Collaboratively Learning Linear Models with Structured Missing Data

Chen Cheng*† chencheng@stanford.edu
Gary Cheng*‡ chenggar@stanford.edu
John Duchi‡† jduchi@stanford.edu

September 11, 2023

Abstract

We study the problem of collaboratively learning least squares estimates for m agents. Each agent observes a different subset of the features—e.g., containing data collected from sensors of varying resolution. Our goal is to determine how to coordinate the agents in order to produce the best estimator for each agent. We propose a distributed, semi-supervised algorithm Collab, consisting of three steps: local training, aggregation, and distribution. Our procedure does not require communicating the labeled data, making it communication efficient and useful in settings where the labeled data is inaccessible. Despite this handicap, our procedure is nearly asymptotically local minimax optimal—even among estimators allowed to communicate the labeled data such as imputation methods. We test our method on real and synthetic data.

1 Introduction

Consider a set of agents that collect data to make predictions, where different agents may collect different features—because of different sensor availability or specialization—but wish to leverage shared structure to achieve better accuracy. Concretely, suppose we have m agents, where each agent $i \in [m]$ observes n samples of (x_{i+}, y) where $x_{i+} \in \mathbb{R}^{d_i}$ is some subset of $x \in \mathbb{R}^d$. We set this as a regression problem where the data (x, y) has the linear relationship $y = \langle x, \theta \rangle + \xi$ for some noise variable ξ. For example, these agents could be a network of satellites, each collecting data with a distinct set of sensors of varying resolution and specialization, with the purpose of estimating quantities like crop-yields [22], biomass [19], and solar-flare intensity [12]. Or these agents could be a group of seismic sensors, using acoustic modalities or accelerometers to predict whether an earthquake will occur [2]. Other examples may include networks of hospitals or phones [13]. In these settings, the agents can share information to collaboratively train a model; however, they are limited by communication bandwidth constraints, a situation satellites and seismic sensors often face due to radio frequency spectrum scarcity and interference [6, 20]. Without being too rigorous, we will define a communication efficient algorithm as one with communication cost that is sublinear in n; this definition is suited for applications with significant data volume but limited communication resources. Can we construct a statistically optimal and communication efficient procedure to estimate θ?

We answer in the affirmative and introduce our estimator Collab. Collab consists of three steps: local training on all agents, aggregation on a coordinating server, and distribution back to all agents. Our algorithm is communication-efficient: each agent $i \in [m]$ syncs twice with a coordinating server and incurs communication cost scaling like $\Theta(d_i^2)$. We prove local minimax lower bounds which prove that Collab is (nearly) instance-optimal. We choose to study this problem in a stylized linear setting so that we can provide stronger guarantees for the algorithms we make. Indeed, our results which pair the exact asymptotic covariance of our estimator Collab with matching asymptotic local minimax lower

*Equal contribution, authors ordered alphabetically by last and first names
†Statistics Department, Stanford University
‡Electrical Engineering Department, Stanford University

bounds heavily rely on the linearity of our problem and would not be possible without strong structural assumptions. Having said this, the theory we develop for linear models does hint at potential methods for non-linear settings, which we discuss in Section 7. We also acknowledge privacy considerations are important for real world systems such as hospitals. We choose to focus instead on sensor settings where privacy is less of a concern. We leave adapting our results to privacy-sensitive settings to future work. We compare our methods to single-imputation methods theoretically and empirically. We choose to baseline against imputation methods for three reasons. First, if we ignore communication constraints, our problem is a missing data problem, where formally the data is "missing at random" (MAR) [16]. MAR problems are well studied, so we know that imputation methods work well theoretically and in practice [25, 15]. Second, because we have instance-optimal lower bounds, we know that imputation methods are also optimal for our problem. Finally, because imputation methods use more information than the method we propose, imputation will serve as a "oracle" baseline of sorts.

Contributions. We briefly summarize our contributions.

1. We design a communication-efficient, distributed learning algorithm Collab which performs a weighted de-biasing procedure on the ordinary least squares estimator of each agent's data.
2. We show Collab is asymptotically locally minimax optimal among estimators which have access to the ordinary least squares estimator of each agent's data. We also show that with some additional assumptions, Collab is also asymptotically locally minimax optimal among estimators that have access to *all* of the training data of all agents.
3. We propose and develop theory for various baseline methods based on imputation. We compare the statistical error and communication cost of Collab against these baseline methods both theoretically and empirically on real and synthetic data.
4. We discuss generalizations of Collab for non-Gaussian feature settings and non-linear settings. We highlight open problems and identify possible directions for future work.

1.1 Related Work

Missing data. If we ignore the communication and computational aspects of our problem, the problem we study reduces to one of estimation with missing data. There has been a lot of work on this topic; please see [17] for an overview. The data in our problem is missing at random (MAR)—the missing pattern does not depend on the value of the data and is known given agent i. There are many approaches to handling missing data such as weighting and model-based methods [24]. Most related to our work are methods on single imputation. Schafer and Schenker [25] shows imputation with conditional mean is nearly optimal with special corrections applied. More recently, Chandrasekher et al. [3] show that single imputation is minimax optimal in the high dimensional setting. Another closely related popular approach is multiple imputation [23, 1]. Previous work [27, 29] has shown that multiple imputation in low dimensional settings produces correct confidence intervals under a more general set of assumptions compared to single imputation settings. However, we choose to focus on single imputation methods for two reasons. First, we are interested in estimation error and not confidence intervals, and our lower bounds show that single imputation has optimal estimation error for our setting. Second, in our problem context, multiple imputation would require more rounds of communication and consequently higher communication cost. Other methods for missing data include weighting and model-based methods.

Distributed learning. Learning with communication constraints is a well studied practical problem. We provide a couple of examples. Suresh et al. [26] study how to perform mean estimation with communication constraints. Duchi et al. [8] develop communication-constrained minimax lower bounds. Distributed convex optimization methods like Hogwild [21] have also been well studied. However, the works mentioned all concern the no-missing-data regime. A more relevant subfield of distributed learning is federated learning. In federated learning, a central server coordinates a collection of client

devices to train a machine learning model. Training data is stored on client devices, and due to communication and privacy constraints, clients are not allowed to share their training data with each other or the central server [13]. In the no-missing-features regime, optimization algorithms for federated optimization are well studied. There is also more theoretical work, which focus on characterizing communication, statistical, and privacy tradeoffs, albeit for a more narrow set of problems such as mean and frequency estimation [4]. More related to the missing data regime we consider is cross-silo federated learning [13] or vertical federated learning [30]. In this paradigm, the datasets on client machines are not only partitioned by samples but also by features. Researchers have studied this problem in the context of trees [5], calculating covariance matrices [14], k-means clustering [28], support vector machines [31], and neural nets [18]. Most related to our work is Gascón et al. [9], Hardy et al. [11]; they study how to privately perform linear regression in a distributed manner. However, unlike our work, these works focus more on developing algorithms with privacy guarantees rather than statistical ones.

2 Mathematical model

We assume we have m agents that observes a subset of the dimensions of the input data $x \in \mathbb{R}^d$. Each agent i has a "view" permutation matrix $\Pi_i^\top := \begin{bmatrix}\Pi_{i+}^\top & \Pi_{i-}^\top\end{bmatrix} \in \mathbb{R}^{d\times d}$. $\Pi_{i+} \in \mathbb{R}^{d_i \times d}$ describes which feature dimensions the agent sees, and $\Pi_{i-} \in \mathbb{R}^{(d-d_i)\times d}$ describes the dimensions the agent does not see. For a feature, label pair (x, y), the i-th agent has data (x_{i+}, y) where $x_{i+} \coloneqq \Pi_{i+}x \in \mathbb{R}^{d_i}$. Each agent has n such observations (independent across agents) denoted as a matrix $X_{i+} \in R^{n\times d_i}$ and vector $y_i \in \mathbb{R}^n$. We let $X_{i-} \in \mathbb{R}^{n\times(d-d_i)}$ denote the unobserved dimensions of the input data x drawn for the i-th agent, and we let $X_i \in \mathbb{R}^{n\times d}$ denote the matrix of input data x drawn for the i-th agent, including the dimensions of x unobserved by the i-th agent. To simplify discussions in the following sections, for any vector $v \in \mathbb{R}^d$ we use the shorthand $v_{i+} = \Pi_{i+}v$ and $v_{i-} = \Pi_{i-}v$. Similarly for any matrix $A \in \mathbb{R}^{d\times d}$ we denote by

$$A_{i+} = \Pi_{i+}A\Pi_{i+}^\top, \qquad A_{i-} = \Pi_{i-}A\Pi_{i-}^\top,$$
$$A_{i\pm} = \Pi_{i+}A\Pi_{i-}^\top, \qquad A_{i\mp} = \Pi_{i-}A\Pi_{i+}^\top.$$

For a p.s.d. matrix A, we let $\|x\|_A = \langle x, Ax\rangle$.

We assume the data from the m agents follow the same linear model. The features vectors x comprising the data matrices $X_1, \ldots, X_m$ are i.i.d. with zero mean and covariance $\Sigma \succ 0$. We will assume that each agent has knowledge of Σ_{i+}—e.g., they have a lot of unlabeled data to use to estimate this quantity. The labels generated follow the linear model

$$y_i = X_i\theta + \xi_i, \qquad \xi_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2 I_n).$$

Throughout this work we consider a fixed ground truth parameter θ.

Objectives. We are interested in proposing a method of using the data of the agents to form an estimate $\hat{\theta}$ which minimizes the full-feature prediction error on a fresh sample $x \in \mathbb{R}^d$

$$\mathbb{E}_x[(\langle x, \hat{\theta}\rangle - \langle x, \theta\rangle)^2] = \|\hat{\theta} - \theta\|_\Sigma^2. \tag{1}$$

We are also interested in forming an estimate $\hat{\theta}_i$ which minimizes the missing-feature prediction error of a fresh sample $x_{i+} \in \mathbb{R}^{d_i}$ for agent i—i.e., $x_{i+} = \Pi_{i+}x$ where $x \in \mathbb{R}^d$ is fresh. Define $T_i \coloneqq \begin{bmatrix}I_{d_i} & \Sigma_{i+}^{-1}\Sigma_{i\pm}\end{bmatrix}\Pi_i$ and the Schur complement $\Gamma_{i-} \coloneqq \Sigma\backslash\Sigma_{i+} \coloneqq \Sigma_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1}\Sigma_{i\pm}$. The local test error is then

$$\mathbb{E}_x[(\langle x_{i+}, \hat{\theta}_i\rangle - \langle x, \theta\rangle)^2] = \|\hat{\theta}_i - T_i\theta\|_{\Sigma_{i+}}^2 + \|\theta_{i-}\|_{\Gamma_{i-}}^2 \tag{2}$$

Here, $\|\theta_{i-}\|_{\Gamma_{i-}}^2$ is irreducible error. The role of the operator T_i is significant as $T_i\theta$ is the best possible estimator for the ith agent[1]. Through this paper, we will also highlight the communication

[1]Maybe surprisingly, $T_i\theta$ is better than naively selecting the subset of θ corresponding to the features observed by agent i—i.e., $\Pi_i\theta$. This is because $T_i\theta$ leverages the correlations between features.

costs of the methods we consider. Recall that we would like our methods to have $o(n)$ communication cost.

3 Our approach

We begin by outlining an approach of solving this problem for general feature distributions. The general approach is not immediately usable because it requires some knowledge of θ, so we need to do some massaging. In Section 3.2, we show how to circumvent this issue in the Gaussian feature setting and introduce our method Collab. Adapting the general approach to other non-Gaussian settings is an open problem, but we discuss some potential approaches in Section 7.

3.1 General approach

Our solution begins with each agent i performing ordinary least squares on their local data

$$\hat{\theta}_i = X_{i+}^{\dagger} y_i \overset{\text{(i)}}{=} (X_{i+}^{\top} X_{i+})^{-1} X_{i+}^{\top} y_i,$$

where $A^{\dagger}$ denotes the Moore–Penrose inverse for a general matrix A, and (i) holds whenever $\mathsf{rank}(X_{i+}) \geq d_i$. Because we focus on the large sample asymptotics regime ($n \gg d_i$), (i) will hold with probability 1.

Then, we aggregate $\hat{\theta}$ using a form of weighted empirical risk minimization parameterized by the positive definite matrices $W_i \in \mathbb{R}^{d_i \times d_i}$

$$\hat{\theta} = \hat{\theta}(W_1, \cdots, W_m) \coloneqq \operatorname*{argmin}_{\theta} \sum_{i=1}^{m} \left\| \theta_{i+} + \Sigma_{i+}^{-1} \Sigma_{i\pm} \theta_{i-} - \hat{\theta}_i \right\|_{W_i}^2. \tag{3}$$

We know by first order stationarity that $\hat{\theta} = \left(\sum_{i=1}^m T_i^{\top} W_i T_i\right)^{-1} \left(\sum_{i=1}^m T_i^{\top} W_i \hat{\theta}_i\right)$. $\hat{\theta}$ is a consistent estimate of θ regardless the choice of weighting matrices W_i. Furthermore, if the features X_i, $\hat{\theta}$ are Gaussian, $\hat{\theta}$ is also unbiased. We show this result in the Appendix in Lemma B.1. While Lemma B.1 shows that the choice of weighting matrices W_i does not affect consistency, the choice of weighting matrices W_i does dictate the asymptotic convergence rate of the estimator. In the next theorem, we show what the best performing choice of weighting matrices are. The proof is in Appendix B.2.

