Causal Inference with Corrupted Data: Measurement Error, Missing Values, Discretization, and Differential Privacy

Anish Agarwal
Amazon Core AI

Rahul Singh*
MIT Economics

Original draft: July 6, 2021. This draft: November 7, 2022

Abstract

The US Census Bureau will deliberately corrupt data sets derived from the 2020 US Census in an effort to maintain privacy, suggesting a painful trade-off between the privacy of respondents and the precision of economic analysis. To investigate whether this trade-off is inevitable, we formulate a semiparametric model of causal inference with high dimensional corrupted data. We propose a procedure for data cleaning, estimation, and inference with data cleaning-adjusted confidence intervals. We prove consistency, Gaussian approximation, and semiparametric efficiency by finite sample arguments, with a rate of $n^{-1/2}$ for semiparametric estimands that degrades gracefully for nonparametric estimands. Our key assumption is that the true covariates are approximately low rank, which we interpret as approximate repeated measurements and validate in the Census. In our analysis, we provide nonasymptotic theoretical contributions to matrix completion, statistical learning, and semiparametric statistics. Calibrated simulations verify the coverage of our data cleaning-adjusted confidence intervals and demonstrate the relevance of our results for 2020 Census data.

Keywords: disclosure avoidance, heterogeneous treatment effect, matrix completion, semiparametric efficiency

*We thank Alberto Abadie, Isaiah Andrews, Josh Angrist, Dmitry Arkhangelsky, David Autor, Abhijit Banerjee, David Bruns-Smith, Victor Chernozhukov, Jon Cohen, Rachel Cummings, Peng Ding, Avi Feller, Bryan Graham, Florian Gunsilius, Chris Harshaw, Skip Hirshberg, Guido Imbens, Simon Jaeger, Michael Jansson, Michael Jordan, Patrick Kline, Lihua Lei, Lexin Li, Anna Mikusheva, Ismael Mourifie, Sendhil Mullainathan, Whitney Newey, Betsy Ogburn, Demian Pouzo, Jim Powell, Ashesh Rambachan, James Robins, Jesse Rothstein, Andrea Rotnitzky, Frank Schilbach, Vira Semenova, Devavrat Shah, Garima Sharma, Dennis Shen, Nakul Singh, Jann Spiess, Liyang Sun, Tavneet Suri, Vasilis Syrgkanis, Panos Toulis, Suhas Vijaykumar, and Bin Yu for helpful comments. We thank Leon Deng, Ajinkya Gundaria, and Caleb Rollins for their contributions through the MIT Undergraduate Research Opportunity Program. Rahul Singh thanks the Jerry Hausman Dissertation Fellowship. Part of this work was done while both authors were visiting the Simons Institute for the Theory of Computing.

1 Introduction

1.1 Motivation and research question

The 2010 US Census inadvertently revealed too much information. In a simulated hack, researchers at the Census Bureau were able to re-identify between 52 and 179 million respondents from ostensibly anonymous summary tables (Hawes, 2021). To protect privacy, the Bureau will inject synthetic noise into forthcoming summary tables of the 2020 Census (Jarmin, 2019) and coarsen wage microdata in the Current Population Survey (CPS) (Benedetto et al., 2022). Techniques like these, called privacy mechanisms in computer science, guarantee a particular notion of privacy called *differential privacy* via deliberate data corruption (Dwork et al., 2006). Differential privacy is widely implemented in the technology sector, e.g. Apple iOS and Google Chrome data. Due to its recent adoption in the government sector, researchers in economics and statistics have warned of a looming trade-off: the privacy of respondents versus the precision of economic analysis (Duchi et al., 2018; Abowd and Schmutte, 2019; Hotz et al., 2022).

We study differential privacy and discretization as modern challenges for causal inference. Economic data continue to suffer from classical types of data corruption in the form of missing values and measurement error (Griliches, 1986). Therefore we analyze a class of data corruptions that encompasses both modern and classical issues *simultaneously*, while remaining agnostic about their relative magnitudes. Our research question is not only how to conduct causal inference, but if it is even possible, with high dimensional economic data that suffer from measurement error, missing values, discretization, and differential privacy.

1.2 Contributions

We study a broad class of causal parameters, including semiparametric estimands such as the average treatment effect, the local average treatment effect, and the average elasticity, as well as nonparametric estimands such as heterogeneous treatment effects, in a nonlinear and high dimensional setting. Our main contribution is a procedure for automatic data cleaning, causal estimation, and finally inference with confidence intervals that account for the bias and variance consequences of data cleaning. Our key assumption is that the true

covariates are approximately low rank, which we validate for US Census data and interpret from a causal perspective. In particular, we argue that covariates collected from the Census include approximate repeated measurements—e.g. disability benefits, medical benefits, and unemployment benefits—which implies that they are approximately low rank. This phenomenon powers our entire analysis. There are three key aspects of our contribution.

First, our relatively simple procedure adapts to the *type* and *level* of data corruption. The same code works in a variety of settings, allowing for classical corruptions such as measurement error and missing values as well as modern corruptions such as discretization and differential privacy. Crucially, the researcher does not need to know in advance the corruption distribution, e.g. its parametric form or covariance structure, and in this way we depart from the error-in-variable Lasso and Dantzig literatures (Loh and Wainwright, 2012; Rosenbaum and Tsybakov, 2013; Datta and Zou, 2017). We depart from previous work on principal component regression (Agarwal et al., 2021, 2020a) by proposing new variants for causal inference. We propose an error-in-variable balancing weight that adapts to the causal parameter of interest—a natural yet original solution. In particular, the error-in-variable balancing weight appears to be the first of its kind. In the way our procedure handles missing values, it shares principles with multiple imputation (Rubin, 1976). In the appendix, we extend our method to handle outcome attrition (Heckman, 1979; Hausman and Wise, 1979; Das et al., 2003; Huber, 2014; Bia et al., 2020; Singh, 2021).

Second, our theoretical analysis allows the rate of data cleaning to be slower than the rate of causal inference, so an analyst can use matrix completion (Candès and Recht, 2009; Candès and Tao, 2010; Keshavan et al., 2009; Hastie et al., 2015; Chatterjee, 2015) for automatic data cleaning of covariates. This key result extends the classic semiparametric framework, where the goal is to obtain $n^{-1/2}$ convergence for the causal parameter despite a possibly slow rate of convergence for a nonparametric regression. Our goal is to obtain $n^{-1/2}$ convergence for the causal parameter despite a possibly slow rate of convergence for high dimensional data cleaning. We build on both the classic semiparametric literature for low dimensional domains (Hasminskii and Ibragimov, 1979; Klaassen, 1987; Robinson, 1988; Bickel et al., 1993; Newey, 1994; Andrews, 1994; Robins et al., 1995; Robins and Rotnitzky, 1995; Ai and Chen, 2003; van der Laan and Rubin, 2006) and more recent developments for high dimensional domains (Zheng and van der Laan, 2011; Chernozhukov et al., 2016, 2018a,

2022a, 2021). En route, we extend semiparametric and nonparametric debiased machine learning theory to i.n.i.d. corrupted data, with new results on nominal and conservative variance estimation that are of independent interest. We also develop an original theory of implicit data cleaning. Altogether, our framework translates slow, on-average data cleaning guarantees into fast causal estimation and inference guarantees.

Third, our empirical results suggest that there exist scenarios in which the trade-off between privacy and precision can be overcome, and others in which it cannot. We replicate and extend Autor et al. (2013)'s seminal paper about the effect of import competition on US labor markets. To begin, we demonstrate the plausibility of our key assumption: Census data products contain many variables that are approximate repeated measurements. Next, we deliberately corrupt the data, injecting synthetic noise calibrated to the privacy level mandated for the 2020 US Census. We implement differential privacy and discretization in a way that belongs to our class of data corruptions, which can therefore be cleaned and adjusted for in the confidence interval. We find that the main results of Autor et al. (2013) can be recovered without losing statistical precision. In this representative setting for economic research, it appears to be possible to achieve both privacy at the individual level and precision at the population level.

The structure of this paper is as follows. Section 2 situates our contributions within the context of related work. Section 3 formalizes our class of data corruptions and our key assumption. Section 4 proposes our procedure and demonstrates its performance in simulations. Section 5 theoretically justifies our procedure, and verifies the key assumption for nonlinear factor models. Section 6 presents the semi-synthetic exercise and discusses limitations. Section 7 concludes. For readability, we reserve certain details for the appendix: additional examples in Appendix A, further empirical results in Appendix B, and nonlinear basis functions in Appendix C.

2 Related work

Semiparametric statistics. We use two insights from classic semiparametric theory. First, a causal parameter typically has regression and balancing weight representations, and both appear in the semiparametrically efficient asymptotic variance (Newey, 1994).

We directly build on this insight: an error-in-variable regression and an error-in-variable balancing weight appear in our data cleaning-adjusted confidence intervals. Both quantities also appear in our doubly robust estimating equation (Robins et al., 1995; Robins and Rotnitzky, 1995; van der Laan and Rubin, 2006; Chernozhukov et al., 2016, 2018a; Foster and Syrgkanis, 2019; Chernozhukov et al., 2022a). Second, sample splitting (Klaassen, 1987) eliminates the restrictive condition that function classes used in estimation must be simple (Zheng and van der Laan, 2011; Chernozhukov et al., 2016, 2018a, 2021). We combine these two classic ideas with implicit data cleaning, which appears to be a new idea.

Low rank models in econometrics. A vast literature studies the identification, estimation, and inference of latent factors (λ_i, μ_j) in models of the form

$$Z_{i,\cdot} = X_{i,\cdot} + H_{i,\cdot}, \quad X_{ij} = \lambda_i^T \mu_j \tag{1}$$

where $Z_{i,\cdot}$ is observed, the ambient dimension $dim(X_{i,\cdot})$ is high and growing, and the latent dimension $dim(\lambda_i)$ is low and fixed (Bai and Ng, 2002; Onatski, 2009; Bai and Ng, 2013; Bai and Wang, 2014). Rather than imposing a fixed linear factor model, we require that the approximate rank of $X_{i,\cdot}$ diverges much more slowly than $dim(X_{i,\cdot})$. The nonlinear factor model $X_{ij} = g(\lambda_i, \mu_j)$, where $dim(\lambda_j)$ may be diverging, is *sufficient* but *unnecessary* for our analysis. Like the factor model literature, we allow for weak correlation and heteroscedastity of measurement errors within units (Bai, 2003, 2009). We further allow for missingness and do not need to identify the latent factors for downstream causal inference.

Whereas we study treatment effects, policy effects, and elasticities in cross sectional data, a rich literature studies panel data settings where the analyst observes many units over time. One strand of this literature considers inference on the latent factors themselves in factor augmented regressions (Stock and Watson, 2002; Bai and Ng, 2006; Wang and Fan, 2017). Another strand considers inference for average treatment on the treated via synthetic control. In particular, works in this strand posit a low rank (Athey et al., 2021; Chernozhukov et al., 2018b; Bai and Ng, 2019; Xiong and Pelger, 2019; Fernández-Val et al., 2020; Agarwal et al., 2020b; Feng, 2020) or approximately low rank (Arkhangelsky et al., 2019; Agarwal et al., 2021) factor model for potential outcomes over units and times. Observed outcomes are interpreted as corrupted potential outcomes with idiosyncratic noise. By contrast, we study general nonseparable causal models, and we interpret observed

covariates as corrupted observations of true covariates that are approximately low rank.

The only previous work to consider both measurement error and missingness in cross sectional treatment effects appears to be Kallus et al. (2018). The authors consider average treatment effect and prove consistency, without inference, for a parametric linear model. In Kallus et al. (2018), the true covariates $X_{i,\cdot}$ are low dimensional and Gaussian; each corrupted covariate Z_{ij} is drawn independently from a known exponential family distribution; and each missing value is drawn i.i.d. Consistency requires $dim(X_{i,\cdot}) \ll n$ and correct specification of the measurement error distribution. By contrast, we consider a broad class of semiparametric and nonparametric causal parameters. Moreover, we prove Gaussian approximation and semiparametric efficiency with data cleaning-adjusted confidence intervals. We do not assume knowledge of the distributions of $X_{i,\cdot}$ and $Z_{i,\cdot}$; we allow $X_{i,\cdot}$ to be high dimensional; and we allow for dependent and i.n.i.d. data corruption.

Error-in-variable regression. We provide a framework to repurpose error-in-variable regression estimators for downstream causal inference. Error-in-variables regression has a vast literature spanning econometrics, statistics, and computer science studying the model

$$Y_i = \gamma_0(X_{i,\cdot}) + \varepsilon_i, \quad Z_{i,\cdot} = X_{i,\cdot} + H_{i,\cdot} \tag{2}$$

where $(X_{i,\cdot}, \varepsilon_i, H_{i,\cdot})$ are mutually independent and $(\varepsilon_i, H_{i,\cdot})$ are mean zero (Schennach and Hu, 2013). We consider a generalization of this setting with missingness, and we define our causal parameter as a scalar summary of γ_0.

Methods in econometrics typically assume auxiliary information: repeated measurements (Li and Vuong, 1998; Li, 2002; Delaigle et al., 2008), instrumental variables (Newey, 2001; Schennach, 2007; Hu and Schennach, 2008; Wang and Hsiao, 2011; Chen et al., 2011; Singh et al., 2019), and negative controls (Miao et al., 2018; Miao and Tchetgen, 2018; Deaner, 2018; Tchetgen Tchetgen et al., 2020; Singh, 2020). We do not require explicit auxiliary information, though there is a deep connection to the repeated measurement model, which we describe in Section 3.

An important class of methods in statistics extends Lasso and the Dantzig selector (Loh and Wainwright, 2012; Rosenbaum and Tsybakov, 2013; Datta and Zou, 2017). The model is high dimensional, linear, and sparse:

$$Y_i = X_{i,\cdot}\beta + \varepsilon_i, \quad Z_{i,\cdot} = (X_{i,\cdot} + H_{i,\cdot}) \odot \pi_{i,\cdot} \tag{3}$$

where $dim(X_{i,\cdot})$ may exceed n, β has $s \ll n$ nonzero coefficients, and $\pi_{i,\cdot} \in \{\texttt{NA}, 1\}^{dim(X_{i,\cdot})}$ encodes missingness. Analysis places three strong assumptions: exact sparsity of β, a restricted eigenvalue bounded away from zero, and knowledge of the covariance of measurement error $H_{i,\cdot}$. By contrast, we assume $X_{i,\cdot}$ are approximately low rank and $H_{i,\cdot}$ are subexponential; the analyst does not need to know the measurement error covariance.

We propose new variants of principal component regression (PCR) for the error-in-variable regression and balancing weight. Previous work studies PCR for error-in-variable regression only, considering models as in (3). Agarwal et al. (2021) perform data cleaning on the training set and test set together, while Agarwal et al. (2020a) perform data cleaning on the training set and test set separately. Explicit data cleaning on the test set induces correlation across observations, and therefore poses a challenge for downstream statistical inference. We use implicit data cleaning on the test set to preserve independence, and we prove fast rates of generalization. Unlike previous work, we simultaneously allow $X_{i,\cdot}$ to be approximately low rank, γ_0 to be nonlinear, and (π_{ij}, π_{ik}) to be dependent.

Missing values. The literature on missing covariates considers a rich variety of assumptions. We focus on the simple case that covariates are missing completely at random (MCAR). We do, however, allow for dependent missingness within a unit. For the richer setting where covariates are missing at random (MAR), popular methods involve multiple imputation (Rubin, 1976), generalized propensity scores (Rosenbaum and Rubin, 1984; D'Agostino Jr and Rubin, 2000), hot deck imputation (Abadie and Imbens, 2012), and doubly robust extensions thereof (Mayer et al., 2020).

3 Model overview

We define the key aspects of the model: the causal parameter, class of data corruptions, and key approximation assumption.

3.1 Causal parameter

For readability, we focus on one causal parameter in the main text: the average treatment effect with i.n.i.d. data, which we denote by

$$\theta_0 = \frac{1}{n}\sum_{i=1}^{n} \theta_i, \quad \theta_i = \mathbb{E}[Y_i^{(1)} - Y_i^{(0)}],$$

where $Y_i^{(d)}$ is the potential outcome for unit i under intervention $D = d$. θ_0 is a sample average because different units may be drawn from different distributions. In Appendix A, we consider a general class of semiparametric and nonparametric causal parameters including the local average treatment effect, the average elasticity, and heterogeneous treatment effects.

We denote the actual outcome by $Y_i \in \mathbb{R}$, the assigned treatment by $D_i \in \{0, 1\}$, and the vector of covariates that determine treatment selection by $X_{i,\cdot} \in \mathbb{R}^p$. In order to express θ_0 in terms of $(Y_i, D_i, X_{i,\cdot})$, we impose some additional structure on the problem. Generalizing a classic assumption in the literature on distribution shift, we assume that the conditional distributions $\mathbb{P}(Y_i|D_i, X_{i,\cdot})$ and $\mathbb{P}(D_i|X_{i,\cdot})$ are common across units; distribution shift is only in the marginal distributions of covariates $\mathbb{P}_i(X_{i,\cdot})$.

Imposing these conditions as well as selection on $X_{i,\cdot}$, we recover two classic formulations of the treatment effect. The outcome formulation is in terms of the outcome mechanism γ_0, also called the regression, which is common across units:

$$\theta_i = \mathbb{E}[\gamma_0(1, X_{i,\cdot}) - \gamma_0(0, X_{i,\cdot})], \quad \gamma_0(D_i, X_{i,\cdot}) = \mathbb{E}[Y_i|D_i, X_{i,\cdot}].$$

The treatment formulation is in terms of the treatment mechanism $\mathbb{E}[D_i|X_{i,\cdot}]$, which is also common across units, and which appears in the denominator of the balancing weight α_0:

$$\theta_i = \mathbb{E}[Y_i \cdot \alpha_0(D_i, X_{i,\cdot})], \quad \alpha_0(D_i, X_{i,\cdot}) = \frac{D_i}{\mathbb{E}[D_i|X_{i,\cdot}]} - \frac{1 - D_i}{1 - \mathbb{E}[D_i|X_{i,\cdot}]}.$$

Our estimation and analysis combine both classic formulations.

3.2 Data corruption

The crux of our problem is that, instead of observing $(Y, D_i, X_{i,\cdot})$, we observe $(Y_i, D_i, Z_{i,\cdot})$ where

$$Y_i = \gamma_0(D_i, X_{i,\cdot}) + \varepsilon_i, \quad Z_{i,\cdot} = (X_{i,\cdot} + H_{i,\cdot}) \odot \pi_{i,\cdot}. \tag{4}$$

Though the outcome Y_i is generated from treatment D_i and true covariates $X_{i,\cdot}$, we do not observe the true covariates; instead, we observe the corrupted covariates $Z_{i,\cdot}$, which are the true covariates $X_{i,\cdot}$ plus additive corruption $H_{i,\cdot}$, multiplied by a masking vector $\pi_{i,\cdot} \in \{\texttt{NA}, 1\}^p$. Our concise model (4) generalizes the models of previous work (1), (2), (3), and it encompasses all four types of corruption.[1] For example, to encode classical measurement error (Schennach, 2016), one could let $Z_{i,\cdot}$ equal $X_{i,\cdot}$ plus a vector of Gaussian noise. To encode missing values, let $Z_{i,\cdot} = X_{i,\cdot} \odot \pi_{i,\cdot}$.

Discretization is a process by which a continuous vector $X_{i,\cdot}$ is mapped to a discrete vector $Z_{i,\cdot}$, and our class encodes several variants. For example, the covariate of interest may be a vector of probabilities $X_{i,\cdot}$, yet we observe actual occurrences $Z_{i,\cdot} \sim \text{Bernoulli}(X_{i,\cdot})$. Another example is randomized rounding, where continuous values are randomly rounded to nearby integers, e.g. $Z_{i,\cdot} = sign(X_{i,\cdot})\text{Poisson}(|X_{i,\cdot}|)$. While our class does not include rounding in the familiar sense, it provides guidance on which types of rounding can be handled in downstream causal inference. As such, it suggests alternative types of discretization for wage data in the CPS which are more favorable for economic research.

Finally, to encode differential privacy, let $Z_{i,\cdot}$ equal $X_{i,\cdot}$ plus a vector of Laplacian noise. What is the connection between differential privacy and Laplacian noise? Differential privacy is a concept from computer science which means plausible deniability that any individual contributed their data to tabular summaries. The canonical mechanism that ensures differential privacy is to add Laplacian noise, calibrating the variance of the Laplacian to the variability of the true values and other properties of the tabular summary statistics (Dwork et al., 2006). In the context of the Census, we consider adding Laplacian noise to data on aggregate units, which we formalize in Section 6. Injecting synthetic noise in this way helps to prevent the kind of attack simulated on the 2010 Census.

Across examples, $H_{i,\cdot}$ is sub-exponential, i.e. its tails are no worse than the tails of an exponential distribution. So are compositions of various types of data corruption since the class of sub-exponential distributions is closed under addition. Therefore our class of data corruptions includes classical and modern issues *simultaneously*. In particular, it allows us to address the trade-off between privacy and precision in the context of measurement error—a major aspect of the problem that is often overlooked (Steed et al., 2022). In Appendix A,

[1] It may also be viewed as a concise nonlinear LISREL model (Jöreskog and Sörbom, 1996).

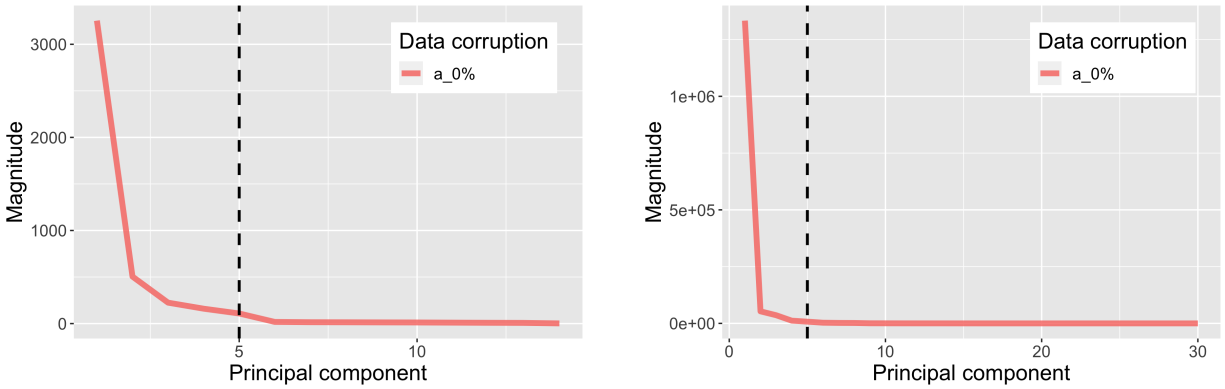
we extend the model to accommodate attrition and privacy of the outcome.

3.3 Key assumption: Approximate repeated measurements

Our key assumption is that the true covariates $X_{i,\cdot}$ are approximately low rank: the rank of the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is approximately $r \ll (n, p)$. Among the n units in the data set, there are approximately only r latent types of unit, i.e., each unit can be approximated as a linear combination of r latent types. Equivalently, among the p covariates in the data set, there are approximately only r latent types of covariate. To interpret this assumption, we propose the intuition of repeated measurements. In the classic repeated measurement model, we have access to two noisy measurements (Z_{i1}, Z_{i2}) of one signal; in our model, we have access to p noisy measurements $(Z_{i1}, ..., Z_{ip})$ that are approximately repeated measurements of only r signals, where both (r, p) grow with sample size n, yet $r \ll (n, p)$.

We place this assumption because it holds in Census data. Previewing our empirical application, consider the commuting zone (CZ) level data set of Autor et al. (2013). The mainland US consists of 722 CZs interpretable as local economies, each of which has a vector of covariates $X_{i,\cdot} \in \mathbb{R}^{13}$ used in the authors' main specification. The variables include percent employment in manufacturing, percent college educated, and percent employment among women—variables that are not precisely repeated measurements but approximately so. We compute the singular value decomposition of $\boldsymbol{X}$ then visualize its singular values, also called its principal components, in Figure 1a. We see that only about five principal components are significantly positive; $r = 5$. In Figure 1b, we confirm similar results with an augmented specification $X_{i,\cdot} \in \mathbb{R}^{30}$ using additional variables from Autor et al. (2013)'s appendix such as average disability, unemployment, and medical benefits. Again, these variables admit interpretation as approximate repeated measurements with $r = 5$. In Appendix B, we verify our key assumption on a broad variety of economic data sets.

Our key assumption is more than statistically convenient; it is causally meaningful. Consider the special case in which the true covariates are exactly low rank, i.e. $r = rank(\boldsymbol{X})$. The singular value decomposition is $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{U} \in \mathbb{R}^{n \times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$, and $\boldsymbol{V} \in \mathbb{R}^{p \times r}$. $\boldsymbol{V}$ consists of r vectors in $\mathbb{R}^p$, called the right singular vectors of $\boldsymbol{X}$, which are also the eigenvectors of the empirical covariance $n^{-1}\boldsymbol{X}^T\boldsymbol{X}$. The span of these vectors is an r



(a) Original specification

(b) Augmented specification

Figure 1: The key assumption holds in Census data

dimensional subspace of $\mathbb{R}^p$, i.e. a low dimensional subset of a high dimensional ambient space. In this scenario, we are assuming that treatment assignment is determined by the subspace. Equivalently, the treatment assignment for unit i depends on the *projection* of $X_{i,\cdot}$ onto this subspace. More generally, when covariates are approximately low rank, $\boldsymbol{X} = \boldsymbol{X}^{(\text{LR})} + \boldsymbol{E}^{(\text{LR})}$, where $\boldsymbol{X}^{(\text{LR})} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is a rank r approximation to $\boldsymbol{X}$ and $\boldsymbol{E}^{(\text{LR})}$ is the approximation residual. We can either assume (i) selection is determined by $\boldsymbol{X}^{(\text{LR})}$ only, i.e. the subspace spanned by $\boldsymbol{V}$; or (ii) selection is determined by both $\boldsymbol{X}^{(\text{LR})}$ and $\boldsymbol{E}^{(\text{LR})}$. To handle the latter case, which is the most general, we keep track of $\Delta_E = \|\boldsymbol{E}^{(\text{LR})}\|_{\max}$ in our theoretical analysis. In this sense, we provide analysis that is robust to violations of the exactly low rank assumption from both a statistical and causal perspective.

4 Data cleaning-adjusted confidence interval

We would like a procedure that estimates causal parameters as if data were uncorrupted, yet adjusts for data cleaning in the confidence interval. Moreover, we would like a procedure that does not require knowledge of the corruption covariance structure in advance, departing from previous work. If such a procedure were to exist, it would in some sense preempt the looming trade-off between privacy and precision.

4.1 Why is inference hard?

We illustrate key concepts with an average treatment effect simulation. By construction, the treatment effect is $\theta_0 = 2.2$. We consider a data generating process, detailed in Appendix I, which satisfies our key assumption: one sample involves a matrix of covariates $\boldsymbol{X} \in \mathbb{R}^{100\times100}$ with rank $r = 5$. See Appendix B for similar results using alternative dimensions of $\boldsymbol{X}$. To make the problem interesting, we allow for nonlinear outcome and treatment mechanisms. Figure 2 plots the principal components of true covariates $\boldsymbol{X}$ in red. As expected, five principal components are nonzero and the rest are zero since $rank(\boldsymbol{X}) = 5$.

As a first pass, we implement ordinary least squares (OLS) with robust standard errors in `Stata`: `regress Y D Z, vce(robust)`. Running OLS on clean data 1000 times, the point estimates $\hat{\theta}$ (Figure 3a) are centered around the true value of 2.2, and appear Gaussian. To evaluate the quality of OLS standard errors, we visualize how the studentized point estimates $(\hat{\theta} - \theta_0)/\hat{\sigma}$ compare the the standard normal density (Figure 3b). OLS works well in the absence of data corruption; there is nothing hidden in the data generating process for clean data.

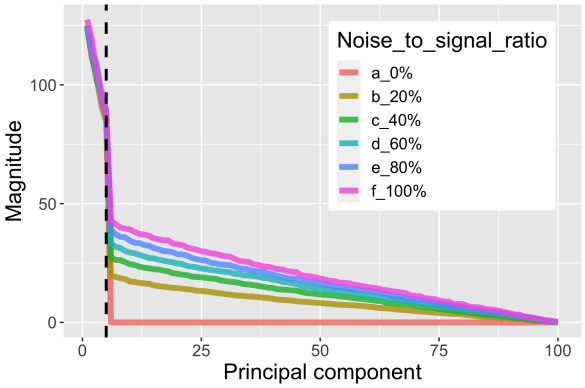
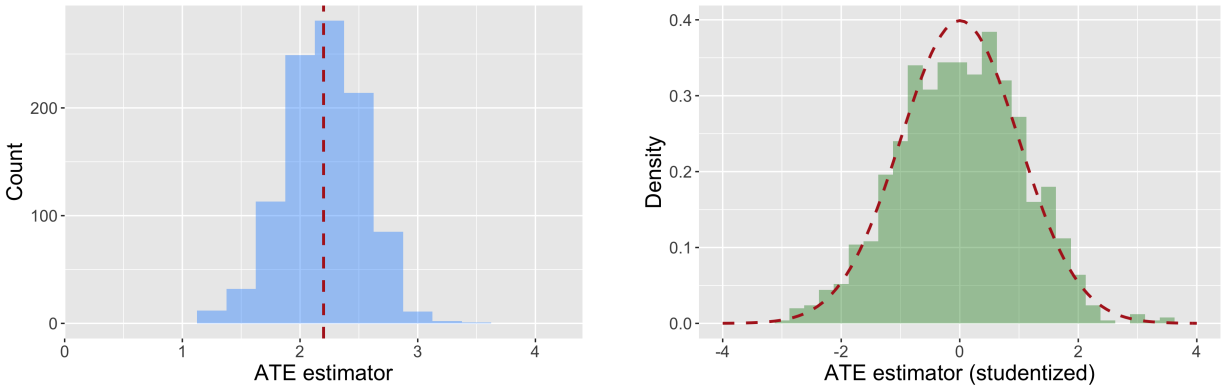


Figure 2: The key assumption holds in simulated data

We repeat this exercise introducing measurement error with variance that is 20% of the variance of the covariates. Inversion of the empirical covariance matrix $n^{-1}\boldsymbol{Z}^T\boldsymbol{Z}$ becomes numerically unstable, manifesting in point estimates that are erratic (Figure 4) and standard errors that are typically `NA`'s. OLS is not well-suited to the combination of high dimensional covariates, (approximate) low rank, and measurement error. We further investigate this phenomenon is Appendix B.

Data corruption can derail causal inference, which motivates filling the `NA`'s, reigning in the extremes, and otherwise de-noising the values in $\boldsymbol{Z}$ in hopes of recovering $\boldsymbol{X}$. These are precisely the goals of matrix completion applied to the matrix $\boldsymbol{Z}$. In this work, we



(a) Estimation (b) Inference

Figure 3: OLS succeeds on clean data

turn to the vast literature on matrix completion (Candès and Recht, 2009; Candès and Tao, 2010; Keshavan et al., 2009; Hastie et al., 2015; Chatterjee, 2015) for automated data cleaning. To select an appropriate matrix completion method, we return to Figure 2 to visualize the principal components of the corrupted covariates $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{H}$ for various noise-to-signal ratios (defined as the noise variance divided by the signal variance). Figure 2 shows that the initial five principal components remain virtually unchanged, while the lower principal components are amplified; signal remains spectrally concentrated while noise is spectrally diffuse. Therefore a natural way to clean the covariates would be to discard the lower principal components—in essence, to perform principal component analysis (PCA), also called hard singular value thresholding.[2]

Why is inference hard after data cleaning? Several challenges arise. First, manual and automated data cleaning may induce strong dependence among observations; which central limit theorem could we use to prove Gaussian approximation? Our answer is to introduce sample splitting, which is a classic idea (Klaassen, 1987), and implicit data cleaning, which is a new idea. Second,

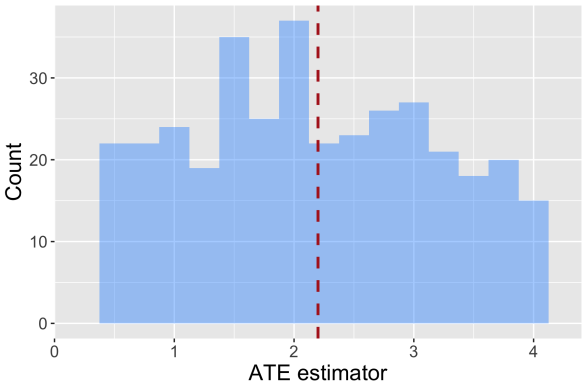


Figure 4: OLS fails on corrupted data

[2]Alternative choices include canonical correlation analysis and partial least squares, which clean $\boldsymbol{Z}$ using Y (Wold, 1982; Wold et al., 1984). We leave these directions to future research.

if we turn to automated data cleaning, the best rates of convergence to the true matrix $\boldsymbol{X}$ are slower than $n^{-1/2}$; how will we possibly obtain a standard error of order $\hat{\sigma}n^{-1/2}$? Our answer is to aim for double rate robustness by introducing a doubly robust estimating equation (Chernozhukov et al., 2018a; van der Laan and Rose, 2018; Rotnitzky et al., 2021). The third issue is a theoretical one to which we will return in Section 5: the best rates of matrix completion are not for recovering specific matrix entries but rather averages across matrix entries; how is this even related to downstream semiparametric inference? Our answer is to develop an algorithmic and analytic framework that forges the connection.

4.2 Overview of the procedure

Split the observations $(Y_i, D_i, Z_{i,\cdot})$ into equally sized TRAIN and TEST sets, each with $m = n/2$ observations. Our procedure consists of four steps, which we state at a high level before filling in the details.

1. Data cleaning: $\hat{\boldsymbol{X}}$ using TRAIN.
2. Error-in-variable regression: $\hat{\gamma}$ using TRAIN.
3. Error-in-variable balancing: $\hat{\alpha}$ using TRAIN.
4. Causal parameter: $\hat{\theta} \pm 1.96\hat{\sigma}n^{-1/2}$ using TEST.

We opt for simplicity at each step, essentially combining PCA and OLS (albeit in new ways). We view these high level steps a template for more complex procedures in future work.

Step 1: Data cleaning. The automated data cleaning procedure is extremely simple: fill in missing values as zeros, scale appropriately, then perform PCA. Importantly, we import the scalings from TRAIN to TEST.

For any mathematical operations to be well defined, the `NA`'s must be filled in somehow. To begin, we tally the likelihood of non-missingness for each covariate $j \in [p]$ in TRAIN:

$$\hat{\rho}_j = \max\left\{\frac{1}{m}\sum_{i\in\text{TRAIN}} \mathbb{1}(Z_{ij} \neq \texttt{NA}), \frac{1}{m}\right\}, \quad \hat{\boldsymbol{\rho}} = diag(\hat{\rho}_1, ..., \hat{\rho}_p) \in \mathbb{R}^{p\times p}.$$

Next, we fill in missing values with a FILL operator defined such that

$$\text{FILL}(Z_{ij}) = \begin{cases} \frac{Z_{ij}}{\hat{\rho}_j} & \text{if } Z_{ij} \neq \texttt{NA} \\ 0 & \text{if } Z_{ij} = \texttt{NA}. \end{cases}$$

The FILL operator may act on either $\boldsymbol{Z}^{\text{TRAIN}}$ or $\boldsymbol{Z}^{\text{TEST}}$, but it always uses the likelihoods $\hat{\boldsymbol{\rho}}$ calculated from $\boldsymbol{Z}^{\text{TRAIN}}$. After filling TRAIN, we project it onto its own principal subspace to calculate the cleaned training covariates $\hat{\boldsymbol{X}}$:

$$\text{FILL}(\boldsymbol{Z}^{\text{TRAIN}}) = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^T, \quad \hat{\boldsymbol{X}} = \hat{\boldsymbol{U}}_k\hat{\boldsymbol{\Sigma}}_k\hat{\boldsymbol{V}}_k^T.$$

Effectively, we are taking the SVD of $\text{FILL}(\boldsymbol{Z}^{\text{TRAIN}})$ and truncating it to include only the top k principal components, where k is a chosen by the analyst. Figure 2 suggests a choice of k, though other popular options include cross-validation and theoretical rules (Chatterjee, 2015; Gavish and Donoho, 2014). Below, we empirically verify that our results are robust to different choices of k. This implementation of PCA preserves the ambient dimension p.

Step 2: Error-in-variable regression. Our error-in-variables regression is also simple: after cleaning TRAIN, perform ordinary least squares (OLS) on TRAIN, then use this OLS coefficient on the filled TEST for prediction. We only fill, and do not clean, the test set.

Nonlinearity can be introduced into the regression to allow for treatment effect heterogeneity. See Appendix C for a characterization of what nonlinearity is allowed, and how nonlinearity manifests in the theoretical guarantees. For the main text, we focus on the interacted dictionary, which allows for heterogeneity and segments the OLS coefficient:

$$b(D_i, \hat{X}_{i,\cdot}) = (D_i\hat{X}_{i,\cdot}, (1-D_i)\hat{X}_{i,\cdot}), \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}^{\text{TREAT}} \\ \hat{\beta}^{\text{UNTREAT}} \end{bmatrix}.$$

Then the OLS coefficient is

$$\hat{\beta} = [\{b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})\}^T b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})]^\dagger [\{b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})\}^T Y^{\text{TRAIN}}],$$

where $\dagger$ means pseudoinverse. The subtlety is in how predictions are constructed from $\hat{\beta}$. Out of sample prediction does *not* involve cleaning the test set: for $i \in \text{TEST}$,

$$\hat{\gamma}(D_i, Z_{i,\cdot}) = b\{D_i, \text{FILL}(Z_{i,\cdot})\}\hat{\beta}.$$

Step 3: Error-in-variable balancing. Our error-in-variable balancing weight generalizes our error-in-variable regression. It avoids the estimation and inversion of propensity scores, which may be numerically unstable in high dimensions. Pleasingly, it achieves exact balance for any finite sample size, in a sense that we formalize below. Moreover, it adapts to the causal parameter of interest, as we explain in Appendix F.

The only difference from the error-in-variable regression is that we replace the sufficient statistic $[\{b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})\}^T Y^{\text{TRAIN}}] \in \mathbb{R}^{p'}$ with another sufficient statistic that we call the counterfactual moment $\hat{M} \in \mathbb{R}^{p'}$. The counterfactual moment resembles the expression $\theta_i = \mathbb{E}[\gamma_0(1, X_{i,\cdot}) - \gamma_0(0, X_{i,\cdot})]$, and it is the *only* aspect of the algorithm that changes for different causal parameters. Formally,

$$\hat{\eta} = [\{b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})\}^T b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})]^\dagger \hat{M}, \quad \hat{M} = [\{b(1, \hat{\boldsymbol{X}})\}^T - \{b(0, \hat{\boldsymbol{X}})\}^T] \, \mathbb{1}_m$$

where $\mathbb{1}_m \in \mathbb{R}^m$ is a vector of ones. As before we do *not* clean the test set: for $i \in \text{TEST}$,

$$\hat{\alpha}(D_i, Z_{i,\cdot}) = b\{D_i, \text{FILL}(Z_{i,\cdot})\}\hat{\eta}.$$

Step 4: Causal estimation and inference. The final step uses the error-in-variable regression $\hat{\gamma}$ and error-in-variable balancing weight $\hat{\alpha}$ learned from TRAIN, and evaluates them on TEST according to the doubly robust estimating equation: for $i \in \text{TEST}$,

$$\hat{\psi}_i = \hat{\gamma}(1, Z_{i,\cdot}) - \hat{\gamma}(1, Z_{i,\cdot}) + \hat{\alpha}(D_i, Z_{i,\cdot})\{Y_i - \hat{\gamma}(D_i, Z_{i,\cdot})\}.$$

This process generates a vector $\hat{\psi} \in \mathbb{R}^m$, with $\hat{\psi}_i$ corresponding to the empirical influence of observation $i \in \text{TEST}$. Reversing the roles of TRAIN and TEST, we generate another such vector. Slightly abusing notation, we concatenate the two to obtain a vector $\hat{\psi} \in \mathbb{R}^n$. Our estimator of the causal parameter $\hat{\theta}$, its variance $\hat{\sigma}^2$, and its data cleaning-adjusted confidence interval are

$$\hat{\theta} = \text{MEAN}(\hat{\psi}), \quad \hat{\sigma}^2 = \text{VAR}(\hat{\psi}), \quad \text{CI} = \hat{\theta} \pm 1.96\hat{\sigma}n^{-1/2}.$$

4.3 Properties of the procedure

Step 1: Data cleaning. We fill missing values by zeroes, and then scale appropriately, for two reasons: (i) the procedure is asymptotically unbiased, and (ii) the procedure avoids unnecessary correlations. We compare our proposal to another that fills missing values with means, which we call FILL-AS-MEANS.

Proposition 4.1 (Filling with zeros is unbiased and simple)**.** *For our proposal,*

$$\mathbb{E}[\text{FILL}(Z_{ij}^{\text{TEST}}) | X_{ij}^{\text{TEST}}, \text{TRAIN}] = X_{ij}^{\text{TEST}} \frac{\rho_j}{\hat{\rho}_j}.$$

The alternative procedure of filling missing values with means from TRAIN *gives*

$$\mathbb{E}[\text{FILL-AS-MEANS}(Z_{ij}^{\text{TEST}})|X_{ij}^{\text{TEST}}, \text{TRAIN}] = X_{ij}^{\text{TEST}}\rho_j + \bar{Z}_j^{\text{TRAIN}}(1-\rho_j).$$

FILL-AS-MEANS gives a convex combination of the signal we wish to recover and the noisy average from TRAIN. The noisy average introduces additional correlations that we avoid with our simpler approach.[3]

After filling missing values, we clean the data by PCA with hyperparameter k. Below, we verify the robustness of the procedure to different choices of k. In practice, we recommend that an analyst should examine a plot of principal components, e.g. Figure 1, and select k that is just after the "elbow" so that it is close to r.

Step 2: Error-in-variable regression. Rather than explicitly cleaning the test covariates, we implicitly clean them, to avoid correlations among predictions. The following result formalizes this implicit data cleaning and what it achieves.

Proposition 4.2 (Implicit data cleaning preserves independence)**.** *For* $i \in$ TEST

$$\hat{\gamma}(D_i, Z_{i,\cdot}) = b(D_i, Z_{i,\cdot})\tilde{\beta}, \quad \tilde{\beta} = \begin{bmatrix} \hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}^{\text{TREAT}} \\ \hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}^{\text{UNTREAT}} \end{bmatrix}$$

where we replace `NA` *with* 0 *in* $Z_{i,\cdot}$*. Therefore for* $(i,j) \in$ TEST*,*

$$\hat{\gamma}(D_i, Z_{i,\cdot}) \perp\!\!\!\perp \hat{\gamma}(D_j, Z_{j,\cdot})|\text{TRAIN}.$$

Remarkably, post-multiplying $b(D_i, Z_{i,\cdot})$ by $\tilde{\beta}$ handles the measurement error, missingness, discretization, and differential privacy of $Z_{i,\cdot}$ while also producing high quality predictions of Y_i. Moreover, since $\tilde{\beta}$ is learned exclusively from TRAIN, it is deterministic conditional on TRAIN, so predictions for observations $(i,j) \in$ TEST preserve their independence. This property will be essential for our inferential theory.

Our new variant of PCR has broader use outside of causal inference. For example, in online learning, a corrupted test observation $Z_{i,\cdot}$ does not need to be explicitly cleaned with respect to TEST or even TRAIN. Instead, it is sufficient to implicitly clean $Z_{i,\cdot}$ by post multiplying it with the coefficient $\tilde{\beta}$. For test observations, data cleaning and prediction can be combined into one step.

[3]Yet another procedure is called HOT-DECK imputation, which introduces correlations with a martingale structure (Abadie and Imbens, 2012). Our simple approach avoids these correlations as well.

Step 3: Error-in-variable balancing. The error-in-variable balancing weight shares the same implicit data cleaning and therefore preserves independence in the same way. In addition, it provides a strong guarantee of finite sample balance between the treated and untreated subpopulations.

Proposition 4.3 (The balancing weight exactly balances covariates)**.** *For any finite sample,*

$$\frac{1}{m}\sum_{i\in\text{TRAIN}}\hat{X}_{i,\cdot}=\frac{1}{m}\sum_{i\in\text{TRAIN}}D_i\hat{X}_{i,\cdot}\cdot\hat{\omega}_i^{\text{TRAIN}}=\frac{1}{m}\sum_{i\in\text{TRAIN}}(1-D_i)\hat{X}_{i,\cdot}\cdot\hat{\omega}_i^{\text{UNTREAT}}$$

where $(\hat{\omega}_i^{\text{TREAT}},\hat{\omega}_i^{\text{UNTREAT}})\in\mathbb{R}$ *are balancing weights computed from* $\hat{\eta}$*: for each* $i\in$ TRAIN*,*

$$\hat{\omega}_i^{\text{TREAT}}=\hat{X}_{i,\cdot}\hat{\eta}^{\text{TREAT}},\quad \hat{\omega}_i^{\text{UNTREAT}}=-\hat{X}_{i,\cdot}\hat{\eta}^{\text{UNTREAT}}.$$

Deterministically, the error-in-variable balancing weight exactly balances the full population, the treated subpopulation, and the untreated subpopulation with respect to their cleaned covariates. It is precisely the reweighting that would ensure comparability of treated and untreated groups in a stratified sampling design. We articulate a more general balancing property for generic causal parameters in Appendix F. We also clarify the sense in which the error-in-variable regression and balancing weight coincide on TRAIN but not TEST.

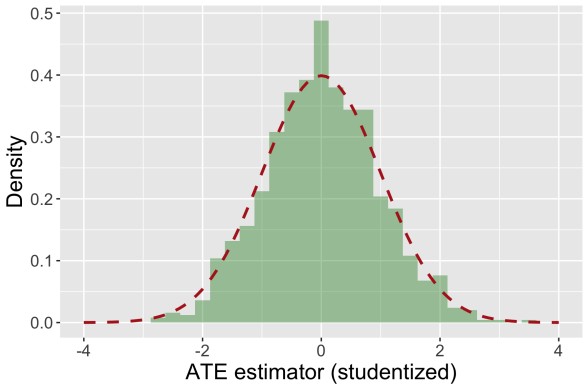
Step 4: Causal estimation and inference. In our estimation procedure, we deal with measurement error bias by cleaning the data. For some special cases, the measurement error bias actually has a closed form solution, which depends on the regression, propensity score, covariate density, and derivatives thereof (Battistin and Chesher, 2014). Our approach avoids estimation of the propensity scores, covariate density, and derivatives, which would be challenging in high dimensions. Instead, we simply combine PCA and OLS.

Multiple imputation (Rubin, 1976) is a popular procedure for handling missing values, and it has important similarities and differences compared to our approach. Both share the goal of estimating and quantifying uncertainty for a downstream scalar parameter. In multiple imputation, the analyst makes, say, two copies of the original data set, then imputes missing values (with some randomness so each imputation may be different). Estimates and standard errors from each copy are then averaged. In our procedure, we split the sample into two folds: TRAIN and TEST. We clean TRAIN and compute estimates and standard errors with TEST, then reverse the roles and take the average. We opt for sample splitting, rather than copying, and we additionally consider measurement error.

4.4 Adapting to the type and level of corruption

Next, we demonstrate that our four step procedure performs well in simulations with a broad variety of data corruptions. We run the same code in every setting; the procedure adapts to the *type* and *level* of data corruption, without prior knowledge of the corruption covariance structure.

To begin, we consider measurement error $Z_{i,\cdot} = X_{i,\cdot} + H_{i,\cdot}$, where $H_{i,\cdot}$ is Gaussian noise, in the average treatment effect simulation described above. Recall that $\theta_0 = 2.2$, $\boldsymbol{X} \in \mathbb{R}^{100\times 100}$, and $r = 5$. We implement our procedure on corrupted data 1000 times, collecting 1000 point estimates $\hat{\theta}$ and 1000 standard errors $\hat{\sigma}$. For a 20% noise-to-signal ratio, we visualize the studentized point estimates $(\hat{\theta} - \theta_0)/\hat{\sigma}$ in Figure 5a. Qualitatively, the histogram closely resembles the standard normal density.



(a) Inference

Noise	PCA	ATE	SE	CI.80	CI.95
20%	k = 5	2.22	0.35	0.81	0.96
60%	k = 5	2.23	0.37	0.81	0.96
100%	k = 5	2.28	0.39	0.82	0.95

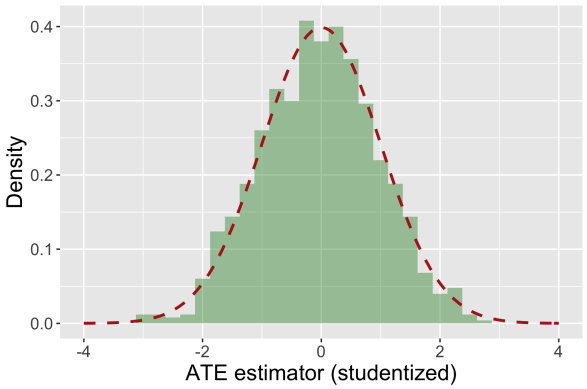
Noise	PCA	ATE	SE	CI.80	CI.95
20%	k = 5	2.22	0.35	0.81	0.96
20%	k = 7	2.21	0.36	0.84	0.96
20%	k = 10	2.22	0.39	0.83	0.97

(b) Coverage

Figure 5: Our approach adapts to measurement error

We quantify performance in coverage tables. In Table 5b, different rows correspond to different noise-to-signal ratios. Initially, we consider the oracle tuning of the PCA hyperparameter $k = r$. For each noise-to-signal ratio, we record the average point estimates, which are close to $\theta_0 = 2.2$. Next, we record the average standard errors, which adaptively increase in length to higher noise levels. Impressively, a 100% noise-to-signal ratio setting corresponds to a confidence interval that is only about 10% longer. These confidence intervals are the correct length, since about 950 of them include the true value $\theta_0 = 2.2$.

Table 5b revisits the issue of tuning the hyperparameter k. This time, we fix the noise-to-signal ratio to 20%. Different rows correspond to different tunings: $k = r$, $k = r + 2$,



(a) Inference

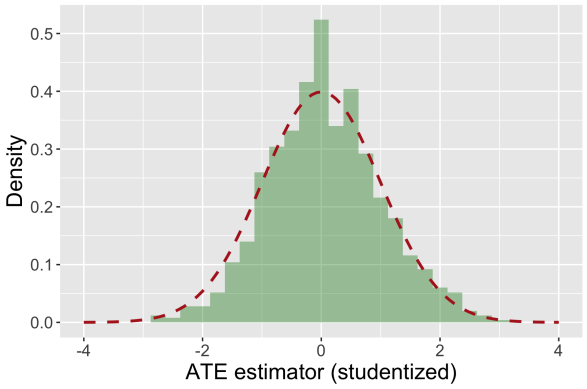
Missingness	PCA	ATE	SE	CI.80	CI.95
10%	k = 5	2.20	0.35	0.81	0.96
30%	k = 5	2.24	0.37	0.81	0.94
50%	k = 5	2.35	0.41	0.79	0.94

Missingness	PCA	ATE	SE	CI.80	CI.95
10%	k = 5	2.20	0.36	0.81	0.96
10%	k = 7	2.20	0.39	0.79	0.95
10%	k = 10	2.19	0.45	0.79	0.95

(b) Coverage

Figure 6: Our approach adapts to missing values

and $k = r + 5$. Point estimates remain close to the true value $\theta_0 = 2.2$. The standard errors adaptively increase in length when k deviates from r, though the length only increases about 10%. The confidence intervals are again the correct length, attaining nominal coverage.



(a) Inference

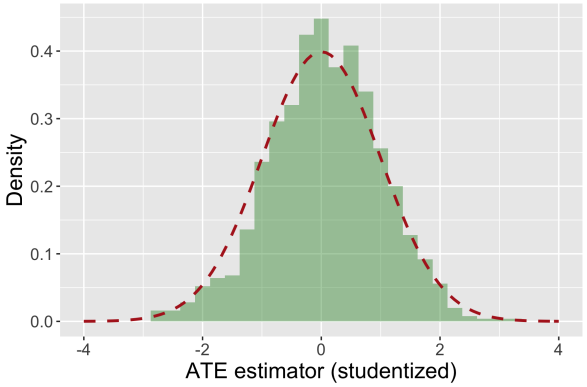
noise	ATE	SE	PCA	CI.80	CI.95
33	2.23	0.36	k = 5	0.81	0.96
33	2.23	0.37	k = 7	0.80	0.95
33	2.23	0.41	k = 10	0.81	0.95

(b) Coverage

Figure 7: Our approach adapts to discretization

We repeat this exercise with other types of data corruption: missing values (Figure 6), discretization (Figure 7), and differential privacy (Figure 8). For missing values $Z_{i,\cdot} = X_{i,\cdot} \odot \pi_{i,\cdot}$, we consider non-response of 10%, 30%, and 50% of *all covariate entries*. Key variables such as income in Census Bureau surveys are missing 40% of the time. Fortunately, our procedure performs well even with this high level of missingness. For discretization, we consider randomized rounding $Z_{i,\cdot} = sign(X_{i,\cdot})\text{Poisson}(|X_{i,\cdot}|)$, which corresponds to a 33% noise-to-signal ratio. Finally, for differential privacy $Z_{i,\cdot} = X_{i,\cdot} + H_{i,\cdot}$, where $H_{i,\cdot}$ is

Laplacian noise, we obtain results that are nearly identical to measurement error. Across settings, our results are robust to hyperparameter tuning.



(a) Inference

Privacy	PCA	ATE	SE	CI.80	CI.95
20%	k = 5	2.19	0.35	0.84	0.97
60%	k = 5	2.23	0.37	0.81	0.96
100%	k = 5	2.29	0.39	0.81	0.95

Privacy	PCA	ATE	SE	CI.80	CI.95
20%	k = 5	2.19	0.35	0.84	0.97
20%	k = 7	2.20	0.36	0.84	0.97
20%	k = 10	2.19	0.39	0.86	0.97

(b) Coverage

Figure 8: Our approach adapts to differential privacy

5 Finite sample analysis

In the previous section, we articulate three reasons why inference after data cleaning is hard. First, data cleaning may induce a great deal of dependence. We introduce implicit data cleaning as an algorithmic solution, yet we still need to provide a theory of implicit data cleaning: why is it okay to not clean the test covariates? Second, the best rates of data cleaning are slower than $n^{-1/2}$. We incorporate the doubly robust estimating equation in the hope of achieving double rate robustness, yet we still need to prove that it works: how is causal inference still possible with standard errors of order $\hat{\sigma}n^{-1/2}$? Third, data cleaning recovers averages across matrix entries. How can we translate guarantees about recovering averages into guarantees about the coverage of data cleaning-adjusted confidence intervals? In this section, we answer these three theoretical questions with finite sample analysis.

We prove four theorems, each corresponding to a step in the procedure.

1. Data cleaning: $\hat{\boldsymbol{X}}$ converges to $\boldsymbol{X}^{\text{TRAIN}}$.

2. Error-in-variable regression: $\hat{\gamma}$ converges to γ_0.

3. Error-in-variable balancing weight: $\hat{\alpha}$ converges to α_0.

4. Causal parameter: $\mathbb{P}\{\theta_0 \in (\hat{\theta} \pm 1.96\hat{\sigma}n^{-1/2})\}$ converges to 0.95.

We have already verified that our key assumption, that covariates are approximately low rank, is reasonable in practice for US Census data. In a corollary, we verify that it is reasonable in theory: it holds for a broad class of linear and nonlinear factor models.

5.1 Step 1: Data cleaning

For the data cleaning guarantee, we place four assumptions on the corrupted data. To lighten notation, we suppress indexing by TRAIN.

Assumption 5.1 (Bounded signal)**.** *There exists an absolute constant $\bar{A} < \infty$ such that for all $i \in [m], j \in [p]$, $|X_{ij}| \leq \bar{A}$.*

Assumption 5.1 is a standard boundedness assumption imposed on true values in the matrix completion literature.

Assumption 5.2 (Measurement error)**.** *Each row of measurement error $H_{i,\cdot}$ is mean zero and subexponential, i.e. $\mathbb{E}[H_{i,\cdot}|X_{i,\cdot}] = 0$ and there exists $a \geq 1$ and $K_a < \infty$ such that $\|H_{i,\cdot}|X_{i,\cdot}\|_{\psi_a} \leq K_a$. Hence there exists $\kappa^2 > 0$ such that $\|\mathbb{E}[H_{i,\cdot}^T H_{i,\cdot}|X_{i,\cdot}]\|_{op} \leq \kappa^2$. We assume measurement error is independent across rows.*

Measurement error is independent across rows, but it may be *dependent* within a given row. If in a given row of $H_{i,\cdot} \in \mathbb{R}^p$ each coordinate is independent, then K_a and κ^2 are constants (i.e. they do not scale with p) (Vershynin, 2018, Lemma 3.4.2). More generally, (K_a, κ) quantify the level of dependence among the entries of $H_{i,\cdot}$ within a row. Our model allows for a great deal of heteroscedasticity. In particular, the results to follow are conditional on $\boldsymbol{X}$ so, for example, the shape of the distribution of H_{ij} may depend on X_{ij} as long as it is mean zero and has tails no wider than those of an exponential distribution. As such, it encompasses discretization and differential privacy.

Assumption 5.3 (Missing values)**.** *In each row of missingness $\pi_{i,\cdot}$, π_{ij} is 1 with probability ρ_j and* `NA` *otherwise. Identifying* `NA` *with* 0*, we assume there exists $\bar{K} < \infty$ such that $\|\pi_{i,\cdot} - (\rho_1, ..., \rho_p)|X_{i,\cdot}\|_{\psi_2} \leq \bar{K}$. We assume missingness is independent across rows.*

Our missingness model generalizes the standard missingness model in the PCR error-in-variables literature in two ways: (i) the missingness of one variable may depend on the missingness of another, and (ii) different variables may be missing with different probabilities. These differences are easily visualized by the contrast

$$\mathbb{E}[\pi_{i,\cdot}^T \pi_{i,\cdot}] = \begin{bmatrix} \rho_1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & \rho_2 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & \rho_p \end{bmatrix} \quad \text{versus} \quad \mathbb{E}[\pi_{i,\cdot}^T \pi_{i,\cdot}] = \begin{bmatrix} \rho & \rho^2 & \cdots & \rho^2 \\ \rho^2 & \rho & \cdots & \rho^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho^2 & \rho^2 & \cdots & \rho \end{bmatrix}$$

where $\rho_{jj'} = \mathbb{E}[\pi_{ij}\pi_{ij'}]$ does not necessarily equal $\rho_j \rho_{j'} = \mathbb{E}[\pi_{ij}]\mathbb{E}[\pi_{ij'}]$. These additional degrees of flexibility are crucial for Census data, where non-responses for different variables are often correlated and where non-response rates of different variables can be vastly different. As with measurement error, missingness is independent across rows, but it may be *dependent* within a given row. If in a given row of $\pi_{i,\cdot} \in \mathbb{R}^p$ each coordinate is independent, then $\rho_{jj'} = \rho_j \rho_{j'}$ and $\bar{K}$ is constant. More generally, $\bar{K}$ quantifies the level of dependence among the entries of $\pi_{i,\cdot}$ within a row. Our model allows for different probabilities of missingness for different variables in a way that can depend on the signal. The results to follow are conditional on $\boldsymbol{X}$ so, for example, the probability ρ_j may depend on $X_{\cdot,j}$. For our theoretical results, we define the additional notation

$$\rho_{\min} \coloneqq \min_{j \in [p]} \rho_j, \quad \boldsymbol{\rho} = diag(\rho_1, ..., \rho_p) \in \mathbb{R}^{p \times p}.$$

Assumption 5.4 (Concentrated signal)**.** *Consider the approximation $\boldsymbol{X}^{(\text{LR})}$ to $\boldsymbol{X}$, with singular values $s_1, ..., s_r$. Assume that $s_1, ..., s_r \geq C\sqrt{\frac{mp}{r}}$, where C is an absolute constant.*

Assumption 5.4 is analogous to incoherence-style conditions in econometrics (Bai and Ng, 2019; Agarwal et al., 2021, 2020a,b) and the notion of pervasiveness in matrix completion (Fan et al., 2018). Similar to a strong factor assumption (Onatski, 2012; Anatolyev and Mikusheva, 2021), it ensures that the explanatory power of $\boldsymbol{X}^{(\text{LR})}$ dominates the explanatory power of various error terms. Specifically, it ensures that signal is spectrally concentrated. A natural setting in which Assumption 5.4 holds is if $X_{ij}^{(\text{LR})} = \Theta(1)$ and $s_1, ...s_r = \Theta(\tau)$, i.e., they are well-balanced. Then, for absolute constants $C, C', C'' > 0$,

$$C \cdot r \cdot \tau^2 = \sum_k s_k^2 = \|\boldsymbol{X}^{(\text{LR})}\|_{Fr}^2 = C' \cdot mp$$

which implies $\tau = C''\sqrt{\frac{mp}{r}}$. We leave to future work the extension of our results to settings with different spectral assumptions on $\boldsymbol{X}^{(\text{LR})}$.

Remark 5.1 (Approximately low rank signal)**.** *We parametrize our rates by the quality of low rank approximation.*

Without loss of generality, $\boldsymbol{X} = \boldsymbol{X}^{(\text{LR})} + \boldsymbol{E}^{(\text{LR})}$, where $\boldsymbol{X}^{(\text{LR})}$ is a low rank approximation to $\boldsymbol{X}$ and $\boldsymbol{E}^{(\text{LR})}$ is the approximation residual. The two key quantities are $r = rank\{\boldsymbol{X}^{(\text{LR})}\}$ and $\Delta_E = \|\boldsymbol{E}^{(\text{LR})}\|_{\max}$. It is *with* loss of generality that r and Δ_E are simultaneously well behaved. Intuitively, as r decreases Δ_E increases (and vice-versa). Indeed, if $\boldsymbol{X}^{(\text{LR})} = \boldsymbol{X}$ then trivially $r \leq (m, p)$ and $\Delta_E = 0$; conversely, if $\boldsymbol{X}^{(\text{LR})} = 0$, then $r = 0$ but $\Delta_E = \bar{A}$. As we show in the corollary, under a nonlinear factor model, both r and Δ_E are well behaved: $r \ll (m, p)$ and $\Delta_E \to 0$. Until that corollary, our rates are parametrized by (r, Δ_E).

Theorem 5.1 (Finite sample data cleaning rate)**.** *Suppose Assumptions 5.1, 5.2, 5.3, and 5.4 hold. Furthermore, suppose that $k = r$ and $\rho_{\min} > \frac{23\log(mp)}{m}$. Then for an absolute constant $C > 0$,*

$$\frac{1}{m}\mathbb{E}\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{2,\infty}^2 \leq C_1 \cdot \frac{r\ln^5(mp)}{\rho_{\min}^4}\left(\frac{1}{m} + \frac{1}{p} + \Delta_E^2\right)$$

where $C_1 = C \cdot \bar{A}^4(K_a + \bar{K})^2(\kappa + K_a + \bar{K})^2$.

The norm of convergence is the so-called $(2, \infty)$ norm:

$$\frac{1}{m}\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{2,\infty}^2 = \max_{j\in[p]}\frac{1}{m}\|\hat{X}_{i,\cdot} - X_{\cdot,j}\|_2^2 = \max_{j\in[p]}\frac{1}{m}\sum_{i=1}^{m}(\hat{X}_{ij} - X_{ij})^2$$

i.e. a maximum across columns and an average across rows. For any given variable $j \in [p]$, Theorem 5.1 guarantees that data cleaning performs well on average across observations $i \in [m]$. Our rate requires both m and p to increase. Rather than a curse, there is *blessing* of dimensionality: more repeated measurements improve the quality of data cleaning. For the bound to be meaningful, (r, Δ_E) must be simultaneously well behaved, which is our key assumption. Recall that $(K_a, \kappa, \bar{K})$ quantify the level of corruption dependence within a row. As long as the dependence is weak, e.g. $(K_a, \kappa, \bar{K})$ scale as some power of $\ln(mp)$, this dependence in negligible. Our downstream results for the error-in-variable regression and balancing weight are predicated on this data cleaning guarantee. As such, the geometry used in data cleaning is the crux of the entire framework: signal is spectrally concentrated, while noise is spectrally diffuse, so we can concentrate out the noise.

5.2 Step 2: Error-in-variable regression

We place three additional assumptions for the error-in-variable regression guarantee.

Assumption 5.5 (Response noise)**.** *The response noise satisfies* $\mathbb{E}[\varepsilon_i] = 0$ *and* $\mathbb{V}[\varepsilon_i] \leq \bar{\sigma}^2$*. It is independent across rows and independent of* $Z_{i,\cdot}$*.*

This weak condition permits measurement error and differential privacy of the outcome Y_i. See Appendix A for outcome attrition. Next we assume TRAIN and TEST each contains a sufficient variety of observations. For a matrix $\boldsymbol{M} \in \mathbb{R}^{m \times p}$, we define its row space as $\text{ROW}(\boldsymbol{M}) = span\{M_{i,\cdot}\}$.

Assumption 5.6 (Row space inclusion)**.** *Assume* $\text{ROW}[b\{\boldsymbol{X}^{(\text{LR}),\text{TRAIN}}\}] = \text{ROW}[b\{\boldsymbol{X}^{(\text{LR}),\text{TEST}}\}]$.

This property permits $\boldsymbol{X}^{(\text{LR}),\text{TRAIN}} \neq \boldsymbol{X}^{(\text{LR}),\text{TEST}}$, and also permits the two matrices to have different SVDs. In Appendix E, we verify that Assumption 5.6 holds with high probability under weak auxiliary conditions. Finally, we place a weak technical condition.

Assumption 5.7 (Well conditioned estimators)**.** *Let* $\hat{s}'_{k'}$ *be the smallest non-zero singular value of* $b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})$*. Assume that* $\hat{s}'_{k'} \gtrsim \frac{\bar{\varepsilon}}{\text{polynomial}(m,p)}$ *where* $\mathbb{E}[\varepsilon_i^8] \leq \bar{\varepsilon}^8$*.*

For $(\hat{\beta}, \hat{\eta})$ to be well conditioned, the singular value $\hat{s}'_{k'}$ should not be too small. In particular, it must be bounded below by an arbitrary negative power of m and p. We view the Assumption 5.7 as a diagnostic tool for empirical practice: choose the PCA hyperparameter k such that the k'-th singular value $\hat{s}'_{k'}$ is not too close to zero.

Before stating the result, we introduce a theoretical device β^* as the coefficient of the best low rank nonlinear approximation to γ_0. In particular, we approximate

$$\gamma_0(D_i, X_{i,\cdot}) = b(D_i, X_{i,\cdot}^{(\text{LR})})\beta^* + \phi_i^{(\text{LR})}$$

where $\phi_i^{(\text{LR})}$ is the approximation error. It turns out to be convenient to keep track of this approximation error by defining $\phi_i := \gamma_0(D_i, X_{i,\cdot}) - b(D_i, X_{i,\cdot})\beta^*$. There will be a trade-off: a richer dictionary b leads to a smaller approximation error in terms of $\|\phi\|_2^2$, but more compounding of measurement error and missingness. The following result helps to characterize how the compounded data corruption magnifies $(\rho_{\min}^{-1}, r, \Delta_E)$ but nothing else.

Remark 5.2 (General dictionaries)**.** *In the main text, we state results that hold for a broad class of dictionaries, with the dictionary specific constant C_b' and the concise notation $(\rho_{\min}', r', \Delta_E')$ in Theorems 5.2 and 5.3. In Appendix C we prove that*

$$C_b' \leq 2^{d_{\max}} \cdot \bar{A}_{\max}^{2d_{\max}} \cdot \|\hat{\boldsymbol{X}}\|_{\max}^{2d_{\max}} \cdot, \quad \frac{1}{\rho_{\min}'} \leq \frac{d_{\max}\bar{A}^{d_{\max}}}{\rho_{\min}}, \quad r' \leq r^{d_{\max}}, \quad \Delta_E' \leq C\bar{A}^{d_{\max}} \cdot d_{\max}\Delta_E$$

where $d_{\max}$ is the degree of the polynomial dictionary. We articulate restrictions on the class of dictionaries in Appendix C. For the interacted dictionary, $d_{\max} = 2$.

Remark 5.3 (Bound on $\|\hat{\boldsymbol{X}}\|_{\max}$)**.** *Under further incoherence style assumptions, we bound $\|\hat{\boldsymbol{X}}\|_{\max} \leq C\sqrt{r}$ in Appendix D. Alternatively, one can bound*

$$\|\hat{\boldsymbol{X}}\|_{\max} \leq \|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{\max} + \|\boldsymbol{X}\|_{\max} \leq \|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_{2,\infty}^2 + \bar{A}$$

then appeal to Theorem 5.1 with high probability. Doing so does not affect the powers of (m, p) in the main results, but does increase the complexity of the pre-factors.

Theorem 5.2 (Finite sample error-in-variable regression rate)**.** *Suppose the conditions of Theorem 5.1 hold, as well as Assumptions 5.5, 5.6, and 5.7.*

$$\textit{If} \quad \rho_{\min}' \gg \tilde{C}\sqrt{r'}\ln^{\frac{3}{2}}(mp)\left\{\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{m}} \vee \Delta_E\right\}, \quad \tilde{C} := C\bar{A}\Big(\kappa + \bar{K} + K_a\Big) \quad \textit{then} \quad (5)$$

$$\begin{aligned}\mathcal{R}(\hat{\gamma}) \leq C_b' C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{(r')^3 \ln^8(mp)}{(\rho_{\min}')^6}\|\beta^*\|_1^2 \left(\frac{1}{m} + \frac{p}{m^2} + \frac{1}{p} + \left(1 + \frac{p}{m}\right)(\Delta_E')^2 + p(\Delta_E')^4\right) \\ + C_2 \cdot \frac{(r')^2 \ln^3(mp)}{(\rho_{\min}')^2}\Delta_\phi\left(1 + (\Delta_E')^2\right)\end{aligned}$$

where $\Delta_\phi = \frac{1}{m}\|\phi^{\text{TRAIN}}\|_2^2 \vee \frac{1}{m}\|\phi^{\text{TEST}}\|_2^2$ and

$$C_1 = C\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2, \quad C_2 = C \cdot \bar{A}^4(\kappa + \bar{K} + K_a)^2.$$

Corollary 5.1 (Simplified regression rate)**.** *Suppose the conditions of Theorem 5.2 hold. Further suppose γ_0 is exactly linear in signal, which is exactly low rank. Then*

$$\mathcal{R}(\hat{\gamma}) \leq C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(mp)}{\rho_{\min}^6}\|\beta^*\|_1^2 \left(\frac{1}{m} + \frac{p}{m^2} + \frac{1}{p}\right).$$

The norm of convergence is a generalized mean square error

$$\mathcal{R}(\hat{\gamma}) = \mathbb{E}\left[\frac{1}{m}\sum_{i \in \text{TEST}}\{\hat{\gamma}(D_i, Z_{i,\cdot}) - \gamma_0(D_i, X_{i,\cdot})\}^2\right],$$

which is a relaxation of mean square error, where the expectation is over randomness in TRAIN and TEST. Two aspects of our problem necessitate this norm: (i) given the $(2, \infty)$ data cleaning guarantee in Theorem 5.1, this is the best we can do; and (ii) for i.n.i.d. data, a population risk is otherwise not well defined.[4] Since the estimator $\hat{\gamma}$ does not involve cleaning TEST, this result is our desired theory of implicit data cleaning. The bound requires both m and p to increase, $p \ll m^2$, and $\rho_{\min} \gg p^{-1/2} \vee m^{-1/2} \vee \Delta_E$. For the bound to be meaningful, (r, Δ_E) must be simultaneously well behaved and the corruption dependence must be weak. Finally, the bound includes the nonlinear approximation error Δ_ϕ and the size of the theoretical device $\|\beta^*\|_1$. $\|\beta^*\|_1$ is well behaved if β^* is approximately sparse. In summary, we keep track of the low rank approximation error Δ_E and the nonlinear sparse approximation error Δ_ϕ. To deal with Δ_E, we demonstrate that nonlinear factor models admit low rank approximation below. Due to our doubly robust approach, inference for the causal parameter θ_0 is robust to non-vanishing Δ_ϕ—a discussion we revisit later.

We make several innovations relative to previous work in the PCR error-in-variables literature. First, we propose an error-in-variable regression estimator that does not clean the test covariates, so we must develop a new theory of implicit data cleaning. Second, we define and analyze a new norm of convergence which we will subsequently use in causal inference. See Appendix E for a comparison of our norm with the norms in previous work. Third, we allow for dependence of missingness across variables and for different probabilities of missingness across variables. This flexibility is realistic for Census data. Fourth, we consider a nonlinear regression function γ_0 that is approximated by a nonlinear dictionary of basis functions b. The dictionary of basis functions is important for causal inference, since it allows for treatment effect heterogeneity, and it requires a novel characterization of which nonlinearities do not compound data corruption too much.

5.3 Step 3: Error-in-variable balancing

We place one final assumptions for the error-in-variable balancing weight.

