Residual Bootstrap Exploration for Stochastic Linear Bandit

Shuang Wu[1] **Chi-Hua Wang**[1,2] **Yuantong Li**[1] **Guang Cheng**[1]

[1]Department of Statistics, University of California, Los Angeles, Los Angeles, California, USA
[2]Department of Statistics, Purdue University, West Lafayette, Indiana, USA

Abstract

We propose a new bootstrap-based online algorithm for stochastic linear bandit problems. The key idea is to adopt residual bootstrap exploration, in which the agent estimates the next step reward by re-sampling the residuals of mean reward estimate. Our algorithm, residual bootstrap exploration for stochastic linear bandit (`LinReBoot`), estimates the linear reward from its re-sampling distribution and pulls the arm with the highest reward estimate. In particular, we contribute a theoretical framework to demystify residual bootstrap-based exploration mechanisms in stochastic linear bandit problems. The key insight is that the strength of bootstrap exploration is based on collaborated optimism between the online-learned model and the re-sampling distribution of residuals. Such observation enables us to show that the proposed `LinReBoot` secure a high-probability $\tilde{O}(d\sqrt{n})$ sub-linear regret under mild conditions. Our experiments support the easy generalizability of the `ReBoot` principle in the various formulations of linear bandit problems and show the significant computational efficiency of `LinReBoot`.

1 INTRODUCTION

Stochastic linear bandit is an online learning problem that the learning agent acts by pulling arms, where each arm is associated with a feature vector, then learning the arms information from the corresponding random rewards. In such problems, the typical goal of a learning agent is to maximize its cumulative reward. Learning more about an arm (explore) or pulling the arm with the highest estimated reward (exploit) leads to the well-known *exploration- exploitation trade-off*, which is the central trade-off captured in many decision-making applications in modern online service industries. Consequently, the design of stochastic linear bandit algorithms demands an easy-generalizable implementation across various contextualize actions and reward generation processes.

In the past decade of bandit literature, such demands have invited researchers to investigate bootstrap-based exploration-exploitation trade-offs and have drawn rising attention [Baransi et al., 2014, Eckles and Kaptein, 2014, Osband and Van Roy, 2015, Vaswani et al., 2018, Hao et al., 2019, Kveton et al., 2019b, Wang et al., 2020]. Yet, prior works on bootstrap-based bandit algorithms focus on provable multi-armed bandit algorithms and only provide a limited empirical evaluation of bootstrap-based stochastic linear bandit algorithms, and their theoretical counterpart remains unknown. Such knowledge gap of bootstrapping stochastic linear bandit persuades our investigation on the provable bootstrap-based stochastic linear bandits: **Can we theoretically and empirically support the validity and easy-generalizability of bootstrapping procedure in stochastic linear bandit algorithms design?** In particular, we aim to deliver a generic framework to demystify the bootstrap optimism in stochastic linear bandit problems and validate the easy generalizability of the bootstrap principle across various contextual linear bandit problems.

Contributions. We introduce `LinReBoot` algorithms that implement Residual Bootstrap Exploration for stochastic linear bandit problem with sub-linear regret. We theoretically show that `LinReBoot` secures $\tilde{O}(d\sqrt{n})$ regret where d is the dimension of features. This sub-linear regret bound matches the regret bound of the same order as those theoretical results of Linear Thompson Sampling algorithms. The key to achieving such sub-linear regret guarantee is to carefully manage and collaborate sample and bootstrap optimism (Section

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

4.1). In particular, by measuring the ”sample-bootstrap optimistic estimated discrepancy ratio” of the optimal arm, `LinReboot` successfully avoids over or under exploration and theoretically secures sub-linear mean regret with high-probability. To our knowledge, this is the first theoretical analysis to support the validity and efficiency of the residual bootstrap-based procedure for stochastic linear bandit problems. We empirically show that `LinReBoot` rivals or exceeds competing algorithms including Linear Thompson Sampling, Linear PHE, Linear GIRO, and Linear UCB under stochastic linear bandit problem as well as more complicated linear bandit settings. These significant results support the easy-generalizability of proposed `LinReBoot`. In summary, our contributions are as follows:

- Propose `LinReBoot` algorithms that implement Residual Bootstrap Exploration in linear bandit problems without boundness assumption of rewards.
- Theoretically show that `LinReBoot` secures $\tilde{O}(d\sqrt{n})$ regret, matching the regret bound of the same order as those theoretical results of Linear Thompson Sampling algorithms.
- Empirically show that `LinReBoot` rivals or exceeds baseline algorithms and supports that `LinReBoot` is easy-generalizable among linear bandit problems.

Related Works. Bootstrap-based contextual bandit algorithms design has been actively studied in the last half-decade and drawn a surge of interest from both theoretical studies and industrial practice [Elmachtoub et al., 2017, Eckles and Kaptein, 2014, Osband et al., 2016, Kveton et al., 2019b, Hao et al., 2019]. Bootstrap-based bandit algorithm design is a paradigm of sequential decision-making based on an exploration mechanism with no pre-defined mean reward model. Such paradigm enjoys a decisive advantage that engineers are free to deploy any reward model of interests without painful adaption to problem structure [Kveton et al., 2019b,a]. `ReBoot` [Wang et al., 2020] provided a theoretical logarithmic regret guarantee for multi-armed bandit (MAB) and empirical investigation to validate the easy generalizability of the `ReBoot` principle. Our work aims to provide a theoretical guarantee for the bootstrap-based linear bandit algorithms and empirically investigate more general contextual linear bandit setting to validate the `ReBoot` principle.

One close related work is [Kveton et al., 2020a] which introduces perturbation of past samples for exploration under stochastic linear bandit problem. The limitation of [Kveton et al., 2020a] is the boundness of rewards, indicating many broader classes of rewards such as Gaussian rewards are not applicable with a theoretical guarantee. In contrast, the proposed `LinReBoot` algorithms relax the boundness reward assumption and thus validate bootstrap-based bandit algorithms in wider bandit environments with a broader class of reward generation processes.

Early works about exploration in bandit problems [Abbasi-Yadkori et al., 2011, Langford and Zhang, 2007, Dani et al., 2008] are practical but no guarantee of the optimality. Some works [Wang et al., 2020, Kveton et al., 2019b,a, Thompson, 1933, Auer et al., 2002] provide well designed exploration for bandit problems and have their own principles for adopting to more general problems. In these works, three principles including `ReBoot`[Wang et al., 2020], `GIRO`[Kveton et al., 2019b] and `PHE`[Kveton et al., 2019a] are devising exploration mechanism based on up-to-now history instead of on pre-defined reward model in the other two principles `TS`[Thompson, 1933] and `UCB`[Auer et al., 2002]. Our work generalizes `ReBoot` into stochastic linear bandit problems.

Notations. Let $[n]$ be set $\{1, 2, ..., n\}$. $\mathbf{1}$ is a vector with all ones and $\boldsymbol{I}$ is the identity matrix. For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|_2$ is 2-norm of $\boldsymbol{v}$ and $\|\boldsymbol{v}\|_{\boldsymbol{A}}^2 := \sqrt{\boldsymbol{v}^\top \boldsymbol{A} \boldsymbol{v}}$ for a semidefinite matrix $\boldsymbol{A}$. Let $\langle \cdot, \cdot \rangle$ be the inner product operation. Denote $\mathcal{F}_t$ as the history of randomness up to round t. $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ is defined as the conditional expectation given $\mathcal{F}_{t-1}$ and $\mathbb{P}_t(\cdot) := \mathbb{P}(\cdot|\mathcal{F}_{t-1})$ is defined as the conditional probability given $\mathcal{F}_{t-1}$. $\mathbb{I}\{\cdot\}$ is indicator function. For a set or event E, we denote its complement as $\bar{E}$. $N(\mu, \sigma^2)$ is Gaussian distribution with mean μ and variance σ^2. We use $\tilde{O}$ for big O notation up to logarithmic factor.

2 STOCHASTIC LINEAR BANDIT

Contextualize Action Set. In stochastic linear bandit problem, we identify the actions with $d-$dimensional features from $\mathcal{A} \subset \mathbb{R}^d$ and assume $|\mathcal{A}|$, the size of the action set, is finite. Let $K := |\mathcal{A}|$ be the number of actions (arms), $\boldsymbol{x}_k \in \mathbb{R}^d$ be the context vector of the k-th arm, that is, $\mathcal{A} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_K\}$.

Reward generating mechanism. The reward function is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$ such that, at time t the agent chooses an action $I_t \in [K]$ with feature $X_t = \boldsymbol{x}_{I_t} \in \mathcal{A}$, the reward is generated by

$$Y_t \equiv \langle X_t, \boldsymbol{\theta} \rangle + \epsilon_t. \tag{1}$$

Specifically, the reward obtained by the agent at round t when pulling arm $I_t = k$ is generated from a distribution with mean $\mu_k := \boldsymbol{x}_k^\top \boldsymbol{\theta}$, conditioning on context $\boldsymbol{x}_k$. The property of noise ϵ_t is described in Assumption 2. Furthermore, denote the recieved reward by r_{I_t} and the reward random variable by Y_t at round t.

Regret. Without loss of generality, assume that arm 1 is the unique optimal arm, that is $\mu_1 > \mu_k \; \forall k \neq 1$.

The optimal gap of the k-th arm is $\Delta_k := \mu_1 - \mu_k \geq 0$. The expected n-round regret is denoted as

$$R_n := \sum_{k=2}^{K} \Delta_k \mathbb{E}[\sum_{t=1}^{n} \mathbb{I}\{I_t = k\}]. \tag{2}$$

The goal of the agent is to maximize the expected cumulative reward in n rounds, which is equivalent to minimizing the expected regret R_n.

Assumption 1. *(Boundness assumptions) True parameter $\boldsymbol{\theta}$ is bounded: $\|\boldsymbol{\theta}\|_2 \leq S_2$.*

Besides, we denote L as the upper bound for context vectors: $\|\boldsymbol{x}_k\|_2 \leq L$ for all $k \in [K]$. Assumption 1 is referred to the boundness assumptions in the stochastic linear bandit literature and is to ensure the regret is bounded if the agent pulls any sub-optimal actions (see Section 5 in [Abbasi-Yadkori et al., 2011]).

Assumption 2. *(Noise Clipping assumption) Noise process $\{\epsilon_t\}_{t=1}^{\infty}$ described in* (1) *satisfies that for some $L_1, L_2 > 0$,*

$$e^{L_1 \eta^2} \leq \mathbb{E}[e^{\eta \epsilon_t} | \mathcal{F}_{t-1}] \leq e^{L_2 \eta^2}, \forall \eta \geq 0, \tag{3}$$

where $\mathcal{F}_{t-1} = \{\epsilon_1, I_1, \cdots, \epsilon_{t-1}, I_{t-1}\}$.

Assumption 2 implies that stochastic process $\{\epsilon_t\}_{t=1}^{\infty}$ is conditionally sub-gaussian with constant L_2. L_1 contributes to the lower bound of moment generating function suggested by [Zhang and Zhou, 2020]. Note that the Assumption 2 allows heteroscedasticity among different arms by choosing L_2 as the largest variance among arms. Such heteroscedasticity consideration arises and has been identified as a challenge in applications of Bayesian optimization [Kirschner, 2021, Cowen-Rivers et al., 2020].

3 RESIDUAL BOOTSTRAP EXPLORATION

3.1 REBOOT PRINCIPLE

This section presents essential proof of concepts to implement ReBoot principle [Wang et al., 2020]. In general, each round of interaction, the decision policy admits four subroutines to implement ReBoot principle: 1) Learning, 2) Fitting, 3) Bootstrapping, and 4) Exploring. Following elaborates on each subroutine:

1) Model Learning. The first subroutine outputs a learned model based on current collected data. Our implementation learns the parameter $\boldsymbol{\theta}$ in Eq.(1) by some user-specified model.

2) Data Fitting. The second subroutine fits the current data set with the learned model in the previous subroutine and then outputs the residual set. Intuitively, the residuals measure the *goodness of fit* of the learned model and should drop a hint on the right amount of exploration. In other words, the residuals should suggest a right magnitude of exploration bonus in decision policy (8). How to manage and integrate uncertainty behind residuals into the exploration mechanism of policy is the main challenge.

3) Residuals Bootstraping. The third subroutine associates the residuals obtained the last subroutine with a bootstrapping distribution. Instead of maintaining a belief distribution on a parameter in the Bayesian approach, ReBoot principle maintains a bootstrapping distribution on the statistical error based on residuals. The challenge is to justify the efficacy of residual-based optimism construction in both theory and practice.

4) Actions Exploring. The fourth subroutines sample the exploration bonus from the bootstrapping distribution and output an index for each action. Such bootstrap procedure is more computationally efficient than prior efforts since this procedure only requires drawing a sample from the bootstrapping distribution. The challenge is to prove that such bootstrap procedure secures sub-linear regret in theory.

3.2 LINREBOOT ALGORITHM

We propose the Linear Residual Bootstrap Exploration algorithm (LinReBoot, Algorithm 1) for stochastic linear bandit problems. This section elaborates the four subroutines in Section 3.1 for the proposed LinReBoot.

1) LinReBoot uses ridge regression procedure, whose learned parameter is $\hat{\boldsymbol{\theta}}_t$ (4b) and estimated mean reward for arm k is $\hat{\mu}_{k,t}$ (4c). Such way to estimate mean reward is easy to manage the confidence [Abbasi-Yadkori et al., 2011]. Thus, we focus on confidence management for the bootstrap-based exploration.

Ridge Regression Procedure. LinReBoot fits linear model at round t as follow,

$$\boldsymbol{V}_t = \boldsymbol{X}_{t-1}^\top \boldsymbol{X}_{t-1} + \lambda \boldsymbol{I}, \tag{4a}$$

$$\hat{\boldsymbol{\theta}}_t = \boldsymbol{V}_t^{-1} \boldsymbol{X}_{t-1}^\top \boldsymbol{Y}_{t-1}, \tag{4b}$$

$$\hat{\mu}_{k,t} = \boldsymbol{x}_k^\top \hat{\boldsymbol{\theta}}_t, \forall k \in [K], \tag{4c}$$

where $\boldsymbol{X}_{t-1} = (X_1, ..., X_{t-1})^\top \in \mathbb{R}^{(t-1)\times d}$. The τ-th row of $\boldsymbol{X}_{t-1}$ is the context $X_\tau^\top$ for $\tau \in [t-1]$, $\boldsymbol{Y}_{t-1} = (Y_1, ..., Y_{t-1})^\top$ is reward vector whose elements are rewards up to round $t-1$. λ denotes the regularization level. $\boldsymbol{V}_t$ denotes the sample covariance matrix up to round t and $\hat{\boldsymbol{\theta}}_t$ is the ridge estimation of target parameter $\boldsymbol{\theta}$ in (1). $\hat{\mu}_{k,t}$ denotes the estimated mean of arm k based on history. Note that the first K rounds in proposed LinReBoot is fully exploring each arm once. In other words, $I_t = t$ when $t \in [K]$, indicating $\boldsymbol{X}_K := (\boldsymbol{x}_1, ..., \boldsymbol{x}_K)^\top \in \mathbb{R}^{K\times d}$. We call this $\boldsymbol{X}_K$ the context matrix with rank $r \leq \min(K, d)$ and singular values $\sigma_1, ..., \sigma_r$. Also define $\sigma^2_{\min} \leq \sigma^2_i \leq \sigma^2_{\max}, \forall i \in [r]$. With these definitions, we make a mild assumption about the shrinkage effect of ridge regression:

Assumption 3. *(Validity of Ridge Regression) The singular value decomposition of context matrix $\boldsymbol{X}_K$ is denoted as $\boldsymbol{X}_K := \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{U}$ where $\boldsymbol{G} \in \mathbb{R}^{K\times K}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{K\times d}$ and $\boldsymbol{U} \in \mathbb{R}^{d\times d}$. Define $\boldsymbol{\Omega} := \boldsymbol{\Sigma}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^\top \in \mathbb{R}^{K\times K}$ and $\boldsymbol{Z} := \boldsymbol{G}\boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{U} \in \mathbb{R}^{K\times d}$. Let $\boldsymbol{z}_1 \in \mathbb{R}^d$ be the first row of $\boldsymbol{Z}$. Given any $\lambda > 0$, there exists a corresponding positive scalar S_1 such that $|\boldsymbol{x}_1^\top\boldsymbol{\theta} - \boldsymbol{z}_1^\top\boldsymbol{\theta}| \geq S_1$ for the θ in* (1).

Remark 1. *Assumption 3 provides a lower bound of the absolute difference between true mean $\boldsymbol{x}_1^\top\boldsymbol{\theta}$ and normalized mean $\boldsymbol{z}_1^\top\boldsymbol{\theta}$ of the optimal arm. Note that if $\lambda \to 0$, then $\boldsymbol{z}_1 \to \boldsymbol{x}_1$ and $S_1 \to 0$. Thus this scalar S_1 measures the small perturbation on the mean of the optimal arm when the ridge regression procedure is applied. This $\boldsymbol{Z}$ can be interpreted as a ridge shrinkage context matrix [Goldstein and Smith, 1974]. One important phenomenon of online ridge regression is that even if the ridge estimator is biased, the shrinkage effect from ridge estimation provides exploration for the agent leading to making a correct decision. The positive scalar S_1 describes the shrinkage effect on the context. That is, the existence of S_1 indicates the ridge procedure is valid and its shrinkage effect exists.*

2) The fitting part of `LinReBoot` outputs the residuals under the linear model framework,

$$e_{k,t,i} = r_{k,i} - \hat{\mu}_{k,t}, \forall i \in [s_{k,t-1}], \tag{5}$$

where $s_{k,t-1} := \sum_{\tau=1}^{t-1} \mathbb{I}\{I_\tau = k\}$ is the number of times pulling arm k by round $t-1$, $r_{k,i}$ is the i-th reward of arm k by round $t-1$. The *goodness of fit* of the learned ridge regression model can be summarised by Residual Sum of Squares(RSS) [Archdeacon, 1994] which is defined as

$$RSS_{k,t} := \sum_{i=1}^{s_{k,t-1}} e^2_{k,t,i}. \tag{6}$$

Such measure plays an important role in the residual bootstrap exploration mechanism.

3) The third part is Residuals Bootstrapping. This subroutine is independent of the model which suggests the power of generalizability of `ReBoot` principle. `ReBoot` principle requires the computation of the exploration bonus [Mammen, 1993], which is $s^{-1}_{k,t-1}\sum_{i=1}^{s_{k,t-1}} \omega_{k,t,i}e_{k,t,i}$, where $\{\omega_{k,t,i}\}_{i=1}^{s_{k,t-1}}$ is residual bootstrap weights for arm k at round t.

Algorithm 1 `LinReBoot`

```
Require: $\lambda$, $s_{1,0} = ... = s_{K,0} = 0$
  for $t = 1, ..., n$ do
    if $t < K + 1$ then
      $I_t \leftarrow t$
    else
      $\boldsymbol{V}_t \leftarrow \boldsymbol{X}_{t-1}^\top\boldsymbol{X}_{t-1} + \lambda\boldsymbol{I}$
      $\hat{\boldsymbol{\theta}}_t \leftarrow \boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top\boldsymbol{Y}_{t-1}$
      for $k = 1, ..., K$ do
        $e_{k,t,i} \leftarrow r_{k,i} - \boldsymbol{x}_k^\top\hat{\boldsymbol{\theta}}_t, \forall i \in \{s_{k,t-1}\}$
        Generate $\{\omega_{k,t,i}\}_{i=1}^{s_{k,t-1}}$
        $\tilde{\mu}_k \leftarrow \boldsymbol{x}_k^\top\hat{\boldsymbol{\theta}}_t + s^{-1}_{k,t-1}\sum_{i=1}^{s_{k,t-1}} \omega_{k,t,i}e_{k,t,i}$
      end for
      $I_t \leftarrow \arg\max_{k\in[K]} \tilde{\mu}_k$
    end if
    $s_{I_t,t} \leftarrow s_{I_t,t-1} + 1$ and $s_{k,t} \leftarrow s_{k,t-1}$. $\forall k \neq I_t$
    Pull arm $I_t$ and get reward $r_{I_t,s_{I_t}}$
    $\boldsymbol{X}_t \leftarrow \begin{bmatrix}\boldsymbol{X}_{t-1}\\ \boldsymbol{x}_{I_t}^\top\end{bmatrix}$ and $\boldsymbol{Y}_t \leftarrow \begin{bmatrix}\boldsymbol{Y}_{t-1}\\ r_{I_t,s_{I_t}}\end{bmatrix}$
  end for
```

Choice of Bootstrapping Weights. The bootstrap weights considered in this work are i.i.d with zero mean and variance σ^2_ω. They are independent of the noise process $\{\epsilon_t\}_{t=1}^\infty$. In the literature of bootstrap procedure [Mammen, 1993] , the choices of bootstrap weights distribution include Gaussian weights, Rademacher weights and skew correcting weights. In `LinReBoot`, we adopt the Gaussian bootstrap weights to enable an efficient implement described at section 3.3.

