Optimal Multitask Linear Regression and Contextual Bandits under Sparse Heterogeneity

Xinmeng Huang* Kan Xu† Donghwan Lee‡

Hamed Hassani§ Hamsa Bastani¶ Edgar Dobriban‖

July 16, 2024

Abstract

Large and complex datasets are often collected from several, possibly heterogeneous sources. Multitask learning methods improve efficiency by leveraging commonalities across datasets while accounting for possible differences among them. Here, we study multitask linear regression and contextual bandits under *sparse heterogeneity*, where the source/task-associated parameters are equal to a global parameter plus a sparse task-specific term. We propose a novel two-stage estimator called MOLAR that leverages this structure by first constructing a covariate-wise weighted median of the task-wise linear regression estimates and then shrinking the task-wise estimates towards the weighted median. Compared to task-wise least squares estimates, MOLAR improves the dependence of the estimation error on the data dimension. Extensions of MOLAR to generalized linear models and constructing confidence intervals are discussed in the paper. We then apply MOLAR to develop methods for sparsely heterogeneous multitask contextual bandits, obtaining improved regret guarantees over single-task bandit methods. We further show that our methods are minimax optimal by providing a number of lower bounds. Finally, we support the efficiency of our methods by performing experiments on both synthetic data and the PISA dataset on student educational outcomes from heterogeneous countries.

Keywords: Multitask Learning, Data Heterogeneity, Contextual Bandit, Minimax Optimality.

*Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania. xinmengh@sas.upenn.edu.

†Department of Information Systems, Arizona State University. kanxu1@asu.edu.

‡Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania. dh7401@sas.upenn.edu.

§Department of Electrical and Systems Engineering, University of Pennsylvania. hassani@seas.upenn.edu.

¶Department of Operations, Information, and Decisions, University of Pennsylvania. hamsab@wharton.upenn.edu.

‖Department of Statistics and Data Science, University of Pennsylvania. dobriban@wharton.upenn.edu.

1 Introduction

Large and complex datasets are often collected from multiple sources—such as from several locations—and with possibly varying data collection methods. This can result in both similarities and differences among the source-specific datasets. While some covariates have consistent effects on the response across all sources, others may have different effects on the response depending on the source. For instance, when predicting students' academic performance, the effects of socioeconomic status, language, and education policies may vary across regions. See Figure 1 for an illustration of the multi-country PISA educational attainment dataset (OECD, 2019), described in more detail in Section 4.2, where the regression coefficients of a few features vary strongly across countries (OECD, 2019). Ignoring this heterogeneity may introduce bias and lead to incorrect predictions and decisions. Similar situations arise in healthcare (Quinonero-Candela et al., 2008), demand prediction (Baardman et al., 2023; Van Herpen et al., 2012), and other areas. Therefore, instead of learning a shared model for all the data, it may help to develop models for each task.

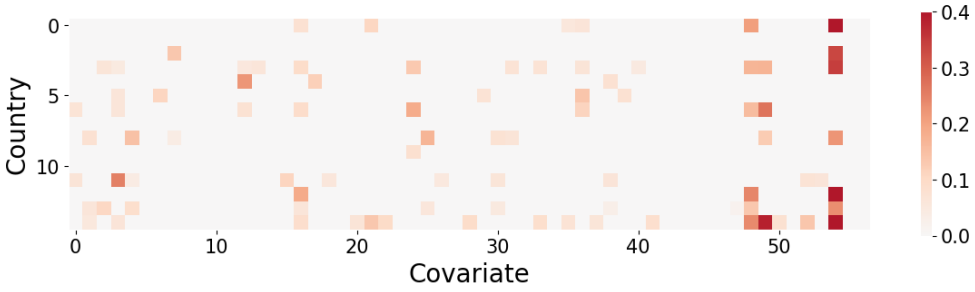

Figure 1: The differences in the least squares estimates of a measure of educational attainment in selected countries for the PISA dataset. See details in Appendix J.

Although complex datasets are often heterogeneous, commonalities also exist across data of the same type. Therefore, it may increase efficiency if we analyze them jointly. This is often referred to as *multitask* analysis. Common multitask methods regularize

the task-associated parameters via penalties (Evgeniou and Pontil, 2004; Evgeniou et al., 2005; Duan and Wang, 2022), as well as cluster or pool datasets based on their similarity (Ben-David et al., 2010; Crammer et al., 2008; Dobriban and Sheng, 2021). While these methods have demonstrated improvements by factors equal to the number of tasks, these methods can perform significantly worse than single-task learning when the heterogeneity across tasks is severe. To achieve larger theoretical improvements, Tripuraneni et al. (2021); Du et al. (2020); Collins et al. (2021) consider low-dimensional shared representations of task-specific models, while Lounici et al. (2009); Singh and Sharma (2020) constrain the task-specific parameters to be sparse with a common support set.

Often, only a small subset of covariates have different effects in different data sources or in predicting different responses. For instance, in the Expedia personalized recommendation dataset, Figure 2 of Bastani (2021) shows that out of 15 customer- and hotel-specific features, only the price has significantly different effects on predicting bookings and clicks. To capture this phenomenon in a broader multitask regime, we consider the following sparse heterogeneity structure. Given M tasks, each task m is associated with a parameter $\beta^{(m)}$, and $\{\beta^{(m)}\}_{m=1}^{M}$ differ only in a small number of coordinates. In particular, they have the form $\beta^{(m)} = \beta^{\star} + \delta^{(m)}$, for some unknown global parameter $\beta^{\star}$ representing the part of the parameters shared across tasks; and for unknown sparse parameters $\delta^{(m)}$—with few nonzero covariates—representing task-specific adjustments.

This structure has garnered interest in a number of prior works: Bastani (2021) combined a large proxy dataset and a small target dataset when the associated parameters differed in a sparse vector; Xu et al. (2021) considered group-wise sparse heterogeneity in matrix factorization of word embeddings; while Xu and Bastani (2021) proposed methods for multiple sparsely heterogeneous contextual bandits.

However, prior work leaves open a number of important problems. In particular, it is not clear what the statistically optimal methods are under sparse heterogeneity, even in linear regression. Prior work has shown that heterogeneity-aware methods improve the dimension-dependence of the estimation error rate, compared to heterogeneity-unaware or single-task methods. However, it remains unclear whether better methods exist. In this paper, we resolve this problem by establishing—to our knowledge—a new lower bound for estimating several models with sparse heterogeneity, in both linear regression and contextual bandits. We also propose novel methods and show that they achieve the lower bounds.

1.1 Contributions

We consider estimating M linear models under s-sparse heterogeneity, in both the offline and online scenarios. We highlight our contributions as follows:

- For linear regression, we propose the MOLAR method—Median-based Multitask Estimator for Linear Regression—to jointly learn heterogeneous models: MOLAR first applies a weighted covariate-wise median to source-specific least squares regression estimators to obtain an estimate shared across tasks, and then obtains estimates of individual tasks by shrinking the individual estimates towards the shared estimate via thresholding. We provide upper bounds on the estimation error for each individual task using MOLAR, showing the benefit of multiple tasks (large total sample size $n_{[M]}$) and the sparse heterogeneity (small s). For balanced datasets, our result improves the rates of pooled ordinary least squares (OLS) and the multitask method of Xu and Bastani (2021) by factors of d/s and $\sqrt{d/s}$, respectively. Beyond estimation in linear models, the applications of MOLAR to generalized linear models, and statistical inference for task-wise parameters are then discussed.

  We also provide matching minimax lower bounds in the multitask setting with sparse

heterogeneity, showing the optimality of MOLAR. Our result generalizes the lower bound for single-task linear regression to allow data points with different noise variances. The ℓ_1 error strengthens an existing lower bound for sparse linear regression (Raskutti et al., 2011).

- We formulate the asynchronous multitask setting for two types of contextual bandit problems where each bandit instance has a certain probability of observing a context and taking an action at any time. We use MOLAR in the multitask bandit setting, leading to improved individual regret bounds, where the scaling in the context dimension d from single-task methods is replaced by the level of heterogeneity. In addition, we provide new minimax lower bounds for asynchronous multitask bandit problems.

- We support our methods with experiments on both synthetic data and the PISA educational attainment dataset. Our empirical results support our theoretical findings, and show an improvement over existing methods.

1.2 Related Works

We review the most closely related works here. More literature is reviewed in Appendix A.

Data Heterogeneity. Complex datasets are often obtained by aggregating data from heterogeneous sources; which may correspond to subpopulations with unique characteristics (Fan et al., 2014; Meinshausen and Bühlmann, 2015; Marron, 2017). Data heterogeneity can reduce the performance of standard methods designed for independent and identically distributed (i.i.d.) data in statistical inference (Guo, 2020; Hu et al., 2022; Yuan et al., 2023) and various learning tasks (Zhao et al., 2016; Gu and Chen, 2022; McMahan et al., 2017). However, it is sometimes possible to mitigate the effects of heterogeneity (Luo et al., 2022a; Yang et al., 2020; Zhang and Wang, 2019; Chen et al., 2022a; Wang et al., 2019).

Multitask Linear Regression & Contextual Bandits. Studying multitask linear regression, Tripuraneni et al. (2021); Du et al. (2020) show improved generalization errors compared to the single-task OLS, by considering a low-dimensional shared representation. A similar result holds for personalized federated learning (Collins et al., 2021). Yang et al. (2019) assume group-wise heterogeneity of parameters and propose to regularize the least squares objective, without a finite-sample theoretical analysis.

Starting from Woodroofe (1979), literature on contextual bandits has developed vastly (see e.g., Sarkar, 1991; Yang and Zhu, 2002; Perchet and Rigollet, 2013; Chen et al., 2021, 2022b,c; Luo et al., 2022b, etc). Multitask contextual bandit methods include regularizing the bandit instance parameters, and pooling data from related bandit instances (see *e.g.*, Soare et al., 2014; Chu et al., 2011; Valko et al., 2013; Cesa-Bianchi et al., 2013; Deshmukh et al., 2017; Gentile et al., 2014, 2017). One can also impose a shared prior distribution over bandit instances (Cella et al., 2020; Kveton et al., 2021; Bastani et al., 2022). However, most resulting regret bounds for individual bandit instances can counter-intuitively deteriorate and can be worse than for single-bandit methods. Furthermore, some methods require bandit instances to appear sequentially to learn the prior (Lazaric et al., 2013).

Recently, Xu and Bastani (2021) propose methods with improved estimation error in linear regression and regret in contextual bandits, when tasks are sparsely heterogeneous. Our matching upper and lower bounds imply that their method is sub-optimal.

Transfer Learning. Transfer learning aims to boost the learning performance on a particular task (typically with a small dataset) when given data from other sources. A few works have studied transfer learning in linear regression (Li et al., 2020) and generalized linear models (Tian and Feng, 2022; Li et al., 2023), but mainly for parameters with ℓ_1-bounded heterogeneity.

1.3 Notation

We introduce necessary notations here and refer to the full definitions in Appendix A. We use $:=$ or $\triangleq$ to introduce definitions. For an integer $d \geq 1$, we write $[d]$ for $\{1, \ldots, d\}$. We use I_d to denote the $d \times d$ identity matrix. For a vector $v \in \mathbb{R}^d$, we denote its entries as $v_1, \ldots, v_d$. We also denote $\|v\|_p = (\sum_{k\in[d]} |v_k|^p)^{1/p}$ for all $p > 0$, with $\|v\|_0$ defined as the number of non-zero entries. For any $\mathcal{I} \subseteq [M]$, given weights $\{w_m\}_{m=1}^M$ (or sample sizes $\{n_m\}_{m=1}^M$), we write $W_\mathcal{I}$ as $\sum_{m\in\mathcal{I}} w_m$ and $n_\mathcal{I}$ for $\sum_{m\in\mathcal{I}} n_m$. For a matrix $A \in \mathbb{R}^{m\times n}$, we denote the (i, j)-th covariate of A by $[A]_{i,j}$ or $A_{i,j}$, and the i-th row (resp., the j-th column) by $A_{i,\cdot}$ (resp., $A_{\cdot,j}$). For two real numbers a and b, we write $a \vee b$ and $a \wedge b$ for $\max\{a, b\}$ and $\min\{a, b\}$, respectively. For an event E, we write $\mathbb{1}(E)$ for the indicator of the event. We use the Bachmann-Landau asymptotic notations $\Omega(\cdot)$, $\Theta(\cdot)$, $O(\cdot)$ to absorb constant factors, and use $\widetilde{\Omega}(\cdot)$, $\widetilde{O}(\cdot)$ to also absorb logarithmic factors. Furthermore, we use probabilistic notations such as $O_P(a_{\{n_m\}_{m=1}^M})$ to denote quantities that are bounded by $a_{\{n_m\}_{m=1}^M}$ with overwhelming probabilities as $\min_{m\in[M]} n_m \to \infty$. For a number $x \in \mathbb{R}$, we use $(x)_+$ to denote its non-negative part, *i.e.*, $x\mathbb{1}(x \geq 0)$.

2 Multitask Linear Regression

In this section, we study multitask linear regression under sparse heterogeneity with $M > 0$ tasks, each of which is associated with an unknown task-specific parameter $\beta^{(m)} \in \mathbb{R}^d$ for $m \in [M]$. For a covariate vector $x_i^{(m)}$ associated with task m, the outcome follows the linear model

$$y_i^{(m)} = \langle x_i^{(m)}, \beta^{(m)} \rangle + \varepsilon_i^{(m)},$$

where $\varepsilon_i^{(m)}$ is noise satisfying conditions specified below. We observe $n_m > 0$ i.i.d. vectors of covariates for each task m; the sample sizes n_m may vary with m. For each task m, we denote by $\mathbf{X}^{(m)} \in \mathbb{R}^{n_m\times d}$ the matrix whose rows are the observed features, and by $Y^{(m)} \in$

$\mathbb{R}^{n_m}$ the vector of corresponding observed outcomes. Our goal is to use $\{(\mathbf{X}^{(m)}, Y^{(m)})\}_{m=1}^M$ to estimate the parameter $\beta^{(m)}$ for each task m. We denote by $\mathbf{X}$ the $n_{[M]} \times d$ concatenation of all matrices $\{\mathbf{X}^{(m)}\}_{m=1}^M$, where $n_{[M]} \triangleq \sum_{m=1}^M n_m$ denotes the total sample size. We consider the following condition, which will enable us to pool information to improve estimation:

Condition 1 (Sparse heterogeneity)**.** *There exists an* unknown $s \in \{0, \ldots, d\}$ *and a common global parameter* $\beta^\star \in \mathbb{R}^d$ *such that for each* $m \in [M]$, $\|\beta^{(m)} - \beta^\star\|_0 \leq s$.

We remark that while a common upper bound s is assumed in Condition 1, most of our methods also adapt to the case where the ℓ_0-heterogeneity between $\beta^{(m)}$ and $\beta^\star$ differs across tasks *i.e.*, $\|\beta^{(m)} - \beta^\star\|_0 \leq s_m$, and results depend on the individual values of s_m as opposed to just their maximum; see Remark 2 for more details. The heterogeneity level s (or $\{s_m\}_{m=1}^M$) and the global parameter $\beta^\star \in \mathbb{R}^d$ may not be identifiable, but our results apply simultaneously to all possible choices of $\beta^\star$ and s (or $\{s_m\}_{m=1}^M$) satisfying Condition 1. We will also require several additional standard assumptions. In particular, in the main text, we consider Gaussian noise to simplify the analysis (Condition 2). We also provide similar results for *Orlicz-norm-bounded noise*, covering sub-Gaussian and sub-exponential noises, in Appendix D at the cost of an auxiliary high-order term in n in our rates.

Condition 2 (Gaussian noise)**.** *For each* $m \in [M]$*, and some* $\sigma_m \geq 0$*,* $m \in [M]$*, the noises* $\{\varepsilon_i^{(m)}\}_{i=1}^{n_m}$ *are i.i.d. random variables with* $\varepsilon_i^{(m)} \sim \mathcal{N}(0, \sigma_m^2)$.

We next introduce a customary condition on task-wise distributions.

Condition 3 (Covariate distribution)**.** *There are constants* $L \geq \mu > 0$ *such that for any* $m \in [M]$*,* $\{x_i^{(m)}\}_{i=1}^{n_m}$ *are the* L*-sub-Gaussian and independently distributed covariates with zero mean and covariance* $\Sigma^{(m)} \succeq \mu I_d$*, where* $\succeq$ *denotes the Loewner order.*

Throughout the paper, we focus on investigating the impact of sample sizes $\{n_m\}_{m=1}^M$, variances $\{\sigma_m^2\}_{m=1}^M$, dimension of covariates d and heterogeneity s, while omitting the dependence on the other problem characteristics, including μ and L. When leveraging heterogeneous datasets, the performance can be hampered by very small and noisy datasets (e.g., Akbani et al., 2004; Kotsiantis et al., 2006; Chawla, 2010, etc). To prove optimal estimation errors, we will require Condition 4 to mildly constrain the sample sizes of the input datasets.

Condition 4 (SAMPLE SIZE CONSTRAINT)**.** *We assume $n_m \geq c\ln(d)d$ for a sufficiently large constant c. For the case of homogeneous variances $\sigma_1^2 = \cdots = \sigma_M^2$, there is a constant $c_{\mathrm{s}} \geq 1$ such that $\ln((d/s) \wedge (n_{[M]}/\min_{m\in[M]} n_m))^2 \max_{m\in[M]} n_m \leq c_{\mathrm{s}} n_{[M]}$ and $\max_{m\in[M]} \sqrt{n_m} \sum_{m=1}^M \sqrt{n_m} \leq c_{\mathrm{s}} n_{[M]}$. For the case of different variances, we require the same inequalities to hold for the rescaled sample sizes $\tilde{n}_m \triangleq n_m/\sigma_m^2$.*

Condition 4 is satisfied by the ideal balanced case where $n_m = \Theta(n)$ and $\sigma_m = \Theta(\sigma)$ for some $n > 0$ and $\sigma > 0$, which implies $\max_{m\in[M]} \sqrt{n_m} \sum_{m=1}^M \sqrt{n_m} = O(n_{[M]})$ and $\ln(n_{[M]}/\min_{m\in[M]} n_m)^2 \max_{m\in[M]} n_m = O(\ln(M)^2 n_{[M]}/M)$, leading to $c_{\mathrm{s}} = O(1)$. However, Condition 4 also covers more skewed sample sizes, including the singly dominant case where $n_1 = \Theta(nM^c)$ with some $c \in [0, 1)$ and $n_m = \Theta(n)$ for $m \geq 1$.

2.1 Algorithm Overview

We now introduce the MOLAR method, which consists of two steps: *collaboration* and *covariate-wise shrinkage*. Here, MOLAR mainly concerns the case where $n_m \gg d$ for all $m \in [M]$ in the main text and a variant applicable to $n_m < d$ is provided in Appendix E. In the first step, we estimate the global parameter $\beta^\star$ via the covariate-wise weighted median of the OLS estimates $\{\widehat{\beta}_{\mathrm{ind}}^{(m)}\}_{m=1}^M$, where the weights are adjusted according to the sample sizes and noise variances of each task. Recall that a weighted median $\mathrm{WMed}(\{z_m\}_{m=1}^M; \{w_m\}_{m=1}^M)$

Algorithm 1 MOLAR: Weighted-median-based Multitask Linear Regressors

Input: $\{(\mathbf{X}^{(m)}, Y^{(m)})\}_{m=1}^M$, thresholds $\{\gamma_m\}_{m=1}^M$, weights $\{w_m\}_{m=1}^M$
for $m \in [M]$ **do**
  Let $\widehat{\beta}_{\text{ind}}^{(m)} = (\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}\mathbf{X}^{(m)\top}Y^{(m)}$ be the OLS estimator for dataset $(\mathbf{X}^{(m)}, Y^{(m)})$
end for
Let $\widehat{\beta}^\star = \text{WMed}(\{\widehat{\beta}_{\text{ind}}^{(m)}\}_{m=1}^M; \{w_m\}_{m=1}^M)$ be the covariate-wise weighted median
for $m \in [M]$ and $k \in [d]$ **do**
  /* Option I: hard thresholding */
  $\widehat{\beta}_{\text{MOLAR},k}^{(m)} = \widehat{\beta}_k^\star$ **if** $|\widehat{\beta}_k^\star - \widehat{\beta}_{\text{ind},k}^{(m)}| \le \gamma_m \sqrt{[(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}}$ **else** $\widehat{\beta}_{\text{ind},k}^{(m)}$
  /* Option II: soft thresholding */
  $\widehat{\beta}_{\text{MOLAR},k}^{(m)} = \widehat{\beta}_k^\star + \mathsf{SoftThresholding}(\widehat{\beta}_{\text{ind},k}^{(m)} - \widehat{\beta}_k^\star; \gamma_m \sqrt{[(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}})$
end for
Output: $\{\widehat{\beta}_{\text{MOLAR}}^{(m)}\}_{m=1}^M$

of the scalar variables $\{z_m\}_{m=1}^M$ and non-negative weights $\{w_m\}_{m=1}^M$ (not necessarily summing up to one) is defined as any of the minimizers to the function $z \mapsto \sum_{m=1}^M w_m|z - z_m|$. Scaling all weights $\{w_m\}_{m=1}^M$ by a common factor leads to the same weighted median. In particular, if $w_1 = \cdots = w_M$, the weighted median recovers the classical median and thus the weights can be omitted for clarity.

There are two key insights in this step. First, since the heterogeneity is sparse, for most coordinates, most OLS estimates are unbiased for $\beta^\star$. Hence, to estimate those covariates of the global parameter, we view the local estimates as potentially perturbed by outliers and leverage robust statistical methods to mitigate their heterogeneity-incurred influence. Second, the weighting mechanism allows OLS estimates from datasets with larger sizes and less noise to contribute more to the global estimate $\widehat{\beta}^\star$. Notably, our work also presents a novel non-asymptotic analysis of weighted medians, which can be of independent interest.

In the second step—covariate-wise shrinkage—we detect mismatched covariates between task-wise OLS estimates $\{\widehat{\beta}_{\text{ind}}^{(m)}\}_{m=1}^M$ and the global estimate $\widehat{\beta}^\star$. Recall that the global and task-wise parameters differ in only a few coordinates. For covariates $k \in [d]$ such that

$|\widehat{\beta}_k^\star - \widehat{\beta}_{\text{ind},k}^{(m)}|$ is below $\gamma_m\sqrt{v_k^{(m)}}$ for some threshold γ_m and $v_k^{(m)} \triangleq \sqrt{[(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}}$, we may expect that $\beta_k^{(m)} = \beta_k^\star$. Also, the estimate $\widehat{\beta}_k^\star$ of $\beta_k^\star$ can be more accurate than $\widehat{\beta}_k^{(m)}$ for $\beta_k^{(m)}$, as it is estimated collaboratively from multiple datasets. As a result, we can assign $\widehat{\beta}_k^\star$ as the final estimate $\widehat{\beta}_{\text{MOLAR},k}^{(m)}$ of $\beta_k^{(m)}$ if $|\widehat{\beta}_k^\star - \widehat{\beta}_{\text{ind},k}^{(m)}| \le \gamma_m\sqrt{v_k^{(m)}}$. On the other hand, for the covariates where the global and local parameters differ, *i.e.*, $\beta_k^{(m)} \neq \beta_k^\star$, the threshold is more likely to be exceeded. In this case, we keep the local estimate $\widehat{\beta}_{\text{ind},k}^{(m)}$ as $\widehat{\beta}_{\text{MOLAR},k}^{(m)}$. This can be viewed as shrinkage of $\widehat{\beta}_{\text{ind},k}^{(m)}$ towards $\widehat{\beta}_k^\star$ via hard thresholding. Alternatively, to allow a smooth transition, we can also conduct a soft thresholding step where the final estimate $\widehat{\beta}_{\text{MOLAR},k}^{(m)}$ shifts $\widehat{\beta}_{\text{ind},k}$ by $\gamma_m\sqrt{v_k^{(m)}}$ when $|\widehat{\beta}_k^\star - \widehat{\beta}_{\text{ind},k}^{(m)}| > \gamma_m\sqrt{v_k^{(m)}}$. The latter can be viewed as shrinkage via the soft thresholding operator $x \mapsto \mathsf{SoftThresholding}(x;\lambda) := \text{sign}(x)(|x|-\lambda)_+$ defined for any $x \in \mathbb{R}$ and $\lambda \ge 0$. The two options have the same theoretical guarantees but can differ slightly in practice.

While using two stages, our method does not require sample splitting. MOLAR requires neither the knowledge of the heterogeneity bounds $\|\beta^{(m)} - \beta^\star\|_0$ nor of the support sets of $\{\beta^{(m)} - \beta^\star\}_{m=1}^M$.

2.2 Analysis of Estimation Error under Gaussian Noise

In this subsection, we provide theoretical results for MOLAR for a Gaussian noise; more general noise is considered in the Appendix. The local OLS estimates $\widehat{\beta}_{\text{ind}}^{(m)}$ can be written as $\widehat{\beta}_{\text{ind}}^{(m)} = (\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}\mathbf{X}^{(m)\top}Y^{(m)} = \beta^{(m)} + (\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}\mathbf{X}^{(m)\top}\boldsymbol{\varepsilon}^{(m)}$, where $\boldsymbol{\varepsilon}^{(m)} \sim \mathcal{N}(0, \sigma_m^2 I_{n_m})$. Thus, denoting $[(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}$ as $v_k^{(m)}$ for any $k \in [d]$ and $m \in [M]$, we have

$$\widehat{\beta}_{\text{ind},k}^{(m)} \mid \mathbf{X}^{(m)} \sim \mathcal{N}(\beta_k^{(m)}, v_k^{(m)}\sigma_m^2). \tag{1}$$

For each $k \in [d]$, let $\mathcal{B}_k \triangleq \{m \in [M] : \beta_k^{(m)} \neq \beta_k^\star\}$ be the set of *unaligned tasks* at covariate k. For any $0 < \eta \le 1$, we define the set of *η-well-aligned covariates* as

$$\mathcal{I}_\eta := \left\{ k \in [d] : \sum_{m=1}^{M} \mathbb{1}(m \in \mathcal{B}_k) w_m < \eta W_{[M]} \right\}, \tag{2}$$

where $W_{[M]} \triangleq \sum_{m=1}^{M} w_m$. The η-well-aligned covariates are only used in the proof, not in our algorithm. At each covariate $k \in [d]$, for $m \in [M] \backslash \mathcal{B}_k$, by (1), $\widehat{\beta}^{(m)}_{\text{ind},k}$ is an unbiased estimate of $\beta^\star_k$. When $W_{\mathcal{B}_k}$ is relatively small compared to $W_{[M]}$, the set $\{\widehat{\beta}^{(m)}_{\text{ind},k}\}_{m \in \mathcal{B}_k}$ of estimates possibly biased for $\beta^\star_k$ have small weights. We will show that the weighted median can estimate $\beta^\star_k$ accurately, despite the biased subset.

We first bound the estimation error of $\widehat{\beta}^\star$ for η-well-aligned covariates. To this end, we need to characterize the estimation error of the weighted median of Gaussian inputs with non-identical means and variances, as shown in Appendix C.1. Applying this to each covariate $k \in \mathcal{I}_\eta$ with $\eta \leq 1/5$, we find the following bounds for the estimation error of $\widehat{\beta}^\star$.

Proposition 1 (Error bound for well-aligned coordinates). *Taking*[1] $w_m = n_m/\sigma_m^2$ *for* $m \in [M]$, *under Conditions 2, 3, and 4, for any* $0 < \eta \leq \frac{1}{5}$, $k \in \mathcal{I}_\eta$, *it holds with probability at least* $1 - O((\min_m \tilde{n}_M/\tilde{n}_{[M]}) \vee (s/d)) - O(Mde^{-c \min_m n_m})$ *that*

$$|\widehat{\beta}^\star_k - \beta^\star_k| = \widetilde{O}\left(\sigma_w \frac{W_{\mathcal{B}_k} + \left(\sum_{m \in \mathcal{B}_k^c} w_m^2\right)^{1/2}}{W_{[M]}} \right),$$

where $\sigma_w := W_{[M]}^{-1} \sum_{m=1}^{M} w_m (\sigma_m/\sqrt{n_m})$ *is the weighted average of standard deviations. In particular, in the regime of balanced variances where* $\sigma_m = \Theta(\sigma)$ *for some* $\sigma > 0$, *we have*

$$|\widehat{\beta}^\star_k - \beta^\star_k| = \widetilde{O}\left(\frac{n_{\mathcal{B}_k}\sigma}{n_{[M]}\sqrt{\max_{m \in [M]} n_m}} + \frac{\sigma}{\sqrt{n_{[M]}}} \right).$$

Remark 1 (Comparison with Xu and Bastani (2021)). *When* $n_m = \Theta(n)$ *and* $\sigma_m = \Theta(\sigma)$ *for some* $n > 0$ *and* $\sigma > 0$ *and all* $m \in [M]$, *Xu and Bastani (2021) estimate* $\beta^\star$ *through a*

[1] Throughout the paper, we assume $\{\sigma_m\}_{m=1}^M$ are known as they can be easily estimated using the OLS-based formula $\widehat{\sigma}_m^2 = \|Y^{(m)} - \mathbf{X}^{(m)}\widehat{\beta}^{(m)}_{\text{ind}}\|_2^2/(n_m - d)$. Experiments with estimated variances can be found in Appendix J.3.1.

trimmed mean of the individual OLS estimates. Then, a Lasso-based shrinkage is used to estimate the local parameters $\{\beta^{(m)}\}_{m=1}^{M}$. *The Lasso step is more computationally expensive than our covariate-wise procedures. Moreover, the trimmed mean is less effective than the median in handling sparse heterogeneity. First, setting the fraction of the data trimmed* ω, *taken as* $\sqrt{s/d}$ *in Xu and Bastani (2021, Corollary 1), requires knowing the sparse heterogeneity level* s. *In contrast, the weighted median is applicable to all potential values of* s. *Second, using a single trimming proportion* ω *is suboptimal, as trimming fewer local estimates for covariates with more aligned tasks can improve accuracy. The best choice of* $\omega = \sqrt{s/d}$ *yields* $\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 = \widetilde{O}_P((d/\sqrt{M} + \sqrt{sd})\sigma/\sqrt{n})$, *which is larger than the minimax optimal rate by a factor of* $\sqrt{d/s}$ *(see Theorem 1 and 2 for optimality).*

Proposition 1 provides an upper bound relating to the weighted frequency $W_{\mathcal{B}_k}/W_{[M]}$ of misalignment. This result cannot be obtained by directly applying concentration inequalities to the estimates, due to heterogeneity. Instead, we analyze the concentration of the empirical weighted $(1/2 \pm W_{\mathcal{B}_k}/W_{[M]})$-quantiles to mitigate the influence of heterogeneity. The constant $1/5$, which restricts the heterogeneity, is not essential and is chosen for clarity. With more cumbersome calculations, it can be replaced with any number below $1/2$.

While estimating $\beta^\star$ is not our main goal, based on Proposition 1, we readily obtain the following bound for estimating $\beta^\star$ by choosing $\eta = \max_{k\in[d]} W_{\mathcal{B}_k}/W_{[M]}$, so $[d] = \mathcal{I}_\eta$, and summing up the errors over all covariates. Noting that $\sum_{k\in[d]} W_{\mathcal{B}_k}/W_{[M]} = s/d$, Corollary 1 reveals that the global parameter can be accurately estimated if heterogeneity happens roughly uniformly across all covariates.

Corollary 1 (Error bound for global parameter). *Taking* $w_m = n_m/\sigma_m^2$ *for all* $m \in [M]$, *under Conditions 2, 3, and 4, let* $\widehat{\beta}^\star$ *be obtained from Algorithm 1 and suppose* $W_{\mathcal{B}_k}/W_{[M]} = O(s/d)$ *for any* $k \in [d]$. *For any* $p \in \{1, 2\}$, *it holds with probability* $1 -$

$O((\tilde{n}_M/\tilde{n}_{[M]}) \vee (s/d)) - O(Mde^{-c\min_m n_m})$ *that*

$$\|\widehat{\beta}^\star - \beta^\star\|_p^p = \widetilde{O}\left(\sigma_w^p\left(\frac{s^p}{d^{p-1}} + \frac{d\left(\sum_{m=1}^M w_m^2\right)^{1/2}}{W_{[M]}}\right)\right),$$

where $\sigma_w := W_{[M]}^{-1}\sum_{m=1}^M w_m(\sigma_m/\sqrt{n_m})$ *. In particular, in the regime of balanced variances where* $\sigma_m^2 = \Theta(\sigma^2)$ *for some* $\sigma > 0$*, we have*

$$\|\widehat{\beta}^\star - \beta^\star\|_p^p = \widetilde{O}\left(\frac{s^p\sigma^p}{\max_{m\in[M]} n_m^{p/2} d^{p-1}} + \frac{d\sigma^p}{n_{[M]}^{p/2}}\right).$$

Proposition 1 above shows that the coefficients of the well-aligned covariates are accurately estimated. Thus one can use the global estimate $\widehat{\beta}^\star$ for the well-aligned covariates where the global parameter $\beta^\star$ also aligns with the local parameter $\beta^{(m)}$. The coefficients of the remaining covariates, which are either poorly aligned or do not satisfy $\beta_k^\star = \beta_k^{(m)}$, could be estimated by individual OLS estimates. In Theorem 1 below, we argue that with high probability, properly chosen thresholds achieve this. The proof is in Appendix C.4.

Theorem 1 (Error bound for task-wise parameters)**.** *Under Conditions 1, 2, 3 and 4, taking* $\gamma_m = c_1\sqrt{\ln((n_{[M]}/n_m)\wedge d)}\sigma_m$ *for all* $m \in [M]$ *with* $c_1 \geq 1$ *being constant, with* $\widehat{\beta}^{(m)}_{\text{MOLAR}}$ *from Algorithm 1 using either Option I or II, it holds for any* $p \in \{1, 2\}$*,* $m \in [M]$ *that*

$$\|\widehat{\beta}^{(m)}_{\text{MOLAR}} - \beta^{(m)}\|_p^p = \widetilde{O}_P\left(\frac{s\sigma_m^p}{n_m^{p/2}} + \frac{d}{\left(\sum_{m=1}^M n_m/\sigma_m^2\right)^{p/2}}\right). \tag{3}$$

In particular, in the regime of balanced variances where $\sigma_m^2 = \Theta(\sigma^2)$ *for some* $\sigma > 0$*, we have*

$$\|\widehat{\beta}^{(m)}_{\text{MOLAR}} - \beta^{(m)}\|_p^p = \widetilde{O}_P\left(\frac{s\sigma^p}{n_m^{p/2}} + \frac{d\sigma^p}{n_{[M]}^{p/2}}\right). \tag{4}$$

Remark 2 (Varying heterogeneity levels)**.** *While we assume the task-wise hetero-*

geneity is constrained by a common parameter s in Condition 1 for simplicity, we remark that similar theoretical results hold under varying heterogeneity levels where $\|\beta^{(m)} - \beta^\star\|_0 \le s_m$ for some $s_m \in \{0, 1, \ldots, d\}$ and all $m \in [M]$. Considering $\sigma_m = \Theta(\sigma)$ for illustration, where we have $\sum_{k\in[d]} W_{\mathcal{B}_k}/W_{[M]} \asymp \sum_{k\in[d]} n_m s_m / n_{[M]} =: \bar{s}_w$, Our theory implies

$$\|\widehat{\beta}^{(m)}_{\mathrm{MOLAR}} - \beta^{(m)}\|_p^p = \widetilde{O}_P\left(\frac{s_m \vee \bar{s}_w \sigma^p}{n_m^{p/2}} + \frac{d\sigma^p}{n_{[M]}^{p/2}}\right).$$

Therefore, a few outlier tasks with large $\|\beta^{(m)} - \beta^\star\|_0$ (at most d) do not heavily impact the ultimate task-wise estimation errors in MOLAR, revealing its robustness.

Remark 3 (Extensions of MOLAR)**.** *MOLAR can be extended beyond parameter estimation in linear models. For example, MOLAR is readily extended to generalized linear models by adjusting the individual maximum likelihood estimates (MLEs) $\{\widehat{\beta}^{(m)}_{\mathrm{ind}}\}_{m=1}^M$ and adjusting the data matrix $\mathbf{X}^{(m)\top}\mathbf{X}^{(m)}$ with the inverse link function. A similar guarantee can be established for sufficiently large sample sizes thanks to the asymptotic normality of task-wise MLEs. For details, we refer to Appendix G.*

In addition, one can construct confidence intervals for task-wise parameters $\{\beta^{(m)}\}_{m=1}^M$ by leveraging the improved concentration of $\widehat{\beta}^\star_k$ for well-aligned covariates k. These can have shorter lengths than the canonical single-task OLS intervals; see Appendix F.

The upper bound in (4) consists of two terms. Taking $p = 2$ for illustration, the first term—s/n_m—is independent of the dimension d, and is a factor d/s smaller than the minimax optimal rate d/n_m of estimation in a single linear regression task (Lehmann and Casella, 1998). Meanwhile, the second, dimension-dependent, term—$d/n_{[M]}$—is $n_{[M]}/n_m$ times smaller than d/n, since it depends on the *total sample size* $n_{[M]}$ used collaboratively. This brings a significant benefit under sparse heterogeneity, *i.e.*, when $s \ll d$. Therefore,

Table 1: Bounds on the estimation error $\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1$ of various methods under balanced variances *i.e.*, $\sigma_m^2 = \Theta(\sigma^2)$ for all $m \in [M]$: $\beta^{(m)}$ is the ground-truth parameter, $\widehat{\beta}^{(m)}$ is the estimator, $\delta^{(m)} = \beta^{(m)} - \beta^\star$ is a non-vanishing measure of heterogeneity. The standard regime shows the results for balanced datasets, *i.e.*, $n_m = \Theta(n)$ for all $m \in [M]$. The data-poor regime shows the results for transfer learning in the $(M+1)$-th task with a potentially small dataset: notably, MOLAR and the minimax lower bound requires $n_{M+1} = O(n_{[M]}(s/d)^2)$ while the others require $n_{M+1} = O(\min_{m\in[M]} n_m/d^2)$. See Xu and Bastani (2021, Sec 3.6) for more details about the baseline methods. Numerical constants and logarithmic factors are omitted for clarity.

Method	Standard Regime	Data-poor Regime
Individual OLS (Lehmann and Casella, 1998)	$\sigma d/\sqrt{n}$	$\sigma d/\sqrt{n_{M+1}}$
Individual LASSO (Tibshirani, 1996)	$\sigma d/\sqrt{n}$	$\sigma d/\sqrt{n_{M+1}}$
Global OLS (Dobriban and Sheng, 2021)	$\|\delta^{(m)}\|_1 + \sigma d/\sqrt{Mn}$	$\|\delta^{(M+1)}\|_1 + \sigma/\sqrt{n_{[M]}}$
Robust Multitask (Xu and Bastani, 2021)	$\sigma\sqrt{sd}/\sqrt{n} + \sigma d/\sqrt{Mn}$	$\sigma s/\sqrt{n_{M+1}}$
MOLAR (Theorem 1)	$\sigma s/\sqrt{n} + \sigma d/\sqrt{Mn}$	$\sigma s/\sqrt{n_{M+1}}$
Lower Bound (Theorem 2)	$\sigma s/\sqrt{n} + \sigma d/\sqrt{Mn}$	$\sigma s/\sqrt{n_{M+1}}$

our method has a factor of $\min\{n_{[M]}/n_m, d/s\}$ improvement in accuracy, compared to the optimal rate for a single linear regression task.

Algorithm 1 is applicable to any value of heterogeneity level s. In particular, when the heterogeneity is dense, *i.e.*, $s = \Omega(d)$, our result recovers the optimal estimation error rate d/n_m in single-task linear regression. Therefore, the collaboration mechanism in Algorithm 1 does not harm the rate, regardless of the level of heterogeneity.

Theorem 1 also implies that the collaborative estimator $\widehat{\beta}^\star$ is useful in transfer learning (Tian and Feng, 2022; Li et al., 2020), where, given a number of "source" tasks with large samples, the goal is to maximize performance on a specific "target" task with a small sample.

Corollary 2 (Transfer learning for a data-poor task)**.** *In the regime of balanced variances, when a new $(M+1)$-th task has a small dataset with $n_{M+1} = O(n_{[M]}(s/d)^{2/p})$, the ℓ_1 and ℓ_2 estimation errors for the task parameter using MOLAR are $\widetilde{O}_P(\sigma s/\sqrt{n_{M+1}})$ and $\widetilde{O}_P(\sigma^2 s/n_{M+1})$, respectively.*

The rates in Corollary 2 do not depend on the feature dimension d; in contrast to a linear dependence for the individual OLS estimate. In Table 1, we compare MOLAR with

the method from Xu and Bastani (2021) and other baseline methods, in terms of the ℓ_1 estimation error, in both the standard regime where $n_1 = \cdots = n_M = n$ and the data-poor regime where n_{M+1} is small. We find that MOLAR outperforms all methods compared in both regimes.

