+1-j}\right) \cdot \frac{\delta_{j_*}}{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \cdot \left(\prod_{j=k-j_*+2}^{k-k'} \boldsymbol{R}_{k+1-j}\right) \boldsymbol{R}_0^{(k')} \\ =& \frac{\delta_{j_*}}{n} \boldsymbol{x}_i^\top \left(\prod_{j=1}^{k-j_*} \boldsymbol{R}_{k+1-j}\right) \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{H}_{k',j_*-1} \boldsymbol{R}_0^{(k')}.\end{aligned}$$

This implies that

$$c_{i,k,k',j_*-1} = \sum_{\boldsymbol{R}_{k+1-j} \in \{\boldsymbol{M}_{k+1-j}, \boldsymbol{M}_{k+1-j,i}\}, 1 \le j \le k-j_*} \frac{\delta_{j_*}}{n} \boldsymbol{x}_i^\top \left(\prod_{j=1}^{k-j_*} \boldsymbol{R}_{k+1-j}\right) \boldsymbol{x}_i,$$

which further tells us

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

We complete the proof of the lemma. □

As a consequence of Lemma 33, we can write

$$\begin{aligned}&\frac{2}{n} \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i}) \boldsymbol{x}_i^\top \nabla_{\boldsymbol{x}_s} \widehat{\boldsymbol{\beta}}_{k+1,-i} \\ =& \frac{2}{n} \sum_{i=1}^n \sum_{k'=0}^k \sum_{r=k'}^k c_{i,k,k',r} (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i}) \boldsymbol{x}_i^\top \boldsymbol{H}_{k',r} \cdot \left(\frac{\delta_{k'}}{n} (y_s - \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_{k',-i}) \boldsymbol{I}_{p+1} - \frac{\delta_{k'}}{n} \boldsymbol{x}_s \widehat{\boldsymbol{\beta}}_{k',-i}^\top \right) \\ =& \sum_{k'=0}^k \sum_{r=k'}^k (\boldsymbol{g}_{k,k',r,s} + \bar{\boldsymbol{g}}_{k,k',r,s}),\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{g}_{k,k',r,s} &:= \frac{2\delta_{k'}}{n^2} \sum_{i=1}^n c_{i,k,k',r} (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i})(y_s - \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_{k',-i}) \boldsymbol{x}_i^\top \boldsymbol{H}_{k',r}, \\ \bar{\boldsymbol{g}}_{k,k',r,s} &:= -\frac{2\delta_{k'}}{n^2} \sum_{i=1}^n c_{i,k,k',r} (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i}) \boldsymbol{x}_i^\top \boldsymbol{H}_{k',r} \boldsymbol{x}_s \widehat{\boldsymbol{\beta}}_{k',-i}^\top.\end{aligned}$$

We define $\boldsymbol{V}_{k,k',r}, \bar{\boldsymbol{V}}_{k,k',r} \in \mathbb{R}^{(p+1)\times n}$ such that the s-th columns correspond to $\boldsymbol{g}_{k,k',r,s}^\top$ and $\bar{\boldsymbol{g}}_{k,k',r,s}^\top$, respectively. We also define $\widetilde{\boldsymbol{V}}_k \in \mathbb{R}^{(p+1)\times n}$ such that the s-th column of this matrix corresponds to $2(y_s - \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_{k+1,-s}) \widehat{\boldsymbol{\beta}}_{k+1,-s}/n$. Inspecting Equation (56), we see that in order to upper bound the Frobenius norm of $\nabla_{\boldsymbol{X}} \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_{k+1})$, it suffices to upper bound the Frobenius norms of matrices $\boldsymbol{V}_{k,k',r}$, $\bar{\boldsymbol{V}}_{k,k',r}$, and $\widetilde{\boldsymbol{V}}_k$, which we analyze in the lemma below.

