Semi-supervised Inference: General Theory and Estimation of Means

Anru Zhang[1], Lawrence D. Brown[2], and T. Tony Cai[2]

University of Wisconsin-Madison and University of Pennsylvania

June 24, 2016

Abstract

We propose a general semi-supervised inference framework focused on the estimation of the population mean. We consider both the ideal semi-supervised setting where infinitely many unlabeled samples are available, as well as the ordinary semi-supervised setting in which only a finite number of unlabeled samples is available. As usual in semi-supervised settings, there exists an unlabeled sample of covariate vectors and a labeled sample consisting of covariate vectors along with real-valued responses ("labels"). Otherwise the formulation is "assumption-lean" in that no major conditions are imposed on the statistical or functional form of the data. Estimators are proposed along with corresponding confidence intervals for the population mean. Theoretical analysis on both the asymptotic behavior and ℓ_2-risk for the proposed procedures are given. Surprisingly, the proposed estimators, based on a simple form of the least squares method, outperform the ordinary sample mean. The method is further extended to a nonparametric setting, in which the oracle rate can be achieved asymptotically. The proposed estimators are further illustrated by simulation studies and a real data example involving estimation of the homeless population.

Keywords: Confidence interval, efficiency, estimation of mean, limiting distribution, semi-supervised Inference.

[1] Department of Statistics, University of Wisconsin-Madison, Madison, WI.

[2] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. The research of Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

1 Introduction

Semi-supervised learning arises naturally in statistics and machine learning when the labels are more difficult or more expensive to acquire than the unlabeled data. While numerous algorithms have been proposed for semi-supervised learning, they are mostly focused on classification, where the labels are discrete values representing the classes to which the samples belong (see, e.g., Zhu (2008); Ando and Zhang (2007); Zhu and Goldberg (2009); Wang et al. (2009)). The analyses typically rely on two types of assumptions, distribution-based and margin-based. The margin-based analysis (see Vapnik (2013); Wang and Shen (2007); Wang et al. (2008, 2009)) generally assumes that the samples with different labels have some separation, and the additional unlabeled samples can help enhance the separation and achieve a better classification result. The distributional approach (see Blum and Mitchell (1998); Ando and Zhang (2005, 2007)) usually relies on some assumptions of a particular type of relation between labels and samples. These assumptions can be difficult to verify in practice. The setting with continuous valued y has also been discussed in the literature, see, e.g., Johnson and Zhang (2008), Lafferty and Wasserman (2008) and Chakrabortty and Cai (2016). For a survey of recent development in semi-supervised learning, readers are referred to Zhu and Goldberg (2009) and the references therein.

The general semi-supervised model can be formulated as follows. Let $(Y, X_1, X_2, \cdots, X_p)$ be a $(p+1)$-dimensional random vector following an unknown joint distribution $P = P(dy, dx_1, \ldots, dx_p)$. Denote by P_X the marginal distribution of $X = (X_1, X_2, \cdots, X_p)$. Suppose one observes n "labeled" samples from P,

$$[\mathbf{Y}, \mathbf{X}] = \{Y_k, X_{k1}, X_{k2}, \cdots, X_{kp}\}_{k=1}^{n}, \tag{1}$$

and, in addition, m "unlabeled" samples from the marginal distribution P_X

$$\mathbf{X}_{\text{add}} = \{X_{k1}, X_{k2}, \cdots, X_{kp}\}_{k=n+1}^{n+m}. \tag{2}$$

In this paper, we focus on estimation and statistical inference for one of the simplest features, namely the population mean $\theta = \mathbb{E}Y$. No specific distributional or marginal assumptions relating X and Y are made.

This inference of population mean under general semi-supervised learning framework has a variety of applications. We discuss the estimation of treatment effect (ATE) in Section 5.1 and a prototypical example involving survey data in Section 5.2. It is noteworthy that for some other problems that do not at first look like mean estimation, one can recast them as mean estimation, possibly after an appropriate transformation. Examples include estimation of the variance of Y or covariance between Y and a given X_i. In work that builds on a portion of the present paper, Azriel et al. (2016) considers construction of linear predictors in semi-supervised learning settings.

To estimate $\theta = \mathbb{E}Y$, the most straight-forward estimator is the sample average $\bar{\mathbf{Y}}$. Surprisingly, as we show later, a simple least-squares-based estimator, which exploits the unknown association of Y and X, outperforms $\bar{\mathbf{Y}}$. We first consider an ideal setting where there are infinitely many unlabeled samples, i.e., $m = \infty$. This is equivalent to the case of known marginal distribution P_X. We refer to this case as **ideal semi-supervised inference**. In this case, our proposed estimator is

$$\hat{\theta} = \bar{\mathbf{Y}} - \hat{\beta}_{(2)}^{\top}(\bar{\mathbf{X}} - \mu), \tag{3}$$

where $\hat{\beta}_{(2)}$ is the p-dimensional least squares estimator for the regression slopes and $\mu = \mathbb{E}X$ is the population mean of X. This estimator is analyzed in detail in Section 2.2. We then consider the more realistic setting where there are a finite number of unlabeled samples, i.e., $m < \infty$. Here one has only partial information about P_X. We call this case **ordinary semi-supervised inference**. In this setting, we propose to estimate θ by

$$\hat{\theta} = \bar{\mathbf{Y}} - \hat{\beta}_{(2)}^{\top}(\bar{\mathbf{X}} - \hat{\mu}), \tag{4}$$

where $\hat{\mu}$ denotes the sample average of both the labeled and unlabeled X's. The detailed analysis of this estimator is given in Section 2.3.

We will investigate the properties of these estimators and in particular establish their asymptotic distributions and the ℓ_2 risk bounds. Both the case of a fixed number of covariates and the case of a growing number of covariates are considered. The basic asymptotic theory in Section 2 begins with a setting in which the dimension, p, of X, is fixed and $n \to \infty$ (see Theorem 2.1). For ordinary semi-supervised learning, the asymptotic results are of non-trivial interest whenever $\liminf_{n\to\infty}(m_n/n) > 0$ (see Theorem 2.3(i)). We then formulate and prove asymptotic results in the setting where p also grows with n. In general, these results require the assumption that $p = o(\sqrt{n})$ (see Theorems 2.2 and 2.3(ii)). The limiting distribution results allow us to construct an asymptotically valid confidence interval based on the proposed estimators that is shorter than the traditional sample-mean-based confidence interval.

In Section 3 we propose a methodology for improving the results of Section 2 by introducing additional covariates as functions of those given in the original problem. We show the proposed estimator achieves an oracle rate asymptotically. This can be viewed as a nonparametric regression estimation procedure.

There are results in the sample-survey literature that are qualitatively related to what we propose. The earliest citation we are aware of is Cochran (1953, Chapter 7). See also Deng and Wu (1987) and more recently Lohr (2009, Chapter 3.2). In these references one collects a finite sample, without replacement, from a (large) finite population. There is a response Y and a single, real covariate, X. The distribution of X within the finite population is known. The sample-survey target of estimation is the mean of Y within the full population. In the case in which the size of this population is infinitely large, sampling without replacement and sampling with replacement are indistinguishable. In that case the results from this sampling theory literature coincide with out results for the ideal semi-supervised scenario with $p = 1$, both in terms of the proposed estimator and its asymptotic variance. Otherwise the sample-survey theory results differ from those within our formulation, although there is a conceptual relationship. In particular the theoretical population mean that is our target is different from the finite population mean that is the target of the sample-survey methods. In addition we allow $p > 1$ and as noted above, we also have asymptotic results for p growing with n. Most notably, our formulation includes the possibility of semi-supervised learning. We believe it should be possible, and sometimes of practical interest, to include semi-supervised sampling within a sampling survey framework, but we do not do so in the present treatment.

The rest of the paper is organized as follows. We introduce the fixed covariate procedures in Section 2. Specifically, ideal semi-supervised learning and ordinary semi-supervised learning are considered respectively in Sections 2.2 and 2.3, where we analyze the asymptotic properties for both estimators. We further give the ℓ_2-risk upper bounds for the two proposed estimators in Section 2.4. We extend the analysis in Section 3 to nonparametric regression model, where we show the proposed procedure achieves an oracle rate asymptotically. Simulation results are reported in Section 4. Applications to the estimation of Average Treatment Effect is discussed in Section 5.1, and Section 5.2 describes a real data illustration involving estimation of the homeless population in a geographical region. The proofs of the main theorems are given in Section 6 and additional technical results are proved in the Appendix.

2 Procedures

We propose in this section a least squares estimator for the population mean in the semi-supervised inference framework. To better characterize the problem, we begin with a brief introduction of the random design regression model. More details of the model can be found in, e.g., Buja et al. (2014).

2.1 A Random Design Regression Model

Let $(Y, X) \sim P$ represent the population response and predictors. Assume all second moments are finite. Denote $\vec{X} = (1, X^\top)^\top \in \mathbb{R}^{p+1}$ as the predictor with intercept. The following is a linear analysis, even though no corresponding linearity assumption is made about the true distribution P of (X, Y). Some notation and definitions are needed. Let

$$\beta = \arg\min_{\gamma \in \mathbb{R}^{p+1}} \mathbb{E}\left(Y - \vec{X}^\top \gamma\right)^2. \tag{5}$$

Here $\beta \in \mathbb{R}^{p+1}$ are referred to as the *population slopes*, and $\delta = Y - \beta^\top \vec{X}$ is called the *total deviation.* We also denote

$$\tau^2 := \mathbb{E}\delta^2, \quad \mu := \mathbb{E}X \in \mathbb{R}^p, \quad \vec{\mu} := \mathbb{E}\vec{X} = (1, \mu^\top)^\top, \quad \vec{\Xi} = \mathbb{E}\vec{X}\vec{X}^\top. \tag{6}$$

Some basic facts about the regression slope and total deviation are summarized in the following lemma.

Lemma 2.1 *Let $(Y, X) \sim P$ have finite second moment, and let the matrix $\vec{\Xi}$ be non-singular. Then*

$$\beta = \vec{\Xi}^{-1}\left(\mathbb{E}\vec{X}Y\right), \quad \mathbb{E}\delta = 0, \quad \mathbb{E}\delta X = 0, \quad \theta = \vec{\mu}^\top \beta.$$

It should be noted that under our general model, there is no independence assumption between X and δ.

For sample of observations $(Y_k, X_{k1}, X_{k2}, \cdots, X_{kp}) \overset{iid}{\sim} P$, $k = 1, \cdots, n$, let $\vec{X}_i = (1, \vec{X}_i^\top)^\top$ and denote the design matrix $\vec{\mathbf{X}} \in \mathbb{R}^{n \times (p+1)}$ as follows

$$\vec{\mathbf{X}} := \begin{bmatrix} \vec{X}_1^\top \\ \cdots \\ \cdots \\ \vec{X}_n^\top \end{bmatrix} := \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}.$$

In our notation, $\vec{\cdot}$ means that the vector/matrix contains the intercept term; boldface indicates that the symbol is related to a multiple sample if observations. Meanwhile, denote the sample response and deviation as $\mathbf{Y} = (Y_1, \cdots, Y_n)^\top$ and $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)^\top$. Now $\mathbf{Y}$ and $\mathbf{X}$ are connected by a regression model:

$$\mathbf{Y} = \vec{\mathbf{X}}\beta + \boldsymbol{\delta}, \quad \text{and} \quad Y_k = \vec{X}_k^\top \beta + \delta_k, \quad k = 1, \cdots, n. \tag{7}$$

Let $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_{p+1})^\top$ be the usual least squares estimator, i.e.

$$\hat{\beta} = (\vec{\mathbf{X}}^\top \vec{\mathbf{X}})^{-1} \vec{\mathbf{X}}^\top \mathbf{Y}. \tag{8}$$

Then $\hat{\beta}$ provides a straightforward estimator for β. β and $\hat{\beta}$ can be further split into two parts,

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_{(2)} \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_{(2)} \end{bmatrix}, \quad \beta_1, \hat{\beta}_1 \in \mathbb{R}, \quad \beta_{(2)}, \hat{\beta}_{(2)} \in \mathbb{R}^p. \tag{9}$$

$\beta_1, \hat{\beta}_1$ and $\beta_{(2)}, \hat{\beta}_{(2)}$ play different roles in the analysis as we will see later. The ℓ_2 risk of the sample average $\bar{\mathbf{Y}}$ about the population mean $\theta = \mathbb{E}Y$ has the following decomposition.

Proposition 2.1 $\bar{\mathbf{Y}}$ *is an unbiased estimator of* θ *and*

$$n\mathbb{E}(\bar{\mathbf{Y}} - \theta)^2 = n\mathrm{Var}(\bar{\mathbf{Y}}) = \tau^2 + \beta_{(2)}^\top \mathbb{E}\left((X-\mu)(X-\mu)^\top\right)\beta_{(2)}. \tag{10}$$

From (10), we can see that as long as $\beta_{(2)} \neq 0$, i.e., there is a significant linear relationship between Y and X, then the risk of $\bar{\mathbf{Y}}$ will be significantly greater than τ^2.

In the next two subsections, we discuss separately under the ideal semi-supervised setting and the ordinary semi-supervised setting.

2.2 Improved Estimator under the Ideal Semi-supervised Setting

We first consider the ideal setting where there are infinitely many unlabeled samples, or equivalently P_X is known. To improve $\bar{\mathbf{Y}}$, we propose the *least squares estimator*,

$$\hat{\theta}_{\mathrm{LS}} := \vec{\mu}^\top \hat{\beta} = \hat{\beta}_1 + \mu^\top \hat{\beta}_{(2)} = \bar{\mathbf{Y}} - \hat{\beta}_{(2)}^\top(\bar{\mathbf{X}} - \mu), \tag{11}$$

where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_{(2)}^\top)^\top$ is defined in (8).

The following theorem provides the asymptotic distribution of the least squares estimator under the minimal conditions that $[Y, X]$ have finite second moments, $\vec{\Xi} = \mathbb{E}\vec{X}\vec{X}^\top$ be non-singular and $\tau^2 = E\delta^2 > 0$.

Theorem 2.1 (Asymptotic Distribution, fixed p) *Let* $(Y_1, X_1), \cdots, (Y_n, X_n)$ *be i.i.d. copies from* P*, and assume that* $[Y, X]$ *has finite second moments,* $\vec{\Xi}$ *is non-singular and* $\tau^2 > 0$*. Then, under the setting that* P *is fixed and* n *grows to infinity,*

$$\frac{\hat{\theta}_{\mathrm{LS}} - \theta}{\tau/\sqrt{n}} \xrightarrow{d} N(0, 1), \tag{12}$$

and

$$MSE/\tau^2 \xrightarrow{d} 1, \quad \textit{where} \quad MSE := \frac{\sum_{i=1}^n (Y_i - \vec{X}_i^\top \hat{\beta})^2}{n - p - 1}. \tag{13}$$

In the more general setting where $P = P_{n,p}$ varies and $p = p_n$ grows, we need stronger conditions to analyze the asymptotic behavior of $\hat{\theta}_{\mathrm{LS}}$. Suppose $\mathbb{E}(X-\mu)(X-\mu)^\top = \Sigma$, we consider the standardization of X as

$$Z \in \mathbb{R}^p, \quad Z = \Sigma^{-1/2}(X - \mu). \tag{14}$$

Clearly, $\mathbb{E}Z = 0, \mathbb{E}ZZ^\top = I_p$. For this setting we assume that Z, δ satisfy the following moment conditions:

$$\text{for some } \kappa > 0, \quad \frac{\mathbb{E}\delta^{2+2\kappa}}{(\mathbb{E}\delta^2)^{1+\kappa}} \leq M_1; \tag{15}$$

$$\forall v \in \mathbb{R}^p, \quad \mathbb{E}|\langle v, Z\rangle|^{2+\kappa} \leq M_2; \tag{16}$$

$$\frac{\mathbb{E}\left(\|Z\|_2^2\delta^2\right)}{\left(\mathbb{E}\|Z\|_2^2\right)\cdot\left(\mathbb{E}\delta^2\right)} \leq M_3. \tag{17}$$

Theorem 2.2 (Asymptotic result, growing p) *Let $(Y_1, X_1), \cdots, (Y_n, X_n)$ be i.i.d. copies from $P = P_{n,p}$, $p = p_n = o(\sqrt{n})$. Assume that the matrix of the second moments of X exists and is non-singular and the standardized random variable Z given in* (14) *satisfies* (15)*,* (16) *and* (17)*, then the asymptotic behavior results* (12) *and* (13) *still hold.*

Based on Theorems 2.1 and 2.2, we can construct the asymptotic $(1-\alpha)$-level confidence interval for θ as

$$\left[\hat{\theta}_{\mathrm{LS}} - z_{1-\alpha/2}\sqrt{\frac{MSE}{n}} \quad , \quad \hat{\theta}_{\mathrm{LS}} + z_{1-\alpha/2}\sqrt{\frac{MSE}{n}}\right]. \tag{18}$$

Remark 2.1 It is not difficult to see that, under the setting in Theorem 2.2,

$$MSE \xrightarrow{d} \tau^2, \quad \hat{\sigma}_Y^2 \xrightarrow{d} \mathrm{Var}(Y) = \tau^2 + \beta_{(2)}^\top \mathbb{E}((X-\mu)(X-\mu)^\top)\beta_{(2)}.$$

Then the traditional z-interval for the mean of Y,

$$\left[\bar{\mathbf{Y}} - z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_Y^2}{n}} \quad , \quad \bar{\mathbf{Y}} + z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_Y^2}{n}}\right], \tag{19}$$

is asymptotically more accurate than (18), which implies that the proposed least squares estimator is asymptotically more accurate than the sample mean.

2.3 Improved Estimator under the Ordinary Semi-supervised Inference Setting

In the last section, we discussed the estimation of θ based on n full observations $Y_k, X_k, k = 1, \cdots, n$ with infinitely many unlabeled samples $\{X_k, k = n+1, \cdots\}$ (or equivalently with known marginal distribution P_X). However, having P_X known is rare in practice. A more realistic practical setting would assume that distribution P_X is unknown and we only have finitely many i.i.d. samples $(X_{i+1}, X_{i+2}, \cdots, X_{i+m})$ without corresponding Y. This problem relates to the one in previous section since we are able to obtain partial information of P_X from the additional unlabeled samples.

When μ or $\vec{\mu}$ is unknown, we estimate by

$$\hat{\mu} = \frac{1}{n+m}\sum_{k=1}^{n+m} X_k, \quad \hat{\vec{\mu}} = (1, \hat{\mu}^\top)^\top. \tag{20}$$

Recall that $\hat{\beta} = (\hat{\beta}_1, \beta_{(2)}^\top)^\top$ is the ordinary least squares estimator. Now, we propose the *semi-supervised least squares estimator* $\hat{\theta}_{\mathrm{SSLS}}$,

$$\hat{\theta}_{\mathrm{SSLS}} = \hat{\vec{\mu}}^\top \hat{\beta} = \bar{\mathbf{Y}} - \hat{\beta}_{(2)}^\top \left(\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n+m} X_i}{n+m} \right). \tag{21}$$

$\hat{\theta}_{\mathrm{SSLS}}$ has the following properties:

- when $m = \infty$, $\hat{\vec{\mu}} = \vec{\mu}$. Then $\hat{\theta}_{\mathrm{SSLS}}$ exactly equals $\hat{\theta}_{\mathrm{LS}}$ in (11);
- when $m = 0$, $\hat{\theta}_{\mathrm{SSLS}}$ exactly equals $\bar{\mathbf{Y}}$. As there are no additional samples of X so that no extra information for P_X is available, it is natural to use $\bar{\mathbf{Y}}$ to estimate θ.
- In the last term of (21), it is important to use $\frac{\sum_{i=1}^{n+m} X_i}{n+m}$ rather than $\frac{\sum_{i=1}^{m} X_i}{m}$, in spite of the fact that the latter might seem more natural because it is independent of the term $\frac{\sum_{i=1}^{n} X_i}{n}$ that precedes it.

Under the same conditions as Theorems 2.1, 2.2, we can show the following asymptotic results for $\hat{\theta}_{\mathrm{SSLS}}$, which relates to the ordinary semi-supervised setting described in the introduction. The labeled sample size $n \to \infty$, the unlabeled sample size is $m = m_n \geq 0$ and the distribution P is fixed (but unknown) which, in particular, implies that p is a fixed dimension, not dependent on n. Let

$$\nu^2 = \sqrt{\tau^2 + \frac{n}{n+m} \beta_{(2)}^\top \Sigma \beta_{(2)}}, \quad \Sigma = \mathbb{E}(X-\mu)(X-\mu)^\top.$$

Theorem 2.3 (Asymptotic distribution of $\hat{\theta}_{\mathrm{SSLS}}$, fixed p) *Let $(Y_1, X_1), \cdots, (Y_n, X_n)$ be i.i.d. labeled samples from P, $X_{n+1}, \cdots, X_{n+m}$ are m additional unlabeled samples from P_X. Suppose $\vec{\Xi}$ is non-singular and $\tau^2 > 0$. If P is fixed and $n \to \infty$ then*

$$\frac{\sqrt{n}(\hat{\theta}_{\mathrm{SSLS}} - \theta)}{\nu} \xrightarrow{d} N(0,1), \tag{22}$$

and

$$\frac{\hat{\nu}^2}{\nu^2} \xrightarrow{d} 1 \tag{23}$$

where $\hat{\nu}^2 = \frac{m}{m+n} MSE + \frac{n}{m+n} \hat{\sigma}_Y^2$ with $MSE = \frac{1}{n-p-1} \sum_{k=1}^{n} (Y_i - \vec{X}_k^\top \hat{\beta})^2$ and $\hat{\sigma}_Y^2 = \frac{1}{n-1} \sum_{k=1}^{n} (Y_i - \bar{\mathbf{Y}})^2$.