2.3 Lower Bound

To complement our upper bounds, we provide minimax lower bounds for our multitask linear regression task under sparse heterogeneity. We consider the best estimators $\{\widehat{\beta}^{(m)}\}_{m=1}^M$ that leverage heterogeneous datasets, in the worst-case sense over all global parameters $\beta^\star \in \mathbb{R}^d$, task-wise parameters $\{\beta^{(m)}\}_{m=1}^M$ each in $\mathbb{B}_s(\beta^\star) := \{\beta \in \mathbb{R}^d : \|\beta - \beta^\star\|_0 \leq s\}$, as well as covariance matrices $\{\Sigma^{(m)}\}_{m=1}^M$ each in $\mathcal{A}_{\mu,L} \triangleq \{\Sigma \in \mathbb{R}^{d\times d} : \Sigma \text{ positive semi-definite}, \mu I_d \preceq \Sigma \preceq L I_d\}$:

$$\mathcal{M}(s, d, \mu, L, m, p, \{n_m\}_{m=1}^M\}, \{\sigma_m\}_{m=1}^M) := \inf_{\widehat{\beta}^{(m)}} \sup_{\substack{\beta^\star \in \mathbb{R}^d,\ \{\beta^{(m)}\}_{m=1}^M \subseteq \mathbb{B}_s(\beta^\star) \\ \{\Sigma^{(m)}\}_{m=1}^M \subseteq \mathcal{A}_{\mu,L}}} \|\widehat{\beta}^{(m)} - \beta^{(m)}\|_p^p. \quad (5)$$

The minimax risk (5) characterizes the best possible worst-case estimation error under our heterogeneous multitask learning model.

Since the supremum is taken over all parameters $\{\beta^{(m)}\}_{m=1}^M \subseteq \mathbb{R}^d$ such that $\|\beta^{(m)} - \beta^\star\|_0 \leq s$ for some $\beta^\star \in \mathbb{R}^d$, we can consider two representative cases. First, the *homogeneous* case where $\beta^{(1)} = \cdots = \beta^{(m)} = \beta^\star$; which reduces to a single linear regression task with $n_{[M]}$ data points and varying noise variances, leading to a lower bound $\Omega_P(d/(\sum_{m=1}^M n_m/\sigma_m^2)^{p/2})$ for the minimax estimation error. Second, in the independent *s-sparse* case where $\beta^\star = 0$ and $\|\beta^{(m)}\|_0 \leq s$ for all $m \in [M]$, clearly $\|\beta^{(m)} - \beta^\star\|_0 \leq s$ and thus $\{\beta^{(m)}\}_{m=1}^M \subseteq \mathbb{B}_s(\beta^\star)$. By constructing independent priors for $\{\beta^{(m)}\}_{m=1}^M$, we show that only data points sampled from the model with parameter $\beta^{(m)}$ are informative for

estimating $\beta^{(m)}$. We then show a lower bound of $\Omega(s\sigma_m^p/n_m^{p/2})$ for the minimax risk. In particular, our lower bound for the ℓ_1 error strengthens the existing lower bound of order $\Omega(s^{1/2}\sigma_m/n_m^{1/2})$ for sparse linear regression (Raskutti et al., 2011). Combining the two cases, we prove the following lower bound. The formal proof is in Appendix C.5.

Theorem 2 (Minimax lower bound for linear regression under sparse heterogeneity)**.** *For any $m \in [M]$, $p \in \{1, 2\}$, it holds that*

$$\mathcal{M}(s, d, \mu, L, m, p, \{n_m\}_{m=1}^M\}, \{\sigma_m\}_{m=1}^M) = \widetilde{\Omega}_P \left(\frac{s\sigma_m^p}{n_m^{p/2}} + \frac{d}{\left(\sum_{m=1}^M n_m/\sigma_m^2\right)^{p/2}} \right).$$

Theorems 2 and 1 imply that MOLAR is minimax optimal up to logarithmic terms.

3 Linear Contextual Bandit

In this section, we study multitask linear contextual bandits as an application of our MOLAR method. A contextual bandit problem (Woodroofe, 1979) consists of data collection over several rounds $t = 1, \ldots, T$, where $T > 0$ is referred to as the time horizon. At each round, an analyst observes a context—covariate—vector, and based on all previous observations, chooses one of K options—referred to as arms—observing the associated reward. The goal is to minimize the *regret* compared to the best possible choices of arms. In linear contextual bandits, two settings are widely studied.

In the first setup, referred to as Model-C by Ren and Zhou (2023), at each round $t \in [T]$, the analyst first observes a set of K d-dimensional contexts $\{x_{t,a}\}_{a\in[K]}$ in which $x_{t,a}$ is associated with arm a. If the analyst selects action $a \in [K]$, then a reward $y_t = \langle x_{t,a}, \beta\rangle + \varepsilon_t$ is earned, where $\beta \in \mathbb{R}^d$ is the unknown parameter vector associated with the bandit and $\{\varepsilon_t\}_{t=0}^\infty$ is a sequence of i.i.d. noise variables. In the second setup, referred to as Model-P by Ren and Zhou (2023), at each round $t \in [T]$, the analyst instead only observes a single

d-dimensional context x_t. Then if action $a \in [K]$ is chosen, the analyst earns the reward $y_t = \langle x_t, \beta_a \rangle + \varepsilon_t$, where $\beta_a \in \mathbb{R}^d$ is the unknown parameter vector associated with arm a and $\{\varepsilon_t\}_{t=0}^{\infty}$ is still a sequence of i.i.d. noise variables.

Both models have been studied: for instance Model-C in Han et al. (2020); Oh et al. (2021); Ren and Zhou (2023) and Model-P in Bastani and Bayati (2020); Bastani et al. (2021). We extend these models to the multitask setting. We present Model-C here, and state a parallel set of results under Model-P in Appendix I. We consider M K-armed bandit instances, where the m-th is associated with a parameter $\beta^{(m)} \in \mathbb{R}^d$. Each bandit $m \in [M]$ has an activation probability $p_m \in [0, 1]$. At each round t within the time horizon T, each bandit m is independently activated with probability p_m; and we observe contexts for the activated bandits. The parameters $\{p_m\}_{m=1}^{M}$ model the frequency of receiving contexts. When $p_1 = \cdots = p_m = 1$, contexts for all bandit instances are always observed. We denote $\sum_{m \in \mathcal{I}} p_m$ as $p_{\mathcal{I}}$ for any $\mathcal{I} \subseteq [M]$. Without loss of generality, we assume $p_1 \geq \cdots \geq p_M$.

The analyst observes a set of K d-dimensional contexts—covariates, feature vectors—$\{x_{t,a}^{(m)}\}_{a \in [K]}$ for each activated bandit instance $m \in \mathcal{S}_t$ in the set $\mathcal{S}_t$ of activated bandit instances at time t. The contexts $\{x_{t,a}^{(m)}\}_{a \in [K]}$, for each $m \in \mathcal{S}_t$ in each round are sampled independently. Using the observed contexts, and all previously observed data, the analyst selects actions $a_t^{(m)} \in [K]$ for each activated bandit instance $m \in \mathcal{S}_t$ and earns the reward $y_t^{(m)} = \langle x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)} \rangle + \varepsilon_t^{(m)} \in \mathbb{R}$, where $\varepsilon_t^{(m)}$ are, for $t \in [T], m \in \mathcal{S}_t$, i.i.d. noise variables. To apply MOLAR in this setting, we require Condition 5 over all bandit parameters. Compared to Condition 1 for linear regression, this requires all parameters to have a bounded ℓ_2 norm, as is common in the area (see *e.g.*, Han et al. (2020); Bastani and Bayati (2020)).

Condition 5 (Sparse heterogeneity & Boundedness)**.** *There is an unknown global parameter $\beta^\star \in \mathbb{R}^d$ and value $s \in \{0, 1, \ldots, d\}$ such that $\|\beta^{(m)} - \beta^\star\|_0 \leq s$ for any $m \in [M]$.*

Further, $\|\beta^{(m)}\|_2 \leq 1$ *for all* $m \in [M]$.

As before, our analysis applies to each $\beta^\star, s$ for which the condition holds.

Condition 6 (Frequency constraint)**.** *We have* $\min_{m\in[M]}(p_1 \vee (p_{[M]}/m))/p_m \leq c_{\mathrm{f}}$ *for an absolute constant* c_{f}*, where* $p_{[M]} \triangleq \sum_{m=1}^{M} p_m$.

Following previous literature studying Model-C, we state other standard conditions as follows. The Gaussian noise Condition 7 is used only for simplicity, as our results can be extended to more general noise with a bounded Orlicz norm by leveraging the results of the offline linear regression with general noise in Appendix D.

Condition 7 (Gaussian noise)**.** *The noise variables* $\{\varepsilon_t^{(m)}\}_{t=1}^{\infty}$ *are i.i.d.* $\mathcal{N}(0,1)$ *variables.*

Next, we require the contexts to be sub-Gaussian. Since bounded contexts are sub-Gaussian, Condition 8 is weaker than the assumption of bounded contexts commonly used in the literature on contextual bandits (Xu and Bastani, 2021; Bastani et al., 2021; Bastani and Bayati, 2020; Kim and Paik, 2019).

Condition 8 (Sub-Gaussianity)**.** *There is* $L \geq 0$*, such that for each* $m \in [M]$*,* $t \in [T]$*, and* $a \in [K]$*,* $x_{t,a}^{(m)}$ *is* L*-sub-Gaussian.*

We next consider Condition 9, which ensures sufficient exploration even with a greedy algorithm (Ren and Zhou, 2023).

Condition 9 (Diverse context)**.** *There are positive constants* μ *and* c_x *such that for any* $\beta \in \mathbb{R}^d$*, vector* $v \in \mathbb{R}^d$*, and* $m \in [M]$*, it holds that* $\mathbb{P}(\langle x_{t,a^\star}^{(m)}, v\rangle^2 \geq \mu) \geq c_x$ *where* $a^\star = \operatorname*{argmax}_{a\in[K]} \langle x_{t,a}^{(m)}, \beta\rangle$ *and the probability is taken over the joint distribution of* $\{x_{t,a}\}_{a=1}^{K}$.

Remark 4. *The "diverse context" condition, which is widely used in the literature on bandits and reinforcement learning, simplifies the algorithmic design and the corresponding proof.*

The condition or equivalent settings have been used in e.g.*, Bastani et al. (2021, Lemma 1), Oh et al. (2021, Condition 4), Cella et al. (2022, Condition 2), Hao et al. (2021, Definition 3.1), Chakraborty et al. (2023, Condition 2.2 (c)), Han et al. (2020). Condition 9 encompasses many classical distributions,* e.g.*, Gaussian contexts where* $x_{t,a} \sim \mathcal{N}(0, \Sigma)$ *with* $\Sigma \succeq 16\mu I_d$ *for each* $a \in [K]$*, and sub-Gaussian distributions. See Ren and Zhou (2023, Section 2.3) for more discussion. In Conditions 8 and 9, we do not require* $\{x_{t,a}^{(m)}\}_{a=1}^{K}$ *to be independent across actions* $a \in [K]$*.*

On the other hand, we remark that the condition is made only for simplicity and to show that MOLAR can be applied in an important online setting. One can readily extend MOLARB to other bandit algorithms with exploration phases (see e.g.*, Bastani and Bayati (2020); Xu and Bastani (2021)).*

3.1 Algorithm Overview

We now introduce our MOLARB algorithm, see Algorithm 2. MOLARB manages multiple bandit instances in a batched way, as in Han et al. (2020); Ren and Zhou (2023). We split the time horizon T into batches that double in length, *i.e.*, $|\mathcal{H}_q| = 2^{q-1}|\mathcal{H}_0|$ for all $q \geq 1$, yielding $Q = O(\log_2(T/|\mathcal{H}_0|))$ batches. Within each batch $\mathcal{H}_q$, each bandit instance m leverages the current estimate $\widehat{\beta}_q^{(m)}$ of the parameter $\beta^{(m)}$ without further exploration. These estimates are updated at the end of a batch, based on all previous observations.

Compared to existing methods, MOLARB has two novel features: *novel estimates* and *fine-grained collaboration*. First, in the single-bandit regime, the estimate $\widehat{\beta}^{(m)}$ is often obtained by using OLS (Goldenshluger and Zeevi, 2013), LASSO (Bastani and Bayati, 2020), and ridge regression (Han et al., 2020). In the multi-bandit regime, we use our MOLAR estimators to improve accuracy based on all instances. If used at each time step, MOLAR induces strong correlations between the observations, as well as between

Algorithm 2 MOLARBandit: Multitask Bandits with MOLAR estimates

Input: Time horizon T, $\widehat{\beta}_{-1}^{(m)} = 0$, $\mathbf{X}_q^{(m)} = \emptyset$, and $Y_q^{(m)} = \emptyset$ for $m \in [M]$, initial batch size H_0 and batch $\mathcal{H}_0 = [H_0]$, number of batches $Q = \lceil \log_2(T/H_0) \rceil$

Define batches $\mathcal{H}_q = \{t : 2^{q-1}H_0 < t \leq \min\{2^q H_0, T\}\}$, for $q = 1, \ldots, Q$

for $t = 1, \cdots, T$ **do**

  for each bandit in parallel **do**

    if $t \in \mathcal{H}_q$ **and** bandit instance m is activated **then**

      Choose $a_t^{(m)} = \arg\max_{a \in [K]} \langle x_{t,a}^{(m)}, \widehat{\beta}_{q-1}^{(m)} \rangle$, breaking ties randomly, and gain reward $y_t^{(m)}$

      Augment observations $\mathbf{X}_q^{(m)} \leftarrow [\mathbf{X}_q^{(m)\top}, x_{t,a_t^{(m)}}^{(m)}]^\top$ and $Y_q^{(m)} \leftarrow [Y_q^{(m)\top}, y_t^{(m)}]^\top$

    end if

  end for

  if $t = 2^q H_0$, *i.e.*, batch $\mathcal{H}_q$ ends **then**

    Let $n_{m,q} = |Y_q^{(m)}|$ and $\mathcal{C}_q = \{m \in [M] : n_{m,q} \geq 2C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))\}$

    Call MOLAR($\{(\mathbf{X}_q^{(m)}, Y_q^{(m)})\}_{m \in \mathcal{C}_q}$) to obtain $\{\widehat{\beta}_q^{(m)}\}_{m \in \mathcal{C}_q}$

    for $m \in [M] \backslash \mathcal{C}_q$ **do**

      Let $\widehat{\beta}_q^{(m)} = \widehat{\beta}_{q-1}^{(m)}$, $\mathbf{X}_{q+1}^{(m)} = \mathbf{X}_q^m$, and $Y_{q+1}^{(m)} = Y_q^m$

    end for

  end if

end for

the estimates $\{\widehat{\beta}_{\mathrm{ind}}^{(m)}\}_{m=1}^M$ from Algorithm 1, across all instances. This makes the median potentially inaccurate (see Lemma C.2). As a remedy, batching ensures that our estimates in the current batch are independent conditional on the observations from previous batches. Batching also reduces computational costs.

Second, as indicated by Condition 3 from Section 2.2, MOLAR requires the eigenvalues of the non-centered empirical covariance matrices of the datasets in the collaboration to be lower bounded. We will show that this holds with high probability even when the arms are adaptively chosen (Lemma H.1 in Appendix H.1) when the sample size $n_{m,q}$ is of order $\widetilde{\Omega}(d)$. However, this may fail within a small batch $\mathcal{H}_q$ for bandit instances that rarely observe contexts because of their small activation probabilities, so $m \notin \mathcal{C}_q$, with

$$\mathcal{C}_q = \{m \in [M] : n_{m,q} \geq 2C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))\},$$

for C_{b} defined in Lemma H.1. Thus we neither involve these instances in MOLAR, nor update their parameter estimates until entering a large batch. In these instances, the observations are not used to update estimates, and are merged into future batches.

3.2 Regret Analysis

Due to the differences in the activation probabilities $\{p_m\}_{m=1}^M$, the number of observed contexts and the regret of each bandit instance can vary greatly. Therefore, we consider the following form of individual regret: given a time horizon $T \geq 1$ and a specific algorithm A that produces action trajectories $\{a_t^{(m)}\}_{t\in[T],m\in[M]}$, we define the cumulative regret for each instance $m \in [M]$ as

$$R_T^{(m)}(A) := \sum_{t=1}^{T} \mathbb{E}\left[\max_{a\in[K]} \langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)}\rangle \mathbb{1}(m \in \mathcal{S}_t)\right]$$

where $\mathcal{S}_t$ is the random set of activated bandits at time t.

After showing that the empirical covariance matrices are well-conditioned with high probability at the end of each batch in Lemma H.1, we leverage results from Section 2.2 to show that the MOLAR estimates $\{\widehat{\beta}_q^{(m)}\}_{m\in\mathcal{C}_q}$ are accurate. This also requires controlling $\{n_{m,q}\}_{m\in\mathcal{C}_q}$ to meet the sample size constraint (Condition 4) with a high probability, based on Condition 6. This result is included in Lemma 1 and is proved in Appendix H.2.

Lemma 1 (Parameter estimation bound for heterogeneous bandits)**.** *Under Conditions 5-9, for any* $0 \leq q < Q$*, and*[2] $\tau = \arg\min_{m\in[M]}(p_1 \vee p_{[M]}/m))/p_m$*, if* $|\mathcal{H}_q| \geq 2C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))/p_\tau$ *with* C_{b} *defined in Lemma H.1, it holds with probability at least* $1 - 2/T$ *(over the randomness of* $\{\mathbf{X}_q^{(m)}\}_{m=1}^M$*) that for all* $m \in \mathcal{C}_q$*,*

$$\mathbb{E}[\|\widehat{\beta}_q^{(m)} - \beta^{(m)}\|_2^2 \mid (\mathbf{X}_q^{(m)}, Y_q^{(m)})_{m\in\mathcal{C}_q}] = \widetilde{O}\left(\frac{1}{|\mathcal{H}_q|}\left(\frac{s}{p_m} + \frac{d}{p_{[M]}}\right)\right),$$

[2]We choose the largest index achieving the minimum.

where logarithmic factors and quantities depending only on c_x, c_{f} are absorbed into $\widetilde{O}(\cdot)$.

Based on Lemma 1, we can bound the individual regret as follows; with a proof in Appendix H.3.

Theorem 3 (Individual regret upper bound for heterogeneous bandits)**.** *Under Conditions 5-9, the expected regret of MOLARB, for any $T \geq 1$ and $1 \leq H_0 \leq d$, is bounded as*

$$\mathbb{E}[R_T^{(m)}] = \widetilde{O}\left(d \wedge (Tp_m) + \sqrt{\left(s + \frac{dp_m}{p_{[M]}}\right) Tp_m}\right) \tag{6}$$

where logarithmic factors as well as quantities depending only on c_x, c_{f}, L, and μ are absorbed into $\widetilde{O}(\cdot)$. In particular, if contexts are observed for all bandits at all times, so $p_1 = \cdots = p_m = 1$, (6) implies

$$\mathbb{E}[R_T^{(m)}] = \widetilde{O}\left(d \wedge T + \sqrt{\left(s + \frac{d}{M}\right) T}\right).$$

For a single contextual bandit without collaboration, when $T = \Omega(d^2)$, the minimax optimal regret bound is $\widetilde{\Theta}(\sqrt{dT})$, up to logarithmic factors (Han et al., 2020; Chu et al., 2011; Auer, 2002). Theorem 3 implies a regret bound of order $\widetilde{O}(\sqrt{(s + d/M)T})$, when $T = \Omega(d^2/(s + d/M))$, which shows a factor of $\min\{M, d/s\}^{1/2}$ improvement. Similarly to Theorem 1, Algorithm 2 is applicable to any heterogeneity level s. When the heterogeneity is non-sparse, *i.e.*, $s = \Omega(d)$, Theorem 2 recovers the optimal regret $\widetilde{\Theta}(\sqrt{dT})$ for a single bandit. Thus, the collaboration mechanism in Algorithm 2 is always benign, regardless of the heterogeneity.

3.3 Lower Bound

To complement our upper bound, we also present a minimax regret lower bound that characterizes the fundamental learning limits in multitask linear contextual bandits under

sparse heterogeneity. Similar to Theorem 2, Theorem 4 also leverages two representative cases: the homogeneous case where $\beta^{(1)} = \cdots = \beta^{(m)} = \beta^\star$ and the s-sparse case where $\beta^\star = 0$ and $\|\beta^{(m)}\|_0 \leq s$ for all $m \in [M]$. The two cases yield a lower bound of orders $\Omega(\sqrt{dTp_m^2/p_{[M]}})$ and $\Omega(\sqrt{sTp_m})$, respectively. For each case, the proof outline is similar to the lower bound from Han et al. (2020) in the single-bandit regime by additionally incorporating the probabilistic activations. We set a uniform prior for the parameters and translate the regret to a measure characterizing the difficulty of distinguishing two distributions. The difficulty—and therefore the regret—is then quantified and bounded via Le Cam's method (Tsybakov, 2008). Since our task-specific regret in the multitask regime is different from the standard regret in the single-task regime, the proof requires some novel steps to handle the activation sets $\mathcal{S}_t$; see Appendix H.4.

Theorem 4 (Individual regret lower bound for heterogeneous bandits)**.** *Given any $1 \leq s \leq d$ and $\{p_m\}_{m=1}^M \subseteq [0,1]$, for any $m \in [M]$, when $T \geq \max\{(d+1)/p_{[M]}, (s+1)/p_m\}/(16L) + 1$, there exist $\{\beta^{(m)}\}_{a\in[K],m\in[M]}$ satisfying Condition 5, and distributions of contexts satisfying Condition 8 and 9, such that for any online Algorithm A and for any $m \in [M]$,*

$$\mathbb{E}[R_T^{(m)}(A)] = \Omega\left(\sqrt{\left(s + \frac{dp_m}{p_{[M]}}\right) Tp_m}\right).$$

In particular, when $p_m = 1$ for all $m \in [M]$, $\mathbb{E}[R_T^{(m)}(A)] = \Omega\left(\sqrt{(s + d/M)\, T}\right)$.

Theorem 4 and 3 imply that MOLARB is minimax optimal up to logarithmic terms.

4 Experiments

In this section, we evaluate the performance of our method in both offline and online scenarios with synthetic and empirical datasets. We provide an overview of the experiments

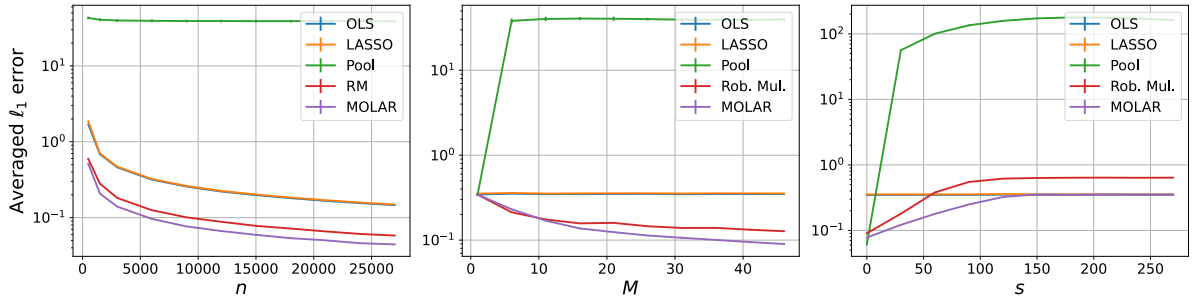

Figure 2: Average ℓ_1 estimation error for multitask linear regression. (Left): Fixing $s = 20$, $M = 30$ and varying n. (Middle): Fixing $s = 20$, $n = 5,000$ and varying M. (Right): Fixing $M = 30$, $n = 5,000$ and varying s. The standard error bars (barely visible) are obtained from ten independent trials.

here and provide the remaining details in Appendix J. For linear regression, we evaluate individual OLS estimates, LASSO estimates, a global OLS estimate via data pooling, the robust multitask estimate (Xu and Bastani, 2021), and our MOLAR estimates, denoted by OLS, LASSO, Pool, RM, and MOLAR below. For contextual bandits, we evaluate the OLS Bandit (Goldenshluger and Zeevi, 2013), LASSO Bandit (Ren and Zhou, 2023), Trace Norm Bandit (Cella et al., 2022), Robust Multitask Bandit (Xu and Bastani, 2021), and our MOLARB methods over multiple Model-C bandit instances[3], denoted by OLSB, LASSOB, TNB, RMB, and MOLARB below. OLSB and LASSOB act by treating M bandit instances independently, either via OLS or LASSO. Trace Norm Bandit is a state-of-the-art multitask bandit method that leverages trace—nuclear—norm regularized estimates, improving accuracy when the parameters span a linear space of rank smaller than d. More experimental details and results can be found in Appendix J.

4.1 Numerical simulations

Linear Regression. We first randomly sample $\beta^\star$ from the uniform distribution over the $(d-1)$-dimensional sphere $\mathbb{S}^{d-1}$ where $d = 300$. From the d covariates of $\beta^\star$, we draw s

[3]A few multitask bandit are not obviously applicable to our setup: Soare et al. (2014); Gentile et al. (2014) aggregate data from similar yet heterogeneous instances, leading to linear growth in regret; (Kveton et al., 2021; Cella et al., 2020; Bastani et al., 2022) consider Bayesian meta-learning that require instances to be observed sequentially rather than simultaneously to construct a prior for instances.

covariates uniformly at random and randomly assign new values sampled from the standard Gaussian distribution with re-normalization to preserve $\|\beta^{(m)}\|_2 = \|\beta^\star\|$ for all $m \in [M]$. We repeat this procedure M times to obtain sparsely perturbed parameters $\{\beta^{(m)}\}_{m=1}^M \subseteq \mathbb{S}^{d-1}$. Then, M datasets $\{x_i^{(1)}\}_{i=1}^n, \ldots, \{x_i^{(m)}\}_{i=1}^n$ with i.i.d. $\mathcal{N}(0, I_d)$ features are sampled for each task $m \in [M]$, each containing n data points[4]. The outcomes $\{y_i^{(1)}\}_{i=1}^n, \ldots, \{y_i^{(m)}\}_{i=1}^n$ are set as $y_i^{(m)} = \langle x_i^{(m)}, \beta^{(m)} \rangle + \varepsilon_i^{(m)}$ where $\varepsilon_i^{(m)}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ noise with $\sigma = 0.1$. We conduct the simulations by varying the sample size n, the number of tasks M, and the number of heterogeneous covariates s. As the datasets have equal sample sizes, we take the averaged ℓ_1 error $\frac{1}{M}\sum_{m=1}^M \|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1$ as the performance metric.

Since we have sparse heterogeneity, we expect RM and MOLAR to outperform baseline methods, which is corroborated by the experimental results from Figure 2. For MOLAR, the estimation error decreases as n and M increase and s decreases, as revealed by Theorem 1. Furthermore, MOLAR outperforms baseline methods for most values of n, M, and s. Other methods outperform MOLAR when s is sufficiently large, as shown in the right panel of Figure 2, which highlights the crucial role of sparse heterogeneity. However, we remark that when s is close to d, the parameters $\{\beta^{(m)}\}_{m=1}^M$ are highly different, and no multitask approaches can provably outperform the individual OLS estimates.

Figure 2 also supports the theoretical predictions from Table 1. OLS does not pool data and thus its estimation error does not vary with M and s. Since the parameters are not sparse, we see no benefit in using LASSO over OLS, and thus, their curves almost overlap. Also, due to the heterogeneity, pooling all datasets introduces a non-vanishing bias when $M > 1$; even for large n and M. Furthermore, these estimation errors grow as s grows, due to the increasing heterogeneity. RM, while addressing heterogeneity, performs worse than

[4]We conduct similar experiments for correlated covariates and disparate task-wise sample sizes in Appendix J.3.2.

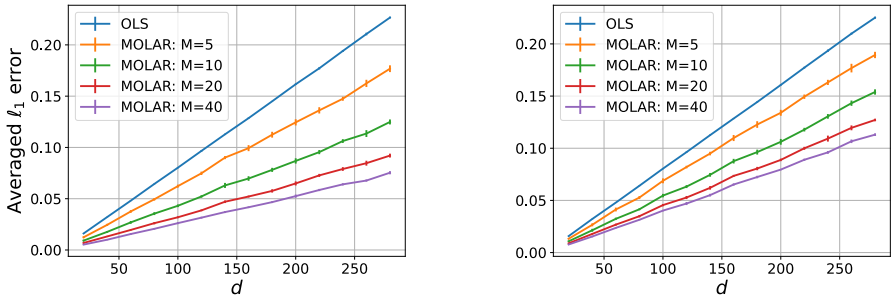

Figure 3: Average ℓ_1 estimation error for multitask linear regression with $n = 10,000$ and $\rho = s/d$ fixed. (Left): $\rho = 0.1$. (Right): $\rho = 0.2$. The standard error bars are obtained from ten independent trials.

MOLAR due to its sub-optimality discussed earlier.

We also perform simulations to support our theoretical results about the rate of the estimation error of MOLAR, presented in Theorem 1. We fix $n = 10,000$ and generate $\{\beta^{(m)}\}_{m=1}^{M}$ in different dimensions d but with a constant ratio $\rho = s/d$. As can be seen in Figure 3, when M and $\rho = s/d$ are fixed, the estimation error of MOLAR grows linearly as d increases. The slopes of the curves, corresponding to $\Theta(\sigma(\rho + 1/\sqrt{M})/\sqrt{n})$ in Theorem 1, decrease as M grows and ρ decays. This aligns with our theoretical results.

Linear Contextual Bandits. We set $(d, s, M, K) = (30, 2, 20, 3)$ and randomly sample the activation probabilities $\{p_m\}_{m=1}^{M}$ from the uniform distribution on $[0, 1]$. We then sample the sparsely heterogeneous parameters $\{\beta^{(m)}\}_{m=1}^{M} \subseteq \mathbb{S}^{d-1}$ as in the linear-regression experiments. For any activated bandit m at time t, we independently sample the contexts $\{x_{t,a}^{(m)}\}_{a\in[K]}$ from $\mathcal{N}(0, I_d)$ and the sample reward noise $\varepsilon_t^{(m)} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$.

We consider instances with a large, medium, and small activation probability respectively, as shown in Figure 4. We observe that MOLARB outperforms all baseline methods. The advantage over OLSB and LASSOB is substantial, as they do not leverage collaboration across tasks. We observe that RMB and TNB outperform OLSB and LASSOB due to regularization. TNB is slightly less accurate because the parameters do not necessarily have a low-rank structure. Moreover, the difference between OLSB and LASSOB is small,

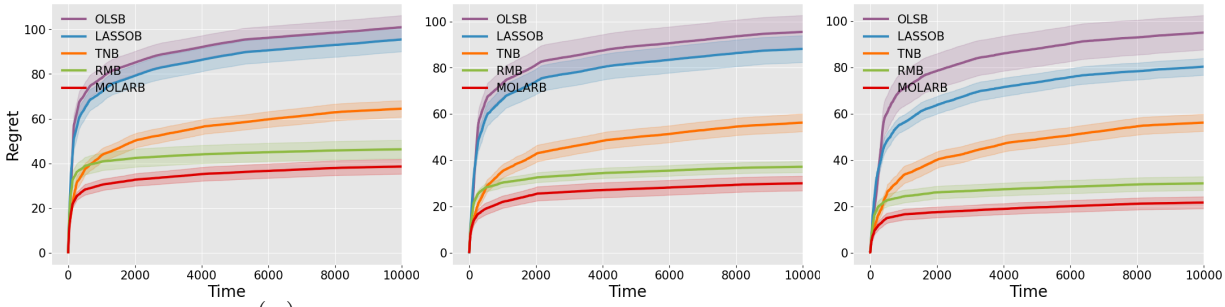

Figure 4: Regret $R_T^{(m)}$ of instances with activation probability 0.778 (Left), 0.466 (Middle), 0.318 (Right), respectively, where shaded regions depict the corresponding 95% normal confidence intervals based on standard errors calculated over twenty independent trials.

as the parameters are not sparse. Also, the cumulative regret of instances with smaller activation probabilities is lower than of the ones with larger activation probabilities, due to fewer rounds of decision-making. Further, MOLARB is computationally more efficient than LASSOB, TNB, and RMB which require solving optimization problems in each update.

4.2 PISA Dataset

The *Programme for International Student Assessment* (PISA) is a large-scale international study conducted by the Organisation for Economic Co-operation and Development (OECD). The study aims to evaluate the quality of education systems around the world by assessing the skills and knowledge of 15-year-old students in reading, mathematics, and science. This dataset has been widely used to gain insights into the impact of factors including teaching practices (OECD, 2019), gender (Stoet and Geary, 2018), and socioeconomic status (Kline et al., 2019) on student academic performance.

In this experiment, we use a part of the PISA2012 data across $M = 15$ countries to learn linear predictors for individual countries, treating each country as a task. After basic preprocessing detailed in the Appendix J, we have 57 student-specific features and a continuous response assessing students' mathematics ability—the variable "PV1MATH", standardized. See Appendix J for additional experimental details including fractions of data used for training, validation, and testing, hyperparameter choices, and robustness

checks. Figure 1 plots the differences in the coefficients across countries. The structure of sparse heterogeneity appears to be reflected in the dataset.

Our experiments include linear regression and contextual bandits. In linear regression, we estimate the linear coefficients of the processed features for predicting the response with the aforementioned estimation methods. We simulate the setup of Model-C as follows. We read the records of two students from each country with an activation probability proportional to its sample size. The goal in each round is to select the student with a better ability in mathematics. Recall that in Model-C, a K-armed bandit observes K contexts at a time, with a shared parameter across the contexts generating the rewards. Accordingly, there are $K = 2$ arms and the reward is the mathematics score. Since the two data points are randomly drawn from the dataset of the same country without replacement, the population parameters are clearly identical.

Table 2: The ℓ_1 estimation errors and the averaged relative error (A.R.E.) on the PISA dataset, over 100 independent random data splits.

Country	OLS	LASSO	Pool	RM	MOLAR
Mexico	1.35 ± 0.02	1.57 ± 0.02	2.11 ± 0.01	$\mathbf{1.22 \pm 0.01}$	1.32 ± 0.01
Italy	2.05 ± 0.04	1.59 ± 0.01	2.34 ± 0.01	$\mathbf{1.53 \pm 0.02}$	1.61 ± 0.02
Spain	2.00 ± 0.04	1.85 ± 0.02	2.64 ± 0.02	$\mathbf{1.66 \pm 0.02}$	$\mathbf{1.67 \pm 0.02}$
Canada	2.09 ± 0.03	2.02 ± 0.02	2.78 ± 0.01	2.05 ± 0.03	$\mathbf{1.78 \pm 0.03}$
Brazil	1.85 ± 0.03	1.80 ± 0.02	2.60 ± 0.01	$\mathbf{1.60 \pm 0.02}$	1.76 ± 0.02
Austrilia	2.52 ± 0.04	1.92 ± 0.02	2.15 ± 0.01	1.99 ± 0.02	$\mathbf{1.76 \pm 0.02}$
UK	2.53 ± 0.03	2.15 ± 0.02	2.32 ± 0.01	1.93 ± 0.02	$\mathbf{1.70 \pm 0.02}$
UAE	2.60 ± 0.05	2.61 ± 0.03	3.08 ± 0.01	2.36 ± 0.03	$\mathbf{2.19 \pm 0.03}$
Switzerland	2.94 ± 0.04	2.66 ± 0.03	3.29 ± 0.01	2.88 ± 0.03	$\mathbf{2.59 \pm 0.03}$
Qatar	2.85 ± 0.04	2.63 ± 0.04	4.24 ± 0.01	$\mathbf{2.49 \pm 0.04}$	$\mathbf{2.56 \pm 0.03}$
Colombia	2.89 ± 0.06	2.36 ± 0.02	2.91 ± 0.01	$\mathbf{2.06 \pm 0.02}$	2.25 ± 0.02
Finland	3.41 ± 0.05	$\mathbf{2.29 \pm 0.02}$	2.97 ± 0.01	2.65 ± 0.03	2.49 ± 0.03
Belgium	3.68 ± 0.06	2.87 ± 0.03	2.91 ± 0.01	2.95 ± 0.04	$\mathbf{2.55 \pm 0.03}$
Denmark	3.48 ± 0.06	2.70 ± 0.03	2.56 ± 0.01	2.23 ± 0.03	$\mathbf{1.87 \pm 0.02}$
Jordan	3.01 ± 0.05	2.57 ± 0.03	2.66 ± 0.01	2.28 ± 0.03	$\mathbf{2.08 \pm 0.03}$
A.R.E.	100%	$85.72 \pm 0.37\%$	$106.05 \pm 0.47\%$	$81.30 \pm 0.33\%$	$\mathbf{76.93 \pm 0.34}\%$

For offline experiments, we show the ℓ_1 estimation errors of all methods on all individual

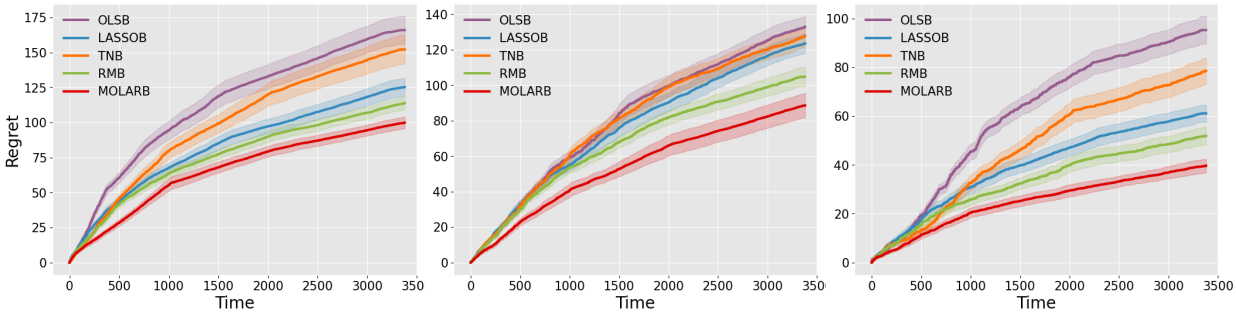

Figure 5: Regret $R_T^{(m)}$ of Canada, UAE, and Denmark of the PISA dataset. The shaded regions depict the corresponding 95% normal confidence intervals based on standard errors calculated over twenty independent trials.

countries in Table 2. The best result, *i.e.*, outperforming all others regardless of standard deviations, is in bold font. For a global comparison, we define the averaged relative error $\sum_{m=1}^{(m)} \|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 / (\sum_{m=1}^{(m)} \|\widehat{\beta}_{\text{ind}}^{(m)} - \beta^{(m)}\|_1)$ where $\widehat{\beta}_{\text{ind}}^{(m)}$ is the individual OLS estimate over task m. MOLAR outperforms other methods on most tasks and is the best in terms of the global error metric.

For online experiments, we present the results for the tasks associated with Canada, UAE, and Denmark—which have activation probabilities $p_m = 0.64,\ 0.32,\ 0.22$, respectively—in Figure 5. More figures can be found in Appendix J.2. Again, we see that MOLARB performs favorably compared to the baselines. This is more pronounced for the tasks with smaller activation probabilities, as suggested by the theory.

5 Conclusion

We consider multitask learning under sparse heterogeneity in both linear regression and linear contextual bandits. For linear regression, we propose the MOLAR algorithm that collaborates on multiple datasets, improving accuracy compared to existing multitask methods. Applying MOLAR to linear contextual bandits, we also improve current regret bounds for individual bandit instances. To complement the upper bounds, we establish lower bounds for multitask linear regression and contextual bandits, justifying the optimality of

the proposed methods. Our methods are also extended to generalized linear models and the construction of confidence intervals. Our experimental results support our theoretical findings. Future directions include investigating problem-specific optimal methods whose rate depends on $\{\beta^{(m)}\}_{m=1}^{M}$.

References

R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *European Conference on Machine Learning*, pages 39–50. Springer, 2004.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

L. Baardman, S. B. Boroujeni, T. Cohen-Hillel, K. Panchamgam, and G. Perakis. Detecting customer trends for optimal promotion targeting. *Manufacturing & Service Operations Management*, 25(2):448–467, 2023.

H. Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

H. Bastani and M. Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

H. Bastani, M. Bayati, and K. Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

H. Bastani, D. Simchi-Levi, and R. Zhu. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 68(3):1865–1881, 2022.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

L. Cella, A. Lazaric, and M. Pontil. Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*, pages 1360–1370. PMLR, 2020.

L. Cella, K. Lounici, and M. Pontil. Multi-task representation learning with stochastic linear bandits. *arXiv preprint arXiv:2202.10066*, 2022.

N. Cesa-Bianchi, C. Gentile, and G. Zappella. A gang of bandits. *Advances in neural information processing systems*, 26, 2013.

S. Chakraborty, S. Roy, and A. Tewari. Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pages 3979–4008. PMLR, 2023.

N. V. Chawla. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pages 875–886, 2010.

C. Chen, W. Xu, and L. Zhu. Distributed estimation in heterogeneous reduced rank regression: With application to order determination in sufficient dimension reduction. *Journal of Multivariate Analysis*, 190:104991, 2022a.

