Learning the Pareto Front Using Bootstrapped Observation Samples

Wonyoung Kim
Chung-Ang University

Garud Iyengar
Columbia University

Assaf Zeevi
Columbia University

Abstract

We consider Pareto front identification (PFI) for linear bandits (PFILin), i.e., the goal is to identify a set of arms with undominated mean reward vectors when the mean reward vector is a linear function of the context. PFILin includes the best arm identification problem and multi-objective active learning as special cases. The sample complexity of our proposed algorithm is optimal up to a logarithmic factor. In addition, the regret incurred by our algorithm during the estimation is within a logarithmic factor of the optimal regret among all algorithms that identify the Pareto front. Our key contribution is a new estimator that in every round updates the estimate for the unknown parameter along *multiple* context directions – in contrast to the conventional estimator that only updates the parameter estimate along the chosen context. This allows us to use low-regret arms to collect information about Pareto optimal arms. Our key innovation is to reuse the exploration samples multiple times; in contrast to conventional estimators that use each sample only once. Numerical experiments demonstrate that the proposed algorithm successfully identifies the Pareto front while controlling the regret.

1 Introduction

Consider a setting where one has to select among a finite set of actions that have multiple different characteristics, see, e.g., Lizotte et al. (2010); Van Moffaert and Nowé (2014); Lin et al. (2019). A classical example is prescribing a drug to a patient, where one needs to consider its efficacy, toxicity, and potentially all its side effects. The efficacy and various side effects typically depend on patient characteristics. Such examples can be found also in online platforms, e-commerce sites, and are pertinent to the design of most recommender systems.

The problem of selecting an action that has multiple attributes is typically modeled using the concept of Pareto optimality, and the learning problem reduces to identifying the Pareto front (Goel et al., 2007), i.e. the set of actions that are not dominated, and therefore, potentially optimal for some user. We consider Pareto front identification (PFI) for linear bandits (PFILin), where the attributes of each action are a linear function of an associated context. PFILin generalizes best arm identification (BAI) problems, PFI for multi armed bandits (MABs), and multi-objective active learning. We propose an algorithm `PFIwR` whose sample complexity is optimal within logarithmic factors.

A "good" PFI algorithm should ideally have a low sample complexity as well as low regret during the identification period. Degenne et al. (2019) and Zhong et al. (2023) discuss the trade-off between regret and sample complexity in the non-contextual single-objective BAI for drug testing. Such considerations are also important for e-commerce platforms where high regret could lead to low customer satisfaction and underexposure of products. While they present the lowest regret bound that all BAI algorithms (which successfully identify the best arm) can achieve, extensions to the multi-objective PFI remain challenging. The BAI problem has only *one* best arm; whereas the PFI algorithm does not know how many arms are on the Pareto front. Therefore, typical PFI algorithms are forced to collect samples from all arms until they identify *all* arms on the Pareto front, which incurs high regret, especially when the number of arms are exponentially large.

In order to minimize regret, the PFI algorithm should minimally explore high-regret actions while learning the parameters. We propose a novel exploration strategy that computes the basis vectors and supports the efficient parameter estimation. Furthermore, we propose the *exploration-mixed estimator* which "recycles" the samples in the exploration phase by bootstrap-

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Table 1: A comparison of the related works in terms of settings and theoretical guarantees.

	Bandit Setting	Multi-objective?	Regret Bound?	PAC bound?
Valko et al. (2013)	Kernel	×	√	×
Soare et al. (2014)	Linear	×	×	√
Zuluaga et al. (2016)	Gaussian Process	√	×	√
Auer et al. (2016)	Multi-armed	√	×	√
Lu et al. (2019)	Generalized linear	√	√	×
Degenne et al. (2019)	Multi-armed	×	√	√
Zhong et al. (2023)	Multi-armed	×	√	√
Our work	Linear	√	√	√

ping and mixing. The recycling allows us to explore high-regret actions only for logarithimically increasing exploration rounds, and exploiting low-regret actions after that. However, recycling samples may cause dependency, and higher estimation error as compared to that of the conventional estimators. We offset the higher error of the exploration-mixed estimator, by using a doubly-robust (DR) estimator (Bang and Robins, 2005), that is robust to the error of the estimator used to impute the rewards for actions that are not selected. These methods ensure we can simultaneously learn rewards for PFI and select arms to minimize regret. We show that our proposed algorithm has optimal sample complexity as well as optimal regret within logarithmic factors among all PFI algorithms that successfully identifies the Pareto front.

The main contributions of this paper are as follows:

(i) We introduce a novel estimation procedure for linear bandit feedback that ensures $\tilde{O}(\sqrt{d/t})$ convergence rate for the reward vectors of *all* arms while largely exploiting low regret arms (Theorem 4.2). This uniform convergence is possible due to three innovations: (i) the novel exploration strategy that finds the basis vectors over the context space (Section 4.1); (ii) the novel *exploration-mixed estimator* that reuses the observations in the past exploration rounds (Section 4.2); and (iii) construction of a DR estimate for unobserved rewards which is robust to the error of the exploration-mixed estimator (Section 4.3).

(ii) We apply the novel estimation paradigm to PFILIn and propose a new algorithm `PFIwR` with sample complexity that is optimal up to logarithmic factors (Theorem 3.3 and 5.1), and has $\tilde{O}(\sqrt{d/t})$ Pareto regret in round t with context dimension d, after $O(d^3 \log dt^2)$ initial exploration rounds independent of the problem complexity (Theorem 5.2). Further, the algorithm is shown to achieves optimal order regret among all PFI algorithms (Theorem 5.3).

(iii) Experimental results clearly show the estimator converges on the rewards of all contexts while exploiting low-regret arms, and `PFIwR` has significantly superior performance to previously known algorithms for *both* PFI and regret minimization.

2 Related Work

The typical approach in multi-objective rewards is to scalarize the problem by either setting the objective to be a weighted combination of all the objective (Roijers et al., 2017, 2018; Wanigasekara et al., 2019), or optimizing one while imposing constraints on the rest (Agrawal and Devanur, 2016; Kim et al., 2023a). While these approaches identify only one action on the Pareto front, we identify all actions on the Pareto front, i.e. identify the set of actions that are potentially optimal for any scalarization approach.

Table 1 compares our contribution with the existing bandit literature. The PFILin problem is a generalization of the BAI (Even-Dar et al., 2002; Soare et al., 2014) and single-objective regret minimization (Auer, 2002a; Valko et al., 2013) to the multi-objective vector rewards. Existing algorithms for multi-objective PFI problems have focused on the Gaussian reward setting (Zuluaga et al., 2016) and non-contextual MAB setting (Auer et al., 2016), and the optimal regret guarantees remain open. Lu et al. (2019) proposed an algorithm that achieves a bound on regret for multi-objective contextual bandits; however the identification of *all* arms in the Pareto front is not established. Similarly, lexicographic linear bandits (Xue et al., 2024) or hyper-volume regret minimizing algorithms (Zhang, 2024) minimize regret and the PFI is not guaranteed. While Degenne et al. (2019) and Zhong et al. (2023) obtained theoretical guarantees for both regret and sample complexity for non-contextual single-objective rewards, extension to linear and multi-objective rewards remains open. We establish bounds on the regret and sample complexity for multi-objective linear bandits.

3 Problem Formulation

In this section, we present the PFILin problem and the problem complexity terms for a lower bound of the sample complexity.

3.1 Pareto Front Identification for Linear Bandits

For a positive integer N, let $[N] := \{1, \ldots, N\}$. In PFILin, an action $k \in [K]$ is associated with a known d-dimensional context vector $x_k \in \mathbb{R}^d$. Let $\mathcal{X} := \{x_1, \ldots, x_K\}$. Without loss of generality, we assume $\|x_k\|_2 \leq 1$ and, as is standard in this literature (e.g., Tao et al. (2018)), we assume that $\mathcal{X}$ spans $\mathbb{R}^d$.

In period t, the decision-maker chooses an $a_t \in [K]$, and observes a sample of the random reward vector $\mathbf{Y}_{a_t,t} = \Theta_\star^\top x_{a_t} + \eta_t$, where $\Theta_\star := (\theta_\star^{\langle 1 \rangle}, \ldots, \theta_\star^{\langle \ell \rangle}) \in \mathbb{R}^{d \times L}$ is the unknown (but fixed) parameters with $\|\theta_\star^{\langle \ell \rangle}\|_2 \leq \theta_{\max}$, for all $l \in [L]$, and $\eta_t \in \mathbb{R}^L$ is a mean-zero, σ-sub-Gaussian random error vector that is independent of actions $\{a_s\}_{s \in [t-1]}$, and other error vectors $\{\eta_s \in \mathbb{R}^L : s \neq t\}$; however, we allow for the L components of η_t to be correlated. Let $y_k := \Theta_\star^\top x_k = \mathbb{E}[\mathbf{Y}_{k,t}] \in \mathbb{R}^L$ denote the true mean reward vector for arm $k \in [K]$. We want to identify the Pareto front of the $\{y_k\}_{k \in [K]}$ defined as follows.

Definition 3.1 (Pareto Front)**.** For vectors $a = (a^{(1)}, \ldots, a^{(L)})$, $b = (b^{(1)}, \ldots, b^{(L)}) \in \mathbb{R}^L$, b **dominates** a (denoted by $a \prec b$) if $a_\ell \leq b_\ell$, for all $\ell \in [L]$ and there exists $\ell \in [L]$ such that $a_\ell < b_\ell$. The **Pareto front** $\mathcal{P}_\star := \{k \in [K] | \nexists k' : y_k \prec y_{k'}\}$ is a set of arms whose mean reward vector is not dominated by the reward of any other arm.

The Pareto front always exists over arbitrary finite set of vectors and thus the reward vectors in our setting. To identify the Pareto front $\mathcal{P}_\star$, one must compute a reasonable estimate for the entire set of reward vectors $\{y_k\}_{k \in [K]}$. Let

$$\begin{aligned} m(k,j) &= \min\left\{\alpha \geq 0 \middle| \exists \ell \in [L] : y_k^{\langle \ell \rangle} + \alpha \geq y_j^{\langle \ell \rangle}\right\} \\ &= \max\left\{0, \ \min_{\ell \in [L]} (y_j^{\langle \ell \rangle} - y_k^{\langle \ell \rangle})\right\}, \end{aligned}$$

denote the amount by which arm j dominates arm k. We have $m(k,j) > 0$ if and only if $y_k^{\langle \ell \rangle} < y_j^{\langle \ell \rangle}$, for all $\ell \in [L]$. Therefore, the distance $\Delta_k^\star := \max_{k_\star \in \mathcal{P}_\star} m(k, k_\star)$ denotes the amount by which each component of the reward vector y_k must be increased to ensure that action k is not dominated by any Pareto optimal action $k_\star \in \mathcal{P}_\star$. By definition, the distance $\Delta_{k_\star}^\star = 0$ for all Pareto optimal actions $k_\star \in \mathcal{P}_\star$. Next, we define the PFI success condition.

Definition 3.2 (PFI success condition)**.** For precision $\epsilon > 0$ and confidence $\delta \in (0,1)$, a PFI algorithm must output a set of arms $\mathcal{P} \subseteq [K]$ such that, with probability at least $1 - \delta$,

$$\mathcal{P}_\star \subseteq \mathcal{P} \ \text{ and } \ \Delta_k^\star \leq \epsilon, \ \text{for all } k \in \mathcal{P} \setminus \mathcal{P}_\star. \tag{1}$$

The first condition in (1) ensures that $\mathcal{P}$ contains the Pareto optimal set $\mathcal{P}_\star$, and the second condition guarantees that the set $\mathcal{P}$ only includes arms sufficiently close to the Pareto front. Achieving this success condition is challenging and cannot be resolved by the multi-objective regret minimizing algorithms such as Xue et al. (2024) and Zhang (2024). For example, in a 2-dimensional instance with reward vectors: $y_1 = (1,0)$, $y_2 = (0,1)$, and $y_3 = (1/2, 0)$, the aforementioned algorithms only need to find *either* arm 1 *or* arm 2 to achieve zero Pareto regret. In contrast, to satisfy the success condition (1), the algorithm must identify *both* arms 1 and 2 while also sampling other arms to determine whether they are Pareto optimal.

Let $\tau_{\epsilon,\delta}$ denote the number of samples required for an algorithm to meet the success condition (1). Then the cumulative regret $R(\tau_{\epsilon,\delta})$ of an algorithm until round $\tau_{\epsilon,\delta}$ is defined as

$$R(\tau_{\epsilon,\delta}) := \sum_{t=1}^{\tau_{\epsilon,\delta}} \Delta_{a_t}^\star := \sum_{t=1}^{\tau_{\epsilon,\delta}} \max_{k^\star \in \mathcal{P}_\star} m(a_t, k^\star), \tag{2}$$

where a_t denotes the action selected by the algorithm. Our goal is to *simultaneously* establish the upper bounds of the sample complexity $\tau_{\epsilon,\delta}$ and the Pareto regret $R(\tau_{\epsilon,\delta})$.

3.2 Required Precision for Pareto Front Identification

For a Pareto sub-optimal arm k, if the estimate $\hat{y}_k$ of the reward vector of arm k has error $\min_{\ell \in [L]} |\hat{y}_k^{\langle \ell \rangle} - y_k^{\langle \ell \rangle}| > \Delta_k^\star$, it can erroneously appear Pareto optimal. Therefore, the required accuracy for a suboptimal arm $k \notin \mathcal{P}_\star$ is $\Delta_k = \Delta_k^\star$. Since the number of arms on the Pareto front is unknown, eliminating suboptimal arms is not sufficient and the algorithm must decide whether the remaining arms are all Pareto optimal or not. Thus, we need another complexity measure,

$$\begin{aligned} M(k,j) &:= \min\left\{\alpha \geq 0 \middle| \forall \ell \in [L] : y_k^{\langle l \rangle} \leq y_j^{\langle l \rangle} + \alpha\right\} \\ &= \max\left\{0, \ \max_{\ell \in [L]} \left(y_k^{\langle l \rangle} - y_j^{\langle l \rangle}\right)\right\}, \end{aligned}$$

which is the amount by which each component of the mean reward of arm j must be increased so that k is weakly dominated by j. Note that $M(k,j) = 0$ if and only if $y_k \preceq y_j$. While the metric $m(j,k)$ measures the shortest coordinate difference between the two arms, the metric $M(j,k)$ measures the maximum coordinate difference between the two arms. The metric $m(j,k)$ is used to measure the distance between a sub-optimal arm and the Pareto front; whereas $M(j,k)$ is used to measure the distance between the arms on the Pareto front.

Fix a Pareto optimal arm k. If the reward for arm k is underestimated by $M(k,j)$ with respect to a Pareto optimal arm j, it may appear weakly dominated by j. Thus, in order to prevent misidentifying the Pareto optimal arm k as a suboptimal arm, the error of the estimator has to be less than $\Delta_k^+ := \min_{j\in\mathcal{P}_\star\setminus\{k\}} \min\{M(k,j), M(j,k)\}$. Next, consider a suboptimal arm j. If the error of the estimator is greater than $M(j,k)+\Delta_j^\star$, the Pareto optimal arm k may appear dominated by suboptimal arm j. In order to distinguish the Pareto optimal arm k from the Pareto suboptimal arms, the error of the estimator has be to less than $\Delta_k^- := \min_{j\notin\mathcal{P}_\star}\{M(j,k)+\Delta_j^\star\}$. In summary, to identify whether arm k is in Pareto front, the estimation error has to be less than

$$\Delta_k := \begin{cases} \Delta_k^\star & k\notin\mathcal{P}_\star \\ \min\{\Delta_k^+,\Delta_k^-\} & k\in\mathcal{P}_\star \end{cases} \tag{3}$$

We index the arms in increasing order of required accuracy, i.e. $\Delta_{(1)}\le\cdots\le\Delta_{(K)}$.

Theorem 3.3 (A lower bound of the sample complexity for PFILin.). *Fix $\epsilon>0$, and let $\Delta_{(k),\epsilon} := \max\{\Delta_{(k)},\epsilon\}$. Suppose the set of context vectors $\mathcal{X}$ spans $\mathbb{R}^d$ and $\|\theta_\star^{\langle\ell\rangle}\|_0 = d$, for all $\ell\in[L]$. Then, for any $\delta\in(0,1/4)$ and $\sigma>0$, there exist a σ-Gaussian distribution for the i.i.d. noise sequence $\{\eta_t\}_{t\ge1}$ such that any algorithm requires at least $(\sigma^2/3)\sum_{k=1}^d \Delta_{(k),\epsilon}^{-2}\log(3L/4\delta)$ rounds to meet the success condition* (1).

Theorem 3.3 generalizes the lower bound in Auer et al. (2016) to the linear bandit setting. Since $\theta_\star^{\langle\ell\rangle}\in\mathbb{R}^d$, the number of rounds required for PFI depends only on the d smallest gaps instead of all K gaps.

4 Proposed Method

Our main contribution is a novel estimation strategy that simultaneously learns rewards of *all* actions while largely exploiting low-regret arms. Our strategy is applicable to more broadly to online learning problems under linear bandit feedback, e.g., BAI (Tao et al., 2018), policy optimization in reinforcement learning (He et al., 2021). We now describe in more detail each ingredient in our approach and its theoretical properties.

4.1 Exploration Strategy with Context Basis

Let $X=[x_1,\ldots,x_K]\in\mathbb{R}^{d\times K}$ denote the matrix of contexts vectors. Using the (reduced) singular value decomposition (SVD), one can compute orthonormal vectors $\{u_i\in\mathbb{R}^d : i\in[d]\}, \{v_i\in\mathbb{R}^K : i\in[d]\}$ and scalars $\{\lambda_i\ge 0 : i=1,\ldots,d\}$ such that $X=\sum_{i=1}^d \lambda_i u_i v_i^\top$. Thus, it follows that $v_i^\top X^\top\theta_\star^{\langle\ell\rangle} = \sqrt{\lambda_i}u_i^\top\theta_\star^{\langle\ell\rangle}$, for $\ell\in[L]$ and $i\in[d]$. For $i\in[d]$, let $\pi^{(i)}\in\mathbb{R}_+^K$ denote the probability mass function $\pi_k^{(i)} = |v_{ik}|/\|v_i\|_1$ over actions $k\in[K]$. Then, for a randomized action $a\sim\pi^{(i)}$, we have

$$\begin{aligned}\mathbb{E}\Big[\|v_i\|_1\mathrm{sign}(v_{ia})Y_{a,s}^{\langle\ell\rangle}\Big] &= \mathbb{E}\Big[\sum_{k=1}^K v_{ik}Y_{k,s}^{\langle\ell\rangle}\Big] \\ &= \sum_{k=1}^K v_{ik}x_k^\top\theta_\star^{\langle\ell\rangle} = v_i^\top X^\top\theta_\star^{\langle\ell\rangle} = (\sqrt{\lambda_i}u_i)^\top\theta_\star^{\langle\ell\rangle},\end{aligned} \tag{4}$$

Thus, $\|v_i\|_1\mathrm{sign}(v_{ia})Y_{a,s}^{\langle\ell\rangle}$ can be viewed as the random reward corresponding to the "context basis" $\sqrt{\lambda_i}u_i$. Sampling $a_i\sim\pi^{(i)}$ for $i\sim\mathrm{unif}([d])$ yields the design matrix $d^{-1}\sum_{i=1}^d\lambda_i u_i u_i^\top = d^{-1}\sum_{k=1}^K x_k x_k^\top$ that satisfies $\max_{k\in[K]}\|x_k\|^2_{(d^{-1}\sum_{k'=1}^K x_{k'}x_{k'}^\top)^{-1}}\le d$ (see Section B.3 for the derivation of the inequality). This design yields a tighter bound than the G-optimal design that is widely used in BAI problems (see e.g., Tao et al. (2018)).

For each $t\ge1$, let $\mathcal{E}_t\subset[t]$ denote the set of rounds reserved for exploration. Fix a confidence level $\delta\in(0,1)$, and let $\gamma_t := Cd^3\log(2dt^2/\delta)$ where C is an absolute constant specified in (29). Define $\mathcal{E}_0:=\emptyset$, and in each round $t\ge1$, sample a basis index $i_t\sim\mathrm{unif}([d])$ and sample the action $\breve{a}_t\sim\pi^{(i_t)}$. Define

$$\mathcal{E}_t := \begin{cases}\mathcal{E}_{t-1}, & \text{if } \sum_{u\in\mathcal{E}_t}\mathbb{I}(\breve{a}_u=\breve{a}_t) > \frac{\gamma_t}{t}\sum_{s=1}^t\mathbb{I}(\breve{a}_s=\breve{a}_t) \\ \mathcal{E}_{t-1}\cup\{t\}, & \text{otherwise.}\end{cases} \tag{5}$$

By construction, the set $\mathcal{E}_t$ satisfies $\sum_{u\in\mathcal{E}_t}\mathbb{I}(\breve{a}_u=\breve{a}_t)\ge(\gamma_t/t)\sum_{s=1}^t\mathbb{I}(\breve{a}_s=\breve{a}_t)$ for all $t\ge T_\gamma := 8Cd^3\big(1+\log\frac{4Cd^3\sqrt{2d}}{e\sqrt{\delta}}\big)\ge\inf\{s\ge1 : s\ge\gamma_s\}$ and the cardinality $|\mathcal{E}_t|\ge\gamma_t$ increases logarithmically in t (See Appendix B.7 for details.) The condition ensures that the number of rounds exploring an arm $\breve{a}_t$ is sufficient enough to represent the distribution of $\breve{a}_s\sim\pi^{(i_s)}$ and $i_s\sim\mathrm{unif}([d])$.

When $t\in\mathcal{E}_t$, i.e, in an exploration round, the algorithm selects the sampled action $a_t=\breve{a}_t$, and when $t\notin\mathcal{E}_t$, the algorithm choose an arm a_t from the set of unidentified arms that has low estimated regret.

4.2 Recycling Reward Samples in the Exploration Phase

For each $t\ge T_\gamma$ and $t\in[t]\setminus\mathcal{E}_t$, i.e., when t is in *exploitation* phase, let a_t denote the (low-regret) arm chosen by the algorithm. Recall that at the beginning of each round t, the context basis index $i_t\sim\mathrm{unif}([d])$, and $\breve{a}_t\sim\pi^{(i_t)}$. In order to learn rewards of *multiple* arms, we "recycle" the reward sample observed in a previous exploration round by bootstrapping as follows. Let

$\mathcal{E}_t(\check{a}_t) = \{s \in \mathcal{E}_t : a_s = \check{a}_t\}$ denote the set of previous exploration rounds where the action $\check{a}_t$ was chosen. Note that, by definition of $\mathcal{E}_t$ in (5), we are guaranteed that $\mathcal{E}_t(\check{a}_t) \neq \emptyset$ for $t \geq T_\gamma$. For the exploitation rounds $\tau \in [t-1] \setminus \mathcal{E}_{t-1}$, let $\check{n}_\tau$ denote time index of the exploration sample "recycled" at exploitation round τ and "mixed" with the chosen action a_τ. We "mix" the action a_t with the exploration sample "recycled" from round

$$\check{n}_t := \underset{n \in \mathcal{E}_t(\check{a}_t)}{\arg\min} \sum_{\tau \in [t-1] \setminus \mathcal{E}_{t-1}} \mathbb{I}\big(\check{n}_\tau = n\big), \tag{6}$$

i.e. we want to balance the reuse choice over the set $\mathcal{E}_t(\check{a}_t)$. We define the exploration-mixed contexts and rewards as follows: for all $\ell \in [L]$,

$$\begin{aligned} \check{X}_{a_t,t} &:= w_t x_{a_t} + \check{w}_t \sqrt{\lambda_{i_t}} u_{i_t}, \\ \check{Y}^{\langle \ell \rangle}_{a_t,t} &:= w_t Y^{\langle \ell \rangle}_{a_t,t} + \check{w}_t \|v_{i_t}\|_1 \mathrm{sign}(v_{i_t,a_{\check{n}_t}}) Y^{\langle \ell \rangle}_{a_{\check{n}_t},\check{n}_t}, \end{aligned} \tag{7}$$

where $w_t, \check{w}_t \sim \mathrm{unif}[-\sqrt{3}, \sqrt{3}]$ are sampled independently. The following properties follow from the linear structure and the distribution of weights $(w_t, \check{w}_t)$ that has mean zero and unit variance, and the definition of the reduced SVD (u_i, v_i, λ_i) for $i \in [d]$.

