Parallelizing Contextual Linear Bandits

Jeffrey Chan\textsuperscript{*} 
UC Berkeley 
chanjed@berkeley.edu

Aldo Pacchiano\textsuperscript{*} 
UC Berkeley 
pacchiano@berkeley.edu

Nilesh Tripuraneni\textsuperscript{*} 
UC Berkeley 
nilesh.tripuraneni@berkeley.edu

Yun S. Song 
UC Berkeley 
yss@berkeley.edu

Peter Bartlett 
UC Berkeley 
peter@berkeley.edu

Michael I. Jordan 
UC Berkeley 
jordan@cs.berkeley.edu

Abstract

Standard approaches to decision-making under uncertainty focus on sequential exploration of the space of decisions. However, simultaneously proposing a batch of decisions, which leverages available resources for parallel experimentation, has the potential to rapidly accelerate exploration. We present a family of (parallel) contextual linear bandit algorithms, whose regret is nearly identical to their perfectly sequential counterparts—given access to the same total number of oracle queries—up to a lower-order “burn-in” term that is dependent on the context-set geometry. We provide matching information-theoretic lower bounds on parallel regret performance to establish our algorithms are asymptotically optimal in the time horizon. Finally, we also present an empirical evaluation of these parallel algorithms in several domains, including materials discovery and biological sequence design problems, to demonstrate the utility of parallelized bandits in practical settings.

1 Introduction

Traditional approaches to adaptive design and search often require resource-intensive modeling and significant time to execute. As the cost of experimentation has steadily dropped, it is the latter that has become a primary constraint in many domains. For example, within the field of biology, advances in synthesis and high-throughput sequencing have shifted the focus from top-down mechanistic modeling towards iterative, data-driven search algorithms [Baker et al., 2018]. In protein engineering, the directed evolution approach [Arnold, 1998]—which mimics evolution via iterative rounds of measurement of modified sequences—has been successful in improving therapeutic antibodies [Hawkins et al., 1992] and changing substrate specificity [Shaikh and Withers, 2008]. Within the realm of information technology, learning in real-world interactive tasks is often well-suited to adaptive, data-driven search; examples include recommender systems which generate movie suggestions for users on the basis of binary feedback for previously recommended movies [Bietti et al., 2018]. An important property of many of these examples is that learning must be done via bandit feedback. Given the state of a system (i.e., a particular protein sequence or user) the corresponding value or feedback signal is only observed for that particular state.

In these settings of limited feedback, there is a critical need to use adaptive approaches to intelligently query the design space. Bandit methods and Bayesian optimization have provided such adaptivity in a variety of applications, providing a principled way to navigate the exploration-exploitation tradeoff in the design space. It can still be infeasible, however, to explore these spaces efficiently, in particular as measured by the overall real time needed for a sequential algorithm to run.

A clear path forward involves making use of parallelism, in particular the massive parallelism that is provided by modern computing hardware. Indeed, in many applications of interest, it is often feasible to perform parallel measurements (or batch queries) simultaneously. In protein engineering, batch queries can be as large as $10^6$ sequences while each query takes months to measure [Sinai and Kelsic, 2020]. Similarly, large-scale recommender systems can often make multiple, concurrent interactions with different users [Agarwal et al., 2016]. Parallelism provides a mechanism to utilize further hardware for progress, while not increasing the amount of real time needed to learn. The drawback of making batch queries at a fixed time is that queries within a batch will necessarily be less informed than sequential queries—since their choice cannot benefit from the results of other experiments within the same batch.

\textsuperscript{*} Equal contribution.
It is thus of value to characterize the utility of parallelism in adaptive design and search. In the current paper, we tackle this problem in a simplified setting that allows us to study some of the fundamental challenges that characterize the problem—that of contextual multi-armed bandits. In particular, we ask: can parallelism provide the same benefit as sequentiality through time in this class of adaptive decision-making problems? Perhaps surprisingly, and in accord with work in related settings, we show that the answer is often yes. We consider a setting in which $P$ distinct processors/machines perform simultaneous queries in batches, $\{x_{t,p}\}_{p=1}^{P}$, of size $P$, over $T$ distinct rounds, to an approximately linear, noisy reward oracle $r_{t,p} \approx x_{t,p}^\top \theta^* + \epsilon_{t,p}$. Importantly, the reward feedback for all $P$ processors in a given round is only observed after the entire batch of all $P$ actions has been selected. In this setting, we consider the price of parallelism with respect to a single, perfectly sequential agent querying the same reward oracle over $TP$ rounds. Our results show that up to a burn-in term independent of time, the worst-case regret of parallel, contextual linear bandit algorithms can nearly match their perfectly sequential counterpart. Informally, under the standard bandit setting where the signal-to-noise ratio is taken to be $\Theta(1)$, and $d$ is the dimension of the action-space, our results show that (parallel) variants of optimistic linear bandit algorithms can achieve the following aggregate regret:

\[
\tilde{O}(P \cdot \kappa + d\sqrt{TP}),
\]

where $\kappa$ is a term capturing the geometric complexity of the sequence of contexts. For arbitrary sequences of contexts we have that $\kappa = \tilde{O}(d)$, but for sequences of contexts with additional geometric structure we provide sharper instance-dependent bounds for $\kappa$. The second term in the former regret matches that of a purely sequential agent playing for $TP$ rounds, $\tilde{O}(d\sqrt{TP})$. The first term is a burn-in term that represents the price of parallelism; it is subleading whenever $T \geq \tilde{O}(P\kappa^2/d^2)$. However, we note for many applications it will be the case that $P \gg T$ in the regimes of interest—so understanding the scaling of $\kappa$ is an important question. Our work is motivated by the class of design problems in which statistical sample complexity is the primary object of interest. In between batch queries at different time steps, the cost of inter-processor communication and computation is often negligible. In this setting we make the following contributions:

- We introduce a family of parallelized linear bandit algorithms applicable in the general setting of contextual bandits (which encompasses potentially changing and infinite action sets), building on classic, optimistic algorithms such as Linear UCB and Linear Thompson sampling.

- We present a unified treatment of these algorithms which allows a fine-grained understanding of the interaction of the scales of the underlying parameters, model misspecification, parallelism, and geometry of the context set. In particular, our analysis highlights the scale of the burn-in term that is proportional to $P$. Although asymptotically negligible as $T \to \infty$, this term can be significant in the regime where we have $TP \gg poly(d)$, but $P \gg T$—which is realistic for many real applications.

- We provide lower bounds on regret for the misspecified, linear bandit problem in the parallel setting, confirming that our algorithms are optimal up to logarithmic factors as the time horizon tends to $T \to \infty$.

- We present a comprehensive experimental evaluation of our parallel algorithms on a suite of synthetic and real problems. Through our theory and experiments, we show the importance of explicitly introducing diversity into the action selections of parallel bandit algorithms since it often leads to practical performance gains.

### 1.1 Related Work

Parallel sequential decision-making problems often occur in high-throughput experimental design for synthetic biology applications. For protein engineering, Romero et al. [2013] employ a batch-mode Gaussian Process Upper Confidence Bound (GP-UCB) algorithm, while Belanger et al. [2019] use the batched Bayesian optimization via parallelized Thompson sampling algorithm proposed in Kandasamy et al. [2018]. Angermueller et al. [2020] propose a portfolio optimization method layered atop an ensemble of optimizers. Finally, Sinai et al. [2020] utilize a heuristic, batched greedy evolutionary search algorithm for design.

Contextual linear bandit algorithms are also often used in dynamic settings on large, multi-user platforms for problems such as ad placement, web search, and recommendation [Swaminathan and Joachims, 2015, Bietti et al., 2018]. As important as such domains are in real-world applications, systematic empirical research has lagged, however, due to the paucity of publicly available nonstationary data sources.

Several investigations of the utility of parallelism in sequential decision-making problems have been conducted in the framework of Gaussian process (GP)-based Bayesian optimization. Desautels et al. [2014] study a Bayesian framework and show that a lazy GP-bandit algorithm initialized with uncertainty sampling can achieve a parallel regret nearly matching...
the corresponding sequential regret, up to a “burn-in” term independent of time. Later, Kathuria et al. [2016] showed that diversity induced by determinantal point processes (DPP) can induce additional, useful exploration in batched/parallel Bayesian optimization in a comparable setting. Kandasamy et al. [2018] establish regret bounds for Bayesian optimization parallelized via Thompson sampling algorithm, obtaining qualitatively similar theoretical results to Kathuria et al. [2016], Desautels et al. [2014].

A related line of work studies bandit learning under delayed reward feedback. Chapelle and Li [2011] provide an early, empirical investigation of Thompson sampling under fixed, delayed feedback. Several theoretical works establish regret bounds for bandit algorithms under known and unknown stochastic delayed feedback models [Joulani et al., 2013, Vernade et al., 2017, 2020, Zhou et al., 2019]. Here the reward information is delayed by a random time interval from the time the action is proposed, which is distinct from our setting in which the batching of the queries results in a fixed time delay. In the setting of finite-armed contextual bandits Dudik et al. [2011], Delayed Policy Elimination algorithm satisfies a regret bound of the form $\tilde{O}(\sqrt{m(T + \tau)})$, where $m$ denotes the number of actions, $T$ the problem horizon and $\tau$ the delay. However, these guarantees require i.i.d. stochastically generated contexts and require access to a cost-sensitive classification oracle for their elimination-based protocol.

A more closely related line of work studies distributed bandit learning under various models of limited communication between agents (which can make decisions in parallel). Hillel et al. [2013] and Szorenyi et al. [2013] study the regret of distributed arm-selection algorithms under restricted models of communication between parallel agents in the setting of multi-armed bandits. Korda et al. [2016] and Wang et al. [2019] provide distributed confidence-ball algorithms for linear bandits in peer-to-peer networks with a focus on limiting communication complexity. However, the algorithms in Korda et al. [2016] and Wang et al. [2019] both require intra-round communication of rewards, which is incompatible with the batch setting that we study in this work. Moreover these works are mostly concerned with the regime $T \gg P$, reducing their practical relevance.

Our work provides a class of algorithms and regret analysis for the general linear contextual bandit problem in the batch setting. In this setting we aim to design theoretically sound and practically useful algorithms for parallel linear bandits. Accordingly, we construct several algorithms that we analyze both theoretically and empirically.

1.2 Preliminaries

Throughout, we use bold lower-case letters (e.g., $x$) to refer to vectors and bold upper-case letters to refer to matrices (e.g., $X$). The norm $\| \cdot \|$ appearing on a vector or matrix refers to its $\ell_2$ norm or spectral norm, respectively. We write $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$ for positive semi-definite matrix $\Sigma$. $\langle x, y \rangle$ is the Euclidean inner product. Generically, we will use “hatted” vectors and matrices (e.g., $\hat{\alpha}$ and $\hat{B}$) to refer to (random) estimators of their underlying population quantities. We also use the bracketed notation $[n] = \{1, \ldots, n\}$. We will use $\gtrsim, \lesssim$ and $\asymp$ to denote greater than, less than, and equal to up to a universal constant and use $\tilde{O}$ to denote an expression that hides polylogarithmic factors in all problem parameters. Our use of $O$, $\Omega$, and $\Theta$ is otherwise standard.

