Environment Invariant Linear Least Squares

Jianqing Fan Cong Fang Yihong Gu Tong Zhang

Princeton University, Peking University, and
The Hong Kong University of Science and Technology

Abstract

This paper considers a multi-environment linear regression model in which data from multiple experimental settings are collected. The joint distribution of the response variable and covariates may vary across different environments, yet the conditional expectations of the response variable, given the unknown set of important variables, are invariant. Such a statistical model is related to the problem of endogeneity, causal inference, and transfer learning. The motivation behind it is illustrated by how the goals of prediction and attribution are inherent in estimating the true parameter and the important variable set. We construct a novel *environment invariant linear least squares (EILLS)* objective function, a multi-environment version of linear least squares regression that leverages the above conditional expectation invariance structure and heterogeneity among different environments to determine the true parameter. Our proposed method is applicable without any additional structural knowledge and can identify the true parameter under a near-minimal identification condition related to the heterogeneity of the environments. We establish non-asymptotic ℓ_2 error bounds on the estimation error for the EILLS estimator in the presence of spurious variables. Moreover, we further show that the ℓ_0 penalized EILLS estimator can achieve variable selection consistency in high-dimensional regimes. These non-asymptotic results demonstrate the sample efficiency of the EILLS estimator and its capability to circumvent the curse of endogeneity in an algorithmic manner without any additional prior structural knowledge. To the best of our knowledge, this paper is the first to realize statistically efficient invariance learning in the general linear model.

Keywords: Least Squares, Endogeneity, Multiple Environments, Invariance, Heterogeneity, Structural Causal Model, Invariant Risk Minimization.

1 Introduction

The development of statistical regression methods dates back to the least squares proposed in the early nineteenth century (Legendre, 1805; Gauss, 1809). The ordinary linear least squares method, also known as *the combination of observations*, uses observations (data) to fit a model predicting the *response variable* as a linear function of several designated *explanatory variables*. At that time, the fitted models for specific tasks, for example, determining lunar motion, were successfully deployed in the real world for predicting the unseen future and were of great commercial and military significance (Stigler, 1986). For example, such a prediction of lunar liberation helps determine the ship's position and facilitates navigating the ocean during the Age of Discovery. This kind of prediction requires only a strong correlation with the response variable. Since then, one ultimate goal of fitting a regression model is to discover the law, which is *invariant* across time and space or more broadly *environments* to some extent, from the data and then ground it in the real world for prediction. The latter requires understanding causation.

With the rapidly growing amount of high-dimension data in the era of big data, there is a surge in demand for making predictions based on numerous explanatory variables (Fan et al., 2014; Wainwright, 2019; Fan et al., 2020). Compared to the circumstances in the nineteenth century, in which the target is to predict the response variable using fixed, carefully chosen explanatory variables, we are now in the stage where the algorithms such as Lasso (Tibshirani, 1997), SCAD (Fan & Li, 2001), Dantzig selector (Candes & Tao, 2007), to name a few, can automatically select tens of important variables responsible for the response variable out of thousands or more of candidates. There is considerable literature on the theoretical analysis of these methods regarding the estimation error and variable selection property (Zhao & Yu, 2006; Candes & Tao, 2007; Bickel et al., 2009; Bühlmann & Van De Geer, 2011; Wainwright, 2019; Fan et al., 2020), demonstrating their successes and promising prospects.

Let us use a thought experiment to illustrate the underlying risks and potential remedies when incorporating more candidate variables. Suppose our task is to fit a model to classify cows and camels based on extracted hierarchical features from some "oracle" agent. We find a dataset $\mathcal{D}$ containing 10k images of cows and camels from the Internet and use 70% of them to train a classifier on top of two features provided by the agent: the back shape x_1, and the background color x_2. It is contemplated that cows often appear on the grass while most camels appear on the sand. This indicates that we can use x_2 to build a classifier that works well on the training data and the remaining 30% test data by classifying whether the background color is green or yellow. Moreover, incorporating x_2 can also increase the accuracy of the classifier built on x_1. However, introducing x_2 is not what we expected: it may result in a disaster when we deploy it in real-world applications, for example, a detector in a place farming camels and cows, in which the background color is fixed. Problems of a similar nature arise readily in realistic applications (Torralba & Efros, 2011; Geirhos et al., 2020) and easily in high dimensions (Fan & Liao, 2014). A natural question is whether there are any purely data-driven methods to address such an endogeneity problem. Consider the case where we have another dataset $\widetilde{\mathcal{D}}$ in which an association between the background color and object label still exists yet slightly perturbs; for example, 60%/90% of the camels stand on sand in the two datasets, respectively. Intuitively, we can combine two different associations between the background color and the object label in these two datasets and infer that x_2 may be a "spurious" variable for prediction or causation.

The above thought experiment demonstrates that we may suffer from the "curse of endogeneity", that the conditional expectation of the response given all the explanatory variables may diverge from the law of interests, when including a lot of variables besides the true important variables before estimation (Fan et al., 2014). Such a problem will deviate our route toward building a decent prediction model grounded in the real world and yield non-robust predictions in other environments. Meanwhile, a potential data-driven strategy is to utilize the *heterogeneity* across datasets. This paper implements the above intuition to the linear regression model in statistical modeling, methodology, and theory. We propose a multi-environment version of linear least squares, whose key idea can be summarized as *the combination of combinations of observations* under heterogeneous environments: it combines the linear least squares (*combination of observations*) solutions across different datasets and uses their differences to determine the true parameter $\boldsymbol{\beta}^*$ in a completely data-driven manner.

1.1 The Problem under Study

In this work, we are interested in predicting the response variable $y \in \mathbb{R}$ with a linear function of the explanatory variable $\boldsymbol{x} \in \mathbb{R}^p$ using data from *multiple environments*. Suppose we have collected data from multiple resources/environments. Let $\mathcal{E}$ be the set of environments. For each *environment* $e \in \mathcal{E}$, we observe n i.i.d. $(\boldsymbol{x}_1^{(e)}, y_1^{(e)}), \ldots, (\boldsymbol{x}_n^{(e)}, y_n^{(e)}) \sim \mu^{(e)}$, typically assumed from the linear model[1]

$$y^{(e)} = (\boldsymbol{\beta}^*_{S^*})^\top \boldsymbol{x}^{(e)}_{S^*} + \varepsilon^{(e)} \qquad \text{with} \qquad \mathbb{E}[\varepsilon^{(e)} | \boldsymbol{x}^{(e)}_{S^*}] \equiv 0, \tag{1.1}$$

where the unknown set of important variables $S^* = \{j : \beta^*_j \neq 0\}$ and the model parameters $\boldsymbol{\beta}^*$ are the same, or *invariant*, across different environments, while $\mu^{(e)}$, the distribution of $(\boldsymbol{x}^{(e)}, y^{(e)})$, may vary. We aim to estimate $\boldsymbol{\beta}^*$ and S^* using the $n \cdot |\mathcal{E}|$ data $\{(\boldsymbol{x}_i^{(e)}, y_i^{(e)})\}_{e \in \mathcal{E}, i \in \{1,\ldots,n\}}$. The model assumptions of multiple environments resemble and slightly relax the assumptions in this paper's predecessors, for example, Peters et al. (2016); Rojas-Carulla et al. (2018); Pfister et al. (2021); Yin et al. (2021).

[1] Our main theoretical results still hold if the conditional expectation is replaced by $\mathbb{E}[\varepsilon^{(e)} \boldsymbol{x}^{(e)}_{S^*}] = 0$, in which $\boldsymbol{\beta}^*$ is the best linear predictor on important variables.

The main challenge behind identifying $\boldsymbol{\beta}^*$ is that the exogeneity condition (Engle et al., 1983) on all predictors, i.e., $\mathbb{E}[\varepsilon^{(e)}|\boldsymbol{x}^{(e)}] \equiv 0$ or at least $\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}] = 0$, no longer holds for each single $e \in \mathcal{E}$. This arises easily in the high-dimensional settings as argued in Fan & Liao (2014) and Fan et al. (2014) where a different data-driven strategy is proposed to solve the problem. Instead, for each environment $e \in \mathcal{E}$, the exogeneity condition only holds on the important variables according to (1.1). The only assumption available now is

$$\forall e \in \mathcal{E}, \qquad \mathbb{E}[y^{(e)}|\boldsymbol{x}_{S^*}^{(e)}] = (\boldsymbol{\beta}_{S^*}^*)^\top \boldsymbol{x}_{S^*}^{(e)}, \tag{1.2}$$

while $\mathbb{E}[y^{(e)}|\boldsymbol{x}^{(e)}]$ is not necessarily equal to $(\boldsymbol{\beta}^*)^\top \boldsymbol{x}^{(e)}$ for each single environment $e \in \mathcal{E}$. This further indicates that $\boldsymbol{\beta}^*$ may not be the best linear predictor for environment $e \in \mathcal{E}$, and the gap can be potentially large. Recall when the covariance matrix $\mathbb{E}[\boldsymbol{x}^{(e)}(\boldsymbol{x}^{(e)})^\top]$ is positive definite, the best linear predictor for environment $e \in \mathcal{E}$ can be written as

$$\boldsymbol{\beta}^{(e)} = \boldsymbol{\beta}^* + \Big(\mathbb{E}[\boldsymbol{x}^{(e)}(\boldsymbol{x}^{(e)})^\top]\Big)^{-1} \mathbb{E}[\boldsymbol{x}^{(e)}\varepsilon^{(e)}]. \tag{1.3}$$

This implies that when only one environment $e \in \mathcal{E}$ is taken into consideration, one can obtain a reduced mean squared error by incorporating those *linear spurious variables*, defined as the variables x_j satisfying $\mathbb{E}[x_j^{(e)}\varepsilon^{(e)}] \neq 0$, and other variables correlated to these variables. They help predict $\varepsilon^{(e)}$ and reduce prediction error under this environment. An example is the variable measuring "background color" in the above thought experiment. However, these variables are unstable because the corresponding association between these variables and y may vary or even be adversarial in other or unseen environments, leading to biased predictions.

Two ultimate goals of fitting a statistical model using data are prediction and attribution. Regarding "prediction", we hope our fitted model can make decent predictions on unseen data grounded in the real world rather than only on the "demo" test data. Regarding "attribution" (Efron, 2020), we wish to attribute the outcome/response variable to the significant variables of the fitted model such that the fitted model can lead to true scientific claims. See also Chapter 1 of Fan et al. (2020). We illustrate that the goal of estimating $\boldsymbol{\beta}^*$ and S^* for the model (1.1) unifies the above two seemingly separate goals, thereby expounding the motivation and importance of the multi-environment linear regression model.

Prediction. Consider the case where the potential distribution of the unseen data $\widetilde{\mu}$ may be different from those $\{\mu^{(e)}\}_{e\in\mathcal{E}}$ yet shares the same conditional expectation structure $\mathbb{E}_{\widetilde{\mu}}[y|\boldsymbol{x}_{S^*}] = (\boldsymbol{\beta}_{S^*}^*)^\top \boldsymbol{x}_{S^*}$ as those of observations in (1.1). Without informing of the unseen data distribution ahead, we can define the out-of-sample L_2 risk in an adversarial manner as

$$\mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}) = \sup_{\substack{\mathbb{E}_\mu[y|\boldsymbol{x}_{S^*}]=(\boldsymbol{\beta}_{S^*}^*)^\top \boldsymbol{x}_{S^*} \\ \mathrm{Var}_\mu[y|\boldsymbol{x}_{S^*}]\vee\max_{1\le j\le p}\mathbb{E}_\mu[x_j^2]\le\sigma^2}} \mathbb{E}_\mu\left[|\boldsymbol{\beta}^\top\boldsymbol{x} - y|^2\right].$$

It follows from Proposition 2.1 in Section 2.1 that $\boldsymbol{\beta}^*$ minimizes $\mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta})$ though it may not be the best linear predictor for a specific environment $\widetilde{\mu}$. This demonstrates that $\boldsymbol{\beta}^*$ is the optimal linear predictor robust to all potential distribution shifts on the unseen data to some extent. Moreover, Proposition 2.1 also implies that the ℓ_2 error $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ can be treated as an surrogate of the excess out-of-sample L_2 risk. This connects the estimation error of $\boldsymbol{\beta}^*$ to the "adversarial" mean squared error on potential unseen data.

Attribution. Let us restrict the multi-environment linear regression model to a specific instance – data with different experimental settings (Didelez et al., 2012; Peters et al., 2016) in the context of causal inference. To be specific, suppose there exists an environment $e_0 \in \mathcal{E}$ with observational data, and the rest are environments with interventional data (He & Geng, 2008), in which some interventions are performed on the variables other than the response variable y. The distribution of the variables in each environment can be encoded as a Structural Causal Model (SCM) (Glymour et al., 2016). Under the modularity (Schölkopf et al., 2012) assumption for SCMs that the intervention on variable x_j only changes the distribution of $\mathbb{P}(x_j|\boldsymbol{x}_{\mathsf{pa(j)}})$ where $\mathsf{pa}(j)$ is the set of direct causes of x_j, we can see that S^* in this instance is exactly the "direct cause" of the response variable. In this case, inferring S^* from data coincides with discovering the direct cause of the variable y, while estimating $\boldsymbol{\beta}^*$ is exactly estimating the true causal coefficients characterizing the mechanism $\mathbb{P}(y|\boldsymbol{x}_{\mathsf{pa(y)}}) = \mathbb{P}(y|\boldsymbol{x}_{S^*})$. See Section 2.3 for additional details.

1.2 Related Works

Multi-environment regression is common in many applications (Meinshausen et al., 2016; Čuklina et al., 2021). There is considerable literature proposing methods to estimate $\boldsymbol{\beta}^*$ and S^* starting from the pioneering work of Peters et al. (2016), in which they propose the "Invariant Causal Prediction" (ICP) to do causal discovery. The key idea behind it is the modularity assumption of SCMs, which is also referred to as invariance, autonomy (Haavelmo, 1944; Aldrich, 1989), and stability (Dawid & Didelez, 2010). To be specific, Peters et al. (2016) considers the multiple environments setting, in which the intervention may be applied to unknown variables other than y, leading to the following multi-environment linear regression model,

$$y^{(e)} = (\boldsymbol{\beta}^*)^\top \boldsymbol{x}^{(e)} + \varepsilon^{(e)} \qquad \text{with} \qquad \varepsilon^{(e)} \sim \mu_\varepsilon,\ \varepsilon^{(e)} \perp\!\!\!\perp \boldsymbol{x}_{S^*}^{(e)},\ \text{and}\ \mathbb{E}[\varepsilon^{(e)}] = 0. \tag{1.4}$$

They exploit the conditional distribution invariance structure, i.e., $\varepsilon^{(e)}|\boldsymbol{x}_{S^*}^{(e)}$, and propose a hypothesis testing procedure which can guarantee the selected set $\widehat{S}$ satisfies $\mathbb{P}(\widehat{S} \subseteq S^*) \ge 1 - \alpha$ for given Type-I error $\alpha > 0$. There is considerable literature extending this idea to other models such as Heinze-Deml et al. (2018); Pfister et al. (2019). Though the Type-I error is guaranteed for their method, these procedures may collapse to conservative solutions, such as $\widehat{S} = \emptyset$, and there is a lack of guarantee in the power of the test.

The conservative nature of ICP methods has sparked the development of numerous optimization-based methods. Built upon the invariance principle, there is also considerable literature (Ghassami et al., 2017; Rothenhäusler et al., 2019, 2021) proposing provably sample-efficient regression methods for estimating the causal parameter $\boldsymbol{\beta}^*$. However, they rely on additional, restrictive structures that simplify the original problem considerably. For instance, the Causal Dantzig (Rothenhäusler et al., 2019) presumes a linear SCM model with the heterogeneity of environments resulting from additive interventions. Implementing these methods in practical scenarios necessitates expert domain knowledge to validate these structures before estimation. This requirement introduces potential risks of model misspecification, as these methods are specifically tailored to the assumed structures.

There is also a considerable literature designing methods for the generic linear model (1.4), for example, Rojas-Carulla et al. (2018); Pfister et al. (2021); Yin et al. (2021). However, these methods tend to be heuristic, and finite sample guarantees are lacking. Inspired by the goal of achieving out-of-distribution generalization, Arjovsky et al. (2019) introduced another heuristic and model-agnostic approach. Their method, Invariant Risk Minimization (IRM), seeks a data representation such that the optimal predictors based on it are invariant across all environments. The ideas of IRM and its variants are widely applied in many machine learning tasks (Sagawa et al., 2020; Zhang et al., 2020; Krueger et al., 2021; Lu et al., 2021). Nevertheless, the theoretical understanding of invariance learning remains sparse, and the performance improvement of these methods over the standard empirical risk minimization is not clear (Rosenfeld et al., 2021; Kamath et al., 2021).

Another critical issue associated with these invariance learning methods is the inadequacy of theoretical insights into their identification conditions, which characterizes when it is possible to identify $\boldsymbol{\beta}^*$ in the model (1.4) using infinite data from *finite* environments. While Peters et al. (2016) delves into this issue, providing sufficient conditions for specific intervention and SCM structure types, it falls short of offering general criteria. Similarly, Arjovsky et al. (2019) also has some preliminary discussions on linear models, yet it stipulates an impractical requirement: $|\mathcal{E}| \ge d$. The understanding of the identification condition for a developed method is of great significance because it elucidates the method's sample efficiency in terms of the number of environments $|\mathcal{E}|$ required: a stronger identification condition may necessitate a potentially increased number of environments $|\mathcal{E}|$ to recover $\boldsymbol{\beta}^*$.

It is worth noting that the invariance structure we can exploit depends on the perturbations we expect in real-world scenarios. Due to its clear causal interpretation, the idea of invariance is also widely adopted in domain adaptation, transfer learning, and out-of-distribution generalization. Numerous invariance forms have been proposed based on various expected perturbations beyond the residual invariance (1.1). There is also considerable literature designing methods to leverage these invariance structures (Muandet et al., 2013; Gong et al., 2016; Heinze-Deml & Meinshausen, 2021) using data from multiple environments. A notable example in classification tasks is the invariance of the label conditional distribution, expressed as

$\mathbb{P}^{(e)}(\phi(\boldsymbol{x})|y) \equiv q(\boldsymbol{x}, y)$, where $\phi(\cdot)$ is an unknown data representation. We refer readers to Chen & Bühlmann (2021); Wang & Veitch (2023) for an overview.

1.3 New Contributions and Comparison with Predecessors

The significance of recovering S^* and $\boldsymbol{\beta}^*$ in the model (1.1), the sample inefficiency of previous methods in terms of n and $|\mathcal{E}|$, and the lack of theoretical understanding in invariance learning raise the following question:

> *Is provably sample-efficient estimation of $\boldsymbol{\beta}^*$ and S^* in the model (1.1) possible under a general, minimal identification condition?*

This paper provides an affirmative "yes" to the above question. In this paper, we propose the *environment invariant linear least squares (EILLS)* estimator that regularizes linear least squares with a *focused linear invariance regularizer* which promotes the invariance or exogeneity on selected variables. In particular, the population-level EILLS objective over the environments $\mathcal{E}$ is defined as

$$\mathsf{Q}_\gamma(\boldsymbol{\beta}) = \underbrace{\sum_{e\in\mathcal{E}} \mathbb{E}\left[|y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}|^2\right]}_{\mathsf{R}(\boldsymbol{\beta})} + \underbrace{\gamma \sum_{j=1}^p \mathbb{1}\{\beta_j \neq 0\} \times \sum_{e\in\mathcal{E}} \left|\mathbb{E}[(y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}) x_j^{(e)}]\right|^2}_{\mathsf{J}(\boldsymbol{\beta})} \tag{1.5}$$

with some hyper-parameter $\gamma > 0$. From a high-level viewpoint, the introduction of the regularizer $\mathsf{J}(\cdot)$ leverages the invariance structure (1.2) while the sum of L_2 loss $\mathsf{R}(\cdot)$ requires a good overall solution and prevents it from collapsing to conservative solutions such as $\boldsymbol{\beta} = 0$. See also Fan & Liao (2014) for a similar method to deal with endogeneity in high-dimensional regression. To get a crude sense of what $\mathsf{J}(\cdot)$ imposes, note that $\mathsf{J}(\cdot)$ discourages selecting variables that have a strong correlation with the fitted residuals in some environments and hence encourages the selected variables to be exogenous, uncorrelated with the fitted residuals for all the environments in a sense that

$$\forall e \in \mathcal{E}, j \text{ with } \bar{\beta}_j \neq 0 \qquad \mathbb{E}[(y^{(e)} - \bar{\boldsymbol{\beta}}^\top \boldsymbol{x}^{(e)}) x_j^{(e)}] = 0.$$

This can be seen when $\gamma \to \infty$ and has a similar spirit of imposing $\mathbb{E}\big[y^{(e)}|\boldsymbol{x}^{(e)}_{\mathrm{supp}(\bar{\boldsymbol{\beta}})}\big] = \bar{\boldsymbol{\beta}}^\top \boldsymbol{x}^{(e)}$ for all the $e \in \mathcal{E}$. As for the empirical counterpart, given the $n \cdot |\mathcal{E}|$ observations from $|\mathcal{E}|$ environments, the EILLS estimator minimizes $\widehat{\mathsf{Q}}_\gamma(\boldsymbol{\beta})$ which substitutes all the expectations in (1.5) by their corresponding empirical means. One can also use the EILLS estimator in high-dimensional case ($p > n$) by adding an ℓ_0 penalty with some pre-defined hyper-parameter $\lambda > 0$ when needed.

We further develop theories in Section 4 characterizing when our proposed EILLS estimator can consistently estimate $\boldsymbol{\beta}^*$ and S^*. We show that our EILLS estimator can identify $\boldsymbol{\beta}^*$ with some large enough γ in the sense that $\boldsymbol{\beta}^*$ is the unique minimizer of the population-level objective (1.5) if

$$\forall S \subseteq \{1, \ldots, p\} \text{ with } \sum_{e\in\mathcal{E}} \mathbb{E}[\varepsilon^{(e)} \boldsymbol{x}_S^{(e)}] \neq 0 \quad \Longrightarrow \quad \exists e, e' \in \mathcal{E}, \;\; \boldsymbol{\beta}^{(e,S)} \neq \boldsymbol{\beta}^{(e',S)} \tag{1.6}$$

where $\boldsymbol{\beta}^{(e,S)} = \mathrm{argmin}_{\mathrm{supp}(\boldsymbol{\beta}) \subseteq S} \mathbb{E}[|y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}|^2]$ is the best linear predictor constrained on S for environment $e \in \mathcal{E}$. Such a general identification condition is minimal if only linear information is used. This demonstrates that the EILLS objective can automatically circumvent the problem caused by endogeneity if incorporating any *pooled linear spurious variable*, defined as the variable x_j satisfying $\sum_{e\in\mathcal{E}} \mathbb{E}[\varepsilon^{(e)} x_j^{(e)}] \neq 0$, will lead to shifts in the least squares solutions among different environments, and it is sample-efficient in terms of $|\mathcal{E}|$ required.

Under the general identification condition (1.6), a series of non-asymptotic results are presented for the EILLS estimator when γ is greater than some critical threshold γ_*, illustrating the sample efficiency of our proposed EILLS estimator. In the low-dimensional regime, the vanilla EILLS estimator can obtain the optimal rate for linear regression. Moreover, the EILLS objective also embodies a variable selection property that can eliminate all the linear spurious variables while keeping all the true important variables provided n is large enough. In the high-dimensional regime where $p > n$, the ℓ_0 regularized EILLS estimator can achieve the variable selection consistency $\{j : \widehat{\beta}_j \neq 0\} = S^*$ with high probability in a similar manner to the standard ℓ_0 regularized least squares. Non-asymptotic upper bounds on γ_* for some concrete models are calculated, illustrating the applicability of our general result.

This paper proposes a provably sample-efficient estimation method for the general model (1.1). To the best of our knowledge, it is the *first* estimator with non-asymptotic guarantees in terms of $|\mathcal{E}|$ and n for the general multi-environment linear regression model (1.1), or a slightly restricted version of it, e.g., (1.4). As a comparison, previous provably sample-efficient estimation methods, like anchor regression (Rothenhäusler et al., 2021) and Causal Dantzig (Rothenhäusler et al., 2019), have predominantly been confined to the linear SCM with additive intervention case, that is, $((\boldsymbol{x}^{(e)})^\top, y^{(e)})^\top \leftarrow \boldsymbol{B}((\boldsymbol{x}^{(e)})^\top, y^{(e)})^\top + g(\boldsymbol{a}^{(e)}, \boldsymbol{\varepsilon}^{(e)})$ with some invariant matrix $\boldsymbol{B} \in \mathbb{R}^{(p+1)\times(p+1)}$. These imposed structures not only restrict the scalable applicability of these methods in practice but also hinder their potential for extension to nonlinear models. While our method primarily addresses the linear model, it demonstrates promise for extension to nonlinear models; see a discussion in Section 6.1. Our approach also shares a similar spirit to the ICP method but necessitates a weaker identification condition to recover the true parameter $\boldsymbol{\beta}^*$. This implies that our method is more sample-efficient than ICP regarding $|\mathcal{E}|$. Furthermore, we conduct a numerical analysis to benchmark our proposed EILLS estimator against various other invariance methods, demonstrating its superior performance in Section 5.

2 Setup and Background

2.1 Multi-Environment Linear Regression

Suppose we are interested in uncovering the "scientific truth" between the response variable y and $\boldsymbol{x}_{S^*}$, a sub-vector of p-dimensional covariate vector $\boldsymbol{x} \in \mathbb{R}^p$. As S^* is unknown, we collect many more variables, but the collected covariates are easily correlated with residuals of y on $\boldsymbol{x}_{S^*}$ (Fan & Liao, 2014; Fan et al., 2014). This gives rise to the following more realistic assumption: The distribution of $\boldsymbol{x}$ and y, μ, satisfies

$$\mu \in \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2} = \left\{ \begin{aligned} \mu :\ & \mathbb{E}_\mu[y|\boldsymbol{x}_{S^*}] = (\boldsymbol{\beta}^*_{S^*})^\top \boldsymbol{x}_{S^*}, \mathrm{Var}_\mu[y|\boldsymbol{x}_{S^*}] \le \sigma^2, \quad \mu\text{-a.s. } \boldsymbol{x}, \\ & \forall j \in [p],\ \mathbb{E}_\mu[x_j^2] \le \sigma^2 \end{aligned} \right\}. \tag{2.1}$$

Here, $S^* \subseteq [p]$ denotes the set of important variables that contribute to explain the "truth", $\boldsymbol{\beta}^*$ is the parameter of interest whose support set $\mathrm{supp}(\boldsymbol{\beta}^*)$ is S^*, and σ^2 is any given positive number.

With only one environment, it is impossible to identify S^*. Consider the data collected from multiple environments: in each environment e, the data $(\boldsymbol{x}^{(e)}, y^{(e)})$ follows from some law $\mu^{(e)}$ in $\mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$. Denote the set of environments by $\mathcal{E}$. For each environment $e \in \mathcal{E}$, we observe $n^{(e)}$ i.i.d. samples $(\boldsymbol{x}_1^{(e)}, y_1^{(e)}), \ldots, (\boldsymbol{x}_{n^{(e)}}^{(e)}, y_{n^{(e)}}^{(e)}) \sim \mu^{(e)}$ with $\mu^{(e)} \in \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$. Let $\mathcal{D}^{(e)} = \{(\boldsymbol{x}_i^{(e)}, y_i^{(e)})\}_{i=1}^{n^{(e)}}$ be the data collected from environment e. Our goal is to use the data $\{\mathcal{D}^{(e)}\}_{e\in\mathcal{E}}$ collected from multiple environments to find a linear predictor $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x}$ that minimizes the out-of-sample L_2 risk that is robust to all potential out-of-sample distributions in $\mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$. That is, find the minimizer of

$$\mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) = \sup_{\mu \in \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mu} \left[|y - \boldsymbol{\beta}^\top \boldsymbol{x}|^2 \right]. \tag{2.2}$$

The following proposition asserts that $\boldsymbol{\beta}^*$ is the unique minimizer of (2.2), whose proof is given in Appendix B.1.

Proposition 2.1 (Properties of $\mathsf{R}_{\mathsf{oos}}$)**.** *We have* $\mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}^*; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) = \sigma^2$*, and for any* $\boldsymbol{\beta} \in \mathbb{R}^p$*,*

$$\begin{aligned} \sigma^2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \le\ & \mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) - \mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}^*; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) \\ & \le \sigma^2 p^2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + 2\sigma^2 p \|\boldsymbol{\beta}_{[p]\setminus S^*}\|_2. \end{aligned} \tag{2.3}$$

As a by-product, Proposition 2.1 also implies that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$ can be seen as a surrogate metric for the out-of-sample L_2 risk (2.2). Proposition 2.1 motivates us to design methods to estimate the variable set S^* and the true parameter $\boldsymbol{\beta}^*$ using data from multiple environments $\mathcal{E}$ which share the same structure $\mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$. Particularly, define the CE-invariant set as follows. We will take advantage of the fact that S^* is CE-invariant across $\mathcal{E}$.

Definition 2.1 (CE-invariant Set). *A variable set $S \subseteq [p]$ is **conditional expectation invariant** (CE-invariant) across environments $\mathcal{E}$ if there exists some $\boldsymbol{\beta}$ with support set S such that*

$$\forall e \in \mathcal{E}, \qquad \mathbb{E}[y^{(e)}|\boldsymbol{x}_S^{(e)}] = \boldsymbol{\beta}_S^\top \boldsymbol{x}_S^{(e)}. \tag{2.4}$$

Necessity of heterogeneous environments. The following proposition argues that introducing multiple environments, i.e., $|\mathcal{E}| \ge 2$, with potentially different distributions is necessary to identify $\boldsymbol{\beta}^*$. The proof is given in Appendix B.2.

Proposition 2.2. *Given fixed $\boldsymbol{\beta}^*$, for any $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\mathrm{supp}(\boldsymbol{\beta}) \setminus S^* \neq \emptyset$, there exists some large enough $\sigma^2 > 0$ and $\mu \in \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$ such that*

$$\mathbb{E}_\mu[y|\boldsymbol{x}] = \boldsymbol{\beta}^\top \boldsymbol{x}.$$

Note that $\mathbb{E}[|\mathbb{E}[y|\boldsymbol{x}] - y|^2] \le \mathbb{E}[|f(\boldsymbol{x}) - y|^2]$ for any measurable function f. Hence Proposition 2.2 implies the population L_2 risk minimizer $f^\mu = \mathrm{argmin}_f\, \mathbb{E}_\mu[|f(\boldsymbol{x}) - y|^2]$ for a specific μ is not necessarily equal to $f^*(\boldsymbol{x}) = (\boldsymbol{\beta}^*)^\top \boldsymbol{x}$. Instead, the bias $\mathbb{E}[|(f^\mu - f^*)(\boldsymbol{x})|^2]$ between the population L_2 risk minimizer and the true regression function f^* can be arbitrarily large. It illustrates that running a linear regression on the data in one environment may not be able to estimate $\boldsymbol{\beta}^*$ well. Instead, running a linear regression on data in one sole environment may introduce some *spurious variables* even in a population aspect. These variables are spurious since incorporating them in the prediction model can increase the prediction performance in one environment. However, the associations between these variables and y are unstable and can lead to much worse prediction performances in other environments.

To estimate $\boldsymbol{\beta}^*$ well, we should use data from *heterogeneous* environments and exploit the invariant structure (2.4). This is the main idea of this paper: we will take advantage of the underlying heterogeneity and invariance to infer the important variable set S^* and the true parameter $\boldsymbol{\beta}^*$. Furthermore, we will show later in our theoretical analysis that only two environments, i.e., $|\mathcal{E}| = 2$, are enough to estimate $\boldsymbol{\beta}^*$ consistently.

2.2 Notations

The following notations will be used throughout this paper. Let $\mathbb{R}^+$ and $\mathbb{N}^+$ be the set of positive real numbers/integers, respectively. Let $|S|$ denote the cardinality of a set S. Define $2^S = \{A : A \subseteq S\}$. Denote by $[m] = \{1, \ldots, m\}$. Let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We use $a(n) \lesssim b(n)$ or $a(n) = O(b(n))$ to represent that there exists some universal constant $C > 0$ such that $a(n) \le C \cdot b(n)$ for all the $n \in \mathbb{N}^+$, and use $a(n) \gtrsim b(n)$ or $a(n) = \Omega(b(n))$ if $a(n) \ge c \cdot b(n)$ with some universal constant $c > 0$ for any $n \in \mathbb{N}^+$. Denote $a(n) \asymp b(n)$ if $a(n) \lesssim b(n)$ and $a(n) \gtrsim b(n)$. We use the notations $a(n) \ll b(n)$, or $b(n) \gg a(n)$, or $a(n) = o(b(n))$ if $\limsup_{n\to\infty}(a(n)/b(n)) = 0$.

We use bold lower case letter $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ to represent a d-dimension vector, let $\|\boldsymbol{x}\|_q = (\sum_{i=1}^d |x_i|^q)^{1/q}$ be it's ℓ_q norm. We use $\mathrm{supp}(\boldsymbol{x}) = \{j \in [d] : x_j \neq 0\}$ to denote the support set of the vector $\boldsymbol{x}$. For any $S = \{j_1, \ldots, j_{|S|}\} \subseteq [d]$ with $j_1 < j_2 < \cdots < j_{|S|}$, we denote $[\boldsymbol{x}]_S = (x_{j_1}, \ldots, x_{j_{|S|}})^\top$ be the $|S|$-dimension sub-vector of $\boldsymbol{x}$ and abbreviate it as $\boldsymbol{x}_S$ when there is no ambiguity. We use bold upper case letter $\boldsymbol{A} = [A_{i,j}]_{i\in[n],j\in[m]}$ to denote a matrix. Denote $\boldsymbol{A}_{S,T} = [A_{i,j}]_{i\in S,j\in T}$ be the sub-matrix of $\boldsymbol{A}$, and abbreviate $\boldsymbol{A}_{S,S}$ as $\boldsymbol{A}_S$. We let $\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{v}\in\mathbb{R}^m,\|\boldsymbol{v}\|_2=1} \|\boldsymbol{A}\boldsymbol{v}\|_2$.

For each $e \in \mathcal{E}$, suppose we have $n^{(e)}$ observations $\{(\boldsymbol{x}_i^{(e)}, y_i^{(e)})\}_{i=1}^{n^{(e)}} \subseteq \mathbb{R}^p \times \mathbb{R}$ drawn i.i.d. from some distribution $\mu^{(e)}$. Let $\mathbb{E}[f(\boldsymbol{x}^{(e)}, y^{(e)})] = \int f(\boldsymbol{x}, y)\mu^{(e)}(d\boldsymbol{x}, dy)$ and $\widehat{\mathbb{E}}[f(\boldsymbol{x}^{(e)}, y^{(e)})] = \frac{1}{n^{(e)}} \sum_{i=1}^{n^{(e)}} f(\boldsymbol{x}_i^{(e)}, y_i^{(e)})$

for a measurable function f. Define the empirical L_2 risk $\widehat{\mathsf{R}}^{(e)}$ and population L_2 risk $\mathsf{R}^{(e)}$ as

$$\widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) = \widehat{\mathbb{E}}\left[|y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}|^2\right] \qquad \text{and} \qquad \mathsf{R}^{(e)}(\boldsymbol{\beta}) = \mathbb{E}\left[|y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}|^2\right]. \tag{2.5}$$

We also define $\varepsilon^{(e)} = y^{(e)} - \mathbb{E}[y^{(e)}|\boldsymbol{x}_{S^*}^{(e)}] = y^{(e)} - (\boldsymbol{\beta}^*)^\top \boldsymbol{x}^{(e)}$ and $\varepsilon_i^{(e)} = y_i^{(e)} - (\boldsymbol{\beta}^*)^\top \boldsymbol{x}_i^{(e)}$. Let $\widehat{\boldsymbol{\Sigma}}^{(e)}$ and $\boldsymbol{\Sigma}^{(e)}$ denote the empirical covariance matrix and population covariance matrix for environment $e \in \mathcal{E}$, respectively, that is, $\widehat{\boldsymbol{\Sigma}}^{(e)} = \widehat{\mathbb{E}}\left[\boldsymbol{x}^{(e)}(\boldsymbol{x}^{(e)})^\top\right]$ and $\boldsymbol{\Sigma}^{(e)} = \mathbb{E}\left[\boldsymbol{x}^{(e)}(\boldsymbol{x}^{(e)})^\top\right]$. When $\boldsymbol{\Sigma}^{(e)}$ is positive definite, for any $S \subseteq [p]$ we can define the $\boldsymbol{\beta}^{(e,S)}$, the population-level best linear predictor constrained on S for environment $e \in \mathcal{E}$ as

$$\boldsymbol{\beta}^{(e,S)} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \mathrm{supp}(\boldsymbol{\beta}) \subseteq S}{\operatorname{argmin}} \mathsf{R}^{(e)}(\boldsymbol{\beta}). \tag{2.6}$$

It is worth noticing that though $\boldsymbol{\beta}^{(e,S)} \in \mathbb{R}^p$, the support set of $\boldsymbol{\beta}^{(e,S)}$ is a subset of S.

2.3 An Example: Structural Causal Model with Different Interventions

We provide a generic statistical model of interest in Section 2.1. Here, we present an instance of the multiple environments setting – structural causal models with different interventions. We will show in this example where such heterogeneity of environments comes from and provide a specific meaning of S^* and $\boldsymbol{\beta}^*$. We first introduce the concept of the Structural Causal Model (Glymour et al., 2016), also called the Structural Equation Model (Bollen, 1989).

A structural causal model (SCM) on $p+1$ variables $\boldsymbol{z} = (z_1, \ldots, z_{p+1}) \in \mathbb{R}^{p+1}$ consists of $p+1$ structural assignments $\{f_j\}_{j=1}^{p+1}$,

$$z_j \leftarrow f_j(\boldsymbol{z}_{\mathtt{pa}(j)}, u_j) \qquad j = 1, \ldots, p+1 \tag{2.7}$$

where $\mathtt{pa}(j) \subseteq [p+1]$ is the set of parents, or *direct cause*, of the variable z_j, and $u_1, \ldots, u_{p+1}$ are independent zero mean *exogenous variables*. We call an SCM a *linear SCM* if all the functions f_j are linear, in which the above assignments (2.7) can be written in a matrix form as $\boldsymbol{z} \leftarrow \boldsymbol{B}\boldsymbol{z} + \boldsymbol{u}$ for some structured matrix $\boldsymbol{B} \in \mathbb{R}^{(p+1)\times(p+1)}$.

The above SCM naturally induces a directed graph $G = (V, E)$, where $V = \{1, \ldots, p+1\}$ is the set of nodes, E is the edge set such that $(i, j) \in E$ if and only if $i \in \mathtt{pa}(j)$. We say there is a directed path from i to j if there exists $(v_1, \cdots, v_k)$ with $k \geq 2$ such that $v_1 = i$, $v_k = j$ and $(v_l, v_{l+1}) \in E$ for any $l \in [k-1]$. We call a directed graph a *directed acyclic graph (DAG)* if there does not exist a direct path from j to j for any $j \in [p+1]$. Throughout this paper, the induced graphs of the SCMs we consider are all DAGs.