Lemma 4.1. *(Exploration-mixed contexts and rewards.) Let $\mathcal{F}_t$ denote the sigma-algebra generated by $\{a_s, \mathbf{Y}_{a_s,s}\}_{s=1}^{t-1}$ and $a_t \in \mathcal{A}$. For any $\ell \in [L]$ and t such that $t \notin \mathcal{E}_t$, $\mathbb{E}[\check{Y}^{\langle \ell \rangle}_{a_t,t} - \check{X}^\top_{a_t,t} \theta^{\langle \ell \rangle}_\star | \mathcal{F}_t] = 0$, and $\mathbb{E}[\check{X}_{a_t,t} \check{X}^\top_{a_t,t} | \mathcal{F}_t] \succeq d^{-1} \sum_{k=1}^K x_k x_k^\top$.*

We can view $\check{X}_{a_t,t}, \check{Y}^{\langle \ell \rangle}_{a_t,t}$ as a stochastic feedback from a new linear bandit problem with the same parameters $\{\theta^{\langle \ell \rangle}_\star\}_{\ell \in [L]}$. Since the random contexts $\check{X}_{a_t,t}$ contains the (randomized) context basis, the (expected) design matrix includes information on all K arms for any selected action a_t. "Recycling" the reward sample $Y^{\langle \ell \rangle}_{a_{\check{n}_t},\check{n}_t}$ allows us to get information on the rewards of the unselected (and hence unobserved) contexts while exploiting low regret action. Next, we define the **exploration-mixed estimator**,

$$\begin{aligned} \check{\theta}^{\langle \ell \rangle}_t := &\Big(\sum_{s \in \mathcal{E}_t} x_{a_s} x_{a_s} + \sum_{s \in [t] \setminus \mathcal{E}_t} \check{X}_{a_s,s} \check{X}^\top_{a_s,s} + \frac{1}{2} I_d \Big)^{-1} \\ &\Big(\sum_{s \in \mathcal{E}_t} x_{a_s} Y^{\langle l \rangle}_{a_s,s} + \sum_{s \in [t] \setminus \mathcal{E}_t} \check{X}_{a_s,s} \check{Y}^{\langle l \rangle}_{a_s,s} \Big). \end{aligned} \tag{8}$$

While the exploration-mixed estimator gains information on the unknown parameter on *multiple contexts*, reusing samples from previous rounds causes dependency that complicates the analysis of the convergence rate of the estimator (see Section B.5 for details). To address this, we apply the doubly-robust (DR) technique from the missing data literature instead of directly using the exploration-mixed estimator, as we explain next.

4.3 Doubly-Robust Estimation

Doubly-robust estimation uses an estimate to impute the missing value, and is robust to the estimation error for the missing value. We defer the details on the robustness of the doubly-robust estimation in Appendix B.6.

In each round t, the unselected rewards $\{Y^{\langle \ell \rangle}_{k,t} : \ell \in [L], k \in [K] \setminus \{a_t\}\}$ are missing. One possible approach is computing a ridge estimator $\theta^{\langle \ell \rangle}_R$ and imputing $x_k^\top \theta^{\langle \ell \rangle}_R$ for $Y^{\langle \ell \rangle}_{k,t}$ to apply doubly-robust estimation, as proposed in Kim et al. (2021, 2022, 2024). However, their approach assumes stochastic contexts that are iid over rounds with finite K, and therefore, not applicable to PFILin where the contexts are fixed and K can be exponentially large. Further, since the ridge estimator only gains information on the selected actions, their DR estimator does not converge while exploiting low regret arms (see Appendix A.2 for detailed comparisons).

We first reduce K rewards into $d+1$ rewards using the basis defined in (4): $\tilde{Y}^{\langle \ell \rangle}_{i,t} := \sum_{k=1}^K v_{i,k} Y^{\langle \ell \rangle}_{k,t}$, corresponding to the d context basis $\sqrt{\lambda_i} u_i$, $i = 1, \ldots, d$, and $\tilde{Y}^{\langle \ell \rangle}_{d+1,t} := Y^{\langle \ell \rangle}_{a_t,t}$. Note that $\mathbb{E}[\tilde{Y}^{\langle \ell \rangle}_{i,t}] = (\sqrt{\lambda_i} u_i)^\top \theta^{\langle \ell \rangle}_\star$ for $i \in [d]$, and we learn $\Theta_\star$ using d context basis vectors $\{\sqrt{\lambda_i} u_i,\ i \in [d]\}$, instead of K contexts. Here, $\{\tilde{Y}^{\langle \ell \rangle}_{i,t} : i \in [d]\}$ is missing data, and only $\tilde{Y}^{\langle \ell \rangle}_{d+1,t}$ is observed. To induce a specified probability of observation, needed to ensure robustness of the DR estimation, we define a probability mass function $\tilde{\pi}$ defined as follows:

$$\tilde{\pi}_i = 1/(2d),\ \forall i = 1, \ldots, d, \qquad \tilde{\pi}_{d+1} = 1/2, \tag{9}$$

and let $\tilde{a}_t \sim \tilde{\pi}$ denote the pseudo-action on $d+1$ arms.

To couple the observed reward $Y^{\langle \ell \rangle}_{a_t,t}$ and the randomly selected reward $\tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t,t}$, we resample both action a_t and pseudo-action $\tilde{a}_t$ until the matching event $\{Y^{\langle \ell \rangle}_{a_t,t} = \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t,t}\} = \{\tilde{a}_t = d+1\}$ happens. For given $\delta' \in (0,1)$, let $\mathcal{M}_t$ denote the event of obtaining the matching $\{Y^{\langle \ell \rangle}_{a_t,t} = \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t,t}\}$ within $\rho_t := \log((t+1)^2/\delta')/\log(2)$ number of resampling so that the event $\mathcal{M}_t$ happens with probability at least $1 - \delta'/(t+1)^2$. If the event $\mathcal{M}_t$ does not happen, we do not update the estimator and use the estimator value obtained in the previous round. While Xu and Zeevi (2020) samples counterfactual pseudo-actions from the policies in previous rounds (which may have zero support on some arms and not applicable to DR estimation), our resampling and coupling method samples the pseudo-action from the (reduced) policy $\tilde{\pi}$ as in (9) that has non-zero support on all arms.

Define new contexts $\tilde{x}_{i,t} := \sqrt{\lambda_i} u_i$ for $i \in [d]$ and $\tilde{x}_{d+1,t} = x_{a_t}$. With the coupled pseudo-action $\tilde{a}^{(t)}$ and its distribution $\tilde{\pi}$, we construct the DR estimate for the reduced missing rewards for $i \in [d+1]$ as:

$$\widehat{Y}_{i,t}^{\langle \ell \rangle} := \tilde{x}_{i,t}^\top \check{\theta}_t^{\langle \ell \rangle} + \frac{\mathbb{I}(\tilde{a}_t = i)}{\tilde{\pi}_i} (\tilde{Y}_{i,t}^{\langle \ell \rangle} - \tilde{x}_{i,t}^\top \check{\theta}_t^{\langle \ell \rangle}). \quad (10)$$

For $i \neq \tilde{a}_t$, we impute a reward $\tilde{x}_{i,t}^\top \check{\theta}_t^{\langle \ell \rangle}$ for the new "context" basis $\tilde{x}_{i,t}$ For $i = \tilde{a}_t$, the second term $(Y_{\tilde{a}_t,t}^{\langle \ell \rangle} - \tilde{x}_{\tilde{a}_t,t}^\top \check{\theta}_t^{\langle \ell \rangle})/\tilde{\pi}_{\tilde{a}_t}$ corrects the imputed reward to ensure unbiasedness of the pseudo-rewards for all arms. Taking the expectation over $\tilde{a}_t$ on both sides of (10) gives $\mathbb{E}[\widehat{Y}_{i,t}^{\langle \ell \rangle}] = \tilde{x}_{i,t}^\top \theta_\star^{\langle \ell \rangle}$ for all $i \in [d+1]$. Therefore, we define our **DR-mix estimator** as a ridge estimator using $\{(\widehat{Y}_{i,s}^{\langle \ell \rangle}, \tilde{x}_{i,s}) : s = 1, \ldots, t, i = 1, \ldots, d+1\}$:

$$\widehat{\theta}_t^{\langle \ell \rangle} = \Big(\sum_{s:\mathbb{I}(\mathcal{M}_s)=1}^{t} \sum_{i=1}^{d+1} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top + I_d \Big)^{-1} \Big(\sum_{s:\mathbb{I}(\mathcal{M}_s)=1}^{t} \sum_{i=1}^{d+1} \tilde{x}_{i,s} \widehat{Y}_{i,s}^{\langle \ell \rangle} \Big). \quad (11)$$

The DR-mix estimator (11) is recursively computable with a rank-1 update of the Gram matrix and summation of the weighted context vectors (see Appendix A.4 for the complexity of the estimator).

A special case of the DR estimation is inverse probability weighted or importance weight estimator used in `EXP3` algorithm (Auer et al., 2002b), which is achieved by setting the imputation estimator $\check{\theta}_t^{\langle \ell \rangle} = 0$. While their analysis involves only the importance-weighted term $\mathbb{I}(\tilde{a}_t = i)/\tilde{\pi}_i \tilde{Y}_{i,t}^{\langle \ell \rangle}$, our method introduces an additional term, $(\mathbb{I}(\tilde{a}_t = i)/\tilde{\pi}_i - 1)\tilde{x}_{i,t}^\top \check{\theta}_t^{\langle \ell \rangle}$ in (10). The handling and analysis of this exploration-mixed term $\check{\theta}_t$ requires a novel matrix martingale inequality (Lemma C.5) and a coupling inequality (Lemma B.3), which is not required in the analysis of `EXP3` algorithms and represents a significant innovation in the proof technique.

Theorem 4.2 (Estimation error bound for the DR-mix estimator)**.** *Let $\widehat{\theta}_t^{\langle \ell \rangle}$ denote the DR-mix estimator* (11) *with the exploration-mixed estimator* (8) *as the imputation estimator and pseudo-rewards* (10)*. Let $F_t := \sum_{k=1}^K x_k x_k^\top + I_d$, Then, for all $k \in [K]$, $\ell \in [L]$, and $t \geq T_\gamma$,*

$$\left| x_k^\top (\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}) \right| \leq 3 \|x_k\|_{F_t^{-1}} \Big(\theta_{\max} + \sigma \sqrt{d \log \frac{Lt}{\delta}} \Big). \quad (12)$$

with probability at least $1 - 7\delta$*. For each $k \in [K]$, with probability at least* $1 - 7\delta$*,*

$$\left| x_k^\top (\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}) \right| \leq 3 \|x_k\|_{F_t^{-1}} \Big(\theta_{\max} + 3\sigma \sqrt{\log \frac{4Lt^2}{\delta}} \Big). \quad (13)$$

The DR-mix estimator has two bounds: (i) a union bound (12) for all K arms that has $\sqrt{d}$ rate and (ii) a bound for each arm (13) without $\sqrt{d}$ rate. When the number of undetermined arms is $\Omega(K)$, we use the union bound (12) instead of the bound (13) which yields $\sqrt{\log(K/\delta)}$ term for all $\Omega(K)$ arms. After eliminating the suboptimal arms and when the number of undetermined arms is $O(d)$, we can use the bound (13) for each remaining arm. The two separate bounds are used to avoid K dependency on the error bound.

The major difference between the conventional estimator and our DR-mix estimator appears in the associated Gram matrix. The Gram matrix $A_t = \sum_{s=1}^t x_{a_s} x_{a_s}^\top + I_d$ of the conventional estimators sums over *only* the selected contexts. In the norm,

$$\Big\| \sum_{s=1}^t \eta_{a_s} x_{a_s} \Big\|_{A_t^{-1}} = \Big\| \sum_{s=1}^t \eta_{a_s} A_t^{-1/2} x_{a_s} \Big\|_2,$$

the a_s and A_t are *not* independent and the conditional expectation of $\eta_{a_s} A_t^{-1/2} x_{a_s}$ is not zero. Thus, one may use the self-normalizing bound in Abbasi-Yadkori et al. (2011) to obtain $\| \sum_{s=1}^t \eta_{a_s} x_{a_s} \|_{A_t^{-1}} = O(\sqrt{d \log(t^2/\delta)})$, where additional $\sqrt{d}$-term appears. In contrast, the *full* Gram matrix $F_t := \sum_{s=1}^t \sum_{k=1}^K x_k x_k^\top + I_d$ of the DR-mix estimator makes the conditional expectation of $\eta_{a_s} F_t^{-1/2} x_{a_s}$ zero because a_s is independent of F_t. Then we can apply the dimension-free martingale inequality (Lemma C.2) to obtain,

$$\Big\| \sum_{s=1}^t \eta_{a_s} x_{a_s} \Big\|_{F_t^{-1}} = \Big\| \sum_{s=1}^t \eta_{a_s} F_t^{-\frac{1}{2}} x_{a_s} \Big\|_2 = O\Big(\sqrt{\log \frac{t^2}{\delta}} \Big).$$

Thus, the difference of using the Gram matrix as A_t and F_t gives $\sqrt{d}$ difference in (13). In Addition, we derive $\max_{k \in [K]} \|x_k\|_{F_t^{-1}} \leq t^{-1/2}$ in Appendix B.3. Thus, with only $\gamma_t = O(d^3 \log dt^2)$ exploration rounds, DR-mix estimator achieves $\widetilde{O}(\sqrt{d/t})$ convergence rate for the reward estimates of *all* arms. Therefore, our estimator enjoys the freedom to choose low-regret arms while simultaneously learning the rewards on all arms.

5 Pareto Front Identification with Regret Minimization

In this section, we apply our novel estimation strategy to PFILin and establish novel algorithm with nearly optimal sample complexity and regret.

5.1 `PFIwR` Algorithm for Linear Bandits

Our proposed algorithm, PFI with regret minimization (`PFIwR`), is displayed in Algorithm 1. While any undetermined arms remains, the algorithm employs our novel estimation strategy to compute the reward

Algorithm 1 Pareto Front Identification with Regret Minimization (PFIwR).

INPUT: context matrix $X = [x_1, \ldots, x_K]$, accuracy parameter $\epsilon > 0$, confidence $\delta > 0$.
Set $\mathcal{A}_0 = [K]$, $\mathcal{P}_0 = \mathcal{E}_0 = \emptyset$ and $\widehat{\theta}_0^{\langle l \rangle} = \mathbf{0}, \forall \ell \in [L]$ and apply reduced SVD on $X = UDV^\top$.
while $\mathcal{A}_t \neq \emptyset$ **do**
  Sample $i_t \sim \text{unif}([d])$ and $\breve{a}_{i_t} \sim \pi^{(i_t)}$ and update $\mathcal{E}_t$ as in (5).
  if $t \in \mathcal{E}_t$: set $a_t = \breve{a}_{i_t}$ **else:** randomly sample a_t over $\{k \in \mathcal{A}_{t-1} : \nexists k' \in \mathcal{A}_{t-1}, \widehat{y}_{k,t} \prec \widehat{y}_{k',t}\}$.
  Sample $\tilde{a}_t$ from the distribution (9), and set $m = 1$
  while $\tilde{a}_t \neq d+1$ and $m \leq \rho_t$ **do**
    Resample a_t and $\tilde{a}_t$, and set $m \leftarrow m + 1$.
  Select action a_t and observe $Y_{a_t,t}^{\langle \ell \rangle}$ for $\ell \in [L]$.
  if $m \leq \rho_t$, (i.e., the resampling succeeds) **then**
    Update the estimators $\{\widehat{\theta}_t^{\langle \ell \rangle}\}_{\ell \in [L]}$ as in (11) then compute $\widehat{y}_{k,t}^{(\ell)} := x_k^\top \widehat{\theta}_t^{\langle \ell \rangle}$ for all $k, k' \in \mathcal{A}_t$,

$$\begin{aligned}\widehat{m}_t(k,k') &:= \min\{\alpha \geq 0 \,|\, \exists \ell \in [L] : \widehat{y}_{k,t}^{(\ell)} + \alpha \geq \widehat{y}_{k',t}^{(\ell)}\},\\ \widehat{M}_t^{2\epsilon}(k,k') &:= \min\{\alpha \geq 0 \,|\, \forall \ell \in [L] : \widehat{y}_{k,t}^{(\ell)} + 2\epsilon \leq \widehat{y}_{k',t}^{(\ell)} + \alpha\}.\end{aligned} \tag{14}$$

    Compute confidence intervals,

$$\beta_{k,t} := \begin{cases} 3\|x_k\|_{F_t^{-1}}\Big(\theta_{\max} + \sigma\sqrt{d \log \frac{7Lt}{\delta}}\Big) & |\mathcal{A}_t| > d \\ 3\|x_k\|_{F_t^{-1}}\Big(\theta_{\max} + 3\sigma\sqrt{\log \frac{56Ldt^2}{\delta}}\Big) & |\mathcal{A}_t| \leq d. \end{cases} \tag{15}$$

    Estimate Pareto front,

$$\begin{aligned}\mathcal{C}_t &:= \{k \in \mathcal{A}_{t-1} | \forall k' \in \mathcal{A}_{t-1} \cup \mathcal{P}_{t-1} : \widehat{m}_t(k,k') \leq \beta_{k,t} + \beta_{k',t}\},\\ \mathcal{P}_t^{(1)} &:= \{k \in \mathcal{C}_t | \forall k' \in \mathcal{C}_t \cup \mathcal{P}_{t-1} \setminus \{k\} : \widehat{M}_t^{2\epsilon}(k,k') \geq \beta_{k,t} + \beta_{k',t}\}.\end{aligned} \tag{16}$$

    Update $\mathcal{P}_t \leftarrow \mathcal{P}_{t-1} \cup \mathcal{P}_t^{(1)}$, $\mathcal{A}_t \leftarrow \mathcal{C}_t \setminus \mathcal{P}_t^{(1)}$.
  else
    Update the estimator $\widehat{\theta}_t^{\langle \ell \rangle} \leftarrow \widehat{\theta}_{t-1}^{\langle \ell \rangle}$ for $\ell \in [L]$.
OUTPUT: $\mathcal{P}_t$

estimates $\widehat{y}_{k,t}^{(\ell)} := x_k^\top \widehat{\theta}_t^{\langle \ell \rangle}$ for $k \in \mathcal{A}_t$ and gap estimates required for PFI. The algorithm uses two different confidence bound in (15) based on the two convergence rates in Theorem 4.2. The first bound uses the union bound (12) because the estimator must converge on all K arms. However, when the number of undetermined arms are less than d, we need at most d reward estimate for $k \in \mathcal{A}_t$ and at most d estimates for the arm in $[K] \setminus \mathcal{A}_t$ that are "nearest" to $\mathcal{A}_t$ and critically affect the PFI. The arms in $[K] \setminus \mathcal{A}_t$ are already distinguishable each other even with the first bound (12) and the tighter bound (13) (without $\sqrt{d}$) applies effectively to the undetermined arms and their nearest arms.

Based on the two confidence bounds, PFIwR computes the set $\mathcal{C}_t$ by eliminating the suboptimal arms that are dominated by other arms by the amount more than the confidence bound $\beta_{k,t} + \beta_{k',t}$. The set $\mathcal{P}_t^{(1)}$ is the current estimate for ϵ-Pareto optimal arms that are not dominated by any other arms. This arm elimination step is simplified compared to that in Auer et al. (2016). The algorithm in Auer et al. (2016) leaves the identified Pareto optimal arm undetermined until all suboptimal arms dominated by the identified arm are eliminated, in order to ensure that a dominated arm is not spuriously declared as Pareto optimal. In contrast, PFIwR does not keep the (identified) dominating arms in $\mathcal{A}_t$. This is because the DR estimate converges on rewards of *all* arms in $\mathcal{A}_t \cup \mathcal{P}_t$, including the identified arms, in contrast to the conventional estimator that does not converge on identified arms unless they are selected. Consequently, the cardinality $|\mathcal{A}_t|$ of the set undetermined arms decreases faster in PFIwR (derived in Appendix B.9), and this allows the algorithm to invoke the $\sqrt{d}$-free confidence bound (15) in earlier rounds.

In addition to efficient estimation for PFI, the proposed PFIwR is able to choose low estimated regret actions after $O(d^3 \log dt)$ exploration rounds. The novel estimator and its convergence (Theorem 4.2) ensure that sampling arms with low estimated regret does not harm the convergence rate of the reward estimates of other arms. Thus, PFIwR is efficient in both for PFI and minimizing regret.

5.2 Sample Complexity and Regret Analysis

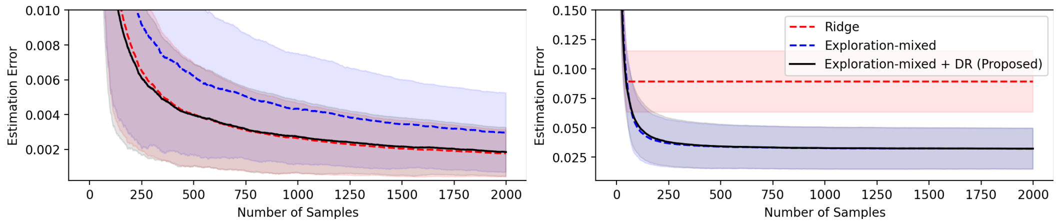
Theorem 5.1 (An upper bound on sample complexity)**.** *Fix $\epsilon > 0$ and $\delta \in (0,1)$. Let $\Delta_{(k),\epsilon} = \max\{\Delta_{(k)}, \epsilon\}$, where $\Delta_{(k)}$ is the ordered gap Δ_k defined in (3) in increasing order. Then the stopping time $\tau_{\epsilon,\delta}$ of PFIwR is bounded above by*

$$\max\Big\{O\Big(\sum_{k=1}^{d} \frac{(\theta_{\max}+\sigma)^2}{\Delta_{(k),\epsilon}^2} \log \frac{(\theta_{\max}+\sigma)dL}{\Delta_{(k),\epsilon}\delta}\Big), T_\gamma\Big\},$$

where T_γ denotes an upper bound on the initial exploration rounds.

The proof and explicit finite-sample bound for $\tau_{\epsilon,\delta}$ is in Appendix B.10. When the contexts are Euclidean basis, our result directly applies to the PFI in the MAB setting studied by Auer et al. (2016). The sample complexity is optimal within a logarithm factor of the lower bound in Theorem 3.3.

Theorem 5.2 (Upper bounds on Pareto regret)**.** *Fix $\epsilon > 0$ and $\delta \in (0,1)$. Let $\Delta_\epsilon^\star := \max\{\epsilon, \min_{k \in [K] \setminus \mathcal{P}_\star} \Delta_k^\star\}$ denote the minimum Pareto regret over suboptimal arms. Then, with probability at least $1-\delta$, the instantaneous Pareto regret, $\Delta_{a_t}^\star \leq 2\max_{j \in \mathcal{A}_{t-1}} \beta_{j,t-1}$, for all $t \notin \mathcal{E}_t$ and $t \geq T_\gamma$, where $\beta_{j,t}$ is the error bound defined in (15). The*



(a) Error $|\widehat{\theta}_1 - \theta_\star^{(1)}|$ for the exploited arm (arm 1) (b) Error $\|\widehat{\theta}_{\{2,3\}} - \theta_\star^{(\{2,3\})}\|_2$ for arm 2 and 3

Figure 1: Estimation errors of the proposed DR-mix estimator (11) with the conventional ridge estimator, and the exploration-mixed estimator (8) for a 3-armed bandit problem. The line and shade represent the average and standard deviation over 1000 independent experiments. The estimators use samples from all arms for $n \in [50]$, and after that, only observe rewards from arm 1.

cumulative Pareto regret of `PFIwR`,

$$R(\tau_{\epsilon,\delta}) = \bar{O}\Big(\theta_{\max} d^3 \log \frac{\theta_{\max} d}{\delta \Delta_{(1),\epsilon}} + \frac{\theta_{\max} d\sigma}{\Delta_\epsilon^\star} \log \frac{\theta_{\max} d\sigma}{\Delta_\epsilon^\star \delta}\Big)$$

with probability at least $1-\delta$, *where* $\bar{O}$ *ignores* $\log\log(\cdot)$ *terms.*

The explicit expression for the finite-sample bound is in Appendix B.11. The first term is the regret from the exploration rounds $\mathcal{E}_t$, whose cardinality $|\mathcal{E}_t| = O(d^3 \log dt)$ for all t. Since the algorithm need to increase $\mathcal{E}_t$ until it identifies all arms on the Pareto front, the bound involves $\Delta_{(1),\epsilon}$, which is the cost for identifying all arms in the Pareto front. When $L = 1$ and the contexts are Euclidean basis, Theorem 5.1 and Theorem 5.2 recovers the sample complexity bound and regret bound for the best arm identification in MAB setting established by Degenne et al. (2019) and Zhong et al. (2023). We show that `PFIwR` establishes nearly optimal regret among algorithms that achieve PFI.

Theorem 5.3 (A lower bound for the regret in PFILin). *For* $\epsilon > 0$, *let* $\Delta_\epsilon^\star := \max\{\epsilon, \min_{k\in[K]\setminus\mathcal{P}_\star} \Delta_k^\star\}$ *denote the minimum Pareto regret over suboptimal arms. Suppose the set of context vectors* $\mathcal{X}$ *span* $\mathbb{R}^d$ *and* $\min_{\ell\in[L]} \|\theta_\star^{\langle\ell\rangle}\|_0 = d$. *Then, for any* $\delta \in (0, 1/4)$ *and* $\sigma > 0$, *there exists a* σ*-sub-Gaussian distribution for the i.i.d. noise sequence* $\{\eta_t\}_{t\geq 1}$ *such that for any PFI algorithms that satisfies PFI success condition (1) with failure probability* δ, *the regret* $R(\tau_{\epsilon,\delta}) \geq \frac{\sqrt{3}d\sigma}{8\Delta_\epsilon^\star} \log \frac{1}{4\delta}$.

