Learning Across Bandits in High Dimension via Robust Statistics

Kan Xu
University of Pennsylvania, Economics, kanxu@sas.upenn.edu

Hamsa Bastani
Wharton School, Operations Information and Decisions, hamsab@wharton.upenn.edu

Decision-makers often face the "many bandits" problem, where one must simultaneously learn across related but heterogeneous contextual bandit instances. For instance, a large retailer may wish to dynamically learn product demand across many stores to solve pricing or inventory problems, making it desirable to learn jointly for stores serving similar customers; alternatively, a hospital network may wish to dynamically learn patient risk across many providers to allocate personalized interventions, making it desirable to learn jointly for hospitals serving similar patient populations. We study the setting where the unknown parameter in each bandit instance can be decomposed into a global parameter plus a sparse instance-specific term. Then, we propose a novel two-stage estimator that exploits this structure in a sample-efficient way by using a combination of robust statistics (to learn across similar instances) and LASSO regression (to debias the results). We embed this estimator within a bandit algorithm, and prove that it improves asymptotic regret bounds in the context dimension d; this improvement is exponential for data-poor instances. We further demonstrate how our results depend on the underlying network structure of bandit instances. Finally, we illustrate the value of our approach on synthetic and real datasets.

Key words: multitask learning, transfer learning, contextual bandits, robust statistics, LASSO, networks

1. Introduction

Contextual bandits are a popular framework for adaptive decision-making and have found numerous applications including personalized content recommendations (Li et al. 2010), mobile health (Tewari and Murphy 2017), targeted COVID-19 screening (Bastani et al. 2021b), dynamic pricing (Qiang and Bayati 2016) and inventory management (Yuan et al. 2021).

While the bandit literature typically considers a single decision-maker solving an isolated bandit instance, decision-makers increasingly face many, simultaneous bandit instances for closely related learning tasks. In these cases, we have an opportunity to not only learn *within* each bandit instance, but also *across* similar instances. To illustrate, consider the following two examples from healthcare and revenue management respectively:

Example 1 (Medical Risk Scoring). Health providers seek to predict patient-specific risk for adverse events (e.g., diabetes) in order to target preventative interventions. Learning this risk score primarily from patient data collected at the target hospital (where decisions are made) is important to account for idiosyncrasies that are specific to the hospital and the patient population it serves. This can include systematic differences in diagnosis/treatment behavior, healthcare utilization, or medical coding (see, e.g., Quiñonero-Candela et al. 2008, Subbaswamy and Saria 2020, Bastani 2021, Mullainathan and Obermeyer 2017). As a result, each hospital faces a distinct bandit learning problem. Yet, we may expect hospitals that serve similar patient populations to have similar underlying predictive models, creating an important opportunity to transfer knowledge across these bandit instances.

Example 2 (Demand Prediction). Large retailers seek to predict store-specific demand for their various products to inform dynamic pricing or inventory management decisions. Learning this demand model primarily from sales data collected at the target store (where decisions are made) is important to account for idiosyncrasies that are specific to the store and the customer population it serves. This can include systematic differences in customer trends/preferences, in-store product placement, or promotion decisions (see, e.g., Baardman et al. 2020, Cohen and Perakis 2018, van Herpen et al. 2012). As a result, each store faces a distinct bandit learning problem. Yet, we may expect stores that serve similar customer populations to have similar underlying demand models, creating an important opportunity to transfer knowledge across these bandit instances.

There are numerous other examples where we wish to learn heterogeneous treatment effects across many simultaneous experiments, ranging from customer promotion targeting, A/B testing on platforms, and identifying promising combination therapies in clinical trials. It is worth noting that bandits are largely used in problems where there is relatively little historical data available, e.g., due to the novelty or nonstationarity of the learning problem, or the limited population size relative to the feature dimension. In such settings, transfer learning from related data sources can be especially valuable to improve performance (Caruana 1997, Pan et al. 2010).

Existing work has proposed general *multitask learning* approaches for such problems, which transfer knowledge across problem instances to improve learning. Unfortunately, existing bandit algorithms targeting the multitask learning setting do not provide better regret bounds — i.e., they do not significantly improve performance compared to treating each bandit instance as its own independent problem. Indeed, in general, transfer or multitask learning cannot improve predictive accuracy without assuming some form of shared structure connecting the different problem instances — intuitively, if the problem instances are unrelated, then learning in one instance cannot significantly improve learning in others (Hanneke and Kpotufe 2020). Our work bridges this gap by imposing a natural structure, motivated by real datasets; by designing an estimator that exploits this structure, we obtain improved regret bounds in the context dimension d.

In particular, each bandit instance j is typically parameterized with a predictive parameter vector β^j — e.g., the parameters of a linear regression model predicting the reward of each arm as a function of the current context. The shared structure we consider is that the β^j have sparse differences relative to one another. In particular, we assume that they have the form

$$\beta^j = \beta^\dagger + \delta^j,$$

for some $\beta^\dagger$ representing the portion of the parameter vector that is "shared" across locally similar problem instances; then, δ^j is a problem-specific vector that represents the idiosyncratic biases specific to problem instance j. Then, we impose that the problem-specific bias δ^j is "small", capturing the notion that the problem instances are largely similar; such an assumption implies that the difference between two problem instances is (statistically) easier to learn than either problem instance by itself (Bastani 2021, Xu et al. 2021). More precisely, we assume that δ^j is *sparse* — i.e., only a few of its components are nonzero. This is often the case when some unknown underlying mechanism systematically affects a subset of the features, e.g., some hospitals under-diagnose certain conditions in claims data compared to others (Bastani 2021).

Even in the static, supervised learning setting, existing multitask learning algorithms (e.g., pooling data or regularizing estimates across problem instances) are not designed to leverage this structure (see §1.1 for an overview of existing methods). Thus, we first propose a novel two-stage robust estimator that exploits this structure in the supervised learning setting; we call the resulting multitask estimator RMEstimator. In the first stage, it leverages the trimmed mean from robust statistics (Rousseeuw 1991, Lugosi and Mendelson 2021) to estimate a "shared" model $\widehat{\beta}^\dagger$ across data collected from similar learning problems.[1] Then, in the second stage, it uses LASSO regression (Chen et al. 1995, Tibshirani 1996) to efficiently learn the problem-specific bias δ^j, which can be combined with our estimate of $\beta^\dagger$ to obtain the problem-specific parameter β^j. We prove finite-sample generalization bounds that show favorable performance compared to existing approaches, especially in terms of the context dimension d.

This estimator (and the tighter confidence bounds it affords) can then be embedded into simultaneous linear contextual bandit algorithms running at each problem instance; we call the resulting multitask bandit algorithm RMBandit. It efficiently manages the bias-variance tradeoff from incorporating auxiliary data from similar bandit instances (multitask learning) in conjunction with the classical exploration-exploitation tradeoff (bandit learning). We derive upper bounds for the cumulative regret of the RMBandit, both for individual problem instances and across all instances. Our multitask learning approach improves regret (compared to running separate bandit instances) in

[1] Note that we do not attempt to estimate the original shared parameter $\beta^\dagger$, since it is not identifiable; rather, as discussed in §3, it suffices to estimate some $\widetilde{\beta}^\dagger$ that lies in an ℓ_0 ball of radius $\mathcal{O}(s)$ around $\beta^\dagger$.

terms of the context dimension d; importantly, this regret improvement is *exponential* for data-poor bandit instances, since they benefit most from transferring knowledge from similar instances.

We also analyze the impact of the underlying network structure on the cumulative regret. Specifically, we assume knowledge of a network that captures the similarity between any pair of bandit instances; this can be inferred based on observed covariates (e.g., geographic distance between hospitals/stores or socio-economic indices of neighborhoods served) or data from past decision-making problems (see, e.g., Crammer et al. 2008). Then, for any given problem instance, we can optimize the "similarity radius" of learning problems from which to transfer knowledge, resulting in regret bounds that scale with the underlying network density.

Finally, we empirically evaluate our approach on both synthetic and real datasets in healthcare and pricing. Indeed, we find that the RMBandit algorithm can substantially speed up learning and improve overall performance compared to existing bandit and multitask learning algorithms.

1.1. Related Literature

Our work relates to the literature on multitask learning and contextual bandits; we contribute on both fronts. Our approach builds on the literature on robust and high-dimensional statistics.

There has been significant interest from the machine learning community on developing methods that combine data from multiple learning problems (typically referred to as tasks). These can be broadly classified into three categories: (i) multitask learning (Caruana 1997), where one aims to learn jointly across a fixed set of similar tasks, (ii) transfer learning (Pan et al. 2010), a special case of multitask learning, where the goal is to maximize performance on a distinguished "target" task, and (iii) meta-learning (Finn et al. 2017), where one aims to learn from historical tasks to improve learning in similar future tasks. Our problem is an instance of multitask learning, since our goal is to learn across a fixed set of bandit instances with related unknown parameters.

Multitask Learning. Naturally, if the tasks are sufficiently different, then learning in one task cannot substantially improve learning in other tasks (Hanneke and Kpotufe 2020). Thus, a common approach in machine learning is to assume that the underlying parameters across tasks are close in ℓ_2 norm. Joint learning can then be operationalized by regularizing the estimated parameters together, e.g., through ridge (Evgeniou and Pontil 2004) or kernel ridge (Evgeniou et al. 2005) regularization. Alternatively, one can employ a shared Bayesian prior across tasks (Raina et al. 2006, Gupta and Kallus 2021) or simply pool data from nearby tasks (Ben-David et al. 2010, Crammer et al. 2008). However, these approaches do not improve performance bounds beyond constants; in general, one must impose (and exploit) additional structure to obtain nontrivial theoretical improvements. Bastani (2021) uses real datasets to motivate the assumption that the parameters across tasks are close in ℓ_0 norm. This structure motivates a two-step estimator of transfer learning using LASSO regression, yielding improved bounds in the feature dimension d for supervised learning (Bastani 2021, Li et al. 2020a, Tian and Feng 2021) and unsupervised learning (Xu et al. 2021). One can further impose that the underlying parameters for each task are sparse, sharing the same support (Lounici et al. 2009) or similar covariance matrices (Li et al. 2020a, Tian and Feng 2021) across tasks; we do not make these assumptions since the applications we consider often have dense underlying parameters (see, e.g., Bastani 2021) and the covariance matrices vary widely across tasks due to covariate shifts (e.g., due to different customer populations at hospitals or stores, see Subbaswamy and Saria 2020).

We build on the last stream of two-step estimators for the multitask learning problem. However, we need a fundamentally different algorithmic approach; as we discuss in §3, the challenge is that the sparse bias terms can be poorly aligned across tasks, and thus classical estimates of the shared model (e.g., via data pooling or model averaging) destroy task-specific sparse structure and therefore cannot be debiased using LASSO (as was the case in prior work). Instead, we take the view that each component where the bias terms align poorly suffers "corruptions" to the shared model; we use a counting argument to show that either the number of corruptions must be small, or the component is one of a small number of well-aligned components. We use robust statistics to overcome corruptions for poorly-aligned components and LASSO to debias well-aligned components. To the best of our knowledge, our work proposes the first such combination of robust statistics and high-dimensional regression, yielding improved bounds for multitask learning.

The first step of our approach (using robust statistics) relates to recent robust machine learning methods that can handle adversarial corruptions to a small fraction of the data (Yin et al. 2018, Konstantinov and Lampert 2019). These approaches do not apply to our setting — as a consequence of our sparse differences assumption, we show that only a few similar *tasks* (as opposed to observations or features) have unknown parameters that are "corrupted" in most dimensions. Rather, we build on the classical trimmed mean estimator (Rousseeuw 1991, Lugosi and Mendelson 2021). The second step (using LASSO) builds on the high-dimensional statistics literature (Tibshirani 1996, Candes and Tao 2007, Bickel et al. 2009, Bühlmann and Van De Geer 2011).

Multitask Bandits. A few recent papers have studied multitask learning across contextual bandit instances; however, to the best of our knowledge, a key drawback of these algorithms is that none of them ultimately improve the regret bounds for any bandit instance beyond constants. Similar to the multitask learning literature discussed above, one strategy is to regularize the learned parameters for a given bandit instance towards parameters for similar bandit instances (Soare et al. 2014). For example, Cesa-Bianchi et al. (2013) and Deshmukh et al. (2017) leverage parameter updates that are similar to kernel ridge regularization, and Gentile et al. (2014) additionally perform a pre-processing step clustering bandit instances prior to such regularization; however, the resulting regret bound for a single bandit instance may actually *increase* in the number of instances N. Another popular approach is to impose a shared Bayesian prior across bandit instances (Cella et al. 2020, Bastani et al. 2021c, Kveton et al. 2021), but they also obtain similar results; furthermore, these algorithms require the more restrictive assumption that bandit instances appear sequentially (rather than simultaneously) in order to learn the prior.

We embed our robust multitask estimator across N linear contextual bandit instances; the specific setting and assumptions we consider are based on Goldenshluger and Zeevi (2013) and Bastani and Bayati (2020). We demonstrate that, unlike prior work, we obtain improved regret bounds for each bandit instance in the context dimension d under the practically-motivated sparse differences assumption; the improvement we obtain is exponential for data-poor instances where shared learning is most helpful. We also study the network structure underlying the bandit instances to better understand how one may choose N. In particular, as we incorporate more instances, we reduce variance (since we have more data) but we increase bias (since we are incorporating observations from disparate sources). We characterize the N that minimizes this bias-variance tradeoff to obtain regret bounds that scale with the density of the underlying bandit network.

1.2. Contributions

We highlight our main technical contributions below:

1. We introduce a new estimator for multitask learning, which leverages a unique combination of robust statistics (for learning a shared model across tasks) and LASSO (for debiasing this shared model for a specific task). We prove upper and lower bounds demonstrating that our estimator outperforms a number of intuitive baseline approaches.
2. We embed our estimator in a multitask bandit algorithm, resulting in regret bounds that exhibit an improved scaling in the context dimension d; notably, we show that this improvement is exponential for data-poor bandit instances.
3. We also examine regret as a function of the underlying network structure, where vertices represent bandit instances and edges capture their pairwise similarity. This sheds light on choosing the number of bandit instances for estimation, minimizing a bias-variance tradeoff.

Finally, we conclude with numerical experiments on synthetic and real datasets.

2. Problem Formulation

Before describing our models and assumptions, we establish some notations. Let $[n]$ denote the index set $\{1, 2, \cdots, n\}$. For any vector $\beta \in \mathbb{R}^d$ and $i \in [d]$, let $\beta_{(i)}$ be the i^{th} element of β; for any index set $I \subseteq [d]$, let β_I denote the vector obtained by replacing the elements of β that are not in I with zero. We use superscripts to index the bandit instance, e.g., the design matrix $\mathbf{X}^j$ corresponds to the covariates observed at bandit instance j. We use a subscript without parentheses to denote the arm, e.g. β_k^j represents the k^{th} arm for bandit instance j.

For any design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ with n observations, let $\widehat{\Sigma} = \frac{\mathbf{X}^\top \mathbf{X}}{n}$ be its sample covariance matrix. Further, for any square matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$, let $\lambda_{\min}(\mathbf{X})$ and $\lambda_{\max}(\mathbf{X})$ denote its minimum and maximum eigenvalues respectively. We use the subscript $(i,\cdot)$ to represent the i^{th} row of a matrix, the subscript $(\cdot,j)$ the j^{th} column, and the subscript (i,j) the element at location (i,j), e.g., $\mathbf{X}_{(i,\cdot)}$ is the i^{th} row of matrix $\mathbf{X}$. We use the standard notation $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$ to characterize the asymptotic growth rate of a function, and $\tilde{\mathcal{O}}(\cdot)$, $\tilde{\Omega}(\cdot)$ and $\tilde{\Theta}(\cdot)$ when logarithmic terms are omitted.

2.1. Model

We consider N distinct service providers, each facing a linear contextual bandit learning problem, e.g., N hospitals in Example 1 or N stores in Example 2 respectively. Keeping with the traditional contextual bandit framework, the decision-maker at each instance has access to the same K potential arms (decisions) with uncertain and context-dependent rewards.

Arrivals. Let T be the overall time horizon across all bandit instances. At each time step t, a new individual arrives for service at one of the N bandit instances, given by the random variable $Z_t \in [N]$. Naturally, some bandit instances may receive more arrivals than others, e.g., service providers with more traffic. Thus, we model the random arrival process as follows: every instance j is associated with a probability p_j such that $\sum_{i=1}^N p_i = 1$. At time t, the new individual arrives at instance j with probability p_j; in other words, Z_t follows a categorical distribution $\text{CG}(\mathbf{p})$ with $\mathbf{p} = \left[p_1 \cdots p_N\right]$. Thus, in expectation, instance j will serve $p_j T$ individuals. We will consider two relevant settings: (1) all instances receive similar traffic (i.e., $p_j = \Theta(1/N)$ for all $j \in [N]$), and (2) a single instance $j \in [N]$ is relatively "data-poor", receiving far less traffic than neighbouring instances (i.e., $p_j = \Theta(p_i/d^2)$ for $i \neq j$). In the data-poor setting, we focus on a single data-poor instance for simplicity; our results generalize straightforwardly to the case where there are a constant number of data-poor instances.

Each individual is also associated with a context vector $X_t \in \mathbb{R}^d$. In practice, different service providers face different customer populations, which will be reflected in the probability distribution of context vectors observed at that instance. Thus, we allow the context distribution to vary as a function of Z_t; that is, if $Z_t = j$, then X_t is drawn i.i.d. from an unknown distribution $\mathcal{P}_X^j$.

Rewards. At each instance, the local decision-maker has access to the same K arms (decisions). However, the rewards of these arms likely vary across instances (e.g., due to systematic differences among service providers or local customer populations, see discussion in Examples 1-2) as well as context distributions. Thus, we model the reward of pulling arm k for an individual with context vector X_t at instance j as

$$X_t^\top \beta_k^j + \epsilon_t.$$

Here, each arm k at instance j is parameterized by an unknown arm parameter $\beta_k^j \in \mathbb{R}^d$, and the corresponding noise ϵ_t given $Z_t = j$ is an i.i.d. σ_j-subgaussian random variable (see Definition 1); note that the variance of the noise term can depend on the instance j.

DEFINITION 1. A random variable $Z \in \mathbb{R}$ with mean $\mu = \mathbb{E}[Z]$ is σ-subgaussian if, for any $\lambda \in \mathbb{R}$, $\mathbb{E}\left[\exp\left(\lambda(Z-\mu)\right)\right] \le \exp\left(\sigma^2\lambda^2/2\right)$.

The formulation above captures *any* N linear contextual bandit instances; we now impose our assumption that these bandit instances are *similar*. As discussed in the introduction, for each arm $k \in [K]$, we assume the arm parameters are sparse relative to one another—i.e., $\beta_k^j - \beta_k^i$ is sparse for each pair $i, j \in [N]$. It is easy to see that an equivalent assumption is that there exists $\beta_k^\dagger \in \mathbb{R}^d$ such that we can write

$$\beta_k^j = \beta_k^\dagger + \delta_k^j,$$

where δ_k^j is sparse (i.e., $\|\delta_k^j\|_0 \le s$ for some $s \in \mathbb{N}$) for all $k \in [K], j \in [N]$. Intuitively, $\beta_k^\dagger$ is a shared vector that captures the similarity across all N bandit instances, and δ_k^j is an instance-specific vector that captures the heterogeneity/idiosyncrasies inherent to learning problem j. This key assumption enables us to learn across instances for a given arm k. Note that the choice of the shared vector $\beta_k^\dagger$ here is not unique — e.g., changes to $\mathcal{O}(s)$ components of $\beta_k^\dagger$ preserves the sparsity of δ_k^j up to constant factors — and therefore is not identifiable. As we describe in §3, it suffices for our purposes to estimate any vector $\widetilde{\beta}_k^\dagger$ that lies in an $\mathcal{O}(s)$ ball in ℓ_0 norm centered around an admissible choice of $\beta_k^\dagger$.

Finally, we note that we do *not* assume that the individual arm parameters $\{\beta_k^j\}_{k\in[K],j\in[N]}$ or the shared models $\{\beta_k^\dagger\}_{k\in[K]}$ are themselves sparse, since these rewards can often depend on the entire set of observed covariates (see, e.g., discussion in Bastani 2021).

Objective. We seek to construct a sequential decision-making policy π that learns the arm parameters $\{\beta_k^j\}_{k\in[K],j\in[N]}$ over time and across instances, in order to maximize expected reward for each arrival. The overall policy π is composed of sub-policies $\pi_t^j : \mathcal{X}^j \to [K]$ at each instance j. With a slight abuse of notation, we use π_t to represent the arm played at time t.

We measure the performance of π by its cumulative expected regret — the standard metric in the analysis of bandit algorithms (Lai and Robbins 1985) — modified naturally to extend across multiple heterogeneous bandit instances. In particular, when $Z_t = j$ (an individual arrives at instance j), we compare ourselves to the oracle policy π_*^j at instance j, which knows the corresponding arm parameters $\{\beta_k^j\}_{k\in[K]}$ in advance. Naturally, π_*^j chooses the arm with the best expected reward, i.e. $\pi_*^j(X_t) = \arg\max_{k\in[K]} X_t^\top \beta_k^j$ for any X_t such that $Z_t = j$. The expected regret incurred by pulling arm $\pi_t = k$ at time t given an arrival at instance j is thus

$$r_t^j = \mathbb{E}\left[\max_{k'\in[K]}(X_t^\top \beta_{k'}^j) - X_t^\top \beta_k^j \,\middle|\, Z_t = j\right],$$

which is simply the difference between the expected reward of using π_*^j and π_t^j. Further taking the expectation over the randomness in where the individual arrives, the expected regret of an overall policy composed of sub-policies $\{\pi_t^j\}_{j\in[N]}$ at time t is

$$r_t = \sum_{j\in[N]} \mathbb{P}\left[Z_t = j\right] r_t^j = \sum_{j\in[N]} p_j r_t^j.$$

Our goal is to derive a policy that minimizes the cumulative expected regret $R_T = \sum_{t=1}^T r_t$ across all instances. We also study the instance-specific cumulative expected regret $R_T^j = \sum_{t=1}^T p_j r_t^j$ for each $j \in [N]$ for standard and data-poor instances.

Network Structure. Finally, we consider the dependence of the regret on the underlying *network structure* of bandit instances, when available. In particular, we consider a fully-connected network with N vertices (each representing a bandit instance) and edge weights $s_{i,j}$ capturing the pairwise similarities between any two instances $(i,j) \in [N] \times [N]$ as measured by our sparse difference metric, i.e., $\|\beta^j - \beta^i\|_0 = s_{i,j}$. Note that this graph is undirected since $s_{i,j} = s_{j,i}$; furthermore, if two instances i and j are unrelated, then they trivially satisfy $s_{i,j} = d$.

Such a graph can be inferred based on observed covariates (e.g., geographic distance between hospitals/stores or socio-economic indices of neighborhoods served) or data from past decision-making problems (see, e.g., the disparity matrix in Crammer et al. 2008). Then, for any given problem instance j, we can optimize the subset of instances $\mathcal{Q}_j \subseteq [N]$ from which to transfer knowledge. For simplicity, we assume a strategy where we fix a threshold $\tilde{s}$, and take all instances with sparsity at most $\tilde{s}$ — i.e.,

$$\mathcal{Q}_j = \{i \in [N] \mid s_{i,j} \leq \tilde{s}\}.$$

We denote the effective number of instances by $\tilde{N} = |\mathcal{Q}_j|$. Under this assumption, there is a tradeoff between choosing smaller $\tilde{s}$, which yields smaller $\tilde{N}$ (resulting in lower bias but larger variance), and larger $\tilde{s}$, which yields larger $\tilde{N}$ (resulting in higher bias but smaller variance). The optimal choice of $\tilde{s}$ (and correspondingly, $\tilde{N}$) depends on the relationship between $\tilde{s}$ and $\tilde{N}$. We consider a natural power law scaling — i.e.,

$$\tilde{s} = \min(\tilde{N}^\alpha, d), \tag{1}$$

for some $\alpha \geq 0$. In other words, as we increase the number of neighbouring instances we include, our sparsity parameter increases by some power law $\tilde{N}^\alpha$ until it eventually hits the maximum possible value d. Our main result allows us to easily compute the optimal choice of $\tilde{N}$, resulting in regret bounds that scale with the network density α.

2.2. Assumptions

We state our assumptions for the static supervised learning setting (RMEstimator) and dynamic online learning setting (RMBandit) respectively as follows.

RMEstimator. Our static setting focuses on a single arm $k \in [K]$ across instances $j \in [N]$ and we drop the index k for this setting. Our first assumption is standard in the literature and states that our features and regression parameters are bounded by a constant.

ASSUMPTION 1 **(Boundedness)**. *There exist a positive constant $x_{\max}$ such that each feature is bounded by $x_{\max}$, i.e., $\|X\|_\infty \le x_{\max}$ for any $X \in \mathcal{X}^j, j \in [N]$, and a positive constant b such that $\|\beta^j\|_1 \le b$ for any $j \in [N]$.*

Our second assumption is that the sample covariance matrix of any instance $j \in [N]$, i.e., $\widehat{\Sigma}^j = \frac{\mathbf{X}^{j\top}\mathbf{X}^j}{n_j}$, is positive-definite, where $\mathbf{X}^j \in \mathbb{R}^{n_j \times d}$ encodes the n_j observed context vectors; thereby, ordinary least squares (OLS) is well-defined. Note that this assumption is not needed in the bandit setting; instead, our bandit algorithm adaptively labels data in a way that ensures that this assumption holds.

ASSUMPTION 2 **(Positive-Definiteness)**. *There exists a positive constant ψ such that for any $j \in [N]$ we have $\lambda_{\min}(\widehat{\Sigma}^j) \ge \psi$.*

RMBandit. Our bandit setting first assumes Assumption 1 on boundedness for any arm $k \in [K]$. As discussed earlier, we embed our robust multitask estimator into the high-dimensional linear contextual bandit setting studied in Bastani and Bayati (2020); therefore, our next three assumptions are directly adapted from this literature. We note that the remaining assumptions in this section are only required for the regret analysis of RMBandit (§4) and *not* for the static performance bounds of our robust multitask estimator (§3).

Our second assumption is a mild margin condition that ensures that the density of the context distribution $\mathcal{P}_X^j$ for each instance j is bounded near a decision boundary (i.e., the intersection of the hyperplane given by $\{X \mid X^\top \beta_{k'}^j = X^\top \beta_k^j\}$ and $\mathcal{X}^j$ for any pair of arms $k' \neq k$). It allows for any bounded, continuous features, as well as any discrete features with a finite number of values.

ASSUMPTION 3 **(Margin Condition)**. *For any arms k and k' of any instance $j \in [N]$, there exists a positive constant L such that $\mathbb{P}\left[|X^\top(\beta_k^j - \beta_{k'}^j)| \le \kappa \mid Z = j\right] \le L\kappa$ for any $\kappa > 0$.*

Our third assumption is that, for each instance $j \in [N]$, our K arms can be split into two mutually exclusive sets:

1. Optimal arms $k \in \mathcal{K}_{\text{opt}}^j$ that are *strictly* optimal in expected reward (by at least h) for any contexts drawn from a set $U_k^j \subset \mathcal{X}^j$ with positive support on $\mathcal{P}_X^j$, i.e., $\mathbb{P}[X \in U_k^j \mid Z = j] \ge p_*$.

2. Sub-optimal arms $k \in \mathcal{K}_{\text{sub}}$ that are *strictly* sub-optimal in expected reward (by at least h) for all contexts in $\mathcal{X}^j$.

In other words, we assume that every arm is either optimal (by at least h) for at least *some* individuals, or sub-optimal for *all* individuals (by at least h). This assumption will ensure that every arm in $\mathcal{K}_{\text{opt}}^j$ will roughly receive at least $p_* p_j T$ samples under a regret-minimizing policy on instance j, ensuring that we can quickly learn accurate parameter estimates for all optimal arms.

ASSUMPTION 4 **(Arm Optimality)**. *All K arms at any given instance j belong to one of two mutually exclusive sets: $\mathcal{K}_{opt}^j$ of optimal arms or $\mathcal{K}_{sub}^j$ of suboptimal arms. There exists some $h > 0$ such that: (i) each $k \in \mathcal{K}_{sub}^j$ satisfies $X^\top \beta_k^j < \max_{k' \neq k} X^\top \beta_{k'}^j - h$ for any context $X \in \mathcal{X}^j$, and (ii) each $k \in \mathcal{K}_{opt}^j$ is optimal on a set of contexts*

$$U_k^j = \{X \in \mathcal{X}^j \mid X^\top \beta_k^j > \max_{k' \neq k} X^\top \beta_{k'}^j + h\},$$

which has positive measure, i.e., $\mathbb{P}\left[X \in U_k^j \mid Z = j\right] \geq p_*$ for some $p_* > 0$.

Our fourth assumption ensures that linear regression is feasible within the set U_k^j; this is a mild assumption since it is with respect to the *true* covariance matrix, which only requires that no features are perfectly collinear in this set. In contrast, the *sample* covariance matrix may not be positive-definite at time t, since we may have observed too few samples from that instance.

ASSUMPTION 5 **(Positive-Definiteness)**. *For every arm $k \in [K]$ and instance $j \in [N]$, the true covariance matrix $\Sigma_k^j = \mathbb{E}[XX^\top | X \in U_k^j, Z = j]$ is positive-definite—i.e., $\lambda_{\min}\left(\Sigma_k^j\right) \geq \psi$ for some $\psi > 0$.*

The assumptions thus far are standard and have been adapted directly from the literature. We now introduce a new assumption motivated by our multitask setting. In general, an arm k can be optimal (belong to $\mathcal{K}_{\text{opt}}^j$) at one bandit instance j and be sub-optimal (belong to $\mathcal{K}_{\text{sub}}^i$) at a neighboring instance i. This implies that we will observe $\mathcal{O}(p_j T)$ samples from arm k at instance j but only $\mathcal{O}(\log(p_i T))$ samples at instance i under a regret-minimizing policy; in other words, instance j cannot effectively transfer knowledge from instance i about arm k. Thus, we impose that if an arm $k \in [K]$ is optimal for *any* instance j, it is also optimal for at least *some* subset of the N neighboring instances so that we have enough observations to enable multitask learning.

ASSUMPTION 6 **(Optimality Density)**. *The set of instances with arm k being an optimal arm, i.e., $\mathcal{W}_k = \left\{j \in [N] \mid k \in \mathcal{K}_{opt}^j\right\}$, has cardinality at least ρN for some $\rho > 0$ for each $k \in [K]$ such that $|\mathcal{W}_k| > 0$.*

3. RMEstimator

First, we study the static supervised learning setting. In this section, we overview (§3.2) and provide intuition (§3.3) on the design of our robust multitask estimator; we provide theoretical performance guarantees in the standard (§3.4) and data-poor (§3.5) regimes, contrasting these guarantees with those of intuitive baseline estimators (§3.6).

3.1. Preliminaries

In what follows, we focus on a single arm $k \in [K]$ across different instances $j \in [N]$; thus, we drop the index k throughout this section. Recall that the reward for context $X_t \in \mathbb{R}^d$ at instance j is

$$Y_t = X_t^\top \beta^j + \epsilon_t,$$

where the noise ϵ_t at instance j are i.i.d. σ_j-subgaussian random variables. For each instance j, let the matrix $\mathbf{X}^j \in \mathbb{R}^{n_j \times d}$ encode the n_j observed context vectors, the vector $Y^j \in \mathbb{R}^{n_j}$ encode the corresponding observed rewards, and the vector $\epsilon^j \in \mathbb{R}^{n_j}$ encode the instance-specific noise. Our goal is to use $\{(\mathbf{X}^j, Y^j)\}_{j\in[N]}$ to estimate the unknown parameter vector β^j for each instance j.

Next, we define and briefly review the *trimmed mean* estimator from the classical robust statistics literature (Rousseeuw 1991, Lugosi and Mendelson 2021), which computes the mean of a distribution $\mathcal{P}$ given samples $\{Z_j\}_{j\in[N]}$. A typical setting is as follows: most of the samples are i.i.d. (i.e., $Z_j \sim \mathcal{P}$), but a small fraction (indexed by the unknown set $\mathcal{J} \subseteq [N]$) are "corrupted" and can be arbitrary. In such settings, the traditional mean can be arbitrarily biased, but the trimmed mean can obtain strong guarantees given a bound on the number of corrupted samples $|\mathcal{J}| < \zeta N$ for some $\zeta < 1/2$. The trimmed mean estimator first sorts the samples in increasing order to obtain $Z_{j_1} \leq \cdots \leq Z_{j_N}$, where the subscript j_ι is the index of the ι^{th} smallest sample. Then, given a hyperparameter $\omega > \zeta$, it removes the top and bottom $N\omega$ values and takes the mean of the remaining ones—i.e.,

$$\text{TrimmedMean}\left(\{Z_j\}_{j\in[N]},\, \omega\right) = \frac{1}{N(1-2\omega)} \sum_{\iota=N\omega+1}^{N(1-\omega)} Z_{j_\iota}.$$

Intuitively, this estimator is robust since either the corruptions are among the deleted values, or they are sufficiently close to the true mean that they do not significantly affect the estimate.

3.2. Algorithm Overview

Our robust multitask estimator is summarized in Algorithm 1. At a high level, the first step combines high-variance OLS estimators across instances using robust statistics to estimate the shared parameter $\beta^\dagger$ (up to $\mathcal{O}(s)$ deviations in ℓ_0 norm); then, the second step uses LASSO regression to debias this estimate for each specific instance $j \in [N]$.

In more detail,

Algorithm 1 Robust Multitask Estimator

Inputs: Instance-specific regularization parameters $\{\lambda_j\}_{j\in[N]}$, trimmed mean hyperparameter ω

for $j \in [N]$ **do**

Let $\widehat{\beta}^j_{\text{ind}} = (\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}\mathbf{X}^\top Y^j$ be the OLS estimator for training data $(\mathbf{X}^j, Y^j)$

end for

for $i \in [d]$ **do**

Let $\widehat{\beta}^\dagger_{\text{RM},(i)} = \text{TrimmedMean}(\{\widehat{\beta}^j_{\text{ind},(i)}\}_{j\in[N]},\, \omega)$ be the element-wise trimmed mean (i.e., mean with the top and bottom ω quantiles removed)

end for

for $j \in [N]$ **do**

Let $\widehat{\beta}^j_{\text{RM}} = \arg\min_\beta \left\{\frac{1}{n_j}\|\mathbf{X}^j\beta - Y^j\|_2^2 + \lambda_j\|\beta - \widehat{\beta}^\dagger_{\text{RM}}\|_1\right\}$

end for

Outputs: $\{\widehat{\beta}^j_{\text{RM}}\}_{j\in[N]}$

- **Step 1 (Estimating $\beta^\dagger$):** We compute the usual OLS estimator

$$\widehat{\beta}^j_{\text{ind}} = (\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}\mathbf{X}^{j\top}Y^j$$

for each instance $j \in [N]$ independently. Then, we combine these estimates using the element-wise trimmed mean to estimate the shared parameter vector $\widehat{\beta}^\dagger_{\text{RM}} \approx \beta^\dagger$—i.e., for each $i \in [d]$,

$$\widehat{\beta}^\dagger_{\text{RM},(i)} = \text{TrimmedMean}\left(\{\widehat{\beta}^j_{\text{ind},(i)}\}_{j\in[N]},\, \omega\right), \tag{2}$$

where $\omega > 0$ is the trimmed mean hyperparameter that we specify later.