Formally, we consider the (parallel) contextual linear bandit setting where at each round $t$, the $p$-th bandit learner receives a context $\mathcal{X}_{t,p} \subset \mathbb{R}^d$ and a master algorithm $\mathcal{A}$ commands each learner to select an action $x_{t,p} \in \mathcal{X}_{t,p}$ on the basis of all the past observations. Given an (unobserved) function $f(\cdot)$ each learner simultaneously receives a noisily generated reward:

$$r_{t,p} = f(x_{t,p}) + \xi_{t,p},$$

(1)

where $\xi_{t,p}$ is an i.i.d. noise process and $f(x_{t,p})$ is approximately linear (i.e., $f(x_{t,p}) \approx x_{t,p}^\top \theta^*$ for some unobserved $\theta^*$). The goal of the master algorithm/processors is to utilize its access to the sequence of rewards $r_{t,p}$ and joint control over the sequence of action selections $x_{t,p}$ (which depends on the past sequence to events) to minimize the parallel regret:

$$\mathcal{R}(T, P) = \sum_{p=1}^P \left( \sum_{t=1}^T f(x_{t,p}^*) - f(x_{t,p}) \right),$$

where $x_{t,p}^* = \arg\max_{x \in \mathcal{X}_{t,p}} f(x)$. We also introduce the notion of the best regret across processors:

$$\mathcal{R}_\star(T) = \min_{p \in [P]} \left( \sum_{t=1}^T f(x_{t,p}^*) - f(x_{t,p}) \right),$$

which captures the performance of the best processor in hindsight. We note the following relationship which follows immediately by definition of the aforementioned regrets:
Remark 1. The parallel regret and best regret satisfy
\[ R_s(T) \leq \frac{R(T, P)}{P}. \]

In the special case that there is a fixed (finite) context for all time across all processors (i.e., \( X_{t,p} = \mathcal{X} \)), as is the case for design problems, the simple regret is useful:
\[ R_s(T, P) = f(x^*) - f(x_{T+1,1}). \]
The simple regret captures the suboptimality of a choice \( x_{T+1} \), given by a next-step policy \( \pi(\cdot) \) at the \((T+1)\)st time against the single best choice \( x^* \in \mathcal{X} \).

Remark 2. There exists a randomized next-step policy \( \pi(\cdot) \) (depending on the sequence of \( x_{t,p} \)) at the \((T+1)\)st timestep such that the simple regret satisfies
\[ E[|R_s(T, P)|] \leq E\left[ \frac{R(T, P)}{TP} \right], \]
when there is a fixed global context \( \mathcal{X} \) for all \( t \in [T], p \in [P] \).

For our analysis, we make the following standard assumptions on the bandit problem in (1).

**Assumption 1 (Subgaussian Noise).** The noise variables \( \xi_{t,p} \) are \( R \)-subgaussian for all \( t \in [T] \) and \( p \in [P] \).

**Assumption 2 (Bounded Covariates).** For all contexts \( X_{t,p}, t \in [T], p \in [P] \), the actions are norm-bounded by a known upper bound:
\[ \|x\| \leq L, \quad \forall x \in X_{t,p}. \]

**Assumption 3 (Almost-Linear Rewards).** The function \( f(\cdot) \) is \( \epsilon \)-close to linear in that for all contexts \( X_{t,p} \) and for all \( x \in X_{t,p} \), there exists a parameter \( \theta^* \) such that
\[ |f(x) - x^T \theta^*| \leq \epsilon. \]
Further, this underlying parameter \( \theta^* \) satisfies the norm bound:
\[ \|\theta^*\| \leq S. \]

In the context of the above conditions we define the signal-to-noise ratio as:
\[ \text{SNR} = \left( \frac{LS}{R} \right)^2, \tag{2} \]
in analogy with the classical setting of offline linear regression. Note that while we allow \( \epsilon \) to be arbitrary, our guarantees are only non-vacuous when \( \epsilon \) is suitably small. Thus \( f(x) \) should be thought of a function that is nearly linear. Although we only consider linear algorithms in this paper, we note that our methods can easily be generalized using finite-dimensional feature expansions and random feature approximations [Rahimi et al., 2007] to increase the flexibility of our model class.

## 2 Parallelizing Linear Bandits

We consider several algorithms for linear bandits which provably achieve a nearly perfect parallel speed-up. Most of these algorithms are inspired by the optimism principle. Such algorithms maintain a running estimate of \( \hat{\theta}_t \approx \theta^* \) and the empirical covariance matrix of queried observations:
\[ V_{t,p} = \lambda I_d + \sum_{a=1}^{t-1} \sum_{b=1}^{P} x_{a,b} x_{a,b}^T + \sum_{k=1}^{p-1} x_{t,k} x_{t,k}^T, \]

1This follows from a similar reduction from sequential regret to simple regret applicable to multi-armed bandits in Lattimore and Szepesvári [2020, Proposition 33.2].
Algorithm 1 Parallel LinUCB

Input: \(P, T, R, S, L, \lambda, \epsilon\), DR Routine.

1. for \(t = 1 : T\) do
2. Compute \(\hat{\theta}_t, V_{t,1},\) and \(\beta_t\) as in (6) and (7).
3. for \(p = 1 : P\) do
4. Given \(X_{t,p}\), compute \(y_{t,p} \leftarrow \arg \max_{x \in X_{t,p}} \max_{\theta \in C_{t,0}} x^\top \theta\) for \(C_{t,0}(\hat{\theta}_t, V_{t,1}, \beta_t, \epsilon)\) as in (6).
5. end for
6. Compute \(\tilde{V}_{t+1,1} = V_{t,1} + \sum_{p=1}^{P} y_{t,p} y_{t,p}^\top\).
7. if \(\tilde{V}_{t+1,1} \preceq 2V_{t,1}\) then
8. for \(p = 1 : P\) do
9. Set \(x_{t,p} \leftarrow y_{t,p}\) and query \(x_{t,p}\) to receive reward \(r_{t,p}\) } Executed In Parallel
10. end for
11. else
12. Set \(\{x_{t,p}\}_{p=1}^P \leftarrow \text{DR}(F_{t,p})\)
13. for \(p = 1 : P\) do
14. Query \(x_{t,p}\) to receive reward \(r_{t,p}\) } Executed In Parallel
15. end for
16. end if
17. end for

Algorithm 2 Parallel Lazy LinUCB

Input: \(P, T, R, S, L, \lambda, \epsilon\), DR Routine.

1. for \(t = 1 : T\) do
2. Compute \(\hat{\theta}_t,\) and \(\beta_t\) as in (6) and (7).
3. for \(p = 1 : P\) do
4. Compute \(\tilde{V}_{t,p} = V_{t,1} + \sum_{k=1}^{p-1} y_{t,k} y_{t,k}^\top\).
5. Given \(X_{t,p}\), compute \(y_{t,p} \leftarrow \arg \max_{x \in X_{t,p}} \max_{\theta \in C_{t,p}} x^\top \theta\) for \(C_{t,p}(\hat{\theta}_t, \tilde{V}_{t,p}, \beta_t, \epsilon)\) as in (6).
6. end for
7. Compute \(\tilde{V}_{t+1,1} = V_{t,1} + \sum_{p=1}^{P} y_{t,p} y_{t,p}^\top\).
8. if \(\tilde{V}_{t+1,1} \preceq 2V_{t,1}\) then
9. for \(p = 1 : P\) do
10. Set \(x_{t,p} \leftarrow y_{t,p}\) and query \(x_{t,p}\) to receive reward \(r_{t,p}\) } Executed In Parallel
11. end for
12. else
13. Set \(\{x_{t,p}\}_{p=1}^P \leftarrow \text{DR}(F_{t,p})\)
14. for \(p = 1 : P\) do
15. Query \(x_{t,p}\) to receive reward \(r_{t,p}\) } Executed In Parallel
16. end for
17. end if
18. end for

which are used to construct a confidence ellipsoid in parameter space where \(\lambda\) is a regularization parameter. The best parameter in this set of plausible parameters—which allows the estimation of a maximum hypothetical reward—is used to optimistically guide exploration in the feature space. We use the notation \(F_{t,p-1}\) to define the filtration of all events that have occurred up until and including the revelation of the context \(X_{t,p}\). Despite the topical differences in these algorithms, there is a common thread which ties together their analyses in the parallel setting in our framework:\(^2\)

\(^2\)The choice of two in the upper bound here is arbitrary and can be replaced with any universal constant \(c > 1\) without changing our results up to constants.
**Condition 1.** [Critical Covariance Inequality] An (estimated) covariance matrix $V_{t,p} \in \mathcal{F}_{t,p-1}$ is said to satisfy the critical covariance inequality at round $t$ for processor $p$ if

$$V_{t,1} \preceq V_{t,p} \preceq 2V_{t,1}. \quad (3)$$

We refer to any round $t$ for which the aforementioned inequality does not hold for any $p \in \{2, \ldots, P\}$ as a doubling round.

For the standard purely sequential setting of linear bands (i.e., $P = 1$), in each interaction round we select $x_t$, gain the information $r_t$, update our model, and iterate. In a parallel setting, on the other hand, we must jointly select $x_{t,p}$ for all $p \in [P]$ before seeing any additional rewards in round $t$. Condition 1 ensures that up a factor of two, there is no direction along which our estimate of the covariance changes too rapidly within a single round. If we imagine unrolling the parallel dimension $p$ sequentially across time (i.e., consider the lexicographic ordering of pairs $(t, p)$), Condition 1 ensures the covariance estimate is quasi-static intra-round. The significance of this simple condition is that once we receive reward $r_{t,p}$ for $p \in [P]$ at the end of the round $t$, it is nearly as if we received the reward $r_{t,p}$ immediately after selecting $x_{t,p}$. We later provide more intuition as how this factors into our analysis and what algorithms satisfy this property.

Finally, in the event a particular round $t$ is a doubling round, we allow our algorithms to call a doubling round routine, $(x_{t,p})_{p=1}^P \leftarrow \text{DR}(\mathcal{F}_{t,p})$, where $\mathcal{F}_{t,p}$ is the $\sigma$–algebra containing all information regarding past contexts, rewards, and selected actions. The doubling round routine allows our algorithm to make a different choice of actions instead of the actions $(y_{t,p})_{p=1}^P$ suggested by the optimistic algorithm if round $t$ is a doubling round. In many cases, this routine can simply be taken to be the identity map and non-trivial parallelism gains are still obtained. However, in Section 3.4 we provide an example of a nontrivial choice of doubling round routine that can exploit the geometry of the context sets for improved performance.

Throughout the following sections our regret upper bounds are stated with high probability. That is, we claim $\mathcal{R}(T, P) \leq \text{RATE}$, with probability at least $1 - O(\delta)$, where $\text{RATE}$ has at most $O(\log(\frac{1}{\delta}))$ dependence on $\delta$. However, under Assumptions 2 and 3, the total regret can always be trivially bounded as $O(\text{LSTP})$. Thus, our high-probability regret bounds can be easily converted to upper bounds in expectation at the cost of only logarithmic factors by setting $\delta \propto (\frac{1}{\text{LSTP}})^2$, for example.

### 2.1 Linear Upper Confidence Bound (UCB) Algorithms

We first show how two natural algorithms, which are parallelized variants of the classic LinUCB algorithm of Abbasi-Yadkori et al. [2011], obtain the optimal sequential regret, up to a burn-in term, when they satisfy Condition 1. In the following we use

$$\theta_t = \text{argmin}_\theta \frac{1}{2} \sum_{a=1}^{t-1} \sum_{b=1}^{P} (r_{a,b} - x_{a,b}^\top \theta)^2 + \lambda||\theta||_2^2 \quad (4)$$

to refer to the least-squares estimator using data until round $t$, and

$$C_{t,p}(\hat{\theta}_t, V_{t,p}, \beta_t, \epsilon) = \{\theta : ||\theta - \hat{\theta}_t||_{V_{t,p}} \leq \sqrt{\beta_t(\delta)} + \sqrt{(t-1)P}\epsilon\} \quad (5)$$

for $V_{t,p} = \lambda I_d + \sum_{a=1}^{t-1} \sum_{b=1}^{P} x_{a,b} x_{a,b}^\top + \sum_{k=1}^{p-1} x_{t,k} x_{t,k}^\top \quad (6)$

$$\sqrt{\beta_t(\delta)} = R \sqrt{\log \left( \frac{\det(V_{t,0})}{\lambda^d \delta^2} \right)} + \sqrt{\lambda} S \leq R \sqrt{d \log \left( \frac{1 + t P L^2 / \lambda}{\delta} \right)} + \sqrt{\lambda} S \quad (7)$$

to refer to a confidence ellipsoid which uses $\hat{\theta}_t$ as its center, but allows the matrix $V_{t,p}$ (which includes intra-round updates) to modulate the exploration directions. As we will argue, these confidence ellipsoids satisfy the “optimism” property in that they contain the unobserved $\theta^*$ with high probability. We note that the confidence ellipsoids also include an addition term proportional to $\epsilon$ to accommodate the nonlinearity of the objective.