Theorem 5.3 is the first result on the trade-off between PFI and Pareto regret minimization. For $L = 1$ and the contexts are Euclidean basis, Theorem 5.3 recovers the lower bound for regret of BAI algorithms developed by Zhong et al. (2023). Note that the lower bound applies only to the algorithms that guarantee PFI; it is possible for an algorithm that does not guarantee PFI to have a regret lower bound that is lower than the one in Theorem 5.3.

6 Experiments

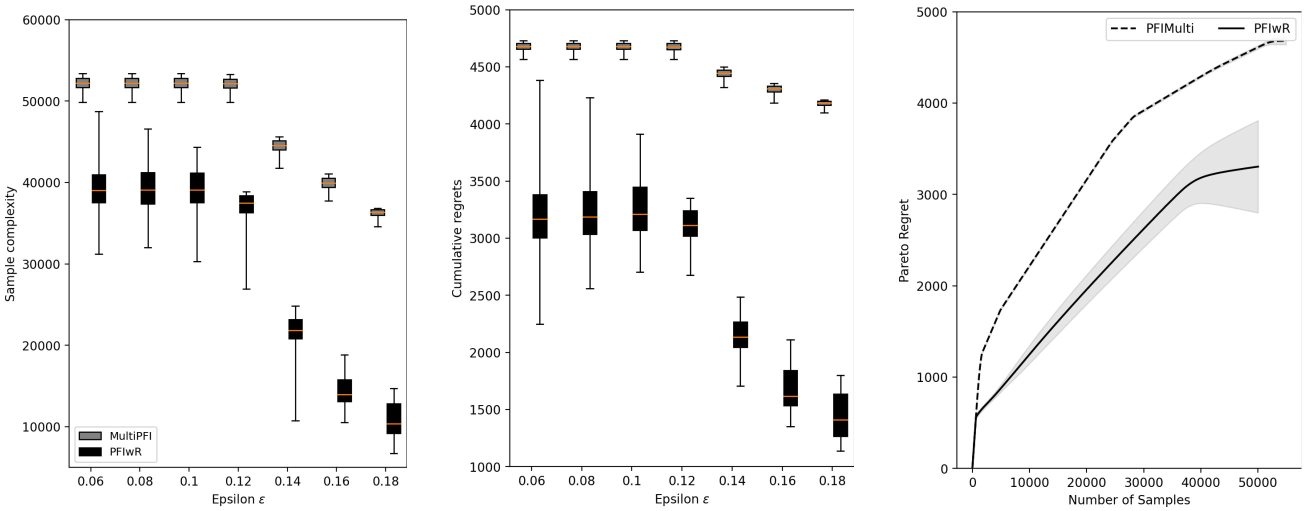
6.1 Consistency of the Proposed Estimator on All Actions

We conduct the following experiment to empirically verify that our proposed DR-mix estimator converges on *all* arms while exploiting low-regret arms. We consider a 3-arm bandit, i.e. the context vectors are the Euclidean basis in $\mathbb{R}^3$. The parameter $\theta_\star^{\langle 1\rangle} = (1, -1, -1)^\top$, and the random error is sampled from centered Gaussian distribution with variance $\sigma^2 = 0.01$. In rounds $n \leq 50$, each of three arms is pulled with equal probability; in rounds $n > 50$, only the optimal arm (arm 1) is pulled.

The plots in Figures 1a and 1b illustrate the reward error of the proposed DR-mix estimator, the conventional ridge estimator, and an exploration-mixed estimator defined in (8) as a function of the number of rounds n. The conventional ridge estimator converges only on the arm that is pulled (arm 1), while the exploration-mixed estimator and the proposed DR-mix estimator converge for all arms, including arms 2 and 3 that are not observable in round $n > 50$. While the exploration-mixed estimator converges as fast as the DR-mix estimator on arm 2 and 3, it converges slower on arm 1; since the DR-mix estimator minimizes on all d context basis while exploration-mixed estimator minimized only one context basis. For further analysis on the estimators, see Appendix A.3.

6.2 Comparison of `MultiPFI` and `PFIwR`

Next, we compare `PFIwR` with `MultiPFI` (Auer et al., 2016) on the SW-LLVM dataset (Zuluaga et al., 2016) (see Appendix A.1 for details). Since our experiments focus on sample complexity and regret minimization specifically for identifying *all* Pareto optimal arms, only `MultiPFI` is directly applicable to our experiments. While the algorithm in Zuluaga et al. (2016) solves the



(a) Number of samples for PFI (b) Total sum of regret at termination (c) Cumulative regret ($\epsilon = 0.06$)

Figure 2: Comparison of `PFIwR` and `MultiPFI` on the SW-LLVM dataset. Both algorithms correctly identify the ϵ-near Pareto optimal arms on all 500 independent experiments.

same problem, it does not guarantee identification of all Pareto optimal arms; see Figure 4 in Auer et al. (2016).

Figure 2 reports the performance of `PFIwR` and `MultiPFI` (Auer et al., 2016) on various $\epsilon = \{0.06, 0.08, \ldots, 0.18\}$. Both algorithms use a fixed δ of 0.1. In Figure 2a, in most cases, `PFIwR` uses fewer samples than `MultiPFI` to satisfy the success condition (1). Even though the number of samples used by `PFIwR` has a larger variance, in most cases, it uses fewer samples for PFI than `MultiPFI`. Figure 2b is a box plot of the cumulative Pareto regret of `PFIwR` and `MultiPFI` at the termination of the algorithm – `PFIwR` has significantly lower regret than `MultiPFI`. Figure 2c display the cumulative Pareto regret of `PFIwR` and `MultiPFI` as a function of rounds n when $\epsilon = 0.06$. Since the number of rounds required for PFI (time horizon) is random, we set the instantaneous regret to zero after the algorithm terminates in each experiment. The regret of `PFIwR` increases slower than `MultiPFI` because it chooses actions that minimize regret in the exploitation phase while learning the rewards.

7 Conclusion

In this paper, we provide a novel algorithm that reuses explored samples to efficiently achieve PFI and regret minimization. We believe this reusing analysis has a wide application reaching beyond the dataset in our experiments and PFI problem. Our future work applies the novel algorithm to another real dataset for PFI problem and other multi-objective optimization problems with bandit feedback.

Acknowledgements

We thank the anonymous referees for offering many helpful comments and valuable feedback. Wonyoung Kim was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00240142) and Garud Iyengar was supported by NSF EFMA-2132142, ARPA-E PERFORM Program, and ONR N000142312374. This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by Korea government (MSIT) (No. RS-2021-II211341, Artificial Intelligence Graduate School Program at Chung-Ang University) and the Chung-Ang University Research Grants in 2025.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

P. Auer, C-K. Chiang, R. Ortner, and M. Drugan. Pareto front identification from stochastic bandit feedback. In *Artificial intelligence and statistics*, pages 939–947. PMLR, 2016.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1): 48–77, 2002b.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Ioannis Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17 (8):1505–1508, 2013.

Rémy Degenne, Thomas Nedelec, Clément Calauzènes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1988–1996. PMLR, 2019.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *COLT*, volume 2, pages 255–270. Springer, 2002.

Tushar Goel, Rajkumar Vaidyanathan, Raphael T Haftka, Wei Shyy, Nestor V Queipo, and Kevin Tucker. Response surface approximation of pareto optimal front in multi-objective optimization. *Computer methods in applied mechanics and engineering*, 196(4-6):879–893, 2007.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Uniform-pac bounds for reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:14188–14199, 2021.

Wonyoung Kim, Gi-Soo Kim, and Myunghee Cho Paik. Doubly robust thompson sampling with linear payoffs. In *Advances in Neural Information Processing Systems*, 2021.

Wonyoung Kim, Kyungbok Lee, and Myunghee Cho Paik. Double doubly robust thompson sampling for generalized linear contextual bandits. *arXiv preprint arXiv:2209.06983*, 2022.

Wonyoung Kim, Garud Iyengar, and Assaf Zeevi. Improved algorithms for multi-period multi-class packing problems with bandit feedback. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 16458–16501. PMLR, 23–29 Jul 2023a.

Wonyoung Kim, Myunghee Cho Paik, and Min-Hwan Oh. Squeeze all: Novel estimator and self-normalized bound for linear contextual bandits. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 3098–3124. PMLR, 25–27 Apr 2023b.

Wonyoung Kim, Garud Iyengar, and Assaf Zeevi. A doubly robust approach to sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2305–2313. PMLR, 2024.

Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proceedings of the 13th ACM Conference on recommender systems*, pages 20–28, 2019.

Daniel J Lizotte, Michael H Bowling, and Susan A Murphy. Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis. In *ICML*, volume 10, pages 695–702, 2010.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3080–3086, 2019.

Diederik M Roijers, Luisa M Zintgraf, and Ann Nowé. Interactive thompson sampling for multi-objective multi-armed bandits. In *Algorithmic Decision Theory: 5th International Conference, ADT 2017, Luxembourg, Luxembourg, October 25–27, 2017, Proceedings 5*, pages 18–34. Springer, 2017.

Diederik M Roijers, Luisa M Zintgraf, Pieter Libin, and Ann Nowé. Interactive multi-objective reinforcement learning in multi-armed bandits for any utility function. In *ALA workshop at FAIM*, volume 8, 2018.

Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85, 1918.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Chao Tao, Saúl Blanco, and Yuan Zhou. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pages 4877–4886. PMLR, 2018.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, 2013.

Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

Nirandika Wanigasekara, Yuxuan Liang, Siong Thye Goh, Ye Liu, Joseph Jay Williams, and David S Rosenblum. Learning multi-objective rewards and user utility function in contextual bandits for personalized ranking. In *IJCAI*, pages 3835–3841, 2019.

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

Bo Xue, Ji Cheng, Fei Liu, Yimu Wang, and Qingfu Zhang. Multiobjective lipschitz bandits under lexicographic ordering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16238–16246, 2024.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

Qiuyi Zhang. Optimal scalarizations for sublinear hypervolume regret. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Zixin Zhong, Wang Chi Cheung, and Vincent Tan. Achieving the pareto frontier of regret minimization and best arm identification in multi-armed bandits. *Transactions on Machine Learning Research*, 2023.

Marcela Zuluaga, Andreas Krause, and Markus Püschel. ε-pal: an active learning approach to the multi-objective optimization problem. *The Journal of Machine Learning Research*, 17(1):3619–3650, 2016.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Learning the Pareto Front Using Bootstrapped Observation Samples: Supplementary Materials

A Supplementary Materials for Experiments

A.1 SW-LLVM Dataset Description

The SW-LLVM dataset (Zuluaga et al., 2016) consists of 1024 2-dimensional reward vectors. We normalized the reward vectors by subtracting the average and dividing by the standard deviation for each component. We created a 16-arm PFI problem using the methodology in Auer et al. (2016): we clustered the reward vectors into 16 groups, with 64 reward vectors in each group. We computed the mean reward $y_i \in \mathbb{R}^2$ for the i-th cluster by taking the average over the i-th cluster, and when the algorithm selects an arm i in any round, we randomly sample a reward vector from the i-th cluster.

A.2 Consistency of the Doubly-Robust Estimator Without the Exploration-Mixed Estimator

The convergence properties of the DR estimator $\widehat{\theta}_t^{\langle \ell \rangle}$ critically depend on the imputation estimator $\check{\theta}_t^{\langle \ell \rangle}$ used in the pseudo-reward (10). In Figure 3 we plot the error of the DR estimators with two different imputation estimators: ridge estimator and exploration-mixed estimator (8) as a function of the number of rounds for a 3-armed bandit problem, or equivalently, a linear bandit problem with the set of context vectors $\mathcal{X}$ given by the Euclidean basis. Using the DR method with the ridge estimator does not guarantee convergence on all arms – it only gets information from the exploited arm that has no information about the rewards of the other arms. In contrast, the DR estimator $\widehat{\theta}_t^{\langle l \rangle}$ with (8) as an imputation estimator, converges on all arms. This is possible because "mixing" contexts and rewards as in (7) transforms the 3-armed bandit data into a linear bandit with stochastic contexts that span $\mathbb{R}^3$ with high probability.

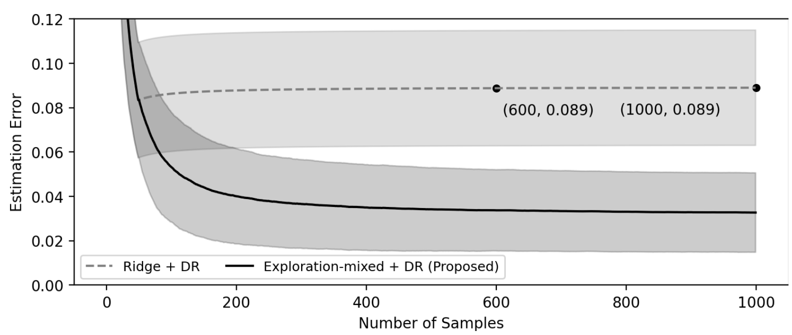


Figure 3: The ℓ_2-error of the reward on the *unexploited* arms (arms 2 and 3) of the proposed estimator and the DR estimator whose imputation estimator is the conventional ridge estimator in the 3-armed bandit problem (for detailed setting, see Section 6.1.) The estimators use samples from all three arms when $t \leq 50$ and only arm 1 when $t > 50$. When constructing a DR estimator, choosing the imputation estimator that learns rewards on all arms is crucial for convergence on all arms.

A.3 Comparison of the Density of the Estimators

The plots in Figure 4 display the evolution of the density of estimates of the three methods for arm 1 and arm 2 for $n = 50, 500, 2000$. For arm 1 (Figure 4a), the ridge estimator and the proposed DR-mix estimator converge faster, i.e., have zero-mean with a lower variance, compared to the exploration-mixed estimator. Since the exploration-mixed estimator (8) creates the context and the associated reward by assigning random weights to the current observation and one from a past exploration round, the reward estimate for the selected arm becomes unstable. In contrast, the DR-mix estimator returns the focus to estimating the reward of the selected arm and converges faster than the exploration-mixed estimator. For arm 2 (Figure 4b), while the ridge estimator diverges with increasing variance, the exploration-mixed estimator and the DR-mix estimator converge. Since there are no new samples from arm 2, the term $\sqrt{n}$ increases the mean and the variance of the density of the ridge estimator. In contrast, the mean of the density of the exploration-mixed estimator and the proposed DR-mix estimator converges to 0, and the variance increases slower than that of the ridge estimator.

The fast convergence of the DR-mix estimator is a consequence of combining the exploration-mixed data and DR technique. The exploration-mixed estimator (8) leverages the linear structure of the mean reward vector to create a pair of "mixed" contexts and rewards by combining the context of the selected arm (arm 1) with randomly selected arms (arms 2 and 3) from the exploration phase. While the "mixing" allows the exploration-mixed estimator to learn all $d = 3$ entries of the parameter vector, it minimizes $\mathbb{E}[|\check{X}_{a_t,t}^\top(\tilde{\theta}_t - \theta_\star)|^2]$ instead of the basis vectors for target contexts $\mathcal{X}$ of interest. Although the exploration-mixed estimator eventually converges to the true parameter, $\theta_\star$, the target of interest $\sum_{k=1}^{3} |x_k^\top(\tilde{\theta}_t - \theta_\star)|^2 = \|\tilde{\theta}_t - \theta_\star\|_2$ converges slower than $\mathbb{E}[|\check{X}_{a_t,t}^\top(\tilde{\theta}_t - \theta_\star)|^2]$. Therefore, we apply the DR method and use pseudo-rewards (10) to move the target to the one of interest by modifying the context from $\check{X}_{a_t,t}$ to $\mathcal{X}$ (equivalently, changing Gram matrix to $\sum_{k=1}^{K} x_k x_k^\top$). Thus, our proposed estimator minimizes the target $\sum_{k=1}^{3} |x_k^\top(\widehat{\theta}_t - \theta_\star)|^2$ directly and estimates the mean rewards of the arms significantly faster.

A.4 Complexity of the Proposed Estimators

The updates to both exploration-mixed estimator and the doubly-robust estimator are rank-1, and therefore, they can be efficiently updated using Sherman-Morrison formula. The overall complexity for the exploration-mixed estimator is $O(d^2 + dL)$ and the doubly-robust estimator is $O(d^3 + d^2L)$ per round. This complexity mainly comes from updating the estimator – the computational cost associated with generating the mixed context and rewards $(\check{X}_t, \check{Y}_t)$ and pseudo rewards $\{(\widehat{Y}_{i,t}^{(\ell)}, \tilde{x}_{i,t}) : i = 1, \ldots, d+1\}$ is minimal. The additional d factor in the complexity of the doubly-robust estimator comes from using $d + 1$ contexts basis vectors instead of only the selected context vector per round.

Specifically, the exploration-mixed estimator consists of three major steps: (i) Sample a pseudo-action $\check{a}_t$ according the exploration strategy; (ii) Identify the exploration round $\check{n}_t$ for mixing; (iii) Computing the mixed context $\check{X}_t$ and $\check{Y}_t$ and update. Step (i) is already set in the initial round and does not need additional computation and step (ii) requires $O(\gamma_t) = O(d^3 \log t)$ memory to save the number of reuse for the samples from the exploration rounds. Then, in step (iii), computing mixed contexts and estimator is $O(d + L)$ complexity and the exploration-mixed estimator can be updated with Sherman-Morrison formula or rank-1 update, which requires $O(d^3 + dL)$ complexity per (exploitation) round.

B Missing Proofs

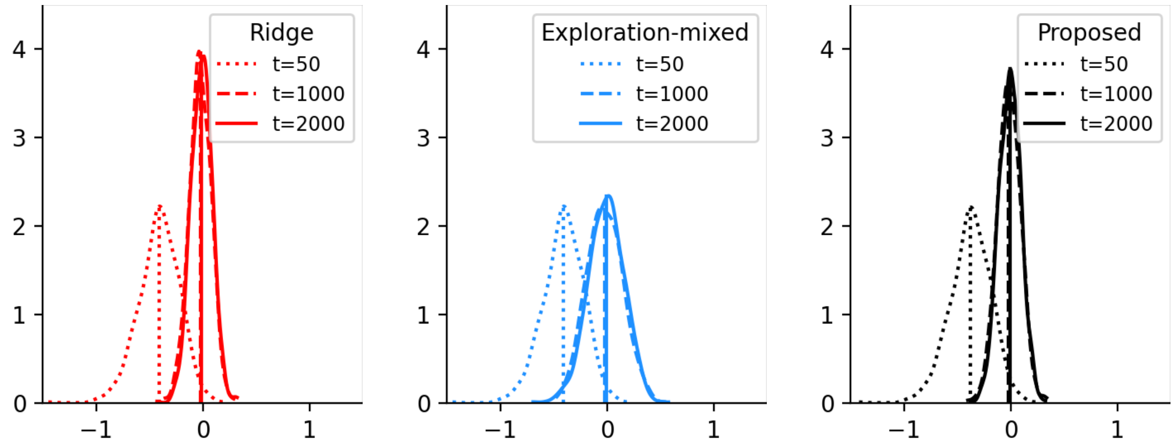
B.1 Proof of Theorem 3.3

Before we prove Theorem 3.3, we present a lower bound for the error of linearly parameterized rewards.

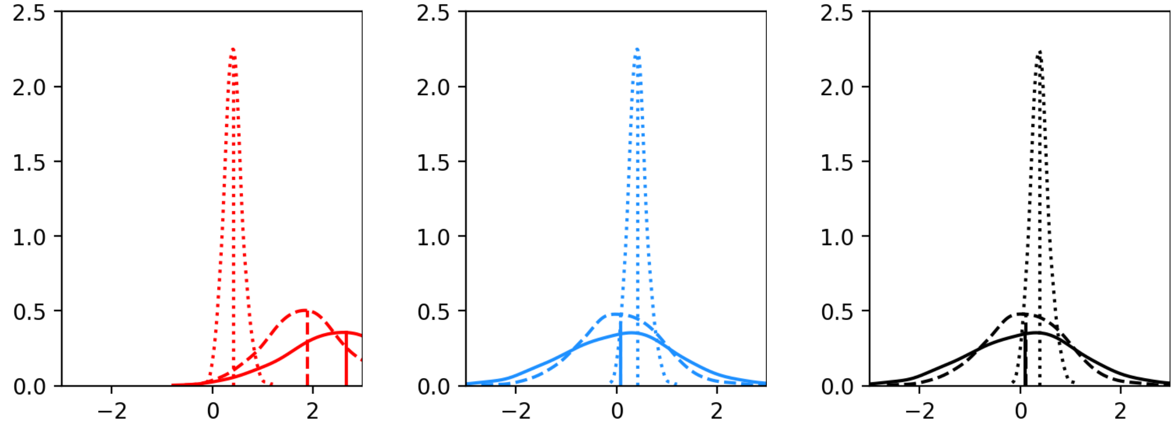
Lemma B.1. *Suppose $\mathcal{X} := \{x_k \in \mathbb{R}^d : k \in [K]\}$ spans $\mathbb{R}^d$. For $k \in [K]$ and $i_k \geq 1$, let $Y^{(i_k)} = \theta_\star^\top x_k + \eta_k^{(i_k)}$, where $\eta_k^{(i_k)}$ is an identically and independently distributed noise. Then, for any $\sigma > 0$ and $\epsilon \in (0, \sigma/4)$. Then there exist mean-zero σ-sub-Gaussian random noises such that the error*

$$\max_{x \in \mathcal{X}} \left| x^\top \left(\widehat{\theta}_n - \theta_\star\right) \right| > \epsilon,$$

with probability at least $\delta \in (0, 1/4)$, for any estimator $\widehat{\theta}_n$ that uses at most $n_1 + \cdots + n_K \leq \frac{d\sigma^2}{3\epsilon^2} \log \frac{1}{4\delta}$ number of



(a) On the *exploited* arm (arm 1)



(b) On the *unexploited* arm (arm 2)

Figure 4: Changes in densities of $\sqrt{n}(\widehat{\theta} - \theta_\star)$ over the number of samples $n = 50, 500, 2000$ on the *exploited* arm (arm 1) and the *unexploited* arm (arm 2). The vertical line represents the average computed from 1000 independent experiments. The proposed DR-mix estimator converges faster with lower variance than the ridge and exploration-mixed estimator on all arms.

independent samples $\cup_{k=1}^{K}\{x^{(i_k)}, Y^{(i_k)}\}_{i_k \in [n_k]}$.

Proof. **Step 1. Constructing a noise distribution:** For $k \in [K]$ the noise

$$\eta_k^{(i_k)} = \begin{cases} -\frac{\sigma}{2\left(1+\frac{2\epsilon}{\sigma}v_k\right)} & \text{w.p. } p_+(v_k) := \frac{1}{2} + \frac{\epsilon}{\sigma}v_k \\ \frac{\sigma}{2\left(1-\frac{2\epsilon}{\sigma}v_k\right)} & \text{w.p. } p_-(v_k) := \frac{1}{2} - \frac{\epsilon}{\sigma}v_k \end{cases},$$

where $v_k \in \{-1, 1\}$. Let $\mathbb{P}_{v_k}$ denote the probability measure for the noise η_k. Since

$$\left|\frac{1}{2\left(1+\frac{2\epsilon}{\sigma}\right)}\right| \leq \left|\frac{1}{2\left(1-\frac{2\epsilon}{\sigma}\right)}\right| \leq 1,$$

for all $\epsilon \in (0, \frac{\sigma}{4})$, we have $|\eta_k^{(i_k)}| \leq \sigma$. It is easy to show that the difference

$$\left|\frac{\sigma}{2\left(1-\frac{2\epsilon}{\sigma}v_k\right)} + \frac{\sigma}{2\left(1+\frac{2\epsilon}{\sigma}v_k\right)}\right| > 2\epsilon,$$

for all $v_k \in \{-1, 1\}$.

Step 2. Reduction to d parameter estimation: For $v \in \{-1,1\}^d$ let $\mathbb{P}_v^{(n)} = \prod_{k=1}^K \prod_{i_k=1}^{n_k} \mathbb{P}_{v_k}^{(i_k)}$ denote joint distribution for $\boldsymbol{\eta}^{(n)} = \{\eta_k^{i_k} : i_k \in [n_k], k \in [K]\}$. Let $\mathcal{S} = \{\tilde{x}_1, \ldots, \tilde{x}_d\} \subset \mathcal{X}$ denote a set of linear independent vectors, and let $y_j := \tilde{x}_j^\top \theta_\star$, $j \in \mathcal{S}$. Let $Y_k = y_k + \eta_k$ denote a sample for arm $k \in [K]$, independent of n initial samples, with $\eta_k \sim \mathbb{P}_{v_k}$. Let $\mathbb{P}_v := \prod_{k=1}^K \mathbb{P}_{v_k} \times \mathbb{P}_v^{(n)}$.