- **Step 2 (Estimating β^j)**: Next, we use LASSO regression to compute $\widehat{\beta}^j_{\text{RM}}$, leveraging our assumption that the instance-specific bias term $\beta^j - \beta^\dagger$ is sparse:

$$\widehat{\beta}^j_{\text{RM}} = \arg\min_\beta \left\{\frac{1}{n_j}\|\mathbf{X}^j\beta - Y^j\|_2^2 + \lambda_j\|\beta - \widehat{\beta}^\dagger_{\text{RM}}\|_1\right\} \tag{3}$$

We make a minor modification in the data-poor regime: we omit the data-poor instance j from the trimmed mean in Step 1 since $\widehat{\beta}^j_{\text{ind}}$ has particularly high variance (see §3.5 for details).

3.3. Design Intuition

We now provide intuition for our design choices relative to alternative strategies; the corresponding error rates are summarized in Table 1 (see §3.6 for precise definitions and more details).

One strategy is to simply use the independent OLS estimator $\widehat{\beta}^j_{\text{ind}}$ (from Step 1) to estimate β^j; this is an unbiased estimator, but has very high variance since it only uses the limited data observed in instance j and does not leverage shared structure across instances. As a result, it has high error when n_j is small (see Table 1).

Estimator	**Estimation Error** *Standard Regime*	*Data-Poor Regime*	**Bound Type**
Independent $\widehat{\beta}^j_{\text{ind}}$	$\frac{d}{\sqrt{n_j}}$	$\frac{d}{\sqrt{n_j}}$	Lower
Averaging $\widehat{\beta}^j_{\text{avg}}$	$\\|\delta^j\\|_1 + \frac{d}{\sqrt{Nn_j}}$	$\\|\delta^j\\|_1 + \frac{1}{\sqrt{Nn_j}}$	Lower
Pooling $\widehat{\beta}^j_{\text{pool}}$	$\\|\delta^j\\|_1 + \frac{d}{\sqrt{Nn_j}}$	$\\|\delta^j\\|_1 + \frac{1}{\sqrt{Nn_j}}$	Lower
Averaging Multitask $\widehat{\beta}^j_{\text{AM}}$	$\frac{\min\{Ns,d\}}{\sqrt{n_j}} + \frac{d}{\sqrt{Nn_j}}$	$\frac{\min\{Ns,d\}}{\sqrt{n_j}}$	Lower
Robust Multitask $\widehat{\beta}^j_{\text{RM}}$	$\sqrt{\frac{sd}{n_j}} + \frac{d}{\sqrt{Nn_j}}$	$\frac{s}{\sqrt{n_j}}$	Upper

Table 1 **Comparison of parameter estimation error $\sup_{\mathcal{G}} \mathbb{E}\left[\|\widehat{\beta}^j - \beta^j\|_1\right]$ (see §3.6 for the precise definitions of these estimators); constants and logarithmic factors are omitted for clarity. The upper bound for our robust multitask estimator outperforms the worst-case lower bounds for intuitive baseline estimators under the same set of problem settings $\mathcal{G}$; our improvement is largest for data-poor instances.**

An alternative strategy is to estimate the shared model $\beta^\dagger$ using data across instances, e.g., the *averaging* estimator takes the model average of the independent estimators:

$$\widehat{\beta}^j_{\text{avg}} = \frac{1}{N}\sum_{i\in[N]} \widehat{\beta}^i_{\text{ind}}.$$

This estimator has low variance since it leverages data across instances, but it is biased since it does not account for the instance-specific idiosyncratic bias term $\delta^j = \beta^j - \beta^\dagger$ (the trimmed mean estimator $\widehat{\beta}^\dagger_{\text{RM}}$ that we actually use in Step 1 of Algorithm 1 suffers the same bias; we will explain the purpose of using the trimmed mean shortly). Similarly, estimating the shared model $\beta^\dagger$ through OLS on data *pooled* across instances suffers the same drawbacks. As shown in Table 1, the error of such estimators never approaches zero due to the bias term δ^j.

Thus, a natural two-step strategy to achieve low variance and low bias is to first compute an estimate $\widehat{\beta}^\dagger$ of the shared parameter, and then try to *debias* it to estimate β^j. Since the bias $\beta^j - \beta^\dagger$ is s-sparse by assumption, it should intuitively be easier to debias $\widehat{\beta}^\dagger$ than to directly estimate β^j.

Along these lines, consider the following *averaging multitask* estimator, denoted by the subscript AM. Here, we estimate the shared parameter via model averaging, $\widehat{\beta}^\dagger_{\text{AM}} = \widehat{\beta}^j_{\text{avg}}$. Then, we use an ℓ_1 penalty on $\beta - \widehat{\beta}^\dagger_{\text{AM}}$ (i.e., LASSO regression) on data from instance j to debias $\widehat{\beta}^\dagger_{\text{AM}}$:

$$\widehat{\beta}^j_{\text{AM}} = \arg\min_{\beta}\left\{\frac{1}{n_j}\|\mathbf{X}^j\beta - Y^j\|_2^2 + \lambda_j\|\beta - \widehat{\beta}^\dagger_{\text{AM}}\|_1\right\}. \tag{4}$$

(Note that this strategy is identical to Algorithm 1, except it uses the traditional mean instead of the trimmed mean in Step 1.) To see why equation (4) helps, suppose we had a perfect estimate of the shared model $\widehat{\beta}^\dagger_{\text{AM}} = \beta^\dagger$; then, $\beta^j - \widehat{\beta}^\dagger_{\text{AM}}$ would be s-sparse, in which case LASSO requires exponentially fewer observations for recovering β^j (relative to $\widehat{\beta}^\dagger_{\text{AM}}$) than traditional OLS.

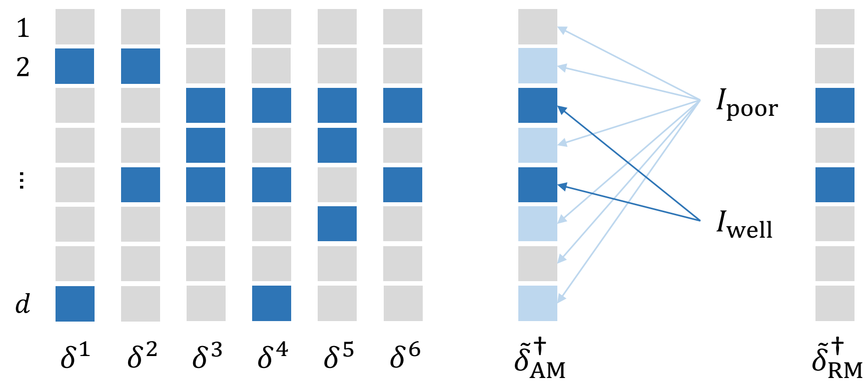


Figure 1 **Illustration of Step 1 of our robust multitask estimator for debiasing data collected from multiple instances. Blue squares depict the support; the shade of blue depicts the magnitude. $\mathcal{I}_{\text{poor}}$ represents the index set which can be debiased using the trimmed mean across instances, while $\mathcal{I}_{\text{well}}$ represents the index set which can be debiased using a subsequent LASSO regression for the target instance.**

The issue with the approach outlined above is that $\beta^j - \widehat{\beta}^\dagger_{\mathrm{AM}}$ is *not* s-sparse, or even "close" to being s-sparse. To illustrate, we can decompose

$$\beta^j - \widehat{\beta}^\dagger_{\mathrm{AM}} = \underbrace{\beta^j - \beta^\dagger}_{s\text{-sparse}} + \underbrace{\beta^\dagger - \widetilde{\beta}^\dagger_{\mathrm{AM}}}_{(Ns)\text{-sparse}} + \underbrace{\widetilde{\beta}^\dagger_{\mathrm{AM}} - \widehat{\beta}^\dagger_{\mathrm{AM}}}_{\text{not sparse but small}}, \quad \text{where } \widetilde{\beta}^\dagger_{\mathrm{AM}} = \frac{1}{N}\sum_{j\in[N]} \beta^j \tag{5}$$

Here, $\widetilde{\beta}^\dagger_{\mathrm{AM}}$ is the value that $\widehat{\beta}^\dagger_{\mathrm{AM}}$ converges to as $n_j \to \infty$, $j \in [N]$. Note that $\widehat{\beta}^\dagger_{\mathrm{AM}}$ does not converge to $\beta^\dagger$; in fact, as noted in the problem formulation, $\beta^\dagger$ is not identifiable. The first term in the decomposition is sparse, and the third term becomes small as $n = \sum_{j\in[N]} n_j$ becomes large (since $\widehat{\beta}^\dagger_{\mathrm{AM}}$ effectively uses all n samples to estimate $\widetilde{\beta}^\dagger_{\mathrm{AM}}$); since LASSO can effectively recover parameters that are approximately sparse, these two terms are not problematic. The key issue is the second term:

$$\widetilde{\delta}^\dagger_{\mathrm{AM}} = \widetilde{\beta}^\dagger_{\mathrm{AM}} - \beta^\dagger = \frac{1}{N}\sum_{j\in[N]}(\beta^j - \beta^\dagger) = \frac{1}{N}\sum_{j\in[N]} \delta^j, \tag{6}$$

which is neither sparse nor small. This is illustrated in Figure 1: since the support of the different bias terms $\{\delta_i\}_{i\in[N]}$ can be "poorly-aligned" (i.e., the idiosyncrasies for each instance affect a different subset of features), the average across instances can result in $\widetilde{\delta}^\dagger_{\mathrm{AM}}$ having as many as $\min\{Ns, d\}$ nonzero components (even as $n_j \to \infty$, $j \in [N]$). This in turn implies that $\beta^j - \widehat{\beta}^\dagger_{\mathrm{AM}}$ is not sparse even for moderate values of N such as $N = \Omega(d/s)$; thus, we cannot use LASSO to efficiently debias $\widehat{\beta}^\dagger_{\mathrm{AM}}$. Other classical estimators of the shared parameter (e.g., data pooling) suffer the same issue.

Our robust multitask estimator addresses this issue by using the trimmed mean $\widehat{\beta}^\dagger_{\mathrm{RM}}$ in Step 1; we will show that this converges to a value $\widetilde{\beta}^\dagger_{\mathrm{RM}}$ (as $n_j \to \infty$, $j \in [N]$) such that

$$\widetilde{\delta}^\dagger_{\mathrm{RM}} = \widetilde{\beta}^\dagger_{\mathrm{RM}} - \beta^\dagger = \mathrm{TrimmedMean}\left(\{\beta^j\}_{j\in[N]} - \beta^\dagger, \omega\right) = \mathrm{TrimmedMean}\left(\{\delta^j\}_{j\in[N]}, \omega\right)$$

is $\mathcal{O}(s)$-sparse. In particular, we have the following decomposition:

$$\beta^j - \widehat{\beta}^\dagger_{\mathrm{RM}} = \underbrace{\beta^j - \beta^\dagger}_{s\text{-sparse}} + \underbrace{\beta^\dagger - \widetilde{\beta}^\dagger_{\mathrm{RM}}}_{O(s)\text{-sparse}} + \underbrace{\widetilde{\beta}^\dagger_{\mathrm{RM}} - \widehat{\beta}^\dagger_{\mathrm{RM}}}_{\text{not sparse but small}} . \tag{7}$$

As discussed above, the third term becomes small as n becomes large. Since the second term $\widetilde{\delta}^\dagger_{\mathrm{RM}}$ is $\mathcal{O}(s)$-sparse, $\beta^j - \widehat{\beta}^\dagger_{\mathrm{RM}}$ is approximately $\mathcal{O}(s)$-sparse; thus, LASSO can efficiently debias $\widehat{\beta}^\dagger_{\mathrm{RM}}$.

We now use a counting argument to illustrate why $\widetilde{\delta}^\dagger_{\mathrm{RM}}$ is $\mathcal{O}(s)$-sparse. As Figure 1 illustrates, we can separate the components $i \in [d]$ into two groups: ones that are "well-aligned" ($i \in \mathcal{I}_{\mathrm{well}}$) and ones that are "poorly-aligned" ($i \in \mathcal{I}_{\mathrm{poor}}$); see Definition 2 below. A poorly-aligned component i is one where very few instances $j \in [N]$ are biased in this component, i.e., $\beta^j_{(i)} \neq \beta^\dagger_{(i)}$. Intuitively, for each such component, the trimmed mean estimator treats these biased instances as "corruptions" to our samples $\{\beta^j_{(i)}\}_{j \in [N]}$, and trims them (with high probability) when computing the average to obtain an unbiased estimate of $\beta^\dagger_{(i)}$. On the other hand, well-aligned components may remain arbitrarily biased. However, the pigeonhole principle implies that there cannot be many well-aligned components; thus, these components (in addition to the components affected by the sparse instance-specific bias term) can be efficiently debiased by LASSO in Step 2. We now formalize this.

Definition 2 (Well- and poorly-aligned components). Given a constant $\zeta \in [0, 1]$, a component $i \in [d]$ is *ζ-poorly-aligned* (denoted $i \in \mathcal{I}^\zeta_{\mathrm{poor}}$) if

$$\frac{|\{j \in [N] \mid \beta^j_{(i)} \neq \beta^\dagger_{(i)}\}|}{N} < \zeta.$$

Otherwise, it is *ζ-well-aligned* (denoted $i \in \mathcal{I}^\zeta_{\mathrm{well}}$).

In other words, a component i is ζ-poorly-aligned if at most ζ fraction of j's satisfy $\beta^j_{(i)} \neq \beta^\dagger_{(i)}$. Now, Step 1 constructs an estimator $\widehat{\beta}^\dagger_{\mathrm{RM}}$ of $\beta^\dagger$ that converges to

$$\widetilde{\beta}^\dagger_{\mathrm{RM},(i)} = \begin{cases} \beta^\dagger_{(i)} & \text{if } i \in \mathcal{I}^\zeta_{\mathrm{poor}} \\ \text{unspecified} & \text{if } i \in \mathcal{I}^\zeta_{\mathrm{well}} \end{cases}$$

as n_j's become large. That is, we aim to correctly estimate all the poorly-aligned components, but the well-aligned components can be anything. We estimate each component $\beta^\dagger_{(i)}$ using the trimmed mean, which is robust to a small fraction ζ of arbitrarily corrupted samples. For a given component i, let the corresponding corrupted instances be

$$\mathcal{J}_i = \{j \in [N] \mid \beta^j_{(i)} \neq \beta^\dagger_{(i)}\}.$$

By definition, for $i \in \mathcal{I}^\zeta_{\mathrm{poor}}$, we have $|\mathcal{J}_i| < N\zeta$. Thus, we can use the trimmed mean estimator to estimate $\beta^\dagger_{(i)}$:

$$\widehat{\beta}^\dagger_{\mathrm{RM},(i)} = \mathrm{TrimmedMean}\left(\{\widehat{\beta}^j_{\mathrm{ind},(i)}\}_{j \in [N]}, \omega\right)$$

for some $\omega > \zeta$. This strategy ensures that $\widehat{\beta}^{\dagger}_{\mathrm{RM},(i)} \approx \beta^{\dagger}_{(i)}$ for each poorly-aligned component as desired. Now, note that there can only be a few well-aligned components. In particular, out of the Nd total components in $\{\beta^j_{(i)}\}_{j\in[N]}$, there are at most Ns components where $\beta^j_{(i)} \neq \beta^{\dagger}_{(i)}$ as a consequence of our problem formulation. Then, by the pigeonhole principle, we have

$$|\mathcal{I}^{\zeta}_{\mathrm{well}}| \leq \frac{Ns}{N\zeta} = \frac{s}{\zeta}.$$

In other words, there are at most s/ζ well-aligned components, so $\widetilde{\delta}^{\dagger}_{\mathrm{RM}}$ is $\mathcal{O}(s)$-sparse as desired (for a constant choice of ζ). Thus, we can efficiently debias our estimate using LASSO in Step 2.

3.4. Theoretical Analysis

Next, we provide bounds on the parameter estimation error of our robust multitask estimator. Our first result bounds the error of the trimmed mean estimator.

PROPOSITION 1. *Suppose we are given N samples $\{Z_j\}_{j\in[N]}$ and a subset $\mathcal{J} \subseteq [N]$ of size $|\mathcal{J}| < N\zeta$ with $\zeta \in (0, 1/2)$, such that $\{Z_j\}_{j\in\mathcal{J}^c}$ are independent σ_j-subgaussian random variables with equal means $\mu = \mathbb{E}[Z_j]$. Then, letting $\widehat{\mu} =$ TrimmedMean$(\{Z_j\}_{j\in[N]}, \omega)$ with $\omega = \zeta + \eta$, we have*

$$\mathbb{P}\left[|\widehat{\mu} - \mu| \geq C_0 \max_{j\in\mathcal{J}^c} \sigma_j\,(3\zeta + 4\eta)\sqrt{\log(\frac{3}{\eta})}\right] \leq 3\exp\left(-\frac{N\eta^2}{9}\right),$$

for any positive η and a constant $C_0 > 2$ satisfying $\eta \leq 1/2 - 1/C_0 - \zeta$.

The proof is provided in Appendix A.1. We use Proposition 1 to show that $\widehat{\beta}^{\dagger}_{\mathrm{RM},(i)}$ is close to the true mean $\beta^{\dagger}_{(i)}$ for poorly-aligned components $i \in \mathcal{I}^{\zeta}_{\mathrm{poor}}$. This result is similar to classical results from robust statistics (see, e.g., Li 2019), but existing results typically assume that the uncorrupted samples are i.i.d., whereas we only require independence (since we wish to apply it to $\{\widehat{\beta}^j_{\mathrm{ind},(i)}\}_{j\in[N]}$, which are not identically distributed).

Next, we have the following error bound for our robust multitask estimator on each instance j.

THEOREM 1. *Given $\zeta \in (0, 1/2)$, the robust multitask estimator $\widehat{\beta}^j_{RM}$ of β^j for any instance $j \in [N]$ computed by Algorithm 1 satisfies*

$$\mathbb{P}\left[\|\widehat{\beta}^j_{RM} - \beta^j\|_1 \geq \frac{6\lambda_j s}{\zeta\psi} + C_0 d\,(3\zeta + 4\eta)\max_{i\in[N]}\sqrt{\frac{\sigma_i^2}{n_i\psi}\log(\frac{3}{\eta})}\right] \leq 3d\exp\left(-\frac{N\eta^2}{9}\right) + 2d\exp\left(-\frac{\lambda_j^2 n_j}{32\sigma_j^2 x_{\max}^2}\right),$$

for any regularization parameter $\lambda_j > 0$ and for any $\eta > 0$ and a constant $C_0 > 2$ satisfying $\eta \leq 1/2 - 1/C_0 - \zeta$.

The proof is provided in Appendix A.2. The following corollary bounds the estimation error in the standard regime where each instance $j \in [N]$ has a similar number of observations n_j:

Corollary 1. *Letting* $\lambda_j = \sqrt{\frac{32\sigma_j^2 x_{\max}^2}{n_j}\log(\frac{4d}{\delta})}$, $\zeta = (\frac{C_0-2}{4C_0})\sqrt{\frac{s}{d}}$, *and* $\eta = \sqrt{\frac{9}{N}\log(\frac{6d}{\delta})}$, *the robust multitask estimator for any instance* $j \in [N]$ *computed by Algorithm 1 satisfies*

$$\begin{aligned}\|\widehat{\beta}^j_{RM} - \beta^j\|_1 \leq &\left(\frac{96C_0}{(C_0-2)\psi}\sqrt{2\sigma_j^2 x_{\max}^2 \log(\frac{4d}{\delta})} + \frac{3C_0}{4}\max_{i\in[N]}\sqrt{\frac{\sigma_i^2 n_j \log(N)}{n_i\psi}}\right)\sqrt{\frac{sd}{n_j}} \\ &+ \left(12C_0 \max_{i\in[N]}\sqrt{\frac{\sigma_i^2 n_j \log(N)\log(\frac{6d}{\delta})}{\psi n_i}}\right)\frac{d}{\sqrt{Nn_j}} \\ =&\tilde{\mathcal{O}}\left(\sqrt{\frac{sd}{n_j}} + \frac{d}{\sqrt{Nn_j}}\right),\end{aligned}$$

with probability at least $1-\delta$ *for any* $\delta \geq \exp\left(-\frac{N}{9}(\frac{1}{2} - \frac{1}{C_0} - \frac{C_0-2}{4C_0}\sqrt{\frac{s}{d}})^2 + \log(6d)\right)$.

The proof is provided in Appendix A.2. Recall that the independent OLS estimator $\widehat{\beta}^j_{\text{ind}}$ on instance j yields an estimation error of $\mathcal{O}(\frac{d}{\sqrt{n_j}})$. In contrast, if the number of instances is at least $N = \Omega(d/s)$, our robust multitask estimator has an estimation error of at most $\tilde{\mathcal{O}}\left(\sqrt{\frac{sd}{n_j}}\right)$ with high probability, i.e., it provides an improvement of $\sqrt{d}$, which can be substantial in high dimension. When we have very few instances from which to share knowledge (i.e., $N = o(d)$), multitask learning is less effective and we obtain the same estimation error as the independent OLS estimator. As we discuss next, our improvement is much larger in the data-poor regime.

3.5. Data-Poor Regime

Multitask learning is especially effective in the data-poor regime, where the target instance j receives substantially fewer observations compared to other instances. In particular, we consider the case where $n_j = \Theta(n_i/d^2)$ for all $i \neq j$. We focus on a single data-poor instance for simplicity; our results generalize straightforwardly to the case where there are a constant number of data-poor instances.

The following corollary bounds the estimation error for data-poor instance j:

Corollary 2. *Letting* $\lambda_j = \sqrt{\frac{32\sigma_j^2 x_{\max}^2}{n_j}\log(\frac{4d}{\delta})}$, $\zeta = \frac{C_0-2}{4C_0}$, *and* $\eta = \sqrt{\frac{9}{N}\log(\frac{6d}{\delta})}$, *the robust multitask estimator for a data-poor instance* j *computed by Algorithm 1 satisfies*

$$\begin{aligned}\|\widehat{\beta}^j_{RM} - \beta^j\|_1 \leq &\left(\frac{96C_0}{(C_0-2)\psi}\sqrt{2\sigma_j^2 x_{\max}^2 \log(\frac{4d}{\delta})}\right)\frac{s}{\sqrt{n_j}} + \left(\frac{3C_0}{4}\max_{i\neq j}\sqrt{\frac{d^2 n_j}{n_i}\frac{\sigma_i^2\log(N)}{\psi}}\right)\frac{1}{\sqrt{n_j}} \\ &+ \left(12C_0\max_{i\neq j}\sqrt{\frac{d^2 n_j}{n_i}\frac{\sigma_i^2 \log(N)\log(\frac{6d}{\delta})}{\psi}}\right)\frac{1}{\sqrt{Nn_j}} \\ =&\tilde{\mathcal{O}}\left(\frac{s}{\sqrt{n_j}}\right),\end{aligned}$$

with probability at least $1-\delta$ *for any* $\delta \geq \exp\left(-\frac{N}{36}(\frac{1}{2} - \frac{1}{C_0})^2 + \log(6d)\right)$.

The proof is provided in Appendix A.2. In this setting, the estimation error of our robust multitask error depends only logarithmically on the context dimension d (as opposed to linearly for independent OLS). In other words, we obtain an *exponential* reduction in estimation error in d, which is especially valuable in high dimension.

3.6. Comparison with Baselines

Finally, we discuss how the estimation error of our robust multitask estimator compares with the baseline approaches discussed in §3.3. In particular, we contrast the upper bounds we derived for our estimator with lower bounds for these baselines in both the standard and data-poor regimes; these bounds are summarized in Table 1. Detailed statements and proofs are provided in Appendix B.

We characterize the estimation error of an estimator $\widehat{\beta}^j$ through the following loss function:

$$\ell(\widehat{\beta}^j, \beta^j) = \sup_{\mathcal{G}} \mathbb{E}\left[\|\widehat{\beta}^j - \beta^j\|_1\right], \tag{8}$$

where $\mathcal{G} = \left\{\{\mathbf{X}^j\}_{j\in[N]}, \{\beta^j\}_{j\in[N]}, \{\mathcal{P}_\epsilon^j\}_{j\in[N]}\right\}$ satisfies our assumptions in §3.1, $\mathcal{P}_\epsilon^j$ is the distribution of the noise terms ϵ^j, and the expectation is taken with respect to ϵ^j's. This choice of $\mathcal{G}$ ensures that our upper and lower bounds are with respect to the same class of problem instances.

We consider the following estimators:

- **Independent OLS (Appendix B.1):** This is the OLS estimator $\widehat{\beta}^j_{\text{ind}} = (\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}\mathbf{X}^\top Y^j$ trained on data only from instance j, i.e., it does not transfer information across instances.
- **Averaging (Appendix B.2):** This estimator $\widehat{\beta}^j_{\text{avg}} = \frac{1}{N}\sum_{i\in[N]}\widehat{\beta}^i_{\text{ind}}$ is a common approach that averages the independent OLS estimates across instances (see, e.g., Dobriban and Sheng 2021).
- **Pooling (Appendix B.3):** This estimator $\widehat{\beta}^j_{\text{pool}} = \left(\sum_{i\in[N]}\mathbf{X}^{i\top}\mathbf{X}^i\right)^{-1}\left(\sum_{i\in[N]}\mathbf{X}^{i\top}Y^i\right)$ is a common approach that pools data across instances to train a single OLS estimator (see, e.g., Crammer et al. 2008, Ben-David et al. 2010):
- **Averaging multitask (Appendix B.4):** This two-step estimator $\widehat{\beta}^j_{\text{AM}}$ is described in detail in §3.3. It is an ablation of our robust multitask estimator that uses the traditional mean rather than the trimmed mean in Step 1.

4. RMBandit Algorithm

Next, we leverage our robust multitask estimator to efficiently learn across N simultaneous linear contextual bandit instances; we extend the single-bandit model with dense (i.e., not sparse) arm parameter vectors studied in Goldenshluger and Zeevi (2013) and Bastani and Bayati (2020). Throughout this section, we drop the subscript RM and denote our robust multitask estimator as $\widehat{\beta}^j_k$ for arm k and instance j.

In this section, we describe our Robust Multitask Bandit (RMBandit) algorithm (§4.1); we demonstrate improved total and instance-specific regret bounds in the standard (§4.2) and data poor regimes (§4.4), along with an overview of the proof strategy (§4.5); we also study the regret dependence on the network structure underlying bandit instances (§4.3).

4.1. Algorithm Description

Our RMBandit algorithm is presented in Algorithm 2. Following prior work, RMBandit manages the exploration-exploitation tradeoff using a small amount ($\mathcal{O}(\log(T))$) of forced random exploration in each instance $j \in [N]$. Furthermore, for each instance j and arm $k \in [K]$, it trades off between (i) an unbiased *forced-sample* estimator, which is trained only on forced random samples, and (ii) a potentially biased *all-sample* estimator, which is trained on all observations for arm k. Instead of using LASSO (Bastani and Bayati 2020) or OLS (Goldenshluger and Zeevi 2013) for these estimators, we use our robust multitask estimator.

This introduces two important challenges. First, our multitask estimator leverages data *across* instances, which induces (previously absent) correlations between our arm parameter estimates $\{\widehat{\beta}_k^j\}_{j\in[N]}$ for a fixed arm k. However, our error bound for the trimmed mean estimator (Proposition 1) requires that our estimates across instances be independent in order to recover a reasonable estimate of the shared model $\beta^\dagger$. Thus, we introduce a new *batching* strategy, where we only perform parameter updates in batches rather than after every time step. This ensures that our arm parameter estimates in the current batch are independent conditioned on the observations from previous batches. Importantly, this batching strategy does not change the convergence rates (and therefore regret), and has the added advantage of being far more computationally tractable.

Second, our robust multitask estimator requires two hyperparameters: the trimming hyperparameter ω and the LASSO regularization parameter λ (see Algorithm 1). We specify a *trimming path* for ω_t to dynamically trade off bias and variance over time, in order to control the convergence of our robust multitask estimators. Intuitively, we trim less for small t (when we have little data) to reduce variance at the cost of admitting "small" corruptions; as t increases (when we have collected more data), we trim more aggressively to eliminate even small corruptions that can bias our estimates. For λ_t, we use the path specified in Bastani and Bayati (2020).

Notation. In more detail, we split the time horizon T into batches $\bigcup_{m\geq 0}\mathcal{B}_m$. The initial batch $\mathcal{B}_0$ has size $q\log(T)$ for some tuning parameter q, and the following batches iteratively double in length (i.e., the m^{th} batch $\mathcal{B}_m$ has size $|\mathcal{B}_m| = 2^{m-1}|\mathcal{B}_0|$), which yields a total of M batches with

$$M = \left\lceil \log_2\left(\frac{T}{q\log(T)}\right)\right\rceil. \tag{9}$$

We define $\mathcal{B}_{\bar{m}} = \bigcup_{l=0}^{m}\mathcal{B}_l$ as all the batches up to batch m. We denote our robust multitask estimator (Algorithm 1) at instance j for arm k as

$$\widehat{\beta}_k^j(\mathcal{B}, \lambda, \omega).$$

The first argument indicates the training sample, i.e., all observations where we pulled arm k in batch $\mathcal{B}$; the remaining arguments are hyperparameters, i.e., the LASSO regularization parameter λ and the trimmed mean parameter ω.

Algorithm 2 Robust Multitask Bandit (RMBandit)

Inputs: Forced-sample estimator hyperparameters $\zeta_0, \eta_0, \{\lambda_{0,j}\}_{j\in[N]}$, initial all-sample estimator hyperparameters $\zeta_{1,0}, \eta_{1,0}, \{\lambda_{1,j,0}\}_{j\in[N]}$, batch size parameter q, time horizon T

Let $\omega_0 = \zeta_0 + \eta_0$, M as in (9), $\mathcal{B}_0 = [q\log(T)]$, $\mathcal{B}_m = \{t \in [T] \mid 2^{m-1}|\mathcal{B}_0| < t \leq 2^m|\mathcal{B}_0|\}$ for $m \in [M]$

for $t \in [T]$ **do**

  Observe an arrival at instance $j = Z_t$, where $Z_t \sim \mathrm{CG}(\mathbf{p})$

  Observe the corresponding context vector X_t for this arrival, where $X_t \sim \mathcal{P}_X^j$

  if $t \in \mathcal{B}_0$ **then**

    Pull arm $\pi_t = \Big((\sum_{r\in[t]} \mathbb{1}(Z_r = j) - 1) \bmod K\Big) + 1$

  else if $t \in \mathcal{B}_m$ **then**

    Let $\mathcal{K} = \Big\{k \in [K] \mid X_t^\top \widehat{\beta}_k^j(\mathcal{B}_0, \lambda_{0,j}, \omega_0) \geq \max_{i\in[K]} X_t^\top \widehat{\beta}_i^j(\mathcal{B}_0, \lambda_{0,j}, \omega_0) - \frac{h}{2}\Big\}$

    Pull arm $\pi_t = \arg\max_{k\in\mathcal{K}} X_t^\top \widehat{\beta}_k^j(\mathcal{B}_{\bar{m}-1}, \lambda_{1,j,\bar{m}-1}, \omega_{1,m-1})$

  end if

  Observe reward $Y_t = X_t^\top \beta_{\pi_t}^j + \epsilon_t$

  if $t = 2^m|\mathcal{B}_0|$ (i.e., when batch $m \in [M]$ ends) **then**

    Update $\zeta_{1,m} = \zeta_{1,0}$, $\eta_{1,m} = \eta_{1,0}\sqrt{\log(d\min_{i\in[N],|\mathcal{B}_m^i|>0}|\mathcal{B}_m^i|)}$, and $\omega_{1,m} = \zeta_{1,m} + \eta_{1,m}$

    Update $\lambda_{1,j,\bar{m}} = \lambda_{1,j,0}\sqrt{\frac{\log(d|\mathcal{B}_{\bar{m}}^j|)}{|\mathcal{B}_{\bar{m}}^j|}}$ for each $j \in [N]$

  end if

end for

Strategy. In our initial batch $\mathcal{B}_0$, we deterministically forced-sample each arm $k \in [K]$ of instance $j \in [N]$ when an individual is observed at instance j (i.e., when $Z_t = j$). At the end of this initial batch, we obtain a forced-sample estimator $\widehat{\beta}_k^j(\mathcal{B}_0, \lambda_{0,j}, \omega_0)$ for each $j \in [N]$ and $k \in [K]$; this forced-sample estimator remains fixed for the entire time horizon T. On the other hand, we also maintain an all-sample estimator $\widehat{\beta}_k^j(\mathcal{B}_{\bar{m}}, \lambda_{1,j,\bar{m}}, \omega_{1,m})$ for each $j \in [N]$ and $k \in [K]$; this estimator is periodically re-trained with updated hyperparameters at the end of each batch m and thereby fixed for the following batch $m+1$. One distinction of this estimator from Algorithm 1 is that the trimmed mean estimator in Step 1 is built upon only data from $\mathcal{B}_m$ such that arm parameter estimates are independent.

The algorithm is executed as follows. If $t \in \mathcal{B}_m$ and a new arrival is observed at instance j, we first use the forced-sample estimators to find the highest estimated reward achievable among the

K arms at instance j. These estimates allow us to identify a subset of arms $\mathcal{K} \subseteq K$ whose rewards are within $h/2$ of the estimated optimal reward. Then, within this set, we pull the arm $k \in \mathcal{K}$ that has the highest estimated reward according to the all-sample estimators. For a more detailed description of this approach, see Bastani and Bayati (2020).

4.2. Main Result: Regret Analysis of RMBandit

We bound the total regret across all N bandit instances (Theorem 2) as well as for an individual bandit instance (Corollary 3). Here, we consider the standard setting where each instance receives similar traffic, i.e., $p_j = \Theta(1/N)$ for all $j \in [N]$;[2] we discuss the data-poor regime in §4.4.

Theorem 2. *When $N = \Omega(\log(d)\log(T))$, the total cumulative expected regret of all instances up to time T implementing Algorithm 2 is*

$$\mathcal{O}\left(Kd(sN+d)\log(N)\log^2(dT/N)\right)$$

for appropriate choices of hyperparameters ζ_0, $\zeta_{1,0}$, η_0, $\eta_{1,0}$, $\lambda_{0,j}$, $\lambda_{1,j,0}$, and q, which we provide in Appendix C.

We provide a proof in Appendix C. Note that we make the mild assumption that d and N are both not too small; our approach is designed for problems where the feature dimension is nontrivial, and there are a reasonable number of instances to allow multitask learning. As we show in Section 5, we obtain improved empirical results in practice even for modest values of d and N.