Lemma 34. *On the set Ω we have*

$$\|\boldsymbol{V}_{k,k',r}\|_F^2 \le \frac{4\delta_{k'}^2 e^{4\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}}{n} \cdot \bar{B}_*^2 \cdot \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0)^2 \cdot (\log n)^2, \tag{59}$$

$$\|\bar{\boldsymbol{V}}_{k,k',r}\|_F^2 \le \frac{4\delta_{k'}^2 e^{4\Delta C_{\Sigma,\zeta}} C_{\Sigma,\zeta}^2 B_*^2}{n} \cdot \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0)^2 \cdot (\log n)^2, \tag{60}$$

$$\|\widetilde{\boldsymbol{V}}_k\|_F^2 \le \frac{4B_*^2}{n} \cdot \mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0)^2 \cdot (\log n)^2. \tag{61}$$

Proof of Lemma 34. We observe that

$$\boldsymbol{V}_{k,k',r} = \frac{2\delta_{k'}}{n^2} \boldsymbol{H}_{k',r} \boldsymbol{X}^\top \boldsymbol{A}_{k,k',r},$$

where $\boldsymbol{A}_{k,k',r} \in \mathbb{R}^{n\times n}$, and $(\boldsymbol{A}_{k,k',r})_{is} = c_{i,k,k',r}(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i})(y_s - \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_{k',-i})$. Note that on Ω, by Corollary 30 and Lemma 31 we have

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

This further implies that

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

As a result,

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

which concludes the proof for the first inequality.

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

$$\begin{aligned} \bar{\boldsymbol{V}}_{k,k',r} &= -\frac{2\delta_{k'}}{n^2} \boldsymbol{Q}_{k,k',r} \boldsymbol{X} \boldsymbol{H}_{k',r} \boldsymbol{X}^\top, \\ \boldsymbol{Q}_{k,k',r} &= \left[\widehat{\boldsymbol{\beta}}_{k,-1} \mid \cdots \mid \widehat{\boldsymbol{\beta}}_{k,-n}\right] \cdot \mathrm{diag}\{(c_{i,k,k',r}(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i}))_{i=1}^n\} \in \mathbb{R}^{(p+1)\times n}. \end{aligned}$$

Therefore,

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

This completes the proof of Equation (60). Finally, we prove Equation (61). By Lemma 29 and Lemma 31 we obtain

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

This is exactly what we aim to prove. □

By triangle inequality,

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

The proof of Lemma 32 now follows by putting together the above upper bound and Lemma 34.

I.3 Upper bounding $\nabla_{\boldsymbol{y}} \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)$

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

More precisely, we will show the following:

Lemma 35 (Bounding norm of gradient with respect to response)**.** *On the set* Ω*,*

$$\begin{aligned}&\|\nabla_{\boldsymbol{y}} \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)\|_2 \\ \leq& \frac{2\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)}{\sqrt{n}} \cdot \sqrt{\log n} + \frac{2\Delta K C_{\Sigma,\zeta} e^{2\Delta C_{\Sigma,\zeta}}}{\sqrt{n}} \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0) \cdot \sqrt{\log n}.\end{aligned} \tag{62}$$

Proof of Lemma 35. For $s \in [n]$, we note that

$$\frac{\partial}{\partial y_s} \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) = \frac{2}{n}(y_s - \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_{k,-s}) - \frac{2}{n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i}) \boldsymbol{x}_i^\top \frac{\partial}{\partial y_s} \widehat{\boldsymbol{\beta}}_{k,-i}. \tag{63}$$

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

$$\begin{aligned}\frac{\partial}{\partial y_s}\widehat{\boldsymbol{\beta}}_{k+1,-i} =& \frac{\partial}{\partial y_s}\widehat{\boldsymbol{\beta}}_{k,-i} + \frac{\delta_k}{n}\boldsymbol{x}_s - \frac{\delta_k}{n}\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j^\top \frac{\partial}{\partial y_s}\widehat{\boldsymbol{\beta}}_{k,-i} \\ =& \boldsymbol{M}_k \frac{\partial}{\partial y_s}\widehat{\boldsymbol{\beta}}_{k,-i} + \boldsymbol{M}_{k,i}\frac{\partial}{\partial y_s}\widehat{\boldsymbol{\beta}}_{k,-i} + \frac{\delta_k}{n}\boldsymbol{x}_s,\end{aligned}$$

where we recall that $\boldsymbol{M}_k = (\boldsymbol{I}_{p+1} - \delta_k \widehat{\boldsymbol{\Sigma}})$ and $\boldsymbol{M}_{k,i} = \delta_k \boldsymbol{x}_i \boldsymbol{x}_i^\top / n$. Invoking the same argument that we employed to derive Lemma 33, we are able to conclude that

$$\frac{\partial}{\partial y_s}\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i} = \sum_{k'=0}^k \sum_{r=k'}^k c_{i,k,k',r}\boldsymbol{x}_i^\top \boldsymbol{H}_{k',r} \cdot \frac{\delta_{k'}}{n}\boldsymbol{x}_s.$$