H. Chen, W. Lu, and R. Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533): 240–255, 2021.

X. Chen, Z. Lai, H. Li, and Y. Zhang. Online statistical inference for contextual bandits via stochastic gradient descent. *arXiv preprint arXiv:2212.14883*, 2022b.

Y. Chen, Y. Wang, E. X. Fang, Z. Wang, and R. Li. Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection. *Journal of the American Statistical Association*, pages 1–13, 2022c.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.

A. A. Deshmukh, U. Dogan, and C. Scott. Multi-task learning for contextual bandits. *Advances in neural information processing systems*, 30, 2017.

E. Dobriban and Y. Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49:918–943, 2021.

S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Y. Duan and K. Wang. Adaptive and robust multi-task learning. *arXiv preprint arXiv:2202.05250*, 2022.

T. Evgeniou and M. Pontil. Regularized multi–task learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004.

T. Evgeniou, C. A. Micchelli, M. Pontil, and J. Shawe-Taylor. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

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

C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765. PMLR, 2014.

C. Gentile, S. Li, P. Kar, A. Karatzoglou, G. Zappella, and E. Etrue. On context-dependent clustering of bandits. In *International Conference on machine learning*, pages 1253–1262. PMLR, 2017.

A. Goldenshluger and A. Zeevi. A linear response bandit problem. *Stochastic Systems*, 3 (1):230–261, 2013.

J. Gu and S. Chen. Weighted distributed estimation under heterogeneity. *arXiv preprint arXiv:2209.06482*, 2022.

Z. Guo. Inference for high-dimensional maximin effects in heterogeneous regression models using a sampling approach. *arXiv preprint arXiv:2011.07568*, 2020.

Y. Han, Z. Zhou, Z. Zhou, J. Blanchet, P. Glynn, and Y. Ye. Sequential batch learning in finite-action linear contextual bandits. *ArXiv*, 2020.

B. Hao, T. Lattimore, C. Szepesvári, and M. Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.

M. Hu, X. Shi, and P. X.-K. Song. Collaborative causal inference with a distributed data-sharing management. *arXiv preprint arXiv:2204.00857*, 2022.

G.-S. Kim and M. C. Paik. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

P. Kline, W. Johnson, L. Ingraham, E. D. Heggestad, J. L. Huang, B. K. Gorman, B. Bray, P. J. Cawley, B. S. Connelly, K. S. Cortina, et al. Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 89 (3):420–460, 2019.

S. Kotsiantis, D. Kanellopoulos, P. Pintelas, et al. Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, 30(1):25–36, 2006.

B. Kveton, M. Konobeev, M. Zaheer, C.-w. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. Meta-thompson sampling. In *International Conference on Machine Learning*, pages 5884–5893. PMLR, 2021.

A. Lazaric, E. Brunskill, et al. Sequential transfer in multi-armed bandit with finite set of models. *Advances in Neural Information Processing Systems*, 26, 2013.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer-Verlag, 1998.

S. Li, T. T. Cai, and H. Li. Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *arXiv preprint arXiv:2006.10593*, 2020.

S. Li, L. Zhang, T. T. Cai, and H. Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, pages 1–12, 2023.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

C. Luo, R. Duan, A. C. Naj, H. R. Kranzler, J. Bian, and Y. Chen. Odach: a one-shot distributed algorithm for cox model with heterogeneous multi-center data. *Scientific reports*, 12(1):6627, 2022a.

Y. Luo, W. W. Sun, and Y. Liu. Contextual dynamic pricing with unknown noise: Explore-then-ucb strategy and improved regrets. In *Advances in Neural Information Processing Systems*, volume 35, pages 37445–37457, 2022b.

J. Marron. Big data in context and robustness against heterogeneity. *Econometrics and Statistics*, 2:73–80, 2017.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

OECD. *Teaching for the future: Effective classroom practices to transform education*. OECD Publishing, 2019.

M.-h. Oh, G. Iyengar, and A. Zeevi. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR, 2021.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *THE ANNALS of STATISTICS*, 41:693–721, 2013.

J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011. doi: 10.1109/TIT.2011.2165799.

Z. Ren and Z. Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science*, 2023.

J. Sarkar. One-armed bandit problems with covariates. *The Annals of Statistics*, pages 1978–2002, 1991.

C. Singh and A. Sharma. Online learning using multiple times weight updating. *Applied Artificial Intelligence*, 34(6):515–536, 2020.

M. Soare, O. Alsharif, A. Lazaric, and J. Pineau. Multi-task linear bandits. In *NIPS2014 workshop on transfer and multi-task learning: theory meets practice*, 2014.

G. Stoet and D. C. Geary. The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, 29(4):581–593, 2018.

Y. Tian and Y. Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 0:1–14, 2022.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

N. Tripuraneni, C. Jin, and M. Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

A. B. Tsybakov. Introduction to nonparametric estimation. In *Springer Series in Statistics*, 2008.

M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

E. Van Herpen, E. Van Nierop, and L. Sloot. The relationship between in-store marketing and observed sales for organic versus fair trade products. *Marketing Letters*, 23:293–308, 2012.

B. Wang, Y. Fang, H. Lian, and H. Liang. Additive partially linear models for massive heterogeneous data. *Electronic Journal of Statistics*, 13(1):391–431, 2019.

M. Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

K. Xu and H. Bastani. Learning across bandits in high dimension via robust statistics. *arXiv preprint arXiv:2112.14233*, 2021.

K. Xu, X. Zhao, H. Bastani, and O. Bastani. Group-sparse matrix factorization for transfer learning of word embeddings. In *International Conference on Machine Learning*, pages 11603–11612. PMLR, 2021.

F. Yang, H. R. Zhang, S. Wu, W. J. Su, and C. Ré. Analysis of information transfer from heterogeneous sources via precise high-dimensional asymptotics. *arXiv preprint arXiv:2010.11750*, 2020.

X. Yang, X. Yan, and J. Huang. High-dimensional integrative analysis with homogeneity and sparsity recovery. *Journal of Multivariate Analysis*, 174:104529, 2019.

Y. Yang and D. Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

K. Yuan, S. A. Alghunaim, and X. Huang. Removing data heterogeneity influence enhances network topology dependence of decentralized sgd. *Journal of Machine Learning Research*, 24(280):1–53, 2023.

X. Zhang and W. Wang. Optimal model averaging estimation for partially linear models. *Statistica Sinica*, 29(2):693–718, 2019.

T. Zhao, G. Cheng, and H. Liu. A partially linear framework for massive heterogeneous data. *Annals of statistics*, 44(4):1400, 2016.

Supplementary Material to "Optimal Multitask Linear Regression and Contextual Bandits under Sparse Heterogeneity"

A More Related Works & Notations

Multitask Learning. When the data has components corresponding to multiple domains (also referred to as tasks or sources), multitask learning aims to develop methods that borrow information across tasks (Caruana, 1998). Multitask learning can be beneficial when the task-associated parameters are close in some sense, *e.g.*, in the ℓ_2 norm, or follow a common prior distribution (Raina et al., 2006; Hanneke and Kpotufe, 2022). Popular multitask methods include regularizing the parameters to be estimated towards a common parameter—through ridge (Evgeniou and Pontil, 2004; Hanzely et al., 2020), ℓ_2 (Duan and Wang, 2022), kernel ridge (Evgeniou et al., 2005) penalties, etc—and clusteringpooling datasets based on similarity metrics (Ben-David et al., 2010; Crammer et al., 2008; Dobriban and Sheng, 2021). One can further leverage certain shared structures to improve rates of estimation. Tripuraneni et al. (2021); Du et al. (2020); Collins et al. (2021) study a low dimensional shared representation of task-specific models. Lounici et al. (2009); Singh and Sharma (2020) consider the parameters for each task to be sparse and share the same support. Bastani (2021); Xu and Bastani (2021); Huang et al. (2022) motivate and study sparse heterogeneity.

Robust Statistics & Learning. In robust statistics and learning (Huber, 1981; Hampel et al., 2011) many methods have been developed that are resilient to unknown data corruption (Rousseeuw, 1991; Minsker, 2013). From the optimization perspective, methods to robustly aggregate gradients of the loss functions have been developed (Su and Vaidya, 2016; Blanchard et al., 2017; Yin et al., 2018). Our setting is different and requires a novel analysis.

Notations. We use $:=$ or $\triangleq$ to introduce definitions. For an integer $d \geq 1$, we write $[d]$ for both $\{1, \ldots, d\}$ and $\{e_1, \ldots, e_d\} \subseteq \mathbb{R}^d$, where e_k is the k-th canonical basis vector of $\mathbb{R}^d$. We use I_d to denote the $d \times d$ identity matrix. We let $\mathbb{B}_d$ denote the unit Euclidean ball centered at the origin in $\mathbb{R}^d$. For a vector $v \in \mathbb{R}^d$, we denote its entries as $v_1, \ldots, v_d$. We also denote $\|v\|_p = (\sum_{k\in[d]} |v_k|^p)^{1/p}$ for all $p > 0$, with $\|v\|_0$ defined as the number of non-zero covariates. For any $\mathcal{I} \subseteq [M]$, given weights $\{w_m\}_{m=1}^M$ (or sample sizes $\{n_m\}_{m=1}^M$), we denote $W_{\mathcal{I}}$ as $\sum_{m\in\mathcal{I}} w_m$ and write $n_{\mathcal{I}}$ for $\sum_{m\in\mathcal{I}} n_m$. We also write $[v]_{\mathcal{I}}$ and $v_{\mathcal{I}}$ as the sub-vector of v with entries in $\mathcal{I}$ For a matrix $A \in \mathbb{R}^{m\times n}$, we denote the (i, j)-th covariate of A by $[A]_{i,j}$ or $A_{i,j}$, and the i-th row (resp., the j-the column) by $A_{i,\cdot}$ (resp., $A_{\cdot,j}$). For two real numbers a and b, we write $a \vee b$ and $a \wedge b$ for $\max\{a, b\}$ and $\min\{a, b\}$, respectively. For $\sigma^2 > 0$, we denote by $\mathsf{subG}(\sigma^2)$ the class of σ^2-sub-Gaussian random variables. For an event E, we write $\mathbb{1}(E)$ for the indicator of the event; so $\mathbb{1}(E)(x) = 1$ if $x \in E$ and $\mathbb{1}(E)(x) = 0$ otherwise. We use the Bachmann-Landau asymptotic notations $\Omega(\cdot)$, $\Theta(\cdot)$, $O(\cdot)$ to absorb constant factors, and use $\tilde{\Omega}(\cdot)$, $\widetilde{O}(\cdot)$ to also absorb logarithmic factors in various problem parameters specified in each

case. Furthermore, we use probabilistic notations such as $O_P(a_{\{n_m\}_{m=1}^M})$ to denote quantities that are bounded by $a_{\{n_m\}_{m=1}^M}$ with overwhelming probabilities as $\min_{m\in[M]} n_m \to \infty$. For a number $x \in \mathbb{R}$, we use $(x)_+$ to denote its non-negative part, *i.e.*, $x\mathbb{1}(x \geq 0)$.

B Technical Lemmas

Lemma B.1 (Tail integral formula for expectation)**.** *For any non-negative, continuous random variable Z with $\mathbb{E}[Z] < \infty$ and any $q \geq 0$, it holds that*

$$\mathbb{E}[Z\mathbb{1}(Z \geq q)] = q\mathbb{P}(Z \geq q) + \int_q^\infty \mathbb{P}(Z \geq t)\mathrm{d}t.$$

Proof. The result is well known (e.g., Exercise 1.2.3 in Vershynin, 2018). □

Lemma B.2 (Maximal Inequalites)**.** *For $\sigma^2 > 0$ and for $1 \leq m \leq M$, let $X_m \sim \mathsf{subG}(\sigma^2)$, not necessarily independent, for all $1 \leq m \leq M$. Then, it holds that*

1. $\mathbb{E}[\max_{1\leq m\leq M} X_m] \leq \sigma\sqrt{2\ln(M)}$;
2. $\mathbb{E}[\max_{1\leq m\leq M} |X_m|] \leq \sigma\sqrt{2\ln(2M)}$;
3. *For any $t \geq 0$, $\mathbb{P}\left(\max_{1\leq m\leq M} X_m \geq t\right) \leq M\exp(-t^2/(2\sigma^2))$.*

Proof. The result is well known (e.g., Koltchinskii and Panchenko, 2002). □

Lemma B.3 (Bernstein's inequality; Uspensky (1937))**.** *Let $Z_1, \ldots, Z_n$ be i.i.d. random variable with $|Z_1 - \mathbb{E}[Z_1]| \leq b$ and $\mathrm{Var}(Z_1) = \sigma^2 > 0$, and let $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$. Then for any $\delta \geq 0$,*

$$\max\{\mathbb{P}(\bar{Z} - \mathbb{E}[\bar{Z}] > \delta), \mathbb{P}(\bar{Z} - \mathbb{E}[\bar{Z}] < -\delta)\} \leq \exp\left(-\frac{n\delta^2}{2(\sigma^2 + b\delta)}\right). \tag{B.1}$$

Lemma B.4 (Properties of Orlicz norm (Smithies, 1962))**.** *For any $\alpha \in (0, 2]$, the following properties hold when $\|Z\|_{\Psi_\alpha}$ exists.*

1. *Normalization: $\mathbb{E}[\Psi_\alpha(|Z|/\|Z\|_{\Psi_\alpha})] \leq 1$.*
2. *Homogeneity: $\|cZ\|_{\Psi_\alpha} = c\|Z\|_{\Psi_\alpha}$ for any $c \in \mathbb{R}$.*
3. *Deviation inequality: $\mathbb{P}(|Z| \geq t) \leq 2\exp(-(t/\|Z\|_{\Psi_\alpha})^\alpha)$.*

Lemma B.5 (Berry-Esseen theorem (Shevtsova, 2010))**.** *Given independent random variables $\{Z_t\}_{i=1}^n$ with $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[Z_i] = \sigma_i^2 \geq 0$, and $\mathbb{E}[|Z_i|^3] = \rho_i < \infty$, let $S_n = \sum_{i=1}^n Z_i/\sqrt{\sum_{i=1}^n \sigma_i^2}$ be the normalized sum, and denote F_n the c.d.f. of S_n, and Φ the c.d.f. of the standard normal distribution. It holds that*

$$\sup_{z\in\mathbb{R}} |F_n(z) - \Phi(z)| \leq 0.6\sum_{t=1}^n \rho_t / \left(\sum_{t=1}^n \sigma_t^2\right)^{3/2}.$$

Lemma B.6 (Generalized Hanson-Wright inequality (Götze et al., 2021; Sambale, 2020))**.** *For any $\alpha \in (0, 2]$, let $Z_1, \ldots, Z_n$ be i.i.d. zero-mean random variables with*

$\|Z_1\|_{\Psi_\alpha} \le \sigma$ *and* $A = (a_{i,j})$ *be a symmetric matrix. Then there is an absolute constant* c_{hw} *such that for any* $t \ge 0$*,*

$$\mathbb{P}\left(\left|\sum_{i,j\in[n]} a_{i,j}Z_iZ_j - \sum_{i\in[n]} a_{i,i}\mathrm{Var}(Z_1)\right| \ge t\right) \le 2\exp\left(-\frac{1}{c_{\mathrm{hw}}}\min\left\{\frac{t^2}{\sigma^4\|A\|_{\mathrm{F}}^2}, \left(\frac{t}{\sigma^2\|A\|_{\mathrm{op}}^2}\right)^{\alpha/2}\right\}\right)$$

where $\|\cdot\|_{\mathrm{F}}$ *and* $\|\cdot\|_{\mathrm{op}}$ *indicates the Frobenius norm and the operator norm, respectively. In particular, when* $A = vv^\top$ *with* $v = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n$*, it holds for any* $t \ge \sigma^2\|v\|_2^2$ *that*

$$\mathbb{P}\left(\left|\left(\sum_{i\in[n]} v_iZ_i\right)^2 - \|v\|_2^2\mathrm{Var}(Z_1)\right| \ge t\right) \le 2\exp\left(-\frac{1}{c_{\mathrm{hw}}}\left(\frac{t}{\sigma^2\|v\|_2^2}\right)^{\alpha/2}\right).$$

Lemma B.7 (Joint convexity of the KL-divergence (Cover and Thomas, 1991))**.** *The Kullback-Leibler divergence* $D_{\mathrm{KL}}(P \,\|\, Q)$ *is jointly convex in its arguments* P *and* Q*: let* P_1*,* P_2*,* Q_1*,* Q_2 *be distributions on a common set* $\mathcal{X}$*, then for any* $\lambda \in [0,1]$*, it holds that*

$$D_{\mathrm{KL}}(\lambda P_1 + (1-\lambda)P_2 \,\|\, \lambda Q_1 + (1-\lambda)Q_2) \le \lambda D_{\mathrm{KL}}(P_1 \,\|\, Q_1) + (1-\lambda)D_{\mathrm{KL}}(P_2 \,\|\, Q_2).$$

More generally, if the parameter θ *follows some prior* R *and if conditioned on the parameter* θ*, the random variables* $X \sim P_\theta$ *and* $Y \sim Q_\theta$*, then*

$$D_{\mathrm{KL}}(P_\theta \circ R \,\|\, Q_\theta \circ R) \le \mathbb{E}_{\theta\sim R}[D_{\mathrm{KL}}(P_\theta \,\|\, Q_\theta)],$$

Lemma B.8 (Lemma 8 of Han et al. (2020))**.** *Suppose that for* $d \ge 2$*,* $Z \in \mathbb{R}^d$ *is uniformly distributed on the source* $(d-1)$*-dimensional sphere, then the absolute moments of the first coordinate* $[Z]_1$ *of* Z *are, for* $k > -1$

$$\mathbb{E}[|[Z]_1|^k] = \frac{\Gamma(\frac{d}{2})\Gamma(\frac{k+1}{2})}{\Gamma(\frac{d+k}{2})\Gamma(\frac{1}{2})}$$

where $\Gamma(\cdot)$ *is the gamma function.*

Definition B.1 (Maximum expectation over small probability sets)**.** *Given a probability space* $A = (\Omega, \mathbb{P}, \mathcal{F})$*, a random variable* Z *over* A*, and* $\delta \in [0,1]$*, define* $\mathbb{E}^\delta[Z]$ *as the maximum expectation of* Z *over all measurable sets with probability at most* δ*:*

$$\mathbb{E}^\delta[Z] := \sup_{A\in\mathcal{F}} \{\mathbb{E}[Z(\omega)\mathbb{1}(\omega \in A)] : \mathbb{P}(A) \le \delta\}. \tag{B.2}$$

Lemma B.9 (Sub-Gaussian integral over small probability sets)**.** *For any* $Z \sim \mathsf{subG}(\sigma^2)$ *with* $\sigma^2 > 0$ *and* $\mathbb{E}[Z] = 0$*, the maximum expectation over small probability sets from* (B.2) *of* $|Z|$ *and* Z^2 *is bounded as*

$$\mathbb{E}^\delta[|Z|] = O\left(\delta\sigma\ln(2/\delta)^{1/2}\right) \tag{B.3}$$

and $\mathbb{E}^\delta[Z^2] = O\left(\delta\sigma^2\ln(2/\delta)\right)$*. Moreover, if* $|Z| \ge |\widetilde{Z}|$ *a.s., then* $\mathbb{E}^\delta[|\widetilde{Z}|] \le \mathbb{E}^\delta[|Z|]$ *for any* $\delta \in [0,1]$*.*

Proof. We first consider a continuous random variable Z over a probability space $A = (\Omega, \mathbb{P}, \mathcal{F})$. Then, from (B.2), $\mathbb{E}^\delta[|Z|]$ is given by the integral over

$$A_\delta = \{\omega \in \Omega : |Z(\omega)| \geq q_\delta\} \quad \text{with} \quad \mathbb{P}(|Z(\omega)| \geq q_\delta) = \delta.$$

Since $Z \sim \mathsf{subG}(\sigma^2)$, by the Chernoff bound, we have $\delta = \mathbb{P}(|Z(\omega)| \geq q_\delta) \leq 2\exp\left(-\frac{q_\delta^2}{2\sigma^2}\right)$, which implies

$$q_\delta \leq \sqrt{2\ln(2/\delta)}\sigma. \tag{B.4}$$

Plugging (B.4) and the Chernoff bound into Lemma B.1, we find

$$\begin{aligned}\mathbb{E}[|Z|\mathbb{1}(|Z| \geq q_\delta)] &= q_\delta \mathbb{P}(|Z| \geq q_\delta) + \int_{q_\delta}^\infty \mathbb{P}(|Z| \geq t)\mathrm{d}t \\ &\leq \delta q_\delta + \int_{q_\delta}^\infty \min\left\{\delta, 2\exp\left(-\frac{t^2}{2\sigma^2}\right)\right\}\mathrm{d}t \end{aligned} \tag{B.5}$$

where the last inequality follows from the Chernoff bound and from $\mathbb{P}(|Z| \geq t) \leq \mathbb{P}(|Z| \geq q_\delta) = \delta$. Now,

$$\begin{aligned}&\int_{q_\delta}^\infty \min\left\{\delta, 2\exp\left(-\frac{t^2}{2\sigma^2}\right)\right\}\mathrm{d}t = \delta\left(\sqrt{2\ln(2/\delta)}\sigma - q_\delta\right) + 2\int_{\sqrt{2\ln(2/\delta)}\sigma}^\infty \exp\left(-\frac{t^2}{2\sigma^2}\right)\mathrm{d}t \\ &= \delta\left(\sqrt{2\ln(2/\delta)}\sigma - q_\delta\right) + 2\sqrt{2\pi}\sigma \cdot \mathbb{P}\left(\mathcal{N}(0,\sigma) \geq \sqrt{2\ln(2/\delta)}\sigma\right). \end{aligned} \tag{B.6}$$

Using (Vershynin, 2018, Proposition 2.1.2), we have

$$\begin{aligned}&\mathbb{P}\left(\mathcal{N}(0,\sigma^2) \geq \sqrt{2\ln(2/\delta)}\sigma\right) = \mathbb{P}\left(\mathcal{N}(0,1) \geq \sqrt{2\ln(2/\delta)}\right) \\ &\leq \frac{1}{\sqrt{2\pi}\sqrt{2\ln(2/\delta)}}\exp\left(-\ln(2/\delta)\right) = \frac{\delta}{4\sqrt{\pi\ln(2/\delta)}}. \end{aligned} \tag{B.7}$$

Combining (B.7) with (B.6) and (B.5), we find (B.3). For random variables Z that are not necessarily continuous, let $Z_\varepsilon := \sqrt{1-\varepsilon}\,Z + \sqrt{\varepsilon}\,Z'$ with independent Gaussian $Z' \sim \mathcal{N}(0,\sigma^2)$ and $0 \leq \varepsilon \leq 1$. Clearly, $Z_\varepsilon \sim \mathsf{subG}(\sigma^2)$ and is continuous. By the result of the continuous case, we have

$$\delta\sigma\left([2\ln(2/\delta)]^{1/2} + [2\ln(2/\delta)]^{-1/2}\right) \geq \mathbb{E}^\delta[|Z_\varepsilon|] \geq \sqrt{1-\varepsilon}\,\mathbb{E}^\delta[|Z|] - \sqrt{\varepsilon}\,\mathbb{E}^\delta[|Z'|]. \tag{B.8}$$

Letting $\varepsilon \to 0$ in the right-hand side of (B.8), (B.3) follows for general Z. The result for $\mathbb{E}^\delta[Z^2]$ follows similarly. □

Lemma B.10. *Given $\eta \in (0,1]$, $r_k \geq 0$ for all $k \in [d]$, $a \geq 0$, and for $p \in \{1,2\}$, consider the functions $f_q : \{x \in \mathbb{R}^d : 0 \leq x_k \leq 1, \forall\, k \in [d] \text{ and } \sum_{k\in[d]} x_k \leq s\} \to \mathbb{R}$, $f_p(x_1,\ldots,x_d) := \sum_{k\in[d]}(x_k \wedge a)^p \mathbb{1}\{x_k < \eta\}) + a^p \mathbb{1}\{x_k \geq \eta\}$. Then it holds that*

$$\max_{x_1,\ldots,x_d} f_p(x_1 \ldots, x_d) \leq a^p\left(\left\lceil \frac{s}{a \wedge \eta}\right\rceil \wedge d\right).$$

Proof. We only prove the result for f_1, and the result for function f_2 follows similarly. Without loss of generality, we assume $1 \geq x_1 \geq x_2 \geq \cdots \geq x_d \geq 0$ and $\lceil s/(a \wedge \eta) \rceil < d$ since $f_p(x_1 \ldots, x_d) \leq da^p$ is clear. In this case, we claim that the maximum can be attained at $x_1 = \cdots = x_{\lfloor s/(a \wedge \eta) \rfloor} = a \wedge \eta$, $x_{\lfloor s/(a \wedge \eta) \rfloor + 1} = s - \eta \lfloor s/(a \wedge \eta) \rfloor$, and $x_k = 0$ for all $k > \lfloor s/\eta \rfloor + 1$. Further, the maximum is upper bounded by $a^p \left(\lceil s/(a \wedge \eta) \rceil \wedge d \right)$. We now use the exchange argument to prove the claim.

S. 1 If there is some k such that $x_k > a \wedge \eta \geq x_{k+1}$, then defining x' by letting $(x'_k, x'_{k+1}) = ((a \wedge \eta), x_k + x_{k+1} - (a \wedge \eta))$ while for other j, $x'_j = x_j$, does not decrease the value of f_1. Therefore, the maximum is attained by x such that for some j, $x_1 = \cdots = x_j = (a \wedge \eta) > x_{j+1} \geq \cdots \geq x_d$.

S. 2 If there is some k such that $(a \wedge \eta) > x_k \geq x_{k+1} > 0$, then defining x' by letting $(x'_k, x'_{k+1}) = (\min\{(a \wedge \eta), x_k + x_{k+1}\}, \max\{0, x_k + x_{k+1} - (a \wedge \eta)\})$ while for other j, $x'_j = x_j$, does not decrease the value of f_1. Therefore, combined with Step 1, the maximum is attained by x such that for some j, $x_1 = \cdots = x_j = a \wedge \eta > x_{j+1} \geq 0$ and $x_k = 0$ for all $k > j + 1$. Thus at most one element lies in $(0, \eta)$.

Combining S. 1 and S. 2 above, we complete the proof of the claim, which further leads to the conclusion. □

C Results on Linear Regression with Gaussian Noise

C.1 Lemma C.2 and its Proof

Lemma C.2 characterizes the estimation error of the median of Gaussian inputs. This is similar to classical results from robust statistics (see, *e.g.*, Lerasle and Oliveira, 2011), but existing results typically assume that the "inlier data" is an i.i.d. sample. In contrast, we only require independence, since we wish to apply it to the non-i.i.d. variables $\{\widehat{\beta}_{\text{ind}}^{(m)}\}_{m=1}^{M}$.

Lemma C.1. *Given independent Gaussian random variables $\{Z_i \sim \mathcal{N}(\mu, \sigma_i^2)\}_{i\in\mathcal{I}}$ with a shared mean but with possibly different variances, and non-negative weights $\{w_i\}_{i\in\mathcal{I}}$ with $W_{\mathcal{I}} \triangleq \sum_{i\in\mathcal{I}} w_i$, then $1/2+\alpha$- weighted population quantile $\mu_{1/2+\alpha}$ is defined such that*

$$\sum_{i\in\mathcal{I}} w_i \Phi\left(\frac{\mu_{1/2+\alpha}-\mu}{\sigma_i}\right) = \left(\frac{1}{2}+\alpha\right) W_{\mathcal{I}}. \tag{C.1}$$

Then it holds for any $|\alpha| < 1/2$ that

$$|\mu_{1/2+\alpha} - \mu| \le C_\alpha \alpha \bar{\sigma}_{\mathcal{I}}$$

where $\bar{\sigma}_{\mathcal{I}} = \sum_{i\in\mathcal{I}} w_i\sigma_i/W_{\mathcal{I}}$, $C_\alpha \triangleq \max_{0<\epsilon<1/2-\alpha}\left\{\phi(\Phi^{-1}(1-\epsilon))(1-\frac{2\alpha}{1-2\epsilon})\right\}^{-1}$, and ϕ, Φ are the density and c.d.f. of the standard normal distribution, respectively.

Proof. We only prove the case where $\alpha \ge 0$ and the other case follows by symmetry. We denote the normalized weight $w_i/W_{\mathcal{I}}$ as $\bar{w}_i$. Clearly, $\mu_{1/2+\alpha} \ge \mu$ for $\alpha \ge 0$, so we can divide $\mathcal{I}$ into to two groups based on weighted probabilities:

$$\mathcal{I}_{\text{small}} := \{i \in \mathcal{I} : \Phi((\mu_{1/2+\alpha}-\mu)/\sigma_i) \le 1-\epsilon\}, \quad \mathcal{I}_{\text{large}} := \mathcal{I}\backslash\mathcal{I}_{\text{small}},$$

where $0 < \epsilon < 1/2-\alpha$ is a real number to be chosen later. Using the mean-value theorem, for all $0 \le z \le \Phi^{-1}(1-\epsilon)$, there exists $\xi \in (0, z)$ such that

$$\Phi(z) = \frac{1}{2} + z\phi(\xi) \ge \frac{1}{2} + z\phi(\Phi^{-1}(1-\epsilon)).$$

where ϕ is the density of the standard normal distribution. We thus have

$$\frac{1}{2}+\alpha \ge \sum_{i\in\mathcal{I}_{\text{small}}} \bar{w}_i\left(\frac{1}{2} + \phi(\Phi^{-1}(1-\epsilon))\sigma_i^{-1}(\mu_{1/2+\alpha}-\mu)\right) + \sum_{i\in\mathcal{I}_{\text{large}}} \bar{w}_i(1-\epsilon), \tag{C.2}$$

leading to

$$\mu_{1/2+\alpha} \le \mu + \alpha \Big/ \left(\phi(\Phi^{-1}(1-\epsilon)) \sum_{i\in\mathcal{I}_{\text{small}}} \bar{w}_i/\sigma_i\right). \tag{C.3}$$

On the other hand, we have from (C.2) that

$$\sum_{i\in\mathcal{I}_{\text{large}}} w_i \le \frac{\alpha}{1/2-\epsilon},$$

which implies

$$\sum_{i\in\mathcal{I}_{\text{small}}} w_i \geq 1 - \frac{\alpha}{1/2-\epsilon}.$$

Consequently, using Hölder's inequality, we obtain

$$\sum_{i\in\mathcal{I}_{\text{small}}} \bar{w}_i/\sigma_i \sum_{i\in\mathcal{I}_{\text{small}}} \bar{w}_i\sigma_i \geq \left(\sum_{i\in\mathcal{I}_{\text{small}}} \bar{w}_i\right)^2 \geq \left(1 - \frac{\alpha}{1/2-\epsilon}\right)^2. \tag{C.4}$$

Combing (C.3) and (C.4), we obtain

$$\mu_{1/2+\alpha} \leq \mu + C_\alpha \sum_{i\in\mathcal{I}_{\text{small}}} \bar{w}_i\sigma_i,$$

where C_α is defined in the statement. □

Lemma C.2. *Given independent Gaussian random variables $\{Z_m \sim \mathcal{N}(\mu_m, \sigma_m^2)\}_{m=1}^M$, any positive weights $\{w_m\}_{m=1}^M$, and some $\mu \in \mathbb{R}$, let $\mathcal{B} \triangleq \{m \in [M] : \mu_m \neq \mu\}$ and $\mathcal{G} \triangleq [M]\backslash\mathcal{B}$. If $|\mathcal{B}| < M$, for any $\delta \geq 0$ such that*

$$\alpha_{\mathcal{B},\delta} \triangleq \sum_{m\in\mathcal{B}} w_m/W_{[M]} + \sqrt{1.01\delta \sum_{m\in\mathcal{G}} w_m^2}/W_{\mathcal{G}} < \frac{1}{2}, \tag{C.5}$$

it holds with probability at least $1 - 2e^{-2\delta}$ that

$$|\text{WMed}(\{Z_m\}_{m\in[M]}; \{w_m\}_{m\in[M]}) - \mu| \leq C_{\alpha_{\mathcal{B},\delta}} \alpha_{\mathcal{B},\delta} \bar{\sigma}_{\mathcal{G}},$$

where $\bar{\sigma}_{\mathcal{G}} \triangleq \sum_{m\in\mathcal{G}} w_m\sigma_m/W_{\mathcal{G}}$ and C_α is the constant depending only on α defined in Lemma C.1.

Proof. Denote $\mathcal{B}^c$ as $\mathcal{G}$ for notational simplicity. For all $z \in \mathbb{R}$, let $\widehat{F}_{\mathcal{G}}(z) := \sum_{m\in\mathcal{G}} w_m \mathbb{1}(Z_n \leq z)/W_{\mathcal{G}}$ and $\widehat{F}_{[M]}(z) := \sum_{m\in[M]} w_m \mathbb{1}(Z_m \leq z)/W_{[M]}$ be the weighted empirical distributions of $\{Z_m\}_{m\in\mathcal{G}}$ and $\{Z_m\}_{m=1}^M$, respectively. Then we have

$$\mathbb{E}[\widehat{F}_{\mathcal{G}}(z)] = \sum_{m\in\mathcal{G}} w_m \mathbb{1}(Z_m \leq z)/W_{\mathcal{G}} = \sum_{m\in\mathcal{G}} w_m \Phi\left(\frac{z-\mu}{\sigma_m}\right)/W_{\mathcal{G}}.$$

By the condition (C.5) on $\alpha_{\mathcal{B},\delta}$, there are unique values z_{high} and z_{low} such that

$$\sum_{m\in\mathcal{G}} w_m \Phi\left(\frac{z_{\text{high}}-\mu}{\sigma_m}\right) = \left(\frac{1}{2} + \alpha_{\mathcal{B},\delta}\right) W_{\mathcal{G}},$$
$$\sum_{m\in\mathcal{G}} w_m \Phi\left(\frac{z_{\text{low}}-\mu}{\sigma_m}\right) = \left(\frac{1}{2} - \alpha_{\mathcal{B},\delta}\right) W_{\mathcal{G}}.$$

By Hoeffding's inequality, for any given $\delta \geq 0$ and $z \in \mathbb{R}$, we have with probability at least $1 - e^{-2\delta}$ that

$$\widehat{F}_{\mathcal{G}}(z) - W_{\mathcal{G}}^{-1} \sum_{m\in\mathcal{G}} w_m \Phi\left(\frac{z-\mu}{\sigma_m}\right) \leq \sqrt{\delta \sum_{m\in\mathcal{G}} w_m^2}/W_{\mathcal{G}}.$$

It is not hard to verify that for all $z \in \mathbb{R}$,

$$\left|\widehat{F}_{\mathcal{G}}(z) - \widehat{F}_{[M]}(z)\right| \leq \frac{W_{\mathcal{B}}}{W_{[M]}}. \tag{C.6}$$

We thus have

$$\begin{aligned}\widehat{F}_{[M]}(z_{\text{high}}) \geq & \widehat{F}_{\mathcal{G}}(z_{\text{high}}) - \frac{W_{\mathcal{B}}}{W_{[M]}} \\ \geq & W_{\mathcal{G}}^{-1} \sum_{m\in\mathcal{G}} w_m \Phi\left(\frac{z-\mu}{\sigma_m}\right) - \frac{W_{\mathcal{B}}}{W_{[M]}} - \sqrt{\delta \sum_{m\in\mathcal{G}} w_m^2}/W_{\mathcal{G}} > \frac{1}{2}.\end{aligned}$$

Similarly, we have $\widehat{F}_{[M]}(z_{\text{low}}) < 1/2$. Further using Lemma C.1 leads to the conclusion. □

C.2 Lemma C.3 and its Proof

Lemma C.3. *Under Conditions 2 and 3, for any $0 < \eta \leq \frac{1}{5}$ and $k \in \mathcal{I}_\eta$, it holds for any $0 \leq \delta \leq W_{[M]}/(21 \max_{m\in[M]} w_m)$ that*

$$\mathbb{P}\left(|\widehat{\beta}_k^\star - \beta_k^\star| \geq 1.25 C_{0.45} \alpha_{\mathcal{B}_k,\delta} \bar{\sigma}_{[M],k} \mid \{\mathbf{X}^{(m)}\}_{m=1}^M\right) \leq 2e^{-2\delta},$$

where $\bar{\sigma}_{[M],k} = \sum_{m\in[M]} w_m \sigma_m \sqrt{[(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}}/W_{[M]}$ and $\alpha_{\mathcal{B}_k,\delta}$ follows from the definition in (C.5).

Proof. Since $W_{\mathcal{B}_k} \leq \eta W_{[M]}$ for any $k \in \mathcal{I}_\eta$ and $\eta \leq \frac{1}{5}$, we have for each $k \in \mathcal{I}_\eta$ and any $0 \leq \delta \leq 1/(25\sum_{m\in\mathcal{G}_k} \bar{w}_m^2)$ that

$$\begin{aligned}\alpha_{\mathcal{B}_k,\delta} = & \frac{W_{\mathcal{B}_k}}{W_{[M]}} + \sqrt{1.01\delta \sum_{m\in\mathcal{G}_k} w_m^2}/W_{\mathcal{G}_k} \leq \frac{W_{\mathcal{B}_k}}{W_{[M]}} + \sqrt{1.01\delta \max_{m\in[M]} w_m}/W_{\mathcal{G}_k}^{1/2} \\ \leq & \frac{W_{\mathcal{B}_k}}{W_{[M]}} + \sqrt{1.01\delta \max_{m\in[M]} w_m}/(0.5W_{[M]})^{1/2} \leq \frac{1}{5} + \frac{1}{4} = 0.45.\end{aligned}$$

Therefore, the condition (C.5) from Lemma C.2 is satisfied with $\alpha = 0.45$. Thus, by Lemma C.2, we have for any $0 \leq \delta \leq W_{[M]}/(21 \max_{m\in[M]} w_m)$ that

$$\mathbb{P}\left(|\widehat{\beta}_k^\star - \beta_k^\star| \geq C_{0.45} \bar{\alpha}_{\mathcal{B}_k,\delta} \sigma_{\mathcal{G}_k} \mid \{\mathbf{X}^{(m)}\}_{m=1}^M\right) \leq 2e^{-2\delta}. \tag{C.7}$$

Furthermore, using $W_{\mathcal{G}_k} \geq 4W_{[M]}/5$, we have

$$\bar{\sigma}_{\mathcal{G}_k} = \sum_{m\in\mathcal{G}_k} w_m \sigma_m / W_{\mathcal{G}_k} \leq 5 \sum_{m\in\mathcal{G}_k} w_m \sigma_m/(4W_{[M]}) = 1.25\bar{\sigma}_{[M],k}. \tag{C.8}$$

Combining (C.7) with (C.8) completes the proof. □

C.3 Proof of Proposition 1

Proof. For simplicity, we only prove the case $n_1 \geq \cdots \geq n_M$ and $\sigma_1 = \cdots = \sigma_M = \sigma$ for some $\sigma > 0$, and thus $w_m = n_m/\sigma^2$ for all $m \in [M]$. The case of heterogeneous variances follows by considering the rescaled sample size $\tilde{n}_m = n_m\sigma^2/\sigma_m^2$ for each $m \in [M]$.

For each $k \in [d]$, let $\bar{\sigma}_k \triangleq \sum_{m\in[M]} w_m\sqrt{v_k^{(m)}}\sigma/W_{[M]}$ where $v_k^{(m)} = \sqrt{[(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}}$. Due to Condition 4, we have

$$W_{[M]}/(21\max_{m\in[M]} w_m) \geq n_{[M]}/(21n_1) \geq (21\ln((n_{[\mathrm{M}]}/n_M) \wedge (d/s))/c_{\mathrm{s}})^2,$$

where the last inequality follows from Condition 4 and the choice of $\{w_m\}_{m=1}^M$ and c_{s} is defined in Condition 4.

Taking $\delta \triangleq (21\ln(n_{[M]}/n_M \wedge (d/s))/c_{\mathrm{s}})^2$ in Lemma C.3, we have

$$\mathbb{P}\left(|\widehat{\beta}_k^\star - \beta_k^\star| \geq 1.25C_{0.45}\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k \mid \{\mathbf{X}^{(m)}\}_{m=1}^M\right) \leq 2e^{-2(21\ln(n_{[M]}/n_M\wedge(d/s))/c_{\mathrm{s}})^2} = O\left(\frac{n_M}{n_{[M]}}\bigvee\frac{s}{d}\right). \tag{C.9}$$

On the other hand, by using a standard ϵ-net argument (see *e.g.*, Vershynin (2018)), one can show that the event

$$E \triangleq \{\mathbf{X}^{(m)\top}\mathbf{X}^{(m)} \gtrsim \mu n_m I_d/2, \forall\, m \in [M]\}$$

holds with probability at least $1 - O(Mde^{-cn_M})$ where c is a constant only depending on μ and L. Since event E implies $\sqrt{v_k^{(m)}} = O(1/\sqrt{n_m})$ and thus

$$\bar{\sigma}_k = O\left(\sum_{m\in[M]}\sqrt{n_m}/\sigma_m/\sum_{m\in[M]} n_m/\sigma_m^2\right),$$

combining with (C.9), we complete the proof.