Next, we consider the regret of RMBandit for a single instance j. To make a direct comparison to existing regret bounds, the expected time horizon[3] for instance j alone should be T. Since we expect $p_j = \Theta(1/N)$ fraction of the total arrivals (across all N instances) to be at instance j, we scale our total horizon as $\frac{T}{p_j} = \Theta(NT)$, which implies an expected time horizon of T for instance j.

Corollary 3. *Consider the same setting as in Theorem 2 with a time horizon of $T/p_j = \Theta(NT)$. The cumulative expected regret of a single instance j is*

$$\mathcal{O}\left(Kd\,(s+d/N)\log(N)\log^2(dT)\right).$$

It is useful to compare the bound above with that of a T-horizon linear contextual bandit instance j in the same setting, but which does *not* leverage knowledge sharing with other simultaneous bandit instances. Prior literature shows that such an instance would achieve regret that scales as $\mathcal{O}(d^2\log^{\frac{3}{2}}(d)\log(T))$ (Bastani and Bayati 2020). In contrast, our upper bound on the regret for

[2] In other words, if there are neighbouring data-poor instances, we do not include data from these instances in our parameter estimation for a standard instance, since they contribute too much variance to improve performance.

[3] Note that, given a fixed time horizon across all N instances, the time horizon (i.e., number of observations) for a single instance j is a random variable since the distribution of arrivals across instances ($\{Z_t\}_{t=1}^T$) is a random process.

instance j using RMBandit (Corollary 3) is smaller by a factor of d, but larger by a factor of $\log(T)$; this is a substantial improvement in high dimension (large d) and underscores the value of learning across bandit instances. We note that the extra factor of $\log(T)$ is likely an analytical limitation that arises because RMBandit leverages the LASSO estimator, e.g., the high-dimensional contextual bandit also attains a regret that scales as $\log^2(T)$ (Bastani and Bayati 2020).

4.3. Bandit Network Structure

Next, we consider the dependence of the regret on the *network structure* of the problem, when available. Recall from §2.1 that we consider a fully connected graph, where the nodes are instances and the edges $(i,j) \in [N] \times [N]$ have weights $s_{i,j} = \|\beta^j - \beta^i\|_0$ that indicate relative sparsity. Then, for any given problem instance j, we can optimize the subset of instances $\mathcal{Q}_j \subseteq [N]$ from which to transfer knowledge, where $\mathcal{Q}_j = \{i \in [N] \mid s_{i,j} \leq \tilde{s}\}$ is the subset of instances that are have an edge weight of no more than $\tilde{s}$ to instance j. Denote the corresponding number of instances $\tilde{N} = |\mathcal{Q}_j|$.

There is a tradeoff between choosing smaller $\tilde{s}$, which restricts the number of instances $\tilde{N}$ from which we share knowledge (resulting in lower bias but larger variance), and larger $\tilde{s}$ which yields larger $\tilde{N}$ (resulting in higher bias but smaller variance). Recall that we consider a natural power law scaling $\tilde{s} = \min(\tilde{N}^\alpha, d)$ for some $\alpha \geq 0$. In other words, as we increase the number of neighbouring instances we include, our sparsity parameter increases by some power law $\tilde{N}^\alpha$ until it eventually hits the maximum possible value d.

In this setting, we can compute the optimal choice $\tilde{N} = d^{\frac{1}{\alpha+1}}$, resulting in the following upper bound on the cumulative regret at instance j; note that it scales with the network density α.

COROLLARY 4. *Consider the same setting as in Corollary 3 with a time horizon of* $T/p_j = \Theta(NT)$. *Under the network structure given by Eq.* (1) *and when there are sufficient bandit instances* $N = \Omega(d^{\frac{1}{\alpha+1}})$, *the cumulative expected regret of a single instance* j *is*

$$\mathcal{O}\left(Kd^{\frac{2\alpha+1}{\alpha+1}} \log(d) \log^2(T)\right),$$

where we choose the optimal value of $\tilde{N} = \Theta(d^{\frac{1}{\alpha+1}})$.

We give a proof in Appendix C.6. As before, we obtain an improvement in the dependence on the context dimension d; in particular, the regret of RMBandit scales as $d^{\frac{2\alpha+1}{\alpha+1}}$, which is always smaller than the d^2 scaling of an independent bandit instance where we do not learn from other instances. The extent of this regret improvement scales with the network density α. When $\alpha \to 0$ (i.e., there are many instances with high similarity to the target instance), we eliminate a factor of d, which can be substantial in high dimension; when $\alpha \to \infty$ (i.e., there are essentially no instances with high similarity to the target instance), our regret improvement disappears.

4.4. Data-Poor Regime

Finally, we turn to the data-poor regime where we expect multitask learning to be most valuable (matching our previous result in §3.5). Again, we consider the case where the target instance j receives substantially fewer observations compared to at least one neighbouring instance $\ell \in [N]$; specifically, we consider $\frac{p_\ell}{p_j} = \Theta(d^2)$ and $\|\beta_k^\ell - \beta_k^j\|_0 \leq s$ for each $k \in [K]$. Once again, to make a direct comparison to existing regret bounds, the expected time horizon for instance j alone should be T. For a data-poor instance, we expect only $p_j = \Theta(\frac{1}{d^2 N})$ fraction of the total arrivals (across all N instances) to be at instance j, so we scale our total horizon as $\frac{T}{p_j} = \Theta(d^2 NT)$, which implies an expected time horizon of T for instance j.

Theorem 3. *Consider the same setting as in Corollary 3 with a time horizon of $T/p_j = \Theta(d^2 NT)$. Suppose there exists an instance $\ell \in [N]$ such that $\frac{p_\ell}{p_j} = \Theta(d^2)$ and $\|\beta_k^\ell - \beta_k^j\|_0 \leq s$ for each $k \in [K]$. Then, taking $\mathcal{Q}_j = \{j, \ell\}$, the cumulative expected regret of the data-poor instance j is*

$$\mathcal{O}\left(Ks^2 \log^2(dT)\right),$$

for appropriate choices of hyperparameters ζ_0, $\zeta_{1,0}$, η_0, $\eta_{1,0}$, $\lambda_{0,j}$, $\lambda_{1,j,0}$, and q, which we provide in Appendix D.

We provide a proof in Appendix D. Theorem 3 shows that RMBandit attains a regret bound that only scales *logarithmically* in the context dimension d — i.e., our multitask learning strategy *exponentially* reduces the regret for the target data-poor instance compared to running an independent bandit that does not leverage multitask learning.

It is worth noting that the regret of instance j scales *as if* the arm parameters $\{\beta_k^j\}$ are s-sparse (see, e.g., Bastani and Bayati 2020). However, our arm parameters are *not* sparse, i.e., $\|\beta_k^j\|_0 = d$. Rather, RMBandit achieves this scaling as a consequence of our multitask learning approach. When a neighbouring instance is data-rich, it provides a good estimate of the shared model $\beta^\dagger$, which allows us to substantially reduce the dimensionality of our estimation problem by focusing on learning only the bias term δ^j (which is s-sparse) rather than β^j (which is dense). This intuition aligns with the offline settings considered in Bastani (2021), Xu et al. (2021).

4.5. Proof Strategy

In this section, we sketch the proof of our main regret bound (Theorem 2). The proof builds on the regret analysis of LASSO Bandit (Bastani and Bayati 2020), but with the confidence intervals afforded by our robust multitask estimator (Theorem 1). As noted earlier, one key added challenge is the requirement that the OLS estimators $\{\widehat{\beta}_{\text{ind}}^j\}_{j \in [N]}$ across different instances be independent in order to invoke our robust multitask estimator; RMBandit achieves this goal using a batching strategy, as highlighted in Lemma 1.

Robust multitask estimator with random design:

We first introduce a variant of our Theorem 1 for the setting where the design matrices $\{\mathbf{X}^j\}_{j\in[N]}$ are random instead of fixed.

PROPOSITION 2. *Given $\zeta \in (0,1/2)$, the robust multitask estimator for any instance $j \in [N]$ from Algorithm 1 satisfies*

$$\mathbb{P}\left[\|\widehat{\beta}^j_{RM} - \beta^j\|_1 \geq \frac{6\lambda_j s}{\zeta\phi} + C_0 d\,(3\zeta + 4\eta) \max_{i\in[N]} \sqrt{\frac{\sigma_i^2}{n_i\phi}\log(\frac{3}{\eta})}\right]$$
$$\leq 3d\exp\left(-\frac{N\eta^2}{9}\right) + 2d\exp\left(-\frac{\lambda_j^2 n_j}{32\sigma_j^2 x_{\max}^2}\right) + \sum_{i\in[N]} \mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}^i) \leq \phi\right],$$

for any $\lambda_j > 0$ and for any positive η and a constant $C_0 > 2$ satisfying $\eta \leq 1/2 - 1/C_0 - \zeta$.

We give a proof in Appendix C.1.

Forced-sample estimator tail inequality:

Next, our algorithm uses a separate forced-sample estimator, which we can guarantee is close to the true parameter with high probability. Our next result provides tail bounds on the error of this estimator.

PROPOSITION 3. *When $N = \Omega(\log(d)\log(T))$, and the hyperparameters ζ_0, η_0, $\lambda_{0,j}$, q are as specified in Theorem 2, the forced-sample estimator $\widehat{\beta}^j_{k,0} = \widehat{\beta}^j_k(\mathcal{B}_0, \lambda_{0,j}, \omega_0)$ of any instance $j \in [N]$ and arm $k \in [K]$ computed by Algorithm 2 satisfies*

$$\mathbb{P}\left[\|\widehat{\beta}^j_{k,0} - \beta^j_k\|_1 \geq \frac{h}{4x_{\max}}\right] \leq \frac{10}{T}.$$

We give a proof in Appendix C.2. At a high level, this result follows directly from Proposition 2, since the forced samples are i.i.d. random variables.

All-sample estimator tail inequality:

Next, we provide a tail inequality for our all-sample estimator for all arms that belong in $\mathcal{K}^j_{opt}$ (it suffices to consider these arms since the forced-sample estimator is sufficiently accurate to exclude arms in $\mathcal{K}^j_{sub}$ from consideration). In contrast to the forced-sample estimator, which is based on $\mathcal{O}(\log(T))$ samples, the all-sample estimator is based on $\mathcal{O}(T)$ samples (since we will show that all optimal arms receive a linear number of samples with high probability). Therefore, the all-sample estimator has smaller error than the forced-sample estimator (the tradeoff is that these samples are adaptively assigned to arms, so they may be collected from biased regions of the covariate space; thus, the i.i.d. samples generated when using the forced-sample estimator are needed to ensure that the all-sample estimator converges). In particular, define the following event, which says that the forced-sample estimators have small error:

$$\mathcal{A} = \left\{\|\widehat{\beta}^j_k(\mathcal{B}_0, \lambda_{0,j}, \omega_0) - \beta^j_k\|_1 \leq \frac{h}{4x_{\max}}, \forall j \in [N], k \in [K]\right\}. \tag{10}$$

This event holds with high probability by Proposition 3. Our next result shows that our all-sample estimator satisfies the following tail inequality conditional on the event $\mathcal{A}$.

PROPOSITION 4. *When $\mathcal{A}$ holds, $T \geq KN$, $N = \Omega(\log(d)\log(T))$, and $\zeta_{1,0}$, $\eta_{1,0}$, $\lambda_{1,j,0}$ take the values in Theorem 2, the all-sample estimator $\widehat{\beta}^j_{k,\bar{m}} = \widehat{\beta}^j_k(\mathcal{B}_{\bar{m}}, \lambda_{1,j,\bar{m}}, \omega_{1,m})$ of any instance $j \in [N]$ and optimal arm $k \in \mathcal{K}^j_{opt}$ computed by Algorithm 2 satisfies*

$$\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}} - \beta^j_k\|_1 \geq C_1\sqrt{\frac{sd\log(dp_j|\mathcal{B}_{\bar{m}}|)}{p_j|\mathcal{B}_{\bar{m}}|}} + C_2\sqrt{\frac{sd\log(\rho N)}{p_j|\mathcal{B}_m|}} + C_3 d\sqrt{\frac{\log(dp_j|\mathcal{B}_m|)\log(\rho N)}{Np_j|\mathcal{B}_m|}}\,\middle|\,\mathcal{A}\right]$$
$$\leq \frac{6}{\min_{i\in\mathcal{W}_k} p_i|\mathcal{B}_m|} + \frac{8}{p_j|\mathcal{B}_{\bar{m}}|} + \frac{6}{TN} + \sum_{i\in\mathcal{W}_k} 2d\exp\left(-\frac{p_* p_i \psi|\mathcal{B}_m|}{32dx^2_{\max}}\right) + \sum_{i\in\mathcal{W}_k} 8\exp\left(-\frac{p_* p_i|\mathcal{B}_m|}{20}\right),$$

where the constants C_1, C_2 and C_3 are provided in Appendix C.3.

Note that compared to Proposition 2, this result has two extra terms in the probability on the right-hand side of the inequality. These terms account for the event that the number of samples received at each optimal arm of each bandit instance scales proportionally with $|\mathcal{B}_m|$.

We give a proof of Proposition 4 in Appendix C.3. As discussed above, because our all-sample estimators are constructed using all available samples, they may not be independent across instances; however, the trimmed mean estimator in Step 1 of our robust multitask estimator requires that the inputs $\{\widehat{\beta}^j_k\}_{j\in[N]}$ for each arm $k \in [K]$ are independent. By using a batching strategy, we ensure that the samples used to train $\widehat{\beta}^j_k$ are independent (which implies that the $\{\widehat{\beta}^j_k\}_{j\in[N]}$ are also independent). In particular, we have the following lemma:

LEMMA 1. *The samples assigned to arm k in batch $\mathcal{B}_m$ (for any $m \geq 1$) are independent across bandit instances conditioned on $\mathcal{F}_{\bar{m-1}} = \sigma(\{X_t, Z_t, Y_t\}_{t\in\mathcal{B}_{\bar{m-1}}})$, the σ-algebra generated by the samples in $\mathcal{B}_{\bar{m-1}}$.*

We give a proof in Appendix C.3. Given this, Proposition 4 follows by applying Proposition 2.

Regret bound: Finally, we describe how the above results enable us to prove Theorem 2. For this regret analysis, we group time steps $t \in [T]$ into three possible cases, and bound the regret across time steps in each case separately:

(I). Time horizon $T \leq KN$, or forced-sample batch ($t \in \mathcal{B}_0$) or the first batch using all-sample estimator ($t \in \mathcal{B}_1$),

(II). All the remaining batches ($t \in \mathcal{B}_m$ for $m > 1$) such that $\mathcal{A}$ does not hold and time horizon $T \geq KN$,

(III). All the remaining batches ($t \in \mathcal{B}_m$ for $m > 1$) such that $\mathcal{A}$ holds and time horizon $T \geq KN$.

For Case I, note that the sizes of the first two batches $\mathcal{B}_0$ and $\mathcal{B}_1$ are both $q\log(T)$ and q scales as $\tilde{\mathcal{O}}(Kd(sN+d))$. In the worst case, the regret for one time step is at most $2bx_{\max}$, so the regret in this case is bounded. For Case II, we have shown that the event $\mathcal{A}$ holds with high probability. Similar to before, in the worst case, the regret for one time step is at most $2bx_{\max}$, so the regret in this case is bounded with high probability. Finally, for Case III when $\mathcal{A}$ holds, Proposition 4 guarantees that the all-sample estimator has small error with high probability, again ensuring that the regret is bounded with high probability. Details are provided in Appendix C.4.

5. Experiments

We now illustrate the value of our approach on synthetic and real datasets. For the latter case, we focus on two well-studied applications of bandit learning: patient risk prediction for personalized interventions, and demand prediction for dynamic pricing. Furthermore, let the average arrival probability across neighbouring instances be $\bar{p} = \frac{1}{N-1}\sum_{i\neq j} p_i$. When possible, matching our theoretical results, we consider two cases for a single instance j: (i) standard with similar arrival probability $p_j \approx \bar{p}$, and (ii) data-poor with arrival probability $p_j \ll \bar{p}$.

In all cases, we simulate the following linear contextual bandit algorithms:

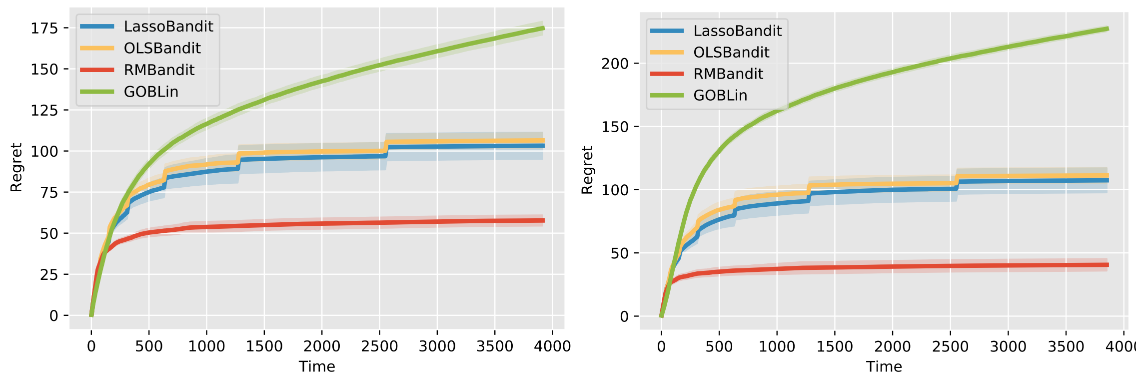
1. LASSO and OLS Bandit (no transfer learning)
2. GOBLin (transfer learning via Laplacian similarity and ℓ_2-regularization)
3. RMBandit (transfer learning via our robust multitask estimator)

The two approaches in the first bullet point operate N independent bandit instances, either via ordinary linear regression (Goldenshluger and Zeevi 2013) or LASSO (Bastani and Bayati 2020). The GOBLin algorithm is a state-of-the-art multitask bandit algorithm that uses a Laplacian matrix and ridge regression to jointly estimate the N parameters, thereby ℓ_2-regularizing both the parameters and their pairwise differences (Cesa-Bianchi et al. 2013). It builds on the OFUL algorithm (Abbasi-Yadkori et al. 2011), which leverages UCB for linear contextual bandits.

Remark 1. There are a few alternative multitask bandit algorithms that are not applicable to our experimental setup. For instance, Soare et al. (2014) and Gentile et al. (2014) simply pool data together across similar instances and ignore instance-specific heterogeneity, which would result in linear regret in our setting. There are also Bayesian meta-learning algorithms (Cella et al. 2020, Bastani et al. 2021c, Kveton et al. 2021), but these require instances to be observed sequentially (rather than simultaneously) in order to construct a prior across instances. Furthermore, Bayesian approaches are often computationally intractable in high dimension (Djolonga et al. 2013).

5.1. Synthetic

Figure 2 shows the expected cumulative regret over time for a single (a) standard and (b) data-poor contextual linear bandit instance. Appendix F.1 provides details on the underlying parameters.



(a) Standard (b) Data-poor

Figure 2 Lines depict the cumulative regret averaged over 15 trials (shaded regions depict the corresponding 95% confidence intervals) of a single linear contextual bandit that receives (a) similar traffic as neighbouring instances, or (b) significantly less traffic than neighbouring instances.

Since our unknown parameters satisfy our assumption of sparse differences, we expect RMBandit to outperform alternative algorithms that do not explicitly leverage this structure. Indeed, RMBandit significantly outperforms OLS and LASSO Bandit, which do not leverage transfer learning. Moreover, since the underlying arm parameters are not sparse, we see no noticeable difference between OLS and LASSO Bandit. Perhaps more surprisingly, we observe that the GOBLin algorithm *increases* cumulative regret compared to not performing transfer learning, which is likely due to two reasons. First, their transfer learning methodology is predicated on the assumption that the arm parameters across instances are close in ℓ_2 norm; this assumption is not met by our synthetic data, and more importantly appears unwarranted in the two real datasets we study in the next two subsections, leading to negative transfer learning. Second, their algorithm builds on the UCB algorithm, which is known to over-explore compared to the OLS Bandit strategy (see, e.g., Russo et al. 2017, Bastani et al. 2021a); it remains an interesting future direction of research to adapt multitask learning methods to other bandit algorithms such as ϵ-greedy or Thompson Sampling. Finally, we note that the cumulative regret in the data-poor instance is lower than the corresponding cumulative regret in the standard instance for RMBandit; in contrast, the data-poor instance suffers similar or worse cumulative regret for the baseline algorithms.

It is also worth noting that RMBandit is significantly more computationally efficient, largely due to the batching strategy which requires only a small number of model updates. To run 15 trials in the standard setting, LASSO and OLS Bandit each took approximately 5-6 minutes, GOBLin took over 11 hours, while our algorithm took less than a minute to run.[4]

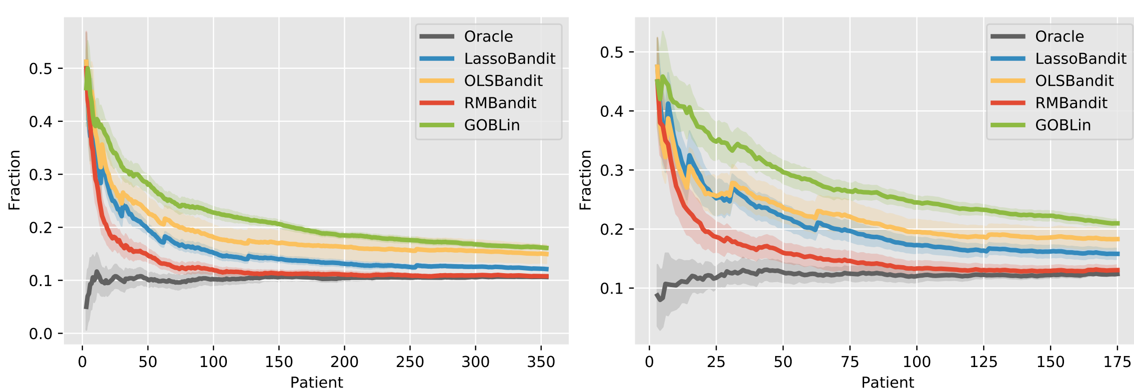
The next two subsections examine real datasets, which may not satisfy our assumptions.

[4] All algorithms were run on the same desktop, with an Intel Core i7-7700K 4.20GHz processor and 16 GB memory.

5.2. Risk Prediction in Health Data

Diabetes is a leading cause of severe health complications such as cardiovascular disease, stroke, and chronic kidney disease (Ismail et al. 2021). Thus, there is significant interest in leveraging machine learning for early detection of (Type II) diabetes, in order to improve treatment outcomes (Zhang et al. 2020). However, significant evidence shows that machine learning models trained on one health system can perform poorly on a different health system (Quiñonero-Candela et al. 2008, Subbaswamy and Saria 2020); this can be due to dataset shifts such as changes in patient demographics, disease prevalence, measurement timing, equipment, and treatment patterns. Thus, it is important to train provider-specific risk models; yet, smaller providers may benefit from additionally leveraging information across providers due to their relatively small patient cohorts.

In this experiment, we use electronic medical record data across $N = 13$ healthcare providers to learn a good diabetes risk model for a single provider. After basic preprocessing, we have approximately 80 patient-specific features constructed from information available before the most recent visit (e.g., past diagnoses, procedures and medications); our outcome is an indicator variable for whether the patient was diagnosed with diabetes during the most recent visit. We aim to learn the best linear classifier online as patient observations accrue, and evaluate different methods based on the classification accuracy over time. We consider a mid-sized provider with 355 unique patients observed during the sample period, as well as a data-poor provider with only 176 unique patients. We additionally fit a linear oracle, which leverages all observed data from the provider in hindsight using a leave-one-out approach, representing the best achievable performance within a linear model family. Figure 3 shows the fraction of incorrect classifications made over time for the (a) standard and (b) data-poor providers. Appendix F.2 provides additional details on the setup.



(a) Standard (355 patients)

(b) Data-Poor (176 patients)

Figure 3 Lines depict the fraction of incorrect classifications averaged over 20 trials (shaded regions depict the corresponding 95% confidence intervals) of a linear contextual bandit operated by (a) a mid-sized provider with 355 unique patients, or (b) a data-poor provider with 176 unique patients.

Once again, we observe that RMBandit performs favorably compared to the other baseline algorithms, and converges to the oracle's classification accuracy much faster. This is especially the case in the data-poor instance, as suggested by the theory.

5.3. Demand Prediction in Retail Data

Contextual bandit algorithms can also naturally be extended to solve dynamic pricing problems with unknown demand (Besbes and Zeevi 2009). We consider such a demand forecasting and price optimization task for food distributors; to this end, we use a publicly available dataset of orders from a meal delivery company.[5]

In this experiment, we use data across $N = 7$ fulfillment centers, serving between six to seven thousand orders each during the sample period. Features include the category and cuisine pertaining to the order, as well as associated promotions. The decision variable is the (continuous) price for the order; rather than arm parameters, there is a single set of unknown parameters (per instance) that aims to predict demand/revenue as a function of price and the observed features. Following the approach of Ban and Keskin (2021), we model the price elasticity of demand as a linear function of the observed features:

$$Y_t = X_t^\top \beta_0^j + p_t \cdot (X_t^\top \beta_1^j) + \epsilon_t.$$

Here, $\{\beta_0^j, \beta_1^j\}$ are the unknown parameters corresponding to instance j; conditioned on an arrival with context X_t at instance $Z_t = j$, Y_t is the observed revenue for the chosen price p_t and noise ϵ_t. Regret is measured with respect to an oracle that knows $\{\beta_0^j, \beta_1^j\}_{j=1}^N$.

We straightforwardly extend RMBandit and the other baseline bandit algorithms to the dynamic pricing setting using a batched explore-then-commit strategy employed by Ban and Keskin (2021). Figure 4 shows the cumulative regret of our algorithm RMX (our dynamic pricing analog of RMBandit) compared to other benchmarks including ILQX and ILSX (the LASSO and OLS based pricing algorithms introduced in Ban and Keskin 2021) and GOLX (our dynamic pricing analog of GOBLin). Note that all 7 fulfillment centers are similarly sized (the smallest has 6072 orders and the largest has 7046 orders); thus, we do not study the data-poor setting in this experiment. Appendix F.3 provides details on the setup, as well as pseudocode for our dynamic pricing algorithms.

Once again, we observe that RMX (the dynamic pricing analog of RMBandit) performs favorably compared to the dynamic pricing analogs of the other baseline bandit algorithms. Thus, our insights on multitask learning carry over to the dynamic pricing context.

[5] `https://datahack.analyticsvidhya.com/contest/genpact-machine-learning-hackathon-1/`

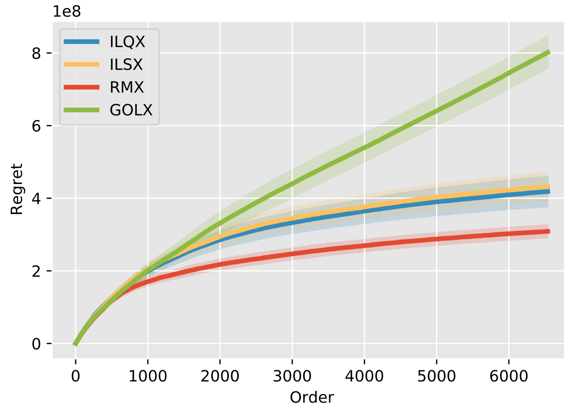


Figure 4 **Lines depict the cumulative regret averaged over 40 trials (shaded regions depict the corresponding 95% confidence intervals) of a linear contextual bandit operated by a single fulfillment center.**

6. Discussion and Conclusions

Decision-makers frequently want to learn heterogeneous treatment effects across many simultaneous experiments. Examples range from learning patient risk across hospitals for personalized interventions (Bastani 2021, Mullainathan and Obermeyer 2017), learning drug effectiveness across combination therapies for clinical trial decisions (Bertsimas et al. 2016), learning COVID-19 risk across travelers for targeting tests (Bastani et al. 2021b), and learning demand across stores for promotion targeting (Baardman et al. 2020, Cohen and Perakis 2018) or dynamic pricing (Bastani et al. 2021c). We propose a novel robust multitask estimator that improves the efficacy of downstream decisions by learning better predictive models with lower sample complexity. To the best of our knowledge, our work proposes the first combination of robust statistics (to learn across similar instances) and LASSO regression (to debias the results) to yield improved bounds for multitask learning. In the online learning setting, these problems translate to running simultaneous contextual bandit algorithms. To this end, we propose the RMBandit algorithm to effectively navigate the exploration-exploitation tradeoff across bandit instances, thereby improving regret bounds in the context dimension d.

We highlight several features of our proposed approach that make it a particularly attractive solution. First, it is well known that data limitations result in worse model performance, which in turn can imply *unfair* decisions, e.g., in healthcare, such biases disproportionately affect protected groups or minorities due to limited representative data (Rajkomar et al. 2018). A natural approach to alleviating unfairness is to improve the performance of our models for data-poor instances (see, e.g., discussion in Hardt et al. 2016). We show that multitask learning can be especially valuable in such settings — our approach leverages data from data-rich instances to provide an exponential improvement in performance for data-poor instances. Thus, we provide one additional tool (among others) for improving fairness in decision-making.

Second, privacy and regulatory constraints prevent granular data sharing in many applications. A growing literature on *federated learning* studies training statistical models over siloed datasets, while keeping data localized (Li et al. 2020b). While our focus is on multitask learning, our approach satisfies the constraints of federated learning, since we only require sharing aggregate statistics (in this case, OLS regression parameters) across instances. All model training is performed locally at the instance-level and does not require any raw data from other instances.

Third, practical deployment of bandits often precludes real-time updates to the model. For instance, many individuals may appear for service simultaneously (Schwartz et al. 2017) and there may be operational constraints or concerns over model reliability (Bastani et al. 2021b). Our RMBandit algorithm employs a batching strategy that only requires a logarithmic number (in the time horizon T) of model updates, while preserving convergence rates (and therefore regret). Furthermore, it has the added advantage of being far more computationally tractable.

References

Abbasi-Yadkori, Yasin, Dávid Pál, Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24**.

Alon, Noga, Joel H Spencer. 2004. *The probabilistic method*. John Wiley & Sons.

Baardman, Lennart, Setareh Borjian Boroujeni, Tamar Cohen-Hillel, Kiran Panchamgam, Georgia Perakis. 2020. Detecting customer trends for optimal promotion targeting. *Manufacturing & Service Operations Management* .

Ban, Gah-Yi, N Bora Keskin. 2021. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science* **67**(9) 5549–5568.

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

Bastani, Hamsa, Mohsen Bayati. 2020. Online decision making with high-dimensional covariates. *Operations Research* **68**(1) 276–294.

Bastani, Hamsa, Mohsen Bayati, Khashayar Khosravi. 2021a. Mostly exploration-free algorithms for contextual bandits. *Management Science* **67**(3) 1329–1349.

Bastani, Hamsa, Kimon Drakopoulos, Vishal Gupta, Ioannis Vlachogiannis, Christos Hadjicristodoulou, Pagona Lagiou, Gkikas Magiorkinis, Dimitrios Paraskevis, Sotirios Tsiodras. 2021b. Efficient and targeted covid-19 border testing via reinforcement learning. *Nature* **599**(7883) 108–113.

Bastani, Hamsa, David Simchi-Levi, Ruihao Zhu. 2021c. Meta dynamic pricing: Transfer learning across experiments. *Management Science* .

Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* **79**(1) 151–175.

Bertsimas, Dimitris, Allison O'Hair, Stephen Relyea, John Silberholz. 2016. An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science* **62**(5) 1511–1531.

Besbes, Omar, Assaf Zeevi. 2009. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* **57**(6) 1407–1420.

Bickel, Peter, Ya'acov Ritov, Alexandre Tsybakov. 2009. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 1705–1732.

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

Candes, Emmanuel, Terence Tao. 2007. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* 2313–2351.

Caruana, Rich. 1997. Multitask learning. *Machine learning* **28**(1) 41–75.

Cella, Leonardo, Alessandro Lazaric, Massimiliano Pontil. 2020. Meta-learning with stochastic linear bandits. *International Conference on Machine Learning*. PMLR, 1360–1370.

Cesa-Bianchi, Nicolo, Claudio Gentile, Giovanni Zappella. 2013. A gang of bandits. *arXiv preprint arXiv:1306.0811* .

Chen, Scott S, David L Donoho, Michael A Saunders. 1995. Atomic decomposition by basis pursuit .

Cohen, Maxime, Georgia Perakis. 2018. Promotion optimization in retail. *Available at SSRN 3194640* .

Crammer, Koby, Michael Kearns, Jennifer Wortman. 2008. Learning from multiple sources. *Journal of Machine Learning Research* **9**(8).

Deshmukh, Aniket Anand, Urun Dogan, Clayton Scott. 2017. Multi-task learning for contextual bandits. *arXiv preprint arXiv:1705.08618* .

Djolonga, Josip, Andreas Krause, Volkan Cevher. 2013. High-dimensional gaussian process bandits. *Advances in neural information processing systems* **26**.

Dobriban, Edgar, Yue Sheng. 2021. Distributed linear regression by averaging. *The Annals of Statistics* **49**(2) 918–943.

Evgeniou, Theodoros, Charles A Micchelli, Massimiliano Pontil, John Shawe-Taylor. 2005. Learning multiple tasks with kernel methods. *Journal of machine learning research* **6**(4).

Evgeniou, Theodoros, Massimiliano Pontil. 2004. Regularized multi–task learning. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 109–117.

Finn, Chelsea, Pieter Abbeel, Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*. PMLR, 1126–1135.

Gentile, Claudio, Shuai Li, Giovanni Zappella. 2014. Online clustering of bandits. *International Conference on Machine Learning*. PMLR, 757–765.

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

Gupta, Vishal, Nathan Kallus. 2021. Data pooling in stochastic optimization. *Management Science* .

Hanneke, Steve, Samory Kpotufe. 2020. A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785* .

Hardt, Moritz, Eric Price, Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* **29** 3315–3323.

Ismail, Leila, Huned Materwala, Juma Al Kaabi. 2021. Association of risk factors with type 2 diabetes: A systematic review. *Computational and Structural Biotechnology Journal* **19** 1759–1785.

Konstantinov, Nikola, Christoph Lampert. 2019. Robust learning from untrusted sources. *International Conference on Machine Learning*. PMLR, 3488–3498.

Kveton, Branislav, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, Csaba Szepesvari. 2021. Meta-thompson sampling. *arXiv preprint arXiv:2102.06129* .

Lai, Tze Leung, Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* **6**(1) 4–22.

Li, Jerry. 2019. Lecture 1: Introduction to robustness. `https://jerryzli.github.io/robust-ml-fall19/lec1.pdf`. [Online; accessed 13-April-2021].

Li, Lihong, Wei Chu, John Langford, Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web*. 661–670.

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

Li, Tian, Anit Kumar Sahu, Ameet Talwalkar, Virginia Smith. 2020b. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine* **37**(3) 50–60.

Lounici, Karim, Massimiliano Pontil, Alexandre B Tsybakov, Sara Van De Geer. 2009. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468* .