Plugging this into Equation (63) leads to the following equality:

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

where

$$\eta_{k,k',r,s} = \frac{2}{n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i}) c_{i,k,k',r}\boldsymbol{x}_i^\top \boldsymbol{H}_{k',r} \cdot \frac{\delta_{k'}}{n}\boldsymbol{x}_s.$$

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

$$\begin{aligned}\boldsymbol{\eta}_{k,k',r} &= \frac{2\delta_{k'}}{n^2}\boldsymbol{X}\boldsymbol{H}_{k',r}\boldsymbol{X}^\top \boldsymbol{q}_{k,k',r}, \\ \boldsymbol{q}_{k,k',r} &= \big(c_{i,k,k',r}(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k+1,-i})\big)_{i=1}^n \in \mathbb{R}^n.\end{aligned}$$

We can upper bound the Euclidean norm of $\boldsymbol{\eta}_{k,k',r}$ using Lemma 31 and 33. More precisely,

$$\|\boldsymbol{\eta}_{k,k',r}\|_2 \leq \frac{2\delta_{k'}}{n^2}\|\boldsymbol{X}^\top\boldsymbol{X}\|_{\mathrm{op}} \cdot \|\boldsymbol{H}_{k',r}\|_{\mathrm{op}} \cdot \|\boldsymbol{q}_{k,k',r}\|_2 \leq \frac{2\delta_{k'}C_{\Sigma,\zeta}e^{2\Delta C_{\Sigma,\zeta}}}{\sqrt{n}} \cdot \mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\sqrt{\log n}.$$

Note that

$$\nabla_{\boldsymbol{y}}\widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) = \frac{2}{n}\boldsymbol{a}_k - \sum_{k'=0}^k\sum_{r=k'}^k \boldsymbol{\eta}_{k,k',r}.$$

Invoking triangle inequality and Lemma 31, we obtain

$$\begin{aligned}&\|\nabla_{\boldsymbol{y}} \widehat{R}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)\|_2\\ \leq & \frac{2\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)}{\sqrt{n}}\cdot\sqrt{\log n}+\frac{2\Delta K C_{\Sigma,\zeta}e^{2\Delta C_{\Sigma,\zeta}}}{\sqrt{n}}\cdot\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\sqrt{\log n}.\end{aligned} \tag{64}$$

This completes the proof. □

J Proof of ?? (for general risk functionals)

J.1 Outline of the proof

A visual schematic for the proof of **??** for general risk functionals is provided in Figure 12.

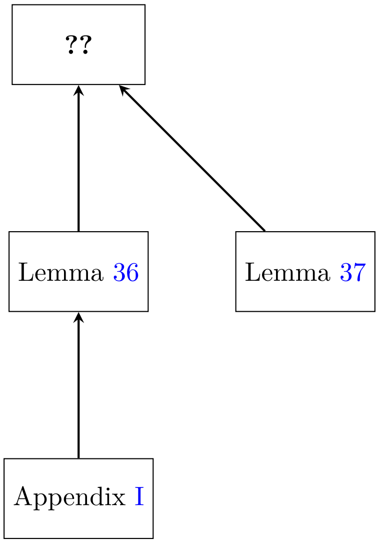

Figure 12: Schematic for the proof of **??** for general risk functionals

Once again, we will work on the set Ω, which we recall is defined in Equation (28). The proof idea is similar to that for the squared loss. More precisely, if we are able to prove Equations (68) to (70) 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:

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

J.2 Concentration analysis

As before, we will first prove that both $\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_{k+1})$ and $\Psi(\widehat{\boldsymbol{\beta}}_{k+1})$ concentrate. To this end, we shall again analyze the gradients and show that they are Lipschitz functions of the input data. Proof for this part is similar to the proof of Lemmas 14 and 15.

