Electronic Journal of Statistics
Vol. 18 (2024) 4119–4197
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2297

Benign overfitting of non-sparse high-dimensional linear regression with correlated noise

Toshiki Tsuda[*]

Yale University,
e-mail: toshiki.tsuda@yale.edu

Masaaki Imaizumi

The University of Tokyo
RIKEN Center for Advanced Intelligence Project,
e-mail: imaizumi@g.ecc.u-tokyo.ac.jp

Abstract: We investigate the high-dimensional linear regression problem in the presence of noise that is correlated with Gaussian covariates. This type of correlation, known as endogeneity in regression models, often results from unobserved variables and other factors. It poses a significant challenge in causal inference and econometrics. In cases where covariates are high-dimensional, it is common to assume sparsity in the true parameters and to estimate them using regularization techniques, even with endogeneity. However, when sparsity is not applicable, controlling both endogeneity and high dimensionality simultaneously has not been well understood. This study demonstrates that an estimator, even without regularization, can achieve consistency, or benign overfitting, under certain assumptions about the covariance matrix. Specifically, our results indicate that the error of this estimator converges to zero when the covariance matrices of the correlated noise and the instrumental variables meet specific conditions related to their eigenvalues. We explore several extensions that relax these conditions and conduct experiments to validate our theoretical findings. As a technical contribution, we employ the convex Gaussian minimax theorem (CGMT) in our dual problem and expand upon CGMT itself.

MSC2020 subject classifications: Primary 62J05.
Keywords and phrases: High-dimension, endogeneity, linear regression, non-sparsity, convex Gaussian minimax theorem.

Received April 2023.

1. Introduction

We consider a high-dimensional linear regression model with correlated noise and the p-dimensional true parameter θ_0:

$$Y_i = \langle X_i, \theta_0 \rangle + \xi_i, \quad \mathbb{E}[X_i \xi_i] \neq 0, \quad i = 1, ..., n,$$

[*] The author was affiliated with the University of Tokyo at the time of writing this paper.

where $n \in \mathbb{N}$ is the number of observations, X_i denotes a p-dimensional centered Gaussian vector representing the observed covariates, ξ_i is a centered Gaussian noise variable, and Y_i is a response variable. In this model, the noise variable ξ_i is correlated with the covariate X_i. We assume that the dimension p is significantly larger than the number of observations n ($p \gg n$) and also the true parameter θ_0 is non-sparse, meaning all p coordinates of θ_0 are non-zero. In this high-dimensional setting, we adopt an *instrumental variable* framework, that is, assuming the existence of a variable Z_i such that $\mathbb{E}[Z_i\xi_i] = 0$, we investigate the risk of a *ridgeless estimator* under certain conditions.

We often encounter real-world situations where covariates and noise variables are correlated. For example, when the covariates in a regression model are partially observed, the partially observed subset may be labeled as X_i and the rest as components of ξ_i. In this case, X_i and ξ_i may be correlated. This scenario often introduces bias in statistical methods that assume independence between covariates X_i and noise variables ξ_i. This phenomenon is known as *endogeneity*, a term particularly common in econometrics. A well-established approach to addressing endogeneity is the use of *instrumental variables*, as discussed in the survey by (Stock *et al.*, 2002). This method employs a variable Z_i that is uncorrelated with the noise variable ξ_i but is related to the covariates X_i. The two critical conditions for instrumental variables Z_i, known as *exclusion restriction* and *relevance restriction*, are essential. This method has been extensively explored in various studies (Söderström and Stoica, 2002; Newey and Powell, 2003; Baiocchi *et al.*, 2014; Andrews *et al.*, 2019).

Instrumental variable estimation has also been extensively investigated in the high-dimensional context, particularly within the *sparse* setting. As data become high-dimensional, the dimensionality of the covariates associated with instrumental variables often exceeds the number of observations, n. To address this, researchers have utilized sparsity, which assumes that most of the p coordinates of the true parameters θ_0 are zero. Consequently, they estimate a limited number of nonzero parameters using lasso-type regularization and its variants. Fan and Liao (2014) proposed a new generalized method of moments estimator for estimation and model selection involving sparse parameters. Belloni *et al.* (2014) and Gautier and Rose (2011) focus on scenarios where θ_0 is (approximately) sparse, employing a lasso-type regularization or the Dantzig selector within the instrumental variable framework. Gold *et al.* (2020) explore a one-step update approach and provide sufficient conditions for inference. Various studies (Belloni *et al.*, 2010, 2012, 2017; Chernozhukov *et al.*, 2015b, 2018; Belloni *et al.*, 2022; Gautier and Tsybakov, 2013) have estimated nuisance parameters to manage their high dimensionality by introducing a new instrumental variable orthogonal to these parameters.

In recent years, high-dimensional statistics featuring non-sparse parameters have rapidly evolved. The development of methods for handling large-scale data, such as those employed in modern machine learning, has led to the proliferation of many non-sparse data sets and models. Among several existing methods, a *ridgeless estimator*, which fits the observed data perfectly without any regularization, has garnered considerable attention. For theoretical analysis, Belkin *et al.* (2019) and Hastie *et al.* (2022) investigate the performance of the ridgeless estimator in high-dimensional linear regression models without sparsity, employing random matrix theory. Bartlett *et al.* (2020) discuss how the effective ranks of a covariance matrix can demonstrate the convergence of the ridgeless estimator in high-dimensional linear regression contexts. These studies have demonstrated several advantages of the ridgeless estimator over regularized estimators in high-dimensional settings (Dobriban and Wager, 2018; Tsigler and Bartlett, 2023). The implications of these findings have been extended to various applications (Bunea *et al.*, 2022; Li *et al.*, 2022; Frei *et al.*, 2022; Nakakita and Imaizumi, 2022). However, despite ongoing developments, these theories remain limited and necessitate independence among noise variables, thus, they are insufficiently flexible to analyze within an instrumental variable framework.

This study investigates a ridgeless estimator in non-sparse, high-dimensional regression models with correlated noise. We assume that the data follow a centered Gaussian distribution and model the correlation between covariates and noise variables using instrumental variables. We assess the estimation error of the ridgeless estimator using a projected residual mean squared error (projected RMSE). This analysis yields two main results: (i) The estimation error has an upper bound that is independent of the dimension p of the covariates. This bound can be expressed through the (normalized) correlation coefficients and the effective rank of the covariance matrix of the instrumental variables. (ii) We identify sufficient conditions on data distributions under which the derived upper bound converges to zero. These conditions require that the covariance matrices of both the instrumental variables and auxiliary variables for covariates have appropriate effective ranks. Several specific covariance matrices meet these conditions. We derive these results for cases where instrumental and auxiliary variables that comprise the covariates are either orthogonal or not.

Our theoretical results suggest the following implications: (i) In the correlated noise setting, the error of the ridgeless estimator does not depend on the dimension p, given the use of instrumental variables. This implies that non-sparse, high-dimensional parameters can be estimated effectively under the instrumental variable framework; in other words, benign overfitting is possible. (ii) In this context, the covariance of the instrumental variables plays a crucial role in determining the risk. Specifically, the covariance matrix of the instrumental variables should possess a certain rank and exhibit decay to some degree to ensure that the sum of eigenvalues does not diverge excessively. Thus, these results support the established notion that instrumental variables should not be weak.

On the technical side, we have developed a proof technique for the evaluation of risks using Gaussian comparison inequalities. The most closely related study (Bartlett *et al.*, 2020) on non-sparse high-dimensional regression relies on matrix concentration inequalities and the leave-one-out method; however, these approaches cannot handle the correlated noise in our setting. Therefore, we have developed a proof using the convex Gaussian minimax theorem (CGMT) (Thrampoulidis *et al.*, 2015, 2018), which accommodates a broader range of models. Rigorously, we rewrite the risk associated with the ridgeless estimator as a minimax optimization problem in dual form and analyze it using CGMT.

This approach was developed by Koehler *et al.* (2021), and we have applied it to the instrumental variable case. Furthermore, we derive an extended CGMT and develop a method to analyze risks in situations where the covariance matrix is not orthogonal.

1.1. Notation

We denote $\|\cdot\|_p$ as $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$. For a square matrix A, we define $\|A\|_{op}$ as an operator norm of A. For a positive semidefinite matrix A, $\|x\|_A^2 := \langle x, Ax\rangle$ denotes the Mahalanobis (semi-)norm. For a set $S \subset \mathbb{R}^p$, we define its radius as $\mathrm{rad}(S) := \sup_{s\in S} \|s\|_2$. Σ^+ denotes the generalized inverse matrix of Σ. $N(\mu, \Sigma)$ denotes a multivariate normal distribution with a mean $\mu \in \mathbb{R}^d$ and a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$. If Σ is positive semi-definite with rank $k < d$, $N(0, \Sigma)$ denotes a distribution of AX' where $A \in \mathbb{R}^{d\times k}$, $\Sigma = AA^\top$, and $X' \sim N(0, I)$ is a k-dimensional normal variable. $\mathbb{1}\{\cdot\}$ denotes an indicator function. For $x, x' \in \mathbb{R}, x \vee x' := \max\{x, x'\}$. For $a, b \in \mathbb{R}$, $\lesssim$ and $\gtrsim$ mean $a \leq Cb$ and $Ca \geq b$ for some absolute constant C, respectively. For real-valued sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, $a_n = O(b_n)$ means $a_n/b_n \leq C$ for any sufficiently large n, $a_n = o(b_n)$ means a_n/b_n converges to zero as $n \to \infty$, $a_n = \upsilon(b_n)$ means a_n/b_n diverges to ∞ as $n \to \infty$, and $a_n = \Theta(b_n)$ means $C_1 b_n \leq a_n \leq C_2 b_n$ holds for any sufficiently large n, where C, C_1, and C_2 are some absolute constants. Let $\xrightarrow{\mathbf{p}}$ denote the convergence in probability.

1.2. Paper organization

Section 2 presents the problem setup and various definitions. Section 3 provides an error analysis of the ridgeless estimator under the assumption that the covariance matrices of the noise and instrumental variables are orthogonal. Section 4 provides an error analysis under a relaxation of orthogonality. Section 5 offers additional error analysis with a generalized norm. Section 6 outlines the proof and explains the technical contributions. Section 7 relates the experiments. Section 8 presents the discussion and conclusion.

2. Preliminary

2.1. Setting

We consider a linear regression problem with dependent noise and instrumental variables. Let $n \in \mathbb{N}$ be the number of data points, $p, k \in \mathbb{N}$ be dimensions of variables, and $\Theta \subset \mathbb{R}^p$ be the parameter space. Suppose that there exist n i.i.d. variables $(X_i, Z_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}^k \times \mathbb{R}$ of the centered variables for $i = 1, \cdots, n$ from the following data generating process

$$Y_i = \langle X_i, \theta_0\rangle + \xi_i, \text{ and } X_i = \Pi_0 Z_i + u_i, \tag{1}$$

where $\theta_0 \in \mathbb{R}^p$ is a true unknown parameter such that $\|\theta_0\|_2 < \infty$, ξ_i is a Gaussian variable from $N(0, \sigma^2)$, $\Pi_0 \in \mathbb{R}^{p \times k}$ is an unknown matrix, and $u_i \in \mathbb{R}^p$ is a (potentially correlated) latent noise vector such that $\mathbb{E}[u_i|Z_i] = 0$. Here, we refer to X_i as a covariate and Z_i as an instrumental variable. We assume that $\mathbb{E}[X_i|Z_i]$ always exists. We define covariance matrices $\Sigma_x = \mathbb{E}[X_i X_i^\top]$, $\Sigma_z = \mathbb{E}[Z_i Z_i^\top]$, and $\Sigma_u = \mathbb{E}[u_i u_i^\top]$. Let $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \xi)$ denote design matrices and vectors $\mathbf{X} = (X_1, ..., X_n)^\top, \mathbf{Y} = (Y_1, ..., Y_n)^\top, \xi = (\xi_1, ..., \xi_n)^\top$, and $\mathbf{Z} = (Z_1, ..., Z_n)^\top$. Note that the covariance matrices Σ_z and Σ_u need not be positive definite, that is, positive semi-definite is sufficient for our analysis.

We describe how these variables are related. We define $\omega \in \mathbb{R}^p$ as the correlation between the covariate X_i and the noise ξ_i:

$$\omega := \mathbb{E}[X_i \xi_i] \neq 0.$$

Further, we assume that the instrument Z_i satisfies the following moment condition:

$$\mathbb{E}[\xi_i|Z_i] = 0,$$

which implies the instrument Z_i and its noise ξ_i are uncorrelated, that is, $\mathbb{E}[Z_i \xi_i] = 0$.

Remark 1 (Modeling with Π_0)**.** We employ the modeling (1), because of the following two reasons. First, this model is often used in applied fields (e.g., econometrics and psychostatistics) Newey and Powell (2003); Chen and Pouzo (2012), that study a specific interpretation of instrumental variables. Second, the usage of the coefficient Π_0 yields the property $\mathbb{E}[u_i|Z_i] = 0$, which simplifies theoretical analysis for an estimation error.

We make an assumption concerning the problem.

Assumption 1 (Gaussianity)**.** *Assume X_i and ξ_i are normally distributed, that is,*

$$X_i \sim N(0, \Sigma_x), \quad \xi_i \sim N(0, \sigma^2).$$

For X_i, Assumption 1 enables us to use the convex Gaussian minimax theorem (CGMT), which is a central tool to derive the upper bound for the risk. A possible way to mitigate Gaussianity includes the application of universality Montanari and Saeed (2022); Han and Shen (2023). As long as X_i is Gaussian, Z_i and u_i do not have to be Gaussian.

Indeed, we can relax the Gaussianity of ξ_i. While we use the Gaussianity of ξ_i to show the concentration of a norm of ξ_i (specifically, Lemma 40 in Appendix), it is possible to extend the concentration to sub-Gaussian vectors. However, since relaxing the Gaussianity is not our primary concern, we use ξ_i as a Gaussian variable here for simplicity.

2.2. *Measure for estimation error*

A goal of the setting is to estimate the true parameter θ_0 in a high-dimensional setting, that is, $p, k \gg n$, without the sparse setting. Specifically, for $\theta \in \Theta$, we consider a residual mean squared error (RMSE) projected on a space of Z:

$$E\left[\left(E[\langle\theta, X\rangle - \langle\theta_0, X\rangle | Z]\right)^2\right], \tag{2}$$

where the random element (X, Z) is an i.i.d. copy of (X_i, Z_i) that follows (1). In the literature of nonparametric instrumental variables, the projected RMSE is often used to evaluate the convergence rate of the estimators (Ai and Chen, 2003; Chen and Pouzo, 2012; Dikkala *et al.*, 2020). It is because we need to deal with ill-posedness in nonparametric instrumental variable estimators. The projected RMSE can sometimes relate to the original, unprojected RMSE. As an extreme case, when X and Z are perfectly correlated, the projected and unprojected RMSEs are identical. For more details, see Chen and Pouzo (2012). In our setting, we use this useful evaluation criterion because we face difficulty evaluating RMSE in non-sparse high-dimensional settings. However, by definition, the Projected RMSE has validity due to being proper to the setting of instrumental variables. Furthermore, it always holds that the projected RMSE is not larger than the RMSE. Hence, our results can be necessary conditions for the convergence of the RMSE.

Note that the projected RMSE can be expressed as a weighted norm $\|\theta - \theta_0\|_{\Xi_z}^2$ with a transformed covariance matrix $\Xi_z := \Pi_0 E[ZZ^\top]\Pi_0^\top$:

$$\begin{aligned}(2) &= (\theta - \theta_0)^\top E\left[E[X|Z]E[X^\top|Z]\right](\theta - \theta_0)\\ &= (\theta - \theta_0)^\top \Pi_0 E\left[ZZ^\top\right]\Pi_0^\top(\theta - \theta_0)\\ &= \|\theta - \theta_0\|_{\Xi_z}^2.\end{aligned}$$

The second equation follows the property $\mathbb{E}[u_i|Z_i] = 0$, which follows the modeling (1).

Remark 2 (Norm weighted by Ξ_z)**.** We discuss some characteristics of the usage of the weighted norm by Ξ_z.

First, the use of norms weighted by covariance matrices is common in the studies for the benign-overfitting. For example, in the linear regression with independent noise case, Hastie *et al.* (2022) and Bartlett *et al.* (2020) study an estimation error in terms of a norm weighted by Σ_x, which corresponds to a predictive risk. In our setting, we use the norm weighted by Ξ_z as an analogy.

Second, the matrix Ξ_z has a reasonable design as a weight, since it is an asymptotically full-rank matrix in our setting. Rigorously, we will introduce a basic condition in Definition 2 and it shows that the effective rank of Ξ_z should diverge to infinity as n increases. Hence, as long as it satisfies the condition, the weighting matrix Ξ_z is asymptotically full rank and encompasses all eigenvalues.

2.3. Ridgeless estimator

We consider an estimator with interpolation, that is, a prediction by an estimator perfectly corresponds to the response in the observed set of data, which always appears when $p \geq n$ holds. Rigorously, with an empirical squared risk

$$\widehat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle X_i, \theta\rangle)^2, \tag{3}$$

the estimator with interpolation is a parameter $\Theta \subset \mathbb{R}^p$ that satisfies $\widehat{L}(\theta) = 0$. As there may be an infinite number of interpolators, we define a ridgeless estimator, also known as a minimal norm interpolator, as

$$\widehat{\theta} = \operatorname*{argmin}_{\theta\in\Theta:\widehat{L}(\theta)=0} \|\theta\|_2 = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^+\mathbf{Y}.$$

Note that we can calculate the minimum norm interpolator only from $(\mathbf{X}, \mathbf{Y})$.

Such estimators have been examined frequently in the context of the linear regression problem. In particular, the motivation for examining the ridgeless estimator (the minimum norm interpolator) is that the gradient descent algorithm for learning parameters converges to a parameter with the smallest norm among parameters that minimize the loss (see Lemma 1 in Hastie *et al.* (2022)).

3. Error analysis: orthogonal case

3.1. Orthogonality assumption

In this section, we consider a setting in which there is orthogonality between the transformed covariance matrix of instrumental variables $\Xi_z = \Pi_0 E[ZZ^\top]\Pi_0^\top$ and the covariance matrix of the latent noise $\Sigma_u = E[uu^\top]$. This situation simplifies our error analysis and is therefore an appropriate first step. This assumption will be relaxed in the next section.

Specifically, we consider the following assumption.

Assumption 2 (Orthogonality Condition)**.** Σ_u *and* Ξ_z *are orthogonal, that is, their sets of eigenvectors* $\{\varphi_j\}_{j=1}^{J_u}, \{\varphi'_j\}_{j=1}^{J_z} \subset \mathbb{R}^p$ *are such that there exists the decompositions* $\Sigma_u = \sum_{j=1}^{J_u} \lambda_j^u \varphi_j\varphi_j^\top$ *and* $\Xi_z = \sum_{j=1}^{J_z} \lambda_j^z \varphi'_j(\varphi'_j)^\top$ *with* $J_u + J_z = p$ *and positive eigenvalues* $\{\lambda_j^u\}_j$ *and* $\{\lambda_j^z\}_j$ *satisfying* $\varphi_j^\top\varphi'_\ell = 0$ *for every* j *and* ℓ*.*

Intuitively, the p-dimensional eigenspaces of Σ_x are divided into J_u-dimensional eigenspaces of Σ_u and J_z-dimensional eigenspaces of Ξ_z, which are orthogonal. We note two points. In this setting, the ranks of Σ_u and Ξ_z are J_u and J_z, respectively; hence they are not full-rank. Consequently, we obtain the following equality:

Lemma. *Assume Assumption 2 holds. Then, the positive semidefinite matrices Ξ_z and Σ_u whose eigenspaces are orthogonal satisfy the following covariance splitting:*

$$\Sigma_x = \Xi_z + \Sigma_u.$$

We will restate this result as Lemma 37 in the supplementary material and offer its proof. This property is essential for our error analysis below, which uses the speed of decay of the eigenvalues.

3.2. Result 1: upper bound on projected RMSE

Here, as the first primary result, we derive an upper bound for the projected RMSE of the ridgeless estimator. As preparation, we introduce a notion of the effective rank for the upper bound.

Definition 1 (Effective Rank)**.** For a positive semidefine matrix Σ, two types of the effective rank are defined as

$$r(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|_{\mathrm{op}}} \quad \textit{and} \quad R(\Sigma) = \frac{\mathrm{tr}(\Sigma)^2}{\mathrm{tr}(\Sigma^2)}.$$

This notion is a more elaborate version of the notion of matrix ranks, which uses the decay speed of the eigenvalues of a matrix to express the complexity of the matrix. Specifically, $r(\Sigma)$ denotes a trace of Σ normalized by its largest eigenvalue, and $R(\Sigma)$ denotes the intrinsic complexity of Σ considering the decay rate of the eigenvalues of Σ. As these effective ranks fully utilize the information of eigenvalues of Σ, they are useful in measuring the complexity of Σ and the stable quantity compared with the usual rank, especially in the high-dimensional setting. This has been used in dealing with concentration of random matrices (Koltchinskii and Lounici, 2017) and has also been applied to the analysis of over-parameterized linear regression with independent noise (Bartlett *et al.*, 2020; Koehler *et al.*, 2021; Tsigler and Bartlett, 2023).

Using the notion of effective rank, we define an auxiliary coefficient as follows. For $\delta \in (0,1)$, we define

$$\eta(\delta) := \sqrt{\log(1/\delta)} \left(\frac{1}{\sqrt{r(\Xi_z)}} + \sqrt{\frac{\mathrm{rank}(\Sigma_u)}{n}} + \frac{n}{R(\Xi_z)} \right).$$

This coefficient $\eta(\delta)$ becomes asymptotically negligible under appropriate conditions, which will be presented in the latter half of this section.

We develop a generic bound for the projected RMSE $\|\widehat{\theta} - \theta\|_{\Xi_z}^2$. With the result of Corollary 10 and Theorem 21, we obtain the following sufficient conditions for benign overfitting. Recall that we define Σ_u^+ as the generalized inverse matrix of Σ_u.

Theorem 1 (Projected-RMSE Bound)**.** *Fix any $\delta \leq 1/2$. Under Assumptions 1-2 with covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$, suppose that n and the effective ranks are such that $R(\Xi_z) \gtrsim \log(1/\delta)^2$ and $\eta(\delta) \leq 1$. Define $\psi(t) = t + t^2$ and $\widetilde{\sigma}^2 := \sigma^2 - \|\omega\|^2_{\Sigma_u^+} \geq 0$. Then, with probability at least $1 - \delta$, it holds that*

$$\|\widehat{\theta} - \theta_0\|^2_{\Xi_z} \lesssim (1 + \eta(\delta))(1 \vee \widetilde{\sigma})\psi\left((\|\Sigma_u^+\omega\|_2 + \|\theta_0\|_2)\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}}\right). \tag{4}$$

This upper bound consists of the following two parts: (i) the coefficient part $(1+\eta(\delta))(1\vee\widetilde{\sigma})$ reflects the asymptotically negligible eigenvalues and noises, and (ii) the principal part $\psi((\|\Sigma_u^+\omega\|_2 + \|\theta_0\|_2)\sqrt{\mathrm{tr}(\Xi_z)/n})$ describes a complexity of the true parameter and the distribution of the data. With this upper bound, an appropriate assumption on Σ_u and Ξ_z guarantees that the projected RMSE converges to zero as $n \to \infty$, which will be explained below. Note that $\widetilde{\sigma}^2 \geq 0$ follows from Lemma 38.

The upper bound for this projected RMSE can sometimes be compared to that of the original unprojected RMSE. As mentioned in Section 2.3, when X and Z are perfectly correlated as an extreme case, the projected and unprojected RMSEs are identical, which means that Theorem 1 can bound the original RMSE. However, in more general case including overparameterization, it is not easy to provide an explicit relation between the projected and unprojected RMSE. Some discussions are mentioned in Chen and Pouzo (2012).

Remark 3 (Comparison with the independent noise case)**.** We compare Theorem 1 with the endogeneity to the result without the endogeneity. Particularly, Koehler *et al.* (2021) develop an upper bound of the mean squared error of the ridgeless estimator as

$$\|\widehat{\theta} - \theta_0\|^2_{\Sigma_x} \lesssim (1 + \eta'(\delta))(1 \vee \sigma)\psi\left(\|\theta_0\|_2\sqrt{\frac{\mathrm{tr}(\Sigma_2)}{n}}\right), \tag{5}$$

where Σ_1 and Σ_2 are some matrices such that $\Sigma_x = \Sigma_1 + \Sigma_2$, and $\eta'(\delta) = \sqrt{\log(1/\delta)}(1/\sqrt{r(\Sigma_2)} + \sqrt{\mathrm{rank}(\Sigma_1)/n} + n/R(\Sigma_2))$. Note that Theorem 1 is not a direct generalization of this result (5), since we do not split Σ_x itself. This result (5) suggests that our bound in Theorem 1 pays an additional cost to handle covariate correlations, such as the replacement of σ with $\widetilde{\sigma}$ and introducing a correlation coefficient $\|\Sigma_u^+\omega\|_2$ in (4).

3.3. Result 2: benign condition for consistency

In this section, we further investigate the upper bound in Theorem 1 and derive sufficient conditions for the upper bound to converge to zero. We also provide several examples of distributions satisfying the condition.

We first provide a basic condition that is widely used for over-parameterized models (e.g., Bartlett *et al.* (2020)).

Definition 2 (Basic condition)**.** This condition requires that the value of the following three limits be zero:

$$\lim_{n\to\infty} \frac{\mathrm{rank}(\Sigma_u)}{n} = \lim_{n\to\infty} \frac{n}{R(\Xi_z)} = \lim_{n\to\infty} \|\theta_0\|_2 \sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}} = 0. \tag{6}$$

Their details are as follows:

(i) (Small latent noise) The first term, $\mathrm{rank}(\Sigma_u)/n$, describes the size of the latent noise vector relative to n, and the condition requires that the latent noise is small.

(ii) (Large effective dimension) The second term, $n/R(\Xi_z)$, decreases as the effective rank $R(\Xi_z)$ is larger than n, which plays the role of an effective dimension in the over-parameterized model.

(iii) (No aliasing) The third term, $\|\theta_0\|_2\sqrt{\mathrm{tr}(\Xi_z)/n}$, represents the magnitude of the error in a noiseless situation and intuitively plays a role similar to bias.

These assumptions are commonly used in the over-parameterized linear regression problem without endogeneity (Bartlett *et al.*, 2020; Koehler *et al.*, 2021; Tsigler and Bartlett, 2023). We will provide examples of covariance matrices that satisfy these assumptions in Section 3.3.1.

We derive a result where the projected RMSE converges to zero. We achieve this result by introducing new assumptions corresponding to the endogeneity in addition to the basic assumptions in Definition 2.

Theorem 2 (Sufficient conditions)**.** *Under Assumptions 1 and 2 with* $\Sigma_x = \Xi_z + \Sigma_u$*, let* $\widehat{\theta}$ *be the ridgeless estimator. Suppose that the basic condition in Definition 2 holds, and the following condition is also satisfied:*

$$\lim_{n\to\infty} \|\Sigma_u^+\omega\|_2 \sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}} = 0. \tag{7}$$

Then, the following holds:

$$\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 \xrightarrow{\mathbf{P}} 0, \ (n \to \infty).$$

This result states that condition (7) is a key factor of the convergence of the projected RMSE to zero in the setting with endogeneity because the basic assumption in Definition 2 is also needed in ordinary regression without endogeneity. Intuitively, condition (7) means that the replacement of σ^2 with $\widetilde{\sigma}^2$ in Theorem 1 is asymptotically negligible. For condition (7), the structure of ω plays an essential role because it is challenging to satisfy (7) with only the property of Σ_u^+. However, we have $\|(\Sigma_u^+)^{1/2}\omega\|_2 \leq \sigma^2$ (Lemma 38), which implies a slow increase of $\|\Sigma_u^+\omega\|_2$. Another implication is about the first term in (6): a strong correlation between X_i and Z_i is necessary for benign overfitting. This is suggested by the fact that $\mathrm{rank}(\Sigma_u) \geq p - \min\{\mathrm{rank}(\Sigma_z), \mathrm{rank}(\Pi_0)\}$ (see Proposition 36).

Remark 4 (Relation to weakness of instrumental variables)**.** Here, we discuss the relation of our results to the study of *weak instrumental variables*. It is known that having instrumental variables with weak correlations reduces the efficiency of estimation (Stock *et al.*, 2002). In our theory, from the result in Theorem 2, one can also claim that the weak instrumental variables reduce the validity of the estimation in the over-parameterized setting. Specifically, the instrumental variable Z_i, with weak correlations, exerts a diminishing effect on the rank of Π_0, thereby exerting influence on the initial two conditions delineated in Definition 2. Firstly, the effective rank $R(\Xi_z)$ will be small in comparison to the sample size, owing to the definition of Ξ_z and its effective rank. Furthermore, the assertion made in Lemma 37 solidifies that under Assumption 2, the equation $\mathrm{rank}(\Sigma_x) = \mathrm{rank}(\Sigma_u) + \mathrm{rank}(\Xi_z)$ holds, consequently resulting in a substantial elevation of the rank of Σ_u, thereby contravening the primary condition outlined in Definition 2. Hence, our result in the over-parameterized setting implies almost the same claim on the weak instrumental variables, while our approach is different from the previous studies.

Remark 5 (Comparison to independent noise case)**.** We compare our results to the results for the independent noise case. Although our results do not constitute a generalization of this case, there are similarities between these cases. One can set $\Pi_0 = I_p$, $Z_i = X_i$ and $u_i = 0$ to establish a scenario where X_i and Z_i are perfectly correlated, which corresponds to the independent noise case. Then, Koehler *et al.* (2021) developed the basic condition for this case:

$$\lim_{n\to\infty} \frac{\mathrm{rank}(\Sigma_1)}{n} = \lim_{n\to\infty} \frac{n}{R(\Sigma_2)} = \lim_{n\to\infty} \|\theta_0\|_2 \sqrt{\frac{\mathrm{tr}(\Sigma_2)}{n}} = 0, \tag{8}$$

by dividing $\mathbf{X}$ into two parts and obtaining $\Sigma_x = \Xi_z = \Sigma_1 + \Sigma_2$ using the orthogonality and the further extension of CGMT. This shows that our basic condition (6) has a form similar to that of the independent noise case.

Remark 6 (Relation to many instrumental variables)**.** The situation where there are many instrumental variables makes it difficult to satisfy our conditions. As the number of instrumental variables increases, it is anticipated that some of them have weak correlations, which raises the problem of weak instrumental variables. This point is also discussed in Remark 4.

Remark 7 (Necessary condition)**.** We discuss a necessary condition for the benign overfitting. When the noise ξ_i is independent of X_i, there is a necessary condition (or rather a necessary and sufficient condition) for the benign overfitting that the eigenvalue decay of Σ_x has a specific rate, which is shown in Theorem 6 in Bartlett *et al.* (2020). In contrast, when the noise is dependent as in our setting, no necessary condition is clarified. This is because the correlation coefficient ω increases the flexibility of the estimation error, and thus the eigenvalues of Σ_x alone cannot describe the necessary condition.

3.3.1. Examples

In this section, we provide examples that satisfy the condition in Theorem 2. The example here uses a matrix derived by Bartlett *et al.* (2020) as a base matrix $\overline{\Sigma}$, then constructs a latent noise covariance matrix Σ_u and of the instrumental variable Ξ_z based on the base matrix $\overline{\Sigma}$. Throughout this section, we assume that $\|\theta_0\|_2 = o(\sqrt{n})$.