Lounici, Karim, Massimiliano Pontil, Sara Van De Geer, Alexandre B Tsybakov. 2011. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics* **39**(4) 2164–2204.

Lugosi, Gabor, Shahar Mendelson. 2021. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics* **49**(1) 393–410.

Mullainathan, Sendhil, Ziad Obermeyer. 2017. Does machine learning automate moral hazard and error? *American Economic Review* **107**(5) 476–80.

Pan, Sinno Jialin, Qiang Yang, et al. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**(10) 1345–1359.

Qiang, Sheng, Mohsen Bayati. 2016. Dynamic pricing with demand covariates. *Available at SSRN 2765257* .

Quiñonero-Candela, Joaquin, Masashi Sugiyama, Anton Schwaighofer, Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

Raina, Rajat, Andrew Y Ng, Daphne Koller. 2006. Constructing informative priors using transfer learning. *Proceedings of the 23rd international conference on Machine learning*. ACM, 713–720.

Rajkomar, Alvin, Michaela Hardt, Michael D Howell, Greg Corrado, Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* **169**(12) 866–872.

Rigollet, Phillippe, Jan-Christian Hütter. 2015. High dimensional statistics. *Lecture notes for course 18S997* **813** 814.

Rousseeuw, Peter J. 1991. Tutorial to robust statistics. *Journal of chemometrics* **5**(1) 1–20.

Russo, Daniel, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen. 2017. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038* .

Schwartz, Eric M, Eric T Bradlow, Peter S Fader. 2017. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science* **36**(4) 500–522.

Soare, Marta, Ouais Alsharif, Alessandro Lazaric, Joelle Pineau. 2014. Multi-task linear bandits. *NIPS2014 Workshop on Transfer and Multi-task Learning: Theory meets Practice*.

Subbaswamy, Adarsh, Suchi Saria. 2020. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics* **21**(2) 345–352.

Tewari, Ambuj, Susan A Murphy. 2017. From ads to interventions: Contextual bandits in mobile health. *Mobile Health*. Springer, 495–517.

Tian, Ye, Yang Feng. 2021. Transfer learning under high-dimensional generalized linear models. *arXiv preprint arXiv:2105.14328* .

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

Tropp, Joel A. 2015. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571* .

van Herpen, Erica, Erjen van Nierop, Laurens Sloot. 2012. The relationship between in-store marketing and observed sales for organic versus fair trade products. *Marketing Letters* **23**(1) 293–308.

Wainwright, Martin J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

Xu, Kan, Xuanyi Zhao, Hamsa Bastani, Osbert Bastani. 2021. Group-sparse matrix factorization for transfer learning of word embeddings. *International Conference on Machine Learning*. PMLR, 11603–11612.

Yin, Dong, Yudong Chen, Ramchandran Kannan, Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. *International Conference on Machine Learning*. PMLR, 5650–5659.

Yuan, Hao, Qi Luo, Cong Shi. 2021. Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science* .

Zhang, Liying, Yikang Wang, Miaomiao Niu, Chongjian Wang, Zhenfei Wang. 2020. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural chinese population: The henan rural cohort study. *Scientific reports* **10**(1) 1–10.

Appendix A: Error Bound for Robust Multitask Estimator

A.1. Trimmed Mean Estimator

Here we prove the tail inequality for our trimmed mean estimator in Proposition 1, which is the key to estimate the global shared parameter $\beta^\dagger$.

Proof of Proposition 1 Recall that the indices of the corrupted samples are denoted by $\mathcal{J} \subseteq [N]$ (so the rest are $\mathcal{J}^c = [N] \setminus \mathcal{J}$). By assumption, $\{Z_j\}_{j\in\mathcal{J}^c}$ are independent and σ_j-subgaussian with mean μ respectively, and $|\mathcal{J}| < N\zeta$ with $\zeta < 1/2$. By Hoeffding's inequality, any uncorrupted sample $j \in \mathcal{J}^c$ satisfies

$$\mathbb{P}\left[|Z_j - \mu| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma_j^2}\right)$$

for any $t > 0$. Letting $t = \sqrt{2\sigma_j^2 \log(\frac{3}{\eta})}$, it follows that

$$Z_j \notin \left[\mu - \sqrt{2\sigma_j^2\log(\frac{3}{\eta})},\ \mu + \sqrt{2\sigma_j^2\log(\frac{3}{\eta})}\right]$$

with a probability of at most $2\eta/3$. Then, again by Hoeffding's inequality, we have

$$\mathbb{P}\left[\sum_{j\in\mathcal{J}^c} \mathbb{1}\left(Z_j \notin I\right) \geq t\right] \leq \exp\left(-\frac{2(t - \sum_{j\in\mathcal{J}^c} p_j)^2}{|\mathcal{J}^c|}\right),$$

where $p_j = \mathbb{P}[Z_j \notin I] \leq 2\eta/3$ and $I = \left[\mu - \max_{j\in\mathcal{J}^c}\sqrt{2\sigma_j^2\log(\frac{3}{\eta})}, \mu + \max_{j\in\mathcal{J}^c}\sqrt{2\sigma_j^2\log(\frac{3}{\eta})}\right]$. Taking $t = \eta|\mathcal{J}^c|$, we have

$$\mathbb{P}\left[\sum_{j\in\mathcal{J}^c} \mathbb{1}\left(Z_j \notin I\right) \geq \eta|\mathcal{J}^c|\right] \leq \exp\left(-\frac{2\eta^2|\mathcal{J}^c|}{9}\right);$$

in other words, with a high probability, at most η fraction of $\mathcal{J}^c$ are outside a neighborhood I of the mean μ. As a consequence, on the event

$$V = \left\{\sum_{j\in\mathcal{J}^c} \mathbb{1}\left(Z_j \notin I\right) \leq \eta|\mathcal{J}^c|\right\},$$

at most $\zeta + \eta$ fraction of the N samples are outside the interval I (recall that ζN samples are corrupted). As a subgaussian distribution might not be symmetric, by trimming the upper and lower ω-quantile of samples, the remaining ones are guaranteed to fall into I.

Suppose the event V holds. Let $\{Z_{j_\iota}\}_{\iota=N\omega+1}^{N(1-\omega)}$ denote the samples after trimming and $\mathcal{T} = \{j_\iota\}_{\iota=N\omega+1}^{N(1-\omega)}$ the corresponding index set. Let $\mathcal{U} = \{j \in [N] \mid Z_j \in I\}$ denote the set of samples that lie in I. Since $\mathcal{T} \subseteq \mathcal{U}$ from our argument above, we have

$$\left|\sum_{j\in\mathcal{T}}(Z_j - \mu)\right| = \left|\sum_{j\in\mathcal{T}\cap\mathcal{U}}(Z_j - \mu)\right| \leq \left|\sum_{j\in\mathcal{T}\cap\mathcal{U}\cap\mathcal{J}^c}(Z_j - \mu)\right| + \left|\sum_{i\in\mathcal{T}\cap\mathcal{U}\cap\mathcal{J}}(Z_j - \mu)\right|. \tag{11}$$

The first term on the RHS of (11) can be upper bounded by

$$\left|\sum_{j\in\mathcal{T}\cap\mathcal{U}\cap\mathcal{J}^c}(Z_j - \mu)\right| \leq \left|\sum_{j\in\mathcal{T}^c\cap\mathcal{U}\cap\mathcal{J}^c}(Z_j - \mu)\right| + \left|\sum_{j\in\mathcal{U}\cap\mathcal{J}^c}(Z_j - \mu)\right|.$$

As we remove $2(\zeta + \eta)$ fraction of the samples, we have

$$\left|\sum_{j\in\mathcal{T}^c\cap\mathcal{U}\cap\mathcal{J}^c}(Z_j - \mu)\right| \leq 2(\zeta + \eta)N\max_{j\in\mathcal{J}^c}\sqrt{2\sigma_j^2\log(\frac{3}{\eta})}. \tag{12}$$

Since those samples in $\mathcal{J}^c$ that lie inside the interval I are independent and bounded, applying Hoeffding's inequality gives

$$\mathbb{P}\left[\left|\frac{1}{|\mathcal{J}^c\cap\mathcal{U}|}\sum_{j\in\mathcal{J}^c\cap\mathcal{U}}(Z_j-\mathbb{E}[Z_j\mid\mathcal{J}^c\cap\mathcal{U}])\right|\geq\chi\cdot\max_{j\in\mathcal{J}^c}\sqrt{\sigma_j^2\log(\frac{3}{\eta})}\right]\leq 2\exp\left(-\frac{|\mathcal{J}^c\cap\mathcal{U}|\chi^2}{4}\right),\tag{13}$$

for any $\chi>0$. The truncation on these samples introduces a bias of at most

$$|\mathbb{E}[Z_j\mid\mathcal{J}^c\cap\mathcal{U}]-\mu|=\left|\frac{\mathbb{E}[(Z_j-\mu)\mathbb{1}(Z_j\notin I)\mid\mathcal{J}^c]}{\mathbb{P}(Z_j\in I\mid\mathcal{J}^c)}\right|\leq\frac{\mathbb{E}[|Z_j-\mu|^k\mid\mathcal{J}^c]^{1/k}\mathbb{P}[Z_j\notin I\mid\mathcal{J}^c]^{1/q}}{\mathbb{P}(Z_j\in I\mid\mathcal{J}^c)},\tag{14}$$

where the last inequality uses Hölder's inequality and k,q are such that $1/k+1/q=1$. Recall that $\mathbb{P}[Z_j\notin I\mid\mathcal{J}^c]\leq 2\eta/3$. In addition, $\mathbb{E}[|Z_j-\mu|^k\mid\mathcal{J}^c]^{1/k}\leq e^{1/e}\sigma_j\sqrt{k}$ for $k\geq 2$ by the property of subgaussian (Rigollet and Hütter 2015). Taking $k=\log(\frac{3}{2\eta})$, inequality (14) gives

$$|\mathbb{E}[Z_j\mid\mathcal{J}^c\cap U]-\mu|\leq\frac{8\sigma_j\eta\sqrt{\log(\frac{3}{2\eta})}}{3-2\eta}.$$

Then, the high probability bound in (13) implies

$$\mathbb{P}\left[|\sum_{j\in\mathcal{J}^c\cap\mathcal{U}}(Z_j-\mu)|\geq|\mathcal{J}^c\cap\mathcal{U}|\left(\chi\cdot\max_{j\in\mathcal{J}^c}\sqrt{\sigma_j^2\log(\frac{3}{\eta})}+\max_{j\in\mathcal{J}^c}\frac{8\sigma_j\eta\sqrt{\log(\frac{3}{2\eta})}}{3-2\eta}\right)\right]\leq 2\exp\left(-\frac{|\mathcal{J}^c\cap\mathcal{U}|\chi^2}{4}\right).$$

Further setting $\chi=\eta$ and by our assumption that $\eta<1/2$, we have

$$\mathbb{P}\left[\left|\sum_{j\in\mathcal{J}^c\cap\mathcal{U}}(Z_j-\mu)\right|\geq 5|\mathcal{J}^c\cap\mathcal{U}|\eta\max_{j\in\mathcal{J}^c}\sqrt{\sigma_j^2\log(\frac{3}{\eta})}\right]\leq 2\exp\left(-\frac{|\mathcal{J}^c\cap\mathcal{U}|\eta^2}{4}\right).\tag{15}$$

Combining (12) and (15), it holds with a high probability that

$$\left|\sum_{j\in\mathcal{T}\cap\mathcal{U}\cap\mathcal{J}^c}(Z_j-\mu)\right|\leq\left(2\sqrt{2}(\zeta+\eta)N+5\eta|\mathcal{J}^c\cap\mathcal{U}|\right)\max_{j\in\mathcal{J}^c}\sqrt{\sigma_j^2\log(\frac{3}{\eta})}$$

Next, the second term on the RHS of (11) has

$$\left|\sum_{j\in\mathcal{T}\cap\mathcal{U}\cap\mathcal{J}}(Z_j-\mu)\right|\leq\zeta N\max_{j\in\mathcal{J}^c}\sqrt{2\sigma_j^2\log(\frac{3}{\eta})}.$$

Thus, we can write

$$\begin{aligned}\left|\frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}Z_j-\mu\right|&\leq\frac{\max_{j\in\mathcal{J}^c}\sigma_j}{(1-2(\zeta+\eta))N}\left(\sqrt{2}(3\zeta+2\eta)N+5\eta|\mathcal{J}^c\cap\mathcal{U}|\right)\sqrt{\log(\frac{3}{\eta})}\\&\leq C_0\max_{j\in\mathcal{J}^c}\sigma_j\,(3\zeta+4\eta)\sqrt{\log(\frac{3}{\eta})}\end{aligned}$$

with a high probability, where we use $|\mathcal{T}|=(1-2(\zeta+\eta))N$, $|\mathcal{J}^c\cap\mathcal{U}|\leq N$ and $\eta\leq 1/2-1/C_0-\zeta$. Since $|\mathcal{J}^c\cap\mathcal{U}|\geq(1-\eta)|\mathcal{J}^c|$ on the event V, we have

$$\mathbb{P}\left[\left|\frac{1}{|\mathcal{T}|}\sum_{j\in\mathcal{T}}Z_j-\mu\right|\geq C_0\max_{j\in\mathcal{J}^c}\sigma_j\,(3\zeta+4\eta)\sqrt{\log(\frac{3}{\eta})}\right]\leq 2\exp\left(-\frac{N\eta^2}{8}\right),$$

where we use $|\mathcal{J}^c|\geq(1-\zeta)N$ and $\eta+\zeta<1/2$. Together with a union bound on the event V, we have

$$\mathbb{P}\left[\left|\frac{1}{|\mathcal{T}|}\sum_{j\in\mathcal{T}}Z_j-\mu\right|\geq C_0\max_{j\in\mathcal{J}^c}\sigma_j\,(3\zeta+4\eta)\sqrt{\log(\frac{3}{\eta})}\right]\leq 3\exp\left(-\frac{N\eta^2}{9}\right).\quad\square$$

A.2. Proof of Theorem 1

We first provide a tail inequality to bound the noises of each instance j, through which our parameter estimation error is also bounded.

LEMMA 2. *Define the event for instance j*

$$\mathcal{H}^j = \left\{ \frac{2}{n_j} \|\mathbf{X}^{j\top} \epsilon^j\|_\infty \leq \frac{\lambda_j}{2} \right\}. \tag{16}$$

Then, we have

$$\mathbb{P}\left[\mathcal{H}^j\right] \geq 1 - 2d \exp\left(-\frac{\lambda_j^2 n_j}{32 \sigma_j^2 x_{\max}^2} \right).$$

Proof of Lemma 2 For any column i of the design matrix $\mathbf{X}^j$, i.e., $\mathbf{X}^j_{(\cdot,i)}$, we have $\|\frac{1}{\sqrt{n_j}} \mathbf{X}^j_{(\cdot,i)}\|_2 \leq x_{\max}$. Then, by Lemma 27, we have

$$\begin{aligned}
\mathbb{P}\left[(\mathcal{H}^j)^c\right] &= \mathbb{P}\left[\max_{i\in[d]} \frac{1}{\sqrt{n_j}} |\mathbf{X}^{j\top}_{(\cdot,i)} \epsilon^j| \geq \frac{\lambda_j \sqrt{n_j}}{4} \right] \\
&\leq d \max_{i\in[d]} \mathbb{P}\left[\frac{1}{\sqrt{n_j}} |\mathbf{X}^{j\top}_{(\cdot,i)} \epsilon^j| \geq \frac{\lambda_j \sqrt{n_j}}{4} \right] \\
&\leq 2d \exp\left(-\frac{\lambda_j^2 n_j}{32 \sigma_j^2 x_{\max}^2} \right). \quad \square
\end{aligned}$$

Now, we prove Theorem 1 by applying Proposition 1 to the OLS estimators of all instances.

Proof of Theorem 1 First, we show that each OLS estimator $\widehat{\beta}^j_{\text{ind}} = (\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}\mathbf{X}^{j\top} Y^j$ constructed in Step 1 is a subgaussian random vector with mean β^j. In particular, the i^{th} component $\widehat{\beta}^j_{\text{ind},(i)}$ of $\widehat{\beta}^j_{\text{ind}}$ is $\sqrt{\frac{\sigma_j^2}{n_j \psi}}$-subgaussian since

$$\begin{aligned}
\mathbb{E}[\exp(\lambda(\widehat{\beta}^j_{\text{ind},(i)} - \beta^j_{(i)}))] &= \mathbb{E}[\exp(\lambda(\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}_{(i,\cdot)} \mathbf{X}^{j\top} \epsilon^j)] \\
&\leq \exp\left(\frac{\lambda^2 \sigma_j^2 \|(\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}_{(i,\cdot)} \mathbf{X}^{j\top}\|_2^2}{2} \right) \\
&\leq \exp\left(\frac{\lambda^2 \sigma_j^2}{2 n_j \psi} \right),
\end{aligned}$$

where the last inequality follows since

$$\|(\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}_{(i,\cdot)} \mathbf{X}^{j\top}\|_2^2 = (\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}_{(i,i)} \leq \lambda_{\max}\left((\mathbf{X}^{j\top}\mathbf{X}^j)^{-1} \right) = \frac{1}{n_j \lambda_{\min}(\widehat{\Sigma}^j)} \leq \frac{1}{n_j \psi}.$$

Now, consider our robust multitask estimator $\{\widehat{\beta}^j_{\text{RM}}\}_{j\in[N]}$ computed by Algorithm 1. Recall that for any poorly-aligned component $i \in \mathcal{I}_{\text{poor}}$, the corresponding corrupted subset of instances is $\mathcal{J}_i = \{j \in [N] \mid \beta^j_{(i)} \neq \beta^\dagger_{(i)}\}$. By definition of $\mathcal{I}_{\text{poor}}$, we have $|\mathcal{J}_i| < N\zeta < N/2$. Since the data from different instances are mutually independent, the vectors $\{\widehat{\beta}^j_{\text{ind}}\}_{j\in[N]}$ are independent. Thus, we can apply Proposition 1 to the trimmed mean of $\{\widehat{\beta}^j_{\text{ind}}\}_{j\in[N]}$, where we use the fact that $\widehat{\beta}^j_{\text{ind}}$ is $\sqrt{\frac{\sigma_j^2}{n_j \psi}}$-subgaussian:

$$\mathbb{P}\left[\left| \widehat{\beta}^\dagger_{\text{RM},(i)} - \beta^\dagger_{(i)} \right| \geq C_0 \left(3\zeta + 4\eta\right) \max_{j\in[N]} \sqrt{\frac{\sigma_j^2}{n_j \psi} \log(\frac{3}{\eta})} \right] \leq 3 \exp\left(-\frac{N\eta^2}{9} \right).$$

By a union bound over $i \in \mathcal{I}_{\text{poor}}$, we have

$$\mathbb{P}\left[\left\|(\widehat{\beta}^{\dagger}_{\text{RM}} - \beta^{\dagger})_{\mathcal{I}_{\text{poor}}}\right\|_1 \geq C_0 d\,(3\zeta + 4\eta) \max_{j\in[N]} \sqrt{\frac{\sigma_j^2}{n_j \psi} \log(\frac{3}{\eta})}\right] \leq 3d \exp\left(-\frac{N\eta^2}{9}\right), \tag{17}$$

where $\beta_{\mathcal{I}}$ for a set $\mathcal{I}$ is described in §2. Next, note that

$$Y^j = \mathbf{X}^j(\beta^{\dagger} + \delta^j) + \epsilon^j = \mathbf{X}^j\left((\beta^{\dagger}_{\mathcal{I}_{\text{poor}}} + \widehat{\beta}^{\dagger}_{\text{RM},\mathcal{I}_{\text{well}}}) + (\beta^{\dagger}_{\mathcal{I}_{\text{well}}} - \widehat{\beta}^{\dagger}_{\text{RM},\mathcal{I}_{\text{well}}} + \delta^j)\right) + \epsilon^j,$$

where $\beta^{\dagger}_{\mathcal{I}_{\text{well}}} - \widehat{\beta}^{\dagger}_{\text{RM},\mathcal{I}_{\text{well}}} + \delta^j$ is $((1/\zeta+1)s)$-sparse — in particular, letting $\bar{\mathcal{I}}_j = \mathcal{I}_{\text{well}} \cup \mathcal{I}_j$, where $\mathcal{I}_j = \{i \in [d] \mid \beta^j_{(i)} \neq \beta^{\dagger}_{(i)}\}$ are the components of β^j that do not equal $\beta^{\dagger}$, then we have $|\bar{\mathcal{I}}_j| \leq (1/\zeta+1)s$. In addition, $\beta^{\dagger}_{\mathcal{I}_{\text{poor}}} + \widehat{\beta}^{\dagger}_{\text{RM},\mathcal{I}_{\text{well}}}$ is closely approximated by $\widehat{\beta}^{\dagger}_{\text{RM}}$ as in (17). Therefore, after computing $\widehat{\beta}^{\dagger}_{\text{RM}}$ in Step 1, we can use LASSO to recover the sparse vector $\beta^{\dagger}_{\mathcal{I}_{\text{well}}} - \widehat{\beta}^{\dagger}_{\text{RM},\mathcal{I}_{\text{well}}} + \delta^j$ to estimate β^j.

The basic inequality of LASSO in the second stage of our algorithm is

$$\frac{1}{n_j}\|\mathbf{X}^j\widehat{\beta}^j_{\text{RM}} - Y^j\|_2^2 + \lambda_j\|\widehat{\beta}^j_{\text{RM}} - \widehat{\beta}^{\dagger}_{\text{RM}}\|_1 \leq \frac{1}{n_j}\|\mathbf{X}^j\beta^j - Y^j\|_2^2 + \lambda_j\|\beta^j - \widehat{\beta}^{\dagger}_{\text{RM}}\|_1.$$

Plugging in $Y^j = \mathbf{X}^j\beta^j + \epsilon^j$, and conditioned on the event $\mathcal{H}^j$ in (16), we obtain

$$\begin{aligned}\frac{1}{n_j}\|\mathbf{X}^j(\widehat{\beta}^j_{\text{RM}} - \beta^j)\|_2^2 + \lambda_j\|\widehat{\beta}^j_{\text{RM}} - \widehat{\beta}^{\dagger}_{\text{RM}}\|_1 &\leq \frac{2}{n_j}\epsilon^{j\top}\mathbf{X}^j(\widehat{\beta}^j_{\text{RM}} - \beta^j) + \lambda_j\|\beta^j - \widehat{\beta}^{\dagger}_{\text{RM}}\|_1\\ &\leq \frac{2}{n_j}\|\mathbf{X}^{j\top}\epsilon^j\|_\infty\|\widehat{\beta}^j_{\text{RM}} - \beta^j\|_1 + \lambda_j\|\beta^j - \widehat{\beta}^{\dagger}_{\text{RM}}\|_1\\ &\leq \frac{\lambda_j}{2}\|\widehat{\beta}^j_{\text{RM}} - \beta^j\|_1 + \lambda_j\|\beta^j - \widehat{\beta}^{\dagger}_{\text{RM}}\|_1.\end{aligned}$$

Decomposing terms based on $\bar{\mathcal{I}}_j$ and $\bar{\mathcal{I}}_j^c$ and rearranging, we have

$$\frac{1}{n_j}\|\mathbf{X}^j(\widehat{\beta}^j_{\text{RM}} - \beta^j)\|_2^2 + \lambda_j\|(\widehat{\beta}^j_{\text{RM}} - \widehat{\beta}^{\dagger}_{\text{RM}})_{\bar{\mathcal{I}}_j^c}\|_1 \leq \frac{3\lambda_j}{2}\|(\widehat{\beta}^j_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j}\|_1 + \frac{\lambda_j}{2}\|(\widehat{\beta}^j_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j^c}\|_1 + \lambda_j\|(\beta^j - \widehat{\beta}^{\dagger}_{\text{RM}})_{\bar{\mathcal{I}}_j^c}\|_1,$$

so it follows that

$$\frac{1}{n_j}\|\mathbf{X}^j(\widehat{\beta}^j_{\text{RM}} - \beta^j)\|_2^2 + \frac{\lambda_j}{2}\|(\widehat{\beta}^j_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j^c}\|_1 \leq \frac{3\lambda_j}{2}\|(\widehat{\beta}^j_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j}\|_1 + 2\lambda_j\|(\widehat{\beta}^{\dagger}_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j^c}\|_1.$$

Then, adding $\frac{\lambda_j}{2}\|(\widehat{\beta}^j_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j}\|_1$ on both sides, we get

$$\begin{aligned}\frac{1}{n_j}\|\mathbf{X}^j(\widehat{\beta}^j_{\text{RM}} - \beta^j)\|_2^2 + \frac{\lambda_j}{2}\|\widehat{\beta}^j_{\text{RM}} - \beta^j\|_1 &\leq 2\lambda_j\|(\widehat{\beta}^j_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j}\|_1 + 2\lambda_j\|(\widehat{\beta}^{\dagger}_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j^c}\|_1\\ &\leq 2\lambda_j\|(\widehat{\beta}^j_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j}\|_1 + 2\lambda_j\|(\widehat{\beta}^{\dagger}_{\text{RM}} - \beta^{\dagger})_{\mathcal{I}_{\text{poor}}}\|_1,\end{aligned} \tag{18}$$

where we use $(\beta^{\dagger}_{\mathcal{I}_{\text{well}}} - \widehat{\beta}^{\dagger}_{\text{RM},\mathcal{I}_{\text{well}}} + \delta^j)_{\bar{\mathcal{I}}_j^c} = \mathbf{0}$. As $\widehat{\Sigma}^j$ is positive definite on $\mathcal{I}^j$, we have

$$\begin{aligned}\|(\widehat{\beta}^j_{\text{RM}} - \beta^j)_{\bar{\mathcal{I}}_j}\|_1 &\leq \sqrt{(1/\zeta+1)s}\|\widehat{\beta}^j_{\text{RM}} - \beta^j\|_2\\ &\leq \sqrt{\frac{(1/\zeta+1)s}{\psi}\frac{1}{n_j}\|\mathbf{X}^j(\widehat{\beta}^j - \beta^j)\|_2^2}.\end{aligned}$$

Then, we derive from inequality (18)

$$\frac{1}{2n_j}\|\mathbf{X}^j(\widehat{\beta}^j_{\text{RM}} - \beta^j)\|_2^2 + \frac{\lambda_j}{2}\|\widehat{\beta}^j_{\text{RM}} - \beta^j\|_1 \leq \frac{2\lambda_j^2(1/\zeta+1)s}{\psi} + 2\lambda_j\|(\widehat{\beta}^{\dagger}_{\text{RM}} - \beta^{\dagger})_{\mathcal{I}_{\text{poor}}}\|_1,$$

where we use the fact that $2ab \le a^2+b^2$. Since $\zeta<1/2$, we have

$$\|\widehat{\beta}^j_{\mathrm{RM}}-\beta^j\|_1 \le \frac{6\lambda_j s}{\zeta\psi}+4\|(\widehat{\beta}^\dagger_{\mathrm{RM}}-\beta^\dagger)_{\mathcal{I}_{\mathrm{poor}}}\|_1.$$

Combining the above with inequality (17) and Lemma 2, we obtain

$$\begin{aligned}\mathbb{P}\left[\left\|\widehat{\beta}^j_{\mathrm{RM}}-\beta^j\right\|_1 \ge \frac{6\lambda_j s}{\zeta\psi}+C_0 d\,(3\zeta+4\eta)\max_{i\in[N]}\sqrt{\frac{\sigma_i^2}{n_i\psi}\log(\frac{3}{\eta})}\right] &\le 3d\exp\left(-\frac{N\eta^2}{9}\right)+\mathbb{P}\left[(\mathcal{H}^j)^c\right]\\ &\le 3d\exp\left(-\frac{N\eta^2}{9}\right)+2d\exp\left(-\frac{\lambda_j^2 n_j}{32\sigma_j^2 x_{\max}^2}\right). \quad\square\end{aligned}$$

Proof of Corollary 1 Taking $\lambda_j=\sqrt{\frac{32\sigma_j^2x_{\max}^2}{n_j}\log(\frac{4d}{\delta})}$ and $\eta=\sqrt{\frac{9}{N}\log(\frac{6d}{\delta})}$ in Theorem 1, and noting that $\delta\le 1$ and $d\ge 1$, we have

$$\sqrt{\log(\frac{3}{\eta})}=\sqrt{\frac{1}{2}\log\left(\frac{N}{\log(\frac{6d}{\delta})}\right)}\le\sqrt{\log(N)}.$$

Therefore, it holds that

$$\|\widehat{\beta}^j_{\mathrm{RM}}-\beta^j\|_1 \le \frac{24s}{\zeta\psi}\sqrt{\frac{2\sigma_j^2x_{\max}^2\log(\frac{4d}{\delta})}{n_j}}+3C_0\zeta d\max_{i\in[N]}\sqrt{\frac{\sigma_i^2\log(N)}{n_i\psi}}+12C_0 d\max_{i\in[N]}\sqrt{\frac{\sigma_i^2\log(N)\log(\frac{6d}{\delta})}{Nn_i\psi}}, \tag{19}$$

with probability at least $1-\delta$. Since n_i's are assumed to be similar in magnitude, choosing $\zeta=\frac{C_0-2}{4C_0}\sqrt{\frac{s}{d}}$ suffices to minimize the scale of the sum of the first two terms in inequality (19). Thus, with probability at least $1-\delta$, we have

$$\|\widehat{\beta}^j_{\mathrm{RM}}-\beta^j\|_1 \le \frac{96C_0}{(C_0-2)\psi}\sqrt{\frac{2\sigma_j^2x_{\max}^2sd\log(\frac{4d}{\delta})}{n_j}}+\frac{3C_0}{4}\max_{i\in[N]}\sqrt{\frac{\sigma_i^2sd\log(N)}{\psi n_i}}+12C_0d\max_{i\in[N]}\sqrt{\frac{\sigma_i^2\log(N)\log(\frac{6d}{\delta})}{\psi Nn_i}}.$$

This holds for any $\delta\ge\exp\left(-\frac{N}{9}(\frac{1}{2}-\frac{1}{C_0}-\frac{C_0-2}{4C_0}\sqrt{\frac{s}{d}})^2+\log(6d)\right)$, as we require $\eta\le 1/2-1/C_0-\zeta$. $\square$

Proof of Corollary 2 Similar to Corollary 1, taking $\lambda_j=\sqrt{\frac{32\sigma_j^2x_{\max}^2}{n_j}\log(\frac{4d}{\delta})}$ and $\eta=\sqrt{\frac{9}{N}\log(\frac{6d}{\delta})}$, we derive from Theorem 1 that

$$\|\widehat{\beta}^j_{\mathrm{RM}}-\beta^j\|_1 \le \frac{24s}{\zeta\psi}\sqrt{\frac{2\sigma_j^2x_{\max}^2\log(\frac{4d}{\delta})}{n_j}}+3C_0\zeta\max_{i\ne j}\sqrt{\frac{d^2\sigma_i^2\log(N)}{n_i\psi}}+12C_0\max_{i\ne j}\sqrt{\frac{d^2\sigma_i^2\log(N)\log(\frac{6d}{\delta})}{Nn_i\psi}},$$

with probability at least $1-\delta$. Since n_j is assumed to be similar to n_i/d^2 for $i\ne j$ in magnitude, choosing ζ to be any constant smaller than $1/2-1/C_0$ suffices to minimize the sum of the first two terms in inequality (19) and thereby we take $\zeta=\frac{C_0-2}{4C_0}$. Thus, with probability at least $1-\delta$, we have

$$\|\widehat{\beta}^j-\beta^j\|_1 \le \frac{96C_0}{(C_0-2)\psi}\sqrt{\frac{2\sigma_j^2x_{\max}^2s^2\log(\frac{4d}{\delta})}{n_j}}+\frac{3C_0}{4}\max_{i\ne j}\sqrt{\frac{d^2\sigma_i^2\log(N)}{\psi n_i}}+12C_0\max_{i\ne j}\sqrt{\frac{d^2\sigma_i^2\log(N)\log(\frac{6d}{\delta})}{Nn_i\psi}}.$$

This holds for any $\delta\ge\exp\left(-\frac{N}{36}(\frac{1}{2}-\frac{1}{C_0})^2+\log(6d)\right)$, as we require $\eta\le 1/2-1/C_0-\zeta$. $\square$

Appendix B: Lower Bounds for Baselines

In this section, we provide detailed statements and proofs for the lower bounds discussed in §3.6 and Table 1. At a high level, our lower bounds follow by exhibiting a concrete instantiation of the parameters β^j and data $\mathbf{X}^j, Y^j$ and establishing a lower bound on the error of the estimator for this instantiation. Recall that our error measure is defined in (8), i.e.,

$$\ell(\widehat{\beta}^j, \beta^j) = \sup_{\mathcal{G}} \mathbb{E}\left[\|\widehat{\beta}^j - \beta^j\|_1\right],$$

where $\mathcal{G} = \left\{\{\mathbf{X}^j\}_{j\in[N]}, \{\beta^j\}_{j\in[N]}, \{\mathcal{P}_\epsilon^j\}_{j\in[N]}\right\}$ satisfies our assumptions in §3, $\mathcal{P}_\epsilon^j$ is the distribution of the noise ϵ^j, and the expectation is taken with respect to ϵ^j's. Since this error measure takes a worst-case scenario over $\mathcal{G}$, it suffices to show the lower bound for a specific case where the assumptions hold.