We let $f^{\psi}_{k+1}(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n) := \widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_{k+1})$, $\widetilde{f}^{\psi}_{k+1} := f^{\psi}_{k+1}\circ h$, $r^{\psi}_{k}(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n) := \Psi(\widehat{\boldsymbol{\beta}}_k)$, and $\widetilde{r}^{\psi}_{k} := r_k\circ h$. Our formal statement is provided below:

Lemma 36 (LOO and risk concentration analysis)**.** *Under the assumptions of Theorem* **??***, 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* $k\in\{0\}\cup[K]$

$$\left|\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)-\mathbb{E}[\widetilde{f}^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]\right|\leq\frac{2\sigma_{\mathsf{T}_2}LK\xi^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot(\log n)^{3/2}}{\sqrt{n}},$$

$$\left|\Psi(\widehat{\boldsymbol{\beta}}_k)-\mathbb{E}[\widetilde{r}^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]\right|\leq\frac{2\sigma_{\mathsf{T}_2}L\bar{\xi}^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)(\log n)^{3/2}}{\sqrt{n}}.$$

In the above display, $\xi^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)$ *and* $\bar{\xi}^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)0$ *are positive constants that depend only on* $(C_{\Sigma,\zeta},\Delta,m,B_0)$*.*

Proof of Lemma 36. We start by writing down the gradient. For all $s \in [n]$, note that

$$\nabla_{\boldsymbol{x}_s} \widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_{k+1}) = -\frac{1}{n}\partial_2\psi(y_s, \boldsymbol{x}_s^\top \widehat{\boldsymbol{\beta}}_{k+1,-s})\widehat{\boldsymbol{\beta}}_{k+1,-s}^\top - \frac{1}{n}\sum_{i=1}^n \partial_2\psi(y_i, \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{k+1,-i})\boldsymbol{x}_i^\top \nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k+1,-i},$$

where ∂_i stands for taking the partial derivative with respect to the i-th input. Here, $i \in \{1, 2\}$. By Lemma 33, on Ω we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \partial_2\psi(y_i, \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{k+1,-i})\boldsymbol{x}_i^\top \nabla_{\boldsymbol{x}_s}\widehat{\boldsymbol{\beta}}_{k+1,-i} \\
&= \sum_{k'=0}^{k}\sum_{r=k'}^{k} c_{i,k,k',r}\frac{1}{n}\sum_{i=1}^n \partial_2\psi(y_i, \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{k+1,-i})\boldsymbol{x}_i^\top \boldsymbol{H}_{k',r}\cdot\left(\frac{\delta_{k'}}{n}(y_s - \boldsymbol{x}_s^\top\widehat{\boldsymbol{\beta}}_{k',-i})\boldsymbol{I}_{p+1} - \frac{\delta_{k'}}{n}\boldsymbol{x}_s\widehat{\boldsymbol{\beta}}_{k',-i}^\top\right) \\
&= \sum_{k'=0}^{k}\sum_{r=k'}^{k}(\boldsymbol{g}_{k,k',r,s}^{\psi} + \bar{\boldsymbol{g}}_{k,k',r,s}^{\psi}),
\end{aligned}$$

where

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

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

$$\|\nabla_{\boldsymbol{X}}\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_{k+1})\|_F \leq \|\widetilde{\boldsymbol{V}}_k^{\psi}\|_F + \sum_{k'=0}^{k}\sum_{r=k'}^{k}\left\{\|\boldsymbol{V}_{k,k',r}^{\psi}\|_F + \|\bar{\boldsymbol{V}}_{k,k',r}^{\psi}\|_F\right\}. \tag{65}$$