Let $\widehat{\theta}_n : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ denote any estimator that estimates $\theta_\star$ using the $n = \sum_{k=1}^K n_k$ data points. Clearly,

$$\mathbb{P}_v \left(\max_{k\in[K]} \left| x_k^\top \left(\widehat{\theta}_n(x_k, Y_k) - \theta_\star \right) \right| > \epsilon \right) \geq \mathbb{P}_v \left(\max_{j\in[d]} \left| \tilde{x}_j^\top \widehat{\theta}_n(\tilde{x}_j, Y_j) - y_j \right| > \epsilon \right).$$

Next,

$$\begin{aligned}
& \mathbb{P}_v \left(\max_{j\in[d]} \left| \tilde{x}_j^\top \widehat{\theta}_n(\tilde{x}_j, Y_j) - y_j \right| \leq \epsilon \right) && (17)\\
\leq \;& \mathbb{P}_v \left(\bigcap_{j=1}^d \left\{ \left| \tilde{x}_j^\top \widehat{\theta}_n(\tilde{x}_j, Y_j) - y_j \right| \leq \epsilon \right\} \right) \\
= \;& \mathbb{P}_v \left(\bigcap_{j=1}^d \left\{ \left| \tilde{x}_j^\top \widehat{\theta}_n(\tilde{x}_j, Y_j) - Y_j - \eta_j \right| \leq \epsilon \right\} \right) \\
= \;& \mathbb{E}_v^{(n)} \left[\prod_{j=1}^d \mathbb{P}_{v_j} \left(\left| \tilde{x}_j^\top \widehat{\theta}_n(\tilde{x}_j, Y_j) - Y_j - \eta_j \right| \leq \epsilon \right) \middle| \boldsymbol{\eta}^{(n)} \right], && (18)
\end{aligned}$$

where the last inequality holds because $\eta_1, \ldots, \eta_d$ are independent. For each $j \in [d]$,

$$\begin{aligned}
& \mathbb{P}_{v_j} \left(\left| \tilde{x}_j^\top \widehat{\theta}_n - Y_j - \eta_j \right| > \epsilon \middle| \boldsymbol{\eta}^{(n)} \right) \\
= \;& \mathbb{P}_{v_j} \left(\left| \tilde{x}_j^\top \widehat{\theta}_n - Y_j + \frac{\sigma}{4p_+(v_j)} \right| > \epsilon \middle| \eta_j = -\frac{\sigma}{4p_+(v_j)}, \boldsymbol{\eta}^{(n)} \right) p_+(v_j) \\
& + \mathbb{P}_{v_j} \left(\left| \tilde{x}_j^\top \widehat{\theta}_n - Y_j - \frac{\sigma}{4p_-(v_j)} \right| > \epsilon \middle| \eta_j = \frac{\sigma}{4p_-(v_j)}, \boldsymbol{\eta}^{(n)} \right) p_-(v_j), \\
\geq \;& \mathbb{P}_{v_j} \left(\left| \tilde{x}_j^\top \widehat{\theta}_n - Y_j + \frac{\sigma}{4p_+(v_j)} \right| > \epsilon \middle| \eta_j = -\frac{\sigma}{4p_+(v_j)}, \boldsymbol{\eta}^{(n)} \right) \min\left\{ p_+(v_j), p_-(v_j) \right\} \\
& + \mathbb{P}_{v_j} \left(\left| \tilde{x}_j^\top \widehat{\theta}_n - Y_j - \frac{\sigma}{4p_-(v_j)} \right| > \epsilon \middle| \eta_j = \frac{\sigma}{4p_-(v_j)}, \boldsymbol{\eta}^{(n)} \right) \min\left\{ p_+(v_j), p_-(v_j) \right\} \\
\geq \;& \min\left\{ p_+(v_j), p_-(v_j) \right\}, && (19)
\end{aligned}$$

where the last inequality holds because for $v_j \in \{-1,1\}$,

$$\left\{ u \in \mathbb{R} : \left| u - Y_j + \frac{\sigma}{4p_+(v_j)} \right| > \epsilon \right\} \cup \left\{ u \in \mathbb{R} : \left| u - Y_j - \frac{\sigma}{4p_-(v_j)} \right| > \epsilon \right\} = \mathbb{R}.$$

Define the function $\hat{f}(x,z) : \mathbb{R} \times \{-1,1\}$ as follows:

$$\hat{f}(x,z) := \begin{cases} x - \frac{\sigma}{4p_+(z)} & \text{if } p_+(z) > p_-(z) \\ x + \frac{\sigma}{4p_-(z)} & \text{o.w.} \end{cases}.$$

Then, the estimator $\hat{f}(Y_i, v_j)$ attains the minimum in (19) since

$$\begin{aligned}
\mathbb{P}_{v_j} \left(\left| \hat{f}(Y_j, v_j) - y_j \right| > \epsilon \right) \;=\;& \mathbb{P}_{v_j} \left(\hat{f}(Y_j, v_j) = Y_j + \frac{\sigma}{4p_-(v_j)} \middle| \eta_j = -\frac{\sigma}{4p_+(v_j)} \right) p_+(v_j) \\
& + \mathbb{P}_{v_j} \left(\hat{f}(Y_j, v_j) = Y_j - \frac{\sigma}{4p_+(v_k)} \middle| \eta_j = \frac{\sigma}{4p_-(v_j)} \right) p_-(v_j) \\
=\;& \mathbb{I}\left(p_+(v_j) \leq p_-(v_j) \right) p_+(v_j) + \mathbb{I}\left(p_+(v_j) > p_-(v_j) \right) p_-(v_j) \\
=\;& \min\left\{ p_+(v_j), p_-(v_j) \right\} \\
\leq\;& \mathbb{P}_{v_j} \left(\left| \tilde{x}_j^\top \widehat{\theta}_n - Y_j - \eta_j \right| > \epsilon \middle| \boldsymbol{\eta}^{(n)} \right).
\end{aligned}$$

Thus, for $j \in [d]$ and $v_j \in \{-1, 1\}$,

$$\mathbb{P}_{v_j}\left(\left|\tilde{x}_j^\top \widehat{\theta}_n - Y_j - \eta_j\right| \leq \epsilon \middle| \boldsymbol{\eta}_n\right) \leq \mathbb{P}_{v_j}\left(\left|\widehat{f}(Y_j, v_j) - y_j\right| \leq \epsilon\right).$$

Plugging this upper bound in (18), we get

$$\begin{aligned}
\mathbb{P}_v\left(\max_{j\in[d]}\left|\tilde{x}_j^\top \widehat{\theta}_n(\tilde{x}_j, Y_j) - y_j\right| \leq \epsilon\right) &\leq \mathbb{E}_v^{(n)}\left[\prod_{j=1}^d \mathbb{P}_{v_j}\left(\left|\hat{f}(Y_j, v_j) - y_j\right| \leq \epsilon\right)\right] \\
&= \mathbb{P}_v\left(\bigcap_{j=1}^d \left\{\left|\hat{f}(Y_j, v_j) - y_j\right| \leq \epsilon\right\}\right).
\end{aligned}$$

Let $n_{\min} = \min_{j\in[d]} n_j$ denote the minimum number of samples over arms in $[d]$. Let $\mathbb{P}_v^{(n_{\min})} = \prod_{i=1}^{n_{\min}} \prod_{j=1}^d \mathbb{P}_{v_j}^{(i)}$ denote the joint probability of the noise associated with $n_{\min}$ samples from each arm $\tilde{x} \in \mathcal{S}$. For $j \in [d]$ and any estimator $\widehat{v}_j$ for v_j that uses $n_{\min}$ samples from arm $j = 1, \ldots, d$,

$$\begin{aligned}
\mathbb{P}_v\left(\bigcap_{j=1}^d \left\{\left|\widehat{f}(Y_j, v_j) - y_j\right| \leq \epsilon\right\}\right) &= \mathbb{P}_v\left(\bigcap_{j=1}^d \left\{\left|\widehat{f}(Y_j, \widehat{v}_j) - y_j\right| \leq \epsilon\right\} \cap \{\widehat{v}_j = v_j\}\right) \\
&\leq \mathbb{P}_v\left(\bigcap_{j=1}^d \{\widehat{v}_j = v_j\}\right) \\
&= \mathbb{P}_v^{(n_{\min})}\left(\bigcap_{j=1}^d \{\widehat{v}_j = v_j\}\right),
\end{aligned}$$

where the last equality follows from the fact that $\widehat{v}_j$ only uses $n_{\min}$ samples from arm $j = 1, \ldots, d$. Therefore,

$$\begin{aligned}
\mathbb{P}_v\left(\max_{k\in[K]}\left|x_k^\top\left(\widehat{\theta}_n(x_k, Y_k) - \theta_\star\right)\right| > \epsilon\right) &\geq \mathbb{P}_v\left(\bigcap_{j=1}^d \left\{\left|\widehat{f}(Y_j, v_j) - y_j\right| > \epsilon\right\}\right) \\
&\geq \mathbb{P}_v^{(n_{\min})}\left(\bigcup_{j=1}^d \{\widehat{v}_j \neq v_j\}\right).
\end{aligned}$$

Step 3. Lower bound for the error probability. Taking maximum over v gives,

$$\begin{aligned}
\sup_{v\in\{-1,1\}^d} \mathbb{P}_v\left(\max_{k\in[K]}\left|x_k^\top\left(\widehat{\theta}_n(x_k, Y_k) - \theta_\star\right)\right| > \epsilon\right) &\geq \sup_{v\in\{-1,1\}^d} \mathbb{P}_v^{(n_{\min})}\left(\widehat{v} \neq v\right) \\
&= \sup_{v\in\{-1,1\}^d} \mathbb{E}_v^{(n_{\min})}\left[\max_{i\in[d]} \mathbb{I}\left(\widehat{v}_i \neq v_i\right)\right] \\
&\geq \sup_{v\in\{-1,1\}^d} \frac{1}{d}\mathbb{E}_v^{(n_{\min})}\left[\sum_{i=1}^d \mathbb{I}\left(\widehat{v}_i \neq v_i\right)\right].
\end{aligned}$$

For two vectors u and w let $u \sim w$ denote that u and w only differ in one coordinate. By Assouad's method (Lemma C.1), there exists at least one $\tilde{v} \in \{-1, 1\}^d$ such that

$$\begin{aligned}
\frac{1}{d}\mathbb{E}_{\tilde{v}}^{(n_{\min})}\left[\sum_{i=1}^d \mathbb{I}\left(\widehat{v}_i \neq \tilde{v}_i\right)\right] &\geq \frac{1}{2}\min_{u,w:u\sim w}\left\|\min\left\{\mathbb{P}_u^{(n_{\min})}, \mathbb{P}_w^{(n_{\min})}\right\}\right\|_1 \\
&= \frac{1}{2}\min_{u,w:u\sim w}\sum_{\boldsymbol{\eta}^{(n_{\min})}}\left|\min\left\{\mathbb{P}_u(\boldsymbol{\eta}^{(n_{\min})}), \mathbb{P}_w(\boldsymbol{\eta}^{(n_{\min})})\right\}\right|
\end{aligned}$$

Thus,

$$\begin{aligned}
\sup_{v\in\{-1,1\}^d} \mathbb{P}_v\left(\max_{k\in[K]}\left|x_k^\top\left(\widehat{\theta}_n(x_k,Y_k)-\theta_\star\right)\right|>\epsilon\right) \geq & \frac{1}{2}\min_{u,w:u\sim w}\left\|\min\left\{\mathbb{P}_u^{(n_{\min})},\mathbb{P}_w^{(n_{\min})}\right\}\right\|_1 \\
=&\frac{1}{2}\min_{u,w:u\sim w}\left\{1-\frac{1}{2}\left\|\mathbb{P}_u^{(n_{\min})}-\mathbb{P}_w^{(n_{\min})}\right\|_1\right\} \\
=&\frac{1}{2}\min_{u,w:u\sim w}\left\{1-TV\left(\mathbb{P}_u^{(n_{\min})},\mathbb{P}_w^{(n_{\min})}\right)\right\},
\end{aligned}$$

where $TV(\mathbb{P}_1,\mathbb{P}_2):=\sup|\mathbb{P}_1(\cdot)-\mathbb{P}_2(\cdot)|$ is the total variation distance between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$. By Bretagnolle-Huber inequality,

$$\begin{aligned}
1-TV\left(\mathbb{P}_u^{(n_{\min})},\mathbb{P}_w^{(n_{\min})}\right) \geq & \frac{1}{2}\exp\left(-KL\left(\mathbb{P}_u^{(n_{\min})},\mathbb{P}_w^{(n_{\min})}\right)\right) \\
=&\frac{1}{2}\exp\left(-n_{\min}KL\left(\mathbb{P}_u,\mathbb{P}_w\right)\right).
\end{aligned}$$

where the last equality uses the chain rule of entropy. Because there exists only one $j\in[d]$ such that $u_j\neq w_j$,

$$\begin{aligned}
KL\left(\mathbb{P}_u,\mathbb{P}_w\right) &= \left(\frac{1}{2}+\frac{\epsilon}{\sigma}\right)\log\frac{\frac{1}{2}+\frac{\epsilon}{\sigma}}{\frac{1}{2}-\frac{\epsilon}{\sigma}}+\left(\frac{1}{2}-\frac{\epsilon}{\sigma}\right)\log\frac{\frac{1}{2}-\frac{\epsilon}{\sigma}}{\frac{1}{2}+\frac{\epsilon}{\sigma}} \\
&= \frac{2\epsilon}{\sigma}\log\frac{1+\frac{2\epsilon}{\sigma}}{1-\frac{2\epsilon}{\sigma}} \\
&\leq \frac{2\epsilon}{\sigma}\left(\frac{2\epsilon}{\sigma}-\log\left(1-\frac{2\epsilon}{\sigma}\right)\right) \\
&\leq \frac{2\epsilon}{\sigma}\left(\frac{2\epsilon}{\sigma}+\frac{\frac{2\epsilon}{\sigma}}{1-\frac{2\epsilon}{\sigma}}\right) \\
&\leq \frac{2\epsilon}{\sigma}\left(\frac{2\epsilon}{\sigma}+\frac{4\epsilon}{\sigma}\right) \\
&= \frac{12\epsilon^2}{\sigma^2},
\end{aligned}$$

where the third inequality holds by $\epsilon\in(0,\sigma/4)$. Thus,

$$\begin{aligned}
\sup_{v\in\{-1,1\}^d} \mathbb{P}_v\left(\max_{k\in[K]}\left|x_k^\top\left(\widehat{\theta}_n(x_k,Y_k)-\theta_\star\right)\right|>\epsilon\right) &\geq \frac{1}{2}\exp\left(-n_{\min}KL\left(\mathbb{P}_u,\mathbb{P}_w\right)\right) \\
&\geq \frac{1}{2}\exp\left(-\frac{12n_{\min}\epsilon^2}{\sigma^2}\right).
\end{aligned}$$

Setting $n_{\min}\leq\frac{\sigma^2}{12\epsilon^2}\log\frac{1}{4\delta}$ gives

$$\sup_{v\in\{-1,1\}^d} \mathbb{P}_v\left(\max_{k\in[K]}\left|x_k^\top\left(\widehat{\theta}_n(x_k,Y_k)-\theta_\star\right)\right|>\epsilon\right)\geq\delta.$$

Step 4. Computing the required number of samples: Recall that $n_{\min}:=\min_{j\in[d]}n_j$ is the minimum number of samples over any d linearly independent contexts. Thus, if $n\leq\frac{d\sigma^2}{12\epsilon^2}\log\frac{1}{4\delta}$ then $n_{\min}\leq\frac{\sigma^2}{12\epsilon^2}\log\frac{1}{4\delta}$ and the lower bound of the probability holds. ☐

Now we are ready to prove the lower sample complexity bound for PFILIn.

Proof of Theorem 3.3. **Step 1. Characterize the failure event:** In order to meet the success condition (1) with $\delta\in(0,1/4)$, the algorithm must produce an estimate $\widehat{y}_k\in\mathbb{R}^L$ for $y_k=\Theta_\star x_k$ such that

$$\mathbb{P}\left(\bigcap_{k=1}^{K}\left\{\left\|\widehat{y}_k-y_k\right\|_\infty\leq\frac{1}{2}\Delta_{(k),\epsilon}\right\}\right)\geq 1-\delta.$$

Note that

$$\bigcap_{k=1}^{K}\left\{\|\widehat{y}_k - y_k\|_\infty \le \frac{1}{2}\Delta_{(k),\epsilon}\right\} \subseteq \bigcap_{k=1}^{d}\left\{\|\widehat{y}_k - y_k\|_\infty \le \frac{1}{2}\Delta_{(k),\epsilon}\right\}$$
$$= \bigcap_{k=1}^{d}\bigcap_{l=1}^{L}\left\{\left|\widehat{y}_k^{(l)} - y_k^{(l)}\right| \le \frac{1}{2}\Delta_{(k),\epsilon}\right\}$$

For $\ell \in [L]$ and $k \in [d]$, let

$$\mathcal{B}_k^{\langle l\rangle} := \left\{|\widehat{y}_k^{\langle \ell\rangle} - y_k^{\langle l\rangle}| > \frac{1}{2}\Delta_{(k)}\right\}.$$

Thus, if $\mathbb{P}(\bigcup_{l=1}^{L}\bigcup_{k=1}^{d}\mathcal{B}_k^{\langle \ell\rangle}) > \delta$, then $\mathbb{P}\left(\bigcap_{k=1}^{K}\{\|\widehat{y}_k - y_k\|_\infty \le \frac{1}{2}\Delta_{(k),\epsilon}\}\right) < 1-\delta$, i.e. the algorithm cannot satisfy the success condition (1).

Step 2. Compute the required number of samples: For each arm $k \in [d]$, suppose the number of observations t_k for the parameters $\{y_k^{\langle l\rangle} : \ell \in [L]\}$ satisfies the upper bound $t_k \le \frac{\sigma^2}{3\Delta_{(k),\epsilon}^2}\log\frac{1}{4p}$. Then by Lemma B.1, for any estimator $\widehat{y}_k^{\langle \ell\rangle}$

$$\left|\widehat{y}_k^{\langle \ell\rangle} - y_k^{\langle l\rangle}\right| > \frac{\Delta_{(k),\epsilon}}{2}, \tag{20}$$

holds with probability at least p, and $\mathbb{P}(\mathcal{B}_k^{\langle \ell\rangle}) \ge p$. Since, for each $k \in [d]$, the estimators $\widehat{y}_k^{\langle 1\rangle}, \dots, \widehat{y}_k^{\langle L\rangle}$ are independent of each other, the events $\mathcal{B}_k^{\langle 1\rangle}, \dots, \mathcal{B}_k^{\langle L\rangle}$ are independent, and therefore,

$$\mathbb{P}\left(\bigcap_{k=1}^{d}\bigcap_{l=1}^{L}(\mathcal{B}_k^{\langle \ell\rangle})^c\right) = \prod_{l=1}^{L}\mathbb{P}\left(\bigcap_{k=1}^{d}(\mathcal{B}_k^{\langle \ell\rangle})^c\right) = \prod_{l=1}^{L}\left\{1 - \mathbb{P}\left(\bigcup_{k=1}^{d}\mathcal{B}_k^{\langle l\rangle}\right)\right\}.$$

Therefore, if the total number of observations $t \le \sum_{k=1}^{d}\frac{\sigma^2}{3\Delta_{(k),\epsilon}^2}\log\frac{1}{4p}$, there exists an arm $k \in [d]$ such that the number of independent samples is less than $\frac{\sigma^2}{3\Delta_{(k),\epsilon}^2}\log\frac{1}{4p}$ and,

$$\mathbb{P}\left(\bigcap_{k=1}^{d}\bigcap_{l=1}^{L}(\mathcal{B}_k^{\langle \ell\rangle})^c\right) \le (1-p)^L \le \frac{1}{1+Lp}.$$

Because $\delta \in (0, 1/4)$, setting $p = \frac{4\delta}{3L} \ge \frac{\delta}{L(1-\delta)}$ gives $\mathbb{P}\left(\cap_{k=1}^{d}\cap_{l=1}^{L}(\mathcal{B}_k^{\langle \ell\rangle})^c\right) \le 1-\delta$. Thus, any algorithm requires at least

$$\sum_{k=1}^{d}\frac{\sigma^2}{3\Delta_k^2}\log\frac{3L}{4\delta}$$

number of rounds to meet the success condition (1). □

B.2 Proof of Lemma 4.1

By definition (7), for all $\ell \in [L]$,

$$\check{Y}_{a_t,t}^{\langle \ell\rangle} - \check{X}_{a_t,t}^\top\theta_\star^{\langle \ell\rangle} = w_t\left(Y_{a_t,t}^{\langle \ell\rangle} - x_{a_t}^\top\theta_\star^{\langle \ell\rangle}\right) + \check{w}_t\left(Y_{a_{\check{n}_t},\check{n}_t}^{\langle \ell\rangle} - x_{a_{\check{n}_t}}^\top\theta_\star^{\langle \ell\rangle}\right).$$

Taking conditional expectations on both sides,

$$\begin{aligned}\mathbb{E}\left[\check{Y}_{a_t,t}^{\langle \ell\rangle} - \check{X}_{a_t,t}^\top\theta_\star^{\langle \ell\rangle}\middle|\mathcal{F}_t\right] =& \mathbb{E}\left[w_t|\mathcal{F}_t\right]\left(Y_{a_t,t}^{\langle \ell\rangle} - x_{a_t}^\top\theta_\star^{\langle \ell\rangle}\right) + \mathbb{E}\left[\check{w}_t\left(Y_{a_{\check{n}_t},\check{n}_t}^{\langle \ell\rangle} - x_{a_{\check{n}_t}}^\top\theta_\star^{\langle \ell\rangle}\right)\middle|\mathcal{F}_t\right]\\ =&\mathbb{E}\left[\check{w}_t\left(Y_{a_{\check{n}_t},\check{n}_t}^{\langle \ell\rangle} - x_{a_{\check{n}_t}}^\top\theta_\star^{\langle \ell\rangle}\right)\middle|\mathcal{F}_t\right]\\ =&\mathbb{E}\left[\mathbb{E}\left[\check{w}_t|\mathcal{F}_t,\check{n}_t\right]\left(Y_{a_{\check{n}_t},\check{n}_t}^{\langle l\rangle} - x_{a_{\check{n}_t}}^\top\theta_\star^{\langle \ell\rangle}\right)\middle|\mathcal{F}_t\right]\\ =&0,\end{aligned}$$

which proves the first identity. For the expected Gram matrix, by definition (8),

$$\begin{aligned}\mathbb{E}\left[\check{X}_{a_t,t}\check{X}_{a_t,t}^\top\middle|\,\mathcal{F}_t,\check{n}_t\right] =&\mathbb{E}\left[w_t^2 x_{a_t}x_{a_t}^\top + w_t\check{w}_t\left(x_{a_t}x_{a_{\check{n}_t}}^\top + x_{a_{\check{n}_t}}x_{a_t}^\top\right) + \check{w}_t^2 x_{a_{\check{n}_t}}x_{a_{\check{n}_t}}^\top\middle|\,\mathcal{F}_t,\check{n}_t\right]\\ =&\mathbb{E}\left[w_t^2\middle|\,\mathcal{F}_t,\check{n}_t\right]x_{a_t}x_{a_t}^\top + \mathbb{E}\left[w_t\check{w}_t\middle|\,\mathcal{F}_t,\check{n}_t\right]\left(x_{a_t}x_{a_{\check{n}_t}}^\top + x_{a_{\check{n}_t}}x_{a_t}^\top\right)\\ &+\mathbb{E}\left[\check{w}_t^2\middle|\,\mathcal{F}_t,\check{n}_t\right]x_{a_{\check{n}_t}}x_{a_{\check{n}_t}}^\top\\ =&x_{a_t}x_{a_t}^\top + x_{a_{\check{n}_t}}x_{a_{\check{n}_t}}^\top.\end{aligned}$$

Taking conditional expectations on both sides,

$$\begin{aligned}\mathbb{E}\left[\check{X}_{a_t,t}\check{X}_{a_t,t}^\top\middle|\,\mathcal{F}_t\right] =&\mathbb{E}\left[\mathbb{E}\left[\check{X}_{a_t,t}\check{X}_{a_t,t}^\top\middle|\,\mathcal{F}_t,\check{n}_t\right]\mathcal{F}_t\right]\\ =&\mathbb{E}\left[x_{a_t}x_{a_t}^\top + x_{a_{\check{n}_t}}x_{a_{\check{n}_t}}^\top\middle|\,\mathcal{F}_t\right].\end{aligned}$$

By definition of n_t, we obtain $a_{n_t} = \check{a}_t$. Because $\check{a}_t \sim \pi^\star$,

$$\begin{aligned}\mathbb{E}\left[x_{a_t}x_{a_t}^\top + x_{a_{\check{n}_t}}x_{a_{\check{n}_t}}^\top\middle|\,\mathcal{F}_t\right] =&x_{a_t}x_{a_t}^\top + \mathbb{E}\left[x_{a_{\check{n}_t}}x_{a_{\check{n}_t}}^\top\middle|\,\mathcal{F}_t\right]\\ =&x_{a_t}x_{a_t}^\top + \sum_{k=1}^{K}\pi_k^\star x_k x_k^\top\\ \succeq&\sum_{k=1}^{K}\pi_k^\star x_k x_k^\top\end{aligned}$$

which completes the proof.