For the remainder of this section, we assume $\epsilon^j \sim \mathcal{N}(\mathbf{0}, \sigma_j^2\mathbf{I})$, and $\widehat{\Sigma}^j = \mathbf{I}$ for $j \in [N]$. Our choices of errors ϵ^j are all gaussian, which ensures the parameter estimates are gaussian as well, thereby enabling us to obtain lower bounds by applying the following lemma:

Lemma 3. *Consider a multivariate gaussian random variable $X \sim \mathcal{N}(\mu, \Sigma) \in \mathbb{R}^d$. We have*

$$\mathbb{E}\left[\|X\|_1\right] \geq \frac{1}{2}\|\mu\|_1 + \frac{1}{\sqrt{2\pi}}\operatorname{tr}(\Sigma^{\frac{1}{2}}).$$

Proof of Lemma 3 Consider the i^{th} component of X, i.e., $X_{(i)}$. Let $\sigma_i^2 = \Sigma_{(i,i)}$. We have $X_{(i)} \sim \mathcal{N}(\mu_{(i)}, \sigma_i^2)$. Without loss of generality, assume $\mu_{(i)} \geq 0$; otherwise, we can consider $-X_{(i)}$ instead and its ℓ_1 norm stays the same. By our gaussian assumption, it holds that

$$\mathbb{E}[|X_{(i)}|] = \int_{-\infty}^{\infty} |x + \mu_{(i)}| \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{x^2}{2\sigma_i^2}}\, dx \geq \int_0^{\infty} (x + \mu_{(i)}) \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{x^2}{2\sigma_i^2}}\, dx = \frac{1}{2}\mu_{(i)} + \frac{1}{\sqrt{2\pi}}\sigma_i.$$

Then, we further have

$$\mathbb{E}[\|X\|_1] = \sum_{i\in[d]} \mathbb{E}[|X_{(i)}|] = \frac{1}{2}\|\mu\|_1 + \frac{1}{\sqrt{2\pi}} \sum_{i\in[d]} \sqrt{\Sigma_{(i,i)}} \geq \frac{1}{2}\|\mu\|_1 + \frac{1}{\sqrt{2\pi}}\operatorname{tr}(\Sigma^{\frac{1}{2}}),$$

where the last step uses $\sqrt{\Sigma_{(i,i)}} = \|\Sigma^{\frac{1}{2}}_{(i,\cdot)}\|_2 \geq \Sigma^{\frac{1}{2}}_{(i,i)}$. $\square$

B.1. Independent Estimator

First, we consider the *independent estimator*, which simply uses ordinary least squares (OLS) independently on each instance (i.e., it does not perform any learning across instances). This estimator is

$$\widehat{\beta}^j_{\text{ind}} = (\mathbf{X}^{j\top}\mathbf{X}^j)^{-1}\mathbf{X}^{j\top}Y^j.$$

Intuitively, this estimator has high variance since it uses relatively little data to estimate β^j. In particular, we have the following result:

Proposition 5. *The estimation error of the independent estimator in both the standard and data-poor regimes satisfies*

$$\ell(\widehat{\beta}^j_{ind}, \beta^j) \geq \frac{d\sigma_j}{\sqrt{2\pi n_j}} = \Omega\left(\frac{d}{\sqrt{n_j}}\right).$$

Proof of Proposition 5 For our choice of $\mathbf{X}^j$ and ϵ^j, the estimation error follows a gaussian distribution:

$$\widehat{\beta}^j_{\text{ind}} - \beta^j \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_j^2}{n_j}\mathbf{I}\right).$$

Therefore, using Lemma 3, we have

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\text{ind}} - \beta^j\|_1\right] \geq \frac{d\sigma_j}{\sqrt{2\pi n_j}}. \quad \square$$

B.2. Averaging Estimator

Next, we consider the *averaging estimator*, which simply takes the average of the independent parameter estimates across instances to reduce variance:

$$\widehat{\beta}^j_{\text{avg}} = \frac{1}{N}\sum_{i\in[N]} \widehat{\beta}^i_{\text{ind}}.$$

Note that this estimator is constant across instances j's; also, it is identical to Step 1 of the averaging multitask estimator described in §3.2 — i.e., $\widehat{\beta}^j_{\text{avg}} = \widehat{\beta}^{\dagger}_{\text{AM}}$.

PROPOSITION 6. *The estimation error of the averaging estimator in the standard regime satisfies*

$$\begin{aligned}
\ell(\widehat{\beta}^j_{avg}, \beta^j) &\geq \frac{1}{2}\left\|\frac{1}{N}\sum_{i\in[N]}(\delta^i - \delta^j)\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\in[N]}\frac{\sigma_i^2 n_j}{n_i}}\frac{d}{\sqrt{Nn_j}} \\
&= \Omega\left(\left\|\frac{1}{N}\sum_{i\in[N]}(\delta^i - \delta^j)\right\|_1 + \frac{d}{\sqrt{Nn_j}}\right),
\end{aligned}$$

and in the data-poor regime satisfies

$$\begin{aligned}
\ell(\widehat{\beta}^j_{avg}, \beta^j) &\geq \frac{1}{2}\left\|\frac{1}{N}\sum_{i\in[N]}(\delta^i - \delta^j)\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\in[N]}\frac{\sigma_i^2 n_j}{n_i}}\frac{d}{\sqrt{Nn_j}} \\
&= \Omega\left(\left\|\frac{1}{N}\sum_{i\neq j}(\delta^i - \delta^j)\right\|_1 + \frac{1}{\sqrt{Nn_j}}\right).
\end{aligned}$$

Proof of Proposition 6 For our choice of $\mathbf{X}^j$'s and ϵ^j's, the estimation error follows a gaussian distribution:

$$\widehat{\beta}^j_{\text{avg}} - \beta^j = \frac{1}{N}\sum_{i\in[N]}(\widehat{\beta}^i_{\text{ind}} - \beta^i) + \frac{1}{N}\sum_{i\in[N]}(\delta^i - \delta^j) \sim \mathcal{N}\left(\frac{1}{N}\sum_{i\in[N]}(\delta^i - \delta^j), \frac{1}{N^2}\sum_{i\in[N]}\frac{\sigma_i^2}{n_i}\mathbf{I}\right).$$

Therefore, by Lemma 3, we have

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\text{avg}} - \beta^j\|_1\right] \geq \frac{1}{2}\left\|\frac{1}{N}\sum_{i\in[N]}(\delta^i - \delta^j)\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\in[N]}\frac{\sigma_i^2 n_j}{n_i}}\frac{d}{\sqrt{Nn_j}}.$$

For data-poor regime, we use all the instances except j as a proxy. Following a similar proof strategy as above, we have

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\text{avg}} - \beta^j\|_1\right] \geq \frac{1}{2}\left\|\frac{1}{N}\sum_{i\neq j}(\delta^i - \delta^j)\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\neq j}\frac{\sigma_i^2 d^2 n_j}{n_i}}\frac{1}{\sqrt{Nn_j}}.$$

B.3. Pooling Estimator

Next, we consider the *pooling estimator*, which pools all the data $\mathbf{X}^j, Y^j$ across instances, and then uses OLS on this pooled dataset:

$$\widehat{\beta}^j_{\text{pool}} = \left(\sum_{i\in[N]}\mathbf{X}^{i\top}\mathbf{X}^i\right)^{-1}\left(\sum_{i\in[N]}\mathbf{X}^{i\top}Y^i\right).$$

As with the averaging estimator, this estimator is constant across instances j's. Intuitively, it performs similarly to the averaging estimator, except it accounts for differences in the covariance matrices $\widehat{\Sigma}^j = \frac{\mathbf{X}^{j\top}\mathbf{X}^j}{n_j}$ across instances.

PROPOSITION 7. *The estimation error of the pooling estimator in the standard regime satisfies*

$$\ell(\widehat{\beta}^j_{pool}, \beta^j) \geq \frac{1}{2}\left\|\frac{\sum_{i\in[N]} n_i(\delta^i - \delta^j)}{\sum_{i\in[N]} n_i}\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{(\sum_{i\in[N]}\sigma_i^2 n_i)Nn_j}{(\sum_{i\in[N]} n_i)^2}}\frac{d}{\sqrt{Nn_j}}$$
$$= \Omega\left(\left\|\frac{\sum_{i\in[N]} n_i(\delta^i - \delta^j)}{\sum_{i\in[N]} n_i}\right\|_1 + \frac{d}{\sqrt{Nn_j}}\right),$$

and in the data-poor regime satisfies

$$\ell(\widehat{\beta}^j_{pool}, \beta^j) \geq \frac{1}{2}\left\|\frac{\sum_{i\neq j} n_i(\delta^i - \delta^j)}{\sum_{i\neq j} n_i}\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{(\sum_{i\neq j}\sigma_i^2 n_i)Nd^2 n_j}{(\sum_{i\neq j} n_i)^2}}\frac{1}{\sqrt{Nn_j}}$$
$$= \Omega\left(\left\|\frac{\sum_{i\neq j} n_i(\delta^i - \delta^j)}{\sum_{i\neq j} n_i}\right\|_1 + \frac{1}{\sqrt{Nn_j}}\right).$$

Proof of Proposition 7 For our choice of $\mathbf{X}^j$'s and ϵ^j's, the estimation error follows a gaussian distribution:

$$\widehat{\beta}^j_{\text{pool}} - \beta^j = \left(\sum_{i\in[N]} \mathbf{X}^{i\top}\mathbf{X}^i\right)^{-1}\left(\sum_{i\in[N]} \mathbf{X}^{i\top}\mathbf{X}^i(\delta^i - \delta^j)\right) + \left(\sum_{i\in[N]} \mathbf{X}^{i\top}\mathbf{X}^i\right)^{-1}\left(\sum_{i\in[N]} \mathbf{X}^{i\top}\epsilon^i\right)$$
$$\sim \mathcal{N}\left(\frac{\sum_{i\in[N]} n_i(\delta^i - \delta^j)}{\sum_{i\in[N]} n_i}, \frac{\sum_{i\in[N]}\sigma_i^2 n_i}{(\sum_{i\in[N]} n_i)^2}\mathbf{I}\right).$$

Therefore, Lemma 3 implies

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\text{pool}} - \beta^j\|_1\right] \geq \frac{1}{2}\left\|\frac{\sum_{i\in[N]} n_i(\delta^i - \delta^j)}{\sum_{i\in[N]} n_i}\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{(\sum_{i\in[N]}\sigma_i^2 n_i)Nn_j}{(\sum_{i\in[N]} n_i)^2}}\frac{d}{\sqrt{Nn_j}}.$$

Note that for data-poor regime, we use all the instances except j as a proxy. Following a similar proof strategy as above, we have

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\text{pool}} - \beta^j\|_1\right] \geq \frac{1}{2}\left\|\frac{\sum_{i\neq j} n_i(\delta^i - \delta^j)}{\sum_{i\neq j} n_i}\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{(\sum_{i\neq j}\sigma_i^2 n_i)Nd^2 n_j}{(\sum_{i\neq j} n_i)^2}}\frac{1}{\sqrt{Nn_j}}. \quad \square$$

B.4. Averaging Multitask Estimator

Finally, we consider the averaging multitask estimator described in §3.2, which uses a traditional estimate of the mean instead of our robust estimate in Step 1. Our lower bound on the error of this estimator demonstrates the importance of robustness. Following the proof of the LASSO lower bound in Theorem 7.1 of Lounici et al. (2011), we assume that λ_j is chosen based on the analysis of the error upper bound. With a similar argument as in Corollary 1, we take $\lambda_j = \sqrt{\frac{32\sigma_j^2 x_{\max}^2}{n_j}\log(\frac{4d}{c})}$ with $\delta = c$ for any $1/2 > c > 0$, which is mainly based on Lemma 2.

PROPOSITION 8. *Let* $b \geq \min\{Ns, d\}\left(\sqrt{\frac{50\sigma_j^2 x_{\max}^2}{n_j}\log(\frac{4d}{c})} + \sqrt{\frac{2}{N^2}\sum_{k\in[N]}\frac{\sigma_k^2}{n_k}\log(\frac{2}{c})}\right)$ *and the regularization parameter* $\lambda_j = \sqrt{\frac{32\sigma_j^2 x_{\max}^2}{n_j}\log(\frac{4d}{c})}$. *The estimation error of the averaging multitask estimator in the standard regime satisfies*

$$\ell(\widehat{\beta}^j_{AM}, \beta^j) = \tilde{\Omega}\left(\frac{\min\{Ns, d\}}{\sqrt{n_j}} + \frac{d}{\sqrt{Nn_j}}\right),$$

and in the data-poor regime satisfies

$$\ell(\widehat{\beta}^j_{AM}, \beta^j) = \tilde{\Omega}\left(\frac{\min\{Ns, d\}}{\sqrt{n_j}}\right).$$

Proof of Proposition 8 The first order condition of problem (4) is

$$\frac{1}{n_j}\mathbf{X}^{j\top}(Y^j-\mathbf{X}^j\widehat{\beta}^j_{\mathrm{AM}})=\lambda_j\partial\|\widehat{\beta}^j_{\mathrm{AM}}-\widehat{\beta}^\dagger_{\mathrm{AM}}\|_1,$$

where $\partial\|\widehat{\beta}^j_{\mathrm{AM}}-\widehat{\beta}^\dagger_{\mathrm{AM}}\|_1$ is the subgradient of ℓ_1 norm at $\widehat{\beta}^j_{\mathrm{AM}}-\widehat{\beta}^\dagger_{\mathrm{AM}}$; in particular, for the i^{th} component,

$$\begin{cases}\frac{1}{n_j}\left(\mathbf{X}^{j\top}(Y^j-\mathbf{X}^j\widehat{\beta}^j_{\mathrm{AM}})\right)_{(i)}=\lambda_j\mathrm{sign}(\widehat{\beta}^j_{\mathrm{AM},(i)}-\widehat{\beta}^\dagger_{\mathrm{AM},(i)}) & \text{if } \widehat{\beta}^j_{\mathrm{AM},(i)}\neq\widehat{\beta}^\dagger_{\mathrm{AM},(i)}\\ \left|\frac{1}{n_j}\left(\mathbf{X}^{j\top}(Y^j-\mathbf{X}^j\widehat{\beta}^j_{\mathrm{AM}})\right)_{(i)}\right|\le\lambda_j & \text{if } \widehat{\beta}^j_{\mathrm{AM},(i)}=\widehat{\beta}^\dagger_{\mathrm{AM},(i)}.\end{cases}\tag{20}$$

Next, on the event $\mathcal{H}^j$, it holds

$$\frac{2}{n_j}|(\mathbf{X}^{j\top}\epsilon^j)_{(i)}|\le\frac{2}{n_j}\|\mathbf{X}^{j\top}\epsilon^j\|_\infty\le\frac{\lambda_j}{2}.$$

Combining it with (20), we have

$$\frac{3\lambda_j}{4}\le\left|\frac{1}{n_j}\left(\mathbf{X}^{j\top}(\mathbf{X}^j\beta^j-\mathbf{X}^j\widehat{\beta}^j_{\mathrm{AM}})\right)_{(i)}\right|=|(\widehat{\beta}^j_{\mathrm{AM}}-\beta^j)_{(i)}|$$

for each i such that $\widehat{\beta}^j_{\mathrm{AM},(i)}\neq\widehat{\beta}^\dagger_{\mathrm{AM},(i)}$, where the last equality is from our assumption $\widehat{\Sigma}^j=\mathbf{I}$. For the rest of the components, note that $\widehat{\beta}^j_{\mathrm{AM},(i)}=\widehat{\beta}^\dagger_{\mathrm{AM},(i)}$. Summing over all $i\in[d]$, we get

$$\|\widehat{\beta}^j_{\mathrm{AM}}-\beta^j\|_1\ge\frac{3|\mathcal{V}|\lambda_j}{4},\tag{21}$$

where $\mathcal{V}=\{i\in[d]\mid\widehat{\beta}^j_{\mathrm{AM},(i)}\neq\widehat{\beta}^\dagger_{\mathrm{AM},(i)}\}$. Next, define $\widehat{\delta}^j_{\mathrm{AM}}=\widehat{\beta}^j_{\mathrm{AM}}-\widehat{\beta}^\dagger_{\mathrm{AM}}$ and $\widetilde{\delta}^j_{\mathrm{AM}}=\beta^j-\widetilde{\beta}^\dagger_{\mathrm{AM}}$, where $\widetilde{\beta}^\dagger_{\mathrm{AM}}$ is defined in (5). Then, $|\mathcal{V}|=\|\widehat{\delta}^j_{\mathrm{AM}}\|_0$. For the remainder of the proof, we use a similar argument as the proof of Theorem 7.1 in Lounici et al. (2011).

First, if $\|\widehat{\delta}^j_{\mathrm{AM}}\|_0<\|\widetilde{\delta}^j_{\mathrm{AM}}\|_0$, then there exists $i\in[d]$ such that $\widehat{\delta}^j_{\mathrm{AM},(i)}=0$ but $\widetilde{\delta}^j_{\mathrm{AM},(i)}\neq0$. By the first order condition (20), for such i, we have

$$|(\widehat{\beta}^\dagger_{\mathrm{AM}}-\beta^j)_{(i)}|=|(\widehat{\beta}^j_{\mathrm{AM}}-\beta^j)_{(i)}|\le\frac{5\lambda_j}{4},$$

and hence

$$|(\beta^j-\widetilde{\beta}^\dagger_{\mathrm{AM}})_{(i)}|\le\frac{5\lambda_j}{4}+|(\widetilde{\beta}^\dagger_{\mathrm{AM}}-\widehat{\beta}^\dagger_{\mathrm{AM}})_{(i)}|.\tag{22}$$

Now, note that

$$\widehat{\beta}^\dagger_{\mathrm{AM}}-\widetilde{\beta}^\dagger_{\mathrm{AM}}=\frac{1}{N}\sum_{i\in[N]}(\widehat{\beta}^i_{\mathrm{ind}}-\beta^i)\sim\mathcal{N}\left(\mathbf{0},\frac{1}{N^2}\sum_{i\in[N]}\frac{\sigma_i^2}{n_i}\mathbf{I}\right).$$

Then, by Hoeffding's inequality, we have for any $t>0$ and $i\in[d]$

$$\mathbb{P}\left[|(\widehat{\beta}^\dagger_{\mathrm{AM}}-\widetilde{\beta}^\dagger_{\mathrm{AM}})_{(i)}|\ge t\right]\le2\exp\left(-\frac{t^2}{\frac{2}{N^2}\sum_{k\in[N]}\frac{\sigma_k^2}{n_k}}\right).$$

Taking $t=\sqrt{\frac{2}{N^2}\sum_{k\in[N]}\frac{\sigma_k^2}{n_k}\log(\frac{2}{c})}$ and given our choice of λ_j, we derive from (22) that

$$\mathbb{P}\left[|(\beta^j-\widetilde{\beta}^\dagger_{\mathrm{AM}})_{(i)}|\le5\sqrt{\frac{2\sigma_j^2x_{\max}^2}{n_j}\log(\frac{4d}{c})}+\sqrt{\frac{2}{N^2}\sum_{k\in[N]}\frac{\sigma_k^2}{n_k}\log(\frac{2}{c})}\right]\ge\mathbb{P}\left[\mathcal{H}^j\right]-c\ge1-2c.$$

Since we consider a worst-case error over all possible $\{\beta^j\}_{j\in[N]}$ that satisfy our assumptions in §3, we focus on special cases in two different situations (i) $Ns \le d$ and (ii) $Ns > d$. When $Ns \le d$, we consider a case where (i) $|\delta^j_{(i)}| > \sqrt{\frac{50\sigma_j^2 x_{\max}^2 N^2}{n_j}\log(\frac{4d}{c})} + \sqrt{2\sum_{k\in[N]}\frac{\sigma_k^2}{n_k}\log(\frac{2}{c})}$ for any $i\in[d]$ and $j\in[N]$ such that $\delta^j_{(i)} \neq 0$, and (ii) $\delta^k_{(i)} = 0$ for any $k \neq j$ and $i \in [d]$ such that $\delta^j_{(i)} \neq 0$ for any $j\in[N]$. When $Ns > d$, we consider a case where (i) $|\delta^j_{(i)}| > \sqrt{\frac{50\sigma_j^2 x_{\max}^2 d^2}{s^2 n_j}\log(\frac{4d}{c})} + \sqrt{\frac{2d^2}{s^2N^2}\sum_{k\in[N]}\frac{\sigma_k^2}{n_k}\log(\frac{2}{c})}$ for any $i\in[d]$ and $j\in[N]$ such that $\delta^j_{(i)} \neq 0$, and (ii) $|\{j\in[N] \mid \delta^j_{(i)} \neq 0\}| = Ns/d$. For the above two cases to satisfy our assumptions in §3, it suffices to have $b \ge \min\{Ns, d\}\left(\sqrt{\frac{50\sigma_j^2 x_{\max}^2}{n_j}\log(\frac{4d}{c})} + \sqrt{\frac{2}{N^2}\sum_{k\in[N]}\frac{\sigma_k^2}{n_k}\log(\frac{2}{c})}\right)$. Then, it leads to a contradiction with probability at least $1-2c$. As a consequence, we must have $\|\widehat{\delta}^j_{\mathrm{AM}}\|_0 \ge \|\widetilde{\delta}^j_{\mathrm{AM}}\|_0$ with a high probability; correspondingly, $|\mathcal{V}| \ge \min\{Ns, d\}$ noting that $\|\widetilde{\delta}^j_{\mathrm{AM}}\|_0 = \min\{Ns, d\}$ given our design above.

On the other hand, it always holds true that

$$\|\widehat{\beta}^j_{\mathrm{AM}} - \beta^j\|_1 \ge \|(\widehat{\beta}^\dagger_{\mathrm{AM}} - \beta^j)_{\mathcal{V}^c}\|_1.$$

By Lemma 3 and given the set $\mathcal{V}$, we have

$$\begin{aligned}
\mathbb{E}\left[\|(\widehat{\beta}^\dagger_{\mathrm{AM}} - \beta^j)_{\mathcal{V}^c}\|_1\right] &\ge \frac{1}{2}\left\|\frac{1}{N}\sum_{i\in[N]}(\delta^i - \delta^j)_{\mathcal{V}^c}\right\|_1 + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\in[N]}\frac{\sigma_i^2 n_j}{n_i}}\frac{|\mathcal{V}^c|}{\sqrt{Nn_j}} \\
&= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\in[N]}\frac{\sigma_i^2 n_j}{n_i}}\frac{|\mathcal{V}^c|}{\sqrt{Nn_j}}, \qquad (23)
\end{aligned}$$

where the last equality holds because the support of $\widehat{\delta}^j_{\mathrm{AM}}$ includes that of $\widetilde{\delta}^j_{\mathrm{AM}}$ as shown in the last paragraph. Given $|\mathcal{V}| \le d$ and the fact that $|\mathcal{V}| \ge \min\{Ns, d\}$ holds with probability at least $1-2c$, we derive from (23) that

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\mathrm{AM}} - \beta^j\|_1\right] \ge \mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\in[N]}\frac{\sigma_i^2 n_j}{n_i}}\frac{d-|\mathcal{V}|}{\sqrt{Nn_j}}\,\middle|\, d \ge |\mathcal{V}| \ge \min\{Ns, d\}\right](1-2c). \qquad (24)$$

Further, from (21), we also have

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\mathrm{AM}} - \beta^j\|_1\right] \ge \mathbb{E}\left[\frac{3|\mathcal{V}|\lambda_j}{4}\right](1-c) \ge \mathbb{E}\left[\frac{3(1-c)|\mathcal{V}|\lambda_j}{4}\,\middle|\, d \ge |\mathcal{V}| \ge \min\{Ns, d\}\right](1-2c). \qquad (25)$$

Combining (24) and (25), we have

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\mathrm{AM}} - \beta^j\|_1\right] \ge \frac{1-2c}{2}\mathbb{E}\left[\frac{3(1-c)|\mathcal{V}|\lambda_j}{4} + \frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\in[N]}\frac{\sigma_i^2 n_j}{n_i}}\frac{d-|\mathcal{V}|}{\sqrt{Nn_j}}\,\middle|\, d \ge |\mathcal{V}| \ge \min\{Ns, d\}\right],$$

where we use $\max\{a, b\} \ge (a+b)/2$. As the above lower bound is linear in $|\mathcal{V}|$, its minimum value is taken at either ends of the interval $[\min\{Ns, d\}, d]$. Therefore, we can write

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{\beta}^j_{\mathrm{AM}} - \beta^j\|_1\right] \ge (1-2c)\min\Bigg\{&\frac{3(1-c)d}{2}\sqrt{\frac{2\sigma_j^2 x_{\max}^2}{n_j}\log(\frac{4d}{c})}, \\
&\frac{3(1-c)\min\{Ns, d\}}{2}\sqrt{\frac{2\sigma_j^2 x_{\max}^2}{n_j}\log(\frac{4d}{c})} + \frac{1}{2\sqrt{2\pi}}\sqrt{\frac{1}{N}\sum_{i\in[N]}\frac{\sigma_i^2 n_j}{n_i}}\frac{\max\{d-Ns, 0\}}{\sqrt{Nn_j}}\Bigg\}.
\end{aligned}$$

Therefore, when $Ns = o(d)$, we can write

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\mathrm{AM}} - \beta^j\|_1\right] = \tilde{\Omega}(\frac{Ns}{\sqrt{n_j}} + \frac{d}{\sqrt{Nn_j}});$$

when $Ns = \Omega(d)$, we get

$$\mathbb{E}\left[\|\widehat{\beta}^j_{\mathrm{AM}} - \beta^j\|_1\right] = \tilde{\Omega}(\frac{d}{\sqrt{n_j}}).$$

The proof for the data-poor regime is similar. $\square$

Appendix C: Regret Analysis of RMBandit Algorithm

In this section, we give a proof for Theorem 2, Corollary 3 and Corollary 4 in §3.4 and §3.5. We list the hyperparameter choices for Theorem 2 to hold as follows. Particularly, we take

$$\zeta_0 = \zeta_{1,0} = \frac{C_0 - 2}{4C_0}\sqrt{\frac{s}{d}}, \quad \eta_0 = \sqrt{\frac{27\log(d)|\mathcal{B}_0|}{qN}}, \quad \eta_{1,0} = \sqrt{\frac{9}{\rho N}},$$
$$\lambda_{0,j} = \sqrt{\frac{96\sigma_j^2 x_{\max}^2 K\log(d)|\mathcal{B}_0|}{q|\mathcal{B}_0^j|}}, \quad \lambda_{1,j,0} = \sqrt{\frac{256\sigma_j^2 x_{\max}^2}{p_*}},$$

and

$$q = \max\left\{\frac{2\cdot 384^3 C_0^2 x_{\max}^4 (\max_{i\in[N]}\sigma_i^2/p_i)Ksd\log(d)}{(C_0-2)^2h^2p_*^2\psi^2}, \frac{576C_0^2x_{\max}^2(\max_{i\in[N]}\sigma_i^2/p_i)Ksd\log(N)}{h^2p_*\psi},\right.$$
$$\left.\frac{6\cdot 192^2C_0^2x_{\max}^2(\max_{i\in[N]}\sigma_i^2/p_i)Kd^2\log(d)\log(N)}{h^2p_*\psi N}, \frac{96x_{\max}^2Kd\log(dN)}{p_*\psi(\min_{i\in[N]}p_i)}, \frac{60K\log(N)}{p_*(\min_{i\in[N]}p_i)}\right\}.$$

C.1. Proof of Proposition 2

We first prove Proposition 2, which extends Theorem 1 to the setting where the design matrices $\mathbf{X}^j$'s are random.

Proof of Proposition 2 The proof follows that of Theorem 1. Define

$$\mathcal{E}^j = \left\{\lambda_{\min}(\widehat{\Sigma}^j) \geq \phi\right\}$$

for any $j \in [N]$ and some $\phi > 0$. Then, on the event $\bigcap_{i\in[N]}\mathcal{E}^i$, Theorem 1 gives

$$\mathbb{P}\left[\|\widehat{\beta}^j_{\mathrm{RM}} - \beta^j\|_1 \geq \frac{6\lambda_j s}{\zeta\phi} + C_0 d\,(3\zeta + 4\eta)\max_{i\in[N]}\sqrt{\frac{\sigma_i^2}{n_i\phi}\log(\frac{3}{\eta})}\,\middle|\, \cap_{i\in[N]}\mathcal{E}^i, \{\mathbf{X}^i\}_{i\in[N]}\right]$$
$$\leq 3d\exp\left(-\frac{N\eta^2}{9}\right) + 2d\exp\left(-\frac{\lambda_j^2 n_j}{32\sigma_j^2x_{\max}^2}\right),$$

where we condition on the design matrices $\{\mathbf{X}^i\}_{i\in[N]}$ since Theorem 1 considers the fixed design. Integrating over $\{\mathbf{X}^i\}_{i\in[N]}$ and using a union bound, we obtain

$$\mathbb{P}\left[\|\widehat{\beta}^j_{\mathrm{RM}} - \beta^j\|_1 \geq \frac{6\lambda_j s}{\zeta\phi} + C_0 d(\max_{i\in[N]}\sqrt{\frac{\sigma_i^2}{n_i\phi}})\,(3\zeta + 4\eta)\sqrt{\log(\frac{3}{\eta})}\right]$$
$$\leq 3d\exp\left(-\frac{N\eta^2}{9}\right) + 2d\exp\left(-\frac{\lambda_j^2 n_j}{32\sigma_j^2x_{\max}^2}\right) + \mathbb{P}\left[\cup_{i\in[N]}(\mathcal{E}^i)^c\right]$$
$$\leq 3d\exp\left(-\frac{N\eta^2}{9}\right) + 2d\exp\left(-\frac{\lambda_j^2 n_j}{32\sigma_j^2x_{\max}^2}\right) + \sum_{i\in[N]}\mathbb{P}\left[(\mathcal{E}^i)^c\right]. \quad \square$$

C.2. Forced-Sample Estimator

Next, we prove Proposition 3, which says that our forced sample estimators have small estimation error with high probability given our batching design. For simplicity, we use ζ, η and λ_j to represent ζ_0, η_0 and $\lambda_{0,j}$ respectively in the following.

We start by defining some additional notations. Let $\mathcal{B}_0^j$ be the index set of those observed at instance j, $\mathcal{B}_{0,k}^j$ be the subset forced sampled at arm k and $\bar{\mathcal{B}}_{0,k}^j$ be the subset of all $t \in \mathcal{B}_{0,k}^j$ such that $X_t \in U_k^j$; in particular,

$$\mathcal{B}_0^j = \{t \in \mathcal{B}_0 \mid Z_t = j\}, \quad \mathcal{B}_{0,k}^j = \left\{ t \in \mathcal{B}_0 \,\middle|\, Z_t = j,\, (k-1) \equiv (\sum_{r \in [t]} \mathbb{1}(Z_r = j) - 1) \bmod K \right\},$$

$$\bar{\mathcal{B}}_{0,k}^j = \left\{ t \in \mathcal{B}_0 \,\middle|\, Z_t = j,\, X_t \in U_k^{Z_t},\, (k-1) \equiv (\sum_{r \in [t]} \mathbb{1}(Z_r = j) - 1) \bmod K \right\}.$$

Note that the distribution of X_t always conditions on the value of Z_t.

Lemma 4. *The forced samples of arm k are independent across bandit instances.*

Proof of Lemma 4 The forced samples of arm k at instance j are $\{(X_t, Y_t)\}_{t \in \mathcal{B}_{0,k}^j}$, where the set of covariates is

$$\left\{ X_t \,\middle|\, t \in \mathcal{B}_{0,k}^j,\, Z_t = j,\, (k-1) \equiv (\sum_{r \in [t]} \mathbb{1}(Z_r = j) - 1) \bmod K \right\}.$$

Since $\{(X_t, Z_t)\}_{t \in \mathcal{B}_0}$ are independent, X_t is independent of $Z_{t'}$ and $X_{t'}$ conditional on Z_t for any $t' \neq t$ and $t' \in \mathcal{B}_0$. Further note that given $Z_t = j$, $\sum_{r \in [t]} \mathbb{1}(Z_r = j) - 1 = \sum_{r \in [t-1]} \mathbb{1}(Z_r = j)$ is also independent of X_t. Thus, X_t's observed at arm k are independent across bandit instances. Similarly, Y_t's are also independent across different instances as the noise ϵ_t also only depends on Z_t by design. As a result, the forced samples of arm k are independent across different instances. $\square$

Now we consider a set of subsamples of arm k at one single bandit instance j.

Lemma 5. *The samples $\{X_t\}_{t \in \mathcal{B}_{0,k}^j}$ are i.i.d. with distribution $\mathcal{P}_X^j$, and its subset $\{X_t\}_{t \in \bar{\mathcal{B}}_{0,k}^j}$ are i.i.d. with distribution $\mathcal{P}^j_{X \mid X \in U_k^j}$.*

Proof of Lemma 5 Using a similar argument in the proof of Lemma 4, we can show that $\{X_t\}_{t \in \mathcal{B}_{0,k}^j}$ are independent. As $\sum_{r \in [t]} \mathbb{1}(Z_r = j)$ is independent of X_t given $Z_t = j$, X_t follows the distribution $\mathcal{P}_X^j$. On the other hand, the subsamples $\{X_t\}_{t \in \bar{\mathcal{B}}_{0,k}^j}$ form the set

$$\left\{ X_t \,\middle|\, t \in \bar{\mathcal{B}}_{0,k}^j,\, X_t \in U_k^{Z_t},\, Z_t = j,\, (k-1) \equiv (\sum_{r \in [t]} \mathbb{1}(Z_r = j) - 1) \bmod K \right\}.$$

Since $\{(X_t, Z_t)\}_{t \in \mathcal{B}_0}$ are independent, X_t is independent of $Z_{t'}$ and $X_{t'}$ conditional on $\left\{X_t \in U_k^{Z_t}, Z_t = j\right\}$ for any $t' \neq t$ and $t' \in \mathcal{B}_0$. Similarly, we can conclude that $\{X_t\}_{t \in \bar{\mathcal{B}}_{0,k}^j}$ are i.i.d. drawn from $\mathcal{P}^j_{X \mid X \in U_k^j}$. $\square$

In the following, we use the notation $\widehat{\Sigma}(\mathcal{B})$ to represent the sample covariance matrix created using the samples $\{X_t\}_{t \in \mathcal{B}}$.