Based on Theorems 2.3 and 2.4, the $(1-\alpha)$-level asymptotic confidence interval for θ can be written as

$$\left[\hat{\theta}_{\mathrm{SSLS}} - z_{1-\alpha/2} \frac{\hat{\nu}}{\sqrt{n}} \quad , \quad \hat{\theta}_{\mathrm{SSLS}} + z_{1-\alpha/2} \frac{\hat{\nu}}{\sqrt{n}} \right]. \tag{24}$$

Since $MSE \le \hat{\sigma}_Y^2$ asymptotically (with equality only when $\beta_{(2)} = 0$), so that when $\beta_{(2)} \neq 0$ the asymptotic CI in (24) is shorter than the traditional sample-mean-based CI (19).

The following statement refers to a setting in which $P = P_n$ and $p = p_n$ may depend on n as $n \to \infty$. Consequently, $\vec{\Xi} = \vec{\Xi}_n$, $\Sigma = \Sigma_n$ and $Z = Z_n$ (defined at (14)) may also depend on n.

Theorem 2.4 (Asymptotic distribution of $\hat{\theta}_{\text{SSLS}}$, growing p) *Let $n \to \infty$, $P = P_n$, and $p = p_n = o(\sqrt{n})$. Suppose $\vec{\Xi}_n$ is non-singular, $\tau_n^2 > 0$ and the standardized random variable Z satisfies* (15), (16) *and* (17). *Then* (22) *and* (23) *hold.*

2.4 ℓ_2 Risk for the Proposed Estimators

In this subsection, we analyze the ℓ_2 risk for both $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$. Since the calculation of the proposed estimators involves the unstable process of inverting the Gram matrix $\vec{\mathbf{X}}^\top \vec{\mathbf{X}}$, for the merely theoretical purpose of obtaining the ℓ_2 risks we again consider the refinement

$$\hat{\theta}_{\text{LS}}^1 := \text{Trun}_{\mathbf{Y}}(\hat{\theta}_{\text{LS}}), \quad \text{and} \quad \hat{\theta}_{\text{SSLS}}^1 := \text{Trun}_{\mathbf{Y}}(\hat{\theta}_{\text{SSLS}}), \tag{25}$$

where

$$\text{Trun}_{\mathbf{Y}}(x) = \begin{cases} (n+1)y_{\max} - ny_{\min}, & \text{if } x > (n+1)y_{\max} - ny_{\min}, \\ x, & \text{if } |x - \frac{y_{\max}+y_{\min}}{2}| \le (n+\frac{1}{2})(y_{\max} - y_{\min}), \\ (n+1)y_{\min} - ny_{\max}, & \text{if } x < (n+1)y_{\min} - ny_{\max}, \end{cases} \tag{26}$$

$y_{\max} = \max_{1\le k\le n} Y_k, y_{\min} = \min_{1\le k\le n} Y_k$. We emphasize that this refinement is mainly for theoretical reasons and is often not necessary in practice.

The regularization assumptions we need for analyzing the ℓ_2 risk are formally stated as below.

1. *(Moment conditions on δ)* There exist $M_1 > 0$ such that

$$\mathbb{E}\delta^4 = \mathbb{E}\delta_n^4 \le M_4; \tag{27}$$

2. *(sub-Gaussian condition)* Suppose $Z = Z_n$ is the standardization of $X = X_n$

$$Z_n \in \mathbb{R}^p, \quad Z_n = \Sigma_n^{-1/2}(X_n - \mu_n), \quad \Sigma_n = \mathbb{E}(X_n - \mu_n)(X_n - \mu_n)^\top,$$

which satisfies

$$\forall u \in \{u \in \mathbb{R}^{p+1} : \|u\|_2 = 1\}, \quad \left\|u^\top Z_n\right\|_{\psi_2} \le M_5. \tag{28}$$

Here $\|\cdot\|_{\psi_2}$ is defined as $\|x\|_{\psi_2} = \sup_{q\ge1} q^{-1/2}(\mathbb{E}|x|^q)^{1/q}$ for any random variable x.

2' *(Bounded condition)* The standardization Z_n satisfies

$$\|Z_n\|_\infty \le M_5, \quad \text{almost surely.} \tag{29}$$

We also note $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^\top$, $\Sigma_{\delta 1} = \mathbb{E}(X-\mu)\delta(X-\mu)^\top$, $\Sigma_{\delta 2} = \mathbb{E}(X-\mu)\delta^2(X-\mu)^\top$. Under the regularization assumptions above, we provide the ℓ_2 risks for $\hat{\theta}^1_{\mathrm{LS}}$ and $\hat{\theta}^1_{\mathrm{SSLS}}$ respectively in the next two theorems.

Theorem 2.5 (ℓ_2 Risk of $\hat{\theta}^1_{\mathrm{LS}}$) *Let $(Y_1, X_1), \cdots, (Y_n, X_n)$ be i.i.d. copies from P_n. Assume that Assumptions 1+2* (27)(28) *or 1+2'* (27)(29) *hold, $p = p_n = o(\sqrt{n})$. Recall $\tau^2 = \tau_n^2 = \mathbb{E}(Y - \vec{X}\beta)^2$ depends on n. Then we have the following estimate for the risk of $\hat{\theta}^1_{\mathrm{LS}}$,*

$$n\mathbb{E}\left(\hat{\theta}^1_{\mathrm{LS}} - \theta\right)^2 = \tau_n^2 + s_n, \tag{30}$$

where

$$s_n = \frac{p^2}{n}A_{n,p} + \frac{p^2}{n^{5/4}}B_{n,p}, \quad \max(|A_{n,p}|, |B_{n,p}|) \le C \tag{31}$$

for a constant C that depends on M_0, M_1 and M_2. The formula for $A_{n,p}$ is

$$\begin{aligned} A_{n,p} = \frac{1}{p^2}\Bigg(& [\mathrm{tr}(\Sigma^{-1}\Sigma_{\delta 1})]^2 + 3\|\Sigma^{-1}\Sigma_{\delta 1}\|_F^2 - \mathrm{tr}(\Sigma^{-1}\Sigma_{\delta 2}) \\ & + 2\mathbb{E}\left(\delta^2(X-\mu)^\top\right) \cdot \mathbb{E}\left(\Sigma^{-1}(X-\mu)(X-\mu)^\top\Sigma^{-1}(X-\mu)\right) + 2p\tau^2\Bigg). \end{aligned} \tag{32}$$

Theorem 2.6 (ℓ_2 risk of $\hat{\theta}^1_{\mathrm{SSLS}}$) *Let $(Y_1, X_1), \cdots, (Y_n, X_n)$ be i.i.d. labeled samples from P, $X_{n+1}, \cdots, X_{n+m}$ are additional m unlabeled samples from P_X. If Assumptions 1+2 or 1+2' in* (27)-(29) *hold, $p = o(\sqrt{n})$, we have the following estimate of the risk for $\hat{\theta}^1_{\mathrm{SSLS}}$,*

$$n\mathbb{E}\left(\hat{\theta}^1_{\mathrm{SSLS}} - \theta\right)^2 = \tau_n^2 + \frac{n}{n+m}\beta_{(2),n}^\top \Sigma_n \beta_{(2),n} + s_{n,m} \tag{33}$$

where

$$|s_{n,m}| \le \frac{Cp^2}{n}. \tag{34}$$

for constant C only depends on M_0, M_1 and M_2 in Assumptions (27)-(29).

Comparing Proposition 2.1 and Theorems 2.5 and Theorem 2.6, we can see as long as

$$\beta_{(2),n}^\top \Sigma_n \beta_{(2),n} > 0,$$

i.e., $\mathbb{E}(Y|X)$ has non-zero correlation with X, $\hat{\theta}^1_{\mathrm{LS}}$ and $\hat{\theta}^1_{\mathrm{SSLS}}$ outperform $\bar{\mathbf{Y}}$ asymptotically in ℓ_2-risk.

Remark 2.2 Comparing Theorems 2.5, 2.6 and Proposition 2.1, we can see the risk of $\hat{\theta}_{\text{SSLS}}$ is approximately a linear combination of $\bar{\mathbf{Y}}$ and $\hat{\theta}_{\text{LS}}$ with weight based on m and n,

$$\mathbb{E}\left(\hat{\theta}^1_{\text{SSLS}}-\theta\right)^2\approx\frac{n}{n+m}\mathbb{E}\left(\bar{\mathbf{Y}}-\theta\right)^2+\frac{m}{m+n}\mathbb{E}\left(\hat{\theta}^1_{\text{LS}}-\theta\right)^2$$

Remark 2.3 (Gaussian Design) Theorems 2.5 and 2.6 only provides upper bound of the ℓ_2 risks since because only moment conditions on the distribution of Y, X are assumed. In fact, under Gaussian design of Y, X, we can obtain an exact expression for the ℓ_2-risk of both $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$. It is noteworthy that the truncation refinement is not necessary for both estimators under Gaussian design. All results are non-asymptotic.

Proposition 2.2 *Assume $X \sim N_p(\mu,\Sigma)$ and $Y|X \sim N_p(X\beta,\tau^2 I)$, where Σ is non-singular. If $\{Y_k,X_k\}_{k=1}^n$ are n i.i.d. copies, then*

$$n\mathbb{E}\left(\hat{\theta}_{\text{LS}}-\theta\right)^2=\tau^2+\frac{p\tau^2}{(n-p-2)}. \tag{35}$$

If we further have m additional unlabeled samples $\{X_k\}_{k=n+1}^{n+m}$, then we also have

$$n\mathbb{E}\left(\hat{\theta}_{\text{SSLS}}-\theta\right)^2=\tau^2+\frac{m}{n+m}\frac{p\tau^2}{n-p-2}+\frac{n}{n+m}\beta_{(2)}^\top\mathbb{E}\left((X-\mu)(X-\mu)^\top\right)\beta_{(2)}. \tag{36}$$

The result in Proposition 2.2 matches with the general expression of (30) and (32) as $\frac{p\tau^2}{(n-p-2)}=\frac{p\tau^2}{n}+O\left(\frac{p^2}{n^2}\right)$ if $p=o(\sqrt{n})$. By comparing (35), (36), we can also see

$$n\mathbb{E}\left(\hat{\theta}_{\text{SSLS}}-\theta\right)^2=\frac{n}{n+m}n\mathbb{E}(\bar{\mathbf{Y}}-\theta)^2+\frac{m}{n+m}n\mathbb{E}(\hat{\theta}_{\text{LS}}-\theta)^2.$$

3 Further Improvements – Oracle Optimality

In the previous sections, we proposed and analyzed $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$ under the semi-supervised learning settings. These estimators are based on linear regression and best linear approximation of Y by X. We consider further improvement in this section. Before we illustrate how the improved estimator works, it is helpful to take a look at the oracle risk for estimating the mean $\theta=\mathbb{E}Y$, which can serve as a benchmark for the performance of the improved estimator.

3.1 Oracle Estimator and Risk

Define $\xi(X)=\mathbb{E}_P(Y|X)$ as the response surface and suppose

$$\xi(x)=\xi_0(x)+c$$

for some unknown constant c. Given samples $\{(Y_k, X_k)\}_{k=1}^n$, our goal is to estimate $\mathbb{E}Y = \theta$. Now assume an oracle has knowledge of $\xi_0(x)$, but not of $\theta = \mathbb{E}(Y)$, c, nor the distribution of $Y - \xi_0(X)$. In this case, the model can be written as

$$\begin{aligned} Y_k - \xi_0(X_k) =& c + \varepsilon_k, \quad k = 1, \cdots, n, \quad \text{where} \quad \mathbb{E}\varepsilon_k = 0; \\ \theta =& \mathbb{E}\xi_0(X) + c. \end{aligned} \tag{37}$$

Under the ideal semi-supervised setting, P_X, ξ_0 and $\mathbb{E}\xi_0(X)$ are known. To estimate θ, the natural idea is to by the following estimator

$$\hat{\theta}^* = \bar{\mathbf{Y}} - \bar{\xi}_0 + \mathbb{E}\xi_0(X) = \frac{1}{n}\sum_{k=1}^n (Y_k - \xi_0(X_k)) + \mathbb{E}\xi_0(X). \tag{38}$$

Clearly $\hat{\theta}^*$ is an unbiased estimator of θ, while

$$\begin{aligned} n\mathbb{E}\left(\hat{\theta}^* - \theta\right)^2 =& n\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - \xi_0(X_i))\right) = \mathrm{Var}\left(Y_i - \xi(X_i)\right) \\ =& \mathbb{E}_X\left(\mathbb{E}_Y\left(Y - \xi(X)|X\right)^2\right) := \sigma^2. \end{aligned} \tag{39}$$

This defines the oracle risk for population mean estimation under the ideal semi-supervised setting as $\sigma^2 = \mathbb{E}_X\left(\mathbb{E}_Y(Y - \mathbb{E}(Y|X))^2\right)$.

For the ordinary semi-supervised setting, where P_X is unknown but m additional unlabeled samples $\{X_k\}_{k=n+1}^{n+m}$ are available, we propose the semi-supervised oracle estimator as

$$\hat{\theta}^*_{\mathrm{ss}} = \bar{\mathbf{Y}} - \frac{1}{n}\sum_{k=1}^n \xi_0(X_k) + \frac{1}{n+m}\sum_{k=1}^{n+m} \xi_0(X_k).$$

Then one can calculate that

$$n\mathbb{E}\left(\hat{\theta}^*_{ss} - \theta\right)^2 = \sigma^2 + \frac{n}{n+m}\mathrm{Var}_{P_X}(\xi(X)). \tag{40}$$

The detailed calculation of (40) is provided in the Appendix.

The preceding motivation for σ^2 and $\sigma^2 + \frac{n}{n+m}\mathrm{Var}_{P_X}(\xi(X))$ as the oracle risks are partly heuristic, based on the arguments in (38) and (39). But it corresponds to a formal minimax statement, as follows.

Proposition 3.1 (Oracle Lower Bound) *Let* $\sigma^2 > 0$,

$$\mathcal{P}_{\xi_0(\cdot),\sigma^2} = \left\{P : \xi_0(x) = \mathbb{E}(Y|X = x) - c, \sigma^2 = \mathbb{E}_X\left(\mathbb{E}_Y(Y - \mathbb{E}(Y|X))^2\right)\right\}.$$

Then based on observations $\{Y_i, X_i\}_{i=1}^n$ *and known marginal distribution* P_X,

$$\inf_{\tilde{\theta}} \sup_{P \in \mathcal{P}_{\xi_0,\sigma^2}} \left[\mathbb{E}_P\left(n\left(\tilde{\theta} - \theta\right)^2\right)\right] = \sigma^2. \tag{41}$$

Let $\sigma^2, \sigma_\xi^2 > 0$, $\xi_0(X)$ be a linear function of X,

$$\mathcal{P}^{\text{ss}}_{\xi_0,\sigma_\xi^2,\sigma^2} = \left\{P : \xi_0(x) = \mathbb{E}(Y|X=x) - c, \sigma_\xi^2 = \text{Var}(\xi(X)), \sigma^2 = \mathbb{E}_X\left(\mathbb{E}_Y(Y - \mathbb{E}(Y|X))^2\right)\right\},$$

based on observations $\{Y_i, X_i\}_{i=1}^n$ and $\{X_i\}_{i=n+1}^{n+m}$,

$$\inf_{\tilde{\theta}} \sup_{P \in \mathcal{P}^{\text{ss}}_{\xi_0,\sigma_\xi^2,\sigma^2}} \left[\mathbb{E}_P\left(n\left(\tilde{\theta} - \theta\right)^2\right)\right] = \sigma^2 + \frac{n}{n+m}\sigma_\xi^2. \tag{42}$$

3.2 Improved Procedure

In order to approach oracle optimality we propose to augment the set of covariates $X_1, \ldots, X_p$ with additional covariates $g_1(X), \ldots, g_q(X)$. (Of course these additional covariates need to be chosen without knowledge of ξ_0. We will discuss their choice later in this section.) In all there are now $p^\bullet = p + q$ covariates, say

$$X^\bullet = (X_1^\bullet, \ldots, X_p^\bullet, X_{p+1}^\bullet, \ldots, X_{p+q}^\bullet) = (X_1, \ldots, X_p, g_1(X), \ldots, g_q(X)).$$

For both ideal and ordinary semi-supervision we propose to let $q = q_n$ as $n \to \infty$, and to use the estimator $\hat{\theta}^\bullet_{\text{LS}}$ and $\hat{\theta}^\bullet_{\text{SSLS}}$. For merely theoretical purpose of ℓ_2 risks we consider the refinement again

$$\hat{\theta}^{\bullet 1}_{\text{LS}} = \text{Trun}_{\mathbf{Y}}(\hat{\theta}^\bullet_{\text{LS}}) \quad \text{and} \quad \hat{\theta}^{\bullet 1}_{\text{SSLS}} = \text{Trun}_{\mathbf{Y}}(\hat{\theta}^\bullet_{\text{SSLS}}),$$

where $\text{Trun}_{\mathbf{Y}}(\cdot)$ is defined as (26). Apply previous theorems for asymptotic distributions and moments. For convenience of statement and proof we assume that the support of X is compact, $\xi(X)$ is bounded and Y is sub-Gaussian. These assumptions can each be somewhat relaxed at the cost of additional technical assumptions and complications. Here is a formal statement of the result.

Theorem 3.1 *Assume the support of X is compact, $\xi(X) = \mathbb{E}(Y|X)$ is bounded, and Y is sub-Gaussian. Consider asymptotics as $n \to \infty$ for the case of both ideal and ordinary semi-supervision. Assume also that either (i) $\xi(X)$ is continuous or (ii) that P_X is absolutely continuous with respect to Lebesgue measure on $\{X\}$. Let $\{g_k(x) : k = 1, \ldots\}$ be a bounded basis for the continuous functions on $\{X\}$ in case (i) and be a bounded basis for the ordinary ℓ_2 Hilbert space on $\{X\}$ in case (ii). There exists a sequence of q_n such that $\lim_{n\to\infty} q_n \to \infty$, and*

- *the estimator $\hat{\theta}^{\bullet 1}_{LS}$ for the problem with observations $\{Y_i, X^\bullet_{p+q_n} : i = 1, \ldots, n\}$ asymptotically achieves the ideal oracle risk, i.e.*

$$\lim_{n\to\infty} n\mathbb{E}\left(\hat{\theta}^{\bullet 1}_{\text{LS}} - \theta\right)^2 = \sigma^2. \tag{43}$$

- *Now we suppose* $\lim_{n\to\infty}\frac{n}{n+m_n} = \rho$ *for some fixed value* $0 \leq \rho \leq 1$. *Applying the estimator* $\hat{\theta}^{\bullet}_{\text{SSLS}}$ *for the problem with observations* $\{Y_i, X^{\bullet}_{p+q_n} : i = 1, \ldots, n\}$ *and* $\{X^{\bullet}_i\}_{i=n+1}^{n+m_n}$. *Then*

$$\lim_{n\to\infty} n\mathbb{E}\left(\hat{\theta}^{\bullet 1}_{\text{SSLS}} - \theta\right)^2 = \sigma^2 + \rho\text{Var}_{P_X}(\xi(X)). \tag{44}$$

Finally, $\hat{\theta}^{\bullet}_{\text{LS}}$ *and* $\hat{\theta}^{\bullet}_{\text{SSLS}}$ *are asymptotically unbiased and normal with the corresponding variances.*

(38) and (44) show that the proposed estimators asymptotically achieve the oracle values in (41) and (42).

4 Simulation Results

In this section, we investigate the numerical performance of the proposed estimators in various settings in terms of estimation errors and coverage probability as well as length of confidence intervals. All the simulations are repeated for 1000 times.

We analyze the linear least squares estimators $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$ proposed in Section 2 in the following three settings.

1. (Gaussian X and quadratic ξ) We generate the design and parameters as follows, $\mu \sim N(0, I_p)$, $\Sigma \in \mathbb{R}^{p\times p}$, $\Sigma_{ij} = I\{i = j\} + \frac{1}{2p}I\{i \neq j\}$, $\beta \sim N(0, I_{p+1})$. Then we draw i.i.d. samples $\mathbf{Y}, \mathbf{X}$ as

$$X_k \sim N(\mu, \Sigma), \quad Y_k = \xi(X_k) + \varepsilon_k,$$

where

$$\xi(X_k) = (\|X_k\|_2^2 - p) + \vec{\mathbf{X}}^\top\beta, \quad \varepsilon_k \sim N\left(0, 2\|X_k\|_2^2/p\right).$$

It is easy to calculate that $\theta = \mathbb{E}Y = \beta_1$ in this setting.

2. (Heavy tailed X and Y) We randomly generate

$$\{X_{ki}\}_{1\leq k\leq n, 1\leq i\leq p} \overset{iid}{\sim} P_3, \quad Y_k = \sum_{i=1}^{p}(\sin(X_{ki}) + X_{ki}) + .5\cdot\varepsilon_k, \quad \varepsilon_k \overset{iid}{\sim} P_3.$$

where P_3 has density $f_{P_3}(x) = \frac{1}{1+|x|^3}$, $-\infty < x < \infty$. Here, the distribution P_3 has no third or higher moments. In this case, $\mu = \mathbb{E}X = 0$, $\theta = \mathbb{E}Y = 0$.