Assumption 5.8 (Row space inclusion)**.** *Assume* $\hat{M} \in \text{ROW}\{b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})\}$.

Whereas Assumption 5.6 is about the low rank approximation of the signal across TRAIN

[4]Interestingly, even with i.i.d. data, (i) necessitates this norm.

and TEST, Assumption 5.8 is about the counterfactual moment in relation to TRAIN after cleaning. With $\hat{M} = [\{b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})\}^T Y^{\text{TRAIN}}]$, which reverts to error-in-variable regression, Assumption 5.8 immediately holds. In other cases, it limits the counterfactual queries that an analyst may ask. Because it is about empirical quantities, we view it as a diagnostic tool that an analyst should use to determine whether the counterfactual can be extrapolated.

As before, we introduce a theoretical device η^* as the coefficient of the best low rank nonlinear approximation to α_0. In particular, we approximate

$$\alpha_0(D_i, Z_{i,\cdot}) = b(D_i, X_{i,\cdot}^{(\text{LR})})\eta^* + \zeta_i^{(\text{LR})}$$

where $\zeta_i^{(\text{LR})}$ is the approximation error.[5] As before, we study this approximation error by defining $\zeta_i := \alpha_0(D_i, Z_{i,\cdot}) - b(D_i, X_{i,\cdot})\eta^*$. There will be a trade-off: a richer dictionary b leads to a smaller approximation error in terms of $\|\zeta\|_2^2$, but amplification of $(\rho_{\min}^{-1}, r, \Delta_E)$.

Remark 5.4 (General causal parameters)**.** *In the main text, we state results that hold for a broad class of causal parameters, with parameter specific constants (C'_m, C''_m) in Theorem 5.3. In Appendix F, we characterize (C'_m, C''_m) for several examples. For ATE with the interacted dictionary, $C'_m = 1$ and $C''_m = \bar{A}$.*

Theorem 5.3 (Finite sample error-in-variable balancing weight rate)**.** *Suppose the conditions of Theorem 5.1 hold, as well as Assumptions 5.6, 5.7, and 5.8. If* (5) *holds and $\|\alpha_0\|_\infty \leq \bar{\alpha}$,*

$$\mathcal{R}(\hat{\alpha}) \leq C_3 \cdot \frac{(r')^5 \ln^{13}(mp)}{(\rho'_{\min})^{10}} \|\eta^*\|_1^2 \cdot \left\{ \frac{1}{m} + \frac{1}{p} + \frac{p}{m^2} + \frac{m}{p^2} + \left(1 + \frac{p}{m} + \frac{m}{p}\right)(\Delta'_E)^2 + (m+p)(\Delta'_E)^4 + mp(\Delta'_E)^6 \right\} + 2\Delta_\zeta$$

where $\Delta_\zeta = \frac{1}{m}\|\zeta^{\text{TRAIN}}\|_2^2 \vee \frac{1}{m}\|\zeta^{\text{TEST}}\|_2^2$ and

$$C_3 = C\bar{A}^{14}(C'_b + \sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^6.$$

Corollary 5.2 (Simplified balancing weight rate)**.** *Suppose the conditions of Theorem 5.3 hold. Further suppose α_0 is exactly linear in signal, which is exactly low rank. Then*

$$\mathcal{R}(\hat{\alpha}) \leq C_3 \cdot \frac{r^5 \ln^{13}(mp)}{\rho_{\min}^{10}} \|\eta^*\|_1^2 \cdot \left\{ \frac{1}{m} + \frac{1}{p} + \frac{p}{m^2} + \frac{m}{p^2} \right\}$$

replacing C'_b with 1 in the definition of C_3.

[5] A further assumption that the treatment mechanism only depends on signal, i.e. $\mathbb{E}[D_i|X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}] = \mathbb{E}[D_i|X_{i,\cdot}]$, implies $\alpha_0(D_i, Z_{i,\cdot}) = \alpha_0(D_i, X_{i,\cdot}) = b(D_i, X_{i,\cdot}^{(\text{LR})})\eta^* + \zeta_i^{(\text{LR})}$.

The norm of convergence is a generalized mean square error

$$\mathcal{R}(\hat{\alpha}) = \mathbb{E}\left[\frac{1}{m}\sum_{i\in\text{TEST}}\{\hat{\alpha}(D_i, Z_{i,\cdot}) - \alpha_0(D_i, Z_{i,\cdot})\}^2\right],$$

which is the same relaxation of mean square error as before. The bound requires both m and p to increase, $m^{1/2} \ll p \ll m^2$, and $\rho_{\min} \gg p^{-1/2} \vee m^{-1/2} \vee \Delta_E$. For the bound to be meaningful, (r, Δ_E) must be simultaneously well behaved and the corruption dependence must be weak. Finally, the bound includes the nonlinear approximation error Δ_ζ and the size of theoretical device $\|\eta^*\|_1$. In summary, we keep track of the low rank approximation error Δ_E and the nonlinear sparse approximation error Δ_ζ. Nonlinear factor models imply bounds on Δ_E. Due to our doubly robust approach, inference for the causal parameter θ_0 is robust to non-vanishing Δ_ζ—a discussion we revisit below.

Theorem 5.3 innovates in all of the ways that Theorem 5.2 does and more. Most importantly, it analyzes a new PCR estimator for a new estimand: the error-in-variable balancing weight. A rich literature proposes balancing weight estimators for causal inference with clean data, but to our knowledge, ours is the first error-in-variable balancing weight estimator for causal inference with corrupted data. As developed in Appendix F, Theorem 5.3 actually holds for a broad class of counterfactual moments and therefore a broad class of causal parameters. Moreover, the counterfactual moment $\hat{M} = [\{b(D^{\text{TRAIN}}, \hat{\boldsymbol{X}})\}^T Y^{\text{TRAIN}}]$ recovers error-in-variable regression. We choose not to simply subsume Theorem 5.2 by Theorem 5.3 for two reasons. First, doing so would require that Y_i and ε_i are bounded, which rules out differential privacy for the outcome; see the Appendix A discussion. Second, Theorem 5.2 has lower powers of $(r, \rho_{\min}^{-1})$ and avoids the term $\frac{m}{p^2}$ so it is typically a tighter bound.

5.4 Step 4: Causal estimation and inference

Before stating our results, we formalize the sense in which the corrupted data problem is an extended semiparametric problem. Let $W_{i,\cdot} = (D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ concatenate the signal and the noise, so that $\mathbb{L}_2(\mathcal{W})$ consists of square integrable functions of the form $\gamma : (D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) \to \mathbb{R}$. Both the true regression $\gamma_0(D_i, X_{i,\cdot})$ and our error-in-variable estimator $\hat{\gamma}(D_i, Z_{i,\cdot})$ belong to this space, which serves as our hypothesis space for semi-

parametric analysis. With this theoretical background, we formalize our distribution shift assumptions.

Assumption 5.9 (Marginal distribution shift)**.** *The extended outcome and treatment mechanisms,* $\mathbb{E}[Y_i|D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}]$ *and* $\mathbb{E}[D_i|X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}]$*, do not vary across observations.*

Assumption 5.9 implies that $\gamma_0(W_{i,\cdot})$ and $\alpha_0(W_{i,\cdot})$ do not vary across observations, though the marginal distributions $\mathbb{P}_i(W_i)$ may vary. Our corruption model implies $\gamma_0(W_{i,\cdot}) = \gamma_0(D_i, X_{i,\cdot})$, and we are agnostic about whether $\alpha_0(W_{i,\cdot}) = \alpha_0(D_i, X_{i,\cdot})$ for the extended hypothesis space.[6] We place one final assumption, mildly strengthening common support.

Assumption 5.10 (Bounded propensity)**.** *The extended propensity score is bounded away from zero and one, i.e.* $1 - \bar{\phi} \leq \mathbb{E}[D_i|X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}] \leq \bar{\phi}$*.*

We introduce some additional notation to state the finite sample Gaussian approximation. Define the oracle influences $\psi_i = \psi(W_{i,\cdot}, \theta_i, \gamma_0, \alpha_0)$ where the influence function is

$$\psi(W_{i,\cdot}, \theta, \gamma, \alpha) = \gamma(1, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) - \gamma(0, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) + \alpha(W_{i,\cdot})\{Y_i - \gamma(W_{i,\cdot})\} - \theta.$$

$\mathbb{E}[\psi_i] = 0$ since $\mathbb{E}[\gamma_0(1, X_{i,\cdot}) - \gamma_0(0, X_{i,\cdot})] = \theta_i$ and $\mathbb{E}[\alpha_0(W_{i,\cdot})\{Y_i - \gamma_0(W_{i,\cdot})\}] = 0$ by law of iterated expectations. We define the higher moments and average higher moments by

$$\sigma_i^2 = \mathbb{E}[\psi_i^2], \quad \xi_i^3 = \mathbb{E}[|\psi_i|^3], \quad \chi_i^4 = \mathbb{E}[\psi_i^4]; \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2, \quad \xi^3 = \frac{1}{n}\sum_{i=1}^n \xi_i^3, \quad \chi^4 = \frac{1}{n}\sum_{i=1}^n \chi_i^4.$$

Remark 5.5 (General causal parameters)**.** *In the main text, we state results that hold for a broad class of causal parameters, with parameter specific constants* $(\bar{Q}, \bar{q})$ *in Theorems 5.4 and 5.5. For ATE,* $\bar{Q} = 2\left(\frac{1}{\phi} + \frac{1}{1-\phi}\right)$ *and* $\bar{q} = 1$ *under Assumptions 5.9 and 5.10. In Appendix G, we characterize* $(\bar{Q}, \bar{q})$ *for several other examples under generalizations of Assumptions 5.9 and 5.10.*

Theorem 5.4 (Finite sample Gaussian approximation)**.** *Suppose Assumptions 5.9 and 5.10 hold,* $\mathbb{E}[\varepsilon_i^2 \mid W_{i,\cdot}] \leq \bar{\sigma}^2$*, and* $\|\alpha_0\|_\infty \leq \bar{\alpha}$*. Further suppose that for* $(i, j) \in \text{TEST}$*,*

$$\hat{\gamma}(W_{i,\cdot}) \perp\!\!\!\perp \hat{\gamma}(W_{j,\cdot})|\text{TRAIN}, \quad \hat{\alpha}(W_{i,\cdot}) \perp\!\!\!\perp \hat{\alpha}(W_{j,\cdot})|\text{TRAIN}.$$

[6] If $\mathbb{E}[D_i|X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}] = \mathbb{E}[D_i|X_{i,\cdot}]$, then $\alpha_0(W_{i,\cdot}) = \alpha_0(D_i, X_{i,\cdot})$.

Then with probability $1-\epsilon$,

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{n^{1/2}}{\sigma}(\hat{\theta}-\theta_0)\le z\right\}-\Phi(z)\right|\le c^{BE}\left(\frac{\xi}{\sigma}\right)^3 n^{-\frac{1}{2}}+\frac{\Delta}{(2\pi)^{1/2}}+\epsilon,$$

where $\Phi(z)$ *is the standard Gaussian cumulative distribution function,* $c^{BE}=0.5600$*, and*

$$\Delta=\frac{3L}{\epsilon\sigma}\left[(\bar{Q}^{1/2}+\bar{\alpha})\{\mathcal{R}(\hat{\gamma})\}^{\bar{q}/2}+\bar{\sigma}\{\mathcal{R}(\hat{\alpha})\}^{1/2}+\{n\mathcal{R}(\hat{\gamma})\mathcal{R}(\hat{\alpha})\}^{1/2}\right].$$

Theorem 5.5 (Finite sample variance estimation)**.** *Suppose Assumptions 5.9 and 5.10 hold,* $\mathbb{E}[\varepsilon_i^2\mid W_{i,\cdot}]\le\bar{\sigma}^2$*, and* $\|\hat{\alpha}\|_\infty\le\bar{\alpha}'$*. Then with probability* $1-\epsilon'$*,*

$$|\hat{\sigma}^2-(\sigma^2+\text{BIAS})|\le\Delta'+\Delta''+3[(\Delta')^{1/2}\{(\Delta'')^{1/2}+\sigma+\Delta_{\text{OUT}}^{1/2}\}+(\Delta'')^{1/2}\{\Delta_{\text{OUT}}^{1/2}+(\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\}+(\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\sigma],$$

where $\text{BIAS}=\Delta_{\text{OUT}}+2\Delta_{\text{OUT}}^{1/2}\sigma$*,* $\Delta_{\text{OUT}}=\frac{1}{n}\sum_{i=1}^n[(\theta_i-\theta_0)^2]$*, and*

$$\Delta'=4(\hat{\theta}-\theta_0)^2+\frac{24L}{\epsilon'}\left[\{\bar{Q}+(\bar{\alpha}')^2\}\mathcal{R}(\hat{\gamma})^{\bar{q}}+\bar{\sigma}^2\mathcal{R}(\hat{\alpha})\right],\quad \Delta''=\left(\frac{2}{\epsilon'}\right)^{1/2}\chi^2 n^{-\frac{1}{2}}.$$

Corollary 5.3 (Confidence interval coverage)**.** *Suppose the conditions of Theorems 5.4 and 5.5 hold. Further assume*

1. *Moment regularity:* $\{(\xi/\sigma)^3+\chi^2\}n^{-\frac{1}{2}}\to 0$*;*
2. *Error-in-variable regression rate:* $\left(\bar{Q}^{1/2}+\bar{\alpha}/\sigma+\bar{\alpha}'\right)\{\mathcal{R}(\hat{\gamma})\}^{\bar{q}/2}\to 0$*;*
3. *Error-in-variable balancing weight rate:* $\bar{\sigma}\{\mathcal{R}(\hat{\alpha})\}^{1/2}\to 0$*;*
4. *Product of rates is fast:* $\{n\mathcal{R}(\hat{\gamma})\mathcal{R}(\hat{\alpha})\}^{1/2}/\sigma\to 0$*.*

Then $\hat{\theta}\xrightarrow{p}\theta_0$, $\hat{\sigma}^2\xrightarrow{p}\sigma^2+\text{BIAS}$, $\mathbb{P}\{\theta_0\in(\hat{\theta}\pm 1.96\hat{\sigma}n^{-1/2})\}\to 0.95+c$, $\text{BIAS}, c\ge 0$.

If in addition $\Delta_{\text{OUT}}\to 0$*, i.e. there are not too many outliers,*

$$\textit{then}\quad \hat{\theta}\xrightarrow{p}\theta_0,\quad \hat{\sigma}^2\xrightarrow{p}\sigma^2,\quad \mathbb{P}\{\theta_0\in(\hat{\theta}\pm 1.96\hat{\sigma}n^{-1/2})\}\to 0.95.$$

Remark 5.6 (General causal parameters and data cleaning)**.** *Corollary 5.3 holds, as stated, for a broad class of semiparametric estimands such as the average elasticity and nonparametric estimands such as heterogeneous treatment effects. Moreover, it holds for not only the data cleaning and estimation procedure that we propose, but for any data cleaning and estimation procedure satisfying its weak conditions.*

The individual rate conditions $\mathcal{R}(\hat{\gamma}) \to 0$ and $\mathcal{R}(\hat{\alpha}) \to 0$ as well as the product rate condition $\{n\mathcal{R}(\hat{\gamma})\mathcal{R}(\hat{\alpha})\}^{1/2} \to 0$ suffice for Gaussian approximation with standard deviation $\sigma n^{-1/2}$, generalizing the main result in Chernozhukov et al. (2021) to the harder setting with corrupted and i.n.i.d. data. These rate conditions are in terms of a more general norm than previous work because of matrix completion in the data cleaning step. Nonetheless, we recover a familiar product rate condition from semiparametric theory. The conditions solve the two remaining theoretical challenges. First, they provide a framework to translate an on-average data cleaning guarantee into a data cleaning-adjusted confidence interval for the causal parameter, by using generalized norms. Second, they ensure that the standard deviation is $\sigma n^{-1/2}$ as long as the *product* of error-in-variable rates (and hence the product of data cleaning rates) is of order $n^{-1/2}$. In summary, they allow for causal inference at rates faster than matrix completion, which will be essential for the leading application, where we desire precision for the population while maintaining privacy for individuals.

A major technical innovation is semiparametric variance estimation in the i.n.i.d. setting, which is essential to the validity of confidence intervals. We define Δ_{OUT} to quantify the frequency of outliers. Since $\theta_i = \mathbb{E}[\gamma_0(1, X_{i,\cdot}) - \gamma_0(0, X_{i,\cdot})]$, Δ_{OUT} quantifies the shift in the marginal distributions of true covariates $\mathbb{P}_i(X_{i,\cdot})$. At best, $\Delta_{\text{OUT}} = 0$ in the i.i.d. case. At worst, Δ_{OUT} is a constant (when individual treatment effects are bounded). The condition $\Delta_{\text{OUT}} \to 0$, i.e. relatively few outliers, suffices for consistent variance estimation and nominal confidence intervals. When $\Delta_{\text{OUT}} \not\to 0$, our variance estimator is asymptotically biased upwards by $\text{BIAS} = \Delta_{\text{OUT}} + 2\Delta_{\text{OUT}}^{1/2}\sigma$, implying conservative confidence intervals. At worst, our confidence intervals are valid but conservative by a theoretically quantifiable amount. Our exact characterization of BIAS may have broader consequences for variance upper bounds and design-based uncertainty, which we pose as a direction for future work.[7]

5.5 Key assumption holds for nonlinear factor models

Finally, we tie together our various results and revisit our key assumption that covariates are approximately low rank. We show that nonlinear factor models (i) encode the intuition of approximate repeated measurements; (ii) imply that covariates are approximately low

[7]We thank Isaiah Andrews for suggesting this connection.

rank; and (iii) satisfy the rate conditions for causal inference. In a nonlinear factor model, $X_{ij} = g(\lambda_i, \mu_j)$ where (λ_i, μ_j) are latent factors corresponding to units and covariates, respectively. We assume that the function g is smooth in its second argument, formalizing the repeated measurement intuition.

Definition 5.1 (Hölder class)**.** *The Hölder class $\mathcal{H}(q, S, C_H)$ on $[0,1)^q$ is the set of functions $g : [0,1)^q \to \mathbb{R}$ whose partial derivatives satisfy*

$$\sum_{s:|s|=\lfloor S \rfloor} \frac{1}{s!} |\nabla_s g(\mu) - \nabla_s g(\mu')| \leq C_H \left\|\mu - \mu'\right\|_{\max}^{S-\lfloor S \rfloor}, \quad \forall \mu, \mu' \in [0,1)^q,$$

where $\lfloor S \rfloor$ denotes the largest integer strictly smaller than S.

Assumption 5.11 (Generalized factor model)**.** *Assume $\boldsymbol{X}$ is generated as $X_{ij} = g(\lambda_i, \mu_j)$, where $\lambda_i, \mu_j \in [0,1)^q$ and $g(\lambda_i, \cdot) \in \mathcal{H}(q, S, C_H)$.*

A linear factor model is a special case of a generalized factor model where $g(\lambda_i, \mu_j) = \lambda_i^T \mu_j$. Such a model satisfies Definition 5.1 for all $S \in \mathbb{N}$ and $C_H = C$, for some absolute positive constant $C < \infty$. Assumption 5.11 also allows for smooth nonlinear factor models, and it implies joint control over (r, Δ_E) as desired. Intuitively, as latent dimension q increases, the rank r increases. As smoothness S increases, the approximation error Δ_E decreases. Our final result demonstrates that, as long as the ratio q/S is small enough, the data cleaning adjusted confidence intervals are valid.

Remark 5.7 (General dictionaries)**.** *In the main text, we state results that hold for a broad class of dictionaries, with the concise notation q' in Corollary 5.4. In Appendix H, we prove that $q' \leq d_{\max} q$, where $d_{\max}$ is the degree of the polynomial dictionary. We articulate restrictions on the class of dictionaries in Appendix C. For the interacted dictionary, $d_{\max} = 2$.*

Corollary 5.4. *Suppose the conditions of Theorems 5.2, 5.3, 5.4 and 5.5 hold, as well as Assumption 5.11. For simplicity, consider the semiparametric case where $\sigma, \bar{\sigma}, \bar{\alpha}, \bar{\alpha}', \bar{Q}$ are bounded above and $\bar{q} = 1$. If in addition*

1. *Moment regularity: $\{(\xi/\sigma)^3 + \chi^2\} n^{-\frac{1}{2}} \to 0$;*

2. *Weak dependence: $(K_a, \kappa, \bar{K}, \rho_{\min}^{-1})$ scale polynomially in $\ln(np)$;*

3. *Nonlinear sparse approximation:* $m\Delta_\phi \leq \|\beta^*\|_1^2, < \infty$*; and* $m\Delta_\zeta \leq \|\eta^*\|_1^2$*;*

4. *Enough repeated measurements:* $n^{\frac{2}{3}} \lesssim p \lesssim n^{\frac{3}{2}}$*, i.e.* $n = p^\upsilon$ *or* $p = n^\upsilon$ *for* $\upsilon \in [1, \frac{3}{2}]$*;*

5. *Small latent dimension to smoothness ratio:* $\frac{q'}{S} < \frac{3}{4} - \frac{\upsilon}{2}$.

Then the conclusions of Corollary 5.3 hold.

In summary, we allow either $n > p$ or $p > n$ as long as (n, p) increase at similar rates. Given (n, p), the ratio of the latent dimension q over smoothness S in the generalized factor model must be sufficiently low. For example, if $n = p$ and $q' = q$ then we require $q < \frac{S}{4}$: the latent dimension must be less than a quarter of the smoothness. A sufficiently low $\frac{q}{S}$ ratio ensures sufficiently fast learning rates $\mathcal{R}(\hat{\gamma})$ and $\mathcal{R}(\hat{\alpha})$ for causal inference with standard error $\hat{\sigma} n^{-1/2}$. For the special case of a linear factor model, the $\frac{q}{S}$ ratio constraint becomes vacuous, and there is no restriction on the latent dimension q. The same is true for a polynomial factor model where $g(\lambda_i, \mu_j) = \text{polynomial}(\lambda_i, \mu_j)$. The doubly robust framework allows us to slightly relax the conditions stated above and still obtain valid inference. In particular, either $\Delta_\phi \not\to 0$ or $\Delta_\zeta \not\to 0$, i.e. γ_0 or α_0 may be incorrectly specified. See Chernozhukov et al. (2022a, Section 2) for further discussion of mis-specification.

6 Case study: Effect of import competition using Census

Equipped with theoretical guarantees, we return to the motivating real world issue: measurement error, missing values, discretization, and differential privacy in the US Census. We replicate a seminal paper in labor economics that uses Census data, while introducing different types and levels of synthetic corruption. In particular, we implement differential privacy at the level mandated for the 2020 Census.

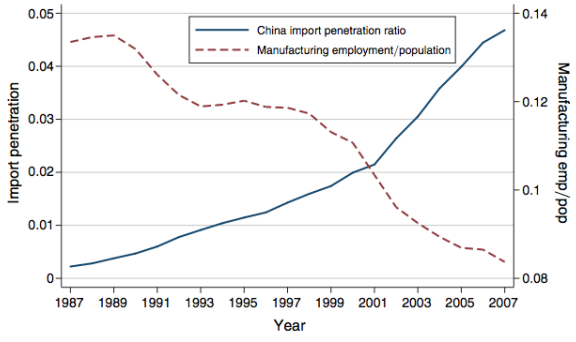


Figure 9: Is correlation causation? Autor et al. (2013, Figure 1)

6.1 Economic research question

We revisit the economic research question studied by Autor et al. (2013): what is the effect of import competition on local labor markets in the US? Figure 9 illustrates how, between the years 1987 and 2007, Chinese imports skyrocketed while US manufacturing employment plummeted. Phrased another way, the research question is whether this correlation is causal. We ask an additional question: can we recover the same effects after introducing various types and levels of synthetic corruption?

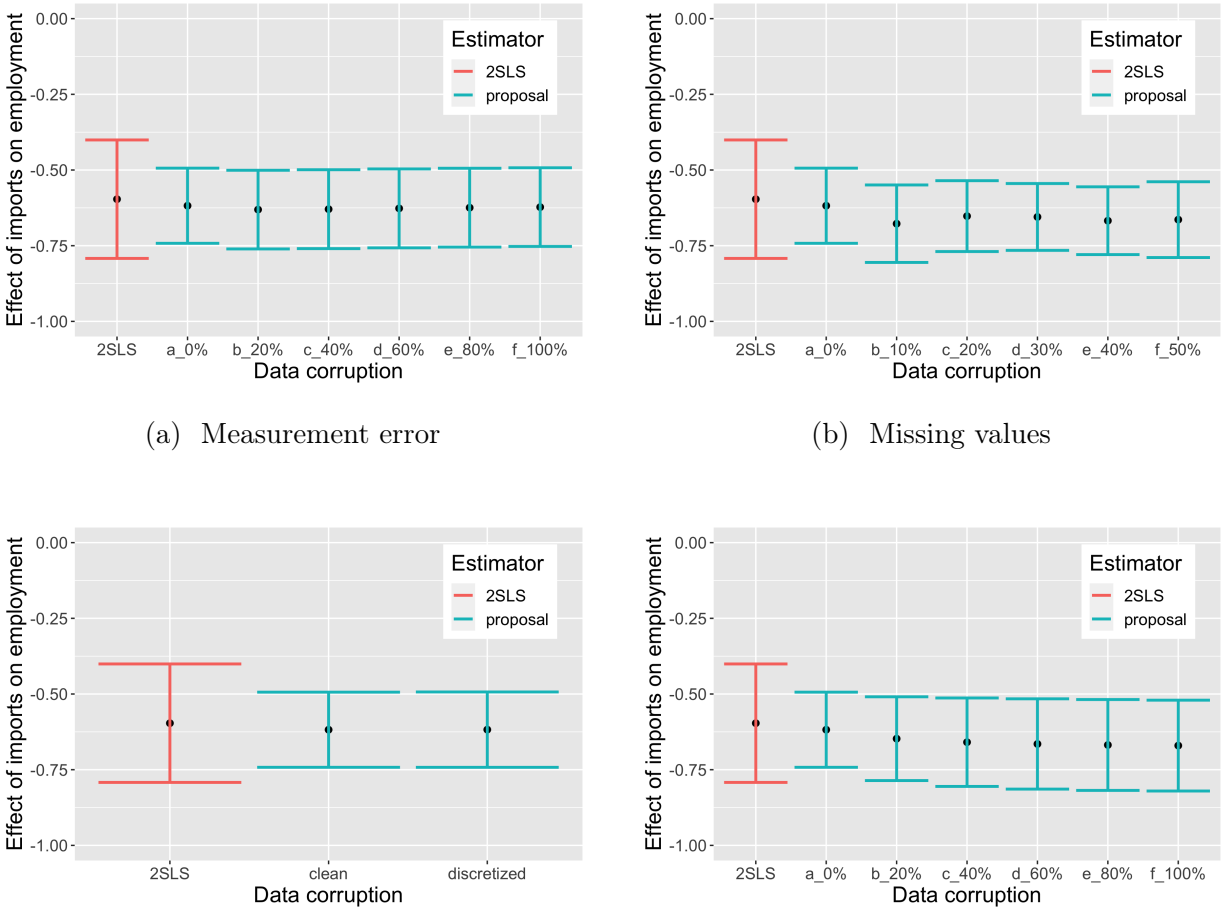
Following the original study, we use Census data at the commuting zone (CZ) level. A CZ is an aggregate unit interpretable as a local economy. 722 CZs make up the mainland US, and CZ data are constructed from individual microdata published by the US government: IPUMS, ACS, BEA, and SSA. The outcome Y_i is percent change in US manufacturing employment; the treatment D_i is percent change in imports from China; the instrument U_i is percent change in imports from China to other countries; and the covariates X_i are CZ characteristics. In more detail, we use data from two periods: 1990-2000 and 2000-2007, for a total of 1,444 observations. (Y_i, D_i, U_i) are changes within a period, e.g. the 2000 level minus the 1990 level, while $X_{i,\cdot}$ are levels at the beginning of a period, e.g. the 1990 level. The causal parameter is the partially linear instrumental variable regression parameter described in Appendix A.

6.2 Can we recover the same effects with data corruption?

In Section 3, we have already verified that the covariates $X_{i,\cdot}$ are approximately low rank. Both the original specification $X_{i,\cdot} \in \mathbb{R}^{14}$ and the augmented specification $X_{i,\cdot} \in \mathbb{R}^{30}$ include approximate repeated measurements, with variables such as percent employment in manufacturing, percent college educated, and percent employment among women. In particular, the approximate rank is five.

Figure 10 presents our first semi-synthetic exercises. For reference, we visualize in red the 2SLS point estimate and confidence interval of Autor et al. (2013), using clean data. Immediately next to Autor et al. (2013)'s results, we visualize our own point estimate and confidence interval with clean data. We recover essentially the same point estimate and a somewhat smaller confidence interval. The true covariates are approximately low rank,

our procedure exploits this fact, and therefore it is is more efficient. Subsequently, we implement our procedure with increasing levels of measurement error: 20%, 40%, 60%, 80%, and 100% noise-to-signal ratio. Our point estimates remain stable, and the standard errors subtly increase in length. We obtain similar results with missing values, discretization, and differential privacy: point estimates remain stable and the standard errors adaptively increase in length for higher noise-to-signal ratios.



(a) Measurement error

(b) Missing values

(c) Discretization

(d) Differential privacy

Figure 10: Synthetic corruption

6.3 Formalizing privacy

Next, we calibrate the semi-synthetic exercises to privacy levels mandated by the US Census Bureau. To do so, we first clarify how our model of causal inference with corrupted data accommodates various notions of differential privacy. With these formal results, we can then

calibrate the variance of the Laplacian noise appropriately. In what follows, we focus on a one-off data release (formally called the non-interactive setting). There are two relevant privacy concepts: central differential privacy of summary tables such as the Census, and per instance differential privacy of microdata such as the current population survey (CPS).

To begin, we define central differential privacy. To be concrete, we maintain the following thought experiment: we are the Census Bureau, and our goal is to publish Autor et al. (2013)'s CZ-level aggregated data set while protecting the privacy of individuals within CZs. In particular, we have access to the individual-level microdata, which we will *not* publicly share; we will only publish the CZ-level summaries for aggregate units. Consider a particular commuting zone $i \in [n]$ with L_i individuals, and denote its individual-level microdata by $\boldsymbol{M}^{(i)} \in \mathbb{R}^{L_i \times p}$. We wish to publish p summary statistics $X_{i,\cdot}$ for this CZ, where $X_{ij} = \frac{1}{L_i} \sum_{\ell=1}^{L_i} M_{\ell j}^{(i)}$, however we wish to maintain plausible deniability that each individual $\ell \in [L_i]$ contributed their data. The simulated attack on the 2010 Census found that Census blocks summary tables did not maintain this plausible deniability.

Definition 6.1 (Central differential privacy for summary tables)**.** *A randomized mechanism $\mathcal{M}$ confers central differential privacy with privacy loss ϵ if and only if for any two possible individual-level data sets $\boldsymbol{M}$ and $\boldsymbol{M}'$ differing in a single row, and for all events E in the range of $\mathcal{M}$,*

$$\frac{\mathbb{P}(\mathcal{M}(\boldsymbol{M}) \in E)}{\mathbb{P}(\mathcal{M}(\boldsymbol{M}') \in E)} \leq e^{\epsilon}$$

where the randomness is with respect to $\mathcal{M}$.

The canonical mechanism that achieves central differential privacy is to publish $\mathcal{M}(\boldsymbol{M}^{(i)}) = X_{i,\cdot} + H_{i,\cdot}$ instead of $X_{i,\cdot}$, where $H_{i,\cdot}$ is Laplacian noise with an appropriately calibrated variance.[8] In addition to the Laplace mechanism (Dwork et al., 2006), the discrete (Duchi et al., 2018) and piece-wise uniform (Wang et al., 2019) mechanisms induce measurement error that is sub-exponential and mean zero, which fits within our framework. For simplicity, we focus on the Laplace mechanism when relating privacy to our theoretical results.

Proposition 6.1 (Strong protections for aggregate data)**.** *Suppose*

1. *Each entry of microdata is bounded, i.e. $|M_{\ell j}^{(i)}| \leq \bar{A}_i$;*

[8] More precisely, $\mathcal{M}_i : \mathbb{R}^{L_i \times p} \to \mathbb{R}^p$.

2. No individual appears in two commuting zones.

Then the mechanism $Z_{ij} = X_{ij} + H_{ij}$ *where* $H_{ij} \overset{i.i.d.}{\sim} Laplace\left(\frac{2\bar{A}_i}{\epsilon}\frac{p}{L_i}\right)$ *confers* ϵ *central differential privacy and the measurement error parameters satisfy* $K_a, \kappa \leq \max_{i\in[n]} \frac{2^{3/2}\bar{A}_i}{\epsilon}\frac{p}{L_i}$.

In summary, the calibrated Laplacian variance depends on the privacy loss ϵ, the most extreme true value $\bar{A}_i$, the number of covariates p, and number of individuals L_i per aggregate unit. The condition $\max_{i\in[n]} \frac{p}{L_i} \lesssim \ln(np)$ implies that the measurement error parameters (K_a, κ) diverge slowly with (n, p), so that our rates of data cleaning and error-in-variable estimation translate into data cleaning adjusted confidence intervals. This auxiliary condition is a practical diagnostic: roughly speaking, the number of published covariates should not greatly exceed the number of individuals per aggregate unit.

Remark 6.1 (Limitations for aggregate data)**.** *The theoretical condition* $\max_{i\in[n]} \frac{p}{L_i} \lesssim \ln(np)$ *sheds new light on limitations. It is perhaps plausible for commuting zones, but implausible for Census blocks, which have fewer individuals per aggregate unit. Fortunately, much empirical economic research studies commuting zones, which we study in our semi-synthetic exercise. An important direction for future research is to empirically investigate, through simulated attacks, how vulnerable various data releases may be for different* $\frac{p}{L_i}$ *regimes.*

In addition to tabular summaries, the Bureau publishes microdata, for which an alternative definition of privacy is necessary. We maintain the following thought experiment: we are the Census Bureau, and our goal is to publish the microdata that is subsequently aggregated by Autor et al. (2013) at the CZ level. Now, the index $i \in [n]$ corresponds to the i-th individual with covariates $X_{i,\cdot} \in \mathbb{R}^p$ that we wish to publish while maintaining plausible deniability that any given covariate profile is contained in the data release.

Definition 6.2 (Per instance differential privacy for microdata)**.** *A randomized mechanism* $\mathcal{M}$ *confers per instance differential privacy with privacy loss* ϵ *if and only if for any two possible individual-level vectors* x *and* x'*, and for all events* E *in the range of* $\mathcal{M}$*,*

$$\frac{\mathbb{P}(\mathcal{M}(x) \in E)}{\mathbb{P}(\mathcal{M}(x') \in E)} \leq e^{\epsilon}$$

where the randomness is with respect to $\mathcal{M}$*.*

A mechanism that achieves per instance differential privacy is to publish $\mathcal{M}(X_{i,\cdot}) = X_{i,\cdot} + H_{i,\cdot}$ instead of $X_{i,\cdot}$, where $H_{i,\cdot}$ is Laplacian noise with an appropriately calibrated variance. We focus on the Laplace mechanism when relating privacy to our theoretical results. In anticipation of a more qualified privacy guarantee, we consider adding noise to only T out of the p covariates.

Proposition 6.2 (Weaker protections for microdata)**.** *Suppose Assumption 5.1 holds. The mechanism* $Z_{ij} = X_{ij} + H_{ij}$ *with* $j \in [T]$*, where* $H_{ij} \overset{i.i.d.}{\sim} Laplace\left(\frac{2\bar{A}T}{\epsilon}\right)$*, confers* ϵ *per instance differential privacy for the initial* T *covariates. Moreover, the measurement error parameters satisfy* $K_a, \kappa \leq \frac{2^{3/2}\bar{A}T}{\epsilon}$.

In summary, the calibrated Laplacian variance depends on the privacy loss ϵ and the subset size T. The condition $T \ll (n, p)$ implies that the measurement error parameters (K_a, κ) diverge slowly with (n, p), so that our rates of data cleaning and error-in-variable estimation translate into data cleaning adjusted confidence intervals. This qualification, that only a subset of variables may be privatized, is an example of event level rather than user level differential privacy: with microdata, only some aspects of an individual's identity are protected, while with summary tables all aspects are protected.

6.4 Privacy calibrated to 2020 US Census levels

As our second semi-synthetic exercise, we implement central differential privacy for Autor et al. (2013)'s CZ-level aggregated data set while protecting the privacy of individuals within CZs. We calibrate the Laplacian variance according to Proposition 6.1, where $\epsilon = 17.14$ according to a Bureau memo, $p = 30$ in the augmented specification, and $(\bar{A}_i, L_i)$ are calculated from the microdata for each CZ. As before, we visualize in red the point estimate and confidence interval of Autor et al. (2013) using clean data in Figure 11. Next to that, we visualize the point estimate and confidence interval we pro-

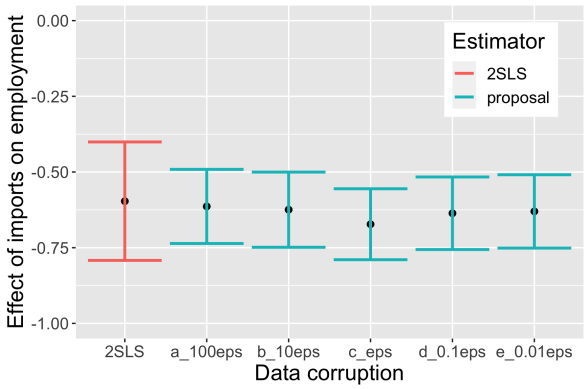


Figure 11: Calibration

pose. To study the robustness of our results to the privacy loss parameter, we consider $(100\epsilon, 10\epsilon, \epsilon, 0.1\epsilon, 0.01\epsilon)$, which corresponds to adding Laplacian noise both below and above the mandated level. Across levels, our point estimates and confidence intervals remain remarkable stable.

7 Conclusion

Recent developments in how the US Census Bureau will publish economic data motivate us to study a class of corruptions that is rich enough to encompass classical types of corruption, such as measurement error and missingness, as well as modern types, such as discretization and differential privacy. Abstractly, our goal is to learn a causal parameter from corrupted data; concretely, our goal is to characterize scenarios in which it is possible to achieve both privacy and precision. To do so, we propose new data cleaning-adjusted confidence intervals that are computationally simple, statistically rigorous, and empirically robust in settings calibrated to empirical economic research. We build a framework to use matrix completion as data cleaning for downstream causal inference, bridging two rich literatures. For future research, we pose the question of how to extend our results to confounded noise and sample selection bias. The modular structure of this paper provides a template to do so.

References

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843.

Abowd, J. M. and Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202.

Agarwal, A., Shah, D., and Shen, D. (2020a). On principal component regression in a high-dimensional error-in-variables setting. *arXiv:2010.14449*.

Agarwal, A., Shah, D., and Shen, D. (2020b). Synthetic interventions. *arXiv:2006.07691*.

Agarwal, A., Shah, D., Shen, D., and Song, D. (2021). On robustness of principal component regression. *Journal of the American Statistical Association*, pages 1–34.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Anatolyev, S. and Mikusheva, A. (2021). Factor models with many assets: Strong factors, weak factors, and the two-pass procedure. *Journal of Econometrics*.

Andrews, D. W. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, pages 43–72.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2019). Synthetic difference in differences. Technical report, National Bureau of Economic Research.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, pages 1–15.

Autor, D. H., Dorn, D., and Hanson, G. H. (2013). The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review*, 103(6):2121–68.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29.

Bai, J. and Ng, S. (2019). Matrix completion, counterfactuals, and factor analysis of missing data. *arXiv:1910.06677*.

Bai, J. and Wang, P. (2014). Identification theory for high dimensional static and dynamic factor models. *Journal of Econometrics*, 178(2):794–804.

Battistin, E. and Chesher, A. (2014). Treatment effect estimation with covariate measurement error. *Journal of Econometrics*, 178(2):707–715.

Ben-Michael, E., Feller, A., Hirshberg, D. A., and Zubizarreta, J. R. (2021). The balancing act in causal inference. *arXiv:2110.14831*.

Benedetto, G., Linse, K., and Parker, E. (2022). Improving disclosure avoidance procedures for the Current Population Survey public use file. *United States Census Bureau*.

Bia, M., Huber, M., and Lafférs, L. (2020). Double machine learning for sample selection models. *arXiv:2012.00745*.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press Baltimore.

Bruns-Smith, D. A. and Feller, A. (2022). Outcome assumptions and duality theory for balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 11037–11055. PMLR.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Chattopadhyay, A. and Zubizarreta, J. R. (2021). On the implied weights of linear regression for causal inference. *arXiv:2104.06581*.

Chen, X., Hong, H., and Nekipelov, D. (2011). Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–37.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1).

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv:1608.00033*.

Chernozhukov, V., Newey, W. K., and Singh, R. (2021). A simple and general debiased machine learning theorem with finite sample guarantees. *arXiv:2105.15197*.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022a). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022b). Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal*.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2018b). A t-test for synthetic controls. *arXiv:1812.10820*.

Chetty, R. and Hendren, N. (2018a). The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3):1107–1162.

Chetty, R. and Hendren, N. (2018b). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228.

D'Agostino Jr, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451):749–759.

Das, M., Newey, W. K., and Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70(1):33–58.

Datta, A. and Zou, H. (2017). Cocolasso for high-dimensional error-in-variables regression. *Annals of Statistics*, 45(6):2400–2426.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

Deaner, B. (2018). Proxy controls and panel data. *arXiv:1810.00283*.

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, 36(2):665–685.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, pages 265–284.

Fan, J., Wang, W., and Zhong, Y. (2018). An l-∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42.

Feng, Y. (2020). Causal inference in possibly nonlinear factor models. *arXiv:2008.13651*.

Fernández-Val, I., Freeman, H., and Weidner, M. (2020). Low-rank approximations of nonseparable panel models. *arXiv:2010.12439*.

Foster, D. J. and Syrgkanis, V. (2019). Orthogonal statistical learning. *arXiv:1901.09036*.

Gavish, M. and Donoho, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053.

Griliches, Z. (1986). Economic data issues. *Handbook of Econometrics*, 3:1465–1514.

Hasminskii, R. Z. and Ibragimov, I. A. (1979). On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank SVD via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402.

Hausman, J. A. and Wise, D. A. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica*, pages 455–473.

Hawes, M. (2021). The Census Bureau's simulated reconstruction-abetted re-identification attack on the 2010 Census. *United States Census Bureau*.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, pages 153–161.

Hotz, V. J., Bollinger, C. R., Komarova, T., Manski, C. F., Moffitt, R. A., Nekipelov, D., Sojourner, A., and Spencer, B. D. (2022). Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences*, 119(31):e2104906119.

Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216.

Huber, M. (2014). Treatment evaluation in the presence of sample selection. *Econometric Reviews*, 33(8):869–905.

Jarmin, R. (2019). Census Bureau adopts cutting edge privacy protections for 2020 Census. *United States Census Bureau*.

Jöreskog, K. G. and Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Kallus, N., Mao, X., and Udell, M. (2018). Causal inference with noisy and missing covariates via matrix factorization. *arXiv:1806.00811*.

Keshavan, R., Montanari, A., and Oh, S. (2009). Matrix completion from noisy entries. *Advances in Neural Information Processing Systems*, 22.

Klaassen, C. A. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620.

Li, T. (2002). Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics*, 110(1):1–26.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65(2):139–165.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Mayer, I., Sverdrup, E., Gauss, T., Moyer, J.-D., Wager, S., and Josse, J. (2020). Doubly robust treatment effect estimation with missing attributes. *Annals of Applied Statistics*, 14(3):1409–1431.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Miao, W. and Tchetgen, E. J. T. (2018). A confounding bridge approach for double negative control inference on causal effects. *arXiv:1808.04945*.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, pages 1349–1382.

Newey, W. K. (2001). Flexible simulated moment estimation of nonlinear errors-in-variables models. *Review of Economics and Statistics*, 83(4):616–627.

Onatski, A. (2009). A formal statistical test for the number of factors in the approximate factor models. *Econometrica*, 77(5):1447–1480.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Poterba, J. M., Venti, S. F., and Wise, D. A. (1996). How retirement saving programs increase saving. *Journal of Economic Perspectives*, 10(4):91–112.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, pages 931–954.

Rosenbaum, M. and Tsybakov, A. B. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 276–290. Institute of Mathematical Statistics.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

Rotnitzky, A., Smucler, E., and Robins, J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Schennach, S. M. (2007). Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica*, 75(1):201–239.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, 8:341–377.

Schennach, S. M. and Hu, Y. (2013). Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association*, 108(501):177–186.

Shevtsova, I. (2010). An improvement of convergence rate estimates in the Lyapunov theorem. In *Doklady Mathematics*, volume 82, pages 862–864. Springer.

Singh, R. (2020). Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. *arXiv:2012.10315*.

Singh, R. (2021). Generalized kernel ridge regression for causal inference with missing-at-random sample selection. *arXiv:2111.05277*.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4595–4607.

Singh, R., Xu, L., and Gretton, A. (2020). Reproducing kernel methods for nonparametric and semiparametric treatment effects. *arXiv:2010.04855*.

Steed, R., Liu, T., Wu, Z. S., and Acquisti, A. (2022). Policy impacts of statistical uncertainty and privacy. *Science*, 377(6609):928–931.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An introduction to proximal causal learning. *arXiv:2009.10982*.

van der Laan, M. J. and Rose, S. (2018). *Targeted Learning in Data Science*. Springer.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press.

Wang, L. and Hsiao, C. (2011). Method of moments estimation and identifiability of semiparametric nonlinear errors-in-variables models. *Journal of Econometrics*, 165(1):30–44.

Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S. C., Shin, H., Shin, J., and Yu, G. (2019). Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649. IEEE.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Annals of Statistics*, 45(3):1342.

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Wold, H. (1982). Soft modelling: The basic design and some extensions. *Systems Under Indirect Observation, Part II*, pages 36–37.

Wold, S., Ruhe, A., Wold, H., and Dunn III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Xiong, R. and Pelger, M. (2019). Large dimensional latent factor modeling with missing observations and applications to causal inference. *arXiv:1910.08273*.

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer Science & Business Media.

Appendix

Table of Contents

A Additional examples

A.1 Semiparametric estimands

We consider the goal of estimation and inference of some causal parameter $\theta_0 \in \mathbb{R}$ which is a scalar summary of the regression γ_0, e.g. a treatment effect, policy effect, or elasticity. We consider a class of causal parameters of the form

$$\theta_0 = \frac{1}{n}\sum_{i=1}^{n} \theta_i, \quad \theta_i = \mathbb{E}[m(W_{i,\cdot}, \gamma_0)]$$

in an i.n.i.d. data generating process of the form

$$Y_i = \gamma_0(D_i, X_{i,\cdot}) + \varepsilon_i, \quad Z_{i,\cdot} = (X_{i,\cdot} + H_{i,\cdot}) \odot \pi_{i,\cdot}, \quad W_{i,\cdot} = (D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}).$$

This model is a restatement of (4). $(D_i, X_{i,\cdot})$ concatenate the various arguments of γ_0, which we hereby call regressors. Rather than observing $X_{i,\cdot}$, we observe $Z_{i,\cdot}$. This model includes the scenario in which some variables are corrupted and other are not. Which regressors are corrupted or uncorrupted constrains the construction of technical regressors; see Appendix C. We concatentate signal and noise as $W_{i,\cdot} = (D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$. A generalization of Assumption 5.9 imposes invariance of the regression γ_0 and generalized balancing weight α_0 across observations, which we formalize in Appendix G.

Example A.1 (Average treatment effect)**.** *Let $(D_i, X_{i,\cdot})$ concatenate treatment $D_i \in \{0, 1\}$ and covariates $X_{i,\cdot} \in \mathbb{R}^p$. Denote $\gamma_0(D_i, X_{i,\cdot}) := \mathbb{E}[Y_i | D_i, X_{i,\cdot}]$. Under the assumption of selection on observables, the average treatment effect is given by*

$$\theta_i = \mathbb{E}[\gamma_0(1, X_{i,\cdot}) - \gamma_0(0, X_{i,\cdot})].$$

With uncorrupted treatment and corrupted covariates, $W_{i,\cdot} = (D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ where $(H_{i,\cdot}, \pi_{i,\cdot})$ are measurement error and missingness for the covariates.[9]

While the true regression $\gamma_0(D_i, X_{i,\cdot})$ is only a function of signal $(D_i, X_{i,\cdot})$, our regression estimator $\hat{\gamma}(D_i, Z_{i,\cdot})$ is a function of both signal and noise $W_{i,\cdot}$. In other words, the hypothesis

[9] More generally, treatment observations may be corrupted as well. For readability, we exposit the simpler and plausible case that treatment is uncorrupted.

space for estimation is the extended space of functions $\mathbb{L}_2(\mathcal{W})$, and we must define an extended functional over $\mathbb{L}_2(\mathcal{W})$. In Example A.1, the extended functional is

$$\gamma \mapsto \mathbb{E}[\gamma(1, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) - \gamma(0, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})].$$

Example A.2 (Local average treatment effect)**.** *Let $(U_i, X_{i,\cdot})$ concatenate instrument $U_i \in \{0, 1\}$ and covariates $X_{i,\cdot} \in \mathbb{R}^p$. Denote $\gamma_0(U_i, X_{i,\cdot}) := \mathbb{E}[Y_i|U_i, X_{i,\cdot}]$ and $\delta_0(U_i, X_{i,\cdot}) := \mathbb{E}[D_i|U_i, X_{i,\cdot}]$. Under standard instrumental variable assumptions, the local average treatment effect for the subpopulation of compliers is given by*

$$\beta_0 = \frac{\theta_0}{\theta_0'}, \quad \theta_i = \mathbb{E}[\gamma_0(1, X_{i,\cdot}) - \gamma_0(0, X_{i,\cdot})], \quad \theta_i' = \mathbb{E}[\delta_0(1, X_{i,\cdot}) - \delta_0(0, X_{i,\cdot})].$$

With uncorrupted instrument and corrupted covariates, $W_{i,\cdot} = (U_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ where $(H_{i,\cdot}, \pi_{i,\cdot})$ are measurement error and missingness for the covariates.

Example A.3 (Average policy effect)**.** *Let $X_{i,\cdot} \in \mathbb{R}^p$ be the covariates. Consider the counterfactual transportation of covariates $x_{i,\cdot} \mapsto t(x_{i,\cdot})$. Denote $\gamma_0(X_{i,\cdot}) := \mathbb{E}[Y_i|X_{i,\cdot}]$. The average policy effect of transporting covariates is given by*

$$\theta_i = \mathbb{E}[\gamma_0\{t(X_{i,\cdot})\} - \gamma_0(X_{i,\cdot})].$$

With corrupted covariates, $W_{i,\cdot} = (X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ where $(H_{i,\cdot}, \pi_{i,\cdot})$ are measurement error and missingness for the covariates.

Example A.4 (Price elasticity of demand)**.** *Let Y_i be price of a particular good. Let $(D_i, X_{i,\cdot})$ concatenate quantities sold of the particular good D_i and other goods $X_{i,\cdot} \in \mathbb{R}^p$. Denote $\gamma_0(D_i, X_{i,\cdot}) = \mathbb{E}[Y_i|D_i, X_{i,\cdot}]$. The average price elasticity of demand is*

$$\theta_i = \mathbb{E}\left[\nabla_d \gamma_0(D_i, X_{i,\cdot})\right].$$

With uncorrupted quantity for the particular good and corrupted quantities for the other goods, $W_{i,\cdot} = (D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ where $(H_{i,\cdot}, \pi_{i,\cdot})$ are measurement error and missingness for the other goods.

A.2 Weighted estimands

In empirical economic research with aggregate units, it is common to weight units by their size. It is also common to consider partially linear models. For example, the estimand

of Autor et al. (2013) may be viewed as a weighted partially linear instrumental variable regression. To bridge theory with practice, we provide these examples next.

A weighted functional $\theta_0 \in \mathbb{R}$ is a scalar that takes the form

$$\theta_0 = \frac{1}{n}\sum_{i=1}^{n} \theta_i, \quad \theta_i = \mathbb{E}[\ell_i m(W_{i,\cdot}, \gamma_0)]$$

where ℓ_i is the weight for aggregate unit i. For simplicity, we take the weights ℓ_i to be known, but their uncertainty can be incorporated as well.

Example A.5 (Weighted partially linear regression)**.** *Let (D_i, X_i) concatenate treatment $D \in \mathbb{R}$ and covariates $X_{i,\cdot} \in \mathbb{R}^p$. Denote $\gamma_0(D_i, X_{i,\cdot}) = \mathbb{E}[Y_i|D_i, X_{i,\cdot}]$. The weighted partially regression coefficient is given by*

$$\theta_i = \mathbb{E}\left[\ell_i\{\gamma_0(d+1, X_{i,\cdot}) - \gamma_0(d, X_{i,\cdot})\}\right].$$

With uncorrupted treatment and corrupted covariates, $W_{i,\cdot} = (D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ where $(H_{i,\cdot}, \pi_{i,\cdot})$ are measurement error and missingness for the covariates.

Example A.6 (Weighted partially linear instrumental variable regression)**.** *Let $(U_i, X_{i,\cdot})$ concatenate instrument $U_i \in \mathbb{R}$ and covariates $X_{i,\cdot} \in \mathbb{R}^p$. Denote $\gamma_0(U_i, X_{i,\cdot}) := \mathbb{E}[Y_i|U_i, X_{i,\cdot}]$ and $\delta_0(U_i, X_{i,\cdot}) := \mathbb{E}[D_i|U_i, X_{i,\cdot}]$. Under standard instrumental variable assumptions, the weighted partially linear instrumental variable regression coefficient is given by*

$$\beta_0 = \frac{\theta_0}{\theta_0'}, \quad \theta_i = \mathbb{E}[\ell_i\{\{\gamma_0(u+1, X_{i,\cdot}) - \gamma_0(u, X_{i,\cdot})\}], \quad \theta_i' = \mathbb{E}[\ell_i\{\delta_0(u+1, X_{i,\cdot}) - \delta_0(u, X_{i,\cdot})\}].$$

With uncorrupted instrument and corrupted covariates, $W_{i,\cdot} = (U_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ where $(H_{i,\cdot}, \pi_{i,\cdot})$ are measurement error and missingness for the covariates.

A.3 Nonparametric estimands

A local functional $\theta_0^{\lim} \in \mathbb{R}$ is a scalar that takes the form

$$\theta_0^{\lim} = \lim_{h \to 0} \theta_0^h, \quad \theta_0^h = \frac{1}{n}\sum_{i=1}^{n} \theta_i^h, \quad \theta_i^h = \mathbb{E}[m_h(W_{i,\cdot}, \gamma_0)] = \mathbb{E}[\ell_h(W_{ij}) m(W_{i,\cdot}, \gamma_0)]$$

where ℓ_h is a Nadaraya Watson weighting with bandwidth h and W_{ij} is a scalar component of $W_{i,\cdot}$. $\theta_0^{\lim}$ is a nonparametric quantity. However, it can be approximated by the sequence

$\{\theta_0^h\}$. Each θ_0^h can be analyzed like a weighted functional as long as we keep track of how certain quantities depend on h. By this logic, finite sample semiparametric theory for θ_0^h translates to finite sample nonparametric theory for $\theta_0^{\lim}$ up to some approximation error. In this sense, our analysis encompasses both semiparametric and nonparametric estimands. As a leading example, we study heterogeneous treatment effects.

Example A.7 (Heterogeneous treatment effect)**.** *Let $(D_i, V_i, X_{i,\cdot})$ concatenate treatment $D_i \in \{0,1\}$, covariate of interest $V_i \in \mathbb{R}$, and other covariates $X_{i,\cdot} \in \mathbb{R}^p$. Denote $\gamma_0(D_i, V_i, X_{i,\cdot}) := \mathbb{E}[Y_i | D_i, V_i, X_{i,\cdot}]$. Under the assumption of selection on observables and identicial distribution of V_i, the heterogeneous treatment effect for the subpopulation with subcovariate value v is given by*

$$\theta_i = \mathbb{E}[\gamma_0(1, V_i, X_{i,\cdot}) - \gamma_0(0, V_i, X_{i,\cdot}) | V_i = v] = \lim_{h \to 0} \mathbb{E}[\ell_h(V_i)\{\gamma_0(1, V_i, X_{i,\cdot}) - \gamma_0(0, V_i, X_{i,\cdot})\}]$$

where

$$\ell_h(V_i) = \frac{K\left\{(V_i - v)/h\right\}}{\omega}, \quad \omega = \mathbb{E}[K\left\{(V_i - v)/h\right\}]$$

and K is the standard kernel function. With uncorrupted treatment, uncorrupted covariate of interest, and corrupted other covariates, $W_{i,\cdot} = (D_i, V_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ where $(H_{i,\cdot}, \pi_{i,\cdot})$ are measurement error and missingness for the other covariates.

In Appendix G, we formalize our general class of semiparametric and nonparametric estimands. We define the class abstractly and verify that each example belongs to the class under generalizations of Assumption 5.10.

A.4 Missing outcomes

So far, we have discussed corruption of the regressors $X_{i,\cdot}$ and taken outcome Y_i to be uncorrupted. Measurement error and differential privacy of Y_i is already captured by response noise ε_i. An important additional issue in empirical research is outcome attrition: for some observations, Y_i is missing. Moreover, the outcome attrition mechanism may depend on the true regressors. Our framework handles this case as well with light modification. The enriched observation model is

$$Y_i = \gamma_0(D_i, X_{i,\cdot}, S_i) + \varepsilon_i, \quad Z_{i,\cdot} = [X_{i,\cdot} + H_{i,\cdot}] \odot \pi_{i,\cdot}, \quad \tilde{Y}_i = Y_i \cdot S_i$$

where $S_i \in \{1, \texttt{NA}\}$ encodes attrition. Instead of observing $(Y_i, D_i, X_{i,\cdot})$ or even $(Y_i, D_i, Z_{i,\cdot})$, the analyst observes $(\tilde{Y}_i, D_i, Z_{i,\cdot})$. In the taxonomy of Rubin (1976), we allow outcome Y_i to be missing at random (MAR) conditional on true regressors $(D_i, X_{i,\cdot})$, of which $X_{i,\cdot}$ may be missing completely at random (MCAR) or may have measurement error. The extended semiparametric model is summarized by

$$\begin{aligned}\mathbb{E}[\tilde{Y}_i | D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}, S_i = 1] &= \mathbb{E}[Y_i | D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}, S_i = 1] \\ &= \mathbb{E}[Y_i | D_i, X_{i,\cdot}, S_i = 1] \\ &= \gamma_0(D_i, X_{i,\cdot}, S_i = 1).\end{aligned}$$

Our framework handles this extension by replacing Y_i with $\tilde{Y}_i$ and replacing $(D, X_{i,\cdot})$ with $(D_i, X_{i,\cdot}, S_i)$, similar to Singh (2021).

B Additional simulations and applications

In this appendix, we provide additional synthetic, semi-synthetic, and real results to argue

1. our procedure is robust to a broad variety of data shapes and sizes;
2. data corruption can flip the sign of OLS and TSLS some of the time;
3. our key assumption holds in popular Census data sets at different levels of granularity.

B.1 Robustness to data dimensions

In the main text, the each sample from the simulated data generating process produces a matrix of covariates $\boldsymbol{X} \in \mathbb{R}^{100 \times 100}$ with rank $r = 5$. In practice, economic data sets come in a variety of shapes and sizes. In this subsection, we ask: how robust is our end-to-end procedure across realistic dimensions of economic data? We consider the following variations of the simulated data generating process: $\boldsymbol{X} \in \mathbb{R}^{50 \times 200}$, $\mathbb{R}^{100 \times 100}$, $\mathbb{R}^{200 \times 50}$, $\mathbb{R}^{500 \times 20}$, and $\mathbb{R}^{1000 \times 10}$. For each choice of sample size n and covariate dimension p, we set the rank to be $r = \{\min(n, p)\}^{1/3}$. Across data dimensions, we introduce measurement error with the fixed noise-to-signal ratio of 20%. We consider the oracle tuning of the PCA hyperparameter $k = r$.

Table 12 quantifies coverage performance. Different rows correspond to different data dimensions. As in the main text, we record the average point estimates, which are close to $\theta_0 = 2.2$. Next, we record the average standard errors, which adaptively decrease in length for larger sample sizes. These confidence intervals are the correct length, since coverage is close to the nominal level. In anticipation of the empirical application, we repeat this exercise for the simulated data generating process with $\boldsymbol{X} \in \mathbb{R}^{722\times 30}$ and rank $r = 5$. Table 12 confirms that our procedure attains coverage close to the nominal level.

n	p	Noise	ATE	SE	CI.80	CI.95
50	200	20%	2.21	0.56	0.82	0.96
100	100	20%	2.21	0.35	0.82	0.95
200	50	20%	2.22	0.21	0.81	0.95
500	20	20%	2.23	0.12	0.78	0.94
1000	10	20%	2.23	0.08	0.76	0.93

n	p	Noise	ATE	SE	CI.80	CI.95
722	30	20%	2.26	0.12	0.78	0.92

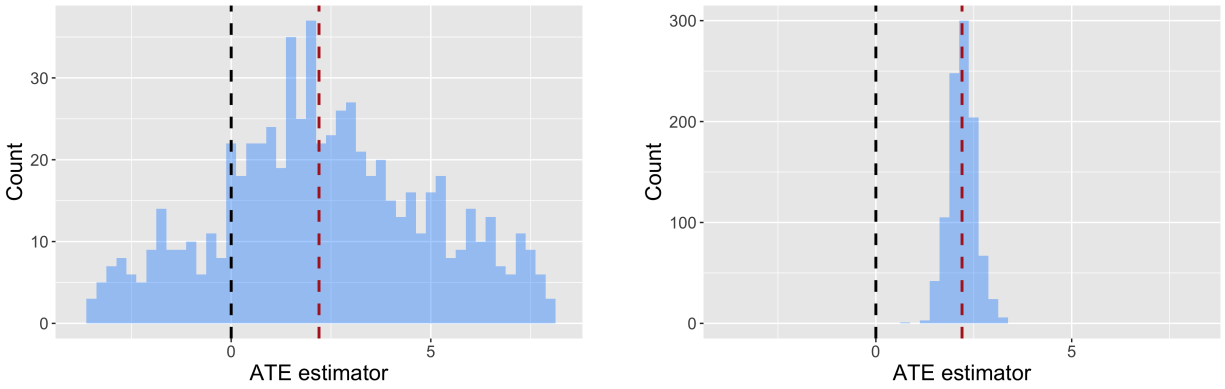
Figure 12: Our approach adapts to data shape

B.2 Can data corruption flip signs?

In the main text, we show that for the simulated data generating process with $\boldsymbol{X} \in \mathbb{R}^{100\times 100}$ and rank $r = 5$, OLS performs well with clean data and performs poorly with corrupted data. In particular, measurement error with a 20% noise-to-signal ratio is enough to pose a problem for OLS. We investigate two follow-up questions. First, can data corruption flip the sign of OLS estimates, i.e. can it lead to negative point estimates when the average treatment effect is $\theta_0 = 2.2 > 0$? Second, can can data corruption flip the sign of OLS and 2SLS estimates in scenarios more similar to our real world example?

As an answer to the first question, we find that data corruption can flip the sign of OLS estimates some of the time. Figure 13a visualizes the histogram of 1000 OLS point estimates from the data generating process with dimensions $\boldsymbol{X} \in \mathbb{R}^{100\times 100}$, rank $r = 5$, and 20% measurement error. Roughly one quarter of the OLS estimates have flipped signs. As Figure 13b shows, the histogram of 1000 point estimates using our procedure remain close to $\theta_0 = 2.2$ with no flipped signs.

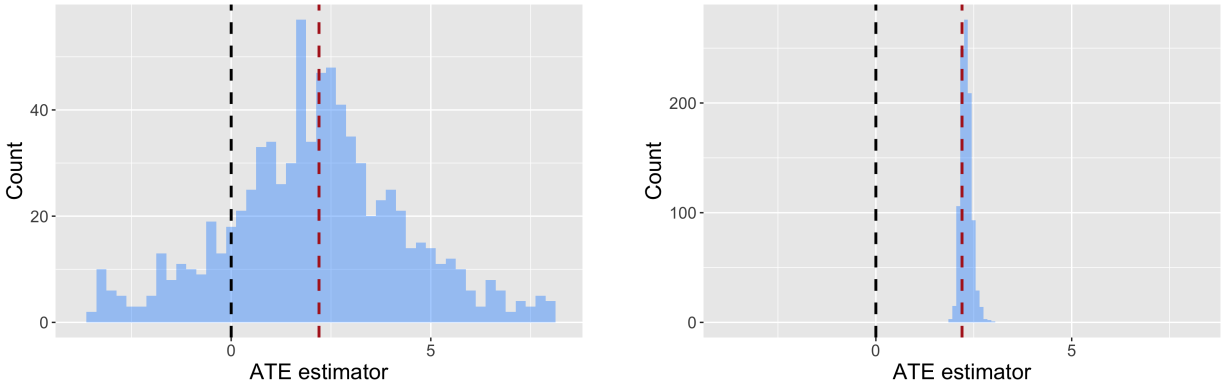
In order to answer the second question, we repeat this exercise for the simulated data generating process with $\boldsymbol{X} \in \mathbb{R}^{722\times 30}$ and rank $r = 5$. Flipping signs requires not only 20% measurement error but also 10% missingness. Figure 14 visualizes the histograms of 1000



(a) OLS (b) Our approach

Figure 13: Data corruption can flip signs: 100×100

point estimates, using OLS versus our procedure. A similar fraction of OLS estimates have flipped signs, while none of our estimates have flipped signs.

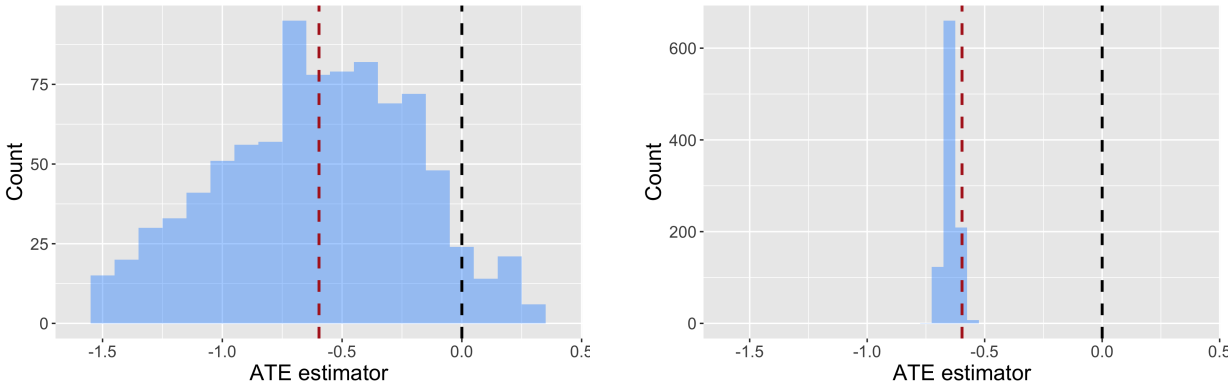


(a) OLS (b) Our approach

Figure 14: Data corruption can flip signs: 722×30

Finally, we conduct a semi-synthetic sign flipping exercise. We consider the Census covariates from Autor et al. (2013) at the commuting zone level. Rather than a synthetic average treatment effect, the estimand is the actual effect of import competition on manufacturing employment in a partially linear instrumental variable model. Flipping signs requires not only 20% measurement error but also 20% missingness. In this thought experiment, we take the reported effect from Autor et al. (2013) as the ground truth, we take the data set from Autor et al. (2013) as clean data, and we generate synthetic measurement error and missingness. Figure 15 visualizes the histograms of 1000 point estimates across 1000

draws of synthetic corruption, using 2SLS versus our procedure. A relatively small fraction of 2SLS estimates have flipped signs, while none of our estimates have flipped signs.



(a) 2SLS

(b) Our approach

Figure 15: Data corruption can flip signs: Census

We summarize the results of these various sign flipping exercises in Table 16. The three rows correspond to (i) synthetic data with $\boldsymbol{X} \in \mathbb{R}^{100\times100}$; (ii) synthetic data with $\boldsymbol{X} \in \mathbb{R}^{722\times30}$; and (iii) semi-synthetic data from Autor et al. (2013). We interpret the OLS and TSLS results as motivation for data cleaning before data analysis. Our procedure may be viewed as an extension of OLS and TSLS with a very simple type of data cleaning that we subsequently account for in our data cleaning-adjusted confidence intervals. An exciting direction for future work is to extend our results to richer types of data cleaning which more closely resemble empirical practice.

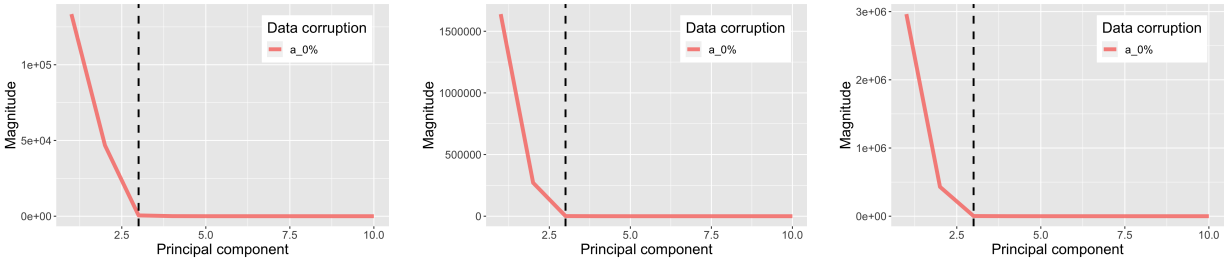
Data	Noise	Missingness	Sign flip
100 x 100	20%	0%	27%
722 x 30	20%	10%	22%
Census	20%	20%	9%

Figure 16: Data corruption can flip signs

B.3 The key assumption holds in many Census data sets

Our key assumption, which powers our entire analysis, is that the true covariates $\boldsymbol{X}$ are approximately low rank. In the main text, we visually confirm that this assumption holds in Census data by plotting the singular values of the covariates from Autor et al. (2013). In this subsection, we argue that this key assumption holds in several Census data sets

that are popular in economic research. We visualize the singular values of covariates from LaLonde (1986); Poterba et al. (1996); Chetty and Hendren (2018a).



(a) NSW (b) PSID (c) CPS

Figure 17: National Supported Work demonstration

We begin with data sets at the individual level. In LaLonde (1986), the author considers the problem of estimating the average treatment effect of the National Supported Work (NSW) demonstration, a randomized job training program. There are three data sets: one using actual NSW participants as the treated and control group; another using NSW participants as the treated group and a PSID sample as the comparison group, and yet another using NSW participants as the treated group and a CPS sample as the comparison group. In Figure 17, we visualize the singular values of covariates for these three data sets. Across data sets, the rank is approximately $r = 3$ while the ambient dimension of covariates is $p = 10$. In Poterba et al. (1996), the authors consider the problem of estimating the (local) average treatment effect of 401(k) participation on savings. Figure 18 visualizes the singular values of covariates, showing that the rank is approximately $r = 2$ while the ambient dimension of covariates is $p = 9$.