□

C.4 Proof of Theorem 1

We provide the proof for $p = 1$ and Option I under identical variances, *i.e.*, $\sigma_1 = \cdots = \sigma_M = \sigma$; the case $p = 2$, Option II, or heterogeneous noise variances follows similarly. Let $\mathcal{I}^{(m)} := \{k \in [d] : \beta_k^{(m)} = \beta_k^\star\}$ and recall $\mathcal{I}_\eta$ from (2). For all $m \in [M]$, we provide a series of bounds for $|\widehat{\beta}_{\mathrm{MOLAR},k}^{(m)} - \beta_k^{(m)}|$ for each $k \in [d]$ in three cases. We denote $\widehat{\beta}_{\mathrm{MOLAR}}^{(m)}$ as $\widehat{\beta}^{(m)}$ below for simplicity. First noting that $v_k^{(m)} = O(t/\sqrt{n_m})$ with probability at least $1 - de^{-cn_mt^2}$ for some constant c (see, *e.g.*, Vershynin (2018)), we have $v_k^{(m)} = O_P(\ln(d)/\sqrt{n_m}), \forall\, m \in [M]$.

Case 1. For any $k \in [d]$, we guarantee that

$$\mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| \mid \{\mathbf{X}^{(m)}\}_{m=1}^M] = \widetilde{O}_P(\sigma/\sqrt{n_m}). \tag{C.10}$$

By definition, for any $m \in [M]$ and $k \in [d]$, $\widehat{\beta}_k^{(m)}$ is either equal to $\widehat{\beta}_{\mathrm{ind},k}^{(m)}$ or $\widehat{\beta}_k^\star$, and the latter happens only when $|\widehat{\beta}_k^\star - \widehat{\beta}_{\mathrm{ind},k}^{(m)}| \le \gamma_m \sqrt{v_k^{(m)}}$. In the latter case, we have

$$|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| = |\widehat{\beta}_k^\star - \beta_k^{(m)}| \le |\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| + |\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \widehat{\beta}_k^\star| \le |\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| + \gamma_m \sqrt{v_k^{(m)}}.$$

Therefore, in both cases,

$$|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| \le |\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}| + \gamma_m \sqrt{v_k^{(m)}}. \tag{C.11}$$

By (1), $\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)} \mid v_k^{(m)} \sim \mathcal{N}(0, \sigma^2 v_k^{(m)})$, we have $\mathbb{E}[|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}| \mid v_k^{(m)}] = O(\sigma v_k^{(m)}) = \widetilde{O}_P(\sigma/\sqrt{n_m})$.

Case 2. When $k \in \mathcal{I}^{(m)} \cap \mathcal{I}_\eta$, we can obtain the improved bound

$$\mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| \mid \{\mathbf{X}^{(m)}\}_{m=1}^M] = \widetilde{O}_P\left(\frac{W_{\mathcal{B}_k}\sigma}{W_{[M]}\sqrt{n_1}} + \frac{\sigma}{\sqrt{n_{[M]}}} + \frac{s\sigma/d}{\sqrt{n_m}}\right). \tag{C.12}$$

Let $\delta = (21 \ln(n_{[M]}/n_M \wedge (d/s))/c_{\mathrm{s}})^2$ and $\bar{\sigma}_k = \sum_{m\in[M]} w_m \sqrt{v_k^{(m)}}\sigma / W_{[M]}$ as stated in Section C.3. Without loss of generality, we consider $1.25 C_{0.45}\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k \le \sigma\sqrt{v_k^{(m)}}$, otherwise (C.12) is implied by (C.10). Define the event $\mathcal{E}_k = \{|\widehat{\beta}_k^\star - \beta_k^\star| \le 1.25 C_{0.45}\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k\}$. By Lemma C.3, we have $\mathbb{P}((\mathcal{E}_k)^c) \le O((n_M/n_{[M]}) \wedge (s/d))$. Furthermore, by the condition $1.25 C_{0.45}\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k \le \sigma\sqrt{v_k^{(m)}}$, we have that the event $\mathcal{E}_k$ implies $|\widehat{\beta}_k^\star - \beta_k^\star| \le 1.25 C_{0.45}\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k$. On the event $\mathcal{E}_k$, if $\widehat{\beta}_k^{(m)} \ne \widehat{\beta}_k^\star$, *i.e.*, $|\widehat{\beta}_k^\star - \widehat{\beta}_{\mathrm{ind},k}^{(m)}| > \gamma_m\sqrt{v_k^{(m)}}$, then, for $k \in \mathcal{I}^{(m)} \cap \mathcal{I}_\eta$,

$$\begin{aligned}|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}| = |\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^\star| \ge& |\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \widehat{\beta}_k^\star| - |\widehat{\beta}_k^\star - \beta_k^\star| > \gamma_m\sigma\sqrt{v_k^{(m)}} - 1.25 C_{0.45}\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k \\ \ge& (\gamma_m - \sigma)\sqrt{v_k^{(m)}}.\end{aligned}$$

Let $\zeta_k^{(m)} \triangleq (\gamma_m - \sigma)\sqrt{v_k^{(m)}}$ and

$$\mathcal{F}_k^{(m)} \triangleq \left\{|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}| \le \zeta_k^{(m)}\sigma\sqrt{v_k^{(m)}}\right\}.$$

The event $\mathcal{F}_k^{(m)} \cap \mathcal{E}_k$ implies that $\widehat{\beta}_k^{(m)} = \widehat{\beta}_k^\star$ for $k \in \mathcal{I}_\eta \cap \mathcal{I}^{(m)}$. Since $\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)} \mid v_k^{(m)} \sim \mathsf{subG}(v_k^{(m)}\sigma^2)$, we have $\mathbb{P}((\mathcal{F}_k^{(m)})^c \mid v_k^{(m)}) \le O((n_M/n_{[M]}) \wedge (s/d))$. Thus, with probability at least $\mathbb{P}(\mathcal{E}_k \cap \mathcal{F}_k^{(m)}) \ge 1 - O((n_M/n_{[M]}) \wedge (s/d))$, it holds that

$$|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| = |\widehat{\beta}_k^\star - \beta_k^{(m)}| \le 1.25 C_{0.45}\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k.$$

Furthermore, using (C.11) and Lemma B.9 and recalling Definition B.1, we have that for any $k \in \mathcal{I}_\eta \cap \mathcal{I}^{(m)}$,

$$\begin{aligned}\mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| \mid \{\mathbf{X}^{(m)}\}_{m=1}^M] &\le \widetilde{O}\left(\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k\right) + \mathbb{E}^{O((n_m/n_{[M]}) \vee (s/d))}\left[|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}| + \gamma_m\sqrt{v_k^{(m)}}\right] \\ &= \widetilde{O}\left(\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k\right) + \widetilde{O}\left((n_M/n_{[M]}) \vee (s/d))\sigma\sqrt{v_k^{(m)}}\right).\end{aligned} \tag{C.13}$$

Now,

$$\bar{\sigma}_k \triangleq \sum_{m\in[M]} w_m \sqrt{v_k^{(m)}}\sigma/W_{[M]} = \widetilde{O}_P\left(\sum_{m\in[M]} \sqrt{n_m}\sigma/n_{[M]}\right) \overset{a}{\leq} \widetilde{O}_P\left(\sigma/\sqrt{n_1}\right) \tag{C.14}$$

and

$$\frac{\sqrt{\sum_{m\in[M]} w_m^2}}{W_M}\bar{\sigma}_k \leq \frac{\sqrt{n_1}}{\sqrt{n_{[M]}}}\widetilde{O}_P\left(\sum_{m\in[M]} \sqrt{n_m}\sigma/n_{[M]}\right) \overset{b}{\leq} \widetilde{O}_P(\sigma/n_{[M]}), \tag{C.15}$$

where inequalities a and b are due to Condition 4. Thus, (C.14) and (C.15) lead to

$$\alpha_{\mathcal{B}_k,\delta}\bar{\sigma}_k = O_P\left(\frac{W_{\mathcal{B}_k}\sigma}{W_{[M]}\sqrt{n_1}} + \frac{\sigma}{\sqrt{n_{[M]}}}\right). \tag{C.16}$$

Combining (C.13) and (C.16), we reach (C.12).

Bounding the summed error. Combining the cases (C.10) and (C.12), we obtain

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 \mid \{\mathbf{X}^{(m)}\}_{m=1}^M] \\
&= \sum_{k\in\mathcal{I}_\eta\cap\mathcal{I}^{(m)}} \mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| \mid \{\mathbf{X}^{(m)}\}_{m=1}^M] + \sum_{k\notin\mathcal{I}^{(m)}\cap\mathcal{I}_\eta} \mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| \mid \{\mathbf{X}^{(m)}\}_{m=1}^M] \\
&\leq \sum_{k\in\mathcal{I}_\eta\cap\mathcal{I}^{(m)}} \widetilde{O}_P\left(\frac{W_{\mathcal{B}_k}\sigma}{W_{[M]}\sqrt{n_1}} + \frac{\sigma}{\sqrt{n_{[M]}}} + \frac{s\sigma/d}{\sqrt{n_m}}\right) + \sum_{k\notin\mathcal{I}^{(m)}\cap\mathcal{I}_\eta} \widetilde{O}\left(\frac{\sigma}{\sqrt{n_m}}\right) \\
&\leq \sum_{k\in\mathcal{I}_\eta} \widetilde{O}_P\left(\frac{W_{\mathcal{B}_k}\sigma}{W_{[M]}\sqrt{n_m}}\right) + \sum_{k\notin\mathcal{I}_\eta} \widetilde{O}_P\left(\frac{\sigma}{\sqrt{n_m}}\right) + \widetilde{O}_P\left(\frac{s\sigma}{\sqrt{n_m}} + \frac{d\sigma}{\sqrt{n_{[M]}}}\right).
\end{aligned} \tag{C.17}$$

where the last inequality is due to $|(\mathcal{I}^{(m)})^c| \leq s$ and $n_1 \geq n_m$ for any $m \in [M]$. Using Lemma B.10 with $a = 1$ and $x_k = W_{\mathcal{B}_k}/W_{[M]}$ for all $k \in [d]$, we have

$$\sum_{k\in[d]} \left(\frac{W_{\mathcal{B}_k}}{W_{[M]}}\mathbb{1}(W_{\mathcal{B}_k}/W_{[M]} < \eta) + \mathbb{1}(W_{\mathcal{B}_k}/W_{[M]} \geq \eta)\right) \leq \lceil s/\eta \rceil. \tag{C.18}$$

Plugging $\eta = 1/5 = O(1)$ into (C.18) and combining (C.17), we have

$$\mathbb{E}[\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 \mid \{\mathbf{X}^{(m)}\}_{m=1}^M] = \widetilde{O}_P\left(\frac{s\sigma}{\sqrt{n_m}} + \frac{d\sigma}{\sqrt{n_{[M]}}}\right).$$

Using Chebyshev's inequality, we obtain

$$\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 = \widetilde{O}_P\left(\frac{s\sigma}{\sqrt{n_m}} + \frac{d\sigma}{\sqrt{n_{[M]}}}\right).$$

Similarly for $p = 2$, we can establish

$$\mathbb{E}[\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2^2 \mid \{\mathbf{X}^{(m)}\}_{m=1}^M] = \widetilde{O}_P\left(\frac{s\sigma^2}{n_m} + \frac{d\sigma^2}{n_{[M]}}\right).$$

C.5 Proof of Theorem 2

Proof. As discussed in Section 2.3, we prove Theorem 2 by considering two special cases of our sparse heterogeneity model:

1. The *homogeneous* case where $\beta^1 = \cdots = \beta^M = \beta^\star \in \mathbb{R}^d$.
2. The *s-sparse* case where $\beta^\star = 0$ and $\|\beta^m\|_0 \le s$ for all $m \in [M]$.

Therefore, clearly

$$\mathcal{M} \ge \inf_{\widehat{\beta}^{(m)}} \sup_{\substack{\beta^\star \in \mathbb{R}^d,\ \{\beta^{(m)}\}_{m=1}^M \subseteq \mathbb{B}_s(\beta^\star) \\ \{\Sigma^{(m)}\}_{m=1}^M \subseteq \mathcal{A}_{\mu,L}}} \|\widehat{\beta}^\star - \beta^\star\|_p^p \triangleq A \tag{C.19}$$

and

$$\mathcal{M} \ge \inf_{\widehat{\beta}^{(m)}} \sup_{\substack{\beta^\star \in \mathbb{R}^d,\ \{\beta^{(m)}\}_{m=1}^M \subseteq \mathbb{B}_s(\beta^\star) \\ \{\Sigma^{(m)}\}_{m=1}^M \subseteq \mathcal{A}_{\mu,L}}} \|\widehat{\beta}^{(m)} - \beta^{(m)}\|_p^p \triangleq B. \tag{C.20}$$

We will show that $A = \widetilde{\Omega}_P(d/(\sum_{m=1}^M n_m/\sigma_m^2)^{p/2})$ and $B = \widetilde{\Omega}_P(s\sigma_m^p/(n_m)^{p/2})$. Then the conclusion follows from the inequality $A \vee B \ge (A+B)/2 = \Omega(A+B)$.

The case where $p = 2$: The proof follows the same idea as Example 8.4.5 of Duchi (2019) and the argument in Duchi and Wainwright (2013). These show a lower bound $\Omega(d\sigma^2/\|\mathbf{X}^\top\mathbf{X}\|_{\mathrm{op}})$ for linear regression with a given covariate matrix $\mathbf{X}$ and i.i.d. $\mathcal{N}(0,\sigma^2)$ noises. Here, we sketch the key ideas for the extension to different noise variances.

Let $\mathcal{V}$ be a packing of $\{-1,1\}^d$ such that $\|v - v'\| \ge d/2$ for distinct elements of $\mathcal{V}$, and $|\mathcal{V}| \ge \exp(d/8)$ as guaranteed by the Gilbert-Varshamov bound (Duchi, 2019, Example 7.5). For fixed $\delta > 0$, if we set $\beta_v^\star = \delta v$, then we have the packing guarantee for distinct elements v, v' that

$$\|\beta_v^\star - \beta_{v'}^\star\|_1 = \delta \sum_{k=1}^d |v_k - v_k'| \ge \delta d/2.$$

Then we have an upper bound on the Kullback-Leibler divergence of the data distributions associated with $\beta_v^\star$, $\beta_{v'}^\star$, a feature vector x, and standard deviation σ:

$$D_{\mathrm{KL}}(\mathcal{N}(x^\top\beta_v^\star, \sigma^2) \,\|\, \mathcal{N}(x^\top\beta_{v'}^\star, \sigma^2)) = \frac{1}{2\sigma^2}\|x^\top(\beta_v^\star - \beta_{v'}^\star)\|_2^2.$$

Consequently, given the independent observations $\mathbf{Y}_v \triangleq \{\{x_i^{(m)\top}\beta_v^\star + \epsilon_i^{(m)}\}_{i=1}^{n_m}\}_{m=1}^M$ where $\epsilon_i^{(m)} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_m^2)$ and $\mathbf{Y}_{v'}$, the Kullback-Leibler divergence is

$$\begin{aligned} D_{\mathrm{KL}}(P(\mathbf{Y}_v \mid \{\mathbf{X}^{(m)}\}_{m=1}^M) \,\|\, P(\mathbf{Y}_{v'} \mid \{\mathbf{X}^{(m)}\}_{m=1}^M)) &= \sum_{m=1}^M \frac{1}{2\sigma_m^2}\|\mathbf{X}^{(m)}(\beta_v^\star - \beta_{v'}^\star)\|_2^2 \\ &\le \sum_{m=1}^M \frac{2d\delta^2}{\sigma_m^2}\|\mathbf{X}^{(m)\top}\mathbf{X}^{(m)}\|_{\mathrm{op}} \end{aligned} \tag{C.21}$$

where the last inequality holds because $\|v - v'\|_2^2 \le 4d$. Now we apply the local Fano method (Duchi, 2019, Proposition 8.43), we obtain

$$A \ge \frac{\delta d}{4}\left(1 - \frac{I(V;\mathbf{Y}) + \ln(2)}{\ln(|\mathcal{V}|)}\right),$$

where $I(V;\mathbf{Y})$ is the mutual information between the index variable $V \sim \mathrm{Unif}(\mathcal{V})$ and the responses $\mathbf{Y} = \{Y^{(m)}\}_{m\in[M]}$. By choosing $\delta = 8^{-1}\left(\sum_{m=1}^M \|\mathbf{X}^\top\mathbf{X}\|_{\mathrm{op}}/\sigma_m^2\right)^{-1/2}$, following (Duchi, 2019, Example 7.5), we find that $1 - (I(V;\mathbf{Y}) + \ln(2))/\ln(|\mathcal{V}|) \ge 1/2$. This leads to $A \mid \{\mathbf{X}^{(m)}\}_{m=1}^M = \Omega(d\left(\sum_{m=1}^M \|\mathbf{X}^\top\mathbf{X}\|_{\mathrm{op}}/\sigma_m^2\right)^{-1/2})$. Finally, noting $\|\mathbf{X}^\top\mathbf{X}\|_{\mathrm{op}} = \widetilde{O}_P(n_m)$, we conclude $A = \widetilde{\Omega}_P(d/(\sum_{m=1}^M n_m/\sigma_m^2)^{p/2})$.

By applying the same argument to a single task with s-dimensional parameters, one can readily show $B = \widetilde{\Omega}_P(s\sigma_m^p/(n_m)^{p/2})$.

The case where $p = 1$: The proof for the ℓ_1 case follows the same workflow by utilizing $\|\beta_v^\star - \beta_{v'}^\star\|_1 = \delta\sum_{k=1}^d |v_k - v_k'| \ge \delta d/2$. We then obtain $A = \Omega(d\sigma^2/(\mu\sum_{\ell\in[M]} n^\ell)^{1/2})$ and thus $B = \Omega(s\sigma/(\mu n_m)^{1/2})$ accordingly.

□

D Results on Linear Regression with General Noise

Lemma C.2, the cornerstone of our previous analysis, relies on the Gaussianity of the noise $\{\varepsilon_i^{(m)}\}_{i=1}^{n_m}$ for each $m \in [M]$. In this section, we argue that the estimation error of $\widehat{\beta}^\star$ holds up to a smaller-order term, even if we relax Gaussianity to bounded skewness and Orlicz norm. While this holds for unbalanced datasets, for simplicity, we only state the results for balanced datasets, *i.e.*, $n_m = n$ and $\sigma_m = \sigma$ for all $m \in [M]$. Recall that for a random variable Z, and for $\alpha \ge 0$, the α-order Orlicz norm of $\varepsilon_i^{(m)}$, $1 \le i \le n$, is

$$\|Z\|_{\Psi_\alpha} := \inf\left\{c > 0 : \mathbb{E}\left[\Psi_\alpha\left(|Z|/c\right)\right] \le 1\right\},$$

where $\Psi_\alpha(u) = e^{u^\alpha} - 1$ for any $u \ge 0$. We make the following technical assumption, replacing Condition 2.

Condition D.1 (Bounded covariates, skewness, and Orlicz norm)**.** *There are constants $c_{\mathrm{sk}}, L \ge \mu \ge 0$, $\alpha \in (0, 2]$, and $\sigma \ge 0$ such that for each $m \in [M]$,*

1. *the covariates satisfy $\mu n I_d \preceq \mathbf{X}^{(m)\top}\mathbf{X}^{(m)} \preceq nLI_d$ and $\|x_i^{(m)}\|_2^2 \le dL$ for all $1 \le i \le n$;*
2. *the noise variables $\varepsilon_i^{(m)}$, $1 \le i \le n$, are i.i.d. with zero mean, satisfying $\mathbb{E}[|\varepsilon_i^{(m)}|^3] \le c_{\mathrm{sk}}(\mathbb{E}[|\varepsilon_i^{(m)}|^2])^{3/2}$;*
3. *the α-order Orlicz norm of $\varepsilon_i^{(m)}$, $1 \le t \le n$, is bounded as $\|\varepsilon_i^{(m)}\|_{\Psi_\alpha} \le \sigma$.*

Remark D.1. *In Condition D.1, the first condition strengthens Condition 3, and is used to control the coefficients involved in a normal approximation via the Berry-Esseen theorem*

(Berry, 1941). This condition holds with a high probability for randomly sampled design matrices with i.i.d. features having a well-conditioned expected covariance matrix (Vershynin, 2018), including the stochastic contexts in bandits that we will use. The second condition, while less standard in the literature, is also needed in the Berry-Esseen theorem. The third condition generalizes the widely used sub-Gaussianity condition (Tripuraneni et al., 2021; Han et al., 2020; Xu and Bastani, 2021; Ren and Zhou, 2023). This allows for many choices of random noise, such as bounded noises (for any $\alpha > 0$), sub-Gaussian noises (for $\alpha = 2$), or sub-exponential noises (for $\alpha = 1$) like Poisson random variables, or noise with heavier tails such as Weibull random variables with shape parameter $\alpha \in (0, 1]$. A smaller α allows for a heavier tail.

Under Condition D.1, we show that the c.d.f. of each normalized OLS estimate is approximately Gaussian up to a vanishing factor of $\sqrt{d/n}$. The key idea is to use Gaussian approximation via the Berry-Esseen theorem (Lemma B.5). Based on this, we control the error of the global estimate $\widehat{\beta}^\star$ in Proposition D.1. Compared to Lemma C.3 obtained under Gaussian noise, the upper bound incurs an additional vanishing term of order $O(\sqrt{d}/n)$. The proof is in Appendix D.2.

Proposition D.1. *Under Condition D.1, for any $k \in [d]$ and $m \in [M]$, the c.d.f. $F_k^{(m)}$ of $\widehat{\beta}_{\mathrm{ind},k}^{(m)}$ (after standardization) satisfies*

$$\sup_{z\in\mathbb{R}} |F_k^{(m)}(z) - \Phi(z)| \le \frac{0.6 c_{\mathrm{sk}} L\sqrt{d}}{\mu\sqrt{n}}.$$

Further, if $n \ge 36d(c_{\mathrm{sk}}L/\mu)^2$, for any $0 < \eta \le \frac{1}{10}$ and $k \in \mathcal{I}_\eta$, it holds for any $0 \le \delta \le M/20$ that

$$\mathbb{P}\left(|\widehat{\beta}_k^\star - \beta_k^\star| \ge 1.25 C_{0.45}\bar{\sigma}_k \alpha'_{\mathcal{B}_k,\delta}\right) \le 4e^{-2\delta},$$

where C_{be} is an abolute constant, $\bar{\sigma}_k = \sigma_{\mathrm{v}} \sum_{m=1}^M \sqrt{v_k^{(m)}}/M$ for each $k \in [d]$ with $\sigma_{\mathrm{v}}^2 \triangleq \mathrm{Var}(\varepsilon_i^{(m)})$, and $\alpha'_{\mathcal{B}_k,\delta} \triangleq |\mathcal{B}_k|/M + \sqrt{1.01\delta/(M - |\mathcal{B}_k|)} + 0.6 c_{\mathrm{sk}} L\sqrt{d}/(\mu\sqrt{n})$.

The Berry-Esseen theorem helps control the estimation error of the median-based estimator $\widehat{\beta}^\star$ with high probability. To control the in-expectation estimation error of the final estimates $\{\widehat{\beta}_{\mathrm{ft}}^{(m)}\}_{m\in[M]}$, we also need the concentration of individual OLS estimates. This is guaranteed by the generalized Hanson-Wright inequality (Lemma B.6), which characterizes the tail of quadratic forms of random variables with bounded Orlicz norm. Combining these two ingredients, we can bound the estimation errors of $\{\widehat{\beta}_{\mathrm{ft}}^{(m)}\}_{m\in[M]}$ in Theorem D.1; with a proof in Appendix D.3.

Theorem D.1. *Under Condition 1 and D.1, for any $p \in \{1, 2\}$, $m \in [M]$, with $\widehat{\beta}_{\mathrm{MOLAR}}^{(m)}$ from Algorithm 1, it holds that*

$$\|\widehat{\beta}_{\mathrm{MOLAR}}^{(m)} - \beta^{(m)}\|_p^p = \widetilde{O}_P\left(\frac{\sigma^p}{n^{p/2}}\left(s + \frac{d}{M^{p/2}} + \frac{d^{1+p/2}}{n^{p/2}}\right)\right), \tag{D.1}$$

where $\widetilde{O}_P(\cdot)$ absorbs a Polylog(M, μ, L) *factor with degree and coefficients depending only on α.*

Consequently, if $\alpha = \Theta(1)$ and $n = \Omega(d(M \vee (d/s)^{2/p}))$, (D.1) matches (4) from Theorem 1.

D.1 Preliminaries

To establish the proofs for Proposition D.1 and Theorem D.1, we first prove some technical lemmas. Below, we view α as an absolute constant. Let σ_{v}^2 be the variance of noise $\varepsilon_i^{(m)}$. Under the assumtption that $\|\varepsilon_i^{(m)}\|_{\Psi_\alpha} \le \sigma$, we first establish the following results.

Lemma D.1 (Relation between σ_{v} and σ). *It holds that $\sigma_{\mathrm{v}}^2 \le 2\Gamma(1+2/\alpha)\sigma^2 = O(\sigma^2)$ where $\Gamma(\cdot)$ is the gamma function.*

Proof. By Lemma B.1, we have $\sigma_{\mathrm{v}}^2 = \mathbb{E}[(\varepsilon_i^{(m)})^2] = \int_0^\infty \mathbb{P}((\varepsilon_i^{(m)})^2 \ge t)\mathrm{d}t = \int_0^\infty \mathbb{P}(|\varepsilon_i^{(m)}| \ge \sqrt{t})\mathrm{d}t$. Using Lemma B.4, we have $\mathbb{P}(|\varepsilon_i^{(m)}| \ge \sqrt{t}) \le 2\exp(-(\sqrt{t}/\sigma)^\alpha)$. Therefore, by the change of variables $u = \sqrt{t}/\sigma$, we have

$$\begin{aligned}\sigma_{\mathrm{v}}^2 =& \int_0^\infty \mathbb{P}(|\varepsilon_i^{(m)}| \ge \sqrt{t})\mathrm{d}t \le 2\int_0^\infty \exp(-(\sqrt{t}/\sigma)^\alpha)\mathrm{d}t \\ =& 4\sigma^2 \int_0^\infty \exp(-u^\alpha)u\mathrm{d}u = 2\Gamma(1+2/\alpha)\sigma^2.\end{aligned}$$

□

Lemma D.2 (Tail and integral of OLS estimate). *For each $m \in [M]$, $k \in [d]$, and $u > 0$, it holds that*

$$\mathbb{P}\left(|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}| \ge (\sigma_{\mathrm{v}} + u\sigma)\sqrt{v_k^{(m)}}\right) \le 2\exp\left(-\frac{u^\alpha}{c_{\mathrm{hw}}}\right). \tag{D.2}$$

where c_{hw} is an absolute constant in Lemma B.6. Furthermore, we have for any $p \in \{1,2\}$ and $\delta \in [0,1]$ with $\ln(2/\delta) \ge 1/\alpha$ that

$$\mathbb{E}^\delta[|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}|^p] = O\left(\delta\sigma^p \ln(2/\delta)^{p/\alpha}\sqrt{v_k^{(m)}}\right). \tag{D.3}$$

Proof. Recall from (D.9) that $\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)} = \sum_{i=1}^n \langle w_k^{(m)}, x_i^{(m)}\rangle \varepsilon_i^{(m)}$ with $w_k^{(m)\top} \triangleq (\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})_{k,\cdot}^{-1}$ and $\sum_{i=1}^n \langle w_k^{(m)}, x_i^{(m)}\rangle^2 = v_k^{(m)}$. By Lemma B.6, we have for any $t \ge \sigma^2 v_k^{(m)}$ that

$$\mathbb{P}\left(\left|\left(\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}\right)^2 - \sigma_{\mathrm{v}}^2 v_k^{(m)}\right| \ge t\right) \le 2\exp\left(-\frac{1}{c_{\mathrm{hw}}}\left(\frac{t}{\sigma^2 v_k^{(m)}}\right)^{\alpha/2}\right).$$

Letting $t = u^2\sigma^2 v_k^{(m)}$ and using $\sigma_{\mathrm{v}} + u\sigma \ge \sqrt{\sigma_{\mathrm{v}}^2 + u^2\sigma^2}$ for any $u \ge 0$, we have

$$\begin{aligned}&\mathbb{P}\left(|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}| \ge (\sigma_{\mathrm{v}} + u\sigma)\sqrt{v_k^{(m)}}\right) \le \mathbb{P}\left(|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}| \ge \sqrt{\sigma_{\mathrm{v}}^2 + u^2\sigma^2}\sqrt{v_k^{(m)}}\right) \\ \le&\mathbb{P}\left(\left|\left(\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \beta_k^{(m)}\right)^2 - \sigma_{\mathrm{v}}^2 v_k^{(m)}\right| \ge u^2\sigma^2 v_k^{(m)}\right) \le 2\exp\left(-\frac{u^\alpha}{c_{\mathrm{hw}}}\right).\end{aligned}$$

We thus obtain (D.2). For (D.3), we only analyze the case $p = 1$ and the case $p = 2$ follows. To this end, we follow the proof of Lemma B.9. By using the smoothing technique

in Lemma B.9, it suffices to consider noise to be continuous, in which case so is $\widehat{\beta}^{(m)}_{\mathrm{ind},k} - \beta^{(m)}_k$. Let $Z \triangleq \widehat{\beta}^{(m)}_{\mathrm{ind},k} - \beta^{(m)}_k$. From (B.2), $\mathbb{E}^\delta[|Z|]$ is given by the integral over the upper δ-level set $A_\delta = \{|Z| \geq q_\delta\}$ with $\mathbb{P}(|Z(\omega)| \geq q_\delta) = \delta$.

From (D.2), we readily find that

$$q_\delta \leq (\sigma_{\mathrm{v}} + c_{\mathrm{hw}}^{1/\alpha} \ln(2/\delta)^{1/\alpha}\sigma)\sqrt{v_k^{(m)}}. \tag{D.4}$$

Plugging (D.4) and (D.2) into Lemma B.1, and using the change of variables $t = (\sigma_{\mathrm{v}} + uc_{\mathrm{hw}}^{1/\alpha}\sigma)\sqrt{v_k^{(m)}}$ and $q_\delta = (\sigma_{\mathrm{v}} + u_0 c_{\mathrm{hw}}^{1/\alpha}\sigma)\sqrt{v_k^{(m)}}$ with $u_0 \leq \ln(2/\delta)^{1/\alpha}$, we obtain

$$\begin{aligned}\mathbb{E}^\delta[|Z|] &= \mathbb{E}[|Z|\mathbb{1}(|Z| \geq q_\delta)] = q_\delta \mathbb{P}(|Z| \geq q_\delta) + \int_{q_\delta}^\infty \mathbb{P}(|Z| \geq t)\mathrm{d}t \\ &\leq \delta q_\delta + c_{\mathrm{hw}}^{1/\alpha}\sigma\sqrt{v_k^{(m)}} \int_{u_0}^\infty \min\left\{\delta, 2\exp\left(-u^\alpha\right)\right\} \mathrm{d}u. \end{aligned} \tag{D.5}$$

where the last inequality follows from (D.2) and from $\mathbb{P}(|Z| \geq t) \leq \mathbb{P}(|Z| \geq q_\delta) = \delta$. Now, we calculate that

$$\int_{u_0}^\infty \min\left\{\delta, 2\exp\left(-u^\alpha\right)\right\} \mathrm{d}u = \delta\left(\ln(2/\delta)^{1/\alpha} - u_0\right) + 2\int_{\ln(2/\delta)^{1/\alpha}}^\infty \exp\left(-u^\alpha\right) \mathrm{d}u \tag{D.6}$$

Note that

$$\int_{\ln(2/\delta)^{1/\alpha}}^\infty e^{-u^\alpha} \mathrm{d}u = \frac{1}{\alpha}\Gamma(1/\alpha, \ln(2/\delta)) \leq \frac{\delta}{2\alpha}\ln(2/\delta)^{1/\alpha}, \tag{D.7}$$

where the inequality is due to $\Gamma(a,b) \leq b^a e^{-b}/(b+1-a) \leq b^a e^{-b}$ for any $b \geq a > 0$. Combining (D.6) and (D.7) with (D.5), we therefore obtain

$$\begin{aligned}\mathbb{E}^\delta[|Z|] &\leq \delta q_\delta + c_{\mathrm{hw}}^{1/\alpha}\sigma\sqrt{v_k^{(m)}}\left(\delta\left(\ln(2/\delta)^{1/\alpha} - u_0\right) + \frac{\delta}{2\alpha}\ln(2/\delta)^{1/\alpha}\right) \\ &= O\left(\delta(\sigma_{\mathrm{v}} + \sigma\ln(2/\delta)^{1/\alpha})\sqrt{v_k^{(m)}}\right) = O\left(\delta\sigma\ln(2/\delta)^{1/\alpha}\sqrt{v_k^{(m)}}\right).\end{aligned}$$

□

Without loss of generality, we consider $\sigma_{\mathrm{v}} > 0$ in the proofs of this section. Otherwise, the setup becomes noiseless, and individual OLS estimates recover the parameters directly.

D.2 Proof of Proposition D.1

Proof. We have that $\widehat{\beta}^{(m)}_{\mathrm{ind}} = (\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}\mathbf{X}^{(m)\top}Y^{(m)} = \beta^{(m)} + (\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}\mathbf{X}^{(m)\top}\boldsymbol{\varepsilon}^{(m)}$ where $\boldsymbol{\varepsilon}^{(m)} = (\varepsilon_1^{(m)}, \ldots, \varepsilon_n^{(m)})^\top$ is the noise vector. Using Condition D.1, we have, for each covariate $k \in [d]$,

$$\mathrm{Var}(\widehat{\beta}^{(m)}_{\mathrm{ind},k}) = \sigma_{\mathrm{v}}^2 \|(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}_{k,\cdot}\mathbf{X}^{(m)\top}\|_2^2 = \sigma_{\mathrm{v}}^2 v_k^{(m)}. \tag{D.8}$$

This also implies

$$\widehat{\beta}_{\mathrm{ind},k}^{(m)} = \beta_k^{(m)} + \sum_{t=1}^{n} \langle w_k^{(m)}, x_i^{(m)} \rangle \varepsilon_i^{(m)} \tag{D.9}$$

where $w_k^{(m)} \triangleq (\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})_{k,\cdot}^{-\top}$ and $w_k^{(m)\top}\mathbf{X}^{(m)\top}\mathbf{X}^{(m)}w_k^{(m)} = v_k^{(m)}$. Using Condition D.1, we have $(\mu n)^{-1} \geq v_k^{(m)} = w_k^{(m)\top}\mathbf{X}^{(m)\top}\mathbf{X}^{(m)}w_k^{(m)} \geq \mu n \|w_k^{(m)}\|_2^2$ and thus $\|w_k^{(m)}\|_2 \leq (\mu n)^{-1}$. Thus, by further using Condition D.1, we obtain

$$\begin{aligned}
&\mathbb{E}[|\langle w_k^{(m)}, x_i^{(m)} \rangle \varepsilon_i^{(m)}|^3] = |\langle w_k^{(m)}, x_i^{(m)} \rangle|^3 \mathbb{E}[|\varepsilon_i^{(m)}|^3] \\
\leq &\|w_k^{(m)}\|_2 \|x_i^{(m)}\|_2 |\langle w_k^{(m)}, x_i^{(m)} \rangle|^2 c_{\mathrm{sk}} \sigma_{\mathrm{v}}^3 \leq \sqrt{dL} c_{\mathrm{sk}} \langle w_k^{(m)}, x_i^{(m)} \rangle^2 \sigma_{\mathrm{v}}^3/(\mu n).
\end{aligned} \tag{D.10}$$

Summing up (D.10) with respect all $t \in [n]$, we find

$$\sum_{i=1}^{n} \mathbb{E}[|\langle w_k^{(m)}, x_i^{(m)} \rangle \varepsilon_i^{(m)}|^3] \leq \sqrt{dL} c_{\mathrm{sk}} \sigma_{\mathrm{v}}^3/(\mu n) \sum_{i=1}^{n} \langle w_k^{(m)}, x_i^{(m)} \rangle^2 = \sqrt{dL} c_{\mathrm{sk}} \sigma_{\mathrm{v}}^3/(\mu n) v_k^{(m)}. \tag{D.11}$$

Therefore, plugging the bounds (D.8) and (D.11) into Lemma B.5, we find

$$\sup_{z \in \mathbb{R}} |F_k^{(m)}(z) - \Phi(z)| \leq C_{\mathrm{be}} \frac{\sum_{t=1}^{n} \mathbb{E}[|\langle w_k^{(m)}, x_t^{(m)} \rangle \varepsilon_i^{(m)}|^3]}{\mathrm{Var}(\widehat{\beta}_{\mathrm{ind},k}^{(m)})^{3/2}} \leq 0.6 \frac{\sqrt{dL} c_{\mathrm{sk}} v_k^{(m)} \sigma_{\mathrm{v}}^3/(\mu n)}{\sigma_{\mathrm{v}}^2 v_k^{(m)} \sigma_{\mathrm{v}}/\sqrt{Ln}} = \frac{0.6 c_{\mathrm{sk}} L \sqrt{d}}{\mu \sqrt{n}}.$$

Next, we use this to control the estimation error of $\widehat{\beta}^\star$. The analysis is similar to the one of Lemma C.2. For each $k \in [d]$ and $z \in \mathbb{R}$, let $\widehat{F}_{\mathcal{G}_k}(z) := \frac{1}{|\mathcal{G}_k|} \sum_{m \in \mathcal{G}_k} \mathbb{1}(\widehat{\beta}_{\mathrm{ind},k}^{(m)} \leq z)$ and $\widehat{F}_{[M]}(z) := \frac{1}{M} \sum_{m \in [M]} \mathbb{1}(\widehat{\beta}_{\mathrm{ind},k}^{(m)} \leq z)$ be the empirical distribution of $\{\widehat{\beta}_{\mathrm{ind},k}^{(m)}\}_{m \in \mathcal{G}_k}$ with $\mathcal{G}_k \triangleq [M] \backslash \mathcal{B}_k$ and $\{\widehat{\beta}_{\mathrm{ind},k}^{(m)}\}_{m \in [M]}$, respectively. Since, for any $m \in [M] \backslash \mathcal{G}_k$, $\widehat{\beta}_{\mathrm{ind},k}^{(m)}$ has mean $\beta_k^\star$ and variance $\sigma_{\mathrm{v}}^2 v_k^{(m)}$, we have

$$\mathbb{E}[\widehat{F}_{\mathcal{G}_k}(z)] = \frac{1}{|\mathcal{G}_k|} \sum_{m \in \mathcal{G}_k} \mathbb{P}(\widehat{\beta}_{\mathrm{ind},k}^{(m)} \leq z) = \frac{1}{|\mathcal{G}_k|} \sum_{m \in [M] \backslash \mathcal{B}_k} F_k^{(m)} \left(\frac{z - \beta_k^\star}{\sigma_{\mathrm{v}} \sqrt{v_k^{(m)}}} \right).$$

Let z_1 be the value such that

$$\frac{1}{|\mathcal{G}_k|} \sum_{m \in \mathcal{G}_k} \Phi \left(\frac{z_1 - \beta_k^\star}{\sigma_{\mathrm{v}} \sqrt{v_k^{(m)}}} \right) = \frac{1}{2} + \alpha'_{\mathcal{B}_k, \delta},$$

where $\alpha'_{\mathcal{B}_k,\delta} \triangleq |\mathcal{B}_k|/M + \sqrt{1.01\delta/(M - |\mathcal{B}_k|)} + 0.6 c_{\mathrm{sk}} L \sqrt{d}/(\mu\sqrt{n})$. By Hoeffding's inequality, for any $\delta \geq 0$ and $z \in \mathbb{R}$, we have with probability at least $1 - 2e^{-2\delta}$ that

$$\left| \widehat{F}_{\mathcal{G}_k}(z) - \frac{1}{|\mathcal{G}_k|} \sum_{m \in \mathcal{G}_k} F_k^{(m)} \left(\frac{z - \beta_k^\star}{\sigma \sqrt{v_k^{(m)}}} \right) \right| \leq \sqrt{\frac{\delta}{M - |\mathcal{B}_k|}}.$$

Similar to (C.6), it is clear that $|\widehat{F}_{\mathcal{G}_k}(z) - \widehat{F}_{[M]}(z)| \le |\mathcal{B}_k|/M$ for all $z \in \mathbb{R}$. Combing the above properties and using a union bound, we have with probability at least $1 - 4e^{-2\delta}$ that

$$\begin{aligned}\widehat{F}_{[M]}(z_1) \ge & \widehat{F}_{[M]}(z_1) - \widehat{F}_{\mathcal{G}_k}(z_1) + \widehat{F}_{\mathcal{G}_k}(z_1) - \mathbb{E}[\widehat{F}_{\mathcal{G}_k}(z_1)] \\ & + \frac{1}{|\mathcal{G}_k|}\sum_{m\in\mathcal{G}_k}\left(F_k^{(m)}\left(\frac{z_1-\beta_k^\star}{\sigma_{\mathrm{v}}\sqrt{v_k^{(m)}}}\right) - \Phi\left(\frac{z_1-\beta_k^\star}{\sigma_{\mathrm{v}}\sqrt{v_k^{(m)}}}\right)\right) + \frac{1}{|\mathcal{G}_k|}\sum_{m\in\mathcal{G}_k}\Phi\left(\frac{z_1-\beta_k^\star}{\sigma_{\mathrm{v}}\sqrt{v_k^{(m)}}}\right) \\ \ge & -\frac{|\mathcal{B}_k|}{M} - \sqrt{\frac{\delta}{M-|\mathcal{B}_k|}} - \frac{0.6c_{\mathrm{sk}}L\sqrt{d}}{\mu\sqrt{n}} + \frac{1}{2} + \alpha'_{\mathcal{B}_k,\delta} > \frac{1}{2}.\end{aligned}$$

This implies $\widehat{\beta}_k^\star = \mathrm{Med}(\{\widehat{\beta}_{\mathrm{ind},k}^{(m)}\}_{m\in[M]}) \le z_1$. By a similar argument to the proof of Lemmas C.2 and C.3, one can verify $\alpha'_{\mathcal{B}_k,\delta} < 0.45$ and thus upper bound z_1 as $\widehat{\beta}_k^\star \le z_1 \le \beta_k^\star + \tilde{\sigma}_k G'_{[M],\mathcal{B}_k,\delta,n} C_\varepsilon$. Similarly, it also holds that $\widehat{\beta}_k^\star \ge \mu - 1.25C_{0.45}\bar{\sigma}_k\alpha'_{\mathcal{B}_k,\delta}$, finishing the proof. □

D.3 Proof of Theorem D.1

Proof. We provide the proof for $p = 1$; the case $p = 2$ follows similarly. Let $\mathcal{I}^{(m)} = \{k \in [d] : \beta_k^{(m)} = \beta_k^\star\}$ for each $m \in [M]$. We denote $\widehat{\beta}_{\mathrm{MOLAR}}^{(m)}$ as $\widehat{\beta}^{(m)}$ below for simplicity. For each $m \in [M]$, we bound $\mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}|]$ for $k \in [d]$ in two cases.