3. (Poisson X and Y) Then we also consider a setting where

$$\{X_{ki}\}_{1\leq k\leq n, 1\leq i\leq p} \overset{iid}{\sim} \text{Poisson}(10), \quad Y_k|X_k \overset{iid}{\sim} \text{Poisson}(10X_{k1}).$$

In this case, $\mu = \mathbb{E}X = (10, \ldots, 10)^\top \in \mathbb{R}^p$, $\theta = \mathbb{E}\mathbf{Y} = 100$.

We compare the average ℓ_2-loss of $\bar{\mathbf{Y}}$, $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$ for various choices of n, p and m. The results are summarized in Table 1. An interesting aspect is even when p grows faster than $n^{1/2}$, $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$ are still preferable estimators to $\bar{\mathbf{Y}}$. It is also noteworthy that although our theoretical analysis for the ℓ_2-risk focused on the refined estimators $\hat{\theta}^1_{\text{LS}}$ and $\hat{\theta}^1_{\text{SSLS}}$ with bounded or sub-Gaussian designs, the refinement and assumptions are for technical needs, which might not be necessary in practice as we can see from this example.

We also compute the 95%-confidence interval for each setting above and list the average length and coverage probability in Table 2. It can be seen that under the condition $p = o(n^{1/2})$, the proposed confidence intervals based on $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$ are valid and shorter on average than the traditional z-confidence interval centered at $\bar{\mathbf{Y}}$.

5 Applications

In this section, we apply the proposed procedures to the average treatment effect estimation and a real data example on homeless population.

5.1 Application to Average Treatment Effect Estimation

We first discuss an application of the proposed least squares estimator to Average Treatment Effect (ATE) estimation. Suppose Y_T and Y_C are the responses for the treatment population and control population respectively, then ATE is then defined as

$$d = \mathbb{E}Y_T - \mathbb{E}Y_C. \tag{45}$$

Under Neyman's paradigm (Splawa-Neyman et al., 1990; Rubin, 1990), a total number of (n_t+n_c) subjects are randomly assigned to the treatment group and control group. Suppose $Y_{t,1}, \cdots, Y_{t,n_t}$ are the responses under treatment, while $Y_{t,1}, \cdots, Y_{t,n_c}$ are the responses of the control group. The straight forward idea for estimating ATE is the sample average treatment effect (SATE), which simply takes the difference of average effects between the two groups. In addition, the covariates associated with the responses are often available and helpful to improve the estimation of ATE.

In the estimation of ATE, we follow the model setting of Pitkin et al. (2013). Suppose n_t, n_c people are from treatment group and control group respectively, where their response and predictor satisfies

$$(Y_t, X_t) \overset{iid}{\sim} P^t, \quad (Y_c, X_c) \overset{iid}{\sim} P^c.$$

Here due to the randomization setting, it is reasonable to assume P^t and P^c share the same marginal distribution of X: $P^t_X = P^c_X = P_X$. There are also m additional samples possibly coming from drop-outs or any other subjects that also represent the population P_X. In

(p, n)	$(\bar{\mathbf{Y}} - \theta)^2$	$(\hat{\theta}_{\text{SSLS}} - \theta)^2$			$(\hat{\theta}_{\text{LS}} - \theta)^2$
		$m = 100$	$m = 1000$	$m = 10000$	
Setting 1: Gaussian X and Quadratic ξ					
(1, 100)	0.304	0.184	0.075	0.063	0.056
(10, 100)	2.73	1.529	0.518	0.313	0.296
(50, 100)	13.397	7.961	3.967	2.988	2.868
(10, 500)	0.526	0.464	0.211	0.067	0.045
(50, 500)	2.668	2.278	1.089	0.373	0.273
(200, 500)	10.743	9.135	4.615	2.345	1.949
Setting 2: Heavy tailed X and Y					
(1, 100)	0.732	0.410	0.244	0.196	0.188
(10, 100)	7.791	5.428	2.505	1.959	1.831
(50, 100)	107.363	47.036	17.754	14.201	13.435
(10, 500)	2.575	2.097	0.988	0.354	0.261
(50, 500)	12.569	10.481	5.619	2.342	1.780
(200, 500)	43.997	36.123	30.856	13.175	9.642
Setting 3: Poisson X and Y					
(1, 100)	97.912	50.510	10.168	2.036	1.015
(10, 100)	98.337	50.772	10.535	2.085	1.061
(50, 100)	94.475	52.166	10.951	3.146	2.100
(10, 500)	20.062	16.765	6.890	1.104	0.186
(50, 500)	19.915	15.793	6.541	1.165	0.225
(200, 500)	20.933	17.639	7.159	1.300	0.333

Table 1: Average squared loss of sample mean estimator $\bar{\mathbf{Y}}$, the least squares estimator $\hat{\theta}_{\text{LS}}$ and the semi-supervised least squares estimators $\hat{\theta}_{\text{SSLS}}$ under different values of (p, n) and various settings.

(p, n)	via $\bar{\mathbf{Y}}$	via $\hat{\theta}_{\text{SSLS}}$			via $\hat{\theta}_{\text{LS}}$
		$m = 100$	$m = 1000$	$m = 10000$	
		Setting 1: Gaussian X and Quadratic ξ			
(1, 100)	1.902(0.055)	1.521(0.046)	1.074(0.049)	0.940(0.061)	0.921(0.064)
(5, 100)	4.430(0.058)	3.301(0.070)	1.911(0.055)	1.467(0.059)	1.400(0.069)
(10, 100)	6.318(0.048)	4.678(0.058)	2.655 (0.063)	2.010(0.076)	1.913(0.084)
(1, 500)	0.845(0.041)	0.793(0.042)	0.608(0.041)	0.451(0.042)	0.413(0.046)
(10, 500)	2.818(0.045)	2.596(0.041)	1.768(0.048)	1.023(0.051)	0.832(0.064)
(25, 500)	4.558(0.051)	4.194(0.039)	2.837(0.054)	1.606(0.058)	1.288(0.078)
		Setting 2: Heavy tailed X and Y			
(1, 100)	3.349(0.039)	2.069(0.059)	1.596(0.061)	1.446(0.044)	1.420(0.038)
(5, 100)	7.332(0.050)	4.885(0.082)	3.384(0.067)	2.920(0.063)	2.847(0.048)
(10, 100)	11.292(0.044)	7.436(0.079)	5.073(0.078)	4.343(0.057)	4.225(0.044)
(1, 500)	1.573(0.046)	1.205(0.055)	0.970(0.077)	0.773(0.066)	0.723(0.058)
(10, 500)	5.947(0.043)	4.427(0.061)	3.217(0.084)	2.180(0.069)	1.904(0.047)
(25, 500)	8.582(0.040)	7.079(0.055)	5.197(0.072)	3.617(0.069)	3.229(0.047)
		Setting 3: Poisson X and Y			
(1, 100)	39.164(0.063)	27.831(0.061)	12.386(0.056)	5.506(0.047)	3.895(0.075)
(5, 100)	39.396(0.053)	28.003(0.043)	12.485(0.067)	5.600(0.062)	4.004(0.070)
(10, 100)	39.143(0.065)	27.832(0.054)	12.443(0.064)	5.655(0.058)	4.105(0.063)
(1, 500)	17.548(0.054)	16.035(0.054)	10.232(0.050)	4.195(0.043)	1.753(0.054)
(10, 500)	17.621(0.053)	16.102(0.062)	10.276(0.048)	4.216(0.050)	1.768(0.043)
(25, 500)	17.632(0.053)	16.113(0.052)	10.285(0.051)	4.229(0.045)	1.795(0.061)

Table 2: Average length and coverage probability (in the parenthesis) 95%-CI based on $\bar{\mathbf{Y}}$, $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$ under different values of (p, n) and various settings.

summary, the available samples include

$$\{(Y_{t,k}, X_{t,k})\}_{k=1}^{n_t}, \quad \{(Y_{c,k}, X_{c,k})\}_{k=1}^{n_c}, \quad \{(X_{a,k})\}_{k=1}^{m}. \tag{46}$$

We again introduce the population slope for both treatment and control group to measure the relationship between Y_t, X_t and Y_c, X_c respectively

$$\beta_t = \arg\min_{\gamma \in \mathbb{R}^{p+1}} \mathbb{E}\left(Y_t - \vec{X}_t^\top \gamma\right)^2, \quad \beta_c = \arg\min_{\gamma \in \mathbb{R}^{p+1}} \mathbb{E}\left(Y_c - \vec{X}_c^\top \gamma\right)^2. \tag{47}$$

Based on Lemma 2.1, β_t, β_c has the following close form when P_t, P_c have non-degenerate second moment:

$$\beta_t = \left(\mathbb{E}\vec{X}_t\vec{X}_t^\top\right)^{-1}\left(\mathbb{E}\vec{X}_t Y_t\right), \quad \beta_c = \left(\mathbb{E}\vec{X}_c\vec{X}_c^\top\right)^{-1}\left(\mathbb{E}\vec{X}_c Y_c\right). \tag{48}$$

Our target, the population ATE, is defined as $d = \mathbb{E}Y_c - \mathbb{E}Y_t$. We propose the corresponding semi-supervised least squares estimator

$$\hat{d}_{\mathrm{SSLS}} = \hat{\vec{\mu}}^\top \left(\hat{\beta}_t - \hat{\beta}_c\right). \tag{49}$$

Here $\hat{\beta}_t, \hat{\beta}_c \in \mathbb{R}^{p+1}$ are the least squares estimators for treatment and control group respectively; $\hat{\vec{\mu}}$ is the mean of all available predictors,

$$\hat{\beta}_t = \left(\vec{\mathbf{X}}_t^\top \vec{\mathbf{X}}_t\right)^{-1} \vec{\mathbf{X}}_t^\top \mathbf{Y}_t, \quad \hat{\beta}_c = \left(\vec{\mathbf{X}}_c^\top \vec{\mathbf{X}}_c\right)^{-1} \vec{\mathbf{X}}_c^\top \mathbf{Y}_c, \tag{50}$$

$$\text{where} \quad \hat{\vec{\mu}} = \begin{pmatrix} 1 \\ \hat{\mu} \end{pmatrix}, \quad \hat{\mu} = \frac{1}{n_t + n_c + m}\left(\sum_{k=1}^{n_t} X_{t,k} + \sum_{k=1}^{n_c} X_{c,k} + \sum_{k=1}^{m} X_{a,k}\right). \tag{51}$$

Based on the analysis we have in the previous section, the proposed $\hat{d}_{\mathrm{SSLS}}$ has the following asymptotic distribution with a fixed p, P^t and P^c.

Theorem 5.1 (Asymptotic behavior of $\hat{d}_{\mathrm{SSLS}}$) *Suppose P^t, P^c are fixed distribution with finite and non-degenerate second moments, then we have the following asymptotic distribution if the sample size n_t, t_c grow to infinity:*

$$\frac{\hat{d}_{\mathrm{SSLS}} - d}{V} \xrightarrow{d} N(0,1), \quad \frac{\hat{V}^2}{V^2} \xrightarrow{d} 1. \tag{52}$$

Here

$$V^2 = \frac{\tau_t^2}{n_t} + \frac{\tau_c^2}{n_c} + \frac{1}{n_t + n_c + m}(\beta_{t,(2)} - \beta_{c,(2)})^\top \mathbb{E}(X-\mu)(X-\mu)^\top(\beta_{t,(2)} - \beta_{c,(2)}), \tag{53}$$

$$\hat{V}^2 = \frac{MSE_t}{n_t} + \frac{MSE_c}{n_c} + \frac{1}{n_t + n_c + m}(\hat{\beta}_t - \hat{\beta}_c)^\top \hat{\mathbf{\Sigma}}_X (\hat{\beta}_t - \hat{\beta}_c), \tag{54}$$

$$MSE_t = \frac{1}{n_t - p - 1}\sum_{k=1}^{n_t}(Y_{t,k} - \vec{X}_{t,k}^\top \hat{\beta}_t)^2, \quad MSE_c = \frac{1}{n_c - p - 1}\sum_{k=1}^{n_c}(Y_{c,k} - \vec{X}_{c,k}^\top \hat{\beta}_c)^2,$$

$$\hat{\mathbf{\Sigma}}_X = \frac{1}{n_t + n_c + m}\Big(\sum_{k=1}^{n_t}(X_{t,k} - \hat{\mu})(X_{t,k} - \hat{\mu})^\top + \sum_{k=1}^{n_c}(X_k - \hat{\mu})(X_k - \hat{\mu})^\top + \sum_{k=1}^{m}(X_k - \hat{\mu})(X_k - \hat{\mu})^\top\Big).$$

Remark 5.1 Similarly to the procedure in Proposition 2.1, we can calculate that for the sample average treatment effect, i.e.,

$$\hat{d} = \sum_{k=1}^{n_t}\frac{Y_{t,k}}{n_t} - \sum_{k=1}^{n_c}\frac{Y_{t,c}}{n_c},$$

$$\mathrm{Var}(\hat{d}) = \frac{\tau_t^2 + \beta_{t,(2)}^\top \mathbb{E}(X-\mu)(X-\mu)^\top \beta_{t,(2)}}{n_t} + \frac{\tau_c^2 + \beta_{c,(2)}^\top \mathbb{E}(X-\mu)(X-\mu)^\top \beta_{t,(2)}}{n_c}.$$

We can check that asymptotically $V^2 \leq \mathrm{Var}(\hat{d})$, which also shows the merit of the proposed semi-supervised least squares estimator.

Remark 5.2 The asymptotic behavior of $\hat{d}_{\mathrm{SSLS}}$ and the ℓ_2 risk for a refined $\hat{d}_{\mathrm{SSLS}}$ for growing p can be elaborated similarly to the previous sections.

5.2 Real Data Example: Estimating Homeless in Los Angeles County

We now consider an application to estimate the number of homeless people in Los Angeles County. Homelessness has been a significant public issue for the United States since nearly a century ago (Rossi, 1991). A natural question for the demographers is to estimate the number of homeless in a certain region. Estimating the number of homeless in metropolitan area is an important but difficult task due to the following reasons. In a typical design of U.S. Census, demographers visit people through their place of residence. In this case, most of the homeless will not be contacted (Rossi, 1991) through this process. Visiting homeless shelter or homeless service center may collect some information of the homeless, but a large number of homeless still cannot be found since they may use the service anonymously or simply not use the service.

The Los Angeles County includes land of 2000 square miles, total population of 10 million and 2,054 census tracts. In 2004-2005, the Los Angeles Homeless Services Authority (LAHSA) conducted the study for the homeless population. Due to the huge cost to perform street visit for all census tracts, the demographers perform a stratified sampling on part of them. First, 244 tracts that are believed to have large amount of homeless are pre-selected and visited. Next for the rest of the tracts, 265 of them are randomly selected

via $\hat{\theta}_{\text{SSLS}}$	95%-CI	via $\bar{\mathbf{Y}}$	95%-CI
53824	[47120, 60529]	52527	[45485, 59570]

Table 3: Estimated total number of homeless in Los Angeles County

and visited. This design leaves 1,545 tracts unvisited. Besides the number of homeless, some predictors are available for all 2,054 tracts. In our analysis, 7 of them are included, `Perc.Industrial`, `Perc.Residential`, `Perc.Vacant`, `Perc.Commercial`, `Perc.OwnerOcc`, `Perc.Minority`, `MedianHouseholdIncome`. These predictors have been used and are known to have high correlation with the response Kriegler and Berk (2010).

Suppose T_{total} is the total number of homeless in Los Angeles, T_{pre} is the number of homeless in 244 pre-selected tracts, θ_{ran} is average number of homeless per tract in all 1,810 non-pre-selected tracts. Clearly,

$$T_{\text{total}} = T_{\text{pre}} + 1810 \cdot \theta_{\text{ran}}. \tag{55}$$

The proposed semi-supervised inference framework fit into the 1,810 samples with 265 labeled and 1,545 unlabeled samples. We can apply the proposed semi-supervised least squares estimator $\hat{\theta}^1_{\text{SSLS}}$ to estimate θ_{ran} and use (55) to calculate the estimate and 95% confidence interval for T_{total}. In contrast, the estimate via sample-mean estimator was also calculated. The results are shown in Table 3. It is easy to see that the estimate via $\hat{\theta}^1_{\text{SSLS}}$ is slightly larger than the one via $\bar{\mathbf{Y}}$.

To further investigate and diagnose, we calculated the least squares estimator $\hat{\beta}$, the average predictor values across all 1,810 non-pre-selected tracts $\bar{\mathbf{X}}_{\text{full}}$ and the average predictor values across 265 randomly selected tracts $\bar{\mathbf{X}}$. These values are listed in Table 5.2.

We can see from Table 5.2 that due to insufficiency of sampling, there is difference between $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}_{\text{full}}$, especially for the predictor `Perc.OwnerOcc`. When there is association between these prectors and reponse, it is more reasonable to adjust for this discrepancy from taking the mean. Recall the proposed estimator

$$\hat{\theta}_{\text{SSLS}} = \bar{\mathbf{Y}} + \hat{\beta}_{(2)}^{\top} \left(\bar{\mathbf{X}}_{\text{full}} - \bar{\mathbf{X}}\right), \quad \text{where} \quad \bar{\mathbf{X}}_{\text{full}} = \frac{1}{n+m}\sum_{k=1}^{n+m} X_k, \bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

The difference between two estimates exactly originated from the adjustment term $\hat{\beta}_{(2)}^{\top}(\bar{\mathbf{X}}_{\text{full}} - \bar{\mathbf{X}})$, which has been justified in both theoretical analysis and simulation studies in the previous sections.

	$\hat{\beta}$	$\bar{\mathbf{X}}_{\text{full}}-\bar{\mathbf{X}}$	$\bar{\mathbf{X}}$	$\bar{\mathbf{X}}_{\text{full}}$
Intercept	21.963			
Perc.Industrial	0.027	0.143	61.293	61.149
Perc.Residential	-0.087	-0.075	4.066	4.141
Perc.Vacant	1.404	-0.075	4.066	4.141
Perc.Commercial	0.338	-0.542	15.130	15.672
Perc.OwnerOcc	-0.233	2.489	54.039	51.550
Perc.Minority	0.058	0.833	50.890	50.057
MedianInc (in \$K)	0.074	0.638	48.805	48.167
	Adjustment: $\hat{\beta}_{(2)}^\top(\bar{\mathbf{X}}_{\text{full}}-\bar{\mathbf{X}})$ = -0.768			

Table 4: Diagnostic Table for Los Angeles Data Example

6 Proofs of The Main Results

We prove the main results in this section. The proofs of other technical results are provided in the Appendix.

6.1 Proofs for the Properties of the Random Design Regression Model

Proof of Lemma 2.1. Since

$$\begin{aligned}
&\mathbb{E}\left(Y-\vec{X}^\top\beta\right)^2 = \mathbb{E}Y^2 + \beta^\top\left(\mathbb{E}\vec{X}\vec{X}^\top\right)\beta - 2\beta^\top\mathbb{E}\left(\vec{X}Y\right)\\
=&\mathbb{E}Y^2 + \left(\beta-(\mathbb{E}\vec{X}\vec{X}^\top)^{-1}\mathbb{E}(\vec{X}Y)\right)^\top\left(\mathbb{E}\vec{X}\vec{X}^\top\right)\left(\beta-(\mathbb{E}\vec{X}\vec{X}^\top)^{-1}\mathbb{E}(\vec{X}Y)\right)\\
&-\mathbb{E}(\vec{X}Y)^\top\left(\mathbb{E}\vec{X}\vec{X}^\top\right)^{-1}\mathbb{E}(\vec{X}Y),
\end{aligned}$$

we know $\beta = \arg\min_\gamma \mathbb{E}(Y-\vec{X}^\top\gamma)^2 = (\mathbb{E}\vec{X}\vec{X}^\top)^{-1}\mathbb{E}(\vec{X}Y)$. Besides,

$$\mathbb{E}(\vec{X}\delta) = \mathbb{E}\vec{X}Y - \mathbb{E}\vec{X}\vec{X}^\top\beta = \mathbb{E}\vec{X}Y - \mathbb{E}\vec{X}\vec{X}^\top\cdot\left(\mathbb{E}\vec{X}\vec{X}^\top\right)^{-1}\mathbb{E}(\vec{X}Y) = 0.$$

Then $\mathbb{E}\delta = 0, \mathbb{E}X\delta = 0$ have been proved since $\vec{X} = (1, X^\top)^\top$. Finally,

$$\begin{aligned}
\vec{\mu}^\top\beta =&\mathbb{E}\vec{X}^\top\left(\mathbb{E}\vec{X}\vec{X}^\top\right)^{-1}\mathbb{E}\vec{X}Y = (1,\mu^\top)\cdot\begin{bmatrix}1 & \mu^\top\\ \mu & \mathrm{Cov}(X)+\mu\mu^\top\end{bmatrix}^{-1}\cdot\begin{pmatrix}EY\\ EXY\end{pmatrix}\\
=&(1,\overbrace{0,\ldots,0}^{p})\begin{pmatrix}EY\\ EXY\end{pmatrix} = EY = \theta,
\end{aligned}$$

which has finished the proof of this lemma. □

Proof of Proposition 2.1. First, $\bar{\mathbf{Y}}$ is the sample mean, which is clearly an unbiased estimator for the population mean θ. In addition, since $\{Y_i\}_{i=1}^n$'s are i.i.d. samples, it can be calculated that

$$\begin{aligned} n\mathrm{Var}(\bar{\mathbf{Y}}) =& \mathrm{Var}(Y_i) = \mathrm{Var}(\delta_i) + \mathrm{Var}(\vec{X}_i\beta) + 2\mathrm{Cov}(\delta_i, \vec{X}_i\beta) \\ \overset{\text{Lemma 2.1}}{=}& \tau^2 + \beta_{(2)}^\top \mathbb{E}(X-\mu)(X-\mu)^\top \beta_{(2)}. \end{aligned} \tag{56}$$

□

6.2 Proofs for Ideal Semi-supervised Inference Estimator $\hat{\theta}_{\mathrm{LS}}$

Proof of Theorem 2.1.