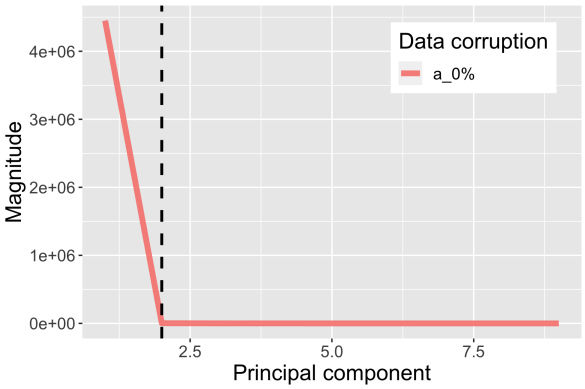
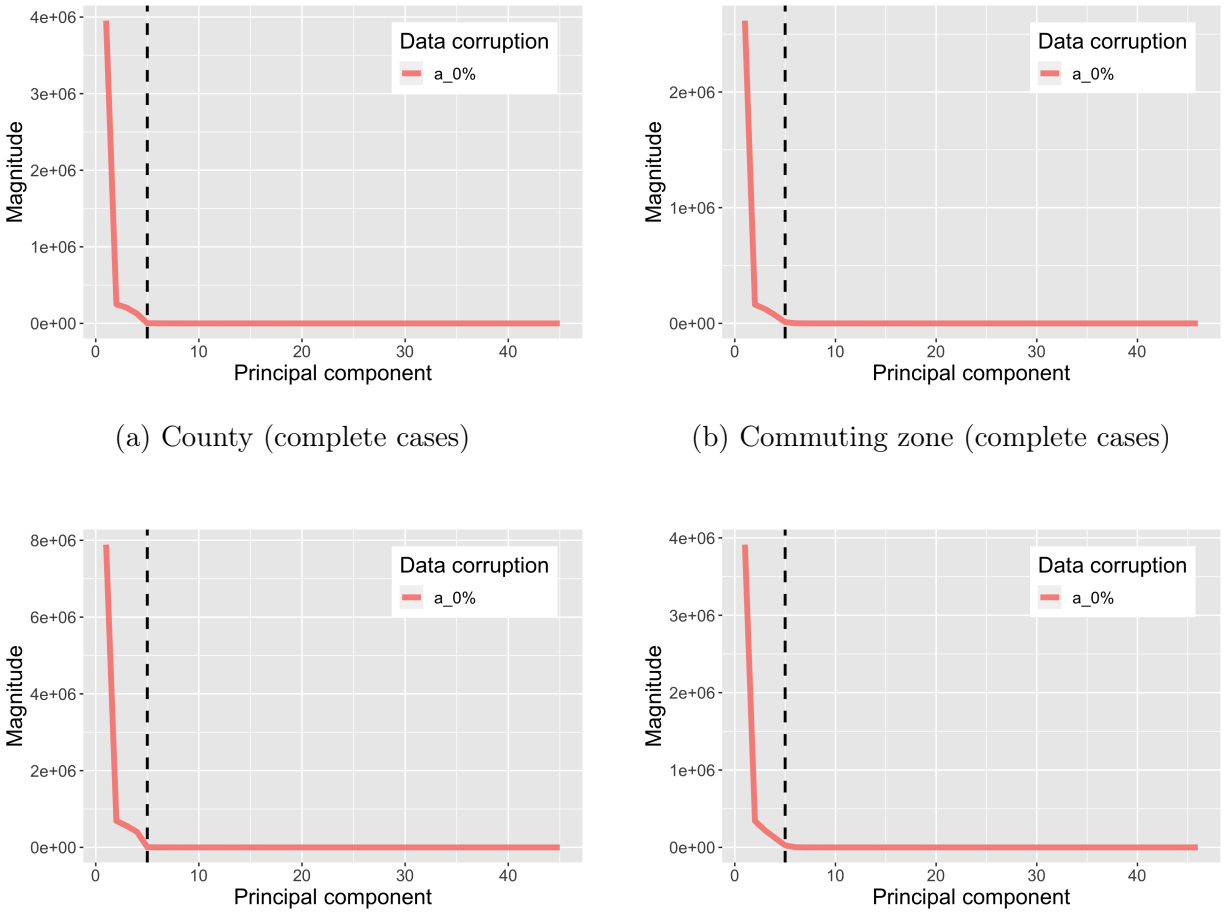


Figure 18: 401(k) participation

Next, we turn to data sets of aggregate units. In Chetty and Hendren (2018a), the authors consider commuting zones to be the aggregate units. In Chetty and Hendren (2018b), a companion paper, the authors consider counties to be the aggregate units. These

two data sets help to evaluate the robustness of the approximate low rank assumption across different levels of geography. In Figure 19, we show that the approximate rank is $r = 5$ while the ambient dimension of covariates is $p = 45$. Since the data set has missing values, we report the singular values using complete cases and using filled cases according to our filling procedure from the main text.



(a) County (complete cases)

(b) Commuting zone (complete cases)

(c) County (filled)

(d) Commuting zone (filled)

Figure 19: Opportunity insights

C Nonlinearity

In this appendix, we characterize the class of nonlinear dictionaries $b : \mathbb{R}^p \to \mathbb{R}^{p'}$ for which our main results go through. We delay proofs until the last subsection of this appendix. We discuss two classes of dictionaries.

C.1 Polynomial dictionary

We refer to the following three simple properties as data cleaning continuity, since they imply that the data cleaning results for original regressors imply similar data cleaning results for technical regressors. We state the properties then verify them for the polynomial dictionary of degree $d_{\max}$.

Assumption C.1 (Dictionary continuity)**.** *The dictionary satisfies three conditions:*

1. *For any two matrices* $\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)} \in \mathbb{R}^{n \times p}$, $\|b(\boldsymbol{M}^{(1)}) - b(\boldsymbol{M}^{(2)})\|_{2,\infty}^2 \leq C_b' \|\boldsymbol{M}^{(1)} - \boldsymbol{M}^{(2)}\|_{2,\infty}^2$;
2. *For any* $\boldsymbol{M} \in \mathbb{R}^{n \times p}$, $rank\{b(\boldsymbol{M})\} \leq \{rank(\boldsymbol{M})\}^{C_b''}$;
3. *For any* $v \in \mathbb{R}^p$, $\|b(v)\|_{\max} \leq (\|v\|_{\max})^{C_b'''}$.

For much of our argument to go through, it suffices that the dictionary exhibits three simple properties: clean original regressors should imply clean technical regressors; low rank original regressors should imply low rank technical regressors; and a bound on the maximum value of original regressors should imply a bound on the maximum value of technical regressors.

Definition C.1 (Polynomial dictionary)**.** *Let* $v = (v_1, v_2, \ldots, v_p) \in \mathbb{R}^p$. *Consider the dictionary* b^{POLY}, *where for* $k \in [p']$, $b_k^{\text{POLY}}(v) = \prod_{\ell=1}^{d(k)} v_\ell$ *with* $v_\ell \in \{v_1, \ldots, v_p\}$.

That is, each basis function $b_k^{\text{POLY}}(v)$ in the dictionary is a polynomial of degree $d(k) \leq d_{\max}$ constructed from coordinates of v, allowing for repeats. This class of dictionaries is commonly used in empirical economic research. It nests as a special case the interacted dictionary studied in the main text, which permits a rich model of heterogeneous treatment effects. Pleasingly, for this class, the dictionary constants (C_b', C_b'', C_b''') have no dependence on p', the number of elements in the dictionary. Rather, (C_b', C_b'', C_b''') only depend on the maximum degree $d_{\max}$ of the polynomial dictionary.

Proposition C.1 (Verifying Assumption C.1)**.** *The polynomial dictionary* b^{POLY} *of degree* $d_{\max}$ *satisfies Assumption C.1 with the following constants*

1. $C_b' \leq 2^{d_{\max}} \cdot \|\boldsymbol{M}^{(1)}\|_{\max}^{2d_{\max}} \cdot \|\boldsymbol{M}^{(2)}\|_{\max}^{2d_{\max}}$;

2. $C_b''' \leq d_{\max}$;

3. $C_b'''' \leq d_{\max}$.

Remarkably, this class of dictionaries also preserves the low rank approximation in the following sense.

Proposition C.2 (Low rank approximation is preserved)**.** *Suppose Assumption 5.1 holds and the true covariates have the low rank approximation* $\boldsymbol{X} = \boldsymbol{X}^{(\text{LR})} + \boldsymbol{E}^{(\text{LR})}$ *where* $r = rank\{\boldsymbol{X}^{(\text{LR})}\}$ *and* $\Delta_E = \|\boldsymbol{E}^{(\text{LR})}\|_{\max}$*. Consider the polynomial dictionary* b^{POLY} *of degree* $d_{\max}$*. Then* $r' := rank\{b(\boldsymbol{X}^{(\text{LR})})\} \leq r^{d_{\max}}$ *and* $\Delta_E' := \|b(\boldsymbol{X}) - b(\boldsymbol{X}^{(\text{LR})})\|_{\max} \leq C\bar{A}^{d_{\max}} \cdot d_{\max}\Delta_E$*.*

The same logic applies for dictionaries applied to $(D_i, X_{i,\cdot})$ rather than $X_{i,\cdot}$.

C.2 Polynomial dictionary with uncorrupted nonlinearity

Assumption C.1 suffices to generalize our data cleaning results. For analysis of the error-in-variables estimators, we impose a further assumption, which constrains which kinds of terms can appear as technical regressors. Consider the polynomial dictionary of degree $d_{\max}$, where the only source of nonlinearity is powers and interactions with regressors known to be uncorrupted.

Definition C.2 (Polynomial dictionary with uncorrupted nonlinearity)**.** *Suppose the observed regressors consist of one uncorrupted regressor* D_i *and several corrupted regressors* $X_{i,\cdot}$*. Consider a polynomial dictionary* b^{POLY} *of degree* $d_{\max}$ *such that each basis function* b_k^{POLY} *is at most linear in the corrupted regressors. By definition,* $p' \leq C \cdot d_{\max}p$*.*

For example, in Example A.1 where D_i is uncorrupted, the interacted dictionary $b : (D_i, X_{i,\cdot}) \mapsto \{D_iX_{i,\cdot}, (1 - D_i)X_{i,\cdot}\}$ satisfies this property. In Example A.4 where D_i is uncorrupted, the nonlinear dictionary $b : (D_i, X_{i,\cdot}) \mapsto (1, D_i, X_{i,\cdot}, D_iX_{i,\cdot}, D_i^2)$ satisfies this property as well since it contains D_i^2 but does not contain X_{ij}^2. Intuitively, this family of dictionaries avoids compounding measurement error because the corrupted regressors are not multiplied with each other. For readability, we focus on the case of one uncorrupted regressor, which can be conceptualized as

$$b : (D_i, X_{i,\cdot}) \mapsto (1, D_i, ..., D_i^{d_{\max}}, X_{i,\cdot}, D_iX_{i,\cdot}, ..., D_i^{d_{\max}-1}X_{i,\cdot}) \tag{6}$$

where D_i is uncorrupted and $X_{i,\cdot}$ are uncorrupted. Definition C.2 naturally generalizes to the case of multiple uncorrupted regressors. We require three properties to hold after the dictionary is applied to the data.

Assumption C.2 (Dictionary is non-collapsing)**.** *The dictionary does not collapse in the following sense.*

1. *Recall that we set* $k := rank(\hat{\boldsymbol{X}})$ *equal to* $r := rank\{\boldsymbol{X}^{(\text{LR})}\}$*. We further assume* $k' := rank\{b(D, \hat{\boldsymbol{X}})\}$ *is equal to* $r' := rank[b\{D, \boldsymbol{X}^{(\text{LR})}\}]$*.*

2. *Assumption 5.4 posits that the smallest singular value of* $\boldsymbol{X}^{(\text{LR})}$ *is* $s_r \geq C\sqrt{\frac{np}{r}}$*. We further posit that the smallest singular value of* $b\{D, \boldsymbol{X}^{(\text{LR})}\}$ *is* $s'_{r'} \geq C\sqrt{\frac{np}{r'}}$*.*

3. *Using the notation of* (6)*, the technical regressors* $(1, D_i, ...D_i^{d_{\max}})$ *are full rank.*

The first property in Assumption C.2 ensures two matrices of equal rank get mapped to two new matrices of equal rank. The second property imposes that singular values, after dictionary mapping, remain well balanced. In particular, we allow for a weaker signal to noise ratio for technical regressors since $r' \geq r$. We do *not* impose $s'_{r'} \geq C\sqrt{\frac{np'}{r'}}$, which is a stronger and less plausible requirement since it implies that the signal to noise ratio increases with the dictionary dimension p'. The third property is a technical assumption which allows the theory of implicit data cleaning to generalize.

C.3 Proofs

We prove Proposition C.1 via the following sequence of lemmas.

Lemma C.1 (C'_b)**.** *For* b^{POLY}*,* $C'_b \leq 2^{d_{\max}} \cdot \|\boldsymbol{M}^{(1)}\|_{\max}^{2d_{\max}} \cdot \|\boldsymbol{M}^{(2)}\|_{\max}^{2d_{\max}}$*.*

Proof. We introduce the notation $[b^{\text{POLY}}(\boldsymbol{M})]_{ik} = \prod_{\{j(k)\}} M_{ij(k)}$, where $j(k) \in [p]$, $M_{ij(k)} \in \{M_{i1}, \ldots, M_{ip}\}$, and $|\{j(k)\}| = d(k)$. We will simplify notation in the following way. Fix k. Let $M_{i\ell}$ refer to the ℓ-th element of the product, where $\ell \in [d(k)]$. Therefore

$$[b^{\text{POLY}}(\boldsymbol{M})]_{ik} = \prod_{\{j(k)\}} M_{ij(k)} = \prod_{\ell=1}^{d(k)} M_{i\ell}.$$

Then for any column $k \in [p']$,

$$\begin{aligned}
\|b(\boldsymbol{M}^{(1)})_{\cdot,k} - b(\boldsymbol{M}^{(2)})_{\cdot,k}\|_2^2 &= \sum_{i=1}^{n} \left(\prod_{\ell=1}^{d(k)} M_{i\ell}^{(1)} - \prod_{\ell=1}^{d(k)} M_{i\ell}^{(2)} \right)^2 \\
&= \sum_{i=1}^{n} \left(\prod_{\ell=1}^{d(k)} M_{i\ell}^{(1)} - \prod_{\ell=1}^{d(k)} M_{i\ell}^{(2)} \pm M_{i1}^{(2)} \prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right)^2 \\
&\leq 2 \sum_{i=1}^{n} \left(\prod_{\ell=1}^{d(k)} M_{i\ell}^{(1)} - M_{i1}^{(2)} \prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right)^2 \\
&\quad + 2 \sum_{i=1}^{n} \left(\prod_{\ell=1}^{d(k)} M_{i\ell}^{(2)} - M_{i1}^{(2)} \prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right)^2 .
\end{aligned}$$

Looking at the first term on the RHS above,

$$\begin{aligned}
\sum_{i=1}^{n} \left(\prod_{\ell=1}^{d(k)} M_{i\ell}^{(1)} - M_{i1}^{(2)} \prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right)^2 &= \sum_{i=1}^{n} \left(M_{i1}^{(1)} - M_{i1}^{(2)} \right)^2 \left(\prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right)^2 \\
&\leq \|\boldsymbol{M}^{(1)}\|_{\max}^{2d_{\max}} \sum_{i=1}^{n} \left(M_{i1}^{(1)} - M_{i1}^{(2)} \right)^2 \\
&\leq \|\boldsymbol{M}^{(1)}\|_{\max}^{2d_{\max}} \|\boldsymbol{M}^{(1)} - \boldsymbol{M}^{(2)}\|_{2,\infty}^2 .
\end{aligned}$$

Now looking at the second term on the RHS,

$$\begin{aligned}
\sum_{i=1}^{n} \left(\prod_{\ell=1}^{d(k)} M_{i\ell}^{(2)} - M_{i1}^{(2)} \prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right)^2 &= \sum_{i=1}^{n} \left(M_{i1}^{(2)} \left(\prod_{\ell=2}^{d(k)} M_{i\ell}^{(2)} - \prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right) \right)^2 \\
&\leq \|\boldsymbol{M}^{(2)}\|_{\max}^2 \sum_{i=1}^{n} \left(\prod_{\ell=2}^{d(k)} M_{i\ell}^{(2)} - \prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right)^2 .
\end{aligned}$$

Continuing forward with $\sum_{i=1}^{n} \left(\prod_{\ell=2}^{d(k)} M_{i\ell}^{(2)} - \prod_{\ell=2}^{d(k)} M_{i\ell}^{(1)} \right)^2$ in a recursive manner leads to the following bound for all $k \in [p']$:

$$\begin{aligned}
&\|b(\boldsymbol{M}^{(1)})_{\cdot,k} - b(\boldsymbol{M}^{(2)})_{\cdot,k}\|_2^2 \\
&\leq \|\boldsymbol{M}^{(1)} - \boldsymbol{M}^{(2)}\|_{2,\infty}^2 \cdot \left(2^{d_{\max}} \cdot \|\boldsymbol{M}^{(1)}\|_{\max}^{2d_{\max}} \cdot \|\boldsymbol{M}^{(2)}\|_{\max}^{2d_{\max}} \right).
\end{aligned}$$

We thus have the desired result. □

Lemma C.2 (C_b''). *For b^{POLY}, $C_b'' \leq d_{\max}$.*

Proof. Fix $\boldsymbol{M} \in \mathbb{R}^{n \times p}$ with rank r. For notational simplicity, let $M_{i\ell}$ refer to the ℓ-th element of the product in $[b^{\text{POLY}}(\boldsymbol{M})]_{ik}$. Observe that $b^{\text{POLY}}(\boldsymbol{M})$ can be equivalently represented as

$$b^{\text{POLY}}(\boldsymbol{M}) = \boldsymbol{B}^{(1)} \odot, ..., \odot \boldsymbol{B}^{(d_{\max})},$$

where $\odot$ means Hadamard product, $\boldsymbol{B}^{(\ell)} \in \mathbb{R}^{n \times p'}$, and for $\ell \in [d_{\max}], i \in [n], k \in [p']$

$$[\boldsymbol{B}^{(\ell)}]_{ik} = \begin{cases} M_{i\ell} & \text{if } \ell \leq d(k) \\ 1 & \text{if } \ell > d(k) \end{cases}.$$

Since each column of each $\boldsymbol{B}^{(\ell)}$ is either a column of $\boldsymbol{M}$ or a column of ones, it has rank at most r. Finally recall that the rank of a Hadamard product is bounded by the product of ranks and so

$$rank\{b^{\text{POLY}}(\boldsymbol{M})\} \leq \prod_{\ell=1}^{d_{\max}} r = r^{d_{\max}}.$$

□

Lemma C.3 (C_b''')**.** *For b^{POLY}, $C_b''' \leq d_{\max}$.*

Proof. Denote $v \in \mathbb{R}^p$ with $\|v\|_\infty \leq \bar{A}$. Note that each basis function is of the form

$$b_k^{\text{POLY}}(v) = \prod_{\ell=1}^{d(k)} v_\ell \leq \prod_{\ell=1}^{d(k)} \bar{A} = \bar{A}^{d(k)} \leq \bar{A}^{d_{\max}}.$$

□

Proof of Proposition C.1. Immediate from Lemmas C.1, C.2, and C.3. □

To begin, we state an important observation about the signal approximation $\boldsymbol{X}^{(\text{LR})}$ that will simplify our subsequent analysis.

Lemma C.4 (Bounded signal approximation)**.** *Suppose Assumption 5.1 holds. Without loss of generality, $\|\boldsymbol{X}^{(\text{LR})}\|_{\max} \leq 3\bar{A}$.*

Since $\|\boldsymbol{X}\|_{\max} \leq \bar{A}$ and $\|\boldsymbol{X}^{(\text{LR})}\|_{\max} \leq 3\bar{A}$, the same constant $C\bar{A}$ handles both objects.

Proof of Lemma C.4. Suppose we have access to some $\boldsymbol{X}^{(\text{LR})}$ with rank r such that $\|\boldsymbol{X}^{(\text{LR})}\|_{\max} > 3\bar{A}$. By reverse triangle inequality

$$\Delta_{E,\boldsymbol{X}^{(\text{LR})}} = \|\boldsymbol{X}^{(\text{LR})} - \boldsymbol{X}\|_{\max} \geq \|\boldsymbol{X}^{(\text{LR})}\|_{\max} - \|\boldsymbol{X}\|_{\max} > 2\bar{A}.$$

We construct a $\boldsymbol{B}^{(\text{LR})}$ with rank r such that $\|\boldsymbol{B}^{(\text{LR})}\|_{\max} \leq 3\bar{A}$ and $\Delta_{E,\boldsymbol{B}^{(\text{LR})}} < \Delta_{E,\boldsymbol{X}^{(\text{LR})}}$. Set

$$\boldsymbol{B}^{(\text{LR})} = \frac{\bar{A}}{\|\boldsymbol{X}^{(\text{LR})}\|_{\max}} \cdot \boldsymbol{X}^{(\text{LR})}.$$

Since $\boldsymbol{B}^{(\text{LR})}$ simply scales $\boldsymbol{X}^{(\text{LR})}$ by a constant, it has the same rank. By construction $\|\boldsymbol{B}^{(\text{LR})}\|_{\max} \leq \bar{A}$. Finally

$$\Delta_{E,\boldsymbol{B}^{(\text{LR})}} = \|\boldsymbol{B}^{(\text{LR})} - \boldsymbol{X}\|_{\max} \leq \|\boldsymbol{B}^{(\text{LR})}\|_{\max} + \|\boldsymbol{X}\|_{\max} \leq 2\bar{A}.$$

□

Proof of Proposition C.2. By definition, $r = rank\{\boldsymbol{X}^{(\text{LR})}\}$. The first result follows directly from Proposition C.1. To see the second result, consider the case where $d_{\max} = 2$. Then any higher order entry of $b(\boldsymbol{X}) - b(\boldsymbol{X}^{(\text{LR})})$ is of the form

$$\begin{aligned} |X_{ij}X_{ik} - X_{ij}^{(\text{LR})}X_{ik}^{(\text{LR})}| &\leq |X_{ij}X_{ik} - X_{ij}^{(\text{LR})}X_{ik}| + |X_{ij}^{(\text{LR})}X_{ik} - X_{ij}^{(\text{LR})}X_{ik}^{(\text{LR})}| \\ &\leq \bar{A}\Delta_E + 3\bar{A}\Delta_E \end{aligned}$$

where the final inequality appeals to Lemma C.4. More generally, there are $d_{\max}$ such terms, and the largest is of the form $(3\bar{A})^{d_{\max}}\Delta_E$. □

D Data cleaning

D.1 Notation and preliminaries

In this appendix, we replace the symbol $X_{i,\cdot}$ with the symbol $A_{i,\cdot}$, so that

$$Z_{i,\cdot} = (A_{i,\cdot} + H_{i,\cdot}) \odot \pi_{i,\cdot}.$$

We suppress indexing by the folds (TRAIN, TEST) to lighten notation. As in Assumption 5.3, we identify `NA` with 0 in $\boldsymbol{Z}$ for the remainder of the appendix. We slightly abuse notation by letting n be the number of observations in TRAIN, departing from the notation of the main text. The entire section is conditional on $\boldsymbol{A}$, which we omit to lighten notation. We write $\|\cdot\| = \|\cdot\|_{op}$, and abbreviate law of iterated expectations (LIE). We denote by C an absolute constant.

Recall $\boldsymbol{A} = \boldsymbol{A}^{(\text{LR})} + \boldsymbol{E}^{(\text{LR})}$ and $r = rank\{\boldsymbol{A}^{(\text{LR})}\}$. We denote the SVDs

$$\boldsymbol{A}^{(\text{LR})} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \quad \hat{\boldsymbol{A}} = \hat{\boldsymbol{U}}_k\hat{\boldsymbol{\Sigma}}_k\hat{\boldsymbol{V}}_k^T, \quad \boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^T.$$

The first k left singular vectors of $\boldsymbol{A}^{(\text{LR})}$ are $\boldsymbol{U}_k$. We denote $s_k = \Sigma_{kk}$ and $\hat{s}_k = \hat{\Sigma}_{kk}$. Recall that $\delta = \frac{1}{1-\sqrt{\frac{22\ln(np)}{n\rho_{\min}}}}$.

Define the unit ball $\mathbb{B}^p$ and unit sphere $\mathbb{S}^{p-1}$ by

$$\mathbb{B}^p = \{v \in \mathbb{R}^p : \|v\|_2 \leq 1\}$$
$$\mathbb{S}^{p-1} = \{v \in \mathbb{R}^p : \|v\|_2 = 1\}$$

Recall that $\hat{\boldsymbol{A}}^{\text{TRAIN}}$ is constructed by taking TRAIN covariates then filling and cleaning them using TRAIN alone. In addition to studying $\hat{\boldsymbol{A}}^{\text{TRAIN}}$, we introduce and study the object $\hat{\boldsymbol{A}}^{\text{TEST}}$, is constructed by taking TEST covariates, filling them using TRAIN, and cleaning them using TEST. It turns out that the analysis does not depend on whether $\hat{\boldsymbol{\rho}}^{\text{TRAIN}}$ or $\hat{\boldsymbol{\rho}}^{\text{TEST}}$ is used when filling in missing values. Rather than writing two nearly identical arguments, we write out one unified argument with formal remarks.

Proof of Proposition 4.1. By LIE,

$$\begin{aligned}
&\mathbb{E}[\text{FILL}(Z_{ij}^{\text{TEST}})|A_{ij}^{\text{TEST}}, \text{TRAIN}] \\
&= \mathbb{E}[\text{FILL}(Z_{ij}^{\text{TEST}})|A_{ij}^{\text{TEST}}, \pi_{ij}^{\text{TEST}} = 1, \text{TRAIN}]\mathbb{P}(\pi_{ij}^{\text{TEST}} = 1|A_{ij}^{\text{TEST}}, \text{TRAIN}) \\
&\quad + \mathbb{E}[\text{FILL}(Z_{ij}^{\text{TEST}})|A_{ij}^{\text{TEST}}, \pi_{ij}^{\text{TEST}} = 0, \text{TRAIN}]\mathbb{P}(\pi_{ij}^{\text{TEST}} = 0|A_{ij}^{\text{TEST}}, \text{TRAIN}) \\
&= \mathbb{E}[Z_{ij}^{\text{TEST}}/\hat{\rho}_j|A_{ij}^{\text{TEST}}, \pi_{ij}^{\text{TEST}} = 1, \text{TRAIN}] \cdot \rho_j + \mathbb{E}[0|A_{ij}^{\text{TEST}}, \pi_{ij}^{\text{TEST}} = 1, \text{TRAIN}] \cdot (1-\rho_j) \\
&= A_{ij}^{\text{TEST}}\frac{\rho_j}{\hat{\rho}_j}.
\end{aligned}$$

Likewise

$$\begin{aligned}
&\mathbb{E}[\text{FILL-AS-MEANS}(Z_{ij}^{\text{TEST}})|A_{ij}^{\text{TEST}}, \text{TRAIN}] \\
&= \mathbb{E}[\text{FILL-AS-MEANS}(Z_{ij}^{\text{TEST}})|A_{ij}^{\text{TEST}}, \pi_{ij}^{\text{TEST}} = 1, \text{TRAIN}]\mathbb{P}(\pi_{ij}^{\text{TEST}} = 1|A_{ij}^{\text{TEST}}, \text{TRAIN}) \\
&\quad + \mathbb{E}[\text{FILL-AS-MEANS}(Z_{ij}^{\text{TEST}})|A_{ij}^{\text{TEST}}, \pi_{ij}^{\text{TEST}} = 0, \text{TRAIN}]\mathbb{P}(\pi_{ij}^{\text{TEST}} = 0|A_{ij}^{\text{TEST}}, \text{TRAIN}) \\
&= \mathbb{E}[Z_{ij}^{\text{TEST}}|A_{ij}^{\text{TEST}}, \pi_{ij}^{\text{TEST}} = 1, \text{TRAIN}] \cdot \rho_j + \mathbb{E}[\bar{Z}_j^{\text{TRAIN}}|A_{ij}^{\text{TEST}}, \pi_{ij}^{\text{TEST}} = 0, \text{TRAIN}] \cdot (1-\rho_j) \\
&= A_{ij}^{\text{TEST}} \cdot \rho_j + \bar{Z}_j^{\text{TRAIN}}(1-\rho_j).
\end{aligned}$$

□

Proposition D.1 (Bound on $\|\hat{\boldsymbol{A}}\|_{\max}$)**.** *Suppose $k = r$ and the corrupted singular values $\hat{s}_1, ..., \hat{s}_r \le C\sqrt{\frac{np}{r}}$. Assume the following incoherence conditions for the corrupted singular vectors: $\|\hat{\boldsymbol{U}}_r\|_{\max} \le Cn^{-1/2}$ and $\|\hat{\boldsymbol{V}}_r\|_{\max} \le Cp^{-1/2}$. Then $\|\hat{\boldsymbol{A}}\|_{\max} \le Cr^{1/2}$.*

The condition $\hat{s}_1, ..., \hat{s}_r \le C\sqrt{\frac{np}{r}}$ can be proven with high probability from the condition $s_1, ..., s_r \le C\sqrt{\frac{np}{r}}$ using Weyl's inequality using an argument similar to Lemma E.9. The condition $s_1, ..., s_r \le C\sqrt{\frac{np}{r}}$ complements Assumption 5.4. The incoherence conditions can be interpreted by recognizing that each left singular vector $U_{\cdot,j} \in \mathbb{R}^n$ and each right singular vector $V_{\cdot,j} \in \mathbb{R}^p$.

Proof. Write

$$\hat{A}_{ij} = \sum_{\ell=1}^{r} \hat{U}_{i\ell}\hat{s}_\ell\hat{V}_{j\ell}.$$

Hence

$$|\hat{A}_{ij}| \le \sum_{\ell=1}^{r} |\hat{U}_{i\ell}| \cdot |\hat{s}_\ell| \cdot |\hat{V}_{j\ell}| \le r \cdot Cn^{-1/2} \cdot C\sqrt{\frac{np}{r}} \cdot Cp^{-1/2} = Cr^{1/2}.$$

□

D.2 High probability events

Define the following beneficial events. We will show each event holds with probability $1 - \frac{2}{n^{10}p^{10}}$.

$$\mathcal{E}_1 = \left\{ \|\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{\rho}\| \le (\sqrt{n} + \sqrt{p})\Delta_{H,op} \right\}, \quad \Delta_{H,op} = C\bar{A}(\kappa + K_a + \bar{K}) \ln^{\frac{3}{2}}(np);$$

$$\mathcal{E}_2 = \left\{ \max_{j\in[p]} \|Z_{\cdot,j} - \rho_j A_{\cdot,j}\|_2^2 \le n\Delta_H \right\}, \quad \Delta_H = C(K_a + \bar{A}\bar{K})^2 \ln^2(np);$$

$$\mathcal{E}_3 = \left\{ \max_{j\in[p]} \|\boldsymbol{U}_k\boldsymbol{U}_k^T(Z_{\cdot,j} - \rho_j A_{\cdot,j})\|_2^2 \le k\Delta_H \right\};$$

$$\mathcal{E}_4 = \left\{ \forall j \in [p],\ \frac{1}{\delta}\rho_j \le \hat{\rho}_j \le \delta\rho_j \right\}, \quad \delta = \frac{1}{1 - \sqrt{\frac{22\ln(np)}{n\rho_{\min}}}};$$

$$\mathcal{E}_5 = \left\{ \max_{j\in[p]} |\hat{\rho}_j - \rho_j| \le C\sqrt{\frac{\ln(np)}{n}} \right\}.$$

Analyzing $\mathcal{E}_1$

Lemma D.1. *Under Assumption 5.3,*

$$\|\mathbb{E}[(\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{\rho})^T(\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{\rho})]\| \le \rho_{\max}(1-\rho_{\min})\left(\max_{j\in[p]}\|A_{\cdot,j}\|_2^2 + \|diag(\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}])\|\right) + \rho_{\max}\|\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}]\|,$$

where $\rho_{\max} \coloneqq \max_{j\in[p]}\rho_j \le 1$.

Proof. To begin, observe that

$$\mathbb{E}[(\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{\rho})^T(\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{\rho})] = \sum_{\ell=1}^{n}\mathbb{E}[(Z_{\ell,\cdot}-A_{\ell,\cdot}\boldsymbol{\rho})\otimes(Z_{\ell,\cdot}-A_{\ell,\cdot}\boldsymbol{\rho})].$$

Let $\boldsymbol{X}=\boldsymbol{A}+\boldsymbol{H}$. We highlight the following relations: for any $(\ell,j)\in[n]\times[p]$,

$$\mathbb{E}[Z_{\ell j}] = \rho_j A_{\ell j}$$
$$\mathbb{E}[Z_{\ell j}^2] = \rho_j\mathbb{E}[X_{\ell j}^2].$$

Now, let us fix a row $\ell\in[n]$ and denote

$$\boldsymbol{W}^{(\ell)} = (Z_{\ell,\cdot}-A_{\ell,\cdot}\boldsymbol{\rho})\otimes(Z_{\ell,\cdot}-A_{\ell,\cdot}\boldsymbol{\rho}).$$

Using the linearity of expectations, the expected value of the (i,j)-th entry of $\boldsymbol{W}^{(\ell)}$ can be written as

$$\mathbb{E}[W_{ij}^{(\ell)}] = \mathbb{E}[Z_{\ell i}Z_{\ell j}] - \rho_j\mathbb{E}[Z_{\ell i}]A_{\ell j} - \rho_i\mathbb{E}[Z_{\ell j}]A_{\ell i} + \rho_i\rho_j A_{\ell i}A_{\ell j}.$$

Suppose $i=j$, then

$$\mathbb{E}[W_{ii}^{(\ell)}] = \rho_i\mathbb{E}[X_{\ell i}^2] - \rho_i^2A_{\ell i}^2 = \rho_i(1-\rho_i)\mathbb{E}[X_{\ell i}^2] + \rho_i^2\mathbb{E}[(X_{\ell i}-A_{\ell i})^2]. \tag{7}$$

On the other hand, if $i\neq j$,

$$\mathbb{E}[W_{ij}^{(\ell)}] \le \sqrt{\rho_i\rho_j}\mathbb{E}[(X_{\ell i}-A_{\ell i})(X_{\ell j}-A_{\ell j})]. \tag{8}$$

since

$$\mathbb{E}[Z_{\ell i}Z_{\ell j}] = \mathbb{E}[\pi_{i\ell}\pi_{\ell j}]\mathbb{E}[X_{\ell i}X_{\ell j}] \le \sqrt{\mathbb{E}[\pi_{\ell i}^2]}\sqrt{\mathbb{E}[\pi_{\ell j}^2]}\mathbb{E}[X_{\ell i}X_{\ell j}] = \sqrt{\rho_i\rho_j}\mathbb{E}[X_{\ell i}X_{\ell j}].$$

Therefore, we can bound $\boldsymbol{W}^{(\ell)}$ as the sum of two matrices where the diagonal components are generated from (7) and the off-diagonal components are generated from (8). That is,

$$\begin{aligned}\mathbb{E}[\boldsymbol{W}^{(\ell)}] &\leq \mathbb{E}\Big(\rho_{\max}(1-\rho_{\min})diag(X_{\ell,\cdot}\otimes X_{\ell,\cdot}) + \rho_{\max}^2 diag(H_{\ell,\cdot}\otimes H_{\ell,\cdot})\Big) \\ &\quad + \mathbb{E}\Big(\rho_{\max}(H_{\ell,\cdot}\otimes H_{\ell,\cdot}) - \rho_{\max}diag(H_{\ell,\cdot}\otimes H_{\ell,\cdot})\Big) \\ &\leq \rho_{\max}(1-\rho_{\min})\mathbb{E}[diag(X_{\ell,\cdot}\otimes X_{\ell,\cdot})] + \rho_{\max}\mathbb{E}[H_{\ell,\cdot}\otimes H_{\ell,\cdot}].\end{aligned}$$

Taking the sum over all rows $\ell \in [n]$ yields

$$\mathbb{E}[(\boldsymbol{Z}-\boldsymbol{A\rho})^T(\boldsymbol{Z}-\boldsymbol{A\rho})] \leq \rho_{\max}(1-\rho_{\min})diag(\mathbb{E}[\boldsymbol{X}^T\boldsymbol{X}]) + \rho_{\max}\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}]. \tag{9}$$

To complete the proof, we apply triangle inequality to (9) to obtain

$$\left\|\mathbb{E}[(\boldsymbol{Z}-\boldsymbol{A\rho})^T(\boldsymbol{Z}-\boldsymbol{A\rho})]\right\| \leq \rho_{\max}(1-\rho_{\min})\left\|diag(\mathbb{E}[\boldsymbol{X}^T\boldsymbol{X}])\right\| + \rho_{\max}\left\|\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}]\right\|.$$

Since $\boldsymbol{H}$ is zero mean, we have

$$\begin{aligned}\left\|diag(\mathbb{E}[\boldsymbol{X}^T\boldsymbol{X}])\right\| &= \left\|diag(\boldsymbol{A}^T\boldsymbol{A}) + diag(\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}])\right\| \\ &\leq \left\|diag(\boldsymbol{A}^T\boldsymbol{A})\right\| + \left\|diag(\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}])\right\|.\end{aligned}$$

Collecting terms completes the proof. □

Lemma D.2 (Lemma H.2 of Agarwal et al. (2021))**.** *Suppose that $X \in \mathbb{R}^n$ and $P \in \{0,1\}^n$ are random vectors. Then for any $a \geq 1$,*

$$\|X \odot P\|_{\psi_a} \leq \|X\|_{\psi_a}.$$

Lemma D.3. *Under Assumptions 5.1, 5.2, and 5.3*

$$\|Z_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\|_{\psi_a} \leq K_a + \bar{A}\bar{K}.$$

Proof. To begin, write

$$\begin{aligned}\|Z_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\|_{\psi_a} &= \|X_{i,\cdot}\odot\pi_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\|_{\psi_a} \\ &= \|X_{i,\cdot}\odot\pi_{i,\cdot} - A_{i,\cdot}\odot\pi_{i,\cdot} + A_{i,\cdot}\odot\pi_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\|_{\psi_a} \\ &\leq \|(X_{i,\cdot}-A_{i,\cdot})\odot\pi_{i,\cdot}\|_{\psi_a} + \|A_{i,\cdot}\odot\pi_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\|_{\psi_a}.\end{aligned}$$

Consider the first term. By Lemma D.2 and Assumption 5.2

$$\|(X_{i,\cdot} - A_{i,\cdot}) \odot \pi_{i,\cdot}\|_{\psi_a} \leq \|(X_{i,\cdot} - A_{i,\cdot})\|_{\psi_a} = \|H_{i,\cdot}\|_{\psi_a} \leq K_a.$$

Consider the second term. By the definition of $\|\cdot\|_{\psi_a}$ and Assumption 5.1

$$\begin{aligned}\|A_{i,\cdot} \odot \pi_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\|_{\psi_a} &= \sup_{u \in \mathbb{B}^p} \left\| \sum_{j=1}^{p} u_j A_{ij}(\pi_{ij} - \rho_j) \right\|_{\psi_a} \\ &= \bar{A} \sup_{u \in \mathbb{B}^p} \left\| \sum_{j=1}^{p} u_j \frac{A_{ij}}{\bar{A}}(\pi_{ij} - \rho_j) \right\|_{\psi_a}\end{aligned}$$

Define the vector with components $v_j = u_j \frac{A_{ij}}{\bar{A}}$. We prove $v \in \mathbb{B}^p$:

$$\|v\|_2^2 = \sum_{j=1}^{p} v_j^2 = \sum_{j=1}^{p} u_j^2 \frac{A_{ij}^2}{\bar{A}^2} \leq \sum_{j=1}^{p} u_j^2 = \|u\|_2^2 \leq 1.$$

Hence,

$$\begin{aligned}\sup_{u \in \mathbb{B}^p} \left\| \sum_{j=1}^{p} u_j \frac{A_{ij}}{\bar{A}}(\pi_{ij} - \rho_j) \right\|_{\psi_a} &\leq \sup_{v \in \mathbb{B}^p} \left\| \sum_{j=1}^{p} v_j(\pi_{ij} - \rho_j) \right\|_{\psi_a} \\ &= \|\pi_{i,\cdot} - (\rho_1, ..., \rho_p)\|_{\psi_a} \\ &\leq \bar{K}.\end{aligned}$$

The last line holds by Assumption 5.3. In summary,

$$\|A_{i,\cdot} \odot \pi_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\|_{\psi_a} \leq \bar{A}\bar{K}.$$

□

Lemma D.4 (Proposition H.1 of Agarwal et al. (2021))**.** *Let $\boldsymbol{W} \in \mathbb{R}^{n \times p}$ be a random matrix whose rows $\boldsymbol{W}_{i,\cdot}$ are independent ψ_a-random vectors for some $a \geq 1$. Then for any $\tau > 0$,*

$$\|\boldsymbol{W}\| \leq \left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|^{1/2} + \sqrt{(1+\tau)p} \max_{i \in [n]} \|\boldsymbol{W}_{i,\cdot}\|_{\psi_a} \left\{1 + \left(2 + \tau\right)\ln(np)\right\}^{\frac{1}{a}} \sqrt{\ln(np)}$$

with probability at least $1 - \frac{2}{n^{1+\tau}p^\tau}$.

Lemma D.5. *Suppose Assumptions 5.1, 5.2, and 5.3 hold. Then $\forall \tau > 0$*

$$\begin{aligned}\|\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{\rho}\| \leq\ & C\sqrt{n}\left(\bar{A} + \kappa + K_a\right) \\ & + \sqrt{1+\tau}\sqrt{p}(K_a + \bar{A}\bar{K})\left\{1 + (2+\tau)\ln(np)\right\}^{\frac{1}{a}} \sqrt{\ln(np)}\end{aligned}$$

w.p. at least $1 - \frac{2}{n^{1+\tau}p^\tau}$.

Proof. Appealing to Lemma D.1

$$\begin{aligned}
&\|\mathbb{E}[(\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{\rho})^T(\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{\rho})]\| \\
&\leq \rho_{\max}(1-\rho_{\min})\left(\max_{j\in[p]}\|A_{\cdot,j}\|_2^2 + \|diag(\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}])\|\right) + \rho_{\max}\|\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}]\| \\
&\leq \max_{j\in[p]}\|A_{\cdot,j}\|_2^2 + \|diag(\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}])\| + \|\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}]\|.
\end{aligned}$$

Analyzing each term

$$\begin{aligned}
\max_{j\in[p]}\|A_{\cdot,j}\|_2^2 &\leq n\bar{A}^2; \\
\|diag(\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}])\| &= \left\|diag\left(\sum_{i\in[n]}\mathbb{E}[H_{i,\cdot}^T H_{i,\cdot}]\right)\right\| \\
&\leq n\max_{i\in[n]}\|diag(\mathbb{E}[H_{i,\cdot}^T H_{i,\cdot}])\| \\
&= n\max_{i\in[n],j\in[p]}|\mathbb{E}[H_{ij}^2]| \\
&\leq nCK_a; \\
\|\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}]\| &= \left\|\sum_{i\in[n]}\mathbb{E}[H_{i,\cdot}^T H_{i,\cdot}]\right\| \\
&\leq n\max_{i\in[n]}\|\mathbb{E}[H_{i,\cdot}^T H_{i,\cdot}]\| \\
&\leq n\kappa^2.
\end{aligned}$$

The result follows by plugging in these results as well as Lemma D.3 into Lemma D.4. □

Proposition D.2 ($\mathcal{E}_1$)**.** *Under Assumptions 5.1, 5.2, and 5.3*

$$\mathbb{P}(\mathcal{E}_1^c) \leq \frac{2}{n^{11}p^{10}} < \frac{2}{n^{10}p^{10}}.$$

Proof. Immediate by Lemma D.5, setting $\tau = 10$ and simplifying the bound. □

Analyzing $\mathcal{E}_2$ and $\mathcal{E}_3$

Lemma D.6 (Lemma H.4 of Agarwal et al. (2021))**.** *Let $X_1, \ldots, X_n$ be independent random variables with mean zero. For $a \geq 1$,*

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_a} \leq C\left(\sum_{i=1}^n \|X_i\|_{\psi_a}^2\right)^{1/2}.$$

Lemma D.7. *Under Assumptions 5.1, 5.2, and 5.3*

$$\|Z_{\cdot,j} - \rho_j A_{\cdot,j}\|_{\psi_a} \le C(K_a + \bar{A}\bar{K}).$$

Proof. Observe that

$$\begin{aligned}
\|Z_{\cdot,j} - \rho_j A_{\cdot,j}\|_{\psi_a} &= \sup_{u\in\mathbb{S}^{n-1}} \left\|u^T\big(Z_{\cdot,j} - \rho_j A_{\cdot,j}\big)\right\|_{\psi_a} \\
&= \sup_{u\in\mathbb{S}^{n-1}} \left\|u^T\big(\boldsymbol{Z} - \boldsymbol{A\rho}\big)e_j\right\|_{\psi_a} \\
&= \sup_{u\in\mathbb{S}^{n-1}} \left\|\sum_{i=1}^n u_i\big(Z_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\big)e_j\right\|_{\psi_a} \\
&\overset{(a)}{\le} C \sup_{u\in\mathbb{S}^{n-1}} \left(\sum_{i=1}^n u_i^2 \left\|\big(Z_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho}\big)e_j\right\|_{\psi_a}^2\right)^{1/2} \\
&\le C \max_{i\in[n]} \|(Z_{i,\cdot} - A_{i,\cdot}\boldsymbol{\rho})e_j\|_{\psi_a},
\end{aligned}$$

where (a) follows from Lemma D.6. Then the conclusion follows from Lemmas D.2 and D.3. □

Lemma D.8 (Lemma I.7 of Agarwal et al. (2021))**.** *Let $W_1, \dots, W_n$ be a sequence of ψ_a-random variables for some $a \ge 1$. For any $t \ge 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n W_i^2 > t\right) \le 2\sum_{i=1}^n \exp\left\{-\left(\frac{t}{n\|W_i\|_{\psi_a}^2}\right)^{a/2}\right\}.$$

Proposition D.3 ($\mathcal{E}_2$)**.** *Under Assumptions 5.1, 5.2, and 5.3*

$$\mathbb{P}(\mathcal{E}_2^c) \le \frac{2}{n^{10}p^{10}}$$

Proof. Fix j. Write

$$\|Z_{\cdot,j} - \rho_j A_{\cdot,j}\|_2^2 = \sum_{i=1}^n W_i^2, \quad W_i = e_i^T(Z_{\cdot,j} - \rho_j A_{\cdot,j}).$$

By Lemmas D.2 and D.7,

$$\|W_i\|_{\psi_a} \le \|Z_{\cdot,j} - \rho_j A_{\cdot,j}\|_{\psi_a} \le C(K_a + \bar{K}\bar{A})$$

By Lemma D.8, the union bound, and appropriate choice of constant C in the definition of Δ_H, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_2^c) &\leq \sum_{j=1}^{p} \mathbb{P}\left(\left\|Z_{\cdot,j} - \rho_j A_{\cdot,j}\right\|_2^2 > n\Delta_H\right) \\ &\leq 2\sum_{j=1}^{p}\sum_{i=1}^{n} \exp\left(-11\ln(np)\right) \\ &= \frac{2}{n^{10}p^{10}}.\end{aligned}$$

□

Proposition D.4 ($\mathcal{E}_3$)**.** *Under Assumptions 5.1, 5.2, and 5.3*

$$\mathbb{P}(\mathcal{E}_3^c) \leq \frac{2}{n^{10}p^{10}}.$$

Proof. The key equality is

$$\|\boldsymbol{U}_k\boldsymbol{U}_k^T(Z_{\cdot,j} - \rho_j A_{\cdot,j})\|_2^2 = \sum_{i=1}^{k} W_i^2, \quad W_i = u_i^T(Z_{\cdot,j} - \rho_j A_{\cdot,j}).$$

To see that it holds, set $v = Z_{\cdot,j} - \rho_j A_{\cdot,j}$. Then

$$\|\boldsymbol{U}_k\boldsymbol{U}_k^T v\|_2^2 = v^T\boldsymbol{U}_k\boldsymbol{U}_k^T\boldsymbol{U}_k\boldsymbol{U}_k^T v = v^T\boldsymbol{U}_k\boldsymbol{U}_k^T v = W^T W.$$

The rest of the argument is analogous to Proposition D.3. □

Analyzing $\mathcal{E}_4$ and $\mathcal{E}_5$

Proposition D.5 ($\mathcal{E}_4$)**.** *Under Assumption 5.3,*

$$\mathbb{P}(\mathcal{E}_4^c) \leq \frac{2}{n^{10}p^{10}}.$$

Proof. Fix $\delta > 1$. Define the event

$$\mathcal{E}_{(j)} = \left\{\frac{1}{\delta}\rho_j \leq \hat{\rho}_j \leq \delta\rho_j\right\}.$$

By the Chernoff bound in for binary random variables (Agarwal et al., 2021, Lemma I.5)

$$\mathbb{P}(\mathcal{E}_{(j)}^c) \leq 2\exp\left(-\frac{(\delta-1)^2}{2\delta^2}n\rho_j\right) \leq 2\exp\left(-\frac{(\delta-1)^2}{2\delta^2}n\rho_{\min}\right).$$

Hence by De Morgan's law and the union bound

$$\mathbb{P}(\mathcal{E}_4^c) = \mathbb{P}\left(\left\{\bigcap_{j\in[p]} \mathcal{E}_{(j)}\right\}^c\right) = \mathbb{P}\left(\bigcup_{j\in[p]} \mathcal{E}_{(j)}^c\right) \leq 2p\exp\left(-\frac{(\delta-1)^2}{2\delta^2}n\rho_{\min}\right).$$

To arrive at the desired result, solve

$$\frac{2}{n^{10}p^{10}} \geq \frac{2}{n^{11}p^{10}} = 2p\exp\left(-\frac{(\delta-1)^2}{2\delta^2}n\rho_{\min}\right)$$

for δ. □

Lemma D.9. *$\rho_{\min} > \frac{23\ln(np)}{n}$ implies $\delta \leq C < \infty$.*

Proof. By definition of δ. □

Proposition D.6 ($\mathcal{E}_5$)**.** *Under Assumption 5.3,*

$$\mathbb{P}(\mathcal{E}_5^c) \leq \frac{2}{n^{10}p^{10}}.$$

Proof. Define the event

$$\mathcal{E}_{(j)} = \{|\hat{\rho}_j - \rho_j| \leq t\}.$$

By Hoeffding's inequality for bounded random variables

$$\mathbb{P}(\mathcal{E}_{(j)}^c) \leq 2\exp\left(-2nt^2\right).$$

Hence by De Morgan's law and the union bound

$$\mathbb{P}(\mathcal{E}_5^c) = \mathbb{P}\left(\left\{\bigcap_{j\in[p]} \mathcal{E}_{(j)}\right\}^c\right) = \mathbb{P}\left(\bigcup_{j\in[p]} \mathcal{E}_{(j)}^c\right) \leq 2p\exp\left(-2nt^2\right).$$

To arrive at the desired result, solve

$$\frac{2}{n^{10}p^{10}} \geq \frac{2}{n^{11}p^{10}} = 2p\exp\left(-2nt^2\right)$$

for t. □

Summary

Define the beneficial event as $\mathcal{E} := \cap_{k=1}^5 \mathcal{E}_k$ and the adverse event as $\mathcal{E}^c = \cup_{k=1}^5 \mathcal{E}_k^c$, where $\mathcal{E}_1$ to $\mathcal{E}_5$ are defined above.

Lemma D.10. *Suppose Assumptions 5.1, 5.2, and 5.3 hold. Then,*

$$\mathbb{P}(\mathcal{E}^c) \leq \frac{10}{n^{10}p^{10}}.$$

Proof. By the union bound as well as Propositions D.2, D.3, D.4, D.5, and D.6 we have

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{k=1}^{5} \mathbb{P}(\mathcal{E}_k^c) \leq \frac{10}{n^{10}p^{10}}.$$

□

D.3 High probability bound

Recall $\boldsymbol{A} = \boldsymbol{A}^{(\text{LR})} + \boldsymbol{E}^{(\text{LR})}$ and $r = rank\{\boldsymbol{A}^{(\text{LR})}\}$. We denote the SVDs

$$\boldsymbol{A}^{(\text{LR})} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \quad \hat{\boldsymbol{A}} = \hat{\boldsymbol{U}}_k\hat{\boldsymbol{\Sigma}}_k\hat{\boldsymbol{V}}_k^T, \quad \boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^T$$

identifying `NA` with 0 in $\boldsymbol{Z}$. We denote $s_k = \Sigma_{kk}$ and $\hat{s}_k = \hat{\Sigma}_{kk}$. Recall that $\delta = \frac{1}{1-\sqrt{\frac{22\ln(np)}{n\rho_{\min}}}}$.

Definition D.1 (Thresholded projection operator)**.** *Consider a matrix $\boldsymbol{B} \in \mathbb{R}^{n\times p}$ with the SVD $\boldsymbol{B} = \sum_{i=1}^{n\wedge p} \sigma_i u_i v_i^T$. With a specific choice of $\lambda \geq 0$, we define a function $\varphi_\lambda^{\boldsymbol{B}} : \mathbb{R}^n \to \mathbb{R}^n$ as follows. For any vector $w \in \mathbb{R}^n$,*

$$\varphi_\lambda^{\boldsymbol{B}}(w) = \sum_{i=1}^{n\wedge p} 1(\sigma_i \geq \lambda) u_i u_i^T w.$$

$\varphi_\lambda^{\boldsymbol{B}}$ is a linear operator that depends on the singular values $\{\sigma_i\}$ and the left singular vectors $\{u_i\}$ of $\boldsymbol{B}$, as well as the threshold λ. If $\lambda = 0$, then use the shorthand $\varphi^{\boldsymbol{B}} = \varphi_0^{\boldsymbol{B}}$.

Lemma D.11 (Eq. 43 of Agarwal et al. (2021))**.** *Take $\lambda^* = \hat{s}_k$, where k is the PCA hyper-parameter. Then*

$$\hat{A}_{\cdot,j} = \frac{1}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}).$$

Lemma D.12. *Suppose we pick the PCA hyper-parameter $k = r$. Then,*

$$\|\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR})}\| \;\Big|\; \{\mathcal{E}_1, \mathcal{E}_4, \mathcal{E}_5\} \leq C\frac{\delta}{\rho_{\min}}\left((\sqrt{n}+\sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\|\right).$$

Proof. To begin, write

$$\begin{aligned}\|\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR})}\| &= \|\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR})}\hat{\boldsymbol{\rho}}\hat{\boldsymbol{\rho}}^{-1}\| \\ &\leq \|\hat{\boldsymbol{\rho}}^{-1}\|\|\boldsymbol{Z} - \boldsymbol{A}^{(\text{LR})}\hat{\boldsymbol{\rho}}\| \\ &= \frac{\|\boldsymbol{Z} - \boldsymbol{A}^{(\text{LR})}\hat{\boldsymbol{\rho}}\|}{\min_j \hat{\rho}_j}.\end{aligned}$$

By triangle inequality

$$\|\boldsymbol{Z} - \boldsymbol{A}^{(\text{LR})}\hat{\boldsymbol{\rho}}\| \leq \|\boldsymbol{Z} - \boldsymbol{A}^{(\text{LR})}\boldsymbol{\rho}\| + \|\boldsymbol{A}^{(\text{LR})}\|\|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\|.$$

Focusing on the first term, under $\mathcal{E}_1$

$$\|\boldsymbol{Z} - \boldsymbol{A}^{(\text{LR})}\boldsymbol{\rho}\| \leq \|\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{\rho}\| + \|\boldsymbol{A}\boldsymbol{\rho} - \boldsymbol{A}^{(\text{LR})}\boldsymbol{\rho}\| \leq (\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\|.$$

Focusing on the second term, under $\mathcal{E}_5$,

$$\|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\| \leq C\sqrt{\frac{\ln(np)}{n}}.$$

Focusing on the denominator, under $\mathcal{E}_4$,

$$\frac{1}{\hat{\rho}_j} \leq \frac{\delta}{\rho_j} \leq \frac{\delta}{\rho_{\min}}.$$

□

Remark D.1 (TRAIN and TEST)**.** *The proof technique does not depend on whether* $\hat{\boldsymbol{\rho}}^{\text{TRAIN}}$ *or* $\hat{\boldsymbol{\rho}}^{\text{TEST}}$ *is used with* $\boldsymbol{Z}^{\text{TRAIN}}$*, since* $\mathcal{E}_4$ *and* $\mathcal{E}_5$ *hold for both empirical scalings.*

Lemma D.13. *Suppose the conditions of Lemma D.12 hold. Then,*

$$\|\boldsymbol{U}\boldsymbol{U}^T - \hat{\boldsymbol{U}}_r\hat{\boldsymbol{U}}_r^T\| \;\Big|\; \{\mathcal{E}_1, \mathcal{E}_4, \mathcal{E}_5\} \leq C\frac{\delta}{\rho_{\min} s_r}\left((\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\|\right).$$

Proof. By Wedin's $\sin\Theta$ Theorem (Davis and Kahan, 1970; Wedin, 1972)

$$\|\boldsymbol{U}\boldsymbol{U}^T - \hat{\boldsymbol{U}}_r\hat{\boldsymbol{U}}_r^T\| \leq \frac{\|\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR})}\|}{s_r}.$$

Finally appeal to Lemma D.12. □

Lemma D.14. *Suppose $k = r$. Then for any $j \in [p]$*

$$\begin{aligned}
&\left\|\hat{A}_{\cdot,j} - A_{\cdot,j}\right\|_2^2 \;\Big|\; \mathcal{E} \\
&\leq \frac{C\delta^4}{\rho_{\min}^4}\left(\frac{(n+p)\Delta_{H,op}^2 + \left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + (\ln(np)/n)\|\boldsymbol{A}^{(\text{LR})}\|^2}{s_r^2}\right)\left(n\Delta_H + \left\|A_{\cdot,j}^{(\text{LR})}\right\|_2^2\right) \\
&\quad + \frac{C\delta^2}{\rho_{\min}^2} r\Delta_H + C\frac{\delta^2}{\rho_{\min}^2}\frac{\ln(np)}{n}\left\|A_{\cdot,j}\right\|_2^2 + C\left\|E_{\cdot,j}^{(\text{LR})}\right\|_2^2.
\end{aligned}$$

Proof. We proceed in steps.

1. Decomposition

   Fix a column index $j \in [p]$. Observe that

   $$\hat{A}_{\cdot,j} - A_{\cdot,j} = \left\{\hat{A}_{\cdot,j} - \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(A_{\cdot,j}\right)\right\} + \left\{\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(A_{\cdot,j}\right) - A_{\cdot,j}\right\}.$$

   By hypothesis, $k = r$. Recall that $\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}} : \mathbb{R}^n \to \mathbb{R}^n$ is the projection operator onto the span of the top r left singular vectors $\{\hat{u}_1, ..., \hat{u}_r\}$ of $\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1}$, which are are also the top r left singular vectors of $\boldsymbol{Z}$ since $\hat{\boldsymbol{\rho}}^{-1}$ is diagonal. Hence

   $$\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j}) - A_{\cdot,j} \in span\{\hat{u}_1, \ldots, \hat{u}_r\}^{\perp}.$$

   By Lemma D.11,

   $$\hat{A}_{\cdot,j} - \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j}) = \frac{1}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}) - \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j}) \in span\{\hat{u}_1, \ldots, \hat{u}_r\}.$$

   Therefore $\langle \hat{A}_{\cdot,j} - \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j}), \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j}) - A_{\cdot,j}\rangle = 0$ and

   $$\left\|\hat{A}_{\cdot,j} - A_{\cdot,j}\right\|_2^2 = \left\|\hat{A}_{\cdot,j} - \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(A_{\cdot,j}\right)\right\|_2^2 + \left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(A_{\cdot,j}\right) - A_{\cdot,j}\right\|_2^2.$$

2. First term

   Again applying Lemma D.11, we can rewrite

   $$\begin{aligned}
   \hat{A}_{\cdot,j} - \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j}) &= \frac{1}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}) - \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j}) \\
   &= \varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(\frac{1}{\hat{\rho}_j}Z_{\cdot,j} - A_{\cdot,j}\right) \\
   &= \frac{1}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j} - \rho_j A_{\cdot,j}) + \frac{\rho_j - \hat{\rho}_j}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j}).
   \end{aligned}$$

Using the parallelogram law (or, equivalently, combining Cauchy-Schwartz and AM-GM inequalities), we obtain

$$\begin{aligned}\left\|\hat{A}_{\cdot,j}-\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j})\right\|_2^2 &= \left\|\frac{1}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})+\frac{\rho_j-\hat{\rho}_j}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j})\right\|_2^2\\ &\le 2\left\|\frac{1}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})\right\|_2^2+2\left\|\frac{\rho_j-\hat{\rho}_j}{\hat{\rho}_j}\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(A_{\cdot,j})\right\|_2^2\\ &\le \frac{2}{\hat{\rho}_j^2}\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})\right\|_2^2+2\Big(\frac{\rho_j-\hat{\rho}_j}{\hat{\rho}_j}\Big)^2\|A_{\cdot,j}\|_2^2\\ &\le \frac{2\delta^2}{\rho_j^2}\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})\right\|_2^2+C\frac{\delta^2}{\rho_j^2}\frac{\ln(np)}{n}\|A_{\cdot,j}\|_2^2. \end{aligned}\tag{10}$$

Here, we have used the fact that $\mathcal{E}_4$ and $\mathcal{E}_5$ imply

$$\frac{1}{\hat{\rho}_j}\le\frac{\delta}{\rho_j},\quad \left(\frac{\rho_j-\hat{\rho}_j}{\hat{\rho}_j}\right)^2\le\frac{\delta^2}{\rho_j^2}(\hat{\rho}_j-\rho_j)^2\le C\frac{\delta^2}{\rho_j^2}\frac{\ln(np)}{n}.$$

The first term of (10) can further be decomposed.

$$\begin{aligned}&\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})\right\|_2^2\\ &\quad\le 2\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})-\varphi^{\boldsymbol{A}^{(\mathrm{LR})}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})\right\|_2^2\\ &\quad+2\left\|\varphi^{\boldsymbol{A}^{(\mathrm{LR})}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})\right\|_2^2.\end{aligned}\tag{11}$$

Finally, we bound (11). Given $k=r$, where k is the PCA hyperparameter, we can represent $\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(w)=\hat{\boldsymbol{U}}_r\hat{\boldsymbol{U}}_r^T w$ and $\varphi^{\boldsymbol{A}^{(\mathrm{LR})}}(w)=\boldsymbol{U}\boldsymbol{U}^T w$ for $w\in\mathbb{R}^n$. By Lemma D.13

$$\begin{aligned}&\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})-\varphi^{\boldsymbol{A}^{(\mathrm{LR})}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})\right\|_2\\ &\le\|\boldsymbol{U}\boldsymbol{U}^T-\hat{\boldsymbol{U}}_r\hat{\boldsymbol{U}}_r^T\|\left\|Z_{\cdot,j}-\rho_j A_{\cdot,j}\right\|_2\\ &\le C\frac{\delta}{\rho_{\min}s_r}\left((\sqrt{n}+\sqrt{p})\Delta_{H,op}+\|\boldsymbol{E}^{(\mathrm{LR})}\|+\sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\mathrm{LR})}\|\right)\left\|Z_{\cdot,j}-\rho_j A_{\cdot,j}\right\|_2.\end{aligned}$$

Combining the inequalities together, we have

$$\begin{aligned}&\left\|\hat{A}_{\cdot,j}-\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\big(A_{\cdot,j}\big)\right\|_2^2\\ &\le\frac{C\delta^4}{\rho_{\min}^2}\left(\frac{(\sqrt{n}+\sqrt{p})\Delta_{H,op}}{\rho_{\min}s_r}+\frac{\|\boldsymbol{E}^{(\mathrm{LR})}\|}{\rho_{\min}s_r}+\frac{\sqrt{\ln(np)/n}\|\boldsymbol{A}^{(\mathrm{LR})}\|}{\rho_{\min}s_r}\right)^2\left\|Z_{\cdot,j}-\rho_j A_{\cdot,j}\right\|_2^2\\ &\quad+\frac{4\delta^2}{\rho_{\min}^2}\left\|\varphi^{\boldsymbol{A}^{(\mathrm{LR})}}(Z_{\cdot,j}-\rho_j A_{\cdot,j})\right\|_2^2+C\frac{\delta^2}{\rho_{\min}^2}\frac{\ln(np)}{n}\|A_{\cdot,j}\|_2^2.\end{aligned}\tag{12}$$

3. Second term

We now bound the second term. Recalling $\boldsymbol{A} = \boldsymbol{A}^{(\mathrm{LR})} + \boldsymbol{E}^{(\mathrm{LR})}$

$$
\begin{aligned}
&\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(A_{\cdot,j}\right) - A_{\cdot,j}\right\|_2^2 \\
&= \left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(A_{\cdot,j}^{(\mathrm{LR})} + E_{\cdot,j}^{(\mathrm{LR})}\right) - A_{\cdot,j}^{(\mathrm{LR})} - E_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2 \\
&\leq 2\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(A_{\cdot,j}^{(\mathrm{LR})}\right) - A_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2 + 2\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(E_{\cdot,j}^{(\mathrm{LR})}\right) - E_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2 \\
&= 2\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(A_{\cdot,j}^{(\mathrm{LR})}\right) - \varphi^{\boldsymbol{A}^{(\mathrm{LR})}}\left(A_{\cdot,j}^{(\mathrm{LR})}\right)\right\|_2^2 + 2\left\|\varphi_{\lambda^*}^{\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1}}\left(E_{\cdot,j}^{(\mathrm{LR})}\right) - E_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2 \\
&\leq 2\left\|\boldsymbol{U}\boldsymbol{U}^T - \hat{\boldsymbol{U}}_r\hat{\boldsymbol{U}}_r^T\right\|^2\left\|A_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2 + 2\left\|E_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2 \\
&\leq C\delta^2\left(\frac{(\sqrt{n}+\sqrt{p})\Delta_{H,op}}{\rho_{\min}s_r} + \frac{\|\boldsymbol{E}^{(\mathrm{LR})}\|}{\rho_{\min}s_r} + \frac{\sqrt{\ln(np)/n}\|\boldsymbol{A}^{(\mathrm{LR})}\|}{\rho_{\min}s_r}\right)^2\left\|A_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2 + 2\left\|E_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2.
\end{aligned}
\tag{13}
$$

where the final inequality appeals to Lemma D.13.

Inserting (12) and (13) back into the decomposition, plugging the bounds in events $\mathcal{E}_2, \mathcal{E}_3$, and combining terms completes the proof.

□

Remark D.2 (TRAIN and TEST)**.** *Since $\{\hat{u}_1, ..., \hat{u}_r\}$ are the left singular vectors of $\boldsymbol{Z}^{\mathrm{TRAIN}}$, $\boldsymbol{Z}^{\mathrm{TRAIN}}\,\hat{\boldsymbol{\rho}}^{\mathrm{TRAIN}}$, and $\boldsymbol{Z}^{\mathrm{TRAIN}}\,\hat{\boldsymbol{\rho}}^{\mathrm{TEST}}$, the argument holds for both cases of interest.*

Lemma D.15. *Suppose $k = r$. Let Assumptions 5.1 and 5.4 hold. If $\rho_{\min} > \frac{23\ln(np)}{n}$ then*

$$
\begin{aligned}
&\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}\right\|_{2,\infty}^2 \;\Big|\; \mathcal{E} \\
&\leq \frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4}\left(r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right)\ln^2(np) + C\left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|_{2,\infty}^2.
\end{aligned}
$$

Proof. By Lemma D.14,

$$
\begin{aligned}
&\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \;\Big|\; \mathcal{E} \\
&\leq \frac{C\delta^4}{\rho_{\min}^4}\left(\frac{(n+p)\Delta_{H,op}^2 + \left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|^2 + (\ln(np)/n)\|\boldsymbol{A}^{(\mathrm{LR})}\|^2}{s_r^2}\right)\left(n\Delta_H + \left\|A_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2\right) \\
&\quad + \frac{C\delta^2}{\rho_{\min}^2}r\Delta_H + C(\delta-1)^2\left\|A_{\cdot,j}\right\|_2^2 + C\left\|E_{\cdot,j}^{(\mathrm{LR})}\right\|_2^2
\end{aligned}
$$

By Lemma D.9

$$\Delta_H = C(K_a + \bar{K}\bar{A})^2 \ln^2(np), \quad \delta = \frac{1}{1 - \sqrt{\frac{22\ln(np)}{n\rho_{\min}}}} \le C.$$

By Assumption 5.1 and Lemma C.4,

$$\|\boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \le Cn\bar{A}^2, \quad \|\boldsymbol{A}^{(\text{LR})}\|^2 \le Cnp\bar{A}^2.$$

The definition of δ implies

$$C\frac{\delta^2}{\rho_{\min}^2}\frac{\ln(np)}{n}\|A_{\cdot,j}\|_2^2 \;\le\; C\frac{\ln(np)}{n\rho_{\min}^2}\|A_{\cdot,j}\|_2^2 \;\le\; C\bar{A}^2\frac{\ln(np)}{\rho_{\min}^2}.$$

The second term dominates the third. Within the first term, $n\Delta_H$ dominates $\left\|A_{\cdot,j}^{(\text{LR})}\right\|_2^2$. Distribute the n, then factor out $\frac{\Delta_H}{\rho_{\min}^4}$ to combine the first term with the second. □

Lemma D.16. *Let the conditions of Lemma D.15 hold. Then*

$$\begin{aligned}&\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}\right\|_{2,\infty}^2 \;\Big|\; \mathcal{E}\\ &\le C\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2 \cdot \frac{r\ln^5(np)}{\rho_{\min}^4}\left(1 + \frac{n}{p} + n\Delta_E^2\right).\end{aligned}$$

Proof. We simplify Lemma D.15 appealing to Assumption 5.4. Recall

$$\begin{aligned}&\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}\right\|_{2,\infty}^2 \;\Big|\; \mathcal{E}\\ &\le \frac{C\bar{A}^2(K_a + \bar{K})^2}{\rho_{\min}^4}\left(r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right)\ln^2(np)\\ &\quad + C\left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^2.\end{aligned}$$

Note that

$$s_r^2 \ge C\frac{np}{r}, \quad \left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 \le np\Delta_E^2, \quad \left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^2 \le n\Delta_E^2, \quad \Delta_{H,op}^2 \le C\cdot\bar{A}^2(\kappa + \bar{K} + K_a)^2\ln^3(np).$$

Then

$$\begin{aligned}&\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}\right\|_{2,\infty}^2 \;\Big|\; \mathcal{E}\\ &\le \frac{C\bar{A}^2(K_a + \bar{K})^2}{\rho_{\min}^4}\left(r + \frac{n(n+p)\cdot\bar{A}^2(\kappa + \bar{K} + K_a)^2\ln^3(np) + n\cdot np\Delta_E^2}{\frac{np}{r}}\right)\ln^2(np)\\ &\quad + Cn\Delta_E^2\\ &\le C(K_a + \bar{K})^2\bar{A}^4(\kappa + \bar{K} + K_a)^2\cdot\frac{r\ln^5(np)}{\rho_{\min}^4}\left(1 + \frac{n}{p} + n\Delta_E^2\right).\end{aligned}$$

□

D.4 Main result

Lemma D.17. *Suppose Assumptions 5.1, 5.2, and 5.3 hold. Then*

$$\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \mathbb{1}\{\mathcal{E}^c\}\right] \leq \Delta_{adv}\frac{1}{n^2p^5}, \quad \Delta_{adv} := C\Big\{\bar{A}^2 + K_a^2 \ln^2(np)\Big\}.$$

Proof. By Cauchy-Schwarz

$$\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \mathbb{1}\{\mathcal{E}^c\}\right] \leq \sqrt{\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4\right]}\sqrt{\mathbb{E}\left[\mathbb{1}^2\{\mathcal{E}^c\}\right]}.$$

Consider the first factor. Note that

$$\max_{j\in[p]} \|\hat{A}_{\cdot,j} - A_{\cdot,j}\|_2 \leq \max_{j\in[p]} \|\hat{A}_{\cdot,j}\|_2 + \max_{j\in[p]} \|A_{\cdot,j}\|_2.$$

In the first term,

$$\begin{aligned}
\max_{j\in[p]} \|\hat{A}_{\cdot,j}\|_2 &\leq \max_{j\in[p]} \frac{1}{\hat{\rho}_j}\|Z_{\cdot,j}\|_2 \\
&\leq n \max_{j\in[p]} \|Z_{\cdot,j}\|_2 \\
&\leq n\cdot\sqrt{n} \max_{i\in[n],j\in[p]} |Z_{ij}| \\
&\leq n\cdot\sqrt{n}(\bar{A} + \max_{i,j}|H_{ij}|).
\end{aligned}$$

In the second term

$$\max_{j\in[p]} \|A_{\cdot,j}\|_2 \leq \sqrt{n}\bar{A}.$$

Collecting results

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4\right] &\leq \mathbb{E}[\{n^{\frac{3}{2}}(\bar{A} + \max_{i,j}|H_{ij}|) + \sqrt{n}\bar{A}\}^4] \\
&\leq \mathbb{E}[\{n^{\frac{3}{2}}(2\bar{A} + \max_{i,j}|H_{ij}|)\}^4] \\
&\leq Cn^6(\bar{A}^4 + \mathbb{E}[\max_{i,j}|H_{ij}|^4]) \\
&\leq Cn^6\{\bar{A}^4 + K_a^4\ln^4(np)\}.
\end{aligned}$$

The final inequality holds because for any $a > 0$ and $\theta \geq 1$, if H_{ij} is a ψ_a-random variable then $\left|H_{ij}\right|^\theta$ is a $\psi_{a/\theta}$-random variable. With the choice of $\theta = 4$, we have that

$$\mathbb{E}[\max_{i,j}|H_{ij}|^4] \leq CK_a^4\ln^{\frac{4}{a}}(np).$$

By Lemma D.10,

$$\mathbb{E}\left[\mathbb{1}^2\{\mathcal{E}^c\}\right] = \mathbb{E}\left[\mathbb{1}\{\mathcal{E}^c\}\right] = \mathbb{P}(\mathcal{E}^c) \leq \frac{C}{n^{10}p^{10}}.$$

Collecting results, we find that

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \mathbb{1}\{\mathcal{E}^c\}\right] &\leq C\sqrt{n^6(\bar{A}^4 + K_a^4 \ln^4(np))}\sqrt{\frac{1}{n^{10}p^{10}}} \\
&= C\sqrt{\bar{A}^4 + K_a^4 \ln^4(np)}\sqrt{\frac{1}{n^4p^{10}}} \\
&\leq C\Big(\bar{A}^2 + K_a^2 \ln^2(np)\Big)\frac{1}{n^2p^5}.
\end{aligned}$$

□

Remark D.3 (TRAIN and TEST)**.** *The result holds for both* $\hat{\boldsymbol{A}}^{\text{TRAIN}}$ *and* $\hat{\boldsymbol{A}}^{\text{TEST}}$ *because* $\hat{\rho}_j^{\text{TRAIN}}, \hat{\rho}_j^{\text{TEST}} \geq \frac{1}{n}$.