Case 1. For any $k \in [d]$ (in particular, for $k \notin \mathcal{I}_\eta$ or $k \notin \mathcal{I}^{(m)}$), since $\gamma_m = \widetilde{O}(\sigma)$, following the argument for Case 1 of Theorem 1 and using Lemma D.1, we readily obtain

$$\mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}|] = O\left(\sigma_{\mathrm{v}}\sqrt{v_k^{(m)}}\right) + \widetilde{O}\left(\sigma\sqrt{v_k^{(m)}}\right) = \widetilde{O}(\sigma/\sqrt{n}), \tag{D.12}$$

where the last inequality is due to Lemma D.1. If $n < 100d(C_{\mathrm{be}}c_{\mathrm{sk}}L/\mu)^2$ or $M \le 20\ln(M) \vee (\alpha^{-1} + \ln(3))$, by summing up (D.12) with respect to all $k \in [d]$, we obtain

$$\mathbb{E}[\|\widehat{\beta}_k^{(m)} - \beta^{(m)}\|_1] = \widetilde{O}(d\sigma/\sqrt{n}).$$

This yields the conclusion in (D.1). Therefore, we next assume $n \ge 100d(C_{\mathrm{be}}c_{\mathrm{sk}}L/\mu)^2$ and $M \ge 20\ln(M) \vee \alpha^{-1}$.

Case 2. When $k \in \mathcal{I}^{(m)} \cap \mathcal{I}_\eta$, we will show

$$\mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}|] = \widetilde{O}\left(\frac{\sigma}{\sqrt{\mu n}}\left(\frac{|\mathcal{B}_k|}{M} + \frac{1}{\sqrt{M}} + \frac{\sqrt{d}}{\sqrt{n}}\right)\right). \tag{D.13}$$

Let $\delta = \ln(M) \vee (\alpha^{-1} + \ln(3)) = \widetilde{O}(1)$. If $1.25C_{0.45}\tilde{\sigma}_k\alpha'_{\mathcal{B}_k,\delta} \ge \sigma_{\mathrm{v}}\sqrt{v_k^{(m)}}$, then from (D.12), we directly conclude (D.13). Otherwise suppose $1.25C_{0.45}\tilde{\sigma}_k\alpha'_{\mathcal{B}_k,\delta} \le \sigma_{\mathrm{v}}\sqrt{v_k^{(m)}}$. Define the event $\mathcal{E}_k = \{|\widehat{\beta}_k^\star - \beta_k^\star| \le 1.25C_{0.45}\tilde{\sigma}_k\alpha'_{\mathcal{B}_k,\delta}\}$. Since $M \ge 20\delta$, using Proposition D.1, we have $\mathbb{P}((\mathcal{E}_k)^c) \le 4e^{-\delta}$. Furthermore, by the condition $G'_{[M],\mathcal{B}_k,\delta,n}\tilde{\sigma}_k \le \sigma_{\mathrm{v}}\sqrt{v_k^{(m)}}$, we have that the

event $\mathcal{E}_k$ implies $|\widehat{\beta}_k^\star - \beta_k^\star| \le \sigma_{\rm v}\sqrt{v_k^{(m)}}$. On the event $\mathcal{E}_k$, if $|\widehat{\beta}_k^\star - \widehat{\beta}_{{\rm ind},k}^{(m)}| > \gamma_m\sqrt{v_k^{(m)}}$, then, for $k \in \mathcal{I}_\eta \cap \mathcal{I}^{(m)}$

$$|\widehat{\beta}_{{\rm ind},k}^{(m)} - \beta_k^{(m)}| = |\widehat{\beta}_{{\rm ind},k}^{(m)} - \beta_k^\star| \ge |\widehat{\beta}_{{\rm ind},k}^{(m)} - \widehat{\beta}_k^\star| - |\widehat{\beta}_k^\star - \beta_k^\star| > (\gamma_m - \sigma_{\rm v})\sqrt{v_k^{(m)}}.$$

Let

$$\mathcal{F}_k^{(m)} \triangleq \left\{ |\widehat{\beta}_{{\rm ind},k}^{(m)} - \beta_k^{(m)}| \le (\gamma_m - \sigma_{\rm v})\sqrt{v_k^{(m)}} \right\}.$$

Since $\gamma_m - \sigma_{\rm v} \ge \sigma_{\rm v} + c_{\rm hw}^{1/\alpha}\delta^{1/\alpha}\sigma$, by (D.2), we have

$$\mathbb{P}(\mathcal{F}_k^{(m)}) \le \mathbb{P}\left(|\widehat{\beta}_{{\rm ind},k}^{(m)} - \beta_k^{(m)}| \ge (\sigma_{\rm v} + c_{\rm hw}^{1/\alpha}\delta^{1/\alpha}\sigma)\sqrt{v_k^{(m)}} \right) \le 2e^{-\delta}.$$

Since event $\mathcal{F}_k^{(m)} \cap \mathcal{E}_k$ implies that $[\widehat{\beta}_{\rm ft}^{(m)}]_k = \widehat{\beta}_k^\star$ for $k \in \mathcal{I}_\eta \cap \mathcal{I}^{(m)}$. In other words, with probability at least $\mathbb{P}(\mathcal{E}_k \cap \mathcal{F}_k^{(m)}) \ge 1 - 6e^{-\delta}$, it holds that

$$|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}| = |\widehat{\beta}_k^\star - \beta_k^{(m)}| \le 1.25 C_{0.45}\tilde{\sigma}_k \alpha'_{\mathcal{B}_k,\delta}.$$

Furthermore, using (C.11) and Lemma D.2 (with $\ln(2/(6e^{-\delta})) \ge 1/\alpha$) and Lemma D.1, we have that for any $k \in \mathcal{I}_\eta \cap \mathcal{I}^{(m)}$,

$$\begin{aligned}
\mathbb{E}[|\widehat{\beta}_k^{(m)} - \beta_k^{(m)}|] &\le \widetilde{O}\left(\tilde{\sigma}_k G'_{[M],\mathcal{B}_k,\delta,n}\right) + \mathbb{E}^{6e^{-\delta}}\left[|\widehat{\beta}_{{\rm ind},k}^{(m)} - \beta_k^{(m)}| + \gamma_m\sqrt{v_k^{(m)}} \right] \\
&= \widetilde{O}\left(\frac{\sigma_{\rm v}}{\sqrt{\mu n}} \left(\frac{|\mathcal{B}_k|}{M} + \frac{1}{\sqrt{M}} + \frac{\sqrt{d}}{\sqrt{n}} \right) \right) + \widetilde{O}\left(\frac{\delta}{\sqrt{n}} (\sigma_{\rm v} + \sigma) \right) \\
&= \widetilde{O}\left(\frac{\sigma}{\sqrt{\mu n}} \left(\frac{|\mathcal{B}_k|}{M} + \frac{1}{\sqrt{M}} + \frac{\sqrt{d}}{\sqrt{n}} \right) \right).
\end{aligned}$$

Bounding the summed error. Combining the cases (D.12), (D.13), and using $|(\mathcal{I}^{(m)})^c| \le s$, we obtain

$$\begin{aligned}
\mathbb{E}[\|\widehat{\beta}_{\rm MOLAR}^{(m)} - \beta^{(m)}\|_1] &\le \frac{\sigma}{\sqrt{\mu n}}\widetilde{O}\left(\sum_{k \in \mathcal{I}_\eta \cap \mathcal{I}^{(m)}} \left(\frac{|\mathcal{B}_k|}{M} + \frac{1}{\sqrt{M}} + \frac{\sqrt{d}}{\sqrt{n}} \right) + |(\mathcal{I}_\eta \cap \mathcal{I}^{(m)})^c| \right) \\
&\le \frac{\sigma}{\sqrt{\mu n}}\widetilde{O}\left(s + \frac{d}{\sqrt{M}} + \frac{\sqrt{d}}{\sqrt{n}} + \sum_{k \in [d]} \left(\frac{|\mathcal{B}_k|}{M} \mathbb{1}(|\mathcal{B}_k|/M < \eta) + \mathbb{1}(|\mathcal{B}_k|/M \ge \eta) \right) \right). \qquad \text{(D.14)}
\end{aligned}$$

Using Lemma B.10 with $a = 1$ and $x_k = |\mathcal{B}_k|/M$ for all $k \in [d]$, we have

$$\sum_{k \in [d]]} \left(\frac{|\mathcal{B}_k|}{M} \mathbb{1}(|\mathcal{B}_k|/M < \eta) + \mathbb{1}(|\mathcal{B}_k|/M \ge \eta) \right) \le \lceil s/\eta \rceil. \qquad \text{(D.15)}$$

Letting $\eta = 1/10 = \Theta(1)$ and plugging (D.15) into (D.14), we find the conlusion.

□

E Extensions to the High-Dimensional Case

Our results in Section 2.2 rely on lower bounds on the singular values of the design matrices, *i.e.*, on $n_m \gg d$ for all $m \in [M]$. When $d \gg n_m$, the design matrices are not invertible, and thus the covariate-wise shrinkage in Algorithm 1 becomes pathological. In this case, given a global estimate $\widehat{\beta}^\star$, one can employ a LASSO-based debiasing step instead of covariate-wise shrinkage as in Xu and Bastani (2021): for all $m \in [M]$,

$$\widehat{\beta}^{(m)} = \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2n_m} \|\mathbf{X}^{(m)}\beta - Y^{(m)}\|_2^2 + \lambda_m \|\beta - \widehat{\beta}^\star\|_1. \qquad \text{(LASSO-based debiasing)}$$

Given a global estimate $\widehat{\beta}^\star$ that is accurate on a support set $\mathcal{G} \subseteq [d]$ (*i.e.*, $\|\widehat{\beta}^\star_{\mathcal{G}} - \beta^\star_{\mathcal{G}}\|_1$ is small), the performance of the LASSO-based debiasing can be analyzed as follows:

Proposition E.1. *Under Conditions 1, 2, and 3, taking* $\lambda_m = c\sigma\sqrt{\ln(dn_m)/n_m}$, *if* $n_m \geq c\ln(d)(s + |\mathcal{G}^c|)$ *for a sufficiently large* c *only depending on* $\Sigma^{(m)}$, *where* $\mathcal{G} \subseteq [d]$ *is a support set, it holds that*

$$\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 = \widetilde{O}_P\left(\frac{\sigma(s + |\mathcal{G}^c|)}{\sqrt{n_m}} + \|\widehat{\beta}^\star_{\mathcal{G}} - \beta^\star_{\mathcal{G}}\|_1\right). \tag{E.1}$$

where for any vector v, $v_{\mathcal{G}}$ *is the sub-vector of* v *with entries in* $\mathcal{G}$.

Proof. Define the event E be the intersection of $\{n_m^{-1}\|\mathbf{X}^{(m)\top}(\mathbf{X}^{(m)}\beta^{(m)} - Y^{(m)})\|_\infty \leq \lambda_m/2\}$ and

$$\left\{\|\mathbf{X}^{(m)}v\|_2^2 \geq n_m c_1 \|v\|_2 \left(\|v\|_2 - c_1\sqrt{\ln(d)/n_m}\|v\|_1\right),\ \forall\, v \in \mathbb{R}^d\right\}, \tag{E.2}$$

where c_1 is a sufficiently large constant that only depends μ and L in Condition 3. Here (E.2) is referred to as the restricted strong convexity and holds with probability at least $1 - \exp(-c_1\mu n_m)$ (Negahban et al., 2012). By the definition of λ_m, $\{n_m^{-1}\|\mathbf{X}^{(m)\top}(\mathbf{X}^{(m)}\beta^{(m)} - Y^{(m)})\|_\infty \leq \lambda_m/2\}$ holds with probability $1 - 1/(n_m d)$. Thus, $\mathbb{P}(E) \to 1$ as $n_m \to \infty$.

Using

$$\frac{1}{2n_m}\|\mathbf{X}^{(m)}\widehat{\beta}^{(m)} - Y^{(m)}\|_2^2 \leq \frac{1}{2n_m}\|\mathbf{X}^{(m)}\beta^{(m)} - Y^{(m)}\|_2^2 + \lambda_m\|\beta^{(m)} - \widehat{\beta}^\star\|_1 - \lambda_m\|\widehat{\beta}^{(m)} - \widehat{\beta}^\star\|_1,$$

we have

$$\begin{aligned}
&\frac{1}{2n_m}\|\mathbf{X}^{(m)}(\widehat{\beta}^{(m)} - \beta^{(m)})\|_2^2 = \frac{1}{2n_m}\|\mathbf{X}^{(m)}\widehat{\beta}^{(m)} - Y^{(m)}\|_2^2 + \frac{1}{2n_m}\|\mathbf{X}^{(m)}\beta^{(m)} - Y^{(m)}\|_2^2 \\
&\quad + \frac{1}{n_m}\langle \mathbf{X}^{(m)}\widehat{\beta}^{(m)} - Y^{(m)}, \mathbf{X}^{(m)}\beta^{(m)} - Y^{(m)}\rangle \\
\leq &\frac{1}{n_m}\langle \widehat{\beta}^{(m)} - \beta^{(m)}, \mathbf{X}^{(m)\top}(\mathbf{X}^{(m)}\beta^{(m)} - Y^{(m)})\rangle + \lambda_m\|\beta^{(m)} - \widehat{\beta}^\star\|_1 - \lambda_m\|\widehat{\beta}^{(m)} - \widehat{\beta}^\star\|_1.
\end{aligned}$$

Thus, letting $\mathcal{J}_m \triangleq \mathcal{G}^c \cap \mathrm{supp}(\beta^{(m)} - \beta^\star)$, we have $|\mathcal{J}_m| \le s + |\mathcal{G}^c|$ by Condition 1. Since $\beta^{(m)}_{\mathcal{J}_m^c} = \beta^\star_{\mathcal{J}_m^c}$, conditioned on the event E, we have

$$
\begin{aligned}
0 \le & \frac{1}{2n_m}\|\mathbf{X}^{(m)}(\widehat{\beta}^{(m)} - \beta^{(m)})\|_2^2 \le \lambda_m\|\beta^{(m)} - \widehat{\beta}^\star\|_1 - \lambda_m\|\widehat{\beta}^{(m)} - \widehat{\beta}^\star\|_1 + \frac{\lambda_m}{2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 \\
=& \lambda_m\|\beta^{(m)}_{\mathcal{J}_m} - \widehat{\beta}^\star_{\mathcal{J}_m}\|_1 + \lambda_m\|\beta^\star_{\mathcal{J}_m^c} - \widehat{\beta}^\star_{\mathcal{J}_m^c}\|_1 - \lambda_m\|\widehat{\beta}^{(m)}_{\mathcal{J}_m} - \widehat{\beta}^\star_{\mathcal{J}_m}\|_1 - \lambda_m\|\widehat{\beta}^{(m)}_{\mathcal{J}_m^c} - \widehat{\beta}^\star_{\mathcal{J}_m^c}\|_1 \\
& + \frac{\lambda_m}{2}\|\widehat{\beta}^{(m)}_{\mathcal{J}_m} - \beta^{(m)}_{\mathcal{J}_m}\|_1 + \frac{\lambda_m}{2}\|\widehat{\beta}^{(m)}_{\mathcal{J}_m^c} - \beta^{(m)}_{\mathcal{J}_m^c}\|_1 \\
\le & \frac{3\lambda_m}{2}\|\widehat{\beta}^{(m)}_{\mathcal{J}_m} - \beta^{(m)}_{\mathcal{J}_m}\|_1 - \frac{\lambda_m}{2}\|\widehat{\beta}^{(m)}_{\mathcal{J}_m^c} - \beta^{(m)}_{\mathcal{J}_m^c}\|_1 + 2\lambda_m\|\widehat{\beta}^\star_{\mathcal{J}_m^c} - \beta^\star_{\mathcal{J}_m^c}\|_1. \qquad \text{(E.3)}
\end{aligned}
$$

Noting $\mathcal{J}_m^c \subseteq \mathcal{G}$, we have

$$
\|\widehat{\beta}^{(m)}_{\mathcal{J}_m^c} - \beta^{(m)}_{\mathcal{J}_m^c}\|_1 \le 3\|\widehat{\beta}^{(m)}_{\mathcal{J}_m} - \beta^{(m)}_{\mathcal{J}_m}\|_1 + 4\|\widehat{\beta}^\star_{\mathcal{G}} - \beta^\star_{\mathcal{G}}\|_1. \qquad \text{(E.4)}
$$

Consequently, it holds that

$$
\begin{aligned}
\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 \le & 4\|\widehat{\beta}^{(m)}_{\mathcal{J}_m} - \beta^{(m)}_{\mathcal{J}_m}\|_1 + 4\|\widehat{\beta}^\star_{\mathcal{G}} - \beta^\star_{\mathcal{G}}\|_1 \\
\le & 4|\mathcal{J}_m|^{1/2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2 + 4\|\widehat{\beta}^\star_{\mathcal{G}} - \beta^\star_{\mathcal{G}}\|_1, \qquad \text{(E.5)}
\end{aligned}
$$

where we use Young's inequality in the last equation. Denote $\|\widehat{\beta}^\star_{\mathcal{G}} - \beta^\star_{\mathcal{G}}\|_1$ as $\mathrm{err}^\star$. Now using the restricted strong convexity and (E.4) in (E.3), we obtain

$$
\begin{aligned}
& 2\lambda_m|\mathcal{J}_m|^{1/2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2 + 2\lambda_m\mathrm{err}^\star \\
\ge & \frac{c_1}{2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2\left(\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2 - 4c_1\sqrt{\frac{\ln(d)|\mathcal{J}_m|}{n_m}}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2 - 4c_1\sqrt{\frac{\ln(d)}{n_m}}\mathrm{err}^\star\right) \\
& + \frac{\lambda_m}{2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1
\end{aligned}
$$

Suppose $n_m \ge c\ln(d)(s + |\mathcal{G}^c|)$ with c sufficiently large such that $4c_1\sqrt{\ln(d)|\mathcal{J}_m|/n_m} \le 1/3$. If $\mathrm{err}^\star \le |\mathcal{J}_m|^{1/2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2$, we obtain

$$
2\lambda_m|\mathcal{J}_m|^{1/2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2 + 2\lambda_m\mathrm{err}^\star \ge \frac{c_1}{6}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2^2 + \frac{\lambda_m}{2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1.
$$

Using Young's inequality $2ab \le a^2 + b^2$, we have $\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 = O\left(\lambda_m|\mathcal{J}_m| + \mathrm{err}^\star\right)$. If $\mathrm{err}^\star > |\mathcal{J}_m|^{1/2}\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_2$, using (E.5), we directly obtain $\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 = O\left(\mathrm{err}^\star\right)$. □

Notably, proposition E.1 does not require $n_m > d$ and directly implies

Corollary E.1. *Suppose*

$$
\|[\widehat{\beta}^\star]_{\mathcal{G}} - \beta^\star]_{\mathcal{G}}\|_1 = \widetilde{O}_P(d\sigma/\sqrt{n_{[M]}} + s\sigma/\sqrt{n_m}) \quad \textit{for some} \quad |\mathcal{G}^c| = O(s). \qquad \text{(E.6)}
$$

If $n_m \ge cs\ln(d)$ with c sufficiently large, LASSO-based debiasing gives the minimax optimal estimation error:

$$
\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 = \widetilde{O}_P\left(\frac{d\sigma}{\sqrt{n_{[M]}}} + \frac{s\sigma}{\sqrt{n_m}}\right).
$$

Algorithm 1 Transfer learning via constrained ℓ_1-minimization

Input: $\{(\mathbf{X}^{(m)}, Y^{(m)})\}_{m=1}^M$, regularization $\{\lambda_m = c\sigma\sqrt{\ln(n_m d)/n_m}\}_{m=1}^M$ and $\lambda_{\text{all}} = c\sigma\sqrt{\ln(p)/n_{[M]}}$

Step 1: Compute an initial estimate $\widehat{\beta}^\star_{\text{init}} = \arg\min_{\beta\in\mathbb{R}^d} \frac{1}{2n_1}\|\mathbf{X}^{(1)}\beta - Y^{(1)}\|_2^2 + \lambda_1\|\beta\|_1$

Step 2: Set

$$\widehat{\beta}^\star, \widehat{\delta}^{(2)}, \cdots, \widehat{\delta}^{(m)} = \underset{\beta,\, \|\delta^{(m)}\|_2\le b}{\text{argmin}} \ \lambda_{\text{all}}\|\beta\|_1 + \sum_{m=2}^M \lambda_m\|\delta^{(m)}\|_1$$

$$\text{s.t.} \begin{cases} \|X^{(m)\top}(X^{(m)}(\beta+\delta^{(m)}) - Y^{(m)})\|_\infty \le \lambda_m, \ \forall\, 2\le m\le M \\ \|\sum_{m=1}^M X^{(m)\top}(X^{(m)}(\beta+\delta^{(m)}) - Y^{(m)})\|_\infty \le \lambda_{\text{all}} \\ \|\beta - \widehat{\beta}^\star_{\text{init}}\|_1 \le \lambda_1^{-1} \end{cases}$$

Output: $\widehat{\beta}^\star$

When $n_m > d$ for all $m \in [M]$, the weighted median-based global estimate given by Algorithm 1 satisfies (E.6). However, the global estimate is not applicable if $n_m < d$ so that the data matrix $\mathbf{X}^{(m)\top}\mathbf{X}^{(m)}$ is not full-rank. Obtaining a good global estimate $\widehat{\beta}^\star$ is challenging when $n_m < d$ as the parameters $\{\beta^{(m)}\}_{m=1}^M$ can be dense. Fortunately, we can make progress by additionally assuming the parameters $\{\beta^{(m)}\}_{m=1}^M$ are sparse (Condition E.1) and the heterogeneity is ℓ_2-bounded.

Condition E.1 (Sparse global parameter)**.** *We have $\|\beta^\star\|_0 \le k$ for some $k \in [s, d]$ and $\max_{m\in[M]}\|\beta^{(m)} - \beta^\star\|_2 \le b$ for some constant $b \ge 0$.*

To this end, we borrow the transGLM method (Li et al., 2023), which leverages multiple datasets with sparse heterogeneity to learn a single generalized linear model. In the linear case, transGLM can be simplified to Algorithm 1. The estimation error of $\widehat{\beta}^\star$ for $\beta^\star$ directly follows from Li et al. (2023, Theorem 3.1). We paraphrase their result in our setup with notations defined in this paper as follows:

Theorem E.1 (Li et al. (2023), Theorem 3.1)**.** *Suppose*

$$n_{[M]} \ge c_1 \max\{k^2\ln(d)^2, Mn_1\}, \quad n_1 \ge c_1 ks\ln(d)^2,$$

where c_1 is some large enough quantity not depending on d, k, s, and $\{n_m\}_{m=1}^M$. Under Conditions 1, 2, and 3, it holds that

$$\|\widehat{\beta}^\star_{\mathcal{G}} - \beta^\star_{\mathcal{G}}\|_1 = \widetilde{O}_P\left(\frac{k\sigma}{\sqrt{n_{[M]}}} + \frac{\sqrt{sk}\sigma}{\sqrt{n_1}}\right),$$

where $\mathcal{G} = [d]\backslash\text{supp}(\beta_1 - \beta)$ with $|\mathcal{G}^c| \le s$.

By Combining Theorem E.1 with Proposition E.1, we have

Corollary E.2 (Transfer Learning + Lasso Debiasing)**.** *Suppose $n_m \ge c_1 ks\ln(d)^2$ for all $1 \le m \le M$ with $n_1/\min_m n_m = O(1)$, and $n_{[M]} \ge c_1 k^2\ln(d)^2$ where c_1 is some large enough value not depending on d, k, s, and $\{n_m\}_{m=1}^M$. Under Conditions 1, 2, and 3, the task-wise*

estimates obtained through applying the LASSO-based debiasing to the global estimate output by Algorithm 1 achieves for $1 \leq m \leq M$*,*

$$\|\widehat{\beta}^{(m)} - \beta^{(m)}\|_1 = \widetilde{O}_P\left(\frac{\sqrt{sk}\sigma}{\sqrt{n_m}} + \frac{k\sigma}{\sqrt{n_{[M]}}}\right).$$

Corollary E.2 bounds the estimation error of the multitask approach with $n_m \geq c_1 ks \ln(d)^2$ and $n_{[M]} \geq c_1 k^2 \ln(d)^2$, which holds in the high-dimensional case so long as $ks \ll d$ and M is large. Furthermore, the multitask approach outperforms the individual LASSO, the single-task minimax optimal method, whose estimation error is $k\sigma/\sqrt{n_m}$.

F Inference for Task-wise Parameters

In this section, we consider statistical inference for $\{\beta^{(m)}\}_{m=1}^M$ in the multi-task learning setting. Notably, the MOLAR estimates obtained in Algorithm 1 and the high-dimensional estimates given in Appendix E are biased. It is tempting to consider debiasing them to facilitate inference as in (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014). However, debiasing may increase the estimation error of our multi-task estimate, and thus likely cannot give confidence intervals with lengths shorter than those of the individual OLS estimates.

To show the potential of narrower confidence intervals in our setup, for simplicity, we consider $n_1 = \cdots = n_M =: n$ and $\sigma_1 = \cdots = \sigma_M =: \sigma$, and assume σ is known throughout the section. The case where $\{\sigma_m\}_{m=1}^M$ or $\{n_m\}_{m=1}^M$ are non-identical can be handled similarly by adjusiting the weights $\{w_m\}_{m=1}^M$ in the collaboration step. We additionally make the following assumption to constrain the distribution of heterogeneity. Again, here the number $1/5$ is taken for simplicity and can be, in principle, replaced with a constant number in $[0, 1/2)$.

Condition F.1 (Bounded entry-wise disagreement)**.** *We assume $|\mathcal{B}_k|/M \le 1/5$ for all $k \in [M]$.*

Given a fixed and properly small tolerance α, for each $k \in [d]$ and $m \in [M]$, we let $\widetilde{I}_k^{(m)} = \left[\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \sigma z_{1-\alpha/2}\sqrt{v_k^{(m)}}, \widehat{\beta}_{\mathrm{ind},k}^{(m)} + \sigma z_{1-\alpha/2}\sqrt{v_k^{(m)}}\right]$ be the interval with coverage $1-\alpha/2$ for $\beta_{\mathrm{ind},k}^{(m)}$ centered at $\widehat{\beta}_{\mathrm{ind},k}^{(m)}$ where $v_k^{(m)} = \sqrt{[\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}}$.We let $I_k^\star$ be an interval with coverage $1-\alpha/2$ for $\beta_k^\star$ centered at $\widehat{\beta}_k^\star$. Based on Lemma C.2 and (C.8), for $\alpha_{\mathcal{B}_k,\delta} < 0.45$, we can set

$$I_k^\star = [\widehat{\beta}_k^\star - 1.25C_{0.45}\alpha_{\mathcal{B},\delta}\bar{\sigma}_k, \widehat{\beta}_k^\star + 1.25C_{0.45}\alpha_{\mathcal{B},\delta}\bar{\sigma}_k]$$

with $\delta = \ln((8/\alpha)\vee(2n))/2$ (*i.e.*, $2e^{-2\delta} = (\alpha/4)\wedge n^{-1}$), $\alpha_{\mathcal{B}_k,\delta} \triangleq |\mathcal{B}_k|/M + \sqrt{1.01\delta/(M-|\mathcal{B}_k|)}$, and $\bar{\sigma}_k = \sigma\sum_{m\in[M]}\sqrt{v_k^{(m)}}/M$. Since $v_k^{(m)} = \widetilde{O}_P(1/n)$, we have $\mathrm{length}(\widetilde{I}_k^{(m)}) = \widetilde{O}_P(1/\sqrt{n})$ and $\mathrm{length}(I_k^\star) = \widetilde{O}_P(\alpha_{\mathcal{B}_k,\delta}/\sqrt{n})$ for all $k \in [d]$ and $m \in [M]$. Noting $\sum_{k=1}^d |\mathcal{B}_k|/M = s/d$, we further have $\sum_{k=1}^d \mathrm{length}(I_k^\star) = O_P(s/\sqrt{n} + d/\sqrt{Mn})$.

Since $\{I_k^\star\}_{k=1}^d$ are narrower than $\{\widetilde{I}_k^{(m)}\}_{k=1}^d$ on average, we use the following strategy to construct the ultimate confidence interval $I_k^{(m)}$ for $\beta_k^{(m)}$ with at least $1-\alpha$ entry-wise coverage. We first compare $\widehat{\beta}_{\mathrm{ind},k}^{(m)}$ and $\widehat{\beta}_k^\star$. If they are close enough such that $|\widehat{\beta}_{\mathrm{ind},k}^{(m)} - \widehat{\beta}_k^\star| < \widetilde{\gamma}_m\sqrt{v_k^{(m)}}$ for some pre-specified $\widetilde{\gamma}_m$, it is likely that $\beta_k^{(m)} = \beta_k^\star$ and we thus adopt the confidence interval $I_k^\star$ as the final interval $I_k^{(m)}$; otherwise we adopt $\widetilde{I}_k^{(m)}$ as $I_k^{(m)}$. Formally, we aim to attain $\mathbb{P}(\beta_k^{(m)} \in I_k^{(m)}) \ge 1-\alpha$ and the total length of entry-wise confidence intervals is $\sum_{k=1}^d \mathrm{length}(I_k^{(m)})$. We show the following guarantee for the confidence intervals $\{I_k^{(m)}\}_{k=1}^d$. Without loss of generality, we assume $M \ge \tilde{c}\ln(n)$ for a sufficiently large constant $\tilde{c}$.

Proposition F.1. *For any $\alpha \in (2e^{-2cM}, 1]$ with some constant c sufficiently small, for any $m \in [M]$, if we set $\widetilde{\gamma}_m = \sqrt{2}\ln((2n)\vee(8/\alpha))\sigma$, it holds for all $k \in [d]$ that*

$$\mathbb{P}\left(\beta_k^{(m)} \in I_k^{(m)}\right) \ge 1-\alpha$$

provided $|\beta_k^{(m)} - \beta_k^\star| > 3\max\{\ln(n \vee (4/\alpha))\sqrt{v_k^{(m)}}\sigma, 1.25C_{\alpha_{\mathcal{B}_k,\delta}}\alpha_{\mathcal{B},\delta}\bar{\sigma}_k\} = \widetilde{\Omega}_P(1/\sqrt{n})$*, where* $\delta = \ln((2n) \vee (8/\alpha))$ *for* $\beta_k^{(m)} \neq \beta_k^\star$*. Furthermore, the total length satisfies*

$$\sum_{k=1}^{d} \text{length}(I_k^{(m)}) = \widetilde{O}_P(s/\sqrt{n} + d/\sqrt{Mn}).$$

Proof. Clearly, $v_k^{(m)} = \widetilde{O}_P(1/n)$ so long as $n \gg d$. We prove the result by considering two cases.

Case 1: $\beta_k^{(m)} = \beta_k^\star$. In this case, given the choice of $\widetilde{\gamma}_m$ and Lemma C.2, we can easily verify $\mathbb{P}(|\widehat{\beta}_{\text{ind},k}^{(m)} - \widehat{\beta}_k^\star| \geq \widetilde{\gamma}_m\sqrt{v_k^{(m)}}) = (\alpha/4) \wedge n^{-1} \leq \alpha/2$. This, combined with $\mathbb{P}(\beta_k^\star \notin I_k^\star) \leq \alpha/2$ and a union bound, leads to

$$\mathbb{P}(\beta_k^{(m)} \in I_k^\star) \geq \mathbb{P}\left(\beta_k^\star \in I_k^\star \text{ and } |\widehat{\beta}_{\text{ind},k}^{(m)} - \widehat{\beta}_k^\star| < \widetilde{\gamma}_m\sqrt{v_k^{(m)}}\right) \geq 1 - \alpha.$$

Furthermore, with probability $\mathbb{P}\left(|\widehat{\beta}_{\text{ind},k}^{(m)} - \widehat{\beta}_k^\star| < \widetilde{\gamma}_m\sqrt{v_k^{(m)}}\right) = 1 - o(1)$ we have that

$$\text{length}(I_k^{(m)}) = \text{length}(I_k^\star) = O(\alpha_{\mathcal{B}_k,\delta}/\sqrt{n}) = \widetilde{O}\left(\left(|\mathcal{B}_k|/M + 1/\sqrt{M}\right)\sqrt{v_k^{(m)}}\right).$$

Case 2: $\beta_k^{(m)} \neq \beta_k^\star$. Denote $|\beta_k^{(m)} - \beta_k^\star|$ as ϵ. In this case, we shall prove that $|\widehat{\beta}_{\text{ind},k}^{(m)} - \widehat{\beta}_k^\star| \geq \widetilde{\gamma}_m\sqrt{v_k^{(m)}}$ with a high probability. We first see that

$$\begin{aligned} |\widehat{\beta}_{\text{ind},k}^{(m)} - \widehat{\beta}_k^\star| =& |\widehat{\beta}_{\text{ind},k}^{(m)} - \beta_k^{(m)} + \beta_k^{(m)} - \beta_k^\star + \beta_k^\star - \widehat{\beta}_k^\star| \\ \geq& \epsilon - |\widehat{\beta}_{\text{ind},k}^{(m)} - \beta_k^{(m)}| - |\beta_k^\star - \widehat{\beta}_k^\star|. \end{aligned} \tag{F.1}$$

Therefore, $|\widehat{\beta}_{\text{ind},k}^{(m)} - \widehat{\beta}_k^\star| < \widetilde{\gamma}_m\sqrt{v_k^{(m)}}$ suggests that one of $|\widehat{\beta}_{\text{ind},k}^{(m)} - \beta_k^{(m)}| \geq (\epsilon - \widetilde{\gamma}_m\sqrt{v_k^{(m)}})/2$ or $|\beta_k^\star - \widehat{\beta}_k^\star| \geq (\epsilon - \widetilde{\gamma}_m\sqrt{v_k^{(m)}})/2$ must hold. However, by the condition on ϵ and $\widetilde{\gamma}_m$, using the sub-Gaussianity of $\widehat{\beta}_{\text{ind},k}^{(m)}$ and Lemma C.2, we have

$$\mathbb{P}\left(|\widehat{\beta}_{\text{ind},k}^{(m)} - \beta_k^{(m)}| \geq (\epsilon - \widetilde{\gamma}_m\sqrt{v_k^{(m)}})/2\right) \leq \mathbb{P}\left(|\widehat{\beta}_{\text{ind},k}^{(m)} - \beta_k^{(m)}| \geq \widetilde{\gamma}_m\sqrt{v_k^{(m)}}\right) \leq \frac{\alpha}{4} \bigwedge \frac{1}{n} \leq \frac{\alpha}{4} \tag{F.2}$$

and

$$\mathbb{P}\left(|\beta_k^\star - \widehat{\beta}_k^\star| \geq (\epsilon - \widetilde{\gamma}_m\sqrt{v_k^{(m)}})/2\right) \leq \mathbb{P}\left(|\beta_k^\star - \widehat{\beta}_k^\star| \geq 1.25C_{\alpha_{\mathcal{B}_k,\delta}}\alpha_{\mathcal{B},\delta}\bar{\sigma}_k\right) \leq \frac{\alpha}{4} \bigwedge \frac{1}{n} \leq \frac{\alpha}{4}. \tag{F.3}$$

where $\delta = \ln((8/\alpha) \vee (2n))/2$ is much smaller than M so that $\alpha_{\mathcal{B}_k,\delta} < 0.45$. Combining (F.2), (F.3) with $\mathbb{P}(\beta_k^{(m)} \notin I_k^{(m)}) \leq \alpha/2$ and a union bound, we find

$$\mathbb{P}\left(\beta_k^{(m)} \in I_k^\star\right) \geq \mathbb{P}(\beta_k^{(m)} \in I_k^{(m)} \text{ and } |\widehat{\beta}_{\text{ind},k}^{(m)} - \widehat{\beta}_k^\star| \geq \widetilde{\gamma}_m\sqrt{v_k^{(m)}}\right) \geq 1 - \alpha.$$

Denoting $\{k \in [d] : \beta_k^{(m)} = \beta_k^\star\}$ as $\mathcal{I}^{(m)}$ with $|\mathcal{I}^{(m)}| = s$, combining the two cases, we have

$$\sum_{k=1}^{d} \text{length}(I_k^{(m)}) = \widetilde{O}_P\left(|\mathcal{I}^{(m)}|/\sqrt{n} + \sum_{k \notin \mathcal{I}^{(m)}} \alpha_{\mathcal{B}_k,\delta}/\sqrt{n}\right).$$

By noting that

$$\sum_{k \notin \mathcal{I}^{(m)}} \alpha_{\mathcal{B}_k,\delta} \leq \sum_{k \in [d]} \alpha_{\mathcal{B}_k,\delta} = \widetilde{O}(s + d/\sqrt{M}),$$

we complete the proof. □

Proposition F.1 shows that our confidence intervals for the entries of the task-wise parameters $\{\beta^{(m)}\}_{m=1}^M$ have total length $\widetilde{O}_P(s/\sqrt{n} + d/\sqrt{Mn})$. When $s \ll d$ and $M \gg 1$, the length is shorter than $\widetilde{O}_P(d/\sqrt{n})$, the length of the standard intervals based on the individual OLS estimates.

However, we remark that this proposition requires that the unequal entries between $\beta^{(m)}$ and $\beta^\star$ be separated by $\Omega(1/\sqrt{n})$. This condition turns out to be necessary to attain confidence intervals with a total length shorter than $\widetilde{O}_P(d/\sqrt{n})$. To see this, we can argue as follows. Even if the global parameter $\beta^\star$ was *exactly known*, corresponding to the case where $M = \infty$, our setup would reduce to constructing a confidence interval for $\beta^{(m)} - \beta^\star$, which is s-sparse. This ideal setting becomes an example of inference for single-task sparse linear regression studied by e.g., Cai and Guo (2017). That work shows that the minimax optimal length of confidence intervals for individual entries is $\Omega_P(1/\sqrt{n})$ when the non-zero entries of the sparse parameter are not constrained to be away from zero (*i.e.*, each entry is either zero or of magnitude $O(1/\sqrt{n})$). One can follow the same idea to show that the total length of confidence intervals with entry-wise coverage is $\Omega_P(d/\sqrt{n})$, irrespective of the parameter's sparsity. The challenge in constructing shorter confidence intervals mainly lies in *identifying the support set of the sparse parameter* when the non-zero entries of the parameter are $O(1/\sqrt{n})$.