Consider the following $|\mathcal{E}|$ SCMs on $p+1$ variables $\boldsymbol{z} = (x_1, \ldots, x_p, y) \in \mathbb{R}^{p+1}$, that for any $e \in \mathcal{E}$, the following assignments holds

$$\begin{aligned} x_j^{(e)} &:= z_j^{(e)} \leftarrow f_j^{(e)}(\boldsymbol{z}_{\mathtt{pa}(j)}^{(e)}, u_j^{(e)}), & j = 1, \ldots, p \\ y^{(e)} &:= z_{p+1}^{(e)} \leftarrow (\boldsymbol{\beta}^*)_{\mathtt{pa}(p+1)}^\top \boldsymbol{x}_{\mathtt{pa}(p+1)}^{(e)} + u_{p+1}^{(e)}, & \end{aligned} \tag{2.8}$$

where all the $\boldsymbol{u}^{(e)} = (u_1^{(e)}, \ldots, u_{p+1}^{(e)})$ are independent across different environments and also have independent, zero mean entries. Here the direct cause relationship $\mathtt{pa} : [p+1] \to 2^{[p+1]}$ and the function $f_{p+1}(\boldsymbol{x}_{\mathtt{pa}(p+1)}^{(e)}, u_{p+1}^{(e)}) = (\boldsymbol{\beta}^*)_{\mathtt{pa}(p+1)}^\top \boldsymbol{x}_{\mathtt{pa}(p+1)}^{(e)} + u_{p+1}^{(e)}$ are both invariant across different $e \in \mathcal{E}$. The structural assignments $f_j^{(e)}$ for variables $x_j^{(e)}$ and the distribution of exogenous variables $u_j^{(e)}$ may vary among different environments. We use the superscript (e) to emphasize this heterogeneity. We use $\boldsymbol{z}$ in the assignments (2.8) to emphasize that y can be the direct cause of some variables x_j. We can see that the heterogeneity may come from performing do-interventions on variables other than y (which will result in different $f_j^{(e)}$ or distribution $u_j^{(e)}$). Fig. 1 provides an example of the model (2.8). Here we consider the case where there are

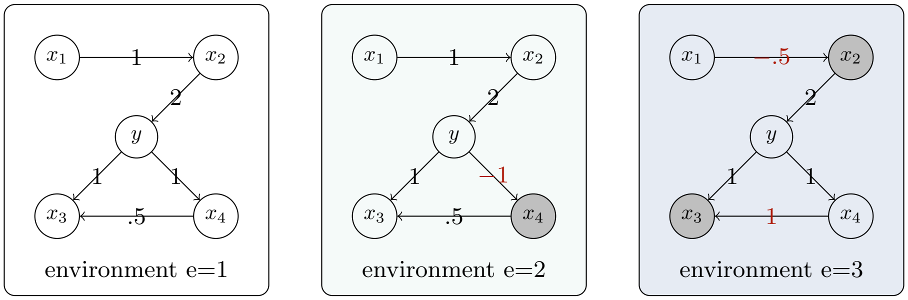


Figure 1: Linear SCMs with different interventions when $p = 4$ and $|\mathcal{E}| = 3$. Here $z_5 = y$, $S^* = \{2\}$, and we omit the dependence on the exogenous variables $u_1, \ldots, u_5$ for a clear presentation. The arrow from node i to node j is marked by $B^{(e)}_{j,i}$ if $B^{(e)}_{j,i} \neq 0$. We can treat $e = 1$ as the observational environment and $e = 2, 3$ as interventional environments. One intervention is performed to the mechanisms of variable x_4 (gray shadowed) in environment $e = 2$, and simultaneously interventions on variable x_2, x_3 (gray shadowed) are applied in environment $e = 3$.

no hidden confounders and the system is not anti-causal, since the model misspecification in either direction cannot be addressed without imposing further structural assumptions.

It is easy to verify that the above model (2.8) is an instance of the problem formulation discussed in Section 2.1 with the important variable set $S^* = \mathtt{pa}(p+1)$ and the true parameter $\boldsymbol{\beta}^*$. Moreover, the set S^* and the parameter $\boldsymbol{\beta}^*$ have precise meanings in the context of SCM. Specifically, S^* is the set of *direct causes* of the response variable. From a traditional viewpoint of causal inference, for each $j \in S^*$, the corresponding coefficient β^*_j measures how the outcome y will change if we only apply intervention on the variable x_j.

3 Methodology

3.1 Focused Linear Invariance Regularizer

As discussed in Proposition 2.2 and Section 2.1, it is necessary to leverage the conditional expectation invariance structure across heterogeneous environments. The straightforward idea is to impose a conditional expectation invariance to the solution $\bar{\boldsymbol{\beta}}$ we find. That is, we wish our solution $\bar{\boldsymbol{\beta}}$ to satisfy

$$\forall e \in \mathcal{E}, \qquad \mathbb{E}\left[y^{(e)} | \boldsymbol{x}^{(e)}_{\mathrm{supp}(\bar{\boldsymbol{\beta}})}\right] \equiv \bar{\boldsymbol{\beta}}^\top \boldsymbol{x}^{(e)}$$

from a population perspective. To implement the idea efficiently, we consider the variational principle that, for any $e \in \mathcal{E}$ and given $\boldsymbol{\beta} \in \mathbb{R}$ with support set S,

$$\begin{aligned}
\mathbb{E}[y^{(e)} | \boldsymbol{x}^{(e)}_S] = \boldsymbol{\beta}_S^\top \boldsymbol{x}^{(e)}_S \quad &\Leftrightarrow \quad \mathbb{E}\left[\left(y^{(e)} - \boldsymbol{\beta}_S^\top \boldsymbol{x}^{(e)}_S\right) g(\boldsymbol{x}^{(e)}_S)\right] = 0 \qquad \forall g \text{ with } \mathbb{E}[|g(\boldsymbol{x}^{(e)}_S)|^2] < \infty \\
&\overset{(a)}{\Rightarrow} \quad \mathbb{E}\left[\left(y^{(e)} - \boldsymbol{\beta}_S^\top \boldsymbol{x}^{(e)}_S\right) g(\boldsymbol{x}^{(e)}_S)\right] = 0 \qquad \forall g \text{ linear} \\
&\Leftrightarrow \quad \mathbb{E}\left[\left(y^{(e)} - \boldsymbol{\beta}_S^\top \boldsymbol{x}^{(e)}_S\right) \boldsymbol{x}^{(e)}_j\right] = 0 \qquad \forall j \in S \\
&\Leftrightarrow \quad \nabla_j \mathsf{R}^{(e)}(\boldsymbol{\beta}) = 0 \qquad \forall j \in S
\end{aligned} \tag{3.1}$$

where the statements connected by "⇔" are equivalent and $\mathsf{R}^{(e)}(\boldsymbol{\beta})$ is defined by (2.5), and we make a relaxation in (a) such that the last statement $\nabla_S \mathsf{R}^{(e)}(\boldsymbol{\beta}) = 0$ is a necessary but not sufficient condition for $\mathbb{E}[y^{(e)} | \boldsymbol{x}^{(e)}_S] = \boldsymbol{\beta}_S^\top \boldsymbol{x}^{(e)}_S$.

Based on the above derivations, we propose to minimize a *focused linear invariance regularizer*, whose population-level form $\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega})$ can be written as

$$\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) = \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\} \sum_{e \in \mathcal{E}} \frac{\omega^{(e)}}{4} |\nabla_j \mathsf{R}^{(e)}(\boldsymbol{\beta})|^2, \tag{3.2}$$

where $\boldsymbol{\omega} = (\omega^{(e)})_{e\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$ are pre-determined weights associated with environments $\mathcal{E}$ satisfying $\sum_{e\in\mathcal{E}} \omega^{(e)} = 1$ and $\omega^{(e)} > 0$ for any $e \in \mathcal{E}$. Some typical choices of $\boldsymbol{\omega}$ can be $\omega^{(e)} = 1/|\mathcal{E}|$ or $\omega^{(e)} = n^{(e)}/(\sum_{e'\in\mathcal{E}} n^{(e')})$. The word "*focused*" emphasizes that we only penalize the gradient in the direction of non-zero coordinates due to the statement (3.1). See Fan & Liao (2014) for a similar idea for dealing with endogeneity.

We first analyze what it implies when $\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega})$ attains its global minima. Since S^* is CE-invariant across environments $\mathcal{E}$ according to our assumption, we have $\mathsf{J}(\boldsymbol{\beta}^*;\omega) = 0$ and this implies the global minima of the loss function $\mathsf{J}(\boldsymbol{\beta};\omega)$ is 0. Moreover, for a fixed $\widetilde{\boldsymbol{\beta}}$ with support set $\widetilde{S}$, it is easy to see that

$$\begin{aligned}\mathsf{J}(\widetilde{\boldsymbol{\beta}};\boldsymbol{\omega}) = 0 \quad &\Leftrightarrow \quad \forall e \in \mathcal{E},\ \nabla_{\widetilde{S}}\mathsf{R}^{(e)}(\widetilde{\boldsymbol{\beta}}) = 0 \\ &\Leftrightarrow \quad \forall e \in \mathcal{E},\ \widetilde{\boldsymbol{\beta}} \in \underset{\mathrm{supp}(\boldsymbol{\beta})\subseteq\widetilde{S}}{\mathrm{argmin}}\ \mathsf{R}^{(e)}(\boldsymbol{\beta}).\end{aligned}$$

Therefore, though $\mathsf{J}(\widetilde{\boldsymbol{\beta}};\omega) = 0$ (attaining its global minima) does not imply that the selected variables set $\widetilde{S}$ is *CE-invariant* across environments $\mathcal{E}$, we argue that $\mathsf{J}(\widetilde{\boldsymbol{\beta}};\boldsymbol{\omega}) = 0$ indeed indicates that the select variable set $\widetilde{S}$ is *LLS-invariant* across environments $\mathcal{E}$, which can be defined as follows.

Definition 3.1 (LLS-invariant Set)**.** *We say a variable set $S \subseteq [p]$ is **linear least squares invariant** (LLS-invariant) across environments $\mathcal{E}$ if there exists some $\boldsymbol{\beta}$ with support set S such that*

$$\forall e \in \mathcal{E}, \qquad \boldsymbol{\beta} \in \underset{\mathrm{supp}(\boldsymbol{\beta})\subseteq S}{\mathrm{argmin}}\ \mathbb{E}\left[|y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}|^2\right]. \tag{3.3}$$

LLS-invariance is weaker than CE-invariance because the CE-invariance of $S^\star$ implies the LLS-invariance of $S^\star$ while the converse is false. In this paper, we only leverage linear information to facilitate LLS-invariance on the solution we find. That is what the word "*linear*" in our regularizer's name emphasizes. According to the above discussion of the relaxation in (a), we can incorporate more hand-crafted nonlinear features other than linear feature x_j in the *focused invariance regularizer* to somewhat strengthen the degree of invariance. For example, consider the following enhanced regularizer that includes another marginal nonlinear feature $h(\cdot)$ in,

$$\mathsf{J}_h(\boldsymbol{\beta};\boldsymbol{\omega}) = \sum_{j=1}^p \mathbb{1}\{\beta_j \neq 0\} \sum_{e\in\mathcal{E}} \omega^{(e)} \left\{\left|\mathbb{E}[\widehat{\varepsilon}_{\boldsymbol{\beta}}^{(e)} \boldsymbol{x}_j^{(e)}]\right|^2 + \left|\mathbb{E}[\widehat{\varepsilon}_{\boldsymbol{\beta}}^{(e)} h(\boldsymbol{x}_j^{(e)})]\right|^2\right\}, \tag{3.4}$$

where $\widehat{\varepsilon}_{\boldsymbol{\beta}}^{(e)} = y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}$. Specifically, we can take $h(u) = u^2$ or $h(u) = \cos(u)$.

Though LLS-invariance is weaker than CE-invariance, these two types of invariance are equivalent under some specific models. For example, suppose $\mu_{\boldsymbol{x},y}^{(e)}$ are all joint Gaussian distribution for any $e \in \mathcal{E}$. In that case, it is easy to verify that if a set S is LLS-invariant, then S is also CE-invariant using the fact that uncorrelatedness implies independence for joint Gaussian random variables. In this case, the enhanced regularizer (3.4) would have no advantages over the most simple linear one (3.2).

3.2 Our Approach: EILLS

Given data $\{\mathcal{D}^{(e)}\}_{e\in\mathcal{E}} = \{\{(\boldsymbol{x}_i^{(e)}, y_i^{(e)})\}_{i=1}^{n^{(e)}}\}_{e\in\mathcal{E}}$ from heterogeneous environments, the empirical-level *focused linear invariance regularizer* with weights $\boldsymbol{\omega}$ can be written as

$$\widehat{\mathsf{J}}(\boldsymbol{\beta};\boldsymbol{\omega}) = \sum_{j=1}^p \mathbb{1}\{\beta_j \neq 0\} \sum_{e\in\mathcal{E}} \omega^{(e)} \left|\widehat{\mathbb{E}}[x_j^{(e)}(y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)})]\right|^2. \tag{3.5}$$

Recall that $\widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta})$ defined in (2.5) is the empirical L_2 loss in environment in $e \in \mathcal{E}$. We also define the following *pooled empirical L_2 loss* over all the environments: for $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\widehat{\mathsf{R}}(\boldsymbol{\beta};\boldsymbol{\omega}) = \sum_{e\in\mathcal{E}} \omega^{(e)} \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) = \sum_{e\in\mathcal{E}} \frac{\omega^{(e)}}{n^{(e)}} \sum_{i=1}^{n^{(e)}} \{y_i^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}_i^{(e)}\}^2. \tag{3.6}$$

The *environment invariant linear least squares (EILLS)* estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$ is defined by minimizing the following objective function:

$$\widehat{\mathsf{Q}}(\boldsymbol{\beta};\gamma,\boldsymbol{\omega}) = \widehat{\mathsf{R}}(\boldsymbol{\beta};\boldsymbol{\omega}) + \gamma\widehat{\mathsf{J}}(\boldsymbol{\beta};\boldsymbol{\omega}), \tag{3.7}$$

which is a linear combination of the pooled empirical L_2 loss $\widehat{\mathsf{R}}(\boldsymbol{\beta};\boldsymbol{\omega})$ and focused linear invariance regularizer $\widehat{\mathsf{J}}(\boldsymbol{\beta};\boldsymbol{\omega})$ with weights $(1,\gamma)$ for some given hyper-parameter $\gamma \in \mathbb{R}^+$. We also define the population analogs of the EILLS objective function as follows:

$$\mathsf{Q}(\boldsymbol{\beta};\gamma,\boldsymbol{\omega}) = \mathsf{R}(\boldsymbol{\beta};\boldsymbol{\omega}) + \gamma\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega}) \quad \text{with} \quad \mathsf{R}(\boldsymbol{\beta};\boldsymbol{\omega}) = \sum_{e\in\mathcal{E}} \omega^{(e)}\mathbb{E}\left[|y^{(e)} - \boldsymbol{x}^\top \boldsymbol{x}^{(e)}|^2\right]. \tag{3.8}$$

The focused regularizer $\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega})$ can screen out all linear spurious variables when γ is large enough. This can lead to a low-dimensional problem and be sufficient for many applications. However, some linear exogenous variables that do not contribute to explaining y can still survive after the screening. They can be eliminated further by introducing an ℓ_0 penalty. This leads us to considering the ℓ_0 regularized EILLS estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{L}}$ that minimizes the following objective:

$$\widehat{\mathsf{L}}(\boldsymbol{\beta};\lambda,\gamma,\boldsymbol{\omega}) = \widehat{\mathsf{Q}}(\boldsymbol{\beta};\gamma,\boldsymbol{\omega}) + \lambda\|\boldsymbol{\beta}\|_0 = \widehat{\mathsf{R}}(\boldsymbol{\beta};\boldsymbol{\omega}) + \gamma\widehat{\mathsf{J}}(\boldsymbol{\beta};\boldsymbol{\omega}) + \lambda\|\boldsymbol{\beta}\|_0 \tag{3.9}$$

with given hyper-parameter λ. This helps reduce variables that are uncorrelated to residuals and y.

At the methodology level, our method has a close connection to the FGMM estimator in Fan & Liao (2014) and the invariant risk minimization (Arjovsky et al., 2019) framework; see detailed connections and differences in Appendix D.

Practical Concerns The current estimator has some combinatorial nature in optimization of (3.5) and involves an additional hyper-parameter γ. For the first concern, we show that one can use the Gumbel (Jang et al., 2016) trick introduced by a follow-up work Gu et al. (2024) to make variants of gradient descent continue to work in practice; see how to adopt Gumbel approximation to transform (3.7) to a continuous optimization program and some simulation studies when $p = 70$ in Appendix D.6. We also provide some guidance on how to tune the hyper-parameter γ in practice; see Appendix D.7.

4 Theory

To simplify the presentation, we consider the case of balanced data with equal weights, that is, $n^{(e)} \equiv n$ and $\omega^{(e)} \equiv 1/|\mathcal{E}|$, and defer the results of varying $(n^{(e)}, \omega^{(e)})$ to Appendix A. Define the pooled covariance matrix $\boldsymbol{\Sigma} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \boldsymbol{\Sigma}^{(e)}$. We first impose some regularity conditions for theoretical analysis.

Condition 4.1. *For each $e \in \mathcal{E}$, $(\boldsymbol{x}_1^{(e)}, y_1^{(e)}), \ldots, (\boldsymbol{x}_n^{(e)}, y_n^{(e)})$ are i.i.d. copies of $(\boldsymbol{x}^{(e)}, y^{(e)}) \sim \mu^{(e)}$, where $\mu^{(e)}$ belongs to $\mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$ for some σ^2. The data from different environments are also independent. We set $\omega^{(e)} \equiv 1/|\mathcal{E}|$.*

Condition 4.2. *There exists some universal constants $\kappa_L \in (0, 1]$ and $\kappa_U \in [1, \infty)$ such that*

$$\forall e \in \mathcal{E}, \qquad \kappa_L \boldsymbol{I}_p \preceq \boldsymbol{\Sigma}^{(e)} \preceq \kappa_U \boldsymbol{I}_p. \tag{4.1}$$

Condition 4.3. *There exists some universal constant $\sigma_x \in [1, \infty)$ such that*

$$\forall e \in \mathcal{E}, \boldsymbol{v} \in \mathbb{R}^p, \qquad \mathbb{E}\left[\exp\left\{\boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}^{(e)}\right\}\right] \le \exp\left(\frac{\sigma_x^2}{2}\cdot\|\boldsymbol{v}\|_2^2\right). \tag{4.2}$$

Condition 4.4. *There exists some universal constant $\sigma_\varepsilon \in \mathbb{R}^+$ such that,*

$$\forall e \in \mathcal{E}, \lambda \in \mathbb{R}, \qquad \mathbb{E}[e^{\lambda\varepsilon^{(e)}}] \le e^{\frac{1}{2}\lambda^2\sigma_\varepsilon^2}. \tag{4.3}$$

Condition 4.1 is the basic setup of our multi-environment linear regression described in Section 2.1. Condition 4.2–4.4 are standard in high-dimensional linear regression analysis. We focus on the centered covariate case that $\mathbb{E}[\boldsymbol{x}^{(e)}] = \mathbf{0}$, and it can be easily generalized to the non-centered counterpart. The lower bound on the smallest eigenvalue, κ_L, is to establish non-asymptotic error bounds on ℓ_2 norm $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. To simplify expressions in our results, we set $\kappa_U \wedge \sigma_x \geq 1$ and $\kappa_L \leq 1$, thus avoiding repeated use of $(\kappa_U \vee 1)$, $(\sigma_x \vee 1)$, and $(\kappa_L^{-1} \vee 1)$. We adopt simple sub-Gaussian conditions to convey our main message. It is possible to relax the above conditions and derive a similar result. For example, one can replace the sub-Gaussian condition of the noise $\varepsilon^{(e)}$ with the sub-Weibull condition (Vladimirova et al., 2020). One can also substitute the joint sub-Gaussian condition of the covariate with the marginal sub-Weibull condition and show that the EILLS objective (3.7) with folded concave penalty function introduced by Fan & Li (2001) has certain oracle property.

4.1 Pooled Linear Spurious Variables and the Bias of Pooled Least Squares

We first define *pooled linear spurious variables* and demonstrate that the vanilla pooled least squares method is an inconsistent estimator in the presence of pooled linear spurious variables.

Definition 4.1 (Pooled Linear Spurious Variables)**.** *We let G be the index set of all pooled linear spurious variables in environments $\mathcal{E}$ concerning the uniform weights $\omega^{(e)} \equiv 1/|\mathcal{E}|$, that is, $G = \{j \in [p] : \sum_{e\in\mathcal{E}} \mathbb{E}[x_j^{(e)}\varepsilon^{(e)}] \neq 0\}$. We say x_j is a pooled linear spurious variable if $j \in G$.*

The following proposition characterizes the properties of pooled least squares.

Proposition 4.1 (Properties of Pooled Least Squares)**.** *Assume Conditions 4.1–4.4 hold. Then, there exists some $\bar{\boldsymbol{\beta}}^{\mathsf{R}} \in \mathbb{R}^p$ satisfying $\frac{1}{\kappa_U}\|\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\|_2 \leq \|\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*\|_2 \leq \frac{1}{\kappa_L}\|\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\|_2$ such that, for any $\boldsymbol{\beta} \in \mathbb{R}^p$,*

$$\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) = \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2^2. \tag{4.4}$$

Moreover, there exist universal constants c_1 and c_2 such that if $n \cdot |\mathcal{E}| \geq c_1\sigma_x^4(p+t)$, then the pooled least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{R}}$ minimizing (3.6) satisfies

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathsf{R}} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2 \leq c_2\sigma_x\left(\sigma_\varepsilon + \sigma_x\|\boldsymbol{\Sigma}^{1/2}(\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*)\|_2\right)\sqrt{\frac{p+t}{n\cdot|\mathcal{E}|}}$$

with probability at least $1 - 2e^{-t}$.

From Proposition 4.1, we observe that $\delta_{\mathsf{b}} = \|\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*\|_2$, which can be interpreted as the bias of the pooled least squares, satisfies $\delta_{\mathsf{b}} \asymp \left\||\mathcal{E}|^{-1}\sum_{e\in\mathcal{E}}\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right\|_2$. Proposition 4.1 thus indicates that, for large enough n,

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{R}} - \boldsymbol{\beta}^*\|_2 \asymp \left\|\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right\|_2.$$

Therefore, the pooled least squares can be a fair estimate of $\boldsymbol{\beta}^*$ only when the biases from each environment happen to cancel: $\sum_{e\in\mathcal{E}}\mathbb{E}[\varepsilon^{(e)}x_j^{(e)}] = o(|\mathcal{E}|)$ for any $j \in \mathcal{E}$. Moreover, the strong convexity with respect to $\bar{\boldsymbol{\beta}}^{\mathsf{R}}$ (4.4) suggests that this issue persists for penalized least squares.

4.2 Local Strong Convexity for Population Loss

We first examine the population EILLS loss (3.8) when $G \neq \emptyset$. Specifically, we show that with a large enough γ, the population EILLS loss satisfies certain strong convexity with respect to $\boldsymbol{\beta}^*$ under the following identification condition.

Condition 4.5 (Identification)**.** *For any $S \subseteq [p]$ satisfying $S \cap G \neq \emptyset$, there exists some $e, e' \in \mathcal{E}$ such that $\boldsymbol{\beta}^{(e,S)} \neq \boldsymbol{\beta}^{(e',S)}$, where $\boldsymbol{\beta}^{(e,S)}$ is defined in (2.6).*

Remark 4.1 (Near Minimal Identification Condition)**.** *It is worth noticing that the above identification condition is minimal if only linear information is considered. To see this, consider the case where $(\boldsymbol{x}^{(e)}, y^{(e)})$ are all multivariate normal distributions, a violation of Condition 4.5 implies that there exist some $\widetilde{\boldsymbol{\beta}}$ with support set $\widetilde{S}$ containing variables outside $S^\star$ such that*

$$\forall e \in \mathcal{E}, \qquad \mathbb{E}[y^{(e)} | \boldsymbol{x}_{\widetilde{S}}^{(e)}] \equiv (\widetilde{\boldsymbol{\beta}}_{\widetilde{S}})^\top \boldsymbol{x}_{\widetilde{S}}^{(e)}.$$

In this case, it is impossible to determine which one among S^* and $\widetilde{S}$ is the true important variable set because they are all CE-invariant across $\mathcal{E}$ and $\widetilde{S} \setminus S^* \neq \emptyset$.

Theorem 4.2 (Strong Convexity with respect to $\boldsymbol{\beta}^*$)**.** *Assume Conditions 4.1–4.2 and 4.5 hold. Then $\boldsymbol{\beta}^*$ is the unique minimizer of $\mathsf{Q}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega})$ for large enough γ: for any $\epsilon \in (0,1)$ and any $\gamma \geq \epsilon^{-1}\gamma^*$ with*

$$\gamma^* = (\kappa_L)^{-3} \sup_{S: S \cap G \neq \emptyset} \left(\mathsf{b}_S / \bar{\mathsf{d}}_S\right), \tag{4.5}$$

where $\mathsf{b}_S = \|\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathbb{E}[\varepsilon^{(e)} \boldsymbol{x}_S^{(e)}]\|_2^2$ and $\bar{\mathsf{d}}_S = \sum_{e\in\mathcal{E}} \frac{1}{|\mathcal{E}|} \|\boldsymbol{\beta}^{(e,S)} - \bar{\boldsymbol{\beta}}^{(S)}\|_2^2$ with $\bar{\boldsymbol{\beta}}^{(S)} = \frac{1}{|\mathcal{E}|}\sum_{e'\in\mathcal{E}} \boldsymbol{\beta}^{(e',S)}$, we have

$$\mathsf{Q}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega}) - \mathsf{Q}(\boldsymbol{\beta}^*; \gamma, \boldsymbol{\omega}) \geq (1-\epsilon)\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\|_2^2 + \kappa_L^2(\gamma - \epsilon^{-1}\gamma^*)\bar{\mathsf{d}}_{\mathrm{supp}(\boldsymbol{\beta})}. \tag{4.6}$$

Note $\bar{\mathsf{d}}_S > 0$ for any $S \cap G \neq \emptyset$ under Condition 4.5. Theorem 4.2 provides a generic condition under which $\boldsymbol{\beta}^*$ is the global optimal of the population-level EILLS objective (3.8) with a non-asymptotic critical threshold for γ. Moreover, a detailed characterization of global optimality (4.6), which is slightly stronger than strong convexity, is also established. The strong convexity with respect to the target function $\boldsymbol{\beta}^*$ lays the foundation to derive faster convergence rate to $\boldsymbol{\beta}^*$ (van der Vaart & Wellner, 1996; Wainwright, 2019).

Remark 4.2 (Interpretation of the Quantities b_S, $\bar{\mathsf{d}}_S$)**.** *Observe that when $S \supseteq S^*$, the bias of the least squares solution in one environment $e \in \mathcal{E}$ is $\boldsymbol{\Delta}^{(e)} = \boldsymbol{\beta}^{(e,S)} - \boldsymbol{\beta}^*$. Here we refer to b_S as bias mean because b_S is the bias of running least squares on X_S using all the data when $S \supseteq S^*$, that is, $\|\widehat{\boldsymbol{\beta}}_{\mathsf{R},S} - \boldsymbol{\beta}^*\|_2^2 \asymp \mathsf{b}_S$ by Proposition 4.1. At the same time, we have $\bar{\mathsf{d}}_S = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \|\boldsymbol{\Delta}^{(e)} - (|\mathcal{E}|^{-1}\sum_{e'\in\mathcal{E}} \boldsymbol{\Delta}^{(e')})\|_2^2$ provided $S \supseteq S^*$. Thus, the quantity $\bar{\mathsf{d}}_S$ can be interpreted as the variance of bias since it measures the variations of the biases $\boldsymbol{\Delta}^{(e)}$ among different environments.*

Remark 4.3 (Interpretation of the Critical Threshold γ^*)**.** *Theorem 4.2 implies that the global optimality of $\boldsymbol{\beta}^*$ is guaranteed when $\gamma > \gamma^*$. γ^* is the ratio of bias mean to bias variance and will affect both the forthcoming estimation error and variable selection property. As an intuitive example, for the previous cow/camel thought experiment, a reasonable value for γ^* is expected when the fractions of cows on grass and camels on sand are both 90% in $e = 0$ and 60% in $e = 1$. Similarly, a moderate γ^* is anticipated when fractions are 51% and 53%. However, with fractions 90% and 89%, γ^* is significantly larger, necessitating more data for both accurate estimation and variable selection.*

Remark 4.4 (Interpretation of Small γ)**.** *In Theorem 4.2, we show that the invariant parameter $\beta^\star$ will uniquely minimize the EILLS objective when γ is larger than some threshold $\gamma^\star$. Yet, it is unclear what is the regularization effect of $\gamma \mathsf{J}$ when γ is small. We provide some intuitions on the impact of the regularizer when γ is small in Appendix D.8.*

We use the toy example below to demonstrate (1) when Condition 4.5 holds, and (2) how γ^* scales in a concrete model, and leave more examples and discussions in Appendix A.4.

Example 4.1. *Consider the following two-environment linear SCMs with $p = 2$:*

$$
\begin{aligned}
x_1^{(e)} &\leftarrow \sqrt{0.5} \cdot \varepsilon_1 \\
y^{(e)} &\leftarrow 1 \cdot x_1^{(e)} + \sqrt{0.5} \cdot \varepsilon_0 \\
x_2^{(e)} &\leftarrow s^{(e)} \cdot y^{(e)} + \varepsilon_2.
\end{aligned}
$$

Here $\varepsilon_0, \varepsilon_1, \varepsilon_2$ are independent, standard normally distributed exogenous variables. The linear model is $y^{(e)} = (\boldsymbol{\beta}^*)^\top \boldsymbol{x}^{(e)} + \varepsilon^{(e)}$ with $\boldsymbol{\beta}^* = (1, 0)$ and $\varepsilon^{(e)} = \sqrt{0.5}\varepsilon_0$. Note that $\mathbb{E}[\varepsilon^{(e)} x_2^{(e)}] = 0.5 s^{(e)}$. When $\omega^{(e)} \equiv 1/2$, the variable x_2 will be a pooled linear spurious variable if $s^{(1)} + s^{(2)} \neq 0$.

We focus on the case where the pooled least squares is not consistent, i.e., $s^{(1)} + s^{(2)} \neq 0$. In the well-conditioned regime where $|s^{(1)}| + |s^{(2)}| = O(1)$, by some calculations, we have

$$
\sqrt{\frac{\mathsf{b}_{\{2\}}}{\bar{\mathsf{d}}_{\{2\}}}} \asymp \frac{|s^{(1)} + s^{(2)}|}{|s^{(2)} - s^{(1)}|} \cdot \frac{1}{|1 - s^{(1)} s^{(2)}|} \qquad \text{and} \qquad \sqrt{\frac{\mathsf{b}_{\{1,2\}}}{\bar{\mathsf{d}}_{\{1,2\}}}} \asymp \frac{|s^{(1)} + s^{(2)}|}{|s^{(2)} - s^{(1)}|}.
$$

First, the Condition 4.5 holds if $s^{(1)} \neq s^{(2)}$ and $s^{(1)} s^{(2)} \neq 1$. The first inequality is easy to understand, without which the underlying distributions in the two environments are identical. The second inequality is strange at first glance. However, we have $\mathbb{E}[y^{(e)} | x_2^{(e)}] \equiv \frac{s}{s^2+1} x_2^{(e)}$ when $s^{(1)} = s = 1/s^{(2)}$. In this case, it is impossible to identify which of $\{1\}$ and $\{2\}$ are the true important variable set because all the sets are CE-invariant across $\mathcal{E}$. This also demonstrates the necessity of taking the supremum over all sets S with $S \cap G \neq \emptyset$ in (4.5).

When $s^{(1)} s^{(2)}$ is away from 1, the critical threshold γ^* satisfies $(\gamma^*)^{1/2} \asymp |s^{(1)} + s^{(2)}| / |s^{(2)} - s^{(1)}|$, where the numerator quantifies the strength of spuriousness, and the denominator quantifies the strength of heterogeneity. Hence, a constant-level γ can be adopted when the strength of heterogeneity is of the same order as the strength of spuriousness. □

4.3 Statistical Analysis of the EILLS Estimator in the Low-dimensional Regime

Our statistical analysis of the EILLS estimator focuses on the regime where it is possible to identify $\boldsymbol{\beta}^*$, i.e., Condition 4.5 holds, and when our choice of γ satisfies $\gamma \geq 3\gamma^* \vee 1$, where γ^* is the quantity defined in Theorem 4.2. We let $\gamma \geq 1$ to simplify the presentation. We are now ready to provide a statistical analysis of the EILLS estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$ minimizing (3.7). The first result is about the sure screening with false positive control.

Theorem 4.3 (Non-asymptotic Variable Selection Property)**.** *Define*

$$
\mathsf{s}_+ = \min_{j \in S^*} |\beta_j^*|^2 \qquad \textit{and} \qquad \mathsf{s}_- = \min_{S \subseteq [p], S \cap G \neq \emptyset} \bar{\mathsf{d}}_S \tag{4.7}
$$

Suppose Conditions 4.1–4.5 hold, and we choose $\gamma \geq 3\gamma^* \vee 1$ where γ^* is defined in Theorem 4.2. There exists some universal constants c_1–c_2 that only depends on $(\kappa_U, \sigma_x, \sigma_\varepsilon)$ such that for any $t > 0$, if $n \geq c_1(\gamma/\kappa_L)(p + \log(|\mathcal{E}|) + t)\{\mathsf{s}_+^{-0.5} + \mathsf{s}_+^{-1} + (\gamma \kappa_L \mathsf{s}_-)^{-0.5}\}$, and $n \cdot |\mathcal{E}| \geq c_2(\gamma/\kappa_L)^2(p + t)\{\mathsf{s}_+^{-1} + (\gamma\kappa_L \mathsf{s}_-)^{-1} + 1\}$, then the EILLS estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$ minimizing (3.7) satisfies

$$
\mathbb{P}\left[S^* \subseteq \mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}) \subseteq G^c\right] \geq 1 - 7e^{-t}. \tag{4.8}
$$

Equation (4.8) reveals that all endogenous variables are screened out when γ is sufficiently large. When the choice of γ and the curvature κ_L are both of constant order, Theorem 4.3 implies that with high probability, the EILLS estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$ can eliminate all the pooled linear spurious variables while keeping all the important variables, i.e., (4.8) holds, if

$$
\frac{n}{p + \log(|\mathcal{E}|)} \gg \mathsf{s}_+^{-1} + \mathsf{s}_-^{-1/2} \left(1 \vee \frac{\mathsf{s}_-^{-1/2}}{|\mathcal{E}|}\right). \tag{4.9}
$$

Here, the quantities s_+ and s_- defined in (4.7) can be interpreted as the signal of true important variables and the signal of heterogeneity, respectively. One can expect more data to differentiate whether it is a signal or noise when one of s_- and s_+ is small.

When the variable selection property (4.8) is satisfied, $\widehat{S}$ does not contain any pooled linear spurious variables while maintaining all the important variables. In this case, one can provide a good estimate of $\boldsymbol{\beta}^*$ by running another least squares constrained on $\widehat{S} = \mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{Q}})$ without regularization. In this case, there is no bias anymore. It then follows from Proposition 4.1 that, with high probability, we have the ℓ_2-error $\{|G^c|/(n \cdot |\mathcal{E}|)\}^{1/2}$ given $\widehat{S} \subseteq G^c$.

When some weak spurious variables exist such that the corresponding absolute value of $\bar{\mathsf{d}}_S$ is small, the EILLS estimator requires a large n to eliminate all these spurious variables. To see this, suppose there exists a weak spurious variable x_j such that $|\mathbb{E}[x_j^{(e)} \varepsilon^{(e)}]| \le \epsilon$ for all the $e \in \mathcal{E}$. Consider the set $\widetilde{S} = S^* \cup \{j\}$, it is easy to verify that $\mathsf{b}_{\widetilde{S}} \le \varepsilon^2$ and $\bar{\mathsf{d}}_{\widetilde{S}} \le \varepsilon^2$. In this case, we require $n \gg n_{*,\mathtt{sel}} \asymp p\epsilon^{-1}(\epsilon^{-1}/|\mathcal{E}|+1)$ to eliminate variable x_j by (4.9). The next theorem claims that a small ℓ_2 estimation error $\|\widehat{\boldsymbol{\beta}}_{\mathsf{Q}} - \boldsymbol{\beta}^*\|_2$ can be obtained in the regime where $p \ll n \ll n_{*,\mathtt{sel}}$ regardless of whether EILLS selects the correct variable.

Theorem 4.4 (Non-asymptotic ℓ_2 Error Bound)**.** *Assume Conditions 4.1–4.5 hold, and we choose* $\gamma \ge 3\gamma^* \vee 1$ *where* γ^* *is defined in Theorem 4.2. There exists some universal constants* c_1*–*c_4 *that only depend on* (κ_U, σ_x) *such that for any* $t > 0$*, if* $n \ge p + \log(2|\mathcal{E}|) + t$ *and* $n \cdot |\mathcal{E}| \ge c_1(\gamma/\kappa_L)(p+t)$*, then* $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$ *minimizing* (3.7) *satisfies*

$$\frac{\|\widehat{\boldsymbol{\beta}}_{\mathsf{Q}} - \boldsymbol{\beta}^*\|_2}{\sigma_\varepsilon(\gamma/\kappa_L)} \le c_2 \left(\sqrt{\frac{p+t}{n \cdot |\mathcal{E}|}} + \frac{p + \log(|\mathcal{E}|) + t}{n}\right) + c_3 \frac{\sqrt{|S^*|}}{n} \cdot \frac{\log(|\mathcal{E}||S^*|) + t}{\min_{j \in S^*} |\beta_j^*|} \tag{4.10}$$

with probability at least $1 - 7e^{-t}$*. Moreover, when the additional conditions in Theorem 4.3 hold, then*

$$\frac{\|\widehat{\boldsymbol{\beta}}_{\mathsf{Q}} - \boldsymbol{\beta}^*\|_2}{\sigma_\varepsilon(\gamma/\kappa_L)} \le c_2 \left(\sqrt{\frac{|G^c| + t}{n \cdot |\mathcal{E}|}} + \frac{|G^c| + \log(|\mathcal{E}|) + t}{n}\right) \tag{4.11}$$

occurs with probability at least $1 - 14e^{-t}$*.*

In the well-conditioned regime where $\min_{j \in S^*} |\beta_j^*| \gtrsim p^{-1/2}$, $\kappa_L \gtrsim 1$, and $\gamma^* \asymp 1$, one can adopt a constant-level hyper-parameter γ such that

$$\|\widehat{\boldsymbol{\beta}}_{\mathsf{Q}} - \boldsymbol{\beta}^*\|_2 \lesssim \sqrt{\frac{p}{n \cdot |\mathcal{E}|}} + \frac{p}{n}$$

with high probability provided $n \gtrsim p$. When the first term dominates ($|\mathcal{E}|$ is not too big), the EILLS estimator achieves an optimal linear regression rate. This implies that the EILLS estimator can estimate $\boldsymbol{\beta}^*$ well in ℓ_2 error when there are some weak spurious variables and the number of data points is not large enough to eliminate all the variables in G.

One technical novelty behind Theorems 4.3–4.4 and Theorem 4.5 later is to apply the localization argument in the analysis of the invariance regularizer to get a faster rate. To see this, we have $\|\widehat{\boldsymbol{\beta}}_{\mathsf{Q}} - \boldsymbol{\beta}^*\|_2^2 \asymp \delta_{n,\mathsf{Q}}^2 \asymp (n \cdot |\mathcal{E}|)^{-1} + n^{-2}$ when $p = O(1)$. The first term is faster than $(n \cdot |\mathcal{E}|)^{-1/2}$ which directly applies uniform concentration, and the second term is faster than $n^{-1} \asymp \sup_{\|\boldsymbol{\beta}\|_2 \le 1} |\mathsf{J}(\boldsymbol{\beta}) - \mathbb{E}[\widehat{\mathsf{J}}(\boldsymbol{\beta})]|$, the (uniform) bias of the invariance regularizer. We use Lemma C.4, a novel one-side instance-dependent error bound for $\mathsf{J}(\boldsymbol{\beta}) - \widehat{\mathsf{J}}(\boldsymbol{\beta}) + \widehat{\mathsf{J}}(\boldsymbol{\beta}^*) - \mathsf{J}(\boldsymbol{\beta}^*)$ to obtain such a faster rate; see Appendix C.3.

4.4 Variable Selection Consistency in the High-dimensional Regime

In the high-dimensional regime, we further define $s^* = |S^*|$, $\beta_{\min} = \min_{j \in S^*} |\beta_j^*|$. We need a condition asserting that the sample size n should be large enough for the given hyper-parameter γ.

Condition 4.6. *Suppose that* $\log(|\mathcal{E}|) \le C \log p$ *and that*
(1) $n \ge c_1(\gamma/\kappa_L)\left\{(s^* + \beta_{\min}^{-2})\log p + (\kappa_L\beta_{\min})^{-1}\sqrt{(s^*+\log p)s^*\log p}\right\}$
(2) $n \cdot |\mathcal{E}| \ge c_2(\gamma/\kappa_L)^2(s^*\log p)\{1 + 1/(\kappa_L\beta_{\min}^2)\}$,
Here c_1*–*c_2 *are positive universal constants that depend only on* $(C, \sigma_x, \kappa_U, \sigma_\varepsilon)$*.*

Theorem 4.5 (Variable Selection Consistency in High Dimensions)**.** *Assume Conditions 4.1–4.6 hold, and we choose* $\gamma \ge 3\gamma^* \vee 1$ *where* γ^* *is defined in Theorem 4.2. Suppose further that the choice of* λ *satisfies*

$$c_1\left\{(\gamma/\kappa_L)^2\frac{s^*\log p}{n\cdot|\mathcal{E}|} + \epsilon(n)\right\} \le \lambda \le c_2\kappa_L\beta_{\min}^2,$$

where $\epsilon(n) = (\gamma/\kappa_L)^2 s^\star(\log p)(s^* + \log p)/n^2 + (\gamma/\kappa_L)\log p\sqrt{n^{-3}(s^*+\log p)}$*, and* c_1, c_2 *are some universal positive constants only depends on* $(C, \kappa_U, \sigma_x, \sigma_\varepsilon)$*. Then the* ℓ_0 *regularized EILLS estimator* $\widehat{\boldsymbol{\beta}}_{\mathsf{L}}$ *minimizing (3.9) satisfies* $\mathbb{P}[\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) = S^*] \ge 1 - p^{-10}$*.*

We can see that $\epsilon(n)$ is negligible when $|\mathcal{E}| \asymp 1$ and $s^* + \log p = o(n)$ because

$$\epsilon(n)\Big/\left\{(\gamma/\kappa_L)^2\frac{s^*\log p}{n\cdot|\mathcal{E}|}\right\} \lesssim |\mathcal{E}|\left\{\frac{s^*+\log p}{n} + \left(\sqrt{\frac{\log p}{n}}\right)\right\} = o(1).$$

Therefore, in the regime where $\gamma, \kappa_L, |\mathcal{E}|$ are fixed while n, p, s^* may grow, the variable selection consistency can be achieved when

$$\frac{s^*\log p}{n} \ll \lambda \ll \beta_{\min}^2.$$

Recall that for standard ℓ_0 regularized least squares with sample size n, the variable selection consistency can be obtained when $n \gg (s^* + \beta_{\min}^{-2})\log p$ and λ satisfies $n^{-1}\log p \ll \lambda \ll \beta_{\min}^2$ (Zhang & Zhang, 2012). Compared to the standard ℓ_0 regularized least squares, the EILLS estimator needs $n \gg s^*\beta_{\min}^{-2}\log p$ in the fixed number of environments setting.