Proof of Theorem 5.1. Define $\mathcal{E} := \cap_{k=1}^5 \mathcal{E}_k$ where $\mathcal{E}_1$ to $\mathcal{E}_5$ are given above.

$$\mathbb{E}\Big[\Big\|\hat{\boldsymbol{A}} - \boldsymbol{A}\Big\|_{2,\infty}^2\Big] = \mathbb{E}\Big[\Big\|\hat{\boldsymbol{A}} - \boldsymbol{A}\Big\|_{2,\infty}^2 \cdot \mathbb{1}\{\mathcal{E}\}\Big] + \mathbb{E}\left[\Big\|\hat{\boldsymbol{A}} - \boldsymbol{A}\Big\|_{2,\infty}^2 \cdot \mathbb{1}\{\mathcal{E}^c\}\right].$$

Focusing on the former term, by Lemma D.16

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \mathbb{1}\{\mathcal{E}\}\right] \\
&\leq C\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2 \cdot \frac{r\ln^5(np)}{\rho_{\min}^4}\left(1 + \frac{n}{p} + n\Delta_E^2\right).
\end{aligned}$$

Focusing on the latter term

$$\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \mathbb{1}\{\mathcal{E}^c\}\right] \leq \Delta_{adv}\frac{1}{n^2p^5}, \quad \Delta_{adv} = C\Big(\bar{A}^2 + K_a^2\ln^2(np)\Big).$$

which is dominated by the former term.

Throughout this section, the arguments given are conditional on $\boldsymbol{A}$. They imply the same rate unconditional on $\boldsymbol{A}$ by the law of iterated expectations. □

Corollary D.1 (Finite sample data cleaning rate)**.** *Suppose the conditions of Theorem 5.1 hold, as well as Assumption C.1. Then*

$$\frac{1}{m}\mathbb{E}\|b(D, \hat{\boldsymbol{A}}) - b(D, \boldsymbol{A})\|_{2,\infty}^2 \leq C_b' C_1 \cdot \frac{r\ln^5(mp)}{\rho_{\min}^4}\left(\frac{1}{m} + \frac{1}{p} + \Delta_E^2\right)$$

where $C_1 = C \cdot \bar{A}^4(K_a + \bar{K})^2(\kappa + K_a + \bar{K})^2$.

Proof of Corollary D.1. Immediate from Theorem 5.1 and the definition of C_b' in Assumption C.1. □

E Error-in-variable regression

The outline of the argument is as follows

1. define TRAIN, TEST, and GENERAL ERROR
2. establish orthogonality properties
3. analyze TRAIN ERROR (more precisely, $\|\hat{\beta} - \beta^*\|_2$)
4. analyze TEST ERROR (more precisely, $\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2$)
5. analyze GENERAL ERROR (more precisely, $\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \gamma_0(\boldsymbol{A}^{\text{TEST}})\|_2^2$)

E.1 Notation and preliminaries

As in Appendix D, we identify `NA` with 0 in $\boldsymbol{Z}$ for the remainder of the appendix. We also use the notation $\boldsymbol{A}$ rather than $\boldsymbol{X}$. Recall that $(\hat{\boldsymbol{\rho}}, \hat{\beta})$ are calculated from TRAIN. We slightly abuse notation by letting n be the number of observations in TRAIN (and also TEST), departing from the notation of the main text. We write $\|\cdot\| = \|\cdot\|_{op}$. We write the proofs without nonlinear dictionaries for clarity. Then we extend our results to allow for nonlinear dictionaries in subsequent remarks. We also let $\bar{A}' = \bar{A}^{d_{\max}}$ and $\rho'_{\min} = \frac{\rho_{\min}}{d_{\max}\bar{A}'}$. Finally, to lighten notation, we abbreviate $b(D_i, A_{i,\cdot})$ as $b(A_{i,\cdot})$ when it is contextually clear.

E.1.1 Errors

Consider the following quantities:

$$\text{TR}\tilde{\text{A}}\text{IN ERROR} = \frac{1}{n}\mathbb{E}\left[\sum_{i\in\text{TRAIN}}\{\hat{A}_{i,\cdot}\hat{\beta} - \gamma_0(A_{i,\cdot})\}^2\right]$$

$$\text{T}\tilde{\text{E}}\text{ST ERROR} = \frac{1}{n}\mathbb{E}\left[\sum_{i\in\text{TEST}}\{\hat{A}_{i,\cdot}\hat{\beta} - \gamma_0(A_{i,\cdot})\}^2\right]$$

$$\text{GENERAL ERROR} = \frac{1}{n}\mathbb{E}\left[\sum_{i\in\text{TEST}}\{\hat{\gamma}_i - \gamma_0(A_{i,\cdot})\}^2\right], \quad \hat{\gamma}_i = Z_{i,\cdot}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} = Z_{i,\cdot}\tilde{\beta}.$$

The TRÃIN ERROR is standard for PCR. The TẼST ERROR is similar to Agarwal et al. (2020a). The GENERAL ERROR is a new quantity introduced in this paper, specific to our

variant of PCR that does not involve cleaning TEST. As we will see, post multiplying by $\tilde{\beta}$ performs a kind of implicit cleaning. By avoiding explicit cleaning, we preserve independence across rows in TEST, which is critical for our inference argument. Therefore the key result is about GENERAL ERROR. En route, we will analyze quantities we refer to as TRAIN ERROR and TEST ERROR, which are closely related to TRÃIN ERROR and TẼST ERROR. When using a dictionary, the updated estimator is $\hat{\gamma}_i = b(D_i, Z_{i,\cdot}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\beta} = b(D_i, Z_{i,\cdot})\tilde{\beta}$ for an updated definition of $\tilde{\beta}$.

Proof of Proposition 4.2. For $i \in \text{TEST}$

$$\begin{aligned}
\hat{\gamma}(D_i, Z_{i,\cdot}) &= b(D_i, Z_{i,\cdot}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\beta} \\
&= \begin{bmatrix} D_i Z_{i,\cdot}\,\hat{\boldsymbol{\rho}}^{-1} & (1-D_i) Z_{i,\cdot}\,\hat{\boldsymbol{\rho}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta}^{\text{TREAT}} \\ \hat{\beta}^{\text{UNTREAT}} \end{bmatrix} \\
&= \begin{bmatrix} D_i Z_{i,\cdot} & (1-D_i) Z_{i,\cdot} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}^{\text{TREAT}} \\ \hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}^{\text{UNTREAT}} \end{bmatrix}.
\end{aligned}$$

Finally, independence holds conditional on TRAIN since $(\hat{\boldsymbol{\rho}}, \hat{\beta})$ are calculated from TRAIN, so the only randomness that remains is in $(D_i, Z_{i,\cdot})$ and $(D_j, Z_{j,\cdot})$ which are i.n.i.d. □

E.1.2 SVDs

Recall that the FILL operator rescales using $\hat{\boldsymbol{\rho}}$ calculated from TRAIN. Denote the SVDs

$$\boldsymbol{A}^{(\text{LR}),\text{TRAIN}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \quad \text{FILL}(\boldsymbol{Z}^{\text{TRAIN}}) = \boldsymbol{Z}^{\text{TRAIN}}\,\hat{\boldsymbol{\rho}}^{-1} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^T, \quad \hat{\boldsymbol{A}}^{\text{TRAIN}} = \hat{\boldsymbol{U}}_k\hat{\boldsymbol{\Sigma}}_k\hat{\boldsymbol{V}}_k^T.$$

In this notation, $\boldsymbol{V}$ is an orthonormal basis for $\text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\}$. Let $\boldsymbol{V}_\perp$ be an orthonormal basis for the orthogonal complement to $\text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\}$. In other words, for any element $v \in \text{ROW}(\boldsymbol{V}^T)$, $\boldsymbol{V}_\perp^T v = 0$. Likewise we define $\hat{\boldsymbol{V}}_{k,\perp}$. Define s_k and $\hat{s}_k$ as the k-th singular values of $\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}$ and $\hat{\boldsymbol{A}}^{\text{TRAIN}}$, respectively.

Next, denote the SVDs

$$\boldsymbol{A}^{(\text{LR}),\text{TEST}} = \boldsymbol{U}'\boldsymbol{\Sigma}'(\boldsymbol{V}')^T, \quad \text{FILL}(\boldsymbol{Z}^{\text{TEST}}) = \boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} = \hat{\boldsymbol{U}}'\hat{\boldsymbol{\Sigma}}'(\hat{\boldsymbol{V}}')^T, \quad \hat{\boldsymbol{A}}^{\text{TEST}} = \hat{\boldsymbol{U}}'_k\hat{\boldsymbol{\Sigma}}'_k(\hat{\boldsymbol{V}}'_k)^T.$$

We define $\boldsymbol{V}'_\perp$ and $\hat{\boldsymbol{V}}'_\perp$ analogously to $\boldsymbol{V}_\perp$. Define s'_k and $\hat{s}'_k$ as the k-th singular values of $\boldsymbol{A}^{(\text{LR}),\text{TEST}}$ and $\hat{\boldsymbol{A}}^{\text{TEST}}$, respectively.

Finally, denote the SVD of the row-wise concatenation of $\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}$ and $\boldsymbol{A}^{(\text{LR}),\text{TEST}}$ as

$$\begin{bmatrix} \boldsymbol{A}^{(\text{LR}),\text{TRAIN}} \\ \boldsymbol{A}^{(\text{LR}),\text{TEST}} \end{bmatrix} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^T.$$

We define $\tilde{\boldsymbol{V}}_\perp$ analogously to $\boldsymbol{V}_\perp$ but with respect to the row-wise concatenation of $\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}$ and $\boldsymbol{A}^{(\text{LR}),\text{TEST}}$.

Remark E.1 (Dictionary)**.** *As discussed in Assumption C.2, to simplify our dictionary discussion, we impose that $r' = k'$. Given the alternative definitions of $(\beta^*, \hat{\beta})$, our analysis requires that we consider the alternative SVDs*

$$b\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \quad b(\hat{\boldsymbol{A}}^{\text{TRAIN}}) = \hat{\boldsymbol{U}}_{r'}\hat{\boldsymbol{\Sigma}}_{r'}\hat{\boldsymbol{V}}_{r'}^T$$

and

$$b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\} = \boldsymbol{U}'\boldsymbol{\Sigma}'(\boldsymbol{V}')^T, \quad b(\hat{\boldsymbol{A}}^{\text{TEST}}) = \hat{\boldsymbol{U}}'_{r'}\hat{\boldsymbol{\Sigma}}'_{r'}(\hat{\boldsymbol{V}}'_{r'})^T.$$

We slightly abuse notation and depart from the main text by re-using the SVD symbols. In particular, we denote the r'-th singular value of $\hat{\boldsymbol{\Sigma}}_{r'}$ by $\hat{s}_{r'}$.

E.2 Orthogonality

The goal of this section is to establish orthogonality properties for the analysis to follow. We begin by verifying that Assumption 5.6 holds with high probability under auxiliary assumptions. In particular, Proposition E.1 provides intuition for how Assumption 5.6 can hold even with i.n.i.d. data. The auxiliary conditions of Proposition E.1 impose more structure than we need for the main argument.

Lemma E.1 (Two sided bound on sub-Gaussian matrices; Theorem 4.6.1 of Vershynin (2018))**.** *Let $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ whose rows $U_{i,\cdot} \in \mathbb{R}^r$ are independent, mean zero, sub-Gaussian, and isotropic with $\|U_{i,\cdot}\|_{\psi_2} \leq K_u$. Then for any $t \geq 0$, w.p. $1 - 2e^{-t^2}$*

$$\sqrt{m} - CK_u^2(\sqrt{r} + t) \leq s_r(\boldsymbol{U}) \leq s_1(\boldsymbol{U}) \leq \sqrt{m} + CK_u^2(\sqrt{r} + t).$$

Proposition E.1 (Verifying row space inclusion)**.** *By hypothesis, $rank\{\boldsymbol{A}^{(\text{LR})}\} = r$, so it admits a representation $A_{ij}^{(\text{LR})} = \langle u_i, v_j, \rangle$ where $u_i, v_j \in \mathbb{R}^r$. Suppose $\{u_i\}$ are independent, mean zero, sub-Gaussian, and isotropic (i.e. $\mathbb{V}(u_i) = I_r$) with $\|u_i\|_{\psi_2} \leq K_u$. Suppose $m \gg K_u^4 \cdot r\ln(mp)$. Then with probability $1 - O\{(mp)^{-10}\}$, $\text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\} = \text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\}$.*

Proof. Consider $\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}$. Let $\boldsymbol{U}$ have rows $\{U_{i,\cdot}\}$. By Lemma E.1 with $t = \ln^{\frac{1}{2}}(mp)$,

$$s_r(\boldsymbol{U}) \geq \sqrt{m} - CK_u^2\{\sqrt{r} + \ln^{\frac{1}{2}}(mp)\} \gg 0.$$

With high probability, $s_r(\boldsymbol{U}) \gg 0$, implying that $\{U_{i,\cdot}\}$ are full rank so that $\text{ROW}(\boldsymbol{U}) = \mathbb{R}^r$.

Now consider $\boldsymbol{A}^{(\text{LR}),\text{TEST}}$. Let $\boldsymbol{U}'$ have rows $\{U'_{i,\cdot}\}$. Fix $i \in \text{TEST}$. Since $U'_{i,\cdot} \in \mathbb{R}^r = \text{ROW}(\boldsymbol{U})$, there exists some $\lambda \in \mathbb{R}^r$ such that $U'_{i,\cdot} = \sum_{k=1}^r \lambda_k U_{k,\cdot}$. Therefore

$$A_{ij}^{(\text{LR}),\text{TEST}} = \langle U'_{i,\cdot}, V_{\cdot,j} \rangle = \left\langle \sum_{k=1}^r \lambda_k U_{k,\cdot}, V_{\cdot,j} \right\rangle = \sum_{k=1}^r \lambda_k \langle U_{k,\cdot}, V_{\cdot,j} \rangle = \sum_{k=1}^r \lambda_k A_{kj}^{(\text{LR}),\text{TRAIN}}.$$

In summary, for any $i \in \text{TEST}$, $A_{i,\cdot}^{(\text{LR}),\text{TEST}} \in \text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\}$. Therefore $\text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\} \subset \text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\}$. Likewise for the other direction. □

Next, we turn to the orthogonality properties of interest. In order to formalize these orthogonality properties, we formally define β^*. To begin, consider the case without a dictionary. We define $\beta^* \in \mathbb{R}^p$ as the unique solution to the following optimization problem across TRAIN and TEST:

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_2 \text{ s.t. } \beta \in \text{argmin} \left\| \begin{bmatrix} \gamma_0(\boldsymbol{A}^{\text{TRAIN}}) \\ \gamma_0(\boldsymbol{A}^{\text{TEST}}) \end{bmatrix} - \begin{bmatrix} \boldsymbol{A}^{(\text{LR}),\text{TRAIN}} \\ \boldsymbol{A}^{(\text{LR}),\text{TEST}} \end{bmatrix} \beta \right\|_2^2.$$

β^* is not the quantity of interest, but rather a theoretical device. It defines the unique, minimal-norm, low-rank, linear approximation to the regression γ_0. The theoretical device β^* generalizes the target quantity of Agarwal et al. (2020a), who study error-in-variables regression in the exactly linear, exactly low rank, no dictionary special case (i.e. $\phi_i^{(\text{LR})} = 0$). Our ultimate goal is to define and analyze an estimator close to $\gamma_0(A_{i,\cdot})$ in generalized mean square error while adhering to the conditional independence criterion of Proposition 4.2.

Remark E.2 (Dictionary)**.** *When using a dictionary, we update our definition of $\beta^* \in \mathbb{R}^{p'}$ as the unique solution to the following optimization problem across* TRAIN *and* TEST*:*

$$\min_{\beta \in \mathbb{R}^{p'}} \|\beta\|_2 \text{ s.t. } \beta \in \text{argmin} \left\| \begin{bmatrix} \gamma_0(\boldsymbol{A}^{\text{TRAIN}}) \\ \gamma_0(\boldsymbol{A}^{\text{TEST}}) \end{bmatrix} - \begin{bmatrix} b\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\} \\ b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\} \end{bmatrix} \beta \right\|_2^2.$$

Lemma E.2 (Orthogonality)**.** *Suppose Assumption 5.6 holds. Then,*

$$\hat{\boldsymbol{V}}_{k,\perp}^T \hat{\beta} = 0, \quad \boldsymbol{V}_\perp^T \beta^* = (\boldsymbol{V}'_\perp)^T \beta^* = 0.$$

Proof. We show each result

1. $\hat{\boldsymbol{V}}_{k,\perp}^T \hat{\beta} = 0$. By Agarwal et al. (2020a, Property 4.1), $\hat{\beta} \in \text{ROW}(\hat{\boldsymbol{A}}^{\text{TRAIN}}) = \text{ROW}(\hat{\boldsymbol{V}}_k^T) = \text{COL}(\hat{\boldsymbol{V}}_k)$. Recall that $\hat{\boldsymbol{V}}_{k,\perp}$ is the basis of $\text{NULL}(\hat{\boldsymbol{V}}_k)$. Therefore by orthogonality of $\text{COL}(\hat{\boldsymbol{V}}_k)$ and $\text{NULL}(\hat{\boldsymbol{V}}_k)$, the result holds.

2. $\boldsymbol{V}_\perp^T \beta^* = (\boldsymbol{V}'_\perp)^T \beta^* = 0$.

   Using the definition of β^* as the minimal ℓ_2-norm solution, and an identical argument to the one above, we have $\beta^* \in \text{COL}(\tilde{\boldsymbol{V}})$.

   Moreover by Assumption 5.6,

   $$\text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\} = \text{ROW}\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\} = \text{ROW}\left\{\begin{bmatrix}\boldsymbol{A}^{(\text{LR}),\text{TRAIN}} \\ \boldsymbol{A}^{(\text{LR}),\text{TEST}}\end{bmatrix}\right\}.$$

   Therefore $\text{ROW}(\boldsymbol{V}^T) = \text{ROW}\{(\boldsymbol{V}')^T\} = \text{ROW}(\tilde{\boldsymbol{V}}^T)$ i.e. $\text{COL}(\boldsymbol{V}) = \text{COL}(\boldsymbol{V}') = \text{COL}(\tilde{\boldsymbol{V}})$. In summary, $\beta^* \in \text{COL}(\boldsymbol{V}) = \text{COL}(\boldsymbol{V}')$.

   Recall that $\boldsymbol{V}_\perp$ is the basis of $\text{NULL}(\boldsymbol{V})$ and likewise $\boldsymbol{V}'_\perp$ is the basis of $\text{NULL}(\boldsymbol{V}')$. Therefore by orthogonality of $\text{COL}(\boldsymbol{V})$ and $\text{NULL}(\boldsymbol{V})$ as well as orthogonality of $\text{COL}(\boldsymbol{V}')$ and $\text{NULL}(\boldsymbol{V}')$, the result holds.

□

Remark E.3 (Dictionary)**.** *Lemma E.2 continues to hold with the updated definitions of the SVDs in Remark E.1.*

E.3 Training error

Recall $Y_i = A_{i,\cdot}^{(\text{LR})} \beta^* + \phi_i^{(\text{LR})} + \varepsilon_i$. Denote by $Y^{\text{TRAIN}} \in \mathbb{R}^n$ the concatenation of $(Y_i)_{i \in \text{TRAIN}}$. Likewise for $\varepsilon^{\text{TRAIN}}$ and $\phi^{(\text{LR}),\text{TRAIN}}$. In the argument for TRAIN ERROR, all objects correspond to TRAIN. For this reason, we suppress superscipt TRAIN.

E.3.1 Decomposition

Lemma E.3. *Deterministically,*

$$\|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\text{LR})}\beta^*\|_2^2 \le C\left\{\|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \|\beta^*\|_1^2 \ \vee \ \|\phi^{(\text{LR})}\|_2^2 \ \vee \ \langle \hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle\right\}.$$

Proof. Write

$$\begin{aligned}\|\hat{\boldsymbol{A}}\hat{\beta} - Y\|_2^2 &= \|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^* - \phi^{(\mathrm{LR})} - \varepsilon\|_2^2 \\ &= \|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^* - \phi^{(\mathrm{LR})}\|_2^2 + \|\varepsilon\|_2^2 - 2\langle\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*, \varepsilon\rangle + 2\langle\phi^{(\mathrm{LR})}, \varepsilon\rangle. \quad (14)\end{aligned}$$

By optimality of $\hat{\beta}$, we have

$$\begin{aligned}\|\hat{\boldsymbol{A}}\hat{\beta} - Y\|_2^2 &\leq \|\hat{\boldsymbol{A}}\beta^* - Y\|_2^2 \\ &= \|(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})})\beta^* - \phi^{(\mathrm{LR})} - \varepsilon\|_2^2 \\ &= \|(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})})\beta^* - \phi^{(\mathrm{LR})}\|_2^2 + \|\varepsilon\|_2^2 - 2\langle(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})})\beta^*, \varepsilon\rangle + 2\langle\phi^{(\mathrm{LR})}, \varepsilon\rangle. \quad (15)\end{aligned}$$

From (14) and (15), we have

$$\|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^* - \phi^{(\mathrm{LR})}\|_2^2 \leq \|(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})})\beta^* - \phi^{(\mathrm{LR})}\|_2^2 + 2\langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle. \quad (16)$$

Now consider

$$\|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^* - \phi^{(\mathrm{LR})}\|_2^2 = \|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2^2 + \|\phi^{(\mathrm{LR})}\|_2^2 - 2\langle\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*, \phi^{(\mathrm{LR})}\rangle; \quad (17)$$

$$\|(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})})\beta^* - \phi^{(\mathrm{LR})}\|_2^2 = \|(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})})\beta^*\|_2^2 + \|\phi^{(\mathrm{LR})}\|_2^2 - 2\langle(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})})\beta^*, \phi^{(\mathrm{LR})}\rangle. \quad (18)$$

Combining (16), (17), and (18)

$$\|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2^2 \leq \|(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})})\beta^*\|_2^2 + 2\langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \phi^{(\mathrm{LR})}\rangle + 2\langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle. \quad (19)$$

By Cauchy-Schwarz

$$\langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \phi^{(\mathrm{LR})}\rangle \leq \|\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*)\|_2 \cdot \|\phi^{(\mathrm{LR})}\|_2. \quad (20)$$

Focusing on the former factor

$$\|\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*)\|_2 \leq \|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2 + \|\hat{\boldsymbol{A}}\beta^* - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2. \quad (21)$$

Let $a = \|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2^2$ and $b = \|\hat{\boldsymbol{A}}\beta^* - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2^2$. Then (19), (20), and (21) imply

$$a \leq b + 2(\sqrt{a} + \sqrt{b})\|\phi^{(\mathrm{LR})}\|_2 + 2\langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle = 2\sqrt{a}\|\phi^{(\mathrm{LR})}\|_2 + c \quad (22)$$

where $c = b + 2\sqrt{b}\|\phi^{(\mathrm{LR})}\|_2 + 2\langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle$. We now analyze the expression $a \leq 2\sqrt{a}\|\phi^{(\mathrm{LR})}\|_2 + c$. Note $a \geq 0$. There are three possible cases.

1. $a \geq 0$, $c \geq 0$, $2\sqrt{a}\|\phi^{(\text{LR})}\|_2 \geq c$:

$$a \leq 4\sqrt{a}\|\phi^{(\text{LR})}\|_2 \implies a \leq 16\|\phi^{(\text{LR})}\|_2^2.$$

2. $a \geq 0$, $c \geq 0$, $2\sqrt{a}\|\phi^{(\text{LR})}\|_2 < c$:

$$a \leq 2c.$$

3. $a \geq 0$, $c < 0$:

$$a < 2\sqrt{a}\|\phi^{(\text{LR})}\|_2 \implies a < 4\|\phi^{(\text{LR})}\|_2^2.$$

(22) and the three cases above imply

$$a \leq 2c \vee 16\|\phi^{(\text{LR})}\|_2^2. \tag{23}$$

Let $d \coloneqq 2b + 4\sqrt{b}\|\phi^{(\text{LR})}\|_2 + 2\|\phi^{(\text{LR})}\|_2^2$. Then

$$d = 2b + 4\sqrt{b}\|\phi^{(\text{LR})}\|_2 + 2\|\phi^{(\text{LR})}\|_2^2 = 2\{\sqrt{b} + \|\phi^{(\text{LR})}\|_2\}^2 \leq 4\{b + \|\phi^{(\text{LR})}\|_2^2\}. \tag{24}$$

Note $2c \leq d + 4\langle \hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon \rangle$. This together with (23) and (24) together implies

$$a \leq C\left\{b \ \vee \ \|\phi^{(\text{LR})}\|_2^2 \ \vee \ \langle \hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon \rangle\right\}.$$

Finally note $b \leq \|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \|\beta^*\|_1^2$. □

Remark E.4 (Dictionary)**.** *The generalization of Lemma E.3 is*

$$\|b(\hat{\boldsymbol{A}})\hat{\beta} - b\{\boldsymbol{A}^{(\text{LR})}\}\beta^*\|_2^2 \leq C\left\{\|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2 \|\beta^*\|_1^2 \ \vee \ \|\phi^{(\text{LR})}\|_2^2 \ \vee \ \langle b(\hat{\boldsymbol{A}})(\hat{\beta} - \beta^*), \varepsilon \rangle\right\}.$$

E.3.2 Parameter

Lemma E.4. *Let the conditions of Lemma E.2 hold. Then*

$$\|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T(\hat{\beta} - \beta^*)\|_2^2 \leq \frac{C}{\hat{s}_k^2}\left\{\|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \|\beta^*\|_1^2 \ \vee \ \|\phi^{(\text{LR})}\|_2^2 \ \vee \ \langle \hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon \rangle\right\}$$

Proof. To begin, note that

$$\|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T(\hat{\beta} - \beta^*)\|_2^2 = \|\hat{\boldsymbol{V}}_k^T(\hat{\beta} - \beta^*)\|_2^2, \tag{25}$$

since $\hat{\boldsymbol{V}}_k$ is an isometry. Recall that $\hat{\boldsymbol{A}} = \hat{\boldsymbol{U}}_k \hat{\boldsymbol{\Sigma}}_k \hat{\boldsymbol{V}}_k^T$. Therefore,

$$\begin{aligned}\|\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*)\|_2^2 &= (\hat{\beta} - \beta^*)^T \hat{\boldsymbol{V}}_k \hat{\boldsymbol{\Sigma}}_k^2 \hat{\boldsymbol{V}}_k^T (\hat{\beta} - \beta^*) \\ &\geq \hat{s}_k^2 \|\hat{\boldsymbol{V}}_k^T (\hat{\beta} - \beta^*)\|_2^2. \end{aligned} \tag{26}$$

Next, consider

$$\begin{aligned}\|\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*)\|_2^2 &\leq 2\|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2^2 + 2\|\boldsymbol{A}^{(\mathrm{LR})}\beta^* - \hat{\boldsymbol{A}}\beta^*\|_2^2 \\ &\leq 2\|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2^2 + 2\|\boldsymbol{A}^{(\mathrm{LR})} - \hat{\boldsymbol{A}}\|_{2,\infty}^2 \|\beta^*\|_1^2. \end{aligned} \tag{27}$$

Using (25), (26), and (27), we have

$$\|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T (\hat{\beta} - \beta^*)\|_2^2 \leq \frac{2}{\hat{s}_k^2} \Big\{ \|\hat{\boldsymbol{A}}\hat{\beta} - \boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2^2 + \|\boldsymbol{A}^{(\mathrm{LR})} - \hat{\boldsymbol{A}}\|_{2,\infty}^2 \|\beta^*\|_1^2 \Big\}.$$

Finally using Lemma E.3, we conclude that

$$\|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T (\hat{\beta} - \beta^*)\|_2^2 \leq \frac{C}{\hat{s}_k^2} \Big\{ \|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}^2 \|\beta^*\|_1^2 \ \vee \ \|\phi^{(\mathrm{LR})}\|_2^2 \ \vee \ \langle \hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon \rangle \Big\}. \tag{28}$$

□

Remark E.5 (Dictionary)**.** *The generalization of Lemma E.4 is*

$$\|\hat{\boldsymbol{V}}_{r'} \hat{\boldsymbol{V}}_{r'}^T (\hat{\beta} - \beta^*)\|_2^2 \leq \frac{C}{\hat{s}_{r'}^2} \Big\{ \|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\mathrm{LR})}\}\|_{2,\infty}^2 \|\beta^*\|_1^2 \ \vee \ \|\phi^{(\mathrm{LR})}\|_2^2 \ \vee \ \langle b(\hat{\boldsymbol{A}})(\hat{\beta} - \beta^*), \varepsilon \rangle \Big\}$$

where the SVDs are as in Remark E.1 and $\hat{s}_{r'}$ is defined accordingly.

Lemma E.5. *Let the conditions of Lemma E.2 hold. Then,*

$$\begin{aligned}&\|\hat{\beta} - \beta^*\|_2^2 \\ &\quad \leq C \left[\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T\|^2 \|\beta^*\|_2^2 + \frac{1}{\hat{s}_k^2} \Big\{ \|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}^2 \|\beta^*\|_1^2 \ \vee \ \|\phi^{(\mathrm{LR})}\|_2^2 \ \vee \ \langle \hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon \rangle \Big\} \right]. \end{aligned}$$

Proof. Write

$$\begin{aligned}\|\hat{\beta} - \beta^*\|_2^2 &= \|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T (\hat{\beta} - \beta^*) + \hat{\boldsymbol{V}}_{k,\perp} \hat{\boldsymbol{V}}_{k,\perp}^T (\hat{\beta} - \beta^*)\|_2^2 \\ &= \|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T (\hat{\beta} - \beta^*)\|_2^2 + \|\hat{\boldsymbol{V}}_{k,\perp} \hat{\boldsymbol{V}}_{k,\perp}^T (\hat{\beta} - \beta^*)\|_2^2 \\ &= \|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T (\hat{\beta} - \beta^*)\|_2^2 + \|\hat{\boldsymbol{V}}_{k,\perp} \hat{\boldsymbol{V}}_{k,\perp}^T \beta^*\|_2^2. \end{aligned} \tag{29}$$

In the last equality we have used Lemma E.2. Next, we bound the two terms in (29).

1. Bounding $\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\beta}-\beta^*)\|_2^2$.

   The bound follows from Lemma E.4.

2. Bounding $\|\hat{\boldsymbol{V}}_{k,\perp}\hat{\boldsymbol{V}}_{k,\perp}^T\beta^*\|_2^2$.

   Write

$$\begin{aligned}\|\hat{\boldsymbol{V}}_{k,\perp}\hat{\boldsymbol{V}}_{k,\perp}^T\beta^*\|_2^2 &= \|(\hat{\boldsymbol{V}}_{k,\perp}\hat{\boldsymbol{V}}_{k,\perp}^T\beta^* - \boldsymbol{V}_\perp\boldsymbol{V}_\perp^T)\beta^*\|_2^2\\ &\le \|\hat{\boldsymbol{V}}_{k,\perp}\hat{\boldsymbol{V}}_{k,\perp}^T - \boldsymbol{V}_\perp\boldsymbol{V}_\perp^T\|^2\|\beta^*\|_2^2\\ &= \|(\boldsymbol{I}-\boldsymbol{V}_\perp\boldsymbol{V}_\perp^T) - (\boldsymbol{I}-\hat{\boldsymbol{V}}_{k,\perp}\hat{\boldsymbol{V}}_{k,\perp}^T)\|^2\|\beta^*\|_2^2\\ &= \|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T\|^2\|\beta^*\|_2^2.\end{aligned}$$

   Here we have used $(\boldsymbol{V}_\perp\boldsymbol{V}_\perp^T)\beta^* = 0$ by Lemma E.2.

□

Remark E.6 (Dictionary)**.** *The generalization of Lemma E.5 is*

$$\begin{aligned}&\|\hat{\beta}-\beta^*\|_2^2\\ &\le C\left[\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T\|^2\|\beta^*\|_2^2 + \frac{1}{\hat{s}_{r'}^2}\left\{\|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\mathrm{LR})}\}\|_{2,\infty}^2\|\beta^*\|_1^2 \ \vee\ \|\phi^{(\mathrm{LR})}\|_2^2 \ \vee\ \langle b(\hat{\boldsymbol{A}})(\hat{\beta}-\beta^*), \varepsilon\rangle\right\}\right]\end{aligned}$$

where the SVDs are as in Remark E.1 and $\hat{s}_{r'}$ is defined accordingly.

E.3.3 High probability events

Lemma E.6 (Weyl's Inequality)**.** *Assume $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times p}$. Let s_k and $\hat{s}_k$ be the k-th singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, in decreasing order and repeated by multiplicities. Then for all $k \in [n \wedge p]$,*

$$|s_k - \hat{s}_k| \le \|\boldsymbol{A}-\boldsymbol{B}\|.$$

Lemma E.7. *Suppose the conditions of Lemma D.12 hold. Then,*

$$|s_r - \hat{s}_r| \ \Big|\ \{\mathcal{E}_1, \mathcal{E}_4, \mathcal{E}_5\} \le C\frac{\delta}{\rho_{\min}}\left\{(\sqrt{n}+\sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\mathrm{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\mathrm{LR})}\|\right\}.$$

Proof. By Weyl's inequality as in Lemma E.6, we obtain

$$|s_r - \hat{s}_r| \le \|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\mathrm{LR})}\|.$$

Apply Lemma D.12 to complete the proof. □

Remark E.7 (Dictionary)**.** *The generalization of Lemma E.7 is*

$$|s_{r'} - \hat{s}_{r'}| \;\Big|\; \{\mathcal{E}_1, \mathcal{E}_4, \mathcal{E}_5\} \leq C\frac{\delta}{\rho'_{\min}} \left((\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\| \right).$$

where $\rho'_{\min} = \frac{\rho_{\min}}{\bar{A}^{d_{\max}} d_{\max}}$.

Proof. We proceed in steps.

1. Decomposition

   With updated SVDs as in Remark E.1, Weyl's inequality implies

   $$|s_{r'} - \hat{s}_{r'}| \leq \|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\text{LR})}\}\| \leq \|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1})\| + \|b(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|.$$

2. Former term

   For a polynomial dictionary with uncorrupted nonlinearity (Definition C.2), the former term simplifies as

   $$\begin{aligned}
   &\|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1})\| \\
   &= \|\{0, 0, ..., 0, (\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}), diag(D)(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}), ..., diag(D)^{d_{\max}-1}(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}})\}\| \\
   &\leq \|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}\| + \bar{A}\|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}\| + ... + \bar{A}^{d_{\max}-1}\|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}\| \\
   &\leq \bar{A}^{d_{\max}} d_{\max}\|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}\|.
   \end{aligned}$$

   Next observe that by Lemma E.7, under the beneficial events,

   $$\|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}\| = \hat{s}_{r+1} = \hat{s}_{r+1} - s_{r+1} \leq C\frac{\delta}{\rho_{\min}} \left\{ (\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\| \right\}.$$

3. Latter term

   By a similar argument,

   $$\|b(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1}) - b\{\boldsymbol{A}^{(\text{LR})}\}\| \leq \bar{A}^{d_{\max}} d_{\max}\|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR})}\|.$$

   By Lemma D.12, under the beneficial events,

   $$\|\boldsymbol{Z}\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR})}\| \leq C\frac{\delta}{\rho_{\min}} \left((\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\| \right).$$

□

Lemma E.8. *Suppose $k = r$. Then*

$$\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\| \;\Big|\; \{\mathcal{E}_1, \mathcal{E}_4, \mathcal{E}_5\} \leq C\frac{\delta}{\rho_{\min} s_r}\left\{(\sqrt{n}+\sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\|\right\}.$$

Proof. Similar to Lemma D.13, we apply Wedin's $\sin\Theta$ Theorem (Davis and Kahan, 1970; Wedin, 1972)) to arrive at the following inequality:

$$\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\| \leq \frac{\|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR})}\|}{s_r}.$$

Apply Lemma D.12 to complete the proof. □

Remark E.8 (Dictionary)**.** *The generalization is*

$$\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T\| \;\Big|\; \{\mathcal{E}_1, \mathcal{E}_4, \mathcal{E}_5\}$$
$$C\frac{\delta}{\rho'_{\min} s_{r'}}\left((\sqrt{n}+\sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\|\right).$$

Proof. Wedin's $\sin\Theta$ Theorem gives

$$\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T\| \leq \frac{\|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|}{s_{r'}}.$$

The numerator is handled in the proof of Remark E.7. □

Lemma E.9. *Suppose Assumptions 5.1, 5.2, 5.3, and 5.4 hold, and $k = r$. If*

$$\rho_{\min} \gg \tilde{C}\sqrt{r}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right), \quad \tilde{C} := C\bar{A}\Big(\kappa + \bar{K} + K_a\Big),$$

then with probability at least $1 - O\{1/(np)^{10}\}$, $\hat{s}_r \gtrsim s_r$.

Proof. The argument is as follows.

1. By Lemma E.7, $|\hat{s}_r - s_r| \leq \Delta$, hence $\hat{s}_r \geq s_r - \Delta$, where Δ is defined below.
2. We want to show $\Delta = o(s_r)$, i.e. $\Delta \leq c_n s_r$ where $c_n \to 0$; it is sufficient to show $\frac{\Delta}{s_r} \to 0$.
3. In such case, $\hat{s}_r \geq s_r - \Delta \geq s_r - c_n s_r = (1 - c_n)s_r$, i.e. $\hat{s}_r \gtrsim s_r$.

By Lemma D.10 and Lemma E.7, with probability at least $1 - O\{1/(np)^{10}\}$

$$\begin{aligned}\Delta &:= C\frac{\delta}{\rho_{\min}}\left\{(\sqrt{n}+\sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\|\right\}\\ &\le C\frac{1}{\rho_{\min}}\left\{\bar{A}\Big(\kappa+\bar{K}+K_a\Big)\Big(\sqrt{n}+\sqrt{p}\Big)\ln^{\frac{3}{2}}(np) + \sqrt{np}\Delta_E + \sqrt{\frac{\ln(np)}{n}}\sqrt{np}\bar{A}\right\}\\ &\le C\frac{\bar{A}\Big(\kappa+\bar{K}+K_a\Big)}{\rho_{\min}}\ln^{\frac{3}{2}}(np)\left(\sqrt{n}+\sqrt{p}+\sqrt{np}\Delta_E\right).\end{aligned}$$

Moreover by Assumption 5.4

$$s_r \ge C\sqrt{\frac{np}{r}}.$$

Therefore a sufficient condition for the lemma statement to hold is

$$\frac{\Delta}{s_r} \le C\frac{\bar{A}\Big(\kappa+\bar{K}+K_a\Big)}{\rho_{\min}}\cdot\sqrt{r}\ln^{\frac{3}{2}}(np)\cdot\left(\frac{1}{\sqrt{p}}+\frac{1}{\sqrt{n}}+\Delta_E\right)\to 0$$

i.e.

$$\rho_{\min} \gg \tilde{C}\sqrt{r}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}}\vee\frac{1}{\sqrt{n}}\vee\Delta_E\right),\quad \tilde{C} := C\bar{A}\Big(\kappa+\bar{K}+K_a\Big).$$

□

Remark E.9 (Dictionary)**.** *The generalization of Lemma E.9 is*

$$\rho'_{\min} := \frac{\rho_{\min}}{\bar{A}^{d_{\max}}d_{\max}} \gg \tilde{C}\sqrt{r'}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}}\vee\frac{1}{\sqrt{n}}\vee\Delta_E\right).$$

Proof. By Remark E.7,

$$\Delta := C\frac{\delta}{\rho'_{\min}}\left((\sqrt{n}+\sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\|\right).$$

Moreover by the generalized Assumption 5.4

$$s_{r'} \ge C\sqrt{\frac{np}{r'}}.$$

□

Lemma E.10. *Suppose Assumption 5.5 holds. Recall k is the PCA hyperparameter. Then*

$$\mathbb{E}\langle\hat{\boldsymbol{A}}(\hat{\beta}-\beta^*),\varepsilon\rangle \le \bar{\sigma}^2 k.$$

Proof. Note that

$$\hat{\beta} = \hat{\boldsymbol{A}}^{\dagger} Y = \hat{\boldsymbol{A}}^{\dagger}\{\boldsymbol{A}^{(\text{LR})}\beta^* + \varepsilon + \phi^{(\text{LR})}\}.$$

Since ε is independent of $\hat{\boldsymbol{A}}$, $\boldsymbol{A}^{(\text{LR})}$, β^*, and $\phi^{(\text{LR})}$ we have

$$\begin{aligned}
\mathbb{E}\langle \hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle &= \mathbb{E}\left\langle \hat{\boldsymbol{A}}\left[\hat{\boldsymbol{A}}^{\dagger}\{\boldsymbol{A}^{(\text{LR})}\beta^* + \varepsilon + \phi^{(\text{LR})}\} - \beta^*\right], \varepsilon\right\rangle \\
&= \mathbb{E}\langle \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\boldsymbol{A}^{(\text{LR})}\beta^*, \varepsilon\rangle + \mathbb{E}\langle \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\varepsilon, \varepsilon\rangle + \mathbb{E}\langle \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\phi^{(\text{LR})}, \varepsilon\rangle - \mathbb{E}\langle \hat{\boldsymbol{A}}\beta^*, \varepsilon\rangle \\
&= \mathbb{E}\langle \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\varepsilon, \varepsilon\rangle.
\end{aligned}$$

Observe that $\mathbb{E}\langle \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\varepsilon, \varepsilon\rangle$ is a scalar. By properties of trace algebra, independence of ε from $\hat{\boldsymbol{A}}$, Assumption 5.5, and the fact that $\hat{\boldsymbol{A}}$ is rank k we obtain

$$\begin{aligned}
\mathbb{E}\langle \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\varepsilon, \varepsilon\rangle &= \mathbb{E}\left[trace\left(\varepsilon^T \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\varepsilon\right)\right] \\
&= \mathbb{E}\left[trace\left(\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\varepsilon\varepsilon^T\right)\right] \\
&= trace\left(\mathbb{E}\left[\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\right]\mathbb{E}\left[\varepsilon\varepsilon^T\right]\right) \\
&\leq \bar{\sigma}^2 trace\left(\mathbb{E}\left[\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\dagger}\right]\right) \\
&= \bar{\sigma}^2 k.
\end{aligned}$$

□

Remark E.10 (Dictionary)**.** *The generalization of Lemma E.10 is*

$$\mathbb{E}\langle b(\hat{\boldsymbol{A}})(\hat{\beta} - \beta^*), \varepsilon\rangle \leq \bar{\sigma}^2 r'.$$

Lemma E.11 (Modified Hoeffding Inequality; Lemma A.3 of Agarwal et al. (2020a))**.** *Let $X \in \mathbb{R}^n$ be random vector with independent mean-zero sub-Gaussian random coordinates with $\|X_i\|_{\psi_2} \leq K$. Let $a \in \mathbb{R}^n$ be another random vector that satisfies $\|a\|_2 \leq b$ almost surely for some constant $b \geq 0$. Then for all $t \geq 0$,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} a_i X_i\Big| \geq t\Big) \leq 2\exp\Big(-\frac{ct^2}{K^2 b^2}\Big),$$

where $c > 0$ is a universal constant.

Lemma E.12 (Modified Hanson-Wright Inequality; Lemma A.4 of Agarwal et al. (2020a))**.** *Let $X \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates with*

$\|X_i\|_{\psi_2} \le K$. *Let* $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ *be a random matrix satisfying* $\|\boldsymbol{B}\| \le a$ *and* $\|\boldsymbol{B}\|_{Fr}^2 \le b$ *almost surely for some* $a, b \ge 0$. *Then for any* $t \ge 0$,

$$\mathbb{P}\left(|X^T\boldsymbol{B}X - \mathbb{E}[X^T\boldsymbol{B}X]| \ge t\right) \le 2\cdot\exp\Big\{-c\min\Big(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\Big)\Big\}.$$

Lemma E.13. *Suppose Assumptions 5.1 and 5.5 hold, and that* $k = r$. *Given* $\hat{\boldsymbol{A}}$, *the following holds with probability at least* $1 - O\{1/(np)^{10}\}$ *with respect to randomness in* ε:

$$\begin{aligned}&\langle \hat{\boldsymbol{A}}(\hat\beta - \beta^*), \varepsilon\rangle \\ &\le C\bar\sigma^2 \ln(np)\left\{r + \|\phi^{(\mathrm{LR})}\|_2 + \|\beta^*\|_1(\sqrt{n}\bar A + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty})\right\}.\end{aligned}$$

Proof. We proceed in steps.

1. Decomposition

   We need to bound $\langle \hat{\boldsymbol{A}}(\hat\beta - \beta^*), \varepsilon\rangle$. To that end, we recall that $\hat\beta = \hat{\boldsymbol{V}}_k\hat{\boldsymbol{\Sigma}}_k^{-1}\hat{\boldsymbol{U}}_k^T Y$, $\hat{\boldsymbol{A}} = \hat{\boldsymbol{U}}_k\hat{\boldsymbol{\Sigma}}_k\hat{\boldsymbol{V}}_k^T$, and $Y = \boldsymbol{A}^{(\mathrm{LR})}\beta^* + \phi^{(\mathrm{LR})} + \varepsilon$. Thus,

   $$\hat{\boldsymbol{A}}\hat\beta = \hat{\boldsymbol{U}}_k\hat{\boldsymbol{\Sigma}}_k\hat{\boldsymbol{V}}_k^T\hat{\boldsymbol{V}}_k\hat{\boldsymbol{\Sigma}}_k^{-1}\hat{\boldsymbol{U}}_k^T Y = \hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\boldsymbol{A}^{(\mathrm{LR})}\beta^* + \hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\phi^{(\mathrm{LR})} + \hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\varepsilon.$$

   Therefore,

   $$\begin{aligned}&\langle \hat{\boldsymbol{A}}(\hat\beta - \beta^*), \varepsilon\rangle \\ &= \langle \hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\boldsymbol{A}^{(\mathrm{LR})}\beta^*, \varepsilon\rangle + \langle \hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\phi^{(\mathrm{LR})}, \varepsilon\rangle + \langle \hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\varepsilon, \varepsilon\rangle - \langle \hat{\boldsymbol{A}}\beta^*, \varepsilon\rangle. \qquad (30)\end{aligned}$$

2. First Hoeffding

   To obtain a high probability bound of the first term, we use Lemma E.11. Note that

   $$\|\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2 \le \|\boldsymbol{A}^{(\mathrm{LR})}\beta^*\|_2 \le \|\boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}\|\beta^*\|_1 \le 3\sqrt{n}\bar A\|\beta^*\|_1$$

   since $\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T$ is a projection matrix and $\|\boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty} \le 3\sqrt{n}\bar A$ due to Lemma C.4.

   It follows that for any $t > 0$

   $$\mathbb{P}\Big(\langle \hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\boldsymbol{A}^{(\mathrm{LR})}\beta^*, \varepsilon\rangle \ge t\Big) \le \exp\Big(-\frac{ct^2}{n\bar A^2\|\beta^*\|_1^2\bar\sigma^2}\Big). \qquad (31)$$

3. Second Hoeffding

To obtain a high probability bound of the second term, we use Lemma E.11. Note that

$$\|\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\phi^{(\text{LR})}\|_2 \le \|\phi^{(\text{LR})}\|_2$$

since $\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T$ is a projection matrix.

It follows that for any $t > 0$

$$\mathbb{P}\Big(\langle\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\phi^{(\text{LR})}, \varepsilon\rangle \ge t\Big) \le \exp\Big(-\frac{ct^2}{\|\phi^{(\text{LR})}\|_2^2\bar{\sigma}^2}\Big). \tag{32}$$

4. Third Hoeffding

To obtain a high probability bound of the fourth term, we use Lemma E.11. Note that

$$\begin{aligned}\|\hat{\boldsymbol{A}}\beta^*\|_2 &\le \|(\hat{\boldsymbol{A}} - \boldsymbol{A})\beta^*\|_2 + \|\boldsymbol{A}\beta^*\|_2 \\ &\le \big(\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty} + \|\boldsymbol{A}\|_{2,\infty}\big)\|\beta^*\|_1 \\ &\le \big(\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty} + \sqrt{n}\bar{A}\big)\|\beta^*\|_1\end{aligned}$$

since $\|\boldsymbol{A}\|_{2,\infty} \le \sqrt{n}\bar{A}$ due to Assumption 5.1.

Therefore, for any $t > 0$

$$\mathbb{P}\Big(\langle\hat{\boldsymbol{A}}\beta^*, \varepsilon\rangle \ge t\Big) \le \exp\left(-\frac{ct^2}{\bar{\sigma}^2(n\bar{A}^2 + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2)\|\beta^*\|_1^2}\right). \tag{33}$$

5. Hanson-Wright

To obtain a high probability bound of the third term, we use Lemma E.12. ε is independent of $\hat{\boldsymbol{U}}_k, \hat{\boldsymbol{\Sigma}}_k, \hat{\boldsymbol{V}}_k$ since $\hat{\boldsymbol{A}}$ is determined by $\boldsymbol{Z}$, which is independent of ε. As a result,

$$\begin{aligned}\mathbb{E}\big[\langle\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\varepsilon, \varepsilon\rangle\big] &= \mathbb{E}\big[\varepsilon^T\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\varepsilon\big] \\ &= \mathbb{E}\big[trace(\varepsilon^T\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\varepsilon)\big] \\ &= \mathbb{E}\big[trace(\varepsilon\varepsilon^T\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T)\big] \\ &= trace(\mathbb{E}\big[\varepsilon\varepsilon^T\big]\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T) \\ &\le trace(\bar{\sigma}^2\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T) \\ &= \bar{\sigma}^2 trace(\hat{\boldsymbol{U}}_k^T\hat{\boldsymbol{U}}_k) \\ &= \bar{\sigma}^2 k.\end{aligned} \tag{34}$$

Since $\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T$ is a projection matrix,

$$\|\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\| \le 1, \quad \|\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\|_{Fr}^2 = trace(\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T) = trace(\hat{\boldsymbol{U}}_k^T\hat{\boldsymbol{U}}_k) = k.$$

Finally, using Lemma E.12 and (34), it follows that for any $t > 0$

$$\mathbb{P}\left(\langle\hat{\boldsymbol{U}}_k\hat{\boldsymbol{U}}_k^T\varepsilon, \varepsilon\rangle \ge \bar{\sigma}^2 k + t\right) \le \exp\Big\{ - c\min\Big(\frac{t^2}{k\bar{\sigma}^4}, \frac{t}{\bar{\sigma}^2}\Big)\Big\}. \tag{35}$$

6. Simplification.

   Set the RHSs of (31), (32), (33), and (35) equal to $1/(np)^{10}$ and solve for t. Combining these results with (30),

$$\begin{aligned}
&\langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle \\
&\le \bar{\sigma}^2 r + C\bar{\sigma}\sqrt{\ln(np)}\left\{\bar{\sigma}\sqrt{r} + \bar{\sigma}\sqrt{\ln(np)} + \|\phi^{(\text{LR})}\|_2 + \|\beta^*\|_1(\sqrt{n}\bar{A} + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty})\right\} \\
&\le C\bar{\sigma}^2\ln(np)\left\{r + \|\phi^{(\text{LR})}\|_2 + \|\beta^*\|_1(\sqrt{n}\bar{A} + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty})\right\}.
\end{aligned}$$

□

Remark E.11 (Dictionary)**.** *The generalization of Lemma E.13 is*

$$\begin{aligned}
&\langle b(\hat{\boldsymbol{A}})(\hat{\beta} - \beta^*), \varepsilon\rangle \\
&\le C\bar{\sigma}^2\ln(np)\left\{r' + \|\phi^{(\text{LR})}\|_2 + \|\beta^*\|_1(\sqrt{n}\bar{A}' + \|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\|_{2,\infty})\right\}.
\end{aligned}$$

E.3.4 Collecting results

Lemma E.14. *Suppose the conditions of Theorem 5.1 hold. Further suppose Assumptions 5.5 and 5.6 hold. With probability at least* $1 - O\{(np)^{-10}\}$, $\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\beta} - \beta^*)\|_2^2 \le (2)$ *where*

$$\begin{aligned}
(2) = &\frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4}\frac{\bar{\sigma}^2\ln^3(np)}{\hat{s}_r^2} \\
&\cdot\left[\|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left\{\frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2} + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2\right\}\right].
\end{aligned}$$

Proof. By Lemma E.4

$$\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\beta} - \beta^*)\|_2^2 \le \frac{C}{\hat{s}_k^2}\Big\{\|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2\|\beta^*\|_1^2 \ \vee \ \|\phi^{(\text{LR})}\|_2^2 \ \vee \ \langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle\Big\}.$$

Note that

$$\|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \leq C\left\{\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2\right\}.$$

By Lemma E.13, with probability at least $1 - O\{(np)^{-10}\}$

$$\langle \hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle \leq C\bar{\sigma}^2 \ln(np)\left\{r + \|\phi^{(\text{LR})}\|_2 + \|\beta^*\|_1(\sqrt{n}\bar{A} + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty})\right\}.$$

Simplifying, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\beta} - \beta^*)\|_2^2 \\
&\quad \leq C\frac{\bar{\sigma}^2\ln(np)}{\hat{s}_r^2}\left[r + \ \|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left\{\sqrt{n}\bar{A}\frac{1}{\|\beta^*\|_1} + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2\right\}\right].
\end{aligned}$$

Note that we have taken $x \vee x^2 \leq x^2$ for $x \in \{\|\phi^{(\text{LR})}\|_2, \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}\|\beta^*\|_1\}$. To justify this simplification, one may consider redefining $\tilde{x} = x \vee 1$, since the objects taken as x are diverging.

By Lemma D.10 and Lemma D.15, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}\right\|_{2,\infty}^2 \\
&\leq \frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4}\left(r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right)\ln^2(np) + C\left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^2.
\end{aligned}$$

In summary, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\beta} - \beta^*)\|_2^2 \\
&\leq \frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4}\frac{\bar{\sigma}^2\ln^3(np)}{\hat{s}_r^2} \\
&\cdot\left[\ \|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left\{\frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2} + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2\right\}\right].
\end{aligned}$$

□

Lemma E.15. *Suppose the conditions of Lemma E.14 hold. With probability at least* $1 - O\{(np)^{-10}\}$, $\|\hat{\beta} - \beta^*\|_2^2 \leq (1) + (2)$ *where*

$$(1) = C\frac{\|\beta^*\|_2^2}{\rho_{\min}^2 s_r^2}\left\{(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\text{LR})}\|^2 + \bar{A}^2\ln(np)p\right\}$$

and (2) is defined above.

Proof. By Lemma E.5

$$\|\hat{\beta} - \beta^*\|_2^2$$
$$\leq C\left[\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T\|^2\|\beta^*\|_2^2 + \frac{1}{\hat{s}_k^2}\left\{\|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2\|\beta^*\|_1^2 \ \vee \ \|\phi^{(\text{LR})}\|_2^2 \ \vee \ \langle\hat{\boldsymbol{A}}(\hat{\beta} - \beta^*), \varepsilon\rangle\right\}\right].$$

By Lemma E.14, it suffices to focus on the first term, since the latter is bounded by (2). By Proposition D.10, Lemma E.8, and since $\delta \leq C$ and $\|\boldsymbol{A}^{(\text{LR})}\| \leq \bar{A}\sqrt{np}$, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\| &\leq C\frac{\delta}{\rho_{\min}s_r}\left\{(\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \sqrt{\frac{\ln(np)}{n}}\|\boldsymbol{A}^{(\text{LR})}\|\right\} \\ &\leq \frac{C}{\rho_{\min}s_r}\left\{(\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \bar{A}\sqrt{\ln(np)}\sqrt{p}\right\}.\end{aligned}$$

Simplifying, with probability at least $1 - O\{(np)^{-10}\}$

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C\frac{\|\beta^*\|_2^2}{\rho_{\min}^2 s_r^2}\left\{(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\text{LR})}\|^2 + \bar{A}^2\ln(np)p\right\} + (2)$$

□

Remark E.12 (Dictionary)**.** *In the generalization of Lemmas E.14 and E.15, note the new appearance of C_b' in (2):*

$$\begin{aligned}(1) &= C\frac{\|\beta^*\|_2^2}{(\rho'_{\min})^2 s_{r'}^2}\left[(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\text{LR})}\|^2 + \bar{A}^2\ln(np)\cdot p\right] \\ (2) &= \frac{CC_b'(K_a + \bar{K}\bar{A})^2}{(\rho'_{\min})^4}\frac{\bar{\sigma}^2\ln^3(np)}{\hat{s}_{r'}^2} \\ &\quad\cdot\Bigg[\ \|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\Bigg\{\frac{\sqrt{n}}{\|\beta^*\|_1} + r' + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2} \\ &\quad + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2 + \|b(\boldsymbol{A}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\Bigg\}\Bigg].\end{aligned}$$

Proof. For simplicity, we prove the result by focusing on $\|\hat{\beta} - \beta^*\|_2^2$. By Lemma E.6

$$\|\hat{\beta} - \beta^*\|_2^2$$
$$\leq C\left[\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T\|^2\|\beta^*\|_2^2 + \frac{1}{\hat{s}_{r'}^2}\left\{\|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\|\beta^*\|_1^2 \ \vee \ \|\phi^{(\text{LR})}\|_2^2 \ \vee \ \langle b(\hat{\boldsymbol{A}})(\hat{\beta} - \beta^*), \varepsilon\rangle\right\}\right].$$

By Remark E.8 and analogous algebra,

$$\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\| \leq \frac{C}{\rho'_{\min}s_{r'}}\left\{(\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\text{LR})}\| + \bar{A}\sqrt{\ln(np)}\sqrt{p}\right\}.$$

Note that

$$\|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2 \le C\left\{\|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\|_{2,\infty}^2 + \|b(\boldsymbol{A}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\right\}.$$

By Remark E.11, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\langle b(\hat{\boldsymbol{A}})(\hat{\beta} - \beta^*), \varepsilon\rangle \\
&\le C\bar{\sigma}^2 \ln(np)\left\{r' + \|\phi^{(\text{LR})}\|_2 + \|\beta^*\|_1(\sqrt{n}\bar{A}' + \|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\|_{2,\infty})\right\}.
\end{aligned}$$

Simplifying, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\|\hat{\beta} - \beta^*\|_2^2 \\
&\quad\le C\frac{\|\beta^*\|_2^2}{(\rho'_{\min})^2 s_{r'}^2}\left\{(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\text{LR})}\|^2 + \bar{A}^2\ln(np)p\right\} \\
&\quad + C\frac{\bar{\sigma}^2\ln(np)}{\hat{s}_{r'}^2}\left[r' + \ \|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left\{\sqrt{n}\bar{A}'\frac{1}{\|\beta^*\|_1} + \|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\|_{2,\infty}^2 + \|b(\boldsymbol{A}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\right\}\right].
\end{aligned}$$

By Lemma D.10 and Lemma D.15, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\left\|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\right\|_{2,\infty}^2 \\
&\le \frac{CC'_b(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4}\left(r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right)\ln^2(np) + C\left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^2.
\end{aligned}$$

In summary, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\|\hat{\beta} - \beta^*\|_2^2 \\
&\le C\frac{\|\beta^*\|_2^2}{(\rho'_{\min})^2 s_{r'}^2}\left\{(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\text{LR})}\|^2 + \bar{A}^2\ln(np)p\right\} \\
&+ \frac{CC'_b(K_a + \bar{K}\bar{A})^2}{(\rho'_{\min})^4}\frac{\bar{\sigma}^2\ln^3(np)}{\hat{s}_{r'}^2} \\
&\cdot\left[\ \|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left\{\frac{\sqrt{n}}{\|\beta^*\|_1} + r' + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right.\right. \\
&\quad\left.\left. + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2 + \|b(\boldsymbol{A}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\right\}\right].
\end{aligned}$$

□

Proposition E.2 (Projected TRAIN ERROR)**.** *Suppose conditions of Theorem 5.1 hold. Further suppose Assumptions 5.5 and 5.6 hold. Let $k = r$ and*

$$\rho_{\min} \gg \tilde{C}\sqrt{r}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right), \quad \tilde{C} := C\bar{A}\Big(\kappa + \bar{K} + K_a\Big).$$

Then with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\|\hat{\boldsymbol{V}}_r \hat{\boldsymbol{V}}_r^T (\hat{\beta} - \beta^*)\|_2^2 \\
&\leq C\bar{A}^4 (K_a + \bar{K})^2 (\kappa + \bar{K} + K_a)^2 \frac{\bar{\sigma}^2}{\rho_{\min}^4} \cdot r \ln^6(np) \\
&\quad \cdot \left\{ \frac{1}{np} \|\phi^{(\mathrm{LR})}\|_2^2 + \|\beta^*\|_1^2 \left(\frac{\sqrt{n}}{\|\beta^*\|_1 np} + \frac{r}{np} + \frac{r}{p^2} + \frac{r}{p} \Delta_E^2 \right) \right\}.
\end{aligned}$$

Proof. We proceed in steps.

1. Recall the inequalities

$$s_r^2 \geq C\frac{np}{r}, \quad \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|^2 \leq np\Delta_E^2, \quad \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|_{2,\infty}^2 \leq n\Delta_E^2.$$

Further,

$$\Delta_{H,op}^2 \leq C \cdot \bar{A}^2 (\kappa + \bar{K} + K_a)^2 \ln^3(np).$$

Moreover, $(n+p)\Delta_{H,op}^2$ dominates $\ln(np) p \bar{A}^2$.

2. Simplifying the RHS of the bound in Lemma E.14

$$\begin{aligned}
(2) &:= \frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4} \frac{\bar{\sigma}^2 \ln^3(np)}{\hat{s}_r^2} \\
&\quad \cdot \left[\|\phi^{(\mathrm{LR})}\|_2^2 + \|\beta^*\|_1^2 \left\{ \frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{n(n+p)\Delta_{H,op}^2 + n \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|^2 + \ln(np) np \bar{A}^2}{s_r^2} + \|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^2 \right\} \right].
\end{aligned}$$

Bounding its latter factor

$$\begin{aligned}
&\|\phi^{(\mathrm{LR})}\|_2^2 + \|\beta^*\|_1^2 \left\{ \frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{n(n+p)\Delta_{H,op}^2 + n \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|^2 + \ln(np) np \bar{A}^2}{s_r^2} + \|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^2 \right\} \\
&\leq \|\phi^{(\mathrm{LR})}\|_2^2 + \|\beta^*\|_1^2 \left\{ \frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{n(n+p)\Delta_{H,op}^2 + n^2 p \Delta_E^2}{s_r^2} + n\Delta_E^2 \right\} \\
&\leq \|\phi^{(\mathrm{LR})}\|_2^2 + \|\beta^*\|_1^2 \left\{ \frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{r}{p}(n+p)\Delta_{H,op}^2 + rn\Delta_E^2 \right\}.
\end{aligned}$$

Hence

$$\begin{aligned}
(2) &\leq C\bar{A}^4 (K_a + \bar{K})^2 (\kappa + \bar{K} + K_a)^2 \ln^6(np) \frac{\bar{\sigma}^2}{\rho_{\min}^4 \hat{s}_r^2} \\
&\quad \cdot \left\{ \|\phi^{(\mathrm{LR})}\|_2^2 + \|\beta^*\|_1^2 \left(\frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{rn}{p} + rn\Delta_E^2 \right) \right\}.
\end{aligned}$$

3. By Lemma E.9,

$$\hat{s}_r^2 \gtrsim s_r^2 \geq C\frac{np}{r}$$

so as long as the regularity condition holds, we can further bound

$$\begin{aligned}
&\frac{1}{\hat{s}_r^2}\left\{ \|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{rn}{p} + rn\Delta_E^2\right)\right\} \\
&\leq r\left\{ \frac{1}{np}\|\phi^{(\text{LR})}\|_2^2 + \frac{1}{np}\|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1} + r + \frac{rn}{p} + rn\Delta_E^2\right)\right\} \\
&= r\left\{ \frac{1}{np}\|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1 np} + \frac{r}{np} + \frac{r}{p^2} + \frac{r}{p}\Delta_E^2\right)\right\}.
\end{aligned}$$

In summary

$$\begin{aligned}
(2) \leq\ & C\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2\bar{\sigma}^2\frac{r\ln^6(np)}{\rho_{\min}^4} \\
&\left\{ \frac{1}{np}\|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1 np} + \frac{r}{np} + \frac{r}{p^2} + \frac{r}{p}\Delta_E^2\right)\right\}.
\end{aligned}$$

□

Remark E.13 (Dictionary)**.** *The generalization of Proposition E.2 is as follows. Suppose*

$$\rho'_{\min} \gg \tilde{C}\sqrt{r'}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right).$$

Then with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\|\hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T(\hat{\beta}-\beta^*)\|_2^2 \\
&\leq CC'_b\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2\frac{\bar{\sigma}^2}{(\rho'_{\min})^4}\cdot r'\ln^6(np) \\
&\cdot\left[\frac{1}{np}\|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_1^2\left\{\frac{\sqrt{n}}{\|\beta^*\|_1 np} + \frac{r'}{np} + \frac{r'}{p^2} + \frac{r'}{p}(\Delta'_E)^2\right\}\right].
\end{aligned}$$

Proof. The only updates are to $(\rho_{\min}^{-1}, r, \Delta_E)$. □

Proposition E.3 (TRAIN ERROR)**.** *Suppose conditions of Proposition E.2 hold. Then with probability at least* $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\|\hat{\beta}-\beta^*\|_2^2 \\
&\leq C\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2\frac{\bar{\sigma}^2}{\rho_{\min}^4}\cdot r\ln^6(np)\cdot\left\{ \frac{1}{np}\|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_2^2\left(\frac{r}{n}+\frac{r}{p}+r\Delta_E^2\right)\right\}.
\end{aligned}$$

Proof. We proceed in steps.

1. Recall the inequalities

$$s_r^2 \geq C\frac{np}{r}, \quad \left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|^2 \leq np\Delta_E^2, \quad \left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|_{2,\infty}^2 \leq n\Delta_E^2.$$

Further,

$$\Delta_{H,op}^2 \leq C \cdot \bar{A}^2(\kappa + \bar{K} + K_a)^2 \ln^3(np).$$

Moreover, $(n+p)\Delta_{H,op}^2$ dominates $\ln(np)p\bar{A}^2$.

2. Simplifying the first term on the RHS of the bound in Lemma E.15

$$\begin{aligned}
(1) &:= C\frac{\|\beta^*\|_2^2}{\rho_{\min}^2 s_r^2}\left\{(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\mathrm{LR})}\|^2 + \bar{A}^2\ln(np)p\right\} \\
&\leq C \cdot \bar{A}^2(\kappa + \bar{K} + K_a)^2 \ln^3(np)\frac{\|\beta^*\|_2^2}{\rho_{\min}^2 s_r^2}\left(n + p + np\Delta_E^2\right) \\
&\leq C \cdot \bar{A}^2(\kappa + \bar{K} + K_a)^2\frac{\|\beta^*\|_2^2}{\rho_{\min}^2} \cdot r\ln^3(np) \cdot \left(\frac{1}{p} + \frac{1}{n} + \Delta_E^2\right).
\end{aligned}$$

3. We have shown, by the arguments above and in the proof of Proposition E.2

$$\begin{aligned}
(1) &\leq C \cdot \bar{A}^2(\kappa + \bar{K} + K_a)^2\frac{\|\beta^*\|_2^2}{\rho_{\min}^2} \cdot r\ln^3(np) \cdot \left(\frac{1}{p} + \frac{1}{n} + \Delta_E^2\right) \\
(2) &\leq C\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2\bar{\sigma}^2\frac{r\ln^6(np)}{\rho_{\min}^4} \\
&\quad \left\{\frac{1}{np}\|\phi^{(\mathrm{LR})}\|_2^2 + \|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1 np} + \frac{r}{np} + \frac{r}{p^2} + \frac{r}{p}\Delta_E^2\right)\right\}.
\end{aligned}$$

We further bound (2), and argue that this further bound dominates (1). Consider the factor

$$\begin{aligned}
&\|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1 np} + \frac{r}{np} + \frac{r}{p^2} + \frac{r}{p}\Delta_E^2\right) \\
&= \|\beta^*\|_1\frac{1}{\sqrt{np}} + \|\beta^*\|_1^2\left(\frac{r}{np} + \frac{r}{p^2} + \frac{r}{p}\Delta_E^2\right) \\
&\leq \|\beta^*\|_2\frac{1}{\sqrt{np}} + \|\beta^*\|_2^2\left(\frac{r}{n} + \frac{r}{p} + r\Delta_E^2\right) \\
&\leq \|\beta^*\|_2^2\left(\frac{r}{n} + \frac{r}{p} + r\Delta_E^2\right)
\end{aligned}$$

where the last line uses $\|\beta^*\|_2 \leq \|\beta^*\|_2^2$ and $\frac{1}{\sqrt{np}} \leq \frac{1}{\min(n,p)} \leq \frac{r}{\min(n,p)} = \frac{r}{n} \vee \frac{r}{p}$. In summary, the further bound is

$$(2) \leq C\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2\bar{\sigma}^2\frac{r\ln^6(np)}{\rho_{\min}^4}\left\{\frac{1}{np}\|\phi^{(\mathrm{LR})}\|_2^2 + \|\beta^*\|_2^2\left(\frac{r}{n} + \frac{r}{p} + r\Delta_E^2\right)\right\}.$$

Clearly this further bound on (2) dominates (1).

□

Remark E.14 (Dictionary)**.** *In the generalization of Proposition E.3,*

$$\rho'_{\min} \gg \tilde{C}\sqrt{r'}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right)$$

and

$$\begin{aligned}&\|\hat{\beta} - \beta^*\|_2^2 \\ &\leq CC'_b\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2\frac{\bar{\sigma}^2}{(\rho'_{\min})^4} \cdot r'\ln^6(np) \cdot \Big[\ \frac{1}{np}\|\phi^{(\text{LR})}\|_2^2 + \|\beta^*\|_2^2\left\{\frac{r'}{n} + \frac{r'}{p} + r'(\Delta'_E)^2\right\}\Big].\end{aligned}$$

Proof. The only updates are to $(\rho_{\min}^{-1}, r, \Delta_E)$. □

For completeness, we state a corollary of Propositions E.2 and E.3 above.

Corollary E.1 (Population projected TRAIN ERROR)**.** *Suppose conditions of Proposition E.2 hold. Then with probability at least* $1 - O\{(np)^{-10}\}$

$$\begin{aligned}&\|\boldsymbol{V}_r\boldsymbol{V}_r^T(\hat{\beta} - \beta^*)\|_2^2 \\ &\leq C\bar{A}^6(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^4\frac{\bar{\sigma}^2}{\rho_{\min}^6} \cdot r\ln^9(np)\{(i) + (ii) + (iii)\}\end{aligned}$$

where

$$\begin{aligned}(i) &= \frac{1}{n}\|\phi^{(\text{LR})}\|_2^2\left(\frac{1}{p} + \frac{r}{p^2} + \frac{r}{np} + \frac{r}{p}\Delta_E^2\right) \\ (ii) &= \|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1 np} + \frac{r}{np} + \frac{r}{p^2} + \frac{r}{p}\Delta_E^2\right) \\ (iii) &= \|\beta^*\|_2^2 \cdot r^2\left\{\frac{1}{n^2} + \frac{1}{p^2} + \frac{1}{np} + \left(\frac{1}{n} + \frac{1}{p}\right)\Delta_E^2 + \Delta_E^4\right\}.\end{aligned}$$

Proof. We proceed in steps.