G Results on GLMs

Given the predictors $x \in \mathbb{R}^d$, if the response y follows the generalized linear models (GLMs), then its conditional distribution takes the form, for all $x \in \mathbb{R}^d$,

$$y \mid x \sim \mathbb{P}(y \mid x) = \rho(y) \exp\left(y\langle x, \beta\rangle - \psi(\langle x, \beta\rangle)\right) \tag{G.1}$$

where $\beta \in \mathbb{R}^d$ is the unknown parameter, and ρ and ψ are some known univariate functions. Two important properties of GLMs are $\mathbb{E}[y \mid x] = \psi'(\langle x, \beta\rangle)$ and $\mathrm{Var}(y \mid x) = \psi''(\langle x, \beta\rangle)$ (McCullagh and Nelder, 1989). In particular, for linear models with Gaussian noise, we have a continuous response y and $\psi(u) = u^2/(2\sigma^2)$ for all $u \in \mathbb{R}$.

In the scenario of multitask GLMs, the individual estimate $\widehat{\beta}^{(m)}_{\mathrm{ind}}$ can be taken as the minimizer of the negative log-likelihood function

$$\widehat{\beta}^{(m)}_{\mathrm{ind}} := \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n_m} \sum_{i=1}^{n_m} \left(-y_i^{(m)} \langle x_i^{(m)}, \beta\rangle + \psi(\langle x_i^{(m)}, \beta\rangle)\right).$$

Due to the nonlinearity of GLMs, to apply the MOLAR method, we also need to replace the inverse data matrix $(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}$ with $(\mathbf{X}^{(m)\top}\widehat{D}^{(m)}\mathbf{X}^{(m)})^{-1}$ where $\widehat{D}^{(m)} \in \mathbb{R}^{n_m \times n_m}$ is the diagonal matrix with elements $\{\psi''(\langle x_i^{(m)}, \widehat{\beta}^{(m)}_{\mathrm{ind}}\rangle)\}_{i=1}^{n_m}$. The two adjustments form Algorithm 2.

Algorithm 2 MOLAR-GLM: Weighted-Median-based Multitask GLM

Input: $\{(\mathbf{X}^{(m)}, Y^{(m)})\}_{m=1}^M$, thresholds $\{\gamma_m\}_{m=1}^M$, weights $\{w_m\}_{m=1}^M$
for $m \in [M]$ **do**
  Let $\widehat{\beta}^{(m)}_{\mathrm{ind}}$ be the individual MLE for $(\mathbf{X}^{(m)}, Y^{(m)})$
end for
Let $\widehat{\beta}^{\star} = \mathrm{WMed}(\{\widehat{\beta}^{(m)}_{\mathrm{ind}}\}_{m=1}^M; \{w_m\}_{m=1}^M)$ be the covariate-wise weighted median
for $m \in [M]$ and $k \in [d]$ **do**
  /* Option I: hard thresholding */
  $\widehat{\beta}^{(m)}_{\mathrm{MOLAR},k} = \widehat{\beta}^{\star}_k$ **if** $|\widehat{\beta}^{\star}_k - \widehat{\beta}^{(m)}_{\mathrm{ind},k}| \le \gamma_m \sqrt{[(\mathbf{X}^{(m)\top}\widehat{D}^{(m)}\mathbf{X}^{(m)})^{-1}]_{k,k}}$ **else** $\widehat{\beta}^{(m)}_{\mathrm{ind},k}$
  /* Option II: soft thresholding */
  $\widehat{\beta}^{(m)}_{\mathrm{MOLAR},k} = \widehat{\beta}^{\star}_k + \mathsf{SoftThresholding}\left(\widehat{\beta}^{(m)}_{\mathrm{ind},k} - \widehat{\beta}^{\star}_k; \gamma_m \sqrt{[(\mathbf{X}^{(m)\top}\widehat{D}^{(m)}\mathbf{X}^{(m)})^{-1})^{-1}]_{k,k}}\right)$
end for
Output: $\{\widehat{\beta}^{(m)}_{\mathrm{MOLAR}}\}_{m=1}^M$

We analyze MOLAR-GLM for sparsely heterogeneous parameters $\{\beta^{(m)}\}_{m=1}^M$ satisfying Condition 1 in the asymptotic sense where sample sizes are sufficiently large. For simplicity, we only consider $n_1 = \cdots = n_M =: n$. Our analysis is built on the asymptotic normality of individual GLM estimates (Van de Geer et al., 2014; Xia et al., 2023). Specifically, suppose the following holds.

Condition G.1 (Conditions for GLMs). *For each $m \in [M]$, the following conditions hold for the GLM model* (G.1).

1. *$\beta^{(m)}$ is the unique maximizer to $\mathbb{E}[y_i^{(m)}\langle x_i^{(m)}, \beta\rangle - \psi(\langle x_i^{(m)}, \beta\rangle]$ and $P(y_i^{(m)} \mid x_i^{(m)})$ is quadratic mean differentiable at $\beta^{(m)}$.*
2. *$\frac{1}{n}\sum_{i=1}^n y_i^{(m)}\langle x_i^{(m)}, \beta\rangle - \psi(\langle x_i^{(m)}\rangle, \beta\rangle$ converges uniformly to $\mathbb{E}[y_i^{(m)}\langle x_i^{(m)}, \beta\rangle - \psi(\langle x_i^{(m)}, \beta\rangle]$ as $n \to \infty$.*
3. *ψ is twice continuously differentiable and ψ is uniformly bounded.*
4. *The population Fisher matrix $\Sigma^{(m)} \triangleq \mathbb{E}[x_i^{(m)} x_i^{(m)\top} \psi''(\langle x_i^{(m)}, \beta^{(m)}\rangle)]$ is positive definite and its eigenvalues are bounded and bounded away from 0,* i.e.*,*

$$cI_d \preceq \mathbb{E}[x_i^{(m)} x_i^{(m)\top} \psi''(\langle x_i^{(m)}, \beta^{(m)}\rangle)] \preceq CI_d$$

for some $\Omega(1) = c \le C = O(1)$.

Condition G.1 is assumed for technical simplicity, which directly facilitates the analysis of maximum likelihood estimates in (Van de Geer et al., 2014) and can be possibly relaxed. Specifically, we can obtain that

Proposition G.1 (Asymptotic normality). *Under Condition G.1, it holds that for each $m \in [M]$ that*

$$\sqrt{n}(\widehat{\beta}_{\text{ind}}^{(m)} - \beta^{(m)}) \xrightarrow{d} \mathcal{N}\left(0,\, (\Sigma^{(m)})^{-1}\right) \text{ and } V^{(m)}/n \xrightarrow{p} \Sigma^{(m)},$$

where $V^{(m)} \triangleq \mathbf{X}^{(m)\top} \widehat{D}^{(m)} \mathbf{X}^{(m)}$.

Proof. Under Condition G.1, it is easily have $V^{(m)}/n \xrightarrow{p} \Sigma^{(m)}$. Since $\frac{1}{n}\sum_{i=1}^n y_i^{(m)}\langle x_i^{(m)}, \beta\rangle - \psi(\langle x_i^{(m)}, \beta\rangle$ converges uniformly to $\mathbb{E}[y_i^{(m)}\langle x_i^{(m)}, \beta\rangle - \psi(\langle x_i^{(m)}, \beta\rangle]$ and $\mathbb{E}[y_i^{(m)}\langle x_i^{(m)}, \beta\rangle - \psi(\langle x_i^{(m)}, \beta\rangle]$ has a unique maximum $\beta^{(m)}$ that is well-separated due to the quadratic mean differentiablility, $\widehat{\beta}_{\text{ind}}^{(m)} \xrightarrow{p} \beta^{(m)}$ by Van de Geer et al. (2014, Theorem 5.7). Then Van de Geer et al. (2014, Theorem 5.39) guarantees $\sqrt{n}(\widehat{\beta}_{\text{ind}}^{(m)} - \beta^{(m)}) \xrightarrow{d} \mathcal{N}(0, (\Sigma^{(m)})^{-1})$ □

Given the asymptotic normality of individual estimates, one can establish the following asymptotic bounds for the tail probability of the global estimate $\widehat{\beta}^\star$.

Lemma G.1. *Under Condition G.1, for any $0 < \eta \le \frac{1}{5}$ and $k \in \mathcal{I}_\eta$, it holds for any $0 \le \delta \le M/21$ and n sufficiently large that*

$$\mathbb{P}\left(|\widehat{\beta}_k^\star - \beta_k^\star| \ge 2C_{0.45}\alpha_{\mathcal{B}_k,\delta}\bar{v}_{[M],k}\right) \le 2e^{-2\delta},$$

where $\bar{v}_{[M],k} = \sum_{m\in[M]}[(\Sigma^{(m)})^{-1/2}]_{k,k}/(\sqrt{n}M)$ and $\alpha_{\mathcal{B}_k,\delta}$ follows from the definition in (C.5).

Proof. By Proposition G.1, we know $\sqrt{n}(\widehat{\beta}^{(m)}_{\text{ind},k} - \beta^{(m)}_k) \xrightarrow{d} \mathcal{N}\left(0,\, [(\Sigma^{(m)})^{-1}]_{k,k}\right)$ for any $k \in [d]$ and $m \in [M]$. Note that the weighted median of $\{\widehat{\beta}^{(m)}_{\text{ind},k}\}_{m=1}^M$ is upper and lower bounded by the $(1/2 + |\mathcal{B}_k|/M)$-weighted-quantile and $(1/2 - |\mathcal{B}_k|/M)$-weighted-quantile of $\{\widehat{\beta}^{(m)}_{\text{ind},k}\}_{m\in\mathcal{G}_k}$ respectively, where $\mathcal{G}_k \triangleq \{m \in [M] : \beta^{(m)}_k = \beta^\star_k\}$. The $(1/2 + |\mathcal{B}_k|/M)$-weighted-quantile of $\{\widehat{\beta}^{(m)}_{\text{ind},k}\}_{m\in\mathcal{G}_k}$ is equivalent to the $(1/2+|\mathcal{B}_k|/M)$-weighted-quantile of $\{\sqrt{n}(\widehat{\beta}^{(m)}_{\text{ind},k} - \beta^{(m)}_k)\}_{m\in\mathcal{G}_k}$ after scaling and translation. We can also leverage Lemma C.1 to cover the case of asymptotic normality: for any $\alpha \in [0, 1/2)$, let $\mu_{1/2+\alpha}$ be value such that

$$\sum_{i\in\mathcal{G}_k} w_i \Phi\left(\frac{\mu_{1/2+\alpha}}{[(\Sigma^{(m)})^{-1}]_{k,k}}\right) = \left(\frac{1}{2} + \alpha\right) W_{\mathcal{G}_k},$$

where Φ is the c.d.f. of the standard normal distribution. By the asymptotic normality of $\sqrt{n}(\widehat{\beta}^{(m)}_{\text{ind},k} - \beta^{(m)}_k)$, we know

$$\sum_{i\in\mathcal{G}_k} w_i \mathbb{P}(\sqrt{n}(\widehat{\beta}^{(m)}_{\text{ind},k} - \beta^{(m)}_k) \le \mu_{1/2+1.01\alpha}) \to \sum_{i\in\mathcal{G}_k} w_i \Phi\left(\frac{\mu_{1/2+1.01\alpha}}{[(\Sigma^{(m)})^{-1}]_{k,k}}\right) = \left(\frac{1}{2} + 1.01\alpha\right) W_{\mathcal{G}_k}.$$

Therefore, for n sufficiently large, we have

$$\sum_{i\in\mathcal{G}_k} w_i \mathbb{P}(\sqrt{n}(\widehat{\beta}^{(m)}_{\text{ind},k} - \beta^{(m)}_k) \le \mu_{1/2+1.01\alpha}) \ge \left(\frac{1}{2} + \alpha\right) W_{\mathcal{G}_k},$$

which implies that the weighted population $(1/2 + \alpha)$-quantile of $\{\sqrt{n}(\widehat{\beta}^{(m)}_{\text{ind},k} - \beta^{(m)}_k)\}_{m\in\mathcal{G}_k}$ is upper bounded by $\mu_{1/2+1.01\alpha}$, which is in turn upper bounded by the one in Lemma C.1. The lower bound can be argued similarly. In summary, we can give a bound for $\beta^\star_k$ similarly to Lemma C.3 for sufficiently large n by relaxing a small number close to C_α to as $C_{1.01\alpha}$, which gives us the result. □

Furthermore, one can easily follow the proof of Theorem 1 to bound the estimation errors of the MOLAR estimates in GLMs for sufficiently large n. We omit the proof here.

Theorem G.1 (Error bound for task-wise parameters)**.** *Under Conditions 1 and G.1, taking $\gamma_m = c_1\sqrt{\ln(M \wedge d)}$ for all $m \in [M]$ with c_1 sufficiently large, with $\widehat{\beta}^{(m)}_{\text{MOLAR}}$ from Algorithm 2 using either Option I or II, it holds for any $p \in \{1, 2\}$, $m \in [M]$ and n sufficiently large that*

$$n^{p/2}\|\widehat{\beta}^{(m)}_{\text{MOLAR}} - \beta^{(m)}\|_p^p = \widetilde{O}_P\left(s + \frac{d}{M^{p/2}}\right).$$

An analytical comparison of MOLAR-GLM with related methods (Tian and Feng, 2022; Li et al., 2023) is in Table 1. Tian and Feng (2022); Li et al. (2023) originally considered sparse parameters, and we adjust their results to dense parameters and state them with notations defined in our manuscript for a clear comparison. We observe that while our method may require larger task-wise sample sizes, it theoretically has faster rates of convergence than existing methods for GLMs (with dense parameters).

Table 1: Bounds on the estimation error $\sqrt{n}\|\widehat{\beta}^{(m)}-\beta^{(m)}\|_1$ for GLMs. Suppose $n_1 = \cdots = n_M =: n$. Constants and logarithmic factors are omitted for clarity. The results Tian and Feng (2022); Li et al. (2023) are adjusted to dense parameters (*i.e.*, $\|\beta^{(m)}\|_0 = \Omega(d)$ for all $m \in [M]$).

Method	Heterogeneity Condition	Rate	Sample Size
Tian and Feng (2022)	$\max_{m\in[M]} \|\beta^{(m)} - \beta^\star\|_1 = h$	$(n^{1/4}h^{1/2}) \wedge (\sqrt{n}h) + d/\sqrt{M}$	$Mn \gg d$, $n \gg h^2$
Li et al. (2023)	$\begin{cases}\max_{m\in[M]} \|\beta^{(m)} - \beta^\star\|_0 = s \\ \max_{m\in[M]} \|\beta^{(m)} - \beta^\star\|_2 = O(1)\end{cases}$	$\sqrt{sd} + d/\sqrt{M}$	$Mn \gg d^2$, $n \gg sd$
MOLAR-GLM	$\max_{m\in[M]} \|\beta^{(m)} - \beta^\star\|_0 = s$	$s + d/\sqrt{M}$	n sufficiently large
Lower Bound	$\max_{m\in[M]} \|\beta^{(m)} - \beta^\star\|_0 = s$	$s + d/\sqrt{M}$	—

Acknowledging the importance of extending the application of the MOLAR method to encompass a broader setup, particularly in the context of generalized linear models for n being small, we think that a detailed investigation and optimality in GLMs particularly for n being small may require entirely different algorithmic designs and should indeed serve as an independent and valuable avenue for future research.

H Results on Contextual Bandits under Model-C

H.1 Lemma H.1 and its Proof

To analyze individual regret, we need that the empirical covariance matrices, at the end of each batch, are well-conditioned with high probability, even when the arms are adaptively chosen. We show Lemma H.1, which guarantees that for any $m \in [M]$, with high probability the singular values of the contexts $\mathbf{X}_q^{(m)}$ are lower bounded. Lemma H.1 is similar to (Han et al., 2020, Lemma 4) and (Ren and Zhou, 2023, Lemma 5) in the single-bandit regime. However, Han et al. (2020) assume Gaussian contexts, which is stronger than our Condition 8 and 9, while Ren and Zhou (2023) consider an s-sparse parameter and sub-Gaussian contexts. The proof of Lemma H.1 relies on using an ε-net argument.

Lemma H.1. *Under Conditions 8 and 9, for some $C_{\mathrm{b}} \geq 2$ only depending on c_x, and for any $0 \leq q < Q$, it holds with probability at least $1 - 1/T$ that $\lambda_{\min}(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}) \geq n_{m,q}\mu c_x/4$, for all $m \in [M]$ with $n_{m,q} \geq C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))$.*

Proof. For $0 \leq q < Q$ and $m \in [M]$, we let $\mathcal{T}_q^{(m)}$ be the set of times when contexts $\mathbf{X}_q^m$ are observed at instance m, *i.e.*, $\mathbf{X}_q^m = (x_{t,a_t^{(m)}}^{(m)})_{t\in\mathcal{T}_q^{(m)}}^\top$. Clearly, we have $|\mathcal{T}_q^{(m)}| = n_{m,q}$; and further $\{x_{t,a_t^{(m)}}^{(m)} : t \in \mathcal{T}_q^{(m)}\}$ are independent, conditioned on $\widehat{\beta}_{q-1}^{(m)}$. The following analysis is conditional on $\{\mathcal{T}_q^{(m)}\}_{m\in[M]}$ and therefore on $\{n_{m,q}\}_{m\in[M]}$.

We first prove an upper bound on $\lambda_{\max}(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})$. For any $t \in [T]$, $a \in [K]$, $m \in [M]$ and any source vector $v \in \mathbb{R}^d$, let $Z_{t,a}^{(m)} = \langle v, x_{t,a}\rangle^2$. Conditioned on $\widehat{\beta}_{q-1}^{(m)}$, for any $\delta > 0$ and $\lambda > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{t\in\mathcal{T}_q^{(m)}} Z_{t,a_t^{(m)}}^{(m)} \geq n_{m,q}\delta \mid \widehat{\beta}_{q-1}^{(m)}\right) \leq e^{-\lambda n_{m,q}\delta}\mathbb{E}\left[\exp\left(\lambda\sum_{t\in\mathcal{T}_q^{(m)}} Z_{t,a_t^{(m)}}^{(m)}\right) \mid \widehat{\beta}_{q-1}^{(m)}\right] \\
=&e^{-\lambda n_{m,q}\delta}\prod_{t\in\mathcal{T}_q^{(m)}}\mathbb{E}\left[\exp\left(\lambda Z_{t,a_t^{(m)}}^{(m)}\right) \mid \widehat{\beta}_{q-1}^{(m)}\right] \leq e^{-\lambda n_{m,q}\delta}\prod_{t\in\mathcal{T}_q^{(m)}}\sum_{a\in[K]}\mathbb{E}\left[\exp\left(\lambda Z_{t,a}^{(m)}\right)\right].
\end{aligned}$$

Taking the expectation with respect to $\widehat{\beta}_{q-1}^{(m)}$, we obtain

$$\mathbb{P}\left(\sum_{t\in\mathcal{T}_q^{(m)}} Z_{t,a_t^{(m)}}^{(m)} \geq n_{m,q}\delta\right) \leq e^{-\lambda n_{m,q}\delta}\prod_{t\in\mathcal{T}_q^{(m)}}\sum_{a\in[K]}\mathbb{E}\left[\exp\left(\lambda Z_{t,a}^{(m)}\right)\right]. \tag{H.1}$$

Since $x_{t,a}^{(m)}$ is assumed to be L-sub-Gaussian and $\|v\|_2 = 1$, $\langle v, x_{t,a}\rangle$ is also L-sub-Gaussian. As a result, $Z_{t,a}^{(m)} = \langle v, x_{t,a}^{(m)}\rangle^2$ is $(4\sqrt{2}L, 4L)$-sub-exponential (Vershynin, 2018). By the sub-Gaussianity of $\langle v, x_{t,a}^{(m)}\rangle$, using Lemma B.9, we have $\mathbb{E}[Z_{t,a}^{(m)}] = \mathbb{E}[\langle v, x_{t,a}^{(m)}\rangle^2] \leq 2L(\ln(2) + 1)$. Applying Bernstein's concentration inequality for sub-exponential variables, for $\delta \geq$

$2L(\ln(2)+1) \geq \mathbb{E}[Z_{t,a}^{(m)}]$ we have

$$\begin{aligned} e^{-\lambda\delta}\mathbb{E}\left[\exp\left(\lambda Z_{t,a}^{(m)}\right)\right] \leq & e^{-\lambda(\delta-\mathbb{E}[Z_{t,a}^{(m)}])}\mathbb{E}\left[\exp\left(\lambda(Z_{t,a}-\mathbb{E}[Z_{t,a}^{(m)}])\right)\right] \\ \leq & \exp\left(-\min\left\{\frac{(\delta-\mathbb{E}[Z_{t,a}^{(m)}])^2}{64L}, \frac{\delta-\mathbb{E}[Z_{t,a}^{(m)}]}{32L}\right\}\right) \\ \leq & \exp\left(-\min\left\{\frac{(\delta-2L(\ln(2)+1))^2}{64L}, \frac{\delta-2L(\ln(2)+1)}{32L}\right\}\right). \end{aligned} \tag{H.2}$$

Combining (H.2) with (H.1), we obtain for $\delta \geq 2L(\ln(2)+1)$,

$$\begin{aligned} &\mathbb{P}\left(v^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})v \geq \delta\right) = \mathbb{P}\left(\sum_{t\in\mathcal{T}_q^{(m)}} Z_{t,a_t^{(m)}}^{(m)} \geq n_{m,q}\delta\right) \\ \leq &\exp\left(-\left(\min\left\{\frac{(\delta-2L(\ln(2)+1))^2}{64L}, \frac{\delta-2L(\ln(2)+1)}{32L}\right\}+\ln(K)\right)n_{m,q}\right) \triangleq p_{q,\delta}^{(m)}. \end{aligned} \tag{H.3}$$

Now, we consider an ε-net $\mathcal{N}_d(\varepsilon)$ of the source ball $\mathbb{B}_d$ with cardinality at most $(1+2/\varepsilon)^d$ (Vershynin, 2018, Corollary 4.2.13). By applying a union bound to (H.3), we have with probability at least $1-(1+2/\varepsilon)^d\sum_{m\in\mathcal{C}_q} p_{q,\delta}^{(m)}$ that $v^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})v \leq \delta$ holds for any $v \in \mathcal{N}_d(\varepsilon)$ and $m \in \mathcal{C}_q$. Now taking any source vector $u \in \mathbb{R}^d$, there exists $v \in \mathcal{N}_d(\varepsilon)$ such that $\|u-v\|_2 \leq \varepsilon$. Furthermore, by symmetry of $(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})$, we have $u^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})v = v^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})u$ and thus

$$\begin{aligned} &u^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})u - v^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})v = (u+v)^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})(u-v) \\ \leq &\varepsilon\left\|\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right)(u+v)\right\|_2 \leq 2\varepsilon\lambda_{\max}\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right). \end{aligned}$$

Rearranging the above inequality gives

$$u^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})u \leq 2\varepsilon\lambda_{\max}\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right)+\delta.$$

Taking the supremum with respect to u, we obtain

$$\lambda_{\max}\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right) \leq \frac{\delta}{1-2\varepsilon}. \tag{H.4}$$

Next we bound $\lambda_{\min}(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})$. For any source vector $v \in \mathbb{R}^d$, we have

$$v^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})v = v^\top\left(\frac{1}{n_{m,q}}\sum_{t\in\mathcal{T}_q^{(m)}} x_{t,a_t^{(m)}}^m x_{t,a_t^{(m)}}^{m\top}\right)v \geq \frac{\mu}{n_{m,q}}\sum_{t\in\mathcal{T}_q^{(m)}} \mathbb{1}(\langle v, x_{t,a_t^{(m)}}^{(m)}\rangle^2 \geq \mu).$$

Since $a_t^{(m)} = \arg\max_{a\in[K]}\langle\widehat{\beta}_{q-1}^{(m)}, x_{t,a}^{(m)}\rangle$ for any $t \in \mathcal{T}_q^{(m)}$ and $m \in [M]$, by Condition 9, we have

$$\mathbb{E}[\mathbb{1}(\langle v, x_{t,a_t^{(m)}}^{(m)}\rangle^2 \geq \mu)] = \mathbb{P}(\langle v, x_{t,a_t^{(m)}}^{(m)}\rangle^2 \geq \mu) \geq c_x.$$

Therefore, by applying the Chernoff bound, we have

$$\mathbb{P}\left(v^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})v \le \mu c_x/2\right) \le e^{-c_x n_{m,q}/8} \triangleq \tilde{p}_q^{(m)}. \tag{H.5}$$

By applying a union bound to (H.5), we have with probability at least $1-(1+2/\varepsilon)^d \sum_{m\in\mathcal{C}_q} \tilde{p}_q^{(m)}$ that $v^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})v \ge \mu c_x/2$ holds for any $v \in \mathcal{N}_d(\varepsilon)$ and $m \in \mathcal{C}_q$. Taking any source vector $u \in \mathbb{R}^d$, there exists $v \in \mathcal{N}_d(\varepsilon)$ such that $\|u-v\|_2 \le \varepsilon$. Furthermore, by the symmetry of $\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}$, we have

$$\begin{aligned}
&u^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})u - v^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})v = (u+v)^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})(u-v)\\
\ge& -\varepsilon \left\|\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right)(u+v)\right\|_2 \ge -2\varepsilon\lambda_{\max}\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right).
\end{aligned}$$

Rearranging the above inequality, we obtain

$$u^\top(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q})u \ge \frac{\mu c_x}{2} - 2\varepsilon\lambda_{\max}\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right). \tag{H.6}$$

Taking the infimum in (H.6) with respect to u and using (H.4), we obtain

$$\lambda_{\min}\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right) \ge \frac{\mu c_x}{2} - \frac{2\varepsilon\delta}{1-2\varepsilon}. \tag{H.7}$$

Finally, letting $\delta = 32\max\{L,\sqrt{L}\}(\ln(K)+1)$ and $\varepsilon = \mu c_x/(8\delta + 2\mu c_x)$, we have $2\varepsilon\delta/(1-2\varepsilon) = \mu c_x/4$ and $p_{q,\delta}^{(m)} \le e^{-n_{m,q}} \le e^{-c_x n_{m,q}/8} = \tilde{p}_q^{(m)}$. Therefore, $\lambda_{\min}\left(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^{(m)}/n_{m,q}\right) \ge \mu c_x/4$ holds for all $m \in \mathcal{C}_q$ with a probability of at least

$$\begin{aligned}
&1-(1+2/\varepsilon)^d \sum_{m\in\mathcal{C}_q}(\tilde{p}_q^{(m)} + p_{q,\delta}^{(m)}) \ge 1 - 2(1+2/\varepsilon)^d \sum_{m\in\mathcal{C}_q} e^{-c_x n_{m,q}/8}\\
=&1 - 2\left(5 + 512\max\{L,\sqrt{L}\}(\ln(K)+1)/(\mu c_x)\right)^d \sum_{m\in\mathcal{C}_q} e^{-c_x n_{m,q}/8}\\
\ge&1 - \exp\left(-\frac{c_x \min_{m\in\mathcal{C}_q} n_{m,q}}{8} + \ln(2M) + d\ln\left(5 + 512\max\{L,\sqrt{L}\}(\ln(K)+1)/(\mu c_x)\right)\right).
\end{aligned}$$

In particular, there exists $C_{\mathrm{b}} \ge 2$ depending only on c_x, such that when $\min_{m\in\mathcal{C}_q} n_{m,q} \ge C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))$, the probability is lower bounded bounded by $1-1/T$. □

H.2 Proof of Lemma 1

Proof. Since $n_{m,0} = \sum_{t\in\mathcal{H}_0}\mathbb{1}(m\in\mathcal{S}_t)$ and $n_{m,q} = n_{m,q-1}\mathbb{1}\{m\notin\mathcal{C}_{q-1}\} + \sum_{t\in\mathcal{H}_q}\mathbb{1}(m\in\mathcal{S}_t) \ge \sum_{t\in\mathcal{H}_q}\mathbb{1}(\in\mathcal{S}_t)$ for any $1\le q<Q$, by using Bernstein's inequality (B.1), we have for each $m\in[\tau]$,

$$\mathbb{P}\left(n_{m,q} < \frac{p_m|\mathcal{H}_q|}{2}\right) \le \exp\left(-\frac{|\mathcal{H}_q|p_m^2}{2(p_m(1-p_m)+p_m/2)}\right) \le \exp\left(-|\mathcal{H}_q|p_m/3\right) \le \frac{1}{2MT}, \tag{H.8}$$

where the last inequality holds because $C_{\mathrm{b}} \geq 2$ and thus $|\mathcal{H}_q| \geq 2C_{\mathrm{b}} \ln(MT)/p_\tau \geq 3\ln(2MT)/p_\tau$. Similarly, we have $\mathbb{P}\left(n_{m,q} > 3p_m|\mathcal{H}_q|/2\right) \geq 1/(2MT)$ for all $m \in [\tau]$. For each $0 \leq q < Q$, define the event

$$\mathcal{F}_q \triangleq \{p_m|\mathcal{H}_q|/2 \leq n_{m,q} \leq 3p_m|\mathcal{H}_q|/2 \text{ for all } m \in [\tau]\}.$$

Using (H.8) and applying the union bound over all $m \in [\tau]$, we have $\mathbb{P}(\mathcal{F}_q) \geq 1 - 1/T$. Furthermore, by the condition on $|\mathcal{H}_q|$, we have for all $m \in [\tau]$ that

$$p_m|\mathcal{H}_q|/2 \geq C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu).$$

Therefore, on the event $\mathcal{F}_q$, we have $[\tau] \subseteq \mathcal{C}_q$, with $\mathcal{C}_q$ from (6). For each $0 \leq q < Q$, define the event

$$\mathcal{E}_q \triangleq \left\{\lambda_{\min}(\mathbf{X}_q^{(m)\top}\mathbf{X}_q^m) \geq n_{m,q}\mu c_x/4 \text{ for all } m \in \mathcal{C}_q\right\}.$$

By Lemma H.1, we have $\mathbb{P}(\mathcal{E}_q) \geq 1 - 1/T$. On the event $\mathcal{E}_q \cap \mathcal{F}_q$, which holds with probability at least $1 - 2/T$, using Condition 6, we have

$$\min_{1\leq m\leq|\mathcal{C}_q|} \frac{n_{1,q} \vee (n_{\mathcal{C}_q,q}/m)}{n_{m,q}} \leq 6 \min_{1\leq m\leq|\mathcal{C}_q|} \frac{p_1 \vee (p_{\mathcal{C}_q}/m)}{p_m} \leq 6c_{\mathrm{f}} = \widetilde{O}(1).$$

Therefore, by applying Theorem 1 with $p = 2$ and using definition of $\mathcal{E}_q$, we have for any $m \in \mathcal{C}_q$ that

$$\mathbb{E}\left[\|\widehat{\beta}_q^m - \beta^{(m)}\|_2^2 \mid (\mathbf{X}_q^{(m)}, Y_q^{(m)})_{m\in\mathcal{C}_q}\right] = \widetilde{O}\left(\frac{1}{\mu}\left(\frac{s}{n_{m,q}} + \frac{d}{n_{\mathcal{C}_q,q}}\right)\right). \tag{H.9}$$

Since event $\mathcal{F}_q$ implies $n_{m,q} \geq p_m|\mathcal{H}_q|/2$ and $[\tau] \subseteq \mathcal{C}_q$ for all $m \in \mathcal{C}_q$, we thus have, using the definition of τ,

$$n_{\mathcal{C}_q,q} \geq n_{[\tau],q} \geq |\mathcal{H}_q|\tau p_\tau/2 = \widetilde{O}\left(|\mathcal{H}_q|p_{[M]}\right). \tag{H.10}$$

Plugging (H.10) into (H.9), we reach the conclusion. □

H.3 Proof of Theorem 3

Proof. For any $t \in \mathcal{H}_q$, $0 \leq q \leq Q$, and $m \in [M]$, it holds that

$$\begin{aligned}
&\max_{a\in[K]}\langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)}\rangle \mathbb{1}(m \in \mathcal{S}_t) \leq \max_{a\in[K],a'\in[K]}\langle x_{t,a}^{(m)} - x_{t,a'}^{(m)}, \beta^{(m)}\rangle \mathbb{1}(m \in \mathcal{S}_t) \\
&= \left(\max_{a\in[K]}\langle x_{t,a}^{(m)}, \beta^{(m)}\rangle + \max_{a'\in[K]}|\langle x_{t,a'}^{(m)}, \beta^{(m)}\rangle|\right)\mathbb{1}(m \in \mathcal{S}_t) \leq 2\max_{a\in[K]}|\langle x_{t,a}^{(m)}, \beta^{(m)}\rangle|\mathbb{1}(m \in \mathcal{S}_t).
\end{aligned} \tag{H.11}$$

Also, by the definition of $a_t^{(m)}$, the instantaneous regret can be bounded as

$$\begin{aligned}
&\max_{a\in[K]}\langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)}\rangle \mathbb{1}(m \in \mathcal{S}_t) \leq \max_{a\in[K]}\langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)} - \widehat{\beta}_{q-1}^m\rangle \mathbb{1}(m \in \mathcal{S}_t) \\
&\leq \max_{a\in[K],a'\in[K]}\langle x_{t,a}^{(m)} - x_{t,a'}^{(m)}, \beta^{(m)} - \widehat{\beta}_{q-1}^m\rangle \mathbb{1}(m \in \mathcal{S}_t) \leq 2\max_{a\in[K]}|\langle x_{t,a}^{(m)}, \beta^{(m)} - \widehat{\beta}_{q-1}^m\rangle|\mathbb{1}(m \in \mathcal{S}_t).
\end{aligned} \tag{H.12}$$

Combinig (H.11) with (H.12), we obtain, for $m \in \mathcal{S}_t$,

$$\max_{a\in[K]}\langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)}\rangle \le 2\left(\max_{a\in[K]}|\langle x_{t,a}^{(m)}, \beta^{(m)}\rangle|\right) \wedge \left(\max_{a\in[K]}|\langle x_{t,a}^{(m)}, \beta^{(m)} - \widehat{\beta}_{q-1}^{m}\rangle|\right) \tag{H.13}$$

Taking the expectation of (H.13) multiplied by $\mathbb{1}(m \in \mathcal{S}_t)$, conditioned on $\widehat{\beta}_{q-1}^{(m)}$, by Condition 8 and Lemma B.2, we have

$$\begin{aligned}
&\mathbb{E}\left[\max_{a\in[K]}\langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)}\rangle \mathbb{1}(m \in \mathcal{S}_t) \mid \widehat{\beta}_{q-1}^{m}\right] \\
\le& 2\left(\|\beta^{(m)}\|_2 \wedge \|\beta^{(m)} - \widehat{\beta}_{q-1}^{m}\|_2\right)\sqrt{2\ln(2K)L}\cdot\mathbb{P}(m \in \mathcal{S}_t) \\
\le& 2p_m\left(1 \wedge \|\beta^{(m)} - \widehat{\beta}_{q-1}^{m}\|_2\right)\sqrt{2\ln(2K)L}.
\end{aligned} \tag{H.14}$$

Taking expectations in (H.14) with respect to $\widehat{\beta}_{q-1}^{(m)}$, we find

$$\mathbb{E}\left[\max_{a\in[K]}\langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)}\rangle \mathbb{1}(m \in \mathcal{S}_t)\right] \le 2\sqrt{2\ln(2K)L}p_m\left(\mathbb{E}[1 \wedge \|\widehat{\beta}_{q-1}^{m} - \beta^{(m)}\|_2]\right). \tag{H.15}$$

Given (H.15), the remaining key step is to bound the estimation error $\mathbb{E}[\|\widehat{\beta}_{q-1}^{m} - \beta^{(m)}\|_2]$ for all $0 \le q \le Q$ and $m \in [M]$. Letting $\tilde{Q} = \lceil \log_2(C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))/(H_0 p_\tau)\rceil + 3$ with C_{b} defined in Lemma H.1, we first bound $\mathbb{E}[\|\widehat{\beta}_{q-1}^{m} - \beta^{(m)}\|_2]$ for $m \in [\tau]$.

Case 1. When $0 \le q < \tilde{Q}$, using $1 \wedge \|\widehat{\beta}_{q-1}^{m} - \beta^{(m)}\|_2 \le 1$ in (H.15), we have

$$\sum_{t\in\mathcal{H}_q}\mathbb{E}\left[\max_{a\in[K]}\langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)}\rangle \mathbb{1}(m \in \mathcal{S}_t)\right] = \widetilde{O}(p_m|\mathcal{H}_q|\sqrt{L}). \tag{H.16}$$

Case 2. When $\tilde{Q} \le q \le Q$, we have by the definition of $\tilde{Q}$,

$$|\mathcal{H}_{q-1}| = 2^{q-2}H_0 \ge 2C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))/p_\tau.$$

Let $\mathcal{G}_{q-1}$ be the event that $[\tau] \subseteq \mathcal{C}_{q-1}$, with $\mathcal{C}_{q-1}$ from (6). From Theorem 1 with $p = 2$, it follows for all $m \in \mathcal{C}_{q-1}$ that

$$\mathbb{E}[1 \wedge \|\widehat{\beta}_{q-1}^{(m)} - \beta^{(m)}\|_2^2 \mid (\mathbf{X}_{q-1}^{(m)}, Y_{q-1}^{(m)})_{m\in\mathcal{C}_{q-1}}] = \widetilde{O}\left(\frac{1}{\mu|\mathcal{H}_{q-1}|}\left(\frac{s}{p_m} + \frac{d}{p_{[M]}}\right)\right).$$

Following the argument from Lemma 1, we know $\mathbb{P}(\mathcal{G}_{q-1}) \ge 1 - 2/T$. Marginalizing over $(\mathbf{X}_{q-1}^{(m)}, Y_{q-1}^{(m)})_{m\in\mathcal{C}_{q-1}}$ and using Jensen's inequality, we have for any $m \in [\tau]$ that

$$\mathbb{E}[1 \wedge \|\widehat{\beta}_{q-1}^{(m)} - \beta^{(m)}\|_2] = \widetilde{O}\left(\frac{1}{\sqrt{\mu|\mathcal{H}_{q-1}|}}\sqrt{\frac{s}{p_m} + \frac{d}{p_{[M]}}} + \frac{1}{T}\right) \tag{H.17}$$

where $1/T$ appears by considering the complement of $\mathcal{G}_{q-1}$, via the bound $1 \wedge \|\widehat{\beta}_{q-1}^{(m)} - \beta^{(m)}\|_2 \leq 2$. Plugging (H.17) into (H.15), we obtain

$$\begin{aligned}
&\sum_{t \in \mathcal{H}_q} \mathbb{E}\left[\max_{a \in [K]} \langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)} \rangle \mathbb{1}(m \in \mathcal{S}_t)\right] \\
=&\widetilde{O}\left(|\mathcal{H}_q| p_m \left(\sqrt{\frac{L}{\mu |\mathcal{H}_{q-1}|}\left(\frac{s}{p_m} + \frac{d}{p_{[M]}}\right)} + \frac{\sqrt{L}}{T}\right)\right) \\
=&\widetilde{O}\left(\sqrt{\frac{L|\mathcal{H}_q|(p_m)^2}{\mu}\left(\frac{s}{p_m} + \frac{d}{p_{[M]}}\right)} + \frac{\sqrt{L}|\mathcal{H}_q| p_m}{T}\right),
\end{aligned} \tag{H.18}$$

where the last equation holds because $|\mathcal{H}_q| \leq 2|\mathcal{H}_{q-1}|$. Combining the bounds (H.18) with (H.16) for the two cases, we obtain for each $m \in [\tau]$ that

$$\begin{aligned}
\mathbb{E}[R_T^{(m)}] =& \sum_{t=1}^{T} \mathbb{E}\left[\max_{a \in [K]} \langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)} \rangle \mathbb{1}(m \in \mathcal{S}_t)\right] \\
=& \left(\sum_{q=0}^{\tilde{Q}-1} + \sum_{q=\tilde{Q}}^{Q}\right) \sum_{t \in \mathcal{H}_q} \mathbb{E}\left[\max_{a \in [K]} \langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)} \rangle \mathbb{1}(m \in \mathcal{S}_t)\right] \\
=&\widetilde{O}\left(\sqrt{L} p_m \sum_{q=0}^{\tilde{Q}-1} |\mathcal{H}_q| + \sum_{q=\tilde{Q}}^{Q}\left(\sqrt{\frac{L|\mathcal{H}_q|(p_m)^2}{\mu}\left(\frac{s}{p_m} + \frac{d}{p_{[M]}}\right)} + \frac{\sqrt{L}|\mathcal{H}_q| p_m}{T}\right)\right).
\end{aligned} \tag{H.19}$$

By direct calculation, we have $\sum_{q=\tilde{Q}}^{Q} |\mathcal{H}_q|/T \leq 1$, while

$$\sum_{q=0}^{\tilde{Q}-1} |\mathcal{H}_q| = O(2^{\tilde{Q}} H_0) = O(C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))/(H_0 p_\tau) \times H_0) = \widetilde{O}(d/p_\tau),$$

and

$$\sum_{q=\tilde{Q}}^{Q} \sqrt{|\mathcal{H}_q|} = O\left(\sum_{q=\tilde{Q}}^{Q} 2^{q/2}\sqrt{H_0}\right) = O\left(2^{Q/2}\sqrt{H_0}\right) = O\left(\sqrt{T}\right).$$

Therefore, from (H.19), we obtain

$$\begin{aligned}
\mathbb{E}[R_T^{(m)}] =&\widetilde{O}\left(\sqrt{L} d p_m/p_\tau + \sqrt{\frac{L}{\mu}\left(s + \frac{d p_m}{p_{[M]}}\right) T p_m}\right) \\
=&\widetilde{O}\left(\sqrt{L} d + \sqrt{\frac{L}{\mu}\left(s + \frac{d p_m}{p_{[M]}}\right) T p_m}\right),
\end{aligned}$$

where the second equation is due to $p_m/p_\tau \leq c_{\mathrm{f}} = \widetilde{O}(1)$.