Next, we upper bound $\|\boldsymbol{V}_{k,k',r}^{\psi}\|_F$, $\|\bar{\boldsymbol{V}}_{k,k',r}^{\psi}\|_F$, and $\|\widetilde{\boldsymbol{V}}_k^{\psi}\|_F$. We observe that

$$\boldsymbol{V}_{k,k',r}^{\psi} = \frac{\delta_{k'}}{n^2}\boldsymbol{H}_{k',r}\boldsymbol{X}^\top\boldsymbol{A}_{k,k',r}^{\psi}, \qquad \bar{\boldsymbol{V}}_{k,k',r}^{\psi} = -\frac{\delta_{k'}}{n^2}\boldsymbol{Q}_{k,k',r}^{\psi}\boldsymbol{X}\boldsymbol{H}_{k',r}\boldsymbol{X}^\top,$$

where

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

We let $\boldsymbol{a}_{k+1}^{\psi} = (\partial_2\psi(y_i, \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_{k+1,-i}))_{i=1}^n$. Recall that $\boldsymbol{a}_{k+1} = (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}_{k+1,-i})_{i=1}^n$. Using triangle inequality, we obtain that $\|\boldsymbol{a}_{k+1}^{\psi}\|_2 \leq 3C_\psi(\|\boldsymbol{a}_{k+1}\|_2 + \|\boldsymbol{y}\|_2) + \sqrt{2n}\bar{C}_\psi$. Invoking Lemma 31, we know that on Ω, $\|\boldsymbol{a}_{k+1}\|_2 \leq \sqrt{n}\mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0)\cdot\sqrt{\log n}$. Furthermore, by definition we know that on Ω, $\|\boldsymbol{y}\|_2 \leq \sqrt{n(m + \log n)}$. By Corollary 30 we see that $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{k,-i}\|_2 \leq \sqrt{n}\bar{B}_*\cdot\sqrt{\log n}$. By Lemma 33, we have $\|c_{i,k,k',r}\boldsymbol{H}_{k',r}\|_{\mathrm{op}} \leq e^{2\Delta C_{\Sigma,\zeta}}$. Putting together all these results, we conclude that

$$\begin{aligned}
\|\boldsymbol{V}_{k,k',r}^{\psi}\|_F \leq & \frac{\delta_{k'}}{n^2}\cdot\|\boldsymbol{H}_{k',r}\|_{\mathrm{op}}\cdot\|\boldsymbol{X}\|_{\mathrm{op}}\cdot\|\boldsymbol{A}_{k,k',r}^{\psi}\|_F \\
\leq & \frac{\delta_{k'}e^{2\Delta C_{\Sigma,\zeta}}C_{\Sigma,\zeta}^{1/2}\bar{B}_*\cdot(3C_\psi\mathcal{G}_2(C_{\Sigma,\zeta}, \Delta, m, B_0) + 3C_\psi\sqrt{m} + \sqrt{2}\bar{C}_\psi)\cdot\log n}{\sqrt{n}}.
\end{aligned} \tag{66}$$

Applying Lemma 29, we deduce that

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

Therefore,

$$\begin{aligned}
\|\bar{\boldsymbol{V}}^{\psi}_{k,k',r}\|_F &\le \frac{\delta_{k'}}{n^2}\cdot\|\boldsymbol{Q}^{\psi}_{k,k',r}\|\cdot\|\boldsymbol{X}\|_{\mathrm{op}}^2\cdot\|\boldsymbol{H}_{k',r}\|_{\mathrm{op}} \\
&\le \frac{\delta_{k'}B_* e^{2\Delta C_{\Sigma,\zeta}}C_{\Sigma,\zeta}\cdot(3C_\psi\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)+3C_\psi\sqrt{m}+\sqrt{2}\bar{C}_\psi)\cdot\log n}{\sqrt{n}}, \\
\|\widetilde{\boldsymbol{V}}^{\psi}_{k}\|_F &\le \frac{1}{n}\|\boldsymbol{a}^{\psi}_{k+1}\|_2\cdot\|\widehat{\boldsymbol{\beta}}_{k+1,-s}\|_2 \le \frac{B_*(3C_\psi\mathcal{G}_2(C_{\Sigma,\zeta},\Delta,m,B_0)+3C_\psi\sqrt{m}+\sqrt{2}\bar{C}_\psi)\cdot\log n}{\sqrt{n}}.
\end{aligned} \tag{67}$$