We first show that $\hat{\theta}_{\mathrm{LS}}$ is invariant under simultaneous affine translation on both $\mathbf{X}$ and μ. Specifically, suppose $X_k = U \cdot Z_k + \alpha$, $(k = 1, \cdots, n)$ for any fixed invertible matrix $U \in \mathbb{R}^{p\times p}$ and vector $\alpha \in \mathbb{R}^p$. Then one has

$$\vec{X}_k = \begin{bmatrix} 1 & 0 \\ \alpha & U \end{bmatrix} \vec{Z}_k, \quad \vec{\mathbf{X}} = \vec{\mathbf{Z}} \begin{bmatrix} 1 & \alpha^\top \\ 0 & U^\top \end{bmatrix},$$

$$\begin{aligned} \hat{\theta}_{\mathrm{LS}} =& \vec{\mu}^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \mathbf{Y} \\ =& (1, \mu^\top) \left(\begin{bmatrix} 1 & 0 \\ \alpha & U \end{bmatrix} \vec{\mathbf{Z}}^\top \vec{\mathbf{Z}} \begin{bmatrix} 1 & \alpha^\top \\ 0 & U^\top \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 \\ \alpha & U \end{bmatrix} \vec{\mathbf{Z}}^\top \mathbf{Y} \\ =& (1, \mu^\top) \begin{bmatrix} 1 & \alpha^\top \\ 0 & U^\top \end{bmatrix}^{-1} \left(\vec{\mathbf{Z}}^\top \vec{\mathbf{Z}}\right)^{-1} \vec{\mathbf{Z}}^\top \mathbf{Y} \\ =& (1, (U^{-1}(\mu - \alpha))^\top) \left(\vec{\mathbf{Z}}^\top \vec{\mathbf{Z}}\right)^{-1} \vec{\mathbf{Z}}^\top \mathbf{Y}. \end{aligned}$$

Since $\mathbb{E}Z_k = U^{-1}(\mu - \alpha)$, we know $\hat{\theta}_{\mathrm{LS}}$ is invariant under simultaneous affine translation on $\mathbf{X}$ and μ.

Based on the affine transformation invariant property, we only need to consider the situation when $\mathbb{E}X = \mu = 0$, $\mathrm{Cov}(X) = I_p$, where I_p is the p-by-p identity matrix. Next we discuss the asymptotic behavior for $\hat{\theta}_{\mathrm{LS}}$. For simplicity, we note $1_n = (\overbrace{1, \cdots, 1}^{n})^\top$, $\mathbb{P}_{\vec{\mathbf{X}}} = \vec{\mathbf{X}}(\vec{\mathbf{X}}^\top \vec{\mathbf{X}})^{-1} \vec{\mathbf{X}}^\top \in \mathbb{R}^{(p+1)\times(p+1)}$ as the projection matrix onto the column space of $\vec{\mathbf{X}}$.

$\bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n} X_k$. Clearly, 1_n lies in the column space of $\vec{\mathbf{X}}$, which means $\mathbb{P}_{\vec{\mathbf{X}}} 1_n = 1_n$. Then,

$$\begin{aligned}
\hat{\theta}_{\mathrm{LS}} - \theta =& \vec{\mu}^\top \hat{\beta} - \theta = \vec{\mu}^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \mathbf{Y} - \theta \\
=& \vec{\mu}^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \left(\vec{\mathbf{X}}\beta + \boldsymbol{\delta}\right) - \theta = \vec{\mu}^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \\
=& \frac{1_n^\top}{n} \vec{\mathbf{X}} \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} - \frac{1_n^\top}{n} (\vec{\mathbf{X}} - 1_n \vec{\mu}^\top) \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \\
=& \frac{1_n^\top \mathbb{P}_{\vec{\mathbf{X}}}}{n} \boldsymbol{\delta} - (0, \frac{1_n^\top}{n} \mathbf{X}) \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \\
=& \frac{1_n^\top}{n} \boldsymbol{\delta} - \left(0, \frac{1_n}{n} \mathbf{X}\right) \left(\frac{1}{n} \vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \left(\frac{1}{n} \vec{\mathbf{X}}^\top \boldsymbol{\delta}\right) \\
=& \bar{\boldsymbol{\delta}} - \left(0, \bar{\mathbf{X}}^\top\right) \left(\frac{1}{n} \vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \left(\frac{1}{n} \vec{\mathbf{X}}^\top \boldsymbol{\delta}\right),
\end{aligned} \tag{57}$$

$$\begin{aligned}
\frac{n-p-1}{n} MSE =& \frac{1}{n} \|\mathbf{Y} - \vec{\mathbf{X}}\hat{\beta}\|_2^2 = \frac{1}{n} \left\| \boldsymbol{\delta} + \vec{\mathbf{X}}\beta - \vec{\mathbf{X}} \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top (\mathbf{X}\beta + \boldsymbol{\delta}) \right\|_2^2 \\
=& \frac{1}{n} \left\| \boldsymbol{\delta} - \vec{\mathbf{X}}(\vec{\mathbf{X}}^\top \vec{\mathbf{X}})^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \right\|_2^2 = \frac{1}{n} \left(\boldsymbol{\delta}^\top \boldsymbol{\delta} - \boldsymbol{\delta}^\top \vec{\mathbf{X}} (\vec{\mathbf{X}}^\top \vec{\mathbf{X}})^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \right) \\
=& \left(\frac{1}{n} \boldsymbol{\delta}^\top \boldsymbol{\delta} - \left(\frac{1}{n} \vec{\mathbf{X}}^\top \boldsymbol{\delta}\right)^\top \left(\frac{1}{n} \vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \left(\frac{1}{n} \vec{\mathbf{X}}^\top \boldsymbol{\delta}\right) \right).
\end{aligned} \tag{58}$$

Since P is fixed and has finite second moment, by law of large number and central limit theorem, one can show as $n \to \infty$, $p = o(n^{1/2})$,

$$\sqrt{n}\bar{\boldsymbol{\delta}} \xrightarrow{d} N(0, \tau^2),$$

$$\frac{1}{n} \boldsymbol{\delta}^\top \boldsymbol{\delta} = \frac{1}{n} \sum_{k=1}^{n} \delta_k^2 \xrightarrow{d} \mathbb{E}\delta^2 = \tau^2,$$

$$\left\| \frac{1_n^\top \mathbf{X}}{n} \right\|_2^2 \xrightarrow{d} \|\mathbb{E}X\|_2^2 = 0, \quad \left\| \frac{\vec{\mathbf{X}}^\top \boldsymbol{\delta}}{n} \right\|_2^2 \xrightarrow{d} \|\mathbb{E}X\delta\|_2^2 = 0,$$

$$\frac{1}{n} \vec{\mathbf{X}}^\top \vec{\mathbf{X}} \xrightarrow{d} \mathbb{E}\vec{X}\vec{X}^\top = \begin{bmatrix} 1 & 0 \\ 0 & \mathrm{Cov}(X) \end{bmatrix}.$$

Since $\mathrm{Cov}(X)$ invertible, we know

$$\left(\frac{1}{n} \vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \xrightarrow{d} \begin{bmatrix} 1 & 0 \\ 0 & \mathrm{Cov}(X)^{-1} \end{bmatrix}.$$

Based on the asymptotic distributions above and (57), (58), we know

$$\sqrt{n}\left(\hat{\theta}_{\mathrm{LS}} - \theta\right) \to N(0, \tau^2)$$

$$\frac{n-p-1}{n}MSE \to \tau^2,$$

in the case that P_X fixed and $n \to \infty$. □

Proof of Theorem 2.2. First, based on the proof of Theorem 2.1, the affine transformation on $\mathbf{X}$ would not affect the property of $\hat{\theta}_{\mathrm{LS}}$. Without loss of generality, we assume that $\mathbb{E}X = 0$, $\mathrm{Var}(X) = I$. In other words, $\mathbf{Z} = \mathbf{X}$. Next, based on formulas (57) and (58), we have

$$\sqrt{n}(\hat{\theta}_{\mathrm{LS}} - \theta)/\tau = \frac{\sqrt{n}\bar{\boldsymbol{\delta}}}{\tau} - \frac{\sqrt{n}}{\tau}(0, \bar{\mathbf{X}}^\top)\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\right),$$

$$\left|\frac{\sqrt{n}}{\tau}(0, \bar{\mathbf{X}}^\top)\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\right)\right| \le \left\|\frac{1_n\mathbf{X}^\top}{n^{3/4}}\right\|_2 \cdot \lambda_{\min}^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right) \cdot \left\|\frac{\vec{\mathbf{X}}^\top\boldsymbol{\delta}}{n^{3/4}\tau}\right\|_2,$$

then we only need to prove the following asymptotic properties in order to finish the proof of Theorem 2.2:

$$\frac{\sqrt{n}\bar{\boldsymbol{\delta}}}{\tau} \xrightarrow{d} N(0,1), \tag{59}$$

$$\left\|\frac{1_n\mathbf{X}}{n^{3/4}}\right\|_2 \xrightarrow{d} 0, \quad \left\|\frac{\vec{\mathbf{X}}^\top\boldsymbol{\delta}}{n^{3/4}}\right\|_2 /\tau \xrightarrow{d} 0, \tag{60}$$

For some uniform $t_1 > t_2 > 0$,

$$P\left(t_1 \ge \lambda_{\max}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right) \ge \lambda_{\min}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right) \ge t_2\right) \to 1. \tag{61}$$

Here $\lambda_{\max}, \lambda_{\min}(\cdot)$ represent the largest and least eigenvalues of the given matrix. Next we will show (59), (60) and (61) separately.

- Based on the assumption of the theorem, $\frac{\delta_1}{\tau}, \cdots, \frac{\delta_n}{\tau}$ are i.i.d. samples with mean 0, variance 1 and bounded $(2+2\varepsilon)$-th moment, (59) holds by Lyapunov's central limit theorem.

- Since $X_1, \cdots, X_k$ are i.i.d. samples with mean 0 and covariance I_p, we can calculate that

  $$\mathbb{E}\left\|\frac{1_n\mathbf{X}^\top}{n^{3/4}}\right\|_2^2 = \frac{1}{n^{3/2}} \cdot n\mathbb{E}\|X\|_2^2 = \frac{p}{n^{1/2}} \to 0, \quad \text{as } n \to 0.$$

  Since $X_1\delta_1, \cdots, X_n\delta_n$ are i.i.d. samples with mean 0 and satisfying (17), we have

  $$\mathbb{E}\left\|\frac{\vec{\mathbf{X}}^\top\boldsymbol{\delta}}{n^{3/4}}\right\|_2^2 = \frac{1}{n^{3/2}} \cdot n\mathbb{E}\|\vec{X}\delta\|_2^2 \le \frac{M_3}{n^{1/2}}\mathbb{E}\|X\|_2^2 \cdot \mathbb{E}\delta^2 = \frac{p}{n^{1/2}}M_3\tau^2$$

  Thus, $\mathbb{E}\|\frac{\vec{\mathbf{X}}^\top\boldsymbol{\delta}}{n^{3/4}}\|_2^2/\tau^2 \to 0$ as $n \to \infty$. Thus, we have (60).

- For (61), since $\mathbb{E}X = 0, \mathrm{Cov}(X) = I_p$ and Assumption (16) holds, (61) is directly implied by Theorem 2 in Yaskov (2014). □

6.3 Proofs for Ordinary Semi-supervised Inference Estimator $\hat{\theta}_{\text{SSLS}}$

Proof of Theorems 2.3 and 2.4. We start with the proof of (23). From the proof of Theorems 2.1 and 2.2, we have proved that

$$\frac{MSE}{\tau^2} \to 1.$$

By the basic property of sample covariance and Proposition 2.1, we also have

$$\frac{n\hat{\sigma}_Y^2}{\mathrm{Var}(Y)} \xrightarrow{d} 1, \quad \mathrm{Var}(Y) = \tau^2 + \mathbb{E}\beta_{(2)}^\top \Sigma \beta_{(2)}.$$

Therefore, under either the settings of Theorems 2.3 or 2.4,

$$\frac{\frac{m}{m+n}MSE + \frac{n}{m+n}\hat{\sigma}_Y^2}{\tau^2 + \frac{n}{n+m}\mathrm{Var}(\beta_{(2)}^\top X)} = \frac{\frac{m}{m+n}MSE + \frac{n}{m+n}\hat{\sigma}_Y^2}{\frac{m}{m+n}\tau^2 + \frac{n}{m+n}\mathrm{Var}(Y)} \xrightarrow{d} 1, \quad \text{as } n \to \infty, \tag{62}$$

which proves (23).

The proof of (22) is more complicated. In the rest of proof, again we use C as constants does not depends on n or m, whose exact value may vary in different scenarios. Again, since $\hat{\theta}_{\text{SSLS}}$ is affine transformation invariant, without loss of generality we can assume that $\mathbb{E}X = 0$, $\mathbb{E}XX^\top = I_p$. Thus, $Z = X$. Similarly as (57), the following decomposition for $\hat{\theta}_{\text{SSLS}} - \theta$ holds,

$$\begin{aligned}
\hat{\theta}_{\text{SSLS}} - \theta =& \hat{\vec{\mu}}^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \mathbf{Y} - \theta = \hat{\vec{\mu}} \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top (\vec{\mathbf{X}}\beta + \boldsymbol{\delta}) - \theta \\
=& (\hat{\vec{\mu}}^\top \beta - \theta) + \hat{\vec{\mu}}(\vec{\mathbf{X}}^\top \vec{\mathbf{X}})^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \\
=& (\hat{\vec{\mu}} - \vec{\mu})^\top \beta + \left(\frac{1_n^\top}{n}\vec{\mathbf{X}}\left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta}\right) + \left(\hat{\vec{\mu}} - \frac{1_n}{n}\vec{\mathbf{X}}\right)^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \\
=& (\hat{\vec{\mu}} - \vec{\mu})^\top \beta + \left(\frac{1_n^\top \mathbb{P}_{\vec{\mathbf{X}}}}{n}\right) \boldsymbol{\delta} + \left(\hat{\vec{\mu}}^\top - \frac{1_n^\top}{n}\vec{\mathbf{X}}\right) \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \\
=& (\hat{\vec{\mu}} - \vec{\mu})^\top \beta + \bar{\boldsymbol{\delta}} - \left(0, \bar{\mathbf{X}} - \hat{\mu}\right)^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta}.
\end{aligned} \tag{63}$$

In order to prove these two theorems, we only need to show the following two asymptotic equalities:

$$\frac{\left(\hat{\vec{\mu}} - \vec{\mu}\right)^\top \beta + \bar{\boldsymbol{\delta}}}{\sqrt{\left(\frac{\tau^2}{n} + \frac{n}{n(n+m)}\beta_{(2)}^\top \mathbb{E}X_c X_c^\top \beta_{(2)}\right)}} \to N(0,1), \tag{64}$$

$$\frac{\left(0, \hat{\mu} - \bar{\mathbf{X}}\right)^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta}}{\sqrt{\tau^2/n}} \xrightarrow{d} 0. \tag{65}$$

We show them separately below under both settings that p is fixed (Theorem 2.3) and p grows (Theorem 2.4). For convenience, we denote $T = \beta_{(2)}^\top \mathbb{E}X_c X_c^\top \beta_{(2)}$, $b_j = X_{j,c}^\top \beta_{(2)}, j = 1, \cdots, m+n$. Clearly $\mathbb{E}b_j^2 = T$.

1. We first show (64). The left hand side of (64) can be further written as

$$
\begin{aligned}
&\left(\hat{\vec{\mu}} - \vec{\mu}\right)^\top \beta + \bar{\boldsymbol{\delta}} = (1-1)\beta_1 + (\hat{\mu} - \mu)\beta_{(2)} + \bar{\boldsymbol{\delta}} \\
=&\sum_{i=1}^{n} \left(-\frac{m}{n(n+m)}(X_i - \mu)^\top \beta_{(2)} + \frac{1}{n}\delta_i \right) + \sum_{i=n+1}^{n+m} \frac{1}{n+m}(X_i - \mu)^\top \beta_{(2)} \\
:=&\sum_{j=1}^{n} A_j^{(n)} + \sum_{j=n+1}^{n+m} B_j^{(n)} := S_n
\end{aligned}
$$

Here $A_j^{(n)} = \frac{m}{n(n+m)} b_j + \frac{1}{n}\delta_j$, $B_j^{(n)} = \frac{1}{n+m} b_j$. It is easy to calculate that $\mathbb{E}A_i^{(n)} = \mathbb{E}B_j^{(n)} = 0$, $\mathbb{E}A_i^{(n)2} = \frac{\tau^2}{n^2} + \frac{m^2}{n^2(n+m)^2}T$, $\mathbb{E}B_j^{(n)2} = \frac{1}{(n+m)^2}T$,

$$
s_n^2 := \mathbb{E}S_n^2 = \sum_{i=1}^{n} \mathbb{E}A_i^{(n)2} + \sum_{j=n+1}^{n+m} \mathbb{E}B_j^{(n)2} = \frac{\tau^2}{n} + \frac{mT}{n(n+m)}. \tag{66}
$$

Next we analyze the asymptotic distribution for S_n separately under both settings when p is fixed and p is growing. Specifically, we use Lindeberg-Feller central limit theorem for the fixed p case under second moment condition and Lyapunov central limit theorem for growing p under $(2+\kappa)$-th moment condition.

- Under the setting of Theorem 2.3, i.e., when p and the distribution $P(Y, X_1, \cdots, X_p)$ is fixed, we check the following Lindeberg-Feller condition:

$$
\begin{aligned}
&\forall \varepsilon > 0, \\
&\lim_{n\to\infty} \frac{1}{s_n^2} \left[\sum_{j=1}^{n} \mathbb{E}\left(A_j^{(n)2} I\{A_j^{(n)2} \geq \varepsilon s_n^2\} \right) + \sum_{j=n+1}^{m+n} \mathbb{E}\left(B_j^{(n)2} I\{B_j^{(n)2} \geq \varepsilon s_n^2\} \right) \right] = 0.
\end{aligned} \tag{67}
$$

  Here $I\{\cdot\}$ is the indicator random variable for given event. Note that, for any $x_1, x_2 \in \mathbb{R}$,

$$
\begin{aligned}
&(x_1 + x_2)^2 I\{(x_1+x_2)^2 \geq s_n^2\} \leq 4\max(x_1^2, x_2^2) I\{4\max(x_1^2, x_2^2) \geq s_n\} \\
\leq& 4x_1^2 I\{x_1^2 \geq s_n^2/4\} + 4x_2^2 I\{x_2^2 \geq s_n^2/4\},
\end{aligned} \tag{68}
$$

  we have

$$
\begin{aligned}
&\mathbb{E}\left(A_j^{(n)2} I\{|A_j^{(n)}|^2 \geq \varepsilon s_n^2\} \right) \leq \mathbb{E}\left(A_j^{(n)2} I\{A_j^{(n)2} \geq \varepsilon\tau^2/n\} \right) \\
\leq& \mathbb{E}\left(4\frac{m^2}{n^2(n+m)^2} b_j^2 I\left\{ \frac{m^2}{n^2(n+m)^2} b_j^2 \geq \frac{\varepsilon\tau^2}{4n} \right\} \right) + \mathbb{E}\left(4\frac{1}{n^2}\delta_j^2 I\left\{ \frac{1}{n^2}\delta_j^2 \geq \frac{\varepsilon\tau^2}{4n} \right\} \right) \\
\leq& \frac{4}{n^2}\left(\mathbb{E}\left(b_j^2 I\left\{ b_j^2 \geq \frac{n\varepsilon\tau^2}{4} \right\} \right) + \mathbb{E}\left(\delta_j^2 I\left\{ \delta_j^2 \geq \frac{n\varepsilon\tau^2}{4} \right\} \right) \right).
\end{aligned}
$$

Similarly one can calculate that

$$\mathbb{E}\left(B_j^{(n)2} I\{|B_j^{(n)}|^2 \geq \varepsilon s_n^2\}\right) \leq \frac{1}{n(n+m)}\mathbb{E}\left(b_j^2 I\left\{b_j^2 \geq \frac{\varepsilon n \tau^2}{4}\right\}\right).$$

Therefore,

$$\begin{aligned}
&\frac{1}{s_n^2}\left[\sum_{j=1}^{n}\mathbb{E}\left(A_j^{(n)2} I\{A_j^{(n)2} \geq \varepsilon s_n^2\}\right) + \sum_{j=m+1}^{m+n}\mathbb{E}\left(B_j^{(n)2} I\{B_j^{(n)2} \geq \varepsilon s_n^2\}\right)\right] \\
\leq & \frac{n}{\tau^2}\cdot\left[\frac{5}{n}\mathbb{E}\left(b_j^2 I\left\{b_j^2 \geq \frac{n\varepsilon\tau^2}{4}\right\}\right) + \frac{4}{n}\mathbb{E}\left(\delta_j^2 I\left\{\delta_j^2 \geq \frac{n\varepsilon\tau^2}{4}\right\}\right)\right] \\
\leq & \frac{5}{\tau^2}\mathbb{E}\left(b^2 I\left\{b^2 \geq n\varepsilon\tau^2/4\right\}\right) + \frac{4}{\tau^2}\mathbb{E}\left(\delta^2 I\left\{\delta^2 \geq n\varepsilon\tau^2/4\right\}\right) \to 0.
\end{aligned}$$

By Lindeberg-Feller CLT, we know $S_n/s_n \to N(0,1)$, which implies (64).