B.3 Design Matrix with Context Basis

We provide a theoretical result on the design matrix $\sum_{i=1}^{d}\lambda_i u_i u_i^\top = \sum_{k=1}^{K} x_k x_k^\top$ constructed by the exploration strategy in Section 4.1.

Lemma B.2 (A bound for the normalized norm of contexts.)**.** *For any $k\in[K]$ and $t\geq 1$, the normalized norm $x_k^\top(\sum_{k'=1}^{K}x_{k'}x_{k'}^\top)^{-1}x_k \leq 1$.*

Proof. For each $k\in[K]$, by Sherman-Morrison formula, for any $\epsilon>0$

$$\begin{aligned}x_k^\top(\sum_{k'=1}^{K}x_{k'}x_{k'}^\top + \epsilon I_d)^{-1}x_k \leq& x_k^\top(x_k x_k^\top + \epsilon I_d)^{-1}x_k\\ =&\frac{x_k^\top x_k}{\epsilon} - \frac{\epsilon^{-2}(x_k^\top x_k)^2}{1+\epsilon^{-1}x_k^\top x_k}\\ =&\frac{x_k^\top x_k}{\epsilon + x_k^\top x_k}\\ \leq&1.\end{aligned}$$

Letting $\epsilon \downarrow 0$ completes the proof. □

Lemma B.2 implies,

$$\max_{k\in[K]} x_k^\top(d^{-1}\sum_{k'=1}^{K}x_{k'}x_{k'}^\top)^{-1}x_k \leq d,$$

which has same bound with G-optimal design (Smith, 1918). Although we reduce K contexts into d context basis vectors, our estimation strategy enjoys the property of the optimal design for all K context vectors.

For the Gram matrix of the DR-mix estimator $F_t := \sum_{s=1}^{t}\sum_{k=1}^{K}x_k x_k^\top + I_d$, Lemma B.2 implies,

$$\|x_k\|_{F_t^{-1}} \leq 1/\sqrt{t},$$

for all $k \in [K]$. The DR-mix estimator imputes the reward on the basis contexts and minimizes $\|\cdot\|_{F_t}$-norm error, which efficiently estimates the rewards on all K arms. In contrast, the exploration-mixed estimator minimizes the $\|\cdot\|_{\sum_{s=1}^{t} \check{X}_{a_s,s}\check{X}_{a_s,s}^\top}$-norm error. Although the expected Gram matrix in the exploration-mixed estimator has $\sum_{k=1}^{K} x_k x_k^\top$ it is discounted by the factor of d because it employs only one of d context basis, not all d context basis, in each round. Therefore the DR-mix estimator converges faster than the exploration-mixed estimator on the rewards of all K arms.

B.4 Coupling with Resampling

We provide the details on the coupling with resampling in the following lemma. The key idea is coupling the event of interest with IID samples and bound the probability with another IID sample.

Lemma B.3. *Let π_t and $\tilde{\pi}_t$ denote the distribution for action on $[K]$ and pseudo-action on $[d+1]$, respectively. Let $\{a_t^{(m)} \sim \pi_t : m \geq 1\}$ and $\{\tilde{a}_t^{(m)} \sim \pi_t : m \geq 1\}$ denote IID samples from the distribution $\pi_t$and $\tilde{\pi}_t$ and for the number of resampling $\rho_t \in \mathbb{N}$, define new contexts*

$$\tilde{x}_{i,t} := \begin{cases} \sqrt{\lambda_i} u_i & \forall i \in [d] \\ x_{a_t} & i = d+1 \end{cases},$$

and the stopping time,

$$m_t := \min \left[\inf \left\{ m \geq 1 : x_{a_t^{(m)}} = \tilde{x}_{\tilde{a}_t^{(m)},t}, Y^{\langle \ell \rangle}_{a_t^{(m)},t} = \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(m)},t}, \forall \ell \in [L] \right\}, \rho_t \right].$$

Then, for any function $f : \mathbb{R}^2 \to \mathbb{R}$ and a real number $x \in \mathbb{R}$,

$$\mathbb{P}\left(f(x_{a_t^{(m_t)}}, Y^{\langle \ell \rangle}_{a_t^{(m_t)},t}) > x \right) \leq \mathbb{P}\left(f(\tilde{x}_{\tilde{a}_t^{(1)},t}, \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(1)},t}) > x \right) + \mathbb{P}\left(m_t > \rho_t\right),$$

$$\mathbb{P}\left(f(\tilde{x}_{\tilde{a}_t^{(m_t)},t}, \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(m_t)},t}) > x \right) \leq \mathbb{P}\left(f(\tilde{x}_{\tilde{a}_t^{(1)},t}, \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(1)},t}) > x \right) + \mathbb{P}\left(m_t > \rho_t\right),$$

for all $\ell \in [L]$.

Proof. For $m \geq 1$, let

$$\mathcal{M}^{(m)} := \left\{ x_{a_t^{(m)}} = \tilde{x}_{\tilde{a}_t^{(m)},t} \right\} \cap \left\{ Y^{\langle \ell \rangle}_{a_t^{(m)},t} = \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(m)},t} \right\}$$

By definition of m_t, the

$$\begin{aligned} \mathbb{P}\left(f(x_{a_t^{(m_t)}}, Y^{\langle \ell \rangle}_{a_t^{(m_t)},t}) > x \right) &\leq \mathbb{P}\left(\left\{ f(x_{a_t^{(m_t)}}, Y^{\langle \ell \rangle}_{a_t^{(m_t)},t}) > x \right\} \cap \mathcal{M}^{(m_t)} \right) + \mathbb{P}\left(\left\{ \mathcal{M}^{(m_t)} \right\}^c \right) \\ &= \mathbb{P}\left(\left\{ f(x_{a_t^{(m_t)}}, Y^{\langle \ell \rangle}_{a_t^{(m_t)},t}) > x \right\} \cap \mathcal{M}^{(m_t)} \right) + \mathbb{P}\left(m_t > \rho_t\right). \end{aligned}$$

On the event $\mathcal{M}^{(m_t)}$,

$$\mathbb{P}\left(\left\{ f(x_{a_t^{(m_t)}}, Y^{\langle \ell \rangle}_{a_t^{(m_t)},t}) > x \right\} \cap \mathcal{M}^{(m_t)} \right) = \mathbb{P}\left(\left\{ f(\tilde{x}_{\tilde{a}_t^{(m_t)},t}, \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(m_t)},t}) > x \right\} \cap \mathcal{M}^{(m_t)} \right).$$

Because the event $\{f(\tilde{x}_{\tilde{a}_t^{(m)},t}, \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(m)},t}) > x\} \cap \mathcal{M}^{(m)}$ is IID over $m \geq 1$ given $\mathcal{F}_t$.

$$\begin{aligned} &\mathbb{P}\left(\left\{ f(\tilde{x}_{\tilde{a}_t^{(m_t)},t}, \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(m_t)},t}) > x \right\} \cap \mathcal{M}^{(m_t)} \right). \\ &= \mathbb{E}\left[\mathbb{P}\left(\left\{ f(\tilde{x}_{\tilde{a}_t^{(m_t)},t}, \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(m_t)},t}) > x \right\} \cap \mathcal{M}^{(m_t)} \middle| \mathcal{F}_t \right) \right] \\ &= \mathbb{E}\left[\mathbb{P}\left(\left\{ f(\tilde{x}_{\tilde{a}_t^{(1)},t}, \tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(1)},t}) > x \right\} \cap \mathcal{M}^{(1)} \middle| \mathcal{F}_t \right) \right] \\ &\leq \mathbb{P}\left(f(\tilde{Y}^{\langle \ell \rangle}_{\tilde{a}_t^{(1)},t}) > x \right). \end{aligned}$$

The second inequality in the lemma can be derived in a similar way. $\square$

B.5 Concerns in Reusing Samples in Exploration-Mixed Estimator

The reuse of the samples from the exploration rounds causes dependency among samples that complicates the analysis. Let $\breve{\eta}_t^{\langle \ell \rangle} := \breve{Y}_{a_t,t} - \breve{X}_{a_t,t}^\top \theta_\star^{\langle \ell \rangle} = w_t \eta_t^{\langle \ell \rangle} + \breve{w}_t \eta_{\breve{n}_t}^{\langle \ell \rangle}$. Then the problematic term of the analysis is:

$$\sum_{s\in[t]\setminus\mathcal{E}_t} \breve{\eta}_s^{\langle \ell \rangle} \breve{X}_{a_s,s} = \sum_{s\in[t]\setminus\mathcal{E}_t} w_s^2 \eta_s^{\langle \ell \rangle} x_{a_s} + w_s \breve{w}_s (\eta_s^{\langle \ell \rangle} x_{a_{\breve{n}_s}} + \eta_{\breve{n}_s}^{\langle \ell \rangle} x_{a_s}) + \breve{w}_s^2 \eta_{\breve{n}_s}^{\langle \ell \rangle} x_{a_{\breve{n}_s}}.$$

In the last term, the $\eta_{\breve{n}_s}^{\langle \ell \rangle} x_{a_{\breve{n}_s}}$ are repeated at most $t/|\mathcal{E}_t|$ times which causes $\tilde{O}(\sqrt{t/|\mathcal{E}_t|})$ error term as in the following lemma.

Lemma B.4 (A self-normalizing bound for the exploration-mixed estimator.)**.** *Fix $\delta \in (0,1)$. Then, for all $t \in \mathbb{N}$ that satisfy $t \geq 4 \cdot 288 d^2 \log \frac{2dt^2}{\delta}$, the exploration-mixed estimator defined in (8) satisfies*

$$\left\| \breve{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right\|_{F_t} \leq 2 \left\| \theta_\star^{\langle \ell \rangle} \right\|_{F_t^{-1}} + 12\sigma \sqrt{3d \log\left(\frac{3Lt}{\delta}\right)} + 24\sigma \sqrt{\frac{3td}{|\mathcal{E}_t|} \log\left(\frac{3Lt}{\delta}\right)},$$

with probability at least $1-4\delta$,

Proof. Let us fix t throughout the proof. For $s \in [t]$ and $\ell \in [L]$ let $\tilde{\eta}_s^{\langle \ell \rangle} := \breve{Y}_{a_s,s}^{\langle \ell \rangle} - \breve{X}_{a_s,s}^\top \theta_\star^{\langle \ell \rangle}$. By definition of $\breve{\theta}_t^{\langle \ell \rangle}$,

$$\begin{aligned}
&\breve{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \\
&= \left(\sum_{s\in\mathcal{E}_t} x_{a_s} x_{a_s}^\top + \sum_{s\notin\mathcal{E}_t} \breve{X}_{a_s,s} \breve{X}_{a_s,s}^\top + \frac{1}{2} I_d \right)^{-1} \left(\sum_{s\in\mathcal{E}_t} x_{a_s} \eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \breve{X}_{a_s,s} \tilde{\eta}_s^{\langle \ell \rangle} - \frac{1}{2} \theta_\star^{\langle \ell \rangle} \right) \\
&= \left(\sum_{s\in\mathcal{E}_t} 2x_{a_s} x_{a_s}^\top + \sum_{s\notin\mathcal{E}_t} 2\breve{X}_{a_s,s} \breve{X}_{a_s,s}^\top + I_d \right)^{-1} \left(\sum_{s\in\mathcal{E}_t} 2x_{a_s} \eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} 2\breve{X}_{a_s,s} \tilde{\eta}_s^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right) \\
&:= \tilde{A}_t^{-1} \left(\sum_{s\in\mathcal{E}_t} 2x_{a_s} \eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} 2\breve{X}_{a_s,s} \tilde{\eta}_s^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right)
\end{aligned}$$

Define the new contexts,

$$\tilde{x}_{i,s} := \begin{cases} \sqrt{\lambda_i} u_i & \forall i \in [d] \\ x_{a_s} & i = d+1 \end{cases}, \quad \breve{X}_{i,s} := \begin{cases} \breve{w}_s \tilde{x}_{i,s} + \breve{w}_s \sqrt{\lambda_{i_s}} u_{i_s} & \forall i \in [d] \\ \breve{w}_s x_{a_s} + \breve{w}_s \sqrt{\lambda_{i_s}} u_{i_s} & i = d+1 \end{cases}.$$

Setting the number of resampling $\rho_s = \log(s^2/\delta)/\log(2)$, the probability of obtaining matching samples $\tilde{x}_{\tilde{a}_s}^{(m_s)} = x_{a_s}^{(m_s)}$ for $m_s \leq \rho_s$ is at least $1-\delta$ for all $s \in [t]$, where m_s is the number of trials until the matching. With the matching pseudo action $\tilde{a}_s^{(m_s)}$,

$$\tilde{A}_t = \sum_{s\in\mathcal{E}_t} \frac{1}{\tilde{\pi}_s(\tilde{a}_s^{(m_s)})} \tilde{x}_{\tilde{a}_s^{(m_s)},s} \tilde{x}_{\tilde{a}_s^{(m_s)},s}^\top + \sum_{s\notin\mathcal{E}_t} \frac{1}{\tilde{\pi}_s(\tilde{a}_s^{(m_s)})} \breve{X}_{\tilde{a}_s^{(m_s)},s} \breve{X}_{\tilde{a}_s^{(m_s)},s}^\top + I_d.$$

By the coupling lemma (LemmaB.3) and , with probability at least $1-\delta$,

$$\begin{aligned}
\tilde{A}_t &= \sum_{s\in\mathcal{E}_t} \frac{1}{\tilde{\pi}_s(\tilde{a}_s)} \tilde{x}_{\tilde{a}_s,s} \tilde{x}_{\tilde{a}_s,s}^\top + \sum_{s\notin\mathcal{E}_t} \frac{1}{\tilde{\pi}_s(\tilde{a}_s)} \breve{X}_{\tilde{a}_s,s} \breve{X}_{\tilde{a}_s,s}^\top + I_d \\
&= \sum_{s\in\mathcal{E}_t} \sum_{i=1}^{d+1} \frac{\mathbb{I}(\tilde{a}_s = i)}{\tilde{\pi}_s(i)} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top + \sum_{s\notin\mathcal{E}_t} \frac{\mathbb{I}(\tilde{a}_s = i)}{\tilde{\pi}_s(i)} \breve{X}_{i,s} \breve{X}_{i,s}^\top + I_d \\
&:= \sum_{s\in\mathcal{E}_t} \sum_{i=1}^{d+1} U_{i,s} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top + \sum_{s\notin\mathcal{E}_t} U_{i,s} \breve{X}_{i,s} \breve{X}_{i,s}^\top + I_d.
\end{aligned}$$

Then the self-normalized norm,

$$\left\|\check{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}\right\|_{F_t} = \left\| \sum_{s \in \mathcal{E}_t} 2 x_{a_s} \eta_s^{(\ell)} + \sum_{s \notin \mathcal{E}_t} 2 \check{X}_{a_s,s} \tilde{\eta}_s^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right\|_{\tilde{A}_t^{-1} F_t \tilde{A}_t^{-1}}.$$

To find the expectation of $\tilde{A}_t$, recall that for $s \in [t]$, the pseudo-action $\tilde{a}_s$ is sampled from $\tilde{\pi}_s$ defined in (9). Let $\mathbb{E}_s[\cdot] := \mathbb{E}[\cdot|\sigma(\cup_{u \notin \mathcal{E}_t, u<s}\{w_s, \check{w}_s, \check{n}_s\} \cup \{\tilde{a}_1, \ldots, \tilde{a}_{s-1}\})]$ denote the conditional expectation at round s. Then $\mathbb{E}_s[U_{i,s}] = 1$. Define

$$D_s := \begin{cases} F_t^{-1/2} \left[\sum_{i=1}^{d+1} (U_{i,s} - 1) \tilde{x}_{i,s} \tilde{x}_{i,s}^\top \right] F_t^{-1/2} & s \in \mathcal{E}_t \\ F_t^{-1/2} \left[\sum_{i=1}^{d+1} \left(U_{i,s} \check{X}_{i,s} \check{X}_{i,s}^\top - \frac{2d+1}{d} \sum_{i=0}^{d} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top - x_{a_s} x_{a_s}^\top \right) \right] F_t^{-1/2} & s \notin \mathcal{E}_t \end{cases}. \tag{21}$$

Note that D_s is martingale difference because for $s \notin \mathcal{E}_t$,

$$\begin{aligned}
&\mathbb{E}_s \left[F_t^{-1/2} \left(\sum_{i=0}^{d+1} U_{i,s} \check{X}_{i,s} \check{X}_{i,s}^\top \right) F_t^{-1/2} \right] \\
&= F_t^{-1/2} \mathbb{E}_s \left[\sum_{i=0}^{d+1} U_{i,s} \check{X}_{i,s} \check{X}_{i,s}^\top \right] F_t^{-1/2} \\
&= F_t^{-1/2} \mathbb{E}_s \left[\sum_{i=0}^{d+1} \check{X}_{i,s} \check{X}_{i,s}^\top \right] F_t^{-1/2} \\
&= F_t^{-1/2} \mathbb{E}_s \left[\sum_{i=0}^{d+1} \left(\check{w}_s \tilde{x}_{i,s} + \check{w}_s \sqrt{\lambda_{i_s}} u_{i_s} \right) \left(\check{w}_s \tilde{x}_{i,s} + \check{w}_s \sqrt{\lambda_{i_s}} u_{i_s} \right)^\top \right] F_t^{-1/2} \\
&= F_t^{-1/2} \mathbb{E}_s \left[\sum_{i=0}^{d+1} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top + (d+1) \lambda_{i_s} u_{i_s} u_{i_s}^\top \right] F_t^{-1/2} \\
&= F_t^{-1/2} \mathbb{E}_s \left[\sum_{i=0}^{d+1} \left(\tilde{x}_{i,s} \tilde{x}_{i,s}^\top + \frac{d+1}{d} \lambda_i u_i u_i^\top \right) \right] F_t^{-1/2} \\
&= F_t^{-1/2} \mathbb{E}_s \left[\frac{2d+1}{d} \sum_{i=0}^{d} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top + x_{a_s} x_{a_s}^\top \right] F_t^{-1/2}
\end{aligned}$$

where the second last equality holds by $w_t, \check{w}_t \sim \text{unif}[-\sqrt{3}, \sqrt{3}]$ and the last equality holds by $i_s \sim \text{unif}[d]$. For $s \in \mathcal{E}_t$,

$$\begin{aligned}
\|D_s\|_2 \leq& 2d \max_{i \in [d+1]} \left\| F_t^{-1/2} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top F_t^{-1/2} \right\|_2 \\
\leq& 2d \max_{i \in [d+1]} \|\tilde{x}_{i,s}\|_{F_t^{-1}}^2 \\
\leq& \frac{2d}{t},
\end{aligned}$$

where the last inequality can be easily found by the fact $F_t \succeq t \sum_{i=1}^{d} \tilde{x}_{i,s} \tilde{x}_{i,s} + I_d = t \sum_{k=1}^{K} x_k x_k^\top + I_d$ and following the proof of Lemma B.2. For $s \notin \mathcal{E}_t$,

$$\begin{aligned}
\|D_s\|_2 \leq& 2d \max_{i \in [d+1]} \left\| F_t^{-1/2} \check{X}_{i,s} \check{X}_{i,s}^\top F_t^{-1/2} \right\|_2 \\
\leq& 2 \left(w_t^2 + \check{w}_t^2 \right) d \max_{i \in [d+1]} \left\| F_t^{-1/2} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top F_t^{-1/2} \right\|_2 \\
\leq& \frac{12d}{t},
\end{aligned}$$

where the last inequality holds by $\check{w}_s, \check{w}_s \sim$unif$[-\sqrt{3}, \sqrt{3}]$. Thus, the eigenvalue of the martingale difference matrix lies in $[-12d/t, 12d/t]$. Then by Hoeffding bound for matrices (Lemma C.5),

$$\mathbb{P}\left(\left\|\sum_{s=1}^{t} D_s\right\|_2 > x\right) \leq 2d \exp\left(-\frac{tx^2}{288d^2}\right).$$

Note that

$$\begin{aligned}
F_t \preceq & t\sum_{k=1}^{K} x_k x_k^\top + \sum_{s=1}^{t} x_{a_s} x_{a_s}^\top + I_d \\
= & \sum_{s=1}^{t}\sum_{i=1}^{d+1} \tilde{x}_{i,s}\tilde{x}_{i,s}^\top + I_d \\
\preceq & \sum_{s\in\mathcal{E}_t}\sum_{i=1}^{d+1} \tilde{x}_{i,s}\tilde{x}_{i,s}^\top + \sum_{s\notin\mathcal{E}_t}\left(\frac{2d+1}{d}\sum_{i=0}^{d} \tilde{x}_{i,s}\tilde{x}_{i,s}^\top + x_{a_s}x_{a_s}^\top\right),
\end{aligned}$$

where the last term appears in the martingale difference (21). Thus,

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}\left(I_d - F_t^{-1/2}\tilde{A}_t F_t^{-1/2}\right) > x\right) &= \mathbb{P}\left(\lambda_{\max}\left(F_t^{-1/2}\left\{F_t - \tilde{A}_t\right\}F_t^{-1/2}\right) > x\right) \\
&\leq \mathbb{P}\left(\lambda_{\max}\left(\sum_{s=1}^{t} D_s\right) > x\right) \\
&\leq 2d\exp\left(-\frac{tx^2}{288d^2}\right).
\end{aligned}$$

Set $x = 1/2$. For $t \in \mathbb{N}$ such that $t \geq 4 \cdot 288 d^2 \log \frac{2dt^2}{\delta}$, with probability at least $1 - \delta/t^2$,

$$\lambda_{\max}\left(I_d - F_t^{-1/2}\tilde{A}_t F_t^{-1/2}\right) \leq \frac{1}{2},$$

which implies

$$I_d - F_t^{-1/2}\tilde{A}_t F_t^{-1/2} \preceq \frac{1}{2} I_d \Rightarrow \frac{1}{2} I_d \preceq F_t^{-1/2}\tilde{A}_t F_t^{-1/2}.$$

Because the matrix $F_t^{-1/2}\tilde{A}_t F_t^{-1/2}$ is symmetric and positive definite, $F_t^{1/2}\tilde{A}_t^{-1}F_t^{1/2} \preceq 2I_d$, and thus,

$$\tilde{A}_t^{-1} F_t \tilde{A}_t^{-1} = F_t^{-1/2}\left(F_t^{-1/2}\tilde{A}_t F_t^{-1/2}\right)^{-2} F_t^{-1/2} \preceq 4F_t^{-1}.$$

Then the self-normalized norm,

$$\begin{aligned}
\left\|\check{\theta}_t^{\langle\ell\rangle} - \theta_\star^{\langle\ell\rangle}\right\|_{F_t} &= \left\|\sum_{s\in\mathcal{E}_t} 2x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} 2\check{X}_{a_s,s}\tilde{\eta}_s^{\langle\ell\rangle} - \theta_\star^{\langle\ell\rangle}\right\|_{\tilde{A}_t^{-1}F_t\tilde{A}_t^{-1}} \\
&\leq \left\|\sum_{s\in\mathcal{E}_t} 2x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} 2\check{X}_{a_s,s}\tilde{\eta}_s^{\langle\ell\rangle} - \theta_\star^{\langle\ell\rangle}\right\|_{4F_t^{-1}} \\
&\leq 4\left\|\sum_{s\in\mathcal{E}_t} x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \check{X}_{a_s,s}\tilde{\eta}_s^{\langle\ell\rangle} - \theta_\star^{\langle\ell\rangle}\right\|_{F_t^{-1}} + 2\left\|\theta_\star^{\langle\ell\rangle}\right\|_{F_t^{-1}}.
\end{aligned} \tag{22}$$

In the first term,

$$\begin{aligned}
\check{X}_{a_s,s}\tilde{\eta}_s^{\langle\ell\rangle} &= \left(w_s x_{a_s} + \check{w}_s x_{a_{\check{n}_s}}\right)\left(w_s\eta_s^{(\ell)} + \check{w}_s\eta_{\check{n}_s}^{(\ell)}\right) \\
&= \eta_s^{(\ell)} u_s^2 x_{a_s} + \eta_{\check{n}_s}^{(\ell)}\check{w}_s^2 x_{a_{\check{n}_s}} + w_s\check{w}_s\left(\eta_s^{(\ell)} x_{a_{\check{n}_s}} + \eta_{\check{n}_t}^{(\ell)} x_{a_s}\right) \\
&= \eta_s^{(\ell)}\left(w_s^2 x_{a_s} + w_s\check{w}_s x_{a_{\check{n}_s}}\right) + \eta_{\check{n}_s}^{(l)}\left(w_s\check{w}_s x_{a_s} + \check{w}_s^2 x_{a_{\check{n}_s}}\right).
\end{aligned}$$