Combining Equations (65) to (67), we see that there exists a constant $\xi_1^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)$ that depends only on $(C_{\Sigma,\zeta},\Delta,m,B_0)$, such that on Ω, for all $k\in\{0\}\cup[K]$ we have

$$\|\nabla_{\boldsymbol{X}}\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_{k+1})\|_F \le \frac{K\xi_1^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n}{\sqrt{n}}.$$

Analogously, we are able to conclude the existence of a non-negative constant $\xi_2^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)$, such that on Ω, it holds that

$$\|\nabla_{\boldsymbol{y}}\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_{k+1})\|_F \le \frac{K\xi_2^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n}{\sqrt{n}}.$$

Hence, we know that $\|\nabla_{\boldsymbol{W}}\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_{k+1})\|_F \le K\xi^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot\log n$ if we set $\xi^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0) = \xi_1^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)+\xi_2^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)$. Following exactly the same steps that we used to derive Lemma 14, we deduce that with probability at least $1-2(n+p)^{-4}-(n\log^2 n)^{-1}m_4-2(K+1)C_{\mathsf{T}_2}n^{-2}$,

$$\left|\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)-\mathbb{E}[\widetilde{f}^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]\right| \le \frac{2\sigma LK\xi^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)\cdot(\log n)^{3/2}}{\sqrt{n}}. \tag{68}$$

Similarly, we are able to 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 $k\in\{0\}\cup[K]$,

$$\left|\Psi(\widehat{\boldsymbol{\beta}}_k)-\mathbb{E}[\widetilde{r}^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]\right| \le \frac{2\sigma L\bar{\xi}^{\psi}(C_{\Sigma,\zeta},\Delta,m,B_0)(\log n)^{3/2}}{\sqrt{n}}. \tag{69}$$

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

J.3 Uniform consistency

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

Lemma 37 (LOO and risk bias analysis)**.** *On the set Ω, it holds that*

$$\begin{aligned}
\sup_{k\in\{0\}\cup[K]}\left|\mathbb{E}[\widetilde{r}^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]-\mathbb{E}[r^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]\right| &= o_n(1), \\
\sup_{k\in\{0\}\cup[K]}\left|\mathbb{E}[\widetilde{f}^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]-\mathbb{E}[f^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]\right| &= o_n(1).
\end{aligned} \tag{70}$$

Proof of Lemma 37. Using Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
\left|\mathbb{E}[r^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]-\mathbb{E}[\widetilde{r}^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]\right| &\le \mathbb{P}(\Omega^c)^{1/2}\mathbb{E}[\psi(y_0,\boldsymbol{x}_0^\top\widehat{\boldsymbol{\beta}}_k)^2]^{1/2}, \\
\left|\mathbb{E}[f^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]-\mathbb{E}[\widetilde{f}^{\psi}_k(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)]\right| &\le \mathbb{P}(\Omega^c)^{1/2}\mathbb{E}[\psi(y_1,\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}}_{k,-1})^2]^{1/2}.
\end{aligned} \tag{71}$$

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 (53) together, we obtain that

$$\mathbb{E}[\psi(y_1,\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}}_{k,-1})^2] \le \phi_\psi\mathbb{E}[\|(y_1,\boldsymbol{x}_1^\top\widehat{\boldsymbol{\beta}}_{k,-1})\|_2^4]+\bar{\phi}_\psi \le \phi_\psi\mathcal{H}(\sigma_\Sigma,\zeta,B_0,m_8,\Delta)^2+\bar{\phi}_\psi. \tag{72}$$

Recall that $\mathbb{P}(\Omega^c) \le 2(n+p)^{-4} + n^{-1}m_4$. Combining this, Equations (71) and (72), we are able to establish Equation (70).