4) The last subroutine is the action exploring based on residual bootstrap. More specifically, for arm k at round t, `LinReBoot` adds exploration bonus from residual bootstrapping on the estimated mean $\hat{\mu}_{k,t}$ as follow,

$$\tilde{\mu}_{k,t} = \hat{\mu}_{k,t} + \frac{1}{s_{k,t-1}}\sum_{i=1}^{s_{k,t-1}} \omega_{k,t,i}e_{k,t,i}, \tag{7}$$

then agent pulls arm with the highest bootstrapped mean,

$$I_t \equiv \arg\max_{k\in[K]} \tilde{\mu}_{k,t}. \tag{8}$$

Note that the variance of bootstrapped mean $\tilde{\mu}_{k,t}$ is $\sigma^2_\omega s^{-2}_{k,t-1} RSS_{k,t}$, indicating an adaptive amount of extra exploration is controlled by $s_{k,t-1}$ and $RSS_{k,t}$.

Short Summary. Our proposed `LinReBoot` has following steps at round $t > K$,

1) Ridge estimation: compute $\boldsymbol{V}_t$, $\hat{\boldsymbol{\theta}}_t$.

2) Finding residuals for each arm: for arm k, compute $\hat{\mu}_{k,t}$ and $\{e_{k,t,i}\}_{i=1}^{s_{k,t-1}}$.

3) Compute Bootstrapped mean for each arm: for arm k, generate $\{\omega_{k,t,i}\}_{i=1}^{s_{k,t-1}}$ and compute $\tilde{\mu}_{k,t}$ (7).

4) Pull arm with the highest $\tilde{\mu}_{k,t}$ then observe reward.

Algorithm 1 describes `LinReBoot`. The strength of `LinReBoot` is its easy generalizability across different bandit problems including linear bandits and even more complicated structured problems (Appendix D.1).

Remark 2. *(LinTS perturbs system parameter estimate, LinReBoot perturbs expected reward estimates) Compare with the LinTS in [Agrawal and Goyal, 2013b], in which LinTS samples a perturbed parameter $\tilde{\boldsymbol{\theta}}_t^{LinTS} = \hat{\boldsymbol{\theta}}_t + \beta_t \boldsymbol{V}_t^{-1/2} \boldsymbol{\eta}_t$ with scaling β_t and appropriate independent noise $\boldsymbol{\eta}_t$ (defined in [Agrawal and Goyal, 2013b]). Our proposed LinReBoot samples a perturbed expected reward $\tilde{\mu}_{k,t}^{LinReBoot} = \langle \hat{\boldsymbol{\theta}}_t, \boldsymbol{x}_k \rangle + \frac{1}{s_{k,t-1}} \sum_{i=1}^{s_{k,t-1}} w_{k,t,i} e_{k,t,i}$. That is, LinReBoot is perturbing the expected reward estimate via prediction error uncertainty, which is supervised by real reward. In contrast, LinTS is perturbing the system parameter, when can be wrong if the system modeling is wrong.*

3.3 EFFICIENT IMPLEMENTATION

By the attractive computational properties of Gaussian distribution, the computational cost of `LinReBoot` can be reduced significantly when Gaussian Bootstrap weights are generated. Formally: assume $\omega_{k,t,i} \sim N(0, \sigma_\omega^2)$, $\forall k, t, i$, recalling (7), for $k \in [K]$ and any $t \geq 1$, bootstrapped mean $\tilde{\mu}_{k,t}$ follows a Gaussian distribution,

$$\tilde{\mu}_{k,t} | \mathcal{F}_{t-1} \sim N(\hat{\mu}_{k,t}, \sigma_\omega^2 s_{k,t-1}^{-2} RSS_{k,t}). \tag{9}$$

Such Gaussian-distributed property of $\tilde{\mu}_{k,t}$ indicates that if we can update $\hat{\mu}_{k,t}$, $s_{k,t-1}$ and $RSS_{k,t}$ incrementally for arm k, this bootstrapped mean $\tilde{\mu}_{k,t}$ can be generated by Gaussian generator without inner loop for generating weights. The first two terms, $\hat{\mu}_{k,t}$ and $s_{k,t-1}$, are naturally updated in incremental manner. For $RSS_{k,t}$, following decomposition ensures an incremental update,

$$RSS_{k,t} = \sum_{i=1}^{s_{k,t-1}} r_{k,i}^2 + s_{k,t-1} \hat{\mu}_{k,t}^2 - 2\hat{\mu}_{k,t} \sum_{i=1}^{s_{k,t-1}} r_{k,i}.$$

Then an efficient generation for $\tilde{\mu}_{k,t} | \mathcal{F}_{t-1}$ is ensured by the incremental updates for $\hat{\mu}_{k,t}$, $s_{k,t-1}$, $\sum_{i=1}^{s_{k,t-1}} r_{k,i}^2$, $\sum_{i=1}^{s_{k,t-1}} r_{k,i}$. Furthermore, since the residual bootstrap weights are generated independently, $\tilde{\mu}_{k,t}$ among arms are also independent given historical randomness and can be sampled from one multivariate Gaussian generation simultaneously. Formally, $\tilde{\boldsymbol{\mu}}^{(t)} = (\tilde{\mu}_{1,t}, \ldots, \tilde{\mu}_{K,t})^\top$ is conditional distributed as

$$\tilde{\boldsymbol{\mu}}^{(t)} | \mathcal{F}_{t-1} \sim N_K(\hat{\boldsymbol{\mu}}^{(t)}, \boldsymbol{\Sigma}_\omega^{(t)}), \tag{10}$$

where $\hat{\boldsymbol{\mu}}^{(t)} = (\hat{\mu}_{1,t}, \ldots, \hat{\mu}_{K,t})^\top$ and $\boldsymbol{\Sigma}_\omega^{(t)}$ is a diagonal matrix with diagonal elements $\sigma_\omega^2 s_{k,t-1}^{-2} RSS_{k,t}$. Detailed steps and more illustration about efficient implementation is provided in Appendix D.7.1. Moreover, an empirical study about computational efficiency is conducted in Appendix D.7.2 and Table.3 provides the computational cost of our proposed `LinReBoot` as well as other baseline algorithms.

4 OPTIMISM DESIGN

Optimistic Estimated Discrepancy. This section identifies and demystifies the technical challenge of implementing `ReBoot` principle in the stochastic linear bandit problem. The key is to conduct a detailed investigation to produce probabilistic control on the behavior of the '**O**ptimistic **E**stimate **D**iscrepancy (**OED**)' of the `LinReBoot` policy (8). In principle, the **OED** is given by

$$\textbf{OED} = \text{Optimism} \times \texttt{Action Context Norm}, \tag{11}$$

where the `Action Context Norm` is given by $\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}$ and Optimism is given by $c_{t,k}$ for the kth action at time t, defined in (14). Design of $c_{t,k}$ will be elaborated in Section 4.1.

Sufficient Explored Arms. We define the concept of *Sufficient Explore Arms* to facilitate the formal regret analysis of `LinReBoot`. Intuitively, an arm is *sufficient explored* if its index produced by the policy (8) is less than the mean reward of the optimal arm. Technically, we say an arm k is *sufficiently explored* at time t if the adopted OED ($c_{t,k}\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}$) is bounded by its optimal gap (Δ_k).

The above notion of sufficient explored arm defines the concept of "set of sufficient explored arms" $\mathcal{S}_t$, formally

$$\mathcal{S}_t := \{k \in [K] : c_{t,k}\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}} < \Delta_k\}, \tag{12}$$

where and $c_{t,k}$ is the collaborated optimism and $c_{t,k}\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}$ is an optimistic estimate of discrepancy of policy index (8).

The key consequence of set (12) is that, any member in $\mathcal{S}_t$ enjoys the property

$$\forall j \in \mathcal{S}_t \cap [K] : \tilde{\mu}_{j,t} < \mu_1; \tag{13}$$

that is, the `LinReBoot` policy always avoids an index (8) from sufficiently explored subset such that the bootstrapped mean of this index is less than the optimal mean reward unless all arm are sufficiently explored. (see equation (82) in the proof of Lemma A.1 at section B.1 for technical details).

4.1 COLLABORATE OPTIMISM

Here we elaborate on the collaborated optimism adopted in the definition of sufficient explored arms (12). Concretely, the collaborated optimism has a form

$$c_{t,k} = c_1(t,k) + c_2(t,k), \tag{14}$$

where $c_1(t,k)$ is called *sample optimism* and $c_2(t,k)$ is called *bootstrap optimism* for arm k at time t.

Sample Optimism. The sample optimism $c_1(t,k)$ serves as a control on the event that "the realized sample estimate discrepancy (ED) is bounded by sample OED":

$$E_{t,k} := \{|\hat{\mu}_{k,t} - \mu_k| \leq c_1(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}, \} \tag{15a}$$

$$E_t := \bigcap_{k=1}^{K} E_{t,k}, \tag{15b}$$

where $c_1(t,k)$ is a constant which can be tuned by our `LinReBoot` algorithm, making the bad event $\bar{E}_{t,k}$ and $\bar{E}$ become unlikely. In fact, this $E_{t,k}$ is the event that the least squared estimation is "close" to the true mean reward for arm k at round t. In section 5, the probability of the bad event $\bar{E}_t$ is controlled by a parameter tuned by users based on lemma 5.1.

Bootstrap Optimism.

The bootstrap optimism $c_2(t,k)$ serves as a control on the event that "the realized bootstrap ED is bounded by bootstrap OED":

$$E'_{t,k} := \{|\tilde{\mu}_{k,t} - \hat{\mu}_{k,t}| \leq c_2(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}\}, \tag{16a}$$

$$E'_t := \bigcap_{k=1}^{K} E'_{t,k}, \tag{16b}$$

where $c_2(t,k)$ is also a constant controlling the conditional probability of the bad event $\bar{E}'_t$. This $c_2(t,k)$ can be tuned by our `LinReBoot` algorithm as well. Similar to $E_{t,k}$, this $E'_{t,k}$ is the event that the residual bootstrap based estimation is "close" to the least squared estimate $\hat{\mu}_{k,t}$ for arm k at round t. In section 5, the probability of bad event $\bar{E}'_t$ is controlled by a parameter tuned by users based on lemma 5.2.

4.2 OPTIMISM DESIGN

Choice of sample optimism (α). The goal of this part is to illustrate how to pick the sample OED such that the event (15) holds with probability at least $1-\alpha$ for a given confidence budget $\alpha \in (0,1)$. Formally, the goal is to find a sample OED function $c_1(t,k)$: $[n] \times [K] \mapsto \mathbb{R}$ such that the event (15a) holds with probability at least $1-\alpha_k$. To meet the purpose of the risk control, we specify the sample OED function with form

$$c_1(t,k) := R_2\sqrt{d\log((1+tL^2/\lambda)/\alpha_k)} + \lambda^{1/2}S_2. \tag{17}$$

Lemma 5.1 gives the formal result on why such choice has confidence budget at most α_k. For regret analysis, define $\alpha_{\min} = \min_{k\in[K]} \alpha_k$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)^\top$.

Choice of bootstrap optimism (β). The goal of this part is to pick bootstrapped OED such that the event (16) holds with probability at least $1-\beta$ for given confidence budget $\beta \in (0,1)$. Formally, the goal is to find a sample OED function $c_2(t,k) : [n] \times [K] \mapsto \mathbb{R}$ such that the event (16a) holds with probability at least $1-\beta_k$. To meet the purpose of the risk control, we specify the bootstrapped OED function with form

$$c_2(t,k) := \sqrt{(2\sigma_\omega^2 RSS_{k,t}\log(2/\beta_k))/s^2_{k,t-1}\|\boldsymbol{x}_k\|^2_{\boldsymbol{V}_t^{-1}}}. \tag{18}$$

Lemma 5.2 gives the formal result on why such choice has a confidence budget at most β_k. For regret analysis, let $\beta_{\min}$ be the smallest β_k, $\forall k \in [K]$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)^\top$.

4.3 OPTIMISM FOR OPTIMAL ARM

Sample-Bootstrap OED ratio of the optimal arm (b). Indicated by the regret analysis in [Kveton et al., 2020a], instead of controlling the exploration independently, the relation between two sources of explorations needs to be considered because this relation is critical for finding the optimal action. To meet such observation, we define a good event,

$$E''_t := \{\tilde{\mu}_{1,t} - \hat{\mu}_{1,t} > c_1(t,1)\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}\}. \tag{19}$$

Given the good event E''_t, the policy index $\tilde{\mu}_{1,t}$ of the optimal arm enjoys further positive bias, hence the agent will have better chance to make optimal action.

In particular, we highlight a constant b used to measure the ratio of the sample optimism (17) to the bootstrap optimism (18); formally, we require b satisfies

$$c_1(t,1)/c_2(t,1) \geq b \cdot \sqrt{2\log{(2/\beta_1)}}. \tag{20}$$

Intuitively, the constant b measures the relation between sample OED and bootstrap OED of the optimal arm. This b plays an important role of the probability lower bound of event (19) (See Lemma 5.3). Note that, if (20) holds, we have the lower bound (26) ; otherwise, we have the lower bound (27). In both cases, we have a lower bound for the event (19).

Notation	Definition
$\zeta_1(n,d)$	$(L_2\sqrt{d\log\left(\frac{1+nL^2/\lambda}{\alpha_{\min}}\right)}+\lambda^{1/2}S_2)\times$ $\sqrt{2(n-K)d\log(1+\sum_{i=1}^{r}\sigma_i^2/d\lambda)}$
$\zeta_2(n,d)$	$\sqrt{2\sigma_\omega^2 log(\frac{2}{\beta_{\min}})}\times$ $\sqrt{2(n-K)d\log(1+\sum_{i=1}^{r}\sigma_i^2/d\lambda)}$
$\zeta_3(n)$	$2K\sqrt{4L_2\sigma_\omega^2\log\left(\frac{2}{\beta_{\min}}\right)}(\log n+1)$
$\zeta_4(n)$	$2S_2L((n-K)(\alpha+\beta)+K-1)$

Table 1: Notations in Regret Analysis

Good event for optimal arm (γ). Here we introduce the event that over exploration and under exploration of the optimal arm have been avoided simultaneously. Formally, the constant γ is the probability that the bandit index (8) is not over-exploration (Event E_t') and also not under-exploration (Event E_t'')

$$\{c_1(t,1)<(\tilde{\mu}_{1,t}-\hat{\mu}_{1,t})/\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}<c_2(t,1)\}. \quad (21)$$

Technically, we can show that the probability of the event (21) is lower bounded by the term

$$\mathbb{P}_t(E_t'')-\mathbb{P}_t(\bar{E}_t'), \quad (22)$$

with probability at least $1-\gamma$ (Lemma 5.4). Such lower bound is translated into an upper bound in regret analysis.

5 FORMAL RESULTS

5.1 REGRET BOUND FOR LINREBOOT

Theorem 5.1. *Under Assumptions 1, 2, 3 and technical conditions* (32) *and* (74)*, with probability at least* $1-(\delta+\gamma)$*, the expected regret of Algorithm 1 is bounded as,*

$$\begin{aligned}R_n \leq & C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\zeta_1(n,d)\\ &+C_2(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma,b,\delta)\zeta_2(n,d)\\ &+C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\zeta_3(n)+\zeta_4(n),\end{aligned} \quad (23)$$

where ζ_1, ζ_2, ζ_3 and ζ_4 are defined in Table.1 and C_1, C_2, M_1, M_2 are described in Table.2.

Proof. See Appendix A.1. □

Corollary 5.2. *Let $\boldsymbol{\alpha}=\boldsymbol{\beta}=\frac{1}{\sqrt{n}}\mathbf{1}$, the order of high probability upper bound in Theorem 5.1 is $\tilde{O}(d\sqrt{n})$.*

Proof. See Appendix A.2. □

Corollary 5.2 shows that our regret bound scales as the regret bound of Linear Thompson sampling [Agrawal and Goyal, 2013b] and Linear PHE [Kveton et al., 2020a].

5.2 VALIDATE SAMPLE OPTIMISM

Lemma 5.1. *Under Assumptions 1, 2, 3 and choose $c_1(t,k)$ as* (17)*, $\mathbb{P}(\bar{E}_{t,k})$, the probability of bad event corresponded to least squared estimation described in (15), is controlled. Formally, $\forall k\in[K]$, $\forall\alpha_k>0$, $\forall t\geq 1$,*

$$\mathbb{P}(|\hat{\mu}_{k,t}-\mu_k|\leq c_1(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}})\geq 1-\alpha_k. \quad (24)$$

Consequently, we have $\mathbb{P}(\bar{E}_t)\leq\alpha:=\sum_{k=1}^K\alpha_k$.

Proof. See Appendix A.3. □

Lemma 5.1 supports that the choice of $c_1(t,k)$ at (17) for the sample optimism event (15) is valid with confidence budget α.

5.3 VALIDATE BOOTSTRAP OPTIMISM

Lemma 5.2. *Suppose bootstrap weights are Gaussian. Pick $c_2(t,k)$ as* (18)*. The conditional probability of bad event corresponding to residual bootstrap exploration described in* (16)*, $\mathbb{P}_t(\bar{E}_{t,k}')$, is controlled. Formally, $\forall k\in[K]$, $\forall\beta_k>0$, $\forall t\geq 1$*

$$\mathbb{P}_t(|\tilde{\mu}_{k,t}-\hat{\mu}_{k,t}|\leq c_2(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}})\geq 1-\beta_k. \quad (25)$$

Consequently, we have $\mathbb{P}_t(\bar{E}_t')\leq\beta:=\sum_{k=1}^K\beta_k$.

Proof. See Appendix A.4. □

Lemma 5.2 supports that the choice of $c_2(t,k)$ at (18) for the sample optimism event (16) is valid with confidence budget β.

5.4 SAMPLE-BOOTSTRAP RATIO

Lemma 5.3. *Under Assumptions 1, 2, 3. Suppose bootstrap weights are Gaussian. The conditional probability of anti-concentration for optimal arm described in*

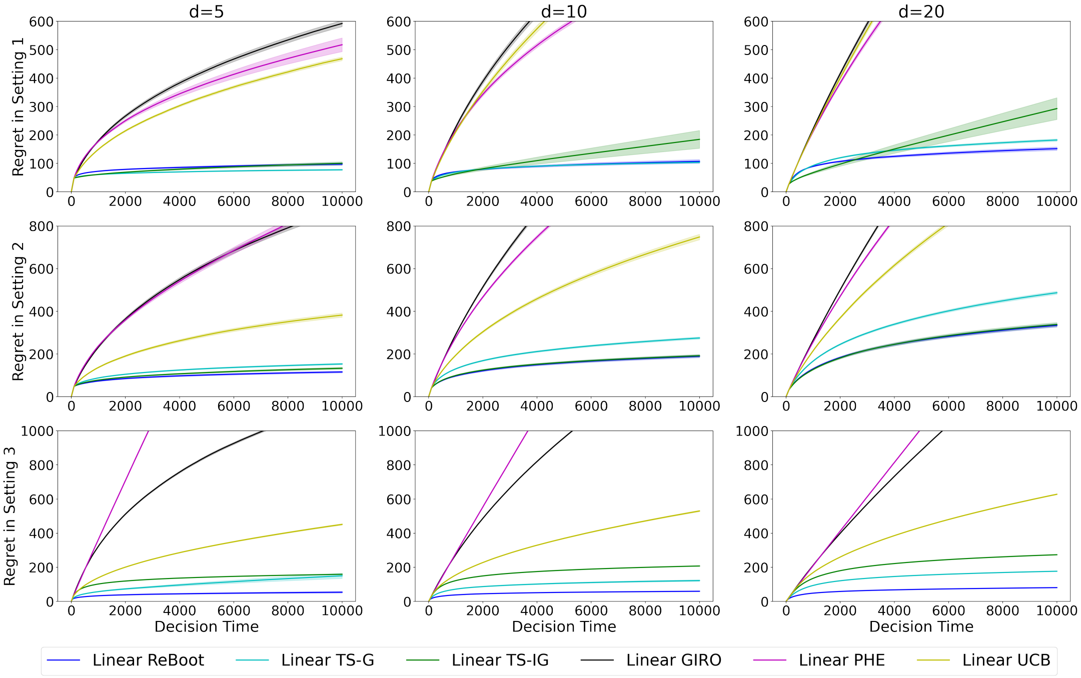

Figure 1: Comparison of `LinReBoot` with Gaussian Bootstrap weights to baselines under three linear bandit problems and three different context dimension d. First row referred to the setting in Section 6.1, second row is for Section 6.2 and the last row is for Section 6.3. Three columns refer to $d = 5$, $d = 10$ and $d = 20$ respectively.