Note that for all exploitation round $s \in [t] \setminus \mathcal{E}_t$, the sampled round for reuse $\check{n}_s \in \mathcal{E}_t$. Thus, we can decompose,

$$
\begin{aligned}
&\sum_{s\in\mathcal{E}_t} x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \check{X}_{a_s,s}\tilde{\eta}_s^{\langle\ell\rangle} \\
&= \sum_{s\in\mathcal{E}_t} x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \eta_s^{(\ell)}\left(w_s^2 x_{a_s} + w_s\check{w}_s x_{a_{\check{n}_s}}\right) + \eta_{\check{n}_s}^{(l)}\left(w_s\check{w}_s x_{a_s} + \check{w}_s^2 x_{a_{\check{n}_s}}\right) \\
&= \sum_{s\in\mathcal{E}_t} x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \eta_s^{(\ell)}\left(w_s^2 x_{a_s} + w_s\check{w}_s x_{a_{\check{n}_s}}\right) \\
&+ \sum_{s\in\mathcal{E}_t}\left\{\left(\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau = s\right)\check{w}_\tau^2\right)x_{a_s} + \sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau = s\right)w_\tau\check{w}_\tau x_{a_\tau}\right\}\eta_s^{(\ell)},
\end{aligned}
$$

and thus,

$$
\begin{aligned}
&\left\|\sum_{s\in\mathcal{E}_t} x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \check{X}_{a_s,s}\tilde{\eta}_s^{\langle} - \theta_\star^{\langle\ell\rangle}\right\|_{F_t^{-1}} \\
&\leq \left\|\sum_{s\in\mathcal{E}_t} x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \eta_s^{(\ell)}\left(w_s^2 x_{a_s} + w_s\check{w}_s x_{a_{\check{n}_s}}\right)\right\|_{F_t^{-1}} \\
&+ \left\|\sum_{s\in\mathcal{E}_t}\left\{\left(\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau = s\right)\check{w}_\tau^2\right)x_{a_s} + \sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau = s\right)w_\tau\check{w}_\tau x_{a_\tau}\right\}\eta_s^{(\ell)}\right\|_{F_t^{-1}}
\end{aligned}
\tag{23}
$$

To bound the first term, for $\lambda \in \mathbb{R}^d$, define

$$
D_s^\lambda := \begin{cases} \frac{\lambda^\top F_t^{-1/2} x_{a_s}\eta_s^{(\ell)}}{\sigma} - \frac{\lambda^\top\lambda}{t} & s\in\mathcal{E}_t \\ \frac{\lambda^\top F_t^{-1/2}\left(w_s^2 x_{a_s} + w_s\check{w}_s x_{a_{\check{n}_s}}\right)\eta_s^{(\ell)}}{3\sigma} - \frac{2\lambda^\top\lambda}{t} & s\notin\mathcal{E}_t \end{cases}.
$$

Let $\mathbb{E}_s$ denote a conditional expectation given errors $\{\eta_\tau^{(\ell)} : \tau\in[s-1], \ell\in[L]\}$, actions $\{a_\tau : \tau\in[s]\}$, random indexes $\{\check{n}_\tau : \tau\in[s]\setminus\mathcal{E}_t\}$ and weights $\{w_\tau, \check{w}_\tau : \tau\in[s]\setminus\mathcal{E}_t\}$. For $s\in\mathcal{E}_t$, because $\eta_s^{(\ell)}$ is σ-sub-Gaussian and $\lambda_{\max}(F_t^{-1/2}x_k x_k^\top F_t^{-1/2}) = x_k^\top F_t^{-1} x_k \leq 1/t$ for all $x\in\mathcal{X}$,

$$
\begin{aligned}
\mathbb{E}_s\left[\exp\left(D_s^\lambda\right)\right] &\leq \mathbb{E}_s\left[\exp\left(\frac{\lambda^\top F_t^{-1/2}x_{a_s}\eta_s^{(\ell)}}{\sigma} - \frac{\lambda^\top F_t^{-1/2}x_{a_s}x_{a_s}^\top F_t^{-1/2}\lambda}{2}\right)\right] \\
&= \mathbb{E}_s\left[\exp\left\{\frac{\lambda^\top F_t^{-1/2}x_{a_s}\eta_s^{(\ell)}}{\sigma} - \frac{\left(\lambda^\top F_t^{-1/2}x_{a_s}\right)^2}{2}\right\}\right] \\
&\leq 1.
\end{aligned}
$$

For $s\notin\mathcal{E}_t$,

$$
\begin{aligned}
&\mathbb{E}_s\left[\exp\left(D_s^\lambda\right)\right] \\
&\leq \exp\left\{\frac{\sigma^2\left(w_s^2\lambda^\top F_t^{-1/2}x_{a_s} + w_s\check{w}_s\lambda^\top F_t^{-1/2}x_{a_{\check{n}_s}}\right)^2}{18\sigma^2} - \left(\lambda^\top F_t^{-1/2}x_{\tilde{a}_s}\right)^2 - \left(\lambda^\top F_t^{-1/2}x_{a_{\check{n}_s}}\right)^2\right\} \\
&\leq \exp\left[\frac{\sigma^2\left\{2w_s^4(\lambda^\top F_t^{-\frac{1}{2}}x_{a_s})^2 + 2w_s^2\check{w}_s^2(\lambda^\top F_t^{-\frac{1}{2}}x_{a_{\check{n}_s}})^2\right\}}{18\sigma^2} - \left(\lambda^\top F_t^{-\frac{1}{2}}x_{\tilde{a}_s}\right)^2 - \left(\lambda^\top F_t^{-\frac{1}{2}}x_{a_{\check{n}_s}}\right)^2\right] \\
&\leq \exp\left(\frac{18\sigma^2\left\{(\lambda^\top F_t^{-1/2}x_{a_s})^2 + (\lambda^\top F_t^{-1/2}x_{a_{\check{n}_s}})^2\right\}}{18\sigma^2} - \left(\lambda^\top F_t^{-1/2}x_{\tilde{a}_s}\right)^2 - \left(\lambda^\top F_t^{-1/2}x_{a_{\check{n}_s}}\right)^2\right) \\
&\leq 1,
\end{aligned}
$$

where the second inequality holds by $(a+b)^2 \le 2a^2 + 2b^2$ and the third inequality holds by $w_s, \breve{w}_s \in [-\sqrt{3}, \sqrt{3}]$ almost surely. Thus, for all $\lambda \in \mathbb{R}^d$,

$$\mathbb{E}\left[\exp\left(\sum_{s=1}^{t} D_s^{\lambda} - 2\lambda^\top \lambda\right)\right] \le 1.$$

Following the proof of Theorem 1 in Abbasi-Yadkori et al. (2011), with probability at least $1-\delta/L$,

$$\begin{aligned}
&\left\| \sum_{s\in\mathcal{E}_t} x_{a_s}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \eta_s^{(\ell)}\left(w_s^2 x_{a_s} + w_s\breve{w}_s x_{a_{\breve{n}_s}}\right)\right\|_{F_t^{-1}} \\
&\le \sqrt{3}\left\| \sum_{s\in\mathcal{E}_t} x_{a_s}F_t^{-1/2}\eta_s^{(\ell)} + \sum_{s\notin\mathcal{E}_t} \eta_s^{(\ell)}\left(w_s^2 F_t^{-1/2}x_{a_s} + w_s\breve{w}_s F_t^{-1/2} x_{a_{\breve{n}_s}}\right)\right\|_{(2I_d+I_d)^{-1}} \\
&\le 3\sigma\sqrt{3d\log\left(\frac{3Lt}{\delta}\right)}.
\end{aligned}$$

To bound the second term in (23), define

$$M_s^{\lambda} := \lambda^\top F_t^{-1/2}\left\{\left(\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\breve{n}_\tau = s\right)\breve{w}_\tau^2\right)x_{a_s} + \sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\breve{n}_\tau = s\right)w_\tau\breve{w}_\tau x_{\tilde{a}_\tau}\right\}\frac{\eta_s^{(\ell)}\sqrt{\gamma_t}}{6\sigma\sqrt{t}},$$

for $\lambda \in \mathbb{R}^d$. Let $\mathbb{E}_s$ denote a conditional expectation given errors $\{\eta_\tau^{(l)} : \tau \in [s-1]\}$, pseudo-actions $\{\tilde{a}_\tau : \tau \in [s]\}$, random indexes $\{\breve{n}_\tau : \tau \in [s]\setminus\mathcal{E}_t\}$. For $s \in \mathcal{E}_t$, for each $s \in \mathcal{E}_t$, by Hoeffding's Lemma, since $w_s \in [-\sqrt{3}, \sqrt{3}]$, for any $v \in \mathbb{R}$,

$$\mathbb{E}_s\left[\exp\left(w_s v\right)\right] \le \mathbb{E}_s\left[\frac{\sqrt{3}-w_s}{2\sqrt{3}}e^{-\sqrt{3}v} + \frac{w_s+\sqrt{3}}{2\sqrt{3}}e^{\sqrt{3}v}\right] = \cosh\left(\sqrt{3}|v|\right).$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\breve{n}_\tau = s\right)w_\tau\breve{w}_\tau\lambda^\top F_t^{-1/2}x_{a_\tau}\frac{\eta_s^{(\ell)}\sqrt{\gamma_t}}{6\sigma\sqrt{t}}\right\}\middle|\, \eta_s^{(\ell)}, \{\breve{n}_\tau, \breve{w}_\tau, a_\tau\}_{\tau\notin\mathcal{E}_t}\right] \\
&\le \prod_{\tau\notin\mathcal{E}_t}\cosh\left(\frac{\sqrt{3\gamma_t}\eta_s^{(\ell)}}{6\sigma\sqrt{t}}\mathbb{I}\left(\breve{n}_\tau = s\right)|\breve{w}_\tau|\left|\lambda^\top F_t^{-1/2}x_{a_\tau}\eta_s^{(\ell)}\right|\right) \\
&\le \prod_{\tau\notin\mathcal{E}_t}\cosh\left(\frac{\sqrt{\gamma_t}}{2\sigma\sqrt{t}}\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|\mathbb{I}\left(\breve{n}_\tau = s\right)\left|\eta_s^{(\ell)}\right|\right)
\end{aligned}$$

Because $\cosh(x)\cosh(y) \le \cosh(x+y)$ for $x, y > 0$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\breve{n}_\tau = s\right)w_\tau\breve{w}_\tau\lambda^\top F_t^{-1/2}x_{a_\tau}\frac{\eta_s^{(\ell)}\sqrt{\gamma_t}}{6\sigma\sqrt{t}}\right\}\middle|\, \eta_s^{(\ell)}, \{\breve{n}_\tau, \breve{w}_\tau, a_\tau\}_{\tau\notin\mathcal{E}_t}\right] \\
&\le \cosh\left(\frac{\sqrt{\gamma_t}}{2\sigma\sqrt{t}}\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_{a_s}\right|\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\breve{n}_\tau = s\right)\left|\eta_s^{(\ell)}\right|\right).
\end{aligned}$$

For $s \in \mathcal{E}_t$, let $N_s := \sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\breve{n}_\tau = s\right)$. Because $\cosh(\cdot)$ is an even function, $\cosh(x|y|) = \cosh(xy)$ for any $x, y \in \mathbb{R}$

and

$$
\begin{aligned}
&\mathbb{E}_s\left[\exp\left(M_s^{\lambda}\right)\right]\\
&\leq \mathbb{E}_s\left[\exp\left(\left(N_s\check{w}_\tau^2\right)\lambda^\top F_t^{-1/2}x_{a_s}\frac{\eta_s^{(\ell)}\sqrt{\gamma_t}}{6\sigma\sqrt{t}}\right)\cosh\left(\frac{\sqrt{\gamma_t}}{2\sigma\sqrt{t}}\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|N_s\eta_s^{(\ell)}\right)\right]\\
&=\frac{1}{2}\mathbb{E}_s\left[\exp\left(\left\{N_s\check{w}_\tau^2\frac{\lambda^\top F_t^{-1/2}x_{a_s}\sqrt{\gamma_t}}{6\sigma\sqrt{t}}+\frac{\sqrt{\gamma_t}}{2\sigma\sqrt{t}}\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|N_s\right\}\eta_s^{(\ell)}\right)\right]\\
&+\frac{1}{2}\mathbb{E}_s\left[\exp\left(\left\{N_s\check{w}_\tau^2\frac{\lambda^\top F_t^{-1/2}x_{a_s}\sqrt{\gamma_t}}{6\sigma\sqrt{t}}-\frac{\sqrt{\gamma_t}}{2\sigma\sqrt{t}}\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|N_s\right\}\eta_s^{(\ell)}\right)\right].
\end{aligned}
$$

Because $\eta_s^{(\ell)}$ is σ-sub-Gaussian and $(a-b)^2+(a+b)^2=2a^2+2b^2$ for $a,b\in\mathbb{R}$,

$$
\begin{aligned}
\mathbb{E}_s\left[\exp\left(M_s^{\lambda}\right)\right]&\leq\exp\left(\frac{\gamma_t}{8t}\left\{\left(N_s\check{w}_\tau^2\frac{\lambda^\top F_t^{-1/2}x_{\tilde{a}_s}}{3}\right)^2+\left(\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|N_s\right)^2\right\}\right)\\
&\leq\exp\left(\frac{\gamma_t}{4t}\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|^2N_s^2\right).
\end{aligned}
$$

By definition of $\check{n}_\tau$ the number of reusing round $s\in\mathcal{E}_t$,

$$
N_s:=\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau=s\right)\leq\frac{\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(a_{\check{n}_\tau}=a_s\right)}{\sum_{u\in\mathcal{E}_t}\mathbb{I}\left(a_u=a_s\right)}+1\leq\frac{\sum_{\tau=1}^{t}\mathbb{I}\left(a_{\check{n}_\tau}=a_s\right)}{\sum_{u\in\mathcal{E}_t}\mathbb{I}\left(a_u=a_s\right)}\leq\frac{t}{\gamma_t},
$$

where the last inequality holds by construction of the exploration set (5). Thus,

$$
\mathbb{E}_s\left[\exp\left(M_s^{\lambda}\right)\right]\leq\exp\left(\frac{1}{4}\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|^2N_s\right)
$$

By the fact that $t=\sum_{s\in\mathcal{E}_t}N_s$

$$
\begin{aligned}
\mathbb{E}\left[\exp\left(\sum_{s\in\mathcal{E}_t}M_s^{\lambda}-2\lambda^\top\lambda\right)\right]&\leq\mathbb{E}\left[\exp\left(\sum_{s\in\mathcal{E}_t}M_s^{\lambda}-t\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|^2\right)\right]\\
&=\mathbb{E}\left[\exp\left(\sum_{s\in\mathcal{E}_t}\left(M_s^{\lambda}-\max_{k\in[K]}\left|\lambda^\top F_t^{-1/2}x_k\right|^2N_s\right)\right)\right]\\
&\leq 1.
\end{aligned}
$$

Following the proof of Theorem 1 in Abbasi-Yadkori et al. (2011), with probability at least $1-\delta/L$,

$$
\begin{aligned}
&\left\|\sum_{s\in\mathcal{E}_t}\left\{\left(\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau=s\right)\check{w}_\tau^2\right)x_{a_s}+\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau=s\right)w_\tau\check{w}_\tau x_{a_\tau}\right\}\eta_s^{(\ell)}\right\|_{F_t^{-1}}\\
&=\sqrt{3}\left\|\sum_{s\in\mathcal{E}_t}\left\{\left(\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau=s\right)\check{w}_\tau^2\right)x_{a_s}+\sum_{\tau\notin\mathcal{E}_t}\mathbb{I}\left(\check{n}_\tau=s\right)w_\tau\check{w}_\tau x_{a_\tau}\right\}\eta_s^{(\ell)}\right\|_{(2I_d+I_d)^{-1}}\\
&\leq 6\sigma\sqrt{\frac{3td}{|\mathcal{E}_t|}\log\left(\frac{3Lt}{\delta}\right)}.
\end{aligned}
$$

In summary, with probability at least $1-4\delta$,

$$
\left\|\check{\theta}_t^{\langle\ell\rangle}-\theta_\star^{\langle\ell\rangle}\right\|_{F_t}\leq 2\left\|\theta_\star^{\langle\ell\rangle}\right\|_{F_t^{-1}}+12\sigma\sqrt{3d\log\left(\frac{3Lt}{\delta}\right)}+24\sigma\sqrt{\frac{3td}{|\mathcal{E}_t|}\log\left(\frac{3Lt}{\delta}\right)},
$$

for all $\ell\in[L]$. □

B.6 Robustness of the Doubly-Robust Estimation

The doubly robust estimation stems from missing data literature in statistics (Bang and Robins, 2005). The term 'doubly-robust' refers to the fact that the estimator for the missing value is unbiased if either (i) the probability of observation or (ii) the model for the missing value is correctly specified. In our PFILin problem, the probability of observation is the probability of choosing an arm, and therefore, is known to the algorithm. And the model for the missing rewards is unknown linear function of known contexts. It is straightforward to show that the pseudo-reward defined in (10) is doubly robust for the missing unselected rewards. We prove a lemma on the robustness of the general DR estimator.

Lemma B.5 (Robustness of DR estimator to the error of imputation estimator.)**.** *For $t \geq 1$ let $F_t := t\sum_{k=1}^{K} x_k x_k^\top + I_d$. For any $x \in \mathcal{X}$ and $\delta \in (0,1)$, the DR estimator $\widehat{\theta}_t^{\langle \ell \rangle}$ employing $\tilde{\theta}^{\langle \ell \rangle} \in \mathbb{R}^d$ as an imputation estimator satisfies*

$$\left| x^\top (\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}) \right| \leq \|x\|_{F_t^{-1}} \left(\theta_{\max} + 2\sigma \sqrt{d \log \frac{Lt}{\delta}} + 3d\sqrt{\frac{2}{t} \log \frac{2dt^2}{\delta}} \left\| \tilde{\theta}^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right\|_{F_t} \right)$$

with probability at least $1-3\delta$ for all $t \geq 1$ and $\ell \in [L]$. For each $x \in \mathcal{X}$,

$$\left| x^\top (\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}) \right| \leq \|x\|_{F_t^{-1}} \left(\theta_{\max} + 8\sigma \sqrt{2 \log \frac{4Lt^2}{\delta}} + 3d\sqrt{\frac{2}{t} \log \frac{2dt^2}{\delta}} \left\| \tilde{\theta}^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right\|_{F_t} \right),$$

with probability at least $1-3\delta$ for all $t \geq 1$ and $\ell \in [L]$.

The first term and second terms correspond to the convergence rate obtained from conventional self-normalizing bound. The third term is the $\|\cdot\|_{F_t}$-error of the estimator $\tilde{\theta}$. The $\|\cdot\|_{F_t}$-error of $\tilde{\theta}$ is multiplied with the $O(dt^{-1/2})$ term, which comes from the fact that in the pseudo-rewards,

$$\widehat{Y}_{i,t}^{\langle \ell \rangle} := \sqrt{\lambda_i} u_i^\top \tilde{\theta}^{\langle \ell \rangle} \left\{ 1 - \frac{\mathbb{I}(\tilde{a}_t = i)}{\tilde{\pi}_i^{(t)}} \right\} + \frac{\mathbb{I}(\tilde{a}_t = i)}{\tilde{\pi}_i^{(t)}} \tilde{Y}_{i,t}^{\langle \ell \rangle}, \ \forall i = 1, \ldots, d+1,$$

the reward estimate $\sqrt{\lambda_i} u_i^\top \tilde{\theta}$ is multiplied by the mean-zero random variable, $\left\{ 1 - \frac{\mathbb{I}(\tilde{a}_t = i)}{\tilde{\pi}_i^{(t)}} \right\}$.

The error of the imputation estimator is normalized by the Gram matrix F_t which consist of all K contexts. Thus, the $\|\cdot\|_{F_t}$ error of the conventional ridge estimator, which uses only selected contexts and rewards in every round, is $\Omega(\sqrt{t})$, which yields slow convergence rate.

Proof. Let us fix $t \geq 1$ throughout the proof. For $s \in [t]$ and $\ell \in [L]$, let $\widehat{\eta}_{i,s}^{\langle \ell \rangle} := \widehat{Y}_{i,s}^{\langle \ell \rangle} - \tilde{x}_{i,s} \theta_\star^{\langle \ell \rangle}$, where

$$\tilde{x}_{i,s} := \begin{cases} \sqrt{\lambda_i} u_i & \forall i \in [d] \\ x_{a_s} & i = d+1 \end{cases}.$$

Let $V_t := \sum_{s=1}^{t} \mathbb{I}(\mathcal{M}_s) \sum_{i=1}^{d+1} \tilde{x}_{i,s} \tilde{x}_{i,s}^\top + I_d$. By the definition of the estimator and the pseudo-reward $\widehat{Y}_{i,s}^{\langle \ell \rangle}$, for $x \in \mathcal{X}$

$$\begin{aligned} x^\top \left(\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right) &= - x^\top V_t^{-1} \theta_\star^{\langle \ell \rangle} + x^\top V_t^{-1} \left\{ \sum_{s=1}^{t} \sum_{i=1}^{d+1} \widehat{\eta}_{i,s}^{\langle \ell \rangle} \tilde{x}_{i,s} \right\} \\ &= - x^\top V_t^{-1} \theta_\star^{\langle \ell \rangle} + x^\top V_t^{-1} \left\{ \sum_{s=1}^{t} \frac{Y_{\tilde{a}_s,s}^{\langle \ell \rangle} - x_{\tilde{a}_s}^\top \theta_\star^{\langle \ell \rangle}}{\tilde{\pi}_{\tilde{a}_s}} \tilde{x}_{\tilde{a}_s,s} \right\} \\ &\quad + x^\top V_t^{-1} \left\{ \sum_{s=1}^{t} \sum_{i=1}^{d+1} \left(1 - \frac{\mathbb{I}(\tilde{a}_s = i)}{\tilde{\pi}_{\tilde{a}_s}} \right) \tilde{x}_{i,s} \tilde{x}_{i,s}^\top \right\} \left(\tilde{\theta}^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right). \end{aligned}$$

On the coupling event $\cap_{s=1}^{t}\mathcal{M}_s$, we have $Y_{\tilde{a}_s,s}^{\langle \ell \rangle} = Y_{a_s,s}^{\langle \ell \rangle}$, $\tilde{x}_{\tilde{a}_s,s} = x_{a_s}$ and $\tilde{\pi}_{\tilde{a}_s} = 1/2$. Thus, the first term,

$$\begin{aligned} x^\top V_t^{-1}\left\{\sum_{s=1}^{t}\frac{Y_{\tilde{a}_s,s}^{\langle \ell \rangle} - x_{\tilde{a}_s}^\top \theta_\star^{\langle \ell \rangle}}{\tilde{\pi}_{\tilde{a}_s}}\tilde{x}_{\tilde{a}_s,s}\right\} =& 2x^\top V_t^{-1}\left\{\sum_{s=1}^{t}\left(Y_{a_s,s}^{\langle \ell \rangle} - x_{a_s}^\top \theta_\star^{\langle \ell \rangle}\right)x_{a_s}\right\} \\ =& 2x^\top V_t^{-1}\left\{\sum_{s=1}^{t}\eta_s^{(\ell)}x_{a_s}\right\} \\ \leq& 2\left\|x\right\|_{V_t^{-1}}\left\|\sum_{s=1}^{t}\eta_s^{(\ell)}x_{a_s}\right\|_{V_t^{-1}}. \end{aligned}$$

By definition of F_t, on the coupling event $\cap_{s=1}^{t}\mathcal{M}_s$,

$$V_t = \sum_{s=1}^{t}\sum_{i=1}^{d+1}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top + I_d \succeq t\sum_{k=1}^{K}x_k x_k^\top + I_d := F_t.$$

Thus,

$$x^\top V_t^{-1}\left\{\sum_{s=1}^{t}\frac{Y_{\tilde{a}_s,s}^{\langle \ell \rangle} - x_{\tilde{a}_s}^\top \theta_\star^{\langle \ell \rangle}}{\tilde{\pi}_{\tilde{a}_s}}\tilde{x}_{\tilde{a}_s,s}\right\} \leq 2\left\|x\right\|_{F_t^{-1}}\left\|\sum_{s=1}^{t}\eta_s^{(\ell)}x_{a_s}\right\|_{F_t^{-1}}$$