Next we bound the estimation errors for $m \notin [\tau]$, and thus also the regret. Let $Q_m = (\lceil \log_2(C_{\mathrm{b}}(\ln(MT) + d\ln(L\ln(K)/\mu))/(H_0 p^m)\rceil + 3) \wedge Q$ for each $m \notin [\tau]$. If p_m is sufficiently small such that $Q_m = Q$, then we have $Tp_m \le 4|\mathcal{H}_{Q-1}|p_m = \widetilde{O}(d)$, which, combined with (H.15), directly implies

$$\sum_{t=1}^{T} \mathbb{E}\left[\max_{a\in[K]} \langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta^{(m)}\rangle \mathbb{1}(m \in \mathcal{S}_t)\right] = O\left(\sqrt{L}Tp_m\right) = O\left(\sqrt{L}\cdot d \wedge (Tp_m)\right).$$

Otherwise if $Q_m < Q$, $Tp_m = \widetilde{\Omega}(d)$. In this case, we show that $m \in \mathcal{C}_{q-1}$ with high probability for $q \ge Q_m$. Using $n_{q-1}^{(m)} \ge \sum_{t\in\mathcal{H}_{q-1}} \mathbb{1}(m \in \mathcal{S}_t)$ for any $q \in [Q]$ and Bernstein's inequality (B.1), we have for each $m \in [\tau]$,

$$\mathbb{P}\left(n_{q-1}^{(m)} < \frac{p_m|\mathcal{H}_{q-1}|}{2}\right) \le \exp\left(-|\mathcal{H}_{q-1}|p_m/3\right) \le \frac{1}{MT}.$$

Letting $\mathcal{F}_{q-1}^{(m)} = \{n_{m,q-1} \ge p_m|\mathcal{H}_{q-1}|/2\}$, we have $\mathbb{P}(\mathcal{F}_{q-1}^{(m)}) \ge 1 - 1/(MT)$ and furthermore $\mathcal{F}_{q-1}^{(m)}$ implies $m \in \mathcal{C}_{q-1}$. Following the arguments in (H.17), (H.18), we similarly have that for any $q \ge Q_m$, $\mathbb{E}\left[1 \wedge \|\widehat{\beta}_{q-1}^m - \beta^{(m)}\|_2\right]$ is bounded by (H.18). Therefore, using $\sum_{q=0}^{Q^{(m)}-1} |\mathcal{H}_q| = \widetilde{O}(d/p^m)$ and $\sum_{q=Q_m}^{Q} \sqrt{|\mathcal{H}_q|} = O(\sqrt{T})$, we can proceed as in (H.19), and then bound

$$\sqrt{L}\sum_{q=0}^{Q_m-1} p_m|\mathcal{H}_q| = \widetilde{O}(\sqrt{L}\cdot d \wedge (Tp_m)).$$

to reach the desired conclusion.

□

H.4 Proof of Theorem 4

Proof. The proof strategy for the term $\Omega(\sqrt{(s + dp_m/p_{[M]})Tp_m})$ is similar to Theorem 2: we prove $\Omega(\sqrt{sTp_m})$ and $\Omega(\sqrt{dTp_m/p_{[M]}})$ by considering two cases: the *homogeneous* case where $\beta^{(1)} = \cdots = \beta^{(M)} = \beta^\star$ and the *s-sparse* case where $\beta^\star = 0$ and $\|\beta^{(m)}\|_0 \le s$ for all $m \in [M]$.

The homogeneous case. The lower bound of $\Omega(\sqrt{dT})$ for a single linear contextual bandit is proved in (Han et al., 2020; Chu et al., 2011). We will follow a similar method to prove the risk is bounded as $\Omega(\sqrt{dTp_m^2/p_{[M]}})$. Since $\beta^{(1)} = \cdots = \beta^{(M)} = \beta^\star$ in this case, we omit the superscript m in $\beta^{(m)}$ for simplicity.

We consider a 2-armed instance, *i.e.*, $K = 2$. Denote by Q the uniform distribution over $\{\beta \in \mathbb{R}^d : \|\beta\|_2 = \Delta\}$ where $\Delta \in [0, 1]$ will be specified later. We let Q be the prior distribution for β, and let $D \triangleq \mathcal{N}(0, L \cdot I_d)$ be the distribution of contexts. By (Ren and Zhou, 2023, Lemma 1), this choice of context distribution satisfies Conditions 8 and 9 with $c_x = \Theta(1)$.

Let $\mathcal{S}_t$ be the set of activated bandits at the t-th round and denote by $P_{\beta,x,t}$ the distribution of the observed rewards $\{\{y_\ell^{(m)}\mathbb{1}(m \in \mathcal{S}_\ell)\}_{m=1}^M\}_{\ell=1}^t$ up to time t, conditioned on

parameter β and contexts $\{x_{\ell,a}^{(m)} : a \in [2], m \in [M]\}_{\ell=1}^{t}$. Since the event $\{m \in \mathcal{S}_t\}$ is independent of the history and of the contexts $\{x_{t,a}^{(m)} : a \in [2]\}$ at the current round, we have

$$\sup_{\beta} \mathbb{E}[R_T^{(m)}(A)] \geq \mathbb{E}_{\beta\sim Q}[R_T^{(m)}(A)] = \sum_{t=1}^{T} \mathbb{E}_Q \left[\mathbb{E}_D \left[\mathbb{E}_{P_{\beta,x,t-1}} \left[p_m \max_{a\in[2]} \langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta\rangle\right]\right]\right] \tag{H.20}$$

where the factor of p_m in (H.20) is due to the integration over the randomness of $\{m \in \mathcal{S}_t\}$. Letting $d_t^{(m)} \triangleq x_{t,2}^{(m)} - x_{t,1}^{(m)}$ for all $m \in [M]$ and $t \geq 1$, we have

$$\max_{a\in[2]} \langle x_{t,a}^{(m)} - x_{t,a_t^{(m)}}^{(m)}, \beta\rangle = \mathbb{1}(a_t^{(m)} = 1)\langle d_t^{(m)}, \beta\rangle_+ + \mathbb{1}(a_t^{(m)} = 2)\langle d_t^{(m)}, \beta\rangle_-$$

where the subscripts $+$ and $-$ denote the positive and negative part respectively, *i.e.*, $u_+ = \max\{u, 0\}$ and $u_- = \max\{-u, 0\}$ for any $u \in \mathbb{R}$. Therefore, from (H.20), we have

$$\sup_{\beta} \mathbb{E}[R_T^{(m)}(A)] \geq p_m \sum_{t=1}^{T} \mathbb{E}_Q \left[\mathbb{E}_D \left[\mathbb{E}_{P_{\beta,x,t-1}} \left[\mathbb{1}(a_t^{(m)} = 1)\langle d_t^{(m)}, \beta\rangle_+ + \mathbb{1}(a_t^{(m)} = 2)\langle d_t^{(m)}, \beta\rangle_-\right]\right]\right]. \tag{H.21}$$

For any $1 \leq t \leq T$, conditioned on $d_t^{(m)}$, we define two new measures $Q_t^{(m)+}$ and $Q_t^{(m)-}$ over $\mathbb{R}^d$ via the Radon–Nikodym derivatives $\mathrm{d}Q_t^{(m)+}(\beta) = \langle d_t^{(m)}, \beta\rangle_+ / Z(d_t^{(m)})\mathrm{d}Q$ and $\mathrm{d}Q_t^{(m)-}(\beta) = \langle d_t^{(m)}, \beta\rangle_- / Z(d_t^{(m)})\mathrm{d}Q$, for all $\beta \in \mathbb{R}^d$, where $Z(d_t^{(m)}) = \mathbb{E}_Q[\langle d_t^{(m)}, \beta\rangle_+] = \mathbb{E}_Q[\langle d_t^{(m)}, \beta\rangle_-]$ is a normalization factor. Here $\mathbb{E}_Q[\langle d_t^{(m)}, \beta\rangle_+] = \mathbb{E}_Q[\langle d_t^{(m)}, \beta\rangle_-]$ due to the symmetry of Q. Plugging the definitions of $Q_t^{(m)+}$ and $Q_t^{(m)-}$ into (H.21), and changing the integration order, we have

$$\begin{aligned}
\sup_{\beta} \mathbb{E}[R_T^{(m)}(A)] \geq & p_m \sum_{t=1}^{T} \mathbb{E}_D \left[Z(d_t^{(m)}) \left(\mathbb{E}_{P_{\beta,x,t-1}\circ Q_t^{(m)+}}[\mathbb{1}(a_t = 1)] + \mathbb{E}_{P_{\beta,x,t-1}\circ Q_t^{(m)-}}[\mathbb{1}(a_t = 2)]\right)\right] \\
\geq & p_m \sum_{t=1}^{T} \mathbb{E}_D \left[Z(d_t^{(m)}) \left(1 - \mathtt{TV}(P_{\beta,x,t-1} \circ Q_t^{(m)+}, P_{\beta,x,t-1} \circ Q_t^{(m)-})\right)\right] \\
\geq & p_m \sum_{t=1}^{T} \mathbb{E}_D \left[Z(d_t^{(m)}) \left(1 - \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_{\beta,x,t-1} \circ Q_t^{(m)+} \,\|\, P_{\beta,x,t-1} \circ Q_t^{(m)-})}\right)\right],
\end{aligned} \tag{H.22}$$

where the second inequality follows the definition of the total variation distance: $P_1(A) + P_2(A^c) \geq 1 - \mathtt{TV}(P_1, P_2)$ with $A = \{a_t = 1\}$, and the third inequality follows from Pinsker's inequality $\mathtt{TV}(P_1, P_2) \leq \sqrt{D_{\mathrm{KL}}(P_1 \,\|\, P_2)/2}$.

Due to the distribution of the contexts, $d_t^{(m)} \neq 0$ with probability one. Hence we can let $u_t^{(m)} = d_t^{(m)}/\|d_t^{(m)}\|_2$ if $d_t^{(m)} \neq 0$; and the zero-probability set where $d_t^{(m)} = 0$ does not affect the result. Further, let

$$\tilde{\beta}_t^m = \beta - 2\langle u_t^{(m)}, \beta\rangle u_t^{(m)}. \tag{H.23}$$

Since $\langle d_t^{(m)}, \tilde{\beta}_t^m \rangle = -\langle d_t^{(m)}, \tilde{\beta}_t^m \rangle$, we have $\langle d_t^{(m)}, \beta \rangle_- = \langle d_t^{(m)}, \tilde{\beta}_t^m \rangle_+$. Furthermore, (H.23) has the inverse transformation

$$\beta = \tilde{\beta}_t^m - 2\langle u_t^{(m)}, \tilde{\beta}_t^m \rangle u_t^{(m)}. \tag{H.24}$$

Since $\langle d_t^{(m)}, \beta \rangle_- = \langle d_t^{(m)}, \tilde{\beta}_t^m \rangle_+$ and Q is reflection-invariant, one can equivalently obtain $\beta \sim Q_t^{(m)-}$ by first generating $\tilde{\beta}_t^{(m)} \sim Q_t^{(m)+}$ and then calculating β via (H.24). Since $P_{\beta,x,t-1} \circ Q_t^{(m)-}$ means drawing β from $Q_t^{(m)-}$ first and then gaining rewards given such β and the independently sampled contexts, it thus holds that

$$\begin{aligned} P_{\beta,x,t-1} \circ (\beta \sim Q_t^{(m)-}) =& P_{\tilde{\beta}_t^{(m)} - 2\langle u_t^{(m)}, \tilde{\beta}_t^{(m)} \rangle u_t^{(m)}, x, t-1} \circ (\tilde{\beta}_t^{(m)} \sim Q_t^{(m)+}) \\ \overset{d}{=}& P_{\beta - 2\langle u_t^{(m)}, \beta \rangle u_t^{(m)}, x, t-1} \circ (\beta \sim Q_t^{(m)+}). \end{aligned}$$

In the second equation above we changed the notation from $\tilde{\beta}_t^{(m)} \mapsto \beta$. From this and (H.22), we have

$$\begin{aligned} &\sup_\beta \mathbb{E}[R_T^{(m)}(A)] \\ \geq& p_m \sum_{t=1}^T \mathbb{E}_D \left[Z(d_t^{(m)}) \left(1 - \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_{\beta,x,t-1} \circ Q_t^{(m)+} \,\|\, P_{\beta - 2\langle u_t^{(m)}, \beta \rangle u_t^{(m)}, x, t-1} \circ Q_t^{(m)+}} \right) \right]. \end{aligned}$$

By Lemma B.7, this is lower bounded by

$$p_m \sum_{t=1}^T \mathbb{E}_D \left[Z(d_t^{(m)}) \left(1 - \sqrt{\frac{1}{2} \mathbb{E}_{\beta \sim Q_t^{(m)+}} \left[D_{\mathrm{KL}}(P_{\beta,x,t-1} \,\|\, P_{\beta - 2\langle u_t^{(m)}, \beta \rangle u_t^{(m)}, x, t-1}) \right]} \right) \right]. \tag{H.25}$$

Since the reward noise follows a $\mathcal{N}(0,1)$ distribution, conditioned on the activation sets $\{\mathcal{S}_\ell\}_{\ell=1}^{t-1}$, we have $P_{\beta,x,t-1} = \otimes_{\ell=1}^{t-1} \otimes_{r \in \mathcal{S}_\ell} \mathcal{N}(\langle \beta, x_{\ell, a_\ell^{(r)}}^{(r)} \rangle, 1)$. Furthermore, by the formula for the Kullback-Leibler divergence between two Gaussian distributions, we have

$$\begin{aligned} &D_{\mathrm{KL}} \left(P_{\beta,x,t-1} \,\|\, P_{\beta - 2\langle u_t^{(m)}, \beta \rangle u_t^{(m)}, x, t-1} \mid \{\mathcal{S}_\ell\}_{\ell=1}^{t-1} \right) \\ =& \frac{1}{2} \sum_{\ell=1}^{t-1} \sum_{r \in \mathcal{S}_\ell} \left(\langle \beta, x_{\ell, a_\ell^{(r)}}^{(r)} \rangle - \left\langle \beta - 2\langle u_t^{(m)}, \beta \rangle u_t^{(m)}, x_{\ell, a_\ell^{(r)}}^{(r)} \right\rangle \right)^2 \\ =& 2\langle u_t^{(m)}, \beta \rangle^2 \sum_{\ell=1}^{t-1} \sum_{r \in \mathcal{S}_\ell} \langle u_t^{(m)}, x_{\ell, a_\ell^{(r)}}^{(r)} \rangle^2 = 2\langle u_t^{(m)}, \beta \rangle^2 \sum_{\ell=1}^{t-1} \sum_{r=1}^{M} \left\langle u_t^{(m)}, x_{\ell, a_\ell^{(r)}}^{(r)} \right\rangle^2 \mathbb{1}(r \in \mathcal{S}_\ell). \end{aligned} \tag{H.26}$$

Since the events $\{r \in \mathcal{S}_\ell\}$ for $\ell \in [t-1]$ and $r \in [M]$ are independent of contexts and the variable β, using (H.26) and Lemma B.7, we obtain

$$\begin{aligned} &D_{\mathrm{KL}}(P_{\beta,x,t-1} \,\|\, P_{\beta - 2\langle u_t^{(m)}, \beta \rangle u_t^{(m)}, x, t-1}) \\ \leq& \mathbb{E}_{\{\mathcal{S}_\ell\}_{\ell=1}^{t-1}} \left[D_{\mathrm{KL}} \left(P_{\beta,x,t-1} \,\|\, P_{\beta - 2\langle u_t^{(m)}, \beta \rangle u_t^{(m)}, x, t-1} \mid \{\mathcal{S}_\ell\}_{\ell=1}^{t-1} \right) \right] \\ =& 2\langle u_t^{(m)}, \beta \rangle^2 \sum_{\ell=1}^{t-1} \sum_{r \in [M]} \langle u_t^{(m)}, x_{\ell, a_\ell^{(r)}}^{(r)} \rangle^2 \mathbb{P}(r \in \mathcal{S}_\ell) = 2\langle u_t^{(m)}, \beta \rangle^2 p_{[M]} \sum_{\ell=1}^{t-1} \langle u_t^{(m)}, x_{\ell, a_\ell^{(r)}}^{(r)} \rangle^2. \end{aligned} \tag{H.27}$$

Therefore, combining (H.25) and (H.27), we have

$$\sup_{\beta} \mathbb{E}[R_T^{(m)}(A)] \tag{H.28}$$

$$\geq p_m \sum_{t=1}^{T} \mathbb{E}_D \left[Z(d_t^{(m)}) \left(1 - \sqrt{\mathbb{E}_{Q_t^{(m)+}}[\langle u_t^{(m)}, \beta \rangle^2] u_t^{m\top} \left(p_{[M]} \sum_{\ell=1}^{t-1} x_{\ell, a_\ell^{(r)}}^{(r)} x_{\ell, a_\ell^{(r)}}^{(r)\top} \right) u_t^m} \right) \right]$$

$$\geq p_m \sum_{t=1}^{T} \mathbb{E}_D \left[Z(d_t^{(m)}) \left(1 - \sqrt{p_{[M]} \mathbb{E}_{Q_t^{(m)+}}[\langle u_t^{(m)}, \beta \rangle^2] u_t^{m\top} \sum_{\ell=1}^{t-1} \left(x_{\ell,1}^{(r)} x_{\ell,1}^{(r)\top} + x_{\ell,2}^{(r)} x_{\ell,2}^{(r)\top} \right) u_t^{(m)}} \right) \right]. \tag{H.29}$$

Taking the expectation of (H.29) with respect to $\{(x_{\ell,1}^r, x_{\ell,2}^r) : r \in [M]\}_{\ell=1}^{t-1}$, each of which is distributed i.i.d. according to $D = \mathcal{N}(0, L \cdot I_{d\times d})$, and using that $\|u_t^{(m)}\|_2 = 1$ with probability one, we have that the above is lower bounded by

$$p_m \sum_{t=1}^{T} \mathbb{E}_{(x_{t,1}^{(m)}, x_{t,2}^{(m)})} \left[Z(d_t^{(m)}) \left(1 - \sqrt{2(t-1) L p_{[M]} \mathbb{E}_{Q_t^{(m)+}}[\langle u_t^{(m)}, \beta \rangle^2]} \right) \right], \tag{H.30}$$

where the outer expectation is only over the randomness of $(x_{t,1}^{(m)}, x_{t,2}^{(m)})$. We next calculate $\mathbb{E}_{Q_t^+}[\langle u_t, \beta \rangle^2]$ and $Z(d_t^{(m)})$. By the definition of Q_t^+, we have

$$\mathbb{E}_{Q_t^{(m)+}}[\langle u_t^{(m)}, \beta \rangle^2] = \frac{\mathbb{E}_Q \left[|\langle u_t^{(m)}, \beta \rangle|^3 \right]}{\mathbb{E}_Q \left[|\langle u_t^{(m)}, \beta \rangle| \right]}. \tag{H.31}$$

By the symmetry of Q, the distribution of $\langle u_t^{(m)}, \beta \rangle$, conditioned on any $u_t^{(m)}$, is identical to that of the first coordinate of β. Therefore, using Lemma B.8, we have

$$\mathbb{E}_Q \left[|\langle u_t^{(m)}, \beta \rangle|^3 \right] = \Delta^3 \frac{\Gamma(\frac{d}{2})\Gamma(2)}{\Gamma(\frac{d+3}{2})\Gamma(\frac{1}{2})} \quad \text{and} \quad \mathbb{E}_Q \left[|\langle u_t^{(m)}, \beta \rangle| \right] = \Delta \frac{\Gamma(\frac{d}{2})\Gamma(1)}{\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})},$$

and thus it follows from (H.31) that

$$\mathbb{E}_{Q_t^{(m)+}}[\langle u_t^{(m)}, \beta \rangle^2] = \frac{2\Delta^2}{d+1}. \tag{H.32}$$

Similarly, we have

$$\mathbb{E}_{(x_{t,1}^{(m)}, x_{t,2}^{(m)})}[Z(d_t^{(m)})] = \mathbb{E}_{(x_{t,1}^{(m)}, x_{t,2}^{(m)})}[\mathbb{E}_Q[\langle d_t^{(m)}, \beta \rangle_+]] = \frac{1}{2} \mathbb{E}_{(x_{t,1}^{(m)}, x_{t,2}^{(m)})}[\mathbb{E}_Q[|\langle d_t^{(m)}, \beta \rangle|]]$$

$$= \frac{1}{2} \mathbb{E}_{(x_{t,1}^{(m)}, x_{t,2}^{(m)})} \left[\|d_t^{(m)}\|_2 \right] \frac{\Delta \Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}} = \frac{1}{2} \mathbb{E} \left[\|x_{t,1}^{(m)} - x_{t,2}^{(m)}\|_2 \right] \frac{\Delta \Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}} = \Omega(\sqrt{L}\Delta). \tag{H.33}$$

Combining (H.30), (H.32), and (H.33), we have

$$\sup_{\beta} \mathbb{E}[R_T^{(m)}(A)] \geq \Omega(p_m \sqrt{L} \Delta T) \cdot \left(1 - \sqrt{\frac{4(t-1)L\Delta^2}{(d+1)} p_{[M]}} \right). \tag{H.34}$$

Choosing $\Delta = \frac{\sqrt{(d+1)}}{4\sqrt{(T-1)L\sum_{r\in[M]} p_r}}$, which satisfies $\Delta \leq 1$ by assumption, in (H.34), we finally establish

$$\sup_{\beta} \mathbb{E}[R_T(A)] = \Omega\left(p_m T\sqrt{L}\Delta\right) = \Omega\left(\sqrt{dTp_m^2/p_{[M]}}\right).$$

The s-sparse case. In this case, we consider $\mathrm{supp}(\beta^{(1)}), \ldots, \mathrm{supp}(\beta^{(M)})$ located in the first s coordinates. If the supports of $\{\beta^{(m)}\}_{m\in[M]}$ are known to the algorithm, then the structure of sparse heterogeneity and the common $\beta^\star$ would be non-informative for estimating $\{\beta^{(m)}\}_{m\in[M]}$. Therefore, in this case, one can obtain the lower bound $\Omega(\sqrt{sTp_m})$ by simply adapting the proof for the homogeneous case with $M = 1$ in s dimensions.

□

I Results on Contextual Bandits under Model-P

Recall that in the single contextual bandit problem under Model-P, we have a set of K parameters $\{\beta^{(a)}\}_{a\in[K]}$ where $\beta^{(a)}$ is associated with arm a. When action a is chosen, a reward $y_{t,a} = \langle x_t, \beta^{(a)}\rangle + \varepsilon_t$ is earned. We extend this to a multitask scenario as follows. We consider M bandit instances, and each bandit m is associated with K arms corresponding to parameters $\{\beta^{(m,a)}\}_{a\in[K]} \subseteq \mathbb{R}^d$, and an activation probability $p_m \in [0,1]$. At any time t, each bandit m is independently activated with probability p_m. The analyst observes an independent d-dimensional context $x_t^{(m)}$ for m in the set $\mathcal{S}_t$ of activated bandit instances. Given all observed contexts, the analyst can select action $a_t^{(m)} \in [K]$ for each activated bandit instance $m \in \mathcal{S}_t$ and earn the reward via $y_t^{(m)} = \langle x_t^{(m)}, \beta^{(m,a_t^{(m)})}\rangle + \varepsilon_t^{(m)} \in \mathbb{R}$, where $\varepsilon_t^{(m)}$ are i.i.d. noise random variables.

To study the proposed multitask scenario under Model-P, we impose the following conditions, which are parallel to Conditions 5, 8, and 9.

Condition I.1 (Sparse heterogeneity & Boundedness)**.** *There is s_a with $0 \le s_a \le d$ such that for each action $a \in [K]$, there is an unknown global parameter $\beta^{\star,(a)} \in \mathbb{R}^d$ with $\|\beta^{(m,a)} - \beta^{\star,(a)}\|_0 \le s$ for any $m \in [M]$. Furthermore, $\|\beta^{(m,a)}\|_2 \le 1$ for all $a \in [K]$ and $m \in [M]$.*

Condition I.2 (Sub-Gaussianity)**.** *For each $t \in [T]$, the marginal distribution of x_t is L-sub-Gaussian.*

Condition I.3 (Diverse covariate)**.** *There are positive constants μ and c_x, such that for any $\{\beta^{(m,a)} : a \in [K]\} \subseteq \mathbb{R}^d$, source vector $v \in \mathbb{R}^d$, and $m \in [M]$, it holds that $\mathbb{P}(\langle x_t^{(m)}, v\rangle^2 \mathbb{1}(a^\star = a) \ge \mu) \ge c_x$ where $a^\star = \arg\max_{a\in[K]} \langle x_t^{(m)}, \beta^{(m,a)}\rangle$ and the probability $\mathbb{P}(\cdot)$ is taken over the distribution of $x_t^{(m)}$.*

Remark I.1. *Condition I.3 ensures sufficient exploration even with a greedy algorithm. (Ren and Zhou, 2023, Lemma 14) proves that Condition I.3 holds when $\mathbb{E}[x_t x_t^\top] \succeq 2\mu I_d$ and $p(x_t) \ge \nu p(-x_t)$ for some $\nu > 0$ where $p(\cdot)$ is the density of x_t.*

I.1 Algorithm & Regret Analysis

Algorithm 3 MOALRBandit: Collaborative Bandits with MOLAR estimates under Model-P

Input: Time horizon T, $\widehat{\beta}_{a,-1}^{(m)} = 0$ for $a \in [K]$ and $m \in [M]$, initial batch size H_0 and batch $\mathcal{H}_0 = [H_0]$; number of batches $Q = \lceil \log_2(T/H_0) \rceil$, $\mathbf{X}_q^{(m,a)} = \emptyset$, and $Y_{a,q}^{(m)} = \emptyset$ for $a \in [K]$, $m \in [M]$, and $0 \le q \le Q$

for $q = 1, \ldots, Q$ **do**

Define batch $\mathcal{H}_q = \{t : 2^{q-1}H_0 < t \le \min\{2^q H_0, T\}\}$

end for

for $t = 1, \cdots, T$ **do**

for each bandit in parallel **do**

Bandit instance m is activated with probability p_m

if $t \in \mathcal{H}_q$ **and** bandit instance m is activated **then**

Choose, breaking ties randomly $a_t^{(m)} = \arg\max_{a \in [K]} \langle x_t^{(m)}, \widehat{\beta}_{q-1}^{(m,a)} \rangle$, and gain reward $y_t^{(m)}$

Augment observations $\mathbf{X}_q^{(m,a_t^{(m)})} \leftarrow [\mathbf{X}_q^{(m,a_t^{(m)})\top}, x_t^{(m)}]^\top$ and $Y_q^{(m,a_t^{(m)})} \leftarrow [Y_q^{(m,a_t^{(m)})\top}, y_t^{(m)}]^\top$

end if

end for

if $t = 2^q H_0$, *i.e.*, batch $\mathcal{H}_q$ ends **then**

Let $n_{m,q} = \sum_{a \in [K]} |Y_q^{(m,a)}|$ and $\mathcal{C}_q = \{m \in [M] : n_{m,q} \ge C'_{\mathrm{b}}(\ln(MKT) + d\ln(L/\mu))\}$ with C'_{b} defined in Lemma I.1

for $a \in [K]$ **do**

Call MOLAR($\{(\mathbf{X}_q^{(m,a)}, Y_q^{(m,a)})\}_{m \in \mathcal{C}_q}$) to obtain $\{\widehat{\beta}_q^{(m,a)}\}_{m \in \mathcal{C}_q}$

for $m \in [M] \backslash \mathcal{C}_{a,q}$ **do**

Let $\widehat{\beta}_q^{(m,a)} = \widehat{\beta}_{q-1}^{(m,a)}$, $\mathbf{X}_{q+1}^{(m,a)} = \mathbf{X}_q^{(m,a)}$, and $Y_{q+1}^{(m,a)} = Y_q^{(m,a)}$

end for

end for

end if

end for

Algorithm 3 describes a variant of the MOLARB algorithm under Model-P. While Algorithm 3 follows the spirit of MOLARB, the difference is that it requires applying MOLAR to all arms with the same index across all bandits due to the nature of Model-P.

We consider the following individual regret metric: given a time horizon $T \ge 1$ and a specific algorithm A that produces action trajectories $\{a_t^{(m)}\}_{t \in [T], m \in [M]}$, we define for each $m \in [M]$ that

$$R_T^{(m)}(A) := \sum_{t=1}^{T} \max_{a \in [K]} \langle x_t^{(m)}, \beta^{(m,a)} - \beta^{(m,a_t^{(m)})} \rangle \mathbb{1}(m \in \mathcal{S}_t).$$

Theorem I.1 establishes a corresponding regret upper bound under Conditions 6, 7, and I.1-I.3. To this end, we first show Lemma I.1, which guarantees that for any given $a \in [K]$ and $0 \le q < Q$, the contexts $\mathcal{X}_{a,q}^{(m)}$ at the end of each batch for all $m \in [M]$ has lower bounded eigenvalues with high probability. Lemma I.1 is similar to (Ren and Zhou, 2023, Lemma 18) in the single-bandit and s-sparse regime.

Lemma I.1. *Under Conditions I.2 and I.3, there is C'_{b} only depending on c_x, such that for any $0 \le q < Q$, it holds with probability at least $1-1/T$ that $\lambda_{\min}(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}) \ge n_{m,q}\mu c_x/4$ for all $a \in [K]$ and $m \in [M]$ with $n_{m,q} \ge C'_{\mathrm{b}}(\ln(MKT) + d\ln(L/\mu))$.*

Proof. The proof is similar to the proof of Lemma H.1. Hence we only sketch the key steps below. For $0 \le q < Q$ and $m \in [M]$, we let $\mathcal{T}_q^{(m)}$ be the set of times when contexts $\mathbf{X}_q^m$ are observed at instance m. Clearly, we have $|\mathcal{T}_q^{(m)}| = n_{m,q}$; and $\{x_t^{(m)} : t \in \mathcal{T}_q^{(m)}\}$ are independent, conditioned on $\{\widehat{\beta}_{q-1}^{(m,a)}\}_{a\in[K]}$. The following analysis is conditional on $\{\mathcal{T}_q^{(m)}\}_{m\in[M]}$ and therefore also on $\{n_{m,q}\}_{m\in[M]}$.

By definition, we have for any $a \in [K]$ and $m \in [M]$,

$$\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)} = \sum_{t\in\mathcal{T}_q^{(m)}} x_t^{(m)} x_t^{(m)\top} \mathbb{1}(a_t^{(m)} = a).$$

We first give an upper bound for $\lambda_{\max}(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q})$. For any source vector $v \in \mathbb{R}^d$, any $t \in [T]$, $a \in [K]$, and $m \in [M]$, let $Z_{t,a}^{(m)} = \langle v, x_t\rangle^2 \mathbb{1}(a_t^{(m)} = a)$. Conditionally on $\{\widehat{\beta}_{q-1}^{(m,a)}\}_{a\in[K]}$, for any $\delta > 0$ and $\lambda > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{t\in\mathcal{T}_q^{(m)}} Z_{t,a}^{(m)} \ge n_{m,q}\delta \mid \{\widehat{\beta}_{q-1}^{(m,a)}\}_{a\in[K]}\right) \le& e^{-\lambda n_{m,q}\delta}\mathbb{E}\left[\exp\left(\lambda\sum_{t\in\mathcal{T}_q^{(m)}} Z_{t,a}^{(m)}\right) \mid \{\widehat{\beta}_{q-1}^{(m,a)}\}_{a\in[K]}\right] \\
=& e^{-\lambda n_{m,q}\delta}\prod_{t\in\mathcal{T}_q^{(m)}} \mathbb{E}\left[\exp\left(\lambda Z_{t,a}^{(m)}\right) \mid \{\widehat{\beta}_{q-1}^{(m,a)}\}_{a\in[K]}\right].
\end{aligned}$$

Since x_t is assumed to be L-sub-Gaussian and $\|v\|_2 = 1$, $Z_{t,a}^{(m)} = \langle v, x_t\rangle^2 \mathbb{1}(a_t^{(m)} = a) \le \langle v, x_t\rangle^2$ is $(4\sqrt{2}L, 4L)$-sub-exponential (Vershynin, 2018). Following the argument in (H.1) and (H.2), we obtain

$$\begin{aligned}
&\mathbb{P}\left(v^\top(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q})v \ge \delta\right) = \mathbb{P}\left(\sum_{t\in\mathcal{T}_q^{(m)}} Z_{t,a}^{(m)} \ge n_{m,q}\delta L\right) \\
\le& \exp\left(-\min\left\{\frac{(\delta - 2L(\ln(2)+1))^2}{64L}, \frac{\delta - 2L(\ln(2)+1)}{32L}\right\} n_{m,q}\right) \triangleq p_{q,\delta}^{(m)}
\end{aligned}$$

for any $\delta \ge 2L(\ln(2)+1)$. Now, following the ε-net-arguments around (H.4), we have with probability at least $1 - (1+2/\varepsilon)^d \sum_{m\in\mathcal{C}_q} p_{q,\delta}^{(m)}$ that

$$\lambda_{\max}\left(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q}\right) \le \frac{\delta}{1-2\varepsilon}. \tag{I.1}$$

Next we bound $\lambda_{\min}(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q})$. For any source vector $v \in \mathbb{R}^d$, we have

$$\begin{aligned} v^\top(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q})v &= v^\top \left(\frac{1}{n_{m,q}} \sum_{t\in\mathcal{T}_q^{(m)}} x_t^m x_t^{m\top} \mathbb{1}(a_t^{(m)} = a) \right) v \\ \geq & \frac{\mu}{n_{m,q}} \sum_{t\in\mathcal{T}_q^{(m)}} \mathbb{1}(\langle v, x_t^{(m)}\rangle^2 \geq \mu)\mathbb{1}(a_t^{(m)} = a). \end{aligned}$$

Since $a_t^{(m)} = \arg\max_{a\in[K]}\langle \widehat{\beta}_{q-1}^{(m,a)}, x_t^{(m)}\rangle$ for any $t \in \mathcal{T}_q^{(m)}$ and $m \in [M]$, by Condition I.3, we have

$$\mathbb{E}[\mathbb{1}(\langle v, x_t^{(m)}\rangle^2 \geq \mu)\mathbb{1}(a_t^{(m)} = a)] = \mathbb{P}(\langle x_t^{(m)}, v\rangle^2 \mathbb{1}(a_t^{(m)} = a) \geq \mu) \geq c_x.$$

Therefore, by applying a Chernoff bound, we have

$$\mathbb{P}\left(v^\top(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q})v \leq \mu c_x/2\right) \leq e^{-c_x n_{m,q}/8} \triangleq \tilde{p}_q^{(m)}. \tag{I.2}$$

Then, using an ε-net argument and applying a union bound to (I.2), we have with probability at least $1 - (1+2/\varepsilon)^d \sum_{m\in\mathcal{C}_q} \tilde{p}_q^{(m)}$ that $v^\top(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q})v \geq \mu c_x/2$ holds for any $v \in \mathcal{N}_d(\varepsilon)$ and $m \in \mathcal{C}_q$. Therefore, following the argument around (H.7), we obtain

$$\lambda_{\min}\left(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q}\right) \geq \frac{\mu c_x}{2} - \frac{2\varepsilon\delta}{1-2\varepsilon}.$$

Finally, letting $\delta = 32\max\{L, \sqrt{L}\}$ and $\varepsilon = \mu c_x/(8\delta + 2\mu c_x)$, we have $2\varepsilon\delta/(1-2\varepsilon) = \mu c_x/4$ and $p_{q,\delta}^{(m)} \leq e^{-n_{a,q}^{(m)}} \leq e^{-c_x n_{m,q}/8} = \tilde{p}_q^{(m)}$. Therefore, $\lambda_{\min}\left(\mathbf{X}_q^{(m,a)\top}\mathbf{X}_q^{(m,a)}/n_{m,q}\right) \geq \mu c_x/4$ holds for all $a \in [K]$ and $m \in \mathcal{C}_q$ with probability at least

$$\begin{aligned} & 1 - (1+2/\varepsilon)^d K \sum_{m\in\mathcal{C}_q} (\tilde{p}_q^{(m)} + p_{q,\delta}^{(m)}) \geq 1 - 2(1+2/\varepsilon)^d K \sum_{m\in\mathcal{C}_q} e^{-c_x n_{m,q}/8} \\ \geq & 1 - \exp\left(-\frac{c_x \min_{m\in\mathcal{C}_q} n_{m,q}}{8} + \ln(2MK) + d\ln\left(5 + 512\max\{L,\sqrt{L}\}/(\mu c_x)\right)\right). \end{aligned}$$

In particular, there is $C_{\mathrm{b}}' \geq 3$ depending only on c_x, such that when $\min_{m\in\mathcal{C}_q} n_{a,q}^{(m)} \geq C_{\mathrm{b}}'(\ln(MKT) + d\ln(L/\mu))$, the probability is lower bounded by $1 - 1/T$. □

Given Lemma I.1, using Theorem 1, we can bound the ℓ_2 estimation error $\mathbb{E}[\max_{a\in[K]} \|\widehat{\beta}_t^{(m,a)} - \beta^{(m,a)}\|_2]$ for all $m \in \mathcal{C}_q$ at the end of batch $\mathcal{H}_q$ as follows.

Lemma I.2. *Under Conditions 6, 7, and I.1-I.3, for any $0 \leq q < Q$, letting $\tau = \arg\min_{m\in[M]}(p^1 \vee \sum_{\ell\in[M]} p^\ell/m))/p^m$, if $|\mathcal{H}_q| \geq 2C_{\mathrm{b}}'(\ln(MKT) + d\ln(L/\mu))/p_\tau$ with C_{b}' defined in Lemma I.1, it holds with probability at least $1 - 2/T$ that for all $a \in [K]$ and $q \in \mathcal{C}_q$,*

$$\mathbb{E}[\max_{a\in[K]} \|\widehat{\beta}_q^{(m,a)} - \beta^{(m,a)}\|_2^2 \mid (\mathbf{X}_q^{(m,a)}, Y_q^{(m,a)})_{a\in[K],m\in\mathcal{C}_q}] = \widetilde{O}\left(\frac{1}{\mu|\mathcal{H}_q|}\left(\frac{s}{p_m} + \frac{d}{p_{[M]}}\right)\right),$$

where the expectation is taken with respect to the randomness of the noise, and logarithmic factors as well as quantities depending only on c_x, c_{f} are absorbed into $\widetilde{O}(\cdot)$.

Proof. The proof is essentially the same as the proof of Lemma 1. We thus omit the proof. □

Based on Lemma I.2, we can bound the individual regret as follows.

Theorem I.1. *Under Conditions 6, 7, and I.1-I.3, for any $T \geq 1$ and $1 \leq H_0 \leq d$, the expected regret of MOLARB under* Model-P*, for any $T \geq 1$, is bounded as*

$$\mathbb{E}[R_T^{(m)}] = \widetilde{O}\left(\cdot d \wedge (Tp_m) + \sqrt{\left(s + \frac{dp_m}{p_{[M]}}\right) Tp_m}\right),$$

where logarithmic factors as well as quantities depending only on c_x, c_{f} are absorbed into $\widetilde{O}(\cdot)$.