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

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,

$$\begin{aligned}
&\left|\mathbb{E}[\Psi(\widehat{\boldsymbol{\beta}}_k)] - \mathbb{E}[\Psi(\widehat{\boldsymbol{\beta}}_{k,-1})]\right| \\
&= \left|\mathbb{E}[r_k^\psi(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)] - \mathbb{E}[f_k^\psi(\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)]\right| \\
&= \left|\mathbb{E}[\psi(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k)] - \mathbb{E}[\psi(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_{k,-1})]\right| \\
&\le \varphi_\psi \mathbb{E}\left[\left(1 + \|(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k)\|_2^2 + \|(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_{k,-1})\|_2^2\right) \cdot |\boldsymbol{x}_0^\top (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-1})|\right] \\
&\le 3\varphi_\psi \mathbb{E}\left[(\boldsymbol{x}_0^\top (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-1}))^2\right]^{1/2} \cdot \mathbb{E}\left[1 + \|(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k)\|_2^4 + \|(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_{k,-1})\|_2^4\right] \\
&\le 3\varphi_\psi (\sigma_\Sigma + 1)^{1/2} \mathbb{E}\left[\|\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{k,-1}\|_2^2\right]^{1/2} \cdot \mathbb{E}\left[1 + \|(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k)\|_2^4 + \|(y_0, \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_{k,-1})\|_2^4\right],
\end{aligned}$$

which by Equations (53) and (55) goes to zero as $n, p \to \infty$. □

K Proof of Theorem 4

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 **??**, then as $n, p \to \infty$ we still have

$$\sup_{k \in \{0\} \cup [K]} |\widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) - \Psi(\widehat{\boldsymbol{\beta}}_k)| \xrightarrow{\text{a.s.}} 0. \tag{73}$$

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

$$\widehat{\Psi}_j^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) = \frac{1}{n}\sum_{i=1}^n g_j(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i}), \qquad \Psi_j(\widehat{\boldsymbol{\beta}}_k) = \mathbb{E}[g_j(y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k) \mid \boldsymbol{X}, \boldsymbol{y}].$$

By **??** we konw that $\sup_{k \in \{0\} \cup [K]} |\widehat{\Psi}_j^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) - \Psi_j(\widehat{\boldsymbol{\beta}}_k)| \xrightarrow{\text{a.s.}} 0$ for every j. Furthermore, notice that $|\widehat{\Psi}_j^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k) - \widehat{\Psi}^{\mathrm{loo}}(\widehat{\boldsymbol{\beta}}_k)| \le 2^{-j}$ and $|\Psi_j(\widehat{\boldsymbol{\beta}}_k) - \Psi(\widehat{\boldsymbol{\beta}}_k)| \le 2^{-j}$ and j is arbitrary, thus completing the proof of Equation (73).

We denote by $\widehat{F}_k$ the c.d.f. of the uniform distribution over $\{y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{k,-i} : i \in [n]\}$, and denote by F_k the c.d.f. of $y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k$ conditioning on $(\boldsymbol{X}, \boldsymbol{y})$. We emphasize that both F_k and $\widehat{F}_k$ are random distributions that depend on $(\boldsymbol{X}, \boldsymbol{y})$. Next, we prove that F_k is Lipschitz continuous.

Lemma 38. *Under the conditions of Theorem 4, F_k is κ_{pdf}-Lipschitz continuous.*

Proof. Note that $y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k = \rho(\boldsymbol{x}_0) - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k + \varepsilon_0$, where ε_0 is independent of $\rho(\boldsymbol{x}_0) - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k$. Since ε_0 has a p.d.f., we see that $y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k$ also has a p.d.f., and we denote it by h. We denote by h_ε the p.d.f. of ε_0 and denote by G the c.d.f. of $\rho(\boldsymbol{x}_0) - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k$, 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 38 and the fact that $y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_k$ has bounded fourth moment (see Equation (53) 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.

L Additional numerical illustrations and setup details

L.1 Setup details for Figure 2

- Feature model: The feature $\boldsymbol{x}_i \in \mathbb{R}^p$ is generated according to

  $$\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2} \boldsymbol{z}_i, \tag{74}$$

  where $\boldsymbol{z}_i \in \mathbb{R}^p$ contains independently sampled entries from a common distribution, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is a positive semidefinite feature covariance matrix. The different distributions that we use for the components of z_i include: (1) Gaussian distribution, (2) Student's t-distribution These represent a mix of both continuous and discrete, and light- and heavy-tailed distributions. We standardize the distributions so that the mean is zero and the variance is one. The different feature covariance matrix structures that we use include: (1) Identity ($\boldsymbol{\Sigma}_{ij} = 1$ when $i = j$ and $\boldsymbol{\Sigma}_{ij} = 0$ when $i \neq j$) and (2) Autoregressive with parameter ρ ($\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$ for all i, j).