Example 1. *Consider the dimension $p \in \mathbb{N} \cup \{\infty\}$ and a base matrix $\overline{\Sigma}$ whose i-th largest eigenvalue has the form*

$$\lambda_i = Ci^{-1} \log^{-\beta}(i+1),\ \ i = 1, ..., p,$$

with some constant $C > 0$ and $\beta > 1$, and also assume condition (7) holds. We further define a truncated version of $\overline{\Sigma}$ with a truncation level $k \leq p$ as $\overline{\Sigma}_{1:k} = U^\top \mathrm{diag}(\lambda_1, ..., \lambda_k, 0, ..., 0)U$, where $U \in \mathbb{R}^{p\times p}$ is an orthogonal matrix generated from a singular value decomposition $\overline{\Sigma} = U^\top \mathrm{diag}(\lambda_1, ..., \lambda_p)U$. Using the notion, we define our truncation level k_n^* as

$$k_n^* := \min\{k \geq 0 : r(\overline{\Sigma} - \overline{\Sigma}_{1:k}) > n\}, \tag{9}$$

which balances the complexities of the latent noise and the instrumental variable. Then, we define the (transformed) covariance matrices of u and z as

$$\Sigma_u = \overline{\Sigma}_{1:k_n^*}, \quad \Xi_z = \overline{\Sigma} - \overline{\Sigma}_{1:k_n^*}. \tag{10}$$

The example is adapted to our setting with endogeneity by considering the example of a covariance matrix by Bartlett *et al.* (2020). Rigorously, we set the covariance matrix by Bartlett *et al.* (2020) as the base matrix $\overline{\Sigma}$ and decompose it under the appropriate cutoff level k_n^* to the (transformed) covariance matrices. Importantly, this example can freely choose the dimension p (even infinite is possible). The following proposition shows that this example yields benign overfitting.

Proposition 3. *Consider Example 1. Assume $\|\theta_0\|_2 = o(\sqrt{n})$. If $\overline{\Sigma}$ and ω satisfy*

$$\lambda_i = Ci^{-1} \log^{-\beta}(i+1), \quad (U\omega)_i = \Theta(i^{-1} \log^{-\beta}(i+1)),$$

where $\beta > 1$ and $C > 0$, then Σ_u and Ξ_z defined in (10) and associated ω as $\|\Sigma_u^+ \omega\|_2 = o(\sqrt{n})$ satisfy all the conditions in Definition 2 and Theorem 2.

Example 2. *We consider the dimension $p = p_n$, which increases faster than n, that is, $\forall c > 0, \exists \bar{n} \in \mathbb{N}, \forall n \geq \bar{n}, p \geq cn$ holds. Furthermore, consider a base matrix $\overline{\Sigma}$ whose i-th largest eigenvalue has the form*

$$\lambda_i = \gamma_i + \varepsilon_n,\ \ i = 1, ..., p,$$

where $\{\gamma_i\}_i$ and $\{\varepsilon_n\}_n$ are sequences such that

$$\gamma_i = \Theta(\exp(-i/\tau)), \quad ne^{-o(n)} = \varepsilon_n p = o(n),$$

with some $\tau > 0$. We further assume condition (7) holds. Similar to Example 1, we use the truncation level k_n^* as (9) and define the (transformed) covariance matrices of u_i and Z_i as

$$\Sigma_u = \overline{\Sigma}_{1:k_n^*}, \quad \Xi_z = \overline{\Sigma} - \overline{\Sigma}_{1:k_n^*}. \tag{11}$$

In the example, we consider the case where p diverges faster than n. In this case, the eigenvalues consist of two terms: an exponentially decaying term, and a term that behaves like noise. The next proposition shows benign overfitting in this setting.

Proposition 4. *Consider Example 2. Set eigenvalues of $\overline{\Sigma}$ as follows:*

$$\lambda_i = \gamma_i + \varepsilon_n,$$

where $\gamma_i = \Theta(\exp(-i/\tau))$ and $\tau > 0$. Assume $\|\theta_0\|_2 = o(\sqrt{n})$. If p and ω satisfy

$$p = \upsilon(n), \quad ne^{-o(n)} = \varepsilon_n p = o(n), \quad (U\omega)_i = \Theta(\exp(-i/\tau)),$$

then Σ_u and Ξ_z defined in (11) and associated ω as $\|\Sigma_u^+\omega\|_2 = o(\sqrt{n})$ satisfy all the conditions in Definition 2 and Theorem 2.

4. Error analysis: non-orthogonal case

In this section, we relax the orthogonality condition of Assumption 2 and study the sufficient conditions for benign overfitting when the covariance matrices Σ_u and Ξ_z are not orthogonal. The approach to derive the conditions is almost the same as in Section 3; we first derive an upper bound for the projected RMSE, then use it to reveal sufficient conditions. To simplify the presentation, we defer the upper bounds to a later section and present only a theorem on the sufficient conditions.

Theorem 5 (Sufficient conditions: Non-Orthogonal Case)**.** *Under Assumption 1, let $\widehat{\theta}$ be the ridgeless estimator. Further, assume $\widetilde{\sigma}^2 := \sigma^2 - \|\omega\|_{\Sigma_u^+}^2 > 0$. Suppose that the basic condition in Definition 2 holds, and the covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$ satisfies the following conditions:*

$$\lim_{n\to\infty} \frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}} = \lim_{n\to\infty} \frac{n}{R(\Xi_z)}\frac{\operatorname{tr}(\Sigma_u\Xi_z)}{\operatorname{tr}(\Xi_z^2)} = \lim_{n\to\infty} \omega^\top\Sigma_u^+\Xi_z\Sigma_u^+\omega = 0. \tag{12}$$

Then, the following holds:

$$\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 \xrightarrow{\mathbf{P}} 0, \ (n \to \infty).$$

In this non-orthogonal case, the above three conditions (12) play a critical role, in addition to Definition 2. We provide explanations of the terms in (12) one by one below.

(i) (Non-degenerated noise) The first condition on $(\|\Sigma_u^+\omega\|_2/\widetilde{\sigma})(\sqrt{\mathrm{tr}(\Xi_z)/n})$ requires that the variance $\widetilde{\sigma}^2$ be non-degenerated and condition (7) hold. Therefore, this condition is a sufficient condition for the condition (7).

(ii) (Effective rank with non-orthogonality) The second condition on the term $(n/R(\Xi_z))(\mathrm{tr}(\Sigma_u\Xi_z)/\mathrm{tr}(\Xi_z^2))$ takes into account the effect of non-orthogonality on the effective rank $R(\Xi_z)$, which already appears in the basic condition in Definition 2. This means that non-orthogonality term $(\mathrm{tr}(\Sigma_u\Xi_z)/\mathrm{tr}(\Xi_z^2))$ has a role in reducing the effective rank $R(\Xi_z)$.

(iii) (Mixed effect) The third condition on $\omega^\top\Sigma_u^+\Xi_z\Sigma_u^+\omega$ includes both the effects of the non-orthogonality and the correlation ω. This condition is asymptotically satisfied as Σ_u and Ξ_z gradually approach orthogonality.

Of the conditions in (12), (i) and (iii) are necessary to handle the endogeneity. In other words, (i) and (iii) are always satisfied when $\omega = 0$ holds. However, condition (ii) is required to achieve benign overfitting under non-orthogonality even in the absence of endogeneity. To make it clear, we reveal a sufficient condition for benign overfitting with non-orthogonality in the setting of ordinary linear regression without endogeneity ($\omega = 0$).

Theorem 6. *(Sufficient conditions: Non-Orthogonal Case when X_i and ξ_i are independent) Under Assumption 1, let $\widehat{\theta}$ be the ridgeless estimator. Suppose that $\omega = 0$ holds. Suppose that the basic condition in Definition 2 holds, and there exists a sequence of covariance $\Sigma_x = \Sigma_1 + \Sigma_2$ such that the following conditions hold:*

$$\lim_{n\to\infty} \frac{n}{R(\Sigma_2)}\left(\frac{\mathrm{tr}(\Sigma_1\Sigma_2)}{\mathrm{tr}(\Sigma_2^2)}\right) = 0. \tag{13}$$

Then, $L(\widehat{\theta})$ converges to σ^2 in probability where $L(\theta) = \mathbb{E}(y - \langle\theta, x\rangle)^2$.

Theorem 6 states that condition (13) is a key factor in RMSE converging to zero in the setting without orthogonality. When we set $\Sigma_1 = \Sigma_u$ and $\Sigma_2 = \Xi_z$, condition (10) is exactly equal to condition (ii) in the above discussion. Intuitively, $\mathrm{tr}(\Sigma_u\Xi_z)$ is the degree of non-orthogonality between Σ_u and Ξ_z, and Theorem 6 requires the degree to be small.

4.1. Example

We provide an example, similar to those provided in Section 3.3.1. That is, we first specify the base matrix $\overline{\Sigma}$, then construct (transformed) covariance matrices based on it. Note that the definition of the dimension and the way of decomposition are slightly different. Throughout this section, we also assume that $\|\theta_0\|_2 = o(\sqrt{n})$.

Example 3 (Non-orthogonal version of Example 1)**.** *Consider the dimension $p = qn$ with some $q > 1$, and a base matrix $\overline{\Sigma}$ whose i-th largest eigenvalue has the form*

$$\lambda_i = Ci^{-1}\log^{-\beta}(i+1),\ \ i = 1, ..., p,$$

with some constant $C > 0$ and $\beta > 0$. We also assume that $\|\Sigma_u^+ \omega\|_2 = o(\sqrt{n})$, $\lim_{n\to\infty}(\sigma^2 - \|\omega\|^2_{\Sigma_u^+}) > 0$, and consider the truncation level as (9). Then, we define the (transformed) covariance matrices with $\alpha > 1$:

$$\Sigma_u := \left(1 - \frac{1}{n^\alpha}\right)\overline{\Sigma}_{1:k_n^*}, \quad \Xi_z := \overline{\Sigma} - \left(1 - \frac{1}{n^\alpha}\right)\overline{\Sigma}_{1:k_n^*}.$$

The base matrix $\overline{\Sigma}$ used in this example is identical to that in Example 1. In contrast, the decomposition to construct the (transformed) covariance matrices is different. The following result demonstrates the validity of this example.

Proposition 7. *Consider Example 3. Suppose $\lim_{n\to\infty}(\sigma^2 - \|\omega\|^2_{\Sigma_u^+}) > 0$ does hold. Under the assumptions $\|\theta_0\|_2 = o(\sqrt{n})$ and $(U\omega)_i = \Theta(i^{-1}\log^{-\beta}(i+1))$ as in Proposition 3, Σ_u, Ξ_z, and ω defined above satisfy all the conditions in Definition 2 and Theorem 5.*

Note that Theorem 6 immediately holds from this proposition by setting $\Sigma_1 = \Sigma_u$ and $\Sigma_2 = \Xi_z$.

5. Extension to general norm

We extend the result of Theorem 2 to the case when θ is measured in terms of a general norm. Let $\|\cdot\|$ be an arbitrary norm. To achieve our aim, we introduce two definitions, the dual norm and effective $\|\cdot\|$-ranks.

Definition 3 (Dual Norm)**.** The dual norm of norm $\|\cdot\|$ on $\mathbb{R}^d$ is $\|u\|_* := \max_{\|v\|=1}\langle v, u\rangle$, and the set of all its sub-gradients with respect to u is $\partial\|u\|_* = \{v : \|v\| = 1, \langle v, u\rangle = \|u\|_*\}$.

Definition 4 (Effective $\|\cdot\|$-rank)**.** The effective $\|\cdot\|$-ranks of a covariance matrix Σ are listed as follows. Let H be normally distributed with mean zero and variance I_d, that is, $H \sim N(0, I_d)$. Denote v^* as $\arg\min_{v\in\partial\|\Sigma^{1/2}H\|_*}\|v\|_\Sigma$. Then, we define

$$r_{\|\cdot\|}(\Sigma) := \left(\frac{E\|\Sigma^{1/2}H\|_*}{\sup_{\|u\|\le 1}\|u\|_\Sigma}\right)^2 \text{ and } R_{\|\cdot\|}(\Sigma) := \left(\frac{E\|\Sigma^{1/2}H\|_*}{E\|v^*\|_\Sigma}\right)^2.$$

Effective $\|\cdot\|$-ranks is a generalization of the effective rank in Definition 2, and the dual norm is necessary to define the general effective rank.

We provide basic conditions for general norm $\|\cdot\|$, which corresponds to Definition 2 and advanced conditions Koehler *et al.* (2021) established.

Definition 5 (Basic condition with general norm)**.** This condition requires that the value of the three limits be zero with respect to a general norm $\|\cdot\|$:

$$\lim_{n\to\infty}\frac{\operatorname{rank}(\Sigma_u)}{n} = \lim_{n\to\infty}\frac{n}{R_{\|\cdot\|}(\Xi_z)} = \lim_{n\to\infty}\frac{\|\theta_0\|\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}} = 0. \tag{14}$$

Each condition in (14) corresponds to conditions in (6). The first condition for small latent noise remains unchanged. For the large effective dimension condition, we replace $R(\Xi_z)$ with the general norm counterpart, $R_{\|\cdot\|}(\Xi_z)$. For the no aliasing condition, θ_0 is measured in terms of any norm $\|\cdot\|$ and $\sqrt{\mathrm{tr}(\Xi_z)}$ is replaced with $\mathbb{E}\|\Xi_z^{1/2}H\|_*$.

Definition 6 (Advanced condition)**.** In addition to Definition 5, we require the following two conditions:

$$\lim_{n\to\infty}\frac{1}{r_{\|\cdot\|}(\Xi_z)}=\lim_{n\to\infty}\mathbb{P}(\|P_u v^*\|^2>1+\eta)=0, \tag{15}$$

for any $\eta>0$.

We provide details of the terms in (15) below.

(i) (Large effective dimension) The first term $1/r_{\|\cdot\|}(\Xi_z)$ decreases as the effective rank $r_{\|\cdot\|}(\Xi_z)$ becomes large as with the second condition in (14). In the Euclidean norm case, $1/r(\Xi_z)$ converges to zero as $n/R(\Xi_z)$ goes toward zero by definition.

(ii) (Contracting ℓ_2 projection condition) This condition implies the projected v^* onto the space spanned by Σ_u is asymptotically smaller than or equal to 1. This condition always holds in the Euclidean norm case because $\|P_u v^*\|_2^2\le\|v^*\|_2^2=1$ holds.

For the projected RMSE to converge to zero, we introduce a new assumption corresponding to condition (7) in Theorem 2 in addition to the conditions in Definitions 5 and 6.

Theorem 8 (Sufficient conditions)**.** *Under Assumptions 1 and 2, let $\widehat{\theta}$ be the ridgeless estimator. Let $\|\cdot\|$ denote an arbitrary norm. Suppose that the basic conditions in Definitions 5 and 6 hold, and the covariance splitting $\Sigma_x=\Xi_z+\Sigma_u$ satisfies the following conditions:*

$$\lim_{n\to\infty}\frac{\|\Sigma_u^+\omega\|\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}}=0.$$

Then, the following holds:

$$\|\widehat{\theta}-\theta_0\|_{\Xi_z}^2\xrightarrow{\mathbf{P}}0,\ \ (n\to\infty).$$

As in the condition in (14), $\Sigma_u^+\omega$ is measured in terms of any norm $\|\cdot\|$ and $\sqrt{\mathrm{tr}(\Xi_z)}$ is replaced with $\mathbb{E}\|\Xi_z^{1/2}H\|_*$. If we consider the general norm, compared to the Euclidean norm case, it is possible we can relax some of the sufficient conditions for benign overfitting, especially the condition in Theorem 2. However, as we must incorporate additional advanced conditions outlined in Definition 6 in conjunction with the basic conditions presented in Definition 5, it remains uncertain whether benign overfitting is more probable.

6. Proof outline

6.1. Approach with CGMT

Our proof relies on two techniques: (i) describing the ridgeless estimator as a solution to an optimization problem and bounding the projected RMSE, and (ii) evaluating the solution by an extended version of the convex Gaussian minimax theorem (CGMT). CGMT was introduced into high-dimensional statistics by Thrampoulidis *et al.* (2015, 2018). Furthermore, Koehler *et al.* (2021) discussed that CGMT can describe benign overfitting by Bartlett *et al.* (2020) in the ordinary regression setting. In this section, we deal with the non-orthogonal case results given in Section 4, which can be easily applied to the orthogonal case in Section 3.

We prepare some notations. We define a normalized correlation coefficient $\rho = (\Sigma_u^{1/2})^+\omega$, which guarantees that $\|\rho\|_2^2 \le \sigma^2$ (see Lemma 38). We also define $\mathbf{X} = (X_1, ..., X_n)^\top$ as an $\mathbb{R}^{n\times p}$-valued random matrix, which has the form

$$\mathbf{X} \overset{\mathcal{D}}{=} \mathbf{W}_1 \Xi_z^{1/2} + \mathbf{W}_2 \Sigma_u^{1/2}, \tag{16}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are $n \times p$ random matrices whose i-th row identically follows a joint distribution of ξ_i such that

$$\begin{pmatrix} W_{1,i} \\ W_{2,i} \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0}_{p\times 1} \\ \mathbf{0}_{p\times 1} \\ 0 \end{pmatrix}, \begin{pmatrix} I_{p\times p} & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times 1} \\ \mathbf{0}_{p\times p} & I_{p\times p} & \rho \\ \mathbf{0}_{p\times 1}^\top & \rho^\top & \sigma^2 \end{pmatrix} \right) \tag{17}$$

for $i = 1, ..., n$. Note that this form follows the Gaussianity from Assumption 1.

6.2. Step (i): bound projected RMSE by optimization form

First, we consider a uniform upper bound for the projected RMSE $E[(E[\langle\theta, X\rangle - \langle\theta_0, X\rangle|Z])^2]$ under the constraint that the estimator $\widehat{\theta}$ is the ridgeless estimator (i.e., $\widehat{L}(\widehat{\theta}) = 0$). Then, we transform it to a maximization problem with a constraint with some compact parameter space $\mathcal{K} \subset \mathbb{R}^p$:

$$\begin{aligned} \max_{\theta\in\mathcal{K}, \widehat{L}(\theta)=0} E\left[(E[\langle\theta, X\rangle - \langle\theta_0, X\rangle|Z])^2 \right] &= \max_{\theta\in\mathcal{K}, \mathbf{X}\theta=\mathbf{Y}} \|\theta - \theta_0\|_{\Xi_z}^2 \\ &= \max_{\theta\in\mathcal{K}, \mathbf{X}(\theta-\theta_0)=\xi} \|\theta - \theta_0\|_{\Xi_z}^2 . \end{aligned}$$

Using the surrogate Gaussians in (16), the upper bound above has the same distribution as the following term:

$$\Phi := \max_{\substack{(\theta_1,\theta_2)\in S, \\ \mathbf{W}_1\theta_1 + \mathbf{W}_2\theta_2 = \xi}} \|\theta_1\|_2^2, \tag{18}$$

where we define $S := \{(\theta_1, \theta_2) : \exists\theta \in \mathcal{K}\ s.t.\ \theta_1 = \Xi_z^{1/2}(\theta - \theta_0)\ and\ \theta_2 = \Sigma_u^{1/2}(\theta - \theta_0))\}$. The details of the derivation are described in the proof of Lemma 12 in the appendix.

Second, we approximate the distribution of the optimization problem (18) using CGMT. CGMT approximates minimax optimization problems by a distribution of their simpler auxiliary problems. Here, we present our variant of CGMT that can deal with correlation between variables, though we also use classical CGMT depending on the situation.

Theorem 9 (Extended CGMT)**.** *Let $\mathbf{W} : n \times d$ be a matrix with i.i.d. $N(0,1)$ entries and suppose $G \sim N(0, I_n)$ and $H \sim N(0, I_d)$ are independent of $\mathbf{W}$ and each other. Let S_W and S_U be non-empty compact sets in $\mathbb{R}^d \times \mathbb{R}^{d'}$ and $\mathbb{R}^n \times \mathbb{R}^{n'}$, respectively, and let $\psi : S_W \times S_U \mapsto \mathbb{R}$ be an arbitrary continuous function. Define the Primary Optimization (PO) problem*

$$\Phi(\mathbf{W}) := \min_{(\omega,\omega')\in S_W} \max_{(u,u')\in S_U} \langle u, \mathbf{W}\omega\rangle + \psi((\omega,\omega'),(u,u')) \tag{19}$$

and the Auxiliary Optimization (AO) problem

$$\phi(G,H) := \min_{(\omega,\omega')\in S_W} \max_{(u,u')\in S_U} \|\omega\|_2\langle G, u\rangle + \|u\|_2\langle H,\omega\rangle + \psi((\omega,\omega'),(u,u')). \tag{20}$$

If we suppose that S_W and S_u are convex sets and $\psi((\omega,\omega'),(u,u'))$ is convex in (ω,ω') and concave in (u,u'), then $\mathbb{P}(\Phi(\mathbf{W}) > c) \leq 2\mathbb{P}(\phi(G,H) \geq c)$ for any $c \in \mathbb{R}$.

This theorem is an extension of the original CGMT to split the variables to be optimized so that it can handle our regression model (1) with the endogeneity. Rigorously, this theorem allows correlation between the covariates and the error terms.

Using the extended CGMT in Theorem 9, we approximate the distribution of the problem (18) by

$$\phi := \max_{(\theta_1,\theta_2)\in S:\|\xi-\mathbf{W}_2\theta_2-G\|\theta_1\|_2\|_2\leq\langle\theta_1,H\rangle} \|\theta_1\|_2^2,$$

where $G \sim N(0, I_n)$ and $H \sim N(0, I_d)$ are Gaussian vectors independent of $\mathbf{W}_1, \mathbf{W}_2, \xi$, and each other. A distribution of this term is tractable because of the relatively simple form. Namely, we obtain the following result. In the case of a Euclidean norm ball, we set $\mathcal{K} := \{\theta \in \mathbb{R}^p | \|\theta\|_2 \leq B\}$. By combining the upper bound of $\mathcal{K}$, we can derive a simpler upper bound for the Euclidean norm.

Corollary 10. *There exists an absolute constant $C_1 \leq 64$ such that the following is true. Assume Assumptions 1 and 2 hold. Pick $\Sigma_x = \Xi_z + \Sigma_u$ and fix $\delta \leq 1/4$. Define $\widetilde{\sigma}^2 := \sigma^2 - \|\omega\|^2_{\Sigma_u^+} \geq 0$ and $g(t_1,t_2) = t_1^2 - t_2^2$. If $B \geq \|\theta_0\|_2$ and n is large enough that $\gamma(\delta) \leq 1$, the following holds with probability at least $1-\delta$:*

$$\max_{\|\theta\|_2\leq B, \mathbf{Y}=\mathbf{X}\theta} \|\theta-\theta_0\|^2_{\Xi_z} \lesssim (1+\gamma(\delta))g\left(B\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}}, \widetilde{\sigma}\right),$$

where we define

$$\gamma(\delta) := \sqrt{\log(1/\delta)} \left(\frac{1}{\sqrt{r(\Xi_z)}} + \sqrt{\frac{\mathrm{rank}(\Sigma_u)}{n}} \right).$$

In this corollary, the radius B of $\mathcal{K}$ plays an important role. That is, the bound in Corollary 10 is valid only when the norm $\|\theta\|_2$ is no more than B. Here, our remaining task is to show that such a B exists. In the next step, we will examine the norm $\|\widehat{\theta}\|_2$ to show the existence of such B.

6.3. Step (ii): bound norm of estimator

As the next step, we specify an upper bound on the norm of the solution, which is equivalent to deriving an upper bound of B that appears in the constraint in Corollary 10. To show the consistency of ridgeless estimators, we need to specify the value of B so that $\mathcal{K}$ includes some parameters.

In the following theorem, we obtain the Euclidean norm bound for the ridgeless estimator. To achieve this result, we again use CGMT from Theorem 9.

Theorem 11 (Euclidean norm bound; special case of Theorem 31)**.** *Fix any $\delta \leq 1/4$. Suppose $\Sigma_x = \Xi_z + \Sigma_u$ and $\widetilde{\sigma}^2 := \sigma^2 - \|\omega\|^2_{\Sigma_u^+} > 0$. If n and the effective ranks are such that $\varepsilon(\delta) \leq 1$ and $R(\Xi_z) \gtrsim \log(1/\delta)^2$, then with probability at least $1-\delta$, it holds that*

$$\|\widehat{\theta}\|_2 \lesssim (1+\varepsilon(\delta))^{1/2} \left(\|\theta_0\|_2 + \|\Sigma_u^+ \omega\|_2 + (2\eta_1 + \widetilde{\sigma} + \eta_2) \sqrt{\frac{n}{\mathrm{tr}(\Xi_z)}} \right),$$

where $\eta_1, \eta_2, \varepsilon \in \mathbb{R}$ are sequences depending on n and δ satisfying

$$\eta_1 \lesssim \sqrt{\frac{n}{R(\Xi_z)}} \|\Xi_z^{1/2} \Sigma_u^+ \omega\|_2,$$

$$\eta_2 \lesssim \sqrt{\left(1 + \sqrt{\frac{2\log(8/\delta)}{r(\Xi_z)}}\right)} \sqrt{\frac{(\mathbb{E}\|\Xi_z^{1/2} H\|_2)^2}{n} \|\Sigma_u^+ \omega\|_2^2 + \|\Xi_z^{1/2} \Sigma_u^+ \omega\|_2^2},$$

$$\varepsilon := \sqrt{\log(1/\delta)} \left(\sqrt{\frac{\mathrm{rank}(\Sigma_u)}{n}} + \left(1 + \frac{\mathrm{tr}(\Sigma_u \Xi_z)}{\mathrm{tr}(\Xi_z^2)}\right) \left(\frac{n}{R(\Xi_z)}\right) + \frac{\|\Sigma_u^+ \omega\|_2}{\widetilde{\sigma}} \sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}} \right).$$

The rigorous definitions of η_1, η_2, and ε will be provided in the appendix. To derive this upper bound, we again use the common uniform upper bound argument. Combining this result with Corollary 10, we derive our primary result on the upper bound of $\|\theta - \theta_0\|^2_{\Xi_z}$.

7. Experiment

We conduct experiments to justify our theoretical results. Specifically, we test whether our derived sufficient conditions in Theorems 2 and 5 lead to benign overfitting. This section contains two experiments: (i) measuring the projected RMSE of the ridgeless estimator, and (ii) comparing the ridgeless estimator to existing high-dimensional operating variable methods.

7.1. Projected RMSE of ridgeless estimator

7.1.1. Setups

We generate $n \in \{200, 300, ..., 1000\}$ independent samples $(X_1, Y_1, Z_1), ..., (X_n, Y_n, Z_n)$ from the regression model (1), and the covariate X_i, noise variable ξ_i, and latent noise u_i follow the distribution

$$\begin{pmatrix} X_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0}_{p\times 1} \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_x & \Sigma_u^{1/2}\rho \\ (\Sigma_u^{1/2}\rho)^\top & \sigma^2 \end{pmatrix} \right), \quad u_i \sim N(\mathbf{0}_{p\times 1}, \Sigma_u). \tag{21}$$

The covariance/coefficient matrices $\Sigma_x, \Sigma_u, \Pi_0$, and Σ_z are determined separately for the following four setups. Through experiments, truncation level k_n^* is determined in the same way as (9).

Setup (i) **Example 1 (Orthogonal Case)**: This setting follows Example 1 with the orthogonal case in Section 3. We set the parameter dimension as $p = 5n$, and set a base matrix $\overline{\Sigma} \in \mathbb{R}^{p\times p}$ such that its i-th largest eigenvalue λ_i is $300i^{-1}(\log(i+1)\exp/2)^{-2}$ for $i = 1, ..., p$. We set the true parameter $\theta_0 \in \mathbb{R}^p$ whose i-th element is $20/\sqrt{i}$ and set ω as $(\Sigma_u^{1/2})^+\omega := U\rho$, where $\rho \in \mathbb{R}^p$ has its i-th element $2/i$ and $U \in \mathbb{R}^{p\times p}$ is an orthogonalized version of $P \in \mathbb{R}^{p\times p}$ such that $P_{j,j'} = \mathbf{1}\{|j - j'| \neq p - 2\}$ for $j, j' = 1, ..., p$. Then, we define Σ_x, Σ_u, and Ξ_z as $\Sigma_x := \Sigma_u + \Xi_z, \Sigma_u := \overline{\Sigma}_{1:k_n^*}$, and $\Xi_z := (\overline{\Sigma} - \overline{\Sigma}_{1:k_n^*})$ as in Example 1. This setting satisfies the sufficient conditions in Theorem 2, and also Ξ_z and Σ_u are orthogonal.

Setup (ii) **Example 2 (Orthogonal Case)**: This setting follows Example 2 with the orthogonal case in Section 3. We set the dimension $p = n^{3/2}$ and a base matrix $\overline{\Sigma}$ as its i-th eigenvalue λ_i being $\lambda_i = \gamma_i + \varepsilon_n$, where $\gamma_i = 10\exp(-(i/2))$ and $\varepsilon_n = \exp\left(-\sqrt{n}\right)/\sqrt{n}$. We also set the true parameter $\theta_0 \in \mathbb{R}^p$ whose i-th element is $20/\sqrt{i}$, and the correlation coefficient ω is defined to satisfy $(\Sigma_u^{1/2})^+\omega := U\rho$ where $\rho \in \mathbb{R}^p$ has $3\exp(-i/4)$ as its i-th element. This setting satisfies the sufficient conditions in Theorem 2, and Ξ_z and Σ_u are orthogonal.

Setup (iii) **Example 1 (Non-orthogonal Case)**: We consider an extension of Example 1 to the non-orthogonal case in Section 4. This is identical to that treated in Example 3. Specifically, $p, \overline{\Sigma}$, and ω are determined as in Setup (i) above. However, Σ_u and Ξ_z are the same as

$$\Sigma_u := \left(1 - \frac{1}{n^{1.01}}\right)\overline{\Sigma}_{1:k_n^*}, \text{ and}$$

$$\Xi_z := (\overline{\Sigma} - \overline{\Sigma}_{1:k_n^*}) + \frac{1}{n^{1.01}} \overline{\Sigma}_{1:k_n^*}, \tag{22}$$

and $\Sigma_x := \Sigma_u + \Xi_z$. In this setting, Ξ_z and Σ_u are non-orthogonal.

Setup (iv) **Example 2 (Non-orthogonal Case)**: We consider an extension of Example 2 to the non-orthogonal case in Section 4. In this setting, $p, \overline{\Sigma}$, and ω are determined as in Setup (ii) above, and Σ_u, Ξ_z, and Σ_x are set as (22). Here, Ξ_z and Σ_u are non-orthogonal.

Setup (v) **Example 1 (Sparse and Orthogonal Case)**: We consider Setup (i) under the sparse setting. The parameters are identical to Setup (i) except the setting of θ_0. When i is not more than 100 and there exists a natural number k such that $i + 4 = 5k$, set the i-th element of θ_0 as $20/\sqrt{i}$. Otherwise, the elements of θ_0 are equal to zero.

Setup (vi) **Example 1 (Sparse and Non-Orthogonal Case)**: We study Setup (iii) in the sparse setting. All the settings are identical to Setup (iii) except the setting of θ_0. We impose the sparsity on θ_0 as in Setup (v).

In addition, beyond our theoretical framework, we also examine the situation when the variable data are non-Gaussian. Specifically, we study the situation where the vector of instrumental variable Z_i follows the multivariate t-distribution with 5 degrees of freedom, and the mean and variance are common.

7.1.2. Results

Figure 1 summarizes the results of each of the setups. The values are means of 50 repetitions. The red line shows the projected RMSE of the ridgeless estimator. The blue lines show the case with the t-distribution.

These results carry several implications: (a) Despite the increase in dimension p being related to n, that is, $p = 5n$ or $p = n^{3/2}$, the projected RMSEs converge to zero. This implies that benign overfitting occurs with this high-dimensional case even with the endogeneity. (b) The convergence occurs even when Z_i is not generated by the Gaussian distribution, which implies that our theoretical results would be applicable to the non-Gaussian case.

7.2. Comparison with related method

7.2.1. Setups

We compare the ridgeless estimator to a regularized estimator for high-dimensions, such as the lasso-type method. Specifically, we consider methods for estimating sparse parameters under high-dimensional covariates and instrumental variables, such as those developed by Belloni *et al.* (2012); Chernozhukov *et al.* (2015a) and many others.