LEMMA 6. *Define the event*

$$\bar{\mathcal{E}}^j_{0,k} = \left\{ \lambda_{\min}(\widehat{\Sigma}(\bar{\mathcal{B}}^j_{0,k})) \geq \frac{\psi}{2} \right\}. \tag{26}$$

Then, given $|\bar{\mathcal{B}}^j_{0,k}|$, *we have*

$$\mathbb{P}\left[\bar{\mathcal{E}}^j_{0,k}\right] \geq 1 - d\exp\left(-\frac{\psi|\bar{\mathcal{B}}^j_{0,k}|}{8dx^2_{\max}}\right).$$

Proof of Lemma 6 Note that $\{X_t X_t^\top\}_{t\in\bar{\mathcal{B}}^j_{0,k}}$ are i.i.d. according to Lemma 5. By Assumption 1, for any $t\in\bar{\mathcal{B}}^j_{0,k}$, $\lambda_{\max}(X_t X_t^\top) \leq \|X_t\|_2^2 \leq dx^2_{\max}$. Therefore, by taking $t = 1/2$ and $L = dx^2_{\max}$, we instantaneously get our result from Lemma 29. □

LEMMA 7. *For any sets* $\mathcal{B}, \bar{\mathcal{B}}$ *with* $\bar{\mathcal{B}} \subseteq \mathcal{B}$, *it holds that*

$$\lambda_{\min}(\widehat{\Sigma}(\mathcal{B})) \geq \frac{|\bar{\mathcal{B}}|}{|\mathcal{B}|}\lambda_{\min}(\widehat{\Sigma}(\bar{\mathcal{B}})) + \frac{|\mathcal{B}\setminus\bar{\mathcal{B}}|}{|\mathcal{B}|}\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}\setminus\bar{\mathcal{B}})).$$

Proof of Lemma 7 By our definition, we have

$$\widehat{\Sigma}(\mathcal{B}) = \frac{|\bar{\mathcal{B}}|}{|\mathcal{B}|}\widehat{\Sigma}(\bar{\mathcal{B}}) + \frac{|\mathcal{B}\setminus\bar{\mathcal{B}}|}{|\mathcal{B}|}\widehat{\Sigma}(\mathcal{B}\setminus\bar{\mathcal{B}}).$$

The result follows by noting that the minimum eigenvalue is concave. □

LEMMA 8. *Given* $|\mathcal{B}^j_{0,k}|$, *it holds that*

$$\mathbb{P}\left[|\bar{\mathcal{B}}^j_{0,k}| \leq \frac{p_*|\mathcal{B}^j_{0,k}|}{2}\right] \leq 2\exp\left(-\frac{p_*|\mathcal{B}^j_{0,k}|}{10}\right).$$

Proof of Lemma 8 Applying Lemma 30 to the indicator random variables $\mathbb{1}\left(t\in\bar{\mathcal{B}}^j_{0,k}\right)$ for all $t\in\mathcal{B}^j_{0,k}$ with $\mu = \mathbb{E}\left[\sum_{t\in\mathcal{B}^j_{0,k}}\mathbb{1}\left(t\in\bar{\mathcal{B}}^j_{0,k}\right)\right] = \sum_{t\in\mathcal{B}^j_{0,k}}\mathbb{P}\left[X_t\in U^j_k \mid Z_t = j\right]$, we have

$$\mathbb{P}\left[||\bar{\mathcal{B}}^j_{0,k}| - \mu| \geq \frac{\mu}{2}\right] \leq 2\exp\left(-\frac{\mu}{10}\right).$$

Noting $\mu \geq p_*|\mathcal{B}^j_{0,k}|$ by Assumption 4, the result then follows. □

LEMMA 9. *It holds that*

$$\mathbb{P}\left[|\mathcal{B}^j_0| \leq \frac{p_j|\mathcal{B}_0|}{2}\right] \leq 2\exp\left(-\frac{p_j|\mathcal{B}_0|}{10}\right).$$

Proof of Lemma 9 Applying Lemma 30 to the indicator random variables $\mathbb{1}(Z_t = j)$ for all $t\in\mathcal{B}_0$ with $\mu = \mathbb{E}\left[\sum_{t\in\mathcal{B}_0}\mathbb{1}(Z_t = j)\right] = \sum_{t\in\mathcal{B}_0}\mathbb{P}[Z_t = j] = p_j|\mathcal{B}_0|$, we have

$$\mathbb{P}\left[||\mathcal{B}^j_0| - p_j|\mathcal{B}_0|| \geq \frac{p_j|\mathcal{B}_0|}{2}\right] \leq 2\exp\left(-\frac{p_j|\mathcal{B}_0|}{10}\right),$$

which implies our result. □

LEMMA 10. *Given* $|\mathcal{B}^j_{0,k}|$, *we have*

$$\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{0,k})) \geq \frac{p_*\psi}{4}\right] \geq 1 - d\exp\left(-\frac{p_*\psi|\mathcal{B}^j_{0,k}|}{16dx^2_{\max}}\right) - 2\exp\left(-\frac{p_*|\mathcal{B}^j_{0,k}|}{10}\right).$$

Proof of Lemma 10 We start by defining the events

$$\bar{\mathcal{D}}^j_{0,k} = \left\{ |\bar{\mathcal{B}}^j_{0,k}| \ge \frac{p_*}{2} |\mathcal{B}^j_{0,k}| \right\}.$$

Conditioned on $\bar{\mathcal{D}}^j_{0,k}$ and $\bar{\mathcal{E}}^j_{0,k}$ in (26), Lemma 7 implies $\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{0,k})) \ge (p_*\psi)/4$. Therefore, we have

$$\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{0,k})) \le \frac{p_*\psi}{4}\right] \le \mathbb{P}\left[(\bar{\mathcal{D}}^j_{0,k})^c \cup (\bar{\mathcal{E}}^j_{0,k})^c\right] \le \mathbb{P}\left[(\bar{\mathcal{E}}^j_{0,k})^c \,\middle|\, \bar{\mathcal{D}}^j_{0,k}\right] + \mathbb{P}\left[(\bar{\mathcal{D}}^j_{0,k})^c\right]. \tag{27}$$

By Lemma 8, the second term above is upper bounded by

$$\mathbb{P}\left[(\bar{\mathcal{D}}^j_{0,k})^c\right] \le 2\exp\left(-\frac{p_*|\mathcal{B}^j_{0,k}|}{10}\right).$$

Lemma 6 upper bounds the first term in inequality (27) with

$$\mathbb{P}\left[(\bar{\mathcal{E}}^j_{0,k})^c \,\middle|\, \bar{\mathcal{D}}^j_{0,k}\right] \le \mathbb{E}\left[d\exp\left(-\frac{\psi|\bar{\mathcal{B}}^j_{0,k}|}{8dx^2_{\max}}\right) \,\middle|\, \bar{\mathcal{D}}^j_{0,k}\right] \le d\exp\left(-\frac{p_*\psi|\mathcal{B}^j_{0,k}|}{16dx^2_{\max}}\right). \quad \square$$

Now we are ready to prove Proposition 3.

Proof of Proposition 3 We apply Proposition 2 to our forced-sample estimators built upon the forced samples in $\{\mathcal{B}^i_{0,k}\}_{i\in[N]}$. Based on Lemma 4 and 5, the forced-sample OLS estimators are subgaussian and independent across instances so the conditions of Proposition 2 are satisfied. Letting $\phi = \frac{p_*\psi}{4}$ in Proposition 2, we have

$$\begin{aligned}
&\mathbb{P}\left[\|\widehat{\beta}^j_{k,0} - \beta^j_k\|_1 \ge \frac{24\lambda_j s}{p_*\zeta\psi} + 2C_0 d(3\zeta+4\eta)\max_{i\in[N]}\sqrt{\frac{\sigma_i^2}{p_*\psi|\mathcal{B}^i_{0,k}|}\log(\frac{3}{\eta})}\right] \\
&\quad \le 3d\exp\left(-\frac{N\eta^2}{9}\right) + 2d\exp\left(-\frac{\lambda_j^2|\mathcal{B}^j_{0,k}|}{32\sigma_j^2 x^2_{\max}}\right) + \sum_{i\in[N]}\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^i_{0,k})) \le \frac{p_*\psi}{4}\right] \\
&\quad \le 3d\exp\left(-\frac{N\eta^2}{9}\right) + 2d\exp\left(-\frac{\lambda_j^2|\mathcal{B}^j_{0,k}|}{32\sigma_j^2 x^2_{\max}}\right) \\
&\qquad + \sum_{i\in[N]} d\exp\left(-\frac{p_*\psi|\mathcal{B}^i_{0,k}|}{16dx^2_{\max}}\right) + \sum_{i\in[N]} 2\exp\left(-\frac{p_*|\mathcal{B}^i_{0,k}|}{10}\right),
\end{aligned} \tag{28}$$

where the last inequality uses Lemma 10. By our design of forced sampling, $|\mathcal{B}^j_{0,k}| = |\mathcal{B}^j_0|/K$. Plugging it into the probability bound (28), we have

$$\begin{aligned}
&\mathbb{P}\left[\|\widehat{\beta}^j_{k,0} - \beta^j_k\|_1 \ge \frac{24\lambda_j s}{p_*\zeta\psi} + 2C_0 d(3\zeta+4\eta)\max_{i\in[N]}\sqrt{\frac{K\sigma_i^2}{p_*\psi|\mathcal{B}^i_0|}\log(\frac{3}{\eta})}\right] \\
&\quad \le 3d\exp\left(-\frac{N\eta^2}{9}\right) + 2d\exp\left(-\frac{\lambda_j^2|\mathcal{B}^j_0|}{32K\sigma_j^2 x^2_{\max}}\right) \\
&\qquad + \sum_{i\in[N]} d\exp\left(-\frac{p_*\psi|\mathcal{B}^i_0|}{16Kdx^2_{\max}}\right) + \sum_{i\in[N]} 2\exp\left(-\frac{p_*|\mathcal{B}^i_0|}{10K}\right).
\end{aligned} \tag{29}$$

Now we configure the parameters ζ, λ_j, η and q (recall that $|\mathcal{B}_0| = q\log(T)$) such that our forced-sample estimator of arm k and instance j has estimation error smaller than $\frac{h}{4x_{\max}}$ with a high probability. Similar to the proof of Corollary 1, we take $\zeta = \frac{C_0-2}{4C_0}\sqrt{\frac{s}{d}}$. We further set

$$\eta = \sqrt{\frac{27\log(d)|\mathcal{B}_0|}{qN}}, \quad \lambda_j = \sqrt{\frac{96\sigma_j^2 x^2_{\max} K\log(d)|\mathcal{B}_0|}{q|\mathcal{B}^j_0|}}.$$

In the following, we will frequently use the inequality

$$3\log(T)\log(x) \ge \log(Tx) \tag{30}$$

for $T>1$, $x>1$. Given our choices of η and λ_j and inequality (30), the sum of the first two probabilities on the RHS of (29) is upper bounded by $5/T$. Further, consider the probability bound (29) on the following events for $j \in [N]$

$$\mathcal{M}_0^j = \left\{ |\mathcal{B}_0^j| \ge \frac{p_j}{2}|\mathcal{B}_0| \right\}. \tag{31}$$

Plugging in our choices of ζ, η and λ_j and using a union bound over $\{\mathcal{M}_0^j\}_{j\in[N]}$ through Lemma 9, we have

$$\begin{aligned}
\mathbb{P}\left[\|\widehat{\beta}_{k,0}^j - \beta_k^j\|_1 \ge \frac{768C_0\sigma_j x_{\max}}{(C_0-2)p_*\psi}\sqrt{\frac{3Ksd\log(d)}{qp_j}} + C_0\left(\frac{3\sqrt{sd}}{2} + 24d\sqrt{\frac{3\log(d)|\mathcal{B}_0|}{qN}}\right)\max_{i\in[N]}\sqrt{\frac{K\sigma_i^2\log(N)}{2p_*\psi p_i|\mathcal{B}_0|}} \right] \\
\le \frac{5}{T} + \sum_{i\in[N]} d\exp\left(-\frac{p_*\psi p_i|\mathcal{B}_0|}{32Kdx_{\max}^2}\right) + \sum_{i\in[N]} 2\exp\left(-\frac{p_*p_i|\mathcal{B}_0|}{20K}\right) + \sum_{i\in[N]} 2\exp\left(-\frac{p_i|\mathcal{B}_0|}{10}\right) \\
\le \frac{5}{T} + \sum_{i\in[N]} d\exp\left(-\frac{p_*\psi p_i|\mathcal{B}_0|}{32Kdx_{\max}^2}\right) + \sum_{i\in[N]} 4\exp\left(-\frac{p_*p_i|\mathcal{B}_0|}{20K}\right),
\end{aligned} \tag{32}$$

where we use $\log(\frac{3}{\eta}) \le \frac{\log(N)}{2}$ and the last inequality follows given $K>1$ and $p_* \le 1$. Next, we choose a sufficiently large q such that

$$\begin{aligned}
\frac{h}{8x_{\max}} &\ge \frac{768C_0\sigma_j x_{\max}}{(C_0-2)p_*\psi}\sqrt{\frac{3Ksd\log(d)}{qp_j}}, \\
\frac{h}{16x_{\max}} &\ge C_0\frac{3\sqrt{sd}}{2}\max_{i\in[N]}\sqrt{\frac{K\sigma_i^2\log(N)}{2p_*\psi p_i|\mathcal{B}_0|}}, \\
\frac{h}{16x_{\max}} &\ge 24C_0 d\sqrt{\frac{3\log(d)|\mathcal{B}_0|}{qN}}\max_{i\in[N]}\sqrt{\frac{K\sigma_i^2\log(N)}{2p_*\psi p_i|\mathcal{B}_0|}}, \\
\frac{1}{T} &\ge \sum_{i\in[N]} d\exp\left(-\frac{p_*\psi p_i|\mathcal{B}_0|}{32Kdx_{\max}^2}\right), \\
\frac{4}{T} &\ge \sum_{i\in[N]} 4\exp\left(-\frac{p_*p_i|\mathcal{B}_0|}{20K}\right).
\end{aligned} \tag{33}$$

With a choice of q satisfying all the above constraints, we can derive from (32) our final result. To meet all the conditions in (C.2) for all $j\in[N]$, it suffices to have

$$\begin{aligned}
q = \max\Bigg\{ &\frac{2\cdot 384^3 C_0^2 x_{\max}^4 (\max_{i\in[N]}\sigma_i^2/p_i)Ksd\log(d)}{(C_0-2)^2h^2p_*^2\psi^2}, \frac{576C_0^2x_{\max}^2(\max_{i\in[N]}\sigma_i^2/p_i)Ksd\log(N)}{h^2p_*\psi}, \\
&\frac{6\cdot 192^2C_0^2x_{\max}^2(\max_{i\in[N]}\sigma_i^2/p_i)Kd^2\log(d)\log(N)}{h^2p_*\psi N}, \frac{96x_{\max}^2Kd\log(dN)}{p_*\psi(\min_{i\in[N]}p_i)}, \frac{60K\log(N)}{p_*(\min_{i\in[N]}p_i)} \Bigg\}.
\end{aligned}$$

Remember we also require $\eta \le 1/2 - 1/C_0 - \zeta$ according to Proposition 2, which is satisfied as long as

$$\log(T) \le \left(\frac{1}{2} - \frac{1}{C_0} - \frac{C_0-2}{4C_0}\sqrt{\frac{s}{d}}\right)^2 \frac{N}{27\log(d)}. \tag{34}$$

The result then follows. $\square$

C.3. All-Sample Estimator

We now prove Proposition 4, which says that our all-sample estimators have small error with high probability. First, the constants in the statement of Proposition 4 are

$$C_1 = \frac{1536 C_0 \sigma_j x_{\max}}{(C_0-2) p_*^{3/2} \psi}, \quad C_2 = \frac{3C_0}{2p_*}(\max_{i\in[N]} \sqrt{\frac{2\sigma_i^2 p_j}{\psi p_i}}), \quad C_3 = \frac{24 C_0}{p_*}(\max_{i\in[N]} \sqrt{\frac{2\sigma_i^2 p_j}{\psi \rho p_i}}),$$

and the hyperparameters for all-sample estimators are

$$\zeta_{1,m} = \zeta_{1,0}, \quad \eta_{1,m} = \eta_{1,0}\sqrt{\log(d \min_{i\in[N], |\mathcal{B}^i_{\tilde{m}}|>0} |\mathcal{B}^i_{\tilde{m}}|)}, \quad \lambda_{1,j,m} = \lambda_{1,j,0}\sqrt{\frac{\log(d|\mathcal{B}^j_{\tilde{m}}|)}{|\mathcal{B}^j_{\tilde{m}}|}}$$

as in Algorithm 2. For simplicity, we use ζ, η and λ_j to represent $\zeta_{1,m}$, $\eta_{1,m}$ and $\lambda_{1,j,m}$ respectively in the following.

We begin with the following bound on the probability of event $\mathcal{A}$ (defined in (10)), which says that the forced sample estimators have small errors.

LEMMA 11. *The event $\mathcal{A}$ holds with at least a probability of $1-10KN/T$.*

Proof of Lemma 11 The result follows by applying a union bound over all arms and bandit instances using Proposition 3. □

Next, we define some additional notations. Let $\mathcal{B}^j_m$ be the index set of those observed at instance j in batch m, $\mathcal{B}^j_{m,k}$ be the subset batch sampled at arm k, and $\bar{\mathcal{B}}^j_{m,k}$ be the subset of all $t\in\mathcal{B}^j_m$ such that $X_t \in U^j_k$ when $\mathcal{A}$ holds; in particular,

$$\mathcal{B}^j_m = \{t\in\mathcal{B}_m \mid Z_t = j\}, \quad \mathcal{B}^j_{m,k} = \left\{t\in\mathcal{B}_m \mid Z_t = j, \pi^{Z_t}_{\bar{m-1}}(X_t) = k\right\}, \quad \bar{\mathcal{B}}^j_{m,k} = \left\{t\in\mathcal{B}_m \mid Z_t = j, X_t\in U^{Z_t}_k, \mathcal{A}\right\},$$

where, with a slight abuse of notation, we use $\pi^{Z_t}_{\bar{m-1}}$ to represent the sub-policy of instance j estimated using the data from $\mathcal{B}_{\bar{m-1}}$. We further define $\mathcal{B}_{\tilde{m}} = \bigcup_{l=1}^m \mathcal{B}_l$, $\mathcal{B}^j_{\tilde{m}} = \bigcup_{l=0}^m \mathcal{B}^j_l$, $\mathcal{B}^j_{\tilde{m}} = \bigcup_{l=1}^m \mathcal{B}^j_l$, $\mathcal{B}^j_{\tilde{m},k} = \bigcup_{l=0}^m \mathcal{B}^j_{l,k}$, $\mathcal{B}^j_{\tilde{m},k} = \bigcup_{l=1}^m \mathcal{B}^j_{l,k}$, and $\bar{\mathcal{B}}^j_{\tilde{m},k} = \bigcup_{l=1}^m \bar{\mathcal{B}}^j_{l,k}$. Next, we prove Lemma 1, which says that the samples assigned to arm k collected in batch $\mathcal{B}_m$ (for any $m\geq 1$) are independent across bandit instances conditioned on $\mathcal{F}_{\bar{m-1}}$ defined in Lemma 1.

Proof of Lemma 1 The collected samples of arm k at instance j in the batch $\mathcal{B}_m$ are $\{(X_t,Y_t)\}_{t\in\mathcal{B}^j_{m,k}}$, where the set of covariates is

$$\left\{X_t \,\middle|\, t\in\mathcal{B}^j_{m,k}, Z_t = j, \pi^{Z_t}_{\bar{m-1}}(X_t) = k\right\}.$$

Note that our estimated policy $\pi^{Z_t}_{\bar{m-1}}$ depends on Z_t and is constructed using samples from $\mathcal{B}_{\bar{m-1}}$. Since $\{(X_t,Z_t,Y_t)\}_{t\in\mathcal{B}_{\tilde{m}}}$ are independent, $\{(X_t,Z_t,\pi^{Z_t}_{\bar{m-1}}(X_t))\}_{t\in\mathcal{B}_m}$ are independent conditional on $\mathcal{F}_{\bar{m-1}}$. Thus, for any $t'\neq t$ and $t'\in\mathcal{B}_m$, X_t is independent of $Z_{t'}$, $X_{t'}$ and $\pi^{Z_{t'}}_{\bar{m-1}}(X_{t'})$ conditional on $\{Z_t = j, \pi^{Z_t}_{\bar{m-1}}(X_t) = k, \mathcal{F}_{\pi_{\bar{m-1}}}\}$. This implies X_t's of arm k in batch m are conditionally independent across bandit instances. Moreover, since the noises ϵ_t's are independent of X_t's and only depends on Z_t's by design, the collected samples of arm k in $\mathcal{B}_m$ are independent across different instances conditional on $\mathcal{F}_{\bar{m-1}}$. □

REMARK 2. Note that the samples across instances are also independent given $\{\mathcal{A},\mathcal{F}_{\bar{m-1}}\}$ since $\mathcal{A}\in\mathcal{F}_{\bar{m-1}}$.

LEMMA 12. *For any $j \in [N]$, $k \in [K]$ and $t \notin \mathcal{B}_0$, if $Z_t = j$, $X_t \in U_k^j$ and the event $\mathcal{A}$ holds, then Algorithm 2 plays the optimal arm k of instance j at time t based on the forced-sample estimator $\widehat{\beta}_{k,0}^j$.*

Proof of Lemma 12 Since $X_t \in U_k^j$, by the definition of U_k^j,

$$X_t^\top \beta_k^j \geq \max_{i \neq k} X_t^\top \beta_i^j + h.$$

Then, for any arm $i \neq k$ on the event $\mathcal{A}$, we have

$$X_t^\top(\widehat{\beta}_{k,0}^j - \widehat{\beta}_{i,0}^j) = X_t^\top(\widehat{\beta}_{k,0}^j - \beta_k^j) - X_t^\top(\widehat{\beta}_{i,0}^j - \beta_i^j) + X_t^\top(\beta_k^j - \beta_i^j) \geq -2x_{\max}\frac{h}{4x_{\max}} + h \geq \frac{h}{2}.$$

Therefore, the optimal arm k for X_t will be pulled. $\square$

LEMMA 13. *(i) $\bar{\mathcal{B}}_{m,k}^j \subseteq \mathcal{B}_{m,k}^j$, (ii) $\{X_t\}_{t \in \mathcal{B}_{m,k}^j}$ are i.i.d. from $\mathcal{P}^j_{X|\pi^j_{\bar{m}-1}(X)=k}$ conditioned on $\mathcal{F}_{\bar{m}-1}$, and its subset $\{X_t\}_{t \in \bar{\mathcal{B}}_{m,k}^j}$ are i.i.d. from $\mathcal{P}^j_{X|X\in U_k^j}$, and (iii) $\{X_t\}_{t \in \bar{\mathcal{B}}_{\widetilde{m},k}^j}$ are i.i.d. from $\mathcal{P}^j_{X|X\in U_k^j}$.*

Proof of Lemma 13 The first claim follows Lemma 12. If $Z_t = j$, $X_t \in U_k^j$ and the event $\mathcal{A}$ holds, then $\pi_{\bar{m}-1}^{Z_t}(X_t) = k$ and hence $t \in \mathcal{B}_{m,k}^j$, i.e., $\bar{\mathcal{B}}_{m,k}^j \subseteq \mathcal{B}_{m,k}^j$.

Using a similar argument as in the proof of Lemma 1, we can show that $\{X_t\}_{t \in \mathcal{B}_{m,k}^j}$ are i.i.d. from distribution $\mathcal{P}^j_{X|\pi^j_{\bar{m}-1}(X)=k}$ conditioned on $\mathcal{F}_{\bar{m}-1}$. On the other hand, note that the event $\mathcal{A}$ only depends on forced samples from $\mathcal{B}_0$ and is therefore independent of $\{(X_t, Z_t)\}_{t \in \mathcal{B}_m}$ for any $m \geq 1$. Thus, X_t for any $t \in \bar{\mathcal{B}}_{m,k}^j$ follows distribution $\mathcal{P}^j_{X|X\in U_k^j}$. Furthermore, since $\{(X_t, Z_t)\}_{t \in \mathcal{B}_m}$ are independent, X_t is independent of $Z_{t'}$ and $X_{t'}$ given $\{Z_t = j, X_t \in U_k^{Z_t}, \mathcal{A}\}$ for any $t' \neq t$ and $t' \in \mathcal{B}_m$. Therefore, $\{X_t\}_{t \in \bar{\mathcal{B}}_{m,k}^j}$ are also independent.

Correspondingly, we can show $\{X_t\}_{t \in \bar{\mathcal{B}}_{\widetilde{m},k}^j}$ are also i.i.d. from $\mathcal{P}^j_{X|X\in U_k^j}$ by noting

$$\bar{\mathcal{B}}_{\widetilde{m},k}^j = \cup_{l=1}^m \bar{\mathcal{B}}_{l,k}^j = \left\{t \in \cup_{l=1}^m \mathcal{B}_l \,\middle|\, Z_t = j,\, X_t \in U_k^{Z_t}, \mathcal{A}\right\}. \quad \square$$

REMARK 3. Note that (ii) and (iii) both hold further conditioned on $\mathcal{A}$. The first statement in (ii) still holds as $\mathcal{A} \in \mathcal{F}_{\bar{m}-1}$ while the second statement in (ii) and the statement in (iii) hold as $\{(X_t, Z_t, Y_t)\}_{\mathcal{B}_0}$ are independent of $\{(X_t, Z_t)\}_{\mathcal{B}_{\widetilde{m}}}$.

LEMMA 14. *Given $|\mathcal{B}_m^j|$ and $|\mathcal{B}_{\widetilde{m}}^j|$ respectively, it holds that*

$$\mathbb{P}\left[|\bar{\mathcal{B}}_{m,k}^j| \geq \frac{p_*|\mathcal{B}_m^j|}{2} \,\middle|\, \mathcal{A}\right] \geq 1 - 2\exp\left(-\frac{p_*|\mathcal{B}_m^j|}{10}\right), \quad \mathbb{P}\left[|\bar{\mathcal{B}}_{\widetilde{m},k}^j| \geq \frac{p_*|\mathcal{B}_{\widetilde{m}}^j|}{2} \,\middle|\, \mathcal{A}\right] \geq 1 - 2\exp\left(-\frac{p_*|\mathcal{B}_{\widetilde{m}}^j|}{10}\right).$$

Proof of Lemma 14 By definition of $\bar{\mathcal{B}}_{m,k}^j$ and given $\mathcal{B}_m^j$, we have

$$|\bar{\mathcal{B}}_{m,k}^j| = \sum_{t \in \mathcal{B}_m^j} \mathbb{1}\left(t \in \bar{\mathcal{B}}_{m,k}^j\right) = \sum_{t \in \mathcal{B}_m^j} \mathbb{1}\left(Z_t = j, X_t \in U_k^{Z_t}\right)\mathbb{1}(\mathcal{A}).$$

Then, take

$$\mu = \mathbb{E}\left[|\bar{\mathcal{B}}_{m,k}^j| \,\middle|\, \mathcal{A}\right] = \mathbb{E}\left[\sum_{t \in \mathcal{B}_m^j} \mathbb{1}\left(Z_t = j, X_t \in U_k^{Z_t}\right)\right] = \sum_{t \in \mathcal{B}_m^j} \mathbb{P}\left[X_t \in U_k^j \mid Z_t = j\right],$$

where the second equality is from the fact that $\{(X_t, Z_t)\}_{t \in \mathcal{B}_m}$ are independent of $\{(X_t, Z_t, Y_t)\}_{t \in \mathcal{B}_0}$. Using Lemma 30, we have

$$\mathbb{P}\left[||\bar{\mathcal{B}}_{m,k}^j| - \mu| \geq \frac{\mu}{2} \,\middle|\, \mathcal{A}\right] \leq 2\exp\left(-\frac{\mu}{10}\right).$$

Therefore, using $\mu \geq p_*|\mathcal{B}_m^j|$ by Assumption 4, we have

$$\mathbb{P}\left[|\bar{\mathcal{B}}_{m,k}^j| \geq \frac{p_*|\mathcal{B}_m^j|}{2} \,\middle|\, \mathcal{A}\right] \geq 1 - 2\exp\left(-\frac{p_*|\mathcal{B}_m^j|}{10}\right).$$

The high probability bound regarding $|\bar{\mathcal{B}}_{\widetilde{m},k}^j|$ can be proved in a same manner. $\square$

LEMMA 15. *It holds that*

$$\mathbb{P}\left[|\mathcal{B}_m^j| \geq \frac{p_j|\mathcal{B}_m|}{2}\right] \geq 1 - 2\exp\left(-\frac{p_j|\mathcal{B}_m|}{10}\right), \quad \mathbb{P}\left[|\mathcal{B}_{\widetilde{m}}^j| \geq \frac{p_j|\mathcal{B}_{\widetilde{m}}|}{2}\right] \geq 1 - 2\exp\left(-\frac{p_j|\mathcal{B}_{\widetilde{m}}|}{10}\right),$$
$$\mathbb{P}\left[\frac{3p_j|\mathcal{B}_{\bar{m}}|}{2} \geq |\mathcal{B}_{\bar{m}}^j| \geq \frac{p_j|\mathcal{B}_{\bar{m}}|}{2}\right] \geq 1 - 2\exp\left(-\frac{p_j|\mathcal{B}_{\bar{m}}|}{10}\right).$$

Proof of Lemma 15 The result follows correspondingly the proof of Lemma 9. □

REMARK 4. Note that the high probability bound of $|\mathcal{B}_m^j|$ also holds conditioned on $\mathcal{A}$ as $\{Z_t\}_{t\in\mathcal{B}_m}$ are independent of $\{(X_t, Z_t, Y_t)\}_{t\in\mathcal{B}_0}$.

LEMMA 16. *Given $|\mathcal{B}_m^j|$, we have*

$$\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}_{m,k}^j)) \geq \frac{p_*\psi}{4} \,\middle|\, \mathcal{A}, \mathcal{F}_{m-1}^-\right] \geq 1 - d\exp\left(-\frac{p_*\psi|\mathcal{B}_m^j|}{16dx_{\max}^2}\right) - 2\exp\left(-\frac{p_*|\mathcal{B}_m^j|}{10}\right).$$

When $T \geq 20KN$ and given $|\mathcal{B}_{\widetilde{m}}^j|$ and $|\mathcal{B}_0^j|$, we futher have

$$\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}_{\widetilde{m},k}^j)) \leq \frac{p_*\psi}{4} \,\middle|\, \mathcal{A}\right] \leq d\exp\left(-\frac{p_*\psi|\mathcal{B}_{\widetilde{m}}^j|}{16dx_{\max}^2}\right) + 2\exp\left(-\frac{p_*|\mathcal{B}_{\widetilde{m}}^j|}{10}\right)$$
$$+ 2d\exp\left(-\frac{p_*\psi|\mathcal{B}_0^j|}{16Kdx_{\max}^2}\right) + 4\exp\left(-\frac{p_*|\mathcal{B}_0^j|}{10K}\right).$$

Proof of Lemma Define the following events analogous to Lemma 10

$$\bar{\mathcal{D}}_{m,k}^j = \left\{|\bar{\mathcal{B}}_{m,k}^j| \geq \frac{p_*}{2}|\mathcal{B}_m^j|\right\}, \quad \bar{\mathcal{E}}_{m,k}^j = \left\{\lambda_{\min}(\widehat{\Sigma}(\bar{\mathcal{B}}_{m,k}^j)) \geq \frac{\psi}{2}\right\}.$$

By Lemma 7 and the fact that $|\mathcal{B}_m^j| \geq |\mathcal{B}_{m,k}^j|$, it holds that $\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}_{m,k}^j)) \geq (p_*\psi)/4$ on the above events $\bar{\mathcal{D}}_{m,k}^j$ and $\bar{\mathcal{E}}_{m,k}^j$. Thus, we have

$$\begin{aligned}\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}_{m,k}^j)) \leq \frac{p_*\psi}{4} \,\middle|\, \mathcal{A}\right] &\leq \mathbb{P}\left[(\bar{\mathcal{D}}_{m,k}^j)^c \cup (\bar{\mathcal{E}}_{m,k}^j)^c \,\middle|\, \mathcal{A}\right] \\ &\leq \mathbb{P}\left[(\bar{\mathcal{E}}_{m,k}^j)^c \,\middle|\, \bar{\mathcal{D}}_{m,k}^j, \mathcal{A}\right] + \mathbb{P}\left[(\bar{\mathcal{D}}_{m,k}^j)^c \,\middle|\, \mathcal{A}\right].\end{aligned}$$

By Lemma 13 and analogous to Lemma 6, the first term above has

$$\mathbb{P}\left[(\bar{\mathcal{E}}_{m,k}^j)^c \,\middle|\, \bar{\mathcal{D}}_{m,k}^j, \mathcal{A}\right] \leq d\exp\left(-\frac{p_*\psi|\mathcal{B}_m^j|}{16dx_{\max}^2}\right).$$

In addition, Lemma 14 implies

$$\mathbb{P}\left[(\bar{\mathcal{D}}_{m,k}^j)^c \,\middle|\, \mathcal{A}\right] \leq 2\exp\left(-\frac{p_*|\mathcal{B}_m^j|}{10}\right).$$

This gives us our first high probability bound.

Next, define the following events correspondingly regarding the batched samples (excluding the forced ones):

$$\bar{\mathcal{D}}_{\widetilde{m},k}^j = \left\{|\bar{\mathcal{B}}_{\widetilde{m},k}^j| \geq \frac{p_*}{2}|\mathcal{B}_{\widetilde{m}}^j|\right\}, \quad \bar{\mathcal{E}}_{\widetilde{m},k}^j = \left\{\lambda_{\min}(\widehat{\Sigma}(\bar{\mathcal{B}}_{\widetilde{m},k}^j)) \geq \frac{\psi}{2}\right\}.$$

On the above events $\bar{\mathcal{D}}_{\widetilde{m},k}^j$ and $\bar{\mathcal{E}}_{\widetilde{m},k}^j$, we have $\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}_{\widetilde{m},k}^j)) \geq (p_*\psi)/4$. With a similar argument as above, we can write

$$\begin{aligned}\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}_{\widetilde{m},k}^j)) \leq \frac{p_*\psi}{4} \,\middle|\, \mathcal{A}\right] &\leq \mathbb{P}\left[(\bar{\mathcal{D}}_{\widetilde{m},k}^j)^c \cup (\bar{\mathcal{E}}_{\widetilde{m},k}^j)^c \,\middle|\, \mathcal{A}\right] \\ &\leq \mathbb{P}\left[(\bar{\mathcal{E}}_{\widetilde{m},k}^j)^c \,\middle|\, \bar{\mathcal{D}}_{\widetilde{m},k}^j, \mathcal{A}\right] + \mathbb{P}\left[(\bar{\mathcal{D}}_{\widetilde{m},k}^j)^c \,\middle|\, \mathcal{A}\right] \\ &\leq d\exp\left(-\frac{p_*\psi|\mathcal{B}_{\widetilde{m}}^j|}{16dx_{\max}^2}\right) + 2\exp\left(-\frac{p_*|\mathcal{B}_{\widetilde{m}}^j|}{10}\right).\end{aligned}$$

Then, note that $\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{\bar{m},k})) \geq (p_*\psi)/4$ when both $\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{\widetilde{m},k})) \geq (p_*\psi)/4$ and $\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{0,k})) \geq (p_*\psi)/4$ hold using Lemma 7. Thus, we have

$$\begin{aligned}
\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{\bar{m},k})) \leq \frac{p_*\psi}{4} \,\middle|\, \mathcal{A}\right] &\leq \mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{\widetilde{m},k})) \leq \frac{p_*\psi}{4} \,\middle|\, \mathcal{A}\right] + \mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{0,k})) \leq \frac{p_*\psi}{4} \,\middle|\, \mathcal{A}\right] \\
&\leq \mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{\widetilde{m},k})) \leq \frac{p_*\psi}{4} \,\middle|\, \mathcal{A}\right] + \frac{\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{0,k})) \leq \frac{p_*\psi}{4}\right]}{\mathbb{P}[\mathcal{A}]} \\
&\leq d\exp\left(-\frac{p_*\psi|\mathcal{B}^j_{\widetilde{m}}|}{16dx^2_{\max}}\right) + 2\exp\left(-\frac{p_*|\mathcal{B}^j_{\widetilde{m}}|}{10}\right) \\
&\qquad + 2d\exp\left(-\frac{p_*\psi|\mathcal{B}^j_0|}{16Kdx^2_{\max}}\right) + 4\exp\left(-\frac{p_*|\mathcal{B}^j_0|}{10K}\right),
\end{aligned}$$

where the third inequality uses Lemma 10 and the fact that $\mathbb{P}[\mathcal{A}] \geq 1/2$ when $T \geq 20KN$. $\square$

Next, we show that Lemma 2 also holds for an adapted sequence of observations so that we can use all samples collected at arm k and instance j to debias in Step 2 for our all-sample estimators. We keep the same notations as in Lemma 2 but use $(\mathbf{X}^j, Y^j)$ to represent all the data collected at arm k and instance j through Algorithm 2 up to batch m, i.e., $\{(X_t, Y_t)\}_{t\in\mathcal{B}^j_{\bar{m},k}}$.