For the second term, define $A_t := \sum_{s=1}^{t}\sum_{i=1}^{d+1}\frac{\mathbb{I}(\tilde{a}_s = i)}{\tilde{\pi}_{\tilde{a}_s}}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top + I_d$. Then,

$$\begin{aligned} &\left|x^\top\left(\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}\right)\right| \\ \leq& \left|x^\top V_t^{-1}\theta_\star^{\langle \ell \rangle}\right| + 2\left\|x\right\|_{V_t^{-1}}\left\|\sum_{s=1}^{t}\eta_s^{(\ell)}x_{a_s}\right\|_{V_t^{-1}} + \left|x^\top V_t^{-1}\left(V_t - A_t\right)\left(\tilde{\theta}^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}\right)\right| \\ \leq& \left\|x\right\|_{F_t^{-1}}\left(\left\|\theta_\star^{\langle \ell \rangle}\right\|_{F_t^{-1}} + 2\left\|\sum_{s=1}^{t}\eta_s^{(\ell)}x_{a_s}\right\|_{F_t^{-1}} + \left\|V_t^{-1/2}\left(V_t - A_t\right)\left(\tilde{\theta}^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}\right)\right\|_2\right) \\ \leq& \left\|x\right\|_{F_t^{-1}}\left(\left\|\theta_\star^{\langle \ell \rangle}\right\|_{F_t^{-1}} + 2\left\|\sum_{s=1}^{t}\eta_s^{(\ell)}x_{a_s}\right\|_{F_t^{-1}} + \left\|V_t^{-1/2}\left(V_t - A_t\right)F_t^{-1/2}\right\|_2\left\|\tilde{\theta}^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}\right\|_{F_t}\right). \end{aligned} \tag{24}$$

Because $V_t \succeq F_t$,

$$\left\|V_t^{-1/2}\left(V_t - A_t\right)F_t^{-1/2}\right\|_2 \leq \left\|F_t^{-1/2}\left(V_t - A_t\right)F_t^{-1/2}\right\|_2$$

In the last term of (24),

$$\left\|F_t^{-1/2}\left(V_t - A_t\right)F_t^{-1/2}\right\|_2 = \left\|\sum_{s=1}^{t}\sum_{i=1}^{d+1}\left\{1 - \frac{\mathbb{I}\left(\tilde{a}_s = i\right)}{\tilde{\pi}_{\tilde{a}_s}}\right\}F_t^{-1/2}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top F_t^{-1/2}\right\|_2$$

For each $s \in [t]$, the matrix

$$\sum_{i=1}^{d+1}\left\{1 - \frac{\mathbb{I}\left(\tilde{a}_s = i\right)}{\tilde{\pi}_{\tilde{a}_s}}\right\}F_t^{-1/2}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top F_t^{-1/2}$$

is symmetric and a martingale difference matrix. Moreover,

$$\begin{aligned} &\left\|\sum_{i=1}^{d+1}\left\{1 - \frac{\mathbb{I}\left(\tilde{a}_s = i\right)}{\tilde{\pi}_{\tilde{a}_s}}\right\}F_t^{-1/2}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top F_t^{-1/2}\right\|_2 \\ \leq& \sum_{i=1}^{d+1}\left|1 - \frac{\mathbb{I}\left(\tilde{a}_s = i\right)}{\tilde{\pi}_{\tilde{a}_s}}\right|\max_{i\in[d+1]}\left\|F_t^{-1/2}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top F_t^{-1/2}\right\|_2 \\ \leq& \left(d + 2d - 1\right)\max_{i\in[d+1]}\left\|F_t^{-1/2}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top F_t^{-1/2}\right\|_2 \\ =& 3d\max_{i\in[d+1]}\left\|F_t^{-1/2}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top F_t^{-1/2}\right\|_2. \end{aligned}$$

Because $\max_{i\in[d+1]} \tilde{x}_{i,s}^\top F_t^{-1}\tilde{x}_{i,s} \le 1/t$ for $s \in [t]$,

$$\left\| \sum_{i=1}^{d+1} \left\{ 1 - \frac{\mathbb{I}\left(\tilde{a}_s = i\right)}{\tilde{\pi}_{\tilde{a}_s}} \right\} F_t^{-1/2}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top F_t^{-1/2} \right\|_2 \le \frac{3d}{t},$$

almost surely. By the Hoeffding bound for the matrix (Lemma C.5), with probability at least $1-\delta/t^2$,

$$\left\| \sum_{s=1}^{t}\sum_{i=1}^{d+1} \left\{ 1 - \frac{\mathbb{I}\left(\tilde{a}_s = i\right)}{\tilde{\pi}_{\tilde{a}_s}} \right\} F_t^{-1/2}\tilde{x}_{i,s}\tilde{x}_{i,s}^\top F_t^{-1/2} \right\|_2 \le 3d\sqrt{\frac{2}{t}\log\frac{2dt^2}{\delta}}. \tag{25}$$

Plugging in (24),

$$\begin{aligned}
&\left| x^\top \left(\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right) \right| \\
&\le \|x\|_{F_t^{-1}} \left(\left\| \theta_\star^{\langle \ell \rangle} \right\|_{F_t^{-1}} + 2\left\| \sum_{s=1}^{t} \eta_s^{(\ell)} x_{a_s} \right\|_{F_t^{-1}} + 3d\sqrt{\frac{2}{t}\log\frac{2dt^2}{\delta}} \left\| \tilde{\theta}^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right\|_{F_t} \right) \\
&\le \|x\|_{F_t^{-1}} \left(\theta_{\max} + 2\left\| \sum_{s=1}^{t} \eta_s^{(\ell)} x_{a_s} \right\|_{F_t^{-1}} + 3d\sqrt{\frac{2}{t}\log\frac{2dt^2}{\delta}} \left\| \tilde{\theta}^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right\|_{F_t} \right).
\end{aligned} \tag{26}$$

Note that F_t is not random and $\eta_s^{(\ell)}$ is σ-sub-Gaussian. Thus, by Lemma 9 in Abbasi-Yadkori et al. (2011), with probability at least $1-\delta/L$,

$$\left\| \sum_{s=1}^{t} \eta_s^{(\ell)} x_{a_s} \right\|_{F_t^{-1}} \le \left\| \sum_{s=1}^{t} \eta_s^{(\ell)} x_{a_s} \right\|_{(\sum_{s=1}^t x_{a_s}x_{a_s}^\top + I_d)^{-1}} \le \sigma\sqrt{\log\frac{L\det\left(\sum_{s=1}^t x_{a_s}x_{a_s}^\top + I_d\right)}{\delta}},$$

for $t \ge 1$. Because $\|x_k\|_2 \le 1$ for all $k \in [K]$,

$$\begin{aligned}
\det\left(\sum_{s=1}^{t} x_{a_s}x_{a_s}^\top + I_d \right) &\le \left\{ \frac{\mathrm{Tr}\left(\sum_{s=1}^t x_{a_s}x_{a_s}^\top + I_d\right)}{d} \right\}^d \\
&\le \left\{ \frac{t+d}{d} \right\}^d \\
&\le t^d,
\end{aligned}$$

which implies,

$$\left\| \sum_{s=1}^{t} \eta_s^{(\ell)} x_{a_s} \right\|_{F_t^{-1}} \le \sigma\sqrt{d\log\frac{Lt}{\delta}},$$

and proves the first bound.

For the second bound, by Lemma **??**, with probability at least $1-\delta/t^2$,

$$\left\| \sum_{s=1}^{t} \eta_s^{(\ell)} x_{a_s} \right\|_{F_t^{-1}} \le \left\| \sum_{s=1}^{t} \eta_s^{(l)} F_t^{-1/2} x_{a_s} \right\|_2 \le 4\sigma\sqrt{2\left(\sum_{s=1}^{t} \left\| F_t^{-1/2}x_{a_s} \right\|_2^2 \right)\frac{4Lt^2}{\delta}}$$

By Lemma B.2,

$$\sum_{s=1}^{t} \left\| F_t^{-1/2}x_{a_s} \right\|_2^2 \le t\max_{k\in[K]} x_k^\top F_t^{-1}x_k \le 1$$

which implies

$$\left\| \sum_{s=1}^{t} \eta_s^{(\ell)} x_{a_s} \right\|_{F_t^{-1}} \le 4\sigma\sqrt{2\log\frac{4Lt^2}{\delta}}.$$

which proves the second bound. $\square$

B.7 Properties of the Exploration Set

In this section, we provide a lemma for a lower bound for the exploration rounds $\mathcal{E}_t$.

Lemma B.6. *Let $\mathcal{E}_t$ denote the set of exploration rounds defined in* (5)*. Then the cardinality of $|\mathcal{E}| \geq \gamma_t$ for all $t \geq T_\gamma$.*

Proof. The inequality can be established using the fact that (5) implies

$$\sum_{u \in \mathcal{E}_t} \mathbb{I}\left(a_u = \check{a}_t\right) \geq \frac{\gamma_t}{t} \sum_{s=1}^{t} \mathbb{I}\left(\check{a}_s = \check{a}_t\right),$$

for all $t \geq T_\gamma$. Taking expectation on both sides given $\check{a}_t$,

$$\sum_{u \in \mathcal{E}_t} \mathbb{P}\left(a_u = \check{a}_t\right) \geq \frac{\gamma_t}{t} \sum_{s=1}^{t} \mathbb{P}\left(\check{a}_s = \check{a}_t\right).$$

Because the distribution of $\check{a}_s$ and a_u for $u \in \mathcal{E}_t$ are the same,

$$\sum_{u \in \mathcal{E}_t} \mathbb{P}\left(a_u = \check{a}_t\right) = |\mathcal{E}_t| \sum_{i=1}^{d} \frac{\pi^{(i)}(\check{a}_t)}{d}, \quad \frac{\gamma_t}{t} \sum_{s=1}^{t} \mathbb{P}\left(\check{a}_s = \check{a}_t\right) = \gamma_t \sum_{i=1}^{d} \frac{\pi^{(i)}(\check{a}_t)}{d},$$

which implies $|\mathcal{E}_t| \geq \gamma_t$. □

B.8 Proof of Theorem 4.2

Theorem B.7 (Theorem 4.2 restated.)**.** *Let $\widehat{\theta}_t^{\langle \ell \rangle}$ denote the DR-mix estimator* (11) *with the exploration-mixed estimator* (8) *as the imputation estimator and pseudo-rewards* (10)*. Let $F_t := \sum_{k=1}^{K} x_k x_k^\top + I_d$, Then, for all $k \in [K]$, $\ell \in [L]$, and $t \geq T_\gamma$,*

$$\left| x_k^\top (\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}) \right| \leq 3 \|x_k\|_{F_t^{-1}} \{\theta_{\max} + \sigma \sqrt{d \log(Lt/\delta)}\}. \tag{27}$$

with probability at least $1 - 7\delta$. For each $k \in [K]$, with probability at least $1 - 7\delta$,

$$\left| x_k^\top (\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle}) \right| \leq 3 \|x_k\|_{F_t^{-1}} \left\{\theta_{\max} + 3\sigma \sqrt{\log(4Lt^2/\delta)}\right\}. \tag{28}$$

Proof. The proof for the two bounds are derived by simple computation using Lemma B.5 and we only prove the second bound. By Lemma B.5, with probability at least $1 - 3\delta$,

$$\left| x^\top \left(\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right) \right| \leq \|x\|_{F_t^{-1}} \left(\theta_{\max} + 8\sigma \sqrt{2 \log \frac{4Lt^2}{\delta}} + 3d \sqrt{\frac{2}{t} \log \frac{2dt^2}{\delta}} \left\| \check{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right\|_{F_t} \right).$$

for all $x \in \mathcal{X}$, $\ell \in [L]$ and $t \geq 1$. By Lemma B.4, with probability at least $1 - 4\delta$,

$$\begin{aligned}
\left\| \check{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right\|_{F_t} \leq & 2 \left\| \theta_\star^{\langle \ell \rangle} \right\|_{F_t^{-1}} + 12\sigma \sqrt{3d \log\left(\frac{3Lt}{\delta}\right)} + 24\sigma \sqrt{\frac{3td}{|\mathcal{E}_t|} \log\left(\frac{3Lt}{\delta}\right)} \\
\leq & 2 \left\| \theta_\star^{\langle \ell \rangle} \right\|_{F_t^{-1}} + 12\sigma \sqrt{3d \log\left(\frac{3Lt}{\delta}\right)} + 24\sigma \sqrt{\frac{3td}{\gamma_t} \log\left(\frac{3Lt}{\delta}\right)} \\
\leq & 2\theta_{\max} + 25\sigma \sqrt{\frac{3td}{\gamma_t} \log \frac{3Lt}{\delta}},
\end{aligned}$$

where the second inequality uses $|\mathcal{E}_t| \geq \gamma_t$ (Lemma B.6) and the last inequality holds by sufficiently large $t \geq 144\gamma_t$. Then,

$$\left| x^\top \left(\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right) \right| \leq \|x\|_{F_t^{-1}} \left(3\theta_{\max} + 8\sigma \sqrt{\log \frac{4Lt^2}{\delta}} + 75\sigma d \sqrt{\frac{6d}{\gamma_t} \log \frac{3Lt}{\delta} \log \frac{2dt^2}{\delta}} \right)$$

for all $\ell \in [L]$. By construction,

$$\gamma_t = C \cdot d^3 \log(2dt^2/\delta) := 6 \cdot (75)^2 d^3 \log(2dt^2/\delta) \tag{29}$$

and

$$\left| x^\top \left(\widehat{\theta}_t^{\langle \ell \rangle} - \theta_\star^{\langle \ell \rangle} \right) \right| \leq 3 \|x\|_{F_t^{-1}} \left(\theta_{\max} + 3\sigma \sqrt{\log \frac{4Lt^2}{\delta}} \right).$$

□

B.9 Instant Arm Elimination of PFIwR

Lemma B.8 (Instant arm elimination in PFIwR.). *For $t \geq T_\gamma$ such that $\max_{k\in[K]} \beta_{k,t} < \Delta_{(k),\epsilon}/4$, the arm k is correctly identified as suboptimal or Pareto optimal, and $k \notin \mathcal{A}_{t+1}$.*

Proof. **Case 1. $k \notin \mathcal{P}_\star$ and $\Delta_k^\star > \epsilon$:** Suppose $k \in [K] \setminus \mathcal{P}_\star$ and $\Delta_{(k),\epsilon}/4 > \beta_t$. If $\Delta_{(k),\epsilon} = \Delta_k^\star$, then there exists $k_\star \in \mathcal{P}_\star \subseteq \mathcal{A}_t \cup \mathcal{P}_t$ such that $\Delta_k^\star = m(k, k^\star)$ and

$$\widehat{m}_t(k, k^\star) \geq m(k, k^\star) - \beta_{k,t} - \beta_{k',t} > 4\beta_t - \beta_{k,t} - \beta_{k',t} > \beta_{k,t} + \beta_{k',t}.$$

Thus, $k \in \mathcal{A}_t \setminus \mathcal{C}_t$, and $k \notin \mathcal{A}_{t+1}$.

Case 2. $k \notin \mathcal{P}_\star$ and $\Delta_k^\star \leq \epsilon$: If $k \notin \mathcal{C}_t$ then $k \notin \mathcal{A}_t$. Consider the case of $k \in \mathcal{C}_t$. Because $\mathcal{P}_\star \subset \mathcal{C}_t \cup \mathcal{P}_{t-1}$, we obtain $\Delta_k^\star = \max_{k' \in \mathcal{C}_t \cup \mathcal{P}_{t-1}} m(k, k')$.Then for all $k' \in \mathcal{C}_t \cup \mathcal{P}_{t-1} \setminus \{k\}$,

$$\begin{aligned}
\widehat{M}_t^{2\epsilon}(k, k') \geq& \max \left\{ 0, 2\epsilon + \max_{\ell \in [L]} (y_k^{\langle \ell \rangle} - y_{k'}^{\langle \ell \rangle}) \right\} - \beta_{k,t} - \beta_{k',t} \\
\geq& 2\epsilon - \min_{\ell \in [L]} (y_{k'}^{\langle \ell \rangle} - y_k^{\langle \ell \rangle}) - \beta_{k,t} - \beta_{k',t} \\
\geq& 2\epsilon - m(k, k') - \beta_{k,t} - \beta_{k',t} \\
\geq& 2\epsilon - \Delta_k^\star - \beta_{k,t} - \beta_{k',t}.
\end{aligned}$$

Because $\Delta_k^\star \leq \epsilon$,

$$\begin{aligned}
\widehat{M}_t^{2\epsilon}(k, k') \geq& \epsilon - \beta_{k,t} - \beta_{k',t} \\
\geq& 4 \max_{k \in [K]} \beta_{k,t} - \beta_{k,t} - \beta_{k',t} \\
\geq& \beta_{k,t} + \beta_{k',t},
\end{aligned}$$

and $k \in \mathcal{P}_t^{(1)}$. Thus, $k \in \mathcal{P}_{t+1}$ and $k \notin \mathcal{A}_{t+1}$.

Case 3. $k \in \mathcal{P}_\star$ and $\Delta_k > \epsilon$: Suppose $k \in \mathcal{P}_\star$. Then for all $k' \in [K]$,

$$\widehat{m}_t(k, k') \leq m(k, k') + \beta_{k,t} + \beta_{k',t} = \beta_{k,t} + \beta_{k',t},$$

and $k \in \mathcal{C}_t$. Suppose $\Delta_{(k),\epsilon}/4 > \beta_t$. For a Pareto optimal arm $k' \in \mathcal{C}_t \cup \mathcal{P}_{t-1} \setminus \{k\}$,

$$\begin{aligned}
\widehat{M}_t^{2\epsilon}(k, k') \geq& M(k, k') - \beta_{k,t} - \beta_{k',t} \\
\geq& \Delta_k^+ - \beta_{k,t} - \beta_{k',t} \\
\geq& \Delta_k - \beta_{k,t} - \beta_{k',t} \\
\geq& \beta_{k,t} + \beta_{k',t}.
\end{aligned}$$

For a suboptimal arm $k^- \in \mathcal{C}_t \cup \mathcal{P}_{t-1} \setminus \{k\}$, there exists $k^+ \in \mathcal{P}_\star$ such that $y_{k^-} + \Delta_{k^-}^\star$ is weakly dominated by y_{k^+}, and

$$\begin{aligned}
\widehat{M}_t^{2\epsilon}(k, k^-) \geq& M(k, k^-) - \beta_{k,t} - \beta_{k^-,t} \\
\geq& M(k, k^+) + \Delta_{k^-}^\star - \beta_{k,t} - \beta_{k^-,t}.
\end{aligned}$$

Consider the case $k^+ \neq k$, then

$$\begin{aligned}\widehat{M}_t^{2\epsilon}(k,k^-) \geq & M(k,k^+) - \beta_{k,t} - \beta_{k^-,t}\\ \geq & \Delta_k^+ - \beta_{k,t} - \beta_{k^-,t}\\ \geq & \beta_{k,t} + \beta_{k^-,t},\end{aligned}$$

and $k \in \mathcal{P}_t^{(1)}$ and $k \notin \mathcal{A}_{t+1}$. For the case of $k^+ = k$,

$$\begin{aligned}\widehat{M}_t^{2\epsilon}(k,k^-) \geq & \Delta_{k^-}^\star - \beta_{k,t} - \beta_{k^-,t}\\ = & M(k^-,k) + \Delta_{k^-}^\star - \beta_{k,t} - \beta_{k^-,t}\\ \geq & \Delta_k^- - \beta_{k,t} - \beta_{k^-,t}\\ \geq & \beta_{k,t} + \beta_{k^-,t}.\end{aligned}$$

Case 4. $k \in \mathcal{P}_\star$ **and** $\Delta_k \leq \epsilon$**:** If $\Delta_{(k),\epsilon} = \epsilon$ then $\Delta_k \leq \epsilon$. Thus, for all $k' \in \mathcal{C}_t \cup \mathcal{P}_{t-1} \setminus \{k\}$, because $k \in \mathcal{P}_\star$, we obtain $\max_{\ell \in [L]}(y_k^{\langle \ell \rangle} - y_{k'}^{\langle \ell \rangle}) \geq 0$ and

$$\widehat{M}_t^{2\epsilon}(k,k') \geq \max\left\{0, 2\epsilon + \max_{\ell\in[L]}(y_k^{\langle \ell \rangle} - y_{k'}^{\langle \ell \rangle})\right\} - \beta_{k,t} - \beta_{k',t} \geq 2\epsilon - \beta_{k,t} - \beta_{k',t} \geq \beta_{k,t} + \beta_{k',t},$$

and $k \in \mathcal{P}_t^{(1)}$. Thus, $k \notin \mathcal{A}_{t+1}$. □

B.10 Proof of Theorem 5.1

Before we prove the sample complexity, we provide an important properties of our proposed `PFIwR` algorithm.

Lemma B.9. *For $t \geq 1$, the Pareto optimal arms are either in $\mathcal{A}_t$ or $\mathcal{P}_t$ in `PFIwR`, i.e., $\mathcal{P}_\star \subseteq \mathcal{A}_t \cup \mathcal{P}_t$.*

Proof. When $t = 0$, the result holds by definition of $\mathcal{A}_0 = [K]$. For $t \geq 1$, suppose $\mathcal{P}_\star \subseteq \mathcal{A}_{t-1} \cup \mathcal{P}_{t-1}$ holds. While updating $\mathcal{A}_t$ and $\mathcal{P}_t$, only arms in $\mathcal{A}_{t-1} \backslash \mathcal{C}_t$ are eliminated. Thus, we prove the results by showing that $\mathcal{A}_{t-1} \backslash \mathcal{C}_t \subseteq \mathcal{P}_\star^c$. For each round t, suppose an arm $k \in \mathcal{A}_{t-1} \backslash \mathcal{C}_t$. Then there exists $k' \in [K]$ such that

$$\widehat{m}_t(k,k') > \beta_{k,t} + \beta_{k',t},$$

which implies

$$y_k^{\langle l \rangle} \leq \widehat{y}_{k,t}^{(l)} + \beta_{k,t} \leq \widehat{y}_{k',t}^{(l)} - \beta_{k',t} \leq y_{k'}^{\langle l \rangle},$$

for all $\ell \in [L]$ and $k \notin \mathcal{P}_\star$. Thus, $\mathcal{A}_{t-1} \backslash \mathcal{C}_t \subseteq \mathcal{P}_\star^c$ is proved. □

Now we are ready to prove the sample complexity of `PFIwR` .

Theorem B.10 (Theorem 5.1 restated)**.** *Fix $\epsilon > 0$ and $\delta \in (0,1)$. Define $\Delta_{(k),\epsilon} = \max\{\epsilon, \Delta_k\}$, where Δ_k is the required accuracy defined in* (3) *with ascending order $\Delta_1 \leq \cdots, \leq \Delta_K$. Then the stopping time $\tau_{\epsilon,\delta}$ of `PFIwR` is bounded by:*

$$\max\left\{O\left(\sum_{k=1}^{d} \frac{d}{\Delta_{(k),\epsilon}^2} \log \frac{dL}{\Delta_\epsilon^2 \delta}\right), T_\gamma\right\}.$$

Proof. **Step 1. Sample complexity for accuracy of the estimator:** For $k \in [K]$, let $\beta_{k,t}$ denote the confidence bound defined in (15). Because $\max_{k\in[K]} x_k^\top F_t^{-1} x_k \leq 1/t$, for $k \in [K]$

$$\beta_{k,t} \leq \beta_t := \begin{cases} \frac{3}{\sqrt{t}}\{\theta_{\max} + \sigma\sqrt{d\log(7Lt/\delta)}\} & |\mathcal{A}_t| > d \\ \frac{3}{\sqrt{t}}\{\theta_{\max} + 3\sigma\sqrt{\log(56Ldt^2/\delta)}\} & |\mathcal{A}_t| \leq d \end{cases}$$

By Theorem 4.2, with probability at least $1-\delta$,

$$\left|y_k^{\langle \ell \rangle} - \widehat{y}_{k,t}^{(\ell)}\right| \leq \beta_{k,t} \leq \beta_t.$$

holds for all $t \geq T_\gamma$, $k \in [K]$ and $\ell \in [L]$ such that $|\mathcal{A}_t| > d$. Thus, for any $\Delta > 0$, if

$$t \geq \frac{9\left(\theta_{\max}+\sigma\right)^2}{\Delta^2} d \log \frac{7Lt}{\delta} \tag{30}$$

then $\left|y_k^{\langle l\rangle} - \widehat{y}_{k,t}^{(l)}\right| \leq \beta_{k,t} \leq \beta_t \leq \Delta$ for all $k \in [K]$ when $|\mathcal{A}_t| > d$. By Lemma C.6, the condition (30) is implied by

$$t \geq \frac{4 \cdot 9d(\theta_{\max}+\sigma)^2}{\Delta^2}\left\{1 + \log \frac{2\cdot 9d(\theta_{\max}+\sigma)^2}{e\Delta^2}\sqrt{\frac{7L}{\delta}}\right\}. \tag{31}$$

For $|\mathcal{A}_t| \leq d$, we only need confidence interval for at most $2d$ arms that affects PFI. Let $\mathcal{N}_t$ denote the arms that are nearest to $\mathcal{A}_t$. Then for $k \in \mathcal{N}_t \cup \mathcal{A}_t$,

$$\left|y_k^{\langle \ell\rangle} - \widehat{y}_{k,t}^{(\ell)}\right| \leq \beta_{k,t} \leq \beta_t$$

holds for $|\mathcal{A}_t| \leq d$. Similarly,

$$t \geq \frac{4 \cdot 9(\theta_{\max}+\sigma)^2}{\Delta^2}\left\{1 + \log \frac{2\cdot 9(\theta_{\max}+3\sigma)^2}{e\Delta^2}\sqrt{\frac{56Ld}{\delta}}\right\}, \tag{32}$$

implies $\left|y_k^{\langle \ell\rangle} - \widehat{y}_{k,t}^{(\ell)}\right| \leq \beta_{k,t} \leq \beta_t \leq \Delta$ for $k \in \mathcal{N}_t \cup \mathcal{A}_t$.