We present our setting. Similar to Section 7.1, we generate $n \in \{100, 200, ..., 1000\}$ observations $(X_1, Y_1, Z_1), ..., (X_n, Y_n, Z_n)$ from the regression model (1)

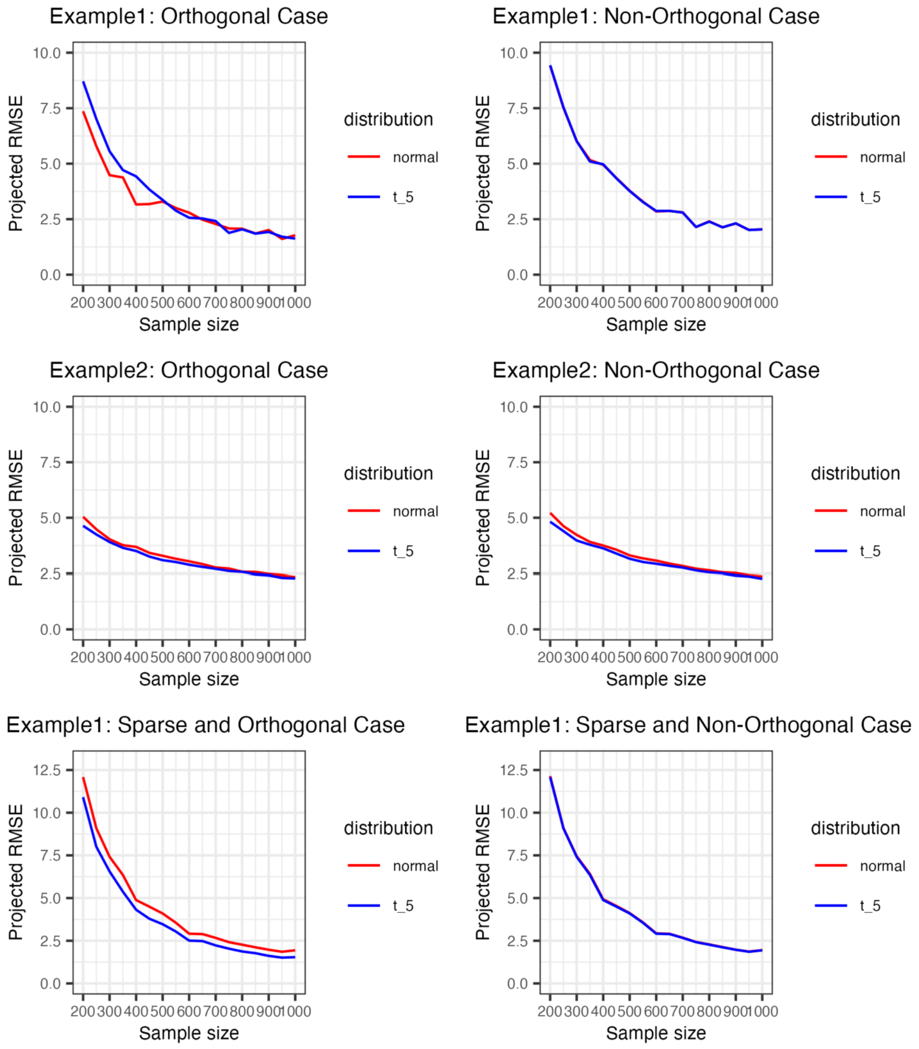

Fig 1. *Projected RMSE of the ridgeless estimator against the sample size n. The red line shows the Gaussian case, and the blue line shows the non-Gaussian case. The dimension p of the parameters is set as $p = 5n$ (Example 1) or $p = n^{3/2}$ (Example 2).*

and the data generating process (21). Here, k denotes a dimension of endogenous variables and we set $k = n/10$.

Setup (vii) **Non-Sparse Case**: We consider the case where the true parameter θ_0 is not sparse. We set $p = 5n$ and set the true parameter $\theta_0 \in \mathbb{R}^p$, which has $20/\sqrt{i}$ as its i-th element. Further, we set the base matrix $\overline{\Sigma} \in \mathbb{R}^{p\times p}$ that has $\lambda_i = 300i^{-1}(\log(i+1)\exp/2)^{-2}$ as its i-th largest eigenvalue. The correlation coefficient $\omega \in \mathbb{R}^p$ with its i-th element is $(2/i)\mathbb{1}\{i \in [1,k]\}$. With these settings, we define Σ_x, Σ_u, and Ξ_z as in (22). The matrices Σ_u and Ξ_z with the correlation term ω satisfy the sufficient conditions in Theorem 5.

Setup (viii) **Partially Sparse Case**: We consider the case where the true parameter θ_0 is less sparse. We set $p = 5n$ and define $\theta_0 \in \mathbb{R}^p$ whose i-th element is $(20/\sqrt{i})\mathbb{1}\{i \leq 0.8n\}$. We define Σ_x, Σ_u, Ξ_z, and ω in the same way as the non-sparse case (Setup (v)).

Setup (ix) **Non-Sparse Case (Rotated)**: We set $p = 5n$ and define Σ_x, Σ_u, Ξ_z, θ_0, and ω in the same way as the non-sparse case (Setup (v)). Further, we set the $(k/5+1)$-th to k-th variables and the (k_n^*+1)-th to $(k_n^* + k/5)$-th variables as endogenous.

For the method to be compared, we utilize the estimator by Chernozhukov *et al.* (2015a) named *LassoIV*. First, we divide the sample in half, then we use one-half of the sample to estimate the parameters of endogenous variables and use the other to estimate the other parameters. We use the R package *hdm* (Chernozhukov *et al.*, 2016) for implementation. For the estimation of exogenous variables, we subtract the endogenous part from the outcome and define the new outcome $\widetilde{Y}_i$, that is,

$$\widetilde{Y}_i := Y_i - \widehat{\beta}^\top W_i,$$

where W_i is a $k \times 1$ endogenous variable and $\widehat{\beta}$ is an estimator by LassoIV. To obtain an estimator for the parameters of exogenous variables, we regress $\widetilde{Y}_i$ on exogenous variables.

7.2.2. Results

Figure 2 summarizes the results of each experiment. We report means of 30 repetitions. When the sample size is small, the projected RMSE by the Lasso method is notably larger than that by the ridgeless estimator. As the sample size grows, though the errors get smaller, the error by the ridgeless estimator is still relatively small. Specifically, in setup (ix), we change the location of endogenous variables. Nevertheless, we can see that the ridgeless estimator provides the smaller projected RMSE.

7.3. Real data analysis

We implement real data analysis in this subsection to exemplify our theoretical result. We used the Current Population Survey (CPS), a monthly survey of U.S. households conducted by the Bureau of the Census of the Bureau of Labor Statistics. Our data consists of the March 2009 survey, including the Asian male individuals who were employed full-time (defined as those who had worked at least 36 hours per week for at least 48 weeks the past year), and excluded those in the military. The sample size is 1,435.

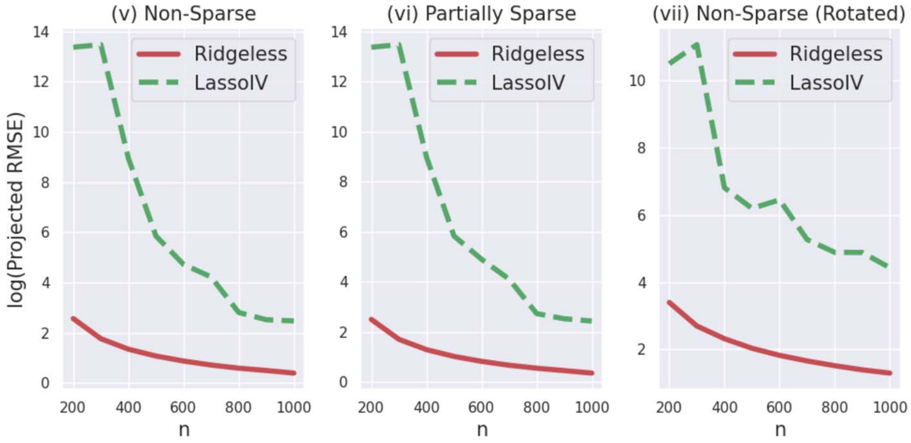

Fig 2. *Projected RMSEs of the ridgeless estimator and LassoIV. Each value is a mean of* 30 *repetitions.*

In this analysis, we set the natural log of hourly wage as the outcome variable Y_i. From the dataset, we use the year of education, the square of the year of education, age, the square of age, and the product of education and age as the covariates. Furthermore, to study a high-dimensional setting, we generate the 20,000-dimensional normal variables X_i^* with the diagonal variance matrix Σ whose ℓ-th diagonal $300\ell^{-1}(\log(\ell+1)\exp/2)^{-2}$ for $\ell = 1, ..., 20,000$. For each $i \in \{1, \cdots, 1435\}$, we have

$$Y_i{=}\beta_1 education_i{+}\beta_2 education_i^2{+}\beta_3 age_i{+}\beta_4 age_i^2{+}\beta_5 education_i age_i{+}\gamma^\top X_i^*{+}\xi_i. \tag{23}$$

As the error term ξ_i included the unobserved ability of an individual that will affect both the natural log of hourly wage and the year of education, the year of education will correlate with the error term ξ_i, that is, the year of education is endogenous.

Under the setting (23), we calculate the sample RMSE. We estimate the interpolator and evaluate the sample RMSE. The sample RMSE is 0.6165. As the estimated RMSE obtained from the LASSO estimator with 5-fold cross validation is 0.4463, this result implies the sample RMSE obtained by the interpolator will approximate RMSE even with the presence of the correlation between the covariates and the noise ξ_i.

8. Discussion and conclusion

We studied the estimation error in the over-parameterized linear regression problem when the covariates are endogenous. In particular, we examined the situation where data are Gaussian and the covariates have a linear model on an instrumental variable. In this setting, we derived sufficient conditions under which the risk of the ridgeless estimator converges to zero. In other words, we show the ridgeless estimator achieves benign overfitting even in the presence of endogeneity in this setting. To show this result, we developed an extended version of CGMT.

An important future challenge for the study of over-parameterization with endogeneity is the development of methods to infer whether our sufficient conditions hold from data. This challenge may be addressed, for example, by estimating a decay rate of eigenvalues of covariance matrices, as in the Hill estimator (Hill, 1975). The development of such practical methods is an important future task.

One limitation of this study depends on the Gaussianity of data. As this is an essential condition for using CGMT, it is not easy to relax. However, there has been some research to extending risks with Gaussian data to those of non-Gaussian data, known as universality (Han and Shen, 2023; Montanari and Saeed, 2022), so it may be a way to analyze non-Gaussian data.

Appendix A: Organization of Appendix

This appendix provides the full proofs of the results in the main body. The first half of the appendix follows the proof outline described in Section 6: (i) a proof of CGMT (Section B), (ii) a proof of an upper bound for the projected RMSE (Section C), and (iii) a proof of an upper bound for the ridgeless estimator (Section D). In Section E, we provide proofs for the primary statement for benign overfitting. In Section F, we independently present the proof for the non-orthogonal case in Section 4. Finally, supportive results are listed in Section G.

Appendix B: Proof of CGMT

We present a proof of Theorem 9 for CGMT. The proof of the standard CGMT is given in Thrampoulidis *et al.* (2015). We extend the standard proof to accommodate partitions of a parameter space. Remember that $\mathbf{W} : n \times d$ is a matrix with i.i.d. $N(0,1)$ entries and suppose $G \sim N(0, I_n)$ and $H \sim N(0, I_d)$ are independent Gaussian vectors.

Proof of Theorem 9. The sets S_ω and S_u are non-empty, compact, and convex by assumption. As the function $\langle u, \mathbf{W}\omega\rangle + \psi((\omega,\omega'),(u,u'))$ is continuous, finite, and convex-concave on $S_\omega \times S_u$, it holds from the minimax result in Rockafellar (1997) (Corollary 37.3.2) that

$$\Phi(\mathbf{W}) = \max_{(u,u')\in S_U} \min_{(\omega,\omega')\in S_W} \langle u, \mathbf{W}\omega\rangle + \psi((\omega,\omega'),(u,u')),$$

where we define $\Phi(\mathbf{W})$ as $\min_{(\omega,\omega')\in S_W} \max_{(u,u')\in S_U} \langle u, \mathbf{W}\omega\rangle + \psi((\omega,\omega'),(u,u'))$. Consequently, the min-max problem in (19) is replaced with a max-min problem. This form implies

$$-\Phi(\mathbf{W}) = \min_{(u,u')\in S_U} \max_{(\omega,\omega')\in S_W} -\langle u, \mathbf{W}\omega\rangle - \psi((\omega,\omega'),(u,u')).$$

By using the symmetry of $\mathbf{W}$, we obtain that for any $c \in \mathbb{R}$,

$$\mathbb{P}(-\Phi(\mathbf{W}) \leq c) = \mathbb{P}\left(\min_{(u,u') \in S_U} \max_{(\omega,\omega') \in S_W} \{\langle u, \mathbf{W}\omega \rangle - \psi((\omega,\omega'),(u,u'))\} \leq c \right).$$

Then, by a variant of the Gaussian minimax theorem (Theorem 10 of Koehler *et al.* (2021)), we have

$$\begin{aligned} &\mathbb{P}(-\Phi(\mathbf{W}) < c) \\ &\leq 2\mathbb{P}\left(\min_{(u,u') \in S_U} \max_{(\omega,\omega') \in S_W} \{\|u\|\langle H, \omega \rangle + \|\omega\|\langle G, u \rangle - \psi((\omega,\omega'),(u,u'))\} \leq c \right) \\ &= 2\mathbb{P}\left(\min_{(u,u') \in S_U} \max_{(\omega,\omega') \in S_W} \{-\|u\|\langle H, \omega \rangle - \|\omega\|\langle G, u \rangle - \psi((\omega,\omega'),(u,u'))\} \leq c \right), \end{aligned}$$

where the last equation follows because of the symmetry of H and G. Note that we have

$$\begin{aligned} &\min_{(u,u') \in S_U} \max_{(\omega,\omega') \in S_W} \{-\|u\|\langle H, \omega \rangle - \|\omega\|\langle G, u \rangle - \psi((\omega,\omega'),(u,u'))\} \\ &= - \max_{(u,u') \in S_U} \min_{(\omega,\omega') \in S_W} \{\|u\|\langle H, \omega \rangle + \|\omega\|\langle G, u \rangle + \psi((\omega,\omega'),(u,u'))\}. \end{aligned}$$

By the minimax inequality (Rockafellar (1997), Lemma 36.1), we obtain that for all G, H,

$$\begin{aligned} &\max_{(u,u') \in S_U} \min_{(\omega,\omega') \in S_W} \{\|\omega\|\langle G, u \rangle + \|u\|\langle H, \omega \rangle + \psi((\omega,\omega'),(u,u'))\} \\ &\leq \min_{(\omega,\omega') \in S_W} \max_{(u,u') \in S_U} \{\|\omega\|\langle G, u \rangle + \|u\|\langle H, \omega \rangle + \psi((\omega,\omega'),(u,u'))\} := \phi(G,H). \end{aligned}$$

Therefore, we have for any $c \in \mathbb{R}$,

$$\mathbb{P}(\Phi(\mathbf{W}) > -c) = \mathbb{P}(-\Phi(\mathbf{W}) < c) \leq 2\mathbb{P}(-\phi(G,H) \leq c) = 2\mathbb{P}(\phi(G,H) \geq -c).$$

□

Appendix C: Upper bound for projected residual mean squared error

In this section, we provide the upper bound for the projected RMSE. Specifically, we prove Corollary 10 in the main body, and then give Corollary 16, which generalized a norm. The objective of this section is to show a general upper bound (Theorem 15). To this end, we analyze the projected RMSE by CGMT using Lemmas 12 and 13. We then analyze the projected RMSE in Lemma 14 to show Theorem 15, leading to Corollaries 10 and 16.

In the following lemma, we rewrite the projected RMSE (2) in the form of an optimization problem to use CGMT. In the statement, we use the empirical squared risk $\widehat{L}(\theta)$ in (3) and the representation of the data matrix $\mathbf{X}$ in (16) and (17). As the ridgeless estimator $\widehat{\theta}$ satisfies $\widehat{L}(\widehat{\theta}) = 0$, we are interested in a parameter θ which satisfies $\widehat{L}(\theta) = 0$.

Lemma 12. *Let $\mathcal{K}$ denote a compact set in $\mathbb{R}^p$. Assume Assumptions 1 and 2 hold. Define the primary optimization problem (PO) as*

$$\Phi := \max_{\substack{(\theta_1,\theta_2)\in S,\\ \mathbf{W}_1\theta_1+\mathbf{W}_2\theta_2=\xi}} \|\theta_1\|_2^2, \tag{18}$$

where we define $S := \{(\theta_1,\theta_2) : \exists\theta \in \mathcal{K}$ s.t. $\theta_1 = \Xi_z^{1/2}(\theta-\theta_0)$ and $\theta_2 = \Sigma_u^{1/2}(\theta-\theta_0))\}$. Then, the following maximized projected RMSE in (2) is equal in distribution to the PO:

$$\max_{\theta\in\mathcal{K},\widehat{L}(\theta)=0} E\left[(E[\langle\theta,X\rangle - \langle\theta_0,X\rangle|Z])^2\right] \overset{\mathcal{D}}{=} \Phi.$$

Proof of Lemma 12. Note that $\widehat{L}(\theta) = 0$ is equivalent to $\mathbf{Y} = \mathbf{X}\theta$. By the definitions of Ξ_z and Σ_u, we have

$$\mathbf{X} \overset{\mathcal{D}}{=} \mathbf{W}_1\Xi_z^{1/2} + \mathbf{W}_2\Sigma_u^{1/2}.$$

Hence, we obtain

$$\begin{aligned}
&\max_{\theta\in\mathcal{K},\widehat{L}(\theta)=0} E\left[(E[\langle\theta,X\rangle - \langle\theta_0,X\rangle|Z])^2\right]\\
&= \max_{\theta\in\mathcal{K},\widehat{L}(\theta)=0} (\theta-\theta_0)^\top\Pi_0\mathbb{E}[ZZ^\top]\Pi_0^\top(\theta-\theta_0)\\
&= \max_{\theta\in\mathcal{K},\mathbf{X}\theta=\mathbf{Y}} \|\theta-\theta_0\|_{\Xi_z}^2\\
&= \max_{\theta\in\mathcal{K},\mathbf{X}(\theta-\theta_0)=\xi} \|\theta-\theta_0\|_{\Xi_z}^2\\
&\overset{\mathcal{D}}{=} \max_{\substack{\theta\in\mathcal{K}-\theta_0\\ (\mathbf{W}_1\Xi_z^{1/2}+\mathbf{W}_2\Sigma_u^{1/2})\theta=\xi}} \|\theta\|_{\Xi_z}^2.
\end{aligned}$$

By the definition of S, we have

$$\max_{\substack{\theta\in\mathcal{K}-\theta_0\\ (\mathbf{W}_1\Xi_z^{1/2}+\mathbf{W}_2\Sigma_u^{1/2})\theta=\xi}} \|\theta\|_{\Xi_z}^2 = \max_{\substack{(\theta_1,\theta_2)\in S\\ \mathbf{W}_1\Xi_z^{1/2}\theta_1+\mathbf{W}_2\Sigma_u^{1/2}\theta_2=\xi}} \|\theta_1\|_2^2.$$

Then, the stated result holds. □

Lemma 13. *Let $G \sim N(0,I_n)$, $H \sim N(0,I_d)$ be Gaussian vectors independent of $\mathbf{W}_1,\mathbf{W}_2,\xi$, and each other. Define the auxiliary optimization problem (AO) as*

$$\phi := \max_{\substack{(\theta_1,\theta_2)\in S\\ \|\xi-\mathbf{W}_2\theta_2-G\|\theta_1\|_2\|_2\le\langle\theta_1,H\rangle}} \|\theta_1\|_2^2. \tag{24}$$

Then, it holds that

$$\mathbb{P}(\Phi > t|\mathbf{W}_2,\xi) \le 2\mathbb{P}(\phi \ge t|\mathbf{W}_2,\xi).$$

Furthermore, by taking expectations, we obtain

$$\mathbb{P}(\Phi > t) \le 2\mathbb{P}(\phi \ge t).$$

Proof of Lemma 13. This lemma is quite similar to Lemma 4 in Koehler *et al.* (2021). The only difference between the two is the objective function of the constrained maximization problems. However, because the objective function (24) does not affect the proof of Lemma 4 in Koehler *et al.* (2021), the result of Lemma 13 also holds. □

We then offer a bound on the projected RMSE. The following lemma is an extension of Lemma 5 in Koehler *et al.* (2021) to the case where the covariates correlate with errors.

As preparation, we define the Gaussian width, which is used in Lemma 14.

Definition 7 (Gaussian width (Vershynin, 2018))**.** The Gaussian width of a set $S \subset \mathbb{R}^p$ is

$$W(S) := \underset{H \sim N(0, I_d)}{E} \left[\sup_{s \in S} |\langle s, H \rangle| \right].$$

Lemma 14. *Let $\beta = 12\sqrt{\frac{\log(32/\delta)}{n}} + 3\sqrt{\frac{\mathrm{rank}(\Sigma_u)}{n}}$. If n is sufficiently large such that $\beta \leq 1$, for every $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\phi \leq \frac{1+\beta}{n} \left\{ W(\Xi_z^{1/2}\mathcal{K}) + \mathrm{rad}(\Xi_z^{1/2}\mathcal{K})\sqrt{2\log(16/\delta)} + \|\theta_0\|_{\Xi_z}\sqrt{2\log(16/\delta)} \right\}^2 - \widetilde{\sigma}^2, \tag{25}$$

where we define

$$\widetilde{\sigma}^2 := \sigma^2 - \|\omega\|^2_{\Sigma_u^+} = \min_{\theta_2 \in \Sigma_u^{1/2}\mathbb{R}^p} \left(\sigma^2 - \|\rho\|^2 + \|\theta_2 - \rho\|_2^2 \right).$$

Proof of Lemma 14. Fix $\delta \in (0,1)$ in this proof. To simplify notations, we define coefficients:

$$\alpha_1 := 2\sqrt{\frac{\log(32/\delta)}{n}} \quad \text{and} \quad \alpha_2 := \sqrt{\frac{\mathrm{rank}(\Sigma_u)+1}{n}} + 2\sqrt{\frac{\log(16/\delta)}{n}}.$$

To prepare for the derivation of the upper bound, we consider a list of the following inequalities, each of which holds with probability at least $1-\delta/8$.

(i) By (90) in Lemma 39, uniformly over all $\theta_2 \in \Sigma_u^{1/2}(\mathcal{K} - \theta_0)$, it holds that

$$|\langle \xi - \mathbf{W}_2\theta_2, G \rangle| \leq \|\xi - \mathbf{W}_2\theta_2\|_2 \|G\|_2 \alpha_2. \tag{26}$$

V, s, and δ in Lemma 39 correspond to G, $\xi - \mathbf{W}_2\theta_2$, and $\delta/8$ in (26), respectively.

(ii) By Lemma 40, it holds that

$$-\alpha_1 \leq \frac{1}{\sqrt{n}}\|G\|_2 - 1 \leq \alpha_1. \tag{27}$$

Moreover, as we obtain the following from (16) that

$$\begin{pmatrix} W_{2,i} \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0}_{p\times 1} \\ 0 \end{pmatrix}, \begin{pmatrix} I_{p\times p} & \rho \\ \rho^T & \sigma^2 \end{pmatrix} \right),$$

we have $\xi_i - W_{2,i}^T\theta_2 \sim N(0, \sigma^2 - 2\rho^T\theta_2 + \|\theta_2\|^2)$. As $\{\xi_i - W_{2,i}^T\theta_2\}_{i=1}^n$ are i.i.d., we have

$$\xi - \mathbf{W}_2\theta_2 \overset{\mathcal{D}}{=} \sqrt{\sigma^2 - 2\rho^T\theta_2 + \|\theta_2\|^2}G.$$

Further, by Lemma 40, we have

$$\begin{aligned} -\alpha_1\sqrt{\sigma^2 - 2\rho^T\theta_2 + \|\theta_2\|^2} &\le \frac{1}{\sqrt{n}}\|\xi - \mathbf{W}_2\theta_2\|_2 - \sqrt{\sigma^2 - 2\rho^T\theta_2 + \|\theta_2\|^2} \\ &\le \alpha_1\sqrt{\sigma^2 - 2\rho^T\theta_2 + \|\theta_2\|^2}. \end{aligned} \tag{28}$$

(iii) By the standard Gaussian tail bound $\mathbb{P}(|Z| \ge t) \le 2e^{-t^2/2}$, it holds that

$$|\langle \Xi_z^{1/2}\theta_0, H\rangle| \overset{\mathcal{D}}{=} |Z'| \le \|\theta_0\|_{\Xi_z}\sqrt{2\log(16/\delta)}, \tag{29}$$

where $Z' \sim N(0, \|\theta_0\|_{\Xi_z}^2)$.

(iv) By Theorem 43, it holds that

$$\max_{\theta_1 \in \Xi_z^{1/2}\mathcal{K}} |\langle \theta_1, H\rangle| \le W(\Xi_z^{1/2}\mathcal{K}) + \mathrm{rad}(\Xi_z^{1/2}\mathcal{K})\sqrt{2\log(16/\delta)} \tag{30}$$

because $\max_{\theta_1 \in \Xi_z^{1/2}\mathcal{K}} |\langle \theta_1, H\rangle|$ is a $\mathrm{rad}(\Xi_z^{1/2}\mathcal{K})$-Lipschitz function of H, and $W(\Xi_z^{1/2}\mathcal{K}) = \mathbb{E}[\sup_{\theta_1 \in \Xi_z^{1/2}\mathcal{K}} |\langle \theta_1, H\rangle|]$.

We further prepare several inequalities. By squaring the last constraint in the definition of the auxiliary optimization problem ϕ, we see that

$$\begin{aligned} \langle \theta_1, H\rangle^2 &\ge \|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|^2 \\ &= \|\xi - \mathbf{W}_2\theta_2\|_2^2 + \|\theta_1\|_2^2\|G\|_2^2 - 2\langle \xi - \mathbf{W}_2\theta_2, \|\theta_1\|_2 G\rangle. \end{aligned}$$

From (26) and the AM-GM inequality ($a^2/2 + b^2/2 \ge ab$), we have

$$\langle \theta_1, H\rangle^2 \ge (1-\alpha_2)[\|\xi - \mathbf{W}_2\theta_2\|_2^2 + \|\theta_1\|_2^2\|G\|_2^2].$$

From the rearrangement of the above inequality, we have

$$\begin{aligned} \|\theta_1\|_2^2 &\le \frac{(1-\alpha_2)^{-1}\langle \theta_1, H\rangle^2 - \|\xi - \mathbf{W}_2\theta_2\|_2^2}{\|G\|_2^2} \\ &\le \frac{(1-\alpha_2)^{-1}\langle \theta_1, H\rangle^2 - \|\xi - \mathbf{W}_2\theta_2\|_2^2}{n(1-\alpha_1)^2} \\ &\le \frac{(1-\alpha_2)^{-1}\langle \theta_1, H\rangle^2 - n(1-\alpha_1)^2(\|\theta_2\|^2 - 2\rho^T\theta_2 + \sigma^2)}{n(1-\alpha_1)^2}, \end{aligned} \tag{31}$$

where the second inequality holds from (27) and the third inequality holds from (28).

Now, we are ready to construct the upper bound on ϕ from the restriction of the optimization problem (24). Plugging (31) into (24), we obtain

$$\begin{aligned}\phi &\leq \max_{\theta_1 \in \Xi_z^{1/2}(\mathcal{K}-\theta_0)} \frac{(1-\alpha_2)^{-1}\langle \theta_1, H\rangle^2}{n(1-\alpha_1)^2} + \max_{\theta_2 \in \Sigma_u^{1/2}(\mathcal{K}-\theta_0)} -(\|\theta_2\|^2 - 2\rho^T\theta_2 + \sigma^2) \\
&\leq \frac{1}{n(1-\alpha_2)(1-\alpha_1)^2}\left(\max_{\theta_1 \in \Xi_z^{1/2}\mathcal{K}} |\langle \theta_1, H\rangle| + |\langle \Xi_z^{1/2}\theta_0, H\rangle|\right)^2 \\
&\quad - \min_{\theta_2 \in \Sigma_u^{1/2}(\mathcal{K}-\theta_0)} \left(\sigma^2 - \|\rho\|^2 + \|\theta_2 - \rho\|_2^2\right) \\
&\leq \frac{1}{n(1-\alpha_2)(1-\alpha_1)^2}\left(\max_{\theta_1 \in \Xi_z^{1/2}\mathcal{K}} |\langle \theta_1, H\rangle| + |\langle \Xi_z^{1/2}\theta_0, H\rangle|\right)^2 \\
&\quad - \min_{\theta_2 \in \Sigma_u^{1/2}\mathbb{R}^p} \left(\sigma^2 - \|\rho\|^2 + \|\theta_2 - \rho\|_2^2\right) \\
&\leq \frac{1}{n(1-\alpha_2)(1-\alpha)^2} \\
&\quad \times \left(W(\Xi_z^{1/2}\mathcal{K}) + \mathrm{rad}(\Xi_z^{1/2}\mathcal{K})\sqrt{2\log(16/\delta)} + \|\theta_0\|_{\Xi_z}^2\sqrt{2\log(16/\delta)}\right)^2 - \widetilde{\sigma}^2,\end{aligned}$$

where the fourth inequality holds from (29) and (30).

We simplify the effect of α_1 and α_2 on the upper bound for ϕ. As $(1-\alpha_1)^2 \geq 1-2\alpha_1$, we have

$$\frac{1}{(1-\alpha_2)(1-\alpha_1)^2} \leq \frac{1}{(1-\alpha_2)(1-2\alpha_1)}.$$

If $\alpha_1 < 1/2$ and $\alpha_2 < 1$,

$$\begin{aligned}(1-2\alpha_1)(1-\alpha_2) &= 1 - \alpha_2 - 2\alpha_1 + 2\alpha_1\alpha_2 \\ &\geq 1 - \alpha_2 - 2\alpha_1.\end{aligned}$$

Assume $\alpha_2 + 2\alpha_1 < 1/2$. By using the inequality $(1-x)^{-1} \leq 1 + 2x$ for $x \in [0, 1/2]$, we can show that

$$\frac{1}{(1-\alpha_2)(1-\alpha_1)^2} \leq \frac{1}{(1-\alpha_2)(1-2\alpha_1)} \leq 1 + 2\alpha_2 + 4\alpha_1.$$

Therefore, if we choose β to satisfy the following inequality:

$$2\alpha_2 + 4\alpha_1 \leq 3\sqrt{\frac{\mathrm{rank}(\Sigma_u)}{n}} + 12\sqrt{\frac{\log(32/\delta)}{n}} := \beta,$$

the stated result holds. □

Finally, we obtain the generalization bound from Lemma 14.