Proof. For any $t \in \mathcal{H}_q$, $0 \leq q \leq Q$, and $m \in [M]$, we have for $m \in \mathcal{S}_t$,

$$\max_{a\in[K]} \langle x_t^{(m)}, \beta^{(m,a)} - \beta^{(m,a_t^{(m)})} \rangle \leq \max_{a,a'\in[K]} \langle x_t^{(m)}, \beta^{(m,a)} - \beta^{(m,a')} \rangle \leq 2 \max_{a\in[K]} |\langle x_t^{(m)}, \beta^{(m,a)} \rangle|. \quad \text{(I.3)}$$

Also, from the definition of $a_t^{(m)}$, we have $\langle x_t^{(m)}, \widehat{\beta}_{q-1}^{(m,a)} \rangle \leq \langle x_t^{(m)}, \widehat{\beta}_{q-1}^{(m,a_t^{(m)})} \rangle$ for any $a \in [K]$. Therefore, the instantaneous regret can be bounded, for $m \in \mathcal{S}_t$, as

$$\begin{aligned}
&\max_{a\in[K]} \langle x_t^{(m)}, \beta^{(m,a)} - \beta^{(m,a_t^{(m)})} \rangle = \max_{a\in[K]} \langle x_t^{(m)}, \beta^{(m,a)} - \widehat{\beta}_{q-1}^{(m,a)} + \widehat{\beta}_{q-1}^{(m,a)} - \widehat{\beta}_{q-1}^{(m,a_t^{(m)})} \rangle \\
\leq &\max_{a\in[K]} \langle x_t^{(m)}, \beta^{(m,a)} - \widehat{\beta}_{q-1}^{(m,a)} + \widehat{\beta}_{q-1}^{(m,a_t^{(m)})} - \widehat{\beta}_{q-1}^{(m,a_t^{(m)})} \rangle \leq 2 \max_{a\in[K]} |\langle x_t^{(m)}, \beta^{(m,a)} - \widehat{\beta}_{q-1}^{(m,a)} \rangle|. \quad \text{(I.4)}
\end{aligned}$$

By combining (I.3), (I.4) and further using Condition I.2 and Lemma B.2, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\max_{a\in[K]} \langle x_t^{(m)}, \beta^{(m,a)} - \beta^{(m,a_t^{(m)})} \rangle \mathbb{1}(m \in \mathcal{S}_t) \mid \{\widehat{\beta}_{q-1}^{(m,a)}\}_{a\in[K]}\right] \\
\leq &2 \max_{a\in[K]} \left(\|\beta^{(m,a)}\|_2 \wedge \|\beta^{(m,a)} - \widehat{\beta}_{q-1}^{(m,a)}\|_2\right) \sqrt{2\ln(2K)L}\mathbb{P}(m \in \mathcal{S}_t) \\
\leq &2p_m \max_{a\in[K]} \left(\|\beta^{(m,a)}\|_2 \wedge \|\beta^{(m,a)} - \widehat{\beta}_{q-1}^{(m,a)}\|_2\right) \sqrt{2\ln(2K)L}. \quad \text{(I.5)}
\end{aligned}$$

Taking expectations of (I.5) with respect to $\{\widehat{\beta}_{q-1}^{(m,a)}\}_{a\in[K]}$, we find

$$\mathbb{E}\left[\max_{a\in[K]} \langle x_t^{(m)}, \beta^{(m,a)} - \beta^{(m,a_t^{(m)})} \rangle \mathbb{1}(m \in \mathcal{S}_t)\right] \leq 2\sqrt{2\ln(2K)L}p_m \mathbb{E}[\max_{a\in[K]} 1 \wedge \|\widehat{\beta}_{q-1}^{(m,a)} - \beta^{(m,a)}\|_2]. \quad \text{(I.6)}$$

Given (I.6), it remains to bound the estimation error $\mathbb{E}[\max_{a\in[K]} \|\widehat{\beta}_{q-1}^{(m,a)} - \beta^{(m,a)}\|_2]$ for all $0 \leq q \leq Q$ and $m \in [M]$. Therefore, the rest follows the argument in the proof of Theorem 3 and uses Lemma I.2.

□

I.2 Lower Bound

We also establish the regret lower bound under Model-P as follows.

Theorem I.2. *Given any $1 \le s \le d$ and $\{p_m\}_{m\in[M]} \subseteq [0,1]$, for any $m \in [M]$, when $T \ge \max\{(d+1)/p_{[M]}, (s+1)/p_m\}/(16L) + 1$, there are $\{\beta^{(m,a)}\}_{a\in[K],m\in[M]}$ satisfying Condition I.1 and distributions of contexts satisfying Condition I.2 and I.3, such that for any online Algorithm A,*

$$\mathbb{E}[R_T^{(m)}(A)] = \Omega\left(\sqrt{\left(s + \frac{dp_m}{p_{[M]}}\right) T p_m}\right).$$

Proof. We consider the two-armed case where $K = 2$ and $\beta^{(m,2)} = -\beta^{(m,1)}$ for all $m \in [M]$. We prove $\Omega(\sqrt{sTp_m})$ and $\Omega(\sqrt{dTp_m/p_{[M]}})$ by considering two cases: the *homogeneous* case where $\beta^{(1,1)} = \cdots = \beta^{(M,1)} = \beta^{\star(1)}$ and the *s-sparse* case where $\beta^{\star(1)} = 0$ and $\|\beta^{(m,1)}\|_0 \le s$ for all $m \in [M]$.

The homogeneous case. Since $\beta^{(1,1)} = \cdots = \beta^{(M,1)} = \beta^{\star(1)}$ in this case, for simplicity, we omit the superscript m and use β_a to denote $\beta^{(m,a)}$ for any $k \in [K]$. Denote by Q the uniform distribution over $\{\beta \in \mathbb{R}^d : \|\beta\|_2 = \Delta\}$ where $\Delta \in [0,1]$ will be determined below. We let Q be the prior distribution of β_1, and let $D \triangleq \mathcal{N}(0, LI_d)$ be the distribution of contexts. Let $\mathcal{S}_t$ be the set of activated bandits at the t-th round, and denote by $P_{\beta_1,x,t}$ the distribution of the observed rewards $\{y_\ell^{(m)}\mathbb{1}(m \in \mathcal{S}_t) : m \in [M]\}_{\ell=1}^t$ up to time t, conditioned on β_1 and the contexts $\{x_\ell^{(m)} : a \in [2], m \in [M]\}_{\ell=1}^t$. Since the event $\{m \in \mathcal{S}_t\}$ is independent of the history and of the contexts $\{x_{t,a}^{(m)} : a \in [2]\}$ at the current round, we have

$$\sup_{\beta_1} \mathbb{E}[R_T^{(m)}(A)] \ge \mathbb{E}_{\beta_1\sim Q}[R_T^{(m)}(A)] = \sum_{t=1}^T \mathbb{E}_Q\left[\mathbb{E}_D\left[\mathbb{E}_{P_{\beta_1,x,t-1}}\left[p_m \max_{a\in[2]} \langle x_t^{(m)}, \beta_a - \beta_{a_t^{(m)}}\rangle\right]\right]\right] \tag{I.7}$$

where the factor of p_m in (H.20) is due to the integration over the randomness of $\{m \in \mathcal{S}_t\}$. Since $\beta_2 = -\beta_1$, we have

$$\max_{a\in[2]} \langle x_t^{(m)}, \beta_a - \beta_{a_t^{(m)}}\rangle = 2\mathbb{1}(a_t^{(m)} = 1)\langle x_t^{(m)}, \beta_1\rangle_- + 2\mathbb{1}(a_t^{(m)} = 2)\langle x_t^{(m)}, \beta_1\rangle_+$$

where the subscripts $+$ and $-$ denote the positive and negative parts, respectively. Therefore, from (I.7), we have that $\sup_\beta \mathbb{E}[R_T^{(m)}(A)]$ is lower bounded by

$$2p_m \sum_{t=1}^T \mathbb{E}_Q\left[\mathbb{E}_D\left[\mathbb{E}_{P_{\beta_1,x,t-1}}\left[\mathbb{1}(a_t^{(m)} = 1)\langle x_t^{(m)}, \beta_1\rangle_- + \mathbb{1}(a_t^{(m)} = 2)\langle x_t^{(m)}, \beta_1\rangle_+\right]\right]\right]. \tag{I.8}$$

For any $1 \le t \le T$, conditionally on $x_t^{(m)}$, we define two measures $Q_t^{(m)+}$ and $Q_t^{(m)-}$ via the Radon–Nikodym derivatives $\mathrm{d}Q_t^{(m)+}(\beta_1) = \langle x_t^{(m)}, \beta_1\rangle_+ / Z(x_t^{(m)})\mathrm{d}Q$ and $\mathrm{d}Q_t^{(m)-}(\beta_1) = \langle x_t^{(m)}, \beta_1\rangle_- / Z(x_t^{(m)})\ \mathrm{d}Q$, where $Z(x_t^{(m)}) = \mathbb{E}_Q[\langle x_t^{(m)}, \beta_1\rangle_+] = \mathbb{E}_Q[\langle x_t^{(m)}, \beta_1\rangle_-]$ is a normalization factor. Here $\mathbb{E}_Q[\langle d_t^{(m)}, \beta_1\rangle_+] = \mathbb{E}_Q[\langle d_t^{(m)}, \beta_1\rangle_-]$ due to the symmetry of Q. Plugging the

definitions of $Q_t^{(m)+}$ and $Q_t^{(m)-}$ into (I.8), and changing the integration order, we have

$$\begin{aligned}\sup_{\beta}\mathbb{E}[R_T^{(m)}(A)] \geq & 2p_m\sum_{t=1}^{T}\mathbb{E}_D\left[Z(x_t^{(m)})\left(\mathbb{E}_{P_{\beta_1,x,t-1}\circ Q_t^{(m)-}}[\mathbb{1}(a_t=1)]+\mathbb{E}_{P_{\beta_1,x,t-1}\circ Q_t^{(m)+}}[\mathbb{1}(a_t=2)]\right)\right]\\ \geq & 2p_m\sum_{t=1}^{T}\mathbb{E}_D\left[Z(x_t^{(m)})\left(1-\mathtt{TV}(P_{\beta_1,x,t-1}\circ Q_t^{(m)-},P_{\beta_1,x,t-1}\circ Q_t^{(m)+})\right)\right]\\ \geq & 2p_m\sum_{t=1}^{T}\mathbb{E}_D\left[Z(x_t^{(m)})\left(1-\sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_{\beta_1,x,t-1}\circ Q_t^{(m)+}\,\|\,P_{\beta_1,x,t-1}\circ Q_t^{(m)-})}\right)\right]\end{aligned} \tag{I.9}$$

where the second inequality follows the definition of the total variation distance: $P_1(A)+P_2(A^c)\geq 1-\mathtt{TV}(P_1,P_2)$ with $A=\{a_t=1\}$, and the third inequality follows from Pinsker's inequality $\mathtt{TV}(P_1,P_2)\leq\sqrt{D_{\mathrm{KL}}(P_1\,\|\,P_2)/2}$.

Due to the distribution of $x_t^{(m)}$, $x_t^{(m)}\neq 0$ with probability one; hence we can set $u_t^{(m)}=x_t^{(m)}/\|x_t^{(m)}\|_2$ and the zero probability event where $x_t^{(m)}=0$ does not affect the result. We also let $\tilde{\beta}_{1,t}^{(m)}=\beta_1-2\langle u_t^{(m)},\beta_1\rangle u_t^{(m)}$, and we have

$$\beta_1=\tilde{\beta}_{1,t}^{(m)}-2\langle u_t^{(m)},\tilde{\beta}_{1,t}^{(m)}\rangle u_t^{(m)}. \tag{I.10}$$

Since $\langle x_t^{(m)},\beta_1\rangle_-=\langle x_t^{(m)},\tilde{\beta}_{1,t}^{(m)}\rangle_+$ and Q is reflection-invariant, one can equivalently obtain $\beta_1\sim Q_t^{(m)-}$ by first generating $\tilde{\beta}_{1,t}^{(m)}\sim Q_t^{(m)+}$ and then calculating β_1 via (I.10). It thus holds that

$$\begin{aligned}P_{\beta_1,x,t-1}\circ(\beta_1\sim Q_t^{(m)-}) =& P_{\tilde{\beta}_{1,t}^{(m)}-2\langle u_t^{(m)},\tilde{\beta}_{1,t}^{(m)}\rangle u_t^{(m)},x,t-1}\circ(\tilde{\beta}_{1,t}^{(m)}\sim Q_t^{(m)+})\\ \overset{d}{=}& P_{\beta_1-2\langle u_t^{(m)},\beta_1\rangle u_t^{(m)},x,t-1}\circ(\beta_1\sim Q_t^{(m)+}).\end{aligned}$$

Following the above argument and (I.9), we have

$$2p_m\sum_{t=1}^{T}\mathbb{E}_D\left[Z(x_t^{(m)})\left(1-\sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_{\beta_1,x,t-1}\circ Q_t^{(m)+}\,\|\,P_{\beta_1-2\langle u_t,\beta_1\rangle u_t,x,t-1}\circ Q_t^{(m)+}}\right)\right].$$

By Lemma B.7, $\sup_\beta\mathbb{E}[R_T^{(m)}(A)]$ is further lower bounded by

$$\begin{aligned}&\sup_{\beta}\mathbb{E}[R_T(A)]\\ \geq & p_m\sum_{t=1}^{T}\mathbb{E}_D\left[Z(x_t^{(m)})\left(1-\sqrt{\frac{1}{2}\mathbb{E}_{\beta_1\sim Q_t^{(m)+}}\left[D_{\mathrm{KL}}(P_{\beta_1,x,t-1}\,\|\,P_{\beta_1-2\langle u_t^{(m)},\beta_1\rangle u_t^{(m)},x,t-1})\right]}\right)\right].\end{aligned} \tag{I.11}$$

Since the reward noise follows the distribution $\mathcal{N}(0,1)$, conditioned on the activation sets $\{\mathcal{S}_\ell\}_{\ell=1}^{t-1}$, we have $P_{\beta_1,x,t-1}=\otimes_{\ell=1}^{t-1}\otimes_{r\in\mathcal{S}_\ell}\mathcal{N}(\langle\beta_{a_\ell^{(r)}},x_\ell^{(r)}\rangle,1)$. Furthermore, by the formula for

the divergence between two Gaussian distributions, we have

$$\begin{aligned}
&D_{\mathrm{KL}}\left(P_{\beta_1,x,t-1}\,\|\,P_{\beta_1-2\langle u_t^{(m)},\beta\rangle u_t^{(m)},x,t-1}\mid\{\mathcal{S}_\ell\}_{\ell=1}^{t-1}\right)\\
=&\frac{1}{2}\sum_{\ell=1}^{t-1}\sum_{r\in\mathcal{S}_\ell}\left(\langle\beta_{a_\ell^{(r)}},x_{\ell,}^r\rangle-\langle\beta_{a_\ell^{(r)}}-2\langle u_t^{(m)},\beta_{a_\ell^{(r)}}\rangle u_t^{(m)},x_\ell^{(r)}\rangle\right)^2\\
=&2\langle u_t^{(m)},\beta_1\rangle^2\sum_{\ell=1}^{t-1}\sum_{r\in\mathcal{S}_\ell}\langle u_t^{(m)},x_\ell^{(r)}\rangle^2=2\langle u_t^{(m)},\beta_1\rangle^2\sum_{\ell=1}^{t-1}\sum_{r\in[M]}\langle u_t^{(m)},x_\ell^{(r)}\rangle^2\mathbb{1}(r\in\mathcal{S}_\ell).
\end{aligned}\tag{I.12}$$

Since the events $\{r\in\mathcal{S}_\ell\}$ for $\ell\in[t-1]$ and $r\in[M]$ are independent of the contexts and of β_1, using (I.12) and Lemma B.7, we obtain

$$\begin{aligned}
&D_{\mathrm{KL}}(P_{\beta_1,x,t-1}\,\|\,P_{\beta_1-2\langle u_t^{(m)},\beta_1\rangle u_t^{(m)},x,t-1})\\
\le&\mathbb{E}_{\{\mathcal{S}_\ell\}_{\ell=1}^{t-1}}\left[D_{\mathrm{KL}}\left(P_{\beta_1,x,t-1}\,\|\,P_{\beta_1-2\langle u_t^{(m)},\beta_1\rangle u_t^{(m)},x,t-1}\mid\{\mathcal{S}_\ell\}_{\ell=1}^{t-1}\right)\right]\\
=&2\langle u_t^{(m)},\beta_1\rangle^2\sum_{\ell=1}^{t-1}\sum_{r\in[M]}\langle u_t^{(m)},x_\ell^{(r)}\rangle^2\mathbb{P}(r\in\mathcal{S}_\ell)=2\langle u_t^{(m)},\beta_1\rangle^2p_{[M]}\sum_{\ell=1}^{t-1}\langle u_t^{(m)},x_\ell^{(r)}\rangle^2.
\end{aligned}\tag{I.13}$$

Therefore, combining (I.11) and (I.13), we have

$$\begin{aligned}
&\sup_\beta\mathbb{E}[R_T^{(m)}(A)]\\
\ge&2p_m\sum_{t=1}^T\mathbb{E}_D\left[Z(x_t^{(m)})\left(1-\sqrt{\mathbb{E}_{Q_t^{(m)+}}[\langle u_t^{(m)},\beta_1\rangle^2]u_t^{m\top}\left(p_{[M]}\sum_{\ell=1}^{t-1}x_\ell^{(r)}x_\ell^{r\top}\right)u_t^m}\right)\right].
\end{aligned}\tag{I.14}$$

Taking expectations in (I.14) with respect to $\{x_\ell^{(r)}:r\in[M]\}_{\ell=1}^{t-1}$, $\sup_\beta\mathbb{E}[R_T^{(m)}(A)]$ is lower bounded by

$$2p_m\sum_{t=1}^T\mathbb{E}_{x_t^{(m)}}\left[Z(x_t^{(m)})\left(1-\sqrt{2(t-1)Lp_{[M]}\mathbb{E}_{Q_t^{(m)+}}[\langle u_t^{(m)},\beta_1\rangle^2]}\right)\right],\tag{I.15}$$

where the outer expectation is only over the randomness of $x_t^{(m)}$. We next calculate $\mathbb{E}_{Q_t^+}[\langle u_t,\beta_1\rangle^2]$ and $Z(x_t^{(m)})$. By (H.32), we have

$$\mathbb{E}_{Q_t^{(m)+}}[\langle u_t^{(m)},\beta_1\rangle^2]=\frac{2\Delta^2}{d+1}.\tag{I.16}$$

We also have

$$\begin{aligned}
&\mathbb{E}_{x_t^{(m)}}[Z(x_t^{(m)})]=\mathbb{E}_{x_t^{(m)}}[\mathbb{E}_Q[\langle x_t^{(m)},\beta_1\rangle_+]]=\frac{1}{2}\mathbb{E}_{x_t^{(m)}}[\mathbb{E}_Q[|\langle u_t^{(m)},\beta\rangle|]]\\
=&\frac{1}{2}\mathbb{E}_{x_t^{(m)}}\left[\|x_t^{(m)}\|_2\right]\frac{\Delta\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\sqrt{\pi}}=\Omega(\sqrt{L}\Delta).
\end{aligned}\tag{I.17}$$

Combining (I.15), (I.16), and (I.17), we have

$$\sup_{\beta} \mathbb{E}[R_T^{(m)}(A)] \geq \Omega(p_m \sqrt{L}\Delta) \sum_{t=1}^{T} \left(1 - \sqrt{\frac{4(t-1)L\Delta^2}{(d+1)} p_{[M]}}\right). \tag{I.18}$$

Plugging $\Delta = \frac{\sqrt{(d+1)}}{4\sqrt{(T-1)L\sum_{r\in[M]} p_r}} \leq 1$ into (I.18), we finally establish

$$\sup_{\beta} \mathbb{E}[R_T(A)] = \Omega\left(p_m T\sqrt{L}\Delta\right) = \Omega\left(\sqrt{dTp_m^2/p_{[M]}}\right).$$

The s-sparse case. In this case, we consider $\mathrm{supp}(\beta^{(1,1)}), \ldots, \mathrm{supp}(\beta^{(M,1)})$ located in the first s coordinates. If the supports of $\{\beta_1^{(m)}\}_{m\in[M]}$ are known by the analyst, then the structure of sparse heterogeneity and the common $\beta^{\star(1)}$ is non-informative for estimating $\{\beta^{(m,1)}\}_{m\in[M]}$. Therefore, in this case, we can obtain the lower bound $\Omega(\sqrt{sTp_m})$ by simply adapting the proof for the homogeneous case with $M = 1$ in s dimensions.

□

J Experimental Details

J.1 Synthetic Experimental Details

To ensure a fair comparison of methods, we set the regularization parameter λ_m of LASSO/ LASSOB, RM/RMB, and TNB, and the threshold parameter γ_m of MOLAR/MOLARB guided by theoretical results to obtain optimal rates with respect to n, d, and M (Xu and Bastani, 2021; Bastani and Bayati, 2020; Cella et al., 2022). Our simulation setup follows conventions from the most closely related statistical literature (Chen et al., 2021; Chan, 2017; Camerlenghi et al., 2019; Xu and Bastani, 2021; Bastani and Bayati, 2020).

Specifically, we set $\lambda_m = c_\lambda \sqrt{\ln(d)/n_m}$ for LASSO/LASSOB and RM/RMB, $\gamma_m = c_\gamma \ln((n_{[M]}/n_m) \wedge d)/n_m)$ for MOLAR/MOLARB, $\lambda_m = c_\lambda \sqrt{(M+d)/n_m}$ for TNB. We only tune the numerical coefficient c_λ and c_γ on a pre-specified grid $\{0.05\sigma, 0.35\sigma, 0.7\sigma, \sigma, 2\sigma\}$, where σ is the standard deviation of noise. We tune these numerical coefficients to lead to the best ℓ_1 estimation errors on independently generated data with $n = 5,000$.

Note that $\sigma = 0.1$ in our offline linear regression setup. After tuning, we take $c_\lambda = 0.005$ for LASSO; we take $c_\lambda = 0.035$ for RM and set the trimming-related parameters to the default values $\zeta = 0.1$, $\eta = 0.1$ suggested by Xu and Bastani (2021); we take $c_\gamma = 0.1$ for MOLAR with the option of hard thresholding, respectively.

In contextual bandits, the noise scale is set as $\sigma = 0.5$. We initialize the first batch size $|H_0| = 1$ to use data efficiently. In the bandit case, the parameters associated with the reported results are $c_\lambda = 0.025$ for LASSOB, $c_\lambda = 1$ for TNB, $(\zeta, \eta, c_\lambda) = (0.1, 0.1, 0.175)$ for RMB, and $c_\gamma = 0.5$ for MOLARB with the option of hard thresholding, respectively.

J.2 PISA Experimental Details

The PISA2012 dataset (OECD, 2019) consists of 485,490 student records collected from 68 countries[1]. However, records associated with many of these countries have more than half the data missing and contain constant features. Thus, we restrict our experiment to the $M = 15$ countries with the largest sample sizes. The sample sizes in these countries range from 7,038 to 33,806, while the data contains about 500 features.

Since many features are highly correlated or are constant across records, to avoid ill-conditioning, we pre-process the data as follows. We create dummy variables to indicate missing values and then fill missing values with zeroes. We apply the LASSO globally to select features, with the regularization hyperparameter selected automatically via 10-fold cross-validation. We then filter out features with pairwise correlations higher than 0.6 among the selected ones, doing this sequentially in the order given by the PISA dictionary. This finally leaves us with 57 features.

[1]It is accessible at `https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm` (official website) and `https://s3.amazonaws.com/udacity-hosted-downloads/ud507/pisa2012.csv.zip` (an exterior csv format).

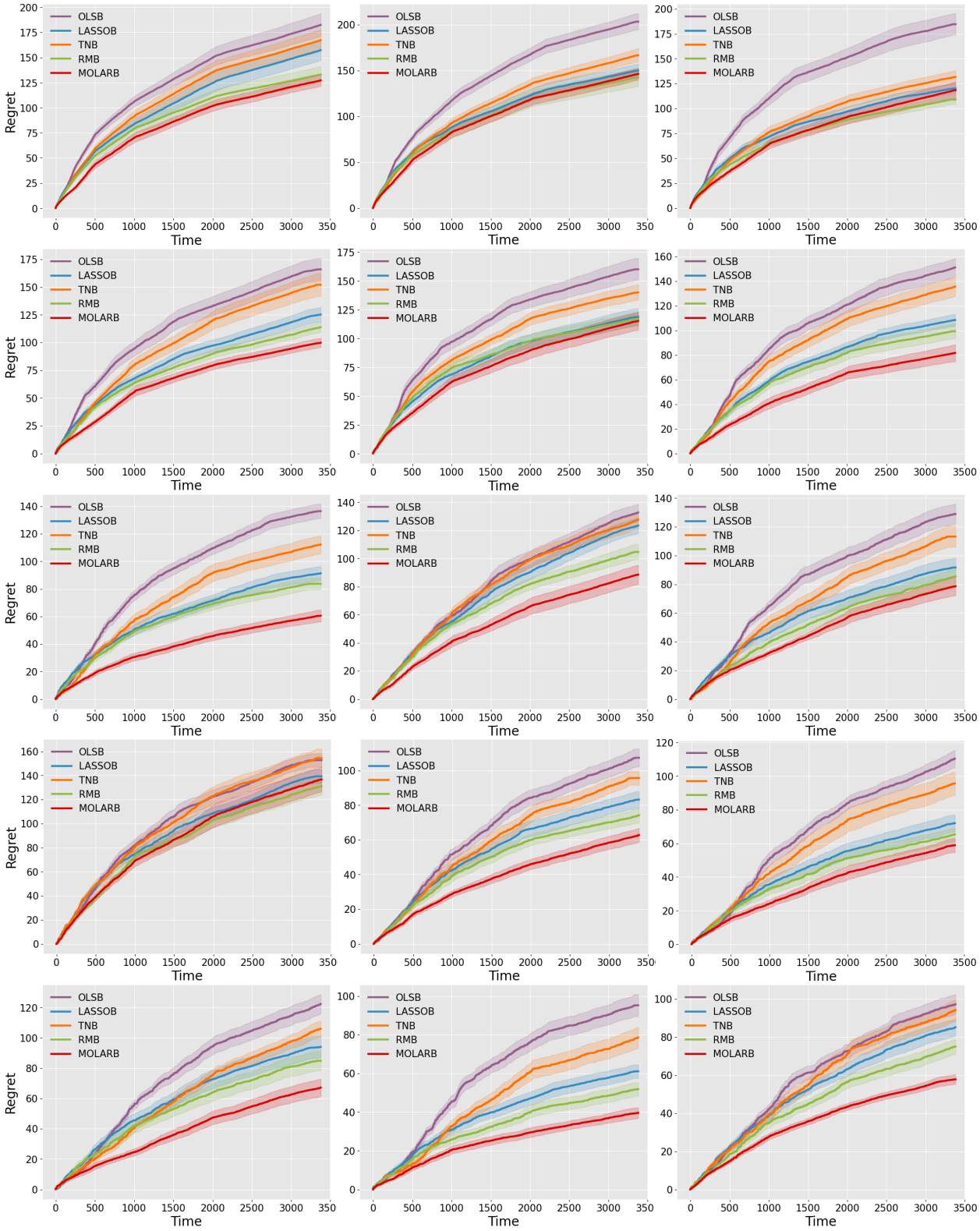

Figure 6: Regret $R_T^{(m)}$ of the top largest 15 countires of the PISA dataset. The shaded regions depict the corresponding 95% normal confidence intervals based on the standard errors from twenty independent trials.

To visibly judge the heterogeneity of the processed datasets, we compute the OLS estimates $\{\widehat{\beta}^{(m)}\}_{m=1}^{M}$ for all countries individually, along with a shared estimate $\widehat{\beta}^{\star}$ obtained by taking the covariate-wise median over the OLS estimates. We then plot the differences $|\delta^{(m)}| = |\widehat{\beta}^{(m)} - \widehat{\beta}^{\star}|$ with values below $6\sqrt{[(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}}$ set as zero. Note that $\sqrt{[(\mathbf{X}^{(m)\top}\mathbf{X}^{(m)})^{-1}]_{k,k}}$ corresponds to the magnitude of the variation of OLS estimates under standard Gaussian noises. The final result is shown in Figure 1. Note that Figure 1 is plotted before splitting the data for the experiments. In Figure 1, we see that the differences in coefficients seem to be consistent with being sparse.

We take $K = 2$ since our arms are to predict the student with a better mathematics score. To enable evaluation, we randomly split the data into a large test set, a small training set, and a small validation set with proportions (90%, 5%, 5%) and (80%, 15%, 5%) for offline and online experiments, respectively. We use the individual OLS estimates of the test set as proxies for true parameters and evaluate methods on the training set with parameters tuned based on the validation set.

Once again, to ensure a fair comparison of methods, we set the hyperparameters λ and γ in the same manner as in the synthetic results, as suggested by theoretical results to achieve optimal rates of convergence. We only tune the numerical coefficients over a pre-specified grid $\{0.05, 0.35, 0.7, 1, 2\}$, and report the optimal results. We run the bandit methods on this processed data in the same manner as in the simulations. Here we run MOLAR and MOLARB with the option of soft thresholding.

We repeat experiments starting from data splitting for 100 and 20 times for offline and online experiments, respectively. The regrets for all 15 countries (Mexico, Italy, Spain, Canada, Brazil, Australia, UK, UAE, Switzerland, Qatar, Colombia, Finland, Belgium, Denmark, and Jordan from top to bottom, left to right) are in Figure 6.

J.3 Ablation Studies

J.3.1 Robustness Examinations for MOLAR

We first examine the robustness of MOLAR to the choice of c_γ. To this end, we repeat the experiments in Figure 2 and simulate MOLAR with empirically estimated noise variances through

$$\hat{\sigma}_m := \sqrt{\|\mathbf{X}^{(m)}\widehat{\beta}^{(m)}_{\text{ind}} - Y^{(m)}\|_2^2/(n_m - d)}$$

and varying c_γ. The results are depicted in Figure 7. By comparing MOLAR with the individual OLS estimates, we observe that MOLAR exhibits advantages for all values of the threshold. In particular, MOLAR is robust for slightly large c_γ).

J.3.2 Correlated Covariates & Disparate Sample Sizes

To supplement the experiments in Figures 2 and 7 where $\Sigma^{(m)} = I_d$ and $n_m = n$ for all $m \in [M]$, we also conduct similar experiments for correlated covariates with disparate tasks-wise covariances and sample sizes. Here, for each task $m \in [M]$, we select the covariance matrix as $\Sigma^{(m)} = Q^{(m)}\text{diag}((1 + 4(k-1)/d)_{k\in[d]})Q^{(m)\top}$ where $Q^{(m)} \in \mathbb{R}^{d\times d}$ is a randomly generated orthonormal matrix. Since $\Sigma^{(m)} \neq I_d$, the covariates are correlated.

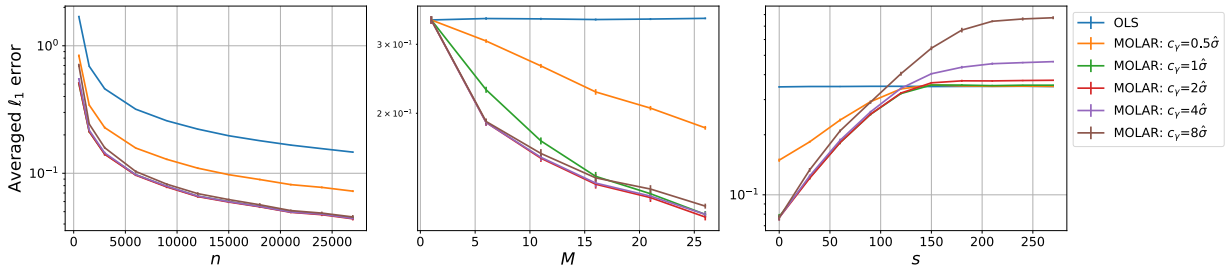

Figure 7: Average ℓ_1 estimation error for MOLAR with varying thresholding parameters. (Left): Fixing $s = 20$, $M = 30$ and varying n. (Middle): Fixing $s = 20$, $n = 5,000$ and varying M. (Right): Fixing $M = 30$, $n = 5,000$ and varying s. The standard error bars are obtained from ten independent trials.

For each pre-specified n, we determine the task-wise sample sizes $\{n_m\}_{m=1}^M$ by first drawing a Dirchlet random vector $(z_1, \ldots, z_d)$ with $0 \leq z_k \leq 1$ and $\sum_{k=1}^d z_k = 1$, and then round Mnz_k to obtain the sample size n_m. By doing this splitting, we roughly maintain $n_{[M]} = \sum_{m=1}^M n_m \approx Mn$ but introduce significant disparity among $\{n_m\}_{m=1}^M$. We thus apply the weighted median to obtain the global estimate $\widehat{\beta}^\star$ with $w_m = n_m$ for all $m \in [M]$ and set other hyperparameters the same as in Section J.1. The results are shown in Figure 8. Again, we observe a significant advantage of MOLAR over other baseline approaches.

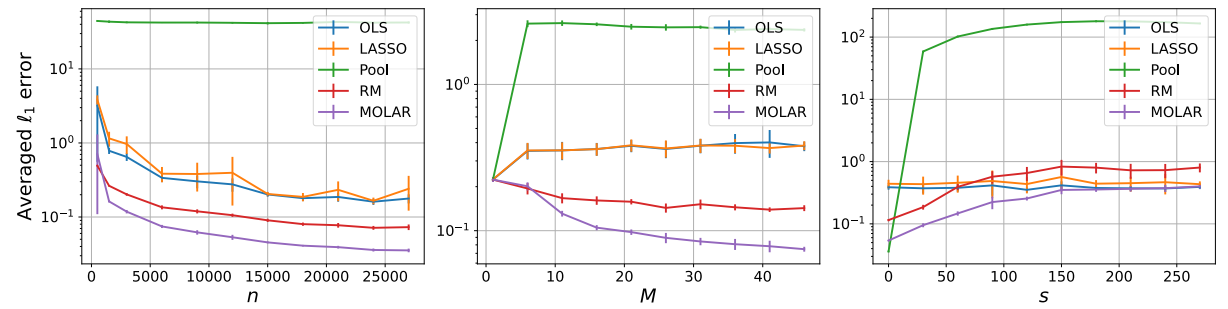

Figure 8: Average ℓ_1 estimation error for multitask linear regression under correlated covariates with disparate task-wise covariances $\{\Sigma^{(m)}\}_{m=1}^M$ and sample sizes $\{n_m\}_{m=1}^M$. (Left): Fixing $s = 20$, $M = 30$ and varying n. (Middle): Fixing $s = 20$, $n = 5,000$ and varying M. (Right): Fixing $M = 30$, $n = 5,000$ and varying s. The standard error bars are obtained from ten independent trials.

J.3.3 Robustness Checks for MOLARB

We also provide a robustness check for MOLARB by varying c_γ and $|H_0|$. We investigate the cumulative expected regret of MOLARB while varying the first batch size $|H_0| \in \{1, 5, 10\}$, and the numerical coefficient $c_\gamma \in \{0.175, 0.35, 0.5, 1\}$. The results, presented in Figure 9, are computed in the same setup as the synthetic bandit simulations. We find that the cumulative regret performance of MOLARB is not substantially impacted by changing the parameters

by up to an order of magnitude. This suggests that MOLARB is quite robust to $|H_0|$ and c_γ in this range.

J.3.4 Usage of Historical Batches

Since MOLARB empties all batch-wise buffers of contexts after using them to update estimates $\{\widehat{\beta}^{(m)}\}_{m=1}^{M}$, we also compare MOLARB with a variant where all historical contexts in each arm are maintained. In this variant, we still use each brand-new batch $\{(\mathbf{X}_q^{(m)}, Y_q^{(m)})\}_{m=1}^{M}$ to collaboratively learn the global estimate $\widehat{\beta}_q^{\star}$, yet the step of covariate-wise shrinkage in obtaining the task-wise estimates $\widehat{\beta}^{(m)}$ leverages all previous batches $(\mathbf{X}_{[q]}^{(m)}, Y_{[q]}^{(m)})$ in the m-th bandit. The results of this variant are marked as "use_hist" in Figure 9. We find that MOLARB does not lose significant sample efficiency, compared to this variant. This finding is consistent with our theoretical results that MOLARB is minimax optimal in this multi-task setup.

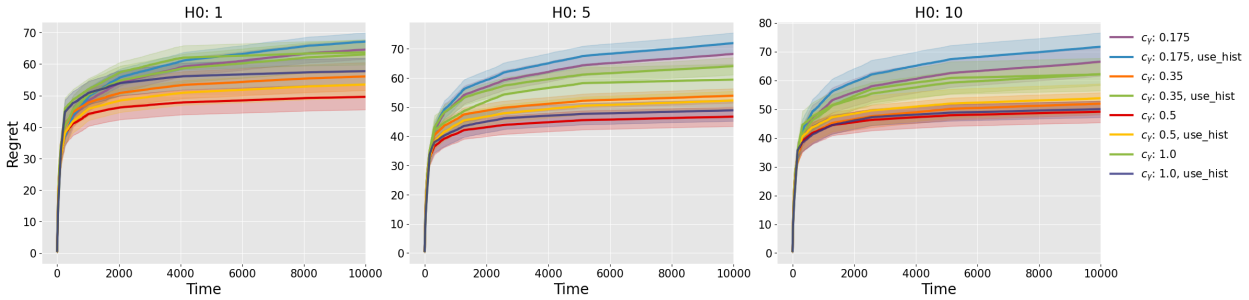

Figure 9: Regret $R_T^{(m)}$ accumulated by MOLARB of an instance with activation probability 0.91 with varying $|H_0|$ and tuning coefficient c_γ, where shaded regions depict the corresponding 95% normal confidence intervals based on standard errors calculated over twenty independent trials.

References

H. Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

H. Bastani and M. Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

A. C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer. Byzantine-tolerant machine learning. *ArXiv*, abs/1703.02757, 2017.

T. T. Cai and Z. Guo. Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

F. Camerlenghi, B. Dumitrascu, F. Ferrari, B. E. Engelhardt, and S. Favaro. Nonparametric bayesian multiarmed bandits for single-cell experiment design. *arXiv: Applications*, 2019.

R. Caruana. *Multitask learning*. Springer, 1998.

L. Cella, K. Lounici, and M. Pontil. Multi-task representation learning with stochastic linear bandits. *arXiv preprint arXiv:2202.10066*, 2022.

H. P. Chan. The multi-armed bandit problem: An efficient non-parametric solution. *arXiv: Statistics Theory*, 2017.

H. Chen, W. Lu, and R. Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.

E. Dobriban and Y. Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49:918–943, 2021.

S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Y. Duan and K. Wang. Adaptive and robust multi-task learning. *arXiv preprint arXiv:2202.05250*, 2022.

J. C. Duchi. *Information Theory and Statistics*, volume 87. Stanford University, 2019.

J. C. Duchi and M. J. Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

T. Evgeniou and M. Pontil. Regularized multi–task learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004.

T. Evgeniou, C. A. Micchelli, M. Pontil, and J. Shawe-Taylor. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

F. Götze, H. Sambale, and A. Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26, 2021.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

Y. Han, Z. Zhou, Z. Zhou, J. Blanchet, P. Glynn, and Y. Ye. Sequential batch learning in finite-action linear contextual bandits. *ArXiv*, 2020.

S. Hanneke and S. Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

F. Hanzely, S. Hanzely, S. Horváth, and P. Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33:2304–2315, 2020.

X. Huang, D. Lee, E. Dobriban, and H. Hassani. Collaborative learning of discrete distributions under heterogeneity and communication constraints. In *Advances in Neural Information Processing Systems*, 2022.

P. J. Huber. Robust statistics. *Wiley Series in Probability and Mathematical Statistics*, 1981.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

V. Koltchinskii and D. Panchenko. A rosenthal-type inequality for subgaussian processes and applications to percolation and concentration of the measure. *The Annals of Probability*, 30(1):245–281, 2002.

M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

S. Li, L. Zhang, T. T. Cai, and H. Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, pages 1–12, 2023.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

P. McCullagh and J. A. Nelder. *Generalized linear models*, volume 39. CRC Press, 1989.

S. Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21:2308–2335, 2013.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27:538–557, 2012.

OECD. *Teaching for the future: Effective classroom practices to transform education*. OECD Publishing, 2019.

R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *International Conference on Machine Learning*, pages 713–720, 2006.

Z. Ren and Z. Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science*, 2023.

P. J. Rousseeuw. Tutorial to robust statistics. *Journal of chemometrics*, 5(1):1–20, 1991.

H. Sambale. Some notes on concentration for α-subexponential random variables. *arXiv preprint arXiv:2002.10761*, 2020.

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

C. Singh and A. Sharma. Online learning using multiple times weight updating. *Applied Artificial Intelligence*, 34(6):515–536, 2020.

F. Smithies. Convex functions and orlicz spaces. *The Mathematical Gazette*, 47:266 – 267, 1962.

L. Su and N. H. Vaidya. Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM symposium on principles of distributed computing*, pages 425–434, 2016.

Y. Tian and Y. Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 0:1–14, 2022.

N. Tripuraneni, C. Jin, and M. Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

J. V. Uspensky. *Introduction to mathematical probability*. McGraw-Hill Book Company, 1937.

S. Van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptoticaly optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3): 1166–1202, 2014.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018.

L. Xia, B. Nan, and Y. Li. Debiased lasso for generalized linear models with a diverging number of covariates. *Biometrics*, 79(1):344–357, 2023.

K. Xu and H. Bastani. Learning across bandits in high dimension via robust statistics. *arXiv preprint arXiv:2112.14233*, 2021.

D. Yin, Y. Chen, R. Kannan, and P. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, 2018.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):217–242, 2014.