Theorem 3.1. *For any weighting matrices W_i, the aggregated estimator $\hat{\theta} = \hat{\theta}(W_1, \cdots, W_m)$ is asymptotically normal*

$$\sqrt{n}\left(\hat{\theta} - \theta\right) = \mathsf{N}(0, C(W_1, \cdots, W_m)),$$

with some covariance matrix $C(W_1, \cdots, W_m)$. The optimal choice of weighting matrices is

$$W_i^{\star} \coloneqq \Sigma_{i+}(\mathbb{E}\left[x_{i+}\theta_{i-}^{\top} z_{i+} z_{i+}^{\top} \theta_{i-} x_{i+}^{\top}\right] + \sigma^2 \Sigma_{i+})^{-1} \Sigma_{i+},$$

where $z_{i+} = x_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+}$. In particular, for all W_i, $C(W_1, \cdots, W_m) \succeq C(W_1^{\star}, \cdots, W_m^{\star}) = \left(\sum_{i=1}^m T_i^{\top} W_i^{\star} T_i\right)^{-1}$.

The main challenge of using Theorem 3.1 is in constructing the optimal weights $W_i^{\star}$, as at face value, they depend on knowledge of θ. While we will discuss high level strategies of bypassing this issue in non-Gaussian data settings in Section 7, we will currently focus our attention on how we can make use of Gaussianity to construct our estimator Collab.

3.2 Collab Estimator - Gaussian feature setting

If X_i are distributed as $\mathsf{N}(0, \Sigma)$, $W_i^{\star}$ has an explicit closed form as

$$W_i^{\star} = W_i^{\mathrm{g}} \coloneqq \frac{\Sigma_{i+}}{\|\theta_{i-}\|_{\Gamma_{i-}}^2 + \sigma^2} = \frac{\Sigma_{i+}}{\mathbb{E}_{x,y}[(\langle x_{i+}, \hat{\theta}_i \rangle - y)^2]},$$

Algorithm 1: COLLAB algorithm

Data: m agents with training data $(X_{1+}, y_1), \ldots, (X_{m+}, y_m)$ each with n datapoints
for *Each agent* $i = 1, \ldots, m$ *in parallel* **do**
  Compute $\hat{\theta}_i = (X_{i+}^\top X_{i+})^{-1} X_{i+}^\top y_i$;
  Compute $\widehat{\Sigma}_i = \frac{1}{n} X_{i+}^\top X_{i+}$ or with additional unlabeled data;
  Compute $R_i = \frac{1}{n} \|X_{i+}\hat{\theta}_i - y\|_2^2$;
  Send $\hat{\theta}_i, \widehat{\Sigma}_i, R_i$ to central server;
end
Central server constructs $\hat{W}_i^{\mathrm{g}} \coloneqq \widehat{\Sigma}_{i+}/R_i$;
Central server computes $\hat{\theta}_i^{\mathrm{clb}} = T_i \hat{\theta}(\hat{W}_1^{\mathrm{g}}, \cdots, \hat{W}_m^{\mathrm{g}})$ and distributes them to respective agents;

where $\Gamma_{i-} = \Sigma_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1}\Sigma_{i\pm}$ is the Schur complement. Recall we assume that each agent has enough unlabeled data to estimate Σ_{i+}. Furthermore, $\frac{1}{n}\|X_{i+}\hat{\theta}_i - y\|_2^2$ is a consistent estimator of $\mathbb{E}_{x,y}[(\langle x_{i+}, \hat{\theta}_i\rangle - y)^2]$. Thus, each agent is able to construct estimates of W_i^{g} by computing

$$\hat{W}_i^{\mathrm{g}} \coloneqq \frac{\Sigma_{i+}}{\frac{1}{n}\|X_{i+}\hat{\theta}_i - y\|_2^2}$$

Now we construct our global and local COLLAB estimators defined respectively as

$$\hat{\theta}^{\mathrm{clb}} \coloneqq \hat{\theta}(\hat{W}_1^{\mathrm{g}}, \cdots, \hat{W}_m^{\mathrm{g}}), \qquad \hat{\theta}_i^{\mathrm{clb}} \coloneqq T_i \hat{\theta}(\hat{W}_1^{\mathrm{g}}, \cdots, \hat{W}_m^{\mathrm{g}}). \tag{4}$$

We summarize the COLLAB algorithm in Algorithm 1. At a high level, $\hat{\theta}^{\mathrm{clb}}$ is an estimate of θ which also minimizes the full-feature prediction error (1) and $\hat{\theta}_i^{\mathrm{clb}}$ minimizes the missing-feature prediction error for agent i (2). Now we show that using the collective "biased wisdom" of local estimates $\hat{\theta}_i$, our collaborative learning approach returns an improved local estimator. The proof is in Appendix B.3.

Corollary 3.2. *Let* $X_i \sim \mathsf{N}(0, \Sigma)$ *and define* $C^{\mathrm{g}} \coloneqq (\sum_{i=1}^m T_i^\top W_i^{\mathrm{g}} T_i)^{-1}$. *The global* COLLAB *estimator* $\hat{\theta}_i^{\mathrm{clb}}$ *and the local* $\hat{\theta}_i^{\mathrm{clb}}$ *on agent* i *are asymptotically normal*

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{clb}} - \theta\right) \xrightarrow{d} \mathsf{N}\left(0, C^{\mathrm{g}}\right) \quad \text{and} \quad \sqrt{n}\left(\hat{\theta}_i^{\mathrm{clb}} - T_i\theta\right) \xrightarrow{d} \mathsf{N}\left(0, T_i C^{\mathrm{g}} T_i^\top\right).$$

The following are true

(i) W_i^{g} *are the optimal choice of weighting matrices i.e.,particular,* $C(W_1, \cdots, W_m) \succeq C(W_1^{\mathrm{g}}, \cdots, W_m^{\mathrm{g}}) = C^{\mathrm{g}}$.

(ii) On agent i, *we have* $\sqrt{n}(\hat{\theta}_i - T_i\theta) \xrightarrow{d} \mathsf{N}(0, (W_i^{\mathrm{g}})^{-1})$. *The asymptotic variance of* $\hat{\theta}_i$ *is larger than that of the* COLLAB *estimator* $\hat{\theta}_i^{\mathrm{clb}}$—*i.e.,* $(W_i^{\mathrm{g}})^{-1} \succeq T_i C^{\mathrm{g}} T_i^\top$.

4 Comparison with other methods

In this section, we compare our collaborative learning procedure with other popular least squares techniques based on imputation and comment on the statistical efficacy and communication cost differences. We summarize our analysis in Table 1. The proofs of the theorems are in Appendix C.

Local imputation w/ collaboration. Suppose a coordinating server collected covariance information Σ_i from each agent and then distributed Σ back to each of them. Then one intuitive strategy is to use this information to impute each agent's local data by replacing X_{i+} with $\mathbb{E}[X_i \mid X_{i+}] = X_{i+}T_i$, before performing local linear regression. In other words, instead of computing $\hat{\theta}_i$, compute

$$\hat{\theta}_i^{\mathrm{imp}} = (T_i^\top X_{i+}^\top X_{i+} T_i)^\dagger T_i^\top X_{i+}^\top y_i$$

Method	Full-feature asymptotic covariance	Missing-feature asymptotic covariance	Communication cost for agent i
Local OLS - $\hat{\theta}_i$	-	$(W_i^{\mathrm{g}})^{-1}$	0
Local imputation w/ collaboration - $\hat{\theta}_i^{\mathrm{imp}}$	$\left(\sum_{i=1}^m T_i^\top W_i^{\mathrm{g}} T_i\right)^{-1}$	$T_i \left(\sum_{i=1}^m T_i^\top W_i^{\mathrm{g}} T_i\right)^{-1} T_i^\top$	$\Theta(d^2)$
Global imputation - $\hat{\theta}_i^{\text{imp-glb}}$	$\left(\sum_{i=1}^m T_i^\top W_i^{\mathrm{g}} T_i\right)^{-1}$	$T_i \left(\sum_{i=1}^m T_i^\top W_i^{\mathrm{g}} T_i\right)^{-1} T_i^\top$	$\Theta(nd_i)$
Collab - $\hat{\theta}_i^{\mathrm{clb}}$	$\left(\sum_{i=1}^m T_i^\top W_i^{\mathrm{g}} T_i\right)^{-1}$	$T_i \left(\sum_{i=1}^m T_i^\top W_i^{\mathrm{g}} T_i\right)^{-1} T_i^\top$	$\Theta(d_i^2)$

Table 1. Full and Missing feature asymptotic covariance and communication cost for agent i. Communication cost is measured by how many real numbers are received and sent from agent i.

to send back to the coordinating server. Note that we use Moore–Penrose inverse here as $T_i^\top X_{i+}^\top X_{i+} T_i$ is in general of rank d_i, and $\hat{\theta}_i^{\mathrm{imp}}$ is then the min-norm interpolant of agent i's data. Similar to Collab, we can use weighted empirical risk minimization parameterized by $W_i \in \mathbb{R}^{d\times d}$ and to aggregate $\hat{\theta}^{\mathrm{imp}}$ via

$$\hat{\theta}^{\mathrm{imp}} = \hat{\theta}(W_1, \cdots, W_m) \coloneqq \operatorname*{argmin}_{\theta} \sum_{i=1}^m \left\| T_i^\top (T_i T_i^\top)^{-1} T_i \theta - \hat{\theta}_i^{\mathrm{imp}} \right\|_{W_i}^2 .$$

The next theorem, in conjunction with Theorem 3.1, implies that under the WERM optimization scheme, aggregation of least squares estimators on imputed local data does not bring additional statistical benefit. In fact, the local imputation estimator is a linearly transformed on local OLS $\hat{\theta}_i$.

Theorem 4.1. *For $\hat{\theta}_i^{\mathrm{imp}}$ from agent i, we have $\hat{\theta}_i^{\mathrm{imp}} = T_i^\top (T_i T_i^\top)^{-1} \hat{\theta}_i$. Given any weighting matrices $W_i \in \mathbb{R}^{d\times d}$, the aggregated imputation estimator $\hat{\theta}^{\mathrm{imp}}$ is consistent and asymptotically normal*

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{imp}} - \theta\right) = \mathsf{N}(0, C^{\mathrm{imp}}(W_1, \cdots, W_m)).$$

Using the same weights $W_i^\star \in \mathbb{R}^{d_i \times d_i}$ as in Theorem 3.1 for aggregated $\hat{\theta}^{\mathrm{imp}}$, we have under p.s.d. cone order, for weights W_i, $C^{\mathrm{imp}}(W_1, \cdots, W_m) \succeq C^\star$, where $C^\star = (\sum_{i=1}^m T_i^\top W_i^\star T_i)^{-1}$. In addition, the equality holds when $W_i = T_i^\top W_i^\star T_i$.

As we will see in Sec. 5 where we provide minimax lower bound for weak observation models, the fact that the weighted imputation does not outperform our Collab approach is because the WERM on local OLS without imputation is already optimal. In fact, having access to the features will not achieve better estimation rate for both the global parameter θ and local parameters $T_i\theta$.

In terms of communication cost, this local imputation method requires more communication than Collab, as a central server needs to communicate Σ to all the hospitals. This amounts to a total of $\Theta(md^2)$ communication cost instead of $\Theta(\sum_{i\in[m]} d_i^2)$ communication cost for Collab.

Global imputation. Finally, we analyze the setting where we allow each agent to send the central server all of their data (X_{i+}, y_i) for $i = 1, \cdots, m$ instead of their local estimators, $\hat{\theta}_i$ or $\hat{\theta}_i^{\mathrm{imp}}$. Having all the data with structured missingness available, a natural idea is to first impute the data, replacing X_{i+} with $\mathbb{E}[X_i \mid X_{i+}] = X_{i+} T_i$, and then performing weighted OLS on *all* of the nm data points. Namely for scalars $\alpha_1, \cdots, \alpha_m > 0$, we take

$$\hat{\theta}^{\text{imp-glb}} = \hat{\theta}^{\text{imp-glb}}(\alpha_1, \cdots, \alpha_m) := \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top X_{i+} T_i\right)^{-1} \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top y_i\right).$$

Surprisingly, in spite of the additional power, $\hat{\theta}^{\text{imp-glb}}$ still does not beat $\hat{\theta}$ in Theorem 3.1.