Theorem 15 (General Bound)**.** *There exists an absolute constant $C_1 \leq 24$ such that the following is true. Assume Assumptions 1 and 2 hold. Let $\mathcal{K}$ denote* *an arbitrary compact set, and take* $\Sigma_x = \Xi_z + \Sigma_u$*. Fixing* $\delta \leq 1/4$*, let* $\beta = C_1 \left(\sqrt{\operatorname{rank}(\Sigma_u)/n} + \sqrt{\log(1/\delta)/n}\right)$*. If* n *is large enough that* $\beta \leq 1$*, then the following holds with probability at least* $1-\delta$*:*

$$\begin{aligned}&\max_{\theta\in\mathcal{K}, Y=X\theta} \|\theta-\theta_0\|_{\Xi_z}^2 \\ &\leq \frac{1+\beta}{n}\left[W(\Xi_z^{1/2}\mathcal{K}) + \left(\operatorname{rad}(\Xi_z^{1/2}\mathcal{K}) + \|\theta_0\|_{\Xi_z}\right)\sqrt{2\log\frac{32}{\delta}}\right]^2 - \widetilde{\sigma}^2.\end{aligned}$$

Proof of Theorem 15. For any $t > 0$, it holds from Lemmas 12 and 13 that

$$\mathbb{P}\left(\max_{\theta\in\mathcal{K}, Y=X\theta} \|\theta-\theta_0\|_{\Xi_z}^2 > t\right) \leq 2\mathbb{P}(\phi \geq t).$$

Lemma 14 implies that the above term is upper bounded by δ if we choose t using the result (25) with δ replaced by $\delta/2$. Then, we obtain the stated result. □

By using the definition of the radius of sets and the Gaussian width, we can reduce the generalization bound in Theorem 15 to a simpler bound:

Corollary 16. *There exists an absolute constant* $C_1 \leq 32$ *such that the following is true. Assume Assumptions 1 and 2 hold. Pick* $\Sigma_x = \Xi_z + \Sigma_u$*, fix* $\delta \leq 1/4$*, and let* $\gamma = C_1(\sqrt{\log(1/\delta)/r_{\|\cdot\|}(\Xi_z)} + \sqrt{\log(1/\delta)/n} + \sqrt{\operatorname{rank}(\Sigma_u)/n})$*. If* $B \geq \|\theta_0\|$ *and* n *is large enough that* $\gamma \leq 1$*, the following holds with probability at least* $1-\delta$*:*

$$\max_{\|\theta\|\leq B, \mathbf{Y}=\mathbf{X}\theta} \|\theta-\theta_0\|_{\Xi_z}^2 \leq (1+\gamma)\frac{\left(B\mathbb{E}\|\Sigma_2^{1/2}H\|_*\right)^2}{n} - \widetilde{\sigma}^2.$$

Proof of Corollary 16. Let $\mathcal{K}$ define $\{\theta : \|\theta\| \leq B\}$ in Theorem 15. By the definition of the Gaussian width and the radius of a set, we have

$$\begin{aligned}W(\Xi_z^{1/2}\mathcal{K}) &= \mathbb{E}\sup_{\|\theta\|\leq B} |\langle \Xi_z^{1/2}\theta, H\rangle| = \mathbb{E}\sup_{\|\theta\|\leq B} |\langle \theta, \Xi_z^{1/2}H\rangle| = B\mathbb{E}\|\Xi_z^{1/2}H\|_*, \\ \operatorname{rad}(\Xi_z^{1/2}\mathcal{K}) &= \sup_{\|\theta\|\leq B} \|\Xi_z^{1/2}\theta\|_2 = B\sup_{\|\theta\|\leq 1}\|\theta\|_{\Xi_z}.\end{aligned}$$

Hence, we obtain

$$r_{\|\cdot\|}(\Xi_z) = \left(\frac{W(\Xi_z^{1/2}\mathcal{K})}{\operatorname{rad}(\Xi_z^{1/2}\mathcal{K})}\right)^2.$$

As we have $\|\theta_0\|_{\Xi_z} = \sqrt{(\theta_0^\top \Xi_z\theta_0/\|\theta_0\|^2)\cdot\|\theta_0\|^2}$, it holds that $\|\theta_0\|_{\Xi_z} \leq \|\theta_0\| \sup_{\|\theta\|\leq 1}\|\theta\|_{\Xi_z}$. By definition, it is clear that $\|\theta_0\|_{\Xi_z} \leq \operatorname{rad}(\Xi_z^{1/2}\mathcal{K})$. Hence,

$$W(\Xi_z^{1/2}\mathcal{K}) + \left(\operatorname{rad}(\Xi_z^{1/2}\mathcal{K}) + \|\theta_0\|_{\Xi_z}\right)\sqrt{2\log\frac{32}{\delta}}$$

$$\begin{aligned}&\leq W(\Xi_z^{1/2}\mathcal{K}) + 2\sqrt{2\log\frac{32}{\delta}}\,\mathrm{rad}(\Xi_z^{1/2}\mathcal{K})\\&= W(\Xi_z^{1/2}\mathcal{K}) + 2\sqrt{\frac{2\log(32/\delta)}{r_{\|\cdot\|}(\Xi_z)}}W(\Xi_z^{1/2}\mathcal{K})\\&= \left(1+2\sqrt{\frac{2\log(32/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)B\mathbb{E}\|\Xi_z^{1/2}H\|_*\end{aligned}$$

holds where the last equality holds by the definition of $r_{\|\cdot\|}(\Xi_z)$. Provided that $\gamma \leq 1$ and $\delta \leq 1/4$, we obtain

$$\begin{aligned}&(1+\beta)\left(1+2\sqrt{\frac{2\log(32/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)^2\\&\leq \left(1+\beta+4\sqrt{\frac{2\log(32/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)\left(1+2\sqrt{\frac{2\log(32/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)\\&\leq 1+\gamma,\end{aligned}$$

where the inequalities follow from using $(1+x)(1+y) \leq 1+x+2y$ for $x \leq 1$. Plugging into Theorem 15 completes the proof. □

When we consider the Euclidean space, we can reduce the main generalization bound to a simpler bound.

Corollary 10 *There exists an absolute constant $C_1 \leq 32$ such that the following is true. Assume Assumptions 1 and 2 hold. Pick $\Sigma_x = \Xi_z + \Sigma_u$, fix $\delta \leq 1/4$, and let $\gamma = C_1\left(\sqrt{\log(1/\delta)/r(\Xi_z)} + \sqrt{\log(1/\delta)/n} + \sqrt{\mathrm{rank}(\Sigma_u)/n}\right)$. If $B \geq \|\theta_0\|_2$ and n is large enough that $\gamma \leq 1$, the following holds with probability at least $1-\delta$:*

$$\max_{\|\theta\|_2\leq B, \mathbf{Y}=\mathbf{X}\theta}\|\theta-\theta_0\|_{\Xi_z}^2 \leq (1+\gamma)\frac{B^2\mathrm{tr}(\Xi_z)}{n} - \widetilde{\sigma}^2.$$

Proof of Corollary 10. By trivial calculation, we have

$$W(\Xi_z^{1/2}\mathcal{K}) \leq B\mathrm{tr}(\Xi_z)^{1/2} \quad \textit{and} \quad \mathrm{rad}(\Xi_z^{1/2}\mathcal{K}) = B\|\Xi_z\|_{\mathrm{op}}^{1/2}.$$

By the definition of $\mathrm{rad}(\Xi_z^{1/2}\mathcal{K})$, we have $\|\theta_0\|_{\Xi_z} \leq \mathrm{rad}(\Xi_z^{1/2}\mathcal{K}) = B\|\Xi_z\|_{op}^{1/2}$. Hence,

$$\begin{aligned}&W(\Xi_z^{1/2}\mathcal{K}) + \left(\mathrm{rad}(\Xi_z^{1/2}\mathcal{K}) + \|\theta_0\|_{\Xi_z}\right)\sqrt{2\log\frac{32}{\delta}}\\&\leq W(\Xi_z^{1/2}\mathcal{K}) + 2\sqrt{2\log\frac{32}{\delta}}\,\mathrm{rad}(\Xi_z^{1/2}\mathcal{K})\end{aligned}$$

$$
\begin{aligned}
&\leq B\mathrm{tr}(\Xi_z)^{1/2} + 2\sqrt{2\log\frac{32}{\delta}}B\|\Xi_z\|_{\mathrm{op}}^{1/2} \\
&= \left(1 + 2\sqrt{\frac{2\log(32/\delta)}{r(\Xi_z)}}\right) B\mathrm{tr}(\Xi_z)^{1/2}
\end{aligned}
$$

holds. The last equality holds by the definition of the effective rank $r(\Xi_z)$ in Definition 1. Under our assumptions that $\gamma \leq 1$ and $\delta \leq 1/4$, we can show that

$$
\begin{aligned}
&(1+\beta)\left(1 + 2\sqrt{\frac{2\log(32/\delta)}{r(\Xi_z)}}\right)^2 \\
&\leq \left(1 + \beta + 4\sqrt{\frac{2\log(32/\delta)}{r(\Xi_z)}}\right)\left(1 + 2\sqrt{\frac{2\log(32/\delta)}{r(\Xi_z)}}\right) \\
&\leq 1 + \gamma,
\end{aligned}
$$

where the inequality follows from using $(1+x)(1+y) \leq 1 + x + 2y$ for $x \leq 1$ and $y \geq 0$. Plugging into Theorem 15 completes the proof. □

Appendix D: Bounds for the ridgeless estimator

In this section, we provide an upper bound of a norm of the ridgeless estimator with the existence of a correlation between the covariates and the error terms. In Lemmas 17 and 18, we rewrite the norm of the estimator to apply CGMT. Lemma 19 bounds an element in the rewritten form of the norm. Then, Theorem 20 develops the desired bound on the norm, and Theorem 21 offers its Euclidean norm case.

First, we formulate the constrained minimization problem with Gaussian covariates.

Lemma 17. *Assume Assumptions 1 and 2 hold. Let $\|\cdot\|$ denote an arbitrary norm. Define the primary optimization problem (PO) as*

$$
\Phi := \min_{\mathbf{W}_1\theta_1 + \mathbf{W}_2\theta_2 = \xi} \|\Sigma_x^{-1/2}(\theta_1 + \theta_2)\|,
$$

where $\theta_1 = \Xi_z^{1/2}\theta$ and $\theta_2 = \Sigma_u^{1/2}\theta$ for $\theta \in \mathbb{R}^p$. Then, for any t, it holds that

$$
\mathbb{P}\left(\min_{\mathbf{X}\theta=\mathbf{Y}} \|\theta\| > t\right) \leq \mathbb{P}\left(\|\theta_0\| + \Phi > t\right).
$$

Proof of Lemma 17. We have $\mathbf{X} \overset{D}{=} \mathbf{W}_1\Xi_z^{1/2} + \mathbf{W}_2\Sigma_u^{1/2}$ by equality in distribution. It follows from the triangle inequality and change of variables that

$$
\min_{\mathbf{X}\theta=\mathbf{Y}} \|\theta\| = \min_{\mathbf{X}\theta=\xi} \|\theta + \theta_0\| \leq \|\theta_0\| + \min_{(\mathbf{W}_1\Xi_z^{1/2} + \mathbf{W}_2\Sigma_u^{1/2})\theta=\xi} \|\theta\|.
$$

As $\Sigma_x^{1/2}\theta = \Xi_z^{1/2}\theta + \Sigma_u^{1/2}\theta$, we have $\theta = \Sigma_x^{-1/2}(\theta_1+\theta_2)$ where $\theta_1 = \Xi_z^{1/2}\theta$ and $\theta_2 = \Sigma_u^{1/2}\theta$. Then, the following inequality holds:

$$\min_{X\theta=Y}\|\theta\| \leq \|\theta_0\| + \min_{\mathbf{W}_1\theta_1+\mathbf{W}_2\theta_2=\xi}\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|.$$

□

As in Lemma 13, we use the result of Theorem 9 to derive the auxiliary optimization problem.

Lemma 18. *In the same setting as Lemma 17, let $G \sim N(0, I_n)$ and $H \sim N(0, I_d)$ be Gaussian vectors independent of $\xi, \mathbf{W}_1, \mathbf{W}_2$, and each other. Define the auxiliary optimization problem (AO) as*

$$\phi := \min_{\|\xi-\mathbf{W}_2\theta_2-\|\theta_1\|_2 G\|_2\leq\langle H,\theta_1\rangle}\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|. \tag{32}$$

Then, it holds that

$$\mathbb{P}(\Phi > t|\xi,\mathbf{W}_2) \leq 2\mathbb{P}(\phi \geq t|\xi,\mathbf{W}_2),$$

and taking the expectations we have

$$\mathbb{P}(\Phi > t) \leq 2\mathbb{P}(\phi \geq t).$$

Proof of Lemma 18. We reformulate Φ to apply the extended CGMT (Theorem 9). By using Lagrangian multipliers, it holds that

$$\begin{aligned}\Phi &= \min_{\theta_1,\theta_2}\max_{\lambda}\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\| + \langle\lambda, \mathbf{W}_1\theta_1+\mathbf{W}_2\theta_2-\xi\rangle\\ &= \min_{\theta_1,\theta_2}\max_{\lambda}\langle\lambda,\mathbf{W}_1\theta_1\rangle + \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\| - \langle\lambda,\xi-\mathbf{W}_2\theta_2\rangle.\end{aligned}$$

As $\mathbf{W}_1$ is independent of $\mathbf{W}_2$ and ξ, the distribution of $\mathbf{W}_1$ is unchanged even though we condition on $\mathbf{W}_2$ and ξ. For any $r, t > 0$, we define

$$\Phi_r(t) := \min_{\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\leq 2t}\max_{\|\lambda\|_2\leq r}\langle\lambda,\mathbf{W}_1\theta_1\rangle + \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\| - \langle\lambda,\xi-\mathbf{W}_2\theta_2\rangle.$$

The corresponding AO is defined as follows:

$$\begin{aligned}&\phi_r(t)\\ &:= \min_{\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\leq 2t}\max_{\|\lambda\|_2\leq r}\|\theta_1\|_2\langle G,\lambda\rangle + \|\lambda\|_2\langle H,\theta_1\rangle\\ &\quad + \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\| - \langle\lambda,\xi-\mathbf{W}_2\theta_2\rangle\\ &= \min_{\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\leq 2t}\max_{\|\lambda\|_2\leq r}\|\lambda\|_2\langle H,\theta_1\rangle - \langle\lambda,\xi-\mathbf{W}_2\theta_2-G\|\theta_1\|_2\rangle\\ &\quad + \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\\ &= \min_{\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\leq 2t}\max_{0\leq\lambda\leq r}\lambda(\langle H,\theta_1\rangle + \|\xi-\mathbf{W}_2\theta_2-G\|\theta_1\|_2\|_2)\end{aligned}$$

$$+ \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|.$$

As two optimization problems $\Phi_r(t)$ and $\phi_r(t)$ are defined on compact sets, we can apply Theorem 9 to those two optimization problems. As an intermediate problem between Φ and $\Phi_r(t)$, we introduce

$$\begin{aligned}\Phi(t) :&= \min_{\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\le 2t} \max_{\lambda} \langle \lambda, \mathbf{W}_1\theta_1\rangle + \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\| - \langle \lambda, \xi - \mathbf{W}_2\theta_2\rangle \\ &= \min_{\substack{\mathbf{W}_1\theta_1+\mathbf{W}_2\theta_2=\xi \\ \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\le 2t}} \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\end{aligned}$$

and also define the corresponding AO as

$$\begin{aligned}&\phi(t) \\ &:= \min_{\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\le 2t} \max_{\lambda\ge 0} \lambda(\langle H, \theta_1\rangle + \|\xi - \mathbf{W}_2\theta_2 - G\|\theta_1\|_2\|_2) \\ &\quad + \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\| \\ &= \min_{\substack{\|\xi - \mathbf{W}_2\theta_2 - G\|\theta_1\|_2\|_2 \le \langle H, \theta_1\rangle \\ \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|\le 2t}} \|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|_2.\end{aligned}$$

By definition, clearly, $\Phi \le \Phi(t)$. Therefore, if $\Phi > t$, $\Phi(t) > t$ holds. If $t \ge \Phi$, then there exists (θ_1^*, θ_2^*) such that $\|\Sigma_x^{-1/2}(\theta_1^*+\theta_2^*)\| \le t$ and $\mathbf{W}_1\theta_1^* + \mathbf{W}_2\theta_2^* = \xi$. As $\|\Sigma_x^{-1/2}(\theta_1^*+\theta_2^*)\| \le 2t$, we obtain

$$\Phi(t) \le \|\Sigma_x^{-1/2}(\theta_1^*+\theta_2^*)\| \le t.$$

Therefore, it holds that

$$\Phi > t \Leftrightarrow \Phi(t) > t.$$

Likewise, $\phi(t) > t$ is equivalent to $\phi > t$.

To establish the result $\mathbb{P}(\Phi > t) \le 2\mathrm{Pr}(\phi > t)$, we need to clarify the relationship between Φ and $\Phi_r(t)$, ϕ and $\phi_r(t)$, respectively, that is,

$$Pr(\Phi > t | \xi, \mathbf{W}_2) \le \lim_{r\to\infty} \mathbb{P}(\Phi_r(t) > t | \xi, \mathbf{W}_2),$$

and

$$\lim_{r\to\infty} \mathbb{P}(\phi_r(t) > t | \xi, \mathbf{W}_2) \le \mathbb{P}(\phi > t | \xi, \mathbf{W}_2).$$

As $\phi_r(t) \le \phi(t)$ for any r, $\mathbb{P}(\phi_r(t) > t | \xi, \mathbf{W}_2) \le \mathbb{P}(\phi(t) > t | \xi, \mathbf{W}_2)$ holds. Then, all we need to show is the following:

$$\Phi_r(t) \to \Phi(t) \quad \text{as } r \to \infty.$$

We consider the following two cases: (i) $\Phi(t) = \infty$ and (ii) $\Phi(t) < \infty$.

Case (i): $\Phi(t) = \infty$, that is, the minimization problem defining $\Phi(t)$ is infeasible. In this case, for all $\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\| \le 2t$, we have

$$\|\mathbf{W}_1\theta_1 + \mathbf{W}_2\theta_2 - \xi\|_2 > 0.$$

By closedness, there exists $\eta = \eta(\mathbf{W}_1, \mathbf{W}_2, \xi)$ such that

$$\|\mathbf{W}_1\theta_1 + \mathbf{W}_2\theta_2 - \xi\|_2 \geq \eta.$$

By definition, η is independent of r. Then, it holds that

$$\Phi_r(t) = \min_{\|\Sigma_x^{-1/2}(\theta_1+\theta_2)\|_2 \leq 2t} \max_{\|\lambda\|_2 \leq r} \langle \lambda, \mathbf{W}_1\theta_1 + \mathbf{W}_2\theta_2 - \xi \rangle + \|\Sigma_x^{-1/2}(\theta_1 + \theta_2)\| \geq r\eta.$$

Therefore, $\Phi_r(t) \to \infty$ as $r \to \infty$.

Case (ii): $\Phi(t) < \infty$, that is, the minimization problem defining $\Phi(t)$ is feasible. By compactness, $\Phi_r(t)$ has solutions for the minimax problem. Let $(\theta_1(r), \theta_2(r))$ be one of solutions for $\Phi_r(t)$. If we take a sequence $\{(\theta_1(r), \theta_2(r))\}_{r=1}^\infty$, there exists a convergent subsequence $\{(\theta_1(r_n), \theta_2(r_n))\}_{n=1}^\infty$ by sequential compactness. Let $\{(\theta_1(\infty), \theta_2(\infty))\}$ be a convergent point of this subsequence. For the sake of contradiction, assume that $\mathbf{W}_1\theta_1(\infty) + \mathbf{W}_2\theta_2(\infty) \neq \xi$. By continuity, there exists η and ε such that, if $\|(\theta_1, \theta_2) - (\theta_1(\infty), \theta_2(\infty))\|_2 \leq \varepsilon$,

$$\|\mathbf{W}_1\theta_1 + \mathbf{W}_2\theta_2 - \xi\|_2 \geq \eta.$$

This implies that for a sufficiently large n, it holds that

$$\|\mathbf{W}_1\theta_1(r_n) + \mathbf{W}_2\theta_2(r_n) - \xi\|_2 \geq \eta.$$

As in the previous section, we have

$$\Phi_{r_n}(t) = \max_{\|\lambda\|_2 \leq r_n} \langle \lambda, \mathbf{W}_1\theta_1(r_n) + \mathbf{W}_2\theta_2(r_n) - \xi \rangle + \|\Sigma_x^{-1/2}(\theta_1(r_n) + \theta_2(r_n))\| \geq r_n\eta.$$

Hence, $\lim_{n\to\infty} \Phi_{r_n}(t) = \infty$. However, this is a contradiction because, for any r, $\Phi_r(t) \leq \Phi(t) < \infty$. Therefore, $\mathbf{W}_1\theta_1(\infty) + \mathbf{W}_2\theta_2(\infty) = \xi$. If we set $\lambda = 0$, we have

$$\Phi_{r_n}(t) \geq \|\Sigma_x^{-1/2}(\theta_1(r_n) + \theta_2(r_n))\|.$$

By continuity, we show that

$$\begin{aligned}\liminf_{n\to\infty} \Phi_{r_n}(t) &\geq \lim_{n\to\infty} \|\Sigma_x^{-1/2}(\theta_1(r_n) + \theta_2(r_n))\| \\ &= \|\Sigma_x^{-1/2}(\theta_1(\infty) + \theta_2(\infty))\| \geq \Phi(t).\end{aligned}$$

As, for any r, $\Phi_r(t) \leq \Phi(t)$, we have

$$\limsup_{n\to\infty} \Phi_{r_n}(t) \leq \Phi(t) \leq \liminf_{n\to\infty} \Phi_{r_n}(t),$$

that is, $\lim_{n\to\infty} \Phi_{r_n}(t) = \Phi(t)$. As $\Phi_r(t)$ is an increasing function in terms of r, we have $\lim_{r\to\infty} \Phi_r(t) = \Phi(t)$.

Through the application of Theorem 9 and two inequalities, $\mathbb{P}(\Phi > t|\xi, \mathbf{W}_2) \leq \lim_{r\to\infty} \mathbb{P}(\Phi_r(t) > t|\xi, \mathbf{W}_2)$ and $\lim_{r\to\infty} \mathbb{P}(\phi_r(t) > t|\xi, \mathbf{W}_2) \leq \mathbb{P}(\phi > t|\xi, \mathbf{W}_2)$, we prove the result $\mathbb{P}(\Phi > t) \leq 2\Pr(\phi > t)$. By the last part of Theorem 9, we have

$$\mathbb{P}(\Phi_r(t) > t|\xi, \mathbf{W}_2) \leq 2\mathbb{P}(\phi_r(t) > t|\xi, \mathbf{W}_2).$$

As $\Phi_r(t)$ monotonically increases to $\Phi(t)$ almost surely, it follows from the continuity of the probability measure that

$$\begin{aligned}\mathbb{P}(\Phi > t|\xi, \mathbf{W}_2) &= \mathbb{P}(\Phi(t) > t|\xi, \mathbf{W}_2)\\ &\le \mathbb{P}(\cup_r \cap_{r' \ge r} \Phi_{r'}(t) > t|\xi, \mathbf{W}_2)\\ &= \mathbb{P}(\underline{\lim}_{r\to\infty} \Phi_r(t) > t|\xi, \mathbf{W}_2)\\ &= \lim_{r\to\infty} \mathbb{P}(\Phi_r(t) > t|\xi, \mathbf{W}_2).\end{aligned}$$

As $\phi_r(t) \le \phi(t)$ holds for any r, $\mathbb{P}(\phi_r(t) > t|\xi, \mathbf{W}_2) \le \mathbb{P}(\phi(t) > t|\xi, \mathbf{W}_2)$. Therefore, we have

$$\mathbb{P}(\Phi > t|\xi, \mathbf{W}_2) \le 2 \lim_{r\to\infty} \mathbb{P}(\phi_r(t) > t|\xi, \mathbf{W}_2) \le 2\mathbb{P}(\phi(t) > t|\xi, \mathbf{W}_2).$$

□

Then, we obtain the general upper bound for the auxiliary optimization problem (AO).

Lemma 19. *Denote P_z and P_u as the orthogonal projection matrix onto the space spanned by Ξ_z and Σ_u, respectively. Let $v_* = \arg\min_{v\in\partial\|\Xi_z^{1/2}H\|} \|v\|_{\Xi_z}$. Assume that there exists $\varepsilon_1, \varepsilon_2 \ge 0$ such that with probability at least $1 - \delta/2$,*

$$\|v^*\|_{\Xi_z} \le (1 + \varepsilon_1)\mathbb{E}\|v^*\|_{\Xi_z}, \tag{33}$$

and

$$\|P_z v^*\|^2 \le 1 + \varepsilon_2. \tag{34}$$

Define ε as

$$\varepsilon := 16\sqrt{\frac{\operatorname{rank}(\Sigma_u)}{n}} + 28\sqrt{\frac{\log(32/\delta)}{n}} + 8\sqrt{\frac{\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}} + 2(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Xi_z)} + 2\varepsilon_2.$$

If n and the effective ranks are sufficiently large such that $\varepsilon \le 1$, then with probability at least $1 - \delta$, it holds that

$$\phi^2 \le \|\Sigma_u^+\omega\|^2 + (1+\varepsilon)\tilde{\sigma}^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}, \tag{35}$$

where we denote $r_{\|\cdot\|}(\Sigma)$ and $R_{\|\cdot\|}(\Sigma)$ as follows:

$$r_{\|\cdot\|}(\Sigma) = \left(\frac{E\|\Sigma^{1/2}H\|_*}{\sup_{\|u\|\le 1}\|u\|_\Sigma}\right)^2 \quad \textit{and } R_{\|\cdot\|}(\Sigma) = \left(\frac{E\|\Sigma^{1/2}H\|_*}{E\|v^*\|_\Sigma}\right)^2.$$

Proof of Lemma 19. Fix $\delta \in (0,1)$ in this proof. To simplify notations, we define coefficients:

$$\alpha_1 := 2\sqrt{\frac{\log(32/\delta)}{n}} \text{ and } \alpha_2 := \sqrt{\frac{\operatorname{rank}(\Sigma_u)+1}{n}} + 2\sqrt{\frac{\log(16/\delta)}{n}}.$$

To prepare for the derivation of the upper bound as in the proof of Lemma 14, we consider the following three inequalities:

(i) By Lemma 39, uniformly over all $\theta_2 \in \Sigma_u^{1/2}(\mathbb{R}^p - \theta_0)$, it holds that

$$|\langle \xi - \mathbf{W}_2\theta_2, G\rangle| \leq \|\xi - \mathbf{W}_2\theta_2\|_2\|G\|_2\alpha_2. \tag{26}$$

(ii) By Lemma 40, it holds that

$$-\alpha_1 \leq \frac{1}{\sqrt{n}}\|G\|_2 - 1 \leq \alpha_1 \tag{27}$$

and

$$\begin{aligned}-\alpha_1\sqrt{\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2} &\leq \frac{1}{\sqrt{n}}\|\xi - \mathbf{W}_2\theta_2\|_2 - \sqrt{\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2}\\ &\leq \alpha_1\sqrt{\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2}.\end{aligned} \tag{28}$$

(iii) By Theorem 43, it holds that

$$\begin{aligned}\|\Xi_z^{1/2}H\|_* &\geq \mathbb{E}\|\Xi_z^{1/2}H\|_* - \sup_{\|u\|\leq 1}\|u\|_{\Xi_z}\sqrt{2\log(8/\delta)}\\ &= \left(1 - \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)\mathbb{E}\|\Xi_z^{1/2}H\|_*,\end{aligned} \tag{36}$$

because $\|\Xi_z^{1/2}H\|_*$ is a $\sup_{\|u\|\leq 1}\|u\|_{\Xi_z}$-Lipschitz continuous function of H.

We construct the upper bound from the restriction of the optimization problem (32). From the restriction of the auxiliary problem, we have

$$\begin{aligned}\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2G\|_2^2 &= \|\xi - \mathbf{W}_2\theta_2\|_2^2 - 2\|\theta_1\|_2\langle\xi - \mathbf{W}_2\theta_2, G\rangle + \|\theta_1\|_2^2\|G\|_2^2\\ &\leq (1+\alpha_2)\left(\|\xi - \mathbf{W}_2\theta_2\|_2^2 + \|\theta_1\|_2^2\|G\|_2^2\right),\end{aligned}$$

where the last inequality follows from (26) and the AM-GM inequality. Combining the results of (27) and (64) yields

$$\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2G\|_2^2 \leq (1+\alpha_2)(1+\alpha_1)^2n\left(\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2 + \theta_1^T\theta_1\right). \tag{37}$$

To consider an upper bound of the ridgeless estimator, we need to choose a suitable θ which satisfies the restriction of the auxiliary problem. We consider the following form of θ:

$$\theta = P_u(\Sigma_u^+\omega) + sP_zv^*.$$

As $\Sigma_u^{1/2}\theta = \Sigma_u^{1/2}\Sigma_u^+\omega = (\Sigma_u^{1/2})^+\omega$ and $\Xi_z^{1/2}\theta = s\Xi_z^{1/2}v^*$, from (37) and the restriction of the auxiliary problem, it suffices to choose s such that

$$\begin{aligned}&(1+\alpha_2)(1+\alpha_1)^2n\left(\sigma^2 - 2\rho^T\Sigma_u^{1/2}\Sigma_u^+\omega + \omega^T\Sigma_u^+\Sigma_u\Sigma_u^+\omega + s^2\|\Xi_z^{1/2}v^*\|_2^2\right)\\ &= (1+\alpha_2)(1+\alpha_1)^2n(\widetilde{\sigma}^2 + s^2\|\Xi_z^{1/2}v^*\|_2^2)\\ &\leq (\langle H, s\Xi_z^{1/2}v^*\rangle)^2\end{aligned}$$

$$= s^2\|\Xi_z^{1/2}H\|_*^2.$$

Solving for s, we can choose

$$s^2 = \widetilde{\sigma}^2 \left(\frac{\|\Xi_z^{1/2}H\|_*^2}{(1+\alpha_2)(1+\alpha_1)^2 n} - \|v^*\|_{\Xi_z}^2 \right)^{-1},$$

under the assumption that

$$\left(\frac{\|\Xi_z^{1/2}H\|_*^2}{(1+\alpha_2)(1+\alpha_1)^2 n} - \|v^*\|_{\Xi_z}^2 \right) > 0. \tag{38}$$

We need to guarantee (38) holds. By (33) and (36), we have

$$\begin{aligned}
&\frac{\|\Xi_z^{1/2}H\|_*^2}{(1+\alpha_2)(1+\alpha_1)^2 n} - \|v^*\|_{\Xi_z}^2 \\
\ge& \frac{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}{(1+\alpha_2)(1+\alpha_1)^2 n} \left(1 - \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)^2 - (1+\varepsilon_1)^2 (\mathbb{E}\|v^*\|_{\Xi_z})^2 \\
\ge& \frac{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}{n} \left(\frac{1}{(1+\alpha_2)(1+\alpha_1)^2} \left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right) - (1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Sigma_2)} \right),
\end{aligned}$$

where the last inequality follows from the definition of $R_{\|\cdot\|}(\cdot)$.

As in the proof of Lemma 14, we linearize the terms including α_1 and α_2 to simplify the upper bound. Provided that $\alpha_1 < 1$, we have

$$\begin{aligned}
(1+\alpha_2)(1+\alpha_1)^2 &= 1 + 2\alpha_1 + \alpha_1^2 + \alpha_2 + 2\alpha_2\alpha_1 + \alpha_2\alpha_1^2 \\
&\le 1 + 3\alpha_1 + 4\alpha_2.
\end{aligned}$$

As $(1-x)^{-1} \ge 1 + x$ for any x, it holds that

$$\begin{aligned}
\frac{1}{(1+\alpha_2)(1+\alpha_1)^2} &\ge (1 + 3\alpha_1 + 4\alpha_2)^{-1} \\
&\ge (1 - (3\alpha_1 + 4\alpha_2)).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\frac{1}{(1+\alpha_2)(1+\alpha_1)^2} \left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right) - (1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Sigma_2)} \\
\ge& (1 - (3\alpha_1 + 4\alpha_2)) \left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right) - (1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Sigma_2)} \\
\ge& 1 - (3\alpha_1 + 4\alpha_2) - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}} - (1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Sigma_2)}
\end{aligned}$$

$$\geq 1-\varepsilon',$$

where

$$\varepsilon' = 8\sqrt{\frac{\operatorname{rank}(\Sigma_u)}{n}} + 14\sqrt{\frac{\log(32/\delta)}{n}} + 4\sqrt{\frac{\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}} + (1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Xi_z)}.$$

Finally, we derive the upper bound of the ridgeless estimator. If $\varepsilon' \leq 1/2$, because $(1-x)^{-1} \leq 1+2x$ for $x \in [0,1/2]$, it holds that

$$s^2 \leq \widetilde{\sigma}^2 \frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}\frac{1}{1-\varepsilon'} \leq \widetilde{\sigma}^2\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}(1+2\varepsilon').$$

Then, it holds from (34) that

$$\phi^2 \leq \|\Sigma_u^+\omega\|^2 + s^2\|P_z v^*\|^2 \leq \|\Sigma_u^+\omega\|^2 + s^2(1+\varepsilon_2).$$

Therefore, we have

$$\phi^2 \leq \|\Sigma_u^+\omega\|^2 + (1+\varepsilon)\widetilde{\sigma}^2\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2},$$

with $\varepsilon = 2\varepsilon' + 2\varepsilon_2$. □

We can now derive the general norm bound in the case where the covariates correlate with errors.