- Response model: Given $\boldsymbol{x}_i$, the response $y_i \in \mathbb{R}$ is generated according to

  $$y_i = \boldsymbol{\beta}_0^\top \boldsymbol{x}_i + \left(\boldsymbol{x}_i^\top \boldsymbol{A} \boldsymbol{x}_i - \operatorname{tr}[\boldsymbol{A} \boldsymbol{\Sigma}] \right)/p + \varepsilon_i, \tag{75}$$

  where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is a fixed signal vector, $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ is a fixed matrix, and $\varepsilon_i \in \mathbb{R}$ is a random noise variable. Note that we have subtracted the mean from the squared nonlinear component and scaled it to keep the variance of the nonlinear component at the same order as the noise variance (see Mei and Montanari (2022) for more details, for example). We again use either Gaussian, Student's t, for the random noise component, which is again standardized so that the mean is zero and the variance is one. We refer to the value of $\boldsymbol{\beta}_0^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_0$ as the effective signal energy. It is worth remarking that even though the regression function above does not satisfy the assumptions of Assumption C, it is easy to see that function is approximately Lipschitz.

L.2 Predictive intervals under a linear response model

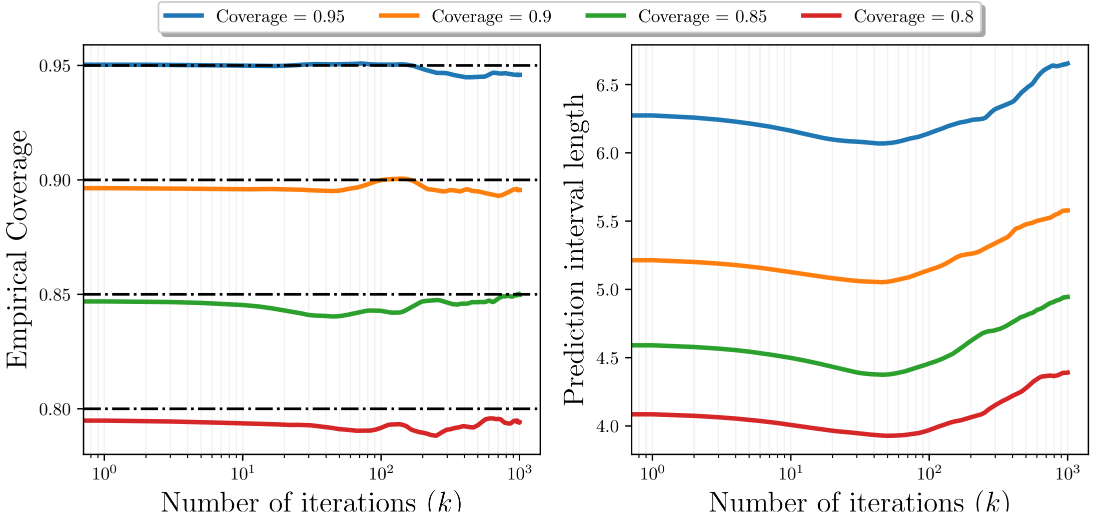

Figure 13: Prediction intervals: length and coverage. We consider an overparameterized regime where the number of observations is $n = 2500$ and the number of features is $p = 5000$. The non-intercept features are Gaussian with a ρ-autoregressive covariance Σ (such that $\Sigma_{ij} = \rho^{|i-j|}$ for all i, j) with $\rho = 0.25$. The response is generated from a linear model with a nonrandom signal vector β_0 that has unit Euclidean norm. We initialize the GD process randomly, and employ a universal step size $\delta = 0.01$. In the left figure we plot the empirical coverage rates with various levels, and in the right plot we plot the length of the prediction intervals. All simulation outcomes are based on one realization of $(\boldsymbol{X}, \boldsymbol{y})$.