Algorithm 1 exploits parallelism in a simple fashion. It finds the best optimistic upper bound on the reward in round $t$ for each context and allocates each of its $P$ parallel resources to querying those arms. Although this seems redundant (if for example all the context sets are equal), this strategy can still provide benefit because when Algorithm 1 queries a common arm $x$, $P$ times, the effective variance in the noise of the observed reward is reduced by a factor of $1/P$. As Theorem 1 shows, even this simple parallelism-enabled noise reduction strategy can provide significant benefit.
Algorithm 2 naturally encourages diversity in its parallel exploration strategy within a given round. Within a round \( t \), Algorithm 2 queries new actions \( x_{t,p} \) in the standard LinUCB fashion via an optimistic approach with an important caveat: while the covariance matrix is sequentially updated using the queried actions, a “stale” mean estimate (with data from the first \( t - 1 \) round) is used in the construction of its corresponding confidence ellipsoid since the rewards \( r_{t,p} \) are not available intra-round. This update strategy exploits a key property of the linear regression estimator used to construct the confidence ellipsoid. The covariance matrix used to modulate exploration across directions does not depend on the rewards \( r_{t,p} \) (although the mean estimate does). One pitfall of such an approach is that shrinking the predicted variance in the absence of corresponding observations can lead to an “overconfident” algorithm which may incorrectly exclude the true parameter \( \theta^* \) from its confidence set. We compensate this aggressive exploration strategy by also using a lazy threshold width \( \sqrt{2\beta_2(\delta)} \) which is inflated by a small multiplicative factor to allay this effect.

Our main result follows which bounds the regret of both Algorithm 1 and Algorithm 2.

**Theorem 1.** Let Assumptions 1, 2 and 3 hold and let \( D_t \) denote the event that round \( t \) is a doubling round (see Condition 1). Then, for any choice of doubling-round routine \( DR \), the regret of both Algorithms 1 and 2 satisfy

\[
\mathcal{R}(T, P) \leq O\left(LSP \cdot \sum_{t=1}^T [D_t]\right) + O\left(\sqrt{dTTP \max(R\sqrt{d} + \sqrt{\lambda S} + \sqrt{TP}\epsilon, LS)}\right),
\]

with probability at least \( 1 - \delta \).

We now make several comments to interpret the result.

- The second term in Theorem 1 represents the near-optimal (up to log factors) regret a single processor could achieve in a total of \( TP \) rounds interacting with a bandit instance in a purely sequential fashion. The first term in Theorem 1 represents the price of parallelization. Theorem 1 hints at the prospect of obtaining a near-optimal worst-case regret (as \( T \to \infty \)) if the estimate of the covariance can be stabilized intra-round (so the algorithms do not suffer too many doubling rounds). Perhaps surprisingly we show how the Condition 1 can be enforced such that the first term in Theorem 1 is independent of \( T \) and \( d \) in the sequel. This can even be done in several cases when the choice of the doubling-round routine \( DR \) is taken to be the identity map.

- Theorem 1 also explicitly represents the scales of noise, covariates and parameters, and model misspecification (i.e., the values in Assumptions 1, 2 and 3) instead of merely asserting that these quantities are \( \Theta(1) \) as is standard in the bandits literature (see Lattimore and Szepesvári [2020]). Explicitly representing these quantities allows a fine-grained understanding of the interplay between parallelism and quantities such as \( \text{SNR} \), which may vary from application to application.

- The cost of misspecification in the regret is high. An \( \epsilon \)-level of misspecification contributes a linearly-scaling regret of \( \epsilon\sqrt{dTP} \) in both \( T \) and \( P \). Balancing the trade-off between the variance \( O(Rd\sqrt{TP}) \) and misspecification bias \( O(\epsilon\sqrt{dTP}) \) is more nuanced in the setting of sequential learning compared to that of i.i.d. supervised learning. In particular Theorem 1 suggests that especially at long time scales and large levels of parallelism, the errors compounded by using an inflexible feature set can easily overwhelm the reduced exploration needed when regressing in a low-dimensional space. If large values of \( P \) are desired the use of a flexible feature expansions (with a higher effective dimension) may be desirable.

Section 3 provides several sufficient conditions under which the covariance stability can be naturally satisfied. Further, Section 4 provides instances that show that in the regime of sufficiently large \( TP \) the regret bounds in Theorem 1 are unimprovable.

### 2.2 Linear Thompson Sampling (TS) Algorithms

Following Section 2.1, we show that two parallel variants of linear Thompson sampling (TS) [Agrawal and Goyal, 2013b,a, Abeille et al., 2017] can obtain optimal sequential regret, up to a burn-in term, when they satisfy Condition 1. Despite being among the oldest bandit methods, and having poorer worst-case theoretical guarantees compared to their deterministic

---

3. This is closely related to the fact that the conditional covariance of jointly Gaussian random variable only depends on the covariance of the original matrix.

4. The standard choice of regularizer in the sequential setting is taken as \( \lambda = L^2 \).

5. Note this is only non-vacuous when \( \epsilon \ll LS/\sqrt{d} \) since the regret can always be trivially bounded by \( O(LSTP) \) under our assumptions.
countershorts, Thompson sampling methods often achieve excellent performance in practice [Thompson, 1933, Russo et al., 2017].

We use the same notation as in Section 2.1 and refer to the least-squares estimator using data until round \( t \) as \( \hat{\theta}_t \), and defined as in Equation 4, to describe our parallel TS variants. Algorithm 3 is the corresponding Thompson sampling version of Algorithm 1. During each round \( p \in [P] \), Algorithm 3 samples \( P \) independent candidate parameters \( \{\tilde{\theta}_{t,p}\}_{p=1}^P \) to induce exploration over the parameter set. Solving the optimization problems, \( \argmax_{x \in \mathcal{X}_{t,p}} x^\top \tilde{\theta}_{t,p} \), over the possibly processor dependent contexts \( \mathcal{X}_{t,p} \), induces \( P \) distinct arm choices for the different processors \( \{x_{t,p}\}_{p=1}^P \). While Algorithm 3 does not update the covariance of the sampling distribution while producing each of the \( P \) candidate models \( \{\tilde{\theta}_{t,p}\}_{p=1}^P \), Algorithm 4 mirrors Algorithm 2. Algorithm 4 proceeds by sampling the model parameters sequentially across the different processors, but updates the sampling covariance matrix in between each sampling step intra-round.

Our main result is the following bound for Algorithms 3 and 4:

**Theorem 2.** Let Assumptions 1, 2 and 3 hold and let \( D_t \) denote the event that round \( t \) is a doubling round (see Condition 1). Then, for any choice of doubling-round routine \( DR \), the regret of both Algorithms 3 and 4 satisfy

\[
\mathcal{R}(T, P) \leq O\left( LSP \cdot \sum_{t=1}^T [D_t] + \tilde{O}\left( d \sqrt{TP} \left( 1 + \frac{L^2}{\lambda} \right) \left( R\sqrt{d} + S\sqrt{\lambda} + \sqrt{TP\epsilon} \right) \right) \right),
\]

with probability at least \( 1 - 3\delta \), whenever \( \delta \leq \frac{1}{6} \).

We make several comments on the result.

- **Theorem 2** has a similar flavor to Theorem 1—allowing for near perfect parallelization (as \( T \to \infty \)) relative to the sequential Thompson sampling algorithm when the first term is independent of \( T \). As before, Condition 1 can be enforced such that the first term in independent of \( T \).
- Note that even in the sequential setting, the regret of linear Thompson sampling suffers an extra multiplicative \( \sqrt{d} \) factor relative to LinUCB [Abeille et al., 2017]. As Theorem 2 shows, the parallel variants of Thompson inherit this \( \sqrt{d} \) factor. Despite this extra dimension-dependent factor (which is needed to maintain the optimism property when using the noisily sampled candidate models for exploration) the performance of the Thompson sampling algorithm is often excellent in practice.

### 3 Stable Covariances

We demonstrate how Condition 1 can be naturally satisfied in a variety of settings. The bounds presented here have a common structure which takes the form:

\[
\mathcal{R}(T, P) \leq \tilde{O}\left( \underbrace{R\sqrt{SNR \cdot P \cdot \kappa}}_{\text{burn-in}} + \mathcal{R}(TP, 1) \right),
\]

where \( \mathcal{R}(TP, 1) \) captures the regret of that learning algorithm operating in purely sequential fashion and \( \kappa \) is a geometry-dependent constant. The price of parallelism is factored into the burn-in term. Although this term is subleading as \( T \to \infty \), for many applications of interest (such as protein engineering) it may be the case that \( P \sim T \) or \( P \gg T \). Thus, understanding the value of \( \kappa \) as a function of the geometry of the context set is a question of interest.

#### 3.1 Arbitrary Contexts

Our first result shows in the general setting of linear contextual bandits, a uniform bound holds on the number of doubling rounds for any sequence of actions selected.

**Lemma 1.** Let \( \{\mathcal{X}_{t,p}\}_{t=1, p=1}^{T,P} \) be an arbitrary sequence of contexts. If Assumption 2 holds and the covariance is estimated as in (7), then almost surely over any sequence \( x_{t,p} \) of selected covariates, the number of total doubling rounds is bounded by at most

\[
\left\lceil \frac{d}{\log(2)} \log \left( 1 + \frac{TPL^2}{d\lambda} \right) \right\rceil.
\]

*The standard bound for TS in the sequential setting can be obtained by setting \( \lambda = L^2 \).*
An application of Lemma 12 shows this implies \( \det(V_{t+1}) > 2 \det(V_{t}) \). So if \( k \) doubling rounds elapse by the end of time \( T \) it must be the case that \( \det(V_{T+1}) > 2^k \det(V_{1}) \). However by Lemma 11, for any sequence of selected covariates satisfying Assumption 2, we have that \( \log \left( \frac{\det(V_{T+1})}{\det(V_{1})} \right) > k \log(2) \). Hence it follows \( k < \left[ \frac{\log(2)}{\log(2)} \right] \).
Lemma 1 ensures that in broad generality, the number of doubling rounds is bounded by $\tilde{O}(d)$. Instantiating our previous results then gives the following.

**Corollary 1.** In the setting of Theorem 1, choosing $\lambda = L^2$ and taking the doubling-routine DR as the identity map, the regret of both Algorithms 1 and 2 satisfy

$$R(T, P) \leq \tilde{O}\left(R \cdot \left(P^{\sqrt{SNR}} \cdot d + \sqrt{dTP(\sqrt{d} + \sqrt{SNR} + \frac{\epsilon}{R}\sqrt{TP})}\right)\right),$$

with probability at least $1 - \delta$. In the setting of Theorem 2 with the choice $\lambda = L^2$ and also taking the doubling-routine DR as the identity map, the regret of Algorithms 3 and 4 satisfy

$$R(T, P) \leq \tilde{O}\left(R \left(P^{\sqrt{SNR}} \cdot d + \sqrt{dTP(\sqrt{d} + \sqrt{SNR} + \frac{\epsilon}{R}\sqrt{TP})}\right)\right),$$

(11)

with probability at least $1 - 3\delta$, whenever $\delta \leq \frac{1}{6}$.