1. Decomposition

   To begin, write

$$\begin{aligned}\|\boldsymbol{V}_r\boldsymbol{V}_r^T(\hat{\beta} - \beta^*)\|_2^2 &\leq 2\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\beta} - \beta^*)\|_2^2 + 2\|(\boldsymbol{V}_r\boldsymbol{V}_r^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T)(\hat{\beta} - \beta^*)\|_2^2 \\ &\leq 2\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\beta} - \beta^*)\|_2^2 + 2\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T - \boldsymbol{V}_r\boldsymbol{V}_r^T\|^2 \cdot \|\hat{\beta} - \beta^*\|_2^2.\end{aligned}$$

2. First term

   By Proposition E.2

$$\begin{aligned}
&\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\beta}-\beta^*)\|_2^2 \\
&\leq C\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2\frac{\bar{\sigma}^2}{\rho_{\min}^4}\cdot r\ln^6(np) \\
&\quad\cdot\left\{\frac{1}{np}\|\phi^{(\mathrm{LR})}\|_2^2+\|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1 np}+\frac{r}{np}+\frac{r}{p^2}+\frac{r}{p}\Delta_E^2\right)\right\}.
\end{aligned}$$

3. Second term

   By Proposition E.3

$$\begin{aligned}
&\|\hat{\beta}-\beta^*\|_2^2 \\
&\leq C\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2\frac{\bar{\sigma}^2}{\rho_{\min}^4}\cdot r\ln^6(np)\cdot\left\{\frac{1}{np}\|\phi^{(\mathrm{LR})}\|_2^2+\|\beta^*\|_2^2\left(\frac{r}{n}+\frac{r}{p}+r\Delta_E^2\right)\right\}.
\end{aligned}$$

   Moreover, by arguments in the proofs of Lemma E.15 and Proposition E.3,

$$\begin{aligned}
\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T-\boldsymbol{V}_r\boldsymbol{V}_r^T\|^2 &\leq \frac{(1)}{\|\beta^*\|_2^2} \\
&\leq C\cdot\bar{A}^2(\kappa+\bar{K}+K_a)^2\frac{1}{\rho_{\min}^2}\cdot r\ln^3(np)\cdot\left(\frac{1}{p}+\frac{1}{n}+\Delta_E^2\right) \\
&= C\cdot\bar{A}^2(\kappa+\bar{K}+K_a)^2\frac{1}{\rho_{\min}^2}\cdot\ln^3(np)\cdot\left(\frac{r}{p}+\frac{r}{n}+\Delta_E^2\right).
\end{aligned}$$

   Therefore the coefficient of the product is

$$C^*=C\bar{A}^6(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^4\frac{\bar{\sigma}^2}{\rho_{\min}^6}\cdot r\ln^9(np)$$

   which dominates the coefficient of the first term. The substantive factor in the product is

$$\begin{aligned}
&\left\{\frac{1}{np}\|\phi^{(\mathrm{LR})}\|_2^2+\|\beta^*\|_2^2\left(\frac{r}{n}+\frac{r}{p}+r\Delta_E^2\right)\right\}\cdot\left(\frac{r}{p}+\frac{r}{n}+r\Delta_E^2\right) \\
&=\frac{1}{np}\|\phi^{(\mathrm{LR})}\|_2^2\left(\frac{r}{p}+\frac{r}{n}+r\Delta_E^2\right)+\|\beta^*\|_2^2\left(\frac{r}{n}+\frac{r}{p}+r\Delta_E^2\right)^2.
\end{aligned}$$

4. Collecting results

   We have established that the coefficient will be C^*. What remains is to determine the dominating terms of

$$\frac{1}{np}\|\phi^{(\mathrm{LR})}\|_2^2+\|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1 np}+\frac{r}{np}+\frac{r}{p^2}+\frac{r}{p}\Delta_E^2\right)$$

and

$$\frac{1}{np}\|\phi^{(\text{LR})}\|_2^2\left(\frac{r}{p}+\frac{r}{n}+r\Delta_E^2\right)+\|\beta^*\|_2^2\left(\frac{r}{n}+\frac{r}{p}+r\Delta_E^2\right)^2.$$

Within the final term,

$$\begin{aligned}\left(\frac{r}{n}+\frac{r}{p}+r\Delta_E^2\right)^2 &= r^2\left(\frac{1}{n}+\frac{1}{p}+\Delta_E^2\right)^2. \\ &\leq Cr^2\left\{\frac{1}{n^2}+\frac{1}{p^2}+\frac{1}{np}+\left(\frac{1}{n}+\frac{1}{p}\right)\Delta_E^2+\Delta_E^4\right\}.\end{aligned}$$

In summary, the bound is

$$\|\boldsymbol{V}_r\boldsymbol{V}_r^T(\hat{\beta}-\beta^*)\|_2^2 \leq C^*\{(i)+(ii)+(iii)\}$$

where

$$\begin{aligned}(i) &= \frac{1}{n}\|\phi^{(\text{LR})}\|_2^2\left(\frac{1}{p}+\frac{r}{p^2}+\frac{r}{np}+\frac{r}{p}\Delta_E^2\right) \\ (ii) &= \|\beta^*\|_1^2\left(\frac{\sqrt{n}}{\|\beta^*\|_1 np}+\frac{r}{np}+\frac{r}{p^2}+\frac{r}{p}\Delta_E^2\right) \\ (iii) &= \|\beta^*\|_2^2\cdot r^2\left\{\frac{1}{n^2}+\frac{1}{p^2}+\frac{1}{np}+\left(\frac{1}{n}+\frac{1}{p}\right)\Delta_E^2+\Delta_E^4\right\}.\end{aligned}$$

□

E.4 Test error

E.4.1 Decomposition

Lemma E.16. *Let Assumption 5.6 hold. Let k, the PCA hyperparameter, equal $r = rank(\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}) = rank(\boldsymbol{A}^{(\text{LR}),\text{TEST}})$. Then,*

$$\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}-\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2 \leq C\sum_{m=1}^{3}\Delta_m$$

where

$$\begin{aligned}\Delta_1 &:= \left\{\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}-\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2+\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2\|\boldsymbol{V}\boldsymbol{V}^T-\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\|^2\right\}\|\hat{\beta}-\beta^*\|_2^2 \\ \Delta_2 &:= \frac{\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2}{\hat{s}_r^2}\left\{\|\hat{\boldsymbol{A}}^{\text{TRAIN}}-\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\|_{2,\infty}^2\|\beta^*\|_1^2 \;\vee\; \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2 \;\vee\; \langle\hat{\boldsymbol{A}}^{\text{TRAIN}}(\hat{\beta}-\beta^*),\varepsilon\rangle\right\} \\ \Delta_3 &:= \|\hat{\boldsymbol{A}}^{\text{TEST}}-\boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2\|\beta^*\|_1^2.\end{aligned}$$

Proof. Consider

$$\begin{aligned}\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2 &= \|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\beta^* + \hat{\boldsymbol{A}}^{\text{TEST}}\beta^* - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2 \\ &\leq 2\|\hat{\boldsymbol{A}}^{\text{TEST}}\big(\hat{\beta} - \beta^*\big)\|_2^2 + 2\|(\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}})\beta^*\|_2^2. \end{aligned} \tag{36}$$

We shall bound the two terms on the right hand side of (36) next. To analyze $\|\hat{\boldsymbol{A}}^{\text{TEST}}\big(\hat{\beta} - \beta^*\big)\|_2^2$, we proceed in steps

1. Decomposition

   Write

   $$\begin{aligned}\|\hat{\boldsymbol{A}}^{\text{TEST}}\big(\hat{\beta} - \beta^*\big)\|_2^2 &= \|\big\{\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{(\text{LR}),\text{TEST}} + \boldsymbol{A}^{(\text{LR}),\text{TEST}}\big\}\big(\hat{\beta} - \beta^*\big)\|_2^2 \\ &\leq 2\|\big\{\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\big\}\big(\hat{\beta} - \beta^*\big)\|_2^2 + 2\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\big(\hat{\beta} - \beta^*\big)\|_2^2.\end{aligned}$$

   We analyze the former and latter term separately.

2. Former term

   Note that $\|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\|$ is the $(r+1)$-st largest singular value of $\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}$. Therefore, by Weyl's inequality (Lemma E.6), we have

   $$\|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\| = \hat{s}'_{r+1} = \hat{s}'_{r+1} - s'_{r+1} \leq \|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|.$$

   In turn, this gives

   $$\begin{aligned}\|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\| &\leq \|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\| + \|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\| \\ &\leq 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|.\end{aligned}$$

   Thus, we have

   $$\begin{aligned}\|\big\{\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\big\}\big(\hat{\beta} - \beta^*\big)\|_2^2 &\leq \|\boldsymbol{A}^{(\text{LR}),\text{TEST}} - \hat{\boldsymbol{A}}^{\text{TEST}}\|^2 \cdot\ \|\hat{\beta} - \beta^*\|_2^2 \\ &\leq 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot\ \|\hat{\beta} - \beta^*\|_2^2.\end{aligned}$$

3. Latter term

   Recall that $\boldsymbol{V}$ and $\boldsymbol{V}_\perp$ span the rowspace and nullspace of $\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}$, respectively. By Assumption 5.6, it follows that $(\boldsymbol{V}')^T\boldsymbol{V}_\perp = 0$ and hence $\boldsymbol{A}^{(\text{LR}),\text{TEST}}\boldsymbol{V}_\perp\boldsymbol{V}_\perp^T = 0$. As

a result,

$$\begin{aligned}\|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\big(\hat{\beta}-\beta^*\big)\|_2^2 &= \|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}(\boldsymbol{V}\boldsymbol{V}^T+\boldsymbol{V}_\perp\boldsymbol{V}_\perp^T)\big(\hat{\beta}-\beta^*\big)\|_2^2\\ &= \|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\boldsymbol{V}\boldsymbol{V}^T\big(\hat{\beta}-\beta^*\big)\|_2^2\\ &\le \|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\|^2\ \|\boldsymbol{V}\boldsymbol{V}^T\big(\hat{\beta}-\beta^*\big)\|_2^2.\end{aligned}$$

Recalling that $\hat{\boldsymbol{V}}_r$ denotes the top r right singular vectors of $\boldsymbol{Z}^{\mathrm{TRAIN}}\,\hat{\boldsymbol{\rho}}^{-1}$, consider

$$\begin{aligned}\|\boldsymbol{V}\boldsymbol{V}^T\big(\hat{\beta}-\beta^*\big)\|_2^2 &= \|\big(\boldsymbol{V}\boldsymbol{V}^T-\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T+\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\big)\big(\hat{\beta}-\beta^*\big)\|_2^2\\ &\le 2\|\boldsymbol{V}\boldsymbol{V}^T-\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\|^2\ \|\hat{\beta}-\beta^*\|_2^2+2\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\big(\hat{\beta}-\beta^*\big)\|_2^2.\end{aligned}$$

Recall from (28) that

$$\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\beta}-\beta^*)\|_2^2 \le \frac{C}{\hat{s}_r^2}\Big\{\|\hat{\boldsymbol{A}}^{\mathrm{TRAIN}}-\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TRAIN}}\|_{2,\infty}^2\|\beta^*\|_1^2\vee\|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2\vee\langle\hat{\boldsymbol{A}}^{\mathrm{TRAIN}}(\hat{\beta}-\beta^*),\varepsilon\rangle\Big\}.$$

Therefore

$$\begin{aligned}&\|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\big(\hat{\beta}-\beta^*\big)\|_2^2\\ &\le C\|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\|^2\|\boldsymbol{V}\boldsymbol{V}^T-\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\|^2\ \|\hat{\beta}-\beta^*\|_2^2\\ &\quad+\frac{C\|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\|^2}{\hat{s}_r^2}\Big\{\|\hat{\boldsymbol{A}}^{\mathrm{TRAIN}}-\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TRAIN}}\|_{2,\infty}^2\|\beta^*\|_1^2\ \vee\ \|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2\ \vee\ \langle\hat{\boldsymbol{A}}^{\mathrm{TRAIN}}(\hat{\beta}-\beta^*),\varepsilon\rangle\Big\}.\end{aligned}$$

Finally, to analyze $\|(\hat{\boldsymbol{A}}^{\mathrm{TEST}}-\boldsymbol{A}^{\mathrm{TEST}})\beta^*\|_2^2$, we appeal to matrix Holder:

$$\|(\hat{\boldsymbol{A}}^{\mathrm{TEST}}-\boldsymbol{A}^{\mathrm{TEST}})\beta^*\|_2^2 \le \|\hat{\boldsymbol{A}}^{\mathrm{TEST}}-\boldsymbol{A}^{\mathrm{TEST}}\|_{2,\infty}^2\|\beta^*\|_1^2.$$

□

Remark E.15 (Dictionary)**.** *Let Assumption 5.6 hold.Let* $r' = rank[b\{\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TRAIN}}\}] = rank[b\{\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\}]$*. Then,*

$$\|b(\hat{\boldsymbol{A}}^{\mathrm{TEST}})\hat{\beta}-b(\boldsymbol{A}^{\mathrm{TEST}})\beta^*\|_2^2 \le C\sum_{m=1}^{3}\Delta_m$$

where

$$\Delta_1 := \Big[\{\bar{A}^{d_{\max}}d_{\max}\|\boldsymbol{Z}^{\mathrm{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}-\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\|\}^2+\|b\{\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\}\|^2\|\boldsymbol{V}\boldsymbol{V}^T-\hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T\|^2\Big]\ \|\hat{\beta}-\beta^*\|_2^2$$

$$\Delta_2 := \frac{\|b\{\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\}\|^2}{\hat{s}_{r'}^2}\Big[\|b(\hat{\boldsymbol{A}}^{\mathrm{TRAIN}})-b\{\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TRAIN}}\}\|_{2,\infty}^2\|\beta^*\|_1^2\ \vee\ \|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2\ \vee\ \langle b(\hat{\boldsymbol{A}}^{\mathrm{TRAIN}})(\hat{\beta}-\beta^*),\varepsilon\rangle\Big]$$

$$\Delta_3 := \|b(\hat{\boldsymbol{A}}^{\mathrm{TEST}})-b(\boldsymbol{A}^{\mathrm{TEST}})\|_{2,\infty}^2\|\beta^*\|_1^2.$$

Proof. The generalized analysis of the former term in Δ_1 is similar to the proof of Remark E.7.

□

E.4.2 High probability events

Define the following events

$$\tilde{\mathcal{E}}_1 \coloneqq \left\{ \|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2, \|\hat{\boldsymbol{A}}^{\text{TRAIN}} - \boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\|_{2,\infty}^2, \leq \tilde{\Delta}_1 \right\}, \quad \tilde{\Delta}_1 \coloneqq C_1 \cdot \frac{r \ln^5(np)}{\rho_{\min}^4} \left(1 + \frac{n}{p} + n\Delta_E^2 \right);$$

$$\tilde{\mathcal{E}}_2 \coloneqq \left\{ \|\boldsymbol{Z}^{\text{TEST}} \hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \leq \tilde{\Delta}_2 \right\}, \quad \tilde{\Delta}_2 \coloneqq C\bar{A}^2(\kappa + \bar{K} + K_a)^2 \frac{\ln^3(np)}{\rho_{\min}^2} \left(n + p + np\Delta_E^2 \right);$$

$$\tilde{\mathcal{E}}_3 \coloneqq \left\{ \|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r \hat{\boldsymbol{V}}_r^T\|^2, \|\boldsymbol{V}'(\boldsymbol{V}')^T - \hat{\boldsymbol{V}}'_r(\hat{\boldsymbol{V}}'_r)^T\|^2 \leq \tilde{\Delta}_3 \right\}, \quad \tilde{\Delta}_3 \coloneqq \frac{r}{np}\tilde{\Delta}_2;$$

$$\tilde{\mathcal{E}}_4 \coloneqq \{\hat{s}_r \gtrsim s_r\};$$

$$\tilde{\mathcal{E}} \coloneqq \cap_{i=1}^4 \tilde{\mathcal{E}}_i;$$

where

$$C_1 = C\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2.$$

Lemma E.17. *Let the conditions of Theorem 5.1 hold. Then $\tilde{\mathcal{E}}_1$ occurs with probability at least $O\{1 - 1/(np)^{10}\}$.*

Proof. By triangle inequality

$$\|\hat{\boldsymbol{A}}^{\text{TRAIN}} - \boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\|_{2,\infty}^2 \leq 2\|\hat{\boldsymbol{A}}^{\text{TRAIN}} - \boldsymbol{A}^{\text{TRAIN}}\|_{2,\infty}^2 + 2n\|\boldsymbol{E}^{(\text{LR}),\text{TRAIN}}\|_{\max}^2. \tag{37}$$

By Lemma D.10 and Lemma D.16 we have the desired result. Note that $\|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2$ behaves like the first term. As remarked earlier, the results in Appendix D are invariant to $\hat{\boldsymbol{\rho}}^{\text{TRAIN}}$ or $\hat{\boldsymbol{\rho}}^{\text{TEST}}$ so they hold for both $\hat{\boldsymbol{A}}^{\text{TRAIN}}$ and $\hat{\boldsymbol{A}}^{\text{TEST}}$ despite asymmetric definitions. □

Remark E.16 (Dictionary)**.** *A generalization Lemma E.17 sufficient for our subsequent analysis is for*

$$\tilde{\mathcal{E}}'_1 \coloneqq \left[\|b(\hat{\boldsymbol{A}}^{\text{TEST}}) - b(\boldsymbol{A}^{\text{TEST}})\|_{2,\infty}^2, \|b(\hat{\boldsymbol{A}}^{\text{TRAIN}}) - b\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\}\|_{2,\infty}^2, \leq \tilde{\Delta}'_1 \right]$$

where

$$\tilde{\Delta}'_1 \coloneqq C'_b C_1 \cdot \frac{r \ln^5(np)}{\rho_{\min}^4} \left\{ 1 + \frac{n}{p} + n(\Delta'_E)^2 \right\}.$$

Proof. Immediate from Assumption C.1. □

Lemma E.18. *Let the conditions of Theorem 5.1 hold. Then $\tilde{\mathcal{E}}_2$ occurs with probability at least $O\{1 - 1/(np)^{10}\}$.*

Proof. From Lemma D.10 and Lemma D.12 with probability at least $O\{1-1/(np)^{10}\}$

$$\begin{aligned}\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}-\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 &\leq C\frac{\delta^2}{\rho_{\min}^2}\left\{(n+p)\Delta_{H,op}^2+\|\boldsymbol{E}^{(\text{LR}),\text{TEST}}\|^2+\frac{\ln(np)}{n}\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2\right\}\\ &\leq C\frac{1}{\rho_{\min}^2}\left\{\bar{A}^2(\kappa+\bar{K}+K_a)^2(n+p)\ln^3(np)+np\Delta_E^2+\frac{\ln(np)}{n}np\bar{A}^2\right\}\\ &\leq C\bar{A}^2(\kappa+\bar{K}+K_a)^2\frac{\ln^3(np)}{\rho_{\min}^2}\left(n+p+np\Delta_E^2\right).\end{aligned}$$

As remarked above, Lemma D.12 is invariant to $\hat{\boldsymbol{\rho}}^{\text{TRAIN}}$ or $\hat{\boldsymbol{\rho}}^{\text{TEST}}$. □

Remark E.17 (Dictionary)**.** *The generalization of Lemma E.18 is for*

$$\tilde{\mathcal{E}}_2 \coloneqq \left[\{\bar{A}^{d_{\max}}d_{\max}\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}-\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|\}^2 \leq \tilde{\Delta}_2'\right]$$

where

$$\tilde{\Delta}_2' \coloneqq C\bar{A}^2(\kappa+\bar{K}+K_a)^2\frac{\ln^3(np)}{(\rho_{\min}')^2}\left(n+p+np\Delta_E^2\right).$$

Proof. See the proof of Remark E.7. □

Lemma E.19. *Let the conditions of Theorem 5.1 hold. Then $\tilde{\mathcal{E}}_3$ occurs with probability at least $O\{1-1/(np)^{10}\}$.*

Proof. By Lemma D.10 and Wedin's $\sin\Theta$ Theorem (Davis and Kahan, 1970; Wedin, 1972),

$$\begin{aligned}\|\boldsymbol{V}\boldsymbol{V}^T-\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\| &\leq \frac{\|\boldsymbol{Z}^{\text{TRAIN}}\,\hat{\boldsymbol{\rho}}^{-1}-\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\|}{s_r},\\ \|\boldsymbol{V}'(\boldsymbol{V}')^T-\hat{\boldsymbol{V}}_r'(\hat{\boldsymbol{V}}_r')^T\| &\leq \frac{\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}-\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|}{s_r'}.\end{aligned}$$

Recall that the argument in Lemma D.12 is invariant to $\hat{\boldsymbol{\rho}}^{\text{TRAIN}}$ or $\hat{\boldsymbol{\rho}}^{\text{TEST}}$. Simplifying as in Lemma E.18, we have the desired result. □

Remark E.18 (Dictionary)**.** *The generalization of Lemma E.19 is about the SVDs in Remark E.1 and $\tilde{\Delta}_3' = \frac{r'}{np}\tilde{\Delta}_2'$.*

Proof. Immediate from Remark E.8 and the extended version of Assumption 5.4. □

Lemma E.20. *Let the conditions of Lemma E.9 hold. Then $\tilde{\mathcal{E}}_4$ occurs with probability at least $O\{1-1/(np)^{10}\}$.*

Proof. Immediate from Lemma E.9. □

Remark E.19 (Dictionary)**.** *The generalization of Lemma E.20 concerns the SVDs in Remark E.1*

Proof. Immediate from Remark E.9. □

Lemma E.21. *Suppose the conditions of Theorem 5.1 hold and*

$$\rho_{\min} \gg \tilde{C}\sqrt{r}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right), \quad \tilde{C} := C\bar{A}\Big(\kappa + \bar{K} + K_a\Big).$$

Then $\mathbb{P}(\tilde{\mathcal{E}}^c) \leq \frac{C}{n^{10}p^{10}}$

Proof. Immediate from Lemmas E.17, E.18, E.19, and E.20 and the union bound. □

Remark E.20 (Dictionary)**.** *The generalization of Lemma E.21 instead uses the condition*

$$\rho'_{\min} \gg \tilde{C}\sqrt{r'}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right).$$

E.4.3 Simplification

Remark E.21 (Dictionary)**.** *The following lemmas are algebraic and generalize in the obvious way: replace* $(\rho_{\min}, r, \Delta_E)$ *with* $(\rho'_{\min}, r', \Delta'_E)$*. We therefore skip the remarks until Proposition E.4. The only subtleties are the presence of* C'_b *as a multiplier of* C_1 *and factors of* $\bar{A}'$ *replacing some factors of* $\bar{A}$*.*

Lemma E.22. *Let the conditions of Proposition E.3 hold. Then,*

$$\begin{aligned}
&\mathbb{E}[\Delta_1 \mid \tilde{\mathcal{E}}] \\
&\leq C_1C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^2\ln^8(np)}{\rho_{\min}^6}\left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right) \cdot \left\{\left(n + p + np\Delta_E^2\right)\|\beta^*\|_2^2 + \left(r + \frac{rn}{p} + rn\Delta_E^2\right)\|\beta^*\|_1^2\right\} \\
&\quad + C_2 \cdot \frac{r^2\ln^3(np)}{\rho_{\min}^2}\left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right)\|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2
\end{aligned}$$

where

$$C_2 := C\bar{A}^4(\kappa + \bar{K} + K_a)^2.$$

Proof. We proceed in steps. The following arguments are all conditional on $\tilde{\mathcal{E}}$. Recall

$$\Delta_1 = \left\{\|\boldsymbol{Z}^{\text{TEST}}\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 + \|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\|^2\right\}\|\hat{\beta} - \beta^*\|_2^2.$$

1. Former factor

   By the definitions of $\tilde{\Delta}_2, \tilde{\Delta}_3$,

$$\begin{aligned}\tilde{\Delta}_2 + np\bar{A}^2\tilde{\Delta}_3 &\le r\bar{A}^2\tilde{\Delta}_2 \\ &\le C\bar{A}^4(\kappa + \bar{K} + K_a)^2 \cdot \frac{r\ln^3(np)}{\rho_{\min}^2}\left(p + n + np\Delta_E^2\right) \\ &= C_2 \cdot \frac{r\ln^3(np)}{\rho_{\min}^2}\left(p + n + np\Delta_E^2\right).\end{aligned}$$

2. Latter factor

   By Lemma E.5,

$$\begin{aligned}\|\hat{\beta} - \beta^*\|_2^2 \le C\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\|^2\|\beta^*\|_2^2 \\ + \frac{C}{\hat{s}_r^2}\Big\{\|\hat{\boldsymbol{A}}^{\text{TRAIN}} - \boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\|_{2,\infty}^2\|\beta^*\|_1^2 \ \vee\ \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2 \ \vee\ \langle\hat{\boldsymbol{A}}^{\text{TRAIN}}(\hat{\beta} - \beta^*), \varepsilon\rangle\Big\}.\end{aligned}$$

   Observe that ε is independent of the event $\tilde{\mathcal{E}}$. Hence, by Lemma E.10

$$\mathbb{E}[\langle\hat{\boldsymbol{A}}^{\text{TRAIN}}(\hat{\beta} - \beta^*), \varepsilon\rangle \mid \tilde{\mathcal{E}}] \le \bar{\sigma}^2 r.$$

   In summary

$$\mathbb{E}[\|\hat{\beta} - \beta^*\|_2^2 \mid \tilde{\mathcal{E}}] \le \tilde{\Delta}_3\|\beta^*\|_2^2 + \frac{\tilde{\Delta}_1\|\beta^*\|_1^2 + \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2 + \bar{\sigma}^2 r}{np/r}.$$

   Focusing on the former term

$$\tilde{\Delta}_3\|\beta^*\|_2^2 = \frac{r}{np}\tilde{\Delta}_2\|\beta^*\|_2^2.$$

   Hence

$$\begin{aligned}&\mathbb{E}[\|\hat{\beta} - \beta^*\|_2^2 \mid \tilde{\mathcal{E}}] \\ &\le \frac{r}{np}\left(\tilde{\Delta}_2\|\beta^*\|_2^2 + \tilde{\Delta}_1\|\beta^*\|_1^2 + \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2 + \bar{\sigma}^2 r\right) \\ &\le C_1 \cdot \bar{\sigma}^2 \cdot \frac{\ln^5(np)}{\rho_{\min}^4}\frac{r}{np}\left\{\left(n + p + np\Delta_E^2\right)\|\beta^*\|_2^2 + \left(r + \frac{rn}{p} + rn\Delta_E^2\right)\|\beta^*\|_1^2\right\} \\ &\quad + \frac{r}{np}\|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2.\end{aligned}$$

3. Collecting results

$$\mathbb{E}[\Delta_1 \mid \tilde{\mathcal{E}}] \le C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^2 \ln^8(np)}{\rho_{\min}^6} \left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right) \cdot \left\{ \left(n + p + np\Delta_E^2\right) \|\beta^*\|_2^2 + \left(r + \frac{rn}{p} + rn\Delta_E^2\right) \|\beta^*\|_1^2 \right\} + C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2} \left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right) \|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2.$$

□

Lemma E.23. *Let the conditions of Proposition E.3 hold. Then,*

$$\mathbb{E}[\Delta_2 \mid \tilde{\mathcal{E}}] \le C_1 \cdot \bar{\sigma}^2 \cdot \bar{A}^2 \cdot \frac{r^2 \ln^5(np)}{\rho_{\min}^4} \|\beta^*\|_1^2 \cdot \left(1 + \frac{n}{p} + n\Delta_E^2\right) + C\bar{A}^2 \cdot r \|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2.$$

Proof. We proceed in steps. The following arguments are all conditional on $\tilde{\mathcal{E}}$. Recall

$$\Delta_2 := \frac{\|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\|^2}{\hat{s}_r^2} \Big\{ \|\hat{\boldsymbol{A}}^{\mathrm{TRAIN}} - \boldsymbol{A}^{(\mathrm{LR}),\mathrm{TRAIN}}\|_{2,\infty}^2 \|\beta^*\|_1^2 \ \vee \ \|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2 \ \vee \ \langle \hat{\boldsymbol{A}}^{\mathrm{TRAIN}}(\hat{\beta} - \beta^*), \varepsilon \rangle \Big\}.$$

1. Former factor

   Note that conditioned on $\tilde{\mathcal{E}}$ and Assumption 5.4,

   $$\frac{\|\boldsymbol{A}^{(\mathrm{LR}),\mathrm{TEST}}\|^2}{\hat{s}_r^2} \le C \frac{r}{np} np\bar{A}^2 = Cr\bar{A}^2.$$

2. Latter factor

   Observing that ε is independent of $\tilde{\mathcal{E}}$. By Lemma E.10

   $$\tilde{\Delta}_1 \|\beta^*\|_1^2 + \|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2 + \bar{\sigma}^2 r \le C_1 \cdot \bar{\sigma}^2 \cdot \frac{r \ln^5(np)}{\rho_{\min}^4} \|\beta^*\|_1^2 \left(1 + \frac{n}{p} + n\Delta_E^2\right) + \|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2.$$

3. Combining results

$$\mathbb{E}[\Delta_2 \mid \tilde{\mathcal{E}}] \le C_1 \cdot \bar{\sigma}^2 \cdot \bar{A}^2 \cdot \frac{r^2 \ln^5(np)}{\rho_{\min}^4} \|\beta^*\|_1^2 \cdot \left(1 + \frac{n}{p} + n\Delta_E^2\right) + C\bar{A}^2 \cdot r \|\phi^{(\mathrm{LR}),\mathrm{TRAIN}}\|_2^2.$$

□

Lemma E.24.

$$\mathbb{E}\left[\Delta_3 \,\middle|\, \tilde{\mathcal{E}}\right] \le C_1 \cdot \frac{r \ln^5(np)}{\rho_{\min}^4} \|\beta^*\|_1^2 \left(1 + \frac{n}{p} + n\Delta_E^2\right).$$

Proof. Recall

$$\Delta_3 := \|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2 \|\beta^*\|_1^2.$$

Using the definition of $\tilde{\mathcal{E}}$, we have

$$\mathbb{E}[\Delta_3 \mid \tilde{\mathcal{E}}] \le \tilde{\Delta}_1 \|\beta^*\|_1^2 \le C_1 \cdot \frac{r \ln^5(np)}{\rho_{\min}^4} \|\beta^*\|_1^2 \left(1 + \frac{n}{p} + n\Delta_E^2\right).$$

□

Lemma E.25. *Let the conditions of Theorem 5.2 hold. Then*

$$\sum_{m=1}^{3} \mathbb{E}[\Delta_m | \tilde{\mathcal{E}}] \le C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(np)}{\rho_{\min}^6} \|\beta^*\|_1^2 \left\{1 + \frac{p}{n} + \frac{n}{p} + (n+p)\Delta_E^2 + np\Delta_E^4\right\}$$
$$+ C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2} \left(1 + \Delta_E^2\right) \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2.$$

Proof. Recall Lemmas E.22, E.23, and E.24:

$$\begin{aligned}
&\mathbb{E}[\Delta_1 \mid \tilde{\mathcal{E}}] \\
&\le C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^2 \ln^8(np)}{\rho_{\min}^6} \left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right) \cdot \left\{ \left(n + p + np\Delta_E^2\right) \|\beta^*\|_2^2 + \left(r + \frac{rn}{p} + rn\Delta_E^2\right) \|\beta^*\|_1^2 \right\} \\
&\quad + C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2} \left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right) \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2 \\
&\mathbb{E}[\Delta_2 \mid \tilde{\mathcal{E}}] \le C_1 \cdot \bar{\sigma}^2 \cdot \bar{A}^2 \cdot \frac{r^2 \ln^5(np)}{\rho_{\min}^4} \|\beta^*\|_1^2 \cdot \left(1 + \frac{n}{p} + n\Delta_E^2\right) + C\bar{A}^2 \cdot r \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2 \\
&\mathbb{E}\left[\Delta_3 \,\middle|\, \tilde{\mathcal{E}}\right] \le C_1 \cdot \frac{r \ln^5(np)}{\rho_{\min}^4} \|\beta^*\|_1^2 \left(1 + \frac{n}{p} + n\Delta_E^2\right).
\end{aligned}$$

Δ_2 dominates Δ_3. Focusing on the first term of Δ_1,

$$\begin{aligned}
(1) &\le C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(np)}{\rho_{\min}^6} \left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right) \cdot \left\{ \left(n + p + np\Delta_E^2\right) \|\beta^*\|_2^2 + \left(1 + \frac{n}{p} + n\Delta_E^2\right) \|\beta^*\|_1^2 \right\} \\
&= C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(np)}{\rho_{\min}^6} \left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right) \cdot \left\{ \left(n + p + np\Delta_E^2\right) \left(\|\beta^*\|_2^2 + \frac{\|\beta^*\|_1^2}{p}\right) \right\} \\
&\le C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(np)}{\rho_{\min}^6} \|\beta^*\|_2^2 \left\{1 + \frac{p}{n} + \frac{n}{p} + (n+p)\Delta_E^2 + np\Delta_E^4\right\}.
\end{aligned}$$

Comparing to the first term of Δ_2, it is sufficient to bound $\|\beta^*\|_2 \le \|\beta^*\|_1$. Focusing on the second term of Δ_1,

$$(2) = C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2} \left(\frac{1}{n} + \frac{1}{p} + \Delta_E^2\right) \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2.$$

Comparing to the second term of Δ_2, it is sufficient to bound $\frac{1}{n} + \frac{1}{p} \leq 2$. In summary, the bound is

$$\sum_{m=1}^{3} \mathbb{E}[\Delta_m | \tilde{\mathcal{E}}] \leq C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(np)}{\rho_{\min}^6} \|\beta^*\|_1^2 \left\{ 1 + \frac{p}{n} + \frac{n}{p} + (n+p)\Delta_E^2 + np\Delta_E^4 \right\} + C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2} \left(1 + \Delta_E^2\right) \|\phi^{(\text{LR}),\text{TRAIN}}\|_2^2.$$

□

E.4.4 Collecting results

Lemma E.26. *Deterministically, for* TRAIN *and* TEST

$$\|\phi^{(\text{LR})}\|_2^2 \leq 2\|\phi\|_2^2 + 2n\Delta_E^2 \|\beta^*\|_1^2.$$

Proof. Write

$$\|\phi^{(\text{LR})}\|_2^2 = \|\phi + \boldsymbol{E}^{(\text{LR})}\beta^*\|_2^2 \leq 2\|\phi\|_2^2 + 2\|\boldsymbol{E}^{(\text{LR})}\beta^*\|_2^2.$$

Focusing on the latter term

$$\|\boldsymbol{E}^{(\text{LR})}\beta^*\|_2^2 \leq \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2 \|\beta^*\|_1^2 \leq n\Delta_E^2 \|\beta^*\|_1^2.$$

□

Lemma E.27. *Let the conditions of Theorem 5.2 hold. Then*

$$\sum_{m=1}^{3} \mathbb{E}[\Delta_m | \tilde{\mathcal{E}}] \leq C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(np)}{\rho_{\min}^6} \|\beta^*\|_1^2 \left\{ 1 + \frac{p}{n} + \frac{n}{p} + (n+p)\Delta_E^2 + np\Delta_E^4 \right\} + C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2} \left(1 + \Delta_E^2\right) \|\phi^{\text{TRAIN}}\|_2^2.$$

Proof. By Lemmas E.25 and E.26

$$\begin{aligned} &\sum_{m=1}^{3} \mathbb{E}[\Delta_m | \tilde{\mathcal{E}}] \\ &\leq C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(np)}{\rho_{\min}^6} \|\beta^*\|_1^2 \left\{ 1 + \frac{p}{n} + \frac{n}{p} + (n+p)\Delta_E^2 + np\Delta_E^4 \right\} \\ &\quad + C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2} \left(1 + \Delta_E^2\right) \left(\|\phi^{\text{TRAIN}}\|_2^2 + n\Delta_E^2 \|\beta^*\|_1^2 \right). \end{aligned}$$

Opening up the product in the latter term

$$\left(1 + \Delta_E^2\right) \left(\|\phi^{\text{TRAIN}}\|_2^2 + n\Delta_E^2 \|\beta^*\|_1^2 \right) = \|\phi^{\text{TRAIN}}\|_2^2 + n\Delta_E^2 \|\beta^*\|_1^2 + \|\phi^{\text{TRAIN}}\|_2^2 \Delta_E^2 + n\Delta_E^4 \|\beta^*\|_1^2.$$

Both $n\Delta_E^2\|\beta^*\|_1^2$ and $n\Delta_E^4\|\beta^*\|_1^2$ are dominated by the first term. Hence

$$\begin{aligned}
&\sum_{m=1}^{3}\mathbb{E}[\Delta_m|\tilde{\mathcal{E}}] \\
&\leq C_1C_2\cdot\bar{\sigma}^2\cdot\frac{r^3\ln^8(np)}{\rho_{\min}^6}\|\beta^*\|_1^2\left\{1+\frac{p}{n}+\frac{n}{p}+(n+p)\Delta_E^2+np\Delta_E^4\right\} \\
&\quad + C_2\cdot\frac{r^2\ln^3(np)}{\rho_{\min}^2}\left(1+\Delta_E^2\right)\|\phi^{\text{TRAIN}}\|_2^2.
\end{aligned}$$

□

Proposition E.4 (TEST ERROR)**.** *Let the conditions of Theorem 5.2 hold. Then*

$$\begin{aligned}
\mathbb{E}[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}-\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}] &\leq C_1C_2\cdot\bar{\sigma}^2\cdot\frac{r^3\ln^8(np)}{\rho_{\min}^6}\|\beta^*\|_1^2\left\{1+\frac{p}{n}+\frac{n}{p}+(n+p)\Delta_E^2+np\Delta_E^4\right\} \\
&\quad + C_2\cdot\frac{r^2\ln^3(np)}{\rho_{\min}^2}\left(1+\Delta_E^2\right)\|\phi^{\text{TRAIN}}\|_2^2.
\end{aligned}$$

Proof. By Lemma E.16,

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}-\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right] &= \mathbb{E}\left[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}-\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2|\tilde{\mathcal{E}}\right]\mathbb{P}(\tilde{\mathcal{E}}) \\
&\leq \mathbb{E}\left[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}-\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2|\tilde{\mathcal{E}}\right] \\
&\leq C\sum_{m=1}^{3}\mathbb{E}[\Delta_m|\tilde{\mathcal{E}}].
\end{aligned}$$

Finally appeal to Lemma E.27. □

Remark E.22 (Dictionary)**.** *The generalization of Proposition E.4 is*

$$\begin{aligned}
&\mathbb{E}[\|b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\beta}-b(\boldsymbol{A}^{\text{TEST}})\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}] \\
&\leq C_b'C_1C_2\cdot\bar{\sigma}^2\cdot\frac{(r')^3\ln^8(np)}{(\rho'_{\min})^6}\|\beta^*\|_1^2\left\{1+\frac{p}{n}+\frac{n}{p}+(n+p)(\Delta'_E)^2+np(\Delta'_E)^4\right\} \\
&\quad + C_2\cdot\frac{(r')^2\ln^3(np)}{(\rho'_{\min})^2}\left\{1+(\Delta'_E)^2\right\}\|\phi^{\text{TRAIN}}\|_2^2.
\end{aligned}$$

E.5 Generalization error

E.5.1 Decomposition

To lighten notation, we define, for $i\in$ TEST

$$\hat{\gamma}_i = Z_{i,\cdot}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta},\quad \gamma_i=\gamma_0(A_{i,\cdot})$$

which form the vectors $\hat{\boldsymbol{\gamma}},\boldsymbol{\gamma}_0\in\mathbb{R}^n$.

Lemma E.28. *Deterministically,*

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2 \le 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2 + 2\|\boldsymbol{A}^{\text{TEST}}\beta^* - \gamma_0(\boldsymbol{A}^{\text{TEST}})\|_2^2.$$

Moreover

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2 \le 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2 + 2\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2.$$

Proof. Write

$$\begin{aligned}
&\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \\
&= \boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \gamma_0(\boldsymbol{A}^{\text{TEST}}) \\
&= \boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} \pm \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} \pm \boldsymbol{A}^{\text{TEST}}\beta^* - \gamma_0(\boldsymbol{A}^{\text{TEST}}).
\end{aligned}$$

□

We analyze each term separately.

1. Approximation error $\|\boldsymbol{A}^{\text{TEST}}\beta^* - \gamma_0(\boldsymbol{A}^{\text{TEST}})\|_2^2 = \|\phi^{\text{TEST}}\|_2^2$.
2. Test error $\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2$.
3. Implicit cleaning error $\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2$.

Approximation error requires no further analysis. Proposition E.4 analyzes TEST ERROR. What remains is an analysis of the final term via projection geometry.

Remark E.23 (Dictionary)**.** *The generalization with a dictionary considers*

1. *Approximation error* $\|b(\boldsymbol{A}^{\text{TEST}})\beta^* - \gamma_0(\boldsymbol{A}^{\text{TEST}})\|_2^2 = \|\phi^{\text{TEST}}\|_2^2$.
2. *Test error* $\|b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\beta} - b(\boldsymbol{A}^{\text{TEST}})\beta^*\|_2^2$.
3. *Implicit cleaning error* $\|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\beta} - b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\beta}\|_2^2$.

E.5.2 Implicit cleaning

Lemma E.29. *Suppose Assumption 5.6 holds and let $k = r$. Then*

$$\begin{aligned}&\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2 \\ &\leq C\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \left\{\|\hat{\beta} - \beta^*\|_2^2 + \|\hat{\boldsymbol{V}}'_r(\hat{\boldsymbol{V}}'_r)^T - \boldsymbol{V}'(\boldsymbol{V}')^T\|^2\|\beta^*\|_2^2\right\}.\end{aligned}$$

Proof. We proceed in steps.

1. Decomposition

   Recall the definitions

   $$\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} = \hat{\boldsymbol{U}}'\hat{\boldsymbol{\Sigma}}'(\hat{\boldsymbol{V}}')^T, \quad \hat{\boldsymbol{A}}^{\text{TEST}} = \hat{\boldsymbol{U}}'_r\hat{\boldsymbol{\Sigma}}'_r(\hat{\boldsymbol{V}}'_r)^T, \quad \hat{\boldsymbol{A}}_\perp^{\text{TEST}} = \hat{\boldsymbol{U}}'_{r,\perp}\hat{\boldsymbol{\Sigma}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T$$

   so that

   $$\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} = \quad \hat{\boldsymbol{A}}^{\text{TEST}} + \hat{\boldsymbol{A}}_\perp^{\text{TEST}}.$$

   Hence

   $$\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} = \hat{\boldsymbol{A}}_\perp^{\text{TEST}}\hat{\beta} = \hat{\boldsymbol{U}}'_{r,\perp}\hat{\boldsymbol{\Sigma}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T\hat{\beta}$$

   and

   $$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2 \leq \|\hat{\boldsymbol{U}}'_{r,\perp}\| \cdot \|\hat{\boldsymbol{\Sigma}}'_{r,\perp}\| \cdot \|(\hat{\boldsymbol{V}}'_{r,\perp})^T\hat{\beta}\|_2 = \|\hat{\boldsymbol{\Sigma}}'_{r,\perp}\| \cdot \|(\hat{\boldsymbol{V}}'_{r,\perp})^T\hat{\beta}\|_2.$$

2. $\|\hat{\boldsymbol{\Sigma}}'_{r,\perp}\|$

   As in the proof of Lemma E.16, we appeal to Weyl's inequality (Lemma E.6):

   $$\|\hat{\boldsymbol{\Sigma}}'_{r,\perp}\| = \hat{s}'_{r+1} = \hat{s}'_{r+1} - s'_{r+1} \leq \|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|.$$

3. $\|(\hat{\boldsymbol{V}}'_{r,\perp})^T\hat{\beta}\|_2$

   Write

   $$\|(\hat{\boldsymbol{V}}'_{r,\perp})^T\hat{\beta}\|_2 = \|\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T\hat{\beta}\|_2 \leq \|\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T(\hat{\beta} - \beta^*)\|_2 + \|\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T\beta^*\|_2.$$

   Focusing on the former term

   $$\|\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T(\hat{\beta} - \beta^*)\|_2 \leq \|\hat{\beta} - \beta^*\|_2.$$

Focusing on the latter term, we appeal to Lemma E.2:

$$\begin{aligned}
\|\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T\beta^*\|_2 &= \|\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T\boldsymbol{V}'(\boldsymbol{V}')^T\beta^*\|_2 \\
&\leq \|\{\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T - \boldsymbol{V}'_{\perp}(\boldsymbol{V}'_{\perp})^T\}\boldsymbol{V}'(\boldsymbol{V}')^T\beta^*\|_2 + \|\boldsymbol{V}'_{\perp}(\boldsymbol{V}'_{\perp})^T\boldsymbol{V}'(\boldsymbol{V}')^T\beta^*\|_2 \\
&= \|\{\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T - \boldsymbol{V}'_{\perp}(\boldsymbol{V}'_{\perp})^T\}\boldsymbol{V}'(\boldsymbol{V}')^T\beta^*\|_2 \\
&= \|\{\hat{\boldsymbol{V}}'_{r,\perp}(\hat{\boldsymbol{V}}'_{r,\perp})^T - \boldsymbol{V}'_{\perp}(\boldsymbol{V}'_{\perp})^T\}\beta^*\|_2 \\
&= \|\{\hat{\boldsymbol{V}}'_{r}(\hat{\boldsymbol{V}}'_{r})^T - \boldsymbol{V}'(\boldsymbol{V}')^T\}\beta^*\|_2 \\
&\leq \|\hat{\boldsymbol{V}}'_{r}(\hat{\boldsymbol{V}}'_{r})^T - \boldsymbol{V}'(\boldsymbol{V}')^T\|\|\beta^*\|_2.
\end{aligned}$$

□

Remark E.24 (Dictionary)**.** *The generalization of Lemma E.29 is*

$$\begin{aligned}
&\|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\beta} - b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\beta}\|_2^2 \\
&\leq C\|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}) - b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\}\|^2 \cdot \left\{\|\hat{\beta} - \beta^*\|_2^2 + \|\hat{\boldsymbol{V}}'_{r'}(\hat{\boldsymbol{V}}'_{r'})^T - \boldsymbol{V}'(\boldsymbol{V}')^T\|^2\|\beta^*\|_2^2\right\}
\end{aligned}$$

using the SVDs in Remark E.1.

Proof. We proceed in steps

1. Decomposition

   For the polynomial dictionary with uncorrupted nonlinearity, denote

$$\begin{aligned}
\boldsymbol{M} &:= b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}) - b(\hat{\boldsymbol{A}}^{\text{TEST}}) \\
&= \{0, 0, ..., 0, (\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}^{\text{TEST}}), diag(D)(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}^{\text{TEST}}), \\
&\quad ..., diag(D)^{d_{\max}-1}(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}^{\text{TEST}})\}.
\end{aligned}$$

   Consider the SVD

$$\boldsymbol{M} = \boldsymbol{U}_M\boldsymbol{\Sigma}_M\boldsymbol{V}_M^T.$$

   Hence

$$\boldsymbol{M}\hat{\beta} = \boldsymbol{U}_M\boldsymbol{\Sigma}_M\boldsymbol{V}_M^T\hat{\beta}$$

   and

$$\|\boldsymbol{M}\hat{\beta}\|_2 \leq \|\boldsymbol{U}_M\| \cdot \|\boldsymbol{\Sigma}_M\| \cdot \|\boldsymbol{V}_M^T\hat{\beta}\|_2 = \|\boldsymbol{\Sigma}_M\| \cdot \|\boldsymbol{V}_M^T\hat{\beta}\|_2.$$

2. $\|\boldsymbol{\Sigma}_M\|$

As argued in Remark E.7, using the decomposition of $\boldsymbol{M}$,

$$\begin{aligned}\|\boldsymbol{\Sigma}_M\| &= \|\boldsymbol{M}\| \\ &\leq \|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}^{\text{TEST}}\| + \bar{A}\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}^{\text{TEST}}\| + ... + \bar{A}^{d_{\max}-1}\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}^{\text{TEST}}\| \\ &\leq \bar{A}^{d_{\max}} d_{\max}\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}^{\text{TEST}}\|.\end{aligned}$$

Next observe that Lemma E.7

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}^{\text{TEST}}\| = \hat{s}'_{r+1} = \hat{s}'_{r+1} - s'_{r+1} \leq \|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|$$

In summary

$$\|\boldsymbol{\Sigma}_M\| \leq \bar{A}^{d_{\max}} d_{\max}\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|$$

which is dominated by the bound on $\|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}) - b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\}\|$.

3. $\|\boldsymbol{V}_M^T\hat{\beta}\|_2$

Write

$$\|\boldsymbol{V}_M^T\hat{\beta}\|_2 = \|\boldsymbol{V}_M\boldsymbol{V}_M^T\hat{\beta}\|_2 \leq \|\boldsymbol{V}_M\boldsymbol{V}_M^T(\hat{\beta} - \beta^*)\|_2 + \|\boldsymbol{V}_M\boldsymbol{V}_M^T\beta^*\|_2.$$

Focusing on the former term

$$\|\boldsymbol{V}_M\boldsymbol{V}_M^T(\hat{\beta} - \beta^*)\|_2 \leq \|\hat{\beta} - \beta^*\|_2.$$

Focusing on the latter term, we appeal to Lemma E.2:

$$\begin{aligned}\|\boldsymbol{V}_M\boldsymbol{V}_M^T\beta^*\|_2 &= \|\boldsymbol{V}_M\boldsymbol{V}_M^T\boldsymbol{V}'(\boldsymbol{V}')^T\beta^*\|_2 \\ &\leq \|\{\boldsymbol{V}_M\boldsymbol{V}_M^T - \boldsymbol{V}'_\perp(\boldsymbol{V}'_\perp)^T\}\boldsymbol{V}'(\boldsymbol{V}')^T\beta^*\|_2 + \|\boldsymbol{V}'_\perp(\boldsymbol{V}'_\perp)^T\boldsymbol{V}'(\boldsymbol{V}')^T\beta^*\|_2 \\ &= \|\{\boldsymbol{V}_M\boldsymbol{V}_M^T - \boldsymbol{V}'_\perp(\boldsymbol{V}'_\perp)^T\}\boldsymbol{V}'(\boldsymbol{V}')^T\beta^*\|_2 \\ &= \|\{\boldsymbol{V}_M\boldsymbol{V}_M^T - \boldsymbol{V}'_\perp(\boldsymbol{V}'_\perp)^T\}\beta^*\|_2 \\ &\leq \|\boldsymbol{V}_M\boldsymbol{V}_M^T - \boldsymbol{V}'_\perp(\boldsymbol{V}'_\perp)^T\|\|\beta^*\|_2\end{aligned}$$

where $\boldsymbol{V}'$ is from the SVD of $b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\}$ and $\boldsymbol{V}_M$ is from the SVD of $\boldsymbol{M} = b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}) - b(\hat{\boldsymbol{A}}^{\text{TEST}})$.

4. Dictionary geometry

To complete the argument, it is sufficient to argue that $\boldsymbol{V}_M \boldsymbol{V}_M^T = \hat{\boldsymbol{V}}'_{r',\perp}(\hat{\boldsymbol{V}}'_{r',\perp})^T$, where $\hat{\boldsymbol{V}}'_{r'}$ is from SVD of $b(\hat{\boldsymbol{A}}^{\text{TEST}})$. In other words, we wish to show

$$\text{NULL}\{b(\hat{\boldsymbol{A}}^{\text{TEST}})\} = \text{ROW}\{b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}) - b(\hat{\boldsymbol{A}}^{\text{TEST}})\}.$$

We can verify this property for the polynomial dictionary with uncorrupted nonlinearity appealing to Assumption C.2. Suppressing superscripts,

$$b(\hat{\boldsymbol{A}}) = \{1, D, ..., D^{d_{\max}}, \hat{\boldsymbol{A}}, diag(D)\hat{\boldsymbol{A}}, ..., diag(D)^{d_{\max}-1}\hat{\boldsymbol{A}}\}$$

with j-th row

$$(1, D_j, ..., D_j^{d_{\max}}, \hat{A}_{j,\cdot}, D_j\hat{A}_{j,\cdot}, ..., D_j^{d_{\max}-1}\hat{A}_{j,\cdot})$$

and

$$\begin{aligned}
&b(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1}) - b(\hat{\boldsymbol{A}}) \\
&= \{0, 0, ..., 0, (\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}), diag(D)(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}}), ..., diag(D)^{d_{\max}-1}(\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1} - \hat{\boldsymbol{A}})\} \\
&= \{0, 0, ..., 0, \hat{\boldsymbol{A}}_\perp, diag(D)\hat{\boldsymbol{A}}_\perp, ..., diag(D)^{d_{\max}-1}\hat{\boldsymbol{A}}_\perp\}
\end{aligned}$$

with i-th row

$$(0, 0, ..., 0, \hat{A}_{\perp i,\cdot}, D_i\hat{A}_{\perp i,\cdot}, ..., D_i^{d_{\max}-1}\hat{A}_{\perp i,\cdot}).$$

The spaces induced by the rows are clearly orthogonal, so

$$\text{ROW}\{b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}) - b(\hat{\boldsymbol{A}}^{\text{TEST}})\} \subset \text{NULL}\{b(\hat{\boldsymbol{A}}^{\text{TEST}})\}.$$

Assumption C.2.3 further implies

$$\text{NULL}\{b(\hat{\boldsymbol{A}}^{\text{TEST}})\} \subset \text{ROW}\{b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}) - b(\hat{\boldsymbol{A}}^{\text{TEST}})\}.$$

□

Proposition E.5 (Implicit cleaning)**.** *Let the conditions of Theorem 5.2 hold. Then*

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}] \\
&\leq C_1C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3\ln^8(np)}{\rho_{\min}^6}\|\beta^*\|_1^2\left\{1 + \frac{p}{n} + \frac{n}{p} + (n+p)\Delta_E^2 + np\Delta_E^4\right\} \\
&\quad + C_2 \cdot \frac{r^2\ln^3(np)}{\rho_{\min}^2}\left(1 + \Delta_E^2\right)\|\phi^{\text{TRAIN}}\|_2^2.
\end{aligned}$$

Proof. To begin, write

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}] &= \mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2|\tilde{\mathcal{E}}\right]\mathbb{P}(\tilde{\mathcal{E}}) \\ &\leq \mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2|\tilde{\mathcal{E}}\right].\end{aligned}$$

By Lemma E.29, it is sufficient to analyze

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \left\{\|\hat{\beta} - \beta^*\|_2^2 + \|\hat{\boldsymbol{V}}_r'(\hat{\boldsymbol{V}}_r')^T - \boldsymbol{V}'(\boldsymbol{V}')^T\|^2\|\beta^*\|_2^2\right\}$$

under the beneficial event $\tilde{\mathcal{E}}$. By Lemma E.5, the bound on the former term dominates the bound on the latter term. Therefore we analyze

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \|\hat{\beta} - \beta^*\|_2^2.$$

By Lemma E.16

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \|\hat{\beta} - \beta^*\|_2^2 \leq \Delta_1 \leq C\sum_{m=1}^{3}\Delta_m$$

so we can use the bound previously used for analyzing TEST ERROR $\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2$. This loose bound is sufficient for our purposes, since the TEST ERROR term will ultimately give this rate. In summary

$$\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2|\tilde{\mathcal{E}}\right] \leq C\sum_{m=1}^{3}\mathbb{E}[\Delta_m|\tilde{\mathcal{E}}].$$

Finally, we appeal to Lemma E.27. □

Remark E.25 (Dictionary)**.** *The generalization of Proposition E.5 is*

$$\begin{aligned}&\mathbb{E}[\|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\beta} - b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\beta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}] \\ &\leq C_b'C_1C_2\cdot\bar{\sigma}^2\cdot\frac{(r')^3\ln^8(np)}{(\rho'_{\min})^6}\|\beta^*\|_1^2\left\{1 + \frac{p}{n} + \frac{n}{p} + (n+p)(\Delta'_E)^2 + np(\Delta'_E)^4\right\} \\ &\quad + C_2\cdot\frac{(r')^2\ln^3(np)}{(\rho'_{\min})^2}\left\{1 + (\Delta'_E)^2\right\}\|\phi^{\text{TRAIN}}\|_2^2.\end{aligned}$$

E.5.3 Bounded estimator moments

For the adverse case, we place a weak technical condition on how the estimator moments scale. We state the technical condition then demonstrate that it is implied by the interpretable condition given in the main text.

Assumption E.1 (Bounded estimator moment)**.**

$$\sqrt{\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i\in\text{TEST}}\hat{\gamma}(W_{i,\cdot})^2\right\}^2\right]} \leq \text{polynomial}(n,p)\cdot C_2\cdot\frac{r^2\ln^3(np)}{\rho_{\min}^2}\left(\|\beta^*\|_1^2+\|\phi^{\text{TRAIN}}\|_2^2\right).$$

where $C_2 = C\cdot\bar{A}^4(\kappa+\bar{K}+K_a)^2$.

Recall from Appendix D that the powers of (n,p) are arbitrary in the probability of the adverse event; $\mathbb{P}(\tilde{\mathcal{E}}^c)\leq\frac{C}{\text{polynomial}(n,p)}$ for any polynomial of (n,p). Therefore the moments of our estimator $\hat{\gamma}(W_{i,\cdot})$ can scale as any arbitrary polynomial of n and p, denoted by polynomial(n,p). We are simply ruling out some extremely adversarial cases. Assumption E.1 is essentially requiring that $\hat{\beta}$ is well conditioned. Indeed, we are able to satisfy the assumption under a simple condition on the smallest singular value used in PCR.

Proposition E.6 (Verifying bounded estimator moment)**.** *Suppose the Assumptions 5.1 and 5.2 hold. Further suppose* $\hat{s}_k\gtrsim\frac{\bar{\varepsilon}}{\text{polynomial}(n,p)}$ *where* $\mathbb{E}[\varepsilon_i^8]\leq\bar{\varepsilon}^8$*. Then Assumption E.1 holds.*

Remark E.26 (Dictionary)**.** *If Assumption C.2 holds then Assumption E.1 becomes*

$$\sqrt{\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i\in\text{TEST}}\hat{\gamma}(W_{i,\cdot})^2\right\}^2\right]} \leq \text{polynomial}(n,p)\cdot C_2\cdot\frac{(r')^2\ln^3(np)}{(\rho'_{\min})^2}\left(\|\beta^*\|_1^2+\|\phi^{\text{TRAIN}}\|_2^2\right).$$

Proposition E.6 generalizes accordingly: if Assumption 5.7 holds then the generalization of Assumption E.1 holds.

We prove Proposition E.6 via a sequence of lemmas.

Lemma E.30. *Suppose Assumptions 5.1 and 5.2 hold. Then*

$$\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\|_{2,\infty}^8\right]\leq C\cdot\bar{A}^8K_a^8\cdot\ln^8(np)n^{12}.$$

Proof. We suppress the superscript to lighten notation. The argument echoes Lemma D.17.

$$\begin{aligned}
\|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1}\|_{2,\infty} &= \max_{j\in[p]} \|\hat{Z}_{\cdot,j}\hat{\rho}_j^{-1}\|_2 \\
&= \max_{j\in[p]} \frac{1}{\hat{\rho}_j}\|Z_{\cdot,j}\|_2 \\
&\le n \max_{j\in[p]} \|Z_{\cdot,j}\|_2 \\
&\le n^{\frac{3}{2}} \max_{i\in[n],j\in[p]} |Z_{ij}| \\
&\le n^{\frac{3}{2}}(\bar{A} + \max_{i,j}|H_{ij}|).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{Z}\,\hat{\boldsymbol{\rho}}^{-1}\|_{2,\infty}^8\right] &\le \mathbb{E}[\{n^{\frac{3}{2}}(\bar{A} + \max_{i,j}|H_{ij}|)\}^8] \\
&\le Cn^{12}(\bar{A}^8 + \mathbb{E}[\max_{i,j}|H_{ij}|^8]) \\
&\le Cn^{12}\{\bar{A}^8 + K_a^8 \ln^8(np)\}.
\end{aligned}$$

The final inequality holds because for any $a > 0$ and $\theta \ge 1$, if H_{ij} is a ψ_a-random variable then $\left|H_{ij}\right|^\theta$ is a $\psi_{a/\theta}$-random variable. With the choice of $\theta = 8$, we have that

$$\mathbb{E}[\max_{i,j}|H_{ij}|^8] \le CK_a^8 \ln^{\frac{8}{a}}(np).$$

□

Lemma E.31. *Suppose Assumption 5.1 holds, $\hat{s}_k \ge \underline{s}$, and $\mathbb{E}[\varepsilon_i^8] \le \bar{\varepsilon}^8$. Then*

$$\mathbb{E}\left[\|\hat{\beta}\|_1^8\right] \le C \cdot \bar{A}^8\bar{\varepsilon}^8\underline{s}^{-8} \cdot p^4(n^4\|\beta^*\|_1^8 + \|\phi^{\text{TRAIN}}\|_2^8).$$

Proof. We suppress the superscript to lighten notation. Recall that $\hat{\beta} = \hat{\boldsymbol{V}}_k\hat{\boldsymbol{\Sigma}}_k^{-1}\hat{\boldsymbol{U}}_k^T Y$, hence

$$\begin{aligned}
\|\hat{\beta}\|_1 &\le \sqrt{p}\|\hat{\beta}\|_2 \\
&\le \sqrt{p}\|\hat{\boldsymbol{V}}_k\| \cdot \|\hat{\boldsymbol{\Sigma}}_k^{-1}\| \cdot \|\hat{\boldsymbol{U}}_k^T\| \cdot \|Y\|_2 \\
&= \sqrt{p}\hat{s}_k^{-1}\|Y\|_2.
\end{aligned}$$

Moreover $Y = \boldsymbol{A}\beta^* + \phi + \varepsilon$ hence

$$\begin{aligned}
\|Y\|_2 &\le \|\boldsymbol{A}\|_{2,\infty}\|\beta^*\|_1 + \|\phi\|_2 + \|\varepsilon\|_2 \\
&\le \sqrt{n}\bar{A}\|\beta^*\|_1 + \|\phi\|_2 + \|\varepsilon\|_2.
\end{aligned}$$

Therefore

$$\|\hat{\beta}\|_1^8 \leq Cp^4\underline{s}^{-8}(n^4\bar{A}^8\|\beta^*\|_1^8 + \|\phi\|_2^8 + \|\varepsilon\|_2^8).$$

and hence

$$\mathbb{E}\left[\|\hat{\beta}\|_1^8\right] \leq Cp^4\underline{s}^{-8}(n^4\bar{A}^8\|\beta^*\|_1^8 + \|\phi\|_2^8 + \mathbb{E}\|\varepsilon\|_2^8).$$

Finally note that

$$\mathbb{E}\|\varepsilon\|_2^8 = \mathbb{E}\left[\left(\sum_{i\in[n]}\varepsilon_i^2\right)^4\right] = n^4\mathbb{E}\left[\left(\frac{1}{n}\sum_{i\in[n]}\varepsilon_i^2\right)^4\right] \leq n^4\mathbb{E}\left[\frac{1}{n}\sum_{i\in[n]}\varepsilon_i^8\right]. \leq n^4\bar{\varepsilon}^8.$$

The first inequality holds since $\{\mathbb{E}_n[A]\}^4 \leq \mathbb{E}_n[A^4]$, taking $A_i = \varepsilon_i^2$. □

Lemma E.32. *Suppose the conditions of Lemmas E.30 and E.31 hold. Then*

$$\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]} \leq C\cdot\bar{A}^4K_a^2\cdot\ln^2(np)\left(\|\beta^*\|_1^2 + \frac{1}{n}\|\phi^{\text{TRAIN}}\|_2^2\right)n^4p\frac{\bar{\varepsilon}^2}{\underline{s}^2}.$$

Proof. By Cauchy-Schwarz, write

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]} &\leq \sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\|_{2,\infty}^4\|\hat{\beta}\|_1^4\right]} \\
&\leq \sqrt{\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\|_{2,\infty}^8\right]}\sqrt{\mathbb{E}\left[\|\hat{\beta}\|_1^8\right]}} \\
&\leq C\cdot\bar{A}^2K_a^2\cdot\ln^2(np)n^3\cdot\bar{A}^2\bar{\varepsilon}^2\underline{s}^{-2}\cdot p(n\|\beta^*\|_1^2 + \|\phi^{\text{TRAIN}}\|_2^2) \\
&= C\cdot\bar{A}^4K_a^2\cdot\ln^2(np)\left(\|\beta^*\|_1^2 + \frac{1}{n}\|\phi^{\text{TRAIN}}\|_2^2\right)n^4p\frac{\bar{\varepsilon}^2}{\underline{s}^2}.
\end{aligned}$$

□

Proof of Proposition E.6. To begin, observe that

$$\sqrt{\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i\in\text{TEST}}\hat{\gamma}(W_{i,\cdot})^2\right\}^2\right]} = \sqrt{\mathbb{E}\left[\frac{1}{n^2}\left\{\sum_{i\in\text{TEST}}\hat{\gamma}(W_{i,\cdot})^2\right\}^2\right]} = \frac{1}{n}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]}.$$

By Lemma E.32

$$\begin{aligned}
\frac{n^{-1}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]}}{n^4p^5} &\leq C\cdot\bar{A}^4K_a^2\cdot\ln^2(np)\left(\|\beta^*\|_1^2 + \frac{1}{n}\|\phi^{\text{TRAIN}}\|_2^2\right)\frac{\bar{\varepsilon}^2}{\underline{s}^2np^4} \\
&\leq C\cdot\bar{A}^4K_a^2\cdot\ln^2(np)\left(\|\beta^*\|_1^2 + \frac{1}{n}\|\phi^{\text{TRAIN}}\|_2^2\right) \\
&\leq C_2\cdot\frac{r^2\ln^3(np)}{\rho_{\min}^2}\left(\|\beta^*\|_1^2 + \|\phi^{\text{TRAIN}}\|_2^2\right)
\end{aligned}$$

where the penultimate inequality holds since $\underline{s} \geq \frac{\bar{\varepsilon}}{\sqrt{np^2}}$ implies $\underline{s}^2 np^4 \geq \bar{\varepsilon}^2$ and the ultimate inequality confirms Assumption E.1. More generally,

$$\begin{aligned}
\frac{\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]}}{\text{polynomial}(n,p)} &\leq C \cdot \bar{A}^4 K_a^2 \cdot \ln^2(np) \left(\|\beta^*\|_1^2 + \frac{1}{n}\|\phi^{\text{TRAIN}}\|_2^2\right) \frac{\bar{\varepsilon}^2}{\underline{s}^2 \cdot \text{polynomial}(n,p)} \\
&\leq C \cdot \bar{A}^4 K_a^2 \cdot \ln^2(np) \left(\|\beta^*\|_1^2 + \frac{1}{n}\|\phi^{\text{TRAIN}}\|_2^2\right) \\
&\leq C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2} \left(\|\beta^*\|_1^2 + \|\phi^{\text{TRAIN}}\|_2^2\right)
\end{aligned}$$

as long as $\underline{s} \geq \frac{\bar{\varepsilon}}{\text{polynomial}(n,p)}$. □

E.5.4 Main result

Proof of Theorem 5.2. We proceed in steps.

1. Decomposition

   By Lemma E.28

   $$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2 \leq 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2 + 2\|\phi^{\text{TEST}}\|_2^2.$$

   Hence

   $$\begin{aligned}
   \mathbb{E}\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2 \leq\ & 2\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right] \\
   & + 2\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right] \\
   & + 2\|\phi^{\text{TEST}}\|_2^2.
   \end{aligned}$$

2. Beneficial case

   By Propositions E.4 and E.5,

   $$\begin{aligned}
   &\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right] \\
   &\leq 2\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right] + 2\mathbb{E}\left[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\beta} - \boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right] \\
   &\leq C_1 C_2 \cdot \bar{\sigma}^2 \cdot \frac{r^3 \ln^8(np)}{\rho_{\min}^6}\|\beta^*\|_1^2 \left\{1 + \frac{p}{n} + \frac{n}{p} + (n+p)\Delta_E^2 + np\Delta_E^4\right\} \\
   &\quad + C_2 \cdot \frac{r^2 \ln^3(np)}{\rho_{\min}^2}\left(1 + \Delta_E^2\right)\|\phi^{\text{TRAIN}}\|_2^2.
   \end{aligned}$$

3. Adverse case

   Write

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}-\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right]\\&\leq 2\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right]+2\mathbb{E}\left[\|\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right].\end{aligned}$$

   Focusing on the latter term,

$$\|\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\leq\|\boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2\|\beta^*\|_1^2\leq n\bar{A}^2\|\beta^*\|_1^2$$

   hence by Lemma E.21

$$\mathbb{E}\left[\|\boldsymbol{A}^{\text{TEST}}\beta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right]\leq n\bar{A}^2\|\beta^*\|_1^2\mathbb{P}(\tilde{\mathcal{E}}^c)\leq C\frac{\bar{A}^2\|\beta^*\|_1^2}{n^9p^{10}}$$

   which is clearly dominated by the bound on the beneficial case.

   Focusing on the former term, Cauchy-Schwarz inequality and Lemma E.21 give

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right]&\leq\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]}\sqrt{\mathbb{E}\left[\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right]}\\&\leq\frac{C}{n^5p^5}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]}\end{aligned}$$

   which is dominated by the bound on the beneficial case if

$$\begin{aligned}&\frac{1}{n^5p^5}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]}\\&\leq C_1C_2\cdot\bar{\sigma}^2\cdot\frac{r^3\ln^8(np)}{\rho_{\min}^6}\|\beta^*\|_1^2\left\{1+\frac{p}{n}+\frac{n}{p}+(n+p)\Delta_E^2+np\Delta_E^4\right\}\\&\quad+C_2\cdot\frac{r^2\ln^3(np)}{\rho_{\min}^2}\left(1+\Delta_E^2\right)\|\phi^{\text{TRAIN}}\|_2^2.\end{aligned}$$

   In the proof of Proposition E.6, we have precisely shown

$$\frac{n^{-1}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\beta}\|_2^4\right]}}{n^4p^5}\leq C_2\cdot\frac{r^2\ln^3(np)}{\rho_{\min}^2}\left(\|\beta^*\|_1^2+\|\phi^{\text{TRAIN}}\|_2^2\right).$$

□

Proof of Corollary 5.1. Identical to the proof of Theorem 5.2, appealing to the previous remarks for the appropriate generalizations $(\rho'_{\min}, r', \Delta'_E)$. □

F Error-in-variable balancing weight

The outline of the argument is as follows

1. define TRAIN, TEST, and GENERAL ERROR
2. exposit the general algorithm
3. establish orthogonality, balance, and equivalence properties
4. analyze TRAIN ERROR(more precisely, $\|\hat{\eta} - \eta^*\|_2$)
5. analyze TEST ERROR(more precisely, $\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2$)
6. analyze GENERAL ERROR(more precisely, $\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \alpha_0(\boldsymbol{W}^{\text{TEST}})\|_2^2$)

F.1 Notation and preliminaries

As in Appendix D, we identify `NA` with 0 in $\boldsymbol{Z}$ for the remainder of the appendix. We also use the notation $\boldsymbol{A}$ rather than $\boldsymbol{X}$. Recall that $(\hat{\boldsymbol{\rho}}, \hat{\beta})$ are calculated from TRAIN. We slightly abuse notation by letting n be the number of observations in TRAIN (and also TEST), departing from the notation of the main text. We write $\|\cdot\| = \|\cdot\|_{op}$. We write the proofs without nonlinear dictionaries for clarity. Then we extend our results to allow for nonlinear dictionaries in subsequent remarks. In doing so, we denote the identity map $b : \mathbb{R}^p \to \mathbb{R}^p$ with components $b_j : \mathbb{R} \to \mathbb{R}$. We also let $\bar{A}' = \bar{A}^{d_{\max}}$, $\rho'_{\min} = \frac{\rho_{\min}}{d_{\max}\bar{A}'}$, and $p' = C \cdot d_{\max} p$. Finally, to lighten notation, we abbreviate $b(D_i, A_{i,\cdot})$ as $b(A_{i,\cdot})$ when it is contextually clear.

F.1.1 Errors and SVDs

Consider the following quantities

$$\text{TRAIN ERROR} = \frac{1}{n}\mathbb{E}\left[\sum_{i\in\text{TRAIN}} \{\hat{A}_{i,\cdot}\hat{\eta} - \alpha_0(W_{i,\cdot})\}^2\right]$$

$$\text{TEST ERROR} = \frac{1}{n}\mathbb{E}\left[\sum_{i\in\text{TEST}} \{\hat{A}_{i,\cdot}\hat{\eta} - \alpha_0(W_{i,\cdot})\}^2\right]$$

$$\text{GENERAL ERROR} = \frac{1}{n}\mathbb{E}\left[\sum_{i\in\text{TEST}} \{\hat{\alpha}_i - \alpha_0(W_{i,\cdot})\}^2\right], \quad \hat{\alpha}_i = Z_{i,\cdot}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} = Z_{i,\cdot}\tilde{\eta}.$$

Each ERROR is new because it corresponds to our new error-in-variable balancing weight. The key result is about GENERAL ERROR, in which we analyze a new variant of PCR that does not involve cleaning TEST. As we will see, post multiplying by $\tilde{\eta}$ performs a kind of implicit cleaning. By avoiding explicit cleaning, we preserve independence across rows in TEST, which is critical for our inference argument. En route, we will analyze TRAIN ERROR and TEST ERROR, which are closely related to TRÃIN ERROR and TẼST ERROR. When using a dictionary, the updated estimator is $\hat{\alpha}_i = b(D_i, Z_{i,\cdot}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\eta} = b(D_i, Z_{i,\cdot})\tilde{\eta}$ for an updated definition of $\tilde{\eta}$. The SVDs are as in Appendix E.

F.1.2 Counterfactual moments

In this exposition, we consider the case with technical regressors; take b to be the identity to return to the case of original regressors. In the main text, we provide the counterfactual moment for Example A.1 with the interacted dictionary. We now introduce a more general notation to describe the counterfactual moments for general parameters and general dictionaries. Recall from Appendix A that we consider causal parameters of the form

$$\theta_0 = \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} \theta_i, \quad \theta_i = \mathbb{E}[m(W_{i,\cdot}, \gamma_0)], \quad W_{i,\cdot} = (A_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}).$$

Given a dictionary $b : \mathbb{R}^p \to \mathbb{R}^{p'}$, define

$$b^{\text{SIGNAL}}(W_{i,\cdot}) = b(A_{i,\cdot}), \quad b^{\text{NOISE}}(W_{i,\cdot}) = b(Z_{i,\cdot}).$$

The general counterfactual moment is calculated as follows.