Theorem 4.2. *For any scalars $\alpha_1, \cdots, \alpha_m > 0$, $\hat{\theta}^{\text{imp-glb}}$ is consistent and asymptotically normal*

$$\sqrt{n}\left(\hat{\theta}^{\text{imp-glb}} - \theta\right) = \mathsf{N}(0, C^{\text{imp-glb}}(\alpha_1, \cdots, \alpha_m)).$$

Recall the lower bound matrix $C^\star := (\sum_{i=1}^m T_i^\top W_i^\star T_i)^{-1}$ in Theorem 3.1. If $X_i \sim \mathsf{N}(0, \Sigma)$, we have under p.s.d. cone order and any $\alpha_i > 0$, $C^{\text{imp-glb}}(\alpha_1, \cdots, \alpha_m) \succeq C^\star$. In addition, the equality holds when $\alpha_i = 1/(\|\theta_{i-}\|^2_{\Gamma_{i-}} + \sigma^2)$.

The communication cost for this method is significantly larger. Having each agent send all of its data to a coordinating server requires $\Theta(\sum_{i\in[m]} d_i n)$ communication cost, as opposed to the $\Theta(\sum_{i\in[m]} d_i^2)$ communication cost for Collab. The fact that communication cost for this method scales with n is a significant disadvantage for the reasons we outlined in the introduction.

5 Asymptotic Local Minimax Lower Bounds

In this section, we prove asymptotic local minimax lower bounds that show Collab is (nearly) optimal. We work in the partially-fixed-design regime. For every sample $x \in \mathbb{R}^d$, $x_{i+} \in \mathbb{R}^{d_i}$ is a fixed vector. We draw x_{i-} from $\mathsf{N}(\mu_{i-}, \Gamma_{i-})$ where μ_{i-} and Γ_{i-} is the conditional mean and variance of x_{i-} given x_{i+}. Here Γ_{i-} is also the Schur complement. We draw x_{i-} from $\mathsf{N}(\mu_{i-}, \Gamma_{i-})$. The samples $x_{i+} \in \mathbb{R}^{d_i}$ comprise the matrices $X_{i+} \in \mathbb{R}^{n\times d_i}$. For all $i \in [m]$, we will assume we have an infinite sequence (w.r.t. n) of matrices X_{i+}. This partially-fixed-design scheme gives the estimators knowledge of the observed features and the distribution of the unobserved features, which is consistent with knowledge that Collab has access to. In this section we fix $\theta \in \mathbb{R}^d$. The corresponding label $y = x_{i+}\theta_{i+} + x_{i-}\theta_{i-} + \xi$, where $\xi \in \mathbb{R}$ is drawn from i.i.d. $\mathsf{N}(0, \sigma^2)$. We use $y_j \in \mathbb{R}^n$ to denote its vector form for the agent j. To model the estimator's knowledge about the labels, we will have two observation models—one weaker and one stronger—which we will specify later when we present our results.

For each observation model, we will have two types of results. The first type of result is a minimax lower bound for full-featured data; i.e., how well can estimator perform on a fresh sample without missing features. This type of result will concern the full-feature asymptotic local minimax risk

$$\liminf_{n\to\infty} \mathfrak{M}_{m,\varepsilon}(\{X_{i+}\}_{i\in[m]}; \mathcal{P}_n, u) := \liminf_{n\to\infty} \inf_{\bar{\theta}} \sup_{P\in\mathcal{P}_n} n\mathbb{E}_{Z\sim P}\langle u, \bar{\theta}(Z, \{X_{i+}\}_{i\in[m]}) - \theta\rangle^2.$$

We will show that there exists a $B \in \mathbb{R}^{d\times d}$ such that the local minimax risk in the previous display is lower bounded by $u^T B u$ for all $u \in \mathbb{R}^d$. In other words, we have lower bounded the asymptotic covariance of our estimator with B (with respect to the p.s.d. cone order). The second type of result is an agent specific minimax lower bound; i.e., what is the best prediction error an estimator (for the given observation model) can possibly have on a fresh sample for a given agent. This type of result will deal with the missing-feature asymptotic local minimax risk, defined as

$$\liminf_{n\to\infty} \mathfrak{M}^{i+}_{m,\varepsilon}(\{X_{i+}\}_{i\in[m]}; \mathcal{P}_n, u) := \liminf_{n\to\infty} \inf_{\bar{\theta}} \sup_{P\in\mathcal{P}_n} n\mathbb{E}_{Z\sim P}\langle u, \bar{\theta}(Z, \{X_{i+}\}_{i\in[m]}) - T_i\theta\rangle^2.$$

Similar to the first minimax error definition, we will again show that there exists a $B_i \in \mathbb{R}^{d_i\times d_i}$ such that the local minimax risk we just defined is lower bounded by $u^T B_i u$ for all $u \in \mathbb{R}^{d_i}$. Recall (2) for discussion surrounding why $T_i\theta$ is the right object to compare against.

5.1 Weak Observation Model: Access only to local models and features

Recall the local least squares estimator $\hat{\theta}_i = (X_{i+}^\top X_{i+})^{-1} X_{i+}^\top y_i$. Let $P_\theta^{\hat{\theta}}$ be a distribution over $\hat{\theta}_1, \ldots, \hat{\theta}_m$ induced by θ and $(\xi_1, \ldots, \xi_m) \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2 I_n)$. We define the following family of distributions $\mathcal{P}^{\hat{\theta}}_{n,c} := \{P^{\hat{\theta}}_{\theta'} : \|\theta' - \theta\|_2 \le cn^{-1/2}\}$ which defines our observation model. Intuitively, in this observation model, we are constructing a lower bound for estimators which have access to the features $X_{1+}, \ldots, X_{m+}$, the population covariance Σ, and access to $\hat{\theta}_1, \ldots, \hat{\theta}_m$. In comparison, our estimator Collab only uses Σ and $\hat{\theta}_1, \ldots \hat{\theta}_m$. We present our first asymptotic local minimax lower bound result here. The proof of this result can be found in Appendix D.1.

Theorem 5.1. *Recall that* $C^{\mathrm{g}} := (\sum_{i=1}^{m} T_i^\top W_i^{\mathrm{g}} T_i)^{-1}$. *For all* $\in [m]$ *and* n *let the rows of* X_{i+} *be drawn i.i.d. from* $\mathsf{N}(0, \Sigma_{i+})$. *Then for all* $u \in \mathbb{R}^d$, *with probability 1, the full-feature asymptotic local minimax risk for* $\mathcal{P}_{n,c}^{\hat{\theta}}$ *is bounded below as,*

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \mathfrak{M}_{m,\varepsilon}(\{X_{i+}\}_{i\in[m]}; \mathcal{P}_{n,c}^{\hat{\theta}}, u) \geq u^\top C^{\mathrm{g}} u.$$

For all $u \in \mathbb{R}^{d_i}$, *with probability 1, the missing-feature asymptotic local minimax risk for* $\mathcal{P}_{n,c}^{\hat{\theta}}$ *is bounded below as*

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \mathfrak{M}_{m,\varepsilon}^{i+}(\{X_{i+}\}_{i\in[m]}; \mathcal{P}_{n,c}^{\hat{\theta}}, u) \geq u^\top T_i C^{\mathrm{g}} T_i^\top u.$$

This exactly matches the upper bound for COLLAB we presented in Corollary 3.2.

5.2 Strong Observation Model: Access to features and labels

Define the family of distributions $\mathcal{P}_{n,c}^{y} := \{P_{\theta'}^{y} : \|\theta' - \theta\|_2 \leq cn^{-1/2}\}$ as the observation model. Intuitively, in this model, we are constructing a lower bound for estimators having access to all of the features $X_{1+}, \ldots, X_{m+}$ and access to $y_1, \ldots y_m$. This observation model is stronger than the previous observation model because estimators now have access to the labels y. We note again that our estimator COLLAB only uses Σ and $\hat{\theta}_1, \ldots \hat{\theta}_m$. The quantities our estimator rely on do not scale with n, making our estimator much weaker than other potential estimators in this observation model, as estimators are allowed to depend on y_i, which grows in size with n. We present our second asymptotic local minimax lower bound result here, starting with defining the strong local lower bound matrix $C^{\mathrm{s}} := (\sum_{i=1}^{m} 2\Sigma/(\|\theta_{i-}\|_{\Gamma_{i-}}^2 + \sigma^2))^{-1}$. The proof of this result is in Appendix D.2.

Theorem 5.2. *For all* $i \in [m]$ *and* n *let the rows of* X_{i+} *be drawn i.i.d. from* $\mathsf{N}(0, \Sigma_{i+})$. *Then for all* $u \in \mathbb{R}^d$, *with probability 1, the full-feature asymptotic local minimax risk for* $\mathcal{P}_{n,c}^{y}$ *is bounded below as*

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \mathfrak{M}_{m,\varepsilon}(\{X_{i+}\}_{i\in[m]}; \mathcal{P}_{n,c}^{y}, u) \geq u^\top C^{\mathrm{s}} u.$$

For all $u \in \mathbb{R}^{d_i}$, *with probability 1, the missing-feature asymptotic local minimax risk for* $\mathcal{P}_{n,c}^{y}$ *is bounded below as*

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \mathfrak{M}_{m,\varepsilon}^{i+}(\{X_{i+}\}_{i\in[m]}; \mathcal{P}_{n,c}^{y}, u) \geq u^\top T_i C^{\mathrm{s}} T_i^\top u.$$

In view of the lower bound in the strong observation model and that of the weak observation model in Theorem 5.1, it is clear that the lower bound in the strong observation setting is in general smaller as

$$\Sigma - T_i^\top \Sigma_{i+} T_i = \Pi_i^\top \begin{bmatrix} 0 & 0 \\ 0 & \Gamma_{i-} \end{bmatrix} \Pi_i \succeq 0,$$

which further implies $C^{\mathrm{g}} \succeq (\sum_{i=1}^{m} \Sigma/(\|\theta_{i-}\|_{\Gamma_{i-}}^2 + \sigma^2))^{-1} \succeq C^{\mathrm{s}}$.

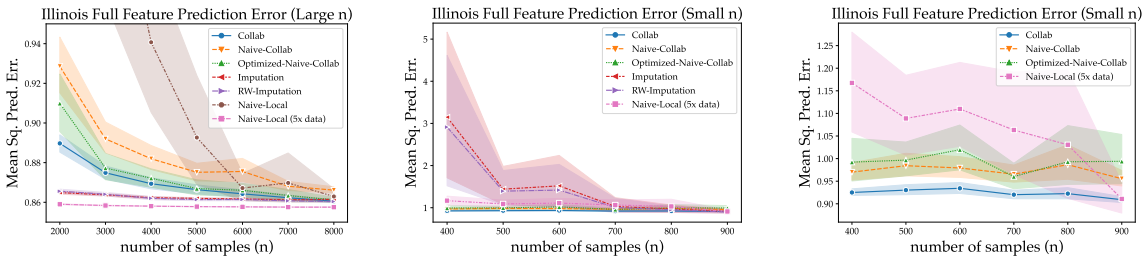
We argue that the two lower bounds are comparable in the missing completely at random [17]. Consider for every agent i, each coordinate is missing independently with probability p. In this case, (d_i, Σ_{i+}, T_i) are i.i.d. random triplets parameterized by p.

Corollary 5.3. *Under the random missingness setup with missing probability* p, *let the eigenvalue of* Σ *be* $\lambda_1(\Sigma) \geq \cdots \geq \lambda_d(\Sigma) > 0$ *and define its condition number* $\kappa = \lambda_1(\Sigma)/\lambda_d(\Sigma)$. *Suppose* $p \leq \frac{1}{2}\kappa^{-1}(1 + \|\theta\|_\Sigma^2/\sigma^2)^{-1}$, *we have the limits* $\lim_{m\to\infty} mC^{\mathrm{g}}$ *and* $\lim_{m\to\infty} mC^{\mathrm{s}}$ *exist and*

$$4 \lim_{m\to\infty} mC^{\mathrm{s}} \succeq \lim_{m\to\infty} mC^{\mathrm{g}} \succeq \lim_{m\to\infty} mC^{\mathrm{s}}.$$

6 Experiments

We perform experiments to empirically test and compare the methods we have discussed in this paper. Our first experiment is on real data with potential distribution shift between agents and models a potentially real setting concerning the US Census. This experiment is meant to show how our methods would perform in practice. The setup of the synthetic experiment is similar to the setup of our theory; due to space, we defer this to Appendix A.2.



(a) Prediction Error for large n (b) Prediction Error for small n (c) Small n omitting Impute algs.

Figure 1: Experimental results for US Census Experiment

6.1 US Census Experiments

We experiment on real US census data modified from the ACSTravelTime dataset from the `folktables` package [7] to test how our methods work on real data, which may contain covariate shift across agents. After dataset preprocessing, described in the Appendix A.1, we have $d = 37$ features. We plot the covariance matrix of the features in Appendix A.1. We compute the covariance from training data across all of the datacenters. We assume we are able to do this because this computation can be done in a distributed manner, without communicating training data points or labels.