(19), $\mathbb{P}_t(\bar{E}_t'')$, has lower bound. Formally, if b satisfies (20),

$$\mathbb{P}_t(E_t'') \geq \frac{b}{\sqrt{2\pi}} \exp\left(-\frac{3c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\mathbf{V}_t^{-1}}}{2\sigma_\omega^2 RSS_{1,t}}\right). \tag{26}$$

Otherwise,

$$\mathbb{P}_t(E_t'') \geq \Phi(-b), \tag{27}$$

where Φ is the CDF of standard normal distribution.

Proof. See Appendix A.5. □

Lemma 5.3 provides the lower bound result for good event E_t''. The result indicates that, if the bootstrap optimism is not 'too large', then the `LinReBoot` procedure can enjoy additional regret reduction.

5.5 VALIDATE GOOD EVENT

Lemma 5.4. *Under Assumptions 1, 2, 3 and suppose Bootstrap weights are Gaussian. Assume b satisfies a technical condition* (74)*. Then, with probability at least $1-\gamma$, $\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t')$ has lower bound,*

$$\frac{b}{\sqrt{2\pi}} \exp\left(-\frac{3s_{1,t-1}^{3/2}c_1^2(t,1)\|\boldsymbol{x}_1\|_2^2}{8\sigma_\omega^2(\sigma_{\min}^2+\lambda)\sqrt{\frac{1}{M_2}\log\left(\frac{M_1}{1-\gamma}\right)}}\right) - \beta, \tag{28}$$

where M_1 and M_2 are defined in Table.2.

Proof. See Appendix A.6. □

Lemma 5.4 provided the a high probability lower bound for the difference between probability of the event for anti-concentration E_t'' and probability of bad event discussed in bootstrap optimism in Section 4.1. This lower bound is also for probability of 'not under and not over exploration' event (21). Lemma 5.4 links the sample optimism and bootstrap optimism and holds a right amount of exploration of the optimal arm.

6 EXPERIMENTS

In this section, we conduct empirical studies under three settings: Stochastic Linear Bandit, Contextual

Linear Bandit and Linear Bandit with Covariates. Our `LinReBoot` is compared to several baselines including `LinTS-G` [Agrawal and Goyal, 2013b, Lattimore and Szepesvári, 2020], `LinTS-IG` [Honda and Takemura, 2014, Riquelme et al., 2018], `LinPHE` [Kveton et al., 2020a], `LinGIRO` [Kveton et al., 2019b] and `LinUCB` [Abbasi-Yadkori et al., 2011, Lattimore and Szepesvári, 2020] . More details about baselines can be found in Appendix D.6.

6.1 STOCHASTIC LINEAR BANDIT

We compare `LinReBoot` to other linear bandit algorithms under stochastic linear bandit described in Section 2. We experiment with several dimensions d including 5, 10 and 20. K is chosen as 100. Synthetic data generation for this setting is deferred to Appendix D.2 in the supplementary material. **Results.** The first row of Figure 1 reports the results for Stochastic Linear Bandit setting. Our `LinReBoot` rivals `LinTS-G` and `LinTS-IG` while substantially exceeds `LinGIRO`, `LinPHE` and `LinUCB`. When d increases, the performance of `LinReBoot` rivals and exceeds the best of other methods.

6.2 CONTEXTUAL LINEAR BANDIT

In the second experiment, we compare `LinReBoot` to other linear bandit algorithms under Contextual Linear Bandit where the contexts are generated from some distributions by arms. Note that this setting matches previous work [Chu et al., 2011]. Linear bandit algorithms can also be applied under this kind of environment. In our experiment, the `LinReBoot` is implemented as Algorithm 2 in Appendix D.1. Like the setting in Section 6.1, the dimension of d is chosen as 5 or 10 or 20 and the synthetic data generation for this setting is described in Appendix D.2. **Results.** The second row of Figure 1 reports the results for Contextual Linear Bandit. Our `LinReBoot` rival `LinTS-G` and substantially exceed `LinTS-IG`, `LinGIRO`, `LinPHE` and `LinUCB`. When d increases, the performance of `LinReBoot` rivals `LinTS-IG` and exceeds others.

6.3 BANDIT WITH COVARIATES

Our last experiment is conducted under the setting of linear bandit with covariates, which is also called linear parametrized bandit by [Rusmevichientong and Tsitsiklis, 2010]. This problem is significantly different from the previous two problems in the following ways. Each arm has its true parameter $\boldsymbol{\theta}_k$. That is, each arm has its estimate $\hat{\boldsymbol{\theta}}_k$ from the ridge regression procedure in Section 3.2. Also, unlike the setting in Section 6.2, the contexts are generated from a distribution that is independent of arms. Thus the overall task in this setting is not only the estimation of the target parameter $\boldsymbol{\theta}$, but also the detection of which arm a context belongs to. This case is also referred to as the online decision-making under covariates [Bastani and Bayati, 2020]. For the `LinReBoot` in this setting, detailed algorithm is provided as Algorithm 3 in Appendix D.1. d is chosen as 5 or 10 or 20 and $K = 10$. Synthetic data generation for this setting is described in Appendix D.2. **Results.** The third row of Figure 1 reports the results for Linear Bandit with Covariates. Our `LinReBoot` exceeds all competing algorithms `LinTS-G`, `LinTS-IG`, `LinGIRO`, `LinPHE` and `LinUCB`.

Summary. From Figure 1, the proposed `LinReBoot` is always the top 3 algorithms under all settings and all choice of dimension d. More specifically, `LinReBoot` is clearly comparable to the state-of-the-art Linear Thompson Sampling algorithms(`LinTS-G`, `LinTS-IG`) or even outperforms them in many cases. Regarding the computational cost, from Table.3, our proposed `LinReBoot` is consistently computational efficient among all settings compared to `LinTS-G`, `LinTS-IG` and `LinUCB` under all three settings.

7 CONCLUSION

We propose `LinReBoot` algorithm for stochastic linear bandit problems. In theory, we prove `LinReBoot` that secures $\tilde{O}(d\sqrt{n})$ high probability expected regret. Empirically, we show `LinReBoot` rivals `LinTS-G`, `LinTS-IG` and exceeds `LinPHE`, `LinGIRO` and `LinUCB`, which supports the easy-generalizability of `ReBoot` principle in [Wang et al., 2020] under various contextual bandit settings including Stochastic Linear Bandit, Contextual Linear Bandit, and Linear Bandit with Covariates.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, Nitin Motgi, Seung-Taek Park, Raghu Ramakrishnan, Scott Roy, and Joe Zachariah. Online models for content optimization. In *Advances in Neural Information Processing Systems*, pages 17–24, 2009.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107. PMLR, 2013a.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013b.

Thomas J Archdeacon. *Correlation and regression analysis: a historian's guide.* Univ of Wisconsin Press, 1994.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–131. Springer, 2014.

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

Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Alexander I Cowen-Rivers, Wenlong Lyu, Rasul Tutunov, Zhi Wang, Antoine Grosnit, Ryan Rhys Griffiths, Hao Jianye, Jun Wang, and Haitham Bou Ammar. An empirical study of assumptions in bayesian optimisation. *arXiv preprint arXiv:2012.03826*, 2020.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Dean Eckles and Maurits Kaptein. Thompson sampling with the online bootstrap. *arXiv preprint arXiv:1410.4009*, 2014.

Adam N Elmachtoub, Ryan McNellis, Sechan Oh, and Marek Petrik. A practical method for solving contextual bandit problems using decision trees. *arXiv preprint arXiv:1706.04687*, 2017.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

M Goldstein and Adrian FM Smith. Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2): 284–291, 1974.

Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng. Bootstrapping upper confidence bound. *Advances in neural information processing systems*, 32, 2019.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

Junya Honda and Akimichi Takemura. Optimality of thompson sampling for gaussian bandits depends on priors. In *Artificial Intelligence and Statistics*, pages 375–383. PMLR, 2014.

Jean Jacod and Albert Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media, 2013.

Johannes Kirschner. *Information-Directed Sampling-Frequentist Analysis and Applications.* PhD thesis, ETH Zurich, 2021.

Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic multi-armed bandits. *arXiv preprint arXiv:1902.10089*, 2019a.

Branislav Kveton, Csaba Szepesvari, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 3601–3610. PMLR, 2019b.

Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In *Uncertainty in Artificial Intelligence*, pages 530–540. PMLR, 2020a.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020b.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20 (1):96–1, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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

Enno Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The annals of statistics*, pages 255–285, 1993.

Ian Osband and Benjamin Van Roy. Bootstrapped thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*, 2015.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29:4026–4034, 2016.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations*, 2018.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 323–332, 2015.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Sharan Vaswani, Branislav Kveton, Zheng Wen, Anup Rao, Mark Schmidt, and Yasin Abbasi-Yadkori. New insights into bootstrapping for bandits. *arXiv preprint arXiv:1805.09793*, 2018.

Chi-Hua Wang, Yang Yu, Botao Hao, and Guang Cheng. Residual bootstrap exploration for bandit algorithms. *arXiv preprint arXiv:2002.08436*, 2020.

Anru R Zhang and Yuchen Zhou. On the non-asymptotic and sharp lower tail bounds of random variables. *Stat*, 9(1):e314, 2020.

Residual Bootstrap Exploration for Stochastic Linear Bandit (Supplementary Materials)

Shuang Wu[1] **Chi-Hua Wang**[1,2] **Yuantong Li**[1] **Guang Cheng**[1]

[1]Department of Statistics, University of California, Los Angeles, Los Angeles, California, USA
[2]Department of Statistics, Purdue University, West Lafayette, Indiana, USA

A PROOFS OF MAIN RESULTS

A.1 PROOF OF THEOREM 5.1

Proof. The regret bound analysis of algorithm 1 involves several key Lemmas and conditions. Inspired by the definition of expected regret, one key Lemma is providing the upper bound for expected optimal gap given the history $\mathcal{F}_{t-1}$ at round t, $\mathbb{E}_t[\Delta_{I_t}]$. This is similar to the proof in other linear bandit algorithms such as `LinPHE` [Kveton et al., 2020a] and `LinUCB` [Abbasi-Yadkori et al., 2011]. Lemma A.1 in the following part gives this result. The other important Lemma is bounding sum of expected 'square root of normalized RSS' which is described in Lemma A.2. The Third key result, Lemma A.3, is an algebra result from [Abbasi-Yadkori et al., 2011] which bounds the sum of action context norms. Moreover, Lemmas in Section 5 play essential roles in regret bound analysis. Lemma 5.1 and Lemma 5.2 control the sample optimism and bootstrap optimism respectively. Lemma 5.3 gives lower bound for the event of anti-concentration, which is necessary lower bound for analyzing exploration in linear bandit algorithms. Another key step is carefully evaluating anti-concentration and its connection to concentration, which is summarised by lemma 5.4. An technical condition about tuning parameter σ_ω^2, which will be discussed later in this proof is also needed for regret analysis. We start from listing the Lemmas and condition and main proof of Theorem 5.1 will be given later.

Lemma A.1. *Assume the same as Theorem 5.1. Suppose $M \geq \max_{k\in[K]} \Delta_k$. When $c_1(t,k), c_2(t,k) \geq 1$ and $\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t') > 0$ for $\forall t > K$ and $\forall k \in [K]$, then on event E_t, almost surely,*

$$\mathbb{E}_t[\Delta_{I_t}] \leq (\frac{2}{\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t')} + 1)(c_1(t,I_t) + c_2(t,I_t))\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}] + M\mathbb{P}(\bar{E}_t') \tag{29}$$

Proof. See appendix B.1 □

Remark 3. *Lemma A.1 provides the upper bound for expected optimal gap given the latest history. This result directly impacts the upper bound of expected regret of `LinReBoot`, which means that each terms in the upper bound given by Lemma A.1 need to be further bounded. As we expect, sample optimism $(c_1(t,I_t)\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}})$ and Bootstrap optimism $(c_2(t,I_t)\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}})$ require further bounding. An interesting observation is the appearance of term $\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t')$ which is the lower bound of probability of E_t'' defined in (21). Intuitively, this event connects the exploration from ridge estimation and the exploration from residual Bootstrapping and iF the lower bound $\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t')$ is too small, then this upper bound in Lemma A.1 becomes trivial, which means our regret analysis become meaningless.*

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Lemma A.2. *Assume the same as Theorem 5.1. With probability at least* $1-\delta$*,*

$$\sum_{t=K+1}^{n}\mathbb{E}[\sqrt{\frac{RSS_{I_t,t}}{s^2_{I_t,t-1}}}] \le \sqrt{2}(L_2\sqrt{r\log(1+\sigma^2_{\max}/\lambda)+2log(\frac{1}{\delta})}+\lambda^{1/2}S_2)\sum_{t=K+1}^{n}\mathbb{E}[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}]+2\sqrt{2}K\sqrt{L_2}(\log n+1) \tag{30}$$

Proof. See appendix B.2 □

Remark 4. *Lemma A.2 is bounding sum of expected 'square root of normalized RSS', that is,* $\sqrt{RSS_{I_t,t}/s^2_{I_t,t-1}}$*. As discussed in Section 4, the RSS contributes additional exploration. As a matter of fact, the 'square root of normalized RSS' is proportional to the variance of Bootstrapped mean. Consequently, this Lemma assists bounding of the magnitude of extra exploration from residual Bootstrapping.*

Lemma A.3. *Assume the same as Theorem 5.1. Then*

$$\sum_{t=K+1}^{n}\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} \le \sqrt{2(n-K)d\log\left(1+\frac{\sum_{i=1}^{r}\sigma_i^2}{d\lambda}\right)} \tag{31}$$

Proof. See appendix B.3 □

Remark 5. *Lemma A.3 bounds the sum of action context norms which is also bounded in regret analysis of most contextual bandit algorithms.*

Technical Condition. Suppose for any $K < t \le n$ and some $\rho > 0$ such that $\rho = \tilde{O}(1)$ with respect to n and d. Then

$$s^{3/2}_{1,t-1}c_1^2(t,1) \le \rho\sigma^2_\omega(\sigma^2_{\min}+\lambda)\sqrt{\frac{1}{M_2}\log\left(\frac{M_1}{1-\gamma}\right)} \tag{32}$$

Remark 6. *This condition indicates that there is a lower bound for* σ^2_ω*, which means the extra exploration contributes to bounding of expected regret. This lower bound strongly supports the necessity of residual Bootstrap exploration. Another observation is that the lower bound is related to the time* t *and the number of pulling of optimal arm, which means that this hyperparameter for exploration* σ^2_ω *should depend on decision round* t*. However, since* σ^2_ω *is also related to some fixed constant related to environment and* ρ *which is a order of logarithm terms of* n *and* t*, it remains hard to determine what is the exact relation between* σ^2_ω *and* n*. This lower bound is only providing the conservative guarantee that the regret bound is sub-linear.*

Main proof of Theorem 5.1.

Following part is the main proof of Theorem 5.1, starting from decomposing regret by events,

$$
\begin{aligned}
R_n &= \sum_{k=2}^{K} \Delta_k \mathbb{E}[\sum_{t=1}^{n} \mathbb{I}\{I_t = k\}] && \text{(33a)}\\
&= \sum_{t=1}^{n} \mathbb{E}[\Delta_{I_t}] && \text{(33b)}\\
&= \sum_{t=K+1}^{n} \mathbb{E}[\Delta_{I_t}] + \sum_{t=1}^{K} \mathbb{E}[\Delta_{I_t}] && \text{(33c)}\\
&\le \sum_{t=K+1}^{n} \mathbb{E}[\Delta_{I_t} \mathbb{I}\{E_t\}] + \sum_{t=K+1}^{n} \mathbb{E}[\Delta_{I_t} \mathbb{I}\{\bar{E}_t\}] + 2S_2L(K-1) \quad \text{(by (34))} && \text{(33d)}\\
&\le \sum_{t=K+1}^{n} \mathbb{E}[\Delta_{I_t} \mathbb{I}\{E_t\}] + 2S_2L(n-K)\mathbb{P}(\bar{E}_t) + 2S_2L(K-1) \quad \text{(by (34))} && \text{(33e)}\\
&= \sum_{t=K+1}^{n} \mathbb{E}[\mathbb{E}_t[\Delta_{I_t} \mathbb{I}\{E_t\}]] + 2S_2L(n-K)\mathbb{P}(\bar{E}_t) + 2S_2L(K-1) && \text{(33f)}\\
&\le \sum_{t=K+1}^{n} \mathbb{E}[(\frac{2}{\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t')} + 1)(c_1(t, I_t) + c_2(t, I_t))\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}]] \\
&\quad + 2S_2L(\sum_{t=K+1}^{n} \mathbb{E}[\mathbb{P}(\bar{E}_t')] + (n-K)\mathbb{P}(\bar{E}_t) + K - 1) \quad \text{(by lemma A.1)} && \text{(33g)}\\
&\le \sum_{t=K+1}^{n} \mathbb{E}[(\frac{2}{\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t')} + 1)(c_1(t, I_t) + c_2(t, I_t))\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}]] \\
&\quad + 2S_2L((n-K)(\alpha + \beta) + K - 1) \quad \text{(by lemma 5.1 and 5.2)} && \text{(33h)}
\end{aligned}
$$

Where (34) is upper bound of optimal gap, that is, $\forall k \in [K]$

$$
\begin{aligned}
\Delta_k &= \boldsymbol{\theta}^\top (\boldsymbol{x}_1 - \boldsymbol{x}_k)\\
&\le \|\boldsymbol{\theta}\|_2 \|\boldsymbol{x}_1 - \boldsymbol{x}_k\|_2\\
&\le \|\boldsymbol{\theta}\|_2 \sqrt{2\|\boldsymbol{x}_1\|_2^2 + 2\|\boldsymbol{x}_k\|_2^2}\\
&\le 2S_2L
\end{aligned} \tag{34}
$$

By lemma 5.4 and the technical condition (32),

$$
\begin{aligned}
\frac{2}{\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t')} &\le \frac{2}{\frac{b}{\sqrt{2\pi}} \exp\left(-\frac{3 s_{1,t-1}^{3/2} c_1^2(t,1) \|\boldsymbol{x}_1\|_2^2}{8\sigma_\omega^2 (\sigma_{\min}^2 + \lambda) \sqrt{\frac{1}{M_2} \log\left(\frac{M_1}{1-\gamma}\right)}}\right) - \beta} && \text{(35a)}\\
&\le \frac{2}{\frac{b}{\sqrt{2\pi}} \exp\left(-\frac{3}{8}\|\boldsymbol{x}_1\|_2^2 \rho\right) - \beta} && \text{(35b)}
\end{aligned}
$$

Where

$$
M_1 := (e-1)^2 \exp\left(\frac{8\sigma_{\max}^2 S_2^2 L_2}{\frac{\lambda^2}{(\sigma_{\max}^2 + \lambda)^2} S_1^2 L_1} - 6\right) \tag{36}
$$

$$
M_2 := \frac{4\sigma_{\max}^2 S_2^2 L_2 - 2\frac{\lambda^2}{(\sigma_{\max}^2 + \lambda)^2} S_1^2 L_1}{(\frac{\lambda^2}{(\sigma_{\max}^2 + \lambda)^2} S_1^2 L_1)^2} \tag{37}
$$