Another question is how much can the ℓ_0 regularized EILLS estimator benefit from growing $|\mathcal{E}|$? When $\gamma \asymp \kappa_L \asymp s^* \asymp 1$, Theorem 4.5 implies that though achieving variable selection consistency still needs $(1 + \beta_{\min}^{-2})\log p = o(n)$ due to Condition 4.6 (1), we can choose a wider range of λ. To be specific, in this case, the variable selection consistency can be achieved when λ satisfies

$$\left\{\frac{1}{|\mathcal{E}|} + \left(\frac{\log p}{n}\right)^{1/2}\right\}\frac{\log p}{n} \ll \lambda \ll \beta_{\min}^2.$$

Hence we can choose any $(n\cdot|\mathcal{E}|)^{-1}\log p \ll \lambda \ll \beta_{\min}^2$ in the regime $(\log p)|\mathcal{E}|^2 = o(n)$. This is the same with running ℓ_0 regularized least squares on total $n\cdot|\mathcal{E}|$ data when $\mathbb{E}[\varepsilon|\boldsymbol{x}] \equiv 0$.

With the variable selection consistency given in Theorem 4.5, we can then attain the optimal ℓ_2-rate by running the least squares on $\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}})$, as if assisted by an oracle.

5 An Illustration by Structural Causal Model

In this section, we use an example of a structural causal model (with potential interventions on covariates $\boldsymbol{x}$) to illustrate how different components of our proposed EILLS objective (3.7) and ℓ_0 regularized EILLS objective (3.9) contribute to either seizing the important variables S^* or eliminating pooled linear spurious variables and unrelated variables. Simulations further support intuitive claims[2].

[2] Scripts to reproduce the simulation result in this section can be found in the supplemental material, see also https://github.com/wmyw96/EILLS.

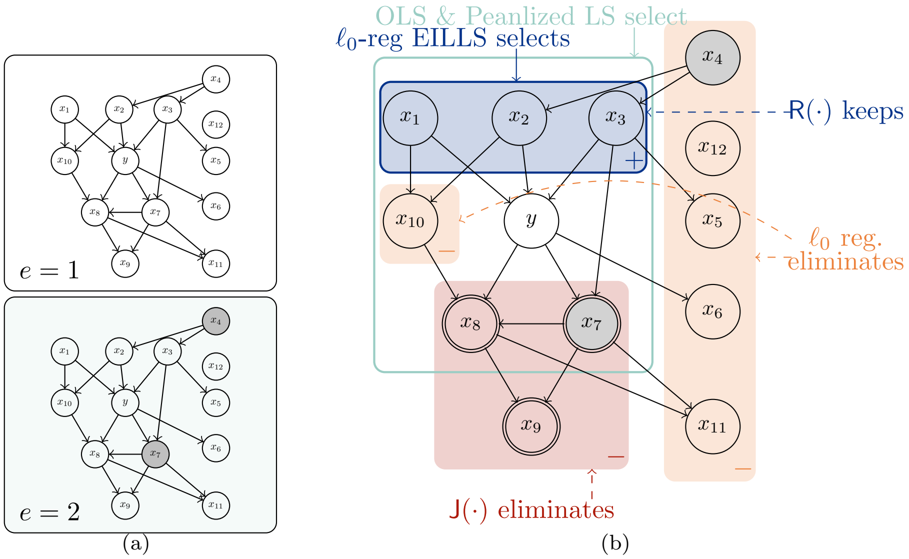


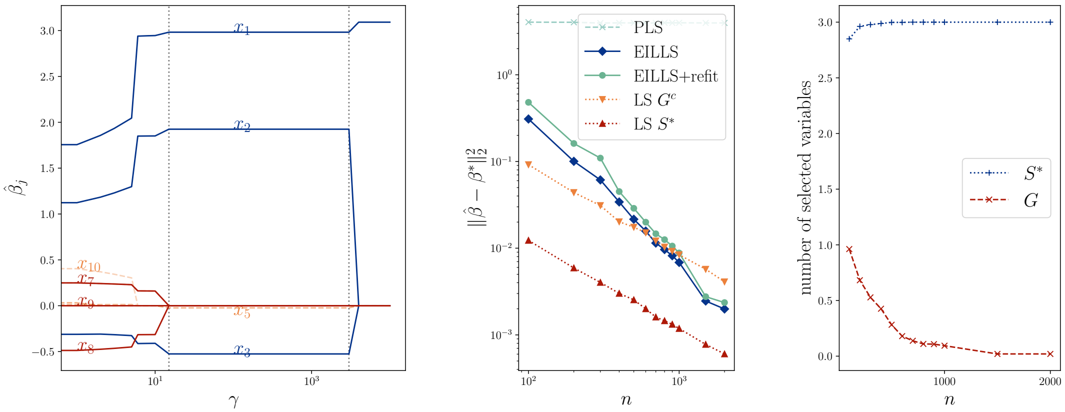
Figure 2: (a) An illustration of the two-environment model, the SCMs in the two environments share the same induced graph, which is also plotted in (b). The arrow from node x to node z indicates that x is the direct cause of z. (b) An illustration of how EILLS works under the two-environment model. The double-circled nodes represent the pooled linear spurious variables.

Model. Consider a two-environment model ($|\mathcal{E}| = 2$). The mechanism of $(\boldsymbol{x}, y)$ in each environment is characterized by an SCM. As shown in Fig. 2 (a), the two SCMs share the same direct cause relationship and most of the structural assignments, except that the mechanisms of x_4 and x_7 are different (gray shadowed); see the detailed structural assignments in Appendix D.4. In this case, the set of important variables S^* is $\{1, 2, 3\}$, corresponding to the set of direct causes of y. The true parameter is $\boldsymbol{\beta}^* = (3, 2, -0.5, 0, \ldots, 0)^\top$. The set of pooled linear spurious variables G is the subset of y's offspring. In this example, $G = \{7, 8, 9\}$ (denote as double circled node in Fig. 2 (b)).

How EILLS rescues. First, although vanilla pooled least squares, which minimizing (3.6), can asymptotically eliminate some uncorrelated variables (for example, $\widehat{\beta}_{\mathsf{R},12} = o_{\mathbb{P}}(1)$), it will asymptotically select some of the pooled linear spurious variables in G together with other variables related to these variables (for example, some of their ancestors and offsprings) according to (1.3). For example, it may asymptotically select variables $S_{\mathsf{R}} = \{1, 2, 3, 7, 8, 10\}$ as shown in the light blue rectangle. That is, $|\widehat{\beta}_{\mathsf{R},j} - \beta_j| = o_{\mathbb{P}}(1)$ with some $\beta_j \neq 0$ will hold for any $j \in S_{\mathsf{R}}$.

As for our EILLS estimator that minimizes (3.7), if the inclusion of pooled linear spurious variables leads to a shift of best linear predictor (Condition 4.5), then, with a large enough γ, our regularizer $\widehat{\mathsf{J}}$ will eliminate all pooled linear spurious variables in red shadows for large enough sample size. We add a red "$-$" in the red shadow to emphasize the regularizer $\widehat{\mathsf{J}}$'s role in eliminating those pooled linear spurious variables. Moreover, including the pooled L_2 risk $\widehat{\mathsf{R}}_{\mathcal{E}}$ will prevent our EILLS objective from collapsing to conservative solutions. Thus it will select all the important variables in blue shadows, in which we also use a blue "$+$" to underline the pooled L_2 risk $\widehat{\mathsf{R}}_{\mathcal{E}}$'s impact on keeping all the important variables. Finally, the objective (3.7) can only guarantee that $\widehat{\beta}_{\mathsf{Q},j} = o_{\mathbb{P}}(1)$ for unrelated variables j in orange shadows, the inclusion of ℓ_0 regularization completes the last step towards the target of variable selection consistency: with a properly chosen hyper-parameter λ, we can eliminate ($-$) these unrelated variables through the ℓ_0 penalty and only keep the important variables in blue rectangle.

Experimental Justification. We support the above intuition via simulations, in which the balanced data setting ($n^{(e)} \equiv n$, $\omega^{(e)} \equiv 1/2$) is adopted. Detailed implementation and experimental configurations are presented in Appendix D.4.



(a) (b) (c)

Figure 3: The simulation results for the model in Fig. 2 (a). (a) depicts how the estimated coefficients for the EILLS estimator vary across hyper-parameter γ in one trial when $n = 300$: we use blue and red solid lines to represent the corresponding coefficients for variables in S^* and G, respectively; and use orange dashed lines to represent the coefficients for other variables. The two gray vertical lines are $\gamma = 15$ and $\gamma = 3 \times 10^3$, respectively. (b) depicts how the average ℓ_2 errors (based on 500 replications, shown in log scale) for different estimators (marked with different shapes) change when n grows: 'LS S' is the estimator that runs least squares on $\boldsymbol{x}_S$ using all the data and 'PLS' is referred to 'LS $[p]$'. (c) depicts the average number of selected variables in S^* (+) and G (×) for the EILLS estimator over 500 replications.

Let us first see how EILLS works in practice by visualizing how the estimates of coefficients change over $\gamma \in [0, 10^4]$ in one trial. As shown in Fig. 3 (a), the pooled least squares running least squares on all the data ($\gamma = 0$) will lead to a biased solution, which significantly selects variables in S_{R}. Meanwhile, the EILLS estimator with proper regularization parameter $\gamma \in [15, 3 \times 10^3]$ selects variables $\{1, 2, 3, 5\}$: it screens out all the variables in G and keeps all the important variables in S^*. Fig. 3 (a) also demonstrates the necessity of incorporating L_2 risk: with a very large γ ($\gamma \geq 4 \times 10^3$) that the L_2 risk is relatively negligible compared with the invariance regularizer, the best (empirical) linear predictor on $\{1, 2, 3\}$ is not preferred than that on $\{1\}$ by the invariance regularizer.

Fig. 3 (b) and Fig. 3 (c) further support the above claims using the average performances over 500 replications, in which we use fixed $\gamma = 20$ for EILLS. As presented in Fig. 3 (b), the average square of ℓ_2 estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$ for pooled least squares estimator (× PLS) does not decrease (remains to be ≈ 4) as n increases, indicating that it converges to a biased solution. At the same time, the average $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$ for the EILLS estimator (◆) decays as n grows (is ≈ 0.1 when $n = 200$) and lies in between that for least squares on $\boldsymbol{x}_{G^c}$ (▼) and least squares on $\boldsymbol{x}_{S^*}$ (▲) when $n \geq 700$. This is expected to happen since the EILLS estimator can not screen out all the uncorrelated variables. We also report the ℓ_2 estimation error for the refitted EILLS estimator, which runs OLS on the variable set $\widehat{S}$ EILLS selects. It is interesting to see that its performance is slightly worse than the vanilla EILLS estimator; we provide a quantitative explanation in Appendix D.5.

The variable selection property for EILLS is further demonstrated in Fig. 3 (c), where the average number of selected variables in S^* (+) and G (×) is plotted across different n. The "+" curve keeps increasing and is almost 3 while the "×" curve decays and approaches 0, implying that the EILLS will select almost all the variables in S^* and screen out all the variables in G when n grows.

Comparison with Other Invariance Approaches. In addition, we compare our EILLS approach with other invariance approaches, including invariance causal prediction (ICP) (Peters et al., 2016), anchor regression (Anchor) (Rothenhäusler et al., 2021), and invariant risk minimization (IRM) (Arjovsky et al., 2019), using the data generating process above. The pooled least squares (PLS) method is also included for comparative purposes. We use EILLS with hyper-parameter $\gamma = 20$. For other invariance approaches, the invariance hyper-parameters are chosen in an oracle manner – we enumerate all the possible hyper-parameters and pick the one that minimizes the ℓ_2 prediction error $\|\bar{\boldsymbol{\Sigma}}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2$. The implementation details can be found

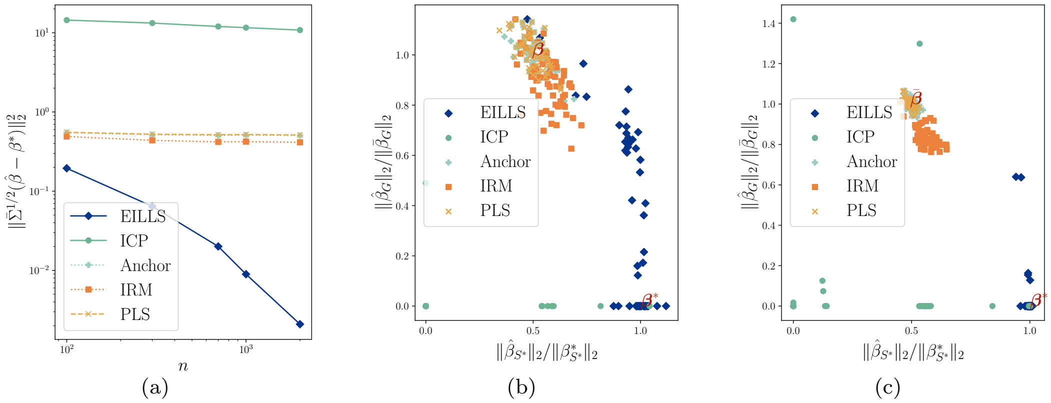


Figure 4: The simulation results for different methods using the data generated from the model in Fig. 2 (a). (a) depicts how the average ℓ_2 prediction errors $\|\bar{\mathbf{\Sigma}}^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2$ (based on 300 replications) for different invariance methods (marked with different shapes and colors) changes when n grows. (b) and (c) visualizes the solutions of different methods in 60 replications when $n = 100$ and $n = 1000$, respectively. The true parameter $\boldsymbol{\beta}^*$ and the population pooled least squares solution $\bar{\boldsymbol{\beta}}$ are also included using red for reference.

in the Supplemental Material.

The results are shown in Fig. 4 (a). It is apparent that among all the estimators evaluated, the EILLS estimator stands out as the only one capable of consistent estimation. We also provide graphical visualizations of the solutions found by different methods for $n = 100$ and $n = 1000$ in Fig. 4 (b) and (c), respectively. In these plots, each point (x, y) with marker m represents a solution $\widehat{\boldsymbol{\beta}}$ that the method m produces. Here, x denotes the relative ℓ_2 norm restricted to the true important variables set S^*, calculated as $\|\widehat{\boldsymbol{\beta}}_{S^*}\|_2/\|\boldsymbol{\beta}^*_{S^*}\|_2$; and y represents the relative ℓ_2 norm restricted to the pooled linear spurious variables set G, expressed as $\|\widehat{\boldsymbol{\beta}}_G\|_2/\|\bar{\boldsymbol{\beta}}_G\|_2$ with $\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{\beta}}^{([12])}$.

As depicted in Fig. 4 (b) and (c), the ICP method demonstrates a very conservative nature, failing to select variables in S^* even when $n = 1000$. This reveals its lack of guarantees in power even when the sample size is large enough. For other optimization-based methods, the solutions obtained by anchor regression are similar to those found by the pooled least squares. Although the IRM method showcases a slight divergence from PLS, with a tendency to push solutions toward the direction of "invariance", this effect remains marginal. In contrast, our EILLS method not only converges to the true parameter $\boldsymbol{\beta}^*$ when n is large but also demonstrates commendable performance when the sample size is moderately large ($n = 100$).

6 Discussion

In this paper, we consider the multi-environment linear regression model. We propose the *environment invariant linear least squares*, an optimization-based method applicable under the generic multi-environment linear regression model without additionally imposed structures. We provide a thorough statistical analysis of the proposed method. Specifically, it is possible to identify the true parameter under a near-minimal population-level condition Condition 4.5. Under such a condition, the EILLS estimator can obtain an optimal linear regression rate in the low-dimensional regime, and the ℓ_0 regularized EILLS estimator can achieve variable selection consistency in the high-dimensional regime.

One key theoretical takeaway from this paper in the invariance field is that a statistically efficient estimation of $\boldsymbol{\beta}^*$ is viable under a general, near-minimal identification condition related to the heterogeneity of the environments. This paper proposes an estimator with a non-convex objective function for the linear model to realize statistically efficient estimation, which is the first in the literature. The understanding of the invariance problem that this paper presents, together with this paper's limitations on the linear model and the computationally inefficient algorithm, opens up several interesting and promising future directions.

6.1 Extension to Nonlinear Models

As illustrated above, one fundamental limitation of previously studied methods is their reliance on specific structures, such as linear SCM with additive intervention. This nature not only restricts them into "structural methods", limiting their scalable uses, but also hinders them from generalizing beyond linear settings, even for generalized linear models. On the contrary, our method can be naturally extended to generalized linear models, and one can use a similar idea to develop approaches in a generic nonparametric setup. Here, we only present a direct extension to the generalized linear model using a similar objective function and leave the extension to the fully nonparametric setup as future studies.

Let $\boldsymbol{x} \in \mathbb{R}^p$ be the covariate vector and $y \in \mathbb{R}$ be the response variable of interest, and we collect data from $|\mathcal{E}|$ environments. Following the setup in Section 2, we assume that there exists some unknown true important variable set S^* and true parameter $\boldsymbol{\beta}^*$ with support set S^* such that

$$\forall e \in \mathcal{E} \qquad \mathbb{E}[y^{(e)}|\boldsymbol{x}_{S^*}^{(e)}] = \varphi((\boldsymbol{\beta}_{S^*}^*)^\top \boldsymbol{x}_{S^*}^{(e)})$$

with some known invertible link function $\varphi(\cdot)$ such as the logistic regression or log-linear model. The population-level *generalized EILLS* objective analogy to (1.5) in this setup can be written as

$$\sum_{e \in \mathcal{E}} \mathbb{E}[\ell(\boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}, y^{(e)})] + \gamma \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\} \left| \mathbb{E}[\{y^{(e)} - \varphi(\boldsymbol{\beta}^\top \boldsymbol{x}^{(e)})\} x_j^{(e)}] \right|^2$$

where $\ell(y, v) = \psi(v) - vy$ with $\varphi(t) = \psi'(t)$, and γ is some hyper-parameter to be determined. Examples include (1) $\psi(t) = \frac{1}{2}t^2, \varphi(t) = t$ for linear regression; (2) $\psi(t) = \log(1+e^t)$, $\varphi(t) = \frac{e^t}{1+e^t}$ for logistic regression. The high-level viewpoints for the two parts in the loss with general φ are similar to linear regression with $\varphi(t) = t$. One is expected to derive theoretical results analogous to Theorem 4.2–4.5; we leave this for future studies.

6.2 The Computational Complexity Concern

Currently, we use an algorithm whose computational complexity scales exponentially with p to search for the global minima of the EILLS objective function. It is natural to ask if we can design a provable algorithm that is both computationally and statistically efficient under the same environment heterogeneity condition Condition 4.5. Even if there are some fundamental limits to this problem and it is impossible to develop both computationally and statistically efficient algorithms, it is still interesting to study if we can develop heuristic search algorithms, like sure screening (Fan & Lv, 2008), forward-and-backward search (Zhang, 2011) and mixed-integer programming (Bertsimas et al., 2016) for variable selection in linear regression, such that it can offer a good solution within an affordable time limit.

Acknowledgement

The authors would like to thank three anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this paper. Y. Gu thank Yiran Jia for pointing out a typo in the proof of Theorem A.4. J. Fan's research is supported by the ONR grants N00014-19-1-2120 and N00014-22-1-2340 and NSF grants DMS-2052926, DMS-2053832 and DMS-2210833. C. Fang's research is supported by National Key R&D Program of China (2022ZD0160301) and NSF China (No. 62376008).

References

Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, 41(1), 15–34.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893.*

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2), 813.

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4), 1705–1732.

Bollen, K. A. (1989). *Structural equations with latent variables*, volume 210. John Wiley & Sons.

Bühlmann, P. & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Candes, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6), 2313 – 2351.

Chen, Y. & Bühlmann, P. (2021). Domain adaptation under structural causal models. *The Journal of Machine Learning Research*, 22(1), 11856–11935.

Čuklina, J., Lee, C. H., Williams, E. G., Sajic, T., Collins, B. C., Rodríguez Martínez, M., Sharma, V. S., Wendt, F., Goetze, S., Keele, G. R., et al. (2021). Diagnostics and correction of batch effects in large-scale proteomic studies: A tutorial. *Molecular systems biology*, 17(8), e10240.

Dawid, A. P. & Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4(none), 184 – 231.

Didelez, V., Dawid, P., & Geneletti, S. (2012). Direct and indirect effects of sequential treatments. *arXiv preprint arXiv:1206.6840*.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, 88, S28–S59.

Engle, R. F., Hendry, D. F., & Richard, J.-F. (1983). Exogeneity. *Econometrica: Journal of the Econometric Society*, (pp. 277–304).

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2), 293–314.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348–1360.

Fan, J., Li, R., Zhang, C.-H., & Zou, H. (2020). *Statistical foundations of data science*. Chapman and Hall/CRC.

Fan, J. & Liao, Y. (2014). Endogeneity in high dimensions. *Annals of statistics*, 42(3), 872.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849–911.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Cambridge University Press; Reissue edition (May 19, 2011).

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673.

Ghassami, A., Salehkaleybar, S., Kiyavash, N., & Zhang, K. (2017). Learning causal structures using regression invariance. *Advances in Neural Information Processing Systems*, 30.

Glymour, M., Pearl, J., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., & Schölkopf, B. (2016). Domain adaptation with conditional transferable components. In *International conference on machine learning* (pp. 2839–2848).: PMLR.

Gu, Y., Fang, C., Bühlmann, P., & Fan, J. (2024). Causality pursuit from heterogeneous environments via neural adversarial invariance learning. *arXiv preprint arXiv:2405.04715*.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, (pp. iii–115).

He, Y.-B. & Geng, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov), 2523–2547.

Heinze-Deml, C. & Meinshausen, N. (2021). Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2), 303–348.

Heinze-Deml, C., Peters, J., & Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2).

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kamath, P., Tangella, A., Sutherland, D., & Srebro, N. (2021). Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics* (pp. 4069–4077).: PMLR.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., & Courville, A. (2021). Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning* (pp. 5815–5826).: PMLR.

Legendre, A.-M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes [New Methods for the Determination of the Orbits of Comets] (in French)*. Paris: F. Didot.

Lu, C., Wu, Y., Hernández-Lobato, J. M., & Schölkopf, B. (2021). Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*.

Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., & Bühlmann, P. (2016). Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27), 7361–7368.

Muandet, K., Balduzzi, D., & Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *International conference on machine learning* (pp. 10–18).: PMLR.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, (pp. 947–1012).

Pfister, N., Bühlmann, P., & Peters, J. (2019). Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527), 1264–1276.

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., & Bühlmann, P. (2021). Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3), 1220–1246.

Rojas-Carulla, M., Schölkopf, B., Turner, R., & Peters, J. (2018). Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1), 1309–1342.

Rosenfeld, E., Ravikumar, P., & Risteski, A. (2021). The risks of invariant risk minimization. *International Conference on Learning Representations*.

Rothenhäusler, D., Bühlmann, P., & Meinshausen, N. (2019). Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3), 1688–1722.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., & Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 83(2), 215–246.

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. (2012). On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4), 385–395.

Torralba, A. & Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011* (pp. 1521–1528).: IEEE.

van der Vaart, A. W. & Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media.

Vladimirova, M., Girard, S., Nguyen, H., & Arbel, J. (2020). Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1), e318.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, Z. & Veitch, V. (2023). The causal structure of domain invariant supervised representation learning. *stat*, 1050, 7.

Yin, M., Wang, Y., & Blei, D. M. (2021). Optimization-based causal estimation from heterogenous environments. *arXiv preprint arXiv:2109.11990*.

Zhang, A., Lyle, C., Sodhani, S., Filos, A., Kwiatkowska, M., Pineau, J., Gal, Y., & Precup, D. (2020). Invariant causal prediction for block mdps. In *International Conference on Machine Learning* (pp. 11214–11224).: PMLR.

Zhang, C.-H. & Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4), 576–593.

Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE transactions on information theory*, 57(7), 4689–4708.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7, 2541–2563.

Supplemental Material

The supplemental materials collect the complete theoretical analysis (Appendix A), all the population-level proofs (Appendix B), finite sample proofs (Appendix C), and omitted discussions in the main text (Appendix D).

A General Theoretical Analysis

A.1 General Notations and Conditions

Define the pooled covariance matrix $\bar{\boldsymbol{\Sigma}} = \sum_{e\in\mathcal{E}} \omega^{(e)}\boldsymbol{\Sigma}^{(e)}$. We first state the standard assumptions used in linear regression, which are analogous to Condition 4.1–4.4. Condition A.2 and Condition A.4 are just copies of Condition 4.2 and Condition 4.4, respectively. Condition A.1 allows for varying $(n^{(e)}, \omega^{(e)})$. Condition A.3 replaces the $\boldsymbol{\Sigma}$ in Condition 4.3 by $\bar{\boldsymbol{\Sigma}}$.

Condition A.1. *For each $e \in \mathcal{E}$, $(\boldsymbol{x}_1^{(e)}, y_1^{(e)}), \ldots, (\boldsymbol{x}_{n^{(e)}}^{(e)}, y_{n^{(e)}}^{(e)})$ are i.i.d. copies of $(\boldsymbol{x}^{(e)}, y^{(e)}) \sim \mu^{(e)}$, where $\mu^{(e)}$ belongs to $\mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$ for some σ^2. The data from different environments are also independent. We have $\omega^{(e)} > 0$ for any $e \in \mathcal{E}$.*

Condition A.2. *There exists some universal constants $\kappa_L \in (0, 1]$ and $\kappa_U \in [1, \infty)$ such that*

$$\forall e \in \mathcal{E}, \qquad \kappa_L \boldsymbol{I}_p \preceq \boldsymbol{\Sigma}^{(e)} \preceq \kappa_U \boldsymbol{I}_p. \tag{A.1}$$

Condition A.3. *There exists some universal constant $\sigma_x \in [1, \infty)$ such that*

$$\forall e \in \mathcal{E}, \boldsymbol{v} \in \mathbb{R}^p, \qquad \mathbb{E}\left[\exp\left\{\boldsymbol{v}^\top \bar{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{x}^{(e)}\right\}\right] \le \exp\left(\frac{\sigma_x^2}{2} \cdot \|\boldsymbol{v}\|_2^2\right). \tag{A.2}$$

Condition A.4. *There exists some universal constant $\sigma_\varepsilon \in \mathbb{R}^+$ such that,*

$$\forall e \in \mathcal{E}, \lambda \in \mathbb{R}, \qquad \mathbb{E}[e^{\lambda\varepsilon^{(e)}}] \le e^{\frac{1}{2}\lambda^2\sigma_\varepsilon^2}. \tag{A.3}$$

We also define several quantities regarding sample size:

$$n_* = \min_{e\in\mathcal{E}} \frac{n^{(e)}}{\omega^{(e)}} \qquad \bar{n} = \left(\sum_{e\in\mathcal{E}} \frac{\omega^{(e)}}{n^{(e)}}\right)^{-1} \qquad n_{\min} = \min_{e\in\mathcal{E}} n^{(e)} \qquad n_\dagger = \left(\sum_{e\in\mathcal{E}} \frac{\omega^{(e)}}{(n^{(e)})^{3/2}}\right)^{-1}. \tag{A.4}$$

It is easy to see that $n_* \ge \bar{n} \ge n_{\min}$. In the case of balanced data with equal weights, i.e., $n^{(e)} \equiv n$ and $\omega^{(e)} \equiv 1/|\mathcal{E}|$, we have $n_* = n \cdot |\mathcal{E}|$, $\bar{n} = n_{\min} = n$ and $n_\dagger = n^{3/2}$.

We also define the *pooled linear spurious variable* concerning the environment weights $\boldsymbol{\omega}$.

Definition A.1 (Pooled Linear Spurious Variables)**.** *We let $G_{\boldsymbol{\omega}}$ be the index set of all pooled linear spurious variables in environments $\mathcal{E}$ concerning the weights $\boldsymbol{\omega}$, that is, $G_{\boldsymbol{\omega}} = \{j \in [p] : \sum_{e\in\mathcal{E}} \omega^{(e)}\mathbb{E}[x_j^{(e)}\varepsilon^{(e)}] \ne 0\}$. We say a variable x_j is a pooled linear spurious variable if $j \in G_{\boldsymbol{\omega}}$.*

We need the following condition related to the heterogeneity of environments in the presence of pooled linear spurious variables to recover $\boldsymbol{\beta}^*$.

Condition A.5 (Identification)**.** *For any $S \subseteq [p]$ satisfying $S \cap G_{\boldsymbol{\omega}} \ne \emptyset$, there exists some $e, e' \in \mathcal{E}$ such that $\boldsymbol{\beta}^{(e,S)} \ne \boldsymbol{\beta}^{(e',S)}$, where $\boldsymbol{\beta}^{(e,S)}$ is defined in (2.6).*

A.2 Strong Convexity with respect to the True Parameter

We are now ready to state the main population-level result, a generalized version of Theorem 4.2 with varying environments coefficients $\boldsymbol{\omega}$.

Theorem A.1 (Strong Convexity with respect to $\boldsymbol{\beta}^*$)**.** *Assume Conditions A.1–A.2 and A.5 hold. Then $\boldsymbol{\beta}^*$ is the unique minimizer of $\mathsf{Q}(\boldsymbol{\beta};\gamma,\boldsymbol{\omega})$ for large enough γ: for any $\epsilon \in (0,1)$ and any $\gamma \ge \epsilon^{-1}\gamma^*$ with*

$$\gamma^* = (\kappa_L)^{-3} \sup_{S: S\cap G_{\boldsymbol{\omega}} \neq \emptyset} \left(\mathsf{b}_S / \bar{\mathsf{d}}_S\right), \tag{A.5}$$

where

$$\mathsf{b}_S = \left\| \sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[\varepsilon^{(e)} \boldsymbol{x}_S^{(e)}] \right\|_2^2 \qquad \text{and} \qquad \bar{\mathsf{d}}_S = \sum_{e\in\mathcal{E}} \omega^{(e)} \|\boldsymbol{\beta}^{(e,S)} - \bar{\boldsymbol{\beta}}^{(S)}\|_2^2 \tag{A.6}$$

with $\bar{\boldsymbol{\beta}}^{(S)} = \sum_{e'\in\mathcal{E}} \omega^{(e')} \boldsymbol{\beta}^{(e',S)}$, we have

$$\mathsf{Q}(\boldsymbol{\beta};\gamma,\boldsymbol{\omega}) - \mathsf{Q}(\boldsymbol{\beta}^*;\gamma,\boldsymbol{\omega}) \ge (1-\epsilon)\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2^2 + \kappa_L^2(\gamma - \epsilon^{-1}\gamma^*)\bar{\mathsf{d}}_{\mathrm{supp}(\boldsymbol{\beta})}. \tag{A.7}$$

It is easy to see that Theorem 4.2 is a direct corollary of the above Theorem A.1 with $\omega^{(e)} \equiv 1/|\mathcal{E}|$.

A.3 Non-asymptotic Analysis

The first proposition characterizes the convergence of the pooled least squares.

Proposition A.2 (Properties of Pooled Least Squares)**.** *Assume Conditions A.1–A.4 hold. Then, there exists some $\bar{\boldsymbol{\beta}}^{\mathsf{R}} \in \mathbb{R}^p$ satisfying $\frac{1}{\kappa_U} \left\|\sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right\|_2 \le \|\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*\|_2 \le \frac{1}{\kappa_L} \left\|\sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right\|_2$ such that, for any $\boldsymbol{\beta} \in \mathbb{R}^p$,*

$$\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) = \|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2^2. \tag{A.8}$$

Moreover, there exist universal constants c_1 and c_2 such that if $n^* \ge c_1\sigma_x^4(p+t)$, then the pooled least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{R}}$ minimizing (3.6) satisfies

$$\|\bar{\boldsymbol{\Sigma}}^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathsf{R}} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2 \le c_2\sigma_x\left(\sigma_\varepsilon + \sigma_x\|\bar{\boldsymbol{\Sigma}}^{1/2}(\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*)\|_2\right)\sqrt{\frac{p+t}{n_*}} \tag{A.9}$$

with probability $1-2e^{-t}$.

The following two theorems are generalized versions of Theorem 4.3 and Theorem 4.4 with varying $n^{(e)}$ and $\omega^{(e)}$.

Theorem A.3 (Non-asymptotic Variable Selection Property)**.** *Define*

$$\mathsf{s}_+ = \min_{j\in S^*} |\beta_j^*|^2 \qquad \text{and} \qquad \mathsf{s}_- = \min_{S\subseteq[p], S\cap G_{\boldsymbol{\omega}}\neq\emptyset} \bar{\mathsf{d}}_S \tag{A.10}$$

Suppose Conditions A.1–A.5 hold, and we choose $\gamma \ge 3\gamma^* \vee 1$ where γ^* is defined in Theorem A.1. There exists some universal constants c_1–c_2 that only depends on $(\kappa_U, \sigma_x, \sigma_\varepsilon)$ such that for any $t > 0$, if $n_{\min} \ge p + \log(2|\mathcal{E}|) + t$, $\bar{n} \ge c_1(\gamma/\kappa_L)(p + \log(|\mathcal{E}|) + t)\{\mathsf{s}_+^{-0.5} + \mathsf{s}_+^{-1} + (\gamma\kappa_L\mathsf{s}_-)^{-0.5}\}$, and $n_* \ge c_2(\gamma/\kappa_L)^2(p+t)\{\mathsf{s}_+^{-1} + (\gamma\kappa_L\mathsf{s}_-)^{-1} + 1\}$, then the EILLS estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$ minimizing (3.7) satisfies

$$\mathbb{P}\left[S^* \subseteq \mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}) \subseteq (G_{\boldsymbol{\omega}})^c\right] \ge 1 - 7e^{-t}. \tag{A.11}$$

Combining Theorem A.3 and Proposition A.2, we can conclude that when $(n_{\min}, \bar{n}, n_*)$ are large enough, one can refit the solution found by the EILLS estimator to obtain a faster rate. To be specific, we can run another pooled least squares on $\boldsymbol{x}_{\widehat{S}}$ with $\widehat{S} = \mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{Q}})$ and obtain the ℓ_2-error $\{|(G_{\boldsymbol{\omega}})^c|/n_*\}^{1/2}$ given that $S^* \subseteq \widehat{S} \subseteq (G_{\boldsymbol{\omega}})^c$.

Theorem A.4 (Non-asymptotic ℓ_2 Error Bound)**.** *Assume Conditions A.1–A.5 hold, and we choose $\gamma \ge 3\gamma^* \vee 1$ where γ^* is defined in Theorem A.1. There exists some universal constants c_1–c_4 that only depend on (κ_U, σ_x) such that for any $t > 0$, if $n_{\min} \ge p + \log(2|\mathcal{E}|) + t$ and $n_* \ge c_1(\gamma/\kappa_L)(p + t)$, then $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$ minimizing (3.7) satisfies*

$$\frac{\|\widehat{\boldsymbol{\beta}}_{\mathsf{Q}} - \boldsymbol{\beta}^*\|_2}{\sigma_\varepsilon(\gamma/\kappa_L)} \le c_2 \left(\sqrt{\frac{p+t}{n_*}} + \frac{p + \log(|\mathcal{E}|) + t}{\bar{n}} \right) + c_3 \frac{\sqrt{|S^*|}}{\bar{n}} \cdot \frac{\log(|\mathcal{E}||S^*|) + t}{\min_{j \in S^*} |\beta_j^*|} \tag{A.12}$$

with probability at least $1 - 7e^{-t}$. Moreover, when the additional conditions in Theorem A.3 hold, then

$$\frac{\|\widehat{\boldsymbol{\beta}}_{\mathsf{Q}} - \boldsymbol{\beta}^*\|_2}{\sigma_\varepsilon(\gamma/\kappa_L)} \le c_2 \left(\sqrt{\frac{p_0+t}{n_*}} + \frac{p_0 + \log(|\mathcal{E}|) + t}{\bar{n}} \right) \qquad \text{with} \;\; p_0 = |(G_{\boldsymbol{\omega}})^c| \tag{A.13}$$

occurs with probability at least $1 - 14e^{-t}$. The dependency of (c_1, c_2, c_3) on (κ_U, σ_x) can be found in (C.19) and (C.20) in Appendix C.4.

In the high-dimensional regime, recall that $s^* = |S^*|$, $\beta_{\min} = \min_{j \in S^*} |\beta_j^*|$. We need the following condition regarding the sample size defined above in (A.4).

Condition A.6. *Suppose that $\log(|\mathcal{E}|) \le C \log p$ and that*
(1) $n_{\min} \ge c_1(s^* + \log p)$,
(2) $n_* \ge c_2(\gamma/\kappa_L)^2 \left(s^* \log p\right) \left\{1 + 1/\left(\kappa_L \beta_{\min}^2\right)\right\}$,
(3) $\bar{n} \ge c_3(\gamma/\kappa_L)(\log p) \left\{s^* + 1/(\beta_{\min}^2)\right\}$ *and* $\bar{n} \ge c_3(\gamma/\kappa_L)\sqrt{(s^* \log p)(s^* + \log p)}/\left(\kappa_L \beta_{\min}\right)$,
(4) $n_\dagger \ge c_4(\gamma/\kappa_L)\left(s^* \log p\right)\sqrt{s^* + \log p}\left\{1 + 1/\left(\sqrt{\kappa_L}\beta_{\min}\right)\right\}$.
Here c_1–c_4 are universal positive constants that depend only on $(C, \sigma_x, \kappa_U, \sigma_\varepsilon)$. See details in (C.23).

The above Condition A.6 matches Condition 4.6 in the case of balanced data with equal weights. The following theorem claims that EILLS can attain variable selection consistency with proper choice of the model selection hyper-parameter λ.

Theorem A.5 (Variable Selection Consistency in High Dimensions)**.** *Assume Conditions A.1–A.6 hold, and we choose $\gamma \ge 3\gamma^* \vee 1$ where γ^* is defined in Theorem A.1. Suppose further that the choice of λ satisfies*

$$c_1 \left\{ (\gamma/\kappa_L)^2 \frac{s^* \log p}{n_*} + \epsilon(\bar{n}, n_\dagger) \right\} \le \lambda \le c_2 \kappa_L \beta_{\min}^2, \tag{A.14}$$

where $\epsilon(\bar{n}, n_\dagger) = (\gamma/\kappa_L)^2(\log p)(s^* + \log p)\{\bar{n}^{-2}s^* + n_\dagger^{-2}(s^*)^2 \log p + (n_\dagger\sqrt{s^* + \log p})^{-1}(\gamma/\kappa_L)^{-1}\}$, and c_1, c_2 are some universal positive constants only depends on $(C, \kappa_U, \sigma_x, \sigma_\varepsilon)$. Then the ℓ_0 regularized EILLS estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{L}}$ solving (3.9) satisfies

$$\mathbb{P}\left[\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) = S^*\right] \ge 1 - p^{-10}.$$

A.4 Examples of Theorem A.1

A.4.1 Linear SCM with Heterogeneous Variance

We start with a toy example similar to Example 4.1.

Example A.1 (Toy Example, Heterogeneous Variances)**.** *Consider the following multi-environment linear model with $\mathcal{E} = \{1, 2\}$ and $p = 2$,*

$$
\begin{aligned}
x_1^{(e)} &\leftarrow \sqrt{v_1} \cdot \varepsilon_1 \\
y^{(e)} &\leftarrow 1 \cdot x_1^{(e)} + \sqrt{v_0^{(e)}} \cdot \varepsilon_0 \\
x_2^{(e)} &\leftarrow h \cdot x_1^{(e)} + s \cdot y^{(e)} + \sqrt{v_2} \cdot \varepsilon_2
\end{aligned}
$$

for some parameters $s, h \in \mathbb{R}$ and $v_1, v_2 \in \mathbb{R}^+$ that is invariant across different environments, and one parameter $v_0^{(e)}$ that will vary for different environments. Here $\varepsilon_0, \varepsilon_1, \varepsilon_2$ are independent, standard normal distributed exogenous variables. We let $\omega^{(e)} \equiv 1/2$.