We (artificially) construct $m = 5$ datacenters (agents), each containing data from one of California, New York, Texas, Florida, and Illinois. The goal is to collaboratively learn a model for each datacenter in a communication efficient way. This setup models potentially real settings where state governments are interested in similar prediction tasks but may not be allowed to directly transfer data about their constituents directly to one another due to privacy or communication constraints. The California datacenter will have access to 37 features, New York to 36, Texas to 35, Florida to 30, and Illinois to 27. This models the feature heterogeneity which varies across geography. Each datacenter will have n datapoints, which we vary in this experiment. The objective to predict people from Illinois's travel time to work given all 37 features. This task models the setting where the datacenter of interest does not have access to labeled full-featured, data to use to predict on full-featured test data.

We compare our method Collab against methods we call Naive-Local, Naive-Collab, Optimized-Naive-Colllab, Imputation, and RW-Imputation. We briefly describe each method here; Appendix A.1 contains a more detailed description of each method. Naive-Local refers to each agent locally perform OLS to construct $\hat{\theta}_i$. Naive-Collab does an equal-weighted average of the agent OLS models—$\sum_{i\in[m]} \Pi_{i+}^{\top}\hat{\theta}_i/m$. Optimized-Naive-Collab uses gradient descent to optimize the choices of weights of Naive-Collab. Optimized-Naive-Collab uses fresh labeled samples without any missing features during gradient descent, so in this sense, Optimized-Naive-Collab is more powerful than our method. Imputation refers to the global imputation estimator $\hat{\theta}^{\text{imp-glb}}$ with $\alpha_i = 1/m$. RW-Imputation is Imputation but with the optimal choice of weights α_i. We also compare against Naive-Local trained with $5n$ datapoints. We choose $5n$ to model the hypothetical scenario setting where all of the other datacenters available contain data (albeit with missing features) from Illinois. For each method that we test, we run 80 trials to form 95% confidence intervals. We see that for $n \leq 800$ in Figures 1(b) and 1(c), Collab performs the best; the imputation methods do the worst, and have much higher variance. In this small n regime, even the Naive-Local method with 5 times the data does worse than Collab. For $n \geq 2000$ in Figure 1(a), the aggregation methods do worse than the imputation methods, and Naive-Local method with 5 times the data is the best performing method. However, Collab remains better than Optimized-Naive-Collab and Naive-Collab. The fact that the performance of the Naive-Collab approaches in much closer to the performance of Collab than in the Synthetic experiment in Appendix A.2 is not surprising, as the covariance of the features is much more isotropic, meaning that the naive aggregation methods will not incur nearly as much bias.

7 Discussion and Future Work

Optimal weights beyond Gaussianity. $\mathbb{E}\left[x_{i+}\theta_{i-}^\top z_{i+} z_{i+}^\top \theta_{i-} x_{i+}^\top\right]$ has a nice closed form in Gaussian setting because z_{i+} and x_{i+} are independent—which is in general not true without Gaussianity. If we can directly sample from the feature distribution $\mathcal{P}$ (e.g., unlabeled data), then we can empirically estimate $\mathbb{E}\left[x_{i+}\theta_{i-}^\top z_{i+} z_{i+}^\top \theta_{i-} x_{i+}^\top\right]$ by sampling from $\mathcal{P}$ and using any consistent plug-in estimate $\hat{\theta}$ (e.g., run Collab with weights $W_i = I_{d_i}$). This will return a good estimate of the optimal weights. An interesting future direction is to prove lower bounds without the Gaussianity assumption.

Generalization to non-linear models. Recall in the Gaussian setting, the optimal weights in Collab are $W_i^{\mathrm{g}} = \Sigma_{i+}/(\mathbb{E}_{x,y}[(\langle x_{i+}, \hat{\theta}_i\rangle - y)^2])$. Then, the optimal loss function in Eq. (3) becomes

$$\sum_{i=1}^m \left\| \theta_{i+} + \Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} - \hat{\theta}_i \right\|_{W_i^{\mathrm{g}}}^2 = \sum_{i=1}^m \frac{\mathbb{E}_{x_{i+}}[(\langle x_{i+}, \hat{\theta}_i\rangle - \langle x_{i+}, T_i\theta\rangle)^2]}{\mathbb{E}_{x,y}[(\langle x_{i+}, \hat{\theta}_i\rangle - y)^2]}.$$

This hints at a generalization to non-linear models. Suppose the local agents train on models $f^i(x_{i+};\theta_i), \mathbb{R}^{d_i}\times\mathbb{R}^{d_i} \mapsto \mathcal{Y}$ and the global model $f(x;\theta), \mathbb{R}^d\times\mathbb{R}^d \mapsto \mathcal{Y}$ satisfies for some mapping $T_i : \mathbb{R}^d \to \mathbb{R}^{d_i}$, $f(x; T_i\theta) = f^i(x_{i+};\theta_i)$. Consider a loss function $\ell(\cdot,\cdot) : \mathcal{Y}\times\mathcal{Y} \to [0,\infty)$. Then we can consider the following way of aggregation inspired by Collab for linear models

$$\hat{\theta} := \operatorname*{argmin}_{\theta} \sum_{i=1}^m \frac{\mathbb{E}_{x_{i+}}\ell(f^i(x_{i+};\hat{\theta}_i), f(x_{i+};T_i\theta))}{\mathbb{E}_{x_{i+},y}\ell(f^i(x_{i+};\hat{\theta}_i), y)}.$$

We can consistently estimate the denominators (weights) using training time loss. An interesting future direction is to investigate the performance of this general approach for non-linear problems.

References

[1] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20, 2011.

[2] David Chas Bolton, Parisa Shokouhi, Bertrand Philippe Gerard Rouet-Leduc, Claudia Hulbert, Jacques Rivière, Chris Marone, and Paul Allan Johnson. Characterizing acoustic signals and searching for precursors during the laboratory seismic cycle using unsupervised machine learning. *Seismological Research Letters*, 2019.

[3] Kabir Aladin Chandrasekher, Ahmed El Alaoui, and Andrea Montanari. Imputation for high-dimensional linear regression. *arXiv preprint arXiv:2001.09180*, 2020.

[4] Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Breaking the communication-privacy-accuracy trilemma. *IEEE Transactions on Information Theory*, 69:1261–1281, 2020.

[5] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, and Qiang Yang. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, 36:87–98, 2019.

[6] Giacomo Curzi, Dario Modenini, and Paolo Tortora. Large constellations of small satellites: A survey of near future challenges and missions. *Aerospace*, 2020.

[7] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Neural Information Processing Systems*, 2021.

[8] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Yuchen Zhang. Optimality guarantees for distributed statistical estimation. *arXiv: Information Theory*, 2014.

[9] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. Privacy-preserving distributed linear regression on high-dimensional data. *Proceedings on Privacy Enhancing Technologies*, 2017:345 – 364, 2017.

[10] Elisabeth Gassiat. Revisiting the van Trees inequality in the spirit of Hájek and Le Cam. 2014.

[11] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *ArXiv*, abs/1711.10677, 2017.

[12] Zhenbang Jiao, Hu Sun, Xiantong Wang, Ward B. Manchester, Tamas I. I. Gombosi, Alfred O. Hero, and Yang Chen. Solar flare intensity prediction with machine learning models. *Space Weather*, 18, 2019.

[13] Peter Kairouz, H. B. McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary B. Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim Y. El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Oluwasanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, R. Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Xiaodong Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14:1–210, 2019.

[14] Alan F. Karr, Xiaodong Lin, Ashish P. Sanil, and Jerome P. Reiter. Privacy-preserving analysis of vertically partitioned data using secure matrix products. *Journal of Official Statistics*, 25:125–138, 2009.

[15] Gayaneh Kyureghian, Oral Capps, and Rodolfo M. Nayga. A missing variable imputation methodology with an empirical application. 2011.

[16] Roderick J. A. Little and Donald B. Rubin. Statistical analysis with missing data. 1988.

[17] Roderick J. A. Little and Donald B. Rubin. Statistical analysis with missing data, third edition. *Wiley Series in Probability and Statistics*, 2019.

[18] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. A secure federated transfer learning framework. *IEEE Intelligent Systems*, 35:70–82, 2020.

[19] Lamin R. Mansaray, Adam Sheka Kanu, Lingbo Yang, Jingfeng Huang, and Fumin Wang. Evaluation of machine learning models for rice dry biomass estimation and mapping using quad-source optical imagery. *GIScience & Remote Sensing*, 57:785 – 796, 2020.

[20] Board on Physics. A strategy for active remote sensing amid increased demand for radio spectrum. 2015.

[21] Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, 2011.

[22] Ritvik Sahajpal, Lucas Fontana, Pedro Lafluf, Guillermo Leale, Estefania Puricelli, Dan O'Neill, Mehdi Hosseini, Mauricio Varela, and Inbal Becker-Reshef. Using machine-learning models for field-scale crop yield and condition modeling in argentina. 2020.

[23] Joseph L Schafer. Multiple imputation: a primer. *Statistical methods in medical research*, 8(1): 3–15, 1999.

[24] Joseph L. Schafer and John W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7 2:147–77, 2002.

[25] Joseph L Schafer and Nathaniel Schenker. Inference with imputed conditional means. *Journal of the American Statistical Association*, 95(449):144–154, 2000.

[26] Ananda Theertha Suresh, Ziteng Sun, Jae Hun Ro, and Felix X. Yu. Correlated quantization for distributed mean estimation and optimization. *ArXiv*, abs/2203.04925, 2022.

[27] Anastasios A. Tsiatis. Semiparametric theory and missing data. 2006.

[28] Jaideep Vaidya and Chris Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *Knowledge Discovery and Data Mining*, 2003.

[29] Naisyin Wang and James M. Robins. Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85:935–948, 1998.

[30] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *arXiv: Artificial Intelligence*, 2019.

[31] Hwanjo Yu, Jaideep Vaidya, and Xiaoqian Jiang. Privacy-preserving svm classification on vertically partitioned data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2006.

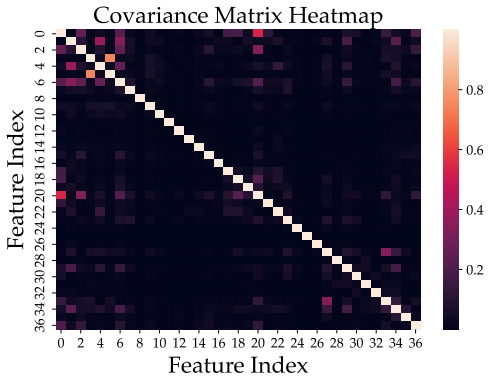
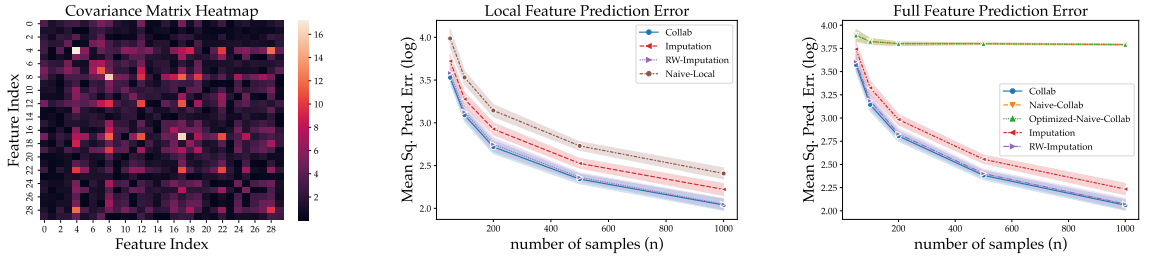


Figure 2: Covariance Heatmap for US Census Experiment



(a) Covariance Matrix Heatmap

(b) Prediction Error on 20-feature Datatcenter

(c) Prediction Error on 30-feature Datatcenter

Figure 3: Experimental results for Synthetic Experiment

A Experimental Details

A.1 Census Experimental Details

We use the 15 of the 17 features in the ACSTravelTime dataset—which include Age, Educational Attainment, Marital Status, Sex, Disability record, Mobility status, Relationship, etc. More specifically, using the notation from [7], we choose to keep the 'AGEP', 'SCHL', 'MAR', 'SEX', 'DIS', 'MIG', 'RELP', 'RAC1P', 'PUMA', 'CIT', 'OCCP', 'JWTR', 'POWPUMA', and 'POVPIP' features. We choose to exclude the State code (ST) and Employment Status of Parents (ESP) as a quick way to bypass low-rank covariance matrix issues. We turn the columns 'MAR', 'SEX', 'DIS', 'MIG', 'RAC1P', 'CIT', 'JWTR' into one-hot vectors. We make use commute time 'JWMNP' as the target variable. We clean our data by making sure AGEP (Age) must be greater than 16, PWGTP (Person weight) must be greater than or equal to 1, ESR (Employment status recode) must be equal to 1 (employed), and JWMNP (Travel time to work) is greater than 0. We normalize our features and targets by centering and dividing by the standard deviation computed from the training data. The California datacenter has access to all of the features. The New York datacenter has access to all categories except 'AGEP'. The Texas datacenter has access to all but 'AGEP', 'SCHL'. The Florida datacenter has access to all but 'AGEP', 'SCHL', 'MAR', 'SEX', and the Illinois datacenter has access to all but 'AGEP', 'SCHL', 'MAR', 'SEX', 'DIS', 'MIG'.