Theorem 20 (General norm bound)**.** *There exists an absolute constant $C_2 \leq 56$ such that the following is true. Under Assumptions 1 and 2 with covariance split $\Sigma_x = \Xi_z + \Sigma_u$, let $\|\cdot\|$ be an arbitrary norm, and fix $\delta \leq 1/4$. Denote the ℓ_2 orthogonal projection matrix onto the space spanned by Ξ_z, Σ_u as P_z, P_u, respectively. Let H be normally distributed with mean zero and variance I_d, that is, $H \sim N(0, I_d)$. Denote v_* as $\arg\min_{v\in\partial\|\Xi_z^{1/2}H\|_*}\|v\|_{\Xi_z}$. Suppose that there exist $\varepsilon_1, \varepsilon_2 \geq 0$ such that with probability at least $1-\delta/4$*

$$\|v^*\|_{\Xi_z} \leq (1+\varepsilon_1)E\|v^*\|_{\Xi_z}$$

and

$$\|Pv^*\|^2 \leq 1+\varepsilon_2.$$

Let ε denote $C_2\left(\sqrt{\frac{\operatorname{rank}(\Sigma_u)}{n}} + \sqrt{\frac{\log(1/\delta)}{r_{\|\cdot\|}(\Xi_z)}} + \sqrt{\frac{\log(1/\delta)}{n}} + (1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Xi_z)} + \varepsilon_2\right)$. Then, if n and the effective ranks are large enough that $\varepsilon \leq 1$, with probability at least $1-\delta$, it holds that

$$\|\widehat{\theta}\| \leq \|\theta_0\| + \|\Sigma_u^+\omega\| + (1+\varepsilon)^{1/2}\widetilde{\sigma}\frac{\sqrt{n}}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)}.$$

Proof of Theorem 20. For any $t > 0$, it holds from Lemmas 17 and 18 that

$$\mathbb{P}(\|\widehat{\theta}\| > t) \leq \mathbb{P}(\Phi > t - \|\theta_0\|) \leq 2\mathbb{P}(\phi \geq t - \|\theta_0\|).$$

Lemma 19 implies that the above term is upper bounded by δ if we choose $t - \|\theta_0\|$ using the result (35) with δ replaced by $\delta/2$. We obtain the stated result by moving $\|\theta_0\|$ to the other side. □

When we consider the Euclidean space, we can reduce the upper bound of the ridgeless estimator to a simpler bound.

Theorem 21 (Euclidean norm bound; special case of Theorem 20)**.** *Fix any $\delta \leq 1/4$. Under Assumptions 1 and 2 with covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$, there exists some $\varepsilon \lesssim \sqrt{\frac{\operatorname{rank}(\Sigma_u)}{n}} + \sqrt{\frac{\log(1/\delta)}{r(\Xi_z)}} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{n\log(1/\delta)}{R(\Xi_z)}$ such that the following is true. If n and the effective ranks are such that $\varepsilon \leq 1$ and $R(\Xi_z) \gtrsim \log(1/\delta)^2$, then with probability at least $1-\delta$, it holds that*

$$\|\widehat{\theta}\|_2 \leq \|\theta_0\|_2 + \|\Sigma_u^+\omega\|_2 + (1+\varepsilon)^{1/2}\widetilde{\sigma}\sqrt{\frac{n}{\operatorname{tr}(\Xi_z)}}, \tag{39}$$

where $\widetilde{\sigma}^2 := \sigma^2 - \omega^T\Sigma_u^+\omega$.

Proof of Theorem 21. Throughout this proof, we simplify the upper bound, especially $n/R_{\|\cdot\|_2}(\Xi_z)$ and $1/r_{\|\cdot\|_2}(\Xi_z)$, in Theorem 20. By the definition of the dual norm and $\partial\|\Xi_x^{1/2}H\|_*$ with Euclidean norm, v^* is equal to $\Xi_z^{1/2}H/\|\Xi_z^{1/2}H\|_2$. Hence, $\|v^*\|_{\Xi_z}$ is $\|\Xi_z H\|_2/\|\Xi_z^{1/2}H\|_2$. From the result of (94), for some constant $c > 0$, we can choose ε_1 such that

$$(1+\varepsilon_1)E\|v^*\|_{\Xi_z} = c\sqrt{\log(16/\delta)\frac{\operatorname{tr}(\Xi_z^2)}{\operatorname{tr}(\Xi_z)}}.$$

If we assume effective rank is sufficiently large, (91) provides that $\left(E\|\Xi_z^{1/2}H\|_2\right)^2 \gtrsim \operatorname{tr}(\Xi_z)$. Therefore, we have

$$(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|_2}(\Xi_z)} = n\frac{(1+\varepsilon_1)^2(E\|v^*\|_{\Xi_z})^2}{\left(E\|\Xi_z^{1/2}H\|_2\right)^2} \lesssim n\log(16/\delta)\frac{\operatorname{tr}(\Xi_z^2)}{\operatorname{tr}(\Xi_z)^2} = \frac{n\log(16/\delta)}{R(\Xi_z)}.$$

Moreover, because P_z is an l_2 projection matrix, let ε_2 be zero. Then, it holds from (92) of Lemma 41 that

$$\varepsilon \lesssim \sqrt{\frac{\operatorname{rank}(\Sigma_u)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{r(\Xi_z)}} + \frac{n\log(1/\delta)}{R(\Xi_z)}.$$

By using the inequality $(1-x)^{-1} \leq 1+2x$ for $x \in [0, 1/2]$ and (91) of Lemma 41, we finally obtain

$$(1+\varepsilon)^{1/2}\widetilde{\sigma}\frac{\sqrt{n}}{E\|\Xi_z^{1/2}H\|_2} \leq (1+\varepsilon)^{1/2}\left(1-\frac{1}{r(\Xi_z)}\right)^{-1/2}\widetilde{\sigma}\frac{\sqrt{n}}{\operatorname{tr}(\Xi_z)}$$

$$\leq (1+\varepsilon)^{1/2}\left(1+\frac{2}{r(\Xi_z)}\right)^{1/2}\widetilde{\sigma}\frac{\sqrt{n}}{\mathrm{tr}(\Xi_z)}$$
$$\leq \left(1+2\varepsilon+\frac{2}{r(\Xi_z)}\right)^{1/2}\widetilde{\sigma}\frac{\sqrt{n}}{\mathrm{tr}(\Xi_z)},$$

with ε replaced by

$$\varepsilon' = 2\varepsilon + \frac{2}{r(\Xi_z)} \lesssim \sqrt{\frac{\mathrm{rank}(\Sigma_u)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{r(\Xi_z)}} + \frac{n\log(1/\delta)}{R(\Xi_z)}.$$

□

Appendix E: Benign overfitting

In this section, we state the primary result on the conditions of benign overfitting by combining the results from the two previous sections. First, we derive the result with an arbitrary norm.

Theorem 22 (Benign Overfitting)**.** *Fix any $\delta \leq 1/2$. Under Assumptions 1 and 2 with covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$, let γ and ε be as defined in Corollary 16 and Theorem 20. Suppose that n and the effective ranks are such that $R(\Xi_z) \gtrsim \log(1/\delta)^2$ and $\gamma, \varepsilon \leq 1$. Then, with probability at least $1-\delta$, it holds that*

$$\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 \leq (1+\gamma)(1+\varepsilon)\left(\widetilde{\sigma} + (\|\Sigma_u^{+}\omega\| + \|\theta_0\|)\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}}\right)^2 - \widetilde{\sigma}^2.$$

Proof of Theorem 22. From the result of Theorem 20, if we adopt

$$B = \|\theta_0\| + \|\Sigma_u^{+}\omega\| + (1+\varepsilon)^{1/2}\widetilde{\sigma}\frac{\sqrt{n}}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)},$$

then $\{\theta : \|\theta\| \leq B\} \cap \{\theta : \mathbf{X}\theta = \mathbf{Y}\}$ is not empty with high probability. Clearly, $B > \|\theta_0\|$. This intersection necessarily includes the ridgeless estimator $\widehat{\theta}$. Therefore, it holds from Corollary 16 that

$$\begin{aligned}
\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 \leq & \max_{\|\theta\|\leq B, \mathbf{Y}=\mathbf{X}\theta} \|\theta-\theta_0\|_{\Xi_z}^2 \\
\leq & (1+\gamma)\frac{B^2(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}{n} - \widetilde{\sigma}^2 \\
= & (1+\gamma)\left((\|\theta_0\| + \|\Sigma_u^{+}\omega\|)\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}} + (1+\varepsilon)^{1/2}\widetilde{\sigma}\right)^2 - \widetilde{\sigma}^2 \\
\leq & (1+\gamma)(1+\varepsilon)\left(\widetilde{\sigma} + (\|\theta_0\| + \|\Sigma_u^{+}\omega\|)\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}}\right)^2 - \widetilde{\sigma}^2.
\end{aligned}$$

Then, we obtain the statement. □

Theorem 8 (Sufficient conditions) *Under Assumptions 1 and 2, let $\widehat{\theta}$ be the ridgeless estimator. Let $\|\cdot\|$ denote an arbitrary norm. Suppose that as n goes to ∞, the covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$ satisfies the following conditions:*

(i) (Small large-variance dimension.)

$$\lim_{n\to\infty} \frac{\mathrm{rank}(\Sigma_u)}{n} = 0.$$

(ii) (Large effective dimension.)

$$\lim_{n\to\infty} \frac{1}{r_{\|\cdot\|}(\Xi_z)} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{n}{R_{\|\cdot\|}(\Xi_z)} = 0.$$

(iii) (No aliasing condition.)

$$\lim_{n\to\infty} \frac{\|\theta_0\| \mathbb{E}\|\Xi_z^{1/2} H\|_*}{\sqrt{n}} = 0.$$

(iv) (Contracting ℓ_2 projection condition.) For any $\eta > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\|P_u v^*\|^2 > 1 + \eta) = 0.$$

(v) (Condition for the minimal interpolation of instrumental variable)

$$\lim_{n\to\infty} \frac{\|\Sigma_u^+ \omega\| \mathbb{E}\|\Xi_z^{1/2} H\|_*}{\sqrt{n}} = 0.$$

Then, $\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2$ converges to 0 in probability.

Proof of Theorem 8. We take advantage of the upper bound on the projected RMSE derived in Theorem 22. To begin with, we reorganize the upper bound in Theorem 22 to elucidate the terms that should be sufficiently small for the projected RMSE to converge. Fix any $\eta > 0$. By trivial calculation, we have

$$\begin{aligned}
&(1+\gamma)(1+\varepsilon)\left(\widetilde{\sigma} + (\|\theta_0\| + \|\Sigma_u^+\omega\|)\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}}\right)^2 - \widetilde{\sigma}^2 \\
&= (1+\gamma)(1+\varepsilon)\Bigg(\left((\|\theta_0\| + \|\Sigma_u^+\omega\|)\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}}\right)^2 \\
&\qquad\qquad + 2\widetilde{\sigma}(\|\theta_0\| + \|\Sigma_u^+\omega\|)\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}}\Bigg) \\
&\quad + (\gamma + \varepsilon + \gamma\varepsilon)\widetilde{\sigma}^2 \\
&\le (1+\gamma)(1+\varepsilon)\Bigg(\left((\|\theta_0\| + \|\Sigma_u^+\omega\|)\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_*}{\sqrt{n}}\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + 2\sigma(\|\theta_0\| + \|\Sigma_u^+ \omega\|) \frac{\mathbb{E}\|\Xi_z^{1/2} H\|_*}{\sqrt{n}} \Bigg) \\
&\quad + (\gamma + \varepsilon + \gamma\varepsilon)\sigma^2 \\
&\leq \eta. \qquad (40)
\end{aligned}
$$

The second to last inequality follows $\widetilde{\sigma}^2 = \sigma^2 - \|\omega\|^2_{\Sigma_u^+} \leq \sigma^2$, and the last inequality holds by selecting sufficiently small γ, ε, and $(\|\theta_0\| + \|\Sigma_u^+ \omega\|)(\mathbb{E}\|\Xi_z^{1/2} H\|_* / \sqrt{n})$.

Fix any $\delta > 0$. Conditions (i) and (ii) in Theorem 8 make γ sufficiently small for large enough n. $(\|\theta_0\| + \|\Sigma_u^+ \omega\|)(\mathbb{E}\|\Xi_z^{1/2} H\|_* / \sqrt{n})$ goes to zero from conditions (iii) and (v). From condition (iv) of Theorem 8, ε_2 in ε can also be arbitrarily small.

Finally, we need to specify the conditions when ε can be sufficiently small. By the definition of $R_{\|\cdot\|(\Xi_z)}$, we have

$$
\sqrt{\frac{n}{R_{\|\cdot\|}(\Xi_z)}} = \mathbb{E}\left[\frac{\|v^*\|_{\Xi_z}}{\mathbb{E}\|\Xi_z^{1/2} H\|_* / \sqrt{n}} \right].
$$

It holds from the Markov inequality that for any $\eta' > 0$,

$$
\begin{aligned}
\mathbb{P}\left(\frac{\|v^*\|_{\Xi_z}}{\mathbb{E}\|\Xi_z^{1/2} H\|_* / \sqrt{n}} > \sqrt{\eta'} \right) &\leq \frac{1}{\sqrt{\eta'}} \mathbb{E}\left[\frac{\|v^*\|_{\Xi_z}}{\mathbb{E}\|\Xi_z^{1/2} H\|_* / \sqrt{n}} \right] \\
&= \frac{1}{\sqrt{\eta'}} \sqrt{\frac{n}{R_{\|\cdot\|}(\Xi_z)}}. \qquad (41)
\end{aligned}
$$

As $n / R_{\|\cdot\|}(\Xi_z)$ converges to zero in its limit, the left-hand side of (41) can be arbitrarily small. Hence, we can pick up ε_1 such that

$$
(1 + \varepsilon_1)\mathbb{E}\|v^*\|_{\Xi_z} = \sqrt{\eta'} \frac{\mathbb{E}\|\Xi_z^{1/2} H\|_*}{\sqrt{n}},
$$

which implies that

$$
(1 + \varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Xi_z)} = \frac{n}{\left(\mathbb{E}\|\Xi_z^{1/2} H\|_*\right)^2} ((1 + \varepsilon_1)\mathbb{E}\|v^*\|_{\Xi_z})^2 = \eta'.
$$

We have shown that γ, ε, and $(\|\theta_0\| + \|\Sigma_u^+ \omega\|)(\mathbb{E}\|\Xi_z^{1/2} H\|_* / \sqrt{n})$ are so small that (40) holds for sufficiently large n. Therefore, we obtain

$$
\mathbb{P}(\|\widehat{\theta} - \theta_0\|^2_{\Xi_z} > \eta) \leq \delta
$$

for any fixed η. As η and δ are arbitrary, we have for any η,

$$
\lim_{n \to \infty} \mathbb{P}(\|\widehat{\theta} - \theta_0\|^2_{\Xi_z} > \eta) = 0.
$$

Then, we obtain the statement. □

Second, we establish sufficient conditions of benign overfitting with the Euclidean norm.

Theorem 1 (Benign Overfitting) *Fix any $\delta \leq 1/2$. Under Assumptions 1 and 2 with covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$, let γ and ε be as defined in Corollary 10 and Theorem 21. Suppose that n and the effective ranks are such that $R(\Xi_z) \gtrsim \log(1/\delta)^2$ and $\gamma, \varepsilon \leq 1$. Then, with probability at least $1-\delta$,*

$$\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 \leq (1+\gamma)(1+\varepsilon)\left(\widetilde{\sigma} + (\|\Sigma_u^+\omega\|_2 + \|\theta_0\|_2)\sqrt{\frac{\text{tr}(\Xi_z)}{n}}\right)^2 - \widetilde{\sigma}^2.$$

Proof of Theorem 1. From the result of Theorem 21, if we adopt

$$B = \|\theta_0\|_2 + \|\Sigma_u^+\omega\|_2 + (1+\varepsilon)^{1/2}\widetilde{\sigma}\sqrt{\frac{n}{\text{tr}(\Xi_z)}},$$

then $\{\theta : \|\theta\|_2 \leq B\} \cap \{\theta : \mathbf{X}\theta = \mathbf{Y}\}$ is not empty with high probability. Clearly, $B > \|\theta_0\|_2$. This intersection necessarily includes the ridgeless estimator $\widehat{\theta}$. Therefore, it holds from Corollary 10 that

$$\begin{aligned}
\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 \leq & \max_{\|\theta\|_2 \leq B, \mathbf{Y}=\mathbf{X}\theta} \|\theta - \theta_0\|_{\Xi_z}^2 \\
\leq & (1+\gamma)\frac{B^2\text{tr}(\Xi_z)}{n} - \widetilde{\sigma}^2 \\
= & (1+\gamma)\left((\|\theta_0\|_2 + \|\Sigma_u^+\omega\|_2)\sqrt{\frac{\text{tr}(\Xi_z)}{n}} + (1+\varepsilon)^{1/2}\widetilde{\sigma}\right)^2 - \widetilde{\sigma}^2 \\
\leq & (1+\gamma)(1+\varepsilon)\left(\widetilde{\sigma} + (\|\theta_0\|_2 + \|\Sigma_u^+\omega\|_2)\sqrt{\frac{\text{tr}(\Xi_z)}{n}}\right)^2 - \widetilde{\sigma}^2.
\end{aligned}$$

□

Theorem 2 (Sufficient conditions) *Under Assumptions 1 and 2, let $\widehat{\theta}$ be the ridgeless estimator. Suppose that as n goes to ∞, the covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$ satisfies the following conditions:*

(i) (Small large-variance dimension.)

$$\lim_{n\to\infty} \frac{\text{rank}(\Sigma_u)}{n} = 0.$$

(ii) (Large effective dimension.)

$$\lim_{n\to\infty} \frac{n}{R(\Xi_z)} = 0.$$

(iii) (No aliasing condition.)

$$\lim_{n\to\infty} \|\theta_0\|_2\sqrt{\frac{\text{tr}(\Xi_z)}{n}} = 0.$$

(iv) (Condition for the minimal interpolation of instrumental variable)

$$\lim_{n\to\infty} \|\Sigma_u^+ \omega\|_2 \sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}} = 0.$$

Then, $\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2$ *converges to* 0 *in probability.*

Proof of Theorem 2. As in the proof of Theorem 8, we rearrange the upper bound derived in Theorem 1 to clarify which terms should be sufficiently small for the projected RMSE to converge. By trivial calculation, we have

$$\begin{aligned}
&(1+\gamma)(1+\varepsilon)\left(\widetilde{\sigma} + (\|\theta_0\|_2 + \|\Sigma_u^+\omega\|_2)\sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}}\right)^2 - \widetilde{\sigma}^2 \\
=&(1+\gamma)(1+\varepsilon)\left(\left((\|\theta_0\|_2+\|\Sigma_u^+\omega\|_2)\sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}}\right)^2 + 2\widetilde{\sigma}(\|\theta_0\|_2+\|\Sigma_u^+\omega\|_2)\sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}}\right) \\
&+ (\gamma+\varepsilon+\gamma\varepsilon)\widetilde{\sigma}^2 \\
\leq&(1+\gamma)(1+\varepsilon)\left(\left((\|\theta_0\|_2+\|\Sigma_u^+\omega\|_2)\sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}}\right)^2 + 2\sigma(\|\theta_0\|_2+\|\Sigma_u^+\omega\|_2)\sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}}\right) \\
&+ (\gamma+\varepsilon+\gamma\varepsilon)\sigma^2. \qquad (42)
\end{aligned}$$

The inequality follows $\widetilde{\sigma}^2 \leq \sigma^2$ as in the proof of Theorem 8.

Fix any $\eta > 0$ and $\delta > 0$. From Lemma 5 of Bartlett *et al.* (2020), it holds that $R(\Xi_z) \leq r(\Xi_z)^2$. If $R(\Xi_z) = \upsilon(n)$ holds as the second condition in Theorem 2, we have $r(\Xi_z) = \upsilon(\sqrt{n}) = \upsilon(1)$, which implies the convergence of $1/r(\Xi_z)$ to zero. Hence, conditions (i) and (ii) in Theorem 1 make γ and ε sufficiently small for large enough n. $(\|\theta_0\|_2 + \|\Sigma_u^+\omega\|_2)\sqrt{\operatorname{tr}(\Xi_z)/n}$ goes to zero from conditions (iii) and (iv). Hence, for sufficiently large n, we obtain that (42) is no more than η. Therefore, we obtain

$$\mathbb{P}(\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 > \eta) \leq \delta,$$

for any fixed η. As η and δ are arbitrary, we have for any η,

$$\lim_{n\to\infty} \mathbb{P}(\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 > \eta) = 0.$$

□

Appendix F: Non-orthogonal case

We present the proof of the non-orthogonal case independently in this section because the case requires additional complicated analysis and is not a simple extension of the orthogonal case.

We present the results in Section 4 for the case where Ξ_z and Σ_u are non-orthogonal. In this section, we use $\Sigma_1 \in \mathbb{R}^{p\times p}$ and $\Sigma_2 \in \mathbb{R}^{p\times p}$ as notations for (potentially non-orthogonal) matrices as the statements in this section can be regarded as generic results for general matrices. In the setting for regression with endogeneity, these notations correspond to Ξ_z and Σ_u, respectively.

First, we introduce auxiliary lemmas for this section.

Lemma 23 (Corollary 2 in Koehler *et al.* (2021))**.** *There exists an absolute constant $C_1 \leq 66$ such that the following is true. Under Assumption 1 with covariance $\Sigma_x = \Sigma_1 + \Sigma_2$, fix $\delta \leq 1/4$ and let $\gamma = C_1(\sqrt{\log(1/\delta)/r(\Sigma_2)} + \sqrt{\log(1/\delta)/n} + \sqrt{\mathrm{rank}(\Sigma_1)/n})$. If $B \geq \|\theta_0\|_2$ and n is large enough that $\gamma \leq 1$, the following holds with probability at least $1-\delta$:*

$$\sup_{\|\theta\|_2 \leq B, \widehat{L}(\theta)=0} L(\theta) \leq (1+\gamma)\frac{B^2 \mathrm{tr}(\Sigma_2)}{n}, \tag{43}$$

where $\widehat{L}(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2/n$ and $L(\theta) := E[(Y_1 - X_1^\top \theta)]^2$.

Lemma 24 (Corollary 4 in Koehler *et al.* (2021))**.** *Suppose $X_1, \cdots, X_n \sim N(0,\Sigma)$ are independent with $\Sigma : p \times p$ a positive semidefinite matrix, $t > 0$, and $n \geq 4(d+t^2)$. Let $\widehat{\Sigma} = \sum_i X_i X_i^\top/n$ be the empirical covariance matrix. Then, with probability at least $1-\delta$,*

$$(1-\varepsilon)\Sigma \preceq \widehat{\Sigma} \preceq (1+\varepsilon)\Sigma,$$

with $\varepsilon = 3\sqrt{d/n} + 3\sqrt{2\log(2/\delta)/n}$.

Lemma 25. *Take any covariance matrix Σ_1, Σ_2. If $\Sigma_1^{1/2}\Sigma_2^{1/2} \neq 0$, it holds that with probability at least $1-\delta$,*

$$1 - \frac{\|\Sigma_1^{1/2}\Sigma_2^{1/2} H\|_2^2}{\mathrm{tr}(\Sigma_1\Sigma_2)} \lesssim \frac{\log(4/\delta)}{\sqrt{R(\Sigma_1\Sigma_2)}}. \tag{44}$$

Moreover, it holds that with probability at least $1-\delta$,

$$\|\Sigma_1^{1/2}\Sigma_2^{1/2} H\|_2^2 \lesssim \log(4/\delta)\mathrm{tr}(\Sigma_1\Sigma_2). \tag{45}$$

Therefore, if $R(\Sigma_2) \gtrsim log(4/\delta)^2$ holds, we have

$$\left(\frac{\|\Sigma_1^{1/2}\Sigma_2^{1/2} H\|_2}{\|\Sigma_2^{1/2} H\|_2}\right)^2 \lesssim \log(4/\delta)\frac{\mathrm{tr}(\Sigma_1\Sigma_2)}{\mathrm{tr}(\Sigma_2)}. \tag{46}$$

Proof of Lemma 25. As $\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}$ is a real symmetric matrix, there exists an orthogonal matrix Q such that $Q\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}Q^\top$ is a diagonal matrix. Further, QH has the normal standard distribution by the definition of H. Therefore, without loss of generality, $\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}$ can be considered as a diagonal matrix that consists of eigenvalues of $\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}$, $\lambda_1, \cdots, \lambda_p$. By the sub-exponential

Bernstein inequality (Vershynin (2018), Theorem 2.8.2), we have with probability at least $1-\delta/2$

$$\begin{aligned}\left|\frac{\|\Sigma_1^{1/2}\Sigma_2^{1/2}H\|_2^2}{\operatorname{tr}(\Sigma_1\Sigma_2)}-1\right| &= \left|\sum_{i=1}^{p}\frac{\lambda_i}{\sum_k \lambda_k}(H_i^2-1)\right| \\ &\lesssim \sqrt{\frac{\log(4/\delta)}{R(\Sigma_1\Sigma_2)}} \vee \frac{\log(4/\delta)}{r(\Sigma_1\Sigma_2)} \le \frac{\log(4/\delta)}{\sqrt{R(\Sigma_1\Sigma_2)}},\end{aligned}$$

where the last inequality follows from the fact $R(\Sigma_1\Sigma_2)\le(r(\Sigma_1\Sigma_2))^2$. By definition, clearly $R(\Sigma_1\Sigma_2)\ge 1$. Therefore, we have

$$\|\Sigma_1^{1/2}\Sigma_2^{1/2}H\|_2^2 \lesssim \log(4/\delta)\operatorname{tr}(\Sigma_1\Sigma_2).$$

Provided $R(\Sigma_2)$ is sufficiently large, we obtain $\|\Sigma_2^{1/2}H\|_2^2 \ge \frac{1}{2}\operatorname{tr}(\Sigma_2)$. Therefore, it holds that

$$\left(\frac{\|\Sigma_1^{1/2}\Sigma_2^{1/2}H\|_2}{\|\Sigma_2^{1/2}H\|_2}\right)^2 \lesssim \log(4/\delta)\frac{\operatorname{tr}(\Sigma_1\Sigma_2)}{\operatorname{tr}(\Sigma_2)}.$$

□

F.1. When X_i and ξ_i are independent

Lemma 26. *Denote P as the projection matrix onto the space spanned by Σ_2. Let v_* denote $\arg\min_{v\in\partial\|\Sigma_2^{1/2}H\|_*}\|v\|_{\Sigma_2}$. Assume that there exist $\varepsilon_1,\varepsilon_2$ and $\varepsilon_3\ge 0$ such that with probability at least $1-\delta/4$,*

$$\|v^*\|_{\Sigma_2} \le (1+\varepsilon_1)\mathbb{E}\|v^*\|_{\Sigma_2}, \tag{47}$$

$$\|Pv^*\| \le 1+\varepsilon_2, \tag{48}$$

and

$$\|\Sigma_1^{1/2}Pv^*\|_2 \le (1+\varepsilon_3)\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2. \tag{49}$$

Define ε as

$$\begin{aligned}\varepsilon :=\, & 84\sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}} + 156\sqrt{\frac{\log(32/\delta)}{n}} + 8\sqrt{\frac{\log(8/\delta)}{r_{\|\cdot\|}(\Sigma_2)}} + 2(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)} \\ & + 2(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2)^2,\end{aligned}$$

where $r_{\|\cdot\|}(\Sigma)$ and $R_{\|\cdot\|}(\Sigma)$ are the effective ranks with general norms as provided in Definition 4. If n and the effective ranks are sufficiently large such that $\varepsilon\le 1$, then with probability at least $1-\delta$, it holds that

$$\phi^2 \le (1+\varepsilon)\sigma^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2} \tag{50}$$

Proof of Lemma 26. Denote α_1, α_2, and α_3 as follows:

$$\begin{aligned}\alpha_1 &:= 2\sqrt{\frac{\log(32/\delta)}{n}},\\ \alpha_2 &:= 3\sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}} + 3\sqrt{\frac{2\log(16/\delta)}{n}},\\ \alpha_3 &:= \sqrt{\frac{\operatorname{rank}(\Sigma_1)+1}{n}} + 2\sqrt{\frac{\log(16/\delta)}{n}}.\end{aligned}$$

To prepare for the derivation of the upper bound, we consider a list of the following inequalities, and each of these holds with probability at least $1-\delta/8$.

(i) By (90) in Lemma 39, uniformly over all $\theta_2 \in \Sigma_1^{1/2}(\mathbb{R}^p)$, it holds that

$$|\langle \xi - \mathbf{W}_2\theta_2, G\rangle| \le \|\xi - \mathbf{W}_2\theta_2\|_2 \|G\|_2 \alpha_3 \tag{51}$$

and

$$|\langle \xi, \mathbf{W}_2\theta_2\rangle| \le \|\xi\|_2 \|\mathbf{W}_2\theta_2\|_2 \alpha_3. \tag{52}$$

For (51), V and s in Lemma 39 correspond to G and $\xi - \mathbf{W}_2\theta_2$, respectively. For (52), V and s in Lemma 39 correspond to ξ and $\mathbf{W}_2\theta_2$, respectively. δ is replaced by $\delta/8$.

(ii) By Lemma 24, uniformly over all $\theta_2 \in \Sigma_1^{1/2}(\mathbb{R}^p)$, it holds that

$$(1-\alpha_2)\|\theta_2\|_2^2 \le \frac{\|\mathbf{W}_2\theta_2\|_2^2}{n} \le (1+\alpha_2)\|\theta_2\|_2^2. \tag{53}$$

Σ and d in Lemma 24 correspond to Σ_1 and $\operatorname{rank}(\Sigma_1)$ in (53), respectively.

(iii) By Lemma 40, it holds that

$$-\alpha_1 \le \frac{1}{\sqrt{n}}\|G\|_2 - 1 \le \alpha_1 \tag{54}$$

and

$$-\alpha_1\sigma \le \frac{1}{\sqrt{n}}\|\xi\|_2 - \sigma \le \alpha_1\sigma. \tag{55}$$

(iv) By Theorem 43, it holds that

$$\begin{aligned}\|\Sigma_2^{1/2}H\|_* &\ge \mathbb{E}\|\Sigma_2^{1/2}H\|_* - \sup_{\|u\|\le 1}\|u\|_{\Sigma_2}\sqrt{2\log(8/\delta)}\\ &= \left(1 - \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\right)\mathbb{E}\|\Sigma_2^{1/2}H\|_*\end{aligned} \tag{56}$$

because $\|\Sigma_2^{1/2}H\|_*$ is a $\sup_{\|u\|\le 1}\|u\|_{\Sigma_2}$-Lipschitz continuous function of H.