In the example, $S^* = \{1\}$, $\boldsymbol{\beta}^* = (1, 0)^\top$, and $\varepsilon^{(e)} = (v_0^{(e)})^{1/2} \cdot \varepsilon_0$. Hence $G_{\boldsymbol{\omega}} = \{2\}$ whenever $s \neq 0$ since $\mathbb{E}[\varepsilon^{(e)} x_2^{(e)}] = s v_0^{(e)}$. In a well-conditioned regime where $|h| + |s| \lesssim 1$ and $v_1 \asymp v_2 \asymp v_0^{(e)} \asymp 1$, we have

$$
\sqrt{\frac{\mathsf{b}_{\{2\}}}{\bar{\mathsf{d}}_{\{2\}}}} \asymp \frac{1}{|v_0^{(1)} - v_0^{(2)}|} \frac{1}{|h(s+h)v_1 + v_2|} \qquad \text{and} \qquad \sqrt{\frac{\mathsf{b}_{\{1,2\}}}{\bar{\mathsf{d}}_{\{1,2\}}}} \asymp \frac{1}{|v_0^{(1)} - v_0^{(2)}|}.
$$

Therefore, Condition A.5 holds if

$$
v_0^{(1)} \neq v_0^{(2)} \qquad \text{and} \qquad h(h+s)v_1 + v_2 \neq 0 \tag{A.15}
$$

and the quantity γ^* is of constant order if

$$
|v_0^{(1)} - v_0^{(2)}| = \Omega(1) \qquad \text{and} \qquad \left| s - \left(-\frac{v_2}{v_1 h} - h \right) \right| = \Omega(1). \tag{A.16}
$$

The first condition, $v_0^{(2)}$ should not be close to $v_0^{(1)}$, is easy to understand. This is because when $v_0^{(2)}$ is very close to $v_0^{(1)}$, the distribution of the two environments might be very similar, which will make it hard to distinguish. The second condition, the direct effect of y on x_2 must be apart from $(-v_2/(v_1 h) - h)$, is strange at first glance. Let us see what it implies when s coincides with $(-v_2/(v_1 h) - h)$. If $s = (-v_2/(v_1 h) - h)$, then the heterogeneity of the environments (from $v_0^{(1)}$ to $v_0^{(2)}$) will not affect the solution $\boldsymbol{\beta}^{(e,\{1,2\})}$ – it will be fixed as $\boldsymbol{\beta}^{(e,\{1,2\})} \equiv (0, s^{-1})^\top$. In other words, though the distributions of $(\boldsymbol{x}^{(1)}, y^{(1)})$ and $(\boldsymbol{x}^{(2)}, y^{(2)})$ differs a lot, the two identity equations hold

$$
\mathbb{E}[y^{(e)} | x_1^{(e)}] \equiv 1 \cdot x_1^{(e)} \qquad \text{and} \qquad \mathbb{E}[y^{(e)} | x_2^{(e)}] \equiv s^{-1} \cdot x_2^{(e)}.
$$

This will be problematic since the two sets $\{1\}$ and $\{2\}$ are all CE-invariant. In this case, one can still show that $\boldsymbol{\beta}^*$ is the unique minimizer of (3.8) if and only if $(v_0^{(1)} + v_0^{(2)})/2 < s^{-2}(h^2 v_1 + v_2) = v_1^2 h^2/(v_1 h^2 + v_2)$, that is, whether the L_2 risk of $\boldsymbol{\beta}^*$ is smaller than the L_2 risk of the spurious solution $(0, s^{-1})^\top$. But it is beyond the scope of our discussion.

Moreover, the condition (A.16) is independent of $|s|$ that measures the magnitude of spuriousness. In particular, we do not need to use a very large γ^* when $|s|$ is very small. This is because both $\bar{\mathsf{d}}_S$ and b_S grow linearly with s^2 around $s = 0$, and the choice of γ^* is their ratio. □

Given the intuitions of the above Example A.1, we are ready to present a clean condition for a generalization of the model in Example A.2 with $p \geq 3$.

Example A.2 (General Linear SCM with intervention on the scale of $\varepsilon^{(e)}$.)**.** *Consider the following two-environment model $\mathcal{E} = \{1, 2\}$ for p-dimensional covariate $\boldsymbol{x}$ and response variable y,*

$$
\begin{aligned}
\boldsymbol{x}^{(e)} &= \boldsymbol{B}\boldsymbol{x}^{(e)} + \boldsymbol{\alpha} y^{(e)} + \boldsymbol{\varepsilon}_x, \\
y^{(e)} &= (\boldsymbol{\beta}^*)^\top \boldsymbol{x}^{(e)} + \varepsilon_0^{(e)}
\end{aligned}
\tag{A.17}
$$

where $\boldsymbol{B} \in \mathbb{R}^p$, $\boldsymbol{\alpha} \in \mathbb{R}^p$ and $\boldsymbol{\beta}^*$ are the same across the two environments, the exogenous variables $(\boldsymbol{\varepsilon}_x, \varepsilon_0^{(e)}) \in \mathbb{R}^p \times \mathbb{R}$ are independent, and the covariance matrix of $\boldsymbol{\varepsilon}_x$, $\boldsymbol{D} = \mathbb{V}[\boldsymbol{\varepsilon}_x]$, is also fixed for the two environments. The only heterogeneity comes from the variance of the noise $\varepsilon_0^{(e)}$: $v_0^{(e)} = \mathbb{V}[\epsilon_0^{(e)}]$ is different for $e \in \{1,2\}$. We assume the induced graph is acyclic, and use $\boldsymbol{W} = (\boldsymbol{I} - \boldsymbol{B} - \boldsymbol{\alpha}(\boldsymbol{\beta}^*)^\top)^{-1}$ to denote the total effect matrix for the covariate $\boldsymbol{x}$, that is, $W_{i,j}$ is the total effect of the variable x_j on variable x_i.

Proposition A.6. *Consider the two-environment model in Example A.2, let $\omega^{(e)} \equiv 1/2$, suppose further that*

$$\max_{i\in[p]} D_{i,i} \vee \|\boldsymbol{W}\|_2 \vee \|\boldsymbol{\alpha}\|_2 \le c_1 \tag{A.18}$$

for some universal constant $c_1 > 0$. For any S, define

$$\xi(S) = 1 - \boldsymbol{\alpha}^\top \boldsymbol{W}_{S,:}^\top \left(\boldsymbol{W}_{S,:}\boldsymbol{D}\boldsymbol{W}_{S,:}^\top\right)^{-1} \boldsymbol{W}_{S,:}\boldsymbol{D}\boldsymbol{W}_{T,:}^\top \boldsymbol{\beta}_T^* \qquad \text{with} \qquad T = S^* \setminus S.$$

If $\xi(S) \neq 0$ for any $S \subseteq [p]$, then Condition A.5 holds, and the property (A.7) in Theorem A.1 holds with

$$\gamma^* = c_2 \left(\min_{i\in[p]} D_{i,i}\right)^{-7} \times \left\{ \frac{(v_0^{(1)} + v_0^{(2)})(1+v_0^{(1)})(1+v_0^{(2)})}{|v_0^{(1)} - v_0^{(2)}|} \times \sup_{S\subseteq[p]} \frac{1}{|\xi(S)|} \right\}^2 \tag{A.19}$$

for some constant $c_2 > 0$ depends only c_1.

When $\min_{i\in[p]} D_{i,i} \gtrsim 1$ and $v_0^{(e)} \lesssim 1$, Proposition A.6 suggests the following requirements

$$|v_0^{(1)} - v_0^{(2)}| = \Omega(1) \qquad \text{and} \qquad \inf_{S\subseteq[p]} |\xi(S)| = \Omega(1)$$

are needed to get a constant order γ^*, here $\xi(S) = 1$ if $S^* \subseteq S$ or $S \cap G_{\boldsymbol{\omega}} = \emptyset$. Compared with the conditions in (A.16), the first condition is the same, and the condition $\inf_{S\subseteq[p]} |\xi(S)| = \Omega(1)$ generalizes the second condition of (A.16) to the multivariate case with complicated and arbitrary variable dependencies. Similar to Proposition A.6, the choice of γ^* is independent of $\|\boldsymbol{\alpha}\|_2$ when $\|\boldsymbol{\alpha}\|_2 \lesssim 1$, which means one do not need to use a large γ^* when the magnitude of spuriousness $\|\boldsymbol{\alpha}\|_2$ is small, i.e., $\|\boldsymbol{\alpha}\|_2 \approx 0$.

B Proofs for Population-level Results

B.1 Proof of Proposition 2.1

We first prove the equality $\mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}^*; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) = \sigma^2$. The definition of $\mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$ implies that,

$$\mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}^*; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) = \sup_{\mu\in\mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mu}\left[|y - \mathbb{E}[y|\boldsymbol{x}_S^*]|^2\right] = \sup_{\mu\in\mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}} \mathrm{var}_\mu[y|\boldsymbol{x}_{S^*}] \le \sigma^2.$$

Moreover, denote $\nu \sim \mathcal{N}(0, \sigma^2 I_p)$ as a Gaussian distribution on $\mathbb{R}^p$. Let $\mu_*(d\boldsymbol{x}, y) = \nu(\boldsymbol{x})\nu_y(d\boldsymbol{y}|\boldsymbol{x})$, where ν_y is also a Gaussian distribution with mean $(\boldsymbol{\beta}^*)^\top \boldsymbol{x}$ and variance σ^2, it is easy to show that $\mu_* \in \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$ and $\mathbb{E}_{\mu_*}[|y - (\boldsymbol{\beta}^*)^\top \boldsymbol{x}|^2] = \sigma^2$.Combining with the above upper bound, we can conclude that $\mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}^*; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) = \sigma^2$.

We next prove the upper bound part of (2.3). For any μ, we have

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{x},y)\sim\mu}\left[|y - \boldsymbol{\beta}^\top \boldsymbol{x}|^2\right] &= \mathbb{E}\left[\left(y - (\boldsymbol{\beta}^*)^\top \boldsymbol{x} + (\boldsymbol{\beta}^*)^\top \boldsymbol{x} - \boldsymbol{\beta}^\top \boldsymbol{x}\right)^2\right] \\
&= \mathbb{E}\left[\left(y - (\boldsymbol{\beta}^*)^\top \boldsymbol{x}\right)^2\right] + \mathbb{E}\left[\left((\boldsymbol{\beta}^*)^\top \boldsymbol{x} - \boldsymbol{\beta}^\top \boldsymbol{x}\right)^2\right] \\
&\qquad\qquad + 2\mathbb{E}\left[\left(y - (\boldsymbol{\beta}^*)^\top \boldsymbol{x}\right)\left((\boldsymbol{\beta}^*)^\top \boldsymbol{x} - \boldsymbol{\beta}^\top \boldsymbol{x}\right)\right] \\
&\overset{(a)}{\le} \sigma^2 + \lambda_{\max}\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right) \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \\
&\qquad\qquad + 2\sigma\sqrt{\lambda_{\max}\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right) \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{[p]\setminus S^*}\|_2^2} \\
&\le \mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}^*; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) + p^2\sigma^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + 2\sigma^2 p\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{[p]\setminus S^*}\|_2
\end{aligned}$$

where (a) follows from the fact $\mathbb{E}[(y - (\boldsymbol{\beta}^*)^\top)\boldsymbol{x}|\boldsymbol{x}_{S^*}] = 0$ and Cauchy-Schwarz inequality. For the lower bound part, Now that $\varepsilon = y - (\boldsymbol{\beta}^*)^\top \boldsymbol{x}$ is independent of $\boldsymbol{x}$ under μ_* yields, for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\begin{aligned}
\mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) \ge \mathbb{E}_{(\boldsymbol{x},y)\sim\mu_*}\left[|y - \boldsymbol{\beta}^\top \boldsymbol{x}|^2\right] &= \mathbb{E}_{(\boldsymbol{x},y)\sim\mu_*}\left[\left(y - (\boldsymbol{\beta}^*)^\top \boldsymbol{x} + (\boldsymbol{\beta}^*)^\top \boldsymbol{x} - \boldsymbol{\beta}^\top \boldsymbol{x}\right)^2\right] \\
&= \mathbb{E}_{(\boldsymbol{x},y)\sim\mu_*}\left[\left(y - (\boldsymbol{\beta}^*)^\top \boldsymbol{x}\right)^2\right] + \sigma^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \\
&= \mathsf{R}_{\mathsf{oos}}(\boldsymbol{\beta}^*; \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}) + \sigma^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2,
\end{aligned}$$

which completes the proof. □

B.2 Proof of Proposition 2.2

Denote $s^* = |S^*|$. Without loss of generality, let $S^* = \{1, \cdots, s\}$ and $\beta_{s+1} \neq 0$. Consider the following matrix

$$\bar{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{I}_{s^*} & \boldsymbol{B} & \boldsymbol{\beta}^*_{S^*} \\ \boldsymbol{B}^\top & \boldsymbol{I}_{p-s^*} + \boldsymbol{B}^\top \boldsymbol{B} & \boldsymbol{\alpha} \\ (\boldsymbol{\beta}^*_{S^*})^\top & \boldsymbol{\alpha}^\top & c \end{bmatrix}$$

where $\boldsymbol{I}_q$ is q by q identity matrix, $\boldsymbol{B} \in \mathbb{R}^{s^* \times (p-s^*)}$ satisfies $B_{j,k} = 1\{k = 1\}(\beta_j^* - \beta_j)/(\beta_k)$, $\boldsymbol{\alpha} = \boldsymbol{B}^\top \boldsymbol{\beta}_{S^*} + (\boldsymbol{I}_{p-s^*} + \boldsymbol{B}^\top \boldsymbol{B})\boldsymbol{\beta}_T$, $c = 1 + \|\boldsymbol{\alpha}\|_2^2 + \|\boldsymbol{\beta}^*_{S^*}\|_2^2$. It is easy to verify that our constructed $\bar{\boldsymbol{\Sigma}}$ satisfies $\bar{\boldsymbol{\Sigma}} \succeq \boldsymbol{I}_{p+1}$. We then argue that the distribution μ that $(\boldsymbol{x}, y) \sim \mathcal{N}(0, \bar{\boldsymbol{\Sigma}})$ satisfies $\mu \in \mathcal{U}_{\boldsymbol{\beta}^*,\sigma^2}$ for some large enough σ^2 and $\mathbb{E}_\mu[y|\boldsymbol{x}] = \boldsymbol{\beta}^\top \boldsymbol{x}$. To this end, we only need to verify that $\mathbb{E}[y|\boldsymbol{x}_{S^*}] = (\boldsymbol{\beta}^*)^\top \boldsymbol{x}$ and $\mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{\beta}^\top \boldsymbol{x}$. For the first condition, our construction of $\bar{\boldsymbol{\Sigma}}$ and the conditional distribution of multivariate Gaussian ensures $\mathbb{E}[y|\boldsymbol{x}_{S^*}] = (\boldsymbol{I}_{s^*})^{-1}(\boldsymbol{\beta}^*_{S^*})^\top \boldsymbol{x}^*_S = (\boldsymbol{\beta}^*_{S^*})^\top \boldsymbol{x}^*_S$. Similarly, in order to show $\mathbb{E}[y|\boldsymbol{x}] = \boldsymbol{\beta}^\top \boldsymbol{x}$, it suffices to verify that

$$\begin{bmatrix} \boldsymbol{I}_{s^*} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{I}_{p-s^*} + \boldsymbol{B}^\top \boldsymbol{B} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}^*_{S^*} \\ \boldsymbol{\alpha} \end{bmatrix},$$

which can be validated by plugging in our construction of $\boldsymbol{B}$ and $\boldsymbol{\alpha}$. This completes the proof. □

B.3 Proof of Theorem A.1

It follows from the definition of $\mathsf{Q}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega})$ and Condition A.1 that

$$\mathsf{Q}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega}) - \mathsf{Q}(\boldsymbol{\beta}^*; \gamma, \boldsymbol{\omega}) = \mathsf{R}(\boldsymbol{\beta}; \boldsymbol{\omega}) - \mathsf{R}(\boldsymbol{\beta}^*; \boldsymbol{\omega}) + \gamma \mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) = \mathsf{T}_1(\boldsymbol{\beta}) + \gamma \mathsf{T}_2(\boldsymbol{\beta}).$$

Denote $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$, $S = \mathrm{supp}(\boldsymbol{\beta})$, $T = S^* \setminus S$, and $\bar{S} = S \cup S^*$. it follows from the model $y^{(e)} = (\boldsymbol{\beta}^*)^\top \boldsymbol{x}^{(e)} + \varepsilon^{(e)}$ that

$$\begin{aligned}
\mathsf{T}_1(\boldsymbol{\beta}) &= \sum_{e\in\mathcal{E}} \omega^{(e)} \left(\mathsf{R}^{(e)}(\boldsymbol{\beta}) - \mathsf{R}^{(e)}(\boldsymbol{\beta}^*)\right) \\
&= \sum_{e\in\mathcal{E}} \omega^{(e)} \left(\mathbb{E}\left[|y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}|^2\right] - \mathbb{E}\left[|y^{(e)} - (\boldsymbol{\beta}^*)^\top \boldsymbol{x}^{(e)}|^2\right]\right) \\
&= \sum_{e\in\mathcal{E}} \omega^{(e)} \left(\mathbb{E}\left[|(\boldsymbol{\beta}^* - \boldsymbol{\beta})^\top \boldsymbol{x}^{(e)} + \varepsilon^{(e)}|^2\right] - \mathbb{E}[|\varepsilon^{(e)}|^2]\right) \\
&= \sum_{e\in\mathcal{E}} \omega^{(e)} \boldsymbol{\Delta}^\top \boldsymbol{\Sigma}^{(e)} \boldsymbol{\Delta} - 2\boldsymbol{\Delta}^\top \mathbb{E}\left[\varepsilon^{(e)} \boldsymbol{x}^{(e)}\right] \\
&= \boldsymbol{\Delta}^\top \left(\sum_{e\in\mathcal{E}} \omega^{(e)} \boldsymbol{\Sigma}^{(e)}\right) \boldsymbol{\Delta} - 2\boldsymbol{\Delta}^\top \left(\sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[\varepsilon^{(e)} \boldsymbol{x}^{(e)}]\right) = \boldsymbol{\Delta}_{\bar{S}}^\top \bar{\boldsymbol{\Sigma}}_{\bar{S}} \boldsymbol{\Delta}_{\bar{S}} - 2\boldsymbol{\Delta}_{\bar{S}}^\top \bar{\boldsymbol{u}}_{\bar{S}}
\end{aligned}$$

It follows from the fact $\bar{\boldsymbol{u}}_T = \mathbf{0}$ and Cauchy-Schwarz inequality that

$$2\boldsymbol{\Delta}_{\bar{S}}^\top \bar{\boldsymbol{u}}_{\bar{S}} = 2\left(\sqrt{\epsilon}\boldsymbol{\Delta}_{\bar{S}}\bar{\boldsymbol{\Sigma}}_{\bar{S}}^{1/2}\right)\left(\epsilon^{-1/2}\bar{\boldsymbol{\Sigma}}_{\bar{S}}^{-1/2}\bar{\boldsymbol{u}}_{\bar{S}}\right) \le \epsilon\boldsymbol{\Delta}_{\bar{S}}^\top\bar{\boldsymbol{\Sigma}}_{\bar{S}}\boldsymbol{\Delta}_{\bar{S}} + \epsilon^{-1}\bar{\boldsymbol{u}}_{\bar{S}}^\top\bar{\boldsymbol{\Sigma}}_{\bar{S}}^{-1}\bar{\boldsymbol{u}}_{\bar{S}} \tag{B.1}$$

Plugging (B.1) back yields

$$\mathsf{T}_1(\boldsymbol{\beta}) \ge (1-\epsilon)\boldsymbol{\Delta}_{\bar{S}}^\top\bar{\boldsymbol{\Sigma}}_{\bar{S}}\boldsymbol{\Delta}_{\bar{S}} - \epsilon^{-1}\bar{\boldsymbol{u}}_{\bar{S}}^\top\bar{\boldsymbol{\Sigma}}_{\bar{S}}^{-1}\bar{\boldsymbol{u}}_{\bar{S}}. \tag{B.2}$$

At the same time, we also have

$$\begin{aligned}
\mathsf{T}_2(\boldsymbol{\beta}) &= \sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\mathbb{E}\left[(y^{(e)} - \boldsymbol{\beta}^\top\boldsymbol{x}^{(e)})\boldsymbol{x}_S^{(e)}\right]\right\|_2^2 \\
&= \sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\mathbb{E}\left[\varepsilon^{(e)}\boldsymbol{x}_{\bar{S}}^{(e)}\right] + (\boldsymbol{\beta}_S^* - \boldsymbol{\beta}_S)^\top\mathbb{E}\left[\boldsymbol{x}_S^{(e)}(\boldsymbol{x}_S^{(e)})^\top\right] + (\boldsymbol{\beta}_T^*)^\top\mathbb{E}\left[\boldsymbol{x}_T^{(e)}(\boldsymbol{x}_S^{(e)})^\top\right]\right\|^2 \\
&= \sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\bar{\boldsymbol{u}}_S + \boldsymbol{\Sigma}_{S,T}^{(e)}\boldsymbol{\beta}_T^* - \boldsymbol{\Sigma}_S^{(e)}\boldsymbol{\Delta}_S\right\|_2^2 \\
&= \sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\boldsymbol{\Sigma}_S(\boldsymbol{\beta}^{(e,S)} - \boldsymbol{\beta}^* - \boldsymbol{\Delta}_S)\right\|^2.
\end{aligned}$$

Putting these pieces together with Condition A.2, we have

$$\begin{aligned}
\mathsf{T}_1(\boldsymbol{\beta}) + \gamma\mathsf{T}_2(\boldsymbol{\beta}) &\ge (1-\epsilon)\|\bar{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\Delta}_{\bar{S}}\|_2^2 - \varepsilon^{-1}\kappa_L^{-1}\|\bar{\boldsymbol{u}}_S\|_2^2 + \gamma\sum_{e\in\mathcal{E}}\omega^{(e)}\kappa_L^2\|\boldsymbol{\beta}^{(e,S)} - \boldsymbol{\beta}^* - \boldsymbol{\Delta}_S\|_2^2 \\
&\ge (1-\epsilon)\|\bar{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\Delta}_{\bar{S}}\|_2^2 - \varepsilon^{-1}\kappa_L^{-1}\mathsf{b}_S + \inf_{\boldsymbol{v}\in\mathbb{R}^{|S|}}\gamma\sum_{e\in\mathcal{E}}\omega^{(e)}\kappa_L^2\|\boldsymbol{\beta}_S^{(e,S)} - \boldsymbol{\beta}_S^* - \boldsymbol{v}\|_2^2
\end{aligned} \tag{B.3}$$

Define

$$L(\boldsymbol{v}) = \sum_{e\in\mathcal{E}}\omega^{(e)}\|\boldsymbol{\beta}_S^{(e,S)} - \boldsymbol{\beta}_S^* - \boldsymbol{v}\|_2^2.$$

It is obvious that $L(\boldsymbol{v})$ is of a quadratic form and is strong convex with curvature 1. So it has the unique global minimizer $\boldsymbol{v}^* = \sum_{e\in\mathcal{E}}\omega^{(e)}\boldsymbol{\beta}^{(e,S)} - \boldsymbol{\beta}_S^*$ with minimum value

$$\inf_{\boldsymbol{v}\in\mathbb{R}^{|S|}} L(\boldsymbol{v}) = \sum_{e\in\mathcal{E}}\omega^{(e)}\|\boldsymbol{\beta}_S^{(e,S)} - \boldsymbol{\beta}_S^* - \boldsymbol{v}^*\|_2^2 = \bar{\mathsf{d}}_S. \tag{B.4}$$

Substituting it back into (B.3) gives

$$\begin{aligned}
\mathsf{T}_1(\boldsymbol{\beta}) + \gamma\mathsf{T}_2(\boldsymbol{\beta}) &\ge (1-\epsilon)\|\bar{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\Delta}\|_2^2 - \epsilon^{-1}\kappa_L^{-1}\mathsf{b}_S + \gamma\kappa_L^2\mathsf{d}_S \\
&= (1-\epsilon)\|\bar{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\Delta}\|_2^2 - \epsilon^{-1}\kappa_L^{-1}\mathsf{b}_S + \epsilon^{-1}\kappa_L^2\bar{\mathsf{d}}_S\gamma^* + (\gamma - \epsilon^{-1}\gamma^*)\kappa_L^2\bar{\mathsf{d}}_S \\
&\ge (1-\epsilon)\|\bar{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\Delta}\|_2^2 + (\gamma - \epsilon^{-1}\gamma^*)\kappa_L^2\bar{\mathsf{d}}_S.
\end{aligned} \tag{B.5}$$

This completes the proof. □

B.4 Proof of Proposition A.6

We first calculate serval quantities of interests and determine the original curvature κ_L.

It follows from the model (A.17) that

$$\boldsymbol{x}^{(e)} = \boldsymbol{B}\boldsymbol{x}^{(e)} + \boldsymbol{\alpha}(\boldsymbol{\beta}^*)^\top\boldsymbol{x}^{(e)} + \boldsymbol{\alpha}\varepsilon_0^{(e)} + \boldsymbol{\varepsilon}_x.$$

Because the induced graph is acyclic, then there exists a permutation $\pi : [p] \to [p]$ such that

$$\left[\boldsymbol{B} + \boldsymbol{\alpha}(\boldsymbol{\beta}^*)^\top\right]_{i,j} = 0 \quad \text{if} \quad \pi(i) < \pi(j),$$

that is, there exists a permutation matrix $\boldsymbol{P}$ such that the matrix

$$\widetilde{\boldsymbol{V}} = \boldsymbol{P}[\boldsymbol{B} + \boldsymbol{\alpha}(\boldsymbol{\beta}^*)^\top]\boldsymbol{P}^\top$$

satisfies $\widetilde{\boldsymbol{V}}$ is a lower triangular matrix with all zeros on the diagonal. Hence, the inverse of the matrix $\boldsymbol{P}\boldsymbol{I}\boldsymbol{P}^\top - \widetilde{\boldsymbol{V}}$ exists and is an upper triangular matrix with all ones on the diagonal, thus implying

$$\boldsymbol{W} = (\boldsymbol{I} - \boldsymbol{B} + \boldsymbol{\alpha}(\boldsymbol{\beta}^*)^\top)^{-1} \ \text{ exists with } \ \nu_{\min}(\boldsymbol{W}) \geq 1.$$

Therefore, we can represent the covariate $\boldsymbol{x}^{(e)}$ in terms of all the exogenous variables $\boldsymbol{\varepsilon}_x$ and $\varepsilon_0^{(e)}$ as

$$\boldsymbol{x}^{(e)} = \boldsymbol{W}(\boldsymbol{\varepsilon}_x + \boldsymbol{\alpha}\varepsilon_0^{(e)}). \tag{B.6}$$

Therefore, denote $\boldsymbol{u} = \boldsymbol{W}\boldsymbol{\alpha}$, we further have

$$\boldsymbol{\Sigma}^{(e)} = \boldsymbol{W}\boldsymbol{D}\boldsymbol{W}^\top + v_0^{(e)}\boldsymbol{u}\boldsymbol{u}^\top, \tag{B.7}$$

$$\mathbb{E}\left[\varepsilon^{(e)}\boldsymbol{x}^{(e)}\right] = v_0^{(e)}\boldsymbol{u}^{(e)} \tag{B.8}$$

Given these quantities, we can write down the bias for any S as

$$\mathsf{b}_S = \left\| \frac{1}{2}\left(\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}_S^{(e)}]\right)(v_0^{(1)} + v_0^{(2)}) \right\|_2^2 = \left(\frac{v_0^{(1)} + v_0^{(2)}}{2}\right)^2 \cdot \|\boldsymbol{u}_S\|_2^2. \tag{B.9}$$

The calculation of bias-difference term $\bar{\mathsf{d}}_S$ is more involved. Denote $T = S^* \setminus S$, then one has

$$\boldsymbol{\beta}_S^{(e,S)} - \boldsymbol{\beta}_S^* = (\boldsymbol{\Sigma}_S^{(e)})^{-1}\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}_S^{(e)}] + (\boldsymbol{\Sigma}_S^{(e)})^{-1}\boldsymbol{\Sigma}_{S,T}^{(e)}\boldsymbol{\beta}_T^* = \mathsf{T}_1^{(e)} + \mathsf{T}_2^{(e)}.$$

Denoting $\boldsymbol{M}_S = \boldsymbol{W}_{S,:}\boldsymbol{D}\boldsymbol{W}_{S,:}^\top$, it follows from Sherman-Morrison formula that

$$\begin{aligned}
\mathsf{T}_1^{(e)} &= (\boldsymbol{M}_S + v_0^{(e)}\boldsymbol{u}_S\boldsymbol{u}_S)^{-1}\boldsymbol{u}_S v_0^{(e)} \\
&= \left\{\boldsymbol{M}_S^{-1} - \frac{\boldsymbol{M}_S^{-1}\boldsymbol{u}_S\boldsymbol{u}_S^\top\boldsymbol{M}_S^{-1}v_0^{(e)}}{1 + \boldsymbol{u}_S\boldsymbol{M}_S^{-1}\boldsymbol{u}_S v_0^{(e)}}\right\}\boldsymbol{u}_S v_0^{(e)} \\
&= \frac{v_0^{(e)}}{1 + \boldsymbol{u}_S\boldsymbol{M}_S^{-1}\boldsymbol{u}_S v_0^{(e)}}\boldsymbol{M}_S^{-1}\boldsymbol{u}_S.
\end{aligned}$$

Observe that $\boldsymbol{u}_T = 0$. This yields

$$\begin{aligned}
\mathsf{T}_2^{(e)} &= \left(\boldsymbol{M}_S + v_0^{(e)}\boldsymbol{u}_S\boldsymbol{u}_S^\top\right)^{-1}(\boldsymbol{W}_{S,:}\boldsymbol{D}\boldsymbol{W}_{T,:}^\top)\boldsymbol{\beta}_T^* \\
&= \boldsymbol{M}_S^{-1}(\boldsymbol{W}_{S,:}\boldsymbol{D}\boldsymbol{W}_{T,:}^\top)\boldsymbol{\beta}_T^* - \frac{\boldsymbol{M}_S^{-1}\boldsymbol{u}_S^{(e)}}{1 + v_0^{(e)}\boldsymbol{u}_S^\top\boldsymbol{M}_S^{-1}\boldsymbol{u}_S}v_0^{(e)}\boldsymbol{u}_S^\top\boldsymbol{M}_S^{-1}(\boldsymbol{W}_{S,:}\boldsymbol{D}\boldsymbol{W}_{T,:}^\top)\boldsymbol{\beta}_T^*
\end{aligned}$$

Putting these pieces together, denote $\iota_S = \boldsymbol{u}_S \boldsymbol{M}_S^{-1} \boldsymbol{u}_S$, we have

$$\begin{aligned}
\bar{\mathsf{d}}_S &= \sum_{e \in \{1,2\}} \frac{1}{2} \left\| \boldsymbol{\beta}^{(e,S)} - \frac{\boldsymbol{\beta}^{(1,S)} + \boldsymbol{\beta}^{(2,S)}}{2} \right\|_2^2 \\
&= \frac{1}{2} \left\| \boldsymbol{\beta}^{(1,S)} - \boldsymbol{\beta}^{(2,S)} \right\|_2^2 \\
&= \frac{1}{2} \|\boldsymbol{M}_S^{-1} \boldsymbol{u}_S\|_2^2 \left| 1 - \boldsymbol{u}_S^\top \boldsymbol{M}_S^{-1} (\boldsymbol{W}_{S,:} \boldsymbol{D} \boldsymbol{W}_{T,:}^\top) \boldsymbol{\beta}_T^* \right|^2 \left| \frac{v_0^{(1)}}{1 + \iota_S v_0^{(1)}} - \frac{v_0^{(2)}}{1 + \iota_S v_0^{(2)}} \right|^2 \\
&\ge \frac{1}{2} c_1^{-2} \|\boldsymbol{u}_S\|_2^2 \frac{|v_0^{(1)} - v_0^{(2)}|^2 \left| 1 - \boldsymbol{u}_S^\top \boldsymbol{M}_S^{-1} (\boldsymbol{W}_{S,:} \boldsymbol{D} \boldsymbol{W}_{T,:}^\top) \boldsymbol{\beta}_T^* \right|^2}{(1 + \iota_S v_0^{(1)})^2 (1 + \iota_S v_0^{(2)})^2} \\
&\ge \frac{C^{-1}}{2} \|\boldsymbol{u}_S\|_2^2 \frac{|v_0^{(1)} - v_0^{(2)}|^2 \left| 1 - \boldsymbol{u}_S^\top \boldsymbol{M}_S^{-1} (\boldsymbol{W}_{S,:} \boldsymbol{D} \boldsymbol{W}_{T,:}^\top) \boldsymbol{\beta}_T^* \right|^2}{(1 + v_0^{(1)})^2 (1 + v_0^{(2)})^2} \left\{ \min_i D_{i,i} \right\}^4
\end{aligned} \tag{B.10}$$

where the last inequality follows from the fact

$$\iota_S \le \|\boldsymbol{M}_S\|^{-1} \|\boldsymbol{u}_S\|^2 \le \{\nu_{\min}(\boldsymbol{W})\}^2 \left\{ \min_i D_{i,i} \right\}^{-1} \|\boldsymbol{W}\|_2^2 \|\boldsymbol{\alpha}\|_2^2 \lesssim \left\{ \min_i D_{i,i} \right\}^{-1}.$$

Plugging the above quantities (B.9) and (B.10) into (A.5) completes the proof.

C Proofs for Non-asymptotic Results

C.1 Preliminaries

We first define some concepts that will be used throughout the proof.

Definition C.1 (Sub-Gaussian Random Variable)**.** *A random variable X is a sub-Gaussian random variable with parameter $\sigma \in \mathbb{R}^+$ if*

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}\left[\exp(\lambda X)\right] \le \exp\left(\frac{\lambda^2}{2} \sigma^2 \right)$$

Definition C.2 (Sub-Exponential Random Variable)**.** *A random variable X is a sub-Exponential random variable with parameter $(\nu, \alpha) \in \mathbb{R}^+ \times \mathbb{R}^+$ if*

$$\forall |\lambda| < 1/\alpha, \qquad \mathbb{E}\left[\exp(\lambda X)\right] \le \exp\left(\frac{\lambda^2}{2} \nu^2 \right).$$

It is easy to verify that the product of two sub-Gaussian random variables is a Sub-Exponential random variable, and the dependence of the parameters can be written as follows.

Lemma C.1 (Product of Two Sub-Gaussian Random Variables)**.** *Suppose X_1 and X_2 are two zero-mean sub-Gaussian random variables with parameters σ_1 and σ_2, respectively. Then $X_1 X_2$ is a sub-exponential random variable with parameter $(c_1 \sigma_1 \sigma_2, c_2 \sigma_1 \sigma_2)$, where $c_1, c_2 > 0$ are some universal constants.*

We also have the following lemma stating the concentration inequality for the sum of independent sub-exponential random variables.

Lemma C.2 (Sum of Independent Sub-exponential Random Variables)**.** *Suppose $X_1, \ldots, X_N$ are independent sub-exponential random variables with parameters $\{(\nu_i, \alpha_i)\}_{i=1}^N$, respectively. There exists some universal constant c_1 such that the following holds,*

$$\mathbb{P}\left[\left| \sum_{i=1}^N (X_i - \mathbb{E}[X_i]) \right| \ge c_1 \left\{ \sqrt{t \times \sum_{i=1}^N \nu_i^2} + t \times \max_{i \in [N]} \alpha_i \right\} \right] \le 2e^{-t}.$$

We also define the following quantity regarding sample size

$$n_{\boldsymbol{\omega}} = \left(\sum_{e\in\mathcal{E}} \frac{(\omega^{(e)})^2}{n^{(e)}}\right)^{-1}. \tag{C.1}$$

It is easy to see that $n_{\boldsymbol{\omega}} \ge n_*$.

We also define

$$\mathcal{B}(S) = \{\boldsymbol{\beta} \in \mathbb{R}^p, \mathrm{supp}(\boldsymbol{\beta}) = S, \|\boldsymbol{\beta}\|_2 = 1\} \qquad \text{and} \qquad \mathcal{B}_s = \bigcup_{S\subseteq[p],|S|\le s} \mathcal{B}(S). \tag{C.2}$$

C.2 Proof of Proposition A.2

It follows from the definition of $\mathsf{R}(\boldsymbol{\beta})$ and $\mathsf{R}^{(e)}(\boldsymbol{\beta})$ that

$$\begin{aligned}
\mathsf{R}(\boldsymbol{\beta}) &= \sum_{e\in\mathcal{E}} \omega^{(e)} \mathsf{R}^{(e)}(\boldsymbol{\beta}) \\
&= \sum_{e\in\mathcal{E}} \omega^{(e)} \left\{ (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}^{(e)} (\boldsymbol{\beta}-\boldsymbol{\beta}^*) - 2(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \mathbb{E}[\boldsymbol{x}^{(e)}\varepsilon^{(e)}] + \mathbb{E}|\varepsilon^{(e)}|^2 \right\} \\
&= (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \bar{\boldsymbol{\Sigma}} (\boldsymbol{\beta}-\boldsymbol{\beta}^*) - 2(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[\boldsymbol{x}^{(e)}\varepsilon^{(e)}] + \sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}|\varepsilon^{(e)}|^2.
\end{aligned}$$

We can see $\mathsf{R}(\boldsymbol{\beta})$ is of quadratic form of $\boldsymbol{\beta}-\boldsymbol{\beta}^*$. Since $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$, the minimizer of $\mathsf{R}(\boldsymbol{\beta})$ is unique and is

$$\bar{\boldsymbol{\beta}}^{\mathsf{R}} = \boldsymbol{\beta}^* + (\bar{\boldsymbol{\Sigma}})^{-1} \sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[\boldsymbol{x}^{(e)}\varepsilon^{(e)}]. \tag{C.3}$$

Combining the above definition of $\bar{\boldsymbol{\beta}}^{\mathsf{R}}$ together with Condition A.2 validates the upper bound and lower bound on $\|\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*\|_2$. Moreover, plugging in above (C.3) into $\mathsf{R}(\boldsymbol{\beta})$ following by some calculations gives

$$\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) = (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \bar{\boldsymbol{\Sigma}} (\boldsymbol{\beta}-\boldsymbol{\beta}^*),$$

which completes the proof of claim (A.8).

Moreover, for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have

$$\begin{aligned}
\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2^2 &= \mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) \\
&= \mathsf{R}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}(\boldsymbol{\beta}) + \widehat{\mathsf{R}}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) + \widehat{\mathsf{R}}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) - \mathsf{R}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}).
\end{aligned}$$

We argue that

$$\begin{aligned}
\mathbb{P}[\mathcal{C}_{1,t}] = \mathbb{P}\Bigg\{ \forall \boldsymbol{\beta} \in \mathbb{R}^p, \quad & \mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) - \left(\widehat{\mathsf{R}}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}(\bar{\boldsymbol{\beta}}^{\mathsf{R}})\right) \\
& \le C\sigma_x \left(\sigma_\varepsilon + \sigma_x \|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta}^* - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2\right) \sqrt{\frac{p+t}{n_*}} \|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2 \\
& + C\sigma_x^2 \sqrt{\frac{p+t}{n_*}} \|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2^2 \Bigg\} \ge 1 - 2e^{-t}
\end{aligned} \tag{C.4}$$

for any $t \in (0, n_* - p]$, which will be validated later. If such a claim holds, then under $\mathcal{C}_{1,t}$ which occurs with probability at least $1 - 2e^{-t}$, we have that, for any $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\begin{aligned}
\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2^2 \le \widehat{\mathsf{R}}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) &+ C\sigma_x^2 \sqrt{\frac{p+t}{n_*}} \|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2^2 \\
&+ C\sigma_x \left(\sigma_\varepsilon + \sigma_x \|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta}^* - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2\right) \sqrt{\frac{p+t}{n_*}} \|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2.
\end{aligned}$$

Suppose further that $n^* \ge (p+t)4C^2\sigma_x^4$, we can apply the above inequality with $\widehat{\boldsymbol{\beta}}_{\mathsf{R}}$ which satisfies $\widehat{\mathsf{R}}(\widehat{\boldsymbol{\beta}}_{\mathsf{R}}) - \widehat{\mathsf{R}}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) \le 0$ and obtain

$$\|\bar{\boldsymbol{\Sigma}}^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathsf{R}} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2^2 \le 2C\sigma_x\left(\sigma_\varepsilon + \sigma_x\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta}^* - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2\right)\sqrt{\frac{p+t}{n_*}}\|\bar{\boldsymbol{\Sigma}}^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathsf{R}} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2.$$

This completes the proof.