A.2 Synthetic Experiments

We start with a synthetic experiment where we generate $m = 30$ agents observing some subset of $d = 30$ features. Ten of the agents will have access to random subsets of 20 of the features. The other twenty agents will have access to random subsets of 15 of the features. Each agent will have n samples which we vary in this experiment. We sample the features from a $\mathsf{N}(0, \Sigma)$ distribution. We generate Σ by first

generating d eigenvalues by sampling d times from a uniform $[0, 1]$ distribution. We randomly select 3 eigenvalues to multiply by 10 and use these eigenvalues to populate the diagonal of a diagonal matrix Λ. Then we use a randomly generated orthogonal matrix W to form $\Sigma \coloneqq W\Lambda W^T$. We plot a heatmap of Σ in Figure 3(a). For each method that we test, we run 20 trials to form 95% confidence intervals.

We compare our method COLLAB, against the Imputation and RW-Imputation methods we outlined in Section 6.1. After we train each of these methods using the data on our 30 agents, we measure how well these methods perform in using the features of a test-agent with access to 20 of the total 30 features to predict outputs. We will also compare our methods against Naive-Local, where we only use the n training datapoints of the 20 features our test-agent has access to, also described in Section 6.1. We plot this result in Figure 3(b).

We also compare our methods in an alternative setting where the test-center of interest has access to all 30 features. This setup models the setting where we are interested making the best possible predictions from all of the features available. In this experiment, we compare against Naive-Collab, Optimized-Naive-Collab, described in Section 6.1. We note that Optimized-Naive-Collab uses fresh labeled samples without any missing features during gradient descent, so in this sense, Optimized-Naive-Collab is more powerful than our method. We plot this result in Figure 3(c).

We see that reweighting is important; this is why COLLAB and RW-Imputation outperform the unweighted Imputation method. Our COLLAB method improves over the Naive-Local approach, meaning that the agents are benefiting from sharing information. COLLAB also matches the performance of the RW-Imputation method, despite only needing to communicate the learned parameters of each agent's model, as opposed to all of the data on each agent. The Naive-Collab approaches level out very quickly, likely reflecting the fact that these methods are biased, as the covariance of our underlying data is far from isotropic.

B Proofs for Section 3

Lemma B.1. *For any positive definite matrices $W_i \in \mathbb{R}^{d_i \times d_i}$, $i = 1, 2, \ldots, m$, the aggregated estimator $\hat{\theta}$ in Eq. (3) is consistent $\hat{\theta} \xrightarrow{p} \theta$. In addition, if $X_i \sim \mathsf{N}(0, \Sigma)$, we have unbiasedness $\mathbb{E}[\hat{\theta}] = \theta$ where $\mathbb{E}$ is over the random data X_i and noise ξ_i.*

B.1 Proof of Lemma B.1

For the general case, identify for $\hat{\theta}_i$, we can write

$$\begin{aligned}\hat{\theta}_i &= (X_{i+}^\top X_{i+})^{-1} X_{i+} y_i = (X_{i+}^\top X_{i+})^{-1} X_{i+}^\top (X_{i+}\theta_{i+} + X_{i-}\theta_{i-} + \xi_i) \\ &= \theta_{i+} + (X_{i+}^\top X_{i+})^{-1}(X_{i+}^\top X_{i-}\theta_{i-} + X_{i+}^\top \xi_i) \\ &= \theta_{i+} + \left(\frac{1}{n} X_{i+}^\top X_{i+}\right)^{-1} \left(\frac{1}{n} X_{i+}^\top X_{i-}\theta_{i-} + \frac{1}{n} X_{i+}^\top \xi_i\right).\end{aligned}$$

The weak law of large numbers implies that $X_{i+}^\top X_{i+}/n \xrightarrow{p} \Sigma_{i+}$, $X_{i+}^\top X_{i-}/n \xrightarrow{p} \Sigma_{i\pm}$ and $\frac{1}{n} X_{i+}^\top \xi_i \xrightarrow{p} 0$. Then Slutsky's theorem gives the consistency guarantee

$$\hat{\theta}_i \xrightarrow{d} \theta_{i+} + \Sigma_{i+}^{-1}(\Sigma_{i\pm}\theta_{i-} + 0) = \theta_{i+} + \Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} = T_i\theta,$$

which is equivalent to $\hat{\theta}_i \xrightarrow{p} T_i\theta$. Substituting back into $\hat{\theta}$, we can obtain again from continuous mapping theorem that

$$\hat{\theta} = \left(\sum_{i=1}^m T_i^\top W_i T_i\right)^{-1} \left(\sum_{i=1}^m T_i^\top W_i \hat{\theta}_i\right) \xrightarrow{p} \left(\sum_{i=1}^m T_i^\top W_i T_i\right)^{-1} \left(\sum_{i=1}^m T_i^\top W_i T_i \theta\right) = \theta.$$

Next, we specialize to Gaussian features and show $\hat{\theta}$ is indeed unbiased in this case. By the tower property, we can write for each local OLS estimator,

$$\mathbb{E}[\hat{\theta}_i] = \mathbb{E}[(X_{i+}^\top X_{i+})^{-1} X_{i+} y_i] = \mathbb{E}\left[\mathbb{E}[(X_{i+}^\top X_{i+})^{-1} X_{i+}^\top (X_{i+}\theta_{i+} + X_{i-}\theta_{i-} + \xi_i) \mid X_{i+}]\right]$$

$$= \theta_{i+} + \mathbb{E}\left[(X_{i+}^\top X_{i+})^{-1} X_{i+}^\top \mathbb{E}[X_{i-} \mid X_{i+}]\right] \theta_{i-}.$$

We want to compute $\mathbb{E}[X_{i-} \mid X_{i+}]$ and the key observation is that with Gaussianity in X_i, we have

$$\begin{aligned}&\mathsf{Cov}(x_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+}, x_{i+}) = \mathsf{Cov}(x_{i-}, x_{i+}) - \Sigma_{i\mp}\Sigma_{i+}^{-1}\mathsf{Cov}(x_{i+}, x_{i+}) \\ &= \Sigma_{i\mp} - \Sigma_{i\mp}\Sigma_{i+}^{-1} \cdot \Sigma_{i+} = 0,\end{aligned}$$

and therefore x_{i+} is independent of $x_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+}$, which further implies that

$$\mathbb{E}\left[X_{i-} \mid X_{i+}\right] = \mathbb{E}\left[X_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm} \mid X_{i+}\right] + \mathbb{E}\left[X_{i-} - X_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm} \mid X_{i+}\right] = X_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm}.$$

Substituting the above property into computing the expectation of local estimates $\hat{\theta}_i$, it then holds

$$\mathbb{E}[\hat{\theta}_i] = \theta_{i+} + \mathbb{E}[(X_{i+}^\top X_{i+})^{-1} X_{i+}^\top X_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm}]\theta_{i-} = \theta_{i+} + \Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} = T_i\theta.$$

We can then conclude the proof as

$$\mathbb{E}[\hat{\theta}] = \left(\sum_{i=1}^m T_i^\top W_i T_i\right)^{-1} \left(\sum_{i=1}^m T_i^\top W_i T_i \theta\right) = \theta.$$

B.2 Proof of Theorem 3.1

We first study the central limit theorem for local OLS estimators $\hat{\theta}_i$. Let the data matrices $X_{i+} = [x_{i+}^1, \ldots, x_{i+}^n]^\top$ and $X_{i-} = [x_{i-}^1, \ldots, x_{i-}^n]$ and the noise vector $\xi_i = [\xi_i^1, \ldots, \xi_i^n]^\top$, we can write out for $\hat{\theta}_i$ that

$$\sqrt{n}\left(\hat{\theta}_i - T_i\theta\right) = \underbrace{\left(X_{i+}^\top X_{i+}/n\right)^{-1}}_{\text{(I)}} \cdot \underbrace{\frac{1}{\sqrt{n}} X_{i\pm}^\top \left\{(X_{i-} - X_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm})\theta_{i-} + \xi_i\right\}}_{\text{(II)}}. \tag{5}$$

For (II), note that

$$\frac{1}{\sqrt{n}} X_{i\pm}^\top \left\{(X_{i-} - X_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm})\theta_{i-} + \xi_i\right\} = \frac{1}{\sqrt{n}} \sum_{k=1}^n x_{i+}^j \left\{(x_{i-}^j - \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+}^j)^\top \theta_{i-} + \xi_i^j\right\}.$$

The summands are independent mean zero random vectors, since

$$\begin{aligned}&\mathbb{E}\left[x_{i+}^j \left\{(x_{i-}^j - \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+}^j)^\top \theta_{i-}\right\}\right] = \left(\mathbb{E}\left[x_{i+}^j {x_{i-}^j}^\top\right] - \mathbb{E}\left[x_{i+}^j {x_{i+}^j}^\top\right]\Sigma_{i+}^{-1}\Sigma_{i\pm}\right)\theta_{i-} \\ &= \left(\Sigma_{i\pm} - \Sigma_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm}\right)\theta_{i-} = 0,\end{aligned}$$

and $\mathbb{E}[x_{i+}^j \xi_i^j] = \mathbb{E}[x_{i+}^j] \cdot \mathbb{E}[\xi_i^j] = 0$. Denote by $z_{i+}^j := x_{i-}^j - \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+}^j$ and we can infer from the above display that x_{i+} and z_{i+} are uncorrelated. (II) is then asymptotically normal by CLT with limiting covariance (suppressing the superscript j below)

$$\begin{aligned}&\mathsf{Cov}\left(x_{i+}\left\{(x_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+})^\top \theta_{i-} + \xi_i\right\}\right) = \mathbb{E}\left[x_{i+}\theta_{i-}^\top z_{i+} z_{i+}^\top \theta_{i-} x_{i+}^\top\right] + \mathbb{E}\left[\xi_i^2 x_{i+} x_{i+}^\top\right] \\ &= \mathbb{E}\left[x_{i+}\theta_{i-}^\top z_{i+} z_{i+}^\top \theta_{i-} x_{i+}^\top\right] + \sigma^2\Sigma_{i+} := Q_i.\end{aligned} \tag{6}$$

If X_i are Gaussian random vectors, we can additionally have independence between z_{i+} and x_{i+} by zero correlation. Therefore

$$\begin{aligned}&\mathbb{E}\left[x_{i+}\theta_{i-}^\top z_{i+} z_{i+}^\top \theta_{i-} x_{i+}^\top\right] = \mathbb{E}\left[x_{i+}\theta_{i-}^\top \mathbb{E}\left[z_{i+} z_{i+}^\top\right]\theta_{i-} x_{i+}^\top\right] \\ &= \theta_{i-}^\top \mathsf{Cov}\left(x_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+}\right)\theta_{i-} \cdot \mathbb{E}\left[x_{i+} x_{i+}^\top\right] = \theta_{i-}^\top \left(\Sigma_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1}\Sigma_{i\pm}\right)\theta_{i-} \cdot \Sigma_{i+} = \|\theta_{i-}\|_{\Gamma_{i-}}^2 \Sigma_{i+},\end{aligned}$$

and $Q_i = (\|\theta_{i-}\|_{\Gamma_{i-}}^2 + \sigma^2)\Sigma_{i+}$.