Algorithm F.1 (Counterfactual moment with data cleaning)**.** *Given corrupted training covariates $\boldsymbol{Z}^{\text{TRAIN}} \in \mathbb{R}^{n \times p}$, the dictionary $b : \mathbb{R}^p \to \mathbb{R}^{p'}$, and the formula $m : \mathcal{W} \times \mathbb{L}_2 \to \mathbb{R}$*

1. *Perform data cleaning on $\boldsymbol{Z}^{\text{TRAIN}}$ to obtain $\hat{\boldsymbol{A}}^{\text{TRAIN}} \in \mathbb{R}^{n \times p}$*
2. *For $i \in$ TRAIN calculate $m_{i,\cdot} = m(W_{i,\cdot}, b^{\text{NOISE}}) \in \mathbb{R}^{p'}$*
3. *For $i \in$ TRAIN, calculate $\hat{m}_{i,\cdot}$ from $m_{i,\cdot}$ by overwriting $Z_{i,\cdot}$ with $\hat{A}_{i,\cdot}$*
4. *Calculate $\hat{M} = \frac{1}{n} \sum_{i \in \text{TRAIN}} \hat{m}_{i,\cdot}$*

To specialize this abstract procedure to a specific setting, it is sufficient to describe $\hat{m}_i$. We provide the explicit expressions for $\hat{m}_i$ in each leading example in the proof of Proposition F.1 below.

As theoretical devices, we introduce several related objects. First, we define the counterfactual vectors $\tilde{m}_{i,\cdot}, \hat{m}_{i,\cdot} \in \mathbb{R}^{p'}$ for observation i. The former vector uses clean data, while the latter uses cleaned data. In particular,

$$\tilde{m}_{i,\cdot} = m(W_{i,\cdot}, b^{\text{SIGNAL}}), \quad \hat{m}_{i,\cdot} = m(W_{i,\cdot}, b^{\text{NOISE}}) \text{ overwriting } Z_{i,\cdot} \text{ with } \hat{A}_{i,\cdot}.$$

We concatenate the vectors $\tilde{m}_{i,\cdot}$ as rows in the matrix $\tilde{\boldsymbol{M}}$. We concatenate the vectors $\hat{m}_{i,\cdot}$ as rows in the matrix $\hat{\boldsymbol{M}}$. We refer to these objects as the counterfactual matrices. We also use the counterfactual vectors to define the counterfactual moments $M^*, \hat{M} \in \mathbb{R}^{p'}$:

$$M^* = \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} \alpha_0(W_{i,\cdot}) b\{A_{i,\cdot}^{(\text{LR})}\}, \quad \hat{M} = \frac{1}{n} \sum_{i \in \text{TRAIN}} \hat{m}_{i,\cdot}.$$

Finally, we introduce notation for the covariance matrices $\boldsymbol{G}^*, \hat{\boldsymbol{G}} \in \mathbb{R}^{p' \times p'}$:

$$\boldsymbol{G}^* = \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} b\{A_{i,\cdot}^{(\text{LR})}\}^T b\{A_{i,\cdot}^{(\text{LR})}\}, \quad \hat{\boldsymbol{G}} = \frac{1}{n} b(\hat{\boldsymbol{A}}^{\text{TRAIN}})^T b(\hat{\boldsymbol{A}}^{\text{TRAIN}}).$$

With this additional notation, we write the generalized coefficient as

$$\hat{\eta} = \hat{\boldsymbol{G}}^\dagger \hat{M}^T.$$

F.2 Data cleaning continuity

A desirable property is that data cleaning guarantees for the corrupted regressors imply data cleaning guarantees of the counterfactual moments. We refer to this property as data cleaning continuity. We define the property, then verify that it holds for the leading examples.

Assumption F.1 (Data cleaning continuity)**.** *There exist $C'_m, C''_m < \infty$ such that*

1. $\|\hat{\boldsymbol{M}} - \tilde{\boldsymbol{M}}\|_{2,\infty}^2 \leq C'_m \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2$;
2. $\max_{j \in [p']} |\tilde{m}_{ij}| \leq C''_m$.

In particular, we explicitly characterize (C'_m, C''_m) for many important causal and structural parameters of interest. Recall the variable definitions in Appendix A.

Proposition F.1 (Verifying data cleaning continuity)**.** *The following conditions verify Assumption F.1 for the leading examples. Suppose Assumption 5.1 holds.*

1. *In Example A.1 with the interacted dictionary,* $\hat{m}_{i,\cdot} = (\hat{A}_{i,\cdot}, -\hat{A}_{i,\cdot})$*,* $C'_m = 1$*, and* $C''_m = \bar{A}$*.*

2. *In Example A.2 with the interacted dictionary,* $\hat{m}_{i,\cdot} = (\hat{A}_{i,\cdot}, -\hat{A}_{i,\cdot})$*,* $C'_m = 1$*, and* $C''_m = \bar{A}$ *for the functionals in the numerator and denominator.*

3. *In Example A.3 with the identity dictionary, suppose the counterfactual policy is of the form* $t : A_{i,\cdot} \mapsto t_1 \odot A_{i,\cdot} + t_2$ *where* $t_1, t_2 \in \mathbb{R}^p$*.* $\hat{m}_{i,\cdot} = [(t_1 - \mathbb{1}^T) \odot \hat{A}_{i,\cdot}] + t_2$*,* $C'_m = (\|t_1\|_{\max} + 1)^2$*, and* $C''_m = (\|t_1\|_{\max} + 1)\bar{A} + \|t_2\|_{\max}$*.*

4. *In Example A.4 with the interacted quadratic dictionary,* $\hat{m}_{i,\cdot} = (0, 1, 2D_i, 0, \hat{A}_{i,\cdot}, 2D_i\hat{A}_{i,\cdot})$*,* $C'_m = 4\bar{A}^2$*, and* $C''_m = 2\bar{A}^2$*.*[10]

5. *In Example A.5 with the partially linear dictionary,* $\hat{m}_{i,\cdot} = (1, 0, ..., 0)$*,* $C'_m = 0$*, and* $C''_m = 1$ [11]

6. *In Example A.6 with the partially linear dictionary,* $\hat{m}_{i,\cdot} = (1, 0, ..., 0)$*,* $C'_m = 0$*, and* $C''_m = 1$ *for the functionals in the numerator and denominator.*

7. *In Example A.7 with the interacted dictionary,* $\hat{m}_{i,\cdot} = (V_i, \hat{A}_{i,\cdot}, -V_i, -\hat{A}_{i,\cdot})$*,* $C'_m = 1$*, and* $C''_m = \bar{A}$*.*[12]

Proof. We verify the result for each example.

1. Example A.1. Recall $b(D_i, Z_{i,\cdot}) = \{D_i Z_{i,\cdot}, (1 - D_i)Z_{i,\cdot}\}$. Write

$$m_{i,\cdot} = b(1, Z_{i,\cdot}) - b(0, Z_{i,\cdot}) = \{Z_{i,\cdot}, 0\} - \{0, Z_{i,\cdot}\} = (Z_{i,\cdot}, -Z_{i,\cdot}).$$

   Hence

$$\hat{m}_{i,\cdot} = (\hat{A}_{i,\cdot}, -\hat{A}_{i,\cdot}), \quad \tilde{m}_{i,\cdot} = (A_{i,\cdot}, -A_{i,\cdot}).$$

[10] Likewise for any polynomial of D_i interacted with $Z_{i,\cdot}$.

[11] Recall that, to estimate a weighted balancing weight, we propose estimating an unweighted balancing weight then applying the weighting. The verification here is for the unweighted balancing weight that will be weighted.

[12] Recall that, to estimate a local balancing weight, we propose estimating a global balancing weight then applying the localization. The verification here is for the global balancing weight that will be localized.

2. Example A.2 is analogous to Example A.1.

3. Example A.3. Recall $b(Z_{i,\cdot}) = Z_{i,\cdot}$. Write

$$m_{i,\cdot} = b\{t(Z_{i,\cdot})\} - b(Z_{i,\cdot}) = t(Z_{i,\cdot}) - Z_{i,\cdot} = t_1 \odot Z_{i,\cdot} + t_2 - Z_{i,\cdot} = [(t_1 - \mathbb{1}^T) \odot Z_{i,\cdot}] + t_2.$$

Hence

$$\hat{m}_{i,\cdot} = [(t_1 - \mathbb{1}^T) \odot \hat{A}_{i,\cdot}] + t_2, \quad \tilde{m}_{i,\cdot} = [(t_1 - \mathbb{1}^T) \odot A_{i,\cdot}] + t_2$$

4. Example A.4. Recall $b(D_i, Z_{i,\cdot}) = (1, D_i, D_i^2, Z_{i,\cdot}, D_i Z_{i,\cdot}, D_i^2 Z_{i,\cdot})$. Write

$$m_{i,\cdot} = \nabla_d b(D_i, Z_{i,\cdot}) = \nabla_d (1, D_i, D_i^2, Z_{i,\cdot}, D_i Z_{i,\cdot}, D_i^2 Z_{i,\cdot}) = (0, 1, 2D_i, 0, Z_{i,\cdot}, 2D_i Z_{i,\cdot}).$$

Hence

$$\hat{m}_{i,\cdot} = (0, 1, 2D_i, 0, \hat{A}_{i,\cdot}, 2D_i \hat{A}_{i,\cdot}), \quad \tilde{m}_{i,\cdot} = (0, 1, 2D_i, 0, A_{i,\cdot}, 2D_i A_{i,\cdot}).$$

5. Example A.5. Let $b(D_i, Z_{i,\cdot}) = \{D_i, \tilde{b}(Z_{i,\cdot})\}$. Write

$$m_{i,\cdot} = b(1, Z_{i,\cdot}) - b(0, Z_{i,\cdot}) = \{1, \tilde{b}(Z_{i,\cdot})\} - \{0, \tilde{b}(Z_{i,\cdot})\} = (1, 0, ..., 0).$$

Hence

$$\hat{m}_{i,\cdot} = (1, 0, ..., 0), \quad \tilde{m}_{i,\cdot} = (1, 0, ..., 0).$$

6. Example A.6 is analogous to Example A.5.

7. Example A.7 is analogous to Example A.1.

□

F.3 Estimator properties

F.3.1 Orthogonality

The goal of this section is to establish orthogonality properties for the analysis to follow. In order to formalize these orthogonality properties, we formally define η^*. To begin, consider the case without a dictionary. We define $\eta^* \in \mathbb{R}^p$ as the unique solution to the following optimization problem across TRAIN and TEST:

$$\min_{\eta \in \mathbb{R}^p} \|\eta\|_2 \text{ s.t. } \eta \in \operatorname{argmin} \left\| \begin{bmatrix} \alpha_0(\boldsymbol{W}^{\text{TRAIN}}) \\ \alpha_0(\boldsymbol{W}^{\text{TEST}}) \end{bmatrix} - \begin{bmatrix} \boldsymbol{A}^{(\text{LR}),\text{TRAIN}} \\ \boldsymbol{A}^{(\text{LR}),\text{TEST}} \end{bmatrix} \eta \right\|_2^2.$$

η^* is not the quantity of interest, but rather a theoretical device. It defines the unique, minimal-norm, low-rank, linear approximation to the balancing weight α_0. The theoretical device η^* is new to the literature since the balancing weight is new to the literature. Our ultimate goal is to define and analyze an estimator close to $\alpha_0(W_{i,\cdot})$ in generalized mean square error while adhering to the conditional independence criterion of Proposition 4.2.

Remark F.1 (Dictionary)**.** *When using a dictionary, we update our definition of* $\eta^* \in \mathbb{R}^{p'}$ *as the unique solution to the following optimization problem across* TRAIN *and* TEST*:*

$$\min_{\eta \in \mathbb{R}^{p'}} \|\eta\|_2 \ \textit{s.t.}\ \eta \in \operatorname{argmin} \left\| \begin{bmatrix} \alpha_0(\boldsymbol{W}^{\text{TRAIN}}) \\ \alpha_0(\boldsymbol{W}^{\text{TEST}}) \end{bmatrix} - \begin{bmatrix} b\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\} \\ b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\} \end{bmatrix} \eta \right\|_2^2 .$$

Lemma F.1. *Suppose Assumptions 5.6 and 5.8 hold. Then,*

$$\hat{\boldsymbol{V}}_{k,\perp}^T \hat{\eta} = 0, \quad \boldsymbol{V}_\perp^T \eta^* = (\boldsymbol{V}'_\perp)^T \eta^* = 0$$

Proof. We show each result

1. $\hat{\boldsymbol{V}}_{k,\perp}^T \hat{\eta} = 0$

   It suffices to show $\hat{\eta} \in \text{ROW}(\hat{\boldsymbol{A}}^{\text{TRAIN}})$ then appeal to the same reasoning as Lemma E.2. This result follows from a generalization of Agarwal et al. (2020a, Property 4.1): $\hat{\eta}$ is the unique solution to the program

   $$\min_{\eta \in \mathbb{R}^p} \|\eta\|_2 \text{ s.t. } \eta \in \operatorname{argmin} -2\hat{M}\eta + \eta^T \hat{\boldsymbol{G}} \eta$$

   where $\hat{M} \in \text{ROW}(\hat{\boldsymbol{A}}^{\text{TRAIN}})$ by Assumption 5.8 and $\text{ROW}(\hat{\boldsymbol{G}}) = \text{ROW}\{(\hat{\boldsymbol{A}}^{\text{TRAIN}})^T \hat{\boldsymbol{A}}^{\text{TRAIN}}\} = \text{ROW}(\hat{\boldsymbol{A}}^{\text{TRAIN}})$. Therefore $\hat{\eta} \in \text{ROW}(\hat{\boldsymbol{A}}^{\text{TRAIN}})$.

2. $\boldsymbol{V}_\perp^T \eta^* = (\boldsymbol{V}'_\perp)^T \eta^* = 0$. See Lemma E.2

□

Remark F.2 (Dictionary)**.** *Lemma E.2 continues to hold with the updated definitions of the SVDs in Remark E.1.*

Lemma F.2. *Deterministically,* $\hat{\boldsymbol{G}}\hat{\eta} = \hat{M}^T$ *and* $\boldsymbol{G}^* \eta^* = (M^*)^T$.

Proof. Immediate from the FOC in the definitions of $(\hat{\eta}, \eta^*)$. □

F.3.2 Balance

Without any further assumptions, we demonstrate that the error-in-variables balancing weight confers a finite sample balancing property. We articulate the balancing property in terms of the coefficient $\hat{\eta}$.

Proposition F.2 (Finite sample balance)**.** *For any finite training sample size n, and any dictionary of basis functions b, the coefficient $\hat{\eta}$ balances the cleaned actual regressors with the corresponding cleaned counterfactuals in the sense that*

$$\frac{1}{n}\sum_{i\in\text{TRAIN}} b(\hat{A}_{i,\cdot})\cdot\hat{\omega}_i = \frac{1}{n}\sum_{i\in\text{TRAIN}} \hat{m}_{i,\cdot}.$$

where $\hat{\omega}_i \in \mathbb{R}$ are balancing weights computed from $\hat{\eta}$: for each $i \in$ TRAIN,

$$\hat{\omega}_i = b(\hat{A}_{i,\cdot})\hat{\eta}.$$

Proof. By Lemma F.2,

$$\frac{1}{n}\sum_{i\in\text{TRAIN}} b(\hat{A}_{i,\cdot})^T b(\hat{A}_{i,\cdot})\hat{\eta} = \hat{\boldsymbol{G}}\hat{\eta} = \hat{M}^T = \frac{1}{n}\sum_{i\in\text{TRAIN}} (\hat{m}_{i,\cdot})^T.$$

□

In words, $\hat{\eta}$ serves to balance actual observations with counterfactual queries. This result is somewhat abstract, so we instantiate it for a leading case. Specifically, Proposition 4.3 considers ATE (Example A.1) with the interacted dictionary.

Proof of Proposition 4.3. By Proposition F.2,

$$\frac{1}{n}\sum_{i\in\text{TRAIN}} b(D_i, \hat{A}_{i,\cdot})\cdot\hat{\omega}_i = \frac{1}{n}\sum_{i\in\text{TRAIN}} \hat{m}_{i,\cdot}.$$

Focusing on the RHS, by Proposition F.1

$$\hat{m}_{i,\cdot} = (\hat{A}_{i,\cdot}, -\hat{A}_{i,\cdot}).$$

Next, turning to the LHS,

$$b(D_i, \hat{A}_{i,\cdot}) = \{D_i\hat{A}_{i,\cdot}, (1-D_i)\hat{A}_{i,\cdot}\}$$

and

$$\hat{\omega}_i = b(D_i, \hat{A}_{i,\cdot})\hat{\eta} = D_i\hat{A}_{i,\cdot}\hat{\eta}^{\text{TREAT}} + (1-D_i)\hat{A}_{i,\cdot}\hat{\eta}^{\text{UNTREAT}} = D_i \cdot \hat{\omega}_i^{\text{TREAT}} - (1-D_i)\hat{\omega}_i^{\text{UNTREAT}}.$$

Therefore

$$\begin{aligned} b(D_i, \hat{A}_{i,\cdot}) \cdot \hat{\omega}_i &= \{D_i\hat{A}_{i,\cdot}, (1-D_i)\hat{A}_{i,\cdot}\} \cdot \{D_i \cdot \hat{\omega}_i^{\text{TREAT}} - (1-D_i)\hat{\omega}_i^{\text{UNTREAT}}\} \\ &= \{D_i\hat{A}_{i,\cdot} \cdot \hat{\omega}_i^{\text{TREAT}}, (1-D_i)\hat{A}_{i,\cdot} \cdot (-\hat{\omega}_i^{\text{UNTREAT}})\}. \end{aligned}$$

In summary, matching components of the LHS and RHS,

$$\begin{aligned} \frac{1}{n}\sum_{i\in\text{TRAIN}} D_i\hat{A}_{i,\cdot} \cdot \hat{\omega}_i^{\text{TREAT}} &= \frac{1}{n}\sum_{i\in\text{TRAIN}} \hat{A}_{i,\cdot}; \\ \frac{1}{n}\sum_{i\in\text{TRAIN}} (1-D_i)\hat{A}_{i,\cdot} \cdot (-\hat{\omega}_i^{\text{UNTREAT}}) &= \frac{1}{n}\sum_{i\in\text{TRAIN}} (-\hat{A}_{i,\cdot}). \end{aligned}$$

□

F.3.3 Equivalence

Finally, we relate the properties of our estimators to an equivalence property that is well documented in the causal inference literature; see e.g. Ben-Michael et al. (2021); Bruns-Smith and Feller (2022) for recent summaries.[13] We demonstrate that a version of the equivalence property holds on TRAIN , in a stronger sense than previously known, but it does not hold on TEST.

Previous work (Robins et al., 2007; Chattopadhyay and Zubizarreta, 2021) shows that a certain equivalence holds for treatment effects when using OLS with the interacted dictionary (without data cleaning). We begin by generalizing this equivalence in three ways: (i) for our entire class of semiparametric and nonparametric estimands, (ii) for any square integrable dictionary, (iii) for estimation with or without data cleaning. In order to document the equivalence, we define, for $i \in$ TRAIN,

$$\tilde{\gamma}(D_i, Z_{i,\cdot}) = b(D, \hat{X}_{i,\cdot})\hat{\beta}, \quad \tilde{\alpha}(D_i, Z_{i,\cdot}) = b(D, \hat{X}_{i,\cdot})\hat{\eta}.$$

To lighten notation we also define the operator $\mathbb{E}_{\text{TRAIN}}[\cdot] = \frac{1}{m}\sum_{i\in\text{TRAIN}}[\cdot]$.

[13] We thank Avi Feller and David Bruns-Smith for suggesting this connection.

Proposition F.3 (Equivalence in TRAIN)**.** *For any linear functional within the class defined by Assumption G.1 and for any square integrable dictionary b, the outcome, balancing weight, and doubly robust estimators coincide on the training set:*

$$\mathbb{E}_{\text{TRAIN}}[m(W_{i,\cdot}, \tilde{\gamma})] = \mathbb{E}_{\text{TRAIN}}[Y_i \tilde{\alpha}(D_i, Z_{i,\cdot})] = \mathbb{E}_{\text{TRAIN}}[m(W_{i,\cdot}, \tilde{\gamma}) + \tilde{\alpha}(D_i, Z_{i,\cdot})\{Y_i - \tilde{\gamma}(D_i, Z_{i,\cdot})\}].$$

We state the result with data cleaning, but it also holds without data cleaning.

Proof. We proceed in steps.

1. To prove the second equality, we appeal to the FOC for $\hat{\eta}$ summarized by Lemma F.2: $\hat{\eta}^T \hat{\boldsymbol{G}} = \hat{M}$. Multiplying the LHS by $\hat{\beta}$,

   $$\hat{\eta}^T \hat{\boldsymbol{G}} \hat{\beta} = \hat{\eta}^T \mathbb{E}_{\text{TRAIN}}[b(D, \hat{X}_{i,\cdot})^T b(D, \hat{X}_{i,\cdot})] \hat{\beta} = \mathbb{E}_{\text{TRAIN}}[\tilde{\alpha}(D_i, Z_{i,\cdot}) \tilde{\gamma}(D_i, Z_{i,\cdot})].$$

   Multiplying the RHS by $\hat{\beta}$,

   $$\hat{M} \hat{\beta} = \mathbb{E}_{\text{TRAIN}}[\hat{m}_{i,\cdot}] \hat{\beta} = \mathbb{E}_{\text{TRAIN}}[m(W_{i,\cdot}, \tilde{\gamma})].$$

   In summary

   $$\mathbb{E}_{\text{TRAIN}}[\tilde{\alpha}(D_i, Z_{i,\cdot}) \tilde{\gamma}(D_i, Z_{i,\cdot})] = \mathbb{E}_{\text{TRAIN}}[m(W_{i,\cdot}, \tilde{\gamma})]$$

   which implies the result.

2. To prove the first equality, we appeal to the FOC for $\hat{\beta}$: $\hat{\beta}^T \hat{\boldsymbol{G}} = \mathbb{E}_{\text{TRAIN}}[Y_i b(D_i, \hat{X}_{i,\cdot})]$. Multiplying the RHS by $\hat{\eta}$,

   $$\mathbb{E}_{\text{TRAIN}}[Y_i b(D_i, \hat{X}_{i,\cdot})] \hat{\eta} = \mathbb{E}_{\text{TRAIN}}[Y_i \tilde{\alpha}(D_i, Z_{i,\cdot})].$$

   Multiplying the LHS by $\hat{\eta}$ and appealing to the previous result

   $$\hat{\beta}^T \hat{\boldsymbol{G}} \hat{\eta} = \hat{\eta}^T \hat{\boldsymbol{G}} \hat{\beta} = \mathbb{E}_{\text{TRAIN}}[\tilde{\alpha}(D_i, Z_{i,\cdot}) \tilde{\gamma}(D_i, Z_{i,\cdot})] = \mathbb{E}_{\text{TRAIN}}[m(W_{i,\cdot}, \tilde{\gamma})].$$

□

However, our estimator involves sample splitting and implicit data cleaning to break dependence, motivated by our goal of inference after data cleaning. In our estimator of the causal parameter, we use, for $i \in$ TEST,

$$\hat{\gamma}(D_i, Z_{i,\cdot}) = b(D, Z_{i,\cdot}) \hat{\beta}, \quad \hat{\alpha}(D_i, Z_{i,\cdot}) = b(D, Z_{i,\cdot}) \hat{\eta}.$$

Proposition F.4 (Non-equivalence in TEST)**.** *For any linear functional within the class defined by Assumption G.1 and for any square integrable dictionary b, the outcome, balancing weight, and doubly robust estimators generically do not coincide on the test set:*

$$\mathbb{E}_{\text{TEST}}[m(Z_{i,\cdot}, \hat{\gamma})] \neq \mathbb{E}_{\text{TEST}}[Y_i \hat{\alpha}(D_i, Z_{i,\cdot})] \neq \mathbb{E}_{\text{TEST}}[m(Z_{i,\cdot}, \hat{\gamma}) + \hat{\alpha}(D_i, Z_{i,\cdot})\{Y_i - \hat{\gamma}(D_i, Z_{i,\cdot})\}].$$

Proof. The FOCs for $(\hat{\beta}, \hat{\eta})$ hold for TRAIN after data cleaning, which is how $(\hat{\beta}, \hat{\eta})$ are estimated. They do not hold for TEST, especially since we do not clean the test covariates. □

As such, the equivalence property is relevant for gaining intuition into the relationship between $(\hat{\beta}, \hat{\eta})$. However, the equivalence does not hold for our final estimator of the causal parameter because of sample splitting and implicit data cleaning.

F.4 Training error

In this argument, all objects indexed by a sample split correspond to TRAIN. For this reason, we suppress superscipt TRAIN. Note that $(M^*, \boldsymbol{G}^*, \eta^*)$ are constructed from (TRAIN, TEST) while $(\hat{M}, \hat{\boldsymbol{G}}, \hat{\eta}, \hat{\boldsymbol{A}}, \boldsymbol{A}^{(\text{LR})})$ are constructed from TRAIN.

F.4.1 Decomposition

Lemma F.3. *Deterministically,*

$$\|\hat{\boldsymbol{A}}\hat{\eta} - \boldsymbol{A}^{(\text{LR})}\eta^*\|_2^2 \leq C \left\{ \|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T (\hat{\eta} - \eta^*)\|_1 \cdot \Delta_{RR} \vee \|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \|\eta^*\|_1^2 \right\}$$

where

$$\Delta_{RR} := n \cdot \left\{ \|\hat{M}^T - (M^*)^T\|_{\max} + \|\boldsymbol{G}^* - \hat{\boldsymbol{G}}\|_{\max} \|\eta^*\|_1 \right\}.$$

Proof. We proceed in steps.

1. Rewrite $\|\hat{\boldsymbol{A}}\hat{\eta} - \boldsymbol{A}^{(\text{LR})}\eta^*\|_2^2$

   Write

   $$\|\hat{\boldsymbol{A}}\hat{\eta} - \boldsymbol{A}^{(\text{LR})}\eta^*\|_2^2 = \|\hat{\boldsymbol{A}}\hat{\eta} \pm \hat{\boldsymbol{A}}\eta^* - \boldsymbol{A}^{(\text{LR})}\eta^*\|_2^2 \leq 2\|\hat{\boldsymbol{A}}(\hat{\eta} - \eta^*)\|_2^2 + 2\|(\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})})\eta^*\|_2^2.$$

2. First term

Next, write

$$\begin{aligned}\frac{1}{n}\|\hat{\boldsymbol{A}}(\hat{\eta}-\eta^*)\|_2^2 &= \frac{1}{n}(\hat{\eta}-\eta^*)^T\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}(\hat{\eta}-\eta^*)\\ &= \frac{1}{n}(\hat{\eta}-\eta^*)^T\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}(\hat{\eta}-\eta^*)\\ &= (\hat{\eta}-\eta^*)^T\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T\hat{\boldsymbol{G}}(\hat{\eta}-\eta^*)\\ &\le \|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_1\cdot\|\hat{\boldsymbol{G}}(\hat{\eta}-\eta^*)\|_{\max}.\end{aligned}$$

The latter factor is bounded as

$$\begin{aligned}&\|\hat{\boldsymbol{G}}(\hat{\eta}-\eta^*)\|_{\max}\\ &= \|\hat{\boldsymbol{G}}\hat{\eta}\pm\hat{M}^T\pm(M^*)^T\pm\boldsymbol{G}^*\eta^*-\hat{\boldsymbol{G}}\eta^*\|_{\max}\\ &\le \|\hat{\boldsymbol{G}}\hat{\eta}-\hat{M}^T\|_{\max}+\|\hat{M}^T-(M^*)^T\|_{\max}+\|(M^*)^T-\boldsymbol{G}^*\eta^*\|_{\max}+\|\boldsymbol{G}^*\eta^*-\hat{\boldsymbol{G}}\eta^*\|_{\max}\\ &\le 0+\|\hat{M}^T-(M^*)^T\|_{\max}+0+\|\boldsymbol{G}^*-\hat{\boldsymbol{G}}\|_{\max}\|\eta^*\|_1\\ &= \|\hat{M}^T-(M^*)^T\|_{\max}+\|\boldsymbol{G}^*-\hat{\boldsymbol{G}}\|_{\max}\|\eta^*\|_1.\end{aligned}$$

where in the second inequality we appeal to Lemma F.2 and we apply Holder inequality row-wise.

3. Second term

   Finally, write

$$\|(\hat{\boldsymbol{A}}-\boldsymbol{A}^{(\mathrm{LR})})\eta^*\|_2^2 \le \|\hat{\boldsymbol{A}}-\boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}^2\|\eta^*\|_1^2.$$

□

Remark F.3 (Dictionary)**.** *The generalization of Lemma F.3 is*

$$\|b(\hat{\boldsymbol{A}})\hat{\eta}-b\{\boldsymbol{A}^{(\mathrm{LR})}\}\eta^*\|_2^2 \le C\left[\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_1\cdot\Delta_{RR}\vee\|b(\hat{\boldsymbol{A}})-b\{\boldsymbol{A}^{(\mathrm{LR})}\}\|_{2,\infty}^2\|\eta^*\|_1^2\right].$$

F.4.2 Parameter

Lemma F.4. *Let the conditions of Lemma F.1 hold. Then*

$$\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 \le C\left\{\frac{1}{\hat{s}_k^2}\|\hat{\boldsymbol{A}}-\boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}^2\|\eta^*\|_1^2+\frac{1}{\hat{s}_k^4}p\cdot\Delta_{RR}^2\right\}.$$

Proof. As in Lemma E.5,

$$\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 \le \frac{2}{\hat{s}_k^2}\Big\{\|\hat{\boldsymbol{A}}\hat{\eta}-\boldsymbol{A}^{(\mathrm{LR})}\eta^*\|_2^2 + \|\boldsymbol{A}^{(\mathrm{LR})}-\hat{\boldsymbol{A}}\|_{2,\infty}^2\|\eta^*\|_1^2\Big\}.$$

Using Lemma F.3, we conclude that

$$\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 \le \frac{C}{\hat{s}_k^2}\Big\{\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_1\cdot\Delta_{RR} \vee \|\boldsymbol{A}^{(\mathrm{LR})}-\hat{\boldsymbol{A}}\|_{2,\infty}^2\|\eta^*\|_1^2\Big\}.$$

There are two cases, in which each of the two terms dominates:

1. $\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 \le C\frac{1}{\hat{s}_k^2}\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_1\cdot\Delta_{RR}$. In this case,

   $$\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 \le C\frac{1}{\hat{s}_k^2}\sqrt{p}\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2\cdot\Delta_{RR}.$$

   Dividing both sides by $\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2$

   $$\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2 \le C\frac{1}{\hat{s}_k^2}\sqrt{p}\cdot\Delta_{RR}$$

   hence

   $$\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 \le C\frac{1}{\hat{s}_k^4}p\cdot\Delta_{RR}^2.$$

2. $\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 \le C\frac{1}{\hat{s}_k^2}\|\hat{\boldsymbol{A}}-\boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}^2\|\eta^*\|_1^2$.

□

Remark F.4 (Dictionary)**.** *The generalization of Lemma F.4 is*

$$\|\hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T(\hat{\eta}-\eta^*)\|_2^2 \le C\left[\frac{1}{\hat{s}_{r'}^2}\|b(\hat{\boldsymbol{A}})-b\{\boldsymbol{A}^{(\mathrm{LR})}\}\|_{2,\infty}^2\|\eta^*\|_1^2 + \frac{1}{\hat{s}_{r'}^4}p'\cdot\Delta_{RR}^2\right]$$

using the SVD of Remark E.1.

Lemma F.5. *Let the conditions of Lemma F.1 hold. Then*

$$\|\hat{\eta}-\eta^*\|_2^2 \le C\left\{\|\boldsymbol{V}\boldsymbol{V}^T-\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T\|^2\|\eta^*\|_2^2 + \frac{1}{\hat{s}_k^2}\|\hat{\boldsymbol{A}}-\boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}^2\|\eta^*\|_1^2 + \frac{1}{\hat{s}_k^4}p\cdot\Delta_{RR}^2\right\}.$$

Proof. As in Lemma E.5

$$\|\hat{\eta}-\eta^*\|_2^2 = \|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 + \|\hat{\boldsymbol{V}}_{k,\perp}\hat{\boldsymbol{V}}_{k,\perp}^T\eta^*\|_2^2.$$

By Lemma F.4,

$$\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta}-\eta^*)\|_2^2 \le C\left\{\frac{1}{\hat{s}_k^2}\|\hat{\boldsymbol{A}}-\boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}^2\|\eta^*\|_1^2 + \frac{1}{\hat{s}_k^4}p\cdot\Delta_{RR}^2\right\}.$$

As in Lemma E.5

$$\|\hat{\boldsymbol{V}}_{k,\perp}\hat{\boldsymbol{V}}_{k,\perp}^T\eta^*\|_2^2 \le \|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T\|^2\|\eta^*\|_2^2.$$

□

Remark F.5 (Dictionary)**.** *The generalization of Lemma F.5 is*

$$\|\hat{\eta} - \eta^*\|_2^2 \le C\left[\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T\|^2\|\eta^*\|_2^2 + \frac{1}{\hat{s}_{r'}^2}\|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\|\eta^*\|_1^2 + \frac{1}{\hat{s}_{r'}^4}p'\cdot\Delta_{RR}^2\right]$$

using the SVD of Remark E.1.

F.4.3 High probability events

We wish to control Δ_{RR}, which in turn depends on

$$\|\hat{M}^T - (M^*)^T\|_{\max}, \quad \|\boldsymbol{G}^* - \hat{\boldsymbol{G}}\|_{\max}.$$

Define $\mathcal{E} := \cap_{k=1}^9\mathcal{E}_k$ where $\mathcal{E}_1$ to $\mathcal{E}_5$ are given in Appendix D, and

$$\mathcal{E}_6 = \left\{\max_{j\in[p]}\left|\frac{1}{n}\sum_{i\in\text{TRAIN}}\{\tilde{m}_{ij} - \mathbb{E}[\tilde{m}_{ij}]\}\right| \le C\cdot C_m''\sqrt{\frac{\log(np)}{n}}\right\};$$

$$\mathcal{E}_7 = \left\{\max_{j\in[p]}\left|\frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}}\{\alpha_0(W_{i,\cdot})A_{ij} - \mathbb{E}[\alpha_0(W_{i,\cdot})A_{ij}]\}\right| \le C\cdot\bar{\alpha}\bar{A}\sqrt{\frac{\log(np)}{n}}\right\};$$

$$\mathcal{E}_8 = \left\{\max_{j,k\in[p]}\left|\frac{1}{n}\sum_{i\in\text{TRAIN}}\{A_{ij}A_{ik} - \mathbb{E}[A_{ij}A_{ik}]\}\right| \le C\cdot\bar{A}^2\sqrt{\frac{\log(np)}{n}}\right\};$$

$$\mathcal{E}_9 = \left\{\max_{j,k\in[p]}\left|\frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}}\{A_{ij}A_{ik} - \mathbb{E}[A_{ij}A_{ik}]\}\right| \le C\cdot\bar{A}^2\sqrt{\frac{\log(np)}{n}}\right\}.$$

We show $\mathcal{E}_j$ hold w.p. $1 - \frac{2}{n^{10}p^{10}}$ en route to controlling $\|\hat{M}^T - (M^*)^T\|_{\max}$ and $\|\boldsymbol{G}^* - \hat{\boldsymbol{G}}\|_{\max}$ and hence Δ_{RR}.

Lemma F.6. *Under Assumption F.1*

$$\mathbb{P}(\mathcal{E}_6^c) \le \frac{2}{n^{10}p^{10}}.$$

Proof. By Assumption F.1, $\tilde{m}_{ij} \le C_m''$. Hence by Hoeffding, for any $j \in [p]$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i\in\text{TRAIN}}\{\tilde{m}_{ij} - \mathbb{E}[\tilde{m}_{ij}\}]\right| \ge t\right) \le 2\exp\left\{-\frac{2n^2t^2}{n(2C_m'')^2}\right\}.$$

Taking the union bound over $j \in [p]$

$$\mathbb{P}\left(\max_{j\in[p]}\left|\frac{1}{n}\sum_{i\in\text{TRAIN}}\{\tilde{m}_{ij}-\mathbb{E}[\tilde{m}_{ij}]\}\right|\geq t\right)\leq 2p\exp\left\{-\frac{2n^2t^2}{n(2C_m'')^2}\right\}=\frac{2}{n^{10}p^{10}}.$$

□

Remark F.6 (Dictionary)**.** *A generalization of Lemma F.6 sufficient for our analysis is for*

$$\mathcal{E}_6'=\left\{\max_{j\in[p']}\left|\frac{1}{n}\sum_{i\in\text{TRAIN}}\{\tilde{m}_{ij}-\mathbb{E}[\tilde{m}_{ij}]\}\right|\leq C\cdot C_m''\sqrt{\frac{\log(np')}{n}}\right\}.$$

Lemma F.7. *Under Assumption 5.1 and* $\|\alpha_0\|_\infty \leq \bar{\alpha}$*,*

$$\mathbb{P}(\mathcal{E}_7^c)\leq\frac{2}{n^{10}p^{10}}.$$

Proof. By Assumption 5.1, $|\alpha_0(W_{i,\cdot})A_{ij}|\leq\bar{\alpha}\bar{A}$. Hence by Hoeffding, for any $j\in[p]$

$$\mathbb{P}\left(\left|\frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}}\{\alpha_0(W_{i,\cdot})A_{ij}-\mathbb{E}[\alpha_0(W_{i,\cdot})A_{ij}]\}\right|\geq t\right)\leq 2\exp\left\{-\frac{2(2n)^2t^2}{(2n)(2\bar{\alpha}\bar{A})^2)}\right\}.$$

Taking the union bound over $j\in[p]$

$$\mathbb{P}\left(\max_{j\in[p]}\left|\frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}}\{\alpha_0(W_{i,\cdot})A_{ij}-\mathbb{E}[\alpha_0(W_{i,\cdot})A_{ij}]\}\right|\geq t\right)\leq 2p\exp\left\{-\frac{2(2n)^2t^2}{(2n)(2\bar{\alpha}\bar{A})^2)}\right\}=\frac{2}{n^{10}p^{10}}.$$

□

Remark F.7 (Dictionary)**.** *A generalization of Lemma F.7 sufficient for our analysis is for*

$$\mathcal{E}_7'=\left\{\max_{j\in[p']}\left|\frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}}\{\alpha_0(W_{i,\cdot})b_j(A_{i,\cdot})-\mathbb{E}[\alpha_0(W_{i,\cdot})b_j(A_{i,\cdot})]\}\right|\leq C\cdot\bar{\alpha}\bar{A}'\sqrt{\frac{\log(np')}{n}}\right\}.$$

Lemma F.8. *Suppose Assumptions 5.1, G.2, G.1, and F.1 hold, and* $\|\alpha_0\|_\infty\leq\bar{\alpha}$*. Then*

$$\|\hat{M}-M^*\|_{\max}|\{\mathcal{E}_6,\mathcal{E}_7\}\leq\Delta_M=\sqrt{C_m'}\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{A}}-\boldsymbol{A}\|_{2,\infty}+C\cdot(C_m''+\bar{\alpha}\bar{A})\sqrt{\frac{\ln(np)}{n}}+\bar{\alpha}\cdot\Delta_E.$$

Proof. We proceed in steps.

1. Decomposition

Write

$$\begin{aligned}\hat{M} - M^* &= \frac{1}{n}\sum_{i\in\text{TRAIN}}(\hat{m}_{i,\cdot} - \tilde{m}_{i,\cdot}) \\ &\quad + \frac{1}{n}\sum_{i\in\text{TRAIN}}\{\tilde{m}_{i,\cdot} - \mathbb{E}[\tilde{m}_{i,\cdot}]\} \\ &\quad + \frac{1}{n}\sum_{i\in\text{TRAIN}}\mathbb{E}[\tilde{m}_{i,\cdot}] - \frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}}\mathbb{E}[\alpha_0(W_{i,\cdot})A_{i,\cdot}] \\ &\quad + \frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}}\{\mathbb{E}[\alpha_0(W_{i,\cdot})A_{i,\cdot}] - \alpha_0(W_{i,\cdot})A_{i,\cdot}\} \\ &\quad + \frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}}\{\alpha_0(W_{i,\cdot})A_{i,\cdot} - \alpha_0(W_{i,\cdot})A_{i,\cdot}^{(\text{LR})}\} \\ &= \sum_{k=1}^{5} R^{(k)}.\end{aligned}$$

By triangle inequality, it suffices to bound the j-th component of each difference in absolute value, i.e. to bound $R_j^{(k)}$.

2. First term

   We analyze

$$\begin{aligned}\{R_j^{(1)}\}^2 &= \left\{\frac{1}{n}\sum_{i\in\text{TRAIN}}(\hat{m}_{ij} - \tilde{m}_{ij})\right\}^2 \\ &\le \frac{1}{n}\sum_{i\in\text{TRAIN}}(\hat{m}_{ij} - \tilde{m}_{ij})^2 \\ &\le \frac{1}{n}\|\hat{\boldsymbol{M}} - \tilde{\boldsymbol{M}}\|_{2,\infty}^2 \\ &\le \frac{1}{n}C'_m\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2\end{aligned}$$

   appealing to Assumption F.1. Hence

$$|R_j^{(1)}| \le \sqrt{C'_m}\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}.$$

3. Second term

   Write

$$R_j^{(2)} = \frac{1}{n}\sum_{i\in\text{TRAIN}}\{\tilde{m}_{ij} - \mathbb{E}[\tilde{m}_{ij}]\},$$

   then appeal to $\mathcal{E}_6$.

4. Third term

   Write

$$R_j^{(3)} = \frac{1}{n}\sum_{i\in\text{TRAIN}} \mathbb{E}[\tilde{m}_{ij}] - \frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}} \mathbb{E}[\alpha_0(W_{i,\cdot})A_{ij}] = 0$$

   by Riesz representation and ex-ante identical distribution of folds in the random partition (TRAIN, TEST). In particular, since $b_j^{\text{SIGNAL}} \in \mathbb{L}_2(\mathcal{W})$,

$$\mathbb{E}[\tilde{m}_{ij}] = \mathbb{E}[m(W_{i,\cdot}, b_j^{\text{SIGNAL}})] = \mathbb{E}[\alpha_0(W_{i,\cdot})b_j^{\text{SIGNAL}}(W_{i,\cdot})] = \mathbb{E}[\alpha_0(W_{i,\cdot})b_j(A_{i,\cdot})].$$

5. Fourth term

   Write

$$-R_j^{(4)} = \frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}} \{\alpha_0(W_{i,\cdot})A_{ij} - \mathbb{E}[\alpha_0(W_{i,\cdot})A_{ij}]\}$$

   then appeal to $\mathcal{E}_7$.

6. Fifth term

   Write the final term as

$$|R_j^{(5)}| = \left|\frac{1}{2n}\sum_{i\in\text{TRAIN,TEST}} \alpha_0(W_{i,\cdot})E_{ij}^{(\text{LR})}\right| \leq \bar{\alpha}\cdot\Delta_E$$

   where $\alpha_0(W_{i,\cdot}) \leq \bar{\alpha}$.

□

Remark F.8 (Dictionary)**.** *The generalization of Lemma F.8 is*

$$\|\hat{M} - M^*\|_{\max}|\{\mathcal{E}_6', \mathcal{E}_7'\} \leq \Delta_M' = \sqrt{C_m'}\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty} + C\cdot(C_m'' + \bar{\alpha}\bar{A}')\sqrt{\frac{\ln(np')}{n}} + \bar{\alpha}\cdot\Delta_E'.$$

Lemma F.9. *Under Assumption 5.1,*

$$\mathbb{P}(\mathcal{E}_8^c) \leq \frac{2}{n^{10}p^{10}}.$$

Proof. By Assumption 5.1, $|A_{ij}A_{ik}| \leq \bar{A}^2$. Hence by Hoeffding, for any $j, k \in [p]$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i\in\text{TRAIN}}\{A_{ij}A_{ik} - \mathbb{E}[A_{ij}A_{ik}]\}\right| \geq t\right) \leq 2\exp\left\{-\frac{2n^2t^2}{n(2\bar{A}^2)^2}\right\}.$$

Taking the union bound over $j, k \in [p]$

$$\mathbb{P}\left(\max_{j,k\in[p]}\left|\frac{1}{n}\sum_{i\in\text{TRAIN}}\{A_{ij}A_{ik} - \mathbb{E}[A_{ij}A_{ik}]\}\right| \geq t\right) \leq 2p^2\exp\left\{-\frac{2n^2t^2}{n(2\bar{A}^2)^2}\right\} = \frac{2}{n^{10}p^{10}}.$$

□

Remark F.9 (Dictionary)**.** *A generalization of Lemma F.9 sufficient for our analysis is for*

$$\mathcal{E}_8' = \left\{ \max_{j,k\in[p']} \left| \frac{1}{n} \sum_{i\in \text{TRAIN}} \{b_j(A_{i,\cdot})b_k(A_{i,\cdot}) - \mathbb{E}[b_j(A_{i,\cdot})b_k(A_{i,\cdot})]\} \right| \leq C \cdot (\bar{A}')^2 \sqrt{\frac{\log(np')}{n}} \right\}.$$

Lemma F.10. *Under Assumption 5.1,*

$$\mathbb{P}(\mathcal{E}_9^c) \leq \frac{2}{n^{10}p^{10}}.$$

Proof. By Assumption 5.1, $|A_{ij}A_{ik}| \leq \bar{A}^2$. Hence by Hoeffding, for any $j, k \in [p]$

$$\mathbb{P}\left(\left| \frac{1}{2n} \sum_{i\in \text{TRAIN,TEST}} \{A_{ij}A_{ik} - \mathbb{E}[A_{ij}A_{ik}]\} \right| \geq t \right) \leq 2\exp\left\{ -\frac{2(2n)^2t^2}{(2n)(2\bar{A}^2)^2} \right\}.$$

Taking the union bound over $j, k \in [p]$

$$\mathbb{P}\left(\left| \frac{1}{2n} \sum_{i\in \text{TRAIN,TEST}} \{A_{ij}A_{ik} - \mathbb{E}[A_{ij}A_{ik}]\} \right| \geq t \right) \leq 2p^2\exp\left\{ -\frac{2(2n)^2t^2}{(2n)(2\bar{A}^2)^2} \right\} = \frac{2}{n^{10}p^{10}}.$$

□

Remark F.10 (Dictionary)**.** *A generalization of Lemma F.10 sufficient for our analysis is for*

$$\mathcal{E}_9' = \left\{ \max_{j,k\in[p']} \left| \frac{1}{2n} \sum_{i\in \text{TRAIN,TEST}} \{b_j(A_{i,\cdot})b_k(A_{i,\cdot}) - \mathbb{E}[b_j(A_{i,\cdot})b_k(A_{i,\cdot})]\} \right| \leq C \cdot (\bar{A}')^2 \sqrt{\frac{\log(np')}{n}} \right\}.$$

Lemma F.11. *Suppose Assumptions 5.1 holds. Then*

$$\|\hat{\boldsymbol{G}} - \boldsymbol{G}^*\|_{\max} | \{\mathcal{E}_8, \mathcal{E}_9\} \leq \Delta_G$$

where

$$\Delta_G = (\bar{A} + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}) \frac{1}{\sqrt{n}} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty} + C \cdot \bar{A}^2 \sqrt{\frac{\ln(np)}{n}} + C \cdot \bar{A}\Delta_E.$$

Proof. We proceed in steps.

1. Decomposition

Write

$$\begin{aligned}
\hat{\boldsymbol{G}} - \boldsymbol{G}^* &= \frac{1}{n} \sum_{i \in \text{TRAIN}} \{\hat{A}_{i,\cdot}^T \hat{A}_{i,\cdot} - \hat{A}_{i,\cdot}^T A_{i,\cdot}\} \\
&+ \frac{1}{n} \sum_{i \in \text{TRAIN}} \{\hat{A}_{i,\cdot}^T A_{i,\cdot} - A_{i,\cdot}^T A_{i,\cdot}\} \\
&+ \frac{1}{n} \sum_{i \in \text{TRAIN}} \{A_{i,\cdot}^T A_{i,\cdot} - \mathbb{E}[A_{i,\cdot}^T A_{i,\cdot}]\} \\
&+ \frac{1}{n} \sum_{i \in \text{TRAIN}} \mathbb{E}[A_{i,\cdot}^T A_{i,\cdot}] - \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} \mathbb{E}[A_{i,\cdot}^T A_{i,\cdot}] \\
&+ \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} \{\mathbb{E}[A_{i,\cdot}^T A_{i,\cdot}] - A_{i,\cdot}^T A_{i,\cdot}\} \\
&+ \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} \{A_{i,\cdot}^T A_{i,\cdot} - A_{i,\cdot}^T A_{i,\cdot}^{(\text{LR})}\} \\
&+ \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} [A_{i,\cdot}^T A_{i,\cdot}^{(\text{LR})} - \{A_{i,\cdot}^{(\text{LR})}\}^T A_{i,\cdot}^{(\text{LR})}] \\
&= \sum_{\ell=1}^{7} S^{(\ell)}.
\end{aligned}$$

By triangle inequality, it suffices to bound the j,k-th component of each difference in absolute value, i.e. to bound $S_{jk}^{(\ell)}$.

2. First term

Write

$$S_{jk}^{(1)} = \frac{1}{n} \sum_{i \in \text{TRAIN}} \hat{A}_{ij}(\hat{A}_{ik} - A_{ik}) \leq \|\hat{\boldsymbol{A}}\|_{\max} \cdot \frac{1}{n} \sum_{i \in \text{TRAIN}} (\hat{A}_{ik} - A_{ik}).$$

Hence

$$\begin{aligned}
\{S_{jk}^{(1)}\}^2 &\leq \|\hat{\boldsymbol{A}}\|_{\max}^2 \cdot \left\{ \frac{1}{n} \sum_{i \in \text{TRAIN}} (\hat{A}_{ik} - A_{ik}) \right\}^2 \\
&\leq \|\hat{\boldsymbol{A}}\|_{\max}^2 \cdot \frac{1}{n} \sum_{i \in \text{TRAIN}} (\hat{A}_{ik} - A_{ik})^2 \\
&\leq \|\hat{\boldsymbol{A}}\|_{\max}^2 \frac{1}{n} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2.
\end{aligned}$$

In summary

$$|S_{jk}^{(1)}| \leq \|\hat{\boldsymbol{A}}\|_{\max} \frac{1}{\sqrt{n}} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}.$$

By Assumption 5.1

$$\|\hat{\boldsymbol{A}}\|_{\max} \leq \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{\max} + \|\boldsymbol{A}\|_{\max} \leq \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty} + \bar{A}.$$

3. Second term

   Write

$$S_{jk}^{(2)} = \frac{1}{n} \sum_{i \in \text{TRAIN}} (\hat{A}_{ij} - A_{ij}) A_{ik} \leq \bar{A} \cdot \frac{1}{n} \sum_{i \in \text{TRAIN}} (\hat{A}_{ij} - A_{ij}).$$

   Hence

$$\begin{aligned} \{S_{jk}^{(2)}\}^2 &\leq \bar{A}^2 \cdot \left\{ \frac{1}{n} \sum_{i \in \text{TRAIN}} (\hat{A}_{ij} - A_{ij}) \right\}^2 \\ &\leq \bar{A}^2 \cdot \frac{1}{n} \sum_{i \in \text{TRAIN}} (\hat{A}_{ij} - A_{ij})^2 \\ &\leq \bar{A}^2 \frac{1}{n} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2. \end{aligned}$$

   In summary

$$|S_{jk}^{(2)}| \leq \bar{A} \frac{1}{\sqrt{n}} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}.$$

4. Third term

   Write

$$S_{jk}^{(3)} = \frac{1}{n} \sum_{i \in \text{TRAIN}} \{A_{ij} A_{ik} - \mathbb{E}[A_{ij} A_{ik}]\}$$

   then appeal to $\mathcal{E}_8$.

5. Fourth term

   Write

$$S_{jk}^{(4)} = \frac{1}{n} \sum_{i \in \text{TRAIN}} \mathbb{E}[A_{ij} A_{ik}] - \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} \mathbb{E}[A_{ij} A_{ik}] = 0$$

   by ex-ante identical distribution of folds in the random partition (TRAIN, TEST).

6. Fifth term

   Write

$$-S_{jk}^{(5)} = \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} \{A_{ij} A_{ik} - \mathbb{E}[A_{ij} A_{ik}]\}$$

   then appeal to $\mathcal{E}_9$.

7. Sixth term

   By Assumption 5.1

$$S_{jk}^{(6)} = \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} A_{ij} E_{ik}^{(\text{LR})} \leq \bar{A} \Delta_E.$$

8. Seventh term

   By Assumption 5.1 and Lemma C.4

$$S_{jk}^{(7)} = \frac{1}{2n} \sum_{i \in \text{TRAIN,TEST}} E^{(\text{LR})} A_{ik}^{(\text{LR})} \leq \|\boldsymbol{E}^{(\text{LR})}\|_{\max} \|\boldsymbol{A}^{(\text{LR})}\|_{\max} \leq 3\bar{A}\Delta_E.$$

□

Remark F.11 (Dictionary)**.** *The generalization of Lemma F.11 is*

$$\|\hat{\boldsymbol{G}} - \boldsymbol{G}^*\|_{\max} | \{\mathcal{E}_8', \mathcal{E}_9'\} \leq \Delta_G'$$

where

$$\Delta_G' = \{\bar{A}' + \|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\|_{2,\infty}\} \frac{1}{\sqrt{n}} \|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\|_{2,\infty} + C \cdot (\bar{A}')^2 \sqrt{\frac{\ln(np')}{n}} + C \cdot \bar{A}' \Delta_E'.$$

Lemma F.12. *Suppose the conditions of Lemmas F.8 and F.11 hold. Then*

$$\begin{aligned} &\Delta_{RR}^2 | \{\mathcal{E}_6, \mathcal{E}_7, \mathcal{E}_8, \mathcal{E}_9\} \leq C n^2 \|\eta^*\|_1^2 \\ &\cdot \left\{ \left(\bar{A} + \sqrt{C_m'}\right)^2 \frac{1}{n} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4 + (C_m'' + \bar{\alpha}\bar{A} + \bar{A}^2)^2 \frac{\ln(np)}{n} + (\bar{A} + \bar{\alpha})^2 \Delta_E^2 \right\}. \end{aligned}$$

Proof. We proceed in steps.

1. Decomposition

   Write

$$\Delta_{RR} = n \cdot \left\{ \|\hat{M}^T - (M^*)^T\|_{\max} + \|\boldsymbol{G}^* - \hat{\boldsymbol{G}}\|_{\max} \|\eta^*\|_1 \right\} \leq n(\Delta_M + \Delta_G \|\eta^*\|_1).$$

   Hence with probability $1 - O\{(np)^{-10}\}$

$$\Delta_{RR}^2 \leq C n^2 (\Delta_M^2 + \Delta_G^2 \|\eta^*\|_1^2).$$

2. Δ_M

   By Lemma F.8, with probability $1 - O\{(np)^{-10}\}$

$$\Delta_M = \sqrt{C_m'} \frac{1}{\sqrt{n}} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty} + C \cdot (C_m'' + \bar{\alpha}\bar{A}) \sqrt{\frac{\ln(np)}{n}} + \bar{\alpha} \cdot \Delta_E.$$

   Note that

$$\Delta_M^2 \leq C \left\{ C_m' \frac{1}{n} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + (C_m'' + \bar{\alpha}\bar{A})^2 \frac{\ln(np)}{n} + \bar{\alpha}^2 \cdot \Delta_E^2 \right\}.$$

3. Δ_G

   By Lemma F.11, with probability $1 - O\{(np)^{-10}\}$

$$\Delta_G = (\bar{A} + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty})\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty} + C \cdot \bar{A}^2\sqrt{\frac{\ln(np)}{n}} + C \cdot \bar{A}\Delta_E.$$

   Note that

$$\Delta_G^2 \leq C\left\{(\bar{A}^2 + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2)\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + \bar{A}^4\frac{\ln(np)}{n} + \bar{A}^2\Delta_E^2\right\}.$$

4. Combining terms

   Certain terms in Δ_M^2 and Δ_G^2 can be combined. In particular,

$$\begin{aligned}
&C'_m\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + (\bar{A}^2 + \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2)\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \\
&\leq (\sqrt{C'_m} + \bar{A})^2\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + \frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4 \\
&\leq (\sqrt{C'_m} + \bar{A})^2\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4
\end{aligned}$$

   using $\|\hat{\boldsymbol{A}}-\boldsymbol{A}\|_{2,\infty}^2 \leq \|\hat{\boldsymbol{A}}-\boldsymbol{A}\|_{2,\infty}^4$. This inequality implicitly appeals to $\|\hat{\boldsymbol{A}}-\boldsymbol{A}\|_{2,\infty}^2 \geq 1$, which we can enforce by taking $1 \vee \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2$ since the latter is a diverging sequence.

□

Remark F.12 (Dictionary)**.** *The generalization of Lemma F.12 is*

$$\begin{aligned}
&\Delta_{RR}^2|\{\mathcal{E}'_6, \mathcal{E}'_7, \mathcal{E}'_8, \mathcal{E}'_9\} \leq Cn^2\|\eta^*\|_1^2 \\
&\cdot \left[\left(\bar{A}'C'_b + \sqrt{C'_m}\right)^2\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4 + \{C''_m + \bar{\alpha}\bar{A}' + (\bar{A}')^2\}^2\frac{\ln(np')}{n} + (\bar{A}' + \bar{\alpha})^2(\Delta'_E)^2\right].
\end{aligned}$$

Proof. When combining terms

$$\begin{aligned}
&C'_m\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + \{(\bar{A}')^2 + \|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\|_{2,\infty}^2\}\frac{1}{n}\|b(\hat{\boldsymbol{A}}) - b(\boldsymbol{A})\|_{2,\infty}^2 \\
&\leq (\sqrt{C'_m} + \sqrt{C'_b}\bar{A})^2\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + \frac{1}{n}(C'_b)^2\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4 \\
&\leq (\sqrt{C'_m} + C'_b\bar{A}')^2\frac{1}{n}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4.
\end{aligned}$$

□

F.4.4 Collecting results

Lemma F.13. *Suppose the conditions of Theorem 5.1 hold. Further suppose Assumptions 5.6, 5.8, 5.9, and 5.10 hold and* $\|\alpha_0\|_\infty \le \bar{\alpha}$. *With probability at least* $1 - O\{(np)^{-10}\}$

$$\|\hat{\boldsymbol{V}}_r \hat{\boldsymbol{V}}_r^T (\hat{\eta} - \eta^*)\|_2^2 \le (2) + (3)$$

defined below.

Proof. By Lemma F.4

$$\|\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^T (\hat{\eta} - \eta^*)\|_2^2 \le C \left\{ \frac{1}{\hat{s}_k^2} \|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \|\eta^*\|_1^2 + \frac{1}{\hat{s}_k^4} p \cdot \Delta_{RR}^2 \right\}.$$

Consider each term with $k = r$, which we bound by (2) and (3), respectively.

1. $\frac{1}{\hat{s}_r^2} \|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \|\eta^*\|_1^2$

   Note that

   $$\|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\text{LR})}\|_{2,\infty}^2 \le C \left\{ \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2 \right\}.$$

   By Lemmas D.10 and D.15, with probability at least $1 - O\{(np)^{-10}\}$,

   $$\begin{aligned} &\left\| \hat{\boldsymbol{A}} - \boldsymbol{A} \right\|_{2,\infty}^2 \\ &\le \frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4} \left\{ r + \frac{n(n+p)\Delta_{H,op}^2 + n \left\| \boldsymbol{E}^{(\text{LR})} \right\|^2 + \ln(np) np \bar{A}^2}{s_r^2} \right\} \ln^2(np) \\ &\quad + C \left\| \boldsymbol{E}^{(\text{LR})} \right\|_{2,\infty}^2. \end{aligned}$$

   In summary

   $$\begin{aligned} (2) = &\frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4} \frac{\ln^2(np)}{\hat{s}_r^2} \|\eta^*\|_1^2 \\ &\cdot \left\{ r + \frac{n(n+p)\Delta_{H,op}^2 + n \left\| \boldsymbol{E}^{(\text{LR})} \right\|^2 + \ln(np) np \bar{A}^2}{s_r^2} + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2 \right\}. \end{aligned}$$

2. $\frac{1}{\hat{s}_r^4} p \cdot \Delta_{RR}^2$

   By Lemma F.12 with probability at least $1 - O\{(np)^{-10}\}$

   $$\begin{aligned} &\Delta_{RR}^2 \\ &\le C n^2 \|\eta^*\|_1^2 \left\{ \left(\bar{A} + \sqrt{C_m'} \right)^2 \frac{1}{n} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4 + (C_m'' + \bar{\alpha}\bar{A} + \bar{A}^2)^2 \frac{\ln(np)}{n} + (\bar{A} + \bar{\alpha})^2 \Delta_E^2 \right\} \\ &\le C(\sqrt{C_m'} + C_m'' + \bar{\alpha}\bar{A} + \bar{A}^2)^2 n^2 \|\eta^*\|_1^2 \left\{ \frac{1}{n} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4 + \frac{\ln(np)}{n} + \Delta_E^2 \right\}. \end{aligned}$$

By Lemma D.10 and Lemma D.15, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}\right\|_{2,\infty}^4 \\
&\leq \frac{C(K_a + \bar{K}\bar{A})^4}{\rho_{\min}^8}\left\{r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right\}^2 \ln^4(np) \\
&\quad + C\left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|_{2,\infty}^4.
\end{aligned}$$

Therefore

$$\begin{aligned}
\Delta_{RR}^2 \leq\ & C(\sqrt{C_m'} + C_m'' + \bar{\alpha}\bar{A} + \bar{A}^2)^2 \frac{(K_a + \bar{K}\bar{A})^4}{\rho_{\min}^8} \ln^4(np) \cdot n^2\|\eta^*\|_1^2 \\
&\left[\frac{1}{n}\left\{r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right\}^2 + \frac{1}{n}\|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^4 + \frac{\ln(np)}{n} + \Delta_E^2\right].
\end{aligned}$$

Hence

$$\begin{aligned}
(3) =\ & C(\sqrt{C_m'} + C_m'' + \bar{\alpha}\bar{A} + \bar{A}^2)^2 \frac{(K_a + \bar{K}\bar{A})^4}{\rho_{\min}^8} \ln^4(np) \cdot \frac{pn^2}{\hat{s}_r^4}\|\eta^*\|_1^2 \\
&\left[\frac{1}{n}\left\{r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right\}^2 + \frac{1}{n}\|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^4 + \frac{\ln(np)}{n} + \Delta_E^2\right].
\end{aligned}$$

□

Lemma F.14. *Suppose the conditions of Lemma F.13 hold. With probability at least* $1 - O\{(np)^{-10}\}$

$$\|\hat{\eta} - \eta^*\|_2^2 \leq (1) + (2) + (3)$$

where (1) *is defined below and* (2), (3) *are defined above.*

Proof. By Lemma F.5

$$\|\hat{\eta} - \eta^*\|_2^2 \leq C\left\{\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T\|^2\|\eta^*\|_2^2 + \frac{1}{\hat{s}_k^2}\|\hat{\boldsymbol{A}} - \boldsymbol{A}^{(\mathrm{LR})}\|_{2,\infty}^2\|\eta^*\|_1^2 + \frac{1}{\hat{s}_k^4}p \cdot \Delta_{RR}^2\right\}.$$

Consider each term, which we bound by (1), (2), and (3), respectively. The second and third term are already bounded in Lemma F.13. Therefore we focus on the first term. As in Lemma E.15, with probability at least $1 - O\{(np)^{-10}\}$

$$\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\| \leq \frac{C}{\rho_{\min}s_r}\left\{(\sqrt{n} + \sqrt{p})\Delta_{H,op} + \|\boldsymbol{E}^{(\mathrm{LR})}\| + \bar{A}\sqrt{\ln(np)}\sqrt{p}\right\}.$$

Hence

$$(1) = C\frac{\|\eta^*\|_2^2}{\rho_{\min}^2 s_r^2}\left\{(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\text{LR})}\|^2 + \bar{A}^2\ln(np)\cdot p\right\}.$$

□

Remark F.13 (Dictionary)**.** *In the generalization of Lemmas F.13 and F.14, note the new appearances of C_b' in (2) and (3):*

$$(1) = C\frac{\|\eta^*\|_2^2}{(\rho_{\min}')^2 s_{r'}^2}\left\{(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\text{LR})}\|^2 + \bar{A}^2\ln(np)\cdot p\right\};$$

$$(2) = \frac{CC_b'(K_a+\bar{K}\bar{A})^2}{\rho_{\min}^4}\frac{\ln^2(np)}{\hat{s}_{r'}^2}\|\eta^*\|_1^2$$
$$\cdot\left[r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2} + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2 + \|b(\boldsymbol{A}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\right];$$

$$(3) = C\{\sqrt{C_m'} + C_m'' + \bar{\alpha}\bar{A}' + C_b'\bar{A}' + (\bar{A}')^2\}^2\frac{(K_a+\bar{K}\bar{A})^4}{\rho_{\min}^8}\ln^4(np)\cdot\frac{p'n^2}{\hat{s}_{r'}^4}\|\eta^*\|_1^2$$
$$\left[\frac{1}{n}\left\{r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right\}^2 + \frac{1}{n}\|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^4 + \frac{\ln(np')}{n} + (\Delta_E')^2\right].$$

Proof. For (1), see Remark E.8. For (2) and (3) recall that Remark F.4 gives

$$\|\hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T(\hat{\eta}-\eta^*)\|_2^2 \le C\left[\frac{1}{\hat{s}_{r'}^2}\|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\|\eta^*\|_1^2 + \frac{1}{\hat{s}_{r'}^4}p'\cdot\Delta_{RR}^2\right].$$

Consider each term, which we bound by (2) and (3), respectively.

1. $\frac{1}{\hat{s}_{r'}^2}\|b(\hat{\boldsymbol{A}}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\|\eta^*\|_1^2$

   Note that

   $$\|b(\hat{\boldsymbol{A}}) - \{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2 \le C\left\{C_b'\|\hat{\boldsymbol{A}}-\boldsymbol{A}\|_{2,\infty}^2 + \|b(\boldsymbol{A}) - b\{\boldsymbol{A}^{(\text{LR})}\}\|_{2,\infty}^2\right\}.$$

   By Lemmas D.10 and D.15, with probability at least $1 - O\{(np)^{-10}\}$,

   $$\left\|\hat{\boldsymbol{A}}-\boldsymbol{A}\right\|_{2,\infty}^2$$
   $$\le \frac{C(K_a+\bar{K}\bar{A})^2}{\rho_{\min}^4}\left\{r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2}\right\}\ln^2(np)$$
   $$+ C\left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^2.$$

In summary

$$(2) = \frac{CC_b'(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4} \frac{\ln^2(np)}{\hat{s}_{r'}^2} \|\eta^*\|_1^2 \cdot \left[r + \frac{n(n+p)\Delta_{H,op}^2 + n \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|^2 + \ln(np) np \bar{A}^2}{s_r^2} + \|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^2 + \|b(\boldsymbol{A}) - b\{\boldsymbol{A}^{(\mathrm{LR})}\}\|_{2,\infty}^2 \right].$$

2. $\frac{1}{\hat{s}_{r'}^4} p' \cdot \Delta_{RR}^2$

By Remark F.12 with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\Delta_{RR}^2 \\
&\le Cn^2 \|\eta^*\|_1^2 \left[\left(\bar{A}' C_b' + \sqrt{C_m'} \right)^2 \frac{1}{n} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4 + \{C_m'' + \bar{\alpha}\bar{A}' + (\bar{A}')^2\}^2 \frac{\ln(np')}{n} + (\bar{A}' + \bar{\alpha})^2 (\Delta_E')^2 \right] \\
&\le C\{\sqrt{C_m'} + C_m'' + \bar{\alpha}\bar{A}' + C_b'\bar{A}' + (\bar{A}')^2\}^2 n^2 \|\eta^*\|_1^2 \left\{ \frac{1}{n} \|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^4 + \frac{\ln(np')}{n} + (\Delta_E')^2 \right\}.
\end{aligned}$$

By Lemma D.10 and Lemma D.15, with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}
&\left\| \hat{\boldsymbol{A}} - \boldsymbol{A} \right\|_{2,\infty}^4 \\
&\le \frac{C(K_a + \bar{K}\bar{A})^4}{\rho_{\min}^8} \left\{ r + \frac{n(n+p)\Delta_{H,op}^2 + n \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|^2 + \ln(np) np \bar{A}^2}{s_r^2} \right\}^2 \ln^4(np) \\
&\quad + C \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|_{2,\infty}^4.
\end{aligned}$$

Therefore

$$\begin{aligned}
\Delta_{RR}^2 \le{}& C\{\sqrt{C_m'} + C_m'' + \bar{\alpha}\bar{A}' + C_b'\bar{A}' + (\bar{A}')^2\}^2 \frac{(K_a + \bar{K}\bar{A})^4}{\rho_{\min}^8} \ln^4(np) \cdot n^2 \|\eta^*\|_1^2 \\
&\left[\frac{1}{n} \left\{ r + \frac{n(n+p)\Delta_{H,op}^2 + n \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|^2 + \ln(np) np \bar{A}^2}{s_r^2} \right\}^2 + \frac{1}{n} \|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^4 + \frac{\ln(np')}{n} + (\Delta_E')^2 \right].
\end{aligned}$$

Hence

$$\begin{aligned}
(3) = {}& C\{\sqrt{C_m'} + C_m'' + \bar{\alpha}\bar{A}' + C_b'\bar{A}' + (\bar{A}')^2\}^2 \frac{(K_a + \bar{K}\bar{A})^4}{\rho_{\min}^8} \ln^4(np) \cdot \frac{p' n^2}{\hat{s}_{r'}^4} \|\eta^*\|_1^2 \\
&\left[\frac{1}{n} \left\{ r + \frac{n(n+p)\Delta_{H,op}^2 + n \left\| \boldsymbol{E}^{(\mathrm{LR})} \right\|^2 + \ln(np) np \bar{A}^2}{s_r^2} \right\}^2 + \frac{1}{n} \|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^4 + \frac{\ln(np')}{n} + (\Delta_E')^2 \right].
\end{aligned}$$

□

Proposition F.5 (Projected TRAIN ERROR)**.** *Suppose the conditions of Theorem 5.1 hold. Further suppose Assumptions 5.6, 5.8, 5.9, and 5.10 hold, and* $\|\alpha_0\|_\infty \le \bar{\alpha}$*. Let* $k = r$ *and*

$$\rho_{\min} \gg \tilde{C}\sqrt{r}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right), \quad \tilde{C} := C\bar{A}\Big(\kappa + \bar{K} + K_a\Big).$$

Then with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\eta} - \eta^*)\|_2^2 \le C\bar{A}^{10}(\sqrt{C_m'} + C_m'' + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^4 \\ \cdot r^4 \cdot \ln^{10}(np) \cdot \frac{\|\eta^*\|_1^2}{\rho_{\min}^8}\left(\frac{1}{np} + \frac{1}{p^2} + \frac{n}{p^3} + \frac{1}{p}\Delta_E^2 + \frac{n}{p}\Delta_E^4\right).\end{aligned}$$

Proof. We proceed in steps.

1. Recall the inequalities

$$s_r^2 \ge C\frac{np}{r}, \quad \left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 \le np\Delta_E^2, \quad \left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^2 \le n\Delta_E^2, \quad \left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^4 \le n^2\Delta_E^4.$$

   Further,

$$\Delta_{H,op}^2 \le C \cdot \bar{A}^2(\kappa + \bar{K} + K_a)^2\ln^3(np)$$

   Moreover, $n(n+p)\Delta_{H,op}^2$ dominates $\ln(np)np\bar{A}^2$.