**Proof of Corollary 1.** Note that by Lemma 1, we must have that $\sum_{t=1}^{T} I[D_t] \leq \left\lceil \frac{d}{\log(2)} \log \left(1 + \frac{TPL^2}{d\lambda}\right)\right\rceil$. An application of Theorems 1 and 2 gives the result after choosing $\lambda = L^2$.

We can interpret the result as follows.

- The baseline regret of Linear UCB and Lazy Linear UCB interacting in a purely sequential fashion (with the standard choice of regularizer $\lambda = L^2$) for $TP$ rounds scales as,

$$R(TP, 1) \leq \tilde{O}\left(R\sqrt{dTP} \cdot (\sqrt{d} + \sqrt{SNR} + \frac{\epsilon}{R}\sqrt{TP})\right),$$

with an analogous expression inflated by an extra $\sqrt{d}$ holding for Thompson sampling and Lazy Thompson sampling. The canonical normalizations for the noise, parameter, and covariates in the literature assume $SNR = R = \Theta(1)$ as well as $\epsilon = 0$, which simplifies to the oft-stated (and optimal) regret $\tilde{O}(d\sqrt{TP})$. In this case, if $T \geq \Omega(P)$, the regret of our parallel algorithms nearly matches the optimal worst-case regret of a single sequential agent.

- The result in Corollary 1 suggests that when we opt for a large choice of $P$, parallelism is particularly beneficial in the small SNR regime. The fact that a low fidelity data-generation process benefits parallelism may seem counter-intuitive. This property arises in part because the algorithms we consider are optimistic – so environments with large SNR have large parameter norms necessitating the usage of large confidence sets which induce more regret.

### 3.2 Arbitrary Contexts with a Stable Initializer

Our next result shows in the general setting of contextual bandits, choosing a large regularizer guarantees a well-conditioned initialization for the empirical covariance that persists for all time.

**Lemma 2.** Let Assumption 2 hold. Then almost surely over any sequence of selected $x_{t,p}$, the covariance estimate in (7) satisfies Condition 1 for all $t \in [T]$ if $\lambda = PL^2$.

**Proof.** $V_{t,p} \succeq V_{t,1}$ follows immediately since $V_{t,p} - V_{t,1} = \sum_{k=1}^{p-1} x_{t,k}x_{t,k}^\top \succeq 0$. To show the second conclusion, consider an arbitrary vector $v \in \mathbb{S}^d$. Then $v^\top V_{t,p} v = v^\top V_{t,1} v + v^\top \sum_{k=1}^{p-1} x_{t,k}x_{t,k}^\top v \leq v^\top V_{t,0} v + pL^2 \leq v^\top V_{t,1} v + v^\top (\lambda I) v \leq 2v^\top V_{t,1} v$ under the setting of the result, which gives the conclusion.

Lemma 2 shows with appropriate choice of regularizer, no round is ever a doubling round for the standard estimate of the empirical covariance. Combining with our previous guarantees gives the following corollary.

**Corollary 2.** In the setting of Theorem 1, if the regularizer is chosen as $\lambda = PL^2$ and the doubling-routine DR taken as the identity map, then the regret of Algorithms 1 and 2 satisfy

$$R(T, P) \leq \tilde{O}\left(R\sqrt{dTP} \cdot (\sqrt{d} + \sqrt{SNR}\sqrt{P} + \frac{\epsilon}{R}\sqrt{TP})\right),$$

with probability at least $1 - \delta$. In the setting of Theorem 2 if $\lambda = PL^2$ and the doubling-routine DR taken as the identity map, the regret of Algorithms 3 and 4 satisfy

$$R(T, P) \leq \tilde{O}\left(Rd\sqrt{TP} \left(\sqrt{d} + \sqrt{SNR}\sqrt{P} + \frac{\epsilon}{R}\sqrt{TP}\right)\right),$$

with probability at least $1 - 3\delta$, whenever $\delta \leq \frac{1}{6}$.

10
**Proof of Corollary 2.** The result follows by combining Lemma 2 with Theorems 1 and 2 since no round is a doubling round for all \( t \in [T] \).

We now make several comments on the result.

- A large regularizer ensures a stable covariance estimate (and no doubling rounds), but comes at the cost of introducing additional bias (depending on \( P \)) into the least-squares estimate. Such bias factors into the second term of the regret since it must be compensated for by using wider confidence sets in the algorithms. Moreover, such a bias results in error that scales multiplicatively in regret as \( \sqrt{T} \).

- For \( P \cdot \text{SNR} \leq \mathcal{O}(d) \), by Corollary 2, the regret of our parallel algorithms nearly matches the optimal worst-case regret of a corresponding single sequential agent.

### 3.3 Finite Context Sets

Next we show how structure in the context set (in this case, finiteness of the action space) can be leveraged to bound the number of doubling rounds without modifying the standard choices of the hyperparameters for the optimistic algorithms considered here.

**Lemma 3.** Let \( \mathcal{X}_{t,p} \subset \mathcal{X} = \{ x_i \}_{i=1}^m \) for all \( t \in [T] \) and \( p \in [P] \), where \( \mathcal{X} \) is a finite set of vectors and \( |\mathcal{X}| = m \). If Assumption 2 holds and the covariance is estimated as in (7), then almost surely over any sequence \( x_{t,p} \) of selected covariates, the number of total doubling rounds is bounded by at most \( m \cdot \log_2(|P|) \).

**Proof.** We first recall from Condition 1 that in a doubling round at time \( t \)

\[
V_{t,P+1} \preceq 2V_{t,1}.
\] (12)

If \( \mathcal{X}_{i,p} \subset \mathcal{X} = \{ x_i \}_{i=1}^m \), then

\[
V_{t,1} = \lambda I_d + \sum_{i=1}^m w_{t,1}(i) x_i x_i^\top
\]

where \( w_{t,1}(i) \) corresponds to the number of times action \( x_i \) has been played by all processors up to and including all \( P \) actions played at time \( t - 1 \). Whenever \( t \) is a doubling round and (12) holds, there must exist \( i \in [m] \) such that

\[
w_{t,P+1}(i) > 2w_{t,1}(i),
\] (13)

since otherwise for all \( i \) it would hold that \( w_{t,P+1}(i) \leq 2w_{t,1}(i) \) implying that \( V_{t,P+1} \preceq 2V_{t,1} \), contradicting (12). Observe that each time (13) holds, \( w_{t,P+1}(i) - w_{t,1}(i) > w_{t,1}(i) \), implying that during round \( t \) arm \( i \) was pulled more times than the total number of times it has been pulled thus far. Then, for all \( i \in [m] \), the difference \( w_{t,P+1}(i) - w_{t,1}(i) \leq P \) and therefore for any \( i \in [m] \) condition (13) cannot hold for more than \( \lfloor \log_2(P) \rfloor \) iterations. Since there are only \( m \) underlying vectors in the contexts the result follows.

This observation implies a bound for our parallel bandit algorithms over finite context sets:

**Corollary 3.** In the setting of Theorem 1, assume the context sets \( \mathcal{X}_{t,p} \subset \mathcal{X} \) for all \( t \in [T], p \in [P] \) where \( |\mathcal{X}| = m \) is a finite set of vectors. Then choosing \( \lambda = L^2 \) and taking the doubling-routine DR as the identity map, the regret of both Algorithms 1 and 2 satisfy

\[
\mathcal{R}(T, P) \leq \mathcal{O} \left( R \left( P \sqrt{\text{SNR}} \cdot m + \sqrt{dTP} \sqrt{\text{SNR}} + \frac{\epsilon}{R} \sqrt{TP} \right) \right),
\]

with probability at least \( 1 - \delta \). In the setting of Theorem 2 with the choice \( \lambda = L^2 \) and also taking the doubling-routine DR as the identity map, the regret of Algorithms 3 and 4 satisfy

\[
\mathcal{R}(T, P) \leq \mathcal{O} \left( R \left( P \sqrt{\text{SNR}} \cdot m + dT \sqrt{\text{SNR}} + \frac{\epsilon}{R} \sqrt{TP} \right) \right),
\] (14)

with probability at least \( 1 - 3\delta \), whenever \( \delta \leq \frac{1}{6} \).

**Proof of Corollary 3.** By Lemma 3, we have that \( \sum_{t=1}^T \mathbb{1}[D_t] \leq m \lfloor \log_2(P) \rfloor \). An application of Theorems 1 and 2 gives the result after choosing \( \lambda = L^2 \).
Algorithm 5 Random Exploration Subroutine

Input: $\pi_{t,p}(\cdot)$.
1: for $p = 1 : P$ do
2: Sample $x_p \sim \pi_{t,p}(\cdot)$
3: end for
4: return $\{x_p\}_{p=1}^P$

We now interpret this result.

- As before, the second terms in Corollary 3 captures the effect of perfect parallel speed-up—it is the regret that a single agent would achieve playing for a total of $TP$ rounds while the first term in Corollary 3 is a burn-in term that bounds the number of rounds which may not be doubling rounds.

- Relative to Corollary 1 the scaling on the burn-in term, which is proportional to $P$, contains a factor $m$ instead of $d$. Thus, if $m \ll d$, Corollary 3 shows how the algorithms considered here can take advantage of additional structure in the context set to mitigate the the cost of parallelism.

- The proof of Corollary 3 does not exploit any correlation structure in the action set to bound the number of doubling rounds—for example, clusters of arms that are tightly bunched together in the global context space under a suitable notion of distance. Under natural conditions on the action set, sharper instance-dependent bounds may be possible.

3.4 Rich Context Sets

Finally we consider a setting in which we have a sequence of context sets $\mathcal{X}_{t,p}$ with regularity structure that are formally defined as follows.

**Definition 1.** The contexts $\mathcal{X}_{t,p}$ and distributions $\pi_{t,p}(\cdot)$ are a pair of rich exploration contexts/distributions if there exist $\chi^2$, $\pi_{\text{max}}^2$ and $\pi_{\text{min}}^2$ such that,

- $x \sim \pi_{t,p}(\cdot)$ satisfies $x \in \mathcal{X}_{t,p}$ almost surely for all $t \in [T], p \in [P]$.

- For any sequence $x_{t,p} \in \mathcal{X}_{t,p}$ selected by the bandit algorithm for $t \in [T]$, $\sum_{p=1}^P x_{t,p} x_{t,p}^\top \leq P \chi^2 I$.

- Given $x_{t,p} \sim \pi_{t,p}(\cdot)$, with population mean and covariance defined as, $E[x_{t,p}] = \mu_{t,p}$ and $E[(x_{t,p} - \mu_{t,p})(x_{t,p} - \mu_{t,p})^\top] = \Sigma_{t,p}$, we have that $\pi_{\text{max}}^2 I \geq \Sigma_{t,p} \geq \pi_{\text{min}}^2 I$ for all $t \in [T], p \in [P]$.

Intuitively, Definition 1 guarantees the sequence of presented contexts are (1) sufficiently similar since there is a common positive semidefinite ordering for the covariances across all contexts, and (2) sufficiently cover all directions in $\mathbb{R}^d$ when the parameters $\chi^2$, $\pi_{\text{max}}^2$ and $\pi_{\text{min}}^2$ are of the same order. If these conditions are satisfied then exploration distributions exist which can uniformly explore all directions of the underlying context sets well. We now provide a simple example of a set of stochastically generated contexts which obey Definition 1. For each processor, let there be a single context set randomly generated as $X_p = \{x: x \sim D(\cdot)\}_{i=1}^N$, where $D(\cdot)$ is a $O(1)$-subgaussian and $O(1)$-bounded distribution in $\mathbb{R}^d$ such that $X_{t,p} = X_p$ for all $t \in [T], p \in [P]$. Now define $\pi_{t,p}(\cdot)$ as the uniform distribution over all the vectors in a given context $X_p$ at time $t$. Then using a simple matrix concentration argument, we can verify there exists $\chi^2 \leq \tilde{O}(1/d)$, $\pi_{\text{max}}^2 \leq \tilde{O}(1/d)$ and $\pi_{\text{min}}^2 \geq \tilde{O}(1/d)$ in Definition 1 when $P \geq \tilde{O}(d)$ (with probability at least $1 - \delta$ over the randomness in $D(\cdot)$ and $\pi(\cdot)$).