We proceed to show $C(W_1, \cdots, W_n) \succeq C^\star$ under general feature distribution $\mathcal{P}$ and $W_i^\star := \Sigma_{i+} Q_i^{-1} \Sigma_{i+}$. By Slutsky theorem, (I) converges to Σ_{i+}^{-1} in probability and we can conclude from Eq. (5) that

$$\sqrt{n}\left(\hat{\theta}_i - T_i\theta\right) \xrightarrow{d} \mathsf{N}\left(0, \Sigma_{i+}^{-1} Q_i \Sigma_{i+}^{-1}\right). \tag{7}$$

Further from $\hat{\theta} = \left(\sum_{i=1}^m T_i^\top W_i T_i\right)^{-1}\left(\sum_{i=1}^m T_i^\top W_i \hat{\theta}_i\right)$, it follows that

$$\sqrt{n}\left(\hat{\theta}_i - \theta\right) = \mathsf{N}(0, C(W_1, \cdots, W_n))$$

where

$$C(W_1, \cdots, W_n) = \left(\sum_{i=1}^m T_i^\top W_i T_i\right)^{-1} \cdot \left(\sum_{i=1}^m T_i^\top W_i W_i^{\star -1} W_i T_i\right) \cdot \left(\sum_{i=1}^m T_i^\top W_i T_i\right)^{-1}. \tag{8}$$

With the choice of $W_i = W_i^\star$, we achieve the claimed lower bound for asymptotic covariance as in this case $C(W_1, \cdots, W_m) = \left(\sum_{i=1}^m T_i^\top W_i^\star T_i\right)^{-1}$. It thus remains to show

$$C(W_1, \cdots, W_n) \succeq \left(\sum_{i=1}^m T_i^\top W_i^\star T_i\right)^{-1} = C^\star.$$

To prove the above claim, we construct auxiliary matrices M_i as

$$M_i = \begin{bmatrix} T_i^\top W_i^\star T_i & T_i^\top W_i T_i \\ T_i^\top W_i T_i & T_i^\top W_i W_i^{\star -1} W_i T_i \end{bmatrix} = \begin{bmatrix} T_i^\top W_i^{\star \frac{1}{2}} \\ T_i^\top W_i W_i^{\star -\frac{1}{2}} \end{bmatrix} \begin{bmatrix} T_i^\top W_i^{\star \frac{1}{2}} \\ T_i^\top W_i W_i^{\star -\frac{1}{2}} \end{bmatrix}^\top \succeq 0.$$

Therefore

$$\sum_{i=1}^m M_i = \begin{bmatrix} C^{\star -1} & \sum_{i=1}^m T_i^\top W_i T_i \\ \sum_{i=1}^m T_i^\top W_i T_i & \sum_{i=1}^m T_i^\top W_i W_i^{\star -1} W_i T_i \end{bmatrix} \succeq 0.$$

As the Schur complement is also p.s.d. we can conclude with

$$\begin{aligned} 0 \preceq C^{\star -1} - \left(\sum_{i=1}^m T_i^\top W_i T_i\right) \cdot \left(\sum_{i=1}^m T_i^\top W_i W_i^{\star -1} W_i T_i\right)^{-1} \\ \cdot \left(\sum_{i=1}^m T_i^\top W_i T_i\right) = C^{\star -1} - C(W_1, \cdots, W_n)^{-1}. \end{aligned}$$

B.3 Proof of Corollary 3.2

We first prove (i) and asymptotic normality of $\sqrt{n}(\hat{\theta}^{\mathrm{clb}} - \theta) \xrightarrow{d} \mathsf{N}\left(0, C^{\mathrm{g}}\right)$. We point out that Theorem 3.1 is not directly applicable as we use estimated weights that reuse the training data. We claim consistency for $\hat{W}_i^{\mathrm{g}} \xrightarrow{p} W^{\mathrm{g}}$, and under this premise, the proof is rather straightforward since we can write

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{clb}} - \theta\right) = \left(\sum_{i=1}^m T_i^\top \hat{W}_i^{\mathrm{g}} T_i\right)^{-1} \left(\sum_{i=1}^m T_i^\top \hat{W}_i^{\mathrm{g}} (\hat{\theta}_i - T_i\theta)\right).$$

With the asymptotic normality established for $\sqrt{n}(\hat{\theta}_i - T_i\theta)$ in Eq. (7), Slutsky's theorem and continuous mapping theorem, we can conclude that $\sqrt{n}(\hat{\theta}^{\mathrm{clb}} - \theta) \xrightarrow{d} \mathsf{N}\left(0, C^{\mathrm{g}}\right)$. Now it remains to showing $\hat{W}_i^{\mathrm{g}} \xrightarrow{p} W^{\mathrm{g}}$, this is from Slutksy's theorem applied to $\hat{W}_i^{\mathrm{g}} = \hat{\Sigma}_{i+}/\hat{R}_i$ and the weak law of large numbers as follows

$$\hat{\Sigma}_{i+} = \frac{X_{i+}^\top X_{i+}}{n} \xrightarrow{p} \Sigma_{i+}, \qquad \hat{R}_i = \frac{1}{n}\|X_{i+}\hat{\theta}_i - y\|_2^2 \xrightarrow{p} \mathbb{E}[\left\|x_{i+}^\top T_i\theta - y_i\right\|_2^2],$$

where

$$\mathbb{E}[\|x_{i+}^\top T_i\theta - y_i\|_2^2] = \mathbb{E}[\|x_{i+}^\top \Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} - x_{i-}^\top\theta_{i-}\|_2^2] + \sigma^2$$
$$= \|\theta_{i-}\|^2_{\mathsf{Cov}\left(x_{i-}-\Sigma_{i\mp}\Sigma_{i+}^{-1}x_{i+}\right)} + \sigma^2 = \|\theta_{i-}\|^2_{\Gamma_{i-}} + \sigma^2.$$

We proceed to prove (ii). Applying delta method to the mapping $\theta \mapsto T_i\theta, \mathbb{R}^d \to \mathbb{R}^{d_i}$ on $\hat{\theta}(W_1^\star, \cdots, W_m^\star)$ immediately yields the asymptotic normality for $\hat{\theta}_i^{\text{clb}}$. It only remains to show $T_iC^\star T_i^\top \preceq W_i^{\star -1}$.

Identify $W_i^{\star -1} - T_iC^\star T_i^\top$ as the Schur complement for the block matrix

$$M = \begin{bmatrix} W_i^{\star -1} & T_i \\ T_i^\top & C^{\star -1} \end{bmatrix},$$

and it suffices to show $M \succeq 0$. This follows from $C^\star = (\sum_{i=1}^m T_i^\top W_i^\star T_i)^{-1}$ and thus

$$M = \begin{bmatrix} W_i^{\star -1} & T_i \\ T_i^\top & \sum_{j=1}^m T_j^\top W_j^\star T_j \end{bmatrix} \succeq \begin{bmatrix} W_i^{\star -1} & T_i \\ T_i^\top & T_i^\top W_i^\star T_i \end{bmatrix} = \begin{bmatrix} W_i^{\star -\frac{1}{2}} \\ T_i^\top W_i^{\star \frac{1}{2}} \end{bmatrix} \begin{bmatrix} W_i^{\star -\frac{1}{2}} \\ T_i^\top W_i^{\star \frac{1}{2}} \end{bmatrix}^\top \succeq 0.$$

C Proofs for Section 4

C.1 Proof of Theorem 4.1

The key part of the proof is showing $\hat{\theta}_i^{\text{imp}} = T_i^\top (T_iT_i^\top)^{-1}\hat{\theta}_i$. If we can have this claim established, we can make use of the following transformation of the loss function

$$\sum_{i=1}^m \left\| T_i^\top (T_iT_i^\top)^{-1}T_i\theta - \hat{\theta}_i^{\text{imp}} \right\|_{W_i}^2 = \sum_{i=1}^m \left\| T_i^\top (T_iT_i^\top)^{-1}T_i\theta - T_i^\top (T_iT_i^\top)^{-1}\hat{\theta}_i \right\|_{W_i}^2$$
$$= \sum_{i=1}^m \left\| T_i\theta - \hat{\theta}_i \right\|^2_{(T_iT_i^\top)^{-1}T_iW_iT_i^\top (T_iT_i^\top)^{-1}}.$$

This reduces the optimization problem into the same one in Eq. (3) up to weight transformation, and the same lower bound for asymptotic covariance in Theorem 3.1 applies. Hence

$$C^{\text{imp-glb}}(\alpha_1, \cdots, \alpha_m) \succeq C^\star.$$

By taking $W_i = T_i^\top W_i^\star T_i$, we have the transformed weights satisfy

$$(T_iT_i^\top)^{-1}T_iW_iT_i^\top (T_iT_i^\top)^{-1} = (T_iT_i^\top)^{-1}T_i^\top W_i^\star T_i(T_iT_i^\top)^{-1} = W_i^\star.$$

From the optimality condition in Theorem 3.1, the equality holds under this choice of W_i's.

It then boils down to proving the claim $\hat{\theta}_i^{\text{imp}} = T_i^\top (T_iT_i^\top)^{-1}\hat{\theta}_i$. We make use of the following two properties of Moore-Penrose pseudo inverse—for $A \in \mathbb{R}^{d_i \times d}$ of rank d_i,

$$(A^\top A)^\dagger = A^\dagger (A^\dagger)^\top, \qquad A^\dagger = A^\top (AA^\top)^{-1}.$$

Substituting $A = (X_{i+}^\top X_{i+})^{\frac{1}{2}}T_i$ into the above displays, we then have

$$\hat{\theta}_i^{\text{imp}} = (T_i^\top X_{i+}^\top X_{i+}T_i)^\dagger T_i^\top X_{i+}^\top y_i$$
$$= T_i^\top (X_{i+}^\top X_{i+})^{\frac{1}{2}} \left((X_{i+}^\top X_{i+})^{\frac{1}{2}}T_iT_i^\top (X_{i+}^\top X_{i+})^{\frac{1}{2}} \right)^{-2} \cdot (X_{i+}^\top X_{i+})^{\frac{1}{2}}T_iT_i^\top X_{i+}^\top y_i$$
$$= T_i^\top (X_{i+}^\top X_{i+})^{\frac{1}{2}} \left((X_{i+}^\top X_{i+})^{-\frac{1}{2}}(T_iT_i^\top)^{-1}(X_{i+}^\top X_{i+})^{-\frac{1}{2}} \right)^{2} \cdot (X_{i+}^\top X_{i+})^{\frac{1}{2}}T_iT_i^\top X_{i+}^\top y_i$$
$$= T_i^\top (T_iT_i^\top)^{-1} \cdot (X_{i+}^\top X_{i+})^{-1} \cdot (T_iT_i^\top)^{-1} \cdot T_iT_i^\top X_{i+}^\top y_i$$
$$= T_i^\top (T_iT_i^\top)^{-1} \cdot (X_{i+}^\top X_{i+})^{-1}X_{i+}^\top y_i = T_i^\top (T_iT_i^\top)^{-1}\hat{\theta}_i.$$

C.2 Proof of Theorem 4.2

By a direct calculation, we have

$$\begin{aligned}
\hat{\theta}^{\text{imp-glb}} - \theta &= \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top X_{i+} T_i\right)^{-1} \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top y_i\right) - \theta \\
&= \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top X_{i+} T_i\right)^{-1} \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top (X_{i+}\theta_{i+} + X_{i-}\theta_{i-} + \xi_i)\right) - \theta \\
&= \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top X_{i+} T_i\right)^{-1} \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top (X_{i+}\theta_{i+} + X_{i-}\theta_{i-} - X_{i+}T_i\theta + \xi_i)\right) \\
&= \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top X_{i+} T_i\right)^{-1} \left(\sum_{i=1}^m \alpha_i T_i^\top X_{i+}^\top (X_{i-}\theta_{i-} - X_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} + \xi_i)\right).
\end{aligned}$$

Consequently

$$\sqrt{n}\left(\hat{\theta}^{\text{imp-glb}} - \theta\right) = \left(\sum_{i=1}^m \alpha_i T_i^\top \cdot \frac{1}{n} X_{i+}^\top X_{i+} \cdot T_i\right)^{-1} \cdot \left(\sum_{i=1}^m \alpha_i T_i^\top \cdot \frac{1}{\sqrt{n}} X_{i+}^\top (X_{i-}\theta_{i-} - X_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} + \xi_i)\right)$$

Following the same proof steps applied to Eq. (5) in Appendix B.2, we can conclude that

$$\begin{aligned}
&\sqrt{n}\left(\hat{\theta}^{\text{imp-glb}} - \theta\right) \\
&\xrightarrow{d} \mathsf{N}\Bigg(0, \underbrace{\left(\sum_{i=1}^m \alpha_i T_i^\top \Sigma_{i+} T_i\right)^{-1} \left(\sum_{i=1}^m \alpha_i^2 T_i^\top Q_i T_i\right) \left(\sum_{i=1}^m \alpha_i T_i^\top \Sigma_{i+} T_i\right)^{-1}}_{:= C^{\text{imp-glb}}(\alpha_1, \cdots, \alpha_m)}\Bigg),
\end{aligned}$$

with the same Q_i's as in Eq. (6), and with Gaussianity of X_i, we also have the explicit form $Q_i = (\|\theta_{i-}\|_{\Gamma_{i-}}^2 + \sigma^2)\Sigma_{i+}$. Note that if $\alpha_i = 1/(\|\theta_{i-}\|_{\Gamma_{i-}}^2 + \sigma^2)$,

$$C^{\text{imp-glb}}(\alpha_1, \cdots, \alpha_m) = \left(\sum_{i=1}^m \frac{T_i^\top \Sigma_{i+} T_i}{\|\theta_{i-}\|_{\Gamma_{i-}}^2 + \sigma^2}\right)^{-1} = C^{\text{g}} = C^\star.$$

Finally, to show $C^{\text{imp-glb}}(\alpha_1, \cdots, \alpha_m) \succeq C^\star$, we identify from Eq. (8) that

$$C^{\text{imp-glb}}(\alpha_1, \cdots, \alpha_m) = C(\alpha_1\Sigma_{1+}, \cdots, \alpha_m\Sigma_{m+}) \succeq C^\star,$$

where the last inequality follows from Theorem 3.1.