We construct the upper bound from the restriction of the optimization problem (32). It holds from (51), (52), and the AM-GM inequality that

$$\begin{aligned}\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2 &\le (1+\alpha_3)(\|\xi - \mathbf{W}_2\theta_2\|_2^2 + \|\theta_1\|_2^2\|G\|_2^2)\\ &\le (1+\alpha_3)((1+\alpha_3)(\|\xi\|^2 + \|\mathbf{W}_2\theta_2\|_2^2) + \|\theta_1\|_2^2\|G\|_2^2).\end{aligned}$$

From the results of (53) (54), and (55), we obtain $\|\mathbf{W}_2\theta_2\|_2^2 \le n(1+\alpha_2)\|\theta_2\|_2^2$, $\|\xi\|_2^2 \le (1+\alpha_1)^2 n\sigma^2$, and $\|G\|^2 \le (1+\alpha_1)^2 n$. Therefore, we have

$$\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2 \le n(1+\alpha_1)^2(1+\alpha_2)(1+\alpha_3)^2(\sigma^2 + \|\theta_2\|_2^2 + \|\theta_1\|_2^2). \quad (57)$$

To consider an upper bound of the ridgeless estimator, we need to choose a suitable θ which satisfies the restriction of the auxiliary problem. We define $\theta := s(Pv^*)$. Then, we have $\theta_1 = s\Sigma_2^{1/2}v^*$ and $\theta_2 = s\Sigma_1^{1/2}Pv^*$. If we consider the value of s that satisfies the inequality:

$$n(1+\alpha_1)^2(1+\alpha_2)(1+\alpha_3)^2(\sigma^2 + \|s\Sigma_1^{1/2}Pv^*\|_2^2 + \|s\Sigma_2^{1/2}v^*\|_2^2) \le s^2\langle H, \Sigma_2^{1/2}v^*\rangle^2,$$

it holds from (57) that

$$\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2 \le (\langle H, \theta_1\rangle)^2.$$

As $\langle H, \theta_1\rangle \ge 0$ by the definition of θ_1, $\theta = s(Pv^*)$ satisfies the restriction of the auxiliary problem in Lemma 18. Solving for s, we can select

$$s^2 = \sigma^2\Bigg(\underbrace{\frac{\langle H, \Sigma_2^{1/2}v^*\rangle^2}{n(1+\alpha_1)^2(1+\alpha_2)(1+\alpha_3)^2} - \|\Sigma_1^{1/2}Pv^*\|_2^2 - \|\Sigma_2^{1/2}v^*\|_2^2}_{=:\Upsilon}\Bigg)^{-1},$$

under the condition that Υ is positive. We derive a lower bound of Υ as in the proof of Lemma 19:

$$\begin{aligned}\Upsilon &= \frac{\|\Sigma_2^{1/2}H\|_*^2}{(1+\alpha_1)^2(1+\alpha_2)(1+\alpha_3)^2 n} - \|v^*\|_{\Sigma_2}^2 - \|\Sigma_1^{1/2}Pv^*\|_2^2\\ &\ge \frac{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}{n}\Bigg(\frac{1}{(1+\alpha_1)^2(1+\alpha_2)(1+\alpha_3)^2}\left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\right)\\ &\qquad -(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)} - \frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}\|\Sigma_1^{1/2}Pv^*\|_2^2\Bigg)\\ &\ge \frac{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}{n}\Bigg(\frac{1}{(1+\alpha_1)^2(1+\alpha_2)(1+\alpha_3)^2}\left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\right)\\ &\qquad -(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)} - (1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2)^2\Bigg).\end{aligned}$$

The equality holds by the definition of $R_{\|\cdot\|}(\Sigma_2)$, the first inequality holds from (47) and (56), and the second inequality follows $\|\Sigma_1^{1/2}Pv^*\|_2 \leq (1+\varepsilon_3)\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2$ by Assumption (49).

We linearize the terms including α_1, α_2, and α_3 to simplify the upper bound. If $\alpha_1 < 1$, $\alpha_2 < 1$, and $\alpha_3 < 1$, it holds that

$$\begin{aligned}(1+\alpha_1)^2(1+\alpha_2)(1+\alpha_3)^2 &= (1+2\alpha_1+\alpha_1^2)(1+\alpha_2)(1+2\alpha_3+\alpha_3^2)\\ &\leq (1+3\alpha_1)(1+\alpha_2)(1+3\alpha_3)\\ &= (1+3\alpha_1+\alpha_2+3\alpha_1\alpha_2)(1+3\alpha_3)\\ &\leq (1+3\alpha_1+4\alpha_2)(1+3\alpha_3)\\ &\leq (1+12\alpha_1+4\alpha_2+15\alpha_3).\end{aligned}$$

As $(1+x)^{-1} \geq (1-x)$ holds for any x, we have

$$\begin{aligned}&\left(\frac{1}{(1+\alpha_1)^2(1+\alpha_2)(1+\alpha_3)^2}\left(1-2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\right)\right.\\ &\qquad\left.-(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)}-(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2)^2\right)\\ &\geq 1-(12\alpha_1+4\alpha_2+15\alpha_3)-2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\\ &\qquad-(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)}-(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2)^2\\ &\geq 1-\varepsilon'\end{aligned}$$

where

$$\begin{aligned}\varepsilon' = 42\sqrt{\frac{\mathrm{rank}(\Sigma_1)}{n}}+78\sqrt{\frac{\log(32/\delta)}{n}}+4\sqrt{\frac{\log(8/\delta)}{r_{\|\cdot\|}(\Sigma_2)}}\\ +(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)}+(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2)^2\end{aligned}$$

If $\varepsilon' \leq 1/2$, because $(1-x)^{-1} \leq 1+2x$ for $x \in [0,1/2]$, it holds that

$$s^2 = \sigma^2\Upsilon^{-1} \leq (1-\varepsilon')^{-1}\sigma^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2} \leq (1+2\varepsilon')\sigma^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}.$$

Therefore, we have

$$\phi^2 \leq s^2\|Pv^*\|^2 \leq (1+\varepsilon_2)(1+2\varepsilon')\sigma^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2} \leq (1+\varepsilon)\sigma^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2}$$

with $\varepsilon = 2\varepsilon' + 2\varepsilon_2$. □

Theorem 27 (General norm bound)**.** *There exists an absolute constant $C_2 \leq 312$ such that the following is true. Under Assumption 1 with covariance split $\Sigma_x = \Sigma_1 + \Sigma_2$, let $\|\cdot\|$ be an arbitrary norm, and fix $\delta \leq 1/4$. Denote the ℓ_2 orthogonal projection matrix onto the space spanned by Σ_2 as P. Let H be normally distributed with mean zero and variance I_d, that is, $H \sim N(0, I_d)$. Denote v_* as $\arg\min_{v \in \partial \|\Sigma_2^{1/2} H\|_*} \|v\|_{\Sigma_2}$. Assume that there exists $\varepsilon_1, \varepsilon_2$ and $\varepsilon_3 \geq 0$ such that with probability at least $1 - \delta/8$,*

$$\|v^*\|_{\Sigma_2} \leq (1 + \varepsilon_1)\mathbb{E}\|v^*\|_{\Sigma_2},$$
$$\|Pv^*\| \leq 1 + \varepsilon_2,$$

and

$$\|\Sigma_1^{1/2} P v^*\|_2 \leq (1 + \varepsilon_3)\mathbb{E}\|\Sigma_1^{1/2} P v^*\|_2.$$

Define ε as

$$\varepsilon := C_2 \left(\sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}} + \sqrt{\frac{\log(1/\delta)}{r_{\|\cdot\|}(\Sigma_2)}} + \sqrt{\frac{\log(1/\delta)}{n}} + (1 + \varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Sigma_2)} \right.$$
$$\left. + (1 + \varepsilon_3)^2 \frac{n}{(\mathbb{E}\|\Sigma_2^{1/2} H\|_*)^2} (\mathbb{E}\|\Sigma_1^{1/2} P v^*\|_2)^2 + \varepsilon_2 \right).$$

If n and the effective ranks are sufficiently large such that $\varepsilon \leq 1$, then with probability at least $1 - \delta$, it holds that

$$\|\widehat{\theta}\| \leq \|\theta_0\| + (1 + \varepsilon)^{1/2} \sigma \frac{\sqrt{n}}{\mathbb{E}\|\Sigma_2^{1/2} H\|_*}.$$

Proof of Theorem 27. For any $t > 0$, it holds from Lemmas 17 and 18 that

$$\mathbb{P}(\|\widehat{\theta}\| > t) \leq \mathbb{P}(\Phi > t - \|\theta_0\|) \leq 2\mathbb{P}(\phi \geq t - \|\theta_0\|).$$

Lemma 26 implies that the above term is upper bounded by δ if we choose $t - \|\theta_0\|$ using the result (50) with δ replaced by $\delta/2$. We obtain the stated result by moving $\|\theta_0\|$ to the other side. □

When we consider the Euclidean space, we can reduce the upper bound of the ridgeless estimator to a simpler bound.

Theorem 28 (Euclidean norm bound; special case of Theorem 27)**.** *Fix any $\delta \leq 1/4$. Under Assumption 1 with covariance $\Sigma_x = \Sigma_1 + \Sigma_2$, there exists some $\varepsilon \lesssim \sqrt{\operatorname{rank}(\Sigma_1)/n} + \sqrt{\log(1/\delta)/r(\Sigma_2)} + \sqrt{\log(1/\delta)/n} + n\log(1/\delta)/R(\Sigma_2)(1 + \operatorname{tr}(\Sigma_1\Sigma_2)/\operatorname{tr}(\Sigma_2^2))$ such that the following is true. If n and the effective ranks are sufficiently large such that $\varepsilon \leq 1$ and $R(\Sigma_2) \gtrsim \log(1/\delta)^2$, then with probability at least $1 - \delta$, it holds that*

$$\|\widehat{\theta}\|_2 \leq \|\theta_0\|_2 + (1 + \varepsilon)^{1/2} \sigma \sqrt{\frac{n}{\operatorname{tr}(\Sigma_2)}}. \tag{58}$$

Proof of Theorem 28. Throughout this proof, we simplify the upper bound, especially $n/R_{\|\cdot\|_2}(\Xi_z)$, $1/r_{\|\cdot\|_2}(\Xi_z)$, and $(n\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2^2)/(\mathbb{E}\|\Sigma_2^{1/2}H\|_*)^2$, in Theorem 27. By the definition of the dual norm and $\partial\|\Xi_x^{1/2}H\|_*$ with Euclidean norm, v^* is equal to $\Sigma_2^{1/2}H/\|\Sigma_2^{1/2}H\|_2$. Hence, $\|v^*\|_{\Xi_z}$ is $\|\Sigma_2 H\|_2/\|\Sigma_2^{1/2}H\|_2$. From the result of (94), for some constant $c_1 > 0$, we can choose ε_1 such that

$$(1+\varepsilon_1)E\|v^*\|_{\Sigma_2} = c_1\sqrt{\log(64/\delta)\frac{\operatorname{tr}(\Sigma_2^2)}{\operatorname{tr}(\Sigma_2)}}.$$

If we assume effective rank is sufficiently large, (91) provides that $\left(E\|\Sigma_2^{1/2}H\|_2\right)^2 \gtrsim \operatorname{tr}(\Sigma_2)$. Therefore, we have

$$\begin{aligned}(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|_2}(\Sigma_2)} &= n\frac{(1+\varepsilon_1)^2(E\|v^*\|_{\Sigma_2})^2}{\left(E\|\Sigma_2^{1/2}H\|_2\right)^2}\\ &\lesssim n\log(64/\delta)\frac{\operatorname{tr}(\Sigma_2^2)}{\operatorname{tr}(\Sigma_2)^2}\\ &= \frac{n\log(64/\delta)}{R(\Sigma_2)}.\end{aligned}$$

It also holds from (46) that for some constant $c_2 > 0$, there exists ε_3 such that

$$(1+\varepsilon_3)\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2 = c_2\sqrt{\log(64/\delta)\frac{\operatorname{tr}(\Sigma_1\Sigma_2)}{\operatorname{tr}(\Sigma_2)}}.$$

By (91), for sufficiently large effective rank, it holds that $(E\|\Sigma_2^{1/2}H\|_2)^2 \gtrsim \operatorname{tr}(\Sigma_2)$. Therefore, we have

$$\begin{aligned}(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Sigma_2^{1/2}H\|_2)^2}(\mathbb{E}\|\Sigma_1^{1/2}Pv^*\|_2)^2 &\lesssim n\log(64/\delta)\frac{\operatorname{tr}(\Sigma_1\Sigma_2)}{\operatorname{tr}(\Sigma_2)^2}\\ &= \frac{n\log(64/\delta)}{R(\Sigma_2)}\frac{\operatorname{tr}(\Sigma_1\Sigma_2)}{\operatorname{tr}(\Sigma_2^2)}.\end{aligned}$$

Finally, we obtain the upper bound of ε. As P is an l_2 projection matrix, let ε_2 be zero. Then, it holds from (92) of Lemma 41 that

$$\varepsilon \lesssim \sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{r(\Sigma_2)}} + \frac{n\log(1/\delta)}{R(\Sigma_2)}\left(1+\frac{\operatorname{tr}(\Sigma_1\Sigma_2)}{\operatorname{tr}(\Sigma_2^2)}\right).$$

By using the inequality $(1-x)^{-1} \le 1+2x$ for $x \in [0,1/2]$ and (91) of Lemma 41, we finally obtain

$$(1+\varepsilon)^{1/2}\sigma\frac{\sqrt{n}}{E\|\Sigma_2^{1/2}H\|_2} \le (1+\varepsilon)^{1/2}\left(1-\frac{1}{r(\Sigma_2)}\right)^{-1/2}\sigma\sqrt{\frac{n}{\operatorname{tr}(\Sigma_2)}}$$

$$\leq (1+\varepsilon)^{1/2}\left(1+\frac{2}{r(\Sigma_2)}\right)^{1/2}\sigma\sqrt{\frac{n}{\operatorname{tr}(\Sigma_2)}}$$
$$\leq \left(1+2\varepsilon+\frac{2}{r(\Sigma_2)}\right)^{1/2}\sigma\sqrt{\frac{n}{\operatorname{tr}(\Sigma_2)}},$$

with ε replaced by

$$\varepsilon' := 2\varepsilon + \frac{2}{r(\Sigma_2)}$$
$$\lesssim \sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{r(\Sigma_2)}} + \frac{n\log(1/\delta)}{R(\Sigma_2)}\left(1+\frac{\operatorname{tr}(\Sigma_1\Sigma_2)}{\operatorname{tr}(\Sigma_2^2)}\right).$$

□

Theorem 29 (Benign Overfitting (Non-orthogonal))**.** *Fix any $\delta \leq 1/2$. Under Assumption 1 with covariance $\Sigma_x = \Sigma_1 + \Sigma_2$, let γ and ε be as defined in Lemma 23 and Theorem 28, respectively. Suppose also that n and the effective ranks are such that $R(\Sigma_2) \gtrsim \log(1/\delta)^2$ and $\gamma, \varepsilon \leq 1$, then, with probability at least $1-\delta$, it holds that*

$$L(\widehat{\theta}) \leq (1+\gamma)(1+\varepsilon)\left(\sigma + \|\theta_0\|_2\sqrt{\frac{\operatorname{tr}(\Sigma_2)}{n}}\right)^2,$$

where we denote $L(\theta)$ as $\mathbb{E}(y - \langle\theta, x\rangle)^2$.

Proof of Theorem 29. From the result of Theorem 28, if we adopt

$$B = \|\theta_0\|_2 + (1+\varepsilon)^{1/2}\sigma\sqrt{\frac{n}{\operatorname{tr}(\Sigma_2)}},$$

then $\{\theta : \|\theta\|_2 \leq B\} \cap \{\theta : \mathbf{X}\theta = \mathbf{Y}\}$ is not empty with high probability. This intersection necessarily contains the ridgeless estimator $\widehat{\theta}$. Clearly, $B > \|\theta_0\|_2$. Therefore, it holds from Lemma 23 that

$$L(\widehat{\theta}) \leq \sup_{\|\theta\|_2 \leq B, \widehat{L}(\theta)=0} L(\theta)$$
$$\leq (1+\gamma)\left(\|\theta_0\|_2 + (1+\varepsilon)^{1/2}\sigma\sqrt{\frac{n}{\operatorname{tr}(\Sigma_2)}}\right)^2\frac{\operatorname{tr}(\Sigma_2)}{n}$$
$$\leq (1+\gamma)(1+\varepsilon)\left(\sigma + \|\theta_0\|_2\sqrt{\frac{\operatorname{tr}(\Sigma_2)}{n}}\right)^2,$$

where we denote $\widehat{L}(\theta)$ as $\|\mathbf{Y} - \mathbf{X}\theta\|_2^2/n$. □

Theorem 6 (Sufficient conditions: Non-Orthogonal Case when X_i and ξ_i are independent) *Under Assumption 1, let $\widehat{\theta}$ be the ridgeless estimator. Suppose also that as n goes to ∞, there exists a sequence of covariance $\Sigma_x = \Sigma_1 + \Sigma_2$ such that the following conditions hold:*

(i) (Small large-variance dimension.)

$$\lim_{n\to\infty} \frac{\operatorname{rank}(\Sigma_1)}{n} = 0.$$

(ii) (Large effective dimension.)

$$\lim_{n\to\infty} \frac{n}{R(\Sigma_2)} = 0.$$

(iii) (No aliasing condition.)

$$\lim_{n\to\infty} \|\theta_0\|_2 \sqrt{\frac{\operatorname{tr}(\Sigma_2)}{n}} = 0.$$

(iv) (The cost of non-orthogonality)

$$\lim_{n\to\infty} \frac{n}{R(\Sigma_2)} \left(\frac{\operatorname{tr}(\Sigma_1 \Sigma_2)}{\operatorname{tr}(\Sigma_2^2)} \right) = 0.$$

Then, $L(\widehat{\theta})$ converges to σ^2 in probability.

Proof of Theorem 6. Fix any $\eta > 0$ and $\delta > 0$. From Lemma 5 of Bartlett *et al.* (2020), it holds that $R(\Sigma_2) \leq r(\Sigma_2)^2$. If $R(\Sigma_2) = \upsilon(n)$ holds as the second condition in Theorem 6, we have $r(\Sigma_2) = \upsilon(\sqrt{n}) = \upsilon(1)$, which implies the convergence of $1/r(\Xi_z)$ to zero. Hence, conditions (i) and (ii) in Theorem 6 make γ sufficiently small for large enough n. Clearly, $\|\theta_0\|_2 \sqrt{\operatorname{tr}(\Sigma_2)/n}$ goes to zero from condition (iii). By the definition of ε, conditions (i) (ii), and (iv) in Theorem 6 imply that ε can be arbitrarily small. Therefore, for sufficiently large n, we obtain

$$(1+\gamma)(1+\varepsilon) \left(\sigma + \|\theta_0\|_2 \sqrt{\frac{\operatorname{tr}(\Sigma_2)}{n}} \right)^2 - \sigma^2 \leq \eta. \tag{59}$$

We have shown that γ, ε, and $\|\theta_0\|_2 \sqrt{\operatorname{tr}(\Sigma_2)/n}$ are so small that equation (59) holds for sufficiently large n. Therefore, we obtain

$$\mathbb{P}(|L(\widehat{\theta}) - \sigma^2| > \eta) \leq \delta$$

for any fixed η. As η and δ are arbitrary, we have for any η,

$$\lim_{n\to\infty} \mathbb{P}(|L(\widehat{\theta}) - \sigma^2| > \eta) = 0.$$

□

F.2. When $\boldsymbol{X_i}$ and $\boldsymbol{\xi_i}$ are dependent

Throughout this subsection, we assume $\widetilde{\sigma}^2 = \sigma^2 - \|\omega\|_{\Sigma_u^+}^2 > 0$ holds.

Lemma 30. *Denote P_z, P_u as the projection matrix onto the space spanned by Ξ_z and Σ_u, respectively. Let $v_* = \arg\min_{v \in \partial \|\Xi_z^{1/2} H\|} \|v\|_{\Xi_z}$. Assume that there exists ε_1, ε_2, and $\varepsilon_3 \geq 0$ such that with probability at least $1 - \delta/4$,*

$$\|v^*\|_{\Xi_z} \leq (1+\varepsilon_1)\mathbb{E}\|v^*\|_{\Xi_z}, \tag{60}$$

$$\|P_z v^*\| \leq 1 + \varepsilon_2, \tag{61}$$

and

$$\|\Sigma_u^{1/2} P_z v^*\|_2 \leq (1+\varepsilon_3)\mathbb{E}\|\Sigma_u^{1/2} P_z v^*\|_2. \tag{62}$$

Denote ε as

$$\begin{aligned}\varepsilon :=& 12\sqrt{\frac{\operatorname{rank}(\Sigma_u)}{n}} + 24\sqrt{\frac{\log(32/\delta)}{n}} + 8\sqrt{\frac{\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}} + 2(1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Xi_z)} \\ &+ 64\frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}} \frac{\mathbb{E}\|\Xi_z^{1/2}H\|_2}{\sqrt{n}}\left(1 + \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right) \\ &+ 2(1+\varepsilon_3)^2 \frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_u^{1/2}P_z v^*\|_2)^2 + 2\varepsilon_2,\end{aligned}$$

where we denote $r_{\|\cdot\|}(\Sigma)$ and $R_{\|\cdot\|}(\Sigma)$ as follows:

$$r_{\|\cdot\|}(\Sigma) = \left(\frac{E\|\Sigma^{1/2}H\|_*}{\sup_{\|u\|\leq 1}\|u\|_\Sigma}\right)^2 \quad \text{and } R_{\|\cdot\|}(\Sigma) = \left(\frac{E\|\Sigma^{1/2}H\|_*}{E\|v^*\|_\Sigma}\right)^2.$$

If n and the effective ranks are sufficiently large such that $\varepsilon \leq 1$, then with probability at least $1-\delta$, the AO defined in (32) is upper bounded as

$$\phi^2 \leq \left(\|\Sigma_u^+\omega\|_2 + (1+\varepsilon)(2\eta_1 + \widetilde{\sigma} + \eta_2)\sqrt{\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}}\right)^2, \tag{63}$$

where we denote η_1 and η_2 as follows:

$$\eta_1 := \sqrt{(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Xi_z)}}\|\Xi_z^{1/2}\Sigma_u^+\omega\|_2,$$

$$\eta_2 := \sqrt{\frac{(\mathbb{E}\|\Xi_z^{1/2}H\|_2)^2}{n}\left(1 + \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right)^2 \|\Sigma_u^+\omega\|_2^2 + \|\Xi_z^{1/2}\Sigma_u^+\omega\|_2^2}.$$

Proof of Lemma 30. This proof has four steps (i) preparation, (ii) introducing a coefficient s, (iii) deriving a bound on the coefficient s, and (iv) developing a bound on ϕ^2 in (32).

Step (i): Preparation. Denote α_1 and α_2 as follows:

$$\alpha_1 := 2\sqrt{\frac{\log(32/\delta)}{n}},$$
$$\alpha_2 := \sqrt{\frac{\operatorname{rank}(\Sigma_u)+1}{n}} + 2\sqrt{\frac{\log(16/\delta)}{n}}.$$

To prepare for the derivation of the upper bound as in the proof of Lemma 14, we consider the following three inequalities:

(i) By Lemma 39, uniformly over all $\theta_2 \in \Sigma_u^{1/2}(\mathbb{R}^p)$, it holds that

$$|\langle \xi - \mathbf{W}_2\theta_2, G\rangle| \leq \|\xi - \mathbf{W}_2\theta_2\|_2 \|G\|_2 \alpha_2. \tag{26}$$

(ii) By Lemma 40, it holds that

$$-\alpha_1 \leq \frac{1}{\sqrt{n}}\|G\|_2 - 1 \leq \alpha_1 \tag{27}$$

and

$$\begin{aligned}-\alpha_1\sqrt{\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2} &\leq \frac{1}{\sqrt{n}}\|\xi - \mathbf{W}_2\theta_2\|_2 - \sqrt{\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2} \\ &\leq \alpha_1\sqrt{\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2}. \end{aligned}\tag{64}$$

(iii) By Theorem 43, it holds that

$$\begin{aligned}\|\Xi_z^{1/2}H\|_* &\geq \mathbb{E}\|\Xi_z^{1/2}H\|_* - \sup_{\|u\|\leq 1}\|u\|_{\Xi_z}\sqrt{2\log(8/\delta)} \\ &= \left(1 - \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)\mathbb{E}\|\Xi_z^{1/2}H\|_*\end{aligned}\tag{36}$$

because $\|\Xi_z^{1/2}H\|_*$ is a $\sup_{\|u\|\leq 1}\|u\|_{\Xi_z}$-Lipschitz continuous function of H. By Theorem 43, it also holds that

$$\begin{aligned}\|\Xi_z^{1/2}H\|_* &\leq \mathbb{E}\|\Xi_z^{1/2}H\|_* + \sup_{\|u\|\leq 1}\|u\|_{\Xi_z}\sqrt{2\log(8/\delta)} \\ &= \left(1 + \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)\mathbb{E}\|\Xi_z^{1/2}H\|_*.\end{aligned}\tag{65}$$

Step (ii): Introducing the coefficient s. To derive an upper bound of the ridgeless estimator, we need to choose a suitable θ which satisfies the restriction of the auxiliary problem ϕ in Lemma 18. We consider the following form of θ:

$$\theta := P_u\Sigma_u^+\omega + sP_z v^*. \tag{66}$$

Here, the coefficient s describes a volume of θ along with the space spanned by Ξ_z. By the setting, we have $\theta_1 = \Xi_z^{1/2}\Sigma_u^+\omega + s\Xi_z^{1/2}v^*$ and $\theta_2 = \Sigma_u^{1/2}\Sigma_u^+\omega +$

$s\Sigma_u^{1/2}P_z v^*$. Hence, we need to choose s that attains the restriction of the auxiliary problem ϕ, that is,

$$\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2 \leq \langle H, \theta_1 \rangle. \tag{32}$$

By the definition of θ_1, we have the following result:

$$\begin{aligned} &(\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2 - \langle H, \Xi_z^{1/2}\Sigma_u^{+}\omega\rangle)^2 \leq s^2\|\Xi_z^{1/2}H\|_*^2 \\ \Rightarrow &\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2 \leq \langle H, \theta_1 \rangle. \end{aligned} \tag{67}$$

Therefore, it is sufficient to consider s satisfying the inequality (67).

For the derivation of the inequality (67), we need to consider the upper bound of $(\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2 - \langle H, \Xi_z^{1/2}\Sigma_u^{+}\omega\rangle)^2$. It holds from (26) and the AM-GM inequality that

$$\begin{aligned} \|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2 &= \|\xi - \mathbf{W}_2\theta_2\|_2^2 - 2\|\theta_1\|_2\langle \xi - \mathbf{W}_2\theta_2, G\rangle + \|\theta_1\|_2^2\|G\|_2^2 \\ &\geq \|\xi - \mathbf{W}_2\theta_2\|_2^2 - 2\alpha_2\|\theta_1\|_2\|\xi - \mathbf{W}_2\theta_2\|_2\|G\|_2 + \|\theta_1\|_2^2\|G\|_2^2 \\ &\geq (1-\alpha_2)\left(\|\xi - \mathbf{W}_2\theta_2\|_2^2 + \|\theta_1\|_2^2\|G\|_2^2\right). \end{aligned}$$

Combining the results of (27) and (64) yields

$$\begin{aligned} &(1-\alpha_2)\left(\|\xi - \mathbf{W}_2\theta_2\|_2^2 + \|\theta_1\|_2^2\|G\|_2^2\right) \\ &\geq (1-\alpha_2)(1-\alpha_1)^2 n\left(\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2 + \theta_1^T\theta_1\right). \end{aligned}$$

Then, we have

$$\begin{aligned} \left(\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2 + \theta_1^T\theta_1\right) &= \left(\sigma^2 - 2\rho^T\Sigma_u^{1/2}\theta + \theta^\top\Sigma_u\theta + \theta^T\Xi_z\theta\right) \\ &\geq \left(\sigma^2 - 2\rho^T\Sigma_u^{1/2}\theta + \theta^\top\Sigma_u\theta\right) \\ &= \sigma^2 - \rho^\top\rho + \|\Sigma_u^{1/2}\theta - \rho\|_2^2 \\ &\geq \sigma^2 - \rho^\top\rho + \min_{\theta\in\mathbb{R}^p}\|\Sigma_u^{1/2}\theta - \rho\|_2^2 \\ &= \widetilde{\sigma}^2 > 0. \end{aligned}$$

Therefore, we have

$$0 < \frac{1}{\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2} \leq \frac{1}{\sqrt{(1-\alpha_2)(1-\alpha_1)^2 n\widetilde{\sigma}^2}}.$$

By trivial calculation, we obtain

$$\begin{aligned} &(\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2 - \langle H, \Xi_z^{1/2}\Sigma_u^{+}\omega\rangle)^2 \\ &\leq \|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2 + 2\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2|\langle H, \Xi_z^{1/2}\Sigma_u^{+}\omega\rangle| \\ &\quad + (\langle H, \Xi_z^{1/2}\Sigma_u^{+}\omega\rangle)^2 \\ &= \|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2\left(1 + 2\frac{|\langle H, \Xi_z^{1/2}\Sigma_u^{+}\omega\rangle|}{\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2}\right) + (\langle H, \Xi_z^{1/2}\Sigma_u^{+}\omega\rangle)^2 \end{aligned}$$

$$
\begin{aligned}
&\leq \|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2 \left(1 + 2\frac{|\langle H, \Xi_z^{1/2}\Sigma_u^+\omega\rangle|}{\sqrt{(1-\alpha_2)(1-\alpha_1)^2 n\widetilde{\sigma}^2}}\right) + (\langle H, \Xi_z^{1/2}\Sigma_u^+\omega\rangle)^2 \\
&= \|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2 (1+\gamma) + (\langle H, \Xi_z^{1/2}\Sigma_u^+\omega\rangle)^2.
\end{aligned}
$$

Combining the results of (27) and (64) yields

$$
\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2^2 \leq (1+\alpha_2)(1+\alpha_1)^2 n\left(\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2 + \theta_1^T\theta_1\right).
$$

Then, it holds that

$$
\begin{aligned}
&(\|\xi - \mathbf{W}_2\theta_2 - \|\theta_1\|_2 G\|_2 - \langle H, \Xi_z^{1/2}\Sigma_u^+\omega\rangle)^2 \\
\leq &(1+\alpha_2)(1+\alpha_1)^2(1+\gamma)n\left(\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2 + \theta_1^T\theta_1\right) + (\langle H, \Xi_z^{1/2}\Sigma_u^+\omega\rangle)^2
\end{aligned}
$$

Therefore, we should choose s that satisfies the subsequent equality:

$$
\begin{aligned}
&s^2\|\Xi_z^{1/2}H\|_*^2 \\
&= (1+\alpha_2)(1+\alpha_1)^2(1+\gamma)n\left(\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2 + \theta_1^T\theta_1\right) + (\langle H, \Xi_z^{1/2}\Sigma_u^+\omega\rangle)^2.
\end{aligned} \tag{68}
$$

We clarify s that satisfies (68). By trivial calculation, we obtain the following results:

$$
\begin{aligned}
&s^2\|\Xi_z^{1/2}H\|_*^2 \\
&\quad= (1+\alpha_2)(1+\alpha_1)^2(1+\gamma)n\left(\sigma^2 - 2\rho^T\theta_2 + \theta_2^T\theta_2 + \theta_1^T\theta_1\right) + (\langle H, \Xi_z^{1/2}\Sigma_u^+\omega\rangle)^2 \\
&\Leftrightarrow \beta_1\left(s + \frac{\beta_2}{\beta_1}\right)^2 - \frac{\beta_2^2}{\beta_1} - (\widetilde{\sigma}^2 - \beta_3) = 0,
\end{aligned}
$$

where we define

$$
\begin{aligned}
\beta_1 &:= \left(\frac{\|\Xi_z^{1/2}H\|_*^2}{(1+\alpha_2)(1+\alpha_1)^2(1+\gamma)n} - \|v^*\|_{\Xi_z}^2 - (v^*)^\top P_z\Sigma_u P_z v^*\right), \\
\beta_2 &:= \left(\rho^\top\Sigma_u^{1/2}P_z v^* - \omega^\top\Sigma_u^+\Sigma_u P_z v^* - v^*\Xi_z\Sigma_u^+\omega\right), \\
\beta_3 &:= -\frac{(\langle H, \Xi_z^{1/2}\Sigma_u^+\omega)\rangle)^2}{(1+\alpha_2)(1+\alpha_1)^2(1+\gamma)n} - (\omega^\top\Sigma_u^+\Xi_z\Sigma_u^+\omega), \\
\gamma &:= 2\frac{|\langle H, \Xi_z^{1/2}\Sigma_u^+\omega\rangle|}{\sqrt{(1-\alpha_2)(1-\alpha_1)^2 n\widetilde{\sigma}^2}}.
\end{aligned}
$$