Define the following notations for simplicity, note that the following constants are independent of n and d,

$$C_1(\alpha_1,\boldsymbol{\beta},\gamma,b) := \frac{2}{\dfrac{b}{\sqrt{2\pi}}\exp\Big(-\frac{3}{8}\|\boldsymbol{x}_1\|_2^2\rho\Big)-\beta}+1 \tag{38a}$$

$$C_2(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma,b,\delta) := C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\times\sqrt{2}(L_2\sqrt{r\log(1+\sigma_{\max}^2/\lambda)+2\log\Big(\frac{1}{\delta}\Big)}+\lambda^{1/2}S_2) \tag{38b}$$

Then, with probability at least $1-\gamma$,

$$\begin{aligned} R_n &\leq C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\sum_{t=K+1}^{n}\mathbb{E}[(c_1(t,I_t)+c_2(t,I_t))\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}]] \\ &\quad+2S_2L((n-K)(\alpha+\beta)+K-1) \end{aligned} \tag{39a}$$

$$\begin{aligned} &= C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\sum_{t=K+1}^{n}\mathbb{E}[c_1(t,I_t)\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}]] \\ &\quad+C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\sum_{t=K+1}^{n}\mathbb{E}[c_2(t,I_t)\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}]] \\ &\quad+2S_2L((n-K)(\alpha+\beta)+K-1) \end{aligned} \tag{39b}$$

$$\begin{aligned} &\leq C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)(L_2\sqrt{d\log\Big(\frac{1+nL^2/\lambda}{\alpha_{\min}}\Big)}+\lambda^{1/2}S_2)\sum_{t=K+1}^{n}\mathbb{E}[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}] \\ &\quad+C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\sum_{t=K+1}^{n}\mathbb{E}[\sqrt{\frac{2\sigma_\omega^2 RSS_{I_t,t}\log\Big(\frac{2}{\beta_{I_t}}\Big)}{s_{I_t,t-1}^2}}] \\ &\quad+2S_2L((n-K)(\alpha+\beta)+K-1) \end{aligned} \tag{39c}$$

$$\begin{aligned} &\leq C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)(L_2\sqrt{d\log\Big(\frac{1+nL^2/\lambda}{\alpha_{\min}}\Big)}+\lambda^{1/2}S_2)\sum_{t=K+1}^{n}\mathbb{E}[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}] \\ &\quad+C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\sqrt{2\sigma_\omega^2\log\Big(\frac{2}{\beta_{\min}}\Big)}\sum_{t=K+1}^{n}\mathbb{E}[\sqrt{\frac{RSS_{I_t,t}}{s_{I_t,t-1}^2}}] \\ &\quad+2S_2L((n-K)(\alpha+\beta)+K-1) \end{aligned} \tag{39d}$$

Further define,

$$\zeta_1(n,d) := (L_2\sqrt{d\log\Big(\frac{1+nL^2/\lambda}{\alpha_{\min}}\Big)}+\lambda^{1/2}S_2)\sqrt{2(n-K)d\log\Big(1+\sum_{i=1}^{r}\sigma_i^2/d\lambda\Big)} \tag{40}$$

$$\zeta_2(n,d) := \sqrt{2\sigma_\omega^2\log\Big(\frac{2}{\beta_{\min}}\Big)}\sqrt{2(n-K)d\log\Big(1+\sum_{i=1}^{r}\sigma_i^2/d\lambda\Big)} \tag{41}$$

$$\zeta_3(n) := 2K\sqrt{4L_2\sigma_\omega^2 log(\frac{2}{\beta_{\min}})}(\log n+1) \tag{42}$$

$$\zeta_4(n) := 2S_2L((n-K)(\alpha+\beta)+K-1) \tag{43}$$

By lemma A.2, with probability at least $1-(\delta+\gamma)$,

$$R_n \leq C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\zeta_1(n,d)+C_2(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma,b,\delta)\zeta_2(n,d)+C_1(\alpha_1,\boldsymbol{\beta},\gamma,b)\zeta_3(n,d)+\zeta_4(n,d) \tag{44}$$

The ζ_1, ζ_2, ζ_3 and ζ_4 can also be found in Table.1 and C_1 and C_2 are summarised in the Table.2.

□

Notation	Definition
M_1	$(e-1)^2\exp\left(\dfrac{8\sigma^2_{\max}S_2^2L_2}{\lambda^2S_1^2L_1/(\sigma^2_{\max}+\lambda)^2}-6\right)$
M_2	$\dfrac{4\sigma^2_{\max}S_2^2L_2-2\lambda^2S_1^2L_1/(\sigma^2_{\max}+\lambda)^2}{(\lambda^2S_1^2L_1/(\sigma^2_{\max}+\lambda)^2)^2}$
C_1	$2\left(\dfrac{b}{\sqrt{2\pi}}\exp\left(-\frac{3}{8}\|\boldsymbol{x}_1\|_2^2\rho\right)-\beta\right)^{-1}+1$
C_2	$C_1\sqrt{2}(L_2\sqrt{r\log(1+\sigma^2_{\max}/\lambda)+2\log\left(\frac{1}{\delta}\right)}+\lambda^{1/2}S_2)$

Table 2: Constants in Analysis

A.2 PROOF OF COROLLARY 5.2

Proof. We will analyze terms C_1, C_2 and ζ_1, ζ_2, ζ_3, ζ_4 one by one in terms of the rate in the big O notation with respect to n and d. Also recall that the notation $\tilde{O}$ is the big O notation up to logarithmic factor with respect to n and d. Following steps include the first step for C_1 and C_2, the second step for ζ_1, ζ_2, ζ_3 and ζ_4 and the last one for combining results.

Step 1 As $\boldsymbol{\beta}$ is chosen as a vector with elements $\frac{1}{\sqrt{n}}$, the term C_1 is actually $O(\rho)$ which is assumed to be $\tilde{O}(1)$. Under stochastic linear bandit that contexts and subgaussian constant L_2 are given, C_2 is also $\tilde{O}(1)$. Note that, other parameters such as δ, λ and b are viewed as constants.

Step 2. From Table.1, as $\boldsymbol{\alpha}$ is chosen as a vector with elements $\frac{1}{\sqrt{n}}$, we can conclude that $\zeta_1(n,d) = O(\sqrt{d\log n}\times\sqrt{nd\log d})$, $\zeta_2(n,d) = O(\sqrt{\log n}\times\sqrt{nd\log d})$, $\zeta_3(n) = O(\log n\sqrt{\log n})$ and $\zeta_4(n) = O(\sqrt{n})$. By the notation of $\tilde{O}$, it can be summarised as $\zeta_1(n,d) = \tilde{O}(d\sqrt{n})$, $\zeta_2(n,d) = \tilde{O}(\sqrt{dn})$, $\zeta_3(n) = \tilde{O}(1)$ and $\zeta_4(n) = \tilde{O}(\sqrt{n})$.

Step 3. As a result, expected regret of our `LinReBoot` in Theorem 5.1 under the choice of tuning parameter mentioned in Corollary 5.2, has high probability upper bound with the order $\tilde{O}(d\sqrt{n})+\tilde{O}(\sqrt{dn})+\tilde{O}(1)+\tilde{O}(\sqrt{n}) = \tilde{O}(d\sqrt{n})$. □

A.3 PROOF OF LEMMA 5.1

Proof. Based on Theorem 2 in [Abbasi-Yadkori et al., 2011] which is Lemma C.1, for all $\alpha \in (0,1)$,

$$\mathbb{P}(\left\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{V}_t} \leq L_2\sqrt{d\log\left(\frac{1+tL^2/\lambda}{\alpha}\right)}+\lambda^{1/2}S_2) \geq 1-\alpha \tag{45}$$

Thus, $\forall\alpha_k \in (0,1)$, with probability at least $1-\alpha_k$

$$\begin{aligned} |\hat{\mu_{k,t}}-\mu_k| &= |\boldsymbol{x}^\top(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta})| &\text{(46a)}\\ &\leq \|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}\left\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}\right\|_{\boldsymbol{V}_t} &\text{(46b)}\\ &\leq L_2\sqrt{d\log\left(\frac{1+tL^2/\lambda}{\alpha}\right)}+\lambda^{1/2}S_2)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}} \quad \text{(lemma C.1)} &\text{(46c)}\end{aligned}$$

That is, let $c_1(t,k) := L_2\sqrt{d\log\left(\frac{1+tL^2/\lambda}{\alpha_k}\right)}+\lambda^{1/2}S_2$,

$$\mathbb{P}(E_{t,k}) \geq 1-\alpha_k \tag{47}$$

Therefore,

$$\mathbb{P}(\bar{E}_t) = \mathbb{P}(\bigcup_{k=1}^K \bar{E}_{t,k}) \leq \sum_{k=1}^K \alpha_k \tag{48}$$

□

A.4 PROOF OF LEMMA 5.2

Proof. Recall the our definition of event $E'_{t,k}$ and $RSS_{k,t}$,

$$E'_{t,k} := \{|\tilde{\mu}_{k,t} - \hat{\mu}_{k,t}| \leq c_2(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}\}$$

$$RSS_{k,t} := \sum_{i=1}^{s_{k,t-1}} e_{k,t,i}^2$$

Then control the probability of the bad event $\bar{E}'_{t,k}$ which indicates a "large" deviation between estimated mean and Bootstrapped mean of the k-th arm at round t. That is, $\forall t \geq K+1, \forall k \in [K]$,

$$\mathbb{P}_t(\bar{E}'_{t,k}) = \mathbb{P}_t(|\tilde{\mu}_{k,t} - \hat{\mu}_{k,t}| > c_2(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}) \tag{49a}$$

$$= \mathbb{P}_t(|\frac{1}{s_{k,t-1}} \sum_{i=1}^{s_{k,t-1}} \omega_{k,t,i} e_{k,t,i}| > c_2(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}) \tag{49b}$$

$$= \mathbb{P}_t(|\sqrt{\frac{\sigma_\omega^2 \sum_{i=1}^{s_{k,t-1}} e_{k,t,i}^2}{s_{k,t-1}^2}} Z| > c_2(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}) \tag{49c}$$

$$= \mathbb{P}_t(|Z| > \frac{c_2(t,k) s_{k,t-1} \|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{k,t}}}) \quad \text{(Define } Z \sim N(0,1)\text{)} \tag{49d}$$

$$\leq \mathbb{P}_t(|Z| > \frac{c_2(t,k) s_{k,t-1} \|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{k,t}}}) \tag{49e}$$

$$\leq 2\exp\left(-\frac{c_2^2(t,k) s_{k,t-1}^2 \|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}^2}{\sigma_\omega^2 RSS_{k,t}}\right) \quad (Z \text{ is subgaussian with constant } 1) \tag{49f}$$

Now let $\beta_k := 2\exp\left(-\frac{c_2^2(t,k) s_{k,t-1}^2 \|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}^2}{\sigma_\omega^2 RSS_{k,t}}\right)$ then

$$c_2(t,k) := \sqrt{\frac{2\sigma_\omega^2 RSS_{k,t} \log\left(\frac{2}{\beta_k}\right)}{s_{k,t-1}^2 \|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}^2}} \tag{50}$$

Therefore,

$$\mathbb{P}_t(|\tilde{\mu}_{k,t} - \hat{\mu}_{k,t}| \leq c_2(t,k)\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}}) \geq 1 - \beta_k \tag{51}$$

□

A.5 PROOF OF LEMMA 5.3

Proof. Follow the same notations in A.4,

$$RSS_{k,t} := \sum_{i=1}^{s_{k,t-1}} e_{k,t,i}^2 \qquad Z \sim N(0,1)$$

Similar to lemma 10 in [Wang et al., 2020], the vanilla Gaussian tail lower bound, lemma C.2, is used. That is, $\forall t$, $\forall b > 0$

$$\begin{aligned}
\mathbb{P}_t(E_t'') &= \mathbb{P}_t(\tilde{\mu}_{1,t} - \hat{\mu}_{1,t} > c_1(t,1)\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}) && \text{(52a)}\\
&= \mathbb{P}_t(\frac{1}{s_{1,t-1}}\sum_{i=1}^{s_{1,t-1}} \omega_{1,i} e_{1,t,i} > c_1(t,1)\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}) && \text{(52b)}\\
&= \mathbb{P}_t(Z > \frac{c_1(t,1)s_{1,t-1}\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{1,t}}}) && \text{(52c)}\\
&\geq \begin{cases} \frac{b}{\sqrt{2\pi}} \exp\left(-\frac{3c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{2\sigma_\omega^2 RSS_{1,t}}\right) & \text{if } \frac{c_1(t,1)s_{1,t-1}\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{1,t}}} \geq b \\ \Phi(-b) & \text{if } 0 < \frac{c_1(t,1)s_{1,t-1}\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{1,t}}} < b \end{cases} && \text{(52d)}
\end{aligned}$$

Where b is the constant chosen by us. This b controlling the sharpness of the lower bound of Gaussian tail. Notice that (20) is equivalent to the condition $\frac{c_1(t,1)s_{1,t-1}\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{1,t}}} \geq b$ by the definition (17) and (18), the above lower bound can be writed as,

$$\mathbb{P}_t(E_t'') \geq \begin{cases} \frac{b}{\sqrt{2\pi}} \exp\left(-\frac{3c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{2\sigma_\omega^2 RSS_{1,t}}\right) & \text{if } \frac{c_1(t,1)}{c_2(t,1)} \geq b\sqrt{2\log\left(\frac{2}{\beta_1}\right)} \\ \Phi(-b) & \text{if } \frac{c_1(t,1)}{c_2(t,1)} < b\sqrt{2\log\left(\frac{2}{\beta_1}\right)} \end{cases} \tag{53}$$

□

A.6 PROOF OF LEMMA 5.4

Proof. Recall our true model:

$$\boldsymbol{Y}_t = \boldsymbol{X}_t\boldsymbol{\theta} + \boldsymbol{\epsilon}_t$$

Further define matrix $\boldsymbol{Q}_{k,t}$ which indicates the RSS decomposition for the k-th arm at time t:

$$[\boldsymbol{Q}_{k,t}]_{ij} = \begin{cases} 1 & i = j \text{ and } I_i = k \\ 0 & \text{otherwise} \end{cases} \forall i,j \in [t] \tag{54}$$

In this proof, we will start from stating lemmas and technical condition, then give main proof which has three steps.

Lemma A.4. *By (54), which is definition of $\boldsymbol{Q}_{k,t}$, RSS_t can be decomposed by arms,*

$$RSS_t := \left\|\boldsymbol{Y}_t - \boldsymbol{X}_t\hat{\boldsymbol{\theta}}_t\right\|_2^2 = \sum_{k=1}^{K} RSS_{k,t} \tag{55}$$

And $RSS_{k,t} := \left\|\boldsymbol{Q}_{k,t}(\boldsymbol{Y}_t - \boldsymbol{X}_t\hat{\boldsymbol{\theta}}_t)\right\|_2^2$ can be re-writed as:

$$\begin{aligned}
RSS_{k,t} =& \left\|\boldsymbol{Q}_{k,t-1}(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{X}_{t-1}\boldsymbol{\theta}\right\|_2^2 \\
&+ \left\|\boldsymbol{Q}_{k,t-1}(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{\epsilon}_{t-1}\right\|_2^2 \\
&+ 2\boldsymbol{\theta}^\top\boldsymbol{X}_{t-1}^\top(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{Q}_{k,t-1}^\top\boldsymbol{Q}_{k,t-1}(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{\epsilon}_{t-1}
\end{aligned} \tag{56}$$

Proof. See appendix B.4. □

Remark. Lemma A.4 provides a decomposition of RSS for arm k at round t.

Lemma A.5. *Stochastic process $\{\epsilon_t\}_{t=1}^{\infty}$ satisfies that for some $R_1, R_2 > 0$,*

$$e^{R_1\eta^2} \leq \mathbb{E}[e^{\eta\epsilon_t}|\mathcal{F}_{t-1}] \leq e^{R_2\eta^2} \quad \forall\eta \geq 0$$

Singular value decomposition of $\boldsymbol{X}_K$ and definition of ridge shrinkage context matrix $\boldsymbol{Z}$ are

$$\begin{aligned}\boldsymbol{X}_K &:= \boldsymbol{G\Sigma U}\\ \boldsymbol{\Omega} &:= \boldsymbol{\Sigma}(\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^\top\\ \boldsymbol{Z} &:= \boldsymbol{G\Omega\Sigma U}\end{aligned}$$

Let $\boldsymbol{z}_1$ be the vector of the first row of matrix $\boldsymbol{Z}$ and suppose $(\boldsymbol{x}_1^\top - \boldsymbol{z}_1^\top\boldsymbol{\theta})^2 \geq S_1^2$. Then $\forall\eta \geq 0$, $\forall t \geq K+1$,

$$exp(\frac{\lambda^2}{(\sigma_{\max}^2 + \lambda)^2}S_1^2L_1\eta^2) \leq \mathbb{E}[e^{\eta\xi_t}] \leq exp(\sigma_{\max}^2S_2^2L_2\eta^2) \tag{57}$$

Where $\xi_t := \frac{1}{\sqrt{s_{1,t-1}}}\boldsymbol{\theta}^\top\boldsymbol{X}_{t-1}^\top(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{Q}_{1,t-1}^\top\boldsymbol{Q}_{1,t-1}(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{\epsilon}_{t-1}$

Proof. See appendix B.5. □

Remark. Lemma A.5 indicates that the random variable ξ_t which is based on noise process $\{\epsilon_\tau\}_{\tau=1}^{t-1}$ also has the clipping noise property. Thus this random variable is also subgaussian. This result supports our application of Lemma A.6 which is given in the next part.

Lemma A.6. *Suppose X is a random variable such that $\exists R_1, R_2 > 0$*

$$\exp\bigl(R_1t^2\bigr) \leq \mathbb{E}[e^{tX}] \leq \exp\bigl(R_2t^2\bigr) \quad \forall t \geq 0 \tag{58}$$

Then

$$\mathbb{P}(X \geq x) \geq C_1\exp\bigl(-C_2x^2\bigr) \tag{59}$$

Where $C_1 := (e-1)^2e^{\frac{8R_2}{R_1}-6}$ and $C_2 := \frac{4R_2-2R_1}{R_1^2}$

Proof. See appendix B.6 □

Remark. This Lemma is inspired by the Theorem 1 and its proof in [Zhang and Zhou, 2020]. This Lemma gives the lower tail bound of random variable X and the only condition is that there is upper and lower bound of the form e^{Ct^2} for the moment generating function of X.