- Under the setting of Theorem 2.4, i.e., when the distribution P is not fixed and p is growing, the proof as we also have $(2+2\kappa)$-moment conditions. In this case, Lyapunov's condition for central limit theorem will be used as the main tool. One can check that

$$\begin{aligned}
\mathbb{E}|A_i|^{2+2\kappa} \leq & C\left(\mathbb{E}\left(\frac{m}{n(n+m)}b_i\right)^{2+2\kappa} + \mathbb{E}\left(\frac{\delta_i}{n}\right)^{2+2\kappa}\right) \\
\overset{(15)}{\leq} & C\left(\frac{m}{n(n+m)}\right)^{2+2\kappa} T^{1+\kappa} + C\left(\frac{\tau}{n}\right)^{2+2\kappa},
\end{aligned}$$

$$\mathbb{E}|B_j|^{2+2\kappa} = \frac{1}{(n+m)^{2+2\kappa}}\mathbb{E}\left((X_i-\mu)^\top \beta_{(2)}\right)^{2+2\kappa} \overset{(16)}{\leq} \frac{1}{(n+m)^{2+2\kappa}}T^{1+\kappa}.$$

Thus,

$$\begin{aligned}
&\sum_{i=1}^{n}\mathbb{E}|A_i^{(n)}|^{2+2\kappa} + \sum_{j=n+1}^{n+m}\mathbb{E}|B_j^{(n)}|^{2+2\kappa} \\
\leq & C\left\{\left(\frac{m}{n(n+m)}\right)^{2+2\kappa} n + \left(\frac{1}{n+m}\right)^{2+2\kappa} m\right\} T^{1+\kappa} + C\frac{\tau^{2+2\kappa}}{n^{1+2\kappa}} \\
\leq & C\frac{m(m^{1+2\kappa}+n^{1+2\kappa})}{n^{1+2\kappa}(n+m)^{2+2\kappa}}T^{1+\kappa} + C\frac{\tau^{2+2\kappa}}{n^{1+2\kappa}} \\
\leq & C\left(\frac{m}{n^{1+2\kappa}(n+m)}T^{1+\kappa} + \frac{\tau^{2+2\kappa}}{n^{1+2\kappa}}\right)
\end{aligned} \tag{69}$$

On the other hand,

$$s_n^{2+2\kappa} = \left(\frac{\tau^2}{n} + \frac{mT}{n(n+m)}\right)^{1+\kappa} \geq \frac{\tau^{2+2\kappa}}{n^{1+\kappa}} + \frac{m^{1+\kappa}T^{1+\kappa}}{n^{1+\kappa}(n+m)^{1+\kappa}}. \tag{70}$$

Since as $n, m \to \infty$,

$$\frac{\frac{m}{n^{1+2\kappa}(n+m)}T^{1+\kappa}}{\frac{m^{1+\kappa}T^{1+\kappa}}{n^{1+\kappa}(n+m)^{1+\kappa}}} = \left(\frac{n+m}{nm}\right)^{\kappa} \to 0, \quad \frac{\tau^{2+2\kappa}/(n^{1+2\kappa})}{\tau^{2+2\kappa}/((n^{1+\kappa})} = \frac{1}{n^{\kappa}} \to 0,$$

combining (69) and (70), we have

$$\lim_{n\to\infty} \frac{1}{s_n^{2+2\kappa}} \left(\sum_{i=1}^{n} \mathbb{E}|A_i^{(n)}|^{2+2\kappa} + \sum_{j=n+1}^{n+m} \mathbb{E}|B_j^{(n)}|^{2+2\kappa} \right).$$

By Lyapunov's central limit theorem, we know

$$\left(\sum_{i=1}^{n} A_i^{(n)} + \sum_{j=n+1}^{n+m} B_j^{(n)} \right) \Bigg/ \sqrt{\frac{m}{n(n+m)}T + \frac{\tau^2}{n}} \to N(0,1),$$

which implies (64).

2. Next, we show (65) under both settings of fixed p and growing p. We can calculate that

$$\frac{\left|(0, \hat{\mu} - \bar{\mathbf{X}})^{\top} \left(\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^{\top}\boldsymbol{\delta}\right|}{\sqrt{\tau^2/n}} \leq \frac{\|\hat{\mu} - \bar{\mathbf{X}}\|_2 \cdot \lambda_{\min}^{-1}(\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}}) \cdot \|\vec{\mathbf{X}}^{\top}\boldsymbol{\delta}\|_2}{\sqrt{\tau^2/n}}$$

$$\leq n^{1/4}\|\hat{\mu} - \bar{\mathbf{X}}\|_2 \cdot \lambda_{\min}^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}}\right) \cdot \left\|\frac{\vec{\mathbf{X}}^{\top}\boldsymbol{\delta}}{n^{3/4}\tau}\right\|_2.$$

- We first consider the simpler case where P is fixed, i.e., the setting in Theorem 2.3. The proof is similar to the one of Theorem 2.1. Note that $\mathbb{E}X_i = 0$, $\mathbb{E}\vec{X}\vec{X}^{\top} = I_{p+1}$, $\mathbb{E}\vec{X}\delta = 0$, thus by law of large number,

$$\frac{1}{n}\vec{\mathbf{X}}^{\top}\boldsymbol{\delta} \xrightarrow{d} 0, \quad \frac{1}{n}\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}} \xrightarrow{d} I_p,$$
$$\hat{\vec{\mu}} = \frac{1}{n+m}\vec{X}_k \to (1, 0, \cdots, 0)^{\top}, \quad \frac{\vec{\mathbf{X}}1_n}{n} = \frac{1}{n}\sum_{k=1}^{n}\vec{X}_k \to (1, 0, \cdots, 0)^{\top}. \tag{71}$$

  These facts together yields (65).

- Now we move to the case that p grows, i.e., the setting in Theorem 2.4. Similarly as the proof of Theorem 2.2, we have

$$\left\|\frac{\sum_{i=1}^{n} X_i}{n^{3/4}}\right\|_2 \xrightarrow{d} 0, \quad \left\|\frac{\sum_{i=1}^{n} \vec{X}_i\delta_i}{n^{3/4}}\right\|_2 /\tau \xrightarrow{d} 0, \quad \left\|\frac{n^{1/4}\sum_{i=1}^{n+m} X_i}{(m+n)}\right\|_2 \xrightarrow{d} 0,$$

$$\exists t_1 \geq t_2 > 0, \quad \text{such that}$$

$$P\left(t_1 \geq \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_i\vec{X}_i^{\top}\right) \geq \lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{X}_i\vec{X}_i^{\top}\right) \geq t_2\right) \to 1.$$

  Similarly these imply (65).

To sum up, we have finished the proof of Theorems 2.3 and 2.4. □

6.4 Proofs for the analysis of ℓ_2-risk

Proof of Theorems 2.5. The idea of the proof for Theorem 2.5 is to first introduce a "good event" Q such that $P(Q^c)$ is exponentially small; then prove that $\mathbb{E}\left[n\left(\hat{\theta}_{\text{LS}}-\theta\right)^2 1_Q\right]$ has upper bound as (30) and (31). For convenience, for any subset $\Omega \subseteq \{1,\dots,n\}$, we introduce the following notations

$$\vec{\boldsymbol{\Xi}} = \frac{1}{n}\sum_{k=1}^{n} \vec{Z}_k \vec{Z}_k^\top, \quad \vec{\boldsymbol{\Xi}}_{-\Omega} = \frac{1}{n}\sum_{k=1, k\notin\Omega}^{n} \vec{Z}_k Z_k^\top. \tag{72}$$

Also, we note poly(n,p) for some polynomial of n and p. We also introduce the following lemmas. The proofs are postponed to the Appendix.

Lemma 6.1 *Suppose* $\vec{\mathbf{Z}} = (\vec{Z}_1,\cdots,\vec{Z}_n)^\top$ *satisfies Assumption 2* (28) *or Assumption 2'* (29).

- **(Theorem 5.39 in Vershynin (2012b))** *We have the following concentration inequality,*

$$P\left(\left\|\frac{1}{n}\sum_{k=1}^{n}\vec{Z}_k\vec{Z}_k^\top - \mathbb{E}\vec{Z}_k\vec{Z}_k^\top\right\| > C\sqrt{\frac{p}{n}} + t\right) \leq 2\exp(-cnt^2). \tag{73}$$

  Here C, c are constants only depending on M_5 in Assumption (28) *or M_6 in Assumption* (29).

- *For all $q \geq 2$, the following moment condition holds for some constant C_q that only depends on q under either Assumption 2* (28) *or Assumption 2'* (29),

$$\mathbb{E}\left\|\sum_{k=1}^{n} Z_k\right\|_2^q \leq C_q\,(pn)^{q/2}. \tag{74}$$

- *The following moment condition holds for $\sum_{k=1}^n \vec{Z}_k\delta_k$ and $2 \leq q < 4$:*

$$\mathbb{E}\left\|\sum_{k=1}^{n}\vec{Z}_k\delta_k\right\|_2^q \leq C_q(pn)^{q/2} \tag{75}$$

  under either Assumption 1+2 ((27), (28)) *or 1+2'* ((27), (29)).

Lemma 6.2 *Suppose A, B are two squared matrices, $A, A+B$ are both invertible. Then for all $q \geq 0$, one has the following expansion for $(A+B)^{-1}$,*

$$(A+B)^{-1} = \sum_{k=0}^{q-1}\left(-A^{-1}B\right)^k A^{-1} + \left(-A^{-1}B\right)^q (A+B)^{-1}. \tag{76}$$

For the proof of Theorem 2.5, we first consider the probability that $\hat{\theta}_{\mathrm{LS}} \neq \hat{\theta}^1_{\mathrm{LS}}$. Note $\bar{\mu} = \frac{\max(\mathbf{Y})+\min(\mathbf{Y})}{2}$, then we have

$$
\begin{aligned}
&P\left(\hat{\theta}_{\mathrm{LS}} \neq \hat{\theta}^1_{\mathrm{LS}}\right) = P\left(|\hat{\theta}_{\mathrm{LS}} - \bar{\mu}| > (n+\frac{1}{2})(\max(\mathbf{Y}) - \min(\mathbf{Y}))\right)\\
\leq &P\left(\left\|\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\right\| \cdot \sqrt{\left\|\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right\|} \cdot \frac{\max(\mathbf{Y}) - \min(\mathbf{Y})}{2} > (n+\frac{1}{2})(\max(\mathbf{Y}) - \min(\mathbf{Y}))\right)\\
\overset{(73)}{\leq} &\exp(-cn), \quad \text{for large } n.
\end{aligned}
$$

Set the event Q as

$$
Q = \left\{\hat{\theta}_{\mathrm{LS}} = \hat{\theta}^1_{\mathrm{LS}}, \quad \max_{1\leq i,j,k\leq n}\left\{\left\|\vec{\boldsymbol{\Xi}} - I_{p+1}\right\|, \left\|\vec{\boldsymbol{\Xi}}_{-\{i,j,k\}} - I_{p+1}\right\|\right\} \leq \frac{C_1}{n^{1/4}}\right\} \tag{77}
$$

for some large constant $C_1 > 0$. Based on Lemma 6.1 and the fact that $\sqrt{p/n} = o(n^{-1/4})$, we have

$$
\begin{aligned}
P\left(Q^c\right) \leq &P\left(\left\|\vec{\Sigma} - I_{p+1}\right\| > C_1 n^{-1/4}\right) + \sum_{i,j,k} P\left(\left\|\vec{\Sigma}_{-\{i,j,k\}} - I_{p+1}\right\| > C_1 n^{-1/4}\right)\\
&+ P\left(\hat{\theta}_{\mathrm{LS}} \neq \hat{\theta}^1_{\mathrm{LS}}\right)\\
\leq &Cn^3 \cdot \exp(-cn^{1/2}) \quad \text{for large } n.
\end{aligned} \tag{78}
$$

Recall the composition of $\hat{\theta}_{\mathrm{LS}} - \theta$ in (57), thus,

$$
\begin{aligned}
&\mathbb{E}\left(1_Q(\hat{\theta}_{\mathrm{LS}} - \theta)^2\right)\\
=&\mathbb{E}1_Q\boldsymbol{\delta}^2 + \mathbb{E}\left[1_Q\left((0, \frac{1_n}{n}\mathbf{Z}^\top)\left(\frac{1}{n}\vec{\mathbf{Z}}^\top\vec{\mathbf{Z}}\right)^{-1}(\frac{1}{n}\vec{\mathbf{Z}}^\top\boldsymbol{\delta})\right)^2\right]\\
&- 2\mathbb{E}\left[1_Q\bar{\boldsymbol{\delta}}(0, \frac{1_n}{n}\mathbf{Z}^\top)\left(\frac{1}{n}\vec{\mathbf{Z}}^\top\vec{\mathbf{Z}}\right)^{-1}(\frac{1}{n}\vec{\mathbf{Z}}^\top\boldsymbol{\delta})\right]\\
=&\mathbb{E}1_Q\bar{\boldsymbol{\delta}}^2 + \mathbb{E}\left[1_Q\left((0, \frac{1_n}{n}\mathbf{Z}^\top)\left(\frac{1}{n}\vec{\mathbf{Z}}^\top\vec{\mathbf{Z}}\right)^{-1}(\frac{1}{n}\vec{\mathbf{Z}}^\top\boldsymbol{\delta})\right)^2\right]\\
&- 2\sum_{k,l,m=1}^n \frac{1}{n^3}\mathbb{E}\left[1_Q\delta_k(0, Z_l^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_m\delta_m\right]\\
=&\mathbb{E}1_Q\bar{\boldsymbol{\delta}}^2 + \mathbb{E}\left[1_Q\left((0, \frac{1_n}{n}\mathbf{Z}^\top)\left(\frac{1}{n}\vec{\mathbf{Z}}^\top\vec{\mathbf{Z}}\right)^{-1}(\frac{1}{n}\vec{\mathbf{Z}}^\top\boldsymbol{\delta})\right)^2\right]\\
&- \frac{2(n-1)}{n^2}\mathbb{E}\left[1_Q\delta_1(0, Z_1^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_2\delta_2\right]\\
&- \frac{2(n-1)}{n^2}\mathbb{E}\left[1_Q\delta_1^2(0, Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_1\right] - \frac{2(n-1)}{n^2}\mathbb{E}\left[1_Q\delta_1(0, Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_2\delta_2\right]\\
&- \frac{2}{n^2}\mathbb{E}\left[1_Q\delta_1^2(0, Z_1^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_1\right] - \frac{2(n-1)(n-2)}{n^2}\mathbb{E}\left[1_Q\delta_1(0, Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_2\delta_3\right].
\end{aligned} \tag{79}
$$

The analyses for each of the seven terms in (79) are relatively complicated, which we postpone to Lemma 6.3 in the Appendix. Based on (79) and Lemma 6.3, one has

$$\begin{aligned}&\mathbb{E}1_Q\left(\hat\theta_{\mathrm{LS}}-\theta\right)^2=\frac{1}{n}\tau^2+O(\mathrm{poly}(p,n)\exp(-cn^{1/2}))\\&+\frac{1}{n^2}\left(2(\mathbb{E}\delta^2Z)^\top\mathbb{E}(ZZ^\top Z)+\Big(\mathrm{tr}(\mathbb{E}Z\delta Z^\top)\Big)^2+3\|\mathbb{E}Z\delta Z^\top\|_F^2-\mathbb{E}\mathrm{tr}(Z\delta^2Z^\top)+2\tau^2\right).\end{aligned}$$

Besides,

$$\begin{aligned}&\mathbb{E}1_{Q^c}\left(\hat\theta^1_{\mathrm{LS}}-\theta\right)^2\leq\mathbb{E}1_{Q^c}(2n\|Y\|_\infty+\mathbb{E}Y)^2\\\leq&\mathrm{poly}(n)(\mathbb{E}Y^{2+2\varepsilon})^{\frac{1}{1+\varepsilon}}\cdot(\mathbb{E}1_{Q^c})^{\frac{\varepsilon}{1+\varepsilon}}\leq\mathrm{poly}(n)\exp(-n^{1/2})\leq\mathrm{poly}(n)\exp(-n^{1/2}).\end{aligned}\tag{80}$$

Our final step gets back to the ℓ_2-risk of $\hat\theta^1_{\mathrm{LS}}$:

$$\begin{aligned}&\mathbb{E}\left(\hat\theta^1_{\mathrm{LS}}-\theta\right)^2=\mathbb{E}1_Q\left(\hat\theta^1_{\mathrm{LS}}-\theta\right)^2+\mathbb{E}1_{Q^c}\left(\hat\theta^1_{\mathrm{LS}}-\theta\right)^2\\=&\frac{1}{n}\tau^2+O(\mathrm{poly}(p,n)\exp(-cn^{1/2}))\\&+\frac{1}{n^2}\left(2(\mathbb{E}\delta^2Z)^\top\mathbb{E}(ZZ^\top Z)+\Big(\mathrm{tr}(\mathbb{E}Z\delta Z^\top)\Big)^2+3\|\mathbb{E}Z\delta Z^\top\|_F^2-\mathbb{E}\mathrm{tr}(Z\delta^2Z^\top)+2\tau^2\right).\end{aligned}$$

In fact, given $Z=\Sigma^{-1/2}X$, we have

$$\begin{aligned}&(\mathbb{E}\delta^2Z)^\top\mathbb{E}(ZZ^\top Z)=(\mathbb{E}\Sigma^{-1/2}X_c\delta^2)^\top\mathbb{E}(\Sigma^{-1/2}X_cZ^\top\Sigma^{-1}X_c)\\=&\mathbb{E}\left(\delta^2X_c\right)^\top\cdot\mathbb{E}\left(\Sigma^{-1}X_cX_c^\top\Sigma^{-1}X_c\right),\\&\mathrm{tr}\left(\mathbb{E}Z\delta Z^\top\right)=\mathrm{tr}\left(\mathbb{E}\Sigma^{-1/2}X\delta X^\top\Sigma^{-1/2}\right)=\mathrm{tr}\left(\Sigma^{-1}\Sigma_{\delta1}\right),\\&\mathbb{E}\mathrm{tr}\left(Z\delta^2Z^\top\right)=\mathbb{E}\mathrm{tr}\left(\Sigma^{-1/2}X\delta^2X^\top\Sigma^{-1/2}\right)=\mathrm{tr}\left(\Sigma^{-1}\Sigma_{\delta2}\right).\end{aligned}$$

Therefore, we have finished the proof of Theorem 2.5. □

References

Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853.

Ando, R. K. and Zhang, T. (2007). Two-view feature generation model for semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 25–32. ACM.

Azriel, D., Brown, L. D., Buja, A., Berk, R., and Zhao, L. (2016). Semi-supervised linear regression. *in preparation.*

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhao, L., and Zhang, K. (2014). Models as approximationsa conspiracy of random predictors and model violations against classical inference in regression. *preprint*.

Chakrabortty, A. and Cai, T. (2016). Efficient and adaptive linear regression in semi-supervised settings. *preprint*.

Chow, Y. S. and Teicher, H. (2012). *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media.

Cochran, W. G. (1953). *Sampling Techniques*. John Wiley And Sons, Inc.; New York.

Deng, L.-Y. and Wu, C. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82(398):568–576.

Johnson, R. and Zhang, T. (2008). Graph-based semi-supervised learning and spectral kernel design. *Information Theory, IEEE Transactions on*, 54(1):275–288.

Kriegler, B. and Berk, R. (2010). Small area estimation of the homeless in los angeles: An application of cost-sensitive stochastic gradient boosting. *The Annals of Applied Statistics*, pages 1234–1255.

Lafferty, J. D. and Wasserman, L. (2008). Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, pages 801–808.

Lohr, S. (2009). *Sampling: design and analysis*. Nelson Education.

Pitkin, E., Berk, R., Brown, L., Buja, A., George, E., Zhang, K., and Zhao, L. (2013). Improved precision in estimating average treatment effects. *arXiv preprint arXiv:1311.0291*.

Rossi, P. H. (1991). Strategies for homeless research in the 1990s. *Housing Policy Debate*, 2(3):1027–1055.

Rubin, D. B. (1990). [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Splawa-Neyman, J., Dabrowska, D., Speed, T., et al. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

Vapnik, V. (2013). *The nature of statistical learning theory.* Springer Science & Business Media.

Vershynin, R. (2012a). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686.

Vershynin, R. (2012b). Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge.

Wang, J. and Shen, X. (2007). Large margin semi-supervised learning. *Journal of Machine Learning Research.*

Wang, J., Shen, X., and Liu, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167.

Wang, J., Shen, X., and Pan, W. (2009). On efficient large margin semisupervised learning: Method and theory. *The Journal of Machine Learning Research*, 10:719–742.

Yaskov, P. (2014). Lower bounds on the smallest eigenvalue of a sample covariance matrix. *Electronic Communications in Probability*, 19:1–10.

Zhu, X. (2008). Semi-supervised learning literature survey. *technical report.*

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130.