Therefore, we choose s such that

$$
s = \left(-\frac{\beta_2}{\beta_1} + \sqrt{\frac{\beta_2^2}{\beta_1^2} + \frac{\widetilde{\sigma}^2 - \beta_3}{\beta_1}}\right),
$$

under the assumption that

$$
\beta_1 > 0 \quad \text{and} \quad \widetilde{\sigma}^2 - \beta_3 \geq 0. \tag{69}
$$

We need to guarantee (69) holds. First, we prove $\widetilde{\sigma}^2 - \beta_3 \geq 0$. By definition, we have $\widetilde{\sigma}^2 > 0$ and $-\beta_3 \geq 0$. Second, we show $\beta_1 > 0$. By (60), (62), and (36), we have

$$\begin{aligned}
\beta_1 \geq & \frac{(\mathbb{E}\|\Xi_z^{1/2} H\|_*)^2}{(1+\alpha_2)(1+\alpha_1)^2(1+\gamma)n} \left(1 - \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)^2 \\
& - (1+\varepsilon_1)^2 (\mathbb{E}\|v^*\|_{\Xi_z})^2 - (1+\varepsilon_3)^2 (\mathbb{E}\|\Sigma_u^{1/2} P_z v^*\|_2)^2 \\
\geq & \frac{(\mathbb{E}\|\Xi_z^{1/2} H\|_*)^2}{n} \left(\frac{1}{(1+\alpha_2)(1+\alpha_1)^2(1+\gamma)} \left(1 - 2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right) \right. \\
& \left. -(1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Sigma_2)} - (1+\varepsilon_3)^2 \frac{n}{(\mathbb{E}\|\Xi_z^{1/2} H\|_*)^2} (\mathbb{E}\|\Sigma_u^{1/2} P_z v^*\|_2)^2 \right).
\end{aligned}$$

We linearize the terms including α_1, α_2, and γ to simplify the upper bound. Provided that $\alpha_1, \alpha_2 < 1/2$, we have

$$\begin{aligned}
(1+\alpha_2)(1+\alpha_1)^2(1+\gamma) &= (1 + 2\alpha_1 + \alpha_1^2 + \alpha_2 + 2\alpha_2\alpha_1 + \alpha_2\alpha_1^2)(1+\gamma) \\
&\leq (1 + 3\alpha_1 + 3\alpha_2)(1+\gamma) \\
&\leq 1 + 3\alpha_1 + 3\alpha_2 + 4\gamma.
\end{aligned}$$

As $(1-x)^{-1} \geq 1 + x$ for any x, it holds that

$$\begin{aligned}
\frac{1}{(1+\alpha_2)(1+\alpha_1)^2(1+\gamma)} &\geq (1 + 3\alpha_1 + 3\alpha_2 + 4\gamma)^{-1} \\
&\geq (1 - (3\alpha_1 + 3\alpha_2 + 4\gamma)).
\end{aligned}$$

Moreover, by the Cauchy-Schwarz inequality, we have

$$\gamma \leq \frac{2\|\Xi_z^{1/2} H\|_2 \|\|\Sigma_u^+ \omega\|_2}{\sqrt{(1-\alpha_2)(1-\alpha_1)^2 n \widetilde{\sigma}^2}}.$$

As $(1-x)^{-1} \leq 1 + 2x$ for $x \in [0, 1/2]$, it holds that

$$\frac{1}{\sqrt{(1-\alpha_2)(1-\alpha_1)^2}} \leq \frac{1}{(1-\alpha_2)(1-\alpha_1)} \leq 4.$$

By (65), we have

$$\frac{\|\Xi_z^{1/2} H\|_2}{\sqrt{n}} \leq \frac{\mathbb{E}\|\Xi_z^{1/2} H\|_2}{\sqrt{n}} \left(1 + \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right).$$

Therefore, it holds that

$$\gamma \leq \frac{8\|\Sigma_u^+ \omega\|_2}{\widetilde{\sigma}} \frac{\mathbb{E}\|\Xi_z^{1/2} H\|_2}{\sqrt{n}} \left(1 + \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right).$$

Hence, we have

$$
\begin{aligned}
&\frac{1}{(1+\alpha_2)(1+\alpha_1)^2(1+\gamma)}\left(1-2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)\\
&\quad -(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)}-(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_u^{1/2}P_zv^*\|_2)^2\\
&\geq(1-(3\alpha_1+3\alpha_2+4\gamma))\left(1-2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}\right)-(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)}\\
&\quad -(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_u^{1/2}P_zv^*\|_2)^2\\
&\geq 1-(3\alpha_1+3\alpha_2+4\gamma)-2\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}-(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Sigma_2)}\\
&\quad -(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_u^{1/2}P_zv^*\|_2)^2\\
&\geq 1-\varepsilon'
\end{aligned}
$$

where we define

$$
\begin{aligned}
\varepsilon' =&6\sqrt{\frac{\operatorname{rank}(\Sigma_u)}{n}}+12\sqrt{\frac{\log(32/\delta)}{n}}+32\frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_2}{\sqrt{n}}\left(1+\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right)\\
&+4\sqrt{\frac{\log(8/\delta)}{r_{\|\cdot\|}(\Xi_z)}}+(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Xi_z)}+(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}(\mathbb{E}\|\Sigma_u^{1/2}P_zv^*\|_2)^2.
\end{aligned}
$$

As ε' is assumed to be less than $1/2$ and $(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2/n$ is positive, we have

$$0<(1-\varepsilon')\frac{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}{n}\leq\beta_1. \tag{70}$$

Step (iii): Bound on the coefficient s. As s is too complicated, we need to obtain a simplified upper bound of s. By trivial calculation, we have

$$s\leq\frac{2|\beta_2|}{\beta_1}+\sqrt{\frac{\widetilde{\sigma}^2}{\beta_1}}+\sqrt{\frac{|\beta_3|}{\beta_1}}.$$

Then, we consider an upper bound of $1/\beta_1$. It holds from (70) that

$$\frac{1}{\beta_1}\leq(1-\varepsilon')^{-1}\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}\leq(1+2\varepsilon')\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}, \tag{71}$$

where the last inequality holds because $(1-x)^{-1}\leq 1+2x$ for $x\in[0,1/2]$.

Next, we show an upper bound of $|\beta_2|$. As $\omega := \Sigma_u^{1/2}\rho$ and $\omega^\top \Sigma_u^+ \Sigma_u = \omega^\top (\Sigma_u^{1/2})^+ \Sigma_u^{1/2}$, we have

$$\begin{aligned}\beta_2 &= \left(\rho^\top \Sigma_u^{1/2} P_z v^* - \omega^\top \Sigma_u^+ \Sigma_u P_z v^* - v^* \Xi_z \Sigma_u^+ \omega\right) \\ &= \left(\omega^\top (\Sigma_u^{1/2})^+ \Sigma_u^{1/2} P_z v^* - \omega^\top \Sigma_u^+ \Sigma_u P_z v^* - v^* \Xi_z \Sigma_u^+ \omega\right) \\ &= - v^* \Xi_z \Sigma_u^+ \omega.\end{aligned}$$

By the Cauchy-Schwarz inequality, it holds that $|\beta_2| \leq \|v^*\|_{\Xi_z} \|\Xi_z^{1/2} \Sigma_u^+ \omega\|_2$. From assumption (60) and the definition of $R_{\|\cdot\|}(\Xi_z)$, we obtain

$$|\beta_2| \leq (1+\varepsilon_1) \sqrt{\frac{(\mathbb{E}\|\Xi_z^{1/2} H\|_*)^2}{n}} \sqrt{\frac{n}{R_{\|\cdot\|}(\Xi_z)}} \|\Xi_z^{1/2} \Sigma_u^+ \omega\|_2. \tag{72}$$

Combining the result (72) with (71), we have

$$\frac{2|\beta_2|}{\beta_1} \leq 2(1+2\varepsilon')(1+\varepsilon_1) \sqrt{\frac{n}{(\mathbb{E}\|\Xi_z^{1/2} H\|_*)^2}} \sqrt{\frac{n}{R_{\|\cdot\|}(\Xi_z)}} \|\Xi_z^{1/2} \Sigma_u^+ \omega\|_2. \tag{73}$$

Finally, we derive an upper bound of β_3 and s. By using the triangular inequality on β_3, we have

$$|\beta_3| \leq \frac{(\langle H, \Xi_z^{1/2} \Sigma_u^+ \omega \rangle)^2}{(1+\alpha_2)(1+\alpha_1)^2 n} + \|\Xi_z^{1/2} \Sigma_u^+ \omega\|_2^2. \tag{74}$$

By the Cauchy-Schwarz inequality, it holds that

$$\frac{(\langle H, \Xi_z^{1/2} \Sigma_u^+ \omega \rangle)^2}{(1+\alpha_2)(1+\alpha_1)^2 n} \leq \frac{\|\Xi_z^{1/2} H\|_2^2}{(1+\alpha_2)(1+\alpha_1)^2 n} \|\Sigma_u^+ \omega\|_2^2. \tag{75}$$

By (65), we have

$$\frac{\|\Xi_z^{1/2} H\|_2^2}{(1+\alpha_2)(1+\alpha_1)^2 n} \leq \frac{(\mathbb{E}\|\Xi_z^{1/2} H\|_2)^2}{n} \frac{1}{(1+\alpha_2)(1+\alpha_1)^2} \left(1 + \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right)^2. \tag{76}$$

Hence, it holds from (71), (74), (75), and (76) that

$$\begin{aligned}\sqrt{\frac{|\beta_3|}{\beta_1}} \leq & \sqrt{(1+2\varepsilon')} \sqrt{\frac{n}{(\mathbb{E}\|\Xi_z^{1/2} H\|_*)^2}} \\ & \times \sqrt{\frac{(\mathbb{E}\|\Xi_z^{1/2} H\|_2)^2}{n} \left(1 + \sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right)^2 \|\Sigma_u^+ \omega\|_2^2 + \|\Xi_z^{1/2} \Sigma_u^+ \omega\|_2^2}.\end{aligned} \tag{77}$$

From (73) and (77), we obtain

$$s \leq \sqrt{\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}}A \tag{78}$$

where A is defined and bounded as follows:

$$\begin{aligned}A &:= 2(1+2\varepsilon')(1+\varepsilon_1)\sqrt{\frac{n}{R_{\|\cdot\|}(\Xi_z)}}\|\Xi_z^{1/2}\Sigma_u^+\omega\|_2 + \sqrt{(1+2\varepsilon')\widetilde{\sigma}^2} \\ &\quad + \sqrt{(1+2\varepsilon')}\sqrt{\frac{(\mathbb{E}\|\Xi_z^{1/2}H\|_2)^2}{n}\left(1+\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right)^2\|\Sigma_u^+\omega\|_2^2 + \|\Xi_z^{1/2}\Sigma_u^+\omega\|_2^2} \\ &\leq 2(1+2\varepsilon')\eta_1 + \sqrt{(1+2\varepsilon')}\widetilde{\sigma} + (1+2\varepsilon')\eta_2,\end{aligned} \tag{79}$$

where

$$\eta_1 := \sqrt{(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|}(\Xi_z)}}\|\Xi_z^{1/2}\Sigma_u^+\omega\|_2,$$

and

$$\eta_2 := \sqrt{\frac{(\mathbb{E}\|\Xi_z^{1/2}H\|_2)^2}{n}\left(1+\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right)^2\|\Sigma_u^+\omega\|_2^2 + \|\Xi_z^{1/2}\Sigma_u^+\omega\|_2^2}.$$

Step (iv): Bound ϕ^2. We simplify the upper bound (78) and derive the upper bound of ϕ^2. By trivial calculation, we utilize the definition of θ as (66) and obtain

$$\begin{aligned}\phi^2 &\leq \|\theta\|^2 \\ &\leq \|P_u\Sigma_u^+\omega + sP_z v^*\|^2 \\ &\leq (\|P_u\Sigma_u^+\omega\| + s\|P_z v^*\|)^2 \\ &\leq (\|\Sigma_u^+\omega\| + s(1+\varepsilon_2))^2 \\ &\leq \left(\|\Sigma_u^+\omega\| + \sqrt{\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}}A(1+\varepsilon_2)\right)^2.\end{aligned}$$

The second to last inequality follows the assumption in (61), and the last inequality follows the upper bound on s as (78). We apply (79) and set $\varepsilon := 2(\varepsilon' + \varepsilon_2)$, and then it holds that

$$\phi^2 \leq \left(\|\Sigma_u^+\omega\| + (1+\varepsilon)(2\eta_1 + \widetilde{\sigma} + \eta_2)\sqrt{\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2}}\right)^2.$$

□

Theorem 31 (General norm bound)**.** *There exists an absolute constant $C_2 \leq 160$ such that the following is true. Under Assumption 1 with $\Sigma_x = \Xi_z + \Sigma_u$, let $\|\cdot\|$ be an arbitrary norm, and fix $\delta \leq 1/4$. Denote the ℓ_2 orthogonal projection matrix onto the space spanned by Ξ_z and Σ_u as P_z and P_u, respectively. Let H be normally distributed with mean zero and variance I_d, that is, $H \sim N(0, I_d)$. Denote v_* as $\arg\min_{v \in \partial \|\Xi_z^{1/2} H\|_*} \|v\|_{\Xi_z}$. Suppose that there exist ε_1 and $\varepsilon_2 \geq 0$ such that with probability at least $1 - \delta/8$,*

$$\|v^*\|_{\Xi_z} \leq (1+\varepsilon_1)\mathbb{E}\|v^*\|_{\Xi_z},$$
$$\|P_z v^*\| \leq 1 + \varepsilon_2,$$

and

$$\|\Sigma_u^{1/2} P_z v^*\|_2 \leq (1+\varepsilon_3)\mathbb{E}\|\Sigma_u^{1/2} P_z v^*\|_2.$$

Denote ε as follows:

$$\begin{aligned}\varepsilon := C_2 \Bigg(& \sqrt{\frac{\operatorname{rank}(\Sigma_u)}{n}} + \sqrt{\frac{\log(1/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}} + \sqrt{\frac{\log(1/\delta)}{n}} + (1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|_2}(\Xi_z)} \\ & + \frac{\|\Sigma_u^+ \omega\|_2}{\widetilde{\sigma}} \frac{\mathbb{E}\|\Xi_z^{1/2} H\|_2}{\sqrt{n}} \left(1 + \sqrt{\frac{\log(1/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right) \\ & + (1+\varepsilon_3)^2 \frac{n}{(\mathbb{E}\|\Xi_z^{1/2} H\|_*)^2} (\mathbb{E}\|\Sigma_u^{1/2} P_z v^*\|_2)^2 + \varepsilon_2 \Bigg).\end{aligned}$$

If n and the effective ranks are large enough that $\varepsilon \leq 1$, with probability at least $1 - \delta$, it holds that

$$\|\widehat{\theta}\| \leq \|\theta_0\| + \|\Sigma_u^+ \omega\| + (1+\varepsilon)(2\eta_1 + \widetilde{\sigma} + \eta_2)\sqrt{\frac{n}{(\mathbb{E}\|\Xi_z^{1/2} H\|_*)^2}}.$$

Proof of Theorem 31. For any $t > 0$, it holds from Lemmas 17 and 18 that

$$\mathbb{P}(\|\widehat{\theta}\| > t) \leq \mathbb{P}(\Phi > t - \|\theta_0\|) \leq 2\mathbb{P}(\phi \geq t - \|\theta_0\|).$$

Lemma 30 implies that the above term is upper bounded by δ if we choose $t - \|\theta_0\|$ using the result (63) with δ replaced by $\delta/2$. We obtain the stated result by moving $\|\theta_0\|$ to the other side. □

Theorem 11 (Euclidean norm bound; special case of Theorem 31) *Fix any $\delta \leq 1/4$. Under the model assumptions with covariance $\Sigma_x = \Xi_z + \Sigma_u$, there exists some $\varepsilon \lesssim \sqrt{\operatorname{rank}(\Sigma_u)/n} + \sqrt{\log(1/\delta)/n} + (1 + (\|\Sigma_u^+ \omega\|_2/\widetilde{\sigma})\sqrt{\operatorname{tr}(\Xi_z)/n})$ $\sqrt{\log(1/\delta)/r(\Xi_z)} + (n\log(1/\delta))/(R(\Xi_z))(1 + \operatorname{tr}(\Sigma_u \Xi_z)/\operatorname{tr}(\Xi_z^2)) + (\|\Sigma_u^+ \omega\|_2/\widetilde{\sigma})$ $\sqrt{\operatorname{tr}(\Xi_z)/n}$ such that the following is true. If n and the effective ranks are such that $\varepsilon \leq 1$ and $R(\Xi_z) \gtrsim \log(1/\delta)^2$, then with probability at least $1 - \delta$, it holds that*

$$\|\widehat{\theta}\|_2 \leq \|\theta_0\|_2 + \|\Sigma_u^+ \omega\|_2 + (1+\varepsilon)^{1/2}(2\eta_1 + \widetilde{\sigma} + \eta_2)\sqrt{\frac{n}{\operatorname{tr}(\Xi_z)}}.$$

Proof of Theorem 11. Throughout this proof, we simplify the upper bound, especially $n/R_{\|\cdot\|_2}(\Xi_z)$, $1/r_{\|\cdot\|_2}(\Xi_z)$ and $(n\mathbb{E}\|\Sigma_u^{1/2}P_zv^*\|_2^2)/(\mathbb{E}\|\Xi_z^{1/2}H\|_*)^2$, in Theorem 31. By the definition of the dual norm and $\partial\|\Xi_x^{1/2}H\|_*$ with Euclidean norm, v^* is equal to $\Xi_z^{1/2}H/\|\Xi_z^{1/2}H\|_2$. Hence, $\|v^*\|_{\Xi_z}$ is $\|\Xi_zH\|_2/\|\Xi_z^{1/2}H\|_2$. From the result of (94), for some constant $c > 0$, we can choose ε_1 such that

$$(1+\varepsilon_1)E\|v^*\|_{\Xi_z} = c\sqrt{\log(64/\delta)\frac{\mathrm{tr}(\Xi_z^2)}{\mathrm{tr}(\Xi_z)}}.$$

If we assume effective rank is sufficiently large, (91) provides that $\left(E\|\Xi_z^{1/2}H\|_2\right)^2 \gtrsim \mathrm{tr}(\Xi_z)$. Therefore, we have

$$(1+\varepsilon_1)^2\frac{n}{R_{\|\cdot\|_2}(\Xi_z)} = n\frac{(1+\varepsilon_1)^2(E\|v^*\|_{\Xi_z})^2}{\left(E\|\Xi_z^{1/2}H\|_2\right)^2} \lesssim n\log(16/\delta)\frac{\mathrm{tr}(\Xi_z^2)}{\mathrm{tr}(\Xi_z)^2} = \frac{n\log(16/\delta)}{R(\Xi_z)}.$$

Moreover, it holds from (46) that for some constant $c_2 > 0$, there exists ε_3 such that

$$(1+\varepsilon_3)\mathbb{E}\|\Sigma_u^{1/2}Pv^*\|_2 = c_2\sqrt{\log(64/\delta)\frac{\mathrm{tr}(\Sigma_u\Xi_z)}{\mathrm{tr}(\Xi_z)}}.$$

By (91), for sufficiently large effective rank, it holds that $(E\|\Xi_z^{1/2}H\|_2)^2 \gtrsim \mathrm{tr}(\Xi_z)$. Therefore, we have

$$\begin{aligned}(1+\varepsilon_3)^2\frac{n}{(\mathbb{E}\|\Xi_z^{1/2}H\|_2)^2}(\mathbb{E}\|\Sigma_u^{1/2}P_zv^*\|_2)^2 &\lesssim n\log(64/\delta)\frac{\mathrm{tr}(\Sigma_u\Xi_z)}{\mathrm{tr}(\Xi_z)^2}\\ &= \frac{n\log(64/\delta)}{R(\Xi_z)}\frac{\mathrm{tr}(\Sigma_u\Xi_z)}{\mathrm{tr}(\Xi_z^2)}.\end{aligned}$$

By trivial calculation, for any covariance matrix Σ, we have

$$\begin{aligned}\mathrm{tr}(\Sigma) = \mathbb{E}\|\Sigma^{1/2}H\|_2^2 &= (\mathbb{E}\|\Sigma^{1/2}H\|_2)^2 + \mathrm{Var}\|\Sigma^{1/2}H\|_2\\ &\geq (\mathbb{E}\|\Sigma^{1/2}H\|_2)^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\frac{\mathbb{E}\|\Xi_z^{1/2}H\|_2}{\sqrt{n}}\left(1+\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right)\\ &\quad\leq \frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}}\left(1+\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right).\end{aligned}$$

Finally, we obtain the upper bound of ε. As P_z is an l_2 projection matrix, let ε_2 be zero. Then, it holds from (92) of Lemma 41 that

$$\varepsilon \lesssim \sqrt{\frac{\mathrm{rank}(\Sigma_u)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \left(1 + \frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}}\right)\sqrt{\frac{\log(1/\delta)}{r(\Xi_z)}}$$

$$+\frac{n\log(1/\delta)}{R(\Xi_z)}\left(1+\frac{\mathrm{tr}(\Sigma_u\Xi_z)}{\mathrm{tr}(\Xi_z^2)}\right)+\frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}}.$$

By using the inequality $(1-x)^{-1}\le 1+2x$ for $x\in[0,1/2]$ and (91) of Lemma 41, we finally obtain

$$\begin{aligned}(1+\varepsilon)\frac{\sqrt{n}}{E\|\Xi_z^{1/2}H\|_2} &\le (1+\varepsilon)\left(1-\frac{1}{r(\Xi_z)}\right)^{-1/2}\sqrt{\frac{n}{\mathrm{tr}(\Xi_z)}}\\ &\le (1+\varepsilon)\left(1+\frac{2}{r(\Xi_z)}\right)^{1/2}\sqrt{\frac{n}{\mathrm{tr}(\Xi_z)}}.\end{aligned}$$

As $\varepsilon\le 1$, we have

$$\begin{aligned}(1+\varepsilon)\left(1+\frac{2}{r(\Xi_z)}\right)^{1/2} &= \left((1+2\varepsilon+\varepsilon^2)\left(1+\frac{2}{r(\Xi_z)}\right)\right)^{1/2}\\ &\le \left((1+3\varepsilon)\left(1+\frac{2}{r(\Xi_z)}\right)\right)^{1/2}\\ &\le \left(1+3\varepsilon+\frac{8}{r(\Xi_z)}\right)^{1/2}.\end{aligned}$$

Therefore, it holds that

$$(1+\varepsilon)\frac{\sqrt{n}}{E\|\Xi_z^{1/2}H\|_2}\le\left(1+3\varepsilon+\frac{8}{r(\Xi_z)}\right)^{1/2}\sqrt{\frac{n}{\mathrm{tr}(\Xi_z)}},$$

and we can replace ε with

$$\begin{aligned}\varepsilon' =&3\varepsilon+\frac{8}{r(\Xi_z)}\\ &\lesssim\sqrt{\frac{\mathrm{rank}(\Sigma_u)}{n}}+\sqrt{\frac{\log(1/\delta)}{n}}+\left(1+\frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}}\right)\sqrt{\frac{\log(1/\delta)}{r(\Xi_z)}}\\ &+\frac{n\log(1/\delta)}{R(\Xi_z)}\left(1+\frac{\mathrm{tr}(\Sigma_u\Xi_z)}{\mathrm{tr}(\Xi_z^2)}\right)+\frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}}.\end{aligned}$$

□

Theorem 32 (Benign Overfitting (Non-orthogonal))**.** *Fix any $\delta\le 1/2$. Under the model assumptions with $\Sigma_x=\Xi_z+\Sigma_u$, let γ and ε be as defined in Corollary 10 and Theorem 11. Suppose that n and the effective ranks are such that $R(\Xi_z)\gtrsim\log(1/\delta)^2$ and $\gamma,\varepsilon\le 1$. Then, with probability at least $1-\delta$,*

$$\|\widehat{\theta}-\theta_0\|_{\Xi_z}^2\le(1+\gamma)(1+\varepsilon)\left((\|\theta_0\|_2+\|\Sigma_u^+\omega\|_2)\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}}+(2\eta_1+\widetilde{\sigma}+\eta_2)\right)^2-\widetilde{\sigma}^2.$$

Proof of Theorem 32. From the result of Theorem 11, if we adopt

$$B = \|\theta_0\|_2 + \|\Sigma_u^+ \omega\|_2 + (1+\varepsilon)^{1/2}(2\eta_1 + \widetilde{\sigma} + \eta_2)\sqrt{\frac{n}{\mathrm{tr}(\Xi_z)}},$$

then $\{\theta : \|\theta\|_2 \leq B\} \cap \{\theta : \mathbf{X}\theta = \mathbf{Y}\}$ is not empty with high probability. This intersection necessarily contains the ridgeless estimator $\widehat{\theta}$. Clearly, $B > \|\theta_0\|_2$ holds. Therefore, it holds from Corollary 10 that

$$\begin{aligned}
&\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 \\
&\leq \max_{\|\theta\|_2 \leq B, \mathbf{Y} = \mathbf{X}\theta} \|\theta - \theta_0\|_{\Xi_z}^2 \\
&\leq (1+\gamma)\frac{B^2 \mathrm{tr}(\Xi_z)}{n} - \widetilde{\sigma}^2 \\
&\leq (1+\gamma)(1+\varepsilon)\left((\|\theta_0\|_2 + \|\Sigma_u^+ \omega\|_2)\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}} + (2\eta_1 + \widetilde{\sigma} + \eta_2)\right)^2 - \widetilde{\sigma}^2.
\end{aligned}$$

□

Theorem 33. *Under Assumption 1, let $\widehat{\theta}$ be the ridgeless estimator. Suppose that as n goes to ∞, the covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$ satisfies the following conditions:*

(i) (Small large-variance dimension.)

$$\lim_{n\to\infty} \frac{\mathrm{rank}(\Sigma_u)}{n} = 0.$$

(ii) (Large effective dimension.)

$$\lim_{n\to\infty} \frac{n}{R(\Xi_z)} = 0.$$

(iii) (No aliasing condition.)

$$\lim_{n\to\infty} \|\theta_0\|_2 \sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}} = 0.$$

(iv) (Condition for the minimal interpolation of instrumental variable in the non-orthogonal case)

$$\lim_{n\to\infty} \frac{\|\Sigma_u^+ \omega\|_2}{\widetilde{\sigma}} \sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}} = 0.$$

(v) (Non-orthogonality)

(a)

$$\lim_{n\to\infty} \eta_1 = 0.$$

(b)

$$\lim_{n\to\infty} \eta_2 = 0.$$

(c)

$$\lim_{n\to\infty} \frac{n}{R(\Xi_z)} \frac{\mathrm{tr}(\Sigma_u \Xi_z)}{\mathrm{tr}(\Xi_z^2)} = 0.$$

Then, $\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2$ converges to 0 in probability.

Proof of Theorem 33. Fix any $\kappa > 0$ and $\delta > 0$. From Lemma 5 of Bartlett *et al.* (2020), it holds that $R(\Xi_z) \le r(\Xi_z)^2$. If $R(\Xi_z) = \upsilon(n)$ holds as the second condition in Theorem 33, we have $r(\Xi_z) = \upsilon(\sqrt{n}) = \upsilon(1)$, which implies the convergence of $1/r(\Xi_z)$ to zero. Hence, conditions (i) and (ii) in Theorem 33 make γ sufficiently small for large enough n. Clearly, $(\|\theta_0\|_2 + \|\Sigma_u^+ \omega\|_2)\sqrt{\mathrm{tr}(\Xi_z)/n}$ goes to zero from conditions (iii) and (iv). By the definition of ε, conditions (i), (ii), (iv), and (v)(c) in Theorem 33 imply that ε can be arbitrarily small. Combined with conditions (v)(a) and (v)(b), for sufficiently large n, we obtain

$$(1+\gamma)(1+\varepsilon)\left((\|\theta_0\|_2 + \|\Sigma_u^+\omega\|_2)\sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}} + (2\eta_1 + \widetilde{\sigma} + 2\eta_2)\right)^2 - \widetilde{\sigma}^2 \le \kappa. \tag{80}$$

We have shown that γ, ε, $(\|\theta_0\|_2 + \|\Sigma_u^+\omega\|_2)\sqrt{\mathrm{tr}(\Xi_z)/n}$, η_1, and η_2 are so small that equation (80) holds for sufficiently large n. Therefore, we obtain

$$\mathbb{P}(\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 > \kappa) \le \delta,$$

for any fixed κ. As κ and δ are arbitrary, we have for any κ,

$$\lim_{n\to\infty} \mathbb{P}(\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2 > \kappa) = 0.$$

□

Lemma 34. *Suppose $\lim_{n\to\infty} \omega^\top \Sigma_u^+ \Xi_z \Sigma_u^+ \omega = 0$ holds. Suppose the second condition of Theorem 33 holds. Then, with probability at least $1-\delta$, $\lim_{n\to\infty} \eta_1 = 0$.*

Proof of Lemma 34. By the definition of η_1, we have

$$\eta_1 = \sqrt{(1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|}(\Xi_z)}} \|\Xi_z^{1/2} \Sigma_u^+ \omega\|_2. \tag{81}$$

By (91), for sufficiently large effective rank, it holds that $\left(E\|\Xi_z^{1/2} H\|_2\right)^2 \gtrsim \mathrm{tr}(\Xi_z)$ and so

$$(1+\varepsilon_1)^2 \frac{n}{R_{\|\cdot\|_2}(\Xi_z)} = n\frac{(1+\varepsilon_1)^2 (E\|v^*\|_{\Xi_z})^2}{\left(E\|\Xi_z^{1/2} H\|_2\right)^2} \lesssim n \log(4/\delta) \frac{\mathrm{tr}(\Xi_z^2)}{\mathrm{tr}(\Xi_z)^2} = \frac{n\log(4/\delta)}{R(\Xi_z)}$$

As $n/R(\Xi_z)$ and $\omega^\top \Sigma_u^+ \Xi_z \Sigma_u^+$ converge to zero, we have $\lim_{n\to\infty} \eta_1 = 0$. □

Lemma 35. *Suppose* $\lim_{n\to\infty} \omega^\top \Sigma_u^+ \Xi_z \Sigma_u^+ \omega = 0$ *holds. The fourth condition of Theorem 33 implies* $\lim_{n\to\infty} \eta_2 = 0$.