Sampling from a rich exploration policy (when it exists) can serve as effective doubling-round subroutine, since it can help stabilize the covariance intra-round in later rounds. Random exploration helps stabilize the covariance in later rounds as a consequence of concentration: given a set of randomly sampled covariates $\{x_{i,p}\}_{i=1,p=1}^{N,P}$, we expect $\frac{1}{NP} \sum_{j=1}^N \sum_{p=1}^P x_{j,p} x_{j,p}^\top \approx \Sigma_\pi$, where $\Sigma_\pi = \Sigma_{t,p}$ for all $t \in [T], p \in [P]$. So if a significant number of doubling rounds occur (during which Algorithm 5 is used as DR), then later rounds are unlikely to be doubling rounds since the covariance matrix will have a large component proportional $\pi_{\text{min}}^2 I \leq \Sigma_\pi$ in its spectrum.

Using this idea we obtain the following corollary.
Theorem 3. For any parallel bandit algorithm, there exists bandit environments satisfying Assumptions 1, 2 and 3, a well-specified model and the second is the cost of misspecification. These come together in the following theorem.

Corollary 4. In the setting of Theorem 1, assume Definition 1 holds for some rich exploration policies, \( \{ \pi_{t,p}(\cdot) \}_{t=1,p=1}^{T,P} \), and context pairs, \( \{ \mathcal{X}_{t,p} \}_{t=1,p=1}^{T,P} \). Then choosing \( \lambda = L^2 \) and taking the doubling-routine DR as Algorithm 5 with these \( \pi_{t,p}(\cdot) \), the regret of both Algorithms 1 and 2 satisfy

\[
R(T,P) \leq \hat{O} \left( R \cdot \left( \sqrt{\text{SNR}} \cdot \left( \frac{L^2 \pi_{\max}^{ \pi_{\min}}}{\pi_{\min}} + P \frac{\lambda_{\max}^{ \pi_{\min}}}{\pi_{\min}} \right) + \sqrt{dTP} (\sqrt{d} + \sqrt{\text{SNR}} + \frac{\epsilon}{R} \sqrt{TP}) \right) ,
\]

with probability at least \( 1 - 2\delta \). In the setting of Theorem 2 with the choice \( \lambda = L^2 \) and also taking the doubling-routine DR as Algorithm 5 with this \( \pi_{t,p}(\cdot) \), the regret of Algorithms 3 and 4 satisfy

\[
R(T,P) \leq \hat{O} \left( R \left( \sqrt{\text{SNR}} \cdot \left( \frac{L^2 \pi_{\max}^{ \pi_{\min}}}{\pi_{\min}} + P \frac{\lambda_{\max}^{ \pi_{\min}}}{\pi_{\min}} \right) + d \sqrt{TP} (\sqrt{d} + \sqrt{\text{SNR}} + \frac{\epsilon}{R} \sqrt{TP}) \right) ,
\]

with probability at least \( 1 - 4\delta \), whenever \( \delta \leq \frac{1}{6} \).

We can provide further interpretation of this result as follows.

- The guarantee presented here resembles that of Corollary 3. However, here the coefficient on the burn-in term (which bounds the number of doubling rounds) scales as \( \frac{L^2 \pi_{\max}^{ \pi_{\min}}}{\pi_{\min}} + P \frac{\lambda_{\max}^{ \pi_{\min}}}{\pi_{\min}} \) and is valid even for infinite context sets.\footnote{Recall the coefficient in Corollary 3 scales as \( m \) where is a bound on the number of distinct vectors in the presented contexts.} The scaling with \( \frac{\lambda_{\max}^{ \pi_{\min}}}{\pi_{\min}} \) is natural. Quantities proportional to \( \frac{\lambda_{\max}^{ \pi_{\min}}}{\pi_{\min}} \) represent the cost of the non-homogeneous geometry of the presented contexts.

- In contrast to Corollaries 1, 2 and 3, Corollary 4 takes advantage of a common geometric structure in the presented contexts \( \mathcal{X}_{t,p} \). Corollary 3 is applicable to general context sets (for which we may have \( \pi_{\min} \approx 0 \)). However, if \( m = \Theta(\exp(d)) \) for example, the guarantee degrades badly. Similarly, Corollary 1 has a burn-in term scaled by \( d \). If the context sets are well-conditioned, in the sense that \( \frac{\lambda_{\max}^{ \pi_{\min}}}{\pi_{\min}} \ll d \), then the parallelism cost here is significantly lower then in the aforementioned cases.

- In the example of rich context sets/distributions pair presented in the text, we can have \( \frac{L^2 \pi_{\max}^{ \pi_{\min}}}{\pi_{\min}} + P \frac{\lambda_{\max}^{ \pi_{\min}}}{\pi_{\min}} \leq \hat{O}(P) \) with high probability when \( P \gtrsim \hat{O}(d) \). As \( P \to \infty \), the burn-in term only scales as \( \hat{O}(P) \), which up to logarithmic factors is free of any explicit dependence on the size of the context sets.

4 Parallel Regret Lower Bounds

Lastly, we show that the upper bounds on the parallel regret provided in the previous sections are nearly matched by complementary information-theoretic lower bounds on the parallel regret under natural parameter scalings. Our regret lower bounds seek to capture two specific terms. The first is the regret induced by the learning and optimally playing \( \theta \) complementary information-theoretic lower bounds on the parallel regret under natural parameter scalings. Our regret lower bounds seek to capture two specific terms. The first is the regret induced by the learning and optimally playing \( \theta^* \) for a well-specified model and the second is the cost of misspecification. These come together in the following theorem.

Theorem 3. For any parallel bandit algorithm, there exists bandit environments satisfying Assumptions 1, 2 and 3,

1. Such that when \( \epsilon = 0 \) there is a single global context set (i.e. \( \mathcal{X} = \mathcal{X}_{t,p} \) for all \( t \in [T], p \in [P] \)) for which,

\[
\mathbb{E}[R(T,P)] \geq \Omega \left( Rd \sqrt{TP} \right)
\]

when \( T \geq d \max(1, \frac{1}{3\sqrt{2\text{SNR}}} ), \) and

2. A single (finite) global context set with \( m \) vectors such that when \( d = \lceil 8 \log(m)/\epsilon^2 \rceil, R = 0 \) and \( LS \geq 1 \),

\[
\mathbb{E}[R(T,P)] \geq \Omega \left( \epsilon \sqrt{\frac{d - 1}{\log(m)} \cdot \min(TP, m - 1)} \right).
\]
The proof of Theorem 3 reposes on three separate parts. The first component is a reduction which argues that the minimax regret of a parallel bandit algorithm, presented with contexts fixed across processors, must be at least as much as its sequential counterpart given access to the same number of total arm queries. The first part of Theorem 3 then follows from a standard construction for lower bounding the sequential regret of a bandit instance when the context set is taken to be a sphere. The second part is a (noiseless) lower bound which uses a probabilistic argument to witness a finite context set, function \( f \), and \( \Theta^* \) for which the misspecification level must show up multiplicatively in the regret. We now make several comments to further interpret these results.

- Together the terms (15) and (16) show the main components of Theorem 1 are unavoidable—in particular the term corresponding to the variance of learning \( Rd\sqrt{TP} \) and the term capturing the magnitude of misspecification \( \epsilon \sqrt{dTP} \).

- The first term (15) in Theorem 3 captures the variance of learning under optimism in the parallel regret. For this portion of the lower bound, the context set is taken to be the sphere, which satisfies the conditions of a rich exploration set when the exploration distribution is taken as the uniform distribution over the spherical shell (here \( \ell = L \)). Hence for SNR \( \lesssim d \) and \( \epsilon = 0 \), the guarantee from Corollary 4 matches this lower bound up to logarithmic factors for sufficiently large \( T \). Similarly in the regime \( P \lesssim d \), the guarantee from Corollary 2 which holds in the absence of any additional structure on the context sets also matches this lower bound up to logarithmic factors when \( \epsilon = 0 \).

- To gain further intuition for (16) it is helpful to consider the high-dimensional scaling limit where \( m = \Theta(d^k) \) for \( k \gg 1 \) and \( TP \ll m \). Then, under the conditions of the result, we can see for any \( \epsilon \) there exists a sufficiently large \( d \) so that the parallel regret satisfies \( \geq \Omega(\epsilon \sqrt{dTP}) \). Hence (in the realistic regime) where the context set contains a large numbers of context vectors and there are not sufficiently many queries to observe all \( m \) of them, the \( \epsilon \sqrt{dTP} \) in the regret within Theorem 1 is unavoidable up to logarithmic factors.

- Capturing the necessity of the “burn-in” terms, which represent the price of parallelism in our upper bounds, is an interesting but challenging research direction. In particular, because in many applications the information-theoretic limits of learning when \( P \sim T \) may be of interest. However, the interplay between the structure of the context set and the burn-in terms in the upper bounds in Corollaries 1, 2, 3 and 4 seems quite nuanced. Any such lower bounds capturing these dependencies will likely need to be constructed on a case-specific basis for different context set geometries as well as be geared towards the small \( T \) regime.

## 5 Experiments

Here we explore the performance of the parallel linear bandit algorithms presented in this paper on several synthetic and real problem instances of increasing complexity. While analysis in the bandits literature is often focused on minimizing regret, in the batch setting best arm identification may be of primary interest for some practical design settings where no cost is incurred for additional arms after the best performing arm within a round. Hence, we explore the performance of our family of algorithms in both parallel regret and best arm identification.

In the synthetic data settings, we investigate both a perfectly linear setting and a misspecified setting generated from the output of a randomly initialized neural network. The real data instances are derived from a material science and biological sequence design applications to provide breadth across a variety of context set geometries and application-specific behaviors. In the real data settings, we consider the performance over parallel variants of all algorithms considered herein against a baseline that is the \( \epsilon \)-greedy algorithm. Note that this baseline makes no structural assumptions on the conditional model \( y_i|x_i \) (and as such is “unbiased”) but also is not able to take advantage of the covariates \( x_i \), since it doesn’t construct a regression model to guide exploration.

For all experimental setups, we fix the total number of arm queries \( TP \) and run with three different levels of parallelism (i.e. \( P = \{1, 10, 30\} \) for the superconductor setting and \( P = \{1, 10, 100\} \) for all others) over 30 separate trials. All algorithms use a doubling round routine which is set to the identity map. As is common, for both Thompson sampling variants, we avoided inflating the confidence set radius by the additional \( \sqrt{d} \) factor so it matches the confidence sets of the other algorithms. The misspecification parameter was set to zero for all experiments. The hyperparameters of each algorithm were tuned via a post-hoc grid search over a logarithmically-spaced grid for the random neural network and real data experiments (see Appendix D for details) as in Foster et al. [2018].