Proof of the Bound (C.4). It follows from the definition of $\widehat{\mathsf{R}}$ and the above identity that

$$\begin{aligned}
&\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) - \left(\widehat{\mathsf{R}}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}(\bar{\boldsymbol{\beta}}^{\mathsf{R}})\right) \\
&\qquad = (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})^\top \bar{\boldsymbol{\Sigma}}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}}) - (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})^\top \left(\sum_{e\in\mathcal{E}} \omega^{(e)}\widehat{\boldsymbol{\Sigma}}^{(e)}\right)(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}}) \\
&\qquad\qquad - 2(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})^\top \sum_{e\in\mathcal{E}} \omega^{(e)}\widehat{\mathbb{E}}\left[\boldsymbol{x}^{(e)}\left((\bar{\boldsymbol{\beta}}^{\mathsf{R}})^\top \boldsymbol{x}^{(e)} - y^{(e)}\right)\right].
\end{aligned}$$

When $p + t \le n_*$, define the following event

$$\begin{aligned}
\mathcal{K}_{1,t} &= \left\{\left\|\boldsymbol{I} - \bar{\boldsymbol{\Sigma}}^{-1/2}\sum_{e\in\mathcal{E}}\omega^{(e)}\widehat{\boldsymbol{\Sigma}}^{(e)}\bar{\boldsymbol{\Sigma}}^{-1/2}\right\|_2 \le C_1\sigma_x^2\sqrt{\frac{p+t}{n_*}}\right\} \\
\mathcal{K}_{2,t} &= \left\{\left\|\sum_{e\in\mathcal{E}}\omega^{(e)}\bar{\boldsymbol{\Sigma}}^{-1/2}\widehat{\mathbb{E}}\left[\boldsymbol{x}^{(e)}\left((\bar{\boldsymbol{\beta}}^{\mathsf{R}})^\top \boldsymbol{x}^{(e)} - y^{(e)}\right)\right]\right\|\right. \\
&\qquad\qquad \left. \le C_2\sigma_x\left(\sigma_\varepsilon + \sigma_x\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta}^* - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2\right)\sqrt{\frac{p+t}{n_*}}\right\}
\end{aligned}$$

for some universal constant C_1–C_2 to be determined. It suffices to prove that $\mathbb{P}[\mathcal{K}_{1,t}] \wedge \mathbb{P}[\mathcal{K}_{2,t}] \ge 1 - e^{-t}$ for any $t > 0$. If the two claims are verified, then we can conclude that under the event $\mathcal{K}_{1,t} \cap \mathcal{K}_{2,t}$ that occurs with probability at least $1 - 2e^{-t}$, the following holds

$$\begin{aligned}
&\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\bar{\boldsymbol{\beta}}^{\mathsf{R}}) - \left(\widehat{\mathsf{R}}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}(\bar{\boldsymbol{\beta}}^{\mathsf{R}})\right) \\
&\quad = \left\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\right\|_2 \left\|\boldsymbol{I} - \bar{\boldsymbol{\Sigma}}^{-1/2}\sum_{e\in\mathcal{E}}\omega^{(e)}\widehat{\boldsymbol{\Sigma}}^{(e)}\bar{\boldsymbol{\Sigma}}^{-1/2}\right\|_2 \left\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\right\|_2 \\
&\qquad + 2\left\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\right\|_2 \left\|\sum_{e\in\mathcal{E}}\omega^{(e)}\bar{\boldsymbol{\Sigma}}^{-1/2}\widehat{\mathbb{E}}\left[\boldsymbol{x}^{(e)}\left((\bar{\boldsymbol{\beta}}^{\mathsf{R}})^\top \boldsymbol{x}^{(e)} - y^{(e)}\right)\right]\right\|_2 \\
&\quad \le C'\sigma_x^2\sqrt{\frac{p+t}{n_*}}\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2^2 \\
&\qquad + C'\sigma_x\left(\sigma_\varepsilon + \sigma_x\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta}^* - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2\right)\sqrt{\frac{p+t}{n_*}}\|\bar{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}^{\mathsf{R}})\|_2,
\end{aligned}$$

which completes the proof of the claim (C.4).

Step 1. High Probability Bound for $\mathcal{K}_{1,t}$. Let $\boldsymbol{A} = \boldsymbol{I} - \bar{\boldsymbol{\Sigma}}^{-1/2}\sum_{e\in\mathcal{E}}\omega^{(e)}\widehat{\boldsymbol{\Sigma}}^{(e)}\bar{\boldsymbol{\Sigma}}^{-1/2}$, similar to the derivation in (C.35), we can construct N pairs of p-dimensional unit vectors $(\boldsymbol{v}_1, \boldsymbol{u}_1), \ldots, (\boldsymbol{v}_N, \boldsymbol{u}_N)$ with $N \le 8100^p$ such that

$$\|\boldsymbol{A}\|_2 \le 4\sup_{k\in[N]} \boldsymbol{v}_k^\top \boldsymbol{A}\boldsymbol{u}_k.$$

For fixed $(\boldsymbol{v}, \boldsymbol{u})$, using the identity

$$\boldsymbol{I} = \sum_{e\in\mathcal{E}} \omega^{(e)} \bar{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\Sigma}^{(e)} \bar{\boldsymbol{\Sigma}}^{-1/2},$$

we obtain

$$\begin{aligned} \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{u} &= \sum_{e\in\mathcal{E}} \sum_{i=1}^{n^{(e)}} \frac{\omega^{(e)}}{n^{(e)}} \left\{ \boldsymbol{u}^\top \bar{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\Sigma}^{(e)} \bar{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{v} - \boldsymbol{u}^\top \bar{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{x}_i^{(e)} (\boldsymbol{x}_i^{(e)})^\top \bar{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{v} \right\} \\ &= \sum_{e\in\mathcal{E}} \sum_{i=1}^{n^{(e)}} \frac{\omega^{(e)}}{n^{(e)}} \left(\mathbb{E}[X_{e,i}] - X_{e,i}\right) \end{aligned}$$

for $X_{e,i} = \boldsymbol{u}^\top \bar{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{x}_i^{(e)} (\boldsymbol{x}_i^{(e)})^\top \bar{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{v}$. It follows from Condition A.3 that $X_{e,i}$ is the product of two zero-mean sub-Gaussian random variables with parameter σ_x. Then it follows from Lemma C.1 and Lemma C.2 that the following holds

$$\mathbb{P}\left[|\boldsymbol{v}_k^\top \boldsymbol{A} \boldsymbol{u}_k| \ge C_3 \sigma_x \left(\sqrt{\frac{u}{n_{\boldsymbol{\omega}}}} + \frac{u}{n_*} \right) \right] \le 2e^{-u}$$

for all the $u > 0$, where C_3 is some universal constant. Applying union bounds further gives

$$\mathbb{P}\left[\sup_{k\in[N]} |\boldsymbol{v}_k^\top \boldsymbol{A} \boldsymbol{u}_k| \ge C_3 \sigma_x \left(\sqrt{\frac{u}{n_{\boldsymbol{\omega}}}} + \frac{u}{n_*} \right) \right] \le 2Ne^{-u}.$$

So it concludes the proof by setting $u = t + \log(2N) = t + C_4 p$ with some $C_4 > 1$ and observing that

$$\sqrt{\frac{t + C_4 p}{n_{\boldsymbol{\omega}}}} + \frac{t + C_4 p}{n_*} \le 2C_4 \sqrt{\frac{t+p}{n_*}}$$

provided $p + t \le n_*$.

Step 2. High Probability Bound for $\mathcal{K}_{2,t}$.

Let $\boldsymbol{b} = \sum_{e\in\mathcal{E}} \omega^{(e)} \bar{\boldsymbol{\Sigma}}^{-1/2} \widehat{\mathbb{E}}\left[\boldsymbol{x}^{(e)} \left((\bar{\boldsymbol{\beta}}^{\mathsf{R}})^\top \boldsymbol{x}^{(e)} - y^{(e)} \right)\right]$, the proof is similar to the first part of Step 5 in the proof of Lemma C.4. To be specific, there exist $N = 90^p$ p-dimensional unit vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ such that $\|\boldsymbol{b}\|_2 \le 2\sup_{k\in[N]} \boldsymbol{u}_k^\top \boldsymbol{b}$. The key here is to decompose $\boldsymbol{u}^\top \boldsymbol{b}$ as the sum of zero-mean independent random variables. Denote

$$X_{e,i} = \boldsymbol{u}^\top \bar{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{x}^{(e)} \left((\boldsymbol{x}^{(e)})^\top (\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*) - \varepsilon^{(e)} \right).$$

We have

$$\begin{aligned} \boldsymbol{u}^\top \boldsymbol{b} = \sum_{e\in\mathcal{E}} \sum_{i=1}^{n^{(e)}} \frac{\omega^{(e)}}{n^{(e)}} X_{e,i} &= \sum_{e\in\mathcal{E}} \sum_{i=1}^{n^{(e)}} \frac{\omega^{(e)}}{n^{(e)}} (X_{e,i} - \mathbb{E}[X_{e,i}]) + \sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[X_{e,1}] \\ &= \sum_{e\in\mathcal{E}} \sum_{i=1}^{n^{(e)}} \frac{\omega^{(e)}}{n^{(e)}} (X_{e,i} - \mathbb{E}[X_{e,i}]) \end{aligned}$$

where the last equality follows from the definition of $\bar{\boldsymbol{\beta}}^{\mathsf{R}}$ (C.3) that

$$\begin{aligned} \sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[X_{e,1}] &= \boldsymbol{u}^\top \bar{\boldsymbol{\Sigma}}^{-1/2} \bar{\boldsymbol{\Sigma}} (\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*) - \bar{\boldsymbol{\Sigma}}^{-1/2} \sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[\boldsymbol{x}^{(e)} \varepsilon^{(e)}] \\ &= \boldsymbol{u}^\top \bar{\boldsymbol{\Sigma}}^{1/2} \left(\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^* - \bar{\boldsymbol{\Sigma}}^{-1} \sum_{e\in\mathcal{E}} \omega^{(e)} \mathbb{E}[\boldsymbol{x}^{(e)} \varepsilon^{(e)}] \right) = 0. \end{aligned}$$

Note that $\boldsymbol{u}^\top \bar{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{x}^{(e)}$ is a zero-mean sub-Gaussian random variable with parameter σ_x, and $(\boldsymbol{x}^{(e)})^\top(\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*) - \varepsilon^{(e)}$ is the sum of two zero-mean sub-Gaussian random variables with parameter σ_ε and $\sigma_x \|\boldsymbol{\Sigma}^{1/2}(\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*)\|_2$, respectively. It then follows from Lemma C.1 and C.2 that

$$\mathbb{P}\left[|\boldsymbol{u}_k^\top \boldsymbol{b}| \ge C_5 \sigma_x \left(\sigma_\varepsilon + \sigma_x \|\boldsymbol{\Sigma}^{1/2}(\bar{\boldsymbol{\beta}}^{\mathsf{R}} - \boldsymbol{\beta}^*)\|_2\right)\left(\sqrt{\frac{u}{n_{\boldsymbol{\omega}}}} + \frac{u}{n_*}\right)\right] \ge 1 - 2e^{-u}.$$

Applying union bounds over all the k and setting $u = t + \log(2N)$ completes the proof. □

C.3 Proof of Theorem A.3

C.3.1 Key Proof Idea

We will use the following decomposition in the proof of Theorem A.3.

$$\begin{aligned}
\mathsf{Q}(\boldsymbol{\beta};\gamma,\boldsymbol{\omega}) - \mathsf{Q}(\boldsymbol{\beta}^*;\gamma,\boldsymbol{\omega}) &= \mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\boldsymbol{\beta}^*) + \gamma\left(\mathsf{J}(\boldsymbol{\beta}) - \mathsf{J}(\boldsymbol{\beta}^*)\right) \\
&= \{\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\boldsymbol{\beta}^*)\} - \left\{\widehat{\mathsf{R}}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}(\boldsymbol{\beta}^*)\right\} \\
&\quad + \left\{\widehat{\mathsf{R}}(\boldsymbol{\beta}) + \gamma\widehat{\mathsf{J}}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}(\boldsymbol{\beta}^*) - \gamma^*\widehat{\mathsf{J}}(\boldsymbol{\beta}^*)\right\} \\
&\quad + \gamma\left[\{\mathsf{J}(\boldsymbol{\beta}) - \mathsf{J}(\boldsymbol{\beta}^*)\} - \left\{\widehat{\mathsf{J}}(\boldsymbol{\beta}) - \widehat{\mathsf{J}}(\boldsymbol{\beta}^*)\right\}\right] \\
&= \mathsf{T}_{\mathsf{R}}(\boldsymbol{\beta},\boldsymbol{\beta}^*) + \mathsf{T}_{\mathsf{J}}(\boldsymbol{\beta},\boldsymbol{\beta}^*) + \widehat{\mathsf{Q}}(\boldsymbol{\beta};\gamma,\boldsymbol{\omega}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*;\gamma,\boldsymbol{\omega}).
\end{aligned} \tag{C.5}$$

The analysis of the first term T_{R} is standard. As stated in the following Lemma, we can provide an instance-dependent two-side bound for it.

Lemma C.3 (Instance-dependent Two-side Bound for R)**.** *Assume Condition A.1–A.4 hold. Define the event*

$$\begin{aligned}
\mathcal{A}_{1,t} = \Bigg\{\forall \boldsymbol{\beta} \in \mathbb{R}^p, \quad & \left|\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\boldsymbol{\beta}^*) - \widehat{\mathsf{R}}(\boldsymbol{\beta}) + \widehat{\mathsf{R}}(\boldsymbol{\beta}^*)\right| \\
& \le c_1\left(\kappa_U \sigma_x^2 \delta_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2^2 + \kappa_U^{1/2}\sigma_x\sigma_\varepsilon\delta_1\|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2\right)\Bigg\}.
\end{aligned} \tag{C.6}$$

with $\delta_1 = \sqrt{\frac{p+t}{n_{\boldsymbol{\omega}}}} + \frac{p+t}{n_*}$ and some universal constants c_1. We have

$$\mathbb{P}[\mathcal{A}_{1,t}] \ge 1 - e^{-t}$$

for any $t > 0$.

Proof of Lemma C.3. It concludes by applying Lemma C.7 with $s = p$. □

However, the analysis of the second term T_{J} is more involved. The following Lemma provides an instance-dependent one-side bound because all the statistical analysis only cares about the upper bound of $\mathsf{T}_{\mathsf{J}}(\boldsymbol{\beta},\boldsymbol{\beta}^*)$, or the lower bound of $-\mathsf{T}_{\mathsf{J}}(\boldsymbol{\beta},\boldsymbol{\beta}^*)$.

Lemma C.4 (Instance-dependent One-side Bound for J)**.** *Assume Condition A.1–A.4 hold. Define the* *event*

$$\begin{aligned}\mathcal{A}_{2,t} = \Bigg\{\forall \boldsymbol{\beta} \in \mathbb{R}^p, \quad & \frac{1}{c_1}\left(\mathsf{J}(\boldsymbol{\beta}) - \mathsf{J}(\boldsymbol{\beta}^*) - \widehat{\mathsf{J}}(\boldsymbol{\beta}) + \widehat{\mathsf{J}}(\boldsymbol{\beta}^*)\right) \\ & \le \|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2^2 \times \kappa_U^2 \sigma_x^2 \sqrt{\frac{p+t}{n_*}} \\ & \quad + \|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2 \times \kappa_U^{3/2} \sigma_x^2 \sigma_\varepsilon \left(\sqrt{\frac{p+t}{n_*}} + \sigma_x \frac{p + \log(2|\mathcal{E}|) + t}{\bar{n}}\right) \\ & \quad + \sqrt{\sum_{e\in\mathcal{E}} \omega^{(e)} \left\|\mathbb{E}[\boldsymbol{x}_{\mathrm{supp}(\boldsymbol{\beta})}^{(e)} \varepsilon^{(e)}]\right\|_2^2} \times \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \sqrt{\frac{p+t}{n_*}} \\ & \quad + |S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| \times \kappa_U \sigma_x^2 \sigma_\varepsilon^2 \frac{t + \log(2|\mathcal{E}||S^*|)}{\bar{n}} + \kappa_U \sigma_x \sigma_\varepsilon^2 \frac{p+t}{n_*} \Bigg\}\end{aligned} \tag{C.7}$$

for some universal constant c_1. *Then we have*

$$\mathbb{P}[\mathcal{A}_{2,t}] \ge 1 - 6e^{-t}. \tag{C.8}$$

for any $t \in (0, n_{\min} - \log(2|\mathcal{E}|) - p]$.

Proof of Lemma C.4. It concludes by applying Lemma C.6 with $s = p$ and observing that $n_{\boldsymbol{\omega}} \ge n_*$. □

We will use the following characterization of the population-level excess risk.

Proposition C.5. *Under Condition A.1, A.2 and A.5. If* $\gamma \ge 3\gamma^*$, *then*

$$\begin{aligned}\mathsf{Q}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega}) - \mathsf{Q}(\boldsymbol{\beta}^*; \gamma, \boldsymbol{\omega}) &\ge \frac{\kappa_L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \frac{\gamma}{3}\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) \\ &\ge \frac{\kappa_L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \frac{\gamma}{6}\kappa_L^2 \bar{\mathsf{d}}_{\mathrm{supp}(\boldsymbol{\beta})} + \frac{\gamma}{6}\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}).\end{aligned}$$

Proof of Proposition C.5. The proof is almost identical to the proof of Theorem A.1. From (B.4) we know $\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) \ge \kappa_L^2 \bar{\mathsf{d}}_{\mathrm{supp}(\boldsymbol{\beta})}$. Following the notations in the proof of Theorem A.1, we have

$$\begin{aligned}\mathsf{Q}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega}) - \mathsf{Q}(\boldsymbol{\beta}^*; \gamma, \boldsymbol{\omega}) &\ge \mathsf{T}_1(\boldsymbol{\beta}) + \gamma \mathsf{T}_2(\boldsymbol{\beta}) \\ &\ge \frac{\kappa_L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - 2\kappa_L^{-1}\mathsf{b}_{\mathrm{supp}(\boldsymbol{\beta})} + \frac{2}{3}\gamma \bar{\mathsf{d}}_{\mathrm{supp}(\boldsymbol{\beta})} + \frac{\gamma}{3}\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) \\ &\ge \frac{\kappa_L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \frac{\gamma}{3}\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega})\end{aligned}$$

This completes the proof. □

With the help of the above claims, we are ready to prove Theorem A.3.

C.3.2 Proof of the Variable Selection Property (A.11)

It follows from our decomposition (C.5) that, for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\widehat{\mathsf{Q}}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*; \gamma, \boldsymbol{\omega}) \ge -\mathsf{T}_{\mathsf{J}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) - \mathsf{T}_{\mathsf{R}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \mathsf{Q}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega}) - \mathsf{Q}(\boldsymbol{\beta}^*; \gamma, \boldsymbol{\omega}). \tag{C.9}$$

The rest of the proof proceeds conditioned on that $n_{\min} > \log(2|\mathcal{E}|) + p + t$ such that the results of Lemma C.4 is applicable. Recall that $\gamma \ge 1 + \kappa_L$. Applying Lemma C.3 and Lemma C.4, we have that, under the event $\mathcal{A}_{1,t} \cap \mathcal{A}_{2,t}$,

$$
\begin{aligned}
\mathsf{T}_{\mathsf{J}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \mathsf{T}_{\mathsf{R}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \le C\gamma \Bigg\{ & \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \times \kappa_U^2 \sigma_x^2 \sqrt{\frac{p+t}{n_*}} \\
& + \|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2 \times \kappa_U^{3/2} \sigma_x^2 \sigma_\varepsilon \left(\sqrt{\frac{p+t}{n_*}} + \sigma_x \frac{p + \log(2|\mathcal{E}|) + t}{\bar{n}} \right) \\
& + \sqrt{\sum_{e \in \mathcal{E}} \omega^{(e)} \left\| \mathbb{E}[\boldsymbol{x}_{\mathrm{supp}(\boldsymbol{\beta})}^{(e)} \varepsilon^{(e)}] \right\|_2^2} \times \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \sqrt{\frac{p+t}{n_*}} \\
& + |S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| \times \kappa_U \sigma_x^2 \sigma_\varepsilon^2 \frac{t + \log(2|\mathcal{E}||S^*|)}{\bar{n}} + \kappa_U \sigma_x \sigma_\varepsilon^2 \frac{p+t}{n_*} \Bigg\} \\
= C\gamma(\mathsf{I}_1 + \mathsf{I}_2 + \mathsf{I}_3 + \mathsf{I}_4). &
\end{aligned} \tag{C.10}
$$

Applying Young's inequality gives

$$
\mathsf{I}_2 \le \frac{\kappa_L}{16C\gamma} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \frac{8C\gamma}{\kappa_L} \kappa_U^3 \sigma_x^4 \sigma_\varepsilon^2 \left\{ \frac{p+t}{n_*} + \sigma_x^2 \left(\frac{p + \log(2|\mathcal{E}|) + t}{\bar{n}} \right)^2 \right\}. \tag{C.11}
$$

At the same time, it follows from Lemma C.10 that

$$
\begin{aligned}
\mathsf{I}_3 &\le \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \sqrt{\frac{p+t}{n_*}} \sqrt{2\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) + 2\kappa_U^2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2} \\
&\le \frac{1}{6C} \mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) + 6C\kappa_U \sigma_x^2 \sigma_\varepsilon^2 \frac{p+t}{n_*} + \frac{\kappa_L}{16C\gamma} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \frac{16C\gamma}{\kappa_L} \kappa_U^3 \sigma_x^2 \sigma_\varepsilon^2 \frac{p+t}{n_*}
\end{aligned}
$$

Using the fact that

$$
\begin{aligned}
|S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| \min_{j \in S^*} |\beta_j^*|^2 &= \sum_{j \in S^*, \beta_j = 0} \min_{j \in S^*} |\beta_j^*|^2 \\
&\le \sum_{j \in S^*, \beta_j = 0} |\beta_j - \beta_j^*|^2 \le \sum_{j=1}^{p} |\beta_j - \beta_j^*|^2 = \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2,
\end{aligned}
$$

we obtain

$$
\begin{aligned}
& |S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| \times \kappa_U \sigma_x^2 \sigma_\varepsilon^2 \frac{t + \log(2|\mathcal{E}||S^*|)}{\bar{n}} \\
&\quad = \left(|S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| \min_{j \in S^*} |\beta_j^*|^2 \right) \times \left\{ \frac{\kappa_U \sigma_x^2 \sigma_\varepsilon^2}{\min_{j \in S^*} |\boldsymbol{\beta}_j^*|^2} \frac{t + \log(2|\mathcal{E}||S^*|)}{\bar{n}} \right\} \\
&\quad \le \|\boldsymbol{\beta} - \boldsymbol{\beta}_*\|_2^2 \left\{ \frac{\kappa_U \sigma_x^2 \sigma_\varepsilon^2}{\min_{j \in S^*} |\boldsymbol{\beta}_j^*|^2} \frac{t + \log(2|\mathcal{E}||S^*|)}{\bar{n}} \right\},
\end{aligned} \tag{C.12}
$$

Putting these pieces together, if

$$
\bar{n} \ge \frac{16C\gamma\kappa_U \sigma_x^2 \sigma_\varepsilon^2 \{t + \log(2|\mathcal{E}||S^*|)\}}{\kappa_L \min_{j \in S^*} |\beta_j^*|^2} \quad \text{and} \quad n_* \ge (p+t) \left(\frac{16C\kappa_U^2 \sigma_x^2}{\kappa_L} \gamma \right)^2, \tag{C.13}
$$

then we find that, under $\mathcal{A}_{1,t} \cap \mathcal{A}_{2,t}$,

$$
\begin{aligned}
\mathsf{T}_{\mathsf{J}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \mathsf{T}_{\mathsf{R}}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \le \frac{\kappa_L}{4} \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \frac{\gamma}{6} \mathsf{J}(\boldsymbol{\beta}; \omega) + C_1 \gamma \Bigg\{ & \frac{\gamma}{\kappa_L} \kappa_U^3 \sigma_x^4 \sigma_\varepsilon^2 \frac{p+t}{n_*} \\
& + \frac{\gamma}{\kappa_L} \kappa_U^3 \sigma_x^6 \sigma_\varepsilon^2 \left(\frac{p + \log(2|\mathcal{E}|) + t}{\bar{n}} \right)^2 \Bigg\}.
\end{aligned}
$$

for another universal constant C_1.

Plugging the above inequality back into (C.9) and applying Proposition C.5 provided our choice of $\gamma \geq 3\gamma^*$, we establish that, if $\mathcal{A}_{1,t} \cap \mathcal{A}_{2,t}$ occurs, then for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\begin{aligned}\widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \geq \frac{\kappa_L}{4}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 &+ \frac{\gamma}{6}\kappa_L^2 \bar{\mathsf{d}}_{\mathrm{supp}(\boldsymbol{\beta})} \\ &- C_2\left\{\frac{\gamma^2}{\kappa_L}\kappa_U^3\sigma_x^6\sigma_\varepsilon^2\left(\frac{p+\log(2|\mathcal{E}|)+t}{\bar{n}}\right)^2 + \frac{\gamma^2}{\kappa_L}\kappa_U^3\sigma_x^4\sigma_\varepsilon^2\frac{p+t}{n_*}\right\}.\end{aligned}$$

The rest of the proof proceeds conditioned on the event $\mathcal{A}_{1,t} \cap \mathcal{A}_{2,t}$. We first argue that if

$$\frac{n_*}{p+t} \geq \frac{18C_2\kappa_U^3\sigma_x^4\sigma_\varepsilon^2}{(\kappa_L^3/\gamma)\mathsf{s}_-} \quad \text{and} \quad \frac{\bar{n}}{(p+\log(2|\mathcal{E}|)+t)} \geq \frac{\sqrt{18C_2}\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon}{\sqrt{(\kappa_L^3/\gamma)\mathsf{s}_-}}, \tag{C.14}$$

with $\mathsf{s}_- = \min_{S \cap G_{\boldsymbol{\omega}} \neq \emptyset} \bar{\mathsf{d}}_S$, then for any $\boldsymbol{\beta}$ with $\mathrm{supp}(\boldsymbol{\beta}) \cap G_{\boldsymbol{\omega}} \neq \emptyset$, the following holds,

$$\widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \geq \frac{\kappa_L}{4}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \left(\frac{\gamma}{6} - 2 \times \frac{\gamma}{18}\right)\kappa_L^2\mathsf{s}_- > 0,$$

which means $\boldsymbol{\beta}$ is not the empirical risk minimizer. This implies that $\mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}) \subseteq G_{\boldsymbol{\omega}}^c$.

Meantime, denote $\mathsf{s}_+ = \min_{j \in S^*} |\beta_j^*|^2$. We then argue that if

$$\frac{n_*}{p+t} \geq \frac{12C_2\gamma^2\kappa_U^3\sigma_x^4\sigma_\varepsilon^2}{\kappa_L^2\mathsf{s}_+} \quad \text{and} \quad \frac{\bar{n}}{(p+\log(2|\mathcal{E}|)+t)} \geq \frac{\sqrt{12C_2}\gamma\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon}{\sqrt{\kappa_L^2\mathsf{s}_+}} \tag{C.15}$$

then for any $\boldsymbol{\beta}$ with $S^* \nsubseteq \mathrm{supp}(\boldsymbol{\beta})$, the following holds

$$\widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \geq \frac{\kappa_L}{4}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - 2 \times \frac{\kappa_L}{12}\min_{j \in S^*}|\beta_j^*|^2 \geq \frac{\kappa_L}{12}\min_{j \in S^*}|\beta_j^*|^2 > 0,$$

which also means that $\boldsymbol{\beta}$ is not the empirical risk minimizer, hence implying the $S^* \subseteq \mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{Q}})$.

Combining the conditions on n_* and $\bar{n}$ in (C.13), (C.14) and (C.15) together, we can conclude that if

$$\frac{\bar{n}}{p+\log(|\mathcal{E}|)+t} \geq \underbrace{(18C_2\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon) \vee (16C\kappa_U\sigma_x^2\sigma_\varepsilon^2)}_{c_1}\frac{\gamma}{\kappa_L}\left(\sqrt{\frac{1}{\mathsf{s}_+ \wedge (\gamma\kappa_L\mathsf{s}_-)}} + \frac{1}{\mathsf{s}_+}\right) \tag{C.16}$$

and

$$\frac{n_*}{p+t} \geq (\gamma/\kappa_L)^2\left\{\underbrace{18C_2\kappa_U^3\sigma_x^4\sigma_\varepsilon^2}_{c_2}\frac{1}{\mathsf{s} \wedge (\gamma\kappa_L\mathsf{s}_-)} + \underbrace{(16C^2)^2\kappa_U^4\sigma_x^4}_{c_3}\right\} \tag{C.17}$$

then under $\mathcal{A}_{1,t} \cap \mathcal{A}_{2,t}$ that occurs with probability at least $1 - e^{-7}$ due to Lemma C.4, Lemma C.3 and union bound, the following holds

$$S^* \subseteq \mathrm{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}) \subseteq G_{\boldsymbol{\omega}}^c. \tag{C.18}$$

This completes the proof. □

C.4 Proof of Theorem A.4

C.4.1 Proof of the Rate (A.12)

The proof is very similar to that of Theorem A.3 but will use Lemma C.4 and the strong convexity of Q around $\boldsymbol{\beta}^*$ in a different way. The proof proceeds conditioned on $\mathcal{A}_{1,t} \cap \mathcal{A}_{2,t}$. It follows from Lemma C.4 and Lemma C.3 that the inequality in (C.10) holds. We will bound the term I_4 differently.

For first term of I_4, it also follows from the fact $xy \le \frac{1}{2}(x^2+y^2)$ that

$$
\begin{aligned}
&|S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| \times \kappa_U \sigma_x^2 \sigma_\varepsilon^2 \frac{t+\log(2|\mathcal{E}||S^*|)}{\bar{n}} \\
&\quad = \left(\sqrt{|S^* \setminus \mathrm{supp}(\boldsymbol{\beta})|} \min_{j\in S^*} |\beta_j^*| \sqrt{\kappa_L/16C\gamma}\right) \times \left\{ \frac{2\sqrt{|S^*|}\kappa_U \sigma_x^2 \sigma_\varepsilon^2}{\sqrt{\kappa_L/16C\gamma}\min_{j\in S^*}|\boldsymbol{\beta}_j^*|} \frac{t+\log(2|\mathcal{E}||S^*|)}{\bar{n}} \right\} \\
&\quad \le \frac{\kappa_L}{16C\gamma} |S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| \min_{j\in S^*} |\boldsymbol{\beta}_j^*|^2 + \frac{16C\gamma\kappa_U^2\sigma_x^4\sigma_\varepsilon^2 |S^*| \{t+\log(2|\mathcal{E}||S^*|)\}^2}{\kappa_L(\min_{j\in S^*}|\beta_j^*|^2)\bar{n}^2} \\
&\quad \le \frac{\kappa_L}{16C\gamma} \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 + 16C\kappa_U^2\sigma_x^4\sigma_\varepsilon^2 \frac{(\gamma/\kappa_L)|S^*|}{\min_{j\in S^*}|\beta_j^*|^2} \left(\frac{t+\log(2|\mathcal{E}||S^*|)}{\bar{n}}\right)^2,
\end{aligned}
$$

At the same time, we also have $\mathsf{I}_1 \le \frac{\kappa_L}{16}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2$ if

$$
n_* \ge \underbrace{(16C\kappa_U\sigma_x^2)}_{c_1}(p+t)\,(\gamma/\kappa_L)^2. \tag{C.19}
$$

Putting these pieces together with the error bounds we have for I_2–I_3 in the proof of Theorem A.3, the following holds for all the $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$
\begin{aligned}
\mathsf{T}_\mathsf{R}(\boldsymbol{\beta},\boldsymbol{\beta}^*) + \mathsf{T}_\mathsf{J}(\boldsymbol{\beta},\boldsymbol{\beta}^*) \le 4 \times \frac{\kappa_L}{16}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 + \frac{\gamma}{6}\mathsf{J}(\boldsymbol{\beta};\omega) \\
+ C_1 \frac{\gamma^2}{\kappa_L}\kappa_U^3\sigma_x^4\sigma_\varepsilon^2 \left\{ \frac{p+t}{n_*} + \sigma_x^2 \left(\frac{p+\log(2|\mathcal{E}|)+t}{\bar{n}}\right)^2 \right\} \\
+ C_1 \frac{\gamma^2}{\kappa_L}\kappa_U^2\sigma_x^4\sigma_\varepsilon^2 \frac{|S^*|}{\min_{j\in S^*}|\beta_j^*|^2} \left(\frac{t+\log(2|\mathcal{E}||S^*|)}{\bar{n}}\right)^2
\end{aligned}
$$

Plugging $\widehat{\boldsymbol{\beta}}_\mathsf{Q}$ into the decomposition (C.5), it then follows from Proposition C.5 that

$$
\begin{aligned}
\frac{\kappa_L}{2}\|\widehat{\boldsymbol{\beta}}_\mathsf{Q}-\boldsymbol{\beta}^*\|_2^2 + \frac{\gamma}{6}\mathsf{J}(\widehat{\boldsymbol{\beta}}_\mathsf{Q};\boldsymbol{\omega}) + 0 &\le \mathsf{Q}(\widehat{\boldsymbol{\beta}}_\mathsf{Q};\gamma,\boldsymbol{\omega}) - \mathsf{Q}(\widehat{\boldsymbol{\beta}}_\mathsf{Q};\gamma,\boldsymbol{\omega}) \\
&= \widehat{\mathsf{Q}}(\widehat{\boldsymbol{\beta}}_\mathsf{Q};\gamma,\boldsymbol{\omega}) - \widehat{\mathsf{Q}}(\widehat{\boldsymbol{\beta}}_\mathsf{Q};\gamma,\boldsymbol{\omega}) + \mathsf{T}_\mathsf{R}(\widehat{\boldsymbol{\beta}}_\mathsf{Q},\boldsymbol{\beta}^*) + \mathsf{T}_\mathsf{J}(\widehat{\boldsymbol{\beta}}_\mathsf{Q},\boldsymbol{\beta}^*) \\
&\le 0 + \mathsf{T}_\mathsf{R}(\widehat{\boldsymbol{\beta}}_\mathsf{Q},\boldsymbol{\beta}^*) + \mathsf{T}_\mathsf{J}(\widehat{\boldsymbol{\beta}}_\mathsf{Q},\boldsymbol{\beta}^*)
\end{aligned}
$$

Combining it with the upper bound on $\mathsf{T}_\mathsf{R}(\boldsymbol{\beta},\boldsymbol{\beta}^*) + \mathsf{T}_\mathsf{J}(\boldsymbol{\beta},\boldsymbol{\beta}^*)$ we derived, we conclude that

$$
\begin{aligned}
\frac{\|\widehat{\boldsymbol{\beta}}_\mathsf{Q}-\boldsymbol{\beta}^*\|_2^2}{\sigma_\varepsilon^2} &\le \underbrace{4C_2\kappa_U^3\sigma_x^6}_{c_2^2} \frac{\gamma^2}{\kappa_L^2}\left(\frac{p+t}{n_*} + \frac{(p+\log(2|\mathcal{E}|)+t)^2}{\bar{n}^2}\right) \\
&\quad + \underbrace{4C_2\kappa_U^2\sigma_x^4}_{c_3^2} \frac{(\gamma^2/\kappa_L^2)|S^*|}{\min_{j\in S^*}|\beta_j^*|^2}\left(\frac{t+\log(2|\mathcal{E}||S^*|)}{\bar{n}}\right)^2
\end{aligned} \tag{C.20}
$$

under $\mathcal{A}_{1,t} \cap \mathcal{A}_{2,t}$. This completes the proof. □

C.4.2 Proof of the Rate (A.13)

We assume that conditions (C.16) and (C.17) are satisfied. Thus, the variable selection property (C.18) also holds. Now that we only focus on the case of no pooled linear spurious variables. The covariate dimension is $p_0 = |(G_{\boldsymbol{\omega}})^c|$. Now we apply Lemma C.3-C.4 with $\boldsymbol{x}_{(G_{\boldsymbol{\omega}})^c}$. When $\mathcal{A}_{1,t} \cap \mathcal{A}_{2,t} \cap \mathcal{A}_{1,t}^{p_0} \cap \mathcal{A}_{2,t}^{p_0}$ occurs, the inequality (C.10) with $p = p_0$ also holds for $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$. Observe that $|S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| = 0$ given the variable selection property (C.18). The rest of the proof follows similarly to that for (A.12).

C.5 Proof of Theorem A.5

We first define some additional notations. Define $s^* = |S^*|$. Let

$$V(s,t) = s\log(ep/s) + s^* + t,$$

and define $0 \times \log(1/0) = 0$. Similar to the low-dimension counterpart, we also need the following two Lemmas.

Lemma C.6 (Instance-dependent One-side Bound for J, High-dimensional)**.** *Assume Conditions A.1 – A.4 hold. For any $1 \le s \le p$, define the following event,*

$$\begin{aligned}\mathcal{A}_{3,t}(s) = \Bigg\{ \forall \boldsymbol{\beta} \in \mathcal{B}_s, \quad & \frac{1}{c_1}\left(\mathsf{J}(\boldsymbol{\beta}) - \mathsf{J}(\boldsymbol{\beta}^*) - \widehat{\mathsf{J}}(\boldsymbol{\beta}) + \widehat{\mathsf{J}}(\boldsymbol{\beta}^*)\right) \\ & \le \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \times \kappa_U^2 \sigma_x^2 \left(\sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*}\right) \\ & \quad + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \times \kappa_U^{3/2} \sigma_x^2 \sigma_\varepsilon \left(\sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*}\right) \\ & \quad + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \times \kappa_U^{3/2} \sigma_x^3 \sigma_\varepsilon \frac{\sqrt{V(s, t + \log(2|\mathcal{E}|)) \cdot V(0, t + \log(2|\mathcal{E}|))}}{\bar{n}} \\ & \quad + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \times \kappa_U^{3/2} \sigma_x^3 \sigma_\varepsilon \frac{V(s, t + \log(2|\mathcal{E}|))\sqrt{V(0, t + \log(2|\mathcal{E}|))}}{n_\dagger} \\ & \quad + \sqrt{\sum_{e \in \mathcal{E}} \omega^{(e)} \left\| \mathbb{E}[\boldsymbol{x}_{\mathrm{supp}(\boldsymbol{\beta})}^{(e)} \varepsilon^{(e)}] \right\|_2^2} \times \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \sqrt{\frac{V(|\mathrm{supp}(\boldsymbol{\beta}) \setminus S^*|, t)}{n_*}} \\ & \quad + |S^* \setminus \mathrm{supp}(\boldsymbol{\beta})| \times \kappa_U \sigma_x^2 \sigma_\varepsilon^2 \frac{t + \log(2s^*|\mathcal{E}|)}{\bar{n}} \\ & \quad + \kappa_U \sigma_x \sigma_\varepsilon^2 \frac{V(|\mathrm{supp}(\boldsymbol{\beta}) \setminus S^*|, t)}{n_*} \Bigg\}\end{aligned} \tag{C.21}$$

for some universal constant c_1. Then we have $\mathbb{P}[\mathcal{A}_{3,t}(s)] \ge 1 - 6e^{-t}$ for any $t \in (0, n_{\min} - \log(2|\mathcal{E}|) - s^*]$.

Proof of Lemma C.6. See Appendix C.6. □

Lemma C.7 (Instance-dependent Two-side Bound for R, High-dimensional)**.** *Assume Conditions A.1 – A.4 hold. Define the event*

$$\begin{aligned}\mathcal{A}_{4,t}(s) = \Bigg\{ \forall \boldsymbol{\beta} \in \mathcal{B}_s, \quad & \left|\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\boldsymbol{\beta}^*) - \widehat{\mathsf{R}}(\boldsymbol{\beta}) + \widehat{\mathsf{R}}(\boldsymbol{\beta}^*)\right| \\ & \le c_1 \left(\kappa_U \sigma_x^2 \delta_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \delta_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\right) \Bigg\}.\end{aligned} \tag{C.22}$$

with $\delta_1 = \sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*}$ *and some universal constants* c_1*. We have* $\mathbb{P}[\mathcal{A}_{4,t}(s)] \geq 1 - e^{-t}$ *for any* $t > 0$*.*

Proof of Lemma C.7. See Appendix C.7. □

We start with a lemma stating that if we choose some large λ, then the $\widehat{S} = \operatorname{supp}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}})$ will satisfy $|\widehat{S}| \leq 2s^*$ with high probability.

Proposition C.8. *Let* $0 < t \leq n_{\min} - \log(2ep|\mathcal{E}|) - s^*$ *be arbitrary. Assume Conditions A.1–A.5 hold, and* $\gamma \geq 3\gamma^* \vee 1$*. Let* $\zeta = \log(ep/s^*) + t$*. Suppose further that* $\log(|\mathcal{E}|) \leq c_1 \log p$ *for some universal constant* $c_1 > 0$*. There exist some universal constants* c_2*–*c_3 *depending only on* C *such that if*

$$\begin{aligned}
\bar{n} &\geq c_2 \kappa_U \sigma_x^2 \gamma (t + \log p) \left\{ s^* + \sigma_\varepsilon^2 / (\kappa_L \min_{j \in S^*} |\beta_j^*|^2) \right\}, \\
n_* &\geq c_2 \kappa_U^4 \sigma_x^4 (\gamma/\kappa_L)^2 \zeta s^*, \\
n_\dagger &\geq c_2 \kappa_U^{3/2} \sigma_x^2 (\gamma/\kappa_L) \zeta s^* \sqrt{\zeta + s^*},
\end{aligned}$$

and the choice of λ *satisfies*

$$\lambda \geq c_3 \sigma_x^2 \sigma_\varepsilon^2 (\gamma/\kappa_L) \zeta \left\{ \frac{\kappa_U^4 \sigma_x^2 (\gamma/\kappa_L)}{n_*} + \frac{\kappa_U^{3/2} \sqrt{\zeta + s^*}}{n_\dagger} + \frac{\kappa_U^3 \sigma_x^2 (\gamma/\kappa_L)(\zeta + s^*)}{\bar{n}^2} \right\}.$$

Then the following holds

$$\mathbb{P}\left[|\widehat{S}| \leq 2s^* \right] \geq 1 - 3e^{-t}.$$

Proof of Proposition C.8. See Appendix C.8. □

We are ready to prove Theorem A.5.