Technical Condition. The difference between $\mathbb{P}_t(E_t'')$ and $\mathbb{P}_t(\bar{E}_t')$ plays a key role in bounding regret when applying the stochastic exploration on least squared framework. The following part is the probabilistic analysis of lower bound of this difference, which will be denoted as $D < \mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t')$ in this proof. First impose some requirements on the tuning parameters β, D, b:

$$D + \beta < \min(\Phi(-b), \frac{b}{\sqrt{2\pi}}e^{-\frac{3}{2}b^2}) \tag{60}$$

This requirement indicates three results:

$$D + \beta < \Phi(-b) \tag{61}$$

$$D + \beta < \frac{b}{\sqrt{2\pi}} \tag{62}$$

$$-\frac{3}{2\log\Bigl(\frac{\sqrt{2\pi}}{b}(D+\beta)\Bigr)} < \frac{1}{b^2} \tag{63}$$

Main proof of lemma 5.4
Step 1: Express event $\{\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t') > D\}$ as an inequality of $RSS_{1,t}$

The idea in this step is starting from decomposing our target event $\{\mathbb{P}_t(Et'') - \mathbb{P}_t(\bar{E}'_t) > D\}$ by the condition mentioned in lemma 5.3. That is,

$$\begin{aligned}
&\mathbb{P}(\mathbb{P}_t(E''_t) - \mathbb{P}_t(\bar{E}'_t) > D) && (64a)\\
\geq&\mathbb{P}(\mathbb{P}_t(E''_t) > D + \beta) \quad \text{(by lemma 5.2)} && (64b)\\
=&\mathbb{P}(\{\mathbb{P}_t(E''_t) > D + \beta\} \cap \{\frac{c_1(t,1)s_{1,t-1}\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{1,t}}} \geq b\})\\
&+ \mathbb{P}(\{\mathbb{P}_t(E''_t) > D + \beta\} \cap \{\frac{c_1(t,1)s_{1,t-1}\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{1,t}}} < b\}) && (64c)\\
\geq&\mathbb{P}(\{\frac{b}{\sqrt{2\pi}}\exp\left(-\frac{3s_{1,t-1}^2 c_1^2(t,1)\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{2\sigma_\omega^2 RSS_{1,t}}\right) > D + \beta\} \cap \{\frac{c_1(t,1)s_{1,t-1}\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{1,t}}} \geq b\})\\
&+ \mathbb{P}(\{\Phi(-b) > D + \beta\} \cap \{\frac{c_1(t,1)s_{1,t-1}\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}}{\sqrt{\sigma_\omega^2 RSS_{1,t}}} < b\}) \quad \text{(by lemma 5.3)} && (64d)
\end{aligned}$$

Then we apply the technical condition described in 60,

$$\begin{aligned}
&\mathbb{P}(\mathbb{P}_t(E''_t) - \mathbb{P}_t(\bar{E}'_t) > D) && (65a)\\
\geq&\mathbb{P}(\{RSS_{1,t} > -\frac{3c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{2\sigma_\omega^2 \log\left(\frac{\sqrt{2\pi}}{b}(D+\beta)\right)}\} \cap \{RSS_{1,t} \leq \frac{c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{\sigma_\omega^2 b^2}\})\\
&+ \mathbb{P}(RSS_{1,t} > \frac{c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{\sigma_\omega^2 b^2}) \quad \text{(by (61) and (62))} && (65b)\\
=&\mathbb{P}(RSS_{1,t} > -\frac{3c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{2\sigma_\omega^2 \log\left(\frac{\sqrt{2\pi}}{b}(D+\beta)\right)}) \quad \text{(by (63))} && (65c)
\end{aligned}$$

Step 2: Apply lemmas to give lower bounds
In this step, three lemmas are used.

$$\begin{aligned}
&\mathbb{P}(RSS_{1,t} > -\frac{3c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{2\sigma_\omega^2 \log\left(\frac{\sqrt{2\pi}}{b}(D+\beta)\right)}) && (66a)\\
\leq&\mathbb{P}(\boldsymbol{\theta}^\top \boldsymbol{X}_{t-1}^\top(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{Q}_{1,t-1}^\top \boldsymbol{Q}_{1,t-1}(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{\epsilon}_{t-1}\\
&\quad > \frac{3c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{8\sigma_\omega^2 \log\left(\frac{b}{\sqrt{2\pi}(D+\beta)}\right)}) \quad \text{(by (67))} && (66b)
\end{aligned}$$

Where (67) is derived directly from lemma A.4,

$$RSS_{1,t} \geq 4\boldsymbol{\theta}^\top \boldsymbol{X}_{t-1}^\top(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{Q}_{1,t-1}^\top \boldsymbol{Q}_{1,t-1}(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{\epsilon}_{t-1} \qquad (67)$$

Denote $\xi_t := \frac{1}{\sqrt{s_{1,t-1}}}\boldsymbol{\theta}^\top \boldsymbol{X}_{t-1}^\top(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{Q}_{1,t-1}^\top \boldsymbol{Q}_{1,t-1}(\boldsymbol{I} - \boldsymbol{X}_{t-1}\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{\epsilon}_{t-1}$. By lemma A.5, moment generating function of random variable ξ_t has upper bound and lower bound,

$$exp(\frac{\lambda^2}{(\sigma_{\max}^2 + \lambda)^2}S_1^2 L_1 \eta^2) \leq \mathbb{E}[e^{\eta\xi_t}] \leq exp(\sigma_{\max}^2 S_2^2 L_2 \eta^2)$$

Then applying lemma A.6,

$$\begin{aligned}
\mathbb{P}(\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t') > D) \geq & \mathbb{P}(RSS_{1,t} > -\frac{3c_1^2(t,1)s_{1,t-1}^2\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{2\sigma_\omega^2\log\left(\frac{\sqrt{2\pi}}{b}(D+\beta)\right)}) && \text{(68a)}\\
\geq & \mathbb{P}(\xi_t > \frac{3c_1^2(t,1)s_{1,t-1}^{3/2}\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{8\sigma_\omega^2\log\left(\frac{b}{\sqrt{2\pi}(D+\beta)}\right)}) && \text{(68b)}\\
\geq & M_1\exp\left(-M_2(\frac{3c_1^2(t,1)s_{1,t-1}^{3/2}\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{8\sigma_\omega^2\log\left(\frac{b}{\sqrt{2\pi}(D+\beta)}\right)})^2\right) && \text{(68c)}
\end{aligned}$$

Where

$$M_1 := (e-1)^2\exp\left(\frac{8\sigma_{max}^2 S_2^2 L_2}{\frac{\lambda^2}{(\sigma_{max}^2+\lambda)^2}S_1^2 L_1} - 6\right) \tag{69}$$

$$M_2 := \frac{4\sigma_{\max}^2 S_2^2 L_2 - 2\frac{\lambda^2}{(\sigma_{\max}^2+\lambda)^2}S_1^2 L_1}{(\frac{\lambda^2}{(\sigma_{\max}^2+\lambda)^2}S_1^2 L_1)^2} \tag{70}$$

Let $1-\gamma := M_1\exp\left(-M_2(\frac{3c_1^2(t,1)s_{1,t-1}^{3/2}\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{8\sigma_\omega^2\log\left(\frac{b}{\sqrt{2\pi}(D+\beta)}\right)})^2\right)$, then

$$D := \frac{b}{\sqrt{2\pi}}\exp\left(-\frac{3c_1^2(t,1)s_{1,t-1}^{3/2}\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{8\sigma_\omega^2\sqrt{\frac{1}{M_2}\log\left(\frac{M_1}{1-\gamma}\right)}}\right) - \beta \tag{71}$$

Thus the connection between concentration and anti-concentration can be described as the following high probability lower bound,

$$\mathbb{P}(\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t') > \frac{b}{\sqrt{2\pi}}\exp\left(-\frac{3c_1^2(t,1)s_{1,t-1}^{3/2}\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}}}{8\sigma_\omega^2\sqrt{\frac{1}{M_2}\log\left(\frac{M_1}{1-\gamma}\right)}}\right) - \beta) \geq 1-\gamma \tag{72}$$

Notice that $\|\boldsymbol{x}_1\|^2_{\boldsymbol{V}_t^{-1}} \leq \frac{\|\boldsymbol{x}_1\|_2^2}{\sigma_{\min}^2+\lambda}$, then $\forall t \geq K+1$, with probability at least $1-\gamma$,

$$\mathbb{P}_t(E_t'') - \mathbb{P}_t(\bar{E}_t') > \frac{b}{\sqrt{2\pi}}\exp\left(-\frac{3s_{1,t-1}^{3/2}c_1^2(t,1)\|\boldsymbol{x}_1\|_2^2}{8\sigma_\omega^2(\sigma_{\min}^2+\lambda)\sqrt{\frac{1}{M_2}\log\left(\frac{M_1}{1-\gamma}\right)}}\right) - \beta \tag{73}$$

Where M_1, M_2 are defined as (69) and (70).

Technical condition on b becomes,

$$\frac{b}{\sqrt{2\pi}}\exp\left(-\frac{3s_{1,t-1}^{3/2}c_1^2(t,1)\|\boldsymbol{x}_1\|_2^2}{8\sigma_\omega^2(\sigma_{\min}^2+\lambda)\sqrt{\frac{1}{M_2}\log\left(\frac{M_1}{1-\gamma}\right)}}\right) < min(\Phi(-b), \frac{b}{\sqrt{2\pi}}e^{-\frac{3}{2}b^2}) \tag{74}$$

□

B PROOFS OF TECHNICAL LEMMAS

B.1 PROOF OF LEMMA A.1

Proof. This proof is mainly adapted from proof of lemma 2 in [Kveton et al., 2020a]. The main extension is to redefine the concept of ”least uncertain undersampled” arm to meet the need of residual bootstrap exploration. First define ’under sampled’ arms,

$$\bar{\mathcal{S}}_t := \{k \in [K] : c_{t,k}\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}} \geq \Delta_k\} \tag{75}$$

Where $c_{t,k} := c_1(t,k) + c_2(t,k)$ and the set of "sufficiently sampled" arms is $\mathcal{S}_t := [K] \setminus \bar{\mathcal{S}}_t$. Also define the "least uncertain" arm at round t,

$$J_t := \arg\min_{k \in \bar{\mathcal{S}}_t} c_{t,k}\|\boldsymbol{x}_k\|_{\boldsymbol{V}_t^{-1}} \tag{76}$$

Then when event E'_t occurs,

$$\begin{aligned}
\Delta_{I_t} &= \mu_1 - \mu_{I_t} + \mu_{J_t} - \mu_{J_t} \qquad &(77\text{a})\\
&= \Delta_{J_t} + \mu_{J_t} - \mu_{I_t} &(77\text{b})\\
&= \Delta_{J_t} + \mu_{J_t} - \tilde{\mu}_{J_t,t} + \tilde{\mu}_{J_t,t} - \tilde{\mu}_{I_t,t} + \tilde{\mu}_{I_t,t} - \mu_{I_t} &(77\text{c})\\
&\leq \Delta_{J_t} + c_{t,J_t}\|\boldsymbol{x}_{J_t}\|_{\boldsymbol{V}_t^{-1}} + c_{t,I_t}\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} + \tilde{\mu}_{J_t,t} - \tilde{\mu}_{I_t,t} \quad (E_t \cap E'_t) &(77\text{d})\\
&\leq \Delta_{J_t} + c_{t,J_t}\|\boldsymbol{x}_{J_t}\|_{\boldsymbol{V}_t^{-1}} + c_{t,I_t}\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} \quad (\tilde{\mu}_{J_t,t} < \tilde{\mu}_{I_t,t}) &(77\text{e})\\
&\leq 2c_{t,J_t}\|\boldsymbol{x}_{J_t}\|_{\boldsymbol{V}_t^{-1}} + c_{t,I_t}\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} \quad (J_t \in \bar{\mathcal{S}}_t) &(77\text{f})
\end{aligned}$$

Thus conditional expected gap can be bounding by the norms of two special arms I_t and J_t at round t,

$$\begin{aligned}
\mathbb{E}_t[\Delta_{I_t}] &= \mathbb{E}_t[\Delta_{I_t}\mathbb{I}\{E'_t\}] + \mathbb{E}_t[\Delta_{I_t}\mathbb{I}\{\bar{E}'_t\}] \qquad &(78\text{a})\\
&\leq \mathbb{E}_t[2c_{t,J_t}\|\boldsymbol{x}_{J_t}\|_{\boldsymbol{V}_t^{-1}} + c_{t,I_t}\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}] + M\mathbb{P}_t(\bar{E}'_t) &(78\text{b})
\end{aligned}$$

Now we need to bound the norm of J_t by the norm of I_t. The key observation to find the relation between I_t and J_t is

$$\mathbb{E}_t[c_{t,I_t}\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}] \geq \mathbb{E}_t[c_{t,I_t}\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}|I_t \in \bar{\mathcal{S}}_t]\mathbb{P}_t(I_t \in \bar{\mathcal{S}}_t) \geq c_{t,J_t}\|\boldsymbol{x}_{J_t}\|_{\boldsymbol{V}_t^{-1}}\mathbb{P}_t(I_t \in \bar{\mathcal{S}}_t) \tag{79}$$

Thus

$$c_{t,J_t}\|\boldsymbol{x}_{J_t}\|_{\boldsymbol{V}_t^{-1}} \leq \frac{\mathbb{E}_t[c_{t,I_t}\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}]}{\mathbb{P}_t(I_t \in \bar{\mathcal{S}}_t)} \tag{80}$$

Now we need to give lower bound of $\mathbb{P}_t(I_t \in \bar{\mathcal{S}}_t)$,

$$\begin{aligned}
\mathbb{P}_t(I_t \in \bar{\mathcal{S}}_t) &= \mathbb{P}_t(\exists k \in \bar{\mathcal{S}}_t \text{ s.t } \tilde{\mu}_{k,t} > \max_{j \in \mathcal{S}_t} \tilde{\mu}_{j,t}) \qquad &(81\text{a})\\
&\geq \mathbb{P}_t(\tilde{\mu}_{1,t} > \max_{j \in \mathcal{S}_t} \tilde{\mu}_{j,t}) \quad (1 \in \bar{\mathcal{S}}_t) &(81\text{b})\\
&\geq \mathbb{P}_t(\{\tilde{\mu}_{1,t} > \max_{j \in \mathcal{S}_t} \tilde{\mu}_{j,t}\} \cap E'_t) &(81\text{c})\\
&\geq \mathbb{P}_t(\{\tilde{\mu}_{1,t} > \mu_1\} \cap E'_t) \quad (\text{by } (82)) &(81\text{d})\\
&\geq \mathbb{P}_t(\tilde{\mu}_{1,t} > \mu_1) - \mathbb{P}_t(\bar{E}'_t) &(81\text{e})\\
&\geq \mathbb{P}_t(E''_t) - \mathbb{P}_t(\bar{E}'_t) \quad (\text{by } (83)) &(81\text{f})
\end{aligned}$$

Where (82), (83) are

$$\begin{aligned}
\forall j \in \mathcal{S}_t \quad \tilde{\mu}_{j,t} \leq \mu_j + c_{t,j}\|\boldsymbol{x}_j\|_{\boldsymbol{V}_t^{-1}} < \mu_j + \Delta_j = \mu\\
\Rightarrow \{\tilde{\mu}_{1,t} > \mu_1\} \subset \{\tilde{\mu}_{1,t} > \tilde{\mu}_{j,t} \quad \forall j \in \mathcal{S}_t\}
\end{aligned} \tag{82}$$

$$\{\tilde{\mu}_{1,t} - \hat{\mu}_{1,t} > c_1(t,1)\|\boldsymbol{x}_1\|_{\boldsymbol{V}_t^{-1}}\} \subset \{\tilde{\mu}_{1,t} > \mu_1\} \quad (\text{since } E_t \text{ occurs}) \tag{83}$$

Therefore,

$$\mathbb{E}_t[\Delta_{I_t}] \leq (\frac{2}{\mathbb{P}_t(E''_t) - \mathbb{P}_t(\bar{E}'_t)} + 1)(c_1(t,I_t) + c_2(t,I_t))\mathbb{E}_t[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}] + M\mathbb{P}(\bar{E}'_t) \tag{84}$$

□

B.2 PROOF OF LEMMA A.2

Proof. First define $\{\epsilon_{I_t,i}\}_{i=1}^{s_{I_t,t-1}}$ for the noise of arm I_t at round t. Note that these $\{\epsilon_{I_t,i}\}_{i=1}^{s_{I_t,t-1}}$ is a subset of the noise vector $\boldsymbol{\epsilon}_{t-1} = (\epsilon_1, ..., \epsilon_{t-1})^\top$ at round t. Also define $\mathcal{F}_{I_t,i}$, the randomness history until the noise $\epsilon_{I_t,i}$ is generated and let $\mathcal{I}_{I_t,t}$ be the set of time stamps when arm I_t is pulled up to round t. For example, suppose arm 1 is pulled at round $1, 11, 21, 25$ up to round 26, then $\mathcal{I}_{1,26} = \{1, 11, 21, 25\}$ and noise set is $\{\epsilon_{1,i}\}_{i=1}^{s_{1,25}} = \{\epsilon_{1,1}, \epsilon_{1,2}, \epsilon_{1,3}, \epsilon_{1,4}\}$. For one of these noises such as $\epsilon_{1,3}$, $\mathcal{F}_{1,3} = \mathcal{F}_{20}$ since $\epsilon_{1,3} = \epsilon_{21}$, indicating $\mathbb{E}[e^{\eta\epsilon_{1,3}}|\mathcal{F}_{20}] \le e^{R_2\eta^2}$, $\forall \eta \ge 0$. As a result, other expressions of residuals and RSS of the arm pulled at round $t \ge K+1$ are

$$e_{I_t,t,i} = \boldsymbol{x}_{I_t}^\top \boldsymbol{\theta} + \epsilon_{I_t,i} - \boldsymbol{x}_{I_t}^\top \hat{\boldsymbol{\theta}}_t \tag{85}$$

$$RSS_{I_t,t} = \sum_{i=1}^{s_{I_t,t-1}} e_{I_t,t,i}^2 = \sum_{i=1}^{s_{I_t,t-1}} (\boldsymbol{x}_{I_t}^\top \boldsymbol{\theta} + \epsilon_{I_t,i} - \boldsymbol{x}_{I_t}^\top \hat{\boldsymbol{\theta}}_t)^2 \tag{86}$$

Starting from ridge estimate $\hat{\boldsymbol{\theta}}_t$,

$$\hat{\boldsymbol{\theta}}_t = \boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top(\boldsymbol{X}_{t-1}\boldsymbol{\theta} + \boldsymbol{\epsilon}_{t-1}) \tag{87a}$$

$$= \boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top\boldsymbol{X}_{t-1}\boldsymbol{\theta} + \boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top\boldsymbol{\epsilon}_{t-1} \tag{87b}$$

$$= \boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top\boldsymbol{\epsilon}_{t-1} + \boldsymbol{V}_t^{-1}(\boldsymbol{X}_{t-1}^\top\boldsymbol{X}_{t-1} + \lambda\boldsymbol{I})\boldsymbol{\theta} - \lambda\boldsymbol{V}_t^{-1}\boldsymbol{\theta} \tag{87c}$$

$$= \boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top\boldsymbol{\epsilon}_{t-1} - \lambda\boldsymbol{V}_t^{-1}\boldsymbol{\theta} + \boldsymbol{\theta} \tag{87d}$$

Thus,

$$\boldsymbol{x}_{I_t}^\top\boldsymbol{\theta} - \boldsymbol{x}_{I_t}^\top\hat{\boldsymbol{\theta}}_t = \boldsymbol{x}_{I_t}^\top\boldsymbol{\theta} - \boldsymbol{x}_{I_t}^\top\boldsymbol{V}_t^{-1}\boldsymbol{X}_{t-1}^\top\boldsymbol{\epsilon}_{t-1} + \lambda\boldsymbol{x}_{I_t}^\top\boldsymbol{V}_t^{-1}\boldsymbol{\theta} - \boldsymbol{x}_{I_t}^\top\boldsymbol{\theta} \tag{88a}$$

$$= \langle \boldsymbol{x}_{I_t}, \boldsymbol{X}_{t-1}^\top\boldsymbol{\epsilon}_{t-1}\rangle_{\boldsymbol{V}_t^{-1}} - \lambda\langle \boldsymbol{x}_{I_t}, \boldsymbol{\theta}\rangle_{\boldsymbol{V}_t^{-1}} \tag{88b}$$

So RSS becomes,

$$\begin{aligned} RSS_{I_t,t} &= \sum_{i=1}^{s_{I_t,t-1}} (\boldsymbol{x}_{I_t}^\top\boldsymbol{\theta} - \boldsymbol{x}_{I_t}^\top\hat{\boldsymbol{\theta}}_t + \epsilon_{I_t,i})^2 \\ &\le 2s_{I_t,t-1}\Big(\langle \boldsymbol{x}_{I_t}, \boldsymbol{X}_{t-1}^\top\boldsymbol{\epsilon}_{t-1}\rangle_{\boldsymbol{V}_t^{-1}} - \lambda\langle \boldsymbol{x}_{I_t}, \boldsymbol{\theta}\rangle_{\boldsymbol{V}_t^{-1}}\Big)^2 + 2\sum_{i=1}^{s_{I_t,t-1}} \epsilon_{I_t,i}^2 \end{aligned} \tag{89}$$

Therefore,

$$\sum_{t=K+1}^{n} \mathbb{E}[\sqrt{\frac{RSS_{I_t,t}}{s_{I_t,t-1}^2}}] \le \sum_{t=K+1}^{n} \mathbb{E}[\sqrt{2\Big(\langle \boldsymbol{x}_{I_t}, \boldsymbol{X}_{t-1}^\top\boldsymbol{\epsilon}_{t-1}\rangle_{\boldsymbol{V}_t^{-1}} - \lambda\langle \boldsymbol{x}_{I_t}, \boldsymbol{\theta}\rangle_{\boldsymbol{V}_t^{-1}}\Big)^2 + \frac{2}{s_{I_t,t-1}^2}\sum_{i=1}^{s_{I_t,t-1}} \epsilon_{I_t,i}^2}] \tag{90a}$$

$$\le \sqrt{2}\sum_{t=K+1}^{n} \boldsymbol{E}_1^{(t)} + \sqrt{2}\sum_{t=K+1}^{n} \boldsymbol{E}_2^{(t)} \tag{90b}$$

where

$$\boldsymbol{E}_1^{(t)} = \mathbb{E}[\langle \boldsymbol{x}_{I_t}, \boldsymbol{X}_{t-1}^\top\boldsymbol{\epsilon}_{t-1}\rangle_{\boldsymbol{V}_t^{-1}} - \lambda\langle \boldsymbol{x}_{I_t}, \boldsymbol{\theta}\rangle_{\boldsymbol{V}_t^{-1}}] \tag{91}$$

$$\boldsymbol{E}_2^{(t)} = \mathbb{E}[\sqrt{\frac{1}{s_{I_t,t-1}^2}\sum_{i=1}^{s_{I_t,t-1}} \epsilon_{I_t,i}^2}] \tag{92}$$

The following part is bounding $\sum_{t=K+1}^{n} \boldsymbol{E}_1^{(t)}$ and $\sum_{t=K+1}^{n} \boldsymbol{E}_2^{(t)}$ respectively.