Appendix: Additional Proofs

Additional Proofs for ℓ_2-risk Analysis

Proof of Theorem 2.6. Similarly to the previous proofs, we can transform X, Y and assume $\mu = 0, \mathrm{Cov}(X) = I_p, X = Z$ without loss of generality. We start by introducing the following notations and decomposition in (63):

$$\mathbf{X} = [X_1 \ \cdots \ X_n]^\top, \quad \mathbf{X}_{\text{full}} = [X_1 \ \cdots \ X_{n+m}]^\top, \quad \mathbf{X}_{\text{add}} = [X_{n+1} \ \cdots \ X_{n+m}]^\top,$$

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad \bar{\mathbf{X}}_{\text{full}} = \frac{1}{n+m}\sum_{k=1}^{n+m} X_k, \quad \bar{\mathbf{X}}_{\text{add}} = \frac{1}{m}\sum_{k=n+1}^{n+m} X_k.$$

$$\begin{aligned}\hat{\theta}_{\text{SSLS}} - \theta = & (\hat{\vec{\mu}} - \vec{\mu})^\top \beta + \bar{\boldsymbol{\delta}} + (\hat{\vec{\mu}} - \bar{\mathbf{X}}^\top)(\vec{\mathbf{X}}^\top \vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}\boldsymbol{\delta} \\ = & \bar{\mathbf{X}}_{\text{full}}^\top \beta_{(2)} + \bar{\boldsymbol{\delta}} + (0 \ \ - \bar{\mathbf{X}} + \bar{\mathbf{X}}_{\text{full}}) \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}\boldsymbol{\delta}.\end{aligned}$$

Again we note

$$\vec{\boldsymbol{\Xi}} = \frac{1}{n}\sum_{k=1}^{n} \vec{X}_k \vec{X}_k^\top, \quad \vec{\boldsymbol{\Xi}}_{-\Omega} = \frac{1}{n}\sum_{k=1, k\notin\Omega}^{n} \vec{X}_k \vec{X}_k^\top, \text{if } \Omega \subseteq \{1, \cdots, n\},$$

and define the "good" event that

$$Q = \left\{\hat{\theta}_{\text{SSLS}} = \hat{\theta}^1_{\text{SSLS}}, \quad \max\left\{\left\|\vec{\boldsymbol{\Xi}} - I_{p+1}\right\|, \left\|\vec{\boldsymbol{\Xi}}_{-\{i,j,k\}} - I_{p+1}\right\| \forall 1 \le i,j,k \le n\right\} \le C_1 n^{-1/4}\right\}.$$

Then,

$$\begin{aligned}&\mathbb{E}(\hat{\theta}^1_{\text{SSLS}} - \theta)^2 = \mathbb{E}1_Q(\hat{\theta}^1_{\text{SSLS}} - \theta)^2 + \mathbb{E}1_{Q^c}(\hat{\theta}^1_{\text{SSLS}} - \theta)^2 \\ =& \mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top \beta_{(2)} + \bar{\boldsymbol{\delta}}\right)^2 1_Q + \mathbb{E}\left((0 \ \ - \bar{\mathbf{X}} + \bar{\mathbf{X}}_{\text{full}}) \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}\boldsymbol{\delta}\right)^2 1_Q \\ &+ 2\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top \beta_{(2)} + \bar{\boldsymbol{\delta}}\right)\left((0, -\bar{\mathbf{X}} + \bar{\mathbf{X}}_{\text{full}}) \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}\boldsymbol{\delta}\right) 1_Q + \mathbb{E}1_{Q^c}(\hat{\theta}^1_{\text{SSLS}} - \theta)^2.\end{aligned} \tag{81}$$

In the analysis below, we analyze the four terms in (81) separately.

- First of all, since δ and $\vec{\mathbf{X}}$ are with mean zero and uncorrelated,

  $$\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top \beta_{(2)} + \bar{\boldsymbol{\delta}}\right)^2 = \mathrm{Var}(\bar{\mathbf{X}}_{\text{full}}^\top \beta_{(2)}) + \mathrm{Var}(\bar{\boldsymbol{\delta}}) = \frac{\tau^2}{n} + \frac{\beta_{(2)}^\top \mathbb{E}(X-\mu)(X-\mu)^\top \beta_{(2)}}{m+n}.$$

  Besides,

  $$\begin{aligned}\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top \beta_{(2)} + \bar{\boldsymbol{\delta}}\right)^2 1_{Q^c} \le & \left(\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top \beta_{(2)} + \bar{\boldsymbol{\delta}}\right)^4\right)^{1/2} (P(Q^c))^{1/2} \\ = & \mathrm{poly}(n)\left(\mathbb{E}(\bar{\mathbf{Y}} - \theta)^4\right)^{1/2} \exp(-cn^{1/2}) \le \mathrm{poly}(n)\exp(-cn^{1/2}).\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{E}\left(\bar{\mathbf{X}}^\top\beta_{(2)}+\bar{\boldsymbol{\delta}}\right)^2 1_Q=\mathbb{E}\left(\bar{\mathbf{X}}^\top\beta_{(2)}+\bar{\boldsymbol{\delta}}\right)^2-\mathbb{E}\left(\bar{\mathbf{X}}^\top\beta_{(2)}+\bar{\boldsymbol{\delta}}\right)^2 1_{Q^c}\\ =&\frac{\tau^2}{n}+\frac{\beta_{(2)}^\top\mathbb{E}(X-\mu)(X-\mu)^\top\beta_{(2)}}{m+n}+O\left(\mathrm{poly}(n)\exp(-cn^{1/2})\right).\end{aligned}\tag{82}$$

- Secondly,

$$\begin{aligned}&\mathbb{E}\left((0\ \ -\bar{\mathbf{X}}+\bar{\mathbf{X}}_{\text{full}})\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{\mathbf{X}}\boldsymbol{\delta}\right)^2 1_Q\\ \leq&2\mathbb{E}\left(\|\bar{\mathbf{X}}\|_2^2+\|\bar{\mathbf{X}}_{\text{full}}\|_2^2\right)\cdot\left(\left\|\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\right\|\right)^2 1_Q\cdot\left\|\vec{\mathbf{X}}\boldsymbol{\delta}\right\|_2^2\\ \leq&C\left(\mathbb{E}\|\bar{\mathbf{X}}\|_2^4+\mathbb{E}\|\bar{\mathbf{X}}_{\text{full}}\|_2^4\right)\cdot\left(\frac{C}{n}\right)^2\cdot\left(\mathbb{E}\|\vec{\mathbf{X}}\boldsymbol{\delta}\|_2^4\right)^{1/2}\\ \overset{\text{Lemma 6.3}}{\leq}&C\left(\frac{p^2}{n^2}\right).\end{aligned}\tag{83}$$

- The analysis of the third term in (81) is more complicated. We first decompose it as

$$\begin{aligned}&\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top\beta_{(2)}+\bar{\boldsymbol{\delta}}\right)\left((0,-\bar{\mathbf{X}}+\bar{\mathbf{X}}_{\text{full}})^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{\mathbf{X}}\boldsymbol{\delta}\right)1_Q\\ =&-\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top\beta_{(2)}+\bar{\boldsymbol{\delta}}\right)\left(0,\frac{m}{n(n+m)}1_n^\top\mathbf{X}\right)\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{\mathbf{X}}\boldsymbol{\delta}1_Q\\ &+\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top\beta_{(2)}+\bar{\boldsymbol{\delta}}\right)\left(0,\frac{1}{n+m}1_m^\top\mathbf{X}_{\text{add}}\right)\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{\mathbf{X}}\boldsymbol{\delta}1_Q\\ =&-\frac{m}{n(n+m)}\sum_{i,j,k=1}^{n}\left(\frac{1}{n+m}X_i^\top\beta_{(2)}+\frac{1}{n}\delta_i\right)(0,X_j)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)\vec{X}_k\delta_k1_Q\\ &-\frac{m}{n(n+m)^2}\sum_{i=n+1}^{n+m}\sum_{j,k=1}^{n}(X_i^\top\beta_{(2)})\,(0,X_j)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)\vec{X}_k\delta_k1_Q\\ &+\frac{1}{n+m}\sum_{i,k=1}^{n}\sum_{j=n+1}^{n+m}\left(\frac{1}{m+n}X_i^\top\beta_{(2)}+\frac{1}{n}\delta_i\right)(0,X_j)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)\vec{X}_k\delta_k1_Q\\ &+\frac{1}{(n+m)^2}\sum_{i,j=n+1}^{n+m}\sum_{k=1}^{n}X_i^\top\beta_{(2)}\,(0,X_j)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)\vec{X}_k\delta_k1_Q.\end{aligned}\tag{84}$$

The evaluation of the four terms above are provided separately in Lemma 6.3. Therefore,

$$\mathbb{E}\left(\bar{\mathbf{X}}^\top\beta_{(2)}+\bar{\boldsymbol{\delta}}\right)\left((0,-\bar{\mathbf{X}}+\bar{\mathbf{X}}_{\text{full}})\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{\mathbf{X}}\boldsymbol{\delta}\right)1_Q\leq C\frac{p^2}{n^2}.\tag{85}$$

- Similarly to (80) in Theorem 2.1, one can show

$$\mathbb{E}1_{Q^c}(\hat{\theta}_{\text{SSLS}}-\theta)^2=\mathrm{poly}(n)\exp(-n^{1/2}).$$

Combining (81) and the separate analyses above, we have finished the proof for this theorem. □

Proof of Propositions 2.2. Similarly to the proofs for the previous theorems, we can linearly transform X and without loss of generality assume $\mathbb{E}X = 0, \mathrm{Var}(X) = I_p$. We then consider $\hat{\theta}_{\mathrm{LS}} - \theta = \vec{\mu}^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \mathbf{Y} - \theta$. Note that

$$\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}} = \begin{bmatrix} 1 & \bar{\mathbf{X}}^\top \\ \bar{\mathbf{X}} & \hat{\mathbf{\Sigma}}_X + \bar{\mathbf{X}}\bar{\mathbf{X}}^\top \end{bmatrix}, \quad \frac{1}{n}\mathbf{X}^\top \mathbf{X} = \hat{\mathbf{\Sigma}}_X + \bar{\mathbf{X}}\bar{\mathbf{X}}^\top,$$

where $\bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^n X_k$, $\hat{\mathbf{\Sigma}}_X = \frac{1}{n}\sum_{k=1}^n (X_k - \bar{\mathbf{X}})(X_k - \bar{\mathbf{X}})^\top$. The block-wise matrix inverse formula yields

$$\left(\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} = \begin{bmatrix} 1 + \bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}} & -\bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \\ -\hat{\mathbf{\Sigma}}_X^{-\top} \bar{\mathbf{X}} & \hat{\mathbf{\Sigma}}_X^{-1} \end{bmatrix}. \tag{86}$$

By the expansion in (57), we have

$$\begin{aligned} \hat{\theta}_{\mathrm{LS}} - \theta &= \bar{\boldsymbol{\delta}} - \left(0, \frac{1_n}{n}\mathbf{X}^\top\right) \left(\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \left(\frac{1}{n}\vec{\mathbf{X}}^\top \boldsymbol{\delta}\right) \\ &= \left(\frac{1_n^\top}{n} + \bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}} \frac{1_n^\top}{n} - \frac{1}{n}\bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \mathbf{X}\right) \boldsymbol{\delta}. \end{aligned}$$

When $\mathbf{X} \in \mathbb{R}^{n \times p}$ are i.i.d. standard normal, it is commonly known that $\bar{\mathbf{X}}$, $\hat{\mathbf{\Sigma}}_X$ and $\boldsymbol{\delta}$ are all independent, and $\bar{\mathbf{X}} \overset{iid}{\sim} N(0, 1/n)$, $\hat{\mathbf{\Sigma}}_X^{-1}$ satisfies inverse-Wishart distribution $n \cdot \mathcal{W}_p^{-1}(I_p, n-1)$, and its expectation is $\mathbb{E}\hat{\mathbf{\Sigma}}_X^{-1} = \frac{n}{n-p-2} I_p$. Therefore,

$$\begin{aligned} &\mathbb{E}\left(\hat{\theta}_{\mathrm{LS}} - \theta\right)^2 = \mathbb{E}\delta^2 \cdot \mathbb{E}\left\| \frac{1_n^\top}{n} + \bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}} \frac{1_n^\top}{n} - \frac{1}{n}\bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \mathbf{X} \right\|_2^2 \\ =&\tau^2 \cdot \mathbb{E}\left(\frac{1}{n}(1 + \bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}})^2 + \frac{1}{n}(\bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} (\bar{\mathbf{X}}\bar{\mathbf{X}}^\top + \hat{\mathbf{\Sigma}}_X) \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}}) - \frac{2}{n}(\bar{\mathbf{X}} \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}} + (\bar{\mathbf{X}} \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}})^2)\right) \\ =&\tau^2 \cdot \mathbb{E}\left(\frac{1}{n} + \frac{1}{n}\bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}}\right) = \frac{\tau^2}{n}\left(1 + \mathrm{tr}\left(\mathbb{E}\hat{\mathbf{\Sigma}}_X^{-1} \cdot \mathbb{E}\bar{\mathbf{X}}\bar{\mathbf{X}}^\top\right)\right) = \frac{\tau^2}{n}\left(1 + \mathrm{tr}\left(\frac{nI_p}{n-p-2} \cdot \frac{I_p}{n}\right)\right) \\ =&\frac{\tau^2}{n} + \frac{p}{n(n-p-2)}\tau^2. \end{aligned}$$

The calculation for $\mathbb{E}(\hat{\theta}_{\mathrm{SSLS}} - \theta)^2$ is similar. Since $\hat{\theta}_{\mathrm{SSLS}} - \theta = \hat{\vec{\mu}}^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \mathbf{Y} - \theta$, by the calculation in Theorem 2.3, we have

$$\begin{aligned} \hat{\theta}_{\mathrm{SSLS}} - \theta &\overset{(63)}{=} (\hat{\vec{\mu}} - \vec{\mu})^\top \beta + \bar{\boldsymbol{\delta}} + \left(\hat{\vec{\mu}}^\top - \frac{1_n^\top}{n}\vec{\mathbf{X}}\right) \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}^\top \boldsymbol{\delta} \\ =&\bar{\mathbf{X}}_{\mathrm{full}}^\top \beta_{(2)} + \left(\frac{1_n^\top}{n} + \frac{m}{m+n}\bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}} \frac{1_n}{n} - \frac{m}{n(m+n)}\bar{\mathbf{X}}^\top \hat{\mathbf{\Sigma}}_X^{-1} \mathbf{X}\right) \boldsymbol{\delta} \\ &+ \frac{1}{m+n} 1_m^\top \mathbf{X}_{\mathrm{add}} (-\hat{\mathbf{\Sigma}}_X^{-1} \bar{\mathbf{X}} \frac{1_n^\top}{n} + \frac{1}{n}\hat{\mathbf{\Sigma}}_X^{-1} \mathbf{X}) \boldsymbol{\delta}. \end{aligned}$$

Since $\mathbf{X}_{\text{add}}$, $\boldsymbol{\delta}$ and $\mathbf{X}$ are all independent with mean 0, it is easy to check that any two of the three terms above are uncorrelated. Thus,

$$
\begin{aligned}
&\mathbb{E}\left(\hat{\theta}_{\text{SSLS}}-\theta\right)^2\\
=&\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top\beta_{(2)}\right)^2+\mathbb{E}\left[\left(\frac{1_n^\top}{n}+\frac{m}{m+n}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\frac{1_n^\top}{n}-\frac{m}{n(m+n)}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\mathbf{X}\right)\boldsymbol{\delta}\right]^2\\
&+\left[\frac{1}{m+n}1_m^\top\mathbf{X}_{\text{add}}(-\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\frac{1_n^\top}{n}+\frac{1}{n}\hat{\boldsymbol{\Sigma}}_X^{-1}\mathbf{X})\boldsymbol{\delta}\right]^2,
\end{aligned}
$$

$$
\mathbb{E}\left(\bar{\mathbf{X}}_{\text{full}}^\top\beta_{(2)}\right)^2=\frac{1}{m+n}\beta_{(2)}^\top\mathbb{E}(X-\mu)(X-\mu)^\top\beta_{(2)},
$$

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1_n^\top}{n}+\frac{m}{m+n}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\frac{1_n^\top}{n}-\frac{m}{n(m+n)}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\mathbf{X}\right)\boldsymbol{\delta}\right]^2\\
=&\tau^2\cdot\mathbb{E}\left\|\frac{1}{n}\left(1+\frac{m}{m+n}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\right)1_n^\top-\frac{m}{n(m+n)}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\mathbf{X}\right\|_2^2\\
=&\tau^2\cdot\mathbb{E}\Bigg\{\frac{1}{n}(1+\frac{m}{n+m}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}})^2+\frac{m^2}{n^2(m+n)^2}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}(n\hat{\boldsymbol{\Sigma}}_X+n\bar{\mathbf{X}}\bar{\mathbf{X}})\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}} \qquad (87)\\
&-\frac{2}{n}\left(1+\frac{m}{m+n}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\right)\cdot\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\cdot\frac{m}{m+n}\Bigg\}\\
=&\tau^2\left(\frac{1}{n}+\frac{m^2}{(n+m)^2n}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\right),
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{m+n}1_m^\top\mathbf{X}_{\text{add}}(-\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\frac{1_n^\top}{n}+\frac{1}{n}\hat{\boldsymbol{\Sigma}}_X^{-1}\mathbf{X})\boldsymbol{\delta}\right]^2\\
=&\frac{\tau^2}{(m+n)^2}\mathbb{E}\left\|1_m^\top\mathbf{X}_{\text{add}}(-\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\frac{1_n^\top}{n}+\frac{1}{n}\hat{\boldsymbol{\Sigma}}_X^{-1}\mathbf{X})\right\|_2^2\\
=&\frac{\tau^2}{(m+n)^2}\mathbb{E}\Bigg[1_m^\top\mathbf{X}_{\text{add}}\Big(\frac{1}{n}\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}-\frac{2}{n}\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\\
&+\frac{1}{n}\hat{\boldsymbol{\Sigma}}_X^{-1}(\hat{\boldsymbol{\Sigma}}_X+\bar{\mathbf{X}}\bar{\mathbf{X}}^\top)\hat{\boldsymbol{\Sigma}}_X^{-1}\Big)\mathbf{X}_{\text{add}}^\top1_m\Bigg]\\
=&\frac{\tau^2}{(m+n)^2n}\mathbb{E}1_m^\top\mathbf{X}_{\text{add}}\hat{\boldsymbol{\Sigma}}_X^{-1}\mathbf{X}_{\text{add}}^\top1_m=\frac{\tau^2m}{(m+n)^2n}\mathbb{E}X_{n+1}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}X_{n+1}.
\end{aligned}
$$

To sum up,

$$
\begin{aligned}
n\mathbb{E}(\hat{\theta}_{\text{SSLS}}-\theta)^2=&\tau^2+\frac{n}{m+n}\beta_{(2)}^\top\mathbb{E}(X-\mu)(X-\mu)^\top\beta_{(2)}\\
&+\frac{m\mathbb{E}\sigma^2(X)}{m+n}\left(\frac{m}{m+n}\left(n\mathbb{E}\bar{\mathbf{X}}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}\bar{\mathbf{X}}\right)+\frac{n}{m+n}\mathbb{E}X_{n+1}^\top\hat{\boldsymbol{\Sigma}}_X^{-1}X_{n+1}\right).
\end{aligned}
$$

Especially when $X \sim N(0, I)$, $\frac{1}{n}\hat{\mathbf{\Sigma}}_X^{-1}$ is independent of $\bar{\mathbf{X}}$ and satisfies the inverse Wishart distribution. At this point, $\mathbb{E}\hat{\mathbf{\Sigma}}_X^{-1} = \frac{nI_p}{n-p-2}$,

$$n\mathbb{E}\left(\hat{\theta}_{\mathrm{SSLS}} - \theta\right)^2 = \tau^2 + \frac{n}{m+n}\beta_{(2)}^\top \mathbb{E}(X-\mu)(X-\mu)^\top \beta_{(2)} + \frac{m}{m+n} \cdot \frac{n}{n-p-2} \cdot \tau^2,$$

which has finished the proof of Proposition 2.2. □

Proofs for Oracle Optimality Setting

Detailed Calculation for (40) **(Oracle risk for $\hat{\theta}^*_{\mathrm{ss}}$).** It is easy to see that $\hat{\theta}^*_{\mathrm{ss}}$ is an unbiased estimator for θ, thus

$$\begin{aligned}
&\mathbb{E}\left(\hat{\theta}^*_{\mathrm{ss}} - \theta\right)^2 = \mathrm{Var}\left(\hat{\theta}^*_{\mathrm{ss}}\right) \\
=&\sum_{k=1}^{n} \mathrm{Var}\left(\frac{Y_k}{n} - \frac{\xi_0(X_k)}{n} + \frac{\xi_0(X_k)}{n+m}\right) + \sum_{k=n+1}^{n+m} \mathrm{Var}\left(\frac{1}{n+m}\xi_0(X_k)\right) \\
=&n\mathbb{E}\left(\mathrm{Var}\left(\frac{Y_k}{n} - \frac{m}{n(n+m)}\xi_0(X_k)\middle| X_k\right)\right) \\
&+ n\mathrm{Var}\left(\frac{\xi(X_k)}{n} - \frac{m}{n(n+m)}\xi_0(X_k)\right) + \frac{m}{(n+m)^2}\mathrm{Var}\left(\xi(X_k)\right) \\
=&n\frac{\sigma^2}{n^2} + n\frac{\sigma_\xi^2}{(n+m)^2} + \frac{m\sigma_\xi^2}{(n+m)^2} \\
=&\frac{\sigma^2}{n} + \frac{1}{n+m}\sigma_\xi^2,
\end{aligned}$$

which has proved (40). □

Proof of Proposition 3.1. We first consider (41). For any given $\sigma^2 > 0$, $\xi_0(\cdot)$ and P_X, we consider the following subset of $\mathcal{P}_{\xi_0(\cdot),\sigma^2}$,

$$\begin{aligned}
P'_{\xi_0, P_X, \sigma^2} = \Big\{P : \int_Y P(Y, X) = P_X, Y = \xi_0(X) + c + \varepsilon,& \\
\varepsilon \text{ is independent from } X, \varepsilon \sim N(0, \sigma^2)\Big\}.&
\end{aligned} \tag{88}$$

Based on sample $\{X_i, Y_i\}_{i=1}^n$, known P_X and $\xi_0(X)$, we can rewrite the model to

$$Y_i - \xi_0(X_i) = c + \varepsilon_i, \quad i = 1, \ldots, n,$$

where Y_i and $\xi(X_i)$ are observable. By classical theory on normal mean estimation with Gaussian noise,

$$\inf_{\tilde{c}} \sup_{c \in \mathbb{R}} n\left(\tilde{c} - c\right)^2 = \sigma^2.$$

Note for the estimating problem in the original proposition, we target on estimating

$$\theta = \mathbb{E}(Y) = c + \mathbb{E}\xi_0(X),$$

where $\mathbb{E}\xi_0(X)$ is known. Thus, estimating θ is equivalent to estimating c, which implies

$$\begin{aligned}
&\inf_{\tilde{\theta}_n} \sup_{P\in\mathcal{P}_{\xi_0,\sigma^2}} \left[\mathbb{E}_P\left(n\left(\tilde{\theta}_n-\theta\right)^2\right)\right] \geq \inf_{\tilde{\theta}_n} \sup_{P\in\mathcal{P}'_{\xi_0,\sigma^2}} \left[\mathbb{E}_P\left(n\left(\tilde{\theta}_n-\theta\right)^2\right)\right] \\
\geq &\inf_{\tilde{c}} \sup_{c\in\mathbb{R}} n\left(\tilde{c}-c\right)^2 = \sigma^2.
\end{aligned}$$

Next we aim at the proof for (42). Suppose we are given fixed $\sigma_\xi^2, \sigma^2 > 0$ and linear function ξ_0. If $\xi_0(X)$ is a constant, ε_ξ always equals 0, the the problem transform to the first situation.