D Proofs for Section 5

We will use the van Trees inequality to prove our lower bound shown. In particular, we will use a slight modification to Theorem 4 of [10], which we state as a corollary below here. Throughout this section, we let $\psi : \mathbb{R}^d \to \mathbb{R}^s$ be an absolutely continuous function. The distribution P_θ in the family $\{P_\theta\}_{\theta \in \mathbb{R}^d}$ is assumed to have density p_θ which satisfies $\int_{\mathbb{R}^d} \|\nabla p_\theta(x)\|_2^2 \, dx < \infty$. Let P_θ^j for $j \in [m]$ denote the distribution over either $\tilde{\theta}_j^n$ or $y_j \in \mathbb{R}^n$. Let $\mathcal{I}_i^n(\theta)$ denote the Fisher Information of P_θ^i, and let $\mathcal{I}^n(\theta) = \sum_{i=1}^m \mathcal{I}_i^n(\theta)$ denote the Fisher Information of P_θ. We note that P_θ is allowed to depend on n.

Corollary D.1 (Gassiat [10])**.** *Let $\psi : \mathbb{R}^d \to \mathbb{R}^s$ be an absolutely continuous function such that $\nabla\psi(\theta)$ is continuous at θ_0. For all n, let all distributions P_θ in the family $\{P_\theta\}_{\theta \in \mathbb{R}^d}$ have density p_θ which*

satisfies $\int_{\mathbb{R}^d} \|\nabla p_\theta(x)\|_2^2 \, dx < \infty$. *If* $\lim_{c\to\infty} \lim_{n\to\infty} \sup_{\|h\|_2<1} \mathcal{I}^n(\theta_0 + ch/\sqrt{n})/n$ *exists almost surely and is positive definite, denote it by* ρ. *Then for all sequences* $(\hat{\theta}_n)_{n\geq 1}$ *of statistics* $S_n : \mathcal{X}^n \to \mathbb{R}^s$ *and for all* $u \in \mathbb{R}^s$

$$\liminf_{c\to\infty} \liminf_{n\to\infty} \sup_{\|h\|<1} \mathbb{E}^n_{\theta_0 + \frac{ch}{\sqrt{n}}} \left[\left\langle \sqrt{n} \left(\hat{\theta}_n - \psi\left(\theta_0 + \frac{ch}{\sqrt{n}}\right)\right), u \right\rangle^2 \right] \geq u^\top \nabla\psi(\theta_0)^\top \rho^{-1} \nabla\psi(\theta_0) u$$

Proof. The main difference between our version of the proof and the one presented in Theorem 4 of Gassiat [10] is that we do not assume $\mathcal{I}^n = n\mathcal{I}$. We also select $\ell(x) = \langle u, x\rangle^2$ in particular. All the steps and notation remain the same except with $n\mathcal{I}$ replaced with $\mathcal{I}^n$ up until equation (13), which we define with a modified choice of $\Gamma_{c,n}$

$$\Gamma_{c,n} \coloneqq \left(\int_{\mathcal{B}_p([0],1)} \nabla\psi(\theta_0 + ch/\sqrt{n}) q(h) dh \right)^\top \left(\frac{1}{c^2}\mathcal{I}_q + \frac{1}{n} \int_{\mathcal{B}_p([0],1)} \mathcal{I}^n(\theta_0 + ch/\sqrt{n}) q(h) dh \right)^{-1} \times \left(\int_{\mathcal{B}_p([0],1)} \nabla\psi(\theta_0 + ch/\sqrt{n}) q(h) dh \right).$$

By definition of ρ, with probability 1,

$$\lim_{c\to\infty} \lim_{n\to\infty} \Gamma_{c,n} = \nabla\psi(\theta_0)^\top \rho^{-1} \nabla\psi(\theta_0)$$

□

D.1 Proof of Theorem 5.1

We will apply Corollary D.1 and apply it to two different choices of ψ to get the full feature minimax bound and missing feature minimax bound respectively. For notational simplicty, let P_θ denote the distribution over $\{\tilde{\theta}_i^n\}_{i\in[m]}$ induced by θ. P_θ is in the exponential family, so the conditions of Corollary D.1 are satisfied.

We begin by computing the Fisher Information. Let P_θ^j for $j \in [m]$ denote the distribution over $\tilde{\theta}_j^n \in \mathbb{R}^{d_j}$. Let $\mathcal{I}_i^n(\theta)$ denote the Fisher Information of P_θ^i, and let $\mathcal{I}^n(\theta) = \sum_{i=1}^m \mathcal{I}_i^n$ denote the Fisher Information of P_θ. Let x_{i+} denote an arbitrary row of X_{i+}. Let x_{i-} be drawn from $\mathsf{N}(\mu_{i-}(x_{i+}), \Gamma_{i-})$. Some straightforward calculations tell us $\mu_{i-}(x_{i+}) = \Sigma_{i\mp}\Sigma_{i+}^{-1} x_{i+}$ and $\Gamma_{i-} = \Sigma_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1}\Sigma_{i\pm}$. From this we can deduce that $\theta_{i-}^T x_{i-}$ is distributed as $\mathsf{N}(\mu_{i-}^T\theta_{i-}, \theta_{i-}^T\Gamma_{i-}\theta_{i-})$; we use μ_{i-} in place of $\mu_{i-}(x_{i-})$ for simplicity. And y_i is distributed as P_θ^i which is $\mathsf{N}(\theta^T\gamma, \theta_{i-}^T\Gamma_{i-}\theta_{i-} + \sigma^2)$ where $\gamma \coloneqq [x_{i+}^T\Pi_{i+}, \mu_{i-}^T\Pi_{i-}]^T$. From this we can deduce that P_θ^i is $\mathsf{N}\left(J_i\Pi_i\theta, \beta_i^{-1}\widehat{\Sigma}_{i+}^{-1}\right)$, where $\beta_i^{-1} \coloneqq \frac{\theta_{i-}\Gamma_{i-}\theta_{i-} + \sigma^2}{n}$; let p_θ^i denote its density. We know $\mathcal{I}^n(\theta) = \sum_{i=1}^m \mathcal{I}_i^n(\theta)$ due to independence. All that remains is to compute $\mathcal{I}_i^n(\theta)$.

$$\mathcal{I}_i^n(\theta) = \int \nabla_\theta \log p_\theta^i(z) [\nabla_\theta \log p_\theta^i(z)]^T p_\theta^i(z) dz.$$

We know that for some constant C,

$$\log p_i^\theta(z) = C + \frac{d_i}{2}\log(\beta_i) - \frac{\beta_i}{2}\left\| \widehat{\Sigma}_{i+}^{\frac{1}{2}}\theta_{i+} + \widehat{\Sigma}_{i+}^{\frac{1}{2}}\Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} - \widehat{\Sigma}_{i+}^{\frac{1}{2}} z \right\|_2^2.$$

Taking derivaties we get that

$$\begin{aligned}
\nabla_{\theta_{i+}} \log p_i^\theta(z) &= -\beta_i \left[\widehat{\Sigma}_{i+}\theta_{i+} + \widehat{\Sigma}_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} - \widehat{\Sigma}_{i+} z \right] \\
\nabla_{\theta_{i-}} \log p_i^\theta(z) &= \left[-\frac{d_i}{n} + \left\| \widehat{\Sigma}_{i+}^{\frac{1}{2}}\theta_{i+} + \widehat{\Sigma}_{i+}^{\frac{1}{2}}\Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} - \widehat{\Sigma}_{i+}^{\frac{1}{2}} z \right\|_2^2 \right] \beta_i \Gamma_{i-}\theta_{i-} \\
&\quad + \left[\Sigma_{i\mp}\Sigma_{i+}^{-1}\widehat{\Sigma}_{i+}\theta_{i+} + \Sigma_{i\mp}\Sigma_{i+}^{-1}\widehat{\Sigma}_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm}\theta_{i-} - \Sigma_{i\mp}\Sigma_{i+}^{-1}\widehat{\Sigma}_{i+} z \right] \beta_i
\end{aligned}$$

Let $b^2 = \left\| \widehat{\Sigma}_{i+}^{\frac{1}{2}} \theta_{i+} + \widehat{\Sigma}_{i+}^{\frac{1}{2}} \Sigma_{i+}^{-1} \Sigma_{i\pm} \theta_{i-} - \widehat{\Sigma}_{i+}^{\frac{1}{2}} z \right\|_2^2$. Now we compute the expectation over outer products:

$$
\begin{aligned}
&\mathbb{E}[\nabla_{\theta_{i+}} \log p_i^\theta(z) \nabla_{\theta_{i+}} \log p_i^\theta(z)^T] = \beta_i \widehat{\Sigma}_{i+} \\
&\mathbb{E}[\nabla_{\theta_{i+}} \log p_i^\theta(z) \nabla_{\theta_{i-}} \log p_i^\theta(z)^T] = \beta_i^2 \widehat{\Sigma}_{i+} \beta_i^{-1} \widehat{\Sigma}_{i+}^{-1} \widehat{\Sigma}_{i+} \Sigma_{i+}^{-1} \Sigma_{i\pm} = \beta_i \widehat{\Sigma}_{i+} \Sigma_{i+}^{-1} \Sigma_{i\pm} \\
&\mathbb{E}[\nabla_{\theta_{i-}} \log p_i^\theta(z) \nabla_{\theta_{i-}} \log p_i^\theta(z)^T] = \beta_i \Sigma_{i\mp} \Sigma_{i+}^{-1} \widehat{\Sigma}_{i+} \Sigma_{i+}^{-1} \Sigma_{i\pm} \\
&\qquad + \left(\frac{d_i^2}{n^2} + \mathbb{E}[b^2] \frac{2d_i}{n} + \mathbb{E}[b^4] \right) \beta_i^2 \Gamma_{i-} \theta_{i-} \theta_{i-}^T \Gamma_{i-} \\
&\quad = \beta_i \Sigma_{i\mp} \Sigma_{i+}^{-1} \widehat{\Sigma}_{i+} \Sigma_{i+}^{-1} \Sigma_{i\pm} + \left(\frac{d_i^2}{n^2} + \frac{2\beta_i^{-1} d_i^2}{n} + \beta_i^{-2}(2d_i + d_i^2) \right) \beta_i^2 \Gamma_{i-} \theta_{i-} \theta_{i-}^T \Gamma_{i-}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{I}_i^n(\theta) &= \int \nabla_\theta \log p_\theta^i(z) [\nabla_\theta \log p_\theta^i(z)]^T p_\theta^i(z) dz \\
&= \int \Pi_i^T \begin{bmatrix} \nabla_{\theta_{i+}} \log p_\theta^i(z) \\ \nabla_{\theta_{i-}} \log p_\theta^i(z) \end{bmatrix} \begin{bmatrix} \nabla_{\theta_{i+}} \log p_\theta^i(z)^T & \nabla_{\theta_{i-}} \log p_\theta^i(z)^T \end{bmatrix} \Pi_i p_\theta^i(z) dz \\
&= \frac{n}{\sigma^2 + \theta_{i-}^T \Gamma \theta_{i-}} \Pi_i^T \begin{bmatrix} \widehat{\Sigma}_{i+} & \widehat{\Sigma}_{i+} \Sigma_{i+}^{-1} \Sigma_{i\pm} \\ \Sigma_{i\mp} \Sigma_{i+}^{-1} \widehat{\Sigma}_{i+} & \Sigma_{i\mp} \Sigma_{i+}^{-1} \widehat{\Sigma}_{i+} \Sigma_{i+}^{-1} \Sigma_{i\pm} \end{bmatrix} \Pi_i \\
&\qquad + \Pi_i^T \begin{bmatrix} 0 & 0 \\ 0 & \left(\frac{d_i^2 \beta_i^2}{n^2} + \frac{2\beta_i d_i^2}{n} + 2d_i + d_i^2 \right) \Gamma_{i-} \theta_{i-} \theta_{i-}^T \Gamma_{i-} \end{bmatrix} \Pi_i \\
&= \frac{n}{\sigma^2 + \theta_{i-}^T \Gamma \theta_{i-}} \left(Q_i + o_n(1) \right)
\end{aligned}
$$

The $o_n(1)$ term is due to strong law of large numbers. From this we know that, with probability 1,

$$
\lim_{c \to \infty} \lim_{n \to \infty} \sup_{\|h\|_2 < 1} \frac{\mathcal{I}^n(\theta_0 + ch/\sqrt{n})}{n} = \sum_{i=1}^m \frac{1}{\sigma^2 + \theta_{i-}^T \Gamma \theta_{i-}} Q_i =: \rho
$$

Applying Corollary D.1 with $\psi \mathbb{R}^d \to \mathbb{R}^d$ as the identity function $\psi(x) = x$ gives the full-feature minimax lower bound. Applying Corollary D.1 with $\psi \mathbb{R}^d \to \mathbb{R}^{d_i}$ as $\psi(x) = T_i x$ gives the missing-feature minimax lower bound.