Bounding $\sum_{t=K+1}^{n} \boldsymbol{E}_1^{(t)}$.

By Cauchy-Schwarz inequality,

$$\begin{aligned}\left(\langle \boldsymbol{x}_{I_t}, \boldsymbol{X}_{t-1}^\top \boldsymbol{\epsilon}_{t-1}\rangle_{\boldsymbol{V}_t^{-1}} - \lambda\langle \boldsymbol{x}_{I_t}, \boldsymbol{\theta}\rangle_{\boldsymbol{V}_t^{-1}}\right)^2 &\le \left(\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} \left\|\boldsymbol{X}_{t-1}^\top \boldsymbol{\epsilon}_{t-1}\right\|_{\boldsymbol{V}_t^{-1}} + \lambda \|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} \|\boldsymbol{\theta}\|_{\boldsymbol{V}_t^{-1}}\right)^2 && \text{(93a)}\\ &\le \left(\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} \left\|\boldsymbol{X}_{t-1}^\top \boldsymbol{\epsilon}_{t-1}\right\|_{\boldsymbol{V}_t^{-1}} + \|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} (\lambda^{1/2} S_2)\right)^2 \quad \text{(by (94))} && \text{(93b)}\\ &= \left(\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} \big(\|\boldsymbol{X}_{t-1}^\top \boldsymbol{\epsilon}_{t-1}\|_{\boldsymbol{V}_t^{-1}} + \lambda^{1/2} S_2\big)\right)^2 && \text{(93c)}\end{aligned}$$

where (94) is

$$\|\boldsymbol{\theta}\|_{\boldsymbol{V}_t^{-1}}^2 \le \lambda_{max}(\boldsymbol{V}_t^{-1}) \|\boldsymbol{\theta}\|_2^2 = \frac{1}{\lambda}\|\boldsymbol{\theta}\|_2^2 \le \frac{1}{\lambda} S_2^2 \tag{94}$$

By lemma C.3, with probability at least $1-\delta$,

$$\begin{aligned}\left(\langle \boldsymbol{x}_{I_t}, \boldsymbol{X}_{t-1}^\top \boldsymbol{\epsilon}_{t-1}\rangle_{\boldsymbol{V}_t^{-1}} - \lambda\langle \boldsymbol{x}_{I_t}, \boldsymbol{\theta}\rangle_{\boldsymbol{V}_t^{-1}}\right)^2 &\le \|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}^2 (L_2 \sqrt{2\log\left(\frac{\det(\boldsymbol{V}_t)^{1/2} \det(\lambda \boldsymbol{I})^{-1/2}}{\delta}\right)} + \lambda^{1/2} S_2)^2 && \text{(95a)}\\ &= \|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}^2 (L_2 \sqrt{2\log\left(\frac{(\lambda^{d-r}\prod_{j=1}^r(\sigma_j^2+\lambda))^{1/2}\lambda^{-d/2}}{\delta}\right)} + \lambda^{1/2} S_2)^2 && \text{(95b)}\\ &\le \|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}^2 (L_2 \sqrt{rlog(1+\sigma_{max}^2/\lambda) + 2\log\left(\frac{1}{\delta}\right)} + \lambda^{1/2} S_2)^2 && \text{(95c)}\end{aligned}$$

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

$$\sum_{t=K+1}^n \boldsymbol{E}_1^{(t)} \le (L_2 \sqrt{r\log(1+\sigma_{max}^2/\lambda) + 2\log\left(\frac{1}{\delta}\right)} + \lambda^{1/2} S_2) \sum_{t=K+1}^n \mathbb{E}[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}] \tag{96a}$$

Bounding $\sum_{t=K+1}^n \boldsymbol{E}_2^{(t)}$.
First separate $\sum_{t=K+1}^n \boldsymbol{E}_2^{(t)}$ by arms,

$$\begin{aligned}\sum_{t=K+1}^n \boldsymbol{E}_2^{(t)} &= \sum_{t=K+1}^n \mathbb{E}[\sqrt{\frac{1}{s_{I_t,t-1}^2} \sum_{i=1}^{s_{I_t,t-1}} \epsilon_{I_t,i}^2}] && \text{(97a)}\\ &\le \sum_{k=1}^K \mathbb{E}[\sum_{t\in\mathcal{I}_{k,n}} \sqrt{\frac{1}{s_{k,t-1}^2} \sum_{i=1}^{s_{k,t-1}} \epsilon_{k,i}^2}] && \text{(97b)}\\ &= \sum_{k=1}^K \mathbb{E}[\sum_{j=1}^{s_{k,n-1}} \sqrt{\frac{1}{j^2}(\epsilon_{k,1}^2 + \cdots + \epsilon_{k,j}^2)}] && \text{(97c)}\end{aligned}$$

For each arm,

$$\mathbb{E}[\sum_{j=1}^{s_{k,n-1}} \sqrt{\frac{1}{j^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)}] \tag{98a}$$

$$=\mathbb{E}[\sum_{j=1}^{s_{k,n-1}} \mathbb{E}[\sqrt{\frac{1}{j^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)}|\mathcal{F}_{k,j}]] \tag{98b}$$

$$\leq\mathbb{E}[\sum_{j=1}^{s_{k,n-1}} \sqrt{\mathbb{E}[\frac{1}{j^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)|\mathcal{F}_{k,j}]}] \tag{98c}$$

$$\leq\mathbb{E}[\sum_{j=2}^{s_{k,n-1}} \sqrt{\frac{1}{j^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j-1}^2)+\frac{1}{j^2}4L_2}+2\sqrt{L_2}] \quad \text{(by lemma C.4)} \tag{98d}$$

$$=\mathbb{E}[\sum_{j=1}^{s_{k,n-1}-1} \sqrt{\frac{1}{(j+1)^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)+\frac{1}{(j+1)^2}4L_2}+2\sqrt{L_2}] \tag{98e}$$

Conditioning on appropriate historical randomness $\mathcal{F}_{k,j}$ again,

$$\mathbb{E}[\sum_{j=1}^{s_{k,n-1}} \sqrt{\frac{1}{j^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)}] \tag{99a}$$

$$=\mathbb{E}[\sum_{j=1}^{s_{k,n-1}-1} \mathbb{E}[\sqrt{\frac{1}{(j+1)^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)+\frac{1}{(j+1)^2}4L_2}|\mathcal{F}_{k,j}]+2\sqrt{L_2}] \tag{99b}$$

$$\leq\mathbb{E}[\sum_{j=1}^{s_{k,n-1}-1} \sqrt{\mathbb{E}[\frac{1}{(j+1)^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)+\frac{1}{(j+1)^2}4L_2|\mathcal{F}_{k,j}]}+2\sqrt{L_2}] \tag{99c}$$

$$\leq\mathbb{E}[\sum_{j=2}^{s_{k,n-1}-1} \sqrt{\frac{1}{(j+1)^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j-1}^2)+\frac{2}{(j+1)^2}4L_2}+\frac{2}{\sqrt{2}}\sqrt{L_2}+2\sqrt{L_2}] \quad \text{(by lemma C.4)} \tag{99d}$$

$$=\mathbb{E}[\sum_{j=1}^{s_{k,n-1}-2} \sqrt{\frac{1}{(j+2)^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)+\frac{2}{(j+2)^2}4L_2}+(1+\frac{1}{2})\times 2\sqrt{L_2}] \tag{99e}$$

Applying conditional expectation given historical randomness until there is no randomness from noise,

$$\mathbb{E}[\sum_{j=1}^{s_{k,n-1}} \sqrt{\frac{1}{j^2}(\epsilon_{k,1}^2+\cdots+\epsilon_{k,j}^2)}] \leq 2\sqrt{L_2}\mathbb{E}[(1+\frac{1}{2}+\cdots+\frac{1}{s_{k,n-1}})] \tag{100a}$$

$$\leq 2\sqrt{L_2}\mathbb{E}[\log(s_{k,n-1})+1] \quad \text{(by (101))} \tag{100b}$$

$$\leq 2\sqrt{L_2}(\log n+1) \tag{100c}$$

where (101) is

$$\sum_{i=1}^{s_{k,n-1}} \frac{1}{i} \leq 1+\int_1^{s_{k,n-1}} \frac{1}{u}du = \log(s_{k,n-1})+1 \tag{101}$$

Consequently,

$$\sum_{t=K+1}^{n} \boldsymbol{E}_2^{(t)} \leq 2K\sqrt{L_2}(\log n+1) \tag{102}$$

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

$$\sum_{t=K+1}^{n} \mathbb{E}[\sqrt{\frac{RSS_{I_t,t}}{s_{I_t,t-1}^2}}] \leq \sqrt{2}(L_2\sqrt{r\log(1+\sigma_{max}^2/\lambda)+2\log\left(\frac{1}{\delta}\right)}+\lambda^{1/2}S_2)\sum_{t=K+1}^{n} \mathbb{E}[\|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}}]+2\sqrt{2}K\sqrt{L_2}(\log n+1) \tag{103}$$

□

B.3 PROOF OF LEMMA A.3

Proof. Similar version of this lemma is proven by [Abbasi-Yadkori et al., 2011] and [Lattimore and Szepesvári, 2020], following part is adapted version based on the notations in this paper. The main adaptation is using the eigenvalues of context matrix $\boldsymbol{X}_K$ under stochastic linear bandit setting. This proof requires proof of two elementary algebraic results,

$$log \frac{det(\boldsymbol{V}_n)}{det(\boldsymbol{V}_{K+1})} = \sum_{t=K+1}^{n} \log\Big(1 + \|\boldsymbol{x}_{I_t}\|^2_{\boldsymbol{V}_t^{-1}}\Big) \tag{104}$$

$$\log \frac{det(\boldsymbol{V}_n)}{det(\boldsymbol{V}_{K+1})} \leq d \log\bigg(\frac{\lambda + n\sum_{i=1}^{r} \sigma_i^2/d}{det(\boldsymbol{V}_{K+1})^{1/d}}\bigg) \tag{105}$$

Step 1: Proof of (104).
Starting from the determinant of $\boldsymbol{V}_n$,

$$\begin{aligned} det(\boldsymbol{V}_n) &= det(\boldsymbol{V}_{n-1} + \boldsymbol{x}_{I_{n-1}}\boldsymbol{x}_{I_{n-1}}^\top) && \text{(106a)}\\ &= det(\boldsymbol{V}_{n-1}^{1/2}(\boldsymbol{I} + \boldsymbol{V}_{n-1}^{-1/2}\boldsymbol{x}_{I_{n-1}}\boldsymbol{x}_{I_{n-1}}^\top \boldsymbol{V}_{n-1}^{-1/2})\boldsymbol{V}_{n-1}^{1/2}) && \text{(106b)}\\ &= det(\boldsymbol{V}_{n-1})(1 + \|\boldsymbol{x}_{I_{n-1}}\|^2_{\boldsymbol{V}_{n-1}^{-1}}) && \text{(106c)}\\ &= det(\boldsymbol{V}_{K+1}) \prod_{t=K+1}^{n} (1 + \|\boldsymbol{x}_{I_{t-1}}\|^2_{\boldsymbol{V}_{t-1}^{-1}}) && \text{(106d)} \end{aligned}$$

Then take logarithm on both side and (104) is obtained.
Step 2: Proof of (105).
By inequality between trace and determinant and notice that eigenvalues of $\boldsymbol{V}_n$ are $\sigma_1^2 + \lambda, ..., \sigma_r^2 + \lambda$ and $d - r\ \lambda$, then,

$$det(\boldsymbol{V}_n) \leq (\frac{1}{d} tr(\boldsymbol{V}_n))^d = (\frac{d\lambda + \sum_{i=1}^{r} \sigma_i^2}{d})^d \tag{107}$$

Thus,

$$\log \frac{det(\boldsymbol{V}_n)}{det(\boldsymbol{V}_{K+1})} \leq \log\bigg(\frac{1}{det(\boldsymbol{V}_{K+1})}(\frac{d\lambda + \sum_{i=1}^{r} \sigma_i^2}{d})^d\bigg) = d\log\bigg(\frac{\lambda + \sum_{i=1}^{r} \sigma_i^2/d}{det(\boldsymbol{V}_{K+1})^{1/d}}\bigg) \tag{108}$$

Step 3: Provide upper bound of sum of norms
By (104) and (105), using a analytic result $x \leq 2log(1+x) \forall x \geq 0$,then sum of the context norm under matrix $\boldsymbol{V}_t^{-1}$ can be bounded,

$$\begin{aligned} \sum_{t=K+1}^{n} \|\boldsymbol{x}_{I_t}\|^2_{\boldsymbol{V}_t^{-1}} &\leq \sum_{t=K+1}^{n} 2\log\Big(1 + \|\boldsymbol{x}_{I_t}\|^2_{\boldsymbol{V}_t^{-1}}\Big) && \text{(109a)}\\ &= 2\log\frac{det(\boldsymbol{V}_n)}{det(\boldsymbol{V}_{K+1})} \quad \text{(by (104))} && \text{(109b)}\\ &\leq 2d\log\bigg(\frac{\lambda + n\sum_{i=1}^{r}\sigma_i^2/d}{det(\boldsymbol{V}_{K+1})^{1/d}}\bigg) \quad \text{(by (105))} && \text{(109c)}\\ &= 2d\log\bigg(\frac{\lambda + \sum_{i=1}^{r}\sigma_i^2/d}{(\lambda^{d-r}\prod_{i=1}^{r}(\sigma_i^2 + \lambda))^{1/d}}\bigg) && \text{(109d)}\\ &\leq 2d\log\bigg(1 + \frac{n\sum_{i=1}^{r}\sigma_i^2}{d\lambda}\bigg) && \text{(109e)} \end{aligned}$$

Therefore, from Cauchy-Schwarz inequality,

$$\sum_{t=K+1}^{n} \|\boldsymbol{x}_{I_t}\|_{\boldsymbol{V}_t^{-1}} \leq \sqrt{(n-K)\sum_{t=K+1}^{n} \|\boldsymbol{x}_{I_t}\|^2_{\boldsymbol{V}_t^{-1}}} \leq \sqrt{2(n-K)d\log\bigg(1 + \frac{\sum_{i=1}^{r}\sigma_i^2}{d\lambda}\bigg)} \tag{110}$$

□

B.4 PROOF OF LEMMA A.4

Proof. For simplicity, focuses on the k-th arm at time t,

$$\boldsymbol{Q} := \boldsymbol{Q}_{k,t-1},\ \boldsymbol{X} := \boldsymbol{X}_{t-1},\ \boldsymbol{Y} := \boldsymbol{Y}_{t-1},\ \boldsymbol{\epsilon} := \boldsymbol{\epsilon}_{t-1},\ \boldsymbol{V} := \boldsymbol{V}_t$$

Therefore,

$$\begin{aligned}
RSS_{k,t} =& \left\| \boldsymbol{Q}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_t) \right\|_2^2 && \text{(111a)}\\
=& \left\| \boldsymbol{Q}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{V}_t^{-1}\boldsymbol{X}^\top \boldsymbol{Y}) \right\|_2^2 && \text{(111b)}\\
=& \left\| \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{V}_t^{-1}\boldsymbol{X}^\top)\boldsymbol{Y} \right\|_2^2 && \text{(111c)}\\
=& \left\| \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{V}_t^{-1}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{V}_t^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon} \right\|_2^2 \quad (\text{by } \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}) && \text{(111d)}\\
=& \left\| \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{V}_t^{-1}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{\theta} \right\|_2^2 + \left\| \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{V}_t^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon} \right\|_2^2 \\
&+ 2\boldsymbol{\theta}^\top \boldsymbol{X}^\top (\boldsymbol{I} - \boldsymbol{X}\boldsymbol{V}_t^{-1}\boldsymbol{X}^\top)\boldsymbol{Q}^\top \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{V}_t^{-1}\boldsymbol{X}^\top)\boldsymbol{\epsilon} && \text{(111e)}
\end{aligned}$$

□

B.5 PROOF OF LEMMA A.5

Proof. Follow the same simplified notations in B.4,

$$\boldsymbol{Q} := \boldsymbol{Q}_{k,t-1},\ \boldsymbol{X} := \boldsymbol{X}_{t-1},\ \boldsymbol{Y} := \boldsymbol{Y}_{t-1},\ \boldsymbol{\epsilon} := \boldsymbol{\epsilon}_{t-1},\ \boldsymbol{V} := \boldsymbol{V}_t$$

In the following part of proof, we overload the notations for singular value decomposition of matrices $\boldsymbol{X}_{t-1}$ and $\boldsymbol{X}_K$, note that this notations are only used in this proof for lemma A.5,

$$\boldsymbol{X} := \boldsymbol{X}_{t-1} = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{U} \text{ and } \boldsymbol{M} := \boldsymbol{I} - \boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top$$

Further denote $s := s_{1,t-1}$ and

$$\boldsymbol{a} := \frac{1}{\sqrt{s}}\boldsymbol{M}\boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{M}\boldsymbol{X}\boldsymbol{\theta} = (a_1, ..., a_{t-1})^\top$$

Step 1: Two sided bounds given $\boldsymbol{a}$
The key observation is that random vector $\boldsymbol{a}$ is deterministic given history $\mathcal{F}_{t-2} \cup \{\{\omega_{k,t-1,i}\}_{i=1}^{s_{k,t-1}}\}_{k=1}^K$. Recalling that noise ϵ_τ is independent of $\omega_{k,t,i}$ for $\forall \tau, k, t, i$, by conditioning on $\mathcal{F}_{t-2}$,

$$\mathbb{E}[e^{\eta\xi_t}] = \mathbb{E}[\mathbb{E}[e^{\eta \boldsymbol{a}^\top \boldsymbol{\epsilon}} | \mathcal{F}_{t-2} \cup \{\{\omega_{k,t-1,i}\}_{i=1}^{s_{k,t-1}}\}_{k=1}^K]] = \mathbb{E}[e^{\eta \sum_{i=1}^{t-2} a_i \epsilon_i} \mathbb{E}[e^{a_{t-1}\epsilon_{t-1}} | \mathcal{F}_{t-2}]] \tag{112}$$

which indicates

$$\mathbb{E}[e^{\eta^2 \sum_{i=1}^{t-2} a_i \epsilon_i} \cdot e^{\eta^2 a_{t-1}^2 L_1}] \leq \mathbb{E}[e^{\eta\xi_t}] \leq \mathbb{E}[e^{\eta^2 \sum_{i=1}^{t-2} a_i \epsilon_i} \cdot e^{\eta^2 a_{t-1}^2 L_2}] \tag{113}$$

Therefore, by conditioning on $\mathcal{F}_{t-2}, \mathcal{F}_{t-3}, ..., \mathcal{F}_1$ consecutively, the partial randomness from vector $\boldsymbol{a}$ is left to integrated by the outside expectation $\mathbb{E}$ and

$$\mathbb{E}[e^{\eta^2 \|\boldsymbol{a}\|_2^2 L_1}] \leq \mathbb{E}[e^{\eta\xi_t}] \leq \mathbb{E}[e^{\eta^2 \|\boldsymbol{a}\|_2^2 L_2}] \tag{114}$$