Proof of Theorem A.5. We prove the Theorem in a more general form, including the tail probability t. Denote $\zeta = \log(ep/s^*) + t$. We first provide a detailed condition for different sample sizes presented in Condition A.6,

$$\begin{aligned}
n_{\min} &\geq C^* (s^* + \log p) \\
n_* &\geq C^* \left\{ \kappa_U^4 \sigma_x^4 (\gamma/\kappa_L)^2 \zeta s^* \right\} \vee \left\{ \kappa_U^3 \sigma_x^4 \sigma_\varepsilon^2 \zeta s^* (\gamma/\kappa_L)^2 (1/\kappa_L \beta_{\min}^2) \right\} \\
\bar{n} &\geq C^* \left\{ \kappa_U \sigma_x^2 \gamma (t + \log p) \left\{ s^* + \sigma_\varepsilon^2 / (\kappa_L \beta_{\min}^2) \right\} \right\} \vee \left\{ \kappa_U^{3/2} \sigma_x^2 \sigma_\varepsilon (\gamma/\kappa_L) \frac{\sqrt{s^* \zeta (s^* + \zeta)}}{\kappa_L \beta_{\min}} \right\} \\
n_\dagger &\geq C^* \kappa_U^{3/2} \sigma_x^2 (\gamma/\kappa_L) s^* \zeta \sqrt{s^* + \zeta} \{1 + \sigma_\varepsilon / (\sqrt{\kappa_L} \beta_{\min})\}
\end{aligned} \tag{C.23}$$

together with a detailed lower bound on the regularization parameter λ

$$\begin{aligned}
\lambda \geq C^* \kappa_U^3 \sigma_x^4 \sigma_\varepsilon^2 (\gamma/\kappa_L)^2 \Bigg\{ & \left(\frac{\kappa_U \zeta}{n_*} \vee \frac{s^* \zeta}{n_*} \right) + \frac{s^* \zeta (s^* + \zeta)}{\bar{n}^2} \\
& + \kappa_U^{-3/2} \sigma_x^{-2} (\gamma/\kappa_L)^{-1} \frac{\zeta \sqrt{s^* + \zeta}}{n_\dagger} + \frac{(s^* \zeta)^2 (s^* + \zeta)}{n_\dagger^2} \Bigg\}.
\end{aligned} \tag{C.24}$$

Define the two events

$$\mathcal{C}_{1,t} = \left\{ |\widehat{S}| \leq 2s^* \right\} \quad \text{and} \quad \mathcal{C}_{2,t} = \mathcal{A}_{3,t}(2s^*) \cap \mathcal{A}_{4,t}(2s^*).$$

We can see that the conditions in Lemma C.6, Lemma C.7, and Proposition C.8 are satisfied by (C.23) and (C.24) with large universal constant $C^* > 0$. Denote $\phi_* = s^* + \log(ep/s^*) + t = s^* + \zeta$ and $\phi_s = s^* \log(ep/s^*) + t \le s^*\zeta$. We follow a similar strategy as that in *Case 1* of Proposition C.8. Under $\mathcal{C}_{1,t} \cap \mathcal{C}_{2,t}$, it follows our choice of γ and Proposition C.5 that, there exists some universal constant $C_1 > 0$,

$$
\begin{aligned}
\widehat{\mathsf{Q}}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) &= \widehat{\mathsf{Q}}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) - \mathsf{Q}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) + \mathsf{Q}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) - \mathsf{Q}(\boldsymbol{\beta}^*) + \mathsf{Q}(\boldsymbol{\beta}^*) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \\
&\ge \frac{\kappa_L}{2}\|\widehat{\boldsymbol{\beta}}_{\mathsf{L}} - \boldsymbol{\beta}^*\|_2^2 + \frac{\gamma}{6}\mathsf{J}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}};\boldsymbol{\omega}) + \widehat{\mathsf{R}}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) - \mathsf{R}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) + \mathsf{R}(\boldsymbol{\beta}^*) - \widehat{\mathsf{R}}(\boldsymbol{\beta}^*) \\
&\qquad - \gamma\left(\mathsf{J}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) - \widehat{\mathsf{J}}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) - \mathsf{J}(\boldsymbol{\beta}^*) + \widehat{\mathsf{J}}(\boldsymbol{\beta}^*)\right) \\
&\overset{(a)}{\ge} \left(\frac{\kappa_L}{2} - C_1\kappa_U^2\sigma_x^2\gamma\sqrt{\frac{\phi_s}{n_*}}\right)\|\widehat{\boldsymbol{\beta}}_{\mathsf{L}} - \boldsymbol{\beta}^*\|_2^2 \\
&\qquad + \frac{\gamma}{6}\mathsf{J}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}};\boldsymbol{\omega}) - C_1\gamma\sqrt{\sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\mathbb{E}[\boldsymbol{x}_{\widehat{S}}^{(e)}\varepsilon^{(e)}]\right\|_2^2}\kappa_U^{1/2}\sigma_x\sigma_\varepsilon\sqrt{\frac{|\widehat{S}\setminus S^*|\log(ep/s^*) + s^* + t}{n_*}} \\
&\qquad - C_1\|\widehat{\boldsymbol{\beta}}_{\mathsf{L}} - \boldsymbol{\beta}^*\|_2 \times \gamma\kappa_U^{3/2}\sigma_x^2\sigma_\varepsilon\sqrt{\frac{\phi_s}{n_*}} \\
&\qquad - C_1\|\widehat{\boldsymbol{\beta}}_{\mathsf{L}} - \boldsymbol{\beta}^*\|_2 \times \gamma\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon\left(\frac{\sqrt{\phi_*}\sqrt{\phi_s}}{\bar{n}} + \frac{\phi_s\sqrt{\phi_*}}{n_\dagger}\right) \\
&\qquad - C_1\gamma|S^*\setminus\widehat{S}|\kappa_U\sigma_x^2\sigma_\varepsilon^2\frac{t+\log p}{\bar{n}} - C_1\gamma\kappa_U\sigma_x\sigma_\varepsilon^2\frac{\phi_s}{n_*} \\
&= \mathsf{I}_1 + \mathsf{I}_2 + \mathsf{I}_3 + \mathsf{I}_4 + \mathsf{I}_5,
\end{aligned}
$$

where (a) follows from the fact that $\mathcal{C}_{1,t} \cap \mathcal{C}_{2,t}$ holds such that we can apply the result of Lemma C.6 and Lemma C.7 with $s = 2s^*$ to $\widehat{\boldsymbol{\beta}}_{\mathsf{L}}$. Denote $\Diamond = 0.5(6C_1)^2(\gamma/\kappa_L)^2\kappa_U^3\sigma_x^4\sigma_\varepsilon^2$. Following a similar strategy of lower bounds on $\mathsf{I}_1 - \mathsf{I}_5$ and using the fact that $\phi_s \le 2s^*\zeta$ and

$$
\mathsf{I}_2 \ge -\frac{\kappa_L}{12}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - 4\Diamond\frac{|\widehat{S}\setminus S^*|\zeta}{n_*},
$$

we obtain

$$
\begin{aligned}
&\widehat{\mathsf{Q}}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) + \lambda\|\widehat{\boldsymbol{\beta}}_{\mathsf{L}}\|_0 - \lambda\|\boldsymbol{\beta}^*\|_0 \\
&\quad = \widehat{\mathsf{Q}}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) + \lambda|\widehat{S}\setminus S^*| - \lambda|S^*\setminus\widehat{S}| \\
&\quad \ge \frac{\kappa_L}{12}\|[\widehat{\boldsymbol{\beta}}_{\mathsf{L}}]_{\widehat{S}} - [\boldsymbol{\beta}^*]_{\widehat{S}}\|_2^2 \\
&\qquad\quad + \frac{\kappa_L}{12}\|[\widehat{\boldsymbol{\beta}}_{\mathsf{L}}]_{S^*\setminus\widehat{S}} - [\boldsymbol{\beta}^*]_{S^*\setminus\widehat{S}}\|_2^2 - \lambda|S^*\setminus\widehat{S}| \\
&\qquad\quad + \frac{\lambda}{2}|\widehat{S}\setminus S^*| - 4\Diamond\cdot\frac{\zeta}{n_*}|\widehat{S}\setminus S^*| \\
&\qquad\quad + \frac{\lambda}{2}|\widehat{S}\setminus S^*| - 2\Diamond\cdot(2s^*)\zeta\left(\frac{1}{n_*} + \frac{\phi_*}{\bar{n}^2} + \frac{2s^*\phi_*\zeta^2}{n_\dagger^2}\right) \\
&\quad \overset{(a)}{\ge} |S^*\setminus\widehat{S}|\left(\frac{\kappa_L}{12}\min_{j\in S^*}|\beta_j^*|^2 - \lambda\right) \\
&\qquad\quad + \frac{\lambda}{2}|\widehat{S}\setminus S^*| - 4\Diamond\cdot s^*\zeta\left(\frac{1}{n_*} + \frac{\phi_*}{\bar{n}^2} + \frac{2s^*\phi_*\zeta}{n_\dagger^2}\right) + \left(\frac{\lambda}{2} - 4\Diamond\cdot\frac{\zeta}{n_*}\right)|\widehat{S}\setminus S^*|,
\end{aligned}
$$

Here (a) follows from the facts $\|[\widehat{\boldsymbol{\beta}}_{\mathsf{L}}]_{\widehat{S}} - [\boldsymbol{\beta}^*]_{\widehat{S}}\|_2^2 \ge 0$, $\|[\widehat{\boldsymbol{\beta}}_{\mathsf{L}}]_{S^*\setminus\widehat{S}} - [\boldsymbol{\beta}^*]_{S^*\setminus\widehat{S}}\|_2^2 \ge |S^*\setminus\widehat{S}|\min_{j\in S^*}|\beta_j^*|$. We argue that if

$$\frac{\kappa_L}{36}\min_{j\in S^*}|\beta_j^*|^2 \ge 4\Diamond\cdot s^*\zeta\left(\frac{1}{n_*}+\frac{\phi_*}{\bar{n}^2}+\frac{2s^*\phi_*\zeta}{n_\dagger^2}\right), \tag{C.25}$$

which is satisfied by Condition A.6, and $\lambda \le \frac{\kappa_L}{36}\min_{j\in S^*}|\beta_j^*|^2$, then $|S^*\backslash\widehat{S}|=0$. This is because if $|S^*\backslash\widehat{S}|\ge 1$, we will have

$$\widehat{\mathsf{Q}}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}})-\widehat{\mathsf{Q}}(\boldsymbol{\beta}^*)+\lambda\|\widehat{\boldsymbol{\beta}}_{\mathsf{L}}\|_0-\lambda\|\boldsymbol{\beta}^*\|_0 \ge |S^*\setminus\widehat{S}|\frac{\kappa_L}{36}\min_{j\in S^*}|\beta_j^*|^2>0.$$

This is contrary to the fact that $\widehat{\boldsymbol{\beta}}_{\mathsf{L}}$ is the minimizer of (3.9). Moreover, we also have $|\widehat{S}\setminus S^*|=0$ if

$$\lambda \ge 9\Diamond\cdot s^*\zeta\left(\frac{1}{n_*}+\frac{\phi_*}{\bar{n}^2}+\frac{2s^*\phi_*\zeta}{n_\dagger^2}\right). \tag{C.26}$$

This is because if $|\widehat{S}\setminus S^*|\ge 1$, we will have

$$\widehat{\mathsf{Q}}(\widehat{\boldsymbol{\beta}}_{\mathsf{L}})-\widehat{\mathsf{Q}}(\boldsymbol{\beta}^*)+\lambda\|\widehat{\boldsymbol{\beta}}_{\mathsf{L}}\|_0-\lambda\|\boldsymbol{\beta}^*\|_0 \ge \frac{\lambda}{2}|\widehat{S}\setminus S^*|-4\Diamond\cdot s^*\zeta\left(\frac{1}{n_*}+\frac{\phi_*}{\bar{n}^2}+\frac{2s^*\phi_*\zeta}{n_\dagger^2}\right)>0,$$

which also contradicts the fact that $\widehat{\boldsymbol{\beta}}_{\mathsf{L}}$ minimizes (3.9). In conclusion, letting $\lambda\le\frac{\kappa_L}{36}\min_{j\in S^*}|\beta_j^*|^2$ and choosing some large enough constant C^* such that (C.25) and (C.26) are satisfied by (C.23) and (C.24), we can then argue that under the event $\mathcal{C}_{1,t}\cap\mathcal{C}_{2,t}$, which occurs with probability $1-11e^{-t}$, one has $|\widehat{S}\setminus S^*|+|S^*\setminus\widehat{S}|=0$, implying that $\widehat{S}=S^*$. This completes the proof via setting $t=10\log(ep)$. □

C.6 Proof of Lemma C.6

Let $S=\mathrm{supp}(\boldsymbol{\beta})$, $I=S\cap S^*$. It follows from the fact that $\mathsf{J}(\boldsymbol{\beta}^*)=0$ and the definition of J that

$$\mathsf{J}(\boldsymbol{\beta})-\mathsf{J}(\boldsymbol{\beta}^*)-\widehat{\mathsf{J}}(\boldsymbol{\beta})+\widehat{\mathsf{J}}(\boldsymbol{\beta}^*)=\sum_{e\in\mathcal{E}}\frac{\omega^{(e)}}{4}\left\{\|\nabla_S\mathsf{R}^{(e)}(\boldsymbol{\beta})\|_2^2-\|\nabla_S\widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta})\|_2^2+\|\nabla_{S^*}\widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}^*)\|_2^2\right\}.$$

Observe that, for any fixed $e \in \mathcal{E}$,

$$
\begin{aligned}
&\frac{1}{8}\left(\|\nabla_S \mathsf{R}^{(e)}(\boldsymbol{\beta})\|_2^2 - \|\nabla_S \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta})\|_2^2 + \|\nabla_{S^*} \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}^*)\|_2^2\right) \\
&= \frac{1}{8}\left\{\|\nabla_S \mathsf{R}^{(e)}(\boldsymbol{\beta})\|_2^2 - \|\nabla_S \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) - \nabla_S \mathsf{R}^{(e)}(\boldsymbol{\beta}) + \nabla_S \mathsf{R}^{(e)}(\boldsymbol{\beta})\|_2^2 + \|\nabla_{S^*} \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}^*)\|_2^2\right\} \\
&= \ -\frac{1}{4}\{\nabla_S \mathsf{R}^{(e)}(\boldsymbol{\beta})\}^\top \left\{\nabla_S \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) - \nabla_S \mathsf{R}^{(e)}(\boldsymbol{\beta})\right\} + \frac{1}{8}\|\nabla_{S^*\setminus S} \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}^*)\|_2^2 \\
&\qquad\qquad - \frac{1}{8}\|\nabla_{S\setminus S^*} \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) - \nabla_{S\setminus S^*} \mathsf{R}^{(e)}(\boldsymbol{\beta})\|_2^2 \\
&\qquad\qquad + \frac{1}{8}\left\{\|\nabla_I \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}^*)\|_2^2 - \|\nabla_I \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) - \nabla_I \mathsf{R}^{(e)}(\boldsymbol{\beta})\|_2^2\right\} \\
&\overset{(a)}{\le} \left(\boldsymbol{\Sigma}_{S,:}^{(e)}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) - \mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}]\right)^\top \Bigg\{\widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}] - \mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}] \\
&\qquad\qquad\qquad - \left\{\widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \boldsymbol{\Sigma}_{S,:}^{(e)}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right\}\Bigg\} \\
&\qquad\qquad\qquad + \left(\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}\varepsilon^{(e)}]\right)^\top \left(\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \boldsymbol{\Sigma}_{I,:}^{(e)}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right) + \frac{1}{2}\left\|\widehat{\mathbb{E}}[\boldsymbol{x}_{S^*\setminus S}^{(e)}\varepsilon^{(e)}]\right\|_2^2 \\
&= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}_{:,S}^{(e)}\left\{\widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}] - \mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}]\right\} \\
&\qquad\quad - \mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}]^\top \left\{\widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}] - \mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}]\right\} \\
&\qquad\quad + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}_{:,S}^{(e)}\left\{\widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \boldsymbol{\Sigma}_{S,:}^{(e)}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right\} \\
&\qquad\quad - \mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}]^\top \left\{\widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \boldsymbol{\Sigma}_{S,:}^{(e)}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right\} \\
&\qquad\quad + \left(\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}\varepsilon^{(e)}]\right)^\top \left(\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \boldsymbol{\Sigma}_{I,:}^{(e)}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right) + \frac{1}{2}\left\|\widehat{\mathbb{E}}[\boldsymbol{x}_{S^*\setminus S}^{(e)}\varepsilon^{(e)}]\right\|_2^2 \\
&= \mathsf{T}_1^{(e)}(\boldsymbol{\beta}) + \mathsf{T}_2^{(e)}(\boldsymbol{\beta}) + \mathsf{T}_3^{(e)}(\boldsymbol{\beta}) + \mathsf{T}_4^{(e)}(\boldsymbol{\beta}) + \mathsf{T}_5^{(e)}(\boldsymbol{\beta}) + \mathsf{T}_6^{(e)}(\boldsymbol{\beta}),
\end{aligned}
\tag{C.27}
$$

where (a) follows from the definition of $\nabla\mathsf{R}(\boldsymbol{\beta})$, $\nabla\widehat{\mathsf{R}}(\boldsymbol{\beta})$ together with the facts

$$
\begin{aligned}
&\frac{1}{8}\left\{\|\nabla_I \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}^*)\|_2^2 - \|\nabla_I \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) - \nabla_I \mathsf{R}^{(e)}(\boldsymbol{\beta})\|_2^2\right\} \\
&= \frac{1}{2}\sum_{j\in I}\left(\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}]\right)^2 - \left(-\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}] + \widehat{\mathbb{E}}[x_j^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \mathbb{E}[x_j^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)]\right)^2 \\
&= \frac{1}{2}\sum_{j\in I}\left(\widehat{\mathbb{E}}[x_j^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \mathbb{E}[x_j^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)]\right) \times \\
&\qquad\qquad\qquad\qquad \left\{2\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}] - \left(\widehat{\mathbb{E}}[x_j^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \mathbb{E}[x_j^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)]\right)\right\} \\
&\le \sum_{j\in I}\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}] \times \left(\widehat{\mathbb{E}}[x_j^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \mathbb{E}[x_j^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)]\right) \\
&= \ \left(\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}\varepsilon^{(e)}]\right)^\top \left(\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}(\boldsymbol{x}^{(e)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^*)] - \boldsymbol{\Sigma}_{I,:}^{(e)}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right),
\end{aligned}
$$

and $-\frac{1}{8}\|\nabla_{S\setminus S^*} \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) - \nabla_{S\setminus S^*} \mathsf{R}^{(e)}(\boldsymbol{\beta})\|_2^2 \le 0$.

The rest of the proof will be divided into several pieces deriving the instance-dependent high-probability upper bounds on $\sum_{e\in\mathcal{E}}\omega^{(e)}\mathsf{T}_k^{(e)}(\boldsymbol{\beta})$ for each $k \in [6]$.

Step 1. Upper bound on $\mathsf{T}_1^{(e)}$. Define the event

$$\mathcal{C}_{1,t} = \left\{ \forall \boldsymbol{\beta} \in \mathcal{B}_s, \quad \sum_{e\in\mathcal{E}} \omega^{(e)} \mathsf{T}_1^{(e)} \le C_1 \kappa_U^{3/2} \sigma_x \sigma_\varepsilon \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \left(\sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*} \right) \right\} \tag{C.28}$$

for some constant C_1 to be determined. We will claim that $\mathbb{P}(\mathcal{C}_{1,t}) \ge 1 - e^{-t}$ in this step. We further assume $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$ since the inequality holds trivially when $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. It then follows from Cauchy-Schwarz inequality that

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\mathcal{B}_s, \boldsymbol{\beta}\neq\boldsymbol{\beta}^*} \frac{\sum_{e\in\mathcal{E}} \omega^{(e)} \mathsf{T}_1^{(e)}}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2} &= \sup_{\boldsymbol{\beta}\in\mathcal{B}_s, \boldsymbol{\beta}\neq\boldsymbol{\beta}^*} \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2} \sum_{e\in\mathcal{E}} \omega^{(e)} \boldsymbol{\Sigma}^{(e)}_{S\cup S^*, S} \left\{ \widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] - \mathbb{E}[\boldsymbol{x}^{(e)} \varepsilon^{(e)}] \right\} \\
&\le \sup_{|S|\le s} \left\| \sum_{e\in\mathcal{E}} \omega^{(e)} \boldsymbol{\Sigma}^{(e)}_{S\cup S^*, S} \left\{ \widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] - \mathbb{E}[\boldsymbol{x}^{(e)} \varepsilon^{(e)}] \right\} \right\|_2 .
\end{aligned}$$

For any $S \subseteq [p]$ with $|S| \le s$, let $\boldsymbol{v}_1^{(S)}, \ldots, \boldsymbol{v}_{N_S}^{(S)}$ be an $1/4-$covering of $\mathcal{B}(S \cup S^*)$, that is, for any $\boldsymbol{v} \in \mathcal{B}(S\cup S^*)$, there exists some $\pi(\boldsymbol{v}) \in [N_S]$ such that

$$\|\boldsymbol{v} - \boldsymbol{v}^{(S)}_{\pi(v)}\|_2 \le 1/4. \tag{C.29}$$

It follows from standard empirical process result that $N_S \le 9^{|S\cup S^*|}$, then

$$\begin{aligned}
N = \sum_{|S|\le s} N_S \le \sum_{|S|\le s} 9^{|S\cup S^*|} &\le \sum_{i=0}^{s} 9^{i+s^*} \binom{p}{i} \\
&\le 9^{s^*} \times \left(\frac{9p}{s}\right)^s \sum_{i=0}^{s} \left(\frac{s}{p}\right)^i \binom{p}{i} \le 9^{s^*} \times \left(\frac{9p}{s}\right)^s \sum_{i=0}^{p} \left(\frac{s}{p}\right)^i \binom{p}{i} \\
&\le 9^{s^*} \left(\frac{9p}{s}\right)^s \left(1 + \frac{s}{p}\right)^p \le 9^{s^*} \left(\frac{9ep}{s}\right)^s .
\end{aligned} \tag{C.30}$$

At the same time, denote $\boldsymbol{\xi} = \sum_{e\in\mathcal{E}} \omega^{(e)} \boldsymbol{\Sigma}^{(e)}_{S\cup S^*, S} \left\{ \widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] - \mathbb{E}[\boldsymbol{x}^{(e)} \varepsilon^{(e)}] \right\}$. For any $S \in [p]$ with $|S| \le s$, it follows from the variational representation of the ℓ_2 norm that

$$\|\boldsymbol{\xi}\|_2 = \sup_{\boldsymbol{v}\in\mathcal{B}(S\cup S^*)} \boldsymbol{v}^\top \boldsymbol{\xi} = \sup_{k\in[N_S]} (\boldsymbol{v}_k^{(S)})^\top \boldsymbol{\xi} + \sup_{\boldsymbol{v}\in\mathcal{B}(S\cup S^*)} (\boldsymbol{v} - \boldsymbol{v}^{(S)}_{\pi(v)})^\top \boldsymbol{\xi} \le \sup_{k\in[N_S]} (\boldsymbol{v}_k^{(S)})^\top \boldsymbol{\xi} + \frac{1}{4}\|\boldsymbol{\xi}\|_2,$$

where the last inequality follows from the Cauchy-Schwarz inequality and our construction of covering in (C.29). This implies $\|\boldsymbol{\xi}\|_2 \le 2 \sup_{k\in[N_S]} (\boldsymbol{v}_k^{(S)})^\top \boldsymbol{\xi}$, thus

$$\begin{aligned}
&\sup_{|S|\le s} \left\| \sum_{e\in\mathcal{E}} \omega^{(e)} \boldsymbol{\Sigma}^{(e)}_{S\cup S^*, S} \left\{ \widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] - \mathbb{E}[\boldsymbol{x}^{(e)} \varepsilon^{(e)}] \right\} \right\|_2 \\
&\qquad \le 2 \sup_{|S|\le s, k\in[N_S]} \underbrace{(\boldsymbol{v}_k^{(S)})^\top \sum_{e\in\mathcal{E}} \omega^{(e)} \boldsymbol{\Sigma}^{(e)}_{S\cup S^*, S} \left\{ \widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] - \mathbb{E}[\boldsymbol{x}^{(e)} \varepsilon^{(e)}] \right\}}_{Z(S,k)} .
\end{aligned} \tag{C.31}$$

For given fixed $\boldsymbol{v}_k^{(S)} \in \mathcal{B}(S \cup S^*)$, $Z(S,k)$ can be written as the sum of independent zero-mean random variables as

$$\sum_{e\in\mathcal{E}} \sum_{i=1}^{n^{(e)}} \frac{\omega^{(e)}}{n^{(e)}} \left(X_{e,i} - \mathbb{E}[X_{e,i}]\right) \quad \text{with} \quad X_{e,i} = \left((\boldsymbol{v}_k^{(S)})^\top \boldsymbol{\Sigma}^{(e)}_{S\cup S^*, S} \boldsymbol{x}_S^{(e)} \right) \left(\varepsilon^{(e)} \right).$$

Observe that $\varepsilon^{(e)}$ is a zero-mean sub-Gaussian random variable with parameter σ_ε by Condition A.4, and $(\boldsymbol{v}_k^{(S)})^\top \boldsymbol{\Sigma}_{S \cup S^*, S}^{(e)} \boldsymbol{x}_S$ is a zero-mean sub-Gaussian random variable with parameter

$$\sigma_1 = \|(\boldsymbol{v}_k^{(S)})^\top \boldsymbol{\Sigma}_{S \cup S^*, S}^{(e)} \bar{\boldsymbol{\Sigma}}_{S,:}^{1/2}\|_2 \sigma_x \le \kappa_U^{3/2} \sigma_x$$

by Condition A.3. It then follows from Lemma C.1 and Lemma C.2 that there exists some universal constant C' such that

$$\mathbb{P}\left[|Z(S,k)| \ge C' \sigma_1 \sigma_\varepsilon \left\{ \sqrt{\sum_{e \in \mathcal{E}} \sum_{i=1}^{n^{(e)}} \left(\frac{\omega^{(e)}}{n^{(e)}} \right)^2 u} + \max_{e \in \mathcal{E}} \frac{\omega^{(e)}}{n^{(e)}} u \right\} \right] \le 2e^{-u}$$

for any $u > 0$. Letting $u = t + \log(2N) \le 3\left(t + s\log(ep/s) + s^*\right)$, we obtain

$$\mathbb{P}\left[\sup_{|S| \le s, k} |Z(S,k)| \ge 3C' \sigma_1 \sigma_\varepsilon \left(\sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*} \right) \right] \le N \times 2e^{-\log(2N) - t} \le e^{-t}.$$

Combining with the argument (C.31) concludes the proof of the claim with $C_1 = 6C'$.

Step 2. Upper Bound on $\mathsf{T}_2^{(e)}$. We claim that $\mathbb{P}(\mathcal{C}_{2,t}) \ge 1 - e^{-t}$ for any $t > 0$, where

$$\begin{aligned} \mathcal{C}_{2,t} = \Bigg\{ \forall \boldsymbol{\beta}, \quad \sum_{e \in \mathcal{E}} \omega^{(e)} \mathsf{T}_2^{(e)} \le C_2 \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \sqrt{\frac{V(|S \setminus S^*|, t)}{n_*}} \times \sqrt{\sum_{e \in \mathcal{E}} \omega^{(e)} \left\| \mathbb{E}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] \right\|_2^2} \\ + C_2 \kappa_U \sigma_x \sigma_\varepsilon^2 \frac{V(|S \setminus S^*|, t)}{n_*} \Bigg\} \end{aligned} \tag{C.32}$$

for some universal constant C_2 to be determined. Note that L.H.S. and R.H.S. of the inequality in (C.32) both depend on S, which is the support set of $\boldsymbol{\beta}$ and satisfies $|S| \le s$. For a fixed S, we can write down $\sum_{e \in \mathcal{E}} \omega^{(e)} \mathsf{T}_2^{(e)}$ as sum of independent random variables as

$$Z(S) = \sum_{e \in \mathcal{E}} \omega^{(e)} \mathsf{T}_2^{(e)} = \sum_{e \in \mathcal{E}} \sum_{i=1}^{n^{(e)}} \frac{\omega^{(e)}}{n^{(e)}} \left(X_{e,i} - \mathbb{E}[X_{e,i}]\right) \quad \text{with} \quad X_{e,i} = \left(\mathbb{E}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}]^\top [\boldsymbol{x}_i^{(e)}]_S \right) (\varepsilon_i^{(e)}).$$

Observe that $X_{i,e}$ are independent sub-exponential random variables because of Lemma C.1 and our assumptions Condition A.3–A.4. It then follows from Lemma C.2 that the following event,

$$\begin{aligned} |Z(S)| &\le C' \kappa_U \sigma_x \sigma_\varepsilon \left\{ \sqrt{\sum_{e \in \mathcal{E}} (\omega^{(e)})^2 \frac{1}{n^{(e)}} \left\| \mathbb{E}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] \right\|_2^2} \times \sqrt{u} + \max_{e \in \mathcal{E}} \frac{\omega^{(e)}}{n^{(e)}} \left\| \mathbb{E}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] \right\|_2^2 \times u \right\} \\ &\le C' \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \left\{ \sqrt{u \times \max_{e' \in \mathcal{E}} \frac{\omega^{(e')}}{n^{(e')}}} \sqrt{\sum_{e \in \mathcal{E}} \omega^{(e)} \left\| \mathbb{E}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] \right\|_2^2} + \frac{u}{n_*} \max_{e \in \mathcal{E}} \left\| \mathbb{E}[\boldsymbol{x}^{(e)} \varepsilon^{(e)}] \right\|_2^2 \right\} \\ &\le C' \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \left\{ \sqrt{\frac{x}{n_*}} \sqrt{\sum_{e \in \mathcal{E}} \omega^{(e)} \left\| \mathbb{E}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] \right\|_2^2} + \kappa_U^{1/2} \sigma_\varepsilon \frac{x}{n_*} \right\} \end{aligned}$$

occurs with probability at least $1 - 2e^{-x}$ for any $u > 0$, where the last inequality follows from Lemma C.11 and the definition of n_* in (A.4). Now, define the event

$$\mathcal{K}_u(r) = \left\{ \forall S, |S \setminus S^*| = r, \quad |Z(S)| \le 2C' \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \left\{ \sqrt{\frac{V(r,u)}{n_*}} \sqrt{\sum_{e \in \mathcal{E}} \omega^{(e)} \left\| \mathbb{E}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] \right\|_2^2} + \kappa_U^{1/2} \sigma_\varepsilon \frac{V(r,u)}{n_*} \right\} \right\}$$

for any $r \geq 1$. The total number of S satisfying $|S \setminus S^*| = r$ can be upper-bounded by

$$N_r = 2^{s^*} \times \binom{p - s^*}{r} \leq 2^{s^*} (ep/r)^r.$$

Applying union bound with $x = u + \log(2N_r) \leq 2(u + r\log(ep/r) + s^*)$ then gives $\mathbb{P}[\mathcal{K}_u(r)] \geq 1 - 2N_r e^{-(u+\log(2N_r))} = 1 - e^{-u}$. Therefore, we can argue that, under $\bigcap_{r=1}^{p} \mathcal{K}_{t+\log p}(r)$, the following holds

$$\forall \boldsymbol{\beta}, \quad Z(S) \leq 4C' \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \left\{ \sqrt{\frac{V(|S \setminus S^*|, t)}{n_*}} \sqrt{\sum_{e \in \mathcal{E}} \omega^{(e)} \left\| \mathbb{E}[\boldsymbol{x}_S^{(e)} \varepsilon^{(e)}] \right\|_2^2} + \kappa_U^{1/2} \sigma_\varepsilon \frac{V(|S \setminus S^*|, t)}{n_*} \right\}$$

by the fact that

$$\forall r \geq 1, \qquad V(r, t + \log p) = r\log(ep/r) + s^* + (\log p) + t \leq 2r\log(ep/r) + s^* + t \leq 2V(r,t).$$

This completes the proof of via setting $C_2 = 4C'$.

Step 3. Upper Bound on $\mathsf{T}_3^{(e)}$. We argue in this step that the event

$$\mathcal{C}_{3,t} = \left\{ \forall \boldsymbol{\beta} \in \mathcal{B}_s, \quad \sum_{e \in \mathcal{E}} \omega^{(e)} \mathsf{T}_3^{(e)} \leq C_3 \kappa_U^2 \sigma_x^2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \left(\sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*} \right) \right\} \tag{C.33}$$

occurs with probability at least $1 - e^{-t}$ for any $t > 0$, where C_3 is some universal constant to be determined. Without loss of generality, let $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$, then it suffices to establish an upper bound for

$$\begin{aligned} &\sup_{\boldsymbol{\beta} \neq \boldsymbol{\beta}^*} \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2} \sum_{e \in \mathcal{E}} \omega^{(e)} \boldsymbol{\Sigma}_{:,S}^{(e)} \left(\widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} (\boldsymbol{x}^{(e)})^\top] - \mathbb{E}[\boldsymbol{x}_S^{(e)} (\boldsymbol{x}^{(e)})^\top] \right) \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)}{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2} \\ &\quad \leq \sup_{|S| \leq s} \left\| \sum_{e \in \mathcal{E}} \omega^{(e)} \boldsymbol{\Sigma}_{S \cup S^*, S}^{(e)} \left(\widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} (\boldsymbol{x}_{S \cup S^*}^{(e)})^\top] - \mathbb{E}[\boldsymbol{x}_S^{(e)} (\boldsymbol{x}_{S \cup S^*}^{(e)})^\top] \right) \right\|_2 = \sup_{|S| \leq s} \|\boldsymbol{A}_S\|_2. \end{aligned} \tag{C.34}$$

We follow a similar strategy as Step 1. For any $S \in \mathcal{B}_s$, let $\{(\boldsymbol{v}_k^{(e)}, \boldsymbol{u}_k^{(e)})\}_{k=1}^{N_S} \in \mathcal{B}(S \cup S^*) \times \mathcal{B}(S \cup S^*) := \mathcal{B}^2(S \cup S^*)$ be a 1/4-covering of $\mathcal{B}^2(S \cup S^*)$ in a sense that for any $(\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{B}^2(S \cup S^*)$, there exists some $\pi(\boldsymbol{u}, \boldsymbol{v}) \in [N_S]$ such that

$$\|\boldsymbol{u} - \boldsymbol{u}_{\pi(\boldsymbol{u},\boldsymbol{v})}^{(S)}\|_2 + \|\boldsymbol{v} - \boldsymbol{v}_{\pi(\boldsymbol{u},\boldsymbol{v})}^{(S)}\|_2 \leq \frac{1}{4}.$$

It follows from standard empirical process theory that $N_S \leq 9^{2|S \cup S^*|}$, then

$$N = \sum_{S \subseteq [p], |S| \leq s} N_S \leq \sum_{S \subseteq [p], |S| \leq s} N_S 9^{2|S \cup S^*|} \leq \sum_{i=0}^{s} 81^{i + s^*} \binom{p}{i} \leq 81^{s^*} \left(\frac{81ep}{s} \right)^s,$$

where the last inequality follows from the same procedure as (C.30).

At the same time, denote $\boldsymbol{u}^\dagger = \boldsymbol{u}_{\pi(\boldsymbol{u},\boldsymbol{v})}^{(S)}$ and $\boldsymbol{v}^\dagger = \boldsymbol{v}_{\pi(\boldsymbol{u},\boldsymbol{v})}^{(S)}$. It follows from the variational representation of the matrix ℓ_2 norm that

$$\begin{aligned} \|\boldsymbol{A}_S\|_2 &= \sup_{(\boldsymbol{u},\boldsymbol{v}) \in \mathcal{B}^2(S \cup S^*)} \boldsymbol{u}^\top \boldsymbol{A}_S \boldsymbol{v} \\ &\leq \sup_{(\boldsymbol{u},\boldsymbol{v}) \in \mathcal{B}^2(S \cup S^*)} (\boldsymbol{u}^\dagger)^\top \boldsymbol{A}_S \boldsymbol{v}^\dagger + \sup_{(\boldsymbol{u},\boldsymbol{v}) \in \mathcal{B}^2(S \cup S^*)} (\boldsymbol{u} - \boldsymbol{u}^\dagger)^\top \boldsymbol{A}_S \boldsymbol{v}^\dagger \\ &\qquad + \sup_{(\boldsymbol{u},\boldsymbol{v}) \in \mathcal{B}^2(S \cup S^*)} (\boldsymbol{u} - \boldsymbol{u}^\dagger)^\top \boldsymbol{A}_S (\boldsymbol{v} - \boldsymbol{v}^\dagger) + \sup_{(\boldsymbol{u},\boldsymbol{v}) \in \mathcal{B}^2(S \cup S^*)} (\boldsymbol{u}^\dagger)^\top \boldsymbol{A}_S (\boldsymbol{v} - \boldsymbol{v}^\dagger) \\ &\leq \sup_{k \in [N_S]} (\boldsymbol{u}_k^{(S)})^\top \boldsymbol{A}_S \boldsymbol{v}_k^{(S)} + \left(\frac{1}{4} + \frac{1}{4} + \frac{1}{16} \right) \|\boldsymbol{A}_S\|_2, \end{aligned}$$

which implies $\|\boldsymbol{A}_S\|_2 \le 4\sup_{k\in[N_S]}(\boldsymbol{u}_k^{(S)})^\top \boldsymbol{A}_S \boldsymbol{v}_k^{(S)}$, thus

$$\sup_{|S|\le s} \|\boldsymbol{A}_S\|_2 \le \sup_{|S|\le s, k\in[N_S]} (\boldsymbol{u}_k^{(e)})^\top \boldsymbol{A}_S(\boldsymbol{v}_k^{(e)}) = \sup_{|S|\le s, k\in[N_S]} Z(S,k). \tag{C.35}$$

For fixed k and S, $Z(S,k)$ can be written as the sum of independent zero-mean random variables as

$$Z(S,k) = \sum_{e\in\mathcal{E}}\sum_{i=1}^{n^{(e)}} \frac{\omega^{(e)}}{n^{(e)}}(X_{e,i} - \mathbb{E}[X_{e,i}]) \qquad \text{with} \qquad X_{e,i} = \left((\boldsymbol{u}_k^{(e)})^\top \Sigma^{(e)}_{S\cup S^*,S}\boldsymbol{x}_S^{(e)}\right)\left((\boldsymbol{x}^{(e)}_{S\cup S^*})^\top \boldsymbol{v}_k^{(e)}\right).$$

Here $X_{e,i}$ is the product of two zero-mean sub-Gaussian random variables with parameter $\kappa_U^{3/2}\sigma_x$ and $\kappa_U^{1/2}\sigma_x$, respectively. It then follows from Lemma C.1 and Lemma C.2 that, for any $u>0$,

$$\mathbb{P}\left[|Z(S,k)| \le C'\kappa_U^2\sigma_x^2\left(\sqrt{\frac{u}{n_{\boldsymbol{\omega}}}} + \frac{u}{n_*}\right)\right] \ge 1 - 2e^{-u},$$

where C' is some universal constant. Finally, we apply union bound with $u = t + \log(2N) \le 6(t + s\log(ep/s) + s^*)$ and obtain

$$\mathbb{P}\left[\sup_{|S|\le s, k\in[N_S]} |Z(S,k)| \le 6C'\kappa_U^2\sigma_x^2\left(\sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*}\right)\right] \ge 1 - N\times 2e^{-\log(2N)+t} = 1 - e^{-t}.$$

Set $C_3 = 24C'$. Combining the above inequality with the suprema arguments (C.35) and (C.34) completes the proof.