If $\xi_0(X) = aX+b$ with $a \neq 0$, since we can always normalize Y, without loss of generality let us assume $\xi_0(X) = X$. We also focus on the situation for $p = 1$ as the proof for $p > 1$ essentially follows. Now we consider the following subset of $\mathcal{P}^{\text{ss}}_{\xi_0,\sigma_Y^2,\sigma^2}$:

$$\begin{aligned}
\mathcal{P}^{\text{ss}'}_{\sigma_\xi^2,\sigma^2} = \Big\{P : &X \sim N(\mu,\sigma_\xi^2), Y = X + c + \varepsilon \text{ for some constants } \mu, c, \\
&\varepsilon \text{ is independent from } X, \varepsilon \sim N(0,\sigma^2)\Big\}.
\end{aligned}$$

In this case, $\mathbb{E}Y = \theta = c + \mu$. In order to calculate the minimax rate for estimating θ, we first consider the Bayes estimator for c and μ under the prior distribution $c, \mu \sim N(0, V^2)$, where $V^2 \to \infty$. It is easy to see that

$$p_0(\mu, c) \propto \exp\left(-\frac{\mu^2+c^2}{2V^2}\right),$$

$$p(Y, X|\mu, c) \propto \exp\left(-\frac{1}{2}\left(\frac{\sigma_\xi^2+\sigma^2}{\sigma_\xi^2\sigma^2}(X-\mu)^2 + \frac{1}{\sigma^2}(Y-\mu-c)^2 - \frac{2}{\sigma^2}(X-\mu)(Y-\mu-c)\right)\right),$$

$$p(X|\mu, c) \propto \exp\left(-\frac{1}{2\sigma_\xi^2}(X-\mu)^2\right).$$

Given observations $\{Y_k, X_k\}_{k=1}^n$ and $\{X_k\}_{k=n+1}^{n+m}$, the posterior distribution for μ and c is

$$\begin{aligned}
&\pi(\mu, c|\{Y_k, X_k\}_{k=1}^n, \{X_k\}_{k=n+1}^{n+m}) \propto \frac{p\left(\{Y_k, X_k\}_{k=1}^n, \{X_k\}_{k=n+1}^{n+m}|\mu, c\right) p_0(\mu, c)}{p\left(\{Y_k, X_k\}_{k=1}^n, \{X_k\}_{k=n+1}^{n+m}\right)} \\
\propto &\prod_{k=1}^n \exp\left(-\frac{1}{2}\left(\frac{\sigma_\xi^2+\sigma^2}{\sigma_\xi^2\sigma^2}(X_k-\mu)^2 + \frac{1}{\sigma^2}(Y_k-\mu-c)^2 - \frac{2}{\sigma^2}(X_k-\mu)(Y_k-\mu-c)\right)\right) \\
&\cdot \exp\left(-\frac{\mu^2+c^2}{V^2}\right) \cdot \prod_{k=n+1}^{n+m} \exp\left(-\frac{1}{2\sigma_\xi^2}(X-\mu)^2\right) / p\left(\{Y_k, X_k\}_{k=1}^n, \{X_k\}_{k=n+1}^{n+m}\right).
\end{aligned}$$

After simplification for the previous equation, when $V^2 \to \infty$, the joint posterior distribution of μ, c is

$$\mu, c|\{Y_k, X_k\}_{k=1}^{n}, \{X_k\}_{k=n+1}^{n+m} \sim N\left(\left(\tfrac{1}{n+m}X_k, \tfrac{1}{n}(Y_k - X_k)\right), \begin{bmatrix} \frac{1}{(n+m)\sigma_\xi^2} & 0 \\ 0 & \frac{1}{n\sigma^2} \end{bmatrix}\right)$$

Therefore, the Bayes estimator for $\theta = \mu + c$ is

$$\hat{\theta}_{bayes} = \mathbb{E}\left(\mu + c \middle| \{Y_k, X_k\}_{k=1}^{n}, \{X_k\}_{k=n+1}^{n+m}\right) = \bar{\mathbf{Y}} - \frac{1}{n}\sum_{k=1}^{n} X_k + \frac{1}{n+m}\sum_{k=1}^{n+m} X_k.$$

Similarly to the calculation for (40), it is easy to check that $\hat{\theta}_{bayes}$ has constant risk for all different values of c and μ:

$$n\mathbb{E}\left(\hat{\theta}_{bayes} - \theta\right)^2 = \sigma^2 + \frac{n}{n+m}\sigma_\xi^2.$$

This implies that θ_{bayes} is the minimax estimator for θ in distribution class $\mathcal{P}^{\mathrm{ss}'}_{\sigma_\xi^2, \sigma^2}$. To sum up, we have finished the proof for this proposition. $\square$

Proof of Theorem 3.1. For any $q \geq 0$, we denote

$$X^{(q)\bullet} = (X_1, \ldots, X_p, g_1(X), \ldots, g_q(X)), \quad \vec{X}^{(q)\bullet} = (1, X_1, \ldots, X_p, g_1(X), \ldots, g_q(X)).$$

Suppose

$$\tau_{(q)}^2 = \underset{\beta^{(q)} \in \mathbb{R}^{1+p+q}}{\arg\min}\ \mathbb{E}_{(Y,X)\sim P}\left(Y - (\beta^{(q)})^\top \vec{X}^{(q)\bullet}\right)^2.$$

Clearly, $\tau_{(q)}^2$ is an non-increasing sequence of q. Based on either Assumption (i) or (ii) of Proposition 3.1,

$$\lim_{q\to\infty} \tau_{(q)}^2 = \mathbb{E}\left(Y - \mathbb{E}(Y|X)\right)^2 = \sigma^2. \tag{89}$$

By Proposition 2.1, $\tau_{(q)}^2 + \mathrm{Var}((\beta^{(q)})^\top \vec{X}^{(q)\bullet}) = \mathrm{Var}(Y)$. By the law of total variance, $\sigma^2 + \mathrm{Var}(\xi(X)) = \mathrm{Var}(Y)$. Suppose $\hat{\theta}_{\mathrm{LS}}^{(q)}$, $\hat{\theta}_{\mathrm{SSLS}}^{(q)}$ are the least squares estimator and semi-supervised least squares estimator with the basis $(X_1, \ldots, X_p, g_1(X), \ldots, g_q(X))$. Corresponding, suppose $(\hat{\theta}_{\mathrm{LS}}^{(q)})^1$ and $(\hat{\theta}_{\mathrm{SSLS}}^{(q)})^1$ as the refined estimators based on (26). Based on Theorems 2.5 and 2.6, for fixed $q > 0$,

$$\limsup_{n\to\infty} n\mathbb{E}\left((\hat{\theta}_{\mathrm{LS}}^{(q)})^1 - \theta\right)^2 = \tau_{(q)}^2,$$

$$\limsup_{n\to\infty} n\mathbb{E}\left((\hat{\theta}_{\mathrm{SSLS}}^{(q)})^1 - \theta\right)^2 = \tau_{(q)}^2 + \rho\mathrm{Var}((\beta^{(q)})^\top \vec{X}^{(q)\bullet}) = (1-\rho)\tau_{(q)}^2 + \rho\mathrm{Var}(Y).$$

By (89),

$$\lim_{q\to\infty} \limsup_{n\to\infty} n\mathbb{E}\left((\hat{\theta}_{\mathrm{LS}}^{(q)})^1 - \theta\right)^2 = \sigma^2,$$

$$\lim_{q\to\infty}\limsup_{n\to\infty} n\mathbb{E}\left((\hat{\theta}^{(q)}_{\mathrm{SSLS}})^1-\theta\right)^2 = (1-\rho)\sigma^2+\rho\mathrm{Var}(Y)=\sigma^2+\rho\mathrm{Var}(\xi(X)).$$

Therefore, there exists sequence $\{q_n\}$ growing slowly enough that guarantees (43) and (44). Finally, the asymptotic distribution results hold similarly which we do no repeat here. □

Proofs for Application in Average Treatment Effect

Proof of Theorem 5.1. We shall note that $\hat{d}_{\mathrm{SSLS}} = \hat{\vec{\mu}}^\top\hat{\beta}_t - \hat{\vec{\mu}}^\top\hat{\beta}_c$. Based on (63), we have the following extensions for these two terms separately

$$\hat{\vec{\mu}}^\top\hat{\beta}_t - \theta_t = \left(\hat{\vec{\mu}}^\top - \vec{\mu}\right)^\top \beta_t + \bar{\boldsymbol{\delta}}_t - \left(0,\ \bar{\mathbf{X}}_t - \hat{\mu}\right)^\top \left(\vec{\mathbf{X}}_t^\top\vec{\mathbf{X}}_t\right)^{-1}\vec{\mathbf{X}}_t^\top\boldsymbol{\delta}_t,$$

$$\hat{\vec{\mu}}^\top\hat{\beta}_c - \theta_c = \left(\hat{\vec{\mu}}^\top - \vec{\mu}\right)^\top \beta_c + \bar{\boldsymbol{\delta}}_c - \left(0,\ \bar{\mathbf{X}}_c - \hat{\mu}\right)^\top \left(\vec{\mathbf{X}}_c^\top\vec{\mathbf{X}}_c\right)^{-1}\vec{\mathbf{X}}_c^\top\boldsymbol{\delta}_c.$$

Thus $\hat{d}_{\mathrm{SSLS}} - d$ has the following decomposition

$$\begin{aligned}
&\hat{d}_{\mathrm{SSLS}} - d = (\hat{\vec{\mu}}^\top\hat{\beta}_t - \theta_t) - (\hat{\vec{\mu}}^\top\hat{\beta}_c - \theta_c)\\
=&\bar{\boldsymbol{\delta}}_t - \bar{\boldsymbol{\delta}}_c + \left(\hat{\vec{\mu}}^\top - \vec{\mu}\right)^\top (\beta_t - \beta_c)\\
&- \left(0,\ \bar{\mathbf{X}}_t - \hat{\mu}\right)^\top \left(\vec{\mathbf{X}}_t^\top\vec{\mathbf{X}}_t\right)^{-1}\vec{\mathbf{X}}_t^\top\boldsymbol{\delta}_t + \left(0,\ \bar{\mathbf{X}}_c - \hat{\mu}\right)^\top \left(\vec{\mathbf{X}}_c^\top\vec{\mathbf{X}}_c\right)^{-1}\vec{\mathbf{X}}_c^\top\boldsymbol{\delta}_c.
\end{aligned} \tag{90}$$

Essentially the same as Theorem 2.1, one can show

$$\frac{\bar{\boldsymbol{\delta}}_t - \bar{\boldsymbol{\delta}}_c + (\hat{\vec{\mu}} - \vec{\mu})^\top(\beta_t - \beta_c)}{V} \to N(0,1), \tag{91}$$

$$\frac{\left(0,\ \bar{\mathbf{X}}_t - \hat{\mu}\right)^\top \left(\vec{\mathbf{X}}_t^\top\vec{\mathbf{X}}_t\right)^{-1}\vec{\mathbf{X}}_t^\top\boldsymbol{\delta}_t}{\sqrt{\tau_t^2/n_t}} \xrightarrow{d} 0, \quad \frac{\left(0,\ \bar{\mathbf{X}}_c - \hat{\mu}\right)^\top \left(\vec{\mathbf{X}}_c^\top\vec{\mathbf{X}}_t\right)^{-1}\vec{\mathbf{X}}_c^\top\boldsymbol{\delta}_t}{\sqrt{\tau_c^2/n_c}} \xrightarrow{d} 0, \tag{92}$$

Combining (91), (92) and (90), we have

$$\frac{\hat{d}_{\mathrm{SSLS}} - d}{V} \to N(0,1).$$

Next we show the asymptotic property for $\hat{V}$. Based on the proof of Theorem 2.3, we have already shown

$$\lim_{n_t\to\infty}\frac{MSE_t}{\tau_t^2} \xrightarrow{d} 1, \quad \lim_{n_c\to\infty}\frac{MSE_c}{\tau_c^2} \xrightarrow{d} 1.$$

Besides, $\hat{\beta}_{t,(2)} \xrightarrow{d} \beta_{t,(2)}$, $\hat{\beta}_{c,(2)} \xrightarrow{d} \beta_{c,(2)}$, $\hat{\mathbf{\Sigma}}_X \xrightarrow{d} \mathbb{E}(X-\mu)(X-\mu)^\top$ as $n_t, n_c \to \infty$. Thus, whenever $V^2 > 0$,

$$\hat{V}^2/V^2 \xrightarrow{d} 1, \quad \text{as } n_t, n_c \to \infty.$$

□

Proof of Technical Lemmas

We collect all technical proofs in this section.

Proof of Lemma 6.1.

- Part 1 directly follows from Theorem 5.39 in Vershynin (2012a).

- For Part 2, it can be calculated that

$$\mathbb{E}\left\|\sum_{k=1}^n Z_k\right\|_2^q = \mathbb{E}\left(\sum_{i=1}^p\left(\sum_{k=1}^n Z_{ki}\right)^2\right)^{q/2} \overset{\text{Hölder's ineq}}{\leq} \mathbb{E}\sum_{i=1}^p\left|\sum_{k=1}^n Z_{ki}\right|^q \cdot p^{q/2-1}.$$

  By Marcinkiewicz-Zygmund inequality (Chow and Teicher, 2012), under either Assumption 2 or 2', we have

$$\begin{aligned}\mathbb{E}\left|\sum_{k=1}^n Z_{ki}\right|^q &\overset{\text{M-Z ineq}}{\leq} C_q\mathbb{E}\left(\sum_{k=1}^n |Z_{ki}|^2\right)^{q/2}\\ &\overset{\text{Hölder's ineq}}{\leq} C_q n^{q/2-1}\sum_{k=1}^n \mathbb{E}|Z_{ki}|^q \leq C_q n^{q/2-1}, \quad i=1,\cdots,p.\end{aligned}$$

  Thus, we conclude that (74) holds.

- Finally we consider Part 3. Recall the fact that $\mathbb{E}\delta = 0$, $\mathbb{E}Z_k\delta = 0$. The proof is similar to Part 2. When $2 \leq q < 4$, under either Assumption 2 or 2',

$$\mathbb{E}\left|Z_{ki}\delta_k\right|^q \overset{\text{Hölder's ineq}}{\leq} \left(\mathbb{E}|Z_{ki}|^{\frac{4q}{4-q}}\right)^{\frac{4-q}{4}}\left(\mathbb{E}\delta_k^4\right)^{\frac{q}{4}} \leq C_q < \infty,$$

$$\mathbb{E}\left|\delta_k\right|^q \leq \left(\mathbb{E}\delta_k^4\right)^{\frac{q}{4}} \leq C_q < \infty.$$

  Thus, by Marcinkiewicz-Zygmund inequality (Chow and Teicher, 2012),

$$\begin{aligned}\mathbb{E}\left|\sum_{k=1}^n Z_{ki}\delta_k\right|^q &\overset{\text{M-Z ineq}}{\leq} C_q\mathbb{E}\left(\sum_{k=1}^n |Z_{ki}\delta_k|^2\right)^{q/2}\\ &\overset{\text{Hölder's ineq}}{\leq} C_q n^{q/2-1}\sum_{k=1}^n \mathbb{E}|Z_{ki}\delta_k|^q \leq C_q n^{q/2-1}, \quad i=1,\cdots,p.\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left|\sum_{k=1}^n \delta_k\right|^q &\overset{\text{M-Z ineq}}{\leq} C_q\mathbb{E}\left(\sum_{k=1}^n |\delta_k|^2\right)^{q/2}\\ &\overset{\text{Hölder's ineq}}{\leq} C_q n^{q/2-1}\sum_{k=1}^n \mathbb{E}|\delta_k|^q \leq C_q n^{q/2-1}, \quad i=1,\cdots,p.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left\|\sum_{k=1}^n \vec{Z}_k\delta_k\right\|_2^q =&\mathbb{E}\left(\left(\sum_{k=1}^n\delta_k\right)^2 + \sum_{i=1}^p\left(\sum_{k=1}^n \vec{Z}_{ki}\delta_k\right)^2\right)^{q/2}\\ \overset{\text{Hölder's ineq}}{\leq}& \mathbb{E}\left(\left|\sum_{k=1}^n\delta_k\right|^q + \sum_{i=1}^p\left|\sum_{k=1}^n Z_{ki}\right|^q\right)\cdot(p+1)^{q/2-1}\\ \leq& C_q(p+1)^{q/2}n^{q/2}\leq C_q(pn)^{q/2},\end{aligned}$$

which has shown (75). □

Proof of Lemma 6.2. Since

$$\begin{aligned}I =&\sum_{k=0}^{q-1}\left((-A^{-1}B)^k-(-A^{-1}B)^{k+1}\right)+(-A^{-1}B)^q\\ =&\sum_{k=0}^{q-1}\left(-A^{-1}B\right)^k(I+A^{-1}B)+(-A^{-1}B)^q\\ =&\sum_{k=0}^{q-1}\left(-A^{-1}B\right)^kA^{-1}(A+B)+\left(-A^{-1}B\right)^q\end{aligned}$$

Right multiply $(A+B)^{-1}$ to the equation above, we obtain (76). □

Lemma 6.3 (Separate Analysis of (79)**)** *Under the setting of the proof for Theorem 2.5, one has*

$$\mathbb{E}\left[1_Q\delta_1^2(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_1\right] = -\frac{1}{n}\mathrm{tr}\left(\left(\mathbb{E}\delta_1^2Z_1\right)^\top\cdot\mathbb{E}\left(Z_2Z_2Z_2^\top\right)\right)-\frac{\tau^2}{n}+O\left(\frac{p^2}{n^{5/4}}\right), \tag{93}$$

$$\mathbb{E}\left[1_Q\delta_1(0,Z_1^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_2\delta_2\right]=O\left(\exp(-cn)\cdot\mathrm{poly}(n,p)\right), \tag{94}$$

$$\mathbb{E}\left[1_Q\delta_1(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_2\delta_2\right]=-\frac{1}{n}\|\mathbb{E}Z\delta Z^\top\|_F^2+O\left(\frac{p^2}{n^{5/4}}\right), \tag{95}$$

$$\mathbb{E}\left[1_Q\delta_1^2(0,Z_1^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_1\delta_1\right]=\mathbb{E}\delta^2Z^\top Z+O\left(\frac{p}{n^{1/4}}\right), \tag{96}$$

$$\mathbb{E}\left[1_Q\delta_1(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_3\delta_3\right]=O\left(\frac{p^4}{n^3}\right), \tag{97}$$

$$\mathbb{E}\left[1_Q\bar{\boldsymbol{\delta}}^2\right]=\frac{\tau^2}{n}+O\left(\exp(-cn^{1/2})\mathrm{poly}(n)\right), \tag{98}$$

$$\begin{aligned}&\mathbb{E}\left[1_Q\left(\left(0,\frac{1_n}{n}\mathbf{Z}^\top\right)\vec{\boldsymbol{\Xi}}^{-1}\left(\frac{1}{n}\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\right)\right)^2\right]\\ =&\frac{1}{n^2}\left(\mathrm{tr}(\mathbb{E}Z\delta^2Z^\top)+\left(\mathrm{tr}(\mathbb{E}Z\delta Z^\top)\right)^2+\|\mathbb{E}Z\delta Z^\top\|_F^2\right)+O\left(\frac{p^2}{n^{2+1/4}}\right).\end{aligned} \tag{99}$$

Proof of Lemma 6.3. We analyze (93) - (99) separately in the next seven parts.