---

\(^5\text{Note that an additive dependence of } \Omega(P) \text{ must at least be necessary, since at time } T = 1 \text{ the algorithm must make } P \text{ queries to an arbitrary set of contexts in the absence of any information about the underlying bandit environment.}\)
5.1 Synthetic Experiments

We begin by testing the ability of our optimistic algorithms to parallelize on simulated data. We consider a problem in $d = 100$ with a linear reward oracle whose underlying parameter $\theta^*/\|\theta^*\|$ for $\theta^* \sim \mathcal{N}(0, I_d)$ subject to Gaussian additive noise $\epsilon \sim \mathcal{N}(0, 1)$. We then generate a fixed, global context set $\mathcal{X} = \{x_i/\|x_i\|_2\}_{i=1}^m$ for $x_i \sim \mathcal{N}(0, I_d)$ with $m = 10^4$ actions. We then set $X_{t,p} = \mathcal{X}$ for all $t \in [T]$ and $p \in [P]$. The hyperparameters of the algorithms were chosen according to their theoretically-motivated values $\lambda = 1, R = 1, S = 1$ with $\delta = 1/T$. As Fig. 1 shows, the parallel versions of each of the algorithms asymptotically achieve a nearly perfect speed-up with respect to parallelism as measured by the regret. As $T \to \infty$ the performance of the different types of base algorithms are comparable.

Next we investigate the parallelism of our methods under changing context sets. We generate a random context set $X_{t,p} = \{x_i/\|x_i\|_2\}_{i=1}^m$ for $x_i \sim \mathcal{N}(0, I_d)$ with $m = 10^4$ actions for each timestep-processor pair $(t, p)$ with hyperparameters set as before. Once again we see in Fig. 2 that each algorithm achieves near perfect speed-up as measured by parallel regret and all base algorithms are asymptotically comparable.

Finally, from our theoretical results we recall the importance of the covariances $V_{t,p}$ remaining quasi-static intra-round. We examine this behavior in the synthetic linear reward setting by determining the minimal doubling round coefficient $\alpha_{t,p}^{\text{min}}$ for each arm query which satisfies $0 \leq \alpha V_{t-1} - V_{t,p}$, where we call $\alpha$ the doubling round coefficient. That is, for any doubling round coefficient $\alpha \geq \alpha_{t,p}^{\text{min}}$ for all $p \in [P]$, the critical covariance inequality is satisfied and no doubling round is triggered for that round. Note that in our theoretical results, we arbitrarily set $\alpha = 2$ for our analysis. As before, we generate a fixed global context and theoretically-motivated hyperparameter values, but this time with $d = 20$ and $m = 10^3$. We then run the algorithms without calling any doubling round routines and observe how the doubling round coefficient changes through time for $P = 100$. At each $(t, p)$ timestep, we compute $\alpha_{t,p}^{\text{min}}$ and plot it against the number of arm queries as shown in Fig. 3. As expected, we qualitatively notice a sawtooth pattern in all algorithms as the minimal doubling round coefficient increases for fixed $t$ as the covariance $V_{t,p}$ gets updated for each additional processor $p$ which then resets back down at the end of each round. Additionally, all algorithms experience a dropoff in minimal doubling round coefficient with each successive round indicating that the covariance gets more stable as more arms are queried with all algorithms having an essentially flat $\alpha_{t,p}^{\text{min}}$ near 1 as $t \geq 2000$. Finally, we observe that in the fixed context setting where diversity of arms isn’t introduced via the context, LinUCB has the highest doubling round coefficients since the algorithm chooses the same arm for all processors leading for significant changes in the shape of the covariance within a round making the algorithm susceptible to overconfidence. Note that the other three algorithms have much smaller doubling round coefficients for a much shorter time due to the increased diversity in actions played in the bandit algorithm.

![Figure 1](image1.png)  
**Figure 1:** Fixed context setting. From left to right: Regret of LinUCB, Lazy LinUCB, LinTS, and Lazy LinTS for varying values of $P$. The mean regret is plotted across 30 runs with the standard deviation as the shaded region. Here $d = 100$, $m = 10^4$.

5.2 Randomly Initialized Neural Network Data

Recent studies (Angermueller et al. [2020], Brookes and Listgarten [2018]) have modeled fitness landscapes from biological sequence design problems with randomly initialized neural networks. Such randomly initialized neural network exhibit nearly linear properties for small parameter values; these properties are essential to the guarantees for our algorithms as demonstrated in the prequel. Moreover, they deviate from a linear model sufficiently to serve as a good testbed for model misspecification. The “biological sequence” input is modeled as a 14-length binary string $x_i \in \{0, 1\}^{14} (m = 16,384)$ sequences) to mimic the combinatorial nature of biological sequences. The fitness landscape $f(x_i)$ is modeled by a feedforward neural network with three hidden layers (of size 128, 256, and 512 hidden units) where each weight is sampled
which we wish to optimize. Note unlike in common use cases for neural networks the initialized weights are never modified.

To model experimental noise, we add Gaussian noise to generate the reward, $y_i$. The output $y_i$ of the randomly initialized feedforward neural network can be thought of as the oracle fitness landscape which we wish to optimize. Note unlike in common use cases for neural networks the initialized weights are never modified. To model experimental noise, we add Gaussian noise to generate the reward, $r_i = y_i + \epsilon_i$, where $\epsilon_i \sim N(0, 0.5^2)$.

Our family of bandit algorithms were run with both linear features only ($d = 14$) and quadratic features ($d = 210$). In this setting, we first verify that a linear model is appropriate. The best fit linear model for linear features and quadratic features had an $R^2$ of 0.7 and 0.87, respectively.

As Fig. 4 demonstrates in the quadratic feature setting, the variance between runs of the parallel regret for the lazy methods tends to be much higher than the non-lazy methods due to the correlation of covariance updates within a batch. LinUCB performs the best upfront in the purely sequential setting due to the high early round cost that Thompson sampling pays upfront, while paying a much lower price in regret in successive rounds. In higher parallelism regimes, Thompson sampling performs the best in terms of parallel regret after the first few hundred arm queries. This indicates that Thompson sampling benefits from encouraging diversity.

In comparison as shown in Fig. 8, the linear features perform demonstrably worse in terms of parallel regret than the quadratic features in just a few hundred arm queries across all levels of parallelization. Furthermore, the performance of LinUCB in the linear feature setting suffers most significantly relative to the other methods further confirming that diversity is important in settings of model misspecification.

### 5.3 Superconductor Data

To assess the utility of the parallel bandit algorithms in a realistic setting we constructed a semi-synthetic problem using the UCI dataset in [Hamidieh et al. 2018](http://example.com) consisting of a collection of superconducting materials along with their maximum superconducting temperature. The dataset consists of $m = 21,263$ superconducting materials, each with a $d = 81$-dimensional feature vector, $x_i$, containing relevant attributes of the materials chemical constituents and a superconducting critical temperature $y_i$. We construct a finite-armed bandit oracle over the $m$ arms which returns a reward $r_i = y_i + \epsilon_i$ for $\epsilon_i \sim N(0, 100^2)$ (since $\max_i y_i = 185.0$). The task in this example is to find the best superconducting material (or arm as measured by $y_i$) given access to a total number of arm queries $\ll m$.

As Fig. 5 shows although $\epsilon$-greedy is a simple algorithm, it can achieve reasonable performance (at the cost of high
Figure 4: Top Left: The histogram of fitness values for the RandomNN dataset. Top right: The parallel regret of the purely sequential setting for 5000 queries with a noise standard deviation of 0.5. Bottom Left: The parallel regret for \( P = 10 \). Bottom Right: The parallel regret for \( P = 100 \). The mean regret and standard deviation are plotted as the solid line and shaded region in all plots.

Figure 5: Leftmost: Fitness Histogram of Landscape. Left to right: Regret of all algorithms for \( P = 1, 10, \) and 30, respectively. Here the best superconducting material (by temperature) as determined by the algorithm at the time is displayed. Curves are also smoothed by a moving-average over a window of size 30 for clarity.

variance), when \( P = 1 \). Indeed prior work has shown that other greedy (linear bandit) algorithms are formidable baselines in setting with diverse covariates [Bayati et al., 2020]. However, in our setting, it is still outperformed by all the linear bandit algorithms studied herein. We also see all algorithms quickly saturate to find superconducting materials with temperatures \( y_i \approx 120 \).

In the cases of \( P = 10 \), and \( P = 30 \) we see all the parallel variants of the linear algorithms studied herein achieve non-trivial parallelism gains; that is the number of sequential rounds needed to discover this best material does not scale linearly with \( P \) for any of the methods. Remarkably, Thompson sampling suffers almost no loss in performance even when \( P = 30 \) in this setting with real data where model misspecification is in full force. Thompson sampling outperforms all other algorithms when \( P = 10 \) and \( P = 30 \). As our results show, explicitly introducing diversity into the selection of actions provides value in this setting.

5.4 Transcription Factor Binding

In order to evaluate the effectiveness of the family of proposed parallelized linear bandit algorithms in a realistic biological sequence design setting, we utilized a fully characterized experimental transcription factor binding affinity dataset from Barrera et al. [2016] (using the software package in Sinai et al. [2020]). Changes in transcription factor binding affinity has been shown to have impact on gene regulatory function and subsequently is associated with disease risk. The dataset experimentally characterizes the binding affinity of all possible 8-length DNA sequence motifs \( (m = 4^8 = 65,536) \) to a transcription factor DNA binding domain providing a good benchmark for our bandit methods in a real biological application with the combinatorial structure common in biological sequence design. In this setting the number of arms \( m \) is \( O(\exp(d)) \).

Often in biological sequence design problems quadratic features are used to model pairwise interactions (referred to as epistasis in biology). We compared linear features with random ReLU features and a quadratic kernel and found linear features work best in and provide additional insight in Appendix D.
The fully characterized landscape allows for exact computation of parallel regret and analysis of the impact of realistic forms of model misspecification. Each arm $x_i$ was one-hot encoded. The scaled binding affinity $y_i \in [0, 1]$ measured the binding of the arm $x_i$ for the SIX6 REF R1 transcription factor target. The finite-armed bandit oracle as in the superconductor setting was modeled as $r_i = y_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 0.3^2)$. The task for this application is to find the sequence with the highest transcription factor binding affinity.

In evaluating the best arm reported across varying levels of parallelism, we can see that LinUCB and $\epsilon$-greedy consistently performs the worst with the other three algorithms (Lazy LinUCB, Thompson sampling, and Lazy Thompson sampling) perform comparably. This implies that diversity of arms is important particularly when the model is misspecified. Note that the right tail of the fitness histogram is rather heavy for this task such that getting to an arm with fitness above 0.9 is rather simple in the noiseless setting and can be optimized in few arm queries as was shown in Angermueller et al. [2020]. However, since biological experiments often have large experimental error we add noise with standard deviation of 0.3 making the problem significantly harder leading to worse performance of algorithms preventing $\epsilon$-greedy from beating the 0.9 threshold at all. Similarly, we find the same relative performance of methods in terms of parallel regret as shown in Appendix D.

Figure 6: TFBinding best arm with linear features. Leftmost: The fitness distribution of the dataset. From left to right: The best smoothed binding affinity for each round with error bars indicating standard deviation with $P = 1, 10,$ and 100, respectively.

6 Conclusions

We have presented several parallel (contextual) linear bandit algorithms inspired by the optimism principle. Utilizing the notion of covariance stability, we provide a unified analysis of their regret in a variety of settings. Our regret upper bounds establish that the performance of these algorithms is nearly identical to that of their sequential counterparts (with the same total number of arm queries), up to a burn-in term which may depend on the geometry of the context set. Finally, we show that the parallelism gains suggested by our theory can also be achieved in several real datasets motivated by practical design problems and demonstrate the importance of diversity in problems that contain model misspecification.