Step 4. Upper Bound on $\mathsf{T}_4^{(e)}$. Our target in this step is to show that, for any $t>0$,

$$\mathbb{P}(\mathcal{C}_{4,t}) = \mathbb{P}\left[\forall\boldsymbol{\beta}\in\mathcal{B}_s,\quad \sum_{e\in\mathcal{E}}\omega^{(e)}\mathsf{T}_4^{(e)} \le C_4\kappa_U^{3/2}\sigma_x^2\sigma_\varepsilon\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\left(\sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*}\right)\right] \ge 1 - e^{-t}. \tag{C.36}$$

The proof is very similar to that in Step 1, we only sketch here and highlight the difference. Following a similar strategy, it suffices to derive an upper bound for the quantity

$$\sup_{|S|\le s, k\in[N_S]} Z(S,k) \qquad \text{with} \qquad Z(S,k) = \sum_{e\in\mathcal{E}}\sum_{i=1}^{n^{(e)}}\frac{\omega^{(e)}}{n^{(e)}}(X_{e,i} - \mathbb{E}[X_{e,i}]) \ \text{ and } \ N = \sum_{S\subseteq[p]} N_S \le 9^{s^*}(9ep/s)^s,$$

where $X_{e,i} = \left((\mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}])^\top[\boldsymbol{x}_i^{(e)}]_S\right)\left([\boldsymbol{x}_i^{(e)}]_{S\cup S^*}^\top \boldsymbol{v}_k^{(S)}\right)$ is the product of two zero-mean sub-Gaussian random variables with parameters $\kappa_U\sigma_x\sigma_\varepsilon$ and $\kappa_U^{1/2}\sigma_x$, respectively. It then follows from Lemma C.2 that, for any $u>0$,

$$\mathbb{P}\left[|Z(S,k)| \le C'\kappa_U^{3/2}\sigma_x^2\sigma_\varepsilon\left(\sqrt{\frac{u}{n_{\boldsymbol{\omega}}}} + \frac{u}{n_*}\right)\right] \ge 1 - 2e^{-u}.$$

So it concludes via applying union bound with $u = t + \log(2N) \le 3V(s,t)$.

Step 5. Upper Bound on $\mathsf{T}_5^{(e)}$. In this step, we claim that the following event

$$\mathcal{C}_{5,t} = \left\{\forall\boldsymbol{\beta}\in\mathbb{R}^p,\quad \sum_{e\in\mathcal{E}}\omega^{(e)}\mathsf{T}_5^{(e)} \le C_5\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon\frac{p + \log(2|\mathcal{E}|) + t}{\bar{n}}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\right\} \tag{C.37}$$

occurs with probability at least $1-e^{-t}$ if $s^* + t + \log(|2\mathcal{E}|) \le n_{\min}$. Define the event $\mathcal{K}_{5,u}(e)$ as

$$\mathcal{K}_{5,u}(e) = \left\{\left\|\widehat{\mathbb{E}}[\boldsymbol{x}_{S^*}^{(e)}\varepsilon^{(e)}]\right\|_2 \le C_1'\kappa_U^{1/2}\sigma_x\sigma_\varepsilon\left(\sqrt{\frac{s^*+u}{n^{(e)}}} + \frac{s^*+u}{n^{(e)}}\right)\right\}$$

for some universal constant C_1'. Observe that $\|\mathbb{E}[\boldsymbol{x}_{S^*}^{(e)}\varepsilon^{(e)}]\|_2 = \sup_{\boldsymbol{v}\in\mathbb{R}^{s^*},\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^\top\widehat{\mathbb{E}}[\boldsymbol{x}_{S^*}^{(e)}\varepsilon^{(e)}]$, and $\boldsymbol{v}^\top\widehat{\mathbb{E}}[\boldsymbol{x}_{S}^{(e)}\varepsilon^{(e)}]$ with given fixed $\boldsymbol{v}\in\mathbb{R}^{s^*}$ is the sum of independent (centered) products of two zero-mean sub-Gaussian random variables with parameter $\kappa_U^{1/2}\sigma_x$ and σ_ε, respectively. Then it follows from Lemma C.1 and Lemma C.2 that, for any fixed $\boldsymbol{v}\in\mathbb{R}^{s^*}$ and $x>0$

$$\mathbb{P}\left[\left|\boldsymbol{v}^\top\widehat{\mathbb{E}}[\boldsymbol{x}_{S^*}^{(e)}\varepsilon^{(e)}]\right| \ge C'\kappa_U^{1/2}\sigma_x\sigma_\varepsilon\left(\sqrt{\frac{x}{n^{(e)}}}+\frac{x}{n^{(e)}}\right)\right] \le 2e^{-x}.$$

Following a similar argument as that of Step 1 and applying union bound gives $\mathbb{P}\left[\mathcal{C}_{5,u}(e)\right] \ge 1-e^{-u}$.

At the same time, let

$$\mathcal{K}_{5,u}'(e) = \left\{\forall\boldsymbol{\beta}\in\mathcal{B}_s,\quad \left\|\widehat{\mathbb{E}}[\boldsymbol{x}_{S^*}^{(e)}(\boldsymbol{x}^{(e)})^\top]-\boldsymbol{\Sigma}_{S^*,:}^{(e)}\right\|_2 \le C_2'\kappa_U\sigma_x^2\left(\sqrt{\frac{V(s,u)}{n^{(e)}}}+\frac{V(s,u)}{n^{(e)}}\right)\right\}$$

for some universal constant $C_2'>0$. It is easy to verify that $\mathbb{P}(\mathcal{C}_{5,u}'(e)) \ge 1-e^{-u}$ for any given fixed $u>0$ and $e\in\mathcal{E}$.

Under the event $\mathcal{K}_u = \bigcap_{e\in\mathcal{E}}\{\mathcal{K}_{5,u}(e)\cap\mathcal{K}_{5,u}'(e)\}$, which occurs with probability $1-2|\mathcal{E}|e^{-u}$, we obtain

$$\begin{aligned}
\forall\boldsymbol{\beta}\in\mathcal{B}_s,\quad \sum_{e\in\mathcal{E}}\omega^{(e)}\mathsf{T}_5^{(e)} &= \sum_{e\in\mathcal{E}}\omega^{(e)}\left(\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}\varepsilon^{(e)}]\right)^\top\left(\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}(\boldsymbol{x}^{(e)})^\top]-\boldsymbol{\Sigma}_{I,:}^{(e)}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\\
&\le \sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}\varepsilon^{(e)}]\right\|_2\left\|\widehat{\mathbb{E}}[\boldsymbol{x}_I^{(e)}(\boldsymbol{x}^{(e)})^\top]-\boldsymbol{\Sigma}_{I,:}^{(e)}\right\|_2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\\
&\le \sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\widehat{\mathbb{E}}[\boldsymbol{x}_{S^*}^{(e)}\varepsilon^{(e)}]\right\|_2\left\|\widehat{\mathbb{E}}[\boldsymbol{x}_{S^*}^{(e)}(\boldsymbol{x}^{(e)})^\top]-\boldsymbol{\Sigma}_{S^*,:}^{(e)}\right\|_2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\\
&\overset{(a)}{\le} C''\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon\sum_{e\in\mathcal{E}}\omega^{(e)}\sqrt{\frac{|S^*|+u}{n^{(e)}}}\left(\sqrt{\frac{V(s,u)}{n^{(e)}}}+\frac{V(s,u)}{n^{(e)}}\right)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\\
&\overset{(b)}{\le} C''\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon\frac{\sqrt{V(s,u)V(0,u)}}{\bar{n}}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\\
&\qquad + C''\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon\frac{V(s,u)\sqrt{V(0,u)}}{n_\dagger}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2
\end{aligned}$$

provided $s^*+u\le n_{\min}$. Here (a) follows from the definition of the events $\mathcal{K}_{5,u}(e)$ and $\mathcal{K}_{5,u}'(e)$ and the fact that $x\le\sqrt{x}$ when $x\in[0,1]$, (b) follows directly from and the definition of $\bar{n}$ in (A.4) and the definition of $n_\dagger$ in (A.4). This completes the proof via letting $u=\log(2|\mathcal{E}|)+t$.

Step 6. Upper Bound on $\mathsf{T}_6^{(e)}$. The goal of this step is to derive a high-probability bound for the event

$$\mathcal{C}_{6,t} = \left\{\forall\boldsymbol{\beta}\in\mathbb{R}^p,\quad \sum_{e\in\mathcal{E}}\omega^{(e)}\mathsf{T}_6^{(e)} \le C_6\kappa_U\sigma_x^2\sigma_\varepsilon^2\frac{t+\log(2|\mathcal{E}||S^*|)}{\bar{n}}|S^*\setminus S|\right\} \tag{C.38}$$

for any $t\in(0,n_{\min}-\log(2|\mathcal{E}||S^*|)]$. Note that both the L.H.S. and R.H.S. of the inequality in (C.38) depends on $\boldsymbol{\beta}$, or more precisely, $S=\mathrm{supp}(\boldsymbol{\beta})$. Denoting $\delta=C_6\kappa_U\sigma_x^2\sigma_\varepsilon^2\{t+\log(2|\mathcal{E}||S^*|)\}/\bar{n}$, we have the following decomposition

$$\begin{aligned}
\mathcal{C}_{6,t} &= \bigcup_{T\subseteq S^*}\left\{\forall\boldsymbol{\beta}\in\mathbb{R}^p, S^*\setminus\mathrm{supp}(\boldsymbol{\beta})=T,\quad \sum_{e\in\mathcal{E}}\omega^{(e)}\mathsf{T}_6^{(e)}\le\delta|T|\right\}\\
&= \bigcup_{T\subseteq S^*}\left\{\forall\boldsymbol{\beta}\in\mathbb{R}^p, S^*\setminus\mathrm{supp}(\boldsymbol{\beta})=T,\quad \sum_{j\in T}\sum_{e\in\mathcal{E}}\omega^{(e)}\left|\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}]\right|^2\le\delta|T|\right\}\\
&= \bigcup_{T\subseteq S^*}\mathcal{K}(T).
\end{aligned}$$

At the same time, given fixed $j \in S^*$ and $e \in \mathcal{E}$, it follows from Lemma C.2 that,

$$\mathbb{P}[\mathcal{K}_{6,x}(e,j)] = \mathbb{P}\left[\left|\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}]\right| \le C'\kappa_U^{1/2}\sigma_x\sigma_\epsilon\left(\sqrt{\frac{x}{n^{(e)}}} + \frac{x}{n^{(e)}}\right)\right] \ge 1 - 2e^{-x}.$$

for some universal constant C'. We claim that

$$\mathcal{K}(T) \subseteq \bigcup_{j\in S^*, e\in\mathcal{E}} \mathcal{K}_{6,t+\log(2s^*|\mathcal{E}|)}(e,j) \tag{C.39}$$

by choosing $C_6 = (C')^2$. This is because under $\bigcup_{j\in S^*,e\in\mathcal{E}} \mathcal{K}_{6,t+\log(2s^*|\mathcal{E}|)}(e,j)$, one has

$$\begin{aligned}\sum_{j\in T}\sum_{e\in\mathcal{E}} \omega^{(e)}\left|\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}]\right|^2 &\le \sum_{j\in T}\sum_{e\in\mathcal{E}} (C')^2\kappa_U\sigma_x^2\sigma_\varepsilon^2\left(\frac{t+\log(2s^*|\mathcal{E}|)}{n^{(e)}}\omega^{(e)}\right) \\ &\le |T|(C')^2\kappa_U\sigma_x^2\sigma_\varepsilon^2\frac{t+\log(2s^*|\mathcal{E}|)}{\bar{n}}\end{aligned}$$

provided $t + \log(2s^*|\mathcal{E}|) \le n_{\min}$, this validates the claim (C.39). Therefore, we have

$$\begin{aligned}\mathbb{P}(\mathcal{C}_{6,t}) = \mathbb{P}\left[\bigcup_{T\subseteq S^*}\mathcal{K}(T)\right] &\ge \mathbb{P}\left[\bigcup_{j\in S^*,e\in\mathcal{E}} \mathcal{K}_{6,t+\log(2s^*|\mathcal{E}|)}(e,j)\right] \\ &\ge 1 - \sum_{e\in\mathcal{E},j\in S^*}\left(1 - \mathbb{P}\left[\mathcal{K}_{6,t+\log(2s^*|\mathcal{E}|)}(e,j)\right]\right) \\ &\ge 1 - 2s^*|\mathcal{E}|e^{-t-\log(2s^*|\mathcal{E}|)} \ge 1 - e^{-t}.\end{aligned}$$

STEP 7. CONCLUSION. We are now ready to conclude the proof by combining results (C.28), (C.32), (C.33), (C.36), (C.37), (C.38) we obtained from STEP 1 to STEP 6. Plugging these upper bounds back into our decomposition in (C.27), we have that, under the event $\bigcap_{k=1}^6 \mathcal{C}_{k,t}$, which occurs with probability at least $1 - 6e^{-t}$, the inequality in (C.21) holds provided $n_{\min} \ge (s^* + \log(2|\mathcal{E}|) + t) \vee (\log(2|\mathcal{E}||S^*|) + t) = s^* + \log(2|\mathcal{E}|) + t$. This completes the proof. □

C.7 Proof of Lemma C.7

It follows from the definition of the pooled L_2 risk that

$$\begin{aligned}&\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\boldsymbol{\beta}^*) - \widehat{\mathsf{R}}(\boldsymbol{\beta}) + \widehat{\mathsf{R}}(\boldsymbol{\beta}^*) \\ &\quad= \sum_{e\in\mathcal{E}} \omega^{(e)}\left\{\mathsf{R}^{(e)}(\boldsymbol{\beta}) - \mathsf{R}^{(e)}(\boldsymbol{\beta}^*) - \left(\widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}) - \widehat{\mathsf{R}}^{(e)}(\boldsymbol{\beta}^*)\right)\right\} \\ &\quad= \sum_{e\in\mathcal{E}} \omega^{(e)}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\left(\boldsymbol{\Sigma}^{(e)} - \widehat{\boldsymbol{\Sigma}}^{(e)}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - 2(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\left(\mathbb{E}[\boldsymbol{x}^{(e)}\varepsilon^{(e)}] - \widehat{\mathbb{E}}[\boldsymbol{x}^{(e)}\varepsilon^{(e)}]\right) \\ &\quad= (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\boldsymbol{A}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - 2(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\boldsymbol{b}.\end{aligned}$$

Recall that $\delta_1 = \sqrt{\frac{V(s,t)}{n_{\boldsymbol{\omega}}}} + \frac{V(s,t)}{n_*}$, we argue that the following two events

$$\begin{aligned}\mathcal{C}_{1,t} &= \left\{\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{B}_s\times\mathcal{B}_s,\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2=1} \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{y} \le C_1\kappa_U\sigma_x^2\delta_1\right\} \\ \mathcal{C}_{2,t} &= \left\{\sup_{\boldsymbol{x}\in\mathcal{B}_s,\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\top\boldsymbol{b} \le C_2\kappa_U^{1/2}\sigma_x\sigma_\varepsilon\delta_1\right\}\end{aligned}$$

satisfies $\mathbb{P}[\mathcal{C}_{1,t}] \wedge \mathbb{P}[\mathcal{C}_{2,t}] \ge 1 - 0.5e^{-t}$ for any $t > 0$, where C_1, C_2 are some universal constants to be determined. If the two claims are verified, then we have, under the event $\mathcal{C}_{1,t} \cap \mathcal{C}_{2,t}$ which occurs with probability at least $1 - e^{-t}$, the following holds

$$\begin{aligned}
&|\mathsf{R}(\boldsymbol{\beta}) - \mathsf{R}(\boldsymbol{\beta}^*) - \widehat{\mathsf{R}}(\boldsymbol{\beta}) + \widehat{\mathsf{R}}(\boldsymbol{\beta}^*)| \\
&\quad \le \left|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \boldsymbol{A}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right| + 2\left|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \boldsymbol{b}\right| \\
&\quad \le \{C_1 \vee (2C_2)\} \left(\kappa_U \sigma_x^2 \delta_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \kappa_U^{1/2} \sigma_x \sigma_\varepsilon \delta_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2\right)
\end{aligned}$$

for some universal constant C'. This completes the proof of Lemma C.3.

Now, we prove the two claims separately.

Proof of the Claim $\mathbb{P}[\mathcal{C}_{1,t}] \ge 1 - 0.5e^{-t}$. The proof strategy is very similar to Step 3 in the proof of Lemma C.4. To be specific, similar to the derivation in (C.35), there exist $N \le 81^{s^*}(81ep/s)^s$ pairs of p-dimensional unit vectors $(\boldsymbol{v}_1, \boldsymbol{u}_1), \ldots, (\boldsymbol{v}_N, \boldsymbol{u}_N)$ such that

$$\sup_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{B}_s \times \mathcal{B}_s, \|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{y} \le 4 \sup_{k \in [N]} \boldsymbol{v}_k^\top \boldsymbol{A} \boldsymbol{u}_k.$$

For fixed $(\boldsymbol{v}_k, \boldsymbol{u}_k)$, it follows from Conditions A.1, A.2 and A.3 that $\boldsymbol{v}_k^\top \boldsymbol{A} \boldsymbol{u}_k$ is the sum of independent variables $X_{e,i} - \mathbb{E}[X_{e,i}]$, and $X_{e,i}$ is the product of two zero-mean sub-Gaussian random variables with parameter $\kappa_U^{1/2}\sigma_x$. Then applying Lemma C.1 and Lemma C.2 gives

$$\mathbb{P}\left[|\boldsymbol{v}_k^\top \boldsymbol{A} \boldsymbol{u}_k| \ge C' \kappa_U^{1/2} \sigma_x \left(\sqrt{\frac{u}{n_{\boldsymbol{\omega}}}} + \frac{u}{n_*}\right)\right] \le 2e^{-u}$$

for any $u > 0$ and some universal constant C'. Using the union bounds over all the k, we have the following event

$$\sup_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{B}_s \times \mathcal{B}_s} \le 4 \sup_{k \in [N]} \boldsymbol{v}_k^\top \boldsymbol{A} \boldsymbol{u}_k \le 4C' \kappa_U^{1/2} \sigma_x \left(\sqrt{\frac{u}{n_{\boldsymbol{\omega}}}} + \frac{u}{n_*}\right)$$

will occurs with probability at least $1 - 2Ne^{-u}$. Letting $u = t + \log(4N) \le 6V(s,t)$ completes the proof.

Proof of the Claim $\mathbb{P}[\mathcal{C}_{2,t}] \ge 1 - 0.5e^{-t}$. The proof strategy is also similar to the first part of Step 5 in the proof of Lemma C.4. To be specific, there exist $N = 90^p$ p-dimensional unit vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ such that $\sup_{\boldsymbol{x} \in \mathcal{B}_s, \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^\top \boldsymbol{b} \le 2 \sup_{k \in [N]} \boldsymbol{u}_k^\top \boldsymbol{b}$. Moreover, for fixed $\boldsymbol{u}_k$, it follows from Conditions A.1 – A.4 such that $\boldsymbol{u}_k^\top \boldsymbol{b}$ is the sum of independent variables $X_{e,i} - \mathbb{E}[X_{e,i}]$, and $X_{e,i}$ is the product of two zero-mean sub-Gaussian random variables with parameters $\kappa_U^{1/2}\sigma_x$ and σ_ε, respectively. Following a similar procedure of applying Lemma C.1, Lemma C.2, and the union bound as above concludes the proof.

C.8 Proof of Proposition C.8

We need the following lemma.

Lemma C.9. *Assume Conditions A.1–A.4 hold. Define the following event*

$$\mathcal{A}_{5,t} = \left\{\widehat{\mathsf{Q}}(\boldsymbol{\beta}^*; \gamma, \boldsymbol{\omega}) \le \sigma_\varepsilon^2 \left(1 + c_1 \sqrt{\frac{t}{n_*}} + c_1 \kappa_U \sigma_x^2 \gamma \frac{s^*(\log(2s^*|\mathcal{E}|) + t)}{\bar{n}}\right)\right\} \tag{C.40}$$

for some universal constant, we have $\mathbb{P}(\mathcal{A}_{5,t}) \ge 1 - 2e^{-t}$.

Proof of Lemma C.9. It follows from the definition that

$$\begin{aligned}
\widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) &= \sum_{e \in \mathcal{E}} \omega^{(e)} \left\{\widehat{\mathbb{E}}|\varepsilon^{(e)}|^2 + \gamma \sum_{j \in S^*} \left(\widehat{\mathbb{E}}[x_j^{(e)} \varepsilon^{(e)}]\right)^2\right\} \\
&= \sum_{e \in \mathcal{E}} \omega^{(e)} \mathbb{E}|\varepsilon^{(e)}|^2 + \sum_{e \in \mathcal{E}} \omega^{(e)} \left(\widehat{\mathbb{E}}|\varepsilon^{(e)}|^2 - \mathbb{E}|\varepsilon^{(e)}|^2\right) + \gamma \sum_{e \in \mathcal{E}} \omega^{(e)} \left(\widehat{\mathbb{E}}[x_j^{(e)} \varepsilon^{(e)}]\right)^2.
\end{aligned}$$

It follows from Condition A.4 that

$$\sum_{e\in\mathcal{E}} \omega^{(e)}\mathbb{E}|\varepsilon^{(e)}|^2 \le \sum_{e\in\mathcal{E}} \omega^{(e)}\sigma_\varepsilon^2 = \sigma_\varepsilon^2, \tag{C.41}$$

and for any $t > 0$,

$$\mathbb{P}\left[\sum_{e\in\mathcal{E}} \omega^{(e)}\left(\widehat{\mathbb{E}}|\varepsilon^{(e)}|^2 - \mathbb{E}|\varepsilon^{(e)}|^2\right) \le C'\sigma_\varepsilon^2\left(\sqrt{\frac{t}{n_{\boldsymbol{\omega}}}} + \frac{t}{n_*}\right)\right] \ge 1 - e^{-x} \tag{C.42}$$

for some universal constant $C' > 0$. At the same time, for any fixed $e \in \mathcal{E}$, it follows from Lemma C.1 and Lemma C.2 that, for any $x > 0$ and $j \in [S^*]$,

$$\mathbb{P}[\mathcal{K}_t(e,j)] = \mathbb{P}\left[\left|\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}] - \mathbb{E}[x_j^{(e)}\varepsilon^{(e)}]\right| \ge C'\kappa_U^{1/2}\sigma_x\sigma_\varepsilon\left(\sqrt{\frac{x}{n^{(e)}}} + \frac{x}{n^{(e)}}\right)\right] \le 1 - 2e^{-x}.$$

Note $\mathbb{E}[x_j^{(e)}\varepsilon^{(e)}] = 0$. Then under $\mathcal{K}_t = \bigcap_{j\in S^*, e\in\mathcal{E}} \mathcal{K}_{t+\log(2s^*|\mathcal{E}|)}(e,j)$, which occurs with probability at least $1 - e^{-t}$, we have

$$\sum_{j\in S^*}\sum_{e\in\mathcal{E}} \omega^{(e)}\left(\widehat{\mathbb{E}}[x_j^{(e)}\varepsilon^{(e)}]\right)^2 \le (C')^2\kappa_U\sigma_x^2\sigma_\varepsilon^2\frac{s^*(\log(2s^*|\mathcal{E}|) + t)}{\bar{n}} \tag{C.43}$$

provided $t + \log(2s^*|\mathcal{E}|) \le n_{\min}$. Putting the three bounds (C.41), (C.42) and (C.43) together completes the proof. □

Now we are ready to prove Proposition C.8.

Proof of Proposition C.8. Observing that

$$\left\{|\widehat{S}| \le 2s^*\right\} \subseteq \left\{\forall\boldsymbol{\beta} \text{ with } \mathrm{supp}(\boldsymbol{\beta}) > 2s^*,\quad \widehat{\mathsf{Q}}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_0 > \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) + \lambda\|\boldsymbol{\beta}^*\|_0\right\} := \mathcal{C}_t,$$

it remains to show that $\mathbb{P}(\mathcal{C}_t) \ge 1 - e^{-t}$. To this end, we use a peeling device. Let $\alpha_\ell = 2^\ell s^*$, then

$$\mathcal{C}_t = \bigcup_{\ell=1}^{\lceil\log_2(p/s^*)\rceil - 1}\left\{\forall\boldsymbol{\beta}\in\mathcal{B}_{\alpha_{\ell+1}}\setminus\mathcal{B}_{\alpha_\ell},\quad \widehat{\mathsf{Q}}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_0 > \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) + \lambda\|\boldsymbol{\beta}^*\|_0\right\} = \bigcup_{\ell=1}^{\lceil\log_2(p/s^*)\rceil - 1}\mathcal{C}_t(\ell).$$

We claim that if λ satisfies the conditions presented in the statement, then

$$\bigcup_{\ell=1}^{\lceil\log_2(p/s^*)\rceil - 1}\mathcal{C}_t(\ell) \subseteq \mathcal{A}_{5,t} \cup \left\{\bigcup_{\ell=1}^{\lceil\log_2(p/s^*)\rceil - 1}\mathcal{A}_{3,u_\ell}(\alpha_{\ell+1}s^*) \cup \mathcal{A}_{4,u_\ell}(\alpha_{\ell+1}s^*)\right\}. \tag{C.44}$$

with $u_\ell = t + \log(\lceil\log_2(p/s^*)\rceil) \le t + \log(ep/s^*)$. Denote $\phi_* = s^* + \log(ep/s^*) + t$ and

$$\widetilde{s} = \frac{n_*}{2\{\log(ep/s^*) + t\}} \wedge \frac{\left(\frac{1}{12C_1}\frac{\kappa_L}{\gamma\kappa_U^2\sigma_x^2}\right)^2 n_*}{\log(ep/s^*) + t} \wedge \frac{n_\dagger}{\{\log(ep/s^*) + t\}\sqrt{\phi_*}(\gamma/\kappa_L)\kappa_U^{3/2}\sigma_x^2},$$

where C_1 is some universal constant to be determined. We prove this claim by considering the following two cases.

Case 1, $\alpha_{\ell+1}s^* \le \widetilde{s}$. Let $S = \mathrm{supp}(\boldsymbol{\beta})$. Denote $\phi_s = \alpha_{\ell+1}s^* \log(ep/s^*) + t$. In this case, there exists some universal constant C_1 such that if $\mathcal{A}_{3,u_\ell}(\alpha_{\ell+1}s^*) \cap \mathcal{A}_{4,u_\ell}(\alpha_{\ell+1}s^*)$ occurs, then we have, for any $\boldsymbol{\beta} \in \mathcal{B}_{\alpha_{\ell+1}}$,

$$\begin{aligned}
\widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) &= \widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \mathsf{Q}(\boldsymbol{\beta}) + \mathsf{Q}(\boldsymbol{\beta}) - \mathsf{Q}(\boldsymbol{\beta}^*) + \mathsf{Q}(\boldsymbol{\beta}^*) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \\
&\overset{(a)}{\ge} \frac{\kappa_L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + \frac{\gamma}{6}\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega}) + \widehat{\mathsf{R}}(\boldsymbol{\beta}) - \mathsf{R}(\boldsymbol{\beta}) + \mathsf{R}(\boldsymbol{\beta}^*) - \widehat{\mathsf{R}}(\boldsymbol{\beta}^*) \\
&\qquad - \gamma\left(\mathsf{J}(\boldsymbol{\beta}) - \widehat{\mathsf{J}}(\boldsymbol{\beta}) - \mathsf{J}(\boldsymbol{\beta}^*) + \widehat{\mathsf{J}}(\boldsymbol{\beta}^*)\right) \\
&\overset{(b)}{\ge} \left(\frac{\kappa_L}{2} - C_1\kappa_U^2\sigma_x^2\gamma\sqrt{\frac{\phi_s}{n_*}}\right)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 \\
&\qquad + \frac{\gamma}{6}\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega}) - C_1\gamma\sqrt{\sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}]\right\|_2^2}\kappa_U^{1/2}\sigma_x\sigma_\varepsilon\sqrt{\frac{\phi_s}{n_*}} \\
&\qquad - C_1\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \times \gamma\kappa_U^{3/2}\sigma_x^2\sigma_\varepsilon\sqrt{\frac{\phi_s}{n_*}} \\
&\qquad - C_1\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \times \gamma\kappa_U^{3/2}\sigma_x^3\sigma_\varepsilon\left(\frac{\sqrt{\phi_*}\sqrt{\phi_s}}{\bar{n}} + \frac{\phi_s\sqrt{\phi_*}}{n_\dagger}\right) \\
&\qquad - C_1\gamma|S^*\setminus S|\kappa_U\sigma_x^2\sigma_\varepsilon^2\frac{t+\log p}{\bar{n}} - C_1\gamma\kappa_U\sigma_x\sigma_\varepsilon^2\frac{\phi_s}{n_*} \\
&= \mathsf{I}_1 + \mathsf{I}_2 + \mathsf{I}_3 + \mathsf{I}_4 + \mathsf{I}_5,
\end{aligned}$$

where (a) follows from Proposition C.5, (b) follows from the bounds in Lemma C.6 & C.7 together with the facts that $\log(|\mathcal{E}|) \lesssim \log p$, $\phi_s/n_* \le \widetilde{s}(\log(ep/s^*) + t)/n_* \le 1$, and $|S\setminus S^*|\log(ep/|S\setminus S^*|) + s^* \le 2|S\setminus S^*|\log(ep/s^*)$ because $|S\setminus S^*| \ge s^*$. For I_1, we have $\mathsf{I}_1 \ge \frac{5}{12}\kappa_L\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$ whenever

$$C_1\kappa_U^2\sigma_x^2\gamma\sqrt{\frac{\phi_s}{n_*}} \le C_1\kappa_U^2\sigma_x^2\gamma\sqrt{\frac{\widetilde{s}\{\log(ep/s^*) + t\}}{n_*}} \le \frac{\kappa_L}{12}.$$

Denote $\lozenge = 0.5(6C_1)^2(\gamma/\kappa_L)^2\kappa_U^3\sigma_x^4\sigma_\varepsilon^2$. For I_2, it follows from Lemma C.10 that

$$\begin{aligned}
\mathsf{I}_2 &\ge \frac{\gamma}{6}\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega}) - C_1\gamma\sqrt{2\kappa_U^2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 + 2\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega})}\kappa_U^{1/2}\sigma_x\sigma_\varepsilon\sqrt{\frac{\phi_s}{n_*}} \\
&\ge \frac{\gamma}{6}\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega}) - \frac{\kappa_L}{12}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - 4\lozenge\cdot\frac{\phi_s}{n_*} - \frac{\gamma}{6}\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega}) \\
&\ge -\frac{\kappa_L}{12}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - 4\lozenge\cdot\frac{(\alpha_{\ell+1}s^*)\{\log(ep/s^*) + t\}}{n_*}
\end{aligned}$$

For I_3 and I_4, it follows from the fact $xy \le \frac{1}{2}(x^2 + y^2)$ that

$$\mathsf{I}_3 \ge -\frac{\kappa_L}{12}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - \lozenge\cdot\frac{(\alpha_{\ell+1}s^*)\{\log(ep/s^*) + t\}}{n_*}$$

and

$$\begin{aligned}
\mathsf{I}_4 \ge &-\frac{\kappa_L}{12}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2 - 2\lozenge\cdot\frac{(\alpha_{\ell+1}s^*)\cdot\phi_*\{\log(ep/s^*) + t\}}{\bar{n}^2} \\
&- 2\lozenge\cdot\frac{(\alpha_{\ell+1}s^*)\cdot\phi_*\widetilde{s}\{\log(ep/s^*) + t\}^2}{n_\dagger^2}.
\end{aligned}$$

For I_5, it follows from (C.12) that

$$\begin{aligned}\mathsf{I}_5 &\geq -C_1\|\boldsymbol{\beta}-\boldsymbol{\beta}_*\|_2^2\gamma\left\{\frac{\kappa_U\sigma_x^2\sigma_\varepsilon^2}{\min_{j\in S^*}|\boldsymbol{\beta}_j^*|^2}\frac{t+\log(es^*p)}{\bar{n}}\right\} - C_1\gamma\kappa_U\sigma_x\sigma_\varepsilon^2\frac{(\alpha_{\ell+1}s^*)\{\log(ep/s^*)+t\}}{n_*} \\ &\geq -\frac{\kappa_L}{12}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 - C_1\gamma\kappa_U\sigma_x\sigma_\varepsilon^2\frac{(\alpha_{\ell+1}s^*)\{\log(ep/s^*)+t\}}{n_*}\end{aligned}$$

if $\bar{n} \geq 12C_1(\gamma/\kappa_L)\kappa_U\sigma_x^2\sigma_\varepsilon^2(t+\log p)/\min_{j\in S^*}|\beta_j^*|^2$. Recall that $\zeta = t + \log(ep/s^*)$. Putting all the pieces together and plugging ζ in, we obtain that, for any $\boldsymbol{\beta} \in \mathcal{B}_{\alpha_{\ell+1}s^*} \setminus \mathcal{B}_{\alpha_\ell s^*}$

$$\begin{aligned}&\lambda\|\boldsymbol{\beta}\|_0 - \lambda\|\boldsymbol{\beta}^*\|_0 + \widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \\ &\overset{(a)}{\geq} \lambda\left(\alpha_\ell - 1\right)s^* + \widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \geq \frac{\lambda}{4}\alpha_{\ell+1}s^* + \widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \\ &\geq \frac{1}{2}\left(\kappa_L - \frac{\kappa_L}{6}\times 5\right)\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 \\ &\qquad + \alpha_{\ell+1}s^*\left\{\frac{\lambda}{4} - 6\Diamond\cdot\left(\frac{\zeta}{n_*} + \frac{\phi_*\cdot\zeta}{\bar{n}^2} + \frac{\phi_*\widetilde{s}\cdot\zeta^2}{n_\dagger^2}\right)\right\},\end{aligned}$$

where (a) follows from the fact that $\boldsymbol{\beta} \notin \mathcal{B}_{\alpha_\ell s^*}$. Recall our choice of $\widetilde{s}$. We can conclude that if

$$\begin{aligned}\lambda &\geq 432C_1^2\kappa_U^3\sigma_x^4(\gamma/\kappa_L)^2\sigma_\varepsilon^2\left(\frac{\zeta}{n_*} + \frac{\zeta\phi_*}{\bar{n}^2}\right) + 432C_1^2\kappa_U^{3/2}\sigma_x^2(\gamma/\kappa_L)\sigma_\varepsilon^2\frac{\zeta\sqrt{\phi_*}}{n_\dagger} \\ &\geq 24\Diamond\cdot\left(\frac{\zeta}{n_*} + \frac{\phi_*\cdot\zeta}{\bar{n}^2} + \frac{\phi_*\widetilde{s}\cdot\zeta^2}{n_\dagger^2}\right),\end{aligned} \tag{C.45}$$

then

$$\underbrace{\left\{\forall\boldsymbol{\beta} \in \mathcal{B}_{\alpha_{\ell+1}s^*} \setminus \mathcal{B}_{\alpha_\ell s^*},\quad \widehat{\mathsf{Q}}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_0 - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) - \lambda\|\boldsymbol{\beta}^*\|_0 > 0\right\}}_{\mathcal{C}_t(\ell)} \subseteq \mathcal{A}_{3,u_\ell}(\alpha_{\ell+1}s^*) \cup \mathcal{A}_{4,u_\ell}(\alpha_{\ell+1}s^*).$$

Case 2, $\alpha_{\ell+1}s^* \geq \widetilde{s}$. Observe that $\alpha_\ell s^* \geq \frac{1}{2}\alpha_{\ell+1}s^* \geq \frac{1}{2}\widetilde{s}$ in this case. Hence the following holds

$$\lambda\|\boldsymbol{\beta}\|_0 - \lambda\|\boldsymbol{\beta}^*\| + \widehat{\mathsf{Q}}(\boldsymbol{\beta}) - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \geq \frac{\lambda}{2}\widetilde{s} - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) - \lambda s^* \geq \frac{\lambda}{4}\widetilde{s} - \widehat{\mathsf{Q}}(\boldsymbol{\beta}^*)$$

provided $4s^* \leq \widetilde{s}$. At the same time, if $t \leq n_*$ and $\bar{n} \geq \kappa_U\sigma_x^2\gamma s^*(\log(2s^*|\mathcal{E}|)+t)$, then it follows from Lemma C.9 that under $\mathcal{A}_{5,t}$,

$$\widehat{\mathsf{Q}}(\boldsymbol{\beta}^*) \leq (2+C_2)\sigma_\varepsilon^2,$$

where $C_2 = c_1$ is the constant in Lemma C.9. At the same time, it follows from our construction of $\widetilde{s}$ that

$$(\widetilde{s})^{-1} \leq (12C_1)^2\kappa_U^4\sigma_x^4(\gamma/\kappa_L)^2\frac{\zeta}{n_*} + \kappa_U^{3/2}\sigma_x^2(\gamma/\kappa_L)\frac{\zeta\sqrt{\phi_*}}{n_\dagger}.$$

Therefore, we can argue that $\mathcal{C}_t(\ell) \subseteq \mathcal{A}_{5,t}$ when

$$\lambda \geq 8(2+C_2)(16C_1)^2\kappa_U^4\sigma_x^4\sigma_\varepsilon^2(\gamma/\kappa_L)^2\frac{\zeta}{n_*} + 8(2+C_2)\kappa_U^{3/2}\sigma_x^2\sigma_\varepsilon^2(\gamma/\kappa_L)\frac{\zeta\sqrt{\phi_*}}{n_\dagger} \geq 8(2+C_2)\sigma_\varepsilon^2\widetilde{s}^{-1}. \tag{C.46}$$

Now, we combine the two cases. According to the above discussion, we find that under the conditions

$$\widetilde{s} \geq 4s^* \qquad \text{and} \qquad \bar{n} \geq C'\kappa_U\sigma_x^2\gamma(t+\log p)\left\{s^* + \sigma_\varepsilon^2/(\kappa_L\min_{j\in S^*}|\beta_j^*|^2)\right\},$$

we have (C.44) holds if our choice of λ satisfies (C.45) and (C.46). It then follows from the union bound that

$$\mathbb{P}\left[\mathcal{C}_t^c\right] \le \sum_{\ell=1}^{\lceil \log_2(p/s^*)\rceil - 1} \{\mathbb{P}[\mathcal{A}_{3,u_\ell}(\alpha_{\ell+1}s^*)^c] + \mathbb{P}[\mathcal{A}_{4,u_\ell}(\alpha_{\ell+1}s^*)]\} + \mathbb{P}(\mathcal{A}_{5,t}^c) \le 3e^{-t}.$$

This completes the proof.

□

C.9 Technical Lemmas

Lemma C.10. *Under Conditions A.1 and A.2, we have*

$$\forall \boldsymbol{\beta}\in\mathbb{R}^p \qquad \sum_{e\in\mathcal{E}} \omega^{(e)} \left\|\mathbb{E}[\boldsymbol{x}^{(e)}_{\mathrm{supp}(\boldsymbol{\beta})}\varepsilon^{(e)}]\right\|_2^2 \le 2\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega}) + 2\kappa_U^2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2$$

Proof of Lemma C.10. Let $S=\mathrm{supp}(\boldsymbol{\beta})$, it follows from the definition of J that

$$\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega}) = \sum_{e\in\mathcal{E}} \omega^{(e)}\|\mathbb{E}[(y^{(e)}-\boldsymbol{\beta}^\top\boldsymbol{x}^{(e)})\boldsymbol{x}_S^{(e)}]\|_2 = \sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\mathbb{E}\left[(\varepsilon^{(e)}+(\boldsymbol{\beta}^*)^\top\boldsymbol{x}^{(e)}-\boldsymbol{\beta}^\top\boldsymbol{x}^{(e)})\boldsymbol{x}_S^{(e)}\right]\right\|_2.$$

Denote $\overline{S}=S^*\cup S$. Then it follows from the fact $(a+b)^2\le 2(a^2+b^2)$ and Condition A.2 that

$$\begin{aligned}\sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}]\right\|_2^2 &= \sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}+\boldsymbol{x}_S^{(e)}(\boldsymbol{x}_{\overline{S}}^{(e)})^\top(\boldsymbol{\beta}^*-\boldsymbol{\beta})_{\overline{S}}]-\boldsymbol{\Sigma}_{S,\overline{S}}^{(e)}(\boldsymbol{\beta}^*-\boldsymbol{\beta})_{\overline{S}}\right\|_2^2\\ &\le 2\sum_{e\in\mathcal{E}}\omega^{(e)}\left\|\mathbb{E}[\boldsymbol{x}_S^{(e)}\varepsilon^{(e)}+\boldsymbol{x}_S^{(e)}(\boldsymbol{x}_{\overline{S}}^{(e)})^\top(\boldsymbol{\beta}^*-\boldsymbol{\beta})_{\overline{S}}]\right\|_2^2 + 2\kappa_U^2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2\\ &\le 2\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega})+2\kappa_U^2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2.\end{aligned}$$

This completes the proof. □

Lemma C.11. *Under Condition A.2 and Condition A.4, we have that, for any $e\in\mathcal{E}$,*

$$\left\|\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right\|_2 \le \sigma_\varepsilon\kappa_U^{1/2}.$$

Proof of Lemma C.1. Let $\boldsymbol{z}=((\boldsymbol{x}^{(e)})^\top,\varepsilon^{(e)})^\top$ be a $p+1$-dimensional random vector. Define the matrix $\boldsymbol{A}=\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$, we have

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{\Sigma}^{(e)} & \mathbb{E}[\boldsymbol{x}^{(e)}\varepsilon^{(e)}]\\ (\mathbb{E}[\boldsymbol{x}^{(e)}\varepsilon^{(e)}])^\top & \mathbb{E}[|\varepsilon^{(e)}|^2]\end{bmatrix}.$$

The matrix is positive semi-definite. Combining this with the fact that $\boldsymbol{\Sigma}^{(e)}$ is invertible indicates that the Schur complement is non-negative, that is,

$$\mathbb{E}[|\varepsilon^{(e)}|^2]-\left(\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right)^\top\left(\boldsymbol{\Sigma}^{(e)}\right)^{-1}\left(\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right)\ge 0.$$

Hence we have

$$\begin{aligned}\left\|\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right\|^2\kappa_U^{-1} &\overset{(a)}{\le} \left\|\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right\|^2\lambda_{\min}\{(\boldsymbol{\Sigma}^{(e)})^{-1}\}\\ &\le \left(\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right)^\top\left(\boldsymbol{\Sigma}^{(e)}\right)^{-1}\left(\mathbb{E}[\varepsilon^{(e)}\boldsymbol{x}^{(e)}]\right)\le\mathbb{E}[|\varepsilon^{(e)}|^2]\overset{(b)}{\le}\sigma_\varepsilon^2.\end{aligned}$$

where (a) follows from Condition A.2 and (b) follows from Condition A.4. This completes the proof. □

D Omitted Discussions in Main Text

D.1 Comparison with Fan & Liao (2014)

Both Fan & Liao (2014) and our work aim to deal with the problem of endogeneity from a technical perspective. Our constructed focused linear invariance regularizer is similar to their developed FGMM criterion function from the view of the *over-identification* idea they proposed.