Lemma 17. *Define the event for an adapted sequence of observations in* $\mathcal{B}^j_{\bar{m},k}$

$$\mathcal{H}^j(\mathcal{B}^j_{\bar{m},k}) = \left\{\frac{2}{|\mathcal{B}^j_{\bar{m},k}|}\|\mathbf{X}^{j\top}\epsilon^j\|_\infty \leq \frac{\lambda_j}{2}\right\}. \tag{35}$$

Then, we have

$$\mathbb{P}\left[\mathcal{H}^j(\mathcal{B}^j_{\bar{m},k})\right] \geq 1 - 2d\exp\left(-\frac{\lambda_j^2|\mathcal{B}^j_{\bar{m},k}|}{32\sigma_j^2 x^2_{\max}}\right).$$

Proof of Lemma 17 Let $\mathcal{F}_t = \sigma(\{X_k, Z_k, Y_k\}_{k\in[t]})$ be the σ-algebra generated by an adapted sequence of random variables up to time t. For the i^{th} element of X_t, i.e., $X_{t,(i)}$, let $D_{t,(i)} = X_{t,(i)}\epsilon_t$ and $\{D_t\}_{t\in\mathcal{B}^j_{\bar{m},k}}$ is a martingale difference sequence adapted to the filtration $\{\mathcal{F}_t\}_{t\in\mathcal{B}^j_{\bar{m},k}}$ respectively as $\mathbb{E}[X_{t,(i)}\epsilon_t \mid \mathcal{F}_{t-1}] = 0$. Besides, $D_{t,(i)}$ is subgaussian adapted to $\mathcal{F}_{t-1}$ as

$$\mathbb{E}[\exp(\lambda D_{t,(i)}) \mid \mathcal{F}_{t-1}] \leq \mathbb{E}[\exp(\lambda^2 X^2_{t,(i)}\sigma_j^2/2) \mid \mathcal{F}_{t-1}] \leq \exp(\lambda^2 x^2_{\max}\sigma_j^2/2).$$

Then, by Lemma 28, we have

$$\begin{aligned}
\mathbb{P}\left[(\mathcal{H}^j(\mathcal{B}^j_{\bar{m},k}))^c\right] &= \mathbb{P}\left[\max_{i\in[d]}\left|\sum_{t\in\mathcal{B}^j_{\bar{m},k}} X_{t,(i)}\epsilon_t\right| \geq \frac{\lambda_j|\mathcal{B}^j_{\bar{m},k}|}{4}\right] \\
&\leq d\max_{i\in[d]}\mathbb{P}\left[\left|\sum_{t\in\mathcal{B}^j_{\bar{m},k}} X_{t,(i)}\epsilon_t\right| \geq \frac{\lambda_j|\mathcal{B}^j_{\bar{m},k}|}{4}\right] \\
&\leq 2d\exp\left(-\frac{\lambda_j^2|\mathcal{B}^j_{\bar{m},k}|}{32\sigma_j^2 x^2_{\max}}\right). \quad \square
\end{aligned}$$

Now we provide the proof of Proposition 4.

Proof of Proposition 4 We follow a similar proof strategy as Proposition 3 to provide an error bound of all-sample estimators conditioned on $\mathcal{A}$. Now we consider learning across a set of instances that has arm k to be an optimal arm, i.e., $\mathcal{W}_k \subseteq [N]$, since suboptimal arms won't observe any users on the event of $\mathcal{A}$. We apply Proposition 2 to our all-sample estimators across $\mathcal{W}_k$.

Based on Lemma 1 and 13, our all-sample OLS estimators are subgaussian and independent across instances conditioned on $\{\mathcal{A}, \mathcal{F}_{\bar{m}-1}\}$ so our trimmed mean estimator in Step 1 is valid. Here we make a small adjustment to Proposition 2 since we use batched data from $\{\mathcal{B}^i_{m,k}\}_{i\in[N]}$ to compute our trimmed mean estimator but all the data in $\mathcal{B}^j_{\bar{m},k}$ to debias for instance j in Step 2. In particular, we bound the event $\mathcal{H}^j(\mathcal{B}^j_{\bar{m},k})$ in (35) in contrast to $\mathcal{H}^j(\mathcal{B}^j_{m,k})$ of only samples from batch m, and bound an extra event that the sample covariance matrix based on data from $\mathcal{B}^j_{\bar{m},k}$ has positive minimum eigenvalue, i.e., $\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{\bar{m},k})) \geq (p_*\psi)/4$. Therefore, we can write

$$
\begin{aligned}
&\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}} - \beta^j_k\|_1 \geq \frac{24\lambda_j s}{p_*\zeta\psi} + 2C_0 d(3\zeta+4\eta)\max_{i\in\mathcal{W}_k}\sqrt{\frac{\sigma_i^2}{p_*\psi|\mathcal{B}^i_{m,k}|}\log(\frac{3}{\eta})}\,\middle|\,\mathcal{A},\mathcal{F}_{\bar{m}-1}\right] \\
&\quad\leq 3d\exp\left(-\frac{\rho N\eta^2}{9}\right) + \mathbb{P}\left[(\mathcal{H}^j(\mathcal{B}^j_{\bar{m},k}))^c \,\middle|\, \mathcal{A},\mathcal{F}_{\bar{m}-1}\right] + \mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^j_{\bar{m},k})) \leq \frac{p_*\psi}{4}\,\middle|\,\mathcal{A},\mathcal{F}_{\bar{m}-1}\right] \\
&\qquad + \sum_{i\in\mathcal{W}_k}\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}(\mathcal{B}^i_{m,k})) \leq \frac{p_*\psi}{4}\,\middle|\,\mathcal{A},\mathcal{F}_{\bar{m}-1}\right], \qquad (36)
\end{aligned}
$$

Note that given $T \geq 20KN$, we have

$$
\mathbb{P}\left[(\mathcal{H}^j(\mathcal{B}^j_{\bar{m},k}))^c \,\middle|\, \mathcal{A}\right] \leq \frac{\mathbb{P}\left[(\mathcal{H}^j(\mathcal{B}^j_{\bar{m},k}))^c\right]}{\mathbb{P}[\mathcal{A}]} \leq 2\mathbb{P}\left[(\mathcal{H}^j(\mathcal{B}^j_{\bar{m},k}))^c\right]. \qquad (37)
$$

Taking expectation over $\mathcal{F}_{\bar{m}-1}$ and using inequality (37), we obtain from (36) that

$$
\begin{aligned}
&\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}} - \beta^j_k\|_1 \geq \frac{24\lambda_j s}{p_*\zeta\psi} + 2C_0 d(3\zeta+4\eta)\max_{i\in\mathcal{W}_k}\sqrt{\frac{\sigma_i^2}{p_*\psi|\mathcal{B}^i_{m,k}|}\log(\frac{3}{\eta})}\,\middle|\,\mathcal{A}\right] \\
&\quad\leq 3d\exp\left(-\frac{\rho N\eta^2}{9}\right) + 4d\exp\left(-\frac{\lambda_j^2|\mathcal{B}^j_{\bar{m},k}|}{32\sigma_j^2 x_{\max}^2}\right) + d\exp\left(-\frac{p_*\psi|\mathcal{B}^j_{\bar{m}}|}{16dx_{\max}^2}\right) + 2\exp\left(-\frac{p_*|\mathcal{B}^j_{\bar{m}}|}{10}\right) \\
&\qquad + 2d\exp\left(-\frac{p_*\psi|\mathcal{B}^j_0|}{16Kdx_{\max}^2}\right) + 4\exp\left(-\frac{p_*|\mathcal{B}^j_0|}{10K}\right) \\
&\qquad + \sum_{i\in\mathcal{W}_k} d\exp\left(-\frac{p_*\psi|\mathcal{B}^i_m|}{16dx_{\max}^2}\right) + \sum_{i\in\mathcal{W}_k} 2\exp\left(-\frac{p_*|\mathcal{B}^i_m|}{10}\right).
\end{aligned}
$$

where we apply Lemma 16 and 17. Then, define the events

$$
\mathcal{D}^j_{m,k} = \left\{|\mathcal{B}^j_{m,k}| \geq \frac{p_*}{2}|\mathcal{B}^j_m|\right\}, \quad \mathcal{D}^j_{\bar{m},k} = \left\{|\mathcal{B}^j_{\bar{m},k}| \geq \frac{p_*}{2}|\mathcal{B}^j_{\bar{m}}|\right\}.
$$

With a union bound over the events $\bigcap_{i\in\mathcal{W}_k}\mathcal{D}^i_{m,k}\cap\mathcal{D}^j_{\bar{m},k}$ conditioned on $\mathcal{A}$, we have

$$
\begin{aligned}
&\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}} - \beta^j_k\|_1 \geq \frac{24\lambda_j s}{p_*\zeta\psi} + 2C_0 d(3\zeta+4\eta)\max_{i\in\mathcal{W}_k}\sqrt{\frac{2\sigma_i^2}{p_*^2\psi|\mathcal{B}^i_m|}\log(\frac{3}{\eta})}\,\middle|\,\mathcal{A}\right] \\
&\quad\leq 3d\exp\left(-\frac{\rho N\eta^2}{9}\right) + 4d\exp\left(-\frac{\lambda_j^2 p_*|\mathcal{B}^j_{\bar{m}}|}{64\sigma_j^2 x_{\max}^2}\right) + d\exp\left(-\frac{p_*\psi|\mathcal{B}^j_{\bar{m}}|}{16dx_{\max}^2}\right) + 4\exp\left(-\frac{p_*|\mathcal{B}^j_{\bar{m}}|}{10}\right) \\
&\qquad + 2d\exp\left(-\frac{p_*\psi|\mathcal{B}^j_0|}{16Kdx_{\max}^2}\right) + 4\exp\left(-\frac{p_*|\mathcal{B}^j_0|}{10K}\right) \\
&\qquad + \sum_{i\in\mathcal{W}_k} d\exp\left(-\frac{p_*\psi|\mathcal{B}^i_m|}{16dx_{\max}^2}\right) + \sum_{i\in\mathcal{W}_k} 4\exp\left(-\frac{p_*|\mathcal{B}^i_m|}{10}\right), \qquad (38)
\end{aligned}
$$

where we use Lemma 14 by noting that $|\mathcal{B}^j_{m,k}| \geq |\bar{\mathcal{B}}^j_{m,k}|$ and $|\mathcal{B}^j_{\bar{m},k}| \geq |\mathcal{B}^j_{\widetilde{m},k}| \geq |\bar{\mathcal{B}}^j_{\widetilde{m},k}|$. Similarly, we take $\zeta = \frac{C_0-2}{4C_0}\sqrt{\frac{s}{d}}$ and set

$$\eta = \sqrt{\frac{9\log(d\min_{i\in\mathcal{W}_k}|\mathcal{B}^i_m|)}{\rho N}}, \quad \lambda_j = \sqrt{\frac{256\sigma_j^2 x_{\max}^2 \log(d|\mathcal{B}^j_{\bar{m}}|)}{p_*|\mathcal{B}^j_{\bar{m}}|}}.$$

Since $|\mathcal{B}^i_m| = 0$ for any $i \in [N]\setminus\mathcal{W}_k$ conditioned on the event $\mathcal{A}$, the value of η is equivalent to

$$\eta = \sqrt{\frac{9\log(d\min_{i\in[N],|\mathcal{B}^i_m|>0}|\mathcal{B}^i_m|)}{\rho N}}.$$

As $\log(d\min_{i\in[N],|\mathcal{B}^i_m|>0}|\mathcal{B}^i_m|) \geq \log(d) \geq 1$, it holds that $\sqrt{\log(\frac{3}{\eta})} \leq \sqrt{\frac{\log(\rho N)}{2}}$. Next, consider $\mathcal{M}^j_0$ defined in (31) and the events

$$\mathcal{M}^j_m = \left\{|\mathcal{B}^j_m| \geq \frac{p_j}{2}|\mathcal{B}_m|\right\}, \quad \mathcal{M}^j_{\widetilde{m}} = \left\{|\mathcal{B}^j_{\widetilde{m}}| \geq \frac{p_j}{2}|\mathcal{B}_{\widetilde{m}}|\right\}, \quad \mathcal{M}^j_{\bar{m}} = \left\{\frac{3p_j}{2}|\mathcal{B}_{\bar{m}}| \geq |\mathcal{B}^j_{\bar{m}}| \geq \frac{p_j}{2}|\mathcal{B}_{\bar{m}}|\right\}.$$

Using a union bound over $\bigcap_{i\in\mathcal{W}_k}\mathcal{M}^i_m \cap \mathcal{M}^j_0 \cap \mathcal{M}^j_{\widetilde{m}} \cap \mathcal{M}^j_{\bar{m}}$ through Lemma 9 and 15 on inequality (38), we obtain

$$\begin{aligned}
&\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}} - \beta^j_k\|_1 \geq C_1\sqrt{\frac{sd\log(dp_j|\mathcal{B}_{\bar{m}}|/2)}{p_j|\mathcal{B}_{\bar{m}}|}} + C_2\sqrt{\frac{sd\log(\rho N)}{p_j|\mathcal{B}_m|}} + C_3 d\sqrt{\frac{\log(dp_j|\mathcal{B}_m|/2)\log(\rho N)}{Np_j|\mathcal{B}_m|}}\,\middle|\,\mathcal{A}\right] \\
&\quad\leq \frac{6}{\min_{i\in\mathcal{W}_k}p_i|\mathcal{B}_m|} + \frac{8}{p_j|\mathcal{B}_{\bar{m}}|} + d\exp\left(-\frac{p_*p_j\psi|\mathcal{B}_{\widetilde{m}}|}{32dx_{\max}^2}\right) + 4\exp\left(-\frac{p_*p_j|\mathcal{B}_{\widetilde{m}}|}{20}\right) + 2d\exp\left(-\frac{p_*p_j\psi|\mathcal{B}_0|}{32Kdx_{\max}^2}\right) \\
&\qquad + 4\exp\left(-\frac{p_*p_j|\mathcal{B}_0|}{20K}\right) + \sum_{i\in\mathcal{W}_k} d\exp\left(-\frac{p_*p_i\psi|\mathcal{B}_m|}{32dx_{\max}^2}\right) + \sum_{i\in\mathcal{W}_k}4\exp\left(-\frac{p_*p_i|\mathcal{B}_m|}{20}\right),
\end{aligned}$$

where C_1, C_2, C_3 are constants that are listed at the beginning of §C.3, and we use the facts that $\log(x)/x$ is monotonically decreasing when $x \geq 3$ and that $|\mathcal{B}_{\widetilde{m}}|/|\mathcal{B}_{\bar{m}}| = 1 - 1/2^{m+1} \geq 3/4$ when $m \geq 1$. Given our choice of q in Proposition 3, the two probability terms regarding forced batch can be upper bounded by

$$2d\exp\left(-\frac{p_*p_j\psi|\mathcal{B}_0|}{32Kdx_{\max}^2}\right) + 4\exp\left(-\frac{p_*p_j|\mathcal{B}_0|}{20K}\right) \leq \frac{6}{TN}.$$

Combined with the fact that $|\mathcal{B}_{\widetilde{m}}| \geq |\mathcal{B}_m|$, we can further simplify the formula into

$$\begin{aligned}
&\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}} - \beta^j_k\|_1 \geq C_1\sqrt{\frac{sd\log(dp_j|\mathcal{B}_{\bar{m}}|/2)}{p_j|\mathcal{B}_{\bar{m}}|}} + C_2\sqrt{\frac{sd\log(\rho N)}{p_j|\mathcal{B}_m|}} + C_3 d\sqrt{\frac{\log(dp_j|\mathcal{B}_m|/2)\log(\rho N)}{Np_j|\mathcal{B}_m|}}\,\middle|\,\mathcal{A}\right] \\
&\qquad\leq \frac{6}{\min_{i\in\mathcal{W}_k}p_i|\mathcal{B}_m|} + \frac{8}{p_j|\mathcal{B}_{\bar{m}}|} + \frac{6}{TN} + \sum_{i\in\mathcal{W}_k}2d\exp\left(-\frac{p_*p_i\psi|\mathcal{B}_m|}{32dx_{\max}^2}\right) + \sum_{i\in\mathcal{W}_k}8\exp\left(-\frac{p_*p_i|\mathcal{B}_m|}{20}\right).
\end{aligned}$$

Then we get our final tail inequality.

In addition, to satisfy $\eta \leq 1/2 - 1/C_0 - \zeta$, we require

$$\log(d|\mathcal{B}_m|) \leq \left(\frac{1}{2} - \frac{1}{C_0} - \frac{C_0-2}{4C_0}\sqrt{\frac{s}{d}}\right)^2\frac{\rho N}{9}.$$

Since $|\mathcal{B}_m| \leq T$, this combined with (34) gives $N = \Omega(\log(d)\log(T))$. $\square$

C.4. Regret Bound for RMBandit

Finally, we prove Theorem 2, which provides an upper bound on the cumulative regret across all bandit instances. We start by providing a worst-case regret bound for Case (I).

LEMMA 18. *The cumulative expected regret from the first two batches $\mathcal{B}_0 \cup \mathcal{B}_1$ or when $T \leq KN$ is at most*

$$2bx_{\max}(2q\log(T) + KN).$$

Proof of Lemma 18 By our design, we have $2q\log(T)$ time steps in total from $\mathcal{B}_0$ and $\mathcal{B}_1$. The worst-case regret per step is $2bx_{\max}$. The result then follows. $\square$

Next, we discuss the worst-case regret bound for Case (II).

LEMMA 19. *When $T \geq KN$ and $\mathcal{A}$ does not hold, the cumulative expected regret from all the batches $\{\mathcal{B}_m\}_{m>1}$ up to time T is at most*

$$20bx_{\max}KN.$$

Proof of Lemma 19 By Lemma 11, the probability of a failure of $\mathcal{A}$ is at most $(10KN)/T$. The worst-case cumulative regret is at most $2bx_{\max}T$ throughout $\{\mathcal{B}_m\}_{m>1}$. The result then follows. $\square$

Before we proceed to the regret analysis of Case (III), we first provide the following helpful lemma, which shows a sufficient amount of data is employed to train the all-sample estimators.

LEMMA 20. *For any $t \in \mathcal{B}_m$ with $m > 1$, we have*

$$|\mathcal{B}_{m-1}| \geq \frac{t}{4}, \quad |\mathcal{B}_{m\bar{-}1}| \geq \frac{t}{2}.$$

Proof of Lemma 20 By our design, $|\mathcal{B}_m| = 2^{m-1}|\mathcal{B}_0|$ for any $m \geq 1$, which implies for any $m > 1$

$$\frac{|\mathcal{B}_{m-1}|}{t} \geq \frac{|\mathcal{B}_{m-1}|}{\sum_{i=0}^{m}|\mathcal{B}_i|} = \frac{1}{4}, \quad \frac{|\mathcal{B}_{m\bar{-}1}|}{t} \geq \frac{\sum_{i=0}^{m-1}|\mathcal{B}_i|}{\sum_{i=0}^{m}|\mathcal{B}_i|} = \frac{1}{2}. \quad \square$$

The second lemma in the following shows that the forced-sample estimators filter out suboptimal arms in Algorithm 2. Therefore, we can apply our high probability bound of all-sample estimators in Proposition 4 to the regret analysis of Case (III).

LEMMA 21. *If $\mathcal{A}$ holds, then the set of arms $\mathcal{K}$ that survive after using the forced-sample estimators contains the optimal arm $k = \arg\max_{i\in[K]} X_t^\top \beta_i^j$ given $Z_t = j$ and no suboptimal arms in $\mathcal{K}_{sub}^j$.*

Proof of Lemma 21 Similar to the proof of Lemma 12, given $\mathcal{A}$, we have for any arm i

$$X_t^\top(\widehat{\beta}_{k,0}^j - \widehat{\beta}_{i,0}^j) \geq -\frac{h}{2} + X_t^\top(\beta_k^j - \beta_i^j) \geq -\frac{h}{2}.$$

Thus, we have

$$X_t^\top \widehat{\beta}_{k,0}^j \geq \max_{i\in[K]} X_t^\top \widehat{\beta}_{i,0}^j - \frac{h}{2}.$$

In other words, the optimal arm will be kept based on the forced-sample estimators in Algorithm 2.

Now consider any suboptimal arm $k' \in \mathcal{K}_{\text{sub}}^j$. By definition, we have $X_t^\top(\beta_i^j - \beta_{k'}^j) \geq h$ for any arm i. Therefore,

$$X_t^\top(\widehat{\beta}_{i,0}^j - \widehat{\beta}_{k',0}^j) \geq -\frac{h}{2} + X_t^\top(\beta_i^j - \beta_{k'}^j) \geq \frac{h}{2},$$

which implies

$$X_t^\top \widehat{\beta}_{k',0}^j \leq \max_{i\in[K]} X_t^\top \widehat{\beta}_{i,0}^j - \frac{h}{2}.$$

In other words, any suboptimal arm k' will be filtered out through the forced-sample estimators and hence not in $\mathcal{K}$. □

Next, we provide a per-period regret bound when a user is observed at time t and instance j in Case (III).

Lemma 22. *When $\mathcal{A}$ holds, $T \geq KN$, $N = \Omega(\log(d)\log(T))$ and $Z_t = j$, the expected regret at time $t \in \mathcal{B}_m$ with $m > 1$ is upper bounded by*

$$\begin{aligned} r_t^j \leq\ & 24x_{\max}^2 LK\left(C_1^2 \frac{sd\log(dp_jt)}{p_jt} + 2C_2^2\frac{sd\log(\rho N)}{p_jt} + 2C_3^2\frac{d^2\log(\rho N)\log(dp_jt)}{Np_jt}\right) \\ & + 4bx_{\max}K\left(\max_{i\in[N]}\frac{24}{p_it} + \frac{16}{p_jt} + \frac{6}{TN} + 2dN\exp\left(-\frac{p_*\psi(\min_{i\in[N]}p_i)t}{128dx_{\max}^2}\right) + 8N\exp\left(-\frac{p_*(\min_{i\in[N]}p_i)t}{80}\right)\right). \end{aligned}$$

Proof of Lemma 22 Without loss of generality, assume arm 1 is optimal for X_t, i.e., $\arg\max_{k\in[K]} X_t^\top\beta_k^j = 1$. Note that here the optimal arm is a function of X_t and hence a random variable, though for simplicity we fix arm 1 as the optimal arm in the following. Consider the conditional expected regret at time $t \in \mathcal{B}_m$

$$r_t^j(X_t) = \mathbb{E}\left[\sum_{k\in\mathcal{K}} X_t^\top(\beta_1^j - \beta_k^j)\mathbb{1}\left(\pi_{m-1}^{j}(X_t) = k\right)\middle| X_t, Z_t = j, \mathcal{A}\right],$$

where the set of arms $\mathcal{K}$ is defined in Algorithm 2. Since $\pi_{m-1}^j(X_t) = k$ implies $X_t^\top\widehat{\beta}_{k,m-1}^j \geq X_t^\top\widehat{\beta}_{1,m-1}^j$, we have

$$r_t^j(X_t) \leq \mathbb{E}\left[\sum_{k\in\mathcal{K}} X_t^\top(\beta_1^j - \beta_k^j)\mathbb{1}\left(X_t^\top\widehat{\beta}_{k,m-1}^j \geq X_t^\top\widehat{\beta}_{1,m-1}^j\right)\middle| X_t, Z_t = j, \mathcal{A}\right]. \tag{39}$$

To bound the above expectation, define the events

$$\mathcal{L}_k^j = \left\{2x_{\max}\delta \leq X_t^\top(\beta_1^j - \beta_k^j)\right\}.$$

Then, we can decompose the upper bound of the regret in (39) into two parts given $\mathcal{L}_k^j$

$$r_t^j(X_t) \leq \sum_{r=1,2} r_{t,r}^j(X_t), \tag{40}$$

where

$$r_{t,1}^j(X_t) = \mathbb{E}\left[\sum_{k\in\mathcal{K}} X_t^\top(\beta_1^j - \beta_k^j)\mathbb{1}\left(\{X_t^\top\widehat{\beta}_{k,m-1}^j \geq X_t^\top\widehat{\beta}_{1,m-1}^j\} \cap \mathcal{L}_k^j\right)\middle| X_t, Z_t = j, \mathcal{A}\right], \tag{41}$$

$$r_{t,2}^j(X_t) = \mathbb{E}\left[\sum_{k\in\mathcal{K}} X_t^\top(\beta_1^j - \beta_k^j)\mathbb{1}\left(\{X_t^\top\widehat{\beta}_{k,m-1}^j \geq X_t^\top\widehat{\beta}_{1,m-1}^j\} \cap (\mathcal{L}_k^j)^c\right)\middle| X_t, Z_t = j, \mathcal{A}\right]. \tag{42}$$

The event $\left\{\{X_t^\top\widehat{\beta}_{k,m-1}^j \geq X_t^\top\widehat{\beta}_{1,m-1}^j\} \cap \mathcal{L}_k^j\right\}$ regarding $r_{t,1}^j(X_t)$ implies

$$X_t^\top(\widehat{\beta}_{k,m-1}^j - \beta_k^j) - X_t^\top(\widehat{\beta}_{1,m-1}^j - \beta_1^j) \geq X_t^\top(\beta_1^j - \beta_k^j) \geq 2x_{\max}\delta.$$

Thus, at least one of $|X_t^\top(\widehat{\beta}^j_{\iota,\bar{m-1}} - \beta^j_\iota)|$ with $\iota \in \{1,k\}$ must be greater than $x_{\max}\delta$, which means

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(\{X_t^\top \widehat{\beta}^j_{k,\bar{m-1}} \geq X_t^\top \widehat{\beta}^j_{1,\bar{m-1}}\} \cap \mathcal{L}^j_k\right) \mid X_t, Z_t = j, \mathcal{A}\right] \\
&\leq \sum_{\iota \in \{1,k\}} \mathbb{P}\left[|X_t^\top(\widehat{\beta}^j_{\iota,\bar{m-1}} - \beta^j_\iota)| \geq x_{\max}\delta \,\middle|\, X_t, Z_t = j, \mathcal{A}\right] \\
&\leq \sum_{\iota \in \{1,k\}} \mathbb{P}\left[\|\widehat{\beta}^j_{\iota,\bar{m-1}} - \beta^j_\iota\|_1 \geq \delta \,\middle|\, \mathcal{A}\right].
\end{aligned}$$

Since Lemma 21 implies both arm 1 and k are not suboptimal, we further upper bound the above probability using our all-sample tail inequality in Proposition 4. Together with Lemma 20, we can write

$$\begin{aligned}
\mathbb{P}\left[\|\widehat{\beta}^j_{\iota,\bar{m-1}} - \beta^j_\iota\|_1 \geq \delta\right] &\leq \max_{i\in[N]} \frac{24}{p_i t} + \frac{16}{p_j t} + \frac{6}{TN} \\
&\quad + 2dN \exp\left(-\frac{p_*\psi(\min_{i\in[N]} p_i)t}{128 d x_{\max}^2}\right) + 8N \exp\left(-\frac{p_*(\min_{i\in[N]} p_i)t}{80}\right),
\end{aligned}$$

for $\iota \in \{1,k\}$, where

$$\delta = C_1\sqrt{\frac{2sd\log(dp_jt)}{p_jt}} + C_2\sqrt{\frac{4sd\log(\rho N)}{p_jt}} + C_3 d\sqrt{\frac{4\log(dp_jt)\log(\rho N)}{Np_jt}}.$$

Then, we can obtain from (41)

$$\begin{aligned}
&\mathbb{E}\left[r^j_{t,1}(X_t) \mid Z_t = j, \mathcal{A}\right] \leq 2bx_{\max}K \sum_{\iota\in\{1,k\}} \mathbb{P}\left[\|\widehat{\beta}^j_{\iota,\bar{m-1}} - \beta^j_\iota\|_1 \geq \delta\right] \\
&\leq 4bx_{\max}K\left(\max_{i\in[N]}\frac{24}{p_it} + \frac{16}{p_jt} + \frac{6}{TN} + 2dN\exp\left(-\frac{p_*\psi(\min_{i\in[N]}p_i)t}{128dx_{\max}^2}\right) + 8N\exp\left(-\frac{p_*(\min_{i\in[N]}p_i)t}{80}\right)\right).
\end{aligned} \tag{43}$$

On the other hand, by Assumption 3, we have for the term $r^j_{t,2}(X_t)$ that

$$\mathbb{E}\left[r^j_{t,2}(X_t) \mid Z_t = j, \mathcal{A}\right] \leq 2x_{\max}\delta K \mathbb{P}\left[(\mathcal{L}^j_k)^c\right] \leq 4x_{\max}^2 LK\delta^2. \tag{44}$$

Combining (43) and (44) with (40), we obtain

$$\begin{aligned}
r^j_t = \mathbb{E}\left[r^j_t(X_t) \,\middle|\, Z_t = j, \mathcal{A}\right] &\leq \sum_{r=1,2} \mathbb{E}\left[r^j_{t,r}(X_t) \mid Z_t = j, \mathcal{A}\right] \\
&\leq 24x_{\max}^2 LK\left(C_1^2\frac{sd\log(dp_jt)}{p_jt} + 2C_2^2\frac{sd\log(\rho N)}{p_jt} + 2C_3^2\frac{d^2\log(\rho N)\log(dp_jt)}{Np_jt}\right) \\
&\quad + 4bx_{\max}K\left(\max_{i\in[N]}\frac{24}{p_it} + \frac{16}{p_jt} + \frac{6}{TN} + 2dN\exp\left(-\frac{p_*\psi(\min_{i\in[N]}p_i)t}{128dx_{\max}^2}\right) + 8N\exp\left(-\frac{p_*(\min_{i\in[N]}p_i)t}{80}\right)\right),
\end{aligned}$$

where we use the inequality that $3(a^2+b^2+c^2) \geq (a+b+c)^2$. $\square$

Now, we provide a cumulative regret bound over time of all bandit instances for Case (III).

LEMMA 23. *When $\mathcal{A}$ holds, $T \geq KN$, $N = \Omega(\log(d)\log(T))$, the cumulative expected regret from all the batches $\{\mathcal{B}_m\}_{m>1}$ up to time T is upper bounded by*

$$\begin{aligned}
\sum_{j\in[N]}\Bigg[&24x_{\max}^2 LK\left(C_1^2 sd\log(dp_jT) + 2C_2^2 sd\log(\rho N) + 2C_3^2\frac{d^2\log(\rho N)\log(dp_jT)}{N}\right)\log(p_jT) \\
&+4bx_{\max}K\left((16 + \max_{i\in[N]}\frac{24p_j}{p_i})\log(p_jT) + \frac{6p_j}{N} + (\max_{i\in[N]}\frac{p_j}{p_i})(\frac{256x_{\max}^2 d^2}{p_*\psi N} + \frac{640}{p_*})\right)\Bigg].
\end{aligned}$$

Proof of Lemma 23 The cumulative expected regret from all instances over $\{\mathcal{B}_m\}_{m>1}$ is

$$\mathbb{E}\left[\sum_{t\in\bigcup_{m>1}\mathcal{B}_m} r_t^{Z_t}(X_t)\,\middle|\,\mathcal{A}\right] = \sum_{t\in\bigcup_{m>1}\mathcal{B}_m} \mathbb{E}\left[\mathbb{E}\left[r_t^{Z_t}(X_t)\,\middle|\,Z_t,\mathcal{A}\right]\middle|\,\mathcal{A}\right] = \sum_{t=2q\log(T)+1}^{T}\sum_{j\in[N]} p_j r_t^j. \tag{45}$$

Note that we have

$$\int_{2q\log(T)}^{T} \frac{1}{p_j t}dt \le \frac{\log(p_jT)}{p_j}.$$

Moreover, given q in Proposition 3, we have

$$\begin{aligned}\int_{2q\log(T)}^{T} dN\exp\left(-\frac{p_*\psi(\min_{i\in[N]}p_i)t}{128dx_{\max}^2}\right)dt \le& \frac{128d^2Nx_{\max}^2}{p_*\psi(\min_{i\in[N]}p_i)}\exp\left(-\frac{p_*\psi(\min_{i\in[N]}p_i)q\log(T)}{64dx_{\max}^2}\right)\\ \le& \frac{128d^2Nx_{\max}^2}{p_*\psi(\min_{i\in[N]}p_i)T^{3K\log(dN)/2}}\\ \le& \frac{128x_{\max}^2d^2}{p_*\psi(\min_{i\in[N]}p_i)N},\end{aligned}$$

where the last inequality holds since $T^{3K\log(dN)/2}\ge N^2$ when $T\ge N$ and $K,d,N>1$. Besides, we have

$$\begin{aligned}\int_{t=2q\log(T)}^{T} N\exp\left(-\frac{p_*(\min_{i\in[N]}p_i)t}{80}\right)dt \le& \frac{80N}{p_*(\min_{i\in[N]}p_i)}\exp\left(-\frac{p_*(\min_{i\in[N]}p_i)q\log(T)}{80}\right)\\ \le& \frac{80N}{p_*(\min_{i\in[N]}p_i)T^{3K\log(N)/4}}\\ \le& \frac{80}{p_*(\min_{i\in[N]}p_i)},\end{aligned}$$

where the last inequality holds since $T^{3K\log(N)/4}\ge N$ when $T\ge N$ and $K,N>1$. Combined all the above with Lemma 22, the cumulative expected regret conditioned on $\mathcal{A}$ in (45) is at most

$$\begin{aligned}\sum_{j\in[N]}\Bigg[&24x_{\max}^2LK\left(C_1^2sd\log(dp_jT)+2C_2^2sd\log(\rho N)+2C_3^2\frac{d^2\log(\rho N)\log(dp_jT)}{N}\right)\log(p_jT)\\ &+4bx_{\max}K\left((16+\max_{i\in[N]}\frac{24p_j}{p_i})\log(p_jT)+\frac{6p_j}{N}+(\max_{i\in[N]}\frac{p_j}{p_i})(\frac{256x_{\max}^2d^2}{p_*\psi N}+\frac{640}{p_*})\right)\Bigg]. \quad\square\end{aligned}$$

Proof of Theorem 2 Summing up the expected regrets of the three cases obtained in Lemma 18, 19 and 23, we can upper bound the total cumulative expected regret up to time T by

$$\begin{aligned}R_T \le\, & 2bx_{\max}(2q\log(T)+KN)+20bx_{\max}KN\\ &+\sum_{j\in[N]}\Bigg[24x_{\max}^2LK\left(C_1^2sd\log(dp_jT)+2C_2^2sd\log(\rho N)+2C_3^2\frac{d^2\log(\rho N)\log(dp_jT)}{N}\right)\log(p_jT)\\ &+4bx_{\max}K\left((16+\max_{i\in[N]}\frac{24p_j}{p_i})\log(p_jT)+\frac{6p_j}{N}+(\max_{i\in[N]}\frac{p_j}{p_i})(\frac{256x_{\max}^2d^2}{p_*\psi N}+\frac{640}{p_*})\right)\Bigg].\end{aligned}$$

In our standard case, $p_i=\Theta(1/N)$ for any $i\in[N]$ and thereby $q=\mathcal{O}(Kd(sN+d)\log(d)\log(N))$. Thus, we have

$$R_T=\mathcal{O}\left(Kd(sN+d)\log(N)\log^2\left(\frac{dT}{N}\right)\right). \quad\square$$

C.5. Single Bandit Instance

Here we prove Corollary 3, which provides the regret for a single bandit instance $j \in [N]$ (while running RMBandit across all bandit instances).