D.2 Proof of Theorem 5.2

We will apply Corollary D.1 and apply it to two different choices of ψ to get the full feature minimax bound and missing feature minimax bound respectively. For notational simplicity, we will use P_θ in place of P_θ^y. P_θ is in the exponential family, so the conditions of Corollary D.1 are satisfied.

We begin by computing the Fisher Information. Let P_θ^j for $j \in [m]$ denote the distribution over $y_j \in \mathbb{R}^n$. Let $\mathcal{I}_i^n(\theta)$ denote the Fisher Information of $\mathbb{P}_\theta^i$, and let $\mathcal{I}^n(\theta) = \sum_{i=1}^m \mathcal{I}_i^n(\theta)$ denote the Fisher Information of P_θ.

Let $x_i^{(k)}, y_i^{(k)}$ be the kth sample from agent i. We will let $\mathcal{I}_i^{(k)}(\theta)$ be the fisher information of $y_i^{(k)}$. We know that $\mathcal{I}_i^n(\theta) = \sum_{k=1}^n \mathcal{I}_i^{(k)}(\theta)$ by independence. Some straightforward calculations tell us that $x_{i-}^{(k)}$ is distributed as $\mathsf{N}(\mu, \Gamma)$ where $\mu = \Sigma_{i\mp} \Sigma_{i+}^{-1} x_{i+}^{(k)}$ and $\Gamma = \Sigma_{i-} - \Sigma_{i\mp} \Sigma_{i+}^{-1} \Sigma_{i\pm}$. From this we can deduce that $\theta_{i-}^T x_{i-}^{(k)}$ is distributed as $\mathsf{N}(\mu^T \theta_{i-}, \theta_{i-}^T \Gamma \theta_{i-})$. And $y_i^{(k)}$ is distributed as $\mathsf{N}(\theta^T \gamma, \theta_{i-}^T \Gamma \theta_{i-} + \sigma^2)$ where $\gamma := \Pi_{i+}^T x_{i+}^{(k)} + \Pi_{i-}^T \mu$.

Let $\phi := \frac{z - \gamma^T \theta}{\sigma^2 + \theta_{i-}^T \Gamma \theta_{i-}}$ and $\Delta := \phi^2 - \frac{1}{\sigma^2 + \theta_{i-}^T \Gamma \theta_{i-}}$. Using p_θ^{ik} denote the density of $x_{i-}^{(k)}, y_i^{(k)}$, we can calculate the derivative of the log density

$$
\nabla_{\theta_{i+}} \log p_\theta^{ik}(z) = \frac{z - \theta_{i+}^T x_{i+}^{(k)} - \theta_{i-}^T \mu}{\sigma^2 + \theta_{i-} \Gamma \theta_{i-}} x_{i+}^{(k)} = \phi x_{i+}^{(k)}
$$

$$\nabla_{\theta_{i-}} \log p_\theta^{ik}(z) = \Delta\Gamma\theta_{i-} + \phi\mu.$$

Using the facts that $\mathbb{E}[\phi] = 0$, $\mathbb{E}[\phi^2] = \frac{1}{\sigma^2+\theta_{i-}\Gamma\theta_{i-}}$, $\mathbb{E}[\phi\Delta] = 0$, and $\mathbb{E}[\Delta^2] = \frac{2}{(\sigma^2+\theta_{i-}\Gamma\theta_{i-})^2}$, where the expectation is an integral over z, we have that

$$\begin{aligned}
\mathcal{I}_i^{(k)}(\theta) &= \int \nabla_\theta \log p_\theta^{ik}(z)[\nabla_\theta \log p_\theta^{ik}(z)]^T p_\theta^{ik}(z)dz \\
&= \int \Pi_i^T \begin{bmatrix} \nabla_{\theta_{i+}} \log p_\theta^{ik}(z) \\ \nabla_{\theta_{i-}} \log p_\theta^{ik}(z) \end{bmatrix} \begin{bmatrix} \nabla_{\theta_{i+}} \log p_\theta^{ik}(z)^T & \nabla_{\theta_{i-}} \log p_\theta^{ik}(z)^T \end{bmatrix} \Pi_i p_\theta^{ik}(z)dz \\
&= \Pi_i^T \begin{bmatrix} \mathbb{E}[\phi^2] x_{i+}^{(k)}(x_{i+}^{(k)})^T & \mathbb{E}[\phi x_{i+}^{(k)}(\Delta\Gamma\theta_{i-} + \phi\mu)^T] \\ \mathbb{E}[(\Delta\Gamma\theta_{i-} + \phi\mu)(\phi x_{i+}^{(k)})^T] & \mathbb{E}[(\Delta\Gamma\theta_{i-} + \phi\mu)(\Delta\Gamma\theta_{i-} + \phi\mu)^T] \end{bmatrix} \Pi_i \\
&= \frac{1}{\sigma^2 + \theta_{i-}\Gamma\theta_{i-}} \Pi_i^T \begin{bmatrix} x_{i+}^{(k)}(x_{i+}^{(k)})^T & x_{i+}^{(k)}\mu^T \\ \mu(x_{i+}^{(k)})^T & \mu\mu^T + \frac{2}{\sigma^2+\theta_{i-}^T\Gamma\theta_{i-}}\Gamma\theta_{i-}\theta_{i-}^T\Gamma \end{bmatrix} \Pi_i \\
&= \frac{1}{\sigma^2 + \theta_{i-}\Gamma\theta_{i-}} \Pi_i^T \begin{bmatrix} x_{i+}^{(k)}(x_{i+}^{(k)})^T & x_{i+}^{(k)}(x_{i+}^{(k)})^T\Sigma_{i+}^{-1}\Sigma_{i\pm} \\ \Sigma_{i\mp}\Sigma_{i+}^{-1}x_{i+}^{(k)}(x_{i+}^{(k)})^T & \Sigma_{i\mp}\Sigma_{i+}^{-1}x_{i+}^{(k)}(x_{i+}^{(k)})^T\Sigma_{i+}^{-1}\Sigma_{i\pm} + \frac{2}{\sigma^2+\theta_{i-}^T\Gamma\theta_{i-}}\Gamma\theta_{i-}\theta_{i-}^T\Gamma \end{bmatrix} \Pi_i.
\end{aligned}$$

From this we can sum over

$$\begin{aligned}
\mathcal{I}_i^n(\theta) &= \sum_{k=1}^n \mathcal{I}_i^{(k)}(\theta) \\
&= \frac{n}{\sigma^2 + \theta_{i-}^T\Gamma\theta_{i-}} \Pi_i^T \begin{bmatrix} \widehat{\Sigma}_{i+} & \widehat{\Sigma}_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm} \\ \Sigma_{i\mp}\Sigma_{i+}^{-1}\widehat{\Sigma}_{i+} & \Sigma_{i\mp}\Sigma_{i+}^{-1}\widehat{\Sigma}_{i+}\Sigma_{i+}^{-1}\Sigma_{i\pm} + \frac{2}{\sigma^2+\theta_{i-}^T\Gamma\theta_{i-}}\Gamma\theta_{i-}\theta_{i-}^T\Gamma \end{bmatrix} \Pi_i \\
&= \frac{n}{\sigma^2 + \theta_{i-}^T\Gamma\theta_{i-}} \left(Q_i + o_n(1) + \Pi_i^T \begin{bmatrix} 0 & 0 \\ 0 & \frac{2}{\sigma^2+\theta_{i-}^T\Gamma\theta_{i-}}\Gamma\theta_{i-}\theta_{i-}^T\Gamma \end{bmatrix} \Pi_i \right)
\end{aligned}$$

The $o_n(1)$ term is due to strong law of large numbers. From this we know that, with probability 1

$$\begin{aligned}
&\lim_{c\to\infty} \lim_{n\to\infty} \sup_{\|h\|_2<1} \frac{\mathcal{I}^n(\theta_0 + ch/\sqrt{n})}{n} \\
&\qquad = \sum_{i=1}^m \frac{1}{\sigma^2 + \theta_{i-}^T\Gamma\theta_{i-}} \left(Q_i + \Pi_i^T \begin{bmatrix} 0 & 0 \\ 0 & \frac{2}{\sigma^2+\theta_{i-}^T\Gamma\theta_{i-}}\Gamma\theta_{i-}\theta_{i-}^T\Gamma \end{bmatrix} \Pi_i \right) =: \rho
\end{aligned}$$

Applying Corollary D.1 with $\psi\mathbb{R}^d \to \mathbb{R}^d$ as the identity function $\psi(x) = x$ gives the full-feature minimax lower bound. Applying Corollary D.1 with $\psi\mathbb{R}^d \to \mathbb{R}^{d_i}$ as $\psi(x) = T_i x$ gives the missing-feature minimax lower bound.

One final transformation remains to get the form of this lower bound to match the one in the theorem statement. We know that from Cauchy-Schwartz that for all $u \in \mathbb{R}^{d-d_i}$

$$\frac{u^T\Gamma\theta_{i-}\theta_{i-}^T\Gamma u}{\theta_{i-}^T\Gamma\theta_{i-}} = \frac{(u^T\Gamma^{\frac{1}{2}}\Gamma^{\frac{1}{2}}\theta_{i-})^2}{\theta_{i-}^T\Gamma\theta_{i-}} \le u^T\Gamma u.$$

Using this fact and the definition of Γ and Q_i we have that

$$\frac{1}{\sigma^2 + \theta_{i-}^T\Gamma\theta_{i-}} \left(Q_i + \Pi_i^T \begin{bmatrix} 0 & 0 \\ 0 & \frac{2}{\sigma^2+\theta_{i-}^T\Gamma\theta_{i-}}\Gamma\theta_{i-}\theta_{i-}^T\Gamma \end{bmatrix} \Pi_i \right) \preceq \frac{2n}{\sigma^2 + \theta_{i-}^T\Gamma\theta_{i-}}\Sigma.$$

Using this bound gives our final result.

D.3 Proof of Corollary 5.3

The existence of the limits is a consequence of strong law of large numbers. To further show the inequality in the limit, we note that

$$\frac{1}{m}\sum_{i=1}^{m}\left(\frac{\Sigma}{\sigma^2+\theta_{i-}^T\Gamma_{i-}\theta_{i-}}-\frac{T_i^\top\Sigma_{i+}T_i}{\sigma^2+\theta_{i-}^T\Gamma_{i-}\theta_{i-}}\right)=\frac{1}{m}\sum_{i=1}^{m}\frac{\Pi_i^\top\begin{bmatrix}0&0\\0&\Gamma_{i-}\end{bmatrix}\Pi_i}{\sigma^2+\theta_{i-}^T\Gamma_{i-}\theta_{i-}}$$
$$\preceq\frac{1}{m}\sum_{i=1}^{m}\frac{\Pi_i^\top\begin{bmatrix}0&0\\0&\Gamma_{i-}\end{bmatrix}\Pi_i}{\sigma^2}\preceq\frac{1}{m}\sum_{i=1}^{m}\frac{\Pi_i^\top\begin{bmatrix}0&0\\0&\Sigma_{i-}\end{bmatrix}\Pi_i}{\sigma^2}\to\frac{p\mathrm{diag}(\Sigma)+p^2(\Sigma-\mathrm{diag}(\Sigma))}{\sigma^2},$$

where the last step holds with probability one by strong law of large numbers. This is true as by our random missing model, Σ_{ij} is not observed with probability p if $i=j$, and p^2 if $i\neq j$. We can further derive that

$$\frac{p\mathrm{diag}(\Sigma)+p^2(\Sigma-\mathrm{diag}(\Sigma))}{\sigma^2}\preceq\frac{p\lambda_1(\Sigma)I}{\sigma^2}\preceq\frac{p\kappa\Sigma}{\sigma^2}\preceq\frac{p\lambda_1(\Sigma)I}{\sigma^2}$$
$$\overset{\text{(i)}}{\preceq}\frac{p\kappa(\sigma^2+\|\theta\|_\Sigma^2)}{\sigma^2}\frac{1}{m}\sum_{i=1}^{m}\frac{\Sigma}{\sigma^2+\theta_{i-}^T\Gamma_{i-}\theta_{i-}}.$$

In (i), we make use of the fact that

$$\Sigma\succeq\Pi_i^\top\begin{bmatrix}0&0\\0&\Gamma_{i-}\end{bmatrix}\Pi_i$$

and therefore $\|\theta\|_\Sigma^2\geq\|\theta_{i-}\|_{\Gamma_{i-}}^2$. By our choice of $p\leq\frac{1}{2}\kappa^{-1}(1+\|\theta\|_\Sigma^2/\sigma^2)^{-1}$, we can conclude that

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\frac{T_i^\top\Sigma_{i+}T_i}{\sigma^2+\theta_{i-}^T\Gamma_{i-}\theta_{i-}}\succeq\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\frac{\Sigma/2}{\sigma^2+\theta_{i-}^T\Gamma_{i-}\theta_{i-}}.$$