Briefly speaking, we say a parameter $\widetilde{\boldsymbol{\beta}}$ is over-identified if there are more restrictions than the degree of freedom. When $|\mathcal{E}| = 1$, there will be exponential number of distinct $\boldsymbol{\beta}$ satisfying $\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) = 0$. To see this, for any $S \subseteq [p]$, one can find some $\boldsymbol{\beta}$ with $\mathrm{supp}(\boldsymbol{\beta}) \subseteq S$ satisfying $|S|$ constraints that $\mathbb{E}[(y^{(1)} - \boldsymbol{\beta}_S^\top \boldsymbol{x}_S^{(1)}) x_j^{(1)}] = 0$ for any $j \in [S]$ because the degree of freedom and the number of constraints are both $|S|$. However, things may be different when $|\mathcal{E}| \geq 2$. In this case, for a given $S \subseteq [p]$, one need to find some $\boldsymbol{\beta}$ with a degree of freedom $|S|$ satisfies $|\mathcal{E}| \cdot |S|$ constraints

$$\forall e \in \mathcal{E}, j \in [S], \qquad \mathbb{E}\left[x_j^{(e)} (y^{(e)} - \boldsymbol{\beta}_S^\top \boldsymbol{x}_S^{(e)})\right] = 0$$

simultaneously to let $\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) = 0$, which does not hold in general.

According to the above discussion, the focused linear invariance regularizer shares a similar spirit with their proposed FGMM from a technical viewpoint since the over-identification of our regularizer comes from multiple environments. In contrast, theirs come from two instrumental variables or two marginal features of the covariate, for example, x_j and x_j^2. However, the statistical models the two papers work on are different. We briefly remark on the differences between our method and theirs using marginal nonlinear features as follows:

1. We are working with multiple environment settings. The necessity of heterogeneous environments and the potential violation of their identification condition are illustrated by Proposition 2.2. In particular, if two marginal features of the covariate are used in their method, the corresponding identification condition is a sufficient condition of the condition that S^* is the only CE-invariant set among $\mathcal{E} = \{1\}$ other than $\emptyset$.

2. We use a linear combination of invariance regularizer and the L_2 loss to avoid collapsing to conservative solutions. At the same time, theirs will suffer from the case where two groups of important variables are independent.

3. We provide a more explainable identification condition in the context of multiple environments and a non-asymptotic upper bound on the critical threshold of the regularization hyper-parameter γ; see the intuition explanations in Appendix D.3 and a formal presentation in Section 4.2. While Fan & Liao (2014) has few discussions on the population-level conditions.

4. The finite sample analyses are completely different because (1) the objective functions are different; and (2) we establish the variable selection consistency for the global minimizer while they only show that some good local minimizer satisfying the variable selection consistency exists. We should carefully deal with the dependence on the number of environments and apply a novel localization argument to obtain a fast rate or a weak condition for variable selection consistency.

D.2 Comparison with IRM

Our constructed EILLS objective is similar to the invariant risk minimization criterion (Arjovsky et al., 2019) by letting $\gamma \to \infty$. To see this, when $\gamma \to \infty$, the population-level EILLS objective becomes

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathsf{R}(\boldsymbol{\beta}; \boldsymbol{\omega}) \qquad s.t. \qquad \mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) = 0.$$

This optimization problem finds the most effective solution in the sense of small L_2 risk among all the LLS-invariant solutions as discussed in Section 3.1. However, it is unclear (1) what it implies when $\mathsf{J}(\boldsymbol{\beta}; \boldsymbol{\omega}) \approx 0$;

and (2) how large γ must be to estimate β^* consistently. Such two concerns are important for the non-asymptotic analysis because of finite-sample data and finite choice of γ. We provide an intuitive explanation in Appendix D.3 addressing the above concerns, which can be treated as an informal presentation of our theory in Section 4.2.

D.3 Debiasing by Regularizer via Bias Differences

Section 3.1 and Appendix D.2 provide a sketchy glimpse at the effects of two losses J and R on the solution. The following concerns remain: (1) we only analyze the implication of focused linear invariance regularizer $\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega})$ on the solution when $\mathsf{J}(\boldsymbol{\beta};\boldsymbol{\omega})=0$, but we can not expect $\mathsf{J}=0$ in practice especially when it comes to the analysis of its empirical analogs with finite sample data; (2) the above paragraph characterizes that property of proposed EILLS estimator in $\gamma\to\infty$, while it is still unknown how large γ is enough. In this part, we will provide a different perspective. In a high-level viewpoint, the L_2 risk will have some bias in the presence of the *pooled linear spurious variables* satisfying $\sum_{e\in\mathcal{E}}\omega^{(e)}\mathbb{E}[\varepsilon^{(e)}x_j^{(e)}]\neq 0$. This is where our proposed regularizer J comes into play: it debiases the pooled least squares loss using the difference of biases between heterogeneous environments, as illuminated below. Such insight also answers the question of how large γ is enough.

We illustrate the phenomenon of debiasing at the population level by considering the simplest case of $|\mathcal{E}|=p=2$ and defer a thorough and rigorous analysis to Section 4.2. Here we let $\omega^{(1)}=\omega^{(2)}=1/2$ and $\boldsymbol{\beta}^*=(1,0)$ such that the first variable x_1 is an important and invariant variable and the second variable x_2 is a linear spurious variable. Moreover, suppose $\lambda(\boldsymbol{\Sigma}^{(e)})\asymp 1$ where $\lambda(\boldsymbol{S})$ represent the eigenvalues of the symmetric matrix $\boldsymbol{S}$.

Under the above toy model, the population-level excess pooled L_2 risk is given

$$\mathsf{R}(\boldsymbol{\beta})-\mathsf{R}(\boldsymbol{\beta}^*)=(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\bar{\boldsymbol{\Sigma}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)-2\times(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top\boldsymbol{b}$$

where $\bar{\boldsymbol{\Sigma}}=\frac{1}{2}(\boldsymbol{\Sigma}^{(1)}+\boldsymbol{\Sigma}^{(2)})$ and $\boldsymbol{b}=(0,b_2)$ with $b_2=\frac{1}{2}\{\mathbb{E}[\varepsilon^{(1)}x_2^{(1)}]+\mathbb{E}[\varepsilon^{(2)}x_2^{(2)}]\}$. This indicates that $\boldsymbol{\beta}^*$ is not the minimizer of $\mathsf{R}(\boldsymbol{\beta})$ when $b_2\neq 0$ and one can further decrease the loss in the direction of $(\bar{\boldsymbol{\Sigma}})^{-1}\boldsymbol{b}$. Furthermore, the decrease in $\mathsf{R}(\boldsymbol{\beta})$ by moving towards this direction is bounded from below by

$$\mathsf{R}(\boldsymbol{\beta})-\mathsf{R}(\boldsymbol{\beta}^*)\ge C_1^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2-C_1\|\boldsymbol{b}\|_2^2$$

for some constant $C_1>0$. Let us see how the invariance regularizer $\mathsf{J}(\boldsymbol{\beta})$ can compromise the bias. Denote $\boldsymbol{b}^{(e)}=(0,\mathbb{E}[\varepsilon^{(e)}x_2^{(e)}])$ be the bias vector in each environment e. When $\boldsymbol{\beta}$ is supported on $\{1,2\}$, we find

$$\begin{aligned}\mathsf{J}(\boldsymbol{\beta})-\mathsf{J}(\boldsymbol{\beta}^*)&=\frac{1}{2}\|\boldsymbol{\Sigma}^{(1)}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)-\boldsymbol{b}^{(1)}\|_2^2+\frac{1}{2}\|\boldsymbol{\Sigma}^{(2)}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)-\boldsymbol{b}^{(2)}\|_2^2\\&\ge 2C_2^{-1}\Big\{\|\boldsymbol{\beta}-\boldsymbol{\beta}^*-(\boldsymbol{\Sigma}^{(1)})^{-1}\boldsymbol{b}^{(1)}\|_2^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}^*-(\boldsymbol{\Sigma}^{(2)})^{-1}\boldsymbol{b}^{(2)}\|_2^2\Big\}\\&\ge C_2^{-1}\|(\boldsymbol{\Sigma}^{(1)})^{-1}\boldsymbol{b}^{(1)}-(\boldsymbol{\Sigma}^{(2)})^{-1}\boldsymbol{b}^{(2)}\|_2\end{aligned}$$

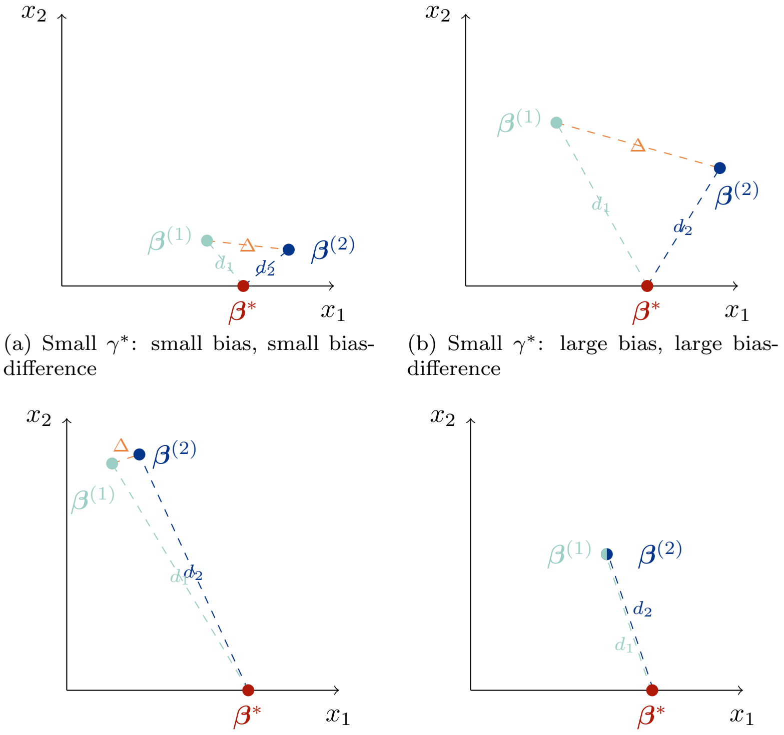
for another constant $C_2>0$. This demonstrates that one needs to pay an extra cost of bias-difference $\ge\gamma C_2^{-1}\|(\boldsymbol{\Sigma}^{(1)})^{-1}\boldsymbol{b}^{(1)}-(\boldsymbol{\Sigma}^{(2)})^{-1}\boldsymbol{b}^{(2)}\|_2$ when adopting pooled linear spurious variables. Such a cost can compromise the gain of the decrease in $\mathsf{R}(\boldsymbol{\beta})$ when a large γ is used. Formally, we have

$$\forall\boldsymbol{\beta}\ \text{ with }\ \|\boldsymbol{\beta}\|_0=2,\qquad\mathsf{Q}(\boldsymbol{\beta})-\mathsf{Q}(\boldsymbol{\beta}^*)\ge C_1^{-1}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2$$

provided

$$\gamma\ge\gamma^*=C_1C_2\frac{\|\boldsymbol{b}\|_2^2}{\|(\boldsymbol{\Sigma}^{(1)})^{-1}\boldsymbol{b}^{(1)}-(\boldsymbol{\Sigma}^{(2)})^{-1}\boldsymbol{b}^{(2)}\|_2}\asymp\frac{\|(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^*)+(\boldsymbol{\beta}^{(2)}-\boldsymbol{\beta}^*)\|_2^2}{\|(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^*)-(\boldsymbol{\beta}^{(2)}-\boldsymbol{\beta}^*)\|_2^2}\tag{D.1}$$

where $\boldsymbol{\beta}^{(e)}=\mathrm{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^2}\,\mathsf{R}^{(e)}(\boldsymbol{\beta})$ is the population risk minimizer for environment $e\in\mathcal{E}$. The R.H.S. of (D.1) follows from the fact that $\boldsymbol{\beta}^{(e)}-\boldsymbol{\beta}^*=(\boldsymbol{\Sigma}^{(e)})^{-1}\boldsymbol{b}^{(e)}$. We present a geometric illustration of the above discussion and how the bias and bias-difference jointly affect the critical threshold γ^* in Fig. 5.



(a) Small γ^*: small bias, small bias-difference

(b) Small γ^*: large bias, large bias-difference

(c) Large γ^*: large bias, small bias-difference

(d) Impossible: non-zero bias, zero bias-difference

Figure 5: A geometric illustration of the bias-difference debiasing idea. We consider the same case where $|\mathcal{E}| = p = 2$, x_1 is the important variable, and x_2 is the pooled linear spurious variable. In each subplot, $\boldsymbol{\beta}^*$ is the true parameter and $\boldsymbol{\beta}^{(e)}$ with $e \in \{1, 2\}$ is the population risk minimizer in each environment. Following the discussion in the text, $d_e = \|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(e)}\|_2 \asymp \|\boldsymbol{b}^{(e)}\|_2$ quantifies the bias of each environment and $\Delta = \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_2$ represents the bias-difference. The four plots demonstrate four cases in which the magnitudes of bias and bias-difference vary, leading to different thresholds γ^* satisfying $\gamma^* \asymp \left(\frac{d_1+d_2}{\Delta}\right)^2$. The above two plots (a) and (b) are the cases where γ^* is of reasonable, constant order. We can see when both bias and bias-difference are relatively small in plot (a) or relatively large in plot (b), and the ratio of the two quantities is within constant order that $d_1 + d_2 \asymp \Delta$, the choice of γ^* is also of constant order. However, when the bias is much larger than the bias difference that $d_1 + d_2 \gg \Delta$ in (c), one needs to use a large γ^* to accommodate the gain in loss decrease from selecting pooled linear spurious variable x_2. (d) present a case where the variable set $\{1, 2\}$ is also LLS-invariance across the two environments because $\beta^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ coincides. In this case, our proposed EILLS approach will fail and converge to the spurious solution $\beta^{(1)} = \boldsymbol{\beta}^{(2)}$ instead of recovering $\boldsymbol{\beta}^*$.

D.4 Implementation Details of Simulations in Section 5

The structural assignment of the SCM in $e = 1$ and $e = 2$ are as follows

$$
\begin{array}{ll}
x_1^{(e)} \leftarrow u_1^{(e)} & \\
x_4^{(1)} \leftarrow u_4^{(1)} & x_4^{(2)} \leftarrow (u_4^{(2)})^2 - 1 \\
x_2^{(e)} \leftarrow \sin(x_4^{(e)}) + u_2^{(e)} & \\
x_3^{(e)} \leftarrow \cos(x_4^{(e)}) + u_3^{(e)} & \\
x_5^{(e)} \leftarrow \sin(x_3^{(e)} + u_5^{(e)}) & \\
x_{10}^{(e)} \leftarrow 2.5x_1^{(e)} + 1.5x_2^{(e)} + u_{10}^{(e)} & \\
y^{(e)} \leftarrow 3x_1^{(e)} + 2x_2^{(e)} - 0.5x_3^{(e)} + u_{13}^{(e)} & \\
x_6^{(e)} \leftarrow 0.8y^{(e)}u_6^{(e)} & \\
x_7^{(1)} \leftarrow 0.5x_3^{(1)} + y^{(1)} + u_7^{(1)} & x_7^{(2)} \leftarrow 4x_3^{(2)} + \tanh(y^{(2)}) + u_7^{(2)} \\
x_8^{(e)} \leftarrow 0.5x_7^{(e)} - y^{(e)} + x_{10}^{(e)} + u_8^{(e)} & \\
x_9^{(e)} \leftarrow \tanh(x_7^{(e)}) + 0.1\cos(x_8^{(e)}) + u_9^{(e)} & \\
x_{11}^{(e)} \leftarrow 0.4(x_7^{(e)} + x_8^{(e)}) * u_{11}^{(e)} & \\
x_{12}^{(e)} \leftarrow u_{12}^{(e)} &
\end{array}
$$

Here $u_1^{(e)}, \ldots, u_{13}^{(e)} \sim \mathcal{N}(\mathbf{0}, I_{13\times 13})$ for all the $e \in \mathcal{E}$.

Implementation. We use brute force search to calculate $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$. To be specific, we enumerate all the possible support set $S \in 2^{[p]}$, for each fixed S, the EILLS objective (3.7) is a quadratic function of $[\boldsymbol{\beta}]_S$, whose minimum value $\widehat{\boldsymbol{\beta}}^{(S)}$ can be found by setting the first order condition to be held, that is,

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(S)} &= \mathop{\mathrm{argmin}}_{\mathrm{supp}(\boldsymbol{\beta})\subseteq S} \sum_{e\in\mathcal{E}} \omega^{(e)} \widehat{\mathbb{E}}[|y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}|^2] + \gamma \sum_{e\in\mathcal{E}} \left\| \widehat{\mathbb{E}}[\{y^{(e)} - \boldsymbol{\beta}^\top \boldsymbol{x}^{(e)}\}\boldsymbol{x}_S^{(e)}] \right\|_2^2 \\
&= \left[\widehat{\boldsymbol{\beta}}_S^{(S)}, \widehat{\boldsymbol{\beta}}_{S^c}^{(S)}\right] \\
&= \left[\left\{ \sum_{e\in\mathcal{E}} \omega^{(e)} \widehat{\boldsymbol{\Sigma}}_S^{(e)} + \gamma \sum_{e\in\mathcal{E}} \omega^{(e)} (\widehat{\boldsymbol{\Sigma}}_S^{(e)})^2 \right\}^{-1} \left\{ \sum_{e\in\mathcal{E}} \omega^{(e)} \widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} y^{(e)}] + \gamma \sum_{e\in\mathcal{E}} \omega^{(e)} \widehat{\boldsymbol{\Sigma}}_S \widehat{\mathbb{E}}[\boldsymbol{x}_S^{(e)} y^{(e)}] \right\}, \mathbf{0} \right].
\end{aligned}
$$

Then $\widehat{\boldsymbol{\beta}}_{\mathsf{Q}}$ is assigned to be $\widehat{\boldsymbol{\beta}}^{(S)}$ with the minimum empirical objective value $\widehat{\mathsf{Q}}$. The total computational complexity is $\mathcal{O}(2^p p^3 + \sum_{e\in\mathcal{E}} n^{(e)} p^2)$.

We argue here it is possible to do some relaxation on the objective (3.7) such that it may be efficiently solved via gradient descent or other methods. However, minimizing $\widehat{\mathsf{Q}}$ is still a non-convex problem. We leave an efficient implementation as future work since this paper focused on a thorough statistical analysis for the multiple environment linear regression.

Implementation for Other Invariance Based Methods. It is noteworthy that all the invariance-based methods have hyper-parameters related to "invariance" besides. For example, the hyper-parameters balancing least squares and invariance regularizer in IRM and Anchor regression, and the Type-I error threshold α in ICP. The criterion for choosing this type of hyper-parameter is fundamentally different from that for hyper-parameters controlling the statistical complexity, such as L_1/L_2 regularization. We can use a train/valid split for the latter and choose the hyper-parameters using the validation dataset. On the contrary, the optimal hyper-parameter for the former should be tuned using the "test dataset". In the comparison experiment, the choice of hyper-parameters for ICP, IRM, and Anchor regression are picked in an oracle manner, that is, we enumerate all the possible hyper-parameters and choose the one that minimizes the ℓ_2

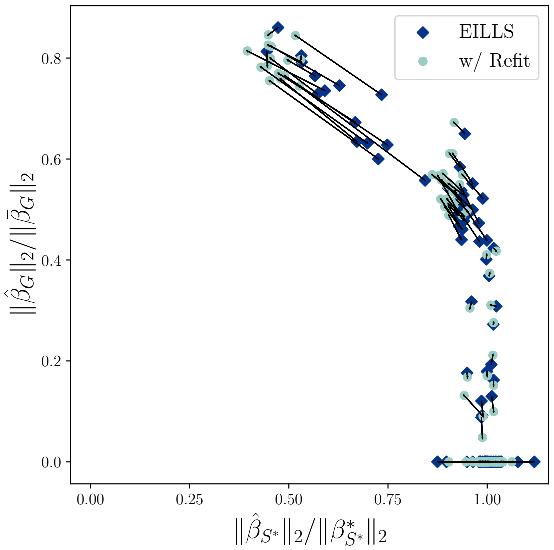


Figure 6: Visualization of solutions EILLS and EILLS with refitting produce in 100 replications when $n = 100$.

prediction error $\|\bar{\mathbf{\Sigma}}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2$. We set the candidate set to be $\{0, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1\}$ for IRM, $\{0, 1, 2, 4, 8, 10, 15, 20, 30, 40, 60, 80, 90, 100, 150, 200, 500, 1000, 5000, 10000\}$ for Anchor regression. For ICP, we try the Type-I error parameter in $\{0.9, 0.95, 0.99, 0.995\}$. We use the dummy variable $1\{E = 1\}$ as the anchor variable for Anchor regression.

Configurations for figures. Fig. 3 (a) plots the estimated coefficients $\widehat{\beta}_j$ over $\gamma \in \{0, 1, 2, 3, 8\} \cup \{5k : k \in [19]\} \cup \{100j : j \in [9]\} \cup \{1000\ell : \ell \in [10]\}$. Fig. 3 (b) (c) enumerates n in $\{100k : k \in [10]\} \cup \{1500, 2000\}$ and all the plotted quantities are the average over 500 replications.

D.5 A Diagnosis on Why Refitting is Worse than Vanilla EILLS

In Fig. 3 (b), we see that the performance of EILLS with refitting is slightly worse than that of the vanilla EILLS estimator. Here, we provide a possible explanation for why there is a noticeable gap when $n = 100$ from the observations in simulation studies and leave the finer analysis for future studies. Fig. 6 visualizes the solution EILLS and EILLS with refitting produced in 100 replications when $n = 100$. In the plot, each point (x, y) with marker m represents a solution $\widehat{\boldsymbol{\beta}}$ that the method m produces. Here, x denotes the relative ℓ_2 norm restricted to the true important variables set S^*, calculated as $\|\widehat{\boldsymbol{\beta}}_{S^*}\|_2/\|\boldsymbol{\beta}^*_{S^*}\|_2$; and y represents the relative ℓ_2 norm restricted to the pooled linear spurious variables set G, expressed as $\|\widehat{\boldsymbol{\beta}}_G\|_2/\|\bar{\boldsymbol{\beta}}_G\|_2$ with $\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{\beta}}^{([12])}$. The solutions in each trial are connected by an arrow from EILLS to EILLS with refitting, indicating that the latter is the refitted version of the former.

As shown in Fig. 6, some solutions falsely select variables in $G = \{7, 8, 9\}$. For these cases, the EILLS solutions tend to be more dependent on the variables in S^* and less dependent on the variables in G than the EILLS with refitting solution. In other words, the EILLS method yields a more robust solution when the variable selection property fails to hold. We guess that will contribute to the ℓ_2 estimation error gap between the two methods.

D.6 EILLS by Gumbel Approximation

In this section, we briefly describe how to use stochastic approximation and Gumbel approximation to handle the combinatorial nature of optimization and let a variant of gradient descent continue to work. This is supported by simulation results when $p = 70$. We follow the notations in (Gu et al., 2024).

The original EILLS objective can be written as

$$(\widehat{\beta}, \widehat{a}) \in \underset{\beta\in\mathbb{R}^p, a\in\{0,1\}^p}{\operatorname{argmin}} \underbrace{\sum_{e\in\mathcal{E}} \widehat{\mathbb{E}}[|Y^{(e)} - (\beta\odot a)^\top X^{(e)}|^2] + \gamma \sum_{j=1}^{p} a_j \left|\widehat{\mathbb{E}}[\{Y^{(e)} - (\beta\odot a)^\top X^{(e)}\}X_j]\right|^2}_{\widehat{\mathsf{Q}}_\gamma(\beta,a)}, \tag{D.2}$$

where $\odot$ is the point-wise multiplication, i.e., $[x\odot y]_j = x_j y_j$ for any $x, y \in \mathbb{R}^d$. This step is used to disentangle the effect of variable selection and parameter estimation, but the combinatorial nature remains. To avoid this, we first rewrite the optimization as a "continuous" optimization:

$$(\widehat{\beta}, \widehat{w}) \in \underset{\beta\in\mathbb{R}^p, w\in\mathbb{R}^p}{\operatorname{argmin}} \mathbb{E}_{B(w)}\left[\widehat{\mathsf{Q}}_\gamma(\beta, B(w))\right],$$

where the j^{th} component of $B(w) \in \{0,1\}^d$ follows an independent Bernoulli with probability of success $\sigma(w_j) = \exp(w_j)/(1+\exp(w_j))$. This is easily seen by taking $\widehat{w} = \operatorname{logit}(\widehat{a}) = \log(\frac{\widehat{a}}{1-\widehat{a}})$ (taking values $\pm\infty$). Note that $B_j(w_j) = I(\operatorname{logit}(U_j) \le w_j)$ is discontinuous in w_j where $U_j \sim$ Uniform[0,1], but can be approximated as

$$B_j(w_j) \approx \frac{1}{1+e^{(\operatorname{logit}(U_j)-w_j))/\tau}} \equiv V_\tau(U_j, w_j) \quad \text{as} \quad \tau \to 0^+, \tag{D.3}$$

for which its gradient can be taken. Let

$$A_\tau(U, w) = (V_\tau(U_1, w_1), \ldots, V_\tau(U_d, w_d))^\top \in \mathbb{R}^d$$

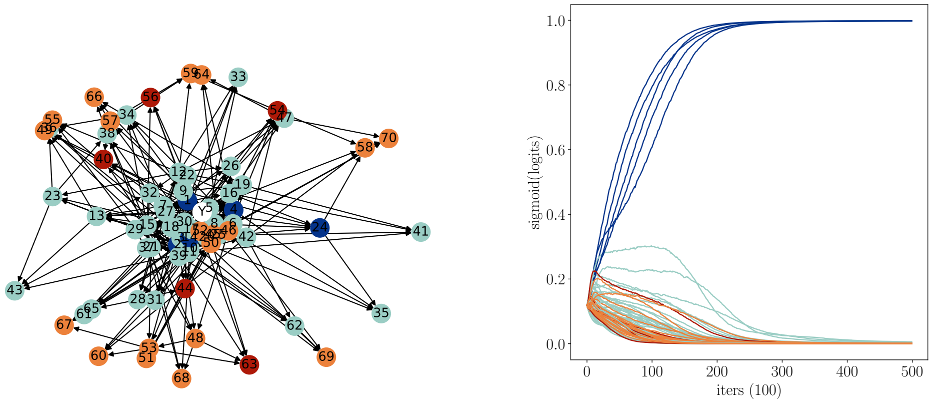
with $\{U_j\}_{j=1}^d$ being i.i.d. uniform random variables. One can approximate the original objective (D.2) by

$$(\widehat{\beta}, \widehat{w}) \underset{\beta\in\mathbb{R}^p, w\in\mathbb{R}^p}{\operatorname{argmin}} \mathbb{E}_U\left[\widehat{\mathsf{Q}}_\gamma(\beta, A_\tau(U, w))\right]. \tag{D.4}$$

Since $\operatorname{logit}(U_j) \overset{d}{=} U_{j,1} - U_{j,2}$ with $\{U_{j,1}, U_{j,2}\}_{j=1}^d$ being i.i.d. Gumbel(0,1) random variables, the approximation (D.3) is also referred to as the Gumbel approximation. Given (D.4), one can adopt the following variants of gradient descent in Algorithm 1.

Algorithm 1 Scaling EILLS to High-dimension by Gumbel Trick

```
Hyper-parameter: number of iterations $T$, hyper-parameter $\gamma$
Gumbel parameters: initial/final temperature $(\tau_0, \tau_T)$, anneal rate $\rho$, anneal iteration $T_\tau$.
Input: data $\{(X_i^{(e)}, Y_i^{(e)})\}_{i\in[n], e\in\mathcal{E}}$
Initialize $\beta, w$
Set $\tau = \tau_0$
for $t \in \{1, \ldots, T\}$ do
    $\tau = \max(\tau_T, \tau \times \rho)$ if $t \bmod T_\tau = 0$.
    Sample $\{U_{j,1}, U_{j,2}\}_{j=1}^p$ from Gumbel(0,1).
    Update $\beta, w$ by descending its gradient
        $\nabla_{(\beta,w)}\left[\widehat{\mathsf{Q}}_\gamma(\beta, A_\tau(U, w))\right]$
end for
Output: estimate $\beta \odot \sigma(w)$.
```



(a) (b)

Figure 7: The visualization of (a) the SCM and (b) the $\sigma(w)$ during training in one trial for the FAIR-Linear estimator. We use different colors to represent the different relationships with Y: blue = parent, red = child, orange = offspring, lightblue = other.

D.6.1 Simulation Results

This section is to illustrate the performance of the above Gumbel-trick optimized EILLS estimator under moderate high dimension $p = 70$, where brute force search is not feasible. The data-generating process and the experimental setting are the same as Section 5.2.1 in Gu et al. (2024).

Data Generating Process. We consider the case where $|\mathcal{E}| = 2$ and the data $(X^{(e)}, Y^{(e)})$ in each environment $e \in \{0, 1\}$ are generated from two SCMs sharing the same causal relationship between variables. For each trial, we first generate the parent-children relationship among the variables. We enumerate all the $i \in [p+1]$. For each $i \in [p+1]$, we randomly pick at most 4 parents for the variable Z_i from $\{Z_1, \ldots, Z_{i-1}\}$, this step ensures that the induced graph is a DAG. We use fixed $p = 70$, and let the variable Z_{36} be Y and the rest variables constitute the covariate X, that is, we let $(Z_1, \ldots, Z_{35}, Z_{36}, Z_{37}, \ldots, Z_{71}) = (X_1, \ldots, X_{35}, Y, X_{36}, \ldots, X_{70})$. We also enforce that Y has at least 5 parents and at least 5 children by adding parents and children when needed. The structural assignment for each variable Z_j is defined as

$$Z_j^{(e)} \leftarrow \sum_{k \in \mathtt{pa}(j)} C_{j,k}^{(e)} f_{j,k}^{(e)}(Z_k^{(e)}) + C_{j,j}^{(e)} \varepsilon_j$$

where $(\varepsilon_1, \ldots, \varepsilon_{71})$ are independent standard normal random variables. For $j \neq 36$, $f_{j,k}^{(e)}$ are sampled randomly from the candidate functions $\{\cos(x), \sin(x), \sin(\pi x), x, 1/(1+e^{-x})\}$, $C_{j,k}^{(e)}$ are sampled from the uniform distribution on $[-1.5, 1.5]$ with $|C_{j,j}^{(e)}| \geq 0.5$. For $j = 36$ and $k < j$, we have $f_{36,k}^{(e)}(x) = x$ and $C_{36,k}^{(0)} \equiv C_{36,k}^{(1)}$ for linearity and invariance. The above data-generating process can be regarded as one observation environment $e = 0$ and an interventional environment $e = 1$ where the random and simultaneous interventions are applied to all the variables other than the variable Y, while the assignment from Y's parent to Y remains and furnishes the target regression function $f^*(x) = \sum_{k \in \mathtt{pa}(36)} C_{36,k}^{(e)} x_k$ in pursuit. In this case, we let $S^* = \mathtt{pa}(36)$ and β^* with support set S^* be such that $\beta_j^* = C_{36,k}^{(0)} = C_{36,k}^{(1)}$ for any $k \in S^*$. We also let the noise variance be different for the two environments, i.e., $C_{36,36}^{(0)} \neq C_{36,36}^{(1)}$. Now, the model only has conditional expectation invariance rather than the full conditional distribution invariance. Fig. 7 (a) visualizes the induced graph in one trial. The complex cause-effect relationships in high-dimensional variables make the problem of estimating β^* very challenging.

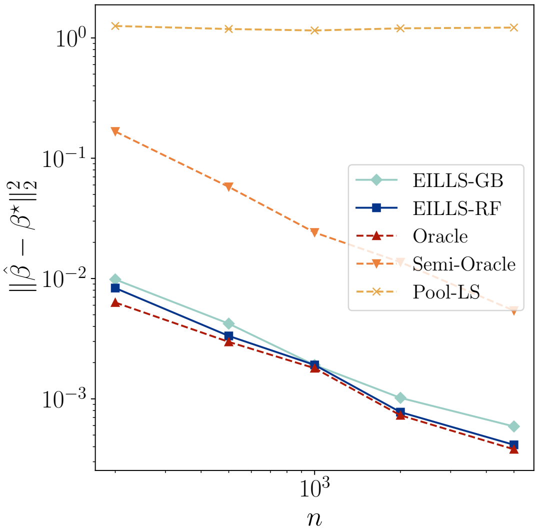


Figure 8: The simulation results for linear models with $p = 70$*. It depicts how the median estimation errors (based on* 50 *replications, shown in log scale) for different estimators (marked with different shapes and colors) change when* n *varies in* $\{200, 500, 1000, 2000, 5000\}$ *respectively.*

Implementation. For the EILLS estimator realized by gradient descent with Gumbel trick, we run gradient descent ascent using Adam optimizer with a learning rate of 1e-3, batch size 64 for $50k$ iterations. We adopt a fixed hyper-parameter $\gamma = 10$ and report the performance of the following estimators using the median of the estimation error $\|\widehat{\beta} - \beta^*\|_2^2$ over 50 replications and varying $n \in \{200, 500, 1000, 2000, 5000\}$.

(1) Pool-LS: it simply runs least squares on the full covariate X using all the data.

(2) EILLS-GB: Our EILLS estimator with Gumbel approximation that outputs $\beta \odot \sigma(w)$.

(3) EILLS-RF: it selects the variables x_j with $\sigma(w_j) > 0.7$ of the fitted model in (2), i.e., $\widehat{S} = \{j : \sigma(w_j) > 0.7\}$, and refits least squares again on $X_{\widehat{S}}$ using all the data.

(4) Oracle: it runs least squares on X_{S^*} using all the data.

(5) Semi-Oracle: it runs least squares on X_{G^c} using all the data, where G is the set of all the descendants of Y. Compared with the ERM, it manually removes all the variables that will lead to a biased estimation, but it will also keep uncorrelated variables compared with the full Oracle estimation.

Fig. 7 (b) visualizes how the Gumbel gate values for different covariables $\sigma(w)$ evolve during training in one trial. We can see that $\sigma(w_j)$ for $j \in S^*$ quickly increases and dominates the values for other variables like children/offspring of Y during the whole training process.

Results. The results are shown in Fig. 8. We can see that the square of the ℓ_2 estimation error $\|\widehat{\beta} - \beta^*\|_2^2$ for the pooled least squares estimator (×) does not decrease and remains to be very large (≈ 1.5) as n increases, indicating that it converges to a biased solution. At the same time, the estimation error for EILLS-GB/EILLS-RF (♦/■) decays as n grows and lies in between that for least squares on X_{G^c} (Semi-Oracle ▼) and least squares on X_{S^*} (Oracle ▲).

D.7 Choice of γ in Practice

We argue that the EILLS estimator is not very sensitive to the choice of γ when the true causal signal is strong enough, thus one can, for example, adopt $\gamma = 36$, or $\gamma = 100$. Such insensitivity is attributed to the discontinuity nature of the loss discussed in Appendix D.8. For example, in the simulation considered in this paper where there exists weak signal $\beta_3 = -0.5$, EILLS discards correct variables only when γ is relatively large $\gamma > 3 \times 10^3$ and it works pretty well in a wide range of $\gamma \in [15, 3 \times 10^3]$ as shown in Fig. 3 (a).

When the target is to produce the best predictions on unseen future data, we can choose a proper γ from a set of candidate γ-values that has the best out-of-sample performance in certain validation environment(s). To be specific, we consider the two cases, for the first case, we have some training environments $\mathcal{E}_{train}$ and also some validation environments $\mathcal{E}_{valid}$ whose associations between y and $\boldsymbol{x}$ are similar to those to be confronted in the future by our belief. We can adopt the following procedure in this case:

Step 1 Set candidate hyper-parameter set Γ.

Step 2 For each $\gamma \in \Gamma$, run EILLS estimator using data in $\mathcal{E}_{train}$ with γ and get the estimated $\widehat{\boldsymbol{\beta}}^{\gamma}$, calculate the worse-case out-of-sample empirical L_2 risk among the validation set as

$$\widehat{R}_{\gamma} = \max_{e \in \mathcal{E}_{valid}} \frac{1}{n^{(e)}} \sum_{i=1}^{n^{(e)}} \left\{ y_i^{(e)} - (\widehat{\boldsymbol{\beta}}^{\gamma})^\top \boldsymbol{x}_i^{(e)} \right\}^2 .$$

Step 3 We choose γ from Γ that corresponds to the minimum out-of-sample empirical L_2 risk, that is, $\widehat{\gamma} = \operatorname{argmin}_{\gamma \in \Gamma} \widehat{R}_{\gamma}$.

For the second case, suppose we only have $\mathcal{E}_{train}$, we can adopt the following leave-one-out procedure.

Step 1 Set candidate hyper-parameter set Γ.

Step 2 For fixed $\gamma \in \Gamma$, enumerate all the $e \in \mathcal{E}_{train}$. For each $e \in \mathcal{E}_{train}$, run EILLS estimator using data in $\mathcal{E}_{train} \setminus \{e\}$ with γ and get the estimated $\widehat{\boldsymbol{\beta}}^{\gamma,e}$, calculate its out-of-sample empirical L_2 risk in e as

$$\widehat{R}_{\gamma,e} = \frac{1}{n^{(e)}} \sum_{i=1}^{n^{(e)}} \left\{ y_i^{(e)} - (\widehat{\boldsymbol{\beta}}^{\gamma,e})^\top \boldsymbol{x}_i^{(e)} \right\}^2 ,$$

and get its maximum value $\widehat{R}_{\gamma} = \max_{e \in \mathcal{E}_{train}} \widehat{R}_{\gamma,e}$.

Step 3 We choose γ from Γ that corresponds to the minimum out-of-sample empirical L_2 risk, that is, $\widehat{\gamma} = \operatorname{argmin}_{\gamma \in \Gamma} \widehat{R}_{\gamma}$.

D.8 The Interpretation of Small γ

When γ is not large enough, i.e., $\gamma < \gamma^*$, the EILLS objective is somewhat similar to running (penalized) pooled least squares on variables excluding some spurious variable whose spuriousness to heterogeneity ratio is smaller than γ. We use the following variant of the thought experiment to illustrate the idea. Suppose we still do the cow/camel classification using three features $x_1 =$ shape, $x_2 =$ backgrouond, and $x_3 =$ whether the object stands or not. In the two-environment training dataset, 95% camels/cows on sand/grass, and 95% camels/cows sit/stand in $\mathcal{D}_1$. Moreover, 90% camels/cows on sand/grass, and 70% camels/cows sit/stand in $\mathcal{D}_2$. Under mild conditions akin to Condition 4.5, the regularization path of the population-level minimizer of EILLS $\boldsymbol{\beta}_\gamma = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathsf{Q}(\boldsymbol{\beta}; \gamma, \boldsymbol{\omega})$ can be interpreted as follows. There are change points $\gamma_2 > \gamma_1 > 0$. When $\gamma \in [0, \gamma_1)$, $\boldsymbol{\beta}_\gamma$ will be similar to regressing Y on (x_1, x_2, x_3) using all the data; it will threshold X_3 and thus be similar to regressing y on (x_1, x_2) using all the data as $\gamma \in (\gamma_1, \gamma_2)$; it will finally recover the ground-truth causal parameter $\boldsymbol{\beta}^*$ with $\operatorname{supp}(\boldsymbol{\beta}^*) = \{1\}$ when $\gamma > \gamma_2$.