Step 2. Finding the sample complexity: From (31), for

$$t \geq \frac{16\cdot 4 \cdot 9d(\theta_{\max}+\sigma)^2}{\Delta^2_{(d+1),\epsilon}}\left\{1 + \log \frac{16\cdot 2\cdot 9d(\theta_{\max}+\sigma)^2}{e\Delta^2_{(d+1),\epsilon}}\sqrt{\frac{7L}{\delta}}\right\}$$

implies $\Delta_{(i),\epsilon} \geq 4\beta_{i,t}$ for $i = d+1, \ldots, K$. Then by Lemma B.8, $|\mathcal{A}_{t+1}| \leq d$. If

$$\begin{aligned} t \geq & \frac{16\cdot 4 \cdot 9d(\theta_{\max}+\sigma)^2}{\Delta^2_{(d+1),\epsilon}}\left\{1 + \log \frac{16\cdot 2\cdot 9d(\theta_{\max}+\sigma)^2}{e\Delta^2_{(d+1),\epsilon}}\sqrt{\frac{7L}{\delta}}\right\} \\ & + \sum_{k=1}^{d} \frac{16\cdot 4 \cdot 9(\theta_{\max}+\sigma)^2}{\Delta^2_{(k),\epsilon}}\left\{1 + \log \frac{16\cdot 2\cdot 9(\theta_{\max}+3\sigma)^2}{e\Delta^2_{(k),\epsilon}}\sqrt{\frac{56Ld}{\delta}}\right\} \end{aligned}$$

then $\Delta_{k,\epsilon} \geq 4\beta_{k,t}$ for all $k \in [K]$ and $\mathcal{A}_{t+1} = \emptyset$ by Lemma B.8. Since Theorem 4.2 requires $t \geq T_\gamma$, the sample complexity is bounded as

$$\begin{aligned} \tau_{\epsilon,\delta} \leq \max & \left[T_\gamma, \frac{16\cdot 4 \cdot 9d(\theta_{\max}+\sigma)^2}{\Delta^2_{(d+1),\epsilon}}\left\{1 + \log \frac{16\cdot 2\cdot 9d(\theta_{\max}+\sigma)^2}{e\Delta^2_{(d+1),\epsilon}}\sqrt{\frac{7L}{\delta}}\right\}\right. \\ & \left. + \sum_{k=1}^{d} \frac{16\cdot 4 \cdot 9(\theta_{\max}+\sigma)^2}{\Delta^2_{(k),\epsilon}}\left\{1 + \log \frac{16\cdot 2\cdot 9(\theta_{\max}+3\sigma)^2}{e\Delta^2_{(k),\epsilon}}\sqrt{\frac{56Ld}{\delta}}\right\}\right]. \end{aligned}$$

The proof completes by the fact that $d\Delta^{-2}_{(d+1),\epsilon} \leq \sum_{k=1}^{d} \Delta^{-2}_{(k),\epsilon}$. □

B.11 Proof of Theorem 5.2

Proof. By Lemma B.9, the Pareto front $\mathcal{P}_\star \subseteq \mathcal{A}_t \cup \mathcal{P}_t$. By definition of $\mathcal{A}_t$ and $\mathcal{P}_t$ in the algorithm,

$$\begin{aligned} \mathcal{A}_t \cup \mathcal{P}_t = & \mathcal{C}_t \cup \mathcal{P}_{t-1} \\ = & \mathcal{C}_t \cup \mathcal{P}_{t-1}^{(2)} \cup \mathcal{P}_{t-2} \\ \subseteq & \mathcal{C}_t \cup \mathcal{C}_{t-1} \cup \mathcal{P}_{t-2} \\ & \vdots \\ \subseteq & \bigcup_{s=1}^{t} \mathcal{C}_s. \end{aligned}$$

Note that $\mathcal{C}_{s+1} \subseteq \mathcal{A}_s \subseteq \mathcal{C}_s$ for $s \geq 1$. Thus, $\mathcal{P}_\star \subseteq \mathcal{C}_1$, and the Pareto regret

$$\Delta^\star_{a_t} = \max_{k \in \mathcal{P}_\star} m(a_t, k) \leq \max_{k \in \mathcal{C}_1} m(a_t, k).$$

For $s \in [t-1]$, suppose $k_s \in \mathcal{C}_s \setminus \mathcal{C}_{s+1}$. Then there exists $k'_s \in \mathcal{A}_s$ such that $\widehat{m}_s(k_s, k'_s) > D_{k_s, k'_2, s}$. By Theorem 4.2, $m(k_s, k'_s) > 0$ with probability at least $1 - \delta$ and k_s is dominated by $k'_s \in \mathcal{A}_s$, which is dominated by the arms in $\mathcal{C}_{s+1}$ by definition of $\mathcal{C}_{s+1}$. Thus,

$$\begin{aligned}
\max_{k \in \mathcal{C}_1} m(a_t, k) &= \max_{k \in \mathcal{C}_1 \setminus \mathcal{C}_2 \cup \mathcal{C}_2} m(a_t, k) \\
&= \max_{k \in \mathcal{C}_2} m(a_t, k) \\
&\vdots \\
&= \max_{k \in \mathcal{C}_t} m(a_t, k).
\end{aligned}$$

By definition of a_t,

$$\begin{aligned}
\Delta^\star_{a_t} &\leq \max_{k \in \mathcal{C}_t} m(a_t, k) \\
&\leq \max_{k \in \mathcal{A}_{t-1}} m(a_t, k) \\
&\leq 2 \max_{j \in \mathcal{A}_{t-1}} \beta_{j,t-1} + \max_{k \in \mathcal{A}_{t-1}} \widehat{m}_{t-1}(a_t, k) \\
&= 2 \max_{j \in \mathcal{A}_{t-1}} \beta_{j,t-1},
\end{aligned}$$

which proves the instantaneous regret bound.

To prove the cumulative regret bound, summing up the regret over $s \in [\tau_{\epsilon,\delta}]$, with probability at least $1 - \delta$,

$$\begin{aligned}
R(\tau_{\epsilon,\delta}) &\leq \sum_{t=1}^{\tau_{\epsilon,\delta}} \mathbb{I}\left(t \in \mathcal{E}_t\right) 2\theta_{\max} + \mathbb{I}\left(t \notin \mathcal{E}_t\right) \Delta^\star_{a_t} \\
&= 2\theta_{\max} \gamma_{\tau_{\epsilon,\delta}} + \sum_{t=1}^{\tau_{\epsilon,\delta}} \mathbb{I}\left(t \notin \mathcal{E}_t\right) \Delta^\star_{a_t} \\
&= \bar{O}\left(\theta_{\max} d^3 \log \frac{\theta_{\max} d}{\delta \Delta_{(1),\epsilon}}\right) + \sum_{t=1}^{\tau_{\epsilon,\delta}} \mathbb{I}\left(t \notin \mathcal{E}_t\right) \Delta^\star_{a_t},
\end{aligned}$$

where $\bar{O}$ ignores $\log\log(\cdot)$ terms and the last equality holds by the sample complexity bound (Theorem 5.1). Because the instantaeneous regret is bounded by $\Delta^\star_{a_t} \leq 2 \max_{k \in \mathcal{A}_{t-1}} \beta_{k,t-1}$, the regret is zero when $2 \max_{k \in \mathcal{A}_{t-1}} \beta_{k,t-1} \leq \min_{k \in [K] \setminus \mathcal{P}_\star} \Delta^\star_k$, which is implied by $4\beta_{t-1} \leq \min_{k \in [K] \setminus \mathcal{P}_\star} \Delta^\star_k$. In addition, by Lemma B.8, the algorithm terminates when $\max_{k \in \mathcal{A}_{t-1}} \beta_{k,t-1} \leq \beta_{t-1} \leq \epsilon/4 \leq \min_{k \in [K]} \Delta_{(k),\epsilon}$. Thus,

$$\begin{aligned}
R(\tau_{\epsilon,\delta}) \leq &\bar{O}\left(\theta_{\max} d^3 \log \frac{\theta_{\max} d}{\delta \Delta_{(1),\epsilon}}\right) + \sum_{t=1}^{\tau_{\epsilon,\delta}} \mathbb{I}\left(4\beta_{t-1} > \max\left\{\min_{k \in [K] \setminus \mathcal{P}_\star} \Delta^\star_k, \epsilon\right\}\right) \mathbb{I}\left(t \notin \mathcal{E}_t\right) \Delta^\star_{a_t} \\
= &\bar{O}\left(\theta_{\max} d^3 \log \frac{\theta_{\max} d}{\delta \Delta_{(1),\epsilon}}\right) + \sum_{t=1}^{\tau_{\epsilon,\delta}} \mathbb{I}\left(4\beta_{t-1} > \Delta^\star_\epsilon\right) \mathbb{I}\left(t \notin \mathcal{E}_t\right) \Delta^\star_{a_t}
\end{aligned}$$

From (31) and (32), define

$$g\left(\frac{\Delta^\star_\epsilon}{4}\right) := \frac{16 \cdot 4 \cdot 9d(\theta_{\max} + \sigma)^2}{(\Delta^\star_\epsilon)^2} \left\{1 + \log \frac{16 \cdot 2 \cdot 9d(\theta_{\max} + \sigma)^2}{e(\Delta^\star_\epsilon)^2} \sqrt{\frac{56L}{\delta}}\right\} \tag{33}$$

Then $t \geq g(\Delta^\star_\epsilon/4)$ implies $\Delta^\star_\epsilon \geq 4\beta_t$ and

$$\begin{aligned}\sum_{t=1}^{\tau_{\epsilon,\delta}} \mathbb{I}\left(4\beta_{t-1} > \Delta^\star_\epsilon\right) \mathbb{I}\left(t \notin \mathcal{E}_t\right) \Delta^\star_{a_t} &\leq \sum_{t=1}^{\tau_{\epsilon,\delta}} \mathbb{I}\left(t-1 \leq g(\Delta^\star_\epsilon/4)\right) \mathbb{I}\left(t \notin \mathcal{E}_t\right) \Delta^\star_{a_t} \\ &\leq \sum_{t=1}^{1+g(\Delta^\star_\epsilon/4)} \Delta^\star_{a_t}.\end{aligned}$$

Because $\Delta^\star_{a_t} \leq 2\max_{k\in\mathcal{A}_{t-1}} \beta_{k,t}$,

$$\begin{aligned}\sum_{t=1}^{1+g(\Delta^\star_\epsilon/4)} \Delta^\star_{a_t} \leq& 2 \sum_{t=1}^{1+g(\Delta^\star_\epsilon/4)} \max_{k\in\mathcal{A}_{t-1}} \beta_{k,t} \\ \leq& 6\left(\theta_{\max}+\sigma\right) \sum_{t=1}^{1+g(\Delta^\star_\epsilon/4)} \frac{\sqrt{d\log(56Lt/\delta)}}{\sqrt{t}} \\ \leq& 12\left(\theta_{\max}+\sigma\right) \sqrt{\{1+g(\Delta^\star_\epsilon/4)\}\log\frac{7L\{1+g(\Delta^\star_\epsilon/4)\}^2}{\delta}}.\end{aligned}$$

Plugging in (33) and ignoring $\log\log(\cdot)$ terms,

$$R(\tau_{\epsilon,\delta}) = \bar{O}\left(\theta_{\max} d^3 \log\frac{\theta_{\max} d}{\delta\Delta_{(1),\epsilon}} + \frac{\theta_{\max} d\sigma^2}{\Delta^\star_\epsilon}\log\frac{\theta_{\max} d\sigma}{\Delta^\star_\epsilon\delta}\right). \tag{34}$$

□

B.12 Proof of Theorem 3.3

Theorem B.11 (Theorem 3.3 restated.)**.** *For $\epsilon > 0$, let $\Delta^\star_\epsilon := \max\{\epsilon, \min_{k\in[K]\setminus\mathcal{P}_\star}\Delta^\star_k\}$ denote the minimum Pareto regret over suboptimal arms. Suppose the set of context vectors $\mathcal{X}$ span $\mathbb{R}^d$ and $\min_{\ell\in[L]}\|\theta^{\langle\ell\rangle}_\star\|_0 = d$. Then, for any $\delta\in(0,1/4)$ and $\sigma>0$, there exists a σ-sub-Gaussian distribution for the i.i.d. noise sequence $\{\eta_t\}_{t\geq1}$ such that for any PFI algorithms that satisfies PFI success condition (1) with failure probability δ,*

$$R(\tau_{\epsilon,\delta}) \geq \frac{\sqrt{3}d\sigma}{8\Delta^\star_\epsilon}\log\frac{1}{4\delta}.$$

Theorem 5.3 shows that `PFIwR` establishes nearly optimal regret among algorithms that achieve PFI and it is the first result on the trade-off between PFI and Pareto regret minimization. For $L=1$ and the contexts are Euclidean basis, Theorem 5.3 recovers the lower bound for regret of BAI algorithms developed by Zhong et al. (2023). Note that the lower bound applies only to the algorithms that guarantee PFI; it is possible for an algorithm that does not guarantee PFI to have a regret lower bound that is lower than the one in Theorem 5.3.

Proof. By Lemma B.1, for $\delta\in(0,1/4)$ and any estimator $\widehat{\theta}^{\langle\ell\rangle}_t$ with round $t \leq d\sigma^2/(12(\Delta^\star_\epsilon)^{-2}\log\frac{3(d+1)L}{4\delta})$ for $\theta^{\langle\ell\rangle}_\star$,

$$\mathbb{P}\left(\max_{\ell\in[L]}\max_{x\in\mathcal{X}}\left|x^\top\left(\widehat{\theta}^{\langle\ell\rangle}_t - \theta^{\langle\ell\rangle}_\star\right)\right| > \Delta^\star_\epsilon\right) \geq 1-\left(1-\frac{\delta}{3(d+1)L}\right)^L \geq \delta,$$

and any estimator cannot find an arm with zero Pareto regret with probability at least $1-\delta$. Thus, we need at least $d\sigma^2/(12(\Delta^\star_\epsilon)^{-2}\log\frac{3(d+1)L}{4\delta})$ number of rounds to ensure that the estimation error is less than the minimum Pareto regret $\Delta^\star_\epsilon$. By Theorem 5.6 in Kim et al. (2023b), for any horizon $T\geq d$, the expected regret is $\Omega(\sqrt{dT})$ in the single objective linear bandit setting where the number of arms is finite and the contexts span $\mathbb{R}^d$. Since the same lower bound applies to the multi-dimensional rewards as well, setting $T = d\sigma^2/(12(\Delta^\star_\epsilon)^{-2}\log\frac{1}{4\delta})$ gives the lower bound. □

C Technical Lemmas

In this section, we provide technical lemmas cited from the literature and novel lemmas (Lemma C.5 and Lemma C.6).

Lemma C.1 (Assouad's method, (Yu, 1997))**.** *For $v \in \{\pm 1\}^d$, let $\mathbb{P}_v$ denote the probability measure on the data space $\mathcal{D}$ whose parameter is v. For any collection of estimators $f = (f_1, \ldots, f_d), f_i : \mathcal{D} \to \{\pm 1\}$ there exists at least one $v \in \{\pm 1\}^d$ such that*

$$\mathbb{E}_v \left[\sum_{j=1}^{d} \mathbb{I}\left(f_j \neq v\right) \right] \geq \frac{d}{2} \min_{v, v' : v \sim v'} \left\| \min(\mathbb{P}_v, \mathbb{P}_{v'}) \right\|_1,$$

where $v \sim v'$ indicates that v and v' only differ in one coordinate.

Lemma C.2 (A dimension-free bound for vector-valued martingales. Lemma C.6 in Kim et al. (2023a).)**.** *Let $\{\mathcal{F}_s\}_{s=0}^t$ be a filtration and $\{\eta_s\}_{s=1}^t$ be a real-valued stochastic process such that η_s is $\mathcal{F}_\tau$-measurable. Let $\{X_s\}_{s=1}^t$ be an $\mathbb{R}^d$-valued stochastic process where X_s is $\mathcal{F}_0$-measurable. Assume that $\{\eta_s\}_{s=1}^t$ are σ-sub-Gaussian given $\{\mathcal{F}_s\}_{s=1}^t$. Then with probability at least $1-\delta$,*

$$\left\| \sum_{s=1}^{t} \eta_s X_s \right\|_2 \leq 4\sigma \sqrt{\sum_{s=1}^{t} \|X_s\|_2^2} \sqrt{2 \log \frac{4t^2}{\delta}}. \tag{35}$$

Remark C.3. While the constant in Kim et al. (2023a) is 12, we prove that the bound also holds with 4 using the following lemma.

Lemma C.4. *Suppose a random variable X satisfies $\mathbb{E}[X] = 0$, and let Y be an σ-sub-Gaussian random variable. If $|X| \leq |Y|$ almost surely, then X is 2σ-sub-Gaussian.*

Lemma C.5 (A Hoeffding bound for the matrices)**.** *Let $\{M_\tau : \tau \in [t]\}$ be a $\mathbb{R}^{d \times d}$-valued stochastic process adapted to the filtration $\{\mathcal{F}_\tau : \tau \in [t]\}$, i.e., M_τ is $\mathcal{F}_\tau$-measurable for $\tau \in [t]$. Suppose the matrix M_τ is symmetric and the eigenvalues of the difference $M_\tau - \mathbb{E}[M_\tau | \mathcal{F}_{\tau-1}]$ lies in $[-b, b]$ for some $b > 0$. Then for $x > 0$,*

$$\mathbb{P}\left(\left\| \sum_{\tau=1}^{t} M_\tau - \mathbb{E}[M_\tau | \mathcal{F}_{\tau-1}] \right\|_2 \geq x \right) \leq 2d \exp\left(-\frac{x^2}{2tb^2} \right)$$

Proof. The proof is an adapted version of Hoeffding's inequality for matrix stochastic process with the argument of Tropp (2012). Let $D_\tau := M_\tau - \mathbb{E}[M_\tau | \mathcal{F}_{\tau-1}]$. Then, for $x > 0$,

$$\mathbb{P}\left(\left\| \sum_{\tau=1}^{t} D_\tau \right\|_2 \geq x \right) \leq \mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau \right) \geq x \right) + \mathbb{P}\left(\lambda_{\max}\left(-\sum_{\tau=1}^{t} D_\tau \right) \geq x \right)$$

We bound the first term and the second term is bounded with similar argument. For any $v > 0$,

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau \right) \geq x \right) &\leq \mathbb{P}\left(\exp\left\{ v \lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau \right) \right\} \geq e^{vx} \right) \\
&\leq e^{-vx} \mathbb{E}\left[\exp\left\{ v \lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau \right) \right\} \right].
\end{aligned}$$

Because $\sum_{\tau=1}^{t} D_\tau$ is a real symmetric matrix,

$$\exp\left\{ v \lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau \right) \right\} = \lambda_{\max}\left\{ \exp\left(v \sum_{\tau=1}^{t} D_\tau \right) \right\} \leq \mathrm{Tr}\left\{ \exp\left(v \sum_{\tau=1}^{t} D_\tau \right) \right\},$$

where the last inequality holds since $\exp(v\sum_{\tau=1}^{t} D_\tau)$ has nonnegative eigenvalues. Taking expectation on both side gives,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{v\lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau\right)\right\}\right] \leq & \mathbb{E}\left[\mathrm{Tr}\left\{\exp\left(v\sum_{\tau=1}^{t} D_\tau\right)\right\}\right] \\
= & \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t} D_\tau\right)\right] \\
= & \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1} D_\tau + \log\exp(vD_t)\right)\right].
\end{aligned}$$

By Lieb's theorem Tropp (2015) the mapping $D \mapsto \exp(H + \log D)$ is concave on positive symmetric matrices for any symmetric positive definite H. By Jensen's inequality,

$$\mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1} D_\tau + \log\exp(vD_t)\right)\right] \leq \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1} D_\tau + \log\mathbb{E}\left[\exp(vD_t)\middle|\,\mathcal{F}_{t-1}\right]\right)\right]$$

By Hoeffding's lemma,

$$e^{vx} \leq \frac{b-x}{2b}e^{-vb} + \frac{x+b}{2b}e^{vb}$$

for all $x \in [-b, b]$. Because the eigenvalue of D_τ lies in $[-b, b]$, we have

$$\begin{aligned}
\mathbb{E}\left[\exp(vD_t)\middle|\,\mathcal{F}_{t-1}\right] \preceq & \mathbb{E}\left[\frac{e^{-vb}}{2b}\left(bI_d - D_t\right) + \frac{e^{vb}}{2b}\left(D_t + bI_d\right)\middle|\,\mathcal{F}_{t-1}\right] \\
= & \frac{e^{-vb}+e^{vb}}{2}I_d \\
\preceq & \exp(\frac{v^2b^2}{2})I_d.
\end{aligned}$$

Recursively,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{v\lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau\right)\right\}\right] \leq & \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1} D_\tau + \log\mathbb{E}\left[\exp(vD_t)\middle|\,\mathcal{F}_{t-1}\right]\right)\right] \\
\leq & \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1} D_\tau + \frac{v^2b^2}{2}I_d\right)\right] \\
\leq & \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-2} D_\tau + \frac{v^2b^2}{2}I_d + \log\mathbb{E}\left[\exp(vD_{t-1})\middle|\,\mathcal{F}_{t-2}\right]\right)\right] \\
\leq & \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-2} D_\tau + \frac{2v^2b^2}{2}I_d\right)\right] \\
& \vdots \\
\leq & \mathrm{Tr}\exp\left((\frac{tv^2b^2}{2})I_d\right) \\
= & \exp\left(\frac{tv^2b^2}{2}\right)\mathrm{Tr}\left(I_d\right) \\
= & d\exp\left(\frac{tv^2b^2}{2}\right).
\end{aligned}$$

Thus we have

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau\right) \geq x\right) \leq d\exp\left(-vx + \frac{tv^2b^2}{2}\right).$$

Minimizing over $v > 0$ gives $v = x/(tb^2)$ and

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t} D_\tau\right) \geq x\right) \leq d \exp\left(-\frac{x^2}{2tb^2}\right),$$

which proves the lemma. $\square$

Lemma C.6 (Threshold for logarithmic inequality.). *For $a > 1/2$ and $b > e^2$, $t \geq 4a\left(1 + \log \frac{2a\sqrt{b}}{e}\right)$ implies $t \geq a \log bt^2$.*

Proof. If $b > e^2$ the function $x \mapsto (a/x) \log bx^2$ has negative derivatives and is decreasing on $x \geq 1$. Then, there exists a unique $\tilde{t} \geq 1$ such that $1 = (a/\tilde{t}) \log b(\tilde{t})^2$. Now, it is sufficient to show that

$$\tilde{t} \leq 2a\left(1 + \sqrt{\log(2a\sqrt{b})/e}\right)^2 \leq 4a\left(1 + \log \frac{2a\sqrt{b}}{e}\right). \tag{36}$$

Let $W : [-1/e, 0) \to \mathbb{R}$ denote the Lambert function which satisfies $W(x)e^{W(x)} = x$ for $x \in [-1/e, 0)$. By definition of $\tilde{t}$,

$$\begin{aligned} -\frac{1}{2a\sqrt{b}} = e^{\log \frac{1}{\sqrt{b\tilde{t}}}} \log \frac{1}{\sqrt{b\tilde{t}}} \Longrightarrow & W\left(-\frac{1}{2a\sqrt{b}}\right) = \log \frac{1}{\sqrt{b\tilde{t}}} \\ \Longrightarrow & \tilde{t} = \frac{1}{\sqrt{b}} \exp\left(-W\left(-\frac{1}{2a\sqrt{b}}\right)\right). \end{aligned}$$

By definition of W,

$$\tilde{t} = -\frac{2a\sqrt{b}}{\sqrt{b}} W\left(-\frac{1}{2a\sqrt{b}}\right) = -2aW\left(-\frac{1}{2a\sqrt{b}}\right).$$

By Theorem 1 in Chatzigeorgiou (2013), $-W(-e^{-u-1}) < 1 + \sqrt{2u} + u \leq (1 + \sqrt{u})^2$ for $u > 0$. Setting $u = \log \frac{2a\sqrt{b}}{e}$ proves (36). $\square$

D Limitation

Although our main contribution is novel and improves current linear bandit algorithms, we found the following limitations are in need to be handled in the future work:

- The number of exploration T_γ can dominate the sample complexity in Theorem 5.1 when the problem complexity gaps are large. The term T_γ does not have problem complexity gap and is not considered as the main term in theoretical analysis; while our proposed estimator may not efficient for the large gaps in practice.
- Our PFI comparison experiments are limited to MAB setting, although we design our algorithm for general contexts with possibly exponentially large number of arms. We choose the MAB setting for the sake of comparison with the previous algorithm (Auer et al., 2016); we believe the superior performance of our algorithm may be drastically visible on general contexts with large number of arms.