Step 2: Two sided bounds for $\|\boldsymbol{a}\|_2^2$
Another key observation is from eigenvalues of $\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top$ under the ridge regression procedure. It can be shown that the eigenvalues of matrix $\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top$ are $\frac{\sigma_1^2}{\sigma_1^2+\lambda}, .., \frac{\sigma_r^2}{\sigma_r^2+\lambda}$ and $t-1-r$ zeros. Thus, spectral decomposition of matrix $\boldsymbol{M}$ is, $\boldsymbol{M} = \boldsymbol{G}(\boldsymbol{I} - \boldsymbol{\Omega})\boldsymbol{G}^\top$ and $\boldsymbol{I} - \boldsymbol{\Omega}$ is diagonal matrix with with diagonal elements $\frac{\lambda}{\sigma_1^2+\lambda}, .., \frac{\lambda}{\sigma_r^2+\lambda}$ and

$t-1-r$ ones. We use $\lambda_{\max}(\boldsymbol{A})$ to denote the maximum eigenvalue of a matrix $\boldsymbol{A}$.
Thus,

$$\begin{aligned}
\|\boldsymbol{a}\|_2^2 &= \frac{1}{s}\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{M}\boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{M}\boldsymbol{M}\boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{M}\boldsymbol{X}\boldsymbol{\theta} && \text{(115a)}\\
&= \frac{1}{s}\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{X}\boldsymbol{\theta} && \text{(115b)}\\
&= \frac{1}{s}\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})^2\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{X}\boldsymbol{\theta} && \text{(115c)}
\end{aligned}$$

For upper bound,

$$\begin{aligned}
\|\boldsymbol{a}\|_2^2 &\le \boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})^2\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{X}\boldsymbol{\theta} \quad (s \ge 1) && \text{(116a)}\\
&\le \lambda_{\max}((\boldsymbol{I}-\boldsymbol{\Omega})^2)\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{X}\boldsymbol{\theta} && \text{(116b)}\\
&= \boldsymbol{\theta}^\top \boldsymbol{U}^\top \boldsymbol{\Sigma}^\top (\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{\Sigma}\boldsymbol{U}\boldsymbol{\theta} \quad (\boldsymbol{X} := \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{U} \text{ and } \lambda_{\max}((\boldsymbol{I}-\boldsymbol{\Omega})^2) = 1) && \text{(116c)}\\
&\le \boldsymbol{\theta}^\top \boldsymbol{U}^\top \boldsymbol{\Sigma}^\top (\boldsymbol{I}-\boldsymbol{\Omega})^2 \boldsymbol{\Sigma}\boldsymbol{U}\boldsymbol{\theta} \quad (\lambda_{\max}(\boldsymbol{Q}) = 1) && \text{(116d)}\\
&\le \boldsymbol{\theta}^\top \boldsymbol{U}^\top \boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}\boldsymbol{U}\boldsymbol{\theta} && \text{(116e)}\\
&\le \sigma_{\max}^2 \boldsymbol{\theta}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\theta} \quad (\lambda_{\max}(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}) = \sigma_{\max}^2) && \text{(116f)}\\
&= \sigma_{\max}^2 \|\boldsymbol{\theta}\|_2^2 && \text{(116g)}\\
&\le \sigma_{\max}^2 S_2^2 && \text{(116h)}
\end{aligned}$$

For lower bound,

$$\begin{aligned}
\|\boldsymbol{a}\|_2^2 &\ge \frac{1}{s}\lambda_{\min}((\boldsymbol{I}-\boldsymbol{\Omega})^2)\boldsymbol{\theta}^\top \boldsymbol{U}^\top \boldsymbol{\Sigma}^\top (\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{\Sigma}\boldsymbol{U}\boldsymbol{\theta} && \text{(117a)}\\
&= \frac{1}{s}(\frac{\lambda}{\sigma_{\max}^2+\lambda})^2 \boldsymbol{\theta}^\top \boldsymbol{U}^\top \boldsymbol{\Sigma}^\top (\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{G}^\top \boldsymbol{Q}\boldsymbol{G}(\boldsymbol{I}-\boldsymbol{\Omega})\boldsymbol{\Sigma}\boldsymbol{U}\boldsymbol{\theta} \quad (\lambda_{\min}((\boldsymbol{I}-\boldsymbol{\Omega})^2) = (\frac{\lambda}{\sigma_{\max}^2+\lambda})^2) && \text{(117b)}\\
&= \frac{1}{s}(\frac{\lambda}{\sigma_{\max}^2+\lambda})^2 \boldsymbol{\theta}^\top (\boldsymbol{X}-\boldsymbol{Z})^\top \boldsymbol{Q}(\boldsymbol{X}-\boldsymbol{Z})\boldsymbol{\theta} \quad (\boldsymbol{Z} := \boldsymbol{G}\boldsymbol{\Omega}\boldsymbol{\Sigma}\boldsymbol{U}) && \text{(117c)}\\
&= (\frac{\lambda}{\sigma_{\max}^2+\lambda})^2 \boldsymbol{\theta}^\top (\boldsymbol{x}_1-\boldsymbol{z}_1)(\boldsymbol{x}_1-\boldsymbol{z}_1)^\top \boldsymbol{\theta} && \text{(117d)}\\
&= (\frac{\lambda}{\sigma_{\max}^2+\lambda})^2 ((\boldsymbol{x}_1-\boldsymbol{z}_1)^\top \boldsymbol{\theta})^2 && \text{(117e)}\\
&\ge (\frac{\lambda}{\sigma_{\max}^2+\lambda})^2 S_1^2 && \text{(117f)}
\end{aligned}$$

Therefore, $\forall \eta \ge 0$,

$$\exp\left(\frac{\lambda^2}{(\sigma_{max}^2+\lambda)^2} S_1^2 L_1 \eta^2\right) \le \mathbb{E}[e^{\eta \xi_t}] \le \exp\left(\sigma_{\max}^2 S_2^2 L_2 \eta^2\right) \tag{118}$$

□

B.6 PROOF OF LEMMA A.6

Proof. This proof is inspired by the Theorem 1 and its proof in [Zhang and Zhou, 2020]. Also, an important lemma, lemma C.5, which is called Paley-Zygmund inequality is used. Since $t = 0$ is the trivial case, in the following part, we assume $t > 0$. Take

$$x := R_1 t - \frac{1}{t} \quad \forall t > 0 \tag{119}$$

Then

$$\begin{aligned}\mathbb{P}(X \geq R_1 t - \frac{1}{t}) &= \mathbb{P}(e^{tX} \geq e^{R_1 t^2 - 1}) && (120a)\\ &\geq \mathbb{P}(e^{tX} \geq e^{-1}\mathbb{E}[e^{tX}]) && (120b)\\ &\geq (1-e^{-1})^2 \frac{(\mathbb{E}[e^{tX}])^2}{\mathbb{E}[e^{2tX}]} \quad \text{(by lemma C.5)} && (120c)\\ &\geq (1-e^{-1})^2 \frac{(e^{R_1 t^2})^2}{e^{4R_2 t^2}} && (120d)\\ &= (1-e^{-1})^2 \exp\bigl(-(4R_2 - 2R_1)t^2\bigr) && (120e)\end{aligned}$$

By (119), t satisfies a quadratic equation $R_1 t^2 - xt - 1 = 0$. Since $t > 0$,

$$t = \frac{x + \sqrt{x^2 + 4R_1}}{2R_1} \tag{121}$$

Therefore,

$$\begin{aligned}\mathbb{P}(X \geq x) &\geq (1-e^{-1})^2 \exp\left(-(4R_2 - 2R_1)(\frac{x + \sqrt{x^2 + 4R_1}}{2R_1})^2\right) && (122a)\\ &= (1-e^{-1})^2 \exp\left(-\frac{2R_2 - R_1}{2R_1^2}(4x^2 + 8R_1)\right) && (122b)\\ &= (e-1)^2 e^{\frac{8R_2}{R_1} - 6} \exp\left(-\frac{4R_2 - 2R_1}{R_1^2}x^2\right) && (122c)\end{aligned}$$

□

C SUPPORTING LEMMAS

C.1 CONFIDENCE ELLIPSOID UNDER LEAST SQUARED ESTIMATION

Lemma C.1. *Under assumptions 1 and 2 and notations from (4), $\forall \alpha > 0$, with probability at least $1 - \alpha$, for all $t \geq 1$, $\boldsymbol{\theta}$ lies in the following confidence ellipsoid,*

$$\mathcal{C}_t := \{\boldsymbol{\theta} \in \mathbb{R}^d : \left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{V}_t} \leq L_2\sqrt{d\log\left(\frac{1 + tL^2/\lambda}{\alpha}\right)} + \lambda^{1/2}S_2\} \tag{123}$$

C.2 LOWER BOUND OF GAUSSIAN TAIL

Lemma C.2. *Set $Z \sim N(0, 1)$. Then, $\forall c > 0$*

$$\mathbb{P}(Z \geq t) \geq \begin{cases} \frac{b}{\sqrt{2\pi}}\exp\left(-\frac{3}{2}t^2\right) & \textit{if } t \geq b \\ \Phi(-c) & \textit{if } 0 < t < b \end{cases} \tag{124}$$

C.3 SELF-NORMALIZED BOUND FOR MARTINGALES

Lemma C.3. *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration and $\{\epsilon_t\}_{t=0}^{\infty}$ be a real-valued stochastic process such that:*
(i) ϵ_t is $\mathcal{F}_t$-measurable
(ii) ϵ_t is conditionally subgaussian with constant R, that is, for some R and $\forall t \geq 0$

$$\mathbb{E}[e^{\lambda\epsilon_t}|\mathcal{F}_{t-1}] \leq e^{\frac{\lambda^2 R}{2}} \quad \forall \lambda \in \mathbb{R}$$

Let $\{X_t\}_{t=0}^{\infty}$ be a $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable and assume $\boldsymbol{V}$ is d by d positive definite matrix. For any t, define

$$\bar{\boldsymbol{V}}_t = \boldsymbol{V} + \sum_{s=1}^{t} X_s X_s^\top \quad S_t = \sum_{s=1}^{t} \epsilon_t X_s$$

Then for any $\delta > 0$ and any $t \geq 0$, with probability at least $1 - \delta$,

$$\|S_t\|^2_{\bar{\boldsymbol{V}}_t^{-1}} \leq 2R\log\left(\frac{det(\bar{\boldsymbol{V}}_t)^{1/2}det(\boldsymbol{V})^{-1/2}}{\delta}\right)$$

C.4 SECOND MOMENT BOUND FOR SUBGAUSSIAN RANDOM VARIABLES

Lemma C.4. *Suppose random variable X is subgaussian with constant R, that is, $\mathbb{E}[e^{tX}] \leq e^{Rt^2}$ $\forall t \in \mathbb{R}$, then*

$$\mathbb{E}[X^2] \leq 4R \tag{125}$$

C.5 PALEY-ZYGMUND INEQUALITY

Lemma C.5. *Suppose X be a random variable, then when $\forall \theta \in [0, 1]$ and $\forall t \geq 0$,*

$$\mathbb{P}(e^{tX} \geq \theta\mathbb{E}[e^{tX}]) \geq (1 - \theta)^2_+ \frac{(\mathbb{E}[e^{tX}])^2}{\mathbb{E}[e^{2tX}]} \tag{126}$$

D SUPPLEMENT TO EXPERIMENTS

D.1 ALGORITHMS FOR LINREBOOT

In the paper, Algorithm 1 implements `LinReBoot` for the stochastic bandit problems. In our experiments, there are two other additional setting with linear reward function for linear bandit problem. We provide other two implementations of `LinReBoot`. The first one is `LinReBoot` for linear contextualized bandit, which is given in Algorithm 2. Another one is `LinReBoot` for linear bandit with covariates, which is given in Algorithm 3.

Algorithm 2 `LinReBoot` in Contextual Linear Bandit

```
Require: $\lambda$, $s_{1,0} = ... = s_{K,0} = 0$
  for $t = 1, ..., n$ do
    if $t < K + 1$ then
      $I_t \leftarrow t$
    else
      Get new contexts $\boldsymbol{x}_1, ..., \boldsymbol{x}_K$
      $\boldsymbol{V}_t \leftarrow \boldsymbol{X}_{t-1}^\top \boldsymbol{X}_{t-1} + \lambda \boldsymbol{I}$
      $\hat{\boldsymbol{\theta}}_t \leftarrow \boldsymbol{V}_t^{-1} \boldsymbol{X}_{t-1}^\top \boldsymbol{Y}_{t-1}$
      for $k = 1, ..., K$ do
        $e_{k,t,i} \leftarrow r_{k,i} - \boldsymbol{x}_k^\top \hat{\boldsymbol{\theta}}_t$, $\forall i \in \{s_{k,t-1}\}$
        Generate $\{\omega_{k,t,i}\}_{i=1}^{s_{k,t-1}}$
        $\tilde{\mu}_k \leftarrow \boldsymbol{x}_k^\top \hat{\boldsymbol{\theta}}_t + s_{k,t-1}^{-1} \sum_{i=1}^{s_{k,t-1}} \omega_{k,t,i} e_{k,t,i}$
      end for
      $I_t \leftarrow \arg\max_{k \in [K]} \tilde{\mu}_k$
    end if
    $s_{I_t,t} \leftarrow s_{I_t,t-1} + 1$ and $s_{k,t} \leftarrow s_{k,t-1}$. $\forall k \neq I_t$
    Pull arm $I_t$ and get reward $r_{I_t, s_{I_t}}$
    $\boldsymbol{X}_t \leftarrow \begin{bmatrix} \boldsymbol{X}_{t-1} \\ \boldsymbol{x}_{I_t}^\top \end{bmatrix}$ and $\boldsymbol{Y}_t \leftarrow \begin{bmatrix} \boldsymbol{Y}_{t-1} \\ r_{I_t, s_{I_t}} \end{bmatrix}$
  end for
```

D.2 EXPERIMENTAL SETTING

This part provides the detailed description of the experimental setting in Section 6. There are three settings in our experiment: Stochastic Linear Bandit, Contextual Linear Bandit and Linear Bandit with Covariates. Each of them has own synthetic data generation procedure which is described in the following parts.

Stochastic Linear Bandit. In the first experiment, we compare `LinReBoot` to other linear bandit algorithms under stochastic linear bandit described in Section 2. The `LinReBoot` is implemented as the efficient version of algorithm 1. Our experiment is conducted under three choice of dimension d including 5, 10 and 20. The number of arm in this setting is 100. True parameter $\boldsymbol{\theta}$ has norm 1 and is generated from uniform distribution by entries. In other word, generate $\theta_i \sim U(-0.5, 0.5), \forall i \in [d]$ and then shrink $\|\boldsymbol{\theta}\|_2 = 1$. Context features $\boldsymbol{x}_1, ..., \boldsymbol{x}_K$ are generated by $\boldsymbol{x}_{ik} \sim U(0, 1), \forall i \in [d], k \in [K]$ and normalized to $\|\boldsymbol{x}_k\|_2 = 1$. By the normalization of $\boldsymbol{\theta}$ and $\{\boldsymbol{x}_k\}_{k=1}^K$, the true mean of reward is bounded by 1, making `LinPHE` and `LinGIRO` become easier to choose a reasonable bounds for reward. Noise ϵ_t is generated from $N(0, 0.1)$. At each choice of d, our results are averaged over 100 randomly chosen environment and we evaluate all algorithms under the exact same environment with horizon length 10000. Regularization parameter λ is chosen as 0.1 through out the experiments. Tuning parameters for each algorithms are described in Appendix D.6.

Contextual Linear Bandit. In the second experiment, we compare `LinReBoot` to other linear bandit algorithms under linear bandit with uncertain/random context. We experiment with several dimensions d including 5, 10 and 20. The number of arm is 100. True parameter is generated by the same way as stochastic linear bandit setting in Section 6.1. Contexts of arm k has distribution $N_d(\boldsymbol{\nu}_k, 1/(2K)\boldsymbol{I})$ where $\boldsymbol{\nu}_k$ is generated by following:

Algorithm 3 `LinReBoot` in Linear Bandit wit Covariates

```
Require: $\lambda$, $s_{1,0} = ... = s_{K,0} = 0$
  for $t = 1, ..., n$ do
    if $t < K + 1$ then
      $I_t \leftarrow t$
    else
      Get new context $\boldsymbol{x}_t$
      for $k = 1, ..., K$ do
        $\boldsymbol{V}_{k,t} \leftarrow \boldsymbol{X}_{k,t-1}^\top \boldsymbol{X}_{k,t-1} + \lambda \boldsymbol{I}$
        $\hat{\boldsymbol{\theta}}_{k,t} \leftarrow \boldsymbol{V}_{k,t}^{-1} \boldsymbol{X}_{k,t-1}^\top \boldsymbol{Y}_{k,t-1}$
        $e_{k,t,i} \leftarrow r_{k,i} - \boldsymbol{x}_k^\top \hat{\boldsymbol{\theta}}_t$, $\forall i \in \{s_{k,t-1}\}$
        Generate $\{\omega_{k,t,i}\}_{i=1}^{s_{k,t-1}}$
        $\tilde{\mu}_k \leftarrow \boldsymbol{x}_k^\top \hat{\boldsymbol{\theta}}_t + s_{k,t-1}^{-1} \sum_{i=1}^{s_{k,t-1}} \omega_{k,t,i} e_{k,t,i}$
      end for
      $I_t \leftarrow \arg\max_{k \in [K]} \tilde{\mu}_k$
    end if
    $s_{I_t,t} \leftarrow s_{I_t,t-1} + 1$ and $s_{k,t} \leftarrow s_{k,t-1}$. $\forall k \neq I_t$
    Pull arm $I_t$ and get reward $r_{I_t, s_{I_t}}$
    $\boldsymbol{X}_{I_t,t} \leftarrow \begin{bmatrix} \boldsymbol{X}_{I_t,t-1} \\ \boldsymbol{x}_t^\top \end{bmatrix}$ and $\boldsymbol{Y}_{I_t,t} \leftarrow \begin{bmatrix} \boldsymbol{Y}_{I_t,t-1} \\ r_{I_t, s_{I_t}} \end{bmatrix}$
  end for
```

$\boldsymbol{\nu}_{ik} \sim U(0,1), \forall i \in [d] \quad k \in [K]$ and normalized to $\|\boldsymbol{\nu}_k\|_2 = 1$. Note that $\boldsymbol{\nu}_k$ are predefined before the simulation. Noise ϵ_t is generated from $N(0, 0.5)$. Remaining environment setting is designed as the same in Section 6.1: number of simulation is 100, horizon length is 10000, regularization parameter $\lambda = 0.1$. Most hyperparameters are chosen as the same as Section 6.1 except for the reward bounds in `LinPHE` and `LinGIRO`. Detailed description is provided in Appendix D.6.

Linear Bandit with Covariates Our last experiment is conducted under the setting of linear bandit with covariates. Again, we experiment with several dimensions d including 5, 10 and 20 while the number of arms is 10 in this setting. True parameter $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K$ are generated one by one and each of them is generated in the following way: (1) Choose an integer $n_- \leq d$ by $n_- \sim Binomial(d, 1/2)$ and randomly sample n_- integers from 1 to d, these n_- integers indicates the entries that has negative direction in $\boldsymbol{\theta}_k$. (2) generate a d-dimensional vector with n_{-1} entries are -1 and remaining $n_+ := d - n_-$ entries are 1 by the n_- integers sampled in the previous step. (3) Each entries will add a random perturbation from $U(-0.95, 0.95)$ to make the magnitude of the each entry is spread between 0.05 to 1. (4) The resulting vector will be normalized by $\|\boldsymbol{\theta}_k\| = \frac{k}{K}$, indicating the norm of the true parameters $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K$ are designed as $\frac{1}{K}, ..., 1$. Contexts are sampled from $N(\boldsymbol{0}, \boldsymbol{I})$ which is independent of arms. Noise ϵ_t is generated from $N(0, 0.1)$. Remaining environment setting is designed as the same in Section 6.1 or Section 6.2: number of repetition is 100 and horizon length is 10000 as well as $\lambda = 0.1$. Reward bounds in `LinPHE` and `LinGIRO` are chosen based on the noise variance and other algorithms are designed as the same as the previous two settings. More specific description is provided in Appendix D.6.