Proof of Lemma 35. By the definition of η_2, we have

$$\eta_2 := \sqrt{\frac{(\mathbb{E}\|\Xi_z^{1/2} H\|_2)^2}{n}\left(1+\sqrt{\frac{2\log(8/\delta)}{r_{\|\cdot\|_2}(\Xi_z)}}\right)\|\Sigma_u^+\omega\|_2^2 + \|\Xi_z^{1/2}\Sigma_u^+\omega\|_2^2}.$$

By trivial calculation, for any covariance matrix Σ, it holds that

$$\operatorname{tr}(\Sigma) = \mathbb{E}\|\Sigma^{1/2}H\|_2^2 = (\mathbb{E}\|\Sigma^{1/2}H\|_2)^2 + \operatorname{Var}(\|\Sigma^{1/2}H\|_2) \geq (\mathbb{E}\|\Sigma^{1/2}H\|_2)^2.$$

From the result of (92), we have

$$\eta_2 \lesssim \sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}\left(1+\sqrt{\frac{\log(1/\delta)}{r(\Xi_z)}}\right)\|\Sigma_u^+\omega\|_2^2 + \|\Xi_z^{1/2}\Sigma_u^+\omega\|_2^2}.$$

As $R(\Xi_z) \to \infty$ implies $r(\Xi_z) \to \infty$, we have the following result:

$$\left(1+\sqrt{\frac{\log(1/\delta)}{r(\Xi_z)}}\right) \to 1.$$

Moreover, the conditions

$$\lim_{n\to\infty} \frac{\|\Sigma_u^+\omega\|_2}{\widetilde{\sigma}}\sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}} = 0 \quad \text{and} \quad \lim_{n\to\infty} \omega^\top \Sigma_u^+ \Xi_z \Sigma_u^+ \omega = 0$$

lead to the conclusion $\lim_{n\to\infty} \eta_2 = 0$. □

Theorem 5 (Sufficient conditions: Non-Orthogonal Case) *Under Assumption 1, let $\widehat{\theta}$ be the ridgeless estimator. Suppose that as n goes to ∞, the covariance splitting $\Sigma_x = \Xi_z + \Sigma_u$ satisfies the following conditions:*

(i) (Small large-variance dimension.)

$$\lim_{n\to\infty} \frac{\operatorname{rank}(\Sigma_u)}{n} = 0.$$

(ii) (Large effective dimension.)

$$\lim_{n\to\infty} \frac{n}{R(\Xi_z)} = 0.$$

(iii) (No aliasing condition.)

$$\lim_{n\to\infty} \|\theta_0\|_2 \sqrt{\frac{\operatorname{tr}(\Xi_z)}{n}} = 0.$$

(iv) (Condition for the minimal interpolation of instrumental variable in the non-orthogonal case)

$$\lim_{n\to\infty} \frac{\|\Sigma_u^+ \omega\|_2}{\widetilde{\sigma}} \sqrt{\frac{\mathrm{tr}(\Xi_z)}{n}} = 0.$$

(v) (Non-orthogonality)

(a)

$$\lim_{n\to\infty} \frac{n}{R(\Xi_z)} \frac{\mathrm{tr}(\Sigma_u \Xi_z)}{\mathrm{tr}(\Xi_z^2)} = 0.$$

(b)

$$\lim_{n\to\infty} \omega^\top \Sigma_u^+ \Xi_z \Sigma_u^+ \omega = 0.$$

Then, $\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2$ converges to 0 in probability.

Proof of Theorem 5. By Lemma 34, it holds that from conditions (ii) and (v)(b) that $\lim_{n\to\infty} \eta_1 = 0$. We also have $\lim_{n\to\infty} \eta_2 = 0$ by Lemma 35. Therefore, from Theorem 33, $\|\widehat{\theta} - \theta_0\|_{\Xi_z}^2$ converges to 0 in probability. □

Appendix G: Supportive result

Proposition 36. *The necessary condition for $\lim_{n\to\infty} \mathrm{rank}(\Sigma_u)/n = 0$ in Theorem 2 is*

$$\mathrm{rank}(\Sigma_u) \geq p - \min\{\mathrm{rank}(\Sigma_z), \mathrm{rank}(\Pi_0)\}.$$

Proof of Proposition 36. As we have $\mathrm{rank}(A) = \mathrm{rank}(A^\top) = \mathrm{rank}(AA^\top) = \mathrm{rank}(A^\top A)$ for any matrix A, we have $\mathrm{rank}(\Pi_0 \Sigma_z^{1/2}) = \mathrm{rank}(\Xi_z)$. From the property of the matrix, we have

$$\mathrm{rank}(\Pi_0 \Sigma_z^{1/2}) \leq \min\{\mathrm{rank}(\Sigma_z^{1/2}), rank(\Pi_0)\}.$$

Under Assumption 2, we have

$$\mathrm{rank}(\Xi_z) + \mathrm{rank}(\Sigma_u) = p.$$

Therefore, we obtain the statement. □

Lemma 37. *Assume Assumption 2 holds. Then, we obtain the following covariance splitting:*

$$\Sigma_x = \Xi_z + \Sigma_u,$$

where Ξ_z and Σ_u are positive semidefinite matrices and subspaces generated from $\Xi_z$andΣ_u, which are orthogonal.

Proof of Lemma 37. By construction, we have

$$X_i = \Pi_0 Z_i + u_i,$$

where $E[u_i|Z_i] = 0$. Therefore, we have

$$\Sigma_x = E[X_i X_i^\top] = \Pi_0 E[Z_i Z_i^\top]\Pi_0^\top + E[u_i u_i^\top] = \Xi_z + \Sigma_u.$$

By construction, clearly Ξ_z, Σ_u are positive semidefinite. By Assumption 2, subspaces generated by Ξ_z, Σ_u are orthogonal. Therefore, we have $\Sigma_x = \Xi_z + \Sigma_u$. □

Lemma 38. *Under Assumptions 1 and 2, we have the following inequality:*

$$\|\rho\|_2^2 = \|\omega\|^2_{\Sigma_u^+} \leq \sigma^2.$$

Furthermore, the covariance matrix of $(W_{1,i}, W_{2,i}, \xi_i)^\top$ *in* (17) *is positive semi-definite.*

Proof of Lemma 38. First, we clarify the necessary and sufficient condition of positive semi-definiteness of the covariance matrix of $(X_i, \xi_i)^\top$. The definition of positive semi-definiteness is

$$(a, b)\begin{pmatrix} \Sigma_x & \Sigma_u^{1/2}\rho \\ \rho^\top \Sigma_u^{1/2} & \sigma^2 \end{pmatrix}(a, b)^\top \geq 0,$$

for any $a \in \mathbb{R}^p$ and $b \in \mathbb{R}$. By trivial calculation, we have

$$(a, b)\begin{pmatrix} \Sigma_x & \Sigma_u^{1/2}\rho \\ \rho^\top \Sigma_u^{1/2} & \sigma^2 \end{pmatrix}(a, b)^\top = a^\top \Sigma_x a + 2ba^\top \Sigma_u^{1/2}\rho + b^2\sigma^2. \tag{82}$$

By solving the first order condition of (82) with respect to $a \in \mathbb{R}^p$, we have the solution $a^* = -b\Sigma_x^{-1}\Sigma_u^{1/2}\rho$. By substituting a^* into (82), it holds that

$$b^2\rho^\top \Sigma_u^{1/2}\Sigma_x^{-1}\Sigma_u^{1/2}\rho - 2b^2\rho^\top\Sigma_u^{1/2}\Sigma_x^{-1}\Sigma_u^{1/2}\rho + b^2\sigma^2 = b^2(\sigma^2 - \rho^\top\Sigma_u^{1/2}\Sigma_x^{-1}\Sigma_u^{1/2}\rho).$$

Therefore, $(\sigma^2 - \rho^\top\Sigma_u^{1/2}\Sigma_x^{-1}\Sigma_u^{1/2}\rho) \geq 0$ is the sufficient and necessary condition for positive semi-definiteness. Likewise, we obtain $(\sigma^2 - \rho^\top\rho) \geq 0$ as the sufficient and necessary condition for positive semi-definiteness for the covariance matrix of $(W_{1,i}, W_{2,i}, \xi_i)^\top$.

Finally, under Assumptions 1 and 2, we show that positive semi-definiteness of the covariance matrix of $(X_i, \xi_i)^\top$ implies that the covariance matrix of $(W_{1,i}, W_{2,i}, \xi_i)^\top$ is positive semi-definite. As ρ is defined as the one that has the minimum norm subject to $\omega = \Sigma_u^{1/2}\rho$, $\rho := \operatorname{argmin}\{\|b\| : \omega = \Sigma_u^{1/2}b\}$. By the property of the generalized inverse matrix, we have

$$\rho = (\Sigma_u^{1/2})^+\omega.$$

Under Assumption 2, $\Sigma_u\Xi_z = 0$. As Σ_u and Ξ_z are symmetric, we have

$$\Sigma_x(\Sigma_u^+ + \Xi_z^+) = (\Sigma_u + \Xi_z)(\Sigma_u^+ + \Xi_z^+)$$

$$
\begin{aligned}
&= \Sigma_u \Sigma_u^+ + \Sigma_u \Xi_z^+ + \Xi_z \Sigma_u^+ + \Xi_z \Xi_z^+ \\
&= I_u + \Sigma_u \Xi_z^\top (\Xi_z \Xi_z^\top)^+ + \Xi_z \Sigma_u^\top (\Sigma_u \Sigma_u^\top)^+ + (I_p - I_u) \\
&= I_p.
\end{aligned}
$$

Hence, $\Sigma_x^{-1} = \Sigma_u^+ + \Xi_z^+$. Then, we have

$$\rho^\top \Sigma_u^{1/2} \Sigma_x^{-1} \Sigma_u^{1/2} \rho = \omega^\top \Sigma_x^{-1} \omega = \omega^\top \Sigma_u^+ \omega + \omega^\top \Xi_z^+ \omega = \omega^\top \Sigma_u^+ \omega = \rho^\top \rho.$$

The third equality holds because $\omega = \Sigma_u^{1/2} \rho$ and Σ_u is orthogonal to Ξ_z. The above discussion suggests

$$\sigma^2 - \rho^\top \Sigma_u^{1/2} \Sigma_x^{-1} \Sigma_u^{1/2} \rho = \sigma^2 - \rho^\top \rho.$$

Therefore, by the positive semi-definiteness of the covariance matrix of $(X_i, \xi_i)^\top$, we have

$$\sigma^2 - \rho^\top \rho = \sigma^2 - \omega^\top \Sigma_x^{-1} \omega \geq 0. \qquad \square$$

Proof of Proposition 3. By Theorem 2 (1) in Bartlett *et al.* (2020), the first and second conditions of Definition 2 are satisfied. As $\mathrm{tr}(\Xi_z)$ converges to a finite value, the third condition also holds under the assumption $\|\theta_0\|_2 = o(\sqrt{n})$.

We show the condition in Theorem 2 is satisfied under the setting of Proposition 3. By the setting, we have

$$\frac{1}{n} \|\Sigma_u^+ \omega\|_2^2 = \frac{1}{n} \sum_{i=1}^{k_n^*} i^2 \log^{2\beta}(i+1) \cdot (U\omega)_i^2 \lesssim \frac{1}{n} \sum_{i=1}^{k_n^*} \frac{i^2 \log^{2\beta}(i+1)}{i^2 \log^{2\beta}(i+1)} = \frac{k_n^*}{n}.$$

As $\lim_{n\to\infty} k_n^*/n = 0$ holds, we have $\lim_{n\to\infty} \|\Sigma_u^+ \omega\|_2^2 / n = 0$. □

Proof of Proposition 4. By Theorem 2 (2) in Bartlett *et al.* (2020), the first and second conditions of Definition 2 are satisfied.

We prove θ_0 and $\mathrm{tr}(\Xi_z)$ satisfy the third condition stated in Definition 2. By the definition of the matrices, we have

$$\mathrm{tr}(\Xi_z) \leq \mathrm{tr}(\Sigma_x) = \sum_{i=1}^{p} (\gamma_i + \varepsilon_n) = \sum_{i=1}^{p} \gamma_i + p\varepsilon_n. \tag{83}$$

As $p\varepsilon_n$ is equal to $ne^{-o(n)}$, the second term of (83) converges to zero. It also holds from the definition that

$$\sum_{i=1}^{p} \gamma_i \lesssim \sum_{i=1}^{p} \exp(-i/\tau).$$

As $\sum_{i=1}^{\infty} \exp(-i/\tau)$ is finite, $\mathrm{tr}(\Xi_z)$ is also finite. Therefore, $\lim_{n\to\infty} \|\theta_0\|_2 \sqrt{\mathrm{tr}(\Xi_z)/n} = 0$ holds.

Finally, we show the setting in Proposition 4 satisfies the condition stated in Theorem 2. By the assumption of Proposition 4, we have

$$\frac{1}{n}\|\Sigma_u^+\omega\|_2^2 = \frac{1}{n}\sum_{i=1}^{k_n^*}\frac{(U\omega)_i^2}{(\gamma_i+\varepsilon_n)^2} \le \frac{1}{n}\sum_{i=1}^{k_n^*}\frac{(U\omega)_i^2}{\gamma_i^2} \lesssim \frac{1}{n}\sum_{i=1}^{k_n^*}\frac{\exp(-2i/\tau)}{\exp(-2i/\tau)} = \frac{k_n^*}{n}.$$

As k_n^*/n goes to zero, $\lim_{n\to\infty}\|\Sigma_u^+\omega\|_2^2/n = 0$ holds. □

Proof of Proposition 7. $\mathrm{rank}(\Sigma_u)$ in Proposition 7 is the same as $\mathrm{rank}(\Sigma_u)$ defined in Proposition 3. Hence, by Theorem 2 (1) in Bartlett *et al.* (2020), the first condition of Definition 2 is satisfied.

To satisfy the second condition of Definition 2, we prove $n/R(\Xi_z)$ goes to zero as n goes to infinity. By the definition of $R(\Xi_z)$, we have

$$\begin{aligned}
\frac{n}{R(\Xi_z)} &= n\frac{\mathrm{tr}(\Xi_z^2)}{(\mathrm{tr}(\Xi_z))^2}\\
&= n\frac{\frac{1}{n^{2\alpha}}\sum_{i=1}^{k_n^*}\lambda_i^2+\sum_{i=k_n^*+1}^{p}\lambda_i^2}{\left(\frac{1}{n^{\alpha}}\sum_{i=1}^{k_n^*}\lambda_i+\sum_{i=k_n^*+1}^{p}\lambda_i\right)^2}\\
&= \frac{\frac{1}{n^{2\alpha}}\sum_{i=1}^{k_n^*}\lambda_i^2+\sum_{i=k_n^*+1}^{p}\lambda_i^2}{\sum_{i=k_n^*+1}^{p}\lambda_i^2}\cdot n\frac{\sum_{i=k_n^*+1}^{p}\lambda_i^2}{\left(\sum_{i=k_n^*+1}^{p}\lambda_i\right)^2}\\
&\quad\cdot\frac{\left(\sum_{i=k_n^*+1}^{p}\lambda_i\right)^2}{\left(\frac{1}{n^{\alpha}}\sum_{i=1}^{k_n^*}\lambda_i+\sum_{i=k_n^*+1}^{p}\lambda_i\right)^2}\\
&= \left(\frac{\sum_{i=1}^{k_n^*}\lambda_i^2}{n^{2\alpha}\sum_{i=k_n^*+1}^{p}\lambda_i^2}+1\right)\cdot n\frac{\sum_{i=k_n^*+1}^{p}\lambda_i^2}{\left(\sum_{i=k_n^*+1}^{p}\lambda_i\right)^2}\cdot\left(\frac{\sum_{i=1}^{k_n^*}\lambda_i}{n^{\alpha}\sum_{i=k_n^*+1}^{p}\lambda_i}+1\right)^{-2}.
\end{aligned}\tag{84}$$

As $\lambda_i = Ci^{-1}\log^{-\beta}(i+1)$ where $\beta>1$, it holds that

$$\lim_{n\to\infty}\sum_{i=1}^{k_n^*}\lambda_i^2 \le \lim_{n\to\infty}\left(\sum_{i=1}^{k_n^*}\lambda_i\right)^2 = \left(\lim_{n\to\infty}\sum_{i=1}^{k_n^*}\lambda_i\right)^2 < \infty. \tag{85}$$

Moreover, we have

$$n^{2\alpha}\sum_{i=k_n^*+1}^{p}\lambda_i^2 \gtrsim n^{2\alpha}(p-k_n^*)\left(\frac{1}{p\log^{\beta}(p+1)}\right)^2 = \frac{1}{q}\left(1-\frac{1}{q}\frac{k_n^*}{n}\right)\left(\frac{n^{\frac{2\alpha-1}{2\beta}}}{\log(qn+1)}\right)^{2\beta}. \tag{86}$$

As $\left(n^{\frac{2\alpha-1}{\beta}}/\log(qn+1)\right)^{2\beta}$ diverges to infinity and k_n^*/n converges to zero as n goes to infinity, it holds from (86) that $n^{2\alpha}\sum_{i=k_n^*+1}^{p}\lambda_i^2$ diverges to infinity.

Therefore, we have

$$\lim_{n\to\infty}\left(\frac{\sum_{i=1}^{k_n^*}\lambda_i^2}{n^{2\alpha}\sum_{i=k_n^*+1}^{p}\lambda_i^2}+1\right)=1. \tag{87}$$

In a similar way to (85) and (86), we obtain

$$\lim_{n\to\infty}\sum_{i=1}^{k_n^*}\lambda_i<\infty,\ \text{and}\ n^{\alpha}\sum_{i=k_n^*+1}^{p}\lambda_i\gtrsim\left(1-\frac{1}{q}\frac{k_n^*}{n}\right)\left(\frac{n^{\frac{\alpha}{\beta}}}{\log(qn+1)}\right)^{\beta}.$$

Hence, we have

$$\lim_{n\to\infty}\left(\frac{\sum_{i=1}^{k_n^*}\lambda_i}{n^{\alpha}\sum_{i=k_n^*+1}^{p}\lambda_i}+1\right)^{-2}=1. \tag{88}$$

Combining the result in the proof of Proposition 3, $\lim_{n\to\infty} n\sum_{i=k_n^*+1}^{p}\lambda_i^2/\left(\sum_{i=k_n^*+1}^{p}\lambda_i\right)^2=0$, with (87) and (88), we have

$$\lim_{n\to\infty}\frac{n}{R(\Xi_z)}=0.$$

We show the third condition in Definition 2 is satisfied under the setting of Proposition 7. By the definition of λ_i, we have

$$\mathrm{tr}(\Xi_z)\le\mathrm{tr}(\Sigma_x)<\infty.$$

Hence, the third condition of Definition 2 clearly holds under the assumption $\|\theta_0\|_2=o(\sqrt{n})$.

We prove Σ_u and ω satisfy the first condition stated in Theorem 5. As $\lim_{n\to\infty}(\sigma^2-\|\omega\|_{\Sigma_u^+}^2)>0$ holds by assumption, it is sufficient to show $\lim_{n\to\infty}\|\Sigma_u^+\omega\|_2\sqrt{\mathrm{tr}(\Xi_z)/n}=0$. By the assumption of Proposition 7, we have

$$\begin{aligned}\frac{1}{n}\|\Sigma_u^+\omega\|_2^2&=\frac{1}{n}\sum_{i=1}^{k_n^*}\frac{n^{2\alpha}}{(n^{\alpha}-1)^2}i^2\log(i+1))^{2\beta}(U\omega)_i^2\\&\lesssim\frac{1}{n}\frac{1}{1-2/n^{\alpha}+1/n^{2\alpha}}\sum_{i=1}^{k_n^*}\frac{i^2\log(i+1))^{2\beta}}{i^2\log(i+1))^{2\beta}}.\end{aligned}$$

From the discussion in the proof of Proposition 3, $\lim_{n\to\infty}\|\Sigma_u^+\omega\|_2^2/n=0$ holds.

For the second condition of Theorem 5, we need to show $\mathrm{tr}(\Sigma_u\Xi_z)/\mathrm{tr}(\Xi_z^2)$ is finite. By the definition of Σ_u and Ξ_z, we have

$$\frac{\mathrm{tr}(\Sigma_u\Xi_z)}{\mathrm{tr}(\Xi_z^2)}=\frac{\left(\frac{n^{\alpha}-1}{n^{2\alpha}}\sum_{i=1}^{k_n^*}\frac{1}{i^2\log(i+1)^{2\beta}}\right)}{\frac{1}{n^{2\alpha}}\sum_{i=1}^{k_n^*}\frac{1}{i^2\log(i+1)^{2\beta}}+\sum_{i=k_n^*+1}^{p}\frac{1}{i^2\log(i+1)^{2\beta}}}$$

$$= \left(\frac{1}{n^\alpha - 1} + \frac{\frac{n^{2\alpha}}{n^\alpha - 1} \sum_{i=k_n^*+1}^{p} \frac{1}{i^2 \log(i+1)^{2\beta}}}{\left(\sum_{i=1}^{k_n^*} \frac{1}{i^2 \log(i+1)^{2\beta}} \right)} \right)^{-1}.$$

By trivial calculation, we have

$$\begin{aligned} \frac{n^{2\alpha}}{n^\alpha - 1} \sum_{i=k_n^*+1}^{p} \frac{1}{i^2 \log(i+1)^{2\beta}} &= \frac{n^\alpha}{n^\alpha - 1} n^\alpha \sum_{i=k_n^*+1}^{p} \frac{1}{i^2 \log(i+1)^{2\beta}} \\ &\gtrsim \frac{n^\alpha}{n^\alpha - 1} \left(1 - \frac{1}{q} \frac{k_n^*}{n} \right) \left(\frac{n^{\frac{\alpha-1}{2\beta}}}{\log(qn+1)} \right)^{2\beta}. \end{aligned}$$

As $\left(n^{\frac{\alpha-1}{\beta}} / \log(qn+1) \right)^{2\beta}$ diverges to infinity and k_n^*/n converges to zero as n goes to infinity, it holds that $\lim_{n \to \infty} \mathrm{tr}(\Sigma_u \Xi_z)/\mathrm{tr}(\Xi_z^2) = 0$.

Finally, we show the setting in Proposition 7 satisfies the last condition stated in Theorem 5. By trivial calculation, we have

$$\omega^\top \Sigma_u^+ \Xi_z \Sigma_u^+ \omega \lesssim \frac{1}{n^\alpha - 1} \sum_{i=1}^{k_n^*} \frac{1}{i \log^\beta (i+1))} \lesssim \frac{k_n^*}{n^\alpha}.$$

Therefore, $\lim_{n \to \infty} \omega^\top \Sigma_u^+ \Xi_z \Sigma_u^+ \omega = 0$ holds. □

Lemma 39 (Application of Theorem 5.1.4 in Vershynin (2018))**.** *Assume S is a subspace of dimension d in $\mathbb{R}^n$ where $n \geq 4$. Let P_S denote the orthogonal projection onto S and let V denote a spherically symmetric random variable. Then, with at least $1 - \delta$ probability, we have*

$$\frac{\|P_S V\|_2}{\|V\|_2} \leq \sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(2/\delta)}{n}}. \tag{89}$$

From this inequality, we also have

$$|\langle s, V \rangle| = |\langle s, P_S V \rangle| \leq \|s\|_2 \|P_S V\|_2 \leq \|s\|_2 \|V\|_2 \left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(2/\delta)}{n}} \right). \tag{90}$$

Lemma 40 (Theorem 3.1.1 in Vershynin (2018))**.** *Suppose that $Z \sim N(0, I_n)$. Then,*

$$\mathbb{P}(|\|Z\|_2 - \sqrt{n}| \geq t) \leq 4e^{-t^2/4}.$$

Lemma 41 (Lemma 9 in Koehler *et al.* (2021))**.** *Let H be normally distributed with mean zero and variance I_d, that is, $H \sim N(0, I_d)$. For any covariance matrix Σ, it holds that*

$$\left(E\|\Sigma^{1/2} H\|_2 \right)^2 \geq \left(1 - \frac{1}{r(\Sigma)} \right) \mathrm{tr}(\Sigma) \tag{91}$$

and

$$\frac{1}{\operatorname{tr}(\Sigma)} \geq \left(1 - \sqrt{\frac{8}{r(\Sigma)}}\right) E\left[\frac{1}{H^T \Sigma H}\right].$$

Consequently, it holds that

$$r(\Sigma) - 1 \leq r_{\|\cdot\|_2}(\Sigma) \leq r(\Sigma) \tag{92}$$

and

$$1 - \frac{4}{\sqrt{r(\Sigma)}} \leq \frac{R_{\|\cdot\|_2}(\Sigma)}{R(\Sigma)} \leq \left(1 - \sqrt{\frac{8}{r(\Sigma^2)}}\right)^{-1},$$

where we define

$$r(\Sigma) = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|_{op}}, \quad R(\Sigma) = \frac{\operatorname{tr}(\Sigma)^2}{\operatorname{tr}(\Sigma^2)}, \quad r_{\|\cdot\|}(\Sigma) = \left(\frac{E\|\Sigma^{1/2}H\|_*}{\sup_{\|u\|\leq 1}\|u\|_\Sigma}\right)^2,$$

and

$$R_{\|\cdot\|}(\Sigma) = \left(\frac{E\|\Sigma^{1/2}H\|_*}{E\|v^*\|_\Sigma}\right)^2.$$

Lemma 42 (Lemma 10 in Koehler *et al.* (2021))**.** *Let H be normally distributed with mean zero and variance I_d, that is, $H \sim N(0, I_d)$. For any covariance matrix Σ, it holds that with probability at least $1 - \delta$*

$$1 - \frac{\|\Sigma^{1/2}H\|_2^2}{\operatorname{tr}(\Sigma)} \lesssim \frac{\log(4/\delta)}{\sqrt{R(\Sigma)}} \tag{93}$$

and

$$\|\Sigma H\|_2^2 \lesssim \log(4/\delta)\operatorname{tr}(\Sigma^2).$$

Therefore, provided that $R(\Sigma) \gtrsim \log(4/\delta)^2$, it holds that

$$\left(\frac{\|\Sigma H\|_2}{\|\Sigma^{1/2}H\|_2}\right)^2 \lesssim \log(4/\delta)\frac{\operatorname{tr}(\Sigma^2)}{\operatorname{tr}(\Sigma)}. \tag{94}$$

Theorem 43 (Theorem 3.25 in van Handel (2014))**.** *Assume f is L-Lipschitz continuous with respect to the Euclidean norm with $L > 0$, that is, for $f : \mathbb{R}^n \to \mathbb{R}$,*

$$|f(x) - f(y)| \leq L\|x - y\|_2,$$

for all $x, y \in \mathbb{R}^n$. Then, we have

$$\mathbb{P}(|f(Z) - E[f(Z)]| \geq t) \leq 2e^{-t^2/2L^2}, \tag{95}$$

where $Z \sim N(0, I_n)$.

Funding

T.Tsuda was supported by JSPS Grant-in-Aid for JSPS Research Fellows (23KJ0713). M.Imaizumi was supported by JSPS KAKENHI (21K11780), JST CREST (JPMJCR21D2), and JST FOREST (JPMJFR216I).

References

Ai, C. and Chen, X. (2003) Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica*, **71**, 1795–1843. MR2015420

Andrews, I., Stock, J. H. and Sun, L. (2019) Weak instruments in instrumental variables regression: Theory and practice, *Annual Review of Economics*, **11**, 727–753.

Baiocchi, M., Cheng, J. and Small, D. S. (2014) Instrumental variable methods for causal inference, *Statistics in medicine*, **33**, 2297–2340. MR3257582

Bartlett, P. L., Long, P. M., Lugosi, G. and Tsigler, A. (2020) Benign overfitting in linear regression, *Proceedings of the National Academy of Sciences*, **117**, 30063–30070. MR4263288

Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off, *Proceedings of the National Academy of Sciences*, **116**, 15849–15854. MR3997901

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012) Sparse models and methods for optimal instruments with an application to eminent domain, *Econometrica*, **80**, 2369–2429. MR3001131

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017) Program evaluation and causal inference with high-dimensional data, *Econometrica*, **85**, 233–298. MR3611771

Belloni, A., Chernozhukov, V. and Hansen, C. (2010) Lasso methods for gaussian instrumental variables models, *arXiv preprint arXiv:1012.1297*.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014) High-dimensional methods and inference on structural and treatment effects, *Journal of Economic Perspectives*, **28**, 29–50.

Belloni, A., Hansen, C. and Newey, W. (2022) High-dimensional linear models with many endogenous variables, *Journal of Econometrics*, **228**, 4–26. MR4393285

Bunea, F., Strimas-Mackey, S. and Wegkamp, M. H. (2022) Interpolating predictors in high-dimensional factor regression., *Journal of Machine Learning Research*, **23**, 10–1. MR4420735

Chen, X. and Pouzo, D. (2012) Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals, *Econometrica*, **80**, 277–321. MR2920758

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal*, **21**, C1–C68. MR3769544

Chernozhukov, V., Hansen, C. and Spindler, M. (2015a) Post-selection and post-regularization inference in linear models with many controls and instruments, *American Economic Review*, **105**, 486–90.

Chernozhukov, V., Hansen, C. and Spindler, M. (2015b) Valid post-selection and post-regularization inference: An elementary, general approach, *Annual Review of Economics*, **7**, 649–688.

Chernozhukov, V., Hansen, C. and Spindler, M. (2016) hdm: High-dimensional metrics, *The R Journal*, **8**, 185.

Dikkala, N., Lewis, G., Mackey, L. and Syrgkanis, V. (2020) Minimax estimation of conditional moment models, *Advances in Neural Information Processing Systems*, **33**, 12248–12262.

Dobriban, E. and Wager, S. (2018) High-dimensional asymptotics of prediction: Ridge regression and classification, *The Annals of Statistics*, **46**, 247–279. MR3766952

Fan, J. and Liao, Y. (2014) Endogeneity in high dimensions, *The Annals of Statistics*, **42**, 872. MR3210990

Frei, S., Chatterji, N. S. and Bartlett, P. (2022) Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data, in *Conference on Learning Theory*, PMLR, pp. 2668–2703.

Gautier, E. and Rose, C. (2011) High-dimensional instrumental variables regression and confidence sets, *arXiv preprint arXiv:1105.2454*.

Gautier, E. and Tsybakov, A. B. (2013) Pivotal estimation in high-dimensional regression via linear programming, in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, Springer, pp. 195–204. MR3236866

Gold, D., Lederer, J. and Tao, J. (2020) Inference for high-dimensional instrumental variables regression, *Journal of Econometrics*, **217**, 79–111. MR4093746

Han, Q. and Shen, Y. (2023) Universality of regularized regression estimators in high dimensions, *The Annals of Statistics*, **51**, 1799–1823. MR4658577

Hastie, T., Montanari, A., Rosset, S. and Tibshirani, R. J. (2022) Surprises in high-dimensional ridgeless least squares interpolation, *The Annals of Statistics*, **50**, 949–986. MR4404925

Hill, B. M. (1975) A simple general approach to inference about the tail of a distribution, *The Annals of Statistics*, pp. 1163–1174. MR0378204

Koehler, F., Zhou, L., Sutherland, D. J. and Srebro, N. (2021) Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting, *Advances in Neural Information Processing Systems*, **34**, 20657–20668.

Koltchinskii, V. and Lounici, K. (2017) Concentration inequalities and moment bounds for sample covariance operators, *Bernoulli*, **23**, 110–133. MR3556768

Li, Z., Su, W. J. and Sejdinovic, D. (2022) Benign overfitting and noisy features, *Journal of the American Statistical Association*, pp. 1–13. MR4681627

Montanari, A. and Saeed, B. N. (2022) Universality of empirical risk minimization, in *Conference on Learning Theory*, PMLR, pp. 4310–4312.

Nakakita, S. and Imaizumi, M. (2022) Benign overfitting in time series linear model with over-parameterization, *arXiv preprint arXiv:2204.08369*.

Newey, W. K. and Powell, J. L. (2003) Instrumental variable estimation of nonparametric models, *Econometrica*, **71**, 1565–1578. MR2000257

Rockafellar, R. T. (1997) *Convex analysis*, vol. 11, Princeton university press. MR1451876

Söderström, T. and Stoica, P. (2002) Instrumental variable methods for system identification, *Circuits, Systems and Signal Processing*, **21**, 1–9. MR1889846

Stock, J. H., Wright, J. H. and Yogo, M. (2002) A survey of weak instruments and weak identification in generalized method of moments, *Journal of Business & Economic Statistics*, **20**, 518–529. MR1973801

Thrampoulidis, C., Abbasi, E. and Hassibi, B. (2018) Precise error analysis of regularized m-estimators in high dimensions, *IEEE Transactions on Information Theory*, **64**, 5592–5628. MR3832326

Thrampoulidis, C., Oymak, S. and Hassibi, B. (2015) Regularized linear regression: A precise analysis of the estimation error, in *Conference on Learning Theory*, PMLR, pp. 1683–1709.

Tsigler, A. and Bartlett, P. L. (2023) Benign overfitting in ridge regression, *Journal of Machine Learning Research*, **24**, 1–76. MR4583284

van Handel, R. (2014) Probability in high dimension: Lecture notes.

Vershynin, R. (2018) *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press. MR3837109