1. Recall $\vec{\boldsymbol{\Xi}}_{-\{1,2\}} = \frac{1}{n}\sum_{k=3}^{n} \vec{X}_k \vec{X}_k^\top$, we also denote $\vec{\boldsymbol{\Xi}}_{1,2} = \frac{1}{n}(\vec{X}_1\vec{X}_1^\top + \vec{X}_2\vec{X}_2^\top)$, $\vec{\boldsymbol{\Xi}}_{1,2,3} = \frac{1}{n}(\vec{X}_1\vec{X}_1^\top + \vec{X}_2\vec{X}_2^\top + \vec{X}_3\vec{X}_3^\top)$. Under the event Q, $\vec{\boldsymbol{\Xi}}^{-1}$ and $\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}$ are invertible. By Lemma 6.2, we can further calculate that

$$\begin{aligned}
&\mathbb{E}\left[1_Q \delta_1^2(0, Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_1\right]\\
=&\mathbb{E}\Big[1_Q\delta_1^2(0, Z_2^\top)\Big(\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1} - \vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\frac{1}{n}(\vec{Z}_1\vec{Z}_1^\top + \vec{Z}_2\vec{Z}_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\\
&+\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{\boldsymbol{\Xi}}_{1,2}\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{\boldsymbol{\Xi}}_{1,2}\vec{\boldsymbol{\Xi}}^{-1}\Big)\vec{Z}_1\Big].
\end{aligned}\tag{100}$$

We will calculate each term in (100) separately below. To get around the difficulty that Q is dependent of Z_1, Z_2, we introduce another event

$$Q' = \left\{\|\vec{\boldsymbol{\Xi}}_{-\{1,2\}} - I\| \leq Cn^{-1/4}\right\}.$$

Based on Lemma 6.1 and $p = o(n^{1/2})$, we have $P(Q') \geq 1 - \exp(-cn^{1/2})$ for some constant $c > 0$, $Q \subseteq Q'$ and Q' is independent of Z_1 and Z_2. Then

$$\begin{aligned}
&\left|\mathbb{E}\left[1_Q\delta_1^2(0, Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\right|\\
\leq&\left|\mathbb{E}\left[1_{Q'}\delta_1^2(0, Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\right| + \left|\mathbb{E}\left[1_{Q'\backslash Q}\delta_1^2(0, Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\right|\\
\leq&\left|\mathbb{E}\left\{\mathbb{E}_{Z_2}\left[1_{Q'}\delta_1^2(0, Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\middle| Z_1, Z_3, \ldots, Z_n\right\}\right|\\
&+\mathbb{E}1_{Q'\backslash Q}^2 \cdot \mathbb{E}\left[1_{Q'\backslash Q}\delta_1\|Z_2\|_2\|\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\|\|\vec{Z}_1\|_2\right]\\
\overset{\text{Cauchy-Schwarz}}{\leq}&0 + \cdot\left(\mathbb{E}\delta^4\right)^{\frac{1}{2}}\left(\mathbb{E}1_{Q'}\right)^{\frac{1}{8}}\left(\mathbb{E}\|Z_2\|^8\right)^{\frac{1}{8}}\mathbb{E}\left(\|\vec{Z}_1\|^8\right)^{\frac{1}{8}}\cdot\left(\mathbb{E}\left[1_{Q'\backslash Q}\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\right]^8\right)^{\frac{1}{8}}\\
\leq&O\left(\exp(-cn)\cdot\text{poly}(n,p)\right).
\end{aligned}\tag{101}$$

Note that

$$\begin{aligned}
(\mathbb{E}\delta_1^2 Z_1)^\top \cdot \mathbb{E}(Z_2Z_2^\top Z_2) + \tau^2 =&(\tau^2, \mathbb{E}\delta_1^2 Z_1)\cdot\begin{pmatrix}1\\ \mathbb{E}Z_2Z_2^\top Z_2\end{pmatrix} = \mathbb{E}\delta_1^2\vec{Z}_1^\top\cdot\mathbb{E}\vec{Z}_2\vec{Z}_2^\top\begin{pmatrix}0\\ Z_2\end{pmatrix}\\
=&\mathbb{E}\delta_1^2(0, Z_2^\top)^\top\vec{Z}_2\vec{Z}_2^\top\vec{Z}_1,
\end{aligned}$$

we also have

$$\begin{aligned}
&\Bigg|\mathbb{E}\left[1_Q\delta_1^2(0,Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\frac{1}{n}\left(\vec{Z}_2\vec{Z}_2^\top\right)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\\
&-\frac{1}{n}\left(\mathbb{E}\delta_1^2 Z_1\right)^\top\cdot\mathbb{E}\left(Z_2Z_2Z_2^\top\right)-\frac{\tau^2}{n}\Bigg|\\
\le&\frac{1}{n}\left|\mathbb{E}\left[1_{Q'}\delta_1^2(0,Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_2\vec{Z}_2^\top\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]-\mathbb{E}\delta_1^2(0,Z_2^\top)\vec{Z}_2\vec{Z}_2^\top\vec{Z}_1\right|\\
&+\frac{1}{n}\left|\mathbb{E}\left[1_{Q'\backslash Q}\delta_1^2(0,Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_2\vec{Z}_2^\top\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\right|\\
\le&\frac{1}{n}\left|\mathbb{E}\left[1_{Q'}\delta_1^2(0,Z_2^\top)(\vec{\boldsymbol{\Xi}}_{-\{1,2\}}-I)\vec{Z}_2\vec{Z}_2^\top\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\right|\\
&+\frac{1}{n}\left|\mathbb{E}\left[1_{Q'}\delta_1(0,Z_2^\top)I\vec{Z}_2\vec{Z}_2^\top(I-\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1})\vec{Z}_1\right]\right|\\
&+\frac{1}{n}\left|\mathbb{E}1_{(Q')^c}\delta_1^2(0,Z_2^\top)\vec{Z}_2\vec{Z}_2^\top\vec{Z}_1\right|+\frac{1}{n}\left|\mathbb{E}\left[1_{Q'\backslash Q}\delta_1^2(0,Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\vec{Z}_1^\top\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\right|.
\end{aligned}$$

Similarly as the procedure before, one can show that the formula above is no more than $O\left(\frac{p^2}{n^{5/4}}\right)+O\left(\exp(-cn)\cdot\mathrm{poly}(n,p)\right)$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[1_Q\delta_1^2(0,Z_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\frac{1}{n}\left(\vec{Z}_1\vec{Z}_1^\top\right)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{Z}_1\right]\\
=&\frac{1}{n}\left(\mathbb{E}\delta_1^2Z_1\right)^\top\cdot\mathbb{E}\left(Z_2Z_2Z_2^\top\right)+\frac{\tau^2}{n}+O\left(\frac{p^2}{n^{5/4}}\right).
\end{aligned}\tag{102}$$

Similarly to the calculation of (101) we can calculate that

$$\mathbb{E}\left[1_Q\delta_1^2(0,Z_2^\top)\left(\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\frac{1}{n}(\vec{Z}_1\vec{Z}_1^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\right)\vec{Z}_1\right]=O\left(\exp(-cn)\mathrm{poly}(n,p)\right).\tag{103}$$

$$\begin{aligned}
&\left|\mathbb{E}\left[1_Q\delta_1^2(0,Z_2^\top)\left(\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{\boldsymbol{\Xi}}_{1,2}\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{\boldsymbol{\Xi}}_{1,2}\vec{\boldsymbol{\Xi}}^{-1}\right)\vec{Z}_1\right]\right|\\
\le&\left|\mathbb{E}\left[1_Q\delta_1^2\|Z_2\|_2(1+cn^{-1/4})^3\|\vec{\boldsymbol{\Xi}}_{1,2}\|^2\|\vec{Z}_1\|_2\right]\right|\le O\left(\frac{p^3}{n^2}\right).
\end{aligned}\tag{104}$$

Summarizing (100), (101), (102), (103) and (104), we obtain (93).

2. Similarly to the calculation of (93), we have

$$\begin{aligned}
&\mathbb{E}\left[1_Q\delta_1(0,Z_1^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_2\delta_2\right]\\
=&\mathbb{E}\Big[1_Q\delta_1(0,Z_1^\top)\Big(\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}-\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\frac{1}{n}(\vec{Z}_1\vec{Z}_1^\top+\vec{Z}_2\vec{Z}_2^\top)\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\\
&+\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{\boldsymbol{\Xi}}_{1,2}\vec{\boldsymbol{\Xi}}_{-\{1,2\}}^{-1}\vec{\boldsymbol{\Xi}}_{1,2}\vec{\boldsymbol{\Xi}}^{-1}\Big)\vec{Z}_2\delta_2\Big].
\end{aligned}\tag{105}$$

We can calculate each term of (105) separately and similarly as the calculation for (93), then finish the proof of (94).

3. Similarly to the calculation of (93) and (94), we have

$$\begin{aligned}
&\mathbb{E}\left[1_Q\delta_1(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_2\delta_2\right]\\
=&\mathbb{E}\Big[1_Q\delta_1(0,Z_2^\top)\Big(\vec{\boldsymbol{\Xi}}^{-1}_{-\{1,2\}}-\vec{\boldsymbol{\Xi}}^{-1}_{-\{1,2\}}\frac{1}{n}(\vec{Z}_1\vec{Z}_1^\top+\vec{Z}_2\vec{Z}_2^\top)\vec{\boldsymbol{\Xi}}^{-1}_{-\{1,2\}}\\
&+\vec{\boldsymbol{\Xi}}^{-1}_{-\{1,2\}}\vec{\boldsymbol{\Xi}}_{1,2}\vec{\boldsymbol{\Xi}}^{-1}_{-\{1,2\}}\vec{\boldsymbol{\Xi}}_{1,2}\vec{\boldsymbol{\Xi}}^{-1}\Big)\vec{Z}_2\delta_2\Big]
\end{aligned}\tag{106}$$

Again based on the decomposition (106), we can similarly prove (95).

4. (96) can be calculated similarly based on the following idea,

$$\begin{aligned}
&\left|\mathbb{E}\left[1_Q\delta_1(0,Z_1^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_1\delta_1\right]-\mathbb{E}\delta^2Z^\top Z\right|\\
\le&\left|\mathbb{E}\left[1_Q\delta_1^2(0,Z_1^\top)\left(\vec{\boldsymbol{\Xi}}^{-1}-I\right)\vec{Z}_1\right]\right|+\left|\mathbb{E}1_{Q^c}\delta^2Z^\top Z\right|\\
\le&O\left(\frac{p}{n^{1/4}}\right)+O\left(\exp(-cn^{1/2})\text{poly}(p,n)\right)\\
=&O\left(\frac{p}{n^{1/4}}\right).
\end{aligned}$$

5. Note that we have the following decomposition,

$$\begin{aligned}
&\mathbb{E}\left[1_Q\delta_1(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_3\delta_3\right]\\
=&\mathbb{E}\Bigg[1_Q\delta_1(0,Z_2^\top)\Big(\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}-\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}+\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\\
&+\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}\Big)\vec{Z}_3\delta_3\Bigg].
\end{aligned}$$

Since $\mathbb{E}\delta_1=0$, $\mathbb{E}Z_2=0$, $\mathbb{E}\vec{Z}_3\delta_3=0$, similarly as the calculation before, we have

$$\begin{aligned}
\mathbb{E}\left[1_Q\delta_1(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{Z}_3\delta_3\right]&=O\left(\exp(-cn^{1/2})\text{poly}(p,n)\right)\\
\mathbb{E}\left[1_Q\delta_1(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{Z}_3\delta_3\right]&=O\left(\exp(-cn^{1/2})\text{poly}(p,n)\right)\\
\mathbb{E}\left[1_Q\delta_1(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{Z}_3\delta_3\right]&=O\left(\exp(-cn^{1/2})\text{poly}(p,n)\right)\\
\mathbb{E}\left[1_Q\delta_1(0,Z_2^\top)\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}_{-\{123\}}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}\vec{Z}_3\delta_3\right]&\\
\le O\left(\frac{p^4}{n^3}\right).&
\end{aligned}$$

6. Since $\mathbb{E}[\bar{\boldsymbol{\delta}}^2]=\frac{\mathbb{E}\delta^2}{n}=\frac{\tau^2}{n}$, we have

$$\begin{aligned}
\left|\mathbb{E}\left[1_Q\bar{\boldsymbol{\delta}}^2\right]-\frac{\tau^2}{n}\right|=\left|\mathbb{E}1_{Q^c}\bar{\boldsymbol{\delta}}^2\right|&\le\sqrt{\mathbb{E}1_{Q^c}^2\cdot\mathbb{E}\bar{\boldsymbol{\delta}}^4}\\
&\le C\exp(-cn^{1/2})\text{poly}(n),
\end{aligned}\tag{107}$$

which implies (98).

7. We can calculate that

$$\begin{aligned}
&\left|\mathbb{E}\left[1_Q\left(\left(0,1_n\mathbf{Z}^\top\right)\vec{\boldsymbol{\Xi}}^{-1}\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\right)^2\right]-\mathbb{E}\left((0,1_n\mathbf{Z}^\top)\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\right)^2\right|\\
\leq&\left|\mathbb{E}\left[1_Q\left((0,1_n\mathbf{Z}^\top)\vec{\boldsymbol{\Xi}}^{-1}\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\right)^2-1_Q\left((0,1_n\mathbf{Z}^\top)I\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\right)^2\right]\right|\\
&+\left|\mathbb{E}1_{Q^c}\left((0,1_n\mathbf{Z}^\top)\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\right)^2\right|\\
\leq&\left|\mathbb{E}\left[1_Q(0,1_n\mathbf{Z}^\top)(\vec{\boldsymbol{\Xi}}^{-1}-I)\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\cdot 1_Q(0,1_n\mathbf{Z}^\top)(\vec{\boldsymbol{\Xi}}^{-1}+I)\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\right]\right|\\
&+\mathbb{E}1_{Q^c}\|1_n\mathbf{Z}^\top\|_2^2\cdot\|\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\|_2^2\\
\leq&\mathbb{E}1_Q\|1_n\mathbf{Z}^\top\|_2^2\cdot\|\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\|_2^2\|\vec{\boldsymbol{\Xi}}^{-1}-I\|\cdot\|\vec{\boldsymbol{\Xi}}^{-1}+I\|+\mathbb{E}1_{Q^c}\|1_n\mathbf{Z}^\top\|_2^2\cdot\|\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\|_2^2\\
\leq&\left(\mathbb{E}\|\vec{Z}^\top\boldsymbol{\delta}\|_2^3\right)^{\frac{2}{3}}\cdot\left(\mathbb{E}1_Q\|\vec{\boldsymbol{\Xi}}^{-1}-I\|^6\|\vec{\boldsymbol{\Xi}}^{-1}+I\|^6\right)^{\frac{1}{6}}\cdot\left(\mathbb{E}\|1_n\mathbf{Z}^\top\|_2^{12}\right)^{\frac{1}{6}}\\
&+(\mathbb{E}1_{Q^c})^{\frac{1}{6}}\left(\mathbb{E}\|\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\|_2^3\right)^{\frac{2}{3}}\left(\mathbb{E}\|1_n\mathbf{Z}^\top\|_2^{12}\right)^{\frac{1}{6}}\\
\leq&C(pn)^2n^{-1/4},
\end{aligned}\tag{108}$$

$$\begin{aligned}
&\mathbb{E}\left((0,1_n\mathbf{Z}^\top)\vec{\mathbf{Z}}^\top\boldsymbol{\delta}\right)^2=\sum_{i,j,k,l=1}^n(0,Z_i^\top)\vec{Z}_j\delta_j(0,Z_k^\top)\vec{Z}_l\delta_l\\
=&\sum_{i=1}^n\mathbb{E}\left(Z_i^\top Z_i\delta_i\right)^2+\sum_{1\leq i\neq j\leq n}\left(Z_i^\top Z_j\delta_jZ_i^\top Z_j\delta_j+Z_i^\top Z_i\delta_iZ_j^\top Z_j\delta_j+Z_i^\top Z_j\delta_jZ_j^\top Z_i\delta_i\right)\\
=&O\left(np^2\right)+n(n-1)\left(\operatorname{tr}\left(\mathbb{E}Z_iZ_i^\top\cdot\mathbb{E}Z_j\delta_j^2Z_j^\top\right)+\left(\operatorname{tr}\left(Z\delta Z^\top\right)\right)^2+\operatorname{tr}\left(\mathbb{E}Z\delta Z^\top\right)^2\right)\\
=&n^2\left(\operatorname{tr}(Z\delta^2Z^\top)+\left(\operatorname{tr}(\mathbb{E}Z\delta Z^\top)\right)^2+\|\operatorname{tr}(\mathbb{E}Z\delta Z^\top)\|_F^2\right)+O\left(np^2\right).
\end{aligned}$$

Combine the two equalities above, we obtain (99). □

Lemma 6.4 (Separate Analysis in proof of Theorem 2.6) *Under the setting in Theorem 2.6, we have*

$$\mathbb{E}\sum_{i,j,k=1}^n X_i^\top\beta_{(2)}\left(0,X_j\right)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_k\delta_k1_Q=O\left(p^2\right),\tag{109}$$

$$\mathbb{E}\sum_{i,j,k=1}^n\delta_i\left(0,X_j\right)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_k\delta_k1_Q=O\left(p^2\right),\tag{110}$$

$$\mathbb{E}\sum_{i=n+1}^{n+m}\sum_{j,k=1}^n(X_i^\top\beta_{(2)})\left(0,X_j\right)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_k\delta_k1_Q=0,\tag{111}$$

$$\mathbb{E}\sum_{i,k=1}^{n}\sum_{j=n+1}^{n+m}(\frac{1}{m+n}X_i^\top\beta_{(2)}+\frac{1}{n}\delta_i)\,(0,X_j)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_k\delta_k 1_Q=0, \tag{112}$$

$$\mathbb{E}\sum_{i,j=n+1}^{n+m}\sum_{k=1}^{n}X_i^\top\beta_{(2)}\,(0,X_j)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_k\delta_k 1_Q=O\left(p^2\right). \tag{113}$$

Proof of Lemma 6.4. We first consider (109). By the fact that $X_1,\cdots,X_n$ are i.i.d. distributed, we have

$$\begin{aligned}
&\mathbb{E}\sum_{i,j,k=1}^{n}\mathbb{E}X_i^\top\beta_{(2)}\,(0,X_j)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_k\delta_k 1_Q\\
=&n(n-1)(n-2)\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_3\delta_3 1_Q\\
&+n\mathbb{E}X_1^\top\beta_{(2)}(0,X_1)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_1\delta_1 1_Q\\
&+n(n-1)\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_1\delta_1 1_Q\\
&+n(n-1)\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_2\delta_2 1_Q\\
&+n(n-1)\mathbb{E}X_1^\top\beta_{(2)}(0,X_1)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_2\delta_2 1_Q.
\end{aligned} \tag{114}$$

Note the expansion of

$$\begin{aligned}
\vec{\boldsymbol{\Xi}}\overset{\text{Lemma 6.2}}{=}\Big(&\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}-\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}+\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\\
&+\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{123}\vec{\boldsymbol{\Xi}}^{-1}\Big),
\end{aligned}$$

we have

$$\begin{aligned}
&\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\left(\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\vec{X}_3\delta_3 1_Q\\
=&\frac{1}{n}\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{X}_3\delta_3 1_Q-\frac{1}{n}\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{X}_3\delta_3 1_Q\\
&+\frac{1}{n}\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{X}_3\delta_3 1_Q\\
&+\frac{1}{n}\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}^{-1}\vec{X}_3\delta_3 1_Q.
\end{aligned}$$

Similarly to the proof of Lemma 6.3, we can compute that

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\mathbf{X}}_3\delta_3 1_Q-\frac{1}{n}\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{X}_3\delta_3 1_Q\\
&+\frac{1}{n}\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{X}_3\delta_3 1_Q=\operatorname{poly}(p,n)\exp(-cn^{1/2}),
\end{aligned}$$

$$\frac{1}{n}\mathbb{E}X_1^\top\beta_{(2)}(0,X_2)^\top\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}_{-\{123\}}^{-1}\vec{\boldsymbol{\Xi}}_{\{123\}}\vec{\boldsymbol{\Xi}}^{-1}\vec{X}_3\delta_3 1_Q=O\left(\frac{p^4}{n^4}\right).$$

Similarly for the other terms in (114), we can compute that

$$\begin{aligned}
&\mathbb{E}X_1^\top \beta_{(2)}(0, X_1)^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{X}_1 \delta_1 1_Q = \frac{1}{n}\mathbb{E}X_1^\top \beta_{(2)}(0, X_1)^\top \vec{\boldsymbol{\Xi}}^{-1} \vec{X}_1 \delta_1 1_Q \\
=&\frac{1}{n}\mathbb{E}X_1^\top \beta_{(2)}(0, X_1)^\top \left(\vec{\boldsymbol{\Xi}}_{-\{1\}}^{-1} - \vec{\boldsymbol{\Xi}}_{-\{1\}}^{-1} \vec{\boldsymbol{\Xi}}_{\{1\}} \vec{\boldsymbol{\Xi}}^{-1}\right) \vec{X}_1 \delta_1 1_Q \\
=&O\left(\frac{p^2}{n^2}\right),
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}X_1^\top \beta_{(2)}(0, X_2)^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}_1 \delta_1 1_Q + \mathbb{E}X_1^\top \beta_{(2)}(0, X_2)^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}_2 \delta_2 1_Q \\
&+ \mathbb{E}X_1^\top \beta_{(2)}(0, X_1)^\top \left(\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \vec{\mathbf{X}}_2 \delta_2 1_Q = O\left(\frac{p^3}{n^3}\right).
\end{aligned}$$

Combining the inequalities above, decomposition (114) along with the fact that $p = o(n^{1/2})$, we can get (109).

Next, the proofs of (110) and (113) are essentially the same as (109), which we do not repeat here. The proofs to (111) and (112) follows from the setting that $\{X_i\}_{i=n+1}^{n+m}$ are with mean zero and independent of $\{(\delta_i, X_i)\}_{i=1}^{n}$. □