2. Simplifying the first term on the RHS of the bound in Lemma F.13

$$\begin{aligned}(2) &= \frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4}\frac{\ln^2(np)}{\hat{s}_r^2}\|\eta^*\|_1^2 \\ &\quad \cdot \left\{r + \frac{n(n+p)\Delta_{H,op}^2 + n\left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 + \ln(np)np\bar{A}^2}{s_r^2} + \|\boldsymbol{E}^{(\text{LR})}\|_{2,\infty}^2\right\} \\ &\le \frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4}\frac{\ln^2(np)}{\hat{s}_r^2}\|\eta^*\|_1^2\left\{r + \frac{n(n+p)\Delta_{H,op}^2 + n^2p\Delta_E^2}{s_r^2} + n\Delta_E^2\right\} \\ &\le \frac{C(K_a + \bar{K}\bar{A})^2}{\rho_{\min}^4}\frac{r\ln^2(np)}{\hat{s}_r^2}\|\eta^*\|_1^2\left\{1 + \frac{(n+p)}{p}\Delta_{H,op}^2 + n\Delta_E^2\right\} \\ &\le \frac{C(K_a + \bar{K})^2\bar{A}^4(\kappa + \bar{K} + K_a)^2}{\rho_{\min}^4}\frac{r\ln^5(np)}{\hat{s}_r^2}\|\eta^*\|_1^2\left(1 + \frac{n}{p} + n\Delta_E^2\right)\end{aligned}$$

   where we bound $(K_a + \bar{K}\bar{A})^2 \le \bar{A}^2(K_a + \bar{K})^2$. By Lemma E.9,

$$\hat{s}_r^2 \gtrsim s_r^2 \ge C\frac{np}{r}$$

so as long as the regularity condition holds,

$$(2) \le \frac{C(K_a+\bar{K})^2\bar{A}^4(\kappa+\bar{K}+K_a)^2}{\rho_{\min}^4}\frac{r^2}{np}\cdot\ln^5(np)\|\eta^*\|_1^2\left(1+\frac{n}{p}+n\Delta_E^2\right)$$
$$\le \frac{C(K_a+\bar{K})^2\bar{A}^4(\kappa+\bar{K}+K_a)^2}{\rho_{\min}^4}r^2\cdot\ln^5(np)\|\eta^*\|_1^2\left(\frac{1}{np}+\frac{1}{p^2}+\frac{1}{p}\Delta_E^2\right).$$

3. Simplifying the second term on the RHS of the bound in Lemma F.13

$$(3) = C(\sqrt{C_m'}+C_m''+\bar{\alpha}\bar{A}+\bar{A}^2)^2\frac{(K_a+\bar{K}\bar{A})^4}{\rho_{\min}^8}\ln^4(np)\cdot\frac{pn^2}{\hat{s}_r^4}\|\eta^*\|_1^2$$
$$\left[\frac{1}{n}\left\{r+\frac{n(n+p)\Delta_{H,op}^2+n\left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|^2+\ln(np)np\bar{A}^2}{s_r^2}\right\}^2+\frac{1}{n}\|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^4+\frac{\ln(np)}{n}+\Delta_E^2\right].$$

Note that $n(n+p)\Delta_{H,op}^2$ dominates $\ln(np)np\bar{A}^2$; $n\left\|\boldsymbol{E}^{(\mathrm{LR})}\right\|^2 \le n^2p\Delta_E^2$; and $\|\boldsymbol{E}^{(\mathrm{LR})}\|_{2,\infty}^4 \le n^2\Delta_E^4$. Moreover,

$$(\sqrt{C_m'}+C_m''+\bar{\alpha}\bar{A}+\bar{A}^2)^2\frac{(K_a+\bar{K}\bar{A})^4}{\rho_{\min}^8} \le \bar{A}^6(\sqrt{C_m'}+C_m''+\bar{\alpha}+\bar{A})^2\frac{(K_a+\bar{K})^4}{\rho_{\min}^8}.$$

Hence

$$(3) \le \bar{A}^6(\sqrt{C_m'}+C_m''+\bar{\alpha}+\bar{A})^2\frac{(K_a+\bar{K})^4}{\rho_{\min}^8}\ln^4(np)\cdot\frac{pn^2}{\hat{s}_r^4}\|\eta^*\|_1^2$$
$$\left[\frac{1}{n}\left\{r+\frac{n(n+p)\Delta_{H,op}^2+n^2p\Delta_E^2}{s_r^2}\right\}^2+\frac{\ln(np)}{n}+\Delta_E^2+n\Delta_E^4\right].$$

Since $s_r^2 \ge C\frac{np}{r}$,

$$(3) \le \bar{A}^6(\sqrt{C_m'}+C_m''+\bar{\alpha}+\bar{A})^2\frac{(K_a+\bar{K})^4}{\rho_{\min}^8}\ln^4(np)\cdot\frac{pn^2}{\hat{s}_r^4}\|\eta^*\|_1^2$$
$$\left[\frac{1}{n}\left\{r+\frac{r(n+p)}{p}\Delta_{H,op}^2+rn\Delta_E^2\right\}^2+\frac{\ln(np)}{n}+\Delta_E^2+n\Delta_E^4\right].$$

Recall that $\Delta_{H,op}^2 \le C\cdot\bar{A}^2(\kappa+\bar{K}+K_a)^2\ln^3(np)$ so that

$$(3) \le \bar{A}^{10}(\sqrt{C_m'}+C_m''+\bar{\alpha}+\bar{A})^2\frac{(K_a+\bar{K})^4(\kappa+\bar{K}+K_a)^4}{\rho_{\min}^8}\ln^4(np)\cdot\frac{pn^2}{\hat{s}_r^4}$$
$$\left[\frac{1}{n}\left\{r+\left(r+\frac{rn}{p}\right)\ln^3(np)+rn\Delta_E^2\right\}^2+\frac{\ln(np)}{n}+\Delta_E^2+n\Delta_E^4\right].$$

Within the final factor

$$\begin{aligned}\frac{1}{n}\left\{r+\left(r+\frac{rn}{p}\right)\ln^3(np)+rn\Delta_E^2\right\}^2 &\le \frac{C}{n}\left\{\left(r+\frac{rn}{p}\right)\ln^3(np)+rn\Delta_E^2\right\}^2\\ &= C\frac{r^2}{n}\left\{\left(1+\frac{n}{p}\right)\ln^3(np)+n\Delta_E^2\right\}^2\\ &\le C\frac{r^2}{n}\left\{\left(1+\frac{n}{p}+\frac{n^2}{p^2}\right)\ln^6(np)+n^2\Delta_E^4\right\}\\ &= Cr^2\left\{\left(\frac{1}{n}+\frac{1}{p}+\frac{n}{p^2}\right)\ln^6(np)+n\Delta_E^4\right\}\end{aligned}$$

which dominates both $\frac{\ln(np)}{n}$ and $n\Delta_E^4$. In summary

$$\begin{aligned}(3) \le \bar{A}^{10}(\sqrt{C'_m}+C''_m+\bar{\alpha}+\bar{A})^2\frac{(K_a+\bar{K})^4(\kappa+\bar{K}+K_a)^4}{\rho_{\min}^8}r^2\cdot\ln^{10}(np)\\ \cdot\frac{pn^2}{\hat{s}_r^4}\|\eta^*\|_1^2\left(\frac{1}{n}+\frac{1}{p}+\frac{n}{p^2}+\Delta_E^2+n\Delta_E^4\right).\end{aligned} \tag{38}$$

By Lemma E.9,

$$\hat{s}_r^4 \gtrsim s_r^4 \ge C\frac{n^2p^2}{r^2}$$

so that

$$\frac{pn^2}{\hat{s}_r^4}\|\eta^*\|_1^2 \le Cr^2\frac{pn^2}{n^2p^2}\|\eta^*\|_1^2 = C\frac{r^2}{p}\|\eta^*\|_1^2.$$

In summary

$$\begin{aligned}(3) \le C\bar{A}^{10}(\sqrt{C'_m}+C''_m+\bar{\alpha}+\bar{A})^2\frac{(K_a+\bar{K})^4(\kappa+\bar{K}+K_a)^4}{\rho_{\min}^8}r^4\cdot\ln^{10}(np)\\ \cdot\|\eta^*\|_1^2\left(\frac{1}{np}+\frac{1}{p^2}+\frac{n}{p^3}+\frac{1}{p}\Delta_E^2+\frac{n}{p}\Delta_E^4\right).\end{aligned}$$

4. We have shown

$$\begin{aligned}(2) &\le \frac{C\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2}{\rho_{\min}^4}r^2\cdot\ln^5(np)\|\eta^*\|_1^2\left(\frac{1}{np}+\frac{1}{p^2}+\frac{1}{p}\Delta_E^2\right);\\ (3) &\le C\bar{A}^{10}(\sqrt{C'_m}+C''_m+\bar{\alpha}+\bar{A})^2\frac{(K_a+\bar{K})^4(\kappa+\bar{K}+K_a)^4}{\rho_{\min}^8}r^4\cdot\ln^{10}(np)\\ &\quad\cdot\|\eta^*\|_1^2\left(\frac{1}{np}+\frac{1}{p^2}+\frac{n}{p^3}+\frac{1}{p}\Delta_E^2+\frac{n}{p}\Delta_E^4\right).\end{aligned}$$

Clearly (3) dominates (2).

□

Remark F.14 (Dictionary). *The generalization of Proposition F.5 is as follows. Suppose*

$$\rho'_{\min} \gg \tilde{C}\sqrt{r'}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right).$$

Then with probability at least $1 - O\{(np)^{-10}\}$

$$\begin{aligned}\|\hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T(\hat{\eta} - \eta^*)\|_2^2 \le C\bar{A}^{10}(C'_b + \sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^4 \\ \cdot (r')^4 \cdot \ln^{10}(np) \cdot \frac{\|\eta^*\|_1^2}{(\rho'_{\min})^8}\left\{\frac{1}{np} + \frac{1}{p^2} + \frac{n}{p^3} + (\Delta'_E)^2 + n(\Delta'_E)^4\right\}.\end{aligned}$$

Proposition F.6 (TRAIN ERROR). *Suppose the conditions of Proposition F.5 hold. Then with probability at least* $1 - O\{(np)^{-10}\}$

$$\begin{aligned}\|\hat{\eta} - \eta^*\|_2^2 \le C\bar{A}^{10}(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^4 \\ \cdot r^4 \cdot \ln^{10}(np) \cdot \frac{\|\eta^*\|_2^2}{\rho_{\min}^8}\left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right).\end{aligned}$$

Proof. We proceed in steps.

1. Recall the inequalities

   $$s_r^2 \ge C\frac{np}{r}, \quad \left\|\boldsymbol{E}^{(\text{LR})}\right\|^2 \le np\Delta_E^2, \quad \left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^2 \le n\Delta_E^2, \quad \left\|\boldsymbol{E}^{(\text{LR})}\right\|_{2,\infty}^4 \le n^2\Delta_E^4.$$

   Further,

   $$\Delta_{H,op}^2 \le C \cdot \bar{A}^2(\kappa + \bar{K} + K_a)^2\ln^3(np)$$

   Moreover, $n(n+p)\Delta_{H,op}^2$ dominates $\ln(np)np\bar{A}^2$.

2. Simplifying the first term on the RHS of the bound in Lemma F.14 as in Proposition E.3

   $$\begin{aligned}(1) &= C\frac{\|\eta^*\|_2^2}{\rho_{\min}^2 s_r^2}\left\{(n+p)\Delta_{H,op}^2 + \|\boldsymbol{E}^{(\text{LR})}\|^2 + \bar{A}^2\ln(np)\cdot p\right\} \\ &\le C \cdot \bar{A}^2(\kappa + \bar{K} + K_a)^2\frac{\|\eta^*\|_2^2}{\rho_{\min}^2} \cdot r\ln^3(np) \cdot \left(\frac{1}{p} + \frac{1}{n} + \Delta_E^2\right).\end{aligned}$$

3. We have shown, by the arguments above and in the proof of Proposition F.5

   $$\begin{aligned}(1) &\le C \cdot \bar{A}^2(\kappa + \bar{K} + K_a)^2\frac{\|\eta^*\|_2^2}{\rho_{\min}^2} \cdot r\ln^3(np) \cdot \left(\frac{1}{p} + \frac{1}{n} + \Delta_E^2\right); \\ (2) &\le \frac{C\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2}{\rho_{\min}^4}r^2 \cdot \ln^5(np)\|\eta^*\|_1^2\left(\frac{1}{np} + \frac{1}{p^2} + \frac{1}{p}\Delta_E^2\right); \\ (3) &\le C\bar{A}^{10}(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2\frac{(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^4}{\rho_{\min}^8}r^4 \cdot \ln^{10}(np) \\ &\quad \cdot \|\eta^*\|_1^2\left(\frac{1}{np} + \frac{1}{p^2} + \frac{n}{p^3} + \frac{1}{p}\Delta_E^2 + \frac{n}{p}\Delta_E^4\right).\end{aligned}$$

   Clearly (3) dominates (2), which dominates (1) after bounding $\|\eta^*\|_1^2 \le p\|\eta^*\|_2^2$.

□

Remark F.15 (Dictionary)**.** *The generalization of Proposition F.6 is as follows. Suppose*

$$\rho'_{\min} \gg \tilde{C}\sqrt{r'}\ln^{\frac{3}{2}}(np)\left(\frac{1}{\sqrt{p}} \vee \frac{1}{\sqrt{n}} \vee \Delta_E\right).$$

Then with probability at least $1 - O\{(np)^{-10}\}$

$$\|\hat{\eta} - \eta^*\|_2^2 \leq C\bar{A}^{10}(C'_b + \sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^4 \\ \cdot (r')^4 \cdot \ln^{10}(np) \cdot \frac{\|\eta^*\|_2^2}{(\rho'_{\min})^8}\left\{\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + (\Delta'_E)^2 + n(\Delta'_E)^4\right\}.$$

F.5 Test error

F.5.1 Decomposition

Lemma F.15. *Let Assumptions 5.6 and 5.8 hold. Let* k *the PCA hyperparameter equal* $r = rank\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\} = rank\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\}$*. Then*

$$\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 \leq C\sum_{m=1}^{3}\Delta_m$$

where

$$\Delta_1 := \left\{\|\boldsymbol{Z}^{\text{TEST}}\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 + \|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\|^2\right\}\|\hat{\eta} - \eta^*\|_2^2;$$
$$\Delta_2 := \frac{\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2}{\hat{s}_r^2}\left\{\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta} - \eta^*)\|_1 \cdot \Delta_{RR} \vee \|\boldsymbol{A}^{(\text{LR}),\text{TRAIN}} - \hat{\boldsymbol{A}}^{\text{TRAIN}}\|_{2,\infty}^2\|\eta^*\|_1^2\right\};$$
$$\Delta_3 := \|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2\|\eta^*\|_1^2.$$

Proof. As in Lemma E.16, consider

$$\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 \leq 2\|\hat{\boldsymbol{A}}^{\text{TEST}}\big(\hat{\eta} - \eta^*\big)\|_2^2 + 2\|(\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}})\eta^*\|_2^2. \tag{39}$$

We shall bound the two terms on the right hand side of (39) next. To analyze $\|\hat{\boldsymbol{A}}^{\text{TEST}}\big(\hat{\eta} - \eta^*\big)\|_2^2$, we proceed in steps.

1. Decomposition

   As in Lemma E.16, write

   $$\|\hat{\boldsymbol{A}}^{\text{TEST}}\big(\hat{\eta} - \eta^*\big)\|_2^2 \leq 2\|\{\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\}\big(\hat{\eta} - \eta^*\big)\|_2^2 + 2\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\big(\hat{\eta} - \eta^*\big)\|_2^2.$$

   We analyze the former and latter term separately.

2. Former term

   As in Lemma E.16,

$$\|\{\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\}\big(\hat{\eta} - \eta^*\big)\|_2^2 \le 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \ \|\hat{\eta} - \eta^*\|_2^2.$$

3. Latter term

   As in Lemma E.16,

$$\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\big(\hat{\eta} - \eta^*\big)\|_2^2 \le \|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \ \|\boldsymbol{V}\boldsymbol{V}^T\big(\hat{\eta} - \eta^*\big)\|_2^2$$

   and

$$\|\boldsymbol{V}\boldsymbol{V}^T\big(\hat{\eta} - \eta^*\big)\|_2^2 \le 2\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\|^2 \ \|\hat{\eta} - \eta^*\|_2^2 + 2\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\big(\hat{\eta} - \eta^*\big)\|_2^2.$$

   Recall from Lemma F.13 that

$$\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\eta} - \eta^*)\|_2^2 \le \frac{C}{\hat{s}_r^2}\Big\{\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta} - \eta^*)\|_1 \cdot \Delta_{RR} \vee \|\boldsymbol{A}^{(\text{LR}),\text{TRAIN}} - \hat{\boldsymbol{A}}^{\text{TRAIN}}\|_{2,\infty}^2\|\eta^*\|_1^2\Big\}.$$

   Therefore

$$\begin{aligned}
&\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\big(\hat{\eta} - \eta^*\big)\|_2^2 \\
&\le C\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T\|^2 \ \|\hat{\eta} - \eta^*\|_2^2 \\
&+ \frac{C\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2}{\hat{s}_r^2}\Big\{\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T(\hat{\eta} - \eta^*)\|_1 \cdot \Delta_{RR} \vee \|\boldsymbol{A}^{(\text{LR}),\text{TRAIN}} - \hat{\boldsymbol{A}}^{\text{TRAIN}}\|_{2,\infty}^2\|\eta^*\|_1^2\Big\}.
\end{aligned}$$

Finally, to analyze $\|(\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}})\eta^*\|_2^2$, we appeal to matrix Holder:

$$\|(\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}})\eta^*\|_2^2 \le \|\hat{\boldsymbol{A}}^{\text{TEST}} - \boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2\|\eta^*\|_1^2.$$

□

Remark F.16 (Dictionary)**.** *Let Assumption 5.6 hold. Let* $r' = rank[b\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\}] = rank[b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\}]$*. Then,*

$$\|b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\eta} - b(\boldsymbol{A}^{\text{TEST}})\eta^*\|_2^2 \le C\sum_{m=1}^{3}\Delta_m$$

where

$$\Delta_1 := \Big[\{\bar{A}^{d_{\max}}d_{\max}\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|\}^2 + \|b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\}\|^2\|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T\|^2\Big] \ \|\hat{\eta} - \eta^*\|_2^2;$$

$$\Delta_2 := \frac{\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2}{\hat{s}_r^2}\Big[\|\hat{\boldsymbol{V}}_{r'}\hat{\boldsymbol{V}}_{r'}^T(\hat{\eta} - \eta^*)\|_1 \cdot \Delta_{RR} \vee \|b\{\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}\} - b\{\hat{\boldsymbol{A}}^{\text{TRAIN}}\}\|_{2,\infty}^2\|\eta^*\|_1^2\Big];$$

$$\Delta_3 := \|b(\hat{\boldsymbol{A}}^{\text{TEST}}) - b(\boldsymbol{A}^{\text{TEST}})\|_{2,\infty}^2\|\eta^*\|_1^2.$$

Proof. The generalized analysis of the former term in Δ_1 is similar to Remark E.15. □

F.5.2 High probability events

Define the new event

$$\begin{aligned}\tilde{\mathcal{E}}_5 &\coloneqq \left\{\Delta_{RR} \le \tilde{\Delta}_5\right\} \\ \tilde{\Delta}_5 &\coloneqq C\bar{A}^5(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})\frac{(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2}{\rho_{\min}^4} r \cdot \ln^5(np) \\ &\quad \cdot n\|\eta^*\|_1 \left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right)^{\frac{1}{2}}.\end{aligned}$$

Set $\tilde{\mathcal{E}} \coloneqq \cap_{k=1}^5 \tilde{\mathcal{E}}_k$ where the remaining events are defined in Appendix E.

Lemma F.16. *Let the conditions of Proposition F.6 hold. Then $\tilde{\mathcal{E}}_5$ occurs with probability at least $O\{1 - 1/(np)^{10}\}$.*

Proof. In the proofs of Lemma F.14 and Proposition F.6, in particular (38), we have shown

$$\begin{aligned}\frac{1}{\hat{s}_r^4} p \cdot \Delta_{RR}^2 &\le (3) \\ &\le \bar{A}^{10}(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2 \frac{(K_a+\bar{K})^4(\kappa+\bar{K}+K_a)^4}{\rho_{\min}^8} r^2 \cdot \ln^{10}(np) \\ &\quad \cdot \frac{pn^2}{\hat{s}_r^4}\|\eta^*\|_1^2 \left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right).\end{aligned}$$

Hence

$$\begin{aligned}\Delta_{RR}^2 &\le \bar{A}^{10}(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2 \frac{(K_a+\bar{K})^4(\kappa+\bar{K}+K_a)^4}{\rho_{\min}^8} r^2 \cdot \ln^{10}(np) \\ &\quad \cdot n^2\|\eta^*\|_1^2 \left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right).\end{aligned}$$

□

Remark F.17 (Dictionary)**.** *The generalization of Lemma F.16 involves the event*

$$\begin{aligned}\tilde{\mathcal{E}}'_5 &\coloneqq \left\{\Delta_{RR} \le \tilde{\Delta}'_5\right\} \\ \tilde{\Delta}'_5 &\coloneqq C\bar{A}^5(C'_b + \sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})\frac{(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2}{(\rho'_{\min})^4} r' \cdot \ln^5(np) \\ &\quad \cdot n\|\eta^*\|_1 \left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + (\Delta'_E)^2 + n(\Delta'_E)^4\right)^{\frac{1}{2}}.\end{aligned}$$

Lemma F.17. *Let the conditions of Proposition F.6 hold. Then $\mathbb{P}(\tilde{\mathcal{E}}^c) \le \frac{C}{n^{10}p^{10}}$.*

Proof. Immediate from Lemmas E.17, E.18, E.19, E.20, and F.16 and the union bound. □

F.5.3 Simplification

Remark F.18 (Dictionary)**.** *The following lemmas are algebraic and generalize in the obvious way: replace* $(\rho_{\min}, r, \Delta_E)$ *with* $(\rho'_{\min}, r', \Delta'_E)$*. We therefore skip the remarks until Proposition F.7. The only subtlety is the presence of* C'_b *in the definition of* C'_3*.*

Lemma F.18. *Let the conditions of Proposition F.6 hold. Then*

$$\mathbb{E}[\Delta_1 \mid \tilde{\mathcal{E}}] \le C_3 \cdot r^5 \ln^{13}(np) \cdot \frac{\|\eta^*\|_2^2}{\rho_{\min}^{10}} \cdot \left\{1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)\Delta_E^2 + (np + n^2)\Delta_E^4 + n^2 p \Delta_E^6\right\}$$

where

$$C_3 = C \cdot \bar{A}^{14}(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2 (K_a + \bar{K})^4 (\kappa + \bar{K} + K_a)^6.$$

Proof. We proceed in steps. The following arguments are all conditional on $\tilde{\mathcal{E}}$. Recall

$$\Delta_1 = \left\{\|\boldsymbol{Z}^{\text{TEST}}\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 + \|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \|\boldsymbol{V}\boldsymbol{V}^T - \hat{\boldsymbol{V}}_r \hat{\boldsymbol{V}}_r^T\|^2\right\} \|\hat{\eta} - \eta^*\|_2^2.$$

1. Former factor

   As in Lemma E.22

   $$\tilde{\Delta}_2 + np\bar{A}^2\tilde{\Delta}_3 \le C\bar{A}^4(\kappa + \bar{K} + K_a)^2 \cdot \frac{r\ln^3(np)}{\rho_{\min}^2}\left(p + n + np\Delta_E^2\right).$$

2. Latter factor

   By Proposition F.6

   $$\|\hat{\eta} - \eta^*\|_2^2 \le C\bar{A}^{10}(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2 (K_a + \bar{K})^4 (\kappa + \bar{K} + K_a)^4 \cdot r^4 \cdot \ln^{10}(np) \cdot \frac{\|\eta^*\|_2^2}{\rho_{\min}^8}\left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right).$$

3. Combining terms

   Note that

   $$p\left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right) = \frac{p}{n} + 1 + \frac{n}{p} + p\Delta_E^2 + np\Delta_E^4;$$
   $$n\left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right) = 1 + \frac{n}{p} + \frac{n^2}{p^2} + n\Delta_E^2 + n^2\Delta_E^4;$$
   $$np\Delta_E^2\left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right) = p\Delta_E^2 + n\Delta_E^2 + \frac{n^2}{p}\Delta_E^2 + np\Delta_E^4 + n^2 p\Delta_E^6.$$

So that the relevant terms are

$$1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)\Delta_E^2 + (np + n^2)\Delta_E^4 + n^2 p \Delta_E^6.$$

□

Lemma F.19. *Let the conditions of Proposition F.6 hold. Then*

$$\begin{aligned}
&\mathbb{E}[\Delta_2 \mid \tilde{\mathcal{E}}] \\
&\leq C\bar{A}^{12}(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^4 \cdot r^4 \cdot \ln^{10}(np) \\
&\quad \cdot \frac{\|\eta^*\|_1^2}{\rho_{\min}^8}\left(1 + \frac{n}{p} + \frac{n^2}{p^2} + n\Delta_E^2 + n^2\Delta_E^4\right).
\end{aligned}$$

Proof. We proceed in steps. The following arguments are all conditional on $\tilde{\mathcal{E}}$. Recall

$$\Delta_2 := \frac{\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2}{\hat{s}_r^2}\Big\{\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\eta} - \eta^*)\|_1 \cdot \Delta_{RR} \vee \|\boldsymbol{A}^{(\text{LR}),\text{TRAIN}} - \hat{\boldsymbol{A}}^{\text{TRAIN}}\|_{2,\infty}^2\|\eta^*\|_1^2\Big\}.$$

1. Former factor

   Note that conditioned on $\tilde{\mathcal{E}}$,

   $$\frac{\|\boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2}{\hat{s}_r^2} \leq C\frac{r}{np}np\bar{A}^2 = Cr\bar{A}^2.$$

2. Latter factor

   Consider the first term.

   By Proposition F.5

   $$\begin{aligned}
   \|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\eta} - \eta^*)\|_1 &\leq \sqrt{p}\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\eta} - \eta^*)\|_2 \\
   &\leq C\bar{A}^5(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2 \\
   &\quad \cdot r^2 \cdot \ln^5(np) \cdot \frac{\|\eta^*\|_1}{\rho_{\min}^4}\left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right)^{\frac{1}{2}}.
   \end{aligned}$$

   By definition of $\tilde{\mathcal{E}}_5$

   $$\begin{aligned}
   \Delta_{RR} &\leq \tilde{\Delta}_5 \\
   &= C\bar{A}^5(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})\frac{(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2}{\rho_{\min}^4} r \cdot \ln^5(np) \cdot n\|\eta^*\|_1 \\
   &\quad \left(\frac{1}{n} + \frac{1}{p} + \frac{n}{p^2} + \Delta_E^2 + n\Delta_E^4\right)^{\frac{1}{2}}.
   \end{aligned}$$

Hence

$$
\begin{aligned}
&\|\hat{\boldsymbol{V}}_r\hat{\boldsymbol{V}}_r^T(\hat{\eta}-\eta^*)\|_1\cdot\Delta_{RR}\\
&\le C\bar{A}^{10}(\sqrt{C'_m}+C''_m+\bar{\alpha}+\bar{A})^2(K_a+\bar{K})^4(\kappa+\bar{K}+K_a)^4\cdot r^3\cdot\ln^{10}(np)\cdot\\
&\quad n\frac{\|\eta^*\|_1^2}{\rho_{\min}^8}\left(\frac{1}{n}+\frac{1}{p}+\frac{n}{p^2}+\Delta_E^2+n\Delta_E^4\right)\\
&= C\bar{A}^{10}(\sqrt{C'_m}+C''_m+\bar{\alpha}+\bar{A})^2(K_a+\bar{K})^4(\kappa+\bar{K}+K_a)^4\cdot r^3\cdot\ln^{10}(np)\\
&\quad\cdot\frac{\|\eta^*\|_1^2}{\rho_{\min}^8}\left(1+\frac{n}{p}+\frac{n^2}{p^2}+n\Delta_E^2+n^2\Delta_E^4\right).
\end{aligned}
$$

Next consider the second term.

$$
\begin{aligned}
&\|\boldsymbol{A}^{(\text{LR}),\text{TRAIN}}-\hat{\boldsymbol{A}}^{\text{TRAIN}}\|_{2,\infty}^2\|\eta^*\|_1^2\\
&\le\tilde{\Delta}_1\|\eta^*\|_1^2\\
&\le C\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2\cdot\frac{r\ln^5(np)}{\rho_{\min}^4}\|\eta^*\|_1^2\left(1+\frac{n}{p}+n\Delta_E^2\right).
\end{aligned}
$$

So the first term dominates the second term.

□

Lemma F.20.

$$
\mathbb{E}\left[\Delta_3\;\middle|\;\tilde{\mathcal{E}}\right]\le C\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2\cdot\frac{r\ln^5(np)}{\rho_{\min}^4}\|\eta^*\|_1^2\left(1+\frac{n}{p}+n\Delta_E^2\right)
$$

Proof. Recall

$$
\Delta_3:=\|\hat{\boldsymbol{A}}^{\text{TEST}}-\boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2\|\eta^*\|_1^2.
$$

Using the definition of $\tilde{\mathcal{E}}$, we have

$$
\mathbb{E}[\Delta_3\mid\tilde{\mathcal{E}}]\le\tilde{\Delta}_1\|\eta^*\|_1^2\le C\bar{A}^4(K_a+\bar{K})^2(\kappa+\bar{K}+K_a)^2\cdot\frac{r\ln^5(np)}{\rho_{\min}^4}\|\eta^*\|_1^2\left(1+\frac{n}{p}+n\Delta_E^2\right).
$$

□

Lemma F.21. *Let the conditions of Theorem 5.3 hold. Then*

$$
\begin{aligned}
&\sum_{m=1}^3\mathbb{E}[\Delta_m|\tilde{\mathcal{E}}]\\
&\le C_3\cdot\frac{r^5\ln^{13}(np)}{\rho_{\min}^{10}}\cdot\|\eta^*\|_1^2\left\{1+\frac{p}{n}+\frac{n}{p}+\frac{n^2}{p^2}+\left(n+p+\frac{n^2}{p}\right)\Delta_E^2+(np+n^2)\Delta_E^4+n^2p\Delta_E^6\right\}.
\end{aligned}
$$

Proof. Recall Lemma F.15, Lemma F.18, Lemma F.19, and Lemma F.20:

$$\begin{aligned}
\mathbb{E}[\Delta_1 \mid \tilde{\mathcal{E}}] &\le C_3 \cdot r^5 \ln^{13}(np) \cdot \frac{\|\eta^*\|_2^2}{\rho_{\min}^{10}} \\
&\quad \cdot \left\{1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)\Delta_E^2 + (np + n^2)\Delta_E^4 + n^2 p \Delta_E^6\right\}; \\
\mathbb{E}[\Delta_2 \mid \tilde{\mathcal{E}}] &\le C\bar{A}^{12}(\sqrt{C_m'} + C_m'' + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^4 \cdot r^4 \cdot \ln^{10}(np) \cdot \frac{\|\eta^*\|_1^2}{\rho_{\min}^8} \\
&\quad \left(1 + \frac{n}{p} + \frac{n^2}{p^2} + n\Delta_E^2 + n^2\Delta_E^4\right); \\
\mathbb{E}\left[\Delta_3 \;\middle|\; \tilde{\mathcal{E}}\right] &\le C\bar{A}^4(K_a + \bar{K})^2(\kappa + \bar{K} + K_a)^2 \cdot \frac{r \ln^5(np)}{\rho_{\min}^4}\|\eta^*\|_1^2 \left(1 + \frac{n}{p} + n\Delta_E^2\right).
\end{aligned}$$

Δ_2 dominates Δ_3. Comparing Δ_1 to Δ_2, it is sufficient to bound $\|\eta^*\|_2 \le \|\eta^*\|_1$. □

F.5.4 Collecting results

Proposition F.7 (TEST ERROR)**.** *Let the conditions of Theorem 5.3 hold. Then*

$$\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 \, \mathbb{1}\{\mathcal{E}\}] \\
&\le C_3 \cdot \frac{r^5 \ln^{13}(np)}{\rho_{\min}^{10}} \cdot \|\eta^*\|_1^2 \left\{1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)\Delta_E^2 + (np + n^2)\Delta_E^4 + n^2 p \Delta_E^6\right\}.
\end{aligned}$$

Proof. By Lemma F.15

$$\begin{aligned}
\mathbb{E}[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 \, \mathbb{1}\{\mathcal{E}\}] &= \mathbb{E}[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 | \tilde{\mathcal{E}}]\mathbb{P}(\mathcal{E}) \\
&\le \mathbb{E}[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 | \tilde{\mathcal{E}}] \\
&\le C \sum_{m=1}^{3} \mathbb{E}[\Delta_m | \tilde{\mathcal{E}}].
\end{aligned}$$

Finally appeal to Lemma F.21. □

Remark F.19 (Dictionary)**.** *The generalization of Proposition F.7 is*

$$\begin{aligned}
&\mathbb{E}[\|b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\eta} - b(\boldsymbol{A}^{\text{TEST}})\eta^*\|_2^2 \, \mathbb{1}\{\mathcal{E}\}] \\
&\le C_3' \cdot \frac{(r')^5 \ln^{13}(np)}{(\rho_{\min}')^{10}} \cdot \|\eta^*\|_1^2 \left\{1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)(\Delta_E')^2 + (np + n^2)(\Delta_E')^4 + n^2 p (\Delta_E')^6\right\}
\end{aligned}$$

where

$$C_3' = C \cdot \bar{A}^{14}(C_b' + \sqrt{C_m'} + C_m'' + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^6.$$

F.6 Generalization error

F.6.1 Decomposition

Lemma F.22. *Deterministically,*

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\|_2^2 \le 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 + 2\|\boldsymbol{A}^{\text{TEST}}\eta^* - \boldsymbol{\alpha}_0\|_2^2.$$

Moreover

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 \le 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2 + 2\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2.$$

Proof. See Lemma E.28 □

We analyze each term separately.

1. Approximation error $\|\boldsymbol{A}^{\text{TEST}}\eta^* - \boldsymbol{\alpha}_0\|_2^2 = \|\zeta^{\text{TEST}}\|_2^2$.
2. Test error $\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2$.
3. Implicit cleaning error $\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2$.

Remark F.20 (Dictionary)**.** *The generalization with a dictionary considers*

1. *Approximation error* $\|b(\boldsymbol{A}^{\text{TEST}})\eta^* - \boldsymbol{\alpha}_0\|_2^2 = \|\zeta^{\text{TEST}}\|_2^2$.
2. *Test error* $\|b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\eta} - b(\boldsymbol{A}^{\text{TEST}})\eta^*\|_2^2$.
3. *Implicit cleaning error* $\|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\eta} - b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\eta}\|_2^2$.

F.6.2 Implicit cleaning

Lemma F.23. *Suppose Assumptions 5.6 and 5.8 hold and let $k = r$. Then*

$$\begin{aligned}&\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2 \\ &\le C\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \left\{\|\hat{\eta} - \eta^*\|_2^2 + \|\hat{\boldsymbol{V}}_r'(\hat{\boldsymbol{V}}_r')^T - \boldsymbol{V}'(\boldsymbol{V}')^T\|^2\|\eta^*\|_2^2\right\}.\end{aligned}$$

Proof. See Lemma E.29. □

Remark F.21 (Dictionary)**.** *The generalization of Lemma F.23 is*

$$\begin{aligned}&\|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\eta} - b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\eta}\|_2^2 \\ &\leq \|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}) - b\{\boldsymbol{A}^{(\text{LR}),\text{TEST}}\}\|^2 \cdot \left\{\|\hat{\eta} - \eta^*\|_2^2 + \|\hat{\boldsymbol{V}}'_{r'}(\hat{\boldsymbol{V}}'_{r'})^T - \boldsymbol{V}'(\boldsymbol{V}')^T\|^2\|\eta^*\|_2^2\right\}\end{aligned}$$

using the SVDs in Remark E.1.

Proof. See Remark E.24. □

Proposition F.8 (Implicit cleaning)**.** *Let the conditions of Theorem 5.3 hold. Then*

$$\begin{aligned}&\mathbb{E}[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}] \\ &\leq C_3 \cdot \frac{r^5\ln^{13}(np)}{\rho_{\min}^{10}} \cdot \|\eta^*\|_1^2\left\{1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)\Delta_E^2 + (np + n^2)\Delta_E^4 + n^2p\Delta_E^6\right\}.\end{aligned}$$

Proof. To begin, write

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}] &= \mathbb{E}[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2|\tilde{\mathcal{E}}]\mathbb{P}(\tilde{\mathcal{E}}) \\ &\leq \mathbb{E}[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2|\tilde{\mathcal{E}}].\end{aligned}$$

By Lemma F.23, it is sufficient to analyze

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \left\{\|\hat{\eta} - \eta^*\|_2^2 + \|\hat{\boldsymbol{V}}'_r(\hat{\boldsymbol{V}}'_r)^T - \boldsymbol{V}'(\boldsymbol{V}')^T\|^2\|\eta^*\|_2^2\right\}$$

under the beneficial event $\tilde{\mathcal{E}}$. By Lemma F.5, the bound on the former term dominates the bound on the latter term. Therefore we analyze

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \|\hat{\eta} - \eta^*\|_2^2.$$

By Lemma F.15

$$\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1} - \boldsymbol{A}^{(\text{LR}),\text{TEST}}\|^2 \cdot \|\hat{\eta} - \eta^*\|_2^2 \leq \Delta_1 \leq C\sum_{m=1}^{3}\Delta_m$$

so we can use the bound previously used for analyzing TEST ERROR$\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2$. This loose bound is sufficient for our purposes, since the TEST ERROR term will ultimately give this rate. In summary,

$$\mathbb{E}[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2|\tilde{\mathcal{E}}] \leq C\sum_{m=1}^{3}\mathbb{E}[\Delta_m|\tilde{\mathcal{E}}].$$

Finally, we appeal to Lemma F.21. □

Remark F.22 (Dictionary)**.** *The generalization of Proposition F.8 is*

$$\mathbb{E}[\|b(\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1})\hat{\eta} - b(\hat{\boldsymbol{A}}^{\text{TEST}})\hat{\eta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}]$$
$$\leq C_3' \cdot \frac{(r')^5 \ln^{13}(np)}{(\rho'_{\min})^{10}} \cdot \|\eta^*\|_1^2 \left\{1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)(\Delta'_E)^2 + (np + n^2)(\Delta'_E)^4 + n^2 p(\Delta'_E)^6\right\}.$$

F.6.3 Bounded estimator moments

For the adverse case, we place a weak technical condition on how the estimator moments scale. We state the technical condition then demonstrate that it is implied by the interpretable condition given in the main text.

Assumption F.2 (Bounded estimator moments)**.**

$$\sqrt{\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i\in\text{TEST}} \hat{\alpha}(W_{i,\cdot})^2\right\}^2\right]} \leq \text{polynomial}(n,p) \cdot C_3 \cdot \frac{r^5 \ln^{13}(np)}{\rho_{\min}^{10}} \cdot \|\eta^*\|_1^2$$

where $C_3 = C\bar{A}^{14}(\sqrt{C'_m} + C''_m + \bar{\alpha} + \bar{A})^2(K_a + \bar{K})^4(\kappa + \bar{K} + K_a)^6$.

Recall from Appendix D that the powers of (n,p) are arbitrary in the probability of the adverse event; $\mathbb{P}(\tilde{\mathcal{E}}^c) \leq \frac{C}{\text{polynomial}(n,p)}$ for any polynomial of (n,p). Therefore the moments of our estimator $\hat{\alpha}(W_{i,\cdot})$ can scale as any arbitrary polynomial of n and p, denoted by polynomial(n,p). We are simply ruling out some extremely adversarial cases. Assumption F.2 is essentially requiring that $\hat{\eta}$ is well conditioned. Indeed, we are able to satisfy the assumption under a simple condition on the smallest singular value used in PCR.

Proposition F.9 (Verifying bounded estimator moment)**.** *Suppose Assumptions 5.1, 5.2, and F.1 hold. Further suppose* $\hat{s}_k \gtrsim \frac{1}{\text{polynomial}(m,p)}$*. Then Assumption F.2 holds.*

Remark F.23 (Dictionary)**.** *If Assumption C.2 holds then Assumption F.2 becomes*

$$\sqrt{\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i\in\text{TEST}} \hat{\alpha}(W_{i,\cdot})^2\right\}^2\right]} \leq \text{polynomial}(n,p) \cdot C_3' \cdot \frac{(r')^5 \ln^{13}(np)}{(\rho'_{\min})^{10}} \cdot \|\eta^*\|_1^2.$$

Proposition F.9 generalizes accordingly: if Assumption 5.7 holds then the generalization of Assumption F.2 holds.

We prove Proposition F.9 via a sequence of lemmas.

Lemma F.24. *Suppose Assumptions 5.1 and 5.2 hold. Then*

$$\mathbb{E}\left[\|\hat{\boldsymbol{A}}^{\text{TRAIN}}\|_{2,\infty}^8\right] \le C \cdot \bar{A}^8 K_a^8 \cdot \ln^8(np) n^{12}.$$

Proof. We suppress the supersript to lighten notation. Write

$$\|\hat{\boldsymbol{A}}\|_{2,\infty} = \max_{j\in[p]} \|\hat{A}_{\cdot,j}\|_2 \le \max_{j\in[p]} \|\hat{Z}_{\cdot,j}\hat{\rho}_j^{-1}\|_2 = \|\boldsymbol{Z}(\hat{\boldsymbol{\rho}})^{-1}\|_{2,\infty}.$$

Finally appeal to Lemma E.30. □

Lemma F.25. *Deterministically,* $\|\hat{M}\|_2^2 \le \frac{p}{n}\|\hat{\boldsymbol{M}}\|_{2,\infty}^2$.

Proof. Recall $[\hat{\boldsymbol{M}}]_{ij} = \hat{m}_{ij}$ and $\hat{M}_j = \frac{1}{n}\sum_{i\in[n]} \hat{m}_{ij}$. Hence

$$\|\hat{M}\|_2^2 = \sum_{j\in[p]} \hat{M}_j^2 \;\; = \sum_{j\in[p]} \left(\frac{1}{n}\sum_{i\in[n]} \hat{m}_{ij}\right)^2 \le \sum_{j\in[p]} \left(\frac{1}{n}\sum_{i\in[n]} \hat{m}_{ij}^2\right) \le \frac{p}{n}\max_{j\in[p]} \sum_{i\in[n]} \hat{m}_{ij}^2 = \frac{p}{n}\|\hat{\boldsymbol{M}}\|_{2,\infty}^2.$$

□

Lemma F.26. *Suppose Assumptions 5.1 and F.1 hold. Then*

$$\|\hat{\boldsymbol{M}}\|_{2,\infty}^2 \le C \cdot \bar{A}^2 C_m'(C_m'')^2(\|\hat{\boldsymbol{A}}^{\text{TRAIN}}\|_{2,\infty}^2 + n).$$

Proof. We suppress the superscript to lighten notation. Write

$$\|\hat{\boldsymbol{M}}\|_{2,\infty}^2 \le 2\|\hat{\boldsymbol{M}} - \tilde{\boldsymbol{M}}\|_{2,\infty}^2 + 2\|\tilde{\boldsymbol{M}}\|_{2,\infty}^2.$$

Focusing on the former term, by Assumption F.1

$$\|\hat{\boldsymbol{M}} - \tilde{\boldsymbol{M}}\|_{2,\infty}^2 \le C_m'\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2.$$

Moreover,

$$\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \le 2\|\hat{\boldsymbol{A}}\|_{2,\infty}^2 + 2\|\boldsymbol{A}\|_{2,\infty}^2 \le 2\|\hat{\boldsymbol{A}}\|_{2,\infty}^2 + 2n\bar{A}^2.$$

In summary,

$$\|\hat{\boldsymbol{M}} - \tilde{\boldsymbol{M}}\|_{2,\infty}^2 \le C \cdot C_m'\bar{A}^2(\|\hat{\boldsymbol{A}}\|_{2,\infty}^2 + n).$$

Focusing on the latter term,

$$\|\tilde{\boldsymbol{M}}\|_{2,\infty}^2 \le n(C_m'')^2.$$

Therefore

$$\|\hat{\boldsymbol{M}}\|_{2,\infty}^2 \le C \cdot C_m'\bar{A}^2(\|\hat{\boldsymbol{A}}\|_{2,\infty}^2 + n) + C(C_m'')^2 n.$$

□

Lemma F.27. *Suppose Assumptions 5.1, 5.2, and F.1 hold, and $\hat{s}_k \geq \underline{s}$. Then*

$$\mathbb{E}\left[\|\hat{\eta}\|_1^8\right] \leq C \cdot \bar{A}^{16}(C'_m)^4(C''_m)^8 K_a^8 \cdot \ln^8(np) \cdot \frac{n^{16}p^8}{\underline{s}^{16}}.$$

Proof. We suppress the superscript to lighten notation. Recall that $\hat{\eta} = \hat{\boldsymbol{V}}_k \hat{\boldsymbol{\Sigma}}_k^{-2} \hat{\boldsymbol{V}}_k^T (n\hat{M})^T$, hence

$$\begin{aligned}
\|\hat{\eta}\|_1 &\leq \sqrt{p}\|\hat{\eta}\|_2 \\
&\leq \sqrt{p}\|\hat{\boldsymbol{V}}_k\|_{op}\|\hat{\boldsymbol{\Sigma}}_k^{-2}\|_{op}\|\hat{\boldsymbol{V}}_k^T\|_{op}\|(n\hat{M})^T\|_2 \\
&= \sqrt{p}n\hat{s}_k^{-2}\|\hat{M}\|_2.
\end{aligned}$$

Therefore by Lemmas F.25 and F.26,

$$\begin{aligned}
\|\hat{\eta}\|_1^8 &\leq p^4 n^8 \underline{s}^{-16}\|\hat{M}\|_2^8 \\
&\leq p^4 n^8 \underline{s}^{-16} \cdot \frac{p^4}{n^4}\|\hat{\boldsymbol{M}}\|_{2,\infty}^8 \\
&\leq p^4 n^8 \underline{s}^{-16} \cdot \frac{p^4}{n^4} \cdot C \cdot \bar{A}^8(C'_m)^4(C''_m)^8(\|\hat{\boldsymbol{A}}\|_{2,\infty}^8 + n^4) \\
&= C \cdot \bar{A}^8(C'_m)^4(C''_m)^8 n^4 p^8 \underline{s}^{-16}(\|\hat{\boldsymbol{A}}\|_{2,\infty}^8 + n^4)
\end{aligned}$$

and hence

$$\mathbb{E}\left[\|\hat{\eta}\|_1^8\right] \leq C \cdot \bar{A}^8(C'_m)^4(C''_m)^8 n^4 p^8 \underline{s}^{-16}\left\{\mathbb{E}\left[\|\hat{\boldsymbol{A}}\|_{2,\infty}^8\right] + n^4\right\}.$$

Finally by Lemma F.24

$$\mathbb{E}\left[\|\hat{\boldsymbol{A}}\|_{2,\infty}^8\right] \leq C \cdot \bar{A}^8 K_a^8 \cdot \ln^8(np) n^{12}$$

which dominates n^4. □

Lemma F.28. *Suppose the conditions of Lemma F.27 hold. Then*

$$\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]} \leq C \cdot \bar{A}^6 C'_m(C''_m)^2 K_a^4 \cdot \ln^4(np)\frac{n^7p^2}{\underline{s}^4}.$$

Proof. By Cauchy-Schwarz and Lemmas E.30 and F.27, write

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]} &\leq \sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\|_{2,\infty}^4\|\hat{\eta}\|_1^4\right]} \\
&\leq \sqrt{\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\|_{2,\infty}^8\right]}\sqrt{\mathbb{E}\left[\|\hat{\eta}\|_1^8\right]}} \\
&\leq C \cdot \bar{A}^2 K_a^2 \cdot \ln^2(np) n^3 \cdot \bar{A}^4 C'_m(C''_m)^2 K_a^2 \cdot \ln^2(np) \cdot \frac{n^4p^2}{\underline{s}^4} \\
&= C \cdot \bar{A}^6 C'_m(C''_m)^2 K_a^4 \cdot \ln^4(np)\frac{n^7p^2}{\underline{s}^4}.
\end{aligned}$$

□

Proof of Proposition F.9. To begin, observe that

$$\sqrt{\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i\in\text{TEST}}\hat{\alpha}(W_{i,\cdot})^2\right\}^2\right]} = \sqrt{\mathbb{E}\left[\frac{1}{n^2}\left\{\sum_{i\in\text{TEST}}\hat{\alpha}(W_{i,\cdot})^2\right\}^2\right]} = \frac{1}{n}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]}.$$

By Lemma F.28

$$\begin{aligned}\frac{n^{-1}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]}}{n^4p^5} &\leq C\cdot\bar{A}^6C'_m(C''_m)^2K_a^4\cdot\ln^4(np)\frac{n^2}{p^3\underline{s}^4}\\ &\leq C\cdot\bar{A}^6C'_m(C''_m)^2K_a^4\cdot\ln^4(np)\\ &\leq C_3\cdot\frac{r^5\ln^{13}(np)}{\rho_{\min}^{10}}\cdot\|\eta^*\|_1^2\end{aligned}$$

where the penultimate inequality holds since $\underline{s}\geq\frac{\sqrt{n}}{p^{\frac{3}{4}}}$ implies $p^3\underline{s}^4\geq n^2$ and the ultimate inequality confirms Assumption F.2. More generally,

$$\begin{aligned}\frac{\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]}}{\text{polynomial}(n,p)} &\leq C\cdot\bar{A}^6C'_m(C''_m)^2K_a^4\cdot\ln^4(np)\frac{1}{\underline{s}^4\cdot\text{polynomial}(n,p)}\\ &\leq C\cdot\bar{A}^6C'_m(C''_m)^2K_a^4\cdot\ln^4(np)\\ &\leq C_3\cdot\frac{r^5\ln^{13}(np)}{\rho_{\min}^{10}}\cdot\|\eta^*\|_1^2\end{aligned}$$

as long as $\underline{s}\geq\frac{1}{\text{polynomial}(n,p)}$.

□

F.6.4 Main result

Proof of Theorem 5.3. We proceed in steps analogous to the proof of Theorem 5.2.

1. Decomposition

   By Lemma F.22

   $$\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0\|_2^2\leq 2\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}-\boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2+2\|\zeta^{\text{TEST}}\|_2^2.$$

   Hence

   $$\begin{aligned}\mathbb{E}\|\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0\|_2^2 \leq{}& 2\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}-\boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right]\\ &+2\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}-\boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right]\\ &+2\|\zeta^{\text{TEST}}\|_2^2.\end{aligned}$$

2. Beneficial case

   By Propositions F.7 and F.8,

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right] \\
&\le 2\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right] + 2\mathbb{E}\left[\|\hat{\boldsymbol{A}}^{\text{TEST}}\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}\}\right] \\
&\le C_3 \cdot \frac{r^5 \ln^{13}(np)}{\rho_{\min}^{10}} \cdot \|\eta^*\|_1^2 \left\{1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)\Delta_E^2 + (np + n^2)\Delta_E^4 + n^2 p \Delta_E^6\right\}.
\end{aligned}$$

3. Adverse case

   Write

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta} - \boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right] \\
&\le 2\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right] + 2\mathbb{E}\left[\|\boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right].
\end{aligned}$$

   Focusing on the latter term,

$$\|\boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2 \le \|\boldsymbol{A}^{\text{TEST}}\|_{2,\infty}^2 \|\eta^*\|_1^2 \le n\bar{A}^2\|\eta^*\|_1^2$$

   hence by Lemma F.17

$$\mathbb{E}\left[\|\boldsymbol{A}^{\text{TEST}}\eta^*\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right] \le n\bar{A}^2\|\eta^*\|_1^2 \mathbb{P}(\tilde{\mathcal{E}}^c) \le C\frac{\bar{A}^2\|\eta^*\|_1^2}{n^9 p^{10}}$$

   which is clearly dominated by the bound on the beneficial case.

   Focusing on the former term, Cauchy-Schwarz inequality and Lemma F.17 give

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^2\,\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right] &\le \sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]}\sqrt{\mathbb{E}\left[\mathbb{1}\{\tilde{\mathcal{E}}^c\}\right]} \\
&\le \frac{C}{n^5 p^5}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]}
\end{aligned}$$

   which is dominated by the bound on the beneficial case if

$$\begin{aligned}
&\frac{1}{n^5 p^5}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]} \\
&\le C_3 \cdot \frac{r^5 \ln^{13}(np)}{\rho_{\min}^{10}} \cdot \|\eta^*\|_1^2 \left\{1 + \frac{p}{n} + \frac{n}{p} + \frac{n^2}{p^2} + \left(n + p + \frac{n^2}{p}\right)\Delta_E^2 + (np + n^2)\Delta_E^4 + n^2 p \Delta_E^6\right\}.
\end{aligned}$$

   In the proof of Proposition F.9, we have precisely shown

$$\frac{n^{-1}\sqrt{\mathbb{E}\left[\|\boldsymbol{Z}^{\text{TEST}}\,\hat{\boldsymbol{\rho}}^{-1}\,\hat{\eta}\|_2^4\right]}}{n^4 p^5} \le C_3 \cdot \frac{r^5 \ln^{13}(np)}{\rho_{\min}^{10}} \cdot \|\eta^*\|_1^2.$$

□

Proof of Corollary 5.2. Identical to the proof of Theorem 5.3, appealing to the previous remarks for the appropriate generalizations $(C_3', \rho_{\min}')$. □

G Data cleaning-adjusted confidence intervals

The outline of this appendix is as follows

1. exposit general semiparametric estimands
2. exposit general nonparametric estimands
3. establish Neyman orthogonality
4. prove Gaussian approximation for the causal parameter
5. prove consistency for the asymptotic variance
6. prove validity for the confidence interval

G.1 From balancing weight to Riesz representer

We consider the goal of estimation and inference of some causal parameter $\theta_0 \in \mathbb{R}$ which is a scalar summary of the regression γ_0, e.g. a treatment effect, policy effect, or elasticity. We consider a class of causal parameters of the form

$$\theta_0 = \frac{1}{n}\sum_{i=1}^{n} \theta_i, \quad \theta_i = \mathbb{E}[m(W_{i,\cdot}, \gamma_0)]$$

in an i.n.i.d. data generating process with some structure. There are two aspects of this structure that we emphasize: (i) mean square continuity and (ii) marginal distribution shift. In particular, we generalize Assumptions 5.10 and 5.9 from the ATE example to the general case. In doing so, we also generalize the balancing weight to a Riesz representer.

Assumption G.1 (Linearity and mean square continuity)**.** *Assume*

1. *The functional $\gamma \mapsto \mathbb{E}[m(W_{i,\cdot}, \gamma)]$ is linear.*

2. There exists $\bar{Q} < \infty$ and $\bar{q} \in (0, 1]$ such that $\forall \gamma \in \Gamma$,

$$\mathbb{E}[m(W_{i,\cdot}, \gamma)^2] \leq \bar{Q} \cdot \{\mathbb{E}[\gamma(W_{i,\cdot})^2]\}^{\bar{q}}.$$

These restrictions generalize the usual bounded propensity score assumptions. For example, for ATE we impose that the propensity score is bounded away from zero and one; Assumption 5.10 in the main text is a special case of Assumption G.1. A consequence of Assumption G.1 is the existence of the balancing weight.

Proposition G.1 (Riesz representation (Chernozhukov et al., 2022b))**.** *Suppose Assumption G.1 holds. Further suppose the restriction $\gamma_0 \in \Gamma$ that could be imposed in estimation. Then there exists a Riesz representer $\alpha_0 \in \mathbb{L}_2(\mathcal{W})$ such that $\forall \gamma \in \Gamma$*

$$\mathbb{E}[m(W_{i,\cdot}, \gamma)] = \mathbb{E}[\alpha_0(W_{i,\cdot})\gamma(W_{i,\cdot})].$$

There exists a unique minimal Riesz representer $\alpha_0 \in \Gamma$ that satisfies this equation. Moreover, denoting by $\bar{M}$ the operator norm of $\gamma \mapsto \mathbb{E}[m(W_{i,\cdot}, \gamma)]$, we have that

$$\{\mathbb{E}[\alpha_0^{\min}(W_{i,\cdot})^2]\}^{\frac{1}{2}} = \bar{M} \leq \bar{Q}^{\frac{1}{2}} < \infty.$$

The balancing weight in the main text is a special case of a Riesz representer. Hereafter, we refer to the Riesz representer as a balancing weight nonetheless, since our estimator $\hat{\alpha}$ achieves balance across examples; see Proposition F.2, which generalizes Proposition 4.3. To lighten notation, we will typically consider the case where $\Gamma = \mathbb{L}_2(\mathcal{W})$ and $\alpha_0^{\min} = \alpha_0$. When we consider the more general case, as in Example A.4, we will use the richer notation.

In general, (γ_0, α_0) could vary for each observation. We impose that these functions do not vary across observations. Such restrictions generalize the usual distribution shift assumptions. For example, for ATE we impose that the marginal distribution of covariates may shift across observations, but the outcome and treatment mechanisms, encoded by the regression function and propensity score, do not vary; Assumption 5.9 in the main text is a special case of Assumption G.2, which we now state.

Assumption G.2 (Marginal distribution shift)**.** *Assume*

1. *The regression γ_0 does not vary across observations: $\mathbb{E}[\gamma_0(W_{i,\cdot})v(W_{i,\cdot})] = \mathbb{E}[Y_i v(W_{i,\cdot})]$ for all $v \in \mathbb{L}_2(\mathcal{W})$ and $i \in [n]$.*

2. *The Riesz representer α_0 does not vary across observations:* $\mathbb{E}[\alpha_0(W_{i,\cdot})u(W_{i,\cdot})] = \mathbb{E}[m(W_{i,\cdot}, u)]$ *for all* $u \in \mathbb{L}_2(\mathcal{W})$ *and* $i \in [n]$.

While Assumptions G.1 and G.2 may appear abstract, we verify that they hold under simple and interpretable conditions for the leading examples.

Proposition G.2 (Verifying Assumptions G.1 and G.2)**.** *The following conditions verify mean square continuity and marginal distribution shift for the leading examples. Recall that* $\|\alpha_0\|_\infty \leq \bar{\alpha}$*, while* $(\bar{Q}, \bar{q})$ *are defined in Assumption G.1.*

1. *In Example A.1,*

$$\alpha_0(W_{i,\cdot}) = \frac{D_i}{\phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})} - \frac{1 - D_i}{1 - \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})}, \quad \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) := \mathbb{E}[D_i | X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}].$$

Suppose the propensity score is bounded away from zero and one, i.e. $0 < \underline{\phi} \leq \phi_0(X_{i,\cdot}, H_{i,\cdot}, \phi_{i,\cdot}) \leq \bar{\phi} < 1$*. Then*

$$\bar{\alpha} = \frac{1}{\underline{\phi}} + \frac{1}{1 - \bar{\phi}}, \quad \bar{Q} = \frac{2}{\underline{\phi}} + \frac{2}{1 - \bar{\phi}}, \quad \bar{q} = 1$$

for $\Gamma = \mathbb{L}_2(\mathcal{W})$*. We impose that the outcome regression and treatment propensity score do not vary across observations.*

2. *In Example A.2,*

$$\alpha_0(W_{i,\cdot}) = \frac{U_i}{\phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})} - \frac{1 - U_i}{1 - \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})}, \quad \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) := \mathbb{E}[U_i | X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}]$$

for the functionals in the numerator and denominator. Suppose the propensity score is bounded away from zero and one, i.e. $0 < \underline{\phi} \leq \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) \leq \bar{\phi} < 1$*. Then*

$$\bar{\alpha} = \frac{1}{\underline{\phi}} + \frac{1}{1 - \bar{\phi}}, \quad \bar{Q} = \frac{2}{\underline{\phi}} + \frac{2}{1 - \bar{\phi}}, \quad \bar{q} = 1$$

for the functionals in the numerator and denominator when $\Gamma = \mathbb{L}_2(\mathcal{W})$*. We impose that the outcome regression and instrument propensity score do not vary across observations.*

3. *In Example A.3,*

$$\alpha_0(W_{i,\cdot}) = \omega(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) - 1, \quad \omega(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) = \frac{f\{t(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})\}}{f(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})}.$$

Suppose the density ratio $\omega(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ is bounded above, i.e. $\omega(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) \leq \bar{\omega} < \infty$. Then

$$\bar{\alpha} = \bar{\omega} + 1, \quad \bar{Q} = 2\bar{\omega} + 2, \quad \bar{q} = 1$$

for $\Gamma = \mathbb{L}_2(\mathcal{W})$. We impose that the outcome regression and covariate density ratio do not vary across observations.

4. *In Example A.4,*

$$\alpha_0(W_{i,\cdot}) = -\nabla_d \ln f(D_i \mid X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}).$$

Suppose the density derivative is bounded above, i.e. $-\nabla_d \ln f(D_i \mid X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) \leq \bar{f} < \infty$. Then

$$\bar{\alpha} = \bar{f}, \quad \bar{Q} = \bar{f}(\bar{\gamma} + \bar{\gamma}'), \quad \bar{q} = 1/2$$

for Γ that consists of functions γ that are twice continuously differentiable in the first argument and that satisfy a Sobolev type condition: $\mathbb{E}[\{\nabla_d \gamma(D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})\}^2] \leq \bar{\gamma}^2 < \infty$ and $\mathbb{E}[\{\partial_d^2 \gamma(D_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})\}^2] \leq (\bar{\gamma}')^2 < \infty$. We impose that the outcome regression and conditional density of goods do not vary across observations.

5. *In Example A.5,*

$$\alpha_0(W_{i,\cdot}) = \ell_i \frac{D_i - \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})}{\mathbb{E}[\{D_i - \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})\}^2]}, \quad \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) := \mathbb{E}[D_i | X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}].$$

Suppose treatment has non-degenerate conditional variance, i.e. $\mathbb{E}[\{D_i - \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})\}^2] > \underline{\phi}$, and the weights are bounded above and below, i.e. $|\ell_i| \leq \bar{\ell}$. Then

$$\bar{\alpha} = \frac{2\bar{\ell}\bar{A}}{\underline{\phi}}, \quad \bar{Q} = \frac{4\bar{\ell}^2\bar{A}^2}{\underline{\phi}^2} \quad \bar{q} = 1$$

for $\Gamma = \mathbb{L}_2(\mathcal{W})$. We impose that the outcome regression and treatment regression do not vary across observations.

6. *In Example A.6,*

$$\alpha_0(W_{i,\cdot}) = \ell_i \frac{U_i - \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})}{\mathbb{E}[\{U_i - \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})\}^2]}, \quad \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) := \mathbb{E}[U_i | X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}]$$

for the functionals in the numerator and denominator. Suppose treatment has non-degenerate conditional variance, i.e. $\mathbb{E}[\{D_i - \phi_0(X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})\}^2] > \underline{\phi}$, and the weights are bounded above and below, i.e. $|\ell_i| \leq \bar{\ell}$. Then

$$\bar{\alpha} = \frac{2\bar{\ell}\bar{A}}{\underline{\phi}}, \quad \bar{Q} = \frac{4\bar{\ell}^2\bar{A}^2}{\underline{\phi}^2} \quad \bar{q} = 1.$$

for the functionals in the numerator and denominator when $\Gamma = \mathbb{L}_2(\mathcal{W})$. We impose that the outcome regression and instrument regression do not vary across observations.

7. *In Example A.7,*

$$\alpha_0(W_{i,\cdot}) = \ell_h(V_i)\left\{\frac{D_i}{\phi_0(V_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})} - \frac{1-D_i}{1-\phi_0(V_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})}\right\}.$$

Suppose the propensity score $\phi_0(V_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot})$ is bounded away from zero and one, i.e. $0 < \underline{\phi} \leq \phi_0(V_i, X_{i,\cdot}, H_{i,\cdot}, \pi_{i,\cdot}) \leq \bar{\phi} < 1$ and other regularity conditions hold given in Lemma G.1 below. Then

$$\bar{\alpha}_h \leq C \cdot \frac{1}{h}\left(\frac{1}{\underline{\phi}} + \frac{1}{1-\bar{\phi}}\right), \quad \bar{Q}_h \leq C \cdot \frac{1}{h^2}\left(\frac{2}{\underline{\phi}} + \frac{2}{1-\bar{\phi}}\right), \quad \bar{q} = 1$$

when $\Gamma = \mathbb{L}_2(\mathcal{W})$. We impose that the outcome regression and treatment propensity score do not vary across observations.

Proof of Proposition G.2. We verify the result for each example. To lighten notation, we suppress the arguments (H, π).

1. Example A.1. To characterize the Riesz representer, write

$$\begin{aligned}\mathbb{E}[m(W_{i,\cdot}, \gamma)] &= \mathbb{E}[\gamma(1, X_{i,\cdot}) - \gamma(0, X_{i,\cdot})] \\ &= \mathbb{E}\left[\frac{D_i}{\phi_0(X_{i,\cdot})}\gamma(D_i, X_{i,\cdot}) - \frac{1-D_i}{1-\phi_0(X_{i,\cdot})}\gamma(D_i, X_{i,\cdot})\right]\end{aligned}$$

so $\alpha_0(D_i, X_{i,\cdot}) = \frac{D_i}{\phi_0(X_{i,\cdot})} - \frac{1-D_i}{1-\phi_0(X_{i,\cdot})}$ and $\bar{\alpha} = \frac{1}{\underline{\phi}} + \frac{1}{1-\bar{\phi}}$. To characterize mean square continuity, write

$$\mathbb{E}[m(W_{i,\cdot}, \gamma)^2] = \mathbb{E}[\{\gamma(1, X_{i,\cdot}) - \gamma(0, X_{i,\cdot})\}^2] \leq 2\mathbb{E}[\gamma(1, X_{i,\cdot})^2] + 2\mathbb{E}[\gamma(0, X_{i,\cdot})^2].$$

Focusing on the former term

$$\mathbb{E}[\gamma(1, X_{i,\cdot})^2] = \mathbb{E}\left[\frac{D_i}{\phi_0(X_{i,\cdot})}\gamma(D_i, X_{i,\cdot})^2\right] \leq \frac{1}{\underline{\phi}}\mathbb{E}\left[\gamma(D_i, X_{i,\cdot})^2\right].$$

where $0 < \underline{\phi} \leq \phi_0(X_{i,\cdot}) \leq \bar{\phi} < 1$ by hypothesis. Likewise

$$\mathbb{E}[\gamma(0, X_{i,\cdot})^2] = \mathbb{E}\left[\frac{1-D_i}{1-\phi_0(X_{i,\cdot})}\gamma(D_i, X_{i,\cdot})^2\right] \leq \frac{1}{1-\bar{\phi}}\mathbb{E}\left[\gamma(D_i, X_{i,\cdot})^2\right].$$

Hence $\bar{Q} = \frac{2}{\underline{\phi}} + \frac{2}{1-\bar{\phi}}$ and $\bar{q} = 1$.

2. Example A.2 is similar to Example A.1.

3. Example A.3. To characterize the Riesz representer, write

$$\mathbb{E}[m(W_{i,\cdot},\gamma)] = \mathbb{E}[\gamma\{t(X_{i,\cdot})\} - \gamma(X_{i,\cdot})] = \mathbb{E}\left[\frac{f\{t(X_{i,\cdot})\}}{f(X_{i,\cdot})}\gamma(X_{i,\cdot}) - \gamma(X_{i,\cdot})\right]$$

so $\alpha_0(X_{i,\cdot}) = \omega(X_{i,\cdot}) - 1$ and $\bar{\alpha} = \bar{\omega} + 1$. To characterize mean square continuity, write

$$\mathbb{E}[m(W_{i,\cdot},\gamma)^2] = \mathbb{E}([\gamma\{t(X_{i,\cdot})\} - \gamma(X_{i,\cdot})]^2) \leq 2\mathbb{E}[\gamma\{t(X_{i,\cdot})\}^2] + 2\mathbb{E}[\gamma(X_{i,\cdot})^2].$$

Focusing on the former term

$$\mathbb{E}[\gamma\{t(X_{i,\cdot})\}^2] = \mathbb{E}\left[\frac{f\{t(X_{i,\cdot})\}}{f(X_{i,\cdot})}\gamma(X_{i,\cdot})^2\right] \leq \bar{\omega} \cdot \mathbb{E}\left[\gamma(X_{i,\cdot})^2\right]$$

where $\omega(X_{i,\cdot}) < \bar{\omega} < \infty$ by hypothesis. Hence $\bar{Q} = 2\bar{\omega} + 2$ and $\bar{q} = 1$.

4. Example A.4. To characterize the Riesz representer, integrate by parts:

$$\begin{aligned}
\mathbb{E}\left[\nabla_d \gamma_0(D_i, X_{i,\cdot})\right] &= \mathbb{E}\left[-\gamma_0(D_i, X_{i,\cdot})\frac{\nabla_d f(D_i, X_{i,\cdot})}{f(D_i, X_{i,\cdot})}\right] \\
&= \left[-\gamma_0(D_i, X_{i,\cdot})\frac{\nabla_d f(D_i|X_{i,\cdot})}{f(D_i|X_{i,\cdot})}\right] \\
&= \left[-\gamma_0(D_i, X_{i,\cdot})\nabla_d \ln f(D_i \mid X_{i,\cdot})\right]
\end{aligned}$$

so $\alpha_0(d, x) = -\nabla_d \ln f(d \mid x)$ and $\bar{\alpha} = \bar{f}$. See Chernozhukov et al. (2021, Lemmas S3 and S4) for mean square continuity.

5. Example A.5. Ignore ℓ_i for simplicity. To characterize the Riesz representer, let $\epsilon_i^D := D_i - \phi_0(X_{i,\cdot})$ be the regression residual of D_i. Then appealing to partial linearity of γ_0 and exogeneity

$$cov(Y_i, \epsilon_i^D) = cov(D\theta_i, \epsilon_i^D)\theta_i = cov(\epsilon_i^D, \epsilon_i^D)\theta_i.$$

Therefore

$$\theta_i = \frac{cov(Y_i, \epsilon_i^D)}{cov(\epsilon_i^D, \epsilon_i^D)} = \frac{\mathbb{E}[Y_i\{D_i - \phi_0(X_{i,\cdot})\}]}{\mathbb{E}[\{D_i - \phi_0(X_{i,\cdot})\}^2]}$$

so $\alpha_0(D_i, X_{i,\cdot}) = \frac{D_i - \phi_0(X_{i,\cdot})}{\mathbb{E}[\{D_i - \phi_0(X_{i,\cdot})\}^2]}$ and $\bar{\alpha} = \frac{2\bar{A}}{\underline{\phi}}$. To characterize mean square continuity,

we use partial linearity to write

$$\begin{aligned}\mathbb{E}[m(W_{i,\cdot},\gamma)^2] &= m(W_{i,\cdot},\gamma)^2 \\ &= \theta_i^2 \\ &= \left[\mathbb{E}\left\{\gamma_0(D_i,X_{i,\cdot})\alpha_0(D_i,X_{i,\cdot})\right\}\right]^2 \\ &\leq \mathbb{E}[\{\gamma_0(D_i,X_{i,\cdot})\alpha_0(D_i,X_{i,\cdot})\}^2] \\ &\leq \bar{\alpha}^2\mathbb{E}[\{\gamma_0(D_i,X_{i,\cdot})\}^2].\end{aligned}$$

Hence $\bar{Q} = \bar{\alpha}^2$ and $\bar{q} = 1$.

6. Example A.6 is similar to Example A.5.

7. Example A.7 is similar to Example A.1. See Chernozhukov et al. (2021, Theorem 2) for the characterization of $(\bar{\alpha}_h, \bar{Q}_h)$ with localization.

□

G.2 From semiparametrics to nonparametrics

A local functional $\theta_0^{\lim} \in \mathbb{R}$ is a scalar that takes the form

$$\theta_0^{\lim} = \lim_{h\to 0}\theta_0^h, \quad \theta_0^h = \frac{1}{n}\sum_{i=1}^n \theta_i^h, \quad \theta_i^h = \mathbb{E}[m_h(W_{i,\cdot},\gamma_0)] = \mathbb{E}[\ell_h(W_{ij})m(W_{i,\cdot},\gamma_0)]$$

where ℓ_h is a Nadaraya Watson weighting with bandwidth h and W_{ij} is a scalar component of $W_{i,\cdot}$. $\theta_0^{\lim}$ is a nonparametric quantity. However, it can be approximated by the sequence $\{\theta_0^h\}$. By this logic, finite sample semiparametric theory for θ_0^h translates to finite sample nonparametric theory for $\theta_0^{\lim}$ up to some approximation error, which we now define.

Definition G.1 (Nonparametric approximation error)**.** *The error from approximating the nonparametric quantity $\theta_0^{\lim}$ with a sequence of semiparametric quantities $\{\theta_0^h\}$ is $\Delta_h = n^{1/2}\sigma^{-1}|\theta_0^h - \theta_0^{\lim}|$.*

Each θ_0^h can be analyzed like θ_0 above as long as we keep track of how certain quantities depend on h, which we preview in Example A.7 of Proposition G.2. We now formalize how these key quantities behave.

Lemma G.1 (Characterization of key quantities; Theorem 2 of Chernozhukov et al. (2021)). *If response noise has finite variance then $\bar{\sigma}^2 < \infty$. Suppose bounded moment and heteroscedasticity conditions hold. Then for global functionals*

$$\xi/\sigma \lesssim \sigma \asymp \bar{M} < \infty, \quad \xi, \chi \lesssim \bar{M}^2 \leq \bar{Q} < \infty, \quad \bar{\alpha} < \infty.$$

Suppose bounded moment, heteroscedasticity, density, and derivative conditions hold. Then for local functionals

$$\xi_h/\sigma_h \lesssim h^{-1/6}, \quad \sigma_h \asymp \bar{M}_h \asymp h^{-1/2}, \quad \xi_h \lesssim h^{-2/3}, \quad \chi_h \lesssim h^{-3/4}, \quad \bar{\alpha}_h \lesssim h^{-1}, \quad \bar{Q}_h \lesssim h^{-2}$$

and $\Delta_h \lesssim n^{1/2} h^{\mathsf{v}+1/2}$ where v is the order of differentiability of the kernel K.

Equipped with this lemma, we prove validity of the data cleaning-adjusted confidence interval for nonparametric quantities.