Proof of Corollary 3 The cumulative expected regret of any target instance j in Case III is

$$\mathbb{E}\left[\sum_{t\in\bigcup_{m>1}\mathcal{B}_m} r_t^j \mathbb{1}\left(Z_t=j\right)\middle|\mathcal{A}\right] = p_j\mathbb{E}\left[\sum_{t\in\bigcup_{m>1}\mathcal{B}_m} r_t^j\middle|\mathcal{A}\right].$$

Corresponding to the proof of Lemma 23, we have

$$\mathbb{E}\left[\sum_{t\in\bigcup_{m>1}\mathcal{B}_m} r_t^j\middle|\mathcal{A}\right] \le 24x_{\max}^2 LK\left(C_1^2 sd\log(dT)+2C_2^2 sd\log(\rho N)+2C_3^2\frac{d^2\log(\rho N)\log(dT)}{N}\right)\log(T)$$
$$+4bx_{\max}K\left((16+\max_{i\in[N]}\frac{24p_j}{p_i})\log(T)+\frac{6p_j}{N}+(\max_{i\in[N]}\frac{p_j}{p_i})(\frac{256x_{\max}^2d^2}{p_*\psi N}+\frac{640}{p_*})\right).$$

Besides, the cumulative expected regret of instance j from Case I and II is simply

$$p_j\left(2bx_{\max}(2q\log(T)+KN)+20bx_{\max}KN\right).$$

Combining all the above with $p_j=\Theta(1/N)$, we have

$$R_T^j=\mathcal{O}\left(Kd(s+d/N)\log(N)\log^2(dT)\right). \quad \square$$

C.6. Network Structure

Besides, we prove Corollary 4, which provides a regret bound for the case where the bandit instances have network structure.

Proof of Corollary 4 Our network structure is an exogenous assumption upon the sparsity s. Therefore, all the previous analyses for Theorem 2 and Corollary 3 still go through for an arbitrary number of selected instances, i.e., $\tilde{N}$. Plugging $s=\tilde{N}^\alpha$ in the regret bound derived in Corollary 3 and optimizing in terms of $\tilde{N}$, we get the optimal value of $\tilde{N}$ to be $\Theta\left(d^{\frac{1}{\alpha+1}}\right)$. Note that the constraint on the time horizon T becomes $T=\Omega\left(d\right)=\mathcal{O}\left(e^{\frac{d^{\frac{1}{\alpha+1}}}{\log(d)}}\right)$ given the above s and $\tilde{N}$. The result then follows. $\square$

Appendix D: RMBandit Algorithm in Data-Poor Regime

In this section, we give hyperparameter choices and a proof for Theorem 3, which bounds the regret across all bandit instances in the data-poor regime. The proof closely follows that of Theorem 2. First, we give the hyperparameter choices for Theorem 3 to hold. We take $\zeta_0=\zeta_{1,0}=1$, $\eta_0=\eta_{1,0}=0$,

$$\lambda_{0,j}=\frac{p_*\psi h}{256x_{\max}s},\quad \lambda_{1,j,0}=\sqrt{\frac{64\sigma_j^2x_{\max}^2}{p_*}},$$

and

$$q=\max\left\{\frac{(128\sqrt{3})^2\sigma_\ell^2x_{\max}^2Kd^2\log d}{h^2p_*\psi p_\ell},\frac{(2048\sqrt{3})^2x_{\max}^4\sigma_j^2Ks^2\log d}{h^2p_jp_*^2\psi^2},\right.$$
$$\left.\frac{96x_{\max}^2Kd\log d}{p_*\psi p_\ell},\frac{4K}{C^2p_*p_j},\frac{20K}{p_*p_j},\frac{12K\log d}{C^2p_*p_j}\right\},$$

where

$$C=\max\left\{\frac{1}{2},\frac{\psi'^2}{512sx_{\max}^2}\right\}.$$

Note that since we are only using a single neighbor of j, we trivially set $\zeta=N=1$ and $\eta=0$, amounting to the transfer learning method.

D.1. Robust Multitask Estimator with Random Design

We begin by proving a variant of the robust multitask estimator for the random design setting, specialized to our setting where $\tilde{N}=2$. Before we do so, we first prove that the compatibility condition holds with high probability for our design.

Definition 3 (Compatibility Condition). For a constant $\phi'>0$, define the set of matrices

$$\mathcal{C}(\mathcal{S},\phi')=\{M\in\mathbb{R}^{d\times d}\mid \forall\|v_{\mathcal{S}^c}\|_1\le 7\|v_{\mathcal{S}}\|_1, \phi'\|v_{\mathcal{S}}\|_1^2\le|\mathcal{S}|v^\top Mv\}.$$

Lemma 24. *The true covariance matrix $\Sigma_k^j=\mathbb{E}_{\mathcal{P}_X^j}[XX^\top|X\in U_k^j], k\in[K]$ of the target instance j satisfies the compatibility condition—i.e., there exists a positive constant ψ' such that $\Sigma_k^j\in\mathcal{C}(\bar{\mathcal{S}}_j,\psi')$ for any $k\in[K]$.*

Proof of Lemma 24 Given Assumption 5, $\lambda_{\min}(\Sigma_k^j)\ge\psi$. Then, for any $v\in\mathbb{R}^d$, we have $\|v_{\mathcal{S}}\|_1\le\sqrt{s}\|v_{\mathcal{S}}\|_2$. Therefore,

$$sv^TMv\ge s\psi\|v\|_2^2\ge\psi s\|v_{\mathcal{S}}\|_2^2\ge\psi\|v_{\mathcal{S}}\|_1^2.\quad\square$$

Now, we state and prove our main proposition for this section.

Proposition 9. *The robust multitask estimator of instance j from Algorithm 1 satisfies the following concentration inequality*

$$\mathbb{P}\left[\|\widehat{\beta}^j-\beta^j\|_1\ge\frac{8\lambda_j s}{\phi'}+4d\sqrt{\frac{\sigma_\ell^2}{n_\ell\phi}}\chi\right]\le 2d\exp\left(-\frac{\chi^2}{2}\right)$$
$$+2d\exp\left(-\frac{\lambda_j^2n_j}{32\sigma_j^2x_{\max}^2}\right)+\mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}^\ell)\le\phi\right]+\mathbb{P}\left[\widehat{\Sigma}^j\notin\mathcal{C}(\bar{\mathcal{S}}_j,\phi')\right],$$

for any $\lambda_j>0$ and $0<\chi$.

Proof of Proposition 9 The proof mainly follows that of Theorem 1 and Proposition 2.

Differently, now we use $\widetilde{\beta}^\ell$ as an estimate of the shared parameter $\beta^\dagger$. On the event $\{\lambda_{\min}(\widehat{\Sigma}^\ell)\ge\phi\}$, note that $\widetilde{\beta}^\ell=(\mathbf{X}^{\ell\top}\mathbf{X}^\ell)^{-1}\mathbf{X}^{\ell\top}Y^\ell$ is a subgaussian random vector with mean β^ℓ. In particular, the i^{th} component of $\widetilde{\beta}$, i.e., $\widetilde{\beta}^\ell_{(i)}$, is $\sqrt{\frac{\sigma_\ell^2}{n_\ell\phi}}$-subgaussian. Therefore,

$$\mathbb{P}\left[|\widetilde{\beta}^\ell_{(i)}-\beta^\ell_{(i)}|\ge\sqrt{\frac{\sigma_\ell^2}{n_\ell\phi}}\chi\right]\le 2\exp\left(-\frac{\chi^2}{2}\right),$$

for any $0<\chi$. Using a union bound on all $i\in\mathcal{S}^c$, we have

$$\mathbb{P}\left[\|(\widehat{\beta}^*-\beta^\dagger)_{(\mathcal{S}^c)}\|_1\ge d\sqrt{\frac{\sigma_\ell^2}{n_\ell\phi}}\chi\right]\le 2d\exp\left(-\frac{\chi^2}{2}\right),$$

since $\widehat{\beta}^*=\widetilde{\beta}^\ell$. The following still holds

$$Y^j=\mathbf{X}^j(\beta^\dagger+\delta^j)+\epsilon^j=\mathbf{X}^j\left((\beta^\dagger_{(\mathcal{S}^c)}+\widehat{\beta}^*_{(\mathcal{S})})+(\beta^\dagger_{(\mathcal{S})}-\widehat{\beta}^*_{(\mathcal{S})}+\delta^j)\right)+\epsilon^j,$$

where $\beta^\dagger_{(\mathcal{S})}-\widehat{\beta}^*_{(\mathcal{S})}+\delta^j$ is now at most $2s$-sparse. Here we follow a proof strategy of transfer learning using LASSO as in Bastani (2021), Xu et al. (2021). Similar to the proof of Theorem 1, we can derive from the basic inequality of LASSO that

$$\frac{1}{n_j}\|\mathbf{X}^j(\widehat{\beta}^j-\beta^j)\|_2^2+\frac{\lambda_j}{2}\|(\widehat{\beta}^j-\beta^j)_{(\bar{\mathcal{S}}_j^c)}\|_1\le\frac{3\lambda_j}{2}\|(\widehat{\beta}^j-\beta^j)_{(\bar{\mathcal{S}}_j)}\|_1+2\lambda_j\|(\widehat{\beta}^*-\beta^\dagger)_{(\mathcal{S}^c)}\|_1.\qquad(46)$$

Now we consider two cases respectively:

(i). $\|(\widehat{\beta}^j - \beta^j)_{(\bar{S}_j)}\|_1 \leq \|(\widehat{\beta}^* - \beta^\dagger)_{(\mathcal{S}^c)}\|_1$

(ii). $\|(\widehat{\beta}^j - \beta^j)_{(\bar{S}_j)}\|_1 > \|(\widehat{\beta}^* - \beta^\dagger)_{(\mathcal{S}^c)}\|_1$.

In the first case, we can obtain from inequality (46) that

$$\|\widehat{\beta}^j - \beta^j\|_1 \leq 8\|(\widehat{\beta}^* - \beta^\dagger)_{(\mathcal{S}^c)}\|_1.$$

In the second case, it holds that $\|(\widehat{\beta}^j - \beta^j)_{(\bar{\mathcal{S}}_j^c)}\|_1 \leq 7\|(\widehat{\beta}^j - \beta^j)_{(\bar{\mathcal{S}}_j)}\|_1$. Therefore, on the event $\{\widehat{\Sigma}^j \in \mathcal{C}(\bar{\mathcal{S}}_j, \phi')\}$, we have

$$\|\widehat{\beta}^j - \beta^j\|_1 \leq \frac{32\lambda_j s}{\phi'}.$$

Combining all the above, we get

$$\|\widehat{\beta}^j - \beta^j\|_1 \leq \frac{32\lambda_j s}{\phi'} + 8\|(\widehat{\beta}^* - \beta^\dagger)_{(\mathcal{S}^c)}\|_1$$

with a high probability. With a union bound, we obtain the following concentration inequality:

$$\mathbb{P}\left[\|\widehat{\beta}^j - \beta^j\|_1 \geq \frac{32\lambda_j s}{\phi'} + 8d\sqrt{\frac{\sigma_\ell^2}{n_\ell \phi}}\chi\right] \leq 2d\exp\left(-\frac{\chi^2}{2}\right) + 2d\exp\left(-\frac{\lambda_j^2 n_j}{32\sigma_j^2 x_{\max}^2}\right).$$

Following a similar argument in the proof of Proposition 2, we get

$$\mathbb{P}\left[\|\widehat{\beta}^j - \beta^j\|_1 \geq \frac{8\lambda_j s}{\phi'} + 4d\sqrt{\frac{\sigma_\ell^2}{n_\ell \phi}}\chi\right] \leq 2d\exp\left(-\frac{\chi^2}{2}\right)$$
$$+ 2d\exp\left(-\frac{\lambda_j^2 n_j}{32\sigma_j^2 x_{\max}^2}\right) + \mathbb{P}\left[\lambda_{\min}(\widehat{\Sigma}^\ell) \leq \phi\right] + \mathbb{P}\left[\widehat{\Sigma}^j \in \mathcal{C}(\bar{\mathcal{S}}_j, \phi')\right]. \quad \square$$

D.2. Forced-Sample Estimator

Lemma 25. *When $|\bar{\mathcal{B}}_{0,k}^j| \geq \frac{3\log(d)}{C^2}$, we have given $|\bar{\mathcal{B}}_{0,k}^j|$*

$$\mathbb{P}\left[\widehat{\Sigma}_k^j(\bar{\mathcal{B}}_{0,k}^j) \in \mathcal{C}(\bar{\mathcal{S}}_j, \frac{\psi'}{\sqrt{2}})\right] \geq 1 - \exp\left(-C^2|\bar{\mathcal{B}}_{0,k}^j|\right),$$

where $C = \max\left\{\frac{1}{2}, \frac{\psi'^2}{512 s x_{\max}^2}\right\}$.

Proof of Lemma 25 See Lemma EC.6 in Bastani and Bayati (2020). $\square$

Proposition 10. *When $\zeta_0, \lambda_{0,j}, \eta_0, q$ take the values in Theorem 3, the forced-sample estimator of instance j and arm k satisfies*

$$\mathbb{P}\left[\|\widehat{\beta}_{k,0}^j - \beta_k^j\|_1 \geq \frac{h}{4x_{\max}}\right] \leq \frac{10}{T}.$$

Proof of Proposition 10 The proof mainly follows that of Proposition 3. Similarly, applying Proposition 9, we have

$$\mathbb{P}\left[\|\widehat{\beta}_{k,0}^j - \beta_k^j\|_1 \geq \frac{32\lambda_j s}{p_*\psi} + 8d\sqrt{\frac{2K\sigma_\ell^2}{p_*\psi p_\ell|\mathcal{B}_0|}}\chi\right]$$
$$\leq 2d\exp\left(-\frac{\chi^2}{2}\right) + 2d\exp\left(-\frac{\lambda_j^2 p_j|\mathcal{B}_0|}{64K\sigma_j^2 x_{\max}^2}\right) + d\exp\left(-\frac{p_*\psi p_\ell|\mathcal{B}_0|}{32Kdx_{\max}^2}\right)$$
$$+ \exp(-\frac{C^2 p_* p_j|\mathcal{B}_0|}{4K}) + \sum_{i=j,l} 2\exp\left(-\frac{p_* p_i|\mathcal{B}_0|}{20K}\right) + \sum_{i=j,l} 2\exp\left(-\frac{p_i|\mathcal{B}_0|}{10}\right). \quad (47)$$

Now we configure the parameters λ_j and χ such that our forced-sample estimator of arm k and instance j has estimation error smaller than $\frac{h}{4x_{\max}}$. Take

$$\lambda_j = \frac{p_*\psi h}{256 x_{\max} s}, \quad \chi = \frac{h}{64 x_{\max} d}\sqrt{\frac{p_*\psi p_\ell |\mathcal{B}_0|}{2K\sigma_\ell^2}}.$$

For the first term on the right hand side of (47) to be less than $2/T$, it suffices to have

$$\chi^2 \geq 6\log(d)\log(T),$$

that is,

$$q \geq \frac{(128\sqrt{3})^2 \sigma_\ell^2 x_{\max}^2 K d^2 \log(d)}{h^2 p_* \psi p_\ell}.$$

Finally, we require

$$q \geq \max\left\{\frac{(2048\sqrt{3})^2 x_{\max}^4 \sigma_j^2 K s^2 \log(d)}{h^2 p_j p_*^2 \psi^2}, \frac{96 x_{\max}^2 K d \log(d)}{p_* \psi p_\ell}, \frac{4K}{C^2 p_* p_j}, \frac{20K}{p_* p_j}, \frac{10}{p_j}\right\}$$

so that the sum of the last four probability terms in inequality (32) is no greater than $8/T$. Moreover, to satisfy $|\bar{\mathcal{B}}_{0,k}^j| \geq \frac{3\log(d)}{C^2}$ in Lemma 25, we also require $q \geq \frac{12K\log(d)}{C^2 p_* p_j}$.

As a result, letting

$$q = \max\left\{\frac{(128\sqrt{3})^2 \sigma_\ell^2 x_{\max}^2 K d^2 \log(d)}{h^2 p_* \psi p_\ell}, \frac{(2048\sqrt{3})^2 x_{\max}^4 \sigma_j^2 K s^2 \log(d)}{h^2 p_j p_*^2 \psi^2}, \right.$$
$$\left. \frac{96 x_{\max}^2 K d \log(d)}{p_* \psi p_\ell}, \frac{4K}{C^2 p_* p_j}, \frac{20K}{p_* p_j}, \frac{12K\log(d)}{C^2 p_* p_j}\right\},$$

we have

$$\mathbb{P}\left[\|\widehat{\beta}_{k,0}^j - \beta_k^j\|_1 \geq \frac{h}{4x_{\max}}\right] \leq \frac{10}{T},$$

for any $k \in [K]$. □

Proposition 11. *When q take the values in Theorem 3, the forced-sample estimator of instance j and arm k satisfies*

$$\mathbb{P}\left[\|\widehat{\beta}_k^\ell(\mathcal{B}_0) - \beta_k^\ell\|_1 \geq \frac{h}{4x_{\max}}\right] \leq \frac{10}{T}.$$

Proof of Proposition 11 For the data-rich instance ℓ, we have

$$\mathbb{P}\left[\|\widetilde{\beta}^\ell - \beta^\ell\|_1 \geq 2d\sqrt{\frac{2K\sigma_\ell^2}{p_*\psi p_\ell|\mathcal{B}_0|}}\chi\right] \leq 2d\exp\left(-\frac{\chi^2}{2}\right)$$
$$+ d\exp\left(-\frac{p_*\psi p_\ell|\mathcal{B}_0|}{32Kdx_{\max}^2}\right) + 2\exp\left(-\frac{p_* p_\ell|\mathcal{B}_0|}{20K}\right) + 2\exp\left(-\frac{p_\ell|\mathcal{B}_0|}{10}\right).$$

Setting the value χ and q as in Propositon 10, we have

$$\mathbb{P}\left[\|\widehat{\beta}_{k,0}^j - \beta_k^j\|_1 \geq \frac{h}{32x_{\max}}\right] \leq \frac{7}{T},$$

for any $k \in [K]$. The result then follows. □

D.3. All-Sample Estimator

Lemma 26. *The event $\mathcal{A}$ holds with at least a probability of $1-\frac{20K}{T}$.*

Proof of Lemma 26 The result follows by applying a union bound over all arms and bandit instances using Proposition 10. $\square$

Assume that the optimal arm of instance ℓ is the same as that of instance j so that $\rho = 1$.

Proposition 12. *When $\mathcal{A}$ holds, and $\zeta_{1,0}, \lambda_{1,j,0}, \eta_{1,0}$ take the values in Theorem 3, the all-sample estimator of instance j and optimal arm $k \in \mathcal{K}^j_{opt}$ using data from the batch $\mathcal{B}_m$ with $m \geq 1$ satisfies*

$$\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}-1} - \beta^j_k\|_1 \geq C_1\sqrt{\frac{s^2\log(dp_j|\mathcal{B}_m|)}{p_j|\mathcal{B}_m|}} + C_2\sqrt{\frac{d^2\log(dp_j|\mathcal{B}_m|)}{p_\ell|\mathcal{B}_m|}}\,\middle|\,\mathcal{A}\right]$$
$$\leq \frac{8}{p_j|\mathcal{B}_m|} + d\exp\left(-\frac{p_*\psi p_\ell|\mathcal{B}_m|}{32dx^2_{\max}}\right) + \exp\left(-\frac{C^2 p_* p_j|\mathcal{B}_m|}{4}\right) + \sum_{i=j,l} 6\exp\left(-\frac{p_* p_i|\mathcal{B}_m|}{20}\right),$$

where

$$C_1 = \frac{256\sqrt{2}\sigma_j x_{\max}}{p_*^{\frac{3}{2}}\psi}, \quad C_2 = \frac{16\sqrt{2}\sigma_\ell}{p_*\psi^{\frac{1}{2}}},$$

and

$$\lambda_{1,j,m} = \lambda_{1,j,0}\sqrt{\frac{\log(|\mathcal{B}^j_m|)}{|\mathcal{B}^j_m|}}, \quad \eta_{1,m} = \eta_{1,0}\sqrt{\log(\min_{i\in[N],|\mathcal{B}^i_m|>0}|\mathcal{B}^i_m|)},$$

as in Algorithm 2.

Proof of Proposition 12 The proof follows that of Proposition 4. Applying Proposition 9, we get

$$\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}} - \beta^j_k\|_1 \geq \frac{32\lambda_j s}{p_*\psi} + 8d\sqrt{\frac{2\sigma^2_\ell}{p_*^2\psi|\mathcal{B}^\ell_m|}}\chi\,\middle|\,\{|\mathcal{B}^j_m|, |\mathcal{B}^\ell_m|\}, \mathcal{A}\right]$$
$$\leq 2d\exp\left(-\frac{\chi^2}{2}\right) + 2d\exp\left(-\frac{\lambda_j^2 p_*|\mathcal{B}^j_m|}{64\sigma_j^2 x^2_{\max}}\right) + d\exp\left(-\frac{p_*\psi|\mathcal{B}^\ell_m|}{16dx^2_{\max}}\right)$$
$$+\exp\left(-\frac{C^2 p_*|\mathcal{B}^j_m|}{2}\right) + \sum_{i=j,l} 4\exp\left(-\frac{p_*|\mathcal{B}^i_m|}{10}\right). \tag{48}$$

Similarly, take

$$\lambda_j = \sqrt{\frac{64\sigma_j^2 x^2_{\max}\log(d|\mathcal{B}^j_m|)}{p_*|\mathcal{B}^j_m|}}, \quad \chi = \sqrt{2\log(d|\mathcal{B}^j_m|)}.$$

Then, define the events

$$\mathcal{M}^i_m = \left\{|\mathcal{B}^i_m| \geq \frac{p_i}{2}|\mathcal{B}_m|\right\}$$

for $i = j, l$. With a union bound, we obtain

$$\mathbb{P}\left[\|\widehat{\beta}^j_{k,\bar{m}} - \beta^j_k\|_1 \geq C_1\sqrt{\frac{s^2\log(dp_j|\mathcal{B}_m|)}{p_j|\mathcal{B}_m|}} + C_2\sqrt{\frac{d^2\log(dp_j|\mathcal{B}_m|)}{p_\ell|\mathcal{B}_m|}}\,\middle|\,\mathcal{A}\right]$$
$$\leq \frac{8}{p_j|\mathcal{B}_m|} + d\exp\left(-\frac{p_*\psi p_\ell|\mathcal{B}_m|}{32dx^2_{\max}}\right) + \exp\left(-\frac{C^2 p_* p_j|\mathcal{B}_m|}{4}\right) + \sum_{i=j,l} 6\exp\left(-\frac{p_* p_i|\mathcal{B}_m|}{20}\right),$$

where

$$C_1 = \frac{256\sqrt{2}\sigma_j x_{\max}}{p_*^{\frac{3}{2}}\psi}, \quad C_2 = \frac{16\sqrt{2}\sigma_\ell}{p_*\psi^{\frac{1}{2}}}. \quad \square$$

D.4. Single Bandit Instance

Proof of Theorem 3 The cumulative expected regret of any target instance j is

$$R_T^j = \mathbb{E}\left[\sum_{t=1}^{T/p_j} r_t^j \mathbb{1}\left(Z_t = j\right)\right].$$

Using a similar argument in the proof of Corollary 3, we have

$$\begin{aligned}R_T^j = p_j \mathbb{E}\left[\sum_{t=1}^{T/p_j} r_t^j\right] &\leq 4bx_{\max} p_j q \log(\frac{T}{p_j}) + 20bx_{\max} p_j K \\ &\quad + 32x_{\max}^2 LK\left((C_1^2 s^2 + C_2^2 \frac{d^2 p_j}{p_\ell}) \log(dT)\log(T)\right) \\ &\quad + 4bx_{\max} K\left(32\log(T) + \frac{128x_{\max}^2 d^2 p_j}{p_* \psi p_\ell} + \frac{16}{C^2 p_*} + \frac{960}{p_*}\right).\end{aligned}$$

Since $\frac{d^2 p_j}{p_\ell} = \mathcal{O}(1)$ and $p_j q = \Theta(Ks^2 \log(d))$, it implies

$$R_T^j = \mathcal{O}\left(Ks^2 \log^2(dT)\right). \quad \square$$

Appendix E: Auxiliary Results

This appendix collects useful results from the literature.

Lemma 27. *Let $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}$ be a vector of n independent σ-subgaussian random variables with mean μ. Then, for any $a \in \mathbb{R}^n$ and $t \geq 0$, it holds that*

$$\mathbb{P}\left[|a^\top (X - \mu)| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sigma^2 \|a\|_2^2}\right).$$

Proof of Lemma 27 See Corollary 1.7 of Rigollet and Hütter (2015).

Lemma 28. *Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty$ be a martingale difference sequence and suppose that $\mathbb{E}[\exp(\lambda D_k) \mid \mathcal{F}_{k-1}] \leq \exp(\sigma^2 \lambda^2 / 2)$ for any $t \in \mathbb{R}$. Then, it holds for any $t > 0$ that*

$$\mathbb{P}\left[|\sum_{k \in [n]} D_k| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2n\sigma^2}\right).$$

Proof of Lemma 28 The result is a special case of Theorem 2.19 of Wainwright (2019).

Lemma 29. *Consider a sequence of independent random symmetric matrices $X_k \in \mathbb{R}^{d \times d}$, $k \in [n]$ with $\lambda_{\min}(X_k) \geq 0$ and $\lambda_{\max}(X_k) \leq L$ for any k. Let $\mu = \lambda_{\min}(\mathbb{E}[\sum_{k \in [n]} X_k])$. We have for $0 < t < 1$ that*

$$\mathbb{P}\left[\lambda_{\min}(\sum_{k \in [n]} X_k) \geq t\mu\right] \geq 1 - d\exp\left(-\frac{(1-t)^2 \mu}{2L}\right).$$

Proof of Lemma 29 See page 61 in Tropp (2015).

Lemma 30. *Suppose $X_1, \cdots, X_n$ are n independent Bernoulli random variables with mean $p_1, \cdots, p_n$ respectively. Let $\mu = \sum_{i \in [n]} p_i$. Then, we have*

$$\mathbb{P}\left[|\sum_{i \in [n]} X_i - \mu| \geq \frac{\mu}{2}\right] \leq 2\exp\left(-\frac{\mu}{10}\right).$$

Proof of Lemma 30 The result follows by taking $\epsilon = 1/2$ in Corollary A.1.14 of Alon and Spencer (2004).

Appendix F: Experimental Details & Results

F.1. Synthetic Experiment Details

In the standard setting, we take the number of instances $N = 10$, the number of arms $K = 10$, the context dimension $d = 20$, and the sparsity $s = 2$. Our total time horizon across instances $T = 40,000$ and the arrival probability $p_j = 1/N$ for all $j \in [N]$; thus, each instance will receive an expected $4,000$ observations. We generate the shared parameters $\{\beta_k^\dagger\}_{k\in[K]}$ by drawing independently from a gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and normalizing them $\|\beta_k^\dagger\|_1 = 1$. We set the first instance's parameters to be equal to the shared parameters; for the remaining instances, we draw the nonzero entries of the bias terms δ_k^j independently from a uniform distribution on $[-0.5, 0.5]$. Note that $\|\beta_k^j\|_1 \leq 2$ for all $j \in [N]$. Next, we draw the context vectors X_t independently from a gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and truncate them so that $\|X_t\|_\infty = 1$. We draw the noise ϵ_t given $Z_t = j$ independently from a gaussian distribution $\mathcal{N}(0, \sigma_j^2)$ with $\sigma_j = 0.05$.

We use the same setup in the data-poor setting, but modify $N = 2$ and take the arrival probability $p_1 = p_2/100$; accordingly, in order to simulate a similar time horizon for the data-poor instance, we increase the time horizon across instances to $T = 400,000$.

To ensure fair comparison, we tune the hyperparameters of all algorithms on a pre-specified grid. Matching the suggestion by Bastani and Bayati (2020), we take $h = 15$, $q = 1$ and $\lambda_{0,j} = \lambda_{1,j,0} = 0.02$ for Lasso and OLS Bandit. We take $\alpha = 1$ for GOBLin. We take $\eta_0 = \eta_{1,0} = 0.2$, $h = 15$ and $q = 50$ for RMBandit in the standard setting; in the data-poor setting, we increase $q = 300$ as suggested by the theory.

F.2. Diabetes Experiment Details

Our original dataset consists of 9948 patients observed from 379 healthcare providers. However, many of these providers observe very few patients, so we restrict our experiment to the $N = 13$ largest hospitals, of which each has at least 150 unique patients (mean 317; median 301) observed during the sample period. We take $K = 2$ since our arms are either to intervene or not intervene on the patient. We consider a simple binary reward that directly evaluates the accuracy of our classification of patients, i.e., the reward is 1 if the prediction is correct, and 0 otherwise.

We perform standard variable selection as a pre-processing step in order to avoid overfitting when computing our linear oracles. In particular, we run a LASSO variable selection procedure by regressing diabetes outcomes against the 184 total features (note that we exclude the healthcare providers that we use in our experiment in this step to avoid overfitting), and we tune the hyperparameters using 10-fold cross-validation. This leaves us with roughly 80 commonly predictive features (depending on the randomness in the cross-validation procedure). Note that this is still a relatively large number of features compared to the number of observations, supporting our argument that arm parameters are likely dense. We fit a linear oracle to data from the target provider in hindsight; to avoid overfitting, we use a leave-one-out approach, i.e., for each patient, we train the best linear model on all data from the target provider excluding the current patient. Our oracle is constructed to provide the best achievable mean squared error within a linear model family.

We run the bandit algorithms on this data in the same manner as in the synthetic setup in the previous subsection. Once again, to ensure fair comparison, we tune the hyperparameters of all algorithms, and we report the optimized results in Figure 3.

F.3. Pricing Experiment Details

Data: The original dataset covers 145 weeks of orders from 77 fulfillment centers across 51 cities. There are 14 different categories (e.g., beverages, snacks) and 4 different cuisines (e.g., Indian, Italian) for meals delivered by the company. To ensure similarity, we restrict our experiment to fulfillment centers in the largest city, and further exclude two centers that do not supply most food varieties. Thus, we have $N = 7$ centers, each processing an average (median) of 6,747 (6,894) orders during the sample period. One order arrives at each time step, and the chosen price is the checkout price, which includes discounts, taxes and delivery charges. The order price in our data ranges from \$55 to \$729; thus, we set $p_{\min} = 0$ and $p_{\max} = 1,000$. Following standard practice, we also normalize the price so it has a similar scale as the other features. Our outcome (demand) is given by the quantity in each order. The contexts are order-specific features including dummy variables capturing the category and cuisine, indicators of email or homepage promotions, and an intercept. Overall, our X_t has dimension 19, and therefore the dimensionality of the unknown parameters of the pricing model $d = 38$.

Algorithm: We now embed our robust multitask estimator within the ILSX/ILQX algorithmic approach proposed in Ban and Keskin (2021) to design the RMX algorithm; similarly, we embed the Laplacian estimator used by GOBLin (Cesa-Bianchi et al. 2013) to design the GOBX algorithm.

Let $\beta^j = \begin{bmatrix} \beta_0^j \\ \beta_1^j \end{bmatrix}$ denote the unknown parameters for instance j. For our forced samples, we fix two experimental prices $p_1 = 200$ and $p_2 = 600$, which we charge in two sets of periods

$$M_i^j = \left\{ t \,\middle|\, \sum_{r \in [t]} \mathbb{1}(Z_r = j) = E^2 + i - 1,\, E = 1, 2, \cdots \right\}$$

for each experimental price $i \in [2]$ and each instance $j \in [N]$. Note that M_i^j is a random set in the multitask setting, since it depends on the realization of arrivals Z_t. Let $M^j = M_1^j \cup M_2^j$ represent the forced price experimentation period, and let $M_t^j = \{r | r \in M^j,\, r < t\}$ be the set of time periods when prices are forced at instance j before time t. We update our estimators at time periods

$$M = \left\{ t \,\middle|\, t = N(E^2 + 1),\, E = 1, 2, \cdots \right\},$$

so that each instance obtains the same number of training observations in expectation as in the single-instance setting. Then, the samples used for estimating the optimal price at time t are $\bigcup_{j \in [N]} M_{\gamma_t}^j$, where $\gamma_t = \max\{r \mid r \in M, r < t\}$.

Note that we now only maintain a single set of estimated parameters for each algorithm. We denote our robust multitask estimator (Algorithm 1) at instance j at time t as

$$\widehat{\beta}^j(\cup_{j \in [N]} M_{\gamma_t}^j, \lambda_{j,t}, \omega_t).$$

The first argument indicates the training data, i.e., all observations with price experimentation before time γ_t (recall that the robust multitask estimators are only updated at $t \in M$); the remaining arguments are hyperparameters. We denote the estimated optimal price of user X_t at instance j at time t as

$$\widehat{p}^j(X_t, \widehat{\beta}^j) = \frac{X_t^\top \widehat{\beta}_0^j}{-2X_t^\top \widehat{\beta}_1^j},$$

Algorithm 3 Robust Multitask Regression with Price Experimentation (RMX)

Inputs: Initial hyperparameters $\zeta_0, \eta_0, \{\lambda_{j,0}\}_{j\in[N]}$
Define $\{M_i^j\}_{i\in[2]}$, M, and $\omega_0 = \zeta_0 + \eta_0$
for $t \in [T]$ **do**
  Observe an arrival at instance $j = Z_t$ and corresponding context vector X_t for this arrival
  if $t \in M_i^j$ **then**
    Charge price $p_t = p_i$
  else
    Charge price $p_t = \widehat{p}^j(X_t, \widehat{\beta}^j)$
  end if
  if $t \in M$ **then**
    Update $\zeta_t = \zeta_0$, $\eta_t = \eta_0 \sqrt{\log(d \min_{j\in[N],|M_{\gamma_t}^j|>0} |M_{\gamma_t}^j|)}$, and $\omega_t = \zeta_t + \eta_t$
    Update $\lambda_{j,t} = \lambda_{j,0} |M_{\gamma_t}^j|^{\frac{1}{4}} \sqrt{\log(d|M_{\gamma_t}^j|)}$
    Estimate parameter $\widehat{\beta}^j(\cup_{j\in[N]} M_{\gamma_t}^j, \lambda_{j,t}, \omega_t)$
  end if
  Observe demand $Y_t = X_t^\top \beta_0^j + p_t \cdot (X_t^\top \beta_1^j) + \epsilon_t$
end for

which is truncated at $p_{\min}$ and $p_{\max}$. We formalize our algorithm in Algorithm 3.

The GOBX algorithm follows exactly as in Algorithm 3, but uses the Laplacian-regularized estimator from (Cesa-Bianchi et al. 2013) instead of our robust multitask estimator. Once again, to ensure fair comparison, we tune the hyperparameters of all algorithms, and we report the optimized results in Figure 4.