Interesting directions for future work including extending the results herein to a suitably defined notion of best-arm identification. Similarly, understanding the impact of parallelism in simple, greedy heuristic algorithms (which nonetheless perform well in practice [Bayati et al., 2020, Sinai et al., 2020]) is another important direction. Another interesting direction is leveraging parallel processors to devise online model selection and hyperparameter tuning strategies. Lastly, in applications such as protein engineering, it is often of interest to discover a diverse set of high-reward sequences under suitable notions of diversity.
A Proofs for Regret Upper Bounds

In this section we provide the proofs of the regret upper bounds for all of the algorithms considered.

A.1 Proofs for Section 2.1

Here we include the Proof of Theorem 1.

**Proof of Theorem 1.** We first decompose the regret for Algorithm 1 by splitting into the linear term and misspecification component.

\[ R(T, P) = \sum_{t=1}^{T} \left( \sum_{p=1}^{P} f(x_{t,p}^*) - f(x_{t,p}) \right) = \sum_{t=1}^{T} \left( \sum_{p=1}^{P} (x_{t,p}^* - x_{t,p}, \Theta^*) \right) + \sum_{t=1}^{T} \left( \sum_{p=1}^{P} f(x_{t,p}) - \langle x_{t,p}^*, \Theta^* \rangle + f(x_{t,p}) - \langle x_{t,p}, \Theta^* \rangle \right). \]

The second term can be immediately upper bounded as,

\[ \sum_{t=1}^{T} \left( \sum_{p=1}^{P} f(x_{t,p}^*) - \langle x_{t,p}^*, \Theta^* \rangle + f(x_{t,p}) - \langle x_{t,p}, \Theta^* \rangle \right) \leq 2\epsilon TP. \]

using Assumption 3. We now approach the linearized reward term. We split this term in each round over the event \( D_t \),

\[ \sum_{t=1}^{T} \left( \sum_{p=1}^{P} (x_{t,p}^* - x_{t,p}, \Theta^*) \right) = \sum_{t=1}^{T} 1[D_t] \left( \sum_{p=1}^{P} (x_{t,p}^* - x_{t,p}, \Theta^*) \right) + \sum_{t=1}^{T} 1[D_t^c] \left( \sum_{p=1}^{P} (x_{t,p}^* - x_{t,p}, \Theta^*) \right) \]

The first term here can be bounded using Assumptions 2 and 3 along with the Cauchy-Schwarz inequality which gives \((\sum_{p=1}^{P} (x_{t,p}^* - x_{t,p}, \Theta^*)) \leq 2LSP \) so:

\[ \sum_{t=1}^{T} 1[D_t] \left( \sum_{p=1}^{P} (x_{t,p}^* - x_{t,p}, \Theta^*) \right) \leq 2LSP \sum_{t=1}^{T} 1[D_t]. \]

Note the above bounds hold for any choices of \( x_{t,p} \in X_{t,p} \) selected by any doubling-round routine. We now turn our attention to the second term. For this term we use essentially the same techniques to bound the instantaneous regret by the exact same value for both Algorithm 1 and Algorithm 2, but separate the analysis into two cases for clarity.

- For Algorithm 1 we refer to the optimistic model of processor \( p \) at round \( t \) as:

\[ \hat{\Theta}_{t,p} = \arg\max_{\Theta \in \Theta_{1,t}, \Theta \in \Theta_{1,t}, 1, \epsilon} \left( \max_{\Theta \in \Theta_{1,t}} \langle x_{t,p}, \Theta \rangle \right) \]

for Algorithm 1. Conditioned on the event in Theorem 8—which we denote \( \mathcal{E}_1 \)–the models \( \hat{\Theta}_{t,p} \) are optimistic:

\[ \langle x_{t,p}, \hat{\Theta}_{t,p} \rangle \mathbb{I}[\mathcal{E}_1] \geq \langle x_{t,p}^*, \Theta^* \rangle \mathbb{I}[\mathcal{E}_1]. \]

Hence,

\[ 1[\mathcal{E}_1] 1[D_t] (x_{t,p}^* - x_{t,p}, \Theta^*) \leq 1[\mathcal{E}_1] 1[D_t] (\hat{\Theta}_{t,p} - \Theta^*) \leq 2\sqrt{21} \mathbb{E}[\mathcal{E}_1] |D_t| |x_{t,p}||\hat{\Theta}_{t,p} - \Theta^*||V_{\pi,t}^{-1}||\hat{\Theta}_{t} - \Theta^*||V_{\pi,t}^{-1} ||\hat{\Theta}_{t} - \Theta^*||V_{\pi,t}^{-1} ||\hat{\Theta}_{t} - \Theta^*||V_{\pi,t}^{-1} \]

using optimism in the first inequality, Cauchy-Schwartz in the second, and the fact that on event \( D_t^c \) round \( t \) is not a generalized doubling round in the in the final inequality. Finally, recall on the event \( \mathcal{E}_1, 1[\mathcal{E}_1]|\hat{\Theta}_t - \Theta^*||V_{\pi,t} \leq (\sqrt{\beta_{t}(\delta)} + \sqrt{(t-1)P\epsilon})1[\mathcal{E}_1] \) and note \( \beta_{t}(\delta) \) is an increasing function of \( t \) so \( \beta_{t}(\delta) \leq \beta_{T}(\delta) \) for all \( t \leq T \). Hence it follows,

\[ 1[\mathcal{E}_1] 1[D_t^c] (x_{t,p}^* - x_{t,p}, \Theta^*) \leq 1[\mathcal{E}_1] 2\sqrt{2} (\sqrt{\beta_{T}(\delta)} + \sqrt{(t-1)P\epsilon})|x_{t,p}||V_{\pi,t}^{-1} \]

additionally relaxing \( 1[D_t^c] \leq 1. \)
The remainder of the proof follows identically for both Algorithms 1 and 2. Without loss of generality we use
accounting for the original misspecification term shows that,
\[ \langle x_{t,p}, \hat{\theta}_{t,p} \rangle \mathbb{1}[E_2] \mathbb{1}[D_t^c] \geq \langle x_{t,p}^*, \theta^* \rangle \mathbb{1}[E_2] \mathbb{1}[D_t^c]. \]

Hence,
\[ \mathbb{1}[E_2] \mathbb{1}[D_t^c] \langle x_{t,p}^*, x_{t,p}, \theta^* \rangle \leq \mathbb{1}[E_2] \mathbb{1}[D_t^c] \langle x_{t,p}, \hat{\theta}_{t,p} - \theta^* \rangle \leq \mathbb{1}[E_2] \mathbb{1}[D_t^c] \langle \hat{\theta}_{t,p} - \theta^* \rangle \mathbb{1}[V_{t,p}^{-1}], \]

using optimism in the first inequality, Cauchy-Schwartz in the second. Finally, on the event \( E_2 \cap D_t^c \), \( \mathbb{1}[E_2] \mathbb{1}[D_t^c] \parallel \hat{\theta}_{t,p} - \theta^* \parallel V_{t,p} \leq 2\sqrt{2} \sqrt{\beta_t(\delta)} \mathbb{1}[E_2] \mathbb{1}[D_t^c] \leq 2\sqrt{2} \sqrt{\beta_t(\delta)} \mathbb{1}[E_2] \mathbb{1}[D_t^c] \). Hence it follows,
\[ \mathbb{1}[E_2] \mathbb{1}[D_t^c] \langle x_{t,p}^*, x_{t,p}, \theta^* \rangle \leq 2\sqrt{2} \mathbb{1}[E_2] ((\sqrt{\beta_t(\delta)} + \sqrt{(t - 1)P \epsilon}) \parallel x_{t,p} \parallel V_{t,p}^{-1}) \]

additionally relaxing \( \mathbb{1}[D_t^c] \leq 1 \).

The remainder of the proof follows identically for both Algorithms 1 and 2. Without loss of generality we use \( E \) to refer to either event \( E_1 \) or \( E_2 \) in the following (note both hold with probability at least \( 1 - \delta \) by Theorems 6 and 7). Recalling that
the instantaneous regret is \( \leq 2LS \) we can combine this bound with the aforementioned bounds to conclude,
\[
\mathbb{1}[E] \cdot \left( \sum_{t=1}^{T} \left( \sum_{p=1}^{P} \langle x_{t,p}^*, x_{t,p}, \theta^* \rangle \right) \right) \\
\overset{(i)}{\leq} \mathbb{1}[E] \cdot \sqrt{TP \left( \sum_{t=1}^{T} \sum_{p=1}^{P} \min((LS)^2, 2(\beta_t(\delta) + (t - 1)P \epsilon)^2) \parallel x_{t,p} \parallel V_{t,p}^{-1} \right)} \\
\overset{(ii)}{\leq} 4\sqrt{2} \mathbb{1}[E] \sqrt{TP \sum_{t=1}^{T} \sum_{p=1}^{P} \min((LS)^2, 2(\beta_t(\delta) + T \epsilon)^2) \parallel x_{t,p} \parallel V_{t,p}^{-1} \right)} \\
\overset{(iii)}{\leq} 4\sqrt{2} \mathbb{1}[E] \sqrt{TP \sum_{t=1}^{T} \sum_{p=1}^{P} \min((LS)^2, 2(\beta_t(\delta) + T \epsilon)^2) \parallel x_{t,p} \parallel V_{t,p}^{-1} \right)} \\
\overset{(iv)}{\leq} 4\sqrt{2} \mathbb{1}[E] \sqrt{TP \min((\sqrt{2} \sqrt{\beta_t(\delta)} + T \epsilon) \parallel x_{t,p} \parallel V_{t,p}^{-1} \right)} \cdot \sqrt{d \log \left( 1 + \frac{TPL^2}{\lambda} \right)}.
\]

Inequality (i) follow by Cauchy-Schwarz. Inequality (ii) employs both our bounds on the instantaneous regret. Inequality
(iii) follows by upper bounding \( t - 1 \leq T \) for the misspecification term. Inequality (iv) follows because for all
\( a, x > 0 \), we have \( \min(a, x) \leq \max(2, a) \log(1 + x) \). Inequality (v) follows because \( \sum_{t=1}^{T} \sum_{p=1}^{P} \log(1 + \parallel x_{t,p} \parallel V_{t,p}^{-1} \right) \leq d \log \left( 1 + \frac{TPL^2}{\lambda} \right) \) by instantiating Lemma 11. Assembling the bounds in the original regret splitting over doubling rounds
and accounting for the misspecification term shows that,
\[ \mathbb{1}[E] \mathbb{R}(T, P) \leq \mathbb{1}[E] \parallel LSP \sum_{t=1}^{T} \parallel D_t \parallel + \sqrt{TP \max(\sqrt{2}(\sqrt{\beta_t(\delta)} + T \epsilon), LS)} \cdot \sqrt{d \log \left( 1 + \frac{TPL^2}{\lambda} \right)} + TP \parallel.
where \( \sqrt{\beta_T(\delta)} \leq R \sqrt{d \log \left( \frac{1+TPt^2/\lambda}{\delta} \right)} + \sqrt{\lambda}S \). This inequality holds on the event \( E \) which occurs with probability at least \( 1 - \delta \) for both Algorithm 1 and Algorithm 2. Inserting this value for \( \beta_T(\delta) \) and hiding logarithmic factors shows on the event \( E \),

\[
R(T, P) \leq O \left( LSP \cdot \sum_{i=1}^{T} I[D_i] + \sqrt{dTP} \max(\sqrt{2}(R\sqrt{d} + \sqrt{\lambda}S + \sqrt{TPe}), LS) \right)
\]

\( \square \)

## A.2 Proofs of Section 2.2

We start by stating a folklore lemma regarding the anti-concentration properties of a Gaussian distribution.