Corollary G.1 (Confidence interval coverage). *Suppose the conditions of Corollary 5.3 and Lemma G.1. Update the rate conditions to be*

1. *Bandwidth regularity: $n^{-1/2} h^{-3/2} \to 0$ and $\Delta_h \to 0$;*
2. *Error-in-variable regression rate: $(h^{-1} + \bar{\alpha}') \{\mathcal{R}(\hat{\gamma})\}^{\bar{q}/2} \to 0$;*
3. *Error-in-variable balancing weight rate: $\bar{\sigma} h^{-1} \{\mathcal{R}(\hat{\alpha})\}^{1/2} \to 0$;*
4. *Product of rates is fast: $h^{-1/2} \{n \mathcal{R}(\hat{\gamma}) \mathcal{R}(\hat{\alpha})\}^{1/2} \to 0$.*

Then the conclusions of Corollary 5.3 hold, replacing $(\hat{\theta}, \theta_0)$ with $(\hat{\theta}^h, \theta_0^{\lim})$.

G.3 Neyman orthogonality

To lighten notation, we suppress indexing by i where possible. Recall

$$\psi_i = \psi(W_{i,\cdot}, \theta_i, \gamma_0, \alpha_0), \quad \psi(w, \theta, \gamma, \alpha) = m(w, \gamma) + \alpha(w)\{y - \gamma(w)\} - \theta,$$

where $\gamma \mapsto m(w, \gamma)$ is linear. We take as given that (γ_0, α_0) exist, though the latter is implied by Assumption G.1.

Definition G.2 (Gateaux derivative)**.** *Let $u(w), v(w)$ be functions and let τ, ζ in $\mathbb{R}$ be scalars. The Gateaux derivative of $\psi(w, \theta, \gamma, \alpha)$ with respect to its argument γ in the direction u is*

$$\{\partial_\gamma \psi(w, \theta, \gamma, \alpha)\}(u) = \frac{\partial}{\partial \tau} \psi(w, \theta, \gamma + \tau u, \alpha)\bigg|_{\tau=0}.$$

The cross derivative of $\psi(w, \theta, \gamma, \alpha)$ with respect to its arguments (γ, α) in the directions (u, v) is

$$\{\partial^2_{\gamma,\alpha} \psi(w, \theta, \gamma, \alpha)\}(u, v) = \frac{\partial^2}{\partial \tau \partial \zeta} \psi(w, \theta, \gamma + \tau u, \alpha + \zeta v)\bigg|_{\tau=0,\zeta=0}.$$

Lemma G.2 (Calculation of derivatives; Proposition S1 of Chernozhukov et al. (2021))**.**

$$\begin{aligned}
\{\partial_\gamma \psi(w, \theta, \gamma, \alpha)\}(u) &= m(w, u) - \alpha(w)u(w); \\
\{\partial_\alpha \psi(w, \theta, \gamma, \alpha)\}(v) &= v(w)\{y - \gamma(w)\}; \\
\{\partial^2_{\gamma,\alpha} \psi(w, \theta, \gamma, \alpha)\}(u, v) &= -v(w)u(w).
\end{aligned}$$

Lemma G.3 (Neyman orthogonality)**.** *Suppose Assumption G.2 holds. For any (u, v),*

$$\mathbb{E}[\partial_\gamma \psi_i](u) = 0, \quad \mathbb{E}[\partial_\alpha \psi_i](v) = 0.$$

Proof. By Lemma G.2 and Assumption G.2,

$$\mathbb{E}[\partial_\gamma \psi_i](u) = \mathbb{E}[m(W_{i,\cdot}, u) - \alpha_0(W_{i,\cdot})u(W_{i,\cdot})] = 0.$$

Likewise

$$\mathbb{E}[\partial_\alpha \psi_i](v) = \mathbb{E}[v(W_{i,\cdot})\{Y_i - \gamma_0(W_{i,\cdot})\}] = 0.$$

□

G.4 Gaussian approximation

Train $(\hat{\gamma}_\ell, \hat{\alpha}_\ell)$ on observations in I_ℓ^c, which serves as TRAIN. Let $m = |I_\ell| = n/L$ be the number of observations in I_ℓ, which serves as TEST. Denote by $\mathbb{E}_\ell[\cdot]$ the average over observations in I_ℓ. This generalized notation allows us to reverse the roles of TRAIN and TEST, and to allow for more than two folds.

Definition G.3 (Foldwise target and oracle)**.**

$$\begin{aligned}
\hat{\theta}_\ell &= \mathbb{E}_\ell[m(W_{i,\cdot}, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W_{i,\cdot})\{Y_i - \hat{\gamma}_\ell(W_{i,\cdot})\}]; \\
\bar{\theta}_\ell &= \mathbb{E}_\ell[m(W_{i,\cdot}, \gamma_0) + \alpha_0(W_{i,\cdot})\{Y_i - \gamma_0(W_{i,\cdot})\}].
\end{aligned}$$

Lemma G.4 (Taylor expansion)**.** *Let* $u = \hat{\gamma}_\ell - \gamma_0$ *and* $v = \hat{\alpha}_\ell - \alpha_0$*. Then* $m^{1/2}(\hat{\theta}_\ell - \bar{\theta}_\ell) = \sum_{j=1}^3 \Delta_{j\ell}$ *where*

$$\begin{aligned}
\Delta_{1\ell} &= m^{1/2}\mathbb{E}_\ell\{m(W_{i,\cdot}, u) - \alpha_0(W_{i,\cdot})u(W_{i,\cdot})\};\\
\Delta_{2\ell} &= m^{1/2}\mathbb{E}_\ell[v(W_{i,\cdot})\{Y_i - \gamma_0(W_{i,\cdot})\}];\\
\Delta_{3\ell} &= \frac{1}{2}m^{1/2}\mathbb{E}_\ell\{-u(W_{i,\cdot})v(W_{i,\cdot})\}.
\end{aligned}$$

Proof of Lemma G.4. An exact Taylor expansion gives

$$\psi(W_{i,\cdot}, \theta_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell) - \psi_i = \{\partial_\gamma \psi_i\}(u) + \{\partial_\alpha \psi_i\}(v) + \frac{1}{2}\{\partial^2_{\gamma,\alpha}\psi_i\}(u, v).$$

Averaging over observations in I_ℓ,

$$\begin{aligned}
\hat{\theta}_\ell - \bar{\theta}_\ell &= \mathbb{E}_\ell[m(W_{i,\cdot}, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W_{i,\cdot})\{Y_i - \hat{\gamma}_\ell(W_{i,\cdot})\}] - \mathbb{E}_\ell[m(W_{i,\cdot}, \gamma_0) + \alpha_0(W_{i,\cdot})\{Y_i - \gamma_0(W_{i,\cdot})\}]\\
&= \mathbb{E}_\ell\{\psi(W_{i,\cdot}, \theta_i, \hat{\gamma}_\ell, \hat{\alpha}_\ell)\} - \mathbb{E}_\ell\{\psi_i\}\\
&= \mathbb{E}_\ell\{\partial_\gamma \psi_i\}(u) + \mathbb{E}_\ell\{\partial_\alpha \psi_i\}(v) + \frac{1}{2}\mathbb{E}_\ell\{\partial^2_{\gamma,\alpha}\psi_i\}(u, v).
\end{aligned}$$

Finally appeal to Lemma G.2. □

Lemma G.5 (Residuals)**.** *Suppose Assumption G.1 holds and*

$$\mathbb{E}[\varepsilon_i^2 | W_{i,\cdot}] \leq \bar{\sigma}^2, \quad \|\alpha_0\|_\infty \leq \bar{\alpha}.$$

Further suppose that for $(i, j) \in I_\ell$*,*

$$\hat{\gamma}_\ell(W_{i,\cdot}) \perp\!\!\!\perp \hat{\gamma}_\ell(W_{j,\cdot}) | I_\ell^c, \quad \hat{\alpha}_\ell(W_{i,\cdot}) \perp\!\!\!\perp \hat{\alpha}_\ell(W_{j,\cdot}) | I_\ell^c.$$

Then with probability $1 - \epsilon/L$*,*

$$\begin{aligned}
|\Delta_{1\ell}| &\leq t_1 = \left(\frac{6L}{\epsilon}\right)^{1/2}\{(\bar{Q} + \bar{\alpha}^2)\mathcal{R}(\hat{\gamma}_\ell)\}^{\bar{q}/2};\\
|\Delta_{2\ell}| &\leq t_2 = \left(\frac{3L}{\epsilon}\right)^{1/2}\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2};\\
|\Delta_{3\ell}| &\leq t_3 = \frac{3L^{1/2}}{2\epsilon}\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.
\end{aligned}$$

Proof. We proceed in steps.

1. Markov inequality implies

$$\mathbb{P}(|\Delta_{1\ell}| > t_1) \leq \frac{\mathbb{E}(\Delta_{1\ell}^2)}{t_1^2};$$
$$\mathbb{P}(|\Delta_{2\ell}| > t_2) \leq \frac{\mathbb{E}(\Delta_{2\ell}^2)}{t_2^2};$$
$$\mathbb{P}(|\Delta_{3\ell}| > t_3) \leq \frac{\mathbb{E}(|\Delta_{3\ell}|)}{t_3}.$$

2. Law of iterated expectations implies

$$\mathbb{E}(\Delta_{1\ell}^2) = \mathbb{E}\{\mathbb{E}(\Delta_{1\ell}^2 \mid I_\ell^c)\};$$
$$\mathbb{E}(\Delta_{2\ell}^2) = \mathbb{E}\{\mathbb{E}(\Delta_{2\ell}^2 \mid I_\ell^c)\};$$
$$\mathbb{E}(|\Delta_{3\ell}|) = \mathbb{E}\{\mathbb{E}(|\Delta_{3\ell}| \mid I_\ell^c)\}.$$

3. Bounding conditional moments

   Conditional on I_ℓ^c, (u, v) are deterministic. Moreover, within fold I_ℓ, $W_{i,\cdot} \perp\!\!\!\perp W_{j,\cdot}$. In particular, $u(W_{i,\cdot}) \perp\!\!\!\perp u(W_{j,\cdot})|I_\ell^c$ where $u(W_{i,\cdot}) = \hat{\gamma}_\ell(W_{i,\cdot}) - \gamma_0(W_{i,\cdot})$ since

$$\hat{\gamma}_\ell(W_{i,\cdot}) \perp\!\!\!\perp \hat{\gamma}_\ell(W_{j,\cdot})|I_\ell^c, \quad \gamma_0(W_{i,\cdot}) \perp\!\!\!\perp \gamma_0(W_{j,\cdot})|I_\ell^c, \quad \hat{\gamma}_\ell(W_{i,\cdot}) \perp\!\!\!\perp \gamma_0(W_{j,\cdot})|I_\ell^c.$$

   Likewise $v(W_{i,\cdot}) \perp\!\!\!\perp v(W_{j,\cdot})|I_\ell^c$ where $v(W_{i,\cdot}) = \hat{\alpha}_\ell(W_{i,\cdot}) - \alpha_0(W_{i,\cdot})$ since

$$v(W_{i,\cdot}) = \hat{\alpha}_\ell(W_{i,\cdot}) - \alpha_0(W_{i,\cdot}), \quad \hat{\alpha}_\ell(W_{i,\cdot}) \perp\!\!\!\perp \hat{\alpha}_\ell(W_{j,\cdot})|I_\ell^c \quad \alpha_0(W_{i,\cdot}) \perp\!\!\!\perp \alpha_0(W_{j,\cdot})|I_\ell^c, \quad \hat{\alpha}_\ell(W_{i,\cdot}) \perp\!\!\!\perp \alpha_0(W_{j,\cdot})|I_\ell^c.$$

   Hence by Lemma G.3

$$\begin{aligned}
\mathbb{E}[\Delta_{1\ell}^2|I_\ell^c] &= \mathbb{E}\left[m\frac{1}{m^2}\sum_{i,j\in I_\ell}\{m(W_{i,\cdot},u) - \alpha_0(W_{i,\cdot})u(W_{i,\cdot})\}\{m(W_{j,\cdot},u) - \alpha_0(W_{j,\cdot})u(W_{j,\cdot})\}\middle| I_\ell^c\right] \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i\in I_\ell}\{m(W_{i,\cdot},u) - \alpha_0(W_{i,\cdot})u(W_{i,\cdot})\}^2\middle| I_\ell^c\right] \\
&= \mathbb{E}\left[\mathbb{E}_\ell[\{m(W_{i,\cdot},u) - \alpha_0(W_{i,\cdot})u(W_{i,\cdot})\}^2]|I_\ell^c\right] \\
&\leq 2\mathbb{E}[\mathbb{E}_\ell[m(W_{i,\cdot},u)^2]|I_\ell^c] + 2\mathbb{E}[\mathbb{E}_\ell[\{\alpha_0(W_{i,\cdot})u(W_{i,\cdot})\}^2]|I_\ell^c] \\
&\leq 2(\bar{Q} + \bar{\alpha}^2)\{\mathbb{E}[\mathbb{E}_\ell[u(W_{i,\cdot})^2]|I_\ell^c]\}^{\bar{q}}.
\end{aligned}$$

In the last line we use Jensen's inequality and $\bar{q} \in (0, 1]$ to argue that

$$\begin{aligned}
\mathbb{E}[\mathbb{E}_\ell[m(W_{i,\cdot}, u)^2]|I_\ell^c] &= \mathbb{E}_\ell[\mathbb{E}[m(W_{i,\cdot}, u)^2|I_\ell^c]] \\
&\leq \bar{Q}\mathbb{E}_\ell[\{\mathbb{E}[u(W_i)^2|I_\ell^c]\}^{\bar{q}}] \\
&\leq \bar{Q}\{\mathbb{E}_\ell[\mathbb{E}[u(W_i)^2|I_\ell^c]]\}^{\bar{q}} \\
&= \bar{Q}\{\mathbb{E}[\mathbb{E}_\ell[u(W_i)^2]|I_\ell^c]\}^{\bar{q}}.
\end{aligned}$$

We also use the fact that $\mathbb{E}[\mathbb{E}_\ell[u(W_{i,\cdot})^2]|I_\ell^c]$ is vanishing and $\bar{q} \in (0, 1]$ to argue that

$$\begin{aligned}
\mathbb{E}[\mathbb{E}_\ell[\{\alpha_0(W_{i,\cdot})u(W_{i,\cdot})\}^2]|I_\ell^c] &\leq \bar{\alpha}^2\mathbb{E}[\mathbb{E}_\ell[u(W_{i,\cdot})\}^2]|I_\ell^c] \\
&\leq \bar{\alpha}^2\{\mathbb{E}[\mathbb{E}_\ell[u(W_{i,\cdot})\}^2]|I_\ell^c]\}^{\bar{q}}.
\end{aligned}$$

Similarly by Lemma G.3

$$\begin{aligned}
\mathbb{E}[\Delta_{2\ell}^2|I_\ell^c] &= \mathbb{E}\left[m\frac{1}{m^2}\sum_{i,j\in I_\ell}\{v(W_{i,\cdot})[Y_i - \gamma_0(W_{i,\cdot})]\}\{v(W_{j,\cdot})[Y_j - \gamma_0(W_{j,\cdot})]\}\middle| I_\ell^c\right] \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i\in I_\ell}\{v(W_{i,\cdot})[Y_i - \gamma_0(W_{i,\cdot})]\}^2\middle| I_\ell^c\right] \\
&= \mathbb{E}\left[\mathbb{E}_\ell[\{v(W_{i,\cdot})[Y_i - \gamma_0(W_{i,\cdot})]\}^2]|I_\ell^c\right] \\
&\leq \bar{\sigma}^2\mathbb{E}\left[\mathbb{E}_\ell[v(W_{i,\cdot})^2]|I_\ell^c\right].
\end{aligned}$$

Finally

$$\begin{aligned}
\mathbb{E}[|\Delta_{3\ell}||I_\ell^c] &= \frac{1}{2}\sqrt{m}\mathbb{E}[|\mathbb{E}_\ell\{-u(W_{i,\cdot})v(W_{i,\cdot})\}||I_\ell^c] \\
&\leq \frac{1}{2}\sqrt{m}\mathbb{E}[\mathbb{E}_\ell\{|u(W_{i,\cdot})v(W_{i,\cdot})|\}|I_\ell^c].
\end{aligned}$$

4. Law of iterated expectations and Jensen's inequality imply

$$\begin{aligned}
\mathbb{E}[\Delta_{1\ell}^2] &\leq \mathbb{E}\left[2(\bar{Q}+\bar{\alpha}^2)\{\mathbb{E}[\mathbb{E}_\ell[u(W_{i,\cdot})^2]|I_\ell^c]\}^{\bar{q}}\right] \\
&\leq 2(\bar{Q}+\bar{\alpha}^2)\{\mathbb{E}\left[\mathbb{E}[\mathbb{E}_\ell[u(W_{i,\cdot})^2]|I_\ell^c]\right]\}^{\bar{q}} \\
&= 2(\bar{Q}+\bar{\alpha}^2)\{\mathbb{E}\left[\mathbb{E}_\ell[u(W_{i,\cdot})^2]\right]\}^{\bar{q}} \\
&= 2(\bar{Q}+\bar{\alpha}^2)\mathcal{R}(\hat{\gamma}_\ell)^{\bar{q}} \\
\mathbb{E}[\Delta_{2\ell}^2] &\leq \mathbb{E}\left[\bar{\sigma}^2\mathbb{E}\left[\mathbb{E}_\ell[v(W_{i,\cdot})^2]|I_\ell^c\right]\right] \\
&= \mathbb{E}\left[\bar{\sigma}^2\mathbb{E}_\ell[v(W_{i,\cdot})^2]\right] \\
&= \bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell) \\
\mathbb{E}|\Delta_{3\ell}| &\leq \frac{1}{2}\mathbb{E}\left[\sqrt{m}\mathbb{E}[\mathbb{E}_\ell\{|u(W_{i,\cdot})v(W_{i,\cdot})|\}|I_\ell^c]\right] \\
&= \frac{1}{2}\sqrt{m}\mathbb{E}[\mathbb{E}_\ell\{|u(W_{i,\cdot})v(W_{i,\cdot})|\}] \\
&\leq \frac{1}{2}\sqrt{m}\sqrt{\mathcal{R}(\hat{\gamma}_\ell)}\sqrt{\mathcal{R}(\hat{\alpha}_\ell)}.
\end{aligned}$$

To verify the last line, use the shorthand $u_i = u(W_{i,\cdot})$ and $v_i = v(W_{i,\cdot})$. Then

$$\begin{aligned}
\mathbb{E}[\mathbb{E}_\ell\{|u(W_{i,\cdot})v(W_{i,\cdot})|\}] &= \frac{1}{m}\mathbb{E}[u^T v] \\
&\leq \frac{1}{m}(\mathbb{E}[u^T u])^{1/2}(\mathbb{E}[v^T v])^{1/2} \\
&= \left(\frac{1}{m}\mathbb{E}[u^T u]\right)^{1/2}\left(\frac{1}{m}\mathbb{E}[v^T v]\right)^{1/2} \\
&= \left(\frac{1}{m}\mathbb{E}[\mathbb{E}_\ell[u_i^2]]\right)^{1/2}\left(\frac{1}{m}\mathbb{E}[\mathbb{E}_\ell[v_i^2]]\right)^{1/2} \\
&= \sqrt{\mathcal{R}(\hat{\gamma}_\ell)}\sqrt{\mathcal{R}(\hat{\alpha}_\ell)}.
\end{aligned}$$

5. Collecting results gives

$$\begin{aligned}
\mathbb{P}(|\Delta_{1\ell}| > t_1) &\leq \frac{2(\bar{Q}+\bar{\alpha}^2)\mathcal{R}(\hat{\gamma}_\ell)^{\bar{q}}}{t_1^2} = \frac{\epsilon}{3L}; \\
\mathbb{P}(|\Delta_{2\ell}| > t_2) &\leq \frac{\bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell)}{t_2^2} = \frac{\epsilon}{3L}; \\
\mathbb{P}(|\Delta_{3\ell}| > t_3) &\leq \frac{m^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}}{2t_3} = \frac{\epsilon}{3L}.
\end{aligned}$$

Therefore with probability $1-\epsilon/L$, the following inequalities hold

$$|\Delta_{1\ell}| \le t_1 = \left(\frac{6L}{\epsilon}\right)^{1/2} (\bar{Q}+\bar{\alpha}^2)^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{\bar{q}/2};$$
$$|\Delta_{2\ell}| \le t_2 = \left(\frac{3L}{\epsilon}\right)^{1/2} \bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2};$$
$$|\Delta_{3\ell}| \le t_3 = \frac{3L}{2\epsilon} m^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{1/2}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.$$

Finally recall $m = n/L$.

□

Definition G.4 (Overall target and oracle)**.** *Let L be the number of folds. Define*

$$\hat{\theta} = \frac{1}{L}\sum_{\ell=1}^{L} \hat{\theta}_\ell, \quad \bar{\theta} = \frac{1}{L}\sum_{\ell=1}^{L} \bar{\theta}_\ell.$$

Lemma G.6 (Oracle approximation)**.** *Suppose the conditions of Lemma G.5 hold as well as Assumption G.2. Then with probability $1-\epsilon$*

$$\frac{n^{1/2}}{\sigma}|\hat{\theta}-\bar{\theta}| \le \Delta = \frac{3L}{\epsilon\sigma}\left[(\bar{Q}^{1/2}+\bar{\alpha})\{\mathcal{R}(\hat{\gamma}_\ell)\}^{\bar{q}/2} + \bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} + \{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}\right].$$

Proof. We confirm Chernozhukov et al. (2021, Proposition S6) generalizes to the new norm.

1. Decomposition.

   By Lemma G.4

   $$n^{1/2}(\hat{\theta}-\bar{\theta}) = \frac{n^{1/2}}{m^{1/2}}\frac{1}{L}\sum_{\ell=1}^{L} m^{1/2}(\hat{\theta}_\ell - \bar{\theta}_\ell) = L^{1/2}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{j=1}^{3}\Delta_{j\ell}.$$

2. Union bound.

   Define the events

   $$\mathcal{E}_\ell = \{\forall j \in \{1,2,3\},\ |\Delta_{j\ell}| \le t_j\}, \quad \mathcal{E} = \cap_{\ell=1}^{L}\mathcal{E}_\ell, \quad \mathcal{E}^c = \cup_{\ell=1}^{L}\mathcal{E}_\ell^c.$$

   Hence by the union bound and Lemma G.5,

   $$\mathbb{P}(\mathcal{E}^c) \le \sum_{\ell=1}^{L}\mathbb{P}(\mathcal{E}_\ell^c) \le L\frac{\epsilon}{L} = \epsilon.$$

3. Collecting results.

   Therefore with probability $1-\epsilon$,

$$n^{1/2}|\hat{\theta}-\bar{\theta}| \leq L^{1/2}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{j=1}^{3}|\Delta_{jk}| \leq L^{1/2}\frac{1}{L}\sum_{\ell=1}^{L}\sum_{j=1}^{3}t_j = L^{1/2}\sum_{j=1}^{3}t_j.$$

   Finally, we simplify (t_1,t_2,t_3) from Lemma G.5. For $a,b>0$, $(a+b)^{1/2} \leq a^{1/2}+b^{1/2}$. Moreover, $3 > 6^{1/2} > 3/2$. Finally, for $\epsilon \leq 1$, $\epsilon^{-1/2} \leq \epsilon^{-1}$. In summary

$$\begin{aligned}
t_1 &= \left(\frac{6L}{\epsilon}\right)^{1/2}(\bar{Q}+\bar{\alpha}^2)^{1/2}\{\mathcal{R}(\hat{\gamma}_\ell)\}^{\bar{q}/2} \leq \frac{3L^{1/2}}{\epsilon}(\bar{Q}^{1/2}+\bar{\alpha})\{\mathcal{R}(\hat{\gamma}_\ell)\}^{\bar{q}/2};\\
t_2 &= \left(\frac{3L}{\epsilon}\right)^{1/2}\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \leq \frac{3L^{1/2}}{\epsilon}\bar{\sigma}\{\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2};\\
t_3 &= \frac{3L^{1/2}}{2\epsilon}\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2} \leq \frac{3L^{1/2}}{\epsilon}\{n\mathcal{R}(\hat{\gamma}_\ell)\mathcal{R}(\hat{\alpha}_\ell)\}^{1/2}.
\end{aligned}$$

□

Lemma G.7 (Berry Esseen Theorem for i.n.i.d. data; Shevtsova (2010))**.** *Suppose B_i are i.n.i.d. random variables with $\mathbb{E}[B_i]=0$, $\mathbb{E}[B_i^2]=\sigma_i^2$, $\mathbb{E}[B_i^3]=\xi_i^3$. Let $\sigma^2=\mathbb{E}_n[\sigma_i^2]$ and $\xi^3=\mathbb{E}_n[\xi_i^3]$. Then*

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{n^{1/2}}{\sigma}\mathbb{E}_n[B_i]\leq z\right\}-\Phi(z)\right| \leq c^{BE}\left(\frac{\xi}{\sigma}\right)^3 n^{-\frac{1}{2}}, \quad c^{BE}=0.5600.$$

Proof of Theorem 5.4. Fix $z\in\mathbb{R}$. First, we show that

$$\mathbb{P}\left\{\frac{n^{1/2}}{\sigma}(\hat{\theta}-\theta_0)\leq z\right\}-\Phi(z) \leq c^{BE}\left(\frac{\xi}{\sigma}\right)^3 n^{-\frac{1}{2}}+\frac{\Delta}{(2\pi)^{1/2}}+\epsilon,$$

where $\Phi(z)$ is the standard Gaussian cumulative distribution function and Δ is defined in Lemma G.6.

1. High probability bound.

   By Lemma G.6, w.p. $1-\epsilon$,

$$\frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta}-\hat{\theta}) \leq \frac{n^{1/2}}{\sigma}|\hat{\theta}-\bar{\theta}| \leq \Delta.$$

   Observe that

$$\mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\hat{\theta}-\theta_0)\leq z\right\} = \mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta}-\theta_0)\leq z+\frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta}-\hat{\theta})\right\} \leq \mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta}-\theta_0)\leq z+\Delta\right\}+\epsilon.$$

2. Mean value theorem.

   There exists some z' such that

$$\Phi(z+\Delta) - \Phi(z) = \nabla_z \Phi(z') \cdot \Delta \leq \frac{\Delta}{\sqrt{2\pi}}.$$

3. Berry Esseen theorem.

   By Lemma G.7

$$\sup_z \left| \mathbb{P}\left\{ \frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \leq z \right\} - \Phi(z) \right| \leq c^{BE} \left(\frac{\xi}{\sigma} \right)^3 n^{-\frac{1}{2}}$$

   taking

$$\begin{aligned}
\bar{\theta} - \theta_0 &= \frac{1}{n} \sum_{i \in [n]} [m(W_i, \gamma_0) + \alpha_0(W_i)\{Y_i - \gamma_0(W_i)\}] - \frac{1}{n} \sum_{i \in [n]} \mathbb{E}[m(W_{i,\cdot}, \gamma_0)] \\
&= \frac{1}{n} \sum_{i \in [n]} [m(W_i, \gamma_0) + \alpha_0(W_i)\{Y_i - \gamma_0(W_i)\} - \theta_i] \\
&= \frac{1}{n} \sum_{i \in [n]} \psi_i.
\end{aligned}$$

   The choice of $B_i = \psi_i$ satisfies the conditions of Lemma G.7 under Assumption G.2.

   Hence

$$\begin{aligned}
&\mathbb{P}\left\{ \frac{n^{\frac{1}{2}}}{\sigma}(\hat{\theta} - \theta_0) \leq z \right\} - \Phi(z) \\
&\leq \mathbb{P}\left\{ \frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta} - \theta_0) \leq z + \Delta \right\} - \Phi(z) + \epsilon \\
&= \mathbb{P}\left\{ \frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta} - \theta_0) \leq z + \Delta \right\} - \Phi(z+\Delta) + \Phi(z+\Delta) - \Phi(z) + \epsilon \\
&\leq c^{BE} \left(\frac{\xi}{\sigma} \right)^3 n^{-\frac{1}{2}} + \frac{\Delta}{\sqrt{2\pi}} + \epsilon.
\end{aligned}$$

Next, we show that

$$\Phi(z) - \mathbb{P}\left\{ \frac{n^{1/2}}{\sigma}(\hat{\theta} - \theta_0) \leq z \right\} \leq c^{BE} \left(\frac{\xi}{\sigma} \right)^3 n^{-\frac{1}{2}} + \frac{\Delta}{(2\pi)^{1/2}} + \epsilon.$$

1. High probability bound.

   By Lemma G.6, w.p. $1 - \epsilon$,

$$\frac{n^{\frac{1}{2}}}{\sigma}(\hat{\theta} - \bar{\theta}) \leq \frac{n^{1/2}}{\sigma} |\hat{\theta} - \bar{\theta}| \leq \Delta$$

hence

$$z - \Delta \leq z - \frac{n^{\frac{1}{2}}}{\sigma}(\hat{\theta} - \bar{\theta}).$$

Observe that

$$\begin{aligned}\mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta} - \theta_0) \leq z - \Delta\right\} &\leq \mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta} - \theta_0) \leq z - \frac{n^{\frac{1}{2}}}{\sigma}(\hat{\theta} - \bar{\theta})\right\} + \epsilon \\ &= \mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\hat{\theta} - \theta_0) \leq z\right\} + \epsilon.\end{aligned}$$

2. Mean value theorem.

   There exists some z' such that

$$\Phi(z) - \Phi(z - \Delta) = \nabla_z \Phi(z') \cdot \Delta \leq \frac{\Delta}{\sqrt{2\pi}}.$$

3. Berry Esseen theorem.

   As argued above,

$$\sup_z \left|\mathbb{P}\left\{\frac{n^{1/2}}{\sigma}(\bar{\theta} - \theta_0) \leq z\right\} - \Phi(z)\right| \leq c^{BE}\left(\frac{\xi}{\sigma}\right)^3 n^{-\frac{1}{2}}.$$

   Hence

$$\begin{aligned}&\Phi(z) - \mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\hat{\theta} - \theta_0) \leq z\right\} \\ &\leq \Phi(z) - \mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta} - \theta_0) \leq z - \Delta\right\} + \epsilon \\ &= \Phi(z) - \Phi(z - \Delta) + \Phi(z - \Delta) - \mathbb{P}\left\{\frac{n^{\frac{1}{2}}}{\sigma}(\bar{\theta} - \theta_0) \leq z - \Delta\right\} + \epsilon \\ &\leq \frac{\Delta}{\sqrt{2\pi}} + c^{BE}\left(\frac{\xi}{\sigma}\right)^3 n^{-\frac{1}{2}} + \epsilon.\end{aligned}$$

□

G.5 Variance estimation

Definition G.5 (Shorter notation)**.** *For $i \in I_\ell$, define*

$$\begin{aligned}\psi_i &= \psi(W_{i,\cdot}, \theta_i, \gamma_0, \alpha_0); \\ \hat{\psi}_i &= \psi(W_{i,\cdot}, \hat{\theta}, \hat{\gamma}_\ell, \hat{\alpha}_\ell).\end{aligned}$$

Lemma G.8 (Foldwise second moment)**.** *Suppose Assumption G.2 holds. Then*

$$\mathbb{E}_\ell\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2\} \le 4\left\{(\hat{\theta} - \theta_0)^2 + \sum_{j=4}^{6} \Delta_{j\ell}\right\},$$

where

$$\begin{aligned}
\Delta_{4\ell} &= \mathbb{E}_\ell\{m(W_{i,\cdot}, u)^2\};\\
\Delta_{5\ell} &= \mathbb{E}_\ell[\{\hat{\alpha}_\ell(W_{i,\cdot})u(W_{i,\cdot})\}^2];\\
\Delta_{6\ell} &= \mathbb{E}_\ell[v(W_{i,\cdot})^2\{Y_i - \gamma_0(W_{i,\cdot})\}^2].
\end{aligned}$$

Proof. Write

$$\begin{aligned}
\hat{\psi}_i - \psi_i &= m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W_i)\{Y_i - \hat{\gamma}_\ell(W_i)\} - \hat{\theta}\\
&\quad - [m(W_i, \gamma_0) + \alpha_0(W_i)\{Y_i - \gamma_0(W_i)\} - \theta_i]\\
&\quad \pm \hat{\alpha}_\ell\{Y_i - \gamma_0(W_i)\}\\
&= (\theta_i - \hat{\theta}) + m(W_i, u) - \hat{\alpha}_\ell(W_i)u(W_i) + v(W_i)\{Y_i - \gamma_0(W_i)\}.
\end{aligned}$$

Hence

$$\hat{\psi}_i - \psi_i + \theta_0 - \theta_i = (\theta_0 - \hat{\theta}) + m(W_i, u) - \hat{\alpha}_\ell(W_i)u(W_i) + v(W_i)\{Y_i - \gamma_0(W_i)\}.$$

Therefore

$$(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2 \le 4\left[(\theta_0 - \hat{\theta})^2 + m(W_i, u)^2 + \{\hat{\alpha}_\ell(W_i)u(W_i)\}^2 + v(W_i)^2\{Y_i - \gamma_0(W_i)\}^2\right].$$

Finally take $\mathbb{E}_\ell[\cdot]$ of both sides. □

Lemma G.9 (Residuals)**.** *Suppose Assumption G.1 holds and*

$$\mathbb{E}[\varepsilon_i^2 | W_{i,\cdot}] \le \bar{\sigma}^2, \quad \|\hat{\alpha}_\ell\|_\infty \le \bar{\alpha}'.$$

Then with probability $1 - \epsilon'/(2L)$*,*

$$\begin{aligned}
\Delta_{4\ell} &\le t_4 = \frac{6L}{\epsilon'}\bar{Q}\mathcal{R}(\hat{\gamma}_\ell)^{\bar{q}};\\
\Delta_{5\ell} &\le t_5 = \frac{6L}{\epsilon'}(\bar{\alpha}')^2\mathcal{R}(\hat{\gamma}_\ell);\\
\Delta_{6\ell} &\le t_6 = \frac{6L}{\epsilon'}\bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell).
\end{aligned}$$

Proof. We proceed in steps analogous to Lemma G.5.

1. Markov inequality implies

$$\mathbb{P}(|\Delta_{4\ell}| > t_4) \leq \frac{\mathbb{E}(|\Delta_{4\ell}|)}{t_4};$$
$$\mathbb{P}(|\Delta_{5\ell}| > t_5) \leq \frac{\mathbb{E}(|\Delta_{5\ell}|)}{t_5};$$
$$\mathbb{P}(|\Delta_{6\ell}| > t_6) \leq \frac{\mathbb{E}(|\Delta_{6\ell}|)}{t_6}.$$

2. Law of iterated expectations implies

$$\mathbb{E}(|\Delta_{4\ell}|) = \mathbb{E}\{\mathbb{E}(|\Delta_{4\ell}| \mid I_\ell^c)\};$$
$$\mathbb{E}(|\Delta_{5\ell}|) = \mathbb{E}\{\mathbb{E}(|\Delta_{5\ell}| \mid I_\ell^c)\};$$
$$\mathbb{E}(|\Delta_{6\ell}|) = \mathbb{E}\{\mathbb{E}(|\Delta_{6\ell}| \mid I_\ell^c)\}.$$

3. Bounding conditional moments

   Note that

$$\mathbb{E}[\Delta_{4\ell}|I_\ell^c] = \mathbb{E}[\mathbb{E}_\ell\{m(W_{i,\cdot}, u)^2\}|I_\ell^c] \leq \bar{Q}\{\mathbb{E}[\mathbb{E}_\ell\{u(W_{i,\cdot})^2\}|I_\ell^c]\}^{\bar{q}}$$

   where in the last expression we use Jensen's inequality and $\bar{q} \in (0, 1]$ to argue that

$$\begin{aligned}\mathbb{E}[\mathbb{E}_\ell[m(W_{i,\cdot}, u)^2]|I_\ell^c] &= \mathbb{E}_\ell[\mathbb{E}[m(W_{i,\cdot}, u)^2|I_\ell^c]] \\ &\leq \bar{Q}\mathbb{E}_\ell[\{\mathbb{E}[u(W_i)^2|I_\ell^c]\}^{\bar{q}}] \\ &\leq \bar{Q}\{\mathbb{E}_\ell[\mathbb{E}[u(W_i)^2|I_\ell^c]]\}^{\bar{q}} \\ &= \bar{Q}\{\mathbb{E}[\mathbb{E}_\ell[u(W_i)^2]|I_\ell^c]\}^{\bar{q}}.\end{aligned}$$

   Similarly

$$\mathbb{E}[\Delta_{5\ell}|I_\ell^c] = \mathbb{E}[\mathbb{E}_\ell[\{\hat{\alpha}_\ell(W_{i,\cdot})u(W_{i,\cdot})\}^2]|I_\ell^c] \leq (\bar{\alpha}')^2\mathbb{E}[\mathbb{E}_\ell\{u(W_{i,\cdot})^2\}|I_\ell^c].$$

   Finally

$$\mathbb{E}[\Delta_{6\ell}|I_\ell^c] = \mathbb{E}[\mathbb{E}_\ell[\{v(W_{i,\cdot})[Y_i - \gamma_0(W_{i,\cdot})]\}^2]|I_\ell^c] \leq \bar{\sigma}^2\mathbb{E}[\mathbb{E}_\ell\{v(W_{i,\cdot})^2\}|I_\ell^c].$$

4. Law of iterated expectations and Jensen's inequality imply

$$\begin{aligned}
\mathbb{E}[\Delta_{4\ell}] &\leq \mathbb{E}\left[\bar{Q}\{\mathbb{E}[\mathbb{E}_\ell\{u(W_{i,\cdot})^2\}|I_\ell^c]\}^{\bar{q}}\right] \\
&\leq \bar{Q}\{\mathbb{E}\left[\mathbb{E}[\mathbb{E}_\ell\{u(W_{i,\cdot})^2\}|I_\ell^c]\right]\}^{\bar{q}} \\
&= \bar{Q}\{\mathbb{E}\left[\mathbb{E}_\ell\{u(W_{i,\cdot})^2\}\right]\}^{\bar{q}} \\
&= \bar{Q}\mathcal{R}(\hat{\gamma}_\ell)^{\bar{q}}; \\
\mathbb{E}[\Delta_{5\ell}] &\leq \mathbb{E}\left[(\bar{\alpha}')^2\mathbb{E}[\mathbb{E}_\ell\{u(W_{i,\cdot})^2\}|I_\ell^c]\right] \\
&= \mathbb{E}\left[(\bar{\alpha}')^2\mathbb{E}_\ell\{u(W_{i,\cdot})^2\}\right] \\
&= (\bar{\alpha}')^2\mathcal{R}(\hat{\gamma}_\ell); \\
\mathbb{E}[\Delta_{6\ell}] &\leq \mathbb{E}\left[\bar{\sigma}^2\mathbb{E}[\mathbb{E}_\ell\{v(W_{i,\cdot})^2\}|I_\ell^c]\right] \\
&= \mathbb{E}\left[\bar{\sigma}^2\mathbb{E}_\ell\{v(W_{i,\cdot})^2\}\right] \\
&= \bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell).
\end{aligned}$$

5. Collecting results gives

$$\begin{aligned}
\mathbb{P}(|\Delta_{4\ell}| > t_4) &\leq \frac{\bar{Q}\mathcal{R}(\hat{\gamma}_\ell)^{\bar{q}}}{t_4} = \frac{\epsilon'}{6L}; \\
\mathbb{P}(|\Delta_{5\ell}| > t_5) &\leq \frac{(\bar{\alpha}')^2\mathcal{R}(\hat{\gamma}_\ell)}{t_5} = \frac{\epsilon'}{6L}; \\
\mathbb{P}(|\Delta_{6\ell}| > t_6) &\leq \frac{\bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell)}{t_6} = \frac{\epsilon'}{6L}.
\end{aligned}$$

Therefore with probability $1 - \epsilon'/(2L)$, the following inequalities hold:

$$\begin{aligned}
|\Delta_{4\ell}| &\leq t_4 = \frac{6L}{\epsilon'}\bar{Q}\mathcal{R}(\hat{\gamma}_\ell)^{\bar{q}}; \\
|\Delta_{5\ell}| &\leq t_5 = \frac{6L}{\epsilon'}(\bar{\alpha}')^2\mathcal{R}(\hat{\gamma}_\ell); \\
|\Delta_{6\ell}| &\leq t_6 = \frac{6L}{\epsilon'}\bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell).
\end{aligned}$$

□

Lemma G.10 (Oracle approximation)**.** *Suppose the conditions of Lemma G.9 as well as Assumption G.2 hold. Then with probability* $1 - \epsilon'/2$

$$\mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2\} \leq \Delta' = 4(\hat{\theta} - \theta_0)^2 + \frac{24L}{\epsilon'}\left[\{\bar{Q} + (\bar{\alpha}')^2\}\mathcal{R}(\hat{\gamma}_\ell)^{\bar{q}} + \bar{\sigma}^2\mathcal{R}(\hat{\alpha}_\ell)\right].$$

Proof. We proceed in steps.

1. Decomposition.

   By Lemma G.8

$$\mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2\} = \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_\ell\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2\}$$
$$\leq 4(\hat{\theta} - \theta_0)^2 + \frac{4}{L}\sum_{\ell=1}^{L}\sum_{j=4}^{6}\Delta_{j\ell}.$$

2. Union bound.

   Define the events

$$\mathcal{E}'_\ell = \{\forall j \in \{4,5,6\},\ |\Delta_{j\ell}| \leq t_j\}, \quad \mathcal{E}' = \cap_{\ell=1}^{L}\mathcal{E}'_\ell, \quad (\mathcal{E}')^c = \cup_{\ell=1}^{L}(\mathcal{E}'_\ell)^c.$$

   Hence by the union bound and Lemma G.9,

$$\mathbb{P}\{(\mathcal{E}')^c\} \leq \sum_{\ell=1}^{L}\mathbb{P}\{(\mathcal{E}'_\ell)^c\} \leq L\frac{\epsilon'}{2L} = \frac{\epsilon'}{2}.$$

3. Collecting results.

   Therefore with probability $1 - \epsilon'/2$,

$$\mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2\} \leq 4(\hat{\theta} - \theta_i)^2 + \frac{4}{L}\sum_{\ell=1}^{L}\sum_{j=4}^{6}|\Delta_{j\ell}|$$
$$\leq 4(\hat{\theta} - \theta_i)^2 + \frac{4}{L}\sum_{\ell=1}^{L}\sum_{j=4}^{6}t_j$$
$$= 4(\hat{\theta} - \theta_i)^2 + 4\sum_{j=4}^{6}t_j.$$

   Finally appeal to Lemma G.9 for (t_4, t_5, t_6).

□

Lemma G.11 (Markov inequality)**.** *Suppose* $\mathbb{E}[\psi_i^4] = \chi_i^4 < \infty$. *Then with probability* $1 - \epsilon'/2$

$$|\mathbb{E}_n(\psi_i^2) - \sigma^2| \leq \Delta'' = \left(\frac{2}{\epsilon'}\right)^{1/2}\frac{\chi^2}{n^{1/2}}.$$

Proof. Recall that

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2, \quad \chi^4 = \frac{1}{n}\sum_{i=1}^{n}\chi_i^4.$$

Let

$$B_i = \psi_i^2, \quad \bar{B} = \mathbb{E}_n[B_i].$$

Note that

$$\begin{aligned}
\mathbb{E}[\bar{B}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[B_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\psi_i^2] = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 = \sigma^2; \\
\mathbb{V}[\bar{B}] &= \frac{\sum_{i=1}^{n}\mathbb{V}(B_i)}{n^2} \leq \frac{\sum_{i=1}^{n}\mathbb{E}[B_i^2]}{n^2} = \frac{\sum_{i=1}^{n}\mathbb{E}[\psi_i^4]}{n^2} = \frac{\sum_{i=1}^{n}\chi_i^4}{n^2} = \frac{\chi^4}{n}.
\end{aligned}$$

By Markov inequality

$$\mathbb{P}(|\bar{B} - \mathbb{E}[\bar{B}]| > t) \leq \frac{\mathbb{V}[\bar{B}]}{t^2} = \frac{\epsilon'}{2}.$$

Solving, the final inequality implies

$$t = \sqrt{\frac{2}{\epsilon'}}\frac{\chi^2}{\sqrt{n}}.$$

□

Proof of Theorem 5.5. We proceed in steps.

1. Decomposition of variance estimator.

   Write

   $$\begin{aligned}
   \hat{\sigma}^2 &= \mathbb{E}_n(\hat{\psi}_i^2) \\
   &= \mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \psi_i)^2\} \\
   &= \mathbb{E}_n\{(\hat{\psi}_i - \psi_i)^2\} + 2\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)\psi_i\} + \mathbb{E}_n(\psi_i^2).
   \end{aligned}$$

   Hence

   $$\hat{\sigma}^2 - \mathbb{E}_n(\psi_i^2) = \mathbb{E}_n\{(\hat{\psi}_i - \psi_i)^2\} + 2\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)\psi_i\}.$$

2. Decomposition of difference.

   Next write

   $$\begin{aligned}
   \hat{\sigma}^2 - (\sigma^2 + \text{BIAS}) &= \{\hat{\sigma}^2 - \mathbb{E}_n(\psi_i^2) - \text{BIAS}\} + \{\mathbb{E}_n(\psi_i^2) - \sigma^2\} \\
   &\leq \{\hat{\sigma}^2 - \mathbb{E}_n(\psi_i^2) - \text{BIAS}\} + \Delta''
   \end{aligned}$$

where the last line holds with probability $1 - \epsilon'/2$ by Lemma G.11. In what follows, we focus on the former term. We solve for $\text{BIAS} = \text{BIAS}_1 + \text{BIAS}_2$ as a function of Δ_{OUT}, using the decomposition

$$\hat{\sigma}^2 - \mathbb{E}_n(\psi_i^2) - \text{BIAS} = \mathbb{E}_n\{(\hat{\psi}_i - \psi_i)^2\} - \text{BIAS}_1 + 2\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)\psi_i\} - \text{BIAS}_2.$$

3. BIAS_1

Write

$$\begin{aligned}
&\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)^2\} \\
&= \mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i + \theta_i - \theta_0)^2\} \\
&= \mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2\} + \mathbb{E}_n\{(\theta_i - \theta_0)^2\} + 2\mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)(\theta_i - \theta_0)\} \\
&\leq \mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2\} + \mathbb{E}_n\{(\theta_i - \theta_0)^2\} + 2[\mathbb{E}_n\{(\hat{\psi}_i - \psi_i + \theta_0 - \theta_i)^2\}]^{1/2}[\mathbb{E}_n\{(\theta_i - \theta_0)^2\}]^{1/2} \\
&\leq \Delta' + \Delta_{\text{OUT}} + 2(\Delta')^{1/2}\Delta_{\text{OUT}}^{1/2}.
\end{aligned}$$

where the last line holds with probability $1 - \epsilon'/2$ appealing to Lemma G.10. Taking $\text{BIAS}_1 = \Delta_{\text{OUT}}$, we have shown

$$\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)^2\} - \text{BIAS}_1 \leq \Delta' + 2(\Delta')^{1/2}\Delta_{\text{OUT}}^{1/2}.$$

4. BIAS_2

Next, write

$$\begin{aligned}
&\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)\psi_i\} \\
&\leq \left[\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)^2\}\right]^{1/2} \left\{\mathbb{E}_n(\psi_i^2)\right\}^{1/2} \\
&\leq \left[\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)^2\}\right]^{1/2} \left\{|\mathbb{E}_n(\psi_i^2) - \sigma^2| + \sigma^2\right\}^{1/2} \\
&\leq \{\Delta' + \Delta_{\text{OUT}} + 2(\Delta')^{1/2}\Delta_{\text{OUT}}^{1/2}\}^{1/2} \cdot \{\Delta'' + \sigma^2\}^{1/2}
\end{aligned}$$

where the last line holds with probability $1 - \epsilon'$ appealing to Lemmas G.10 and G.11 as well as the analysis for BIAS_1. In summary,

$$\begin{aligned}
2\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)\psi_i\} &\leq 2\{\Delta' + \Delta_{\text{OUT}} + 2(\Delta')^{1/2}\Delta_{\text{OUT}}^{1/2}\}^{1/2} \cdot \{\Delta'' + \sigma^2\}^{1/2} \\
&\leq 2\{(\Delta')^{1/2} + \Delta_{\text{OUT}}^{1/2} + 2^{1/2}(\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\} \cdot \{(\Delta'')^{1/2} + \sigma\}
\end{aligned}$$

Taking $\text{BIAS}_2 = 2\Delta_{\text{OUT}}^{1/2}\sigma$, we have shown

$$2\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)\psi_i\} - \text{BIAS}_2 \leq 2(\Delta')^{1/2}\{(\Delta'')^{1/2} + \sigma\} + 2\Delta_{\text{OUT}}^{1/2}(\Delta'')^{1/2} + 2^{3/2}(\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\{(\Delta'')^{1/2} + \sigma\}.$$

5. Collecting results

In summary, with probability $1 - \epsilon'$.

$$\begin{aligned}
&\hat{\sigma}^2 - (\sigma^2 + \text{BIAS}) \\
&\leq \{\hat{\sigma}^2 - \mathbb{E}_n(\psi_i^2) - \text{BIAS}\} + \Delta'' \\
&= \mathbb{E}_n\{(\hat{\psi}_i - \psi_i)^2\} - \text{BIAS}_1 + 2\mathbb{E}_n\{(\hat{\psi}_i - \psi_i)\psi_i\} - \text{BIAS}_2 + \Delta'' \\
&\leq \Delta' + 2(\Delta')^{1/2}\Delta_{\text{OUT}}^{1/2} \\
&\quad + 2(\Delta')^{1/2}\{(\Delta'')^{1/2} + \sigma\} + 2\Delta_{\text{OUT}}^{1/2}(\Delta'')^{1/2} + 2^{3/2}(\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\{(\Delta'')^{1/2} + \sigma\} \\
&\quad + \Delta'' \\
&= \Delta' + \Delta'' \\
&\quad + 2(\Delta')^{1/2}\{(\Delta'')^{1/2} + \sigma + \Delta_{\text{OUT}}^{1/2}\} \\
&\quad + 2\Delta_{\text{OUT}}^{1/2}(\Delta'')^{1/2} \\
&\quad + 2^{3/2}(\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\{(\Delta'')^{1/2} + \sigma\} \\
&\leq \Delta' + \Delta'' \\
&\quad + 3(\Delta')^{1/2}\{(\Delta'')^{1/2} + \sigma + \Delta_{\text{OUT}}^{1/2}\} \\
&\quad + 3(\Delta'')^{1/2}\{\Delta_{\text{OUT}}^{1/2} + (\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\} \\
&\quad + 3(\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\sigma \\
&= \Delta' + \Delta'' + 3[(\Delta')^{1/2}\{(\Delta'')^{1/2} + \sigma + \Delta_{\text{OUT}}^{1/2}\} + (\Delta'')^{1/2}\{\Delta_{\text{OUT}}^{1/2} + (\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\} + (\Delta')^{1/4}\Delta_{\text{OUT}}^{1/4}\sigma].
\end{aligned}$$

Combining terms yields the desired result.

□

G.6 Confidence interval

Proof of Corollary 5.3. Immediately from Δ in Theorem 5.4, $\hat{\theta} \xrightarrow{p} \theta_0$ and

$$\lim_{n\to\infty} \mathbb{P}\left(\theta_0 \in \left[\hat{\theta} \pm \frac{\sigma}{n^{1/2}}\right]\right) = 1 - a.$$

For the desired result, it is sufficient that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2 + \text{BIAS}$, which follows from Δ' and Δ'' in Theorem 5.5. □

Proof of Corollary G.1. By Lemma G.1, write the regularity condition on moments as By Lemma G.1, write the regularity condition on moments as By Lemma G.1, write the regularity condition on moments as By Lemma G.1, write the regularity condition on moments as

$$\left\{(\kappa/\sigma)^3 + \zeta^2\right\} n^{-1/2} \lesssim \left\{\left(h^{-1/6}\right)^3 + (h^{-3/4})^2\right\} n^{-1/2} \lesssim h^{-3/2} n^{-1/2}.$$

By Lemma G.1, write the first learning rate condition as

$$\left(\bar{Q}^{1/2} + \bar{\alpha}/\sigma + \bar{\alpha}'\right) \{\mathcal{R}(\hat{\gamma})\}^{1/2} \lesssim \left(h^{-1} + h^{-1}/h^{-1/2} + \bar{\alpha}'\right) \{\mathcal{R}(\hat{\gamma})\}^{1/2} \lesssim \left(h^{-1} + \bar{\alpha}'\right) \{\mathcal{R}(\hat{\gamma})\}^{1/2}.$$

By Chernozhukov et al. (2021, Lemma S.9), write the second learning rate condition as

$$\bar{\sigma}\{\mathcal{R}(\hat{\alpha}^h)\}^{1/2} \lesssim \bar{\sigma} h^{-1} \{\mathcal{R}(\hat{\alpha})\}^{1/2}.$$

By Lemma G.1 and Chernozhukov et al. (2021, Lemma S.9), write the third learning rate condition as

$$\{n\mathcal{R}(\hat{\gamma})\mathcal{R}(\hat{\alpha}^h)\}^{1/2}/\sigma \lesssim \{n\mathcal{R}(\hat{\gamma})\mathcal{R}(\hat{\alpha})\}^{1/2} h^{-1}/h^{-1/2} = h^{-1/2}\{n\mathcal{R}(\hat{\gamma})\mathcal{R}(\hat{\alpha})\}^{1/2}.$$

□

H Nonlinear factor model

H.1 Notation and preliminaries

We lighten notation by denoting $\mathcal{R}_\gamma = \mathcal{R}(\hat{\gamma}_\ell)$ and $\mathcal{R}_\alpha = \mathcal{R}(\hat{\alpha}_\ell)$. Note that the distinction between n and $m = \frac{n}{2}$ is irrelevant in the context of $(\mathcal{R}_\gamma, \mathcal{R}_\alpha)$ due to the absolute constant C.

Lemma H.1 (Low rank approximation (Agarwal et al., 2021)). *Suppose Assumption 5.11 holds for some fixed $\mathcal{H}(q, S, C_H)$. Then for any small $\delta > 0$, there exists $\boldsymbol{A}^{(lr)}$ such that*

$$r = rank(\boldsymbol{A}^{(lr)}) \leq C \cdot \delta^{-q}, \quad \Delta_E = \|\boldsymbol{A} - \boldsymbol{A}^{(lr)}\|_{\max} \leq C_H \cdot \delta^S$$

where C is allowed to depend on (q, S).

H.2 Main result

Proof of Corollary 5.4. From Lemma H.1

$$r \leq C \cdot \delta^{-q}, \quad \Delta_E \leq C \cdot \delta^S.$$

The conditions Corollary 5.4 imply $(\sigma, \bar{\sigma}, \bar{\alpha}, \bar{\alpha}', \bar{Q})$ are irrelevant, so we wish to verify the following simplified rate conditions from Corollary 5.3:

$$\mathcal{R}_\gamma \to 0, \quad \mathcal{R}_\alpha \to 0, \quad \sqrt{n\mathcal{R}_\gamma \mathcal{R}_\alpha} \to 0.$$

Furthermore, under the conditions of Corollary 5.4, the relevant terms in $(\mathcal{R}_\gamma, \mathcal{R}_\alpha)$ simplify. For $\mathcal{R}_\gamma$, the relevant terms from Theorem 5.2 are

$$\mathcal{R}_\gamma \leq Cr^3 \left\{ \frac{1}{n} + \frac{p}{n^2} + \frac{1}{p} + \left(1 + \frac{p}{n}\right) \Delta_E^2 + p\Delta_E^4 \right\}.$$

For $\mathcal{R}_\alpha$, the relevant terms from Theorem 5.3 are

$$\mathcal{R}_\alpha \leq Cr^5 \left\{ \frac{1}{n} + \frac{1}{p} + \frac{p}{n^2} + \frac{n}{p^2} + \left(1 + \frac{p}{n} + \frac{n}{p}\right) \Delta_E^2 + (n+p)\Delta_E^4 + np\Delta_E^6 \right\}.$$

There are two cases.

1. $n \geq p$. In particular, $n = p^\upsilon$ with $\upsilon \geq 1$. Then

$$\mathcal{R}_\gamma \leq Cr^3 \left(\frac{1}{p} + \Delta_E^2 + p\Delta_E^4 \right) \leq C\delta^{-3q} \left(\frac{1}{p} + \delta^{2S} + p\delta^{4S} \right).$$

The three terms are equalized with $\delta^{2S} = p^{-1}$. Hence

$$\mathcal{R}_\gamma \leq C\delta^{-3q} \frac{1}{p} = Cp^{\frac{3q}{2S}} \frac{1}{p} = Cp^{\frac{3q}{2S}-1}.$$

Similarly

$$\mathcal{R}_\alpha \leq Cr^5 \left(\frac{n}{p^2} + \frac{n}{p}\Delta_E^2 + n\Delta_E^4 + np\Delta_E^6 \right) \leq C\delta^{-5q} \left(\frac{n}{p^2} + \frac{n}{p}\delta^{2S} + n\delta^{4S} + np\delta^{6S} \right).$$

The four terms are equalized with $\delta^{2S} = p^{-1}$. Hence

$$\mathcal{R}_\alpha \leq C\delta^{-5q} \frac{n}{p^2} = Cp^{\frac{5q}{2S}} \frac{n}{p^2} = Cp^{\frac{5q}{2S}-2} n.$$

To satisfy $\mathcal{R}_\gamma \leq \mathcal{R}_\alpha \to 0$, it is sufficient that

$$p^{\frac{5q}{2S}-2+\upsilon} \to 0 \iff \frac{5q}{2S} - 2 + \upsilon < 0 \iff \frac{q}{S} < \frac{2}{5}(2 - \upsilon).$$

To satisfy $\sqrt{n\mathcal{R}_\gamma\mathcal{R}_\alpha} \to 0$, it is sufficient that

$$n^{\frac{1}{2}}p^{\frac{3q}{4S}-\frac{1}{2}}p^{\frac{5q}{4S}-1}n^{\frac{1}{2}} = np^{\frac{2q}{S}-\frac{3}{2}} \to 0 \iff \frac{2q}{S} - \frac{3}{2} + \upsilon < 0 \iff \frac{q}{S} < \frac{1}{2}\left(\frac{3}{2} - \upsilon\right).$$

In summary, a sufficient generalized factor model is one in which

$$\frac{q}{S} < \frac{2}{5}(2-\upsilon) \wedge \frac{1}{2}\left(\frac{3}{2} - \upsilon\right), \quad \upsilon \leq \frac{3}{2}.$$

2. $n \leq p$. In particular, $p = n^\upsilon$ with $\upsilon \geq 1$. Then

$$\mathcal{R}_\gamma \leq Cr^3\left(\frac{p}{n^2} + \frac{p}{n}\Delta_E^2 + p\Delta_E^4\right) \leq C\delta^{-3q}\left(\frac{p}{n^2} + \frac{p}{n}\delta^{2S} + p\delta^{4S}\right).$$

The three terms are equalized with $\delta^{2S} = n^{-1}$. Hence

$$\mathcal{R}_\gamma \leq C\delta^{-3q}\frac{p}{n^2} = Cn^{\frac{3q}{2S}}\frac{p}{n^2} = Cn^{\frac{3q}{2S}-2}p.$$

Similarly

$$\mathcal{R}_\alpha \leq Cr^5\left(\frac{p}{n^2} + \frac{p}{n}\Delta_E^2 + p\Delta_E^4 + np\Delta_E^6\right) \leq C\delta^{-5q}\left(\frac{p}{n^2} + \frac{p}{n}\delta^{2S} + p\delta^{4S} + np\delta^{6S}\right).$$

The four terms are equalized with $\delta^{2S} = n^{-1}$. Hence

$$\mathcal{R}_\alpha \leq C\delta^{-5q}\frac{p}{n^2} = Cn^{\frac{5q}{2S}}\frac{p}{n^2} = Cn^{\frac{5q}{2S}-2}p.$$

To satisfy $\mathcal{R}_\gamma \leq \mathcal{R}_\alpha \to 0$, it is sufficient that

$$n^{\frac{5q}{2S}-2+\upsilon} \to 0 \iff \frac{5q}{2S} - 2 + \upsilon < 0 \iff \frac{q}{S} < \frac{2}{5}(2-\upsilon).$$

To satisfy $\sqrt{n\mathcal{R}_\gamma\mathcal{R}_\alpha} \to 0$, it is sufficient that

$$n^{\frac{1}{2}}n^{\frac{3q}{4S}-1}p^{\frac{1}{2}}n^{\frac{5q}{4S}-1}p^{\frac{1}{2}} = n^{\frac{2q}{S}-\frac{3}{2}}p \to 0 \iff \frac{2q}{S} - \frac{3}{2} + \upsilon < 0 \iff \frac{q}{S} < \frac{1}{2}\left(\frac{3}{2} - \upsilon\right).$$

In summary, a sufficient generalized factor model is one in which

$$\frac{q}{S} < \frac{2}{5}(2-\upsilon) \wedge \frac{1}{2}\left(\frac{3}{2} - \upsilon\right), \quad \upsilon \leq \frac{3}{2}.$$

Note that the latter condition binds for $1 \leq \upsilon \leq \frac{3}{2}$. In conclusion, a sufficient generalized factor model is one in which $n = p^\upsilon$ or $p = n^\upsilon$ and

$$\frac{q}{S} < \frac{3}{4} - \frac{\upsilon}{2}, \quad 1 \leq \upsilon \leq \frac{3}{2}.$$

□

H.3 Nonlinearity

Remark H.1 (Dictionary)**.** *Under the conditions of Corollary 5.4, the relevant terms in* $(\mathcal{R}_\gamma, \mathcal{R}_\alpha)$ *are as before, instead using* (r', Δ'_E)*. To lighten notation, define* $q' = d_{\max} \cdot q$*. Then*

$$r' \leq C \cdot r^{d_{\max}} \leq C \cdot \delta^{-qd_{\max}} = C \cdot \delta^{-q'}$$

and

$$\Delta'_E \leq C\bar{A}^{d_{\max}} \cdot d_{\max}\Delta_E \leq C \cdot \Delta_E \leq C \cdot \delta^S.$$

Therefore the proof of Corollary 5.4 remains the same, updating q *as* $q' = d_{\max} \cdot q$*.*

I Simulation and application

I.1 Simulation design

Consider the following simulation design adapted from Agarwal et al. (2020a); Singh et al. (2020), with fixed (n, p, r). We focus on average treatment effect with corrupted covariates (Example A.1). A single observation consists of the triple $(Y_i, D_i, Z_{i,\cdot})$ for outcome, treatment, and corrupted covariates where $Y \in \mathbb{R}$, $D_i \in \{0, 1\}$, and $Z_{i,\cdot} \in \mathbb{R}^p$. A single observation is generated is as follows.

First, we generate signal from a factor model. Sample $\boldsymbol{U} \sim \mathcal{N}(0, \boldsymbol{I}_{n\times r})$ and $\boldsymbol{V} \sim \mathcal{N}(0, \boldsymbol{I}_{p\times r})$. Then set $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$. By construction,

$$\mathbb{E}[X_{ij}] = \mathbb{E}\left[\sum_{s=1}^{r} U_{is}V_{sj}\right] = \sum_{s=1}^{r} \mathbb{E}\left[U_{is}\right]\mathbb{E}\left[V_{sj}\right] = 0;$$
$$\mathbb{V}[X_{ij}] = \mathbb{V}\left[\sum_{s=1}^{r} U_{is}V_{sj}\right] = \sum_{s=1}^{r} \mathbb{V}\left[U_{is}\right]\mathbb{V}\left[V_{sj}\right] = r.$$

Draw response noise as $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Define the vector $\beta \in \mathbb{R}^p$ by $\beta_j = j^{-2}$. Then set

$$D_i \sim \text{Bernoulli}\{\Lambda(0.25X^T\beta)\}$$
$$Y_i = 2.2D_i + 1.2X_{i,\cdot}\beta + D_iX_{1i} + \varepsilon_i$$

where $\Lambda(t) = (0.95 - 0.05)\frac{\exp(t)}{1+\exp(t)} + 0.05$ is the truncated logistic function. The average treatment effect $\theta_0 = 2.2$ by construction.

Rather than observing the signal covariate $X_{i,\cdot}$, we observe the corrupted covariate

$$Z_{i,\cdot} = [X_{i,\cdot} + H_{i,\cdot}] \odot \pi_{i,\cdot}.$$

$H_{ij} \overset{i.i.d.}{\sim} F_H$ is drawn i.i.d. with mean zero and variance σ_H^2. π_{ij} is 1 with probability ρ and `NA` with probability $1 - \rho$. We consider different choices of the measurement error distribution F_H to corresponding to classical measurement error, discretization, and differential privacy. In summary, the three data corruption parameters are (F_H, σ_H, ρ). The remaining design parameters are (n, p, r) corresponding to the sample size, dimension of covariates, and rank of the signal.

For classical measurement error, $F_H = \mathcal{N}(0, \sigma_H^2)$. For discretization, we generate $Z_{ij} = sign(X_{ij}) \cdot Poisson(|X_{ij}|)$ and implicitly define F_H by $H_{ij} = Z_{ij} - X_{ij}$. Note that

$$\mathbb{E}[Z_{ij}|X_{ij}] = sign(X_{ij})\mathbb{E}[Poisson(|X_{ij}|)|X_{ij}] = sign(X_{ij})|X_{ij}| = X_{ij}$$

as desired. Below, we show that $\sigma_H^2 = \mathbb{V}[H_{ij}] = 1.7$ in this construction. In other words, we consider discretization with about a third as much variance as the signal. For differential privacy, $F_H = Laplace(0, \frac{\sigma_H}{\sqrt{2}})$.

Proposition I.1 (Discretization noise-to-signal ratio)**.** *Given some random variable X, define $P = Poisson(|X|)$. Suppose $Z = sign(X) \cdot P$. Define $H = Z - X$. Then $\mathbb{E}[H] = 0$ and $\mathbb{V}[H] = \mathbb{E}[|X|]$.*

Proof. To begin, write

$$\mathbb{E}[Z|X] = sign(X) \cdot \mathbb{E}[P|X] = sign(X) \cdot |X| = X.$$

By the law of total variance

$$\mathbb{V}[H] = \mathbb{E}[\mathbb{V}[H|X]] + \mathbb{V}[\mathbb{E}[H|X]].$$

Focusing on the latter term

$$\mathbb{E}[H|X] = \mathbb{E}[Z - X|X] = \mathbb{E}[Z|X] - X = 0.$$

Focusing on the former term

$$\mathbb{V}[H|X] = \mathbb{V}[Z|X] = \mathbb{E}[Z^2|X] - \{\mathbb{E}[Z|X]\}^2.$$

Moreover

$$\mathbb{E}[Z^2|X] = \mathbb{E}[P^2|X] = \mathbb{V}[P|X] + \{\mathbb{E}[P|X]\}^2 = |X| + X^2.$$

In summary

$$\mathbb{V}[H|X] = |X| + X^2 - X^2 = |X|,$$

and hence

$$\mathbb{V}[H] = \mathbb{E}[|X|] + \mathbb{V}[0].$$

□

I.2 Formalizing privacy

Proof of Proposition 6.1. Fix the commuting zone $i \in [n]$. We refer to the construction of the summary statistic

$$X_{ij} = f_j(\boldsymbol{M}^{(i)}) = \frac{1}{L_i} \sum_{\ell=1}^{L_i} M_{\ell j}^{(i)}$$

as the j-th query f_j about $\boldsymbol{M}^{(i)}$, where $j \in [p]$. To ensure privacy level ϵ_j for query f_j, a possible mechanism is, according to Dwork et al. (2006, Proposition 3.3)

$$Z_{ij} = X_{ij} + H_{ij}, \quad X_{ij} = f_j(\boldsymbol{M}^{(i)}), \quad H_{ij} \overset{i.i.d.}{\sim} \text{Laplace}(S(f_j)/\epsilon_j).$$

$S(f_j)$ is a quantity called the sensitivity of the query, to which we return below. If no individual appears in two commuting zones, the Bureau can achieve privacy level ϵ while publishing all $j \in [p]$ variables for this commuting zone by setting $\epsilon_j = \epsilon/p$.

We wish to characterize the resulting sub-exponential parameters. They are, by independence of the Laplacians,

$$K_a = \|H_{i,\cdot}\|_{\psi_a} = \max_{j \in [p]} \|H_{ij}\|_{\psi_a} = \max_j \sqrt{2} \cdot S(f_j)/\epsilon_j = \sqrt{2}/\epsilon \cdot p \max_j S(f_j);$$

$$\kappa^2 = \|\mathbb{E}[H_{i,\cdot}^T H_{i,\cdot}]\|_{op} = \max_{ij} \mathbb{V}(H_{ij}) = 2 \max_j S(f_j)^2/\epsilon_j^2 = 2/\epsilon^2 \cdot p^2 \max_j S(f_j)^2.$$

What remains is to define and characterize the the sensitivity $S(f_j)$. The sensitivity of the query f_j is the most that the query may vary if one individual in the microdata were replaced. Formally,

$$\max_{\boldsymbol{M}^{(i)}, \boldsymbol{M}^{(i')}} |f_j(\boldsymbol{M}^{(i)}) - f_j(\boldsymbol{M}^{(i')})| \leq S(f_j)$$

where $\boldsymbol{M}^{(i)}$ and $\boldsymbol{M}^{(i')}$ are two possible data sets of L_i individuals that differ in one individual.

In what follows, we suppress indexing by i to lighten notation. By hypothesis, each entry of microdata is bounded: $|M_{\ell j}| \leq \bar{A}$. This fact, together with the fact that the query f_j is a sample mean, provides a bound on the sensitivity $S(f_j)$. To begin, write

$$f_j(\boldsymbol{M}) = \frac{1}{L}\left\{\sum_{\ell=1}^{L} M_{\ell j}\right\} = \frac{1}{L}\left\{\sum_{\ell=1}^{L-1} M_{\ell j} + M_{\ell L}\right\}.$$

Therefore without loss of generality

$$f_j(\boldsymbol{M}) - f_j(\boldsymbol{M}') = \frac{1}{L}(M_{\ell L} - M'_{\ell L})$$

and hence

$$S(f_j) = \max_{\boldsymbol{M},\boldsymbol{M}'} |f_j(\boldsymbol{M}) - f_j(\boldsymbol{M}')| = \max_{\boldsymbol{M},\boldsymbol{M}'} \left|\frac{1}{L}(M_{\ell L} - M'_{\ell L})\right| \leq \frac{2\bar{A}}{L}.$$

□

Proof of Proposition 6.2. Fix the individual $i \in [n]$. The query is $X_{ij} = f_j(X_{i,\cdot})$. To ensure privacy level ϵ_j for query f_j, a possible mechanism is, according to Dwork et al. (2006, Proposition 3.3)

$$Z_{ij} = X_{ij} + H_{ij}, \quad \{H_{ij}\}_{j\in[T]} \overset{i.i.d.}{\sim} \text{Laplace}(S(f_j)/\epsilon_j).$$

The Bureau can achieve privacy level ϵ while publishing $j \in [T]$ variables for this individual by setting $\epsilon_j = \epsilon/T$.

We wish to characterize the resulting sub-exponential parameters. They are, by independence of the Laplacians,

$$K_a = \|H_{i,\cdot}\|_{\psi_a} = \max_{j\in[p]} \|H_{ij}\|_{\psi_a} = \max_j \sqrt{2}\cdot S(f_j)/\epsilon_j = \sqrt{2}/\epsilon \cdot T \max_j S(f_j);$$
$$\kappa^2 = \|\mathbb{E}[H_{i,\cdot}^T H_{i,\cdot}]\|_{op} = \max_{ij} \mathbb{V}(H_{ij}) = 2\max_j S(f_j)^2/\epsilon_j^2 = 2/\epsilon^2 \cdot T^2 \max_j S(f_j)^2.$$

What remains is to characterize the the sensitivity $S(f_j)$. The sensitivity of the query f_j is the most that the query may vary if one entry in the microdata were replaced. Formally,

$$\max_{X_{i,\cdot}, X'_{i,\cdot}} |f_j(X_{i,\cdot}) - f_j(X'_{i,\cdot})| \leq S(f_j)$$

where $X_{i,\cdot}$ and $X'_{i,\cdot}$ are two descriptions of an individual that differ in one characteristic. By hypothesis, each entry of microdata is bounded: $|X_{ij}| \leq \bar{A}$. Therefore

$$S(f_j) = \max_{X_{i,\cdot}, X'_{i,\cdot}} |f_j(X_{i,\cdot}) - f_j(X'_{i,\cdot})| = \max_{X_{i,\cdot}, X'_{i,\cdot}} |X_{ij} - X'_{ij}| \leq 2\bar{A}.$$

□

I.3 Empirical application

The variable definitions follow Autor et al. (2013). In the authors' original specification (Autor et al., 2013, Table 3, column 6), $X_{i,\cdot} \in \mathbb{R}^{14}$ consists of: a constant, an indicator for the 2000-2007 period, percentage of employment in manufacturing, percentage of college educated population, percentage of foreign-born population, percentage of employment among women, percentage of employment in routine occupations, average offshorability index of occupations, and Census division dummies.

In our augmented specification $X_{i,\cdot} \in \mathbb{R}^{30}$ consists of variables from the original specification as well as additional variables in (Autor et al., 2013, Appendix Table 2). These include percentages of the working age population: employed in manufacturing, employed in non-manufacturing, unemployed, not in the labor force, receiving disability benefits; average log weekly wages: manufacturing, non-manufactuing; average benefits per capita: individual transfers, retirement, disability, medical, federal income assistance, unemployment, TAA; and average household income per working age adult: total, wage and salary.