D.3 `LINREBOOT` IN STOCHASTIC LINEAR BANDIT

The algorithm of `LinReBoot` is described in Algorithm 1 and steps of our `LinReBoot` and its efficient implementation under Gaussian bootstrap weights are summarized in Section 3. For the parameter tuning of `LinReBoot`, our first step candidate set for σ_ω in `LinReBoot` is $\{0.05, 0.1, 0.2, 0.5, 1.0\}$. The following result, figure 2, shows that the values 0.05, 0.1, 0.2 are not enough for resampling exploration under all three choice of context dimension. However, we notice that too large σ_ω leads to slow convergence even if it is indeed sub-linear. Thus 0.5 is the best result under our stochastic linear bandit setting. We decide to do the further fined tuning, using the candidate set $\{0.3, 0.4, 0.6, 0.7\}$ and the result is shown in figure 3. It is clear that $\sigma_\omega = 0.3$ is the best choice when $d = 5$ while $\sigma_\omega = 0.4$ is the best choice under the setting of $d = 10$. When $d = 20$, we conclude that $\sigma_\omega = 0.5$ is better than other candidates. As a result, our experiment in Section 6 choose $\sigma_\omega = 0.3$ for $d = 5$, choose $\sigma_\omega = 0.4$ for $d = 10$

and choose $\sigma_\omega = 0.5$ for $d = 20$.

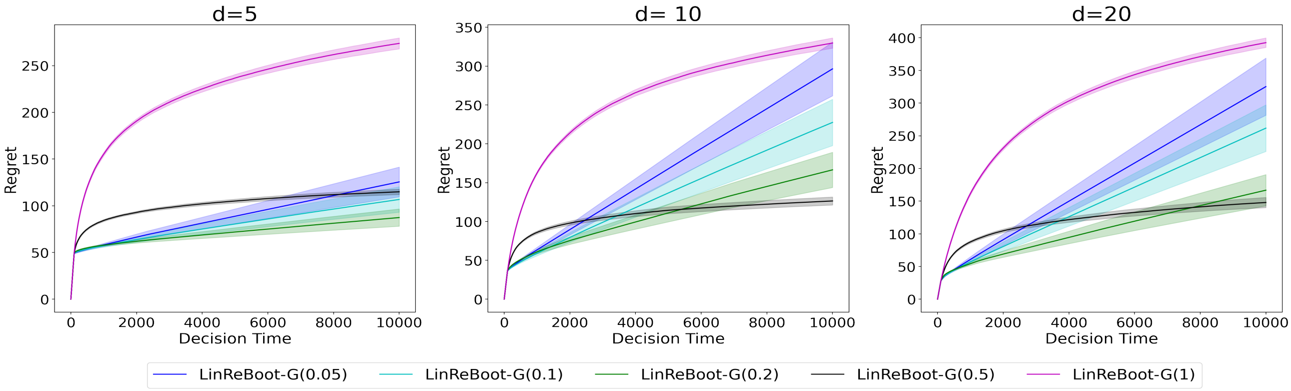

Figure 2: First Step Tuning for `LinReBoot-G` under Stochastic Linear Bandit. The x axis is round t and y axis is cumulative regret. The candidate set for σ_ω is $\{0.05, 0.1, 0.2, 0.5, 1.0\}$ and these three plots from left to right corresponds to $d = 5$, $d = 10$ and $d = 20$ respectively.

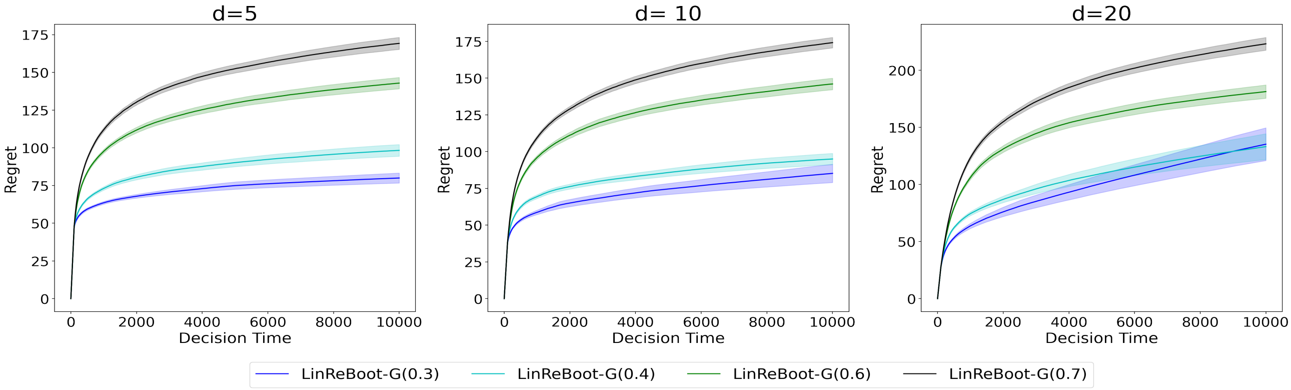

Figure 3: Second Step Tuning for `LinReBoot-G` under Stochastic Linear Bandit. The x axis is round t and y axis is cumulative regret. The candidate set for σ_ω is $\{0.3, 0.4, 0.6, 0.7\}$ and these three plots from left to right corresponds to $d = 5$, $d = 10$ and $d = 20$ respectively.

D.4 LINREBOOT IN CONTEXTUAL LINEAR BANDIT

The algorithm 2 is `LinReBoot` under Contextual Linear Bandit. It is almost the same as algorithm 1 while the algorithm new requires the random contexts from each arm at each round t. For the parameter tuning of `LinReBoot`, our candidate set is designed as $\{0.05, 0.1, 0.2, 0.5, 1.0\}$ and the following result shows that $\sigma_\omega = 0.05$ is the best choice for all three design of context dimension d. Thus our experiment choose $\sigma_\omega = 0.05$ for three possible d under this setting of Contextual Linear Bandit.

D.5 LINREBOOT IN LINEAR BANDIT WITH COVARIATES

The last version of `LinReBoot` is `LinReBoot` under Linear Bandit with Covariates which is provided as algorithm 3. This algorithm is different from the previous two version due to the different task under linear bandit with covariates which requires the algorithm not only the estimation of the target parameter $\boldsymbol{\theta}$, but also detection of which arm a context belongs to. For the parameter tuning of `LinReBoot`, our candidate set is designed as $\{0.05, 0.1, 0.2, 0.5, 1.0\}$ and the following result shows that $\sigma_\omega = 1$ is the best choice for the cases including $d = 5$ and $d = 10$. When $d = 20$, $\sigma_\omega = 1$ is still acceptable while $\sigma_\omega = 0.5$ might be preferred one. In fact, it must be pointed out that when d becomes larger, the performances among difference choice of σ_ω becomes smaller

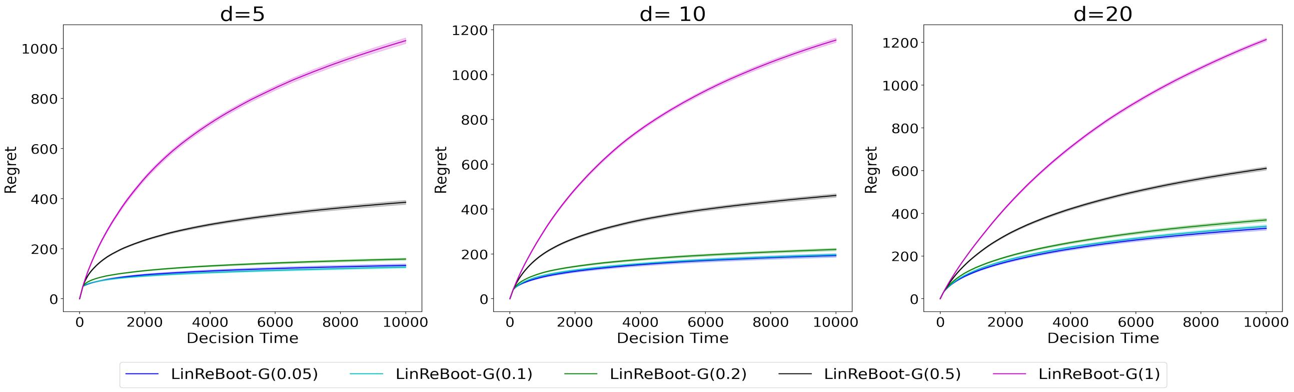

Figure 4: Tuning for `LinReBoot-G` under Contextual Linear Bandit. The x axis is round t and y axis is cumulative regret. The candidate set for σ_ω is $\{0.05, 0.1, 0.2, 0.5, 1.0\}$ and these three plots from left to right corresponds to $d = 5$, $d = 10$ and $d = 20$ respectively.

and larger σ_ω might be worse for larger d. At the end, our experiment choose $\sigma_\omega = 1$ for $d = 5$ and $d = 10$ and $\sigma_\omega = 0.5$ for $d = 20$.

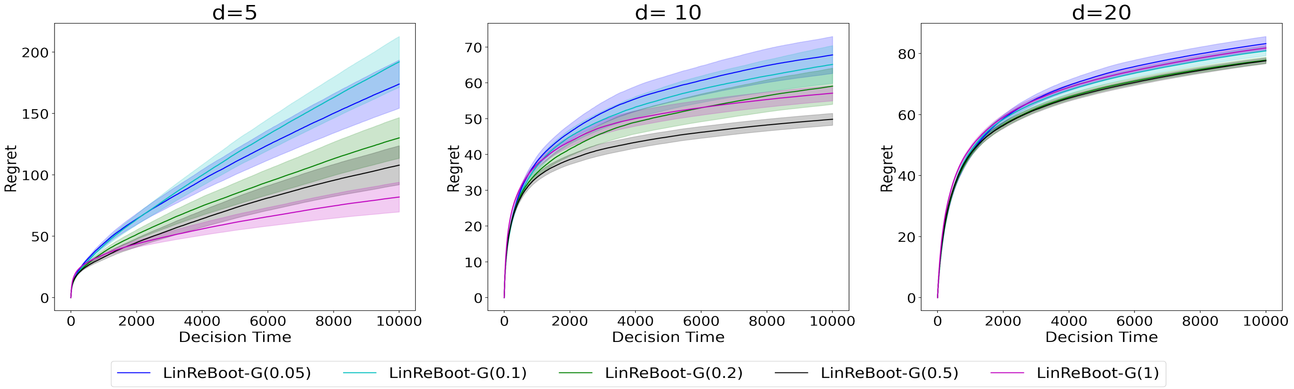

Figure 5: Tuning for `LinReBoot-G` under Linear Bandit with Covariates. The x axis is round t and y axis is cumulative regret. The candidate set for σ_ω is $\{0.05, 0.1, 0.2, 0.5, 1.0\}$ and these three plots from left to right corresponds to $d = 5$, $d = 10$ and $d = 20$ respectively.

D.6 OTHER LINEAR BANDIT ALGORITHMS

Linear Thompson Sampling with Gaussian Prior (`LinTS-G`). Thompson Sampling is a classic algorithm [Thompson, 1933] which requires only that one can sample from the posterior distribution over plausible problem instances (for example, values or rewards). Linear Thompson sampling is a Bayesian linear bandit algorithm which has studied by lots of previous works such as [Agrawal and Goyal, 2013a,b, Riquelme et al., 2018, Russo et al., 2018]. In our experiment, we mainly depends on [Agrawal and Goyal, 2013b, Lattimore and Szepesvári, 2020] for implementing Linear Thompson sampling with Gaussian prior. There is almost the same among three different settings in our work. The only difference is that stochastic linear bandit and Contextual Linear Bandit is estimating/sampling parameter shared among arms while parameters are estimated/sampled using the rewards and contexts from only one arm in the setting of linear bandit with covariates. As mentioned in section 6, the Gaussian prior variance is chosed as $\frac{1}{\lambda} = 10$ by Bayesian perspective of ridge regression model.

Linear Thompson Sampling with Inverse Gamma Prior (`LinTS-IG`). Another version of Thompson sampling under linear bandit is adding inverse gamma prior [Honda and Takemura, 2014, Riquelme et al., 2018, Bishop, 2006]. We implement this inverse gamma version based on the detail suggested as [Riquelme et al., 2018]. Similar to `LinTS-G`, three settings share almost the same `LinTS-IG` and only difference is the parameters in linear bandit with covariates setting are estimated/sampled using the data from one arm. Moreover, Gaussian prior parameter is designed as $\frac{1}{10}$ which match our overall design for regularization $\lambda = 0.1$ and the inverse gamma prior parameters is suggest by [Riquelme et al., 2018]. More specifically, by $\sigma_0^2 \approx \alpha/(\alpha-1)$ where $\sigma_0^2\tau^2 = 10$ is the initial variance on diagonal for sampling our target parameter $\boldsymbol{\theta}$, $\tau^2 = 5$ is Gaussian prior parameter and $\alpha = 2$ is the prior parameter for inverse gamma.

Linear Perturbed-History Exploration (`LinPHE`). A well designed algorithm for stochastic linear bandit under bounded reward is `LinPHE` [Kveton et al., 2020a]. The idea is also inspired from successfully adding exploration under Multi-armed bandit setting [Kveton et al., 2019a]. Our experiments use the suggested hyperparameter $a = 0.5$. However, since the original work is only designed for stochastic linear bandit with bounded rewards, we extended it to more general settings with Gaussian rewards. The detail is provided as follow. In stochastic linear bandit setting, based on our experimental design, true mean of each arm is bounded by 1 and noise variance is set as 0.1, indicating that we have high probability that the reward will be bounded by $1 + 3/\sqrt{10}$ on both sides. In the setting of Contextual Linear Bandit, the original efficient implementation from [Kveton et al., 2020a] can not be used. But we modified by drawing a number from Binomial distribution $Binomial(\lceil a(t-1)\rceil, 1/2)$ at round t and divided this number into $t-1$ parts randomly which are added as perturbation of rewards. The reward is bounded by $1 + 3/\sqrt{2}$. For the last setting, linear bandit with covariates, similar to previous setting, we modify by using Binomial distribution to adapt the non-integer value of a but this time we need to apply the perturbed history by arm, that is using $Binomial(\lceil as_{k,t-1}\rceil, 1/2)$ for all $k \in [K]$. The reward is bounded by 1.3.

Linear Garbage In Reward Out (`LinGIRO`). Garbage In, Reward Out(`GIRO`) is a bootstrapping based algorithm designed for multi-armed bandit with bounded reward [Kveton et al., 2019b]. Since its idea of bootstrapping and perturbation on mean estimation is highly related to our residual bootstrapping exploration, it is worthy to compare with this classical bootstrapping based algorithm. But like `PHE`, it is originally designed for multi-armed bandit and we need to extend it to linear bandit setting with unbounded reward and then apply it to three settings in our experiment. Previous work [Kveton et al., 2019b, Wang et al., 2020] suggest the conservative choice of a is 1, indicating adding one high pseudo reward and one low pseudo reward at each round. The detail, which is almost the same as previous modification for `LinPHE`, is provided as follow. In stochastic linear bandit and linear bandti with random context settings, we bootstrapping the previous reward-context pair and use the new sample to do least squared estimation. After pulling arm, $2a$ pseudo reward-context pairs are added: one is current context with reward upper bound and the other one is current context with reward lower bound. For the last setting, linear bandit with covariates, the only difference is that the bootstrapping is conducted by arm and the pseudo reward-context pairs are added to one arm at each round. The reward bound is chosen as $1 + 3/\sqrt{10}$ for stochastic linear bandit and $1 + 3/\sqrt{2}$ for the setting of Contextual Linear Bandit while 1.3 is chosen for linear bandit with covariates setting.

Linear Upper Confidence Bound (`LinUCB`). Upper Confidence Bound(`UCB`) is a important type of bandit algorithms which is widely used. `LinUCB` is the version extended to linear bandit setting [Abbasi-Yadkori et al., 2011, Chu et al., 2011]. Since its popularity and usage, we believe it should be involved in our experiment and we implement `LinUCB` mainly relying on [Abbasi-Yadkori et al., 2011, Lattimore and Szepesvári, 2020]. The confidence level is chosen as 95% which matches the traditional statistical sense. Moreover, `LinUCB` is almost the same among three different setting. The only difference is stochastic linear bandit and Contextual Linear Bandit are using the rewards and contexts to estimate one target parameter, like (4) in our paper while the last setting, linear bandit with covariates, requires the least squared estimation to be done by arms.

D.7 COMPUTATION EFFICIENCY

D.7.1 Efficient Implementation of `LinReBoot-G`

Section 3.3 discusses about why `LinReBoot-G` can be implemented efficiently. This section provides a further illustration and implementation in practice. First recall $\tilde{\boldsymbol{\mu}}^{(t)} = (\tilde{\mu}_{1,t}, \ldots, \tilde{\mu}_{K,t})^\top$ is conditional distributed as

$$\tilde{\boldsymbol{\mu}}^{(t)} | \mathcal{F}_{t-1} \sim N_K(\hat{\boldsymbol{\mu}}^{(t)}, \boldsymbol{\Sigma}_\omega^{(t)}) \tag{127}$$

where $\hat{\boldsymbol{\mu}}^{(t)} = (\hat{\mu}_{1,t}, \dots, \hat{\mu}_{K,t})^\top = \boldsymbol{X}_K \hat{\boldsymbol{\theta}}_t$ and $\boldsymbol{\Sigma}_\omega^{(t)}$ is a diagonal matrix with diagonal elements $\sigma_\omega^2 s_{k,t-1}^{-2} RSS_{k,t}$. Note that $\boldsymbol{\Sigma}^{(t)}$ can be computed by $\hat{\boldsymbol{\mu}}^{(t)}$ and vectors,

$$\begin{aligned}
\boldsymbol{r}_1^{(t)} &:= (\sum_{i=1}^{s_{1,t-1}} r_{1,i}, \dots, \sum_{i=1}^{s_{K,t-1}} r_{K,i})^\top, \\
\boldsymbol{r}_2^{(t)} &:= (\sum_{i=1}^{s_{1,t-1}} r_{1,i}^2, \dots, \sum_{i=1}^{s_{K,t-1}} r_{K,i}^2)^\top, \\
\boldsymbol{s}^{(t)} &:= (s_{1,t-1}, \dots, s_{K,t-1})^\top.
\end{aligned}$$

These vectors can be updated incrementally by the above illustration. To sum up, when bootstrap weights are Gaussian, the efficient implementation for computing $\tilde{\mu}_{k,t}$ at round t has steps as follow,

- Compute $\boldsymbol{V}_t$, $\hat{\boldsymbol{\theta}}_t$ and $\hat{\boldsymbol{\mu}}^{(t)} = \boldsymbol{X}_K \hat{\boldsymbol{\theta}}_t$
- Compute $\boldsymbol{\Sigma}^{(t)}$ using $\hat{\boldsymbol{\mu}}^{(t)}$, $\boldsymbol{r}_1^{(t)}$, $\boldsymbol{r}_2^{(t)}$ and $\boldsymbol{s}^{(t)}$
- Sample $\tilde{\boldsymbol{\mu}}^{(t)} \sim N_K(\hat{\boldsymbol{\mu}}^{(t)}, \boldsymbol{\Sigma}^{(t)})$
- Pull arm I_t and get its corresponding reward r_{I_t}
- Update $\boldsymbol{r}_1^{(t+1)}$, $\boldsymbol{r}_2^{(t+1)}$ and $\boldsymbol{s}^{(t+1)}$

D.7.2 Computational Cost

The computation cost of linear bandit algorithms involved in our experiment are listed in the following table. Each running time is for one horizon with length 10000. The settings are also provided in Appendix D.2 and the description of algorithms are provided in Appendix D.6.

Model		Run time (seconds)					
Setting	d	`LinReBoot`	`LinTS-G`	`LinTS-IG`	`LinGIRO`	`LinPHE`	`LinUCB`
Stochastic Linear Bandit	5	3.2	1.8	2.2	6.5	4.0	6.2
Stochastic Linear Bandit	10	3.5	2.1	2.5	10.3	4.7	6.6
Stochastic Linear Bandit	20	4.8	3.9	3.8	24.6	5.6	7.4
Contextualized Linear Bandit	5	3.3	1.8	2.2	6.5	4.0	6.3
Contextualized Linear Bandit	10	3.5	2.1	2.5	10.2	4.7	6.6
Contextualized Linear Bandit	20	3.8	3.1	3.6	24.1	5.2	6.9
Linear Bandit with Covariates	5	1.4	7.8	12.9	10.3	5.2	1.2
Linear Bandit with Covariates	10	1.5	9.4	14.1	11.5	5.9	1.4
Linear Bandit with Covariates	20	1.6	14.2	18.9	15.2	7.4	1.5

Table 3: Computational Cost for Linear Bandit Algorithms