### A.2.1 Concentration and Anti-Concentration properties of the Gaussian distribution

**Lemma 4.** Let \( X \) be a random variable distributed according to \( \mathcal{N}(\mu, \sigma^2) \), a one dimensional Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \). The following holds:

\[
\mathbb{P} (X - \mu \geq \tau) \geq \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{\tau^2}{2\sigma^2} \right).
\]

We will also make use of the following concentration inequality for Lipschitz functions of Gaussian vectors:

**Theorem 4** (Theorem 2.4 in Wainwright [2019]). Let \( \eta \sim \mathcal{N}(0, I_d) \) be a standard Gaussian vector and let \( f: \mathbb{R}^d \to \mathbb{R} \) be \( L \)-Lipschitz with respect to the Euclidean norm. Then the variable \( f(\eta) - \mathbb{E} [f(\eta)] \) is subgaussian with parameter at most \( L \) and hence:

\[
\mathbb{P} (f(X) \geq \mathbb{E} [f(X)] + t) \leq \exp \left( -\frac{t^2}{2L^2} \right)
\]

We’ll make use of these two results to prove Lemma 5 which we restate for the reader’s convenience:

**Lemma 5.** The Gaussian distribution satisfies (anticoncentration) for every \( \nu \in \mathbb{R}^d \) with \( \|\nu\| = 1 \):

\[
\mathbb{P}_{\eta \sim \mathcal{N}(0, I_d)} (\nu^T \eta \geq 1) \geq \frac{1}{4}.
\]

And (concentration), \( \forall \delta \in (0, 1) \):

\[
\mathbb{P}_{\eta \sim \mathcal{N}(0, I_d)} (\|\eta\| \leq \sqrt{d} + \sqrt{2\log \left( \frac{1}{\delta} \right)}) \geq 1 - \delta.
\]

**Proof.** Equation 17 is a simple consequence of the following two observations:

1. For any unit norm vector \( \nu \in \mathbb{R}^d \) the random variable \( X = \nu^T \eta \) is distributed as a one dimensional Gaussian with unit variance \( \mathcal{N}(0, 1) \).

2. Setting parameters \( \mu = 0, \sigma = 1, \) and \( \tau = 1 \) Lemma 4 implies that \( \mathbb{P} (X \geq 1) \geq \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2} \exp(-\frac{1}{2}) \geq \).

Equation 18 instead follows from Theorem 4. Since the function \( f(\cdot) = \| \cdot \| \) is \( 1 \)-Lipschitz and \( \mathbb{E}_{\eta \sim \mathcal{N}(0, I_d)} [\|\eta\|] \leq (\mathbb{E}_{\eta \sim \mathcal{N}(0, I_d)} [\|\eta\|^2])^{1/2} = \sqrt{d} \):

\[
\mathbb{P}_{\eta \sim \mathcal{N}(0, I_d)} (\|\eta\| \geq \sqrt{d} + \sqrt{2\log \left( \frac{1}{\delta} \right)}) \leq \delta.
\]

The result follows. \( \square \)

Recall that:

\[
\sqrt{\beta_T(\delta)} = R \sqrt{\log \left( \frac{\det(\mathbf{V}_{t-1,0})}{\lambda^d \delta^2} \right)} + \sqrt{\lambda}S \leq R \sqrt{d \log \left( \frac{1+tPL^2/\lambda}{\delta} \right)} + \sqrt{\lambda}S
\]
A.2.2 Concentration of $\hat{\theta}_{t,p}$

The main objective of this section is to show that with high probability the sampled parameter $\hat{\theta}_{t,p}$ is not too far from the true parameter $\theta$, for all times $t$ and processors $p$. The result is encapsulated by Lemma 6.

Lemma 6. Let:

$$
\gamma_{t,p}(\delta) := \sqrt{2}(\beta_t(\delta) + \sqrt{(t-1)P\epsilon}) \left( \sqrt{d} + 2 \sqrt{\log \left( \frac{(P-1)+p}{\delta} \right)} + 1 \right).
$$

The following conditional probability bound holds:

$$
P \left( \left| \left\{ F_{t,p} \cap \mathcal{E} \right\} \right| \| \hat{\theta}_{t,p} - \theta \| \geq \gamma_{t,p}(\delta) \right| F_{t,p-1} \right) \leq \frac{\delta}{2(t(P-1)+p)^2}
$$

And therefore with probability at least $1 - 2\delta$ and for all $t \in N$ simultaneously:

$$
\| \hat{\theta}_{t,p} - \theta \| \leq \gamma_{t,p}(\delta)
$$

We refer to this event as $\mathcal{E}'$.

In order to prove Lemma 6 let’s start by showing that for any time-step $t$ and processor $t$ the sample $\hat{\theta}_{t-1,p}$ is close to $\theta$, with high probability:

Lemma 7. The following conditional probability bound holds:

$$
P \left( \| \hat{\theta}_{t,p} - \hat{\theta}_t \| \geq \sqrt{2}(\beta_t(\delta) + \sqrt{(t-1)P\epsilon}) \left( \sqrt{d} + 2 \sqrt{\log \left( \frac{(P-1)+p}{\delta} \right)} \right) \right| F_{t,p-1} \right) \leq \frac{\delta}{(t(P-1)+p)^2}
$$

Where $F_{t,p-1}$ corresponds to the sigma algebra generated by all the events up to and including the reveal of contexts $X_{t,p}$. And therefore with probability at least $1 - \delta$ simultaneously and unconditionally for all $t \in N$:

$$
\| \hat{\theta}_{t,p} - \hat{\theta}_t \| \leq \sqrt{2}(\beta_t(\delta) + \sqrt{(t-1)P\epsilon}) \left( \sqrt{d} + 2 \sqrt{\log \left( \frac{(P-1)+p}{\delta} \right)} \right)
$$

Proof. In order to bound $\| \hat{\theta}_{t,p} - \hat{\theta}_t \|$, we make use of Lemma 5. Observe that by definition:

$$
\| \hat{\theta}_{t,p} - \hat{\theta}_t \| \leq \sqrt{2}(\beta_t(\delta) + \sqrt{(t-1)P\epsilon}) \| \eta_{t,p} \|.
$$

Therefore a simple use of Lemma 5 implies that (concentration):

$$
P_{\eta_{t,p}\sim N(0,\sigma^2)} \left( \| \eta_{t,p} \| \leq \sqrt{d} + 2 \sqrt{\log \left( \frac{(P-1)+p}{\delta} \right)} \right| F_{t,p-1} \right) \leq \frac{\delta}{(t(P-1)+p)^2}.
$$

And therefore as a consequence of Equation 21:

$$
P \left( \| \hat{\theta}_{t,p} - \hat{\theta}_t \| \geq \sqrt{2}(\beta_t(\delta) + \sqrt{(t-1)P\epsilon}) \left( \sqrt{d} + 2 \sqrt{\log \left( \frac{(P-1)+p}{\delta} \right)} \right) \right| F_{t,p-1} \right) \leq \frac{\delta}{(t(P-1)+p)^2}
$$

Furthermore, a simple union bound implies that for all $t \in N$:

$$
P \left( \exists t \text{ s.t. } \| \eta_{t,p} \| \geq \sqrt{d} + 2 \sqrt{\log \left( \frac{(P-1)+p}{\delta} \right)} \right) \leq \frac{\delta}{2} \sum_{t=1}^{\infty} \sum_{p=1}^{P} \frac{1}{(t(P-1)+p)^2} \leq \delta
$$

Combining equations 21 and 22 yields:

$$
P \left( \| \hat{\theta}_{t,p} - \hat{\theta}_t \| \leq \sqrt{2}(\beta_t(\delta) + \sqrt{(t-1)P\epsilon}) \left( \sqrt{d} + 2 \sqrt{\log \left( \frac{(P-1)+p}{\delta} \right)} \right) \right) \leq \delta
$$

Lemma 7, conditioning on the $1 - \delta$ probability event $\mathcal{E}$, a simple use of the triangle inequality along with the identity $\sum_{p=1}^{P} \frac{1}{(t(P-1)+p)^2} = \frac{\pi^2}{2} < 2$ finalizes the proof of Lemma 6. From now on we will denote as $\mathcal{E}'$ to the $1 - 2\delta$ probability event defined by Lemma 6.
The main objective of this section will be to prove the following upper bound for the instantaneous regret $\theta^\top x_{t,p} - \theta^\top x_{t,p}$ in the event the round is not a doubling round and $E'$ holds. This is one of the main components of the proof of Theorem 2.

**Lemma 8.** Let $\hat{\theta}_{t,p}$ be a copy of $\hat{\theta}_{t,p}$, equally distributed to $\hat{\theta}_{t,p}$ and independent of it conditionally on $F_{t,p-1}$. We call $x_{t,p}'$ to the resulting argmax action $\arg\max_{x \in X_{t,p}} (\hat{\theta}_{t,p}' x)$. The following inequality holds:

$$\mathbb{I}[\mathcal{D}_t^c \cap E'] (\theta^\top x_{t,p} - \theta^\top x_{t,p}) \leq \frac{2\gamma_{t,p}(\delta)}{\delta \frac{2}{(t(t-1)+1)p^2}} \mathbb{I}[\mathcal{D}_t^c \cap E'] \mathbb{E}\left[\|x_{t,p}'\|_{V_{t,p}} F_{t,p-1}\right] + \gamma_{t,p}(\delta) \mathbb{I}[\mathcal{D}_t^c \cap E']\|x_{t,p}\|_{V_{t,p}}^{-1}.$$

Before proving Lemma 8 we show that with constant probability, the estimated value of the action taken at time and processor tuple $(t,p)$ is optimistic with constant probability:

**Lemma 9.** For all $t \in N$:

$$\mathbb{P}\left(\hat{\theta}_{t,p}' x_{t,p} \geq \theta^\top x_{t,p} \mid F_{t,p-1}, E\right) \geq \frac{1}{4}$$

**Proof.** Recall that whenever $E$ holds:

$$\|\hat{\theta}_t - \theta\|_{V_{t,p}} \leq \sqrt{2\sqrt{\beta_t(\delta)} + \sqrt{(t-1)P\epsilon}}.$$

Notice that by definition $x_{t,p}$ satisfies:

$$\hat{\theta}_{t,p}' x_{t,p} \geq \hat{\theta}_{t,p}' x_{t,p}.$$

Therefore:

$$\hat{\theta}_{t,p}' x_{t,p} = \left(\hat{\theta}_{t,p} - \hat{\theta}_t\right)^\top x_{t,p} + \left(\hat{\theta}_t - \theta\right)^\top x_{t,p} + \theta^\top x_{t,p}$$

$$\geq \sqrt{2\sqrt{\beta_t(\delta)} + \sqrt{(t-1)P\epsilon}} \eta_{t,p} V_{t,p}^{-1/2} x_{t,p} - \|\hat{\theta}_t - \theta\|_{V_{t,p}} \|x_{t,p}\|_{V_{t,p}^{-1}} + \theta^\top x_{t,p}$$

$$\geq \sqrt{2\sqrt{\beta_t(\delta)} + \sqrt{(t-1)P\epsilon}} \eta_{t,p} V_{t,p}^{-1/2} x_{t,p} - 2\sqrt{\beta_t(\delta)} + \sqrt{(t-1)P\epsilon}) \|x_{t,p}\|_{V_{t,p}^{-1}} + \theta^\top x_{t,p}$$

$$= \sqrt{2\sqrt{\beta_t(\delta)} + \sqrt{(t-1)P\epsilon}} \left(\eta_{t,p} V_{t,p}^{-1/2} x_{t,p} - \|V_{t,p}^{-1/2} x_{t,p}\|_2\right).$$

Inequality (i) holds as a consequence of Cauchy Schwartz inequality. Inequality (ii) holds by conditioning on $E$ and because $t \in N$.

By Equation 17 in Lemma 5, and by noting that $x_{t,p}'$ is conditionally independent of $\eta_{t,p}$, we can infer that

$$\eta_{t,p} V_{t,p}^{-1/2} x_{t,p}' \geq \|V_{t,p}^{-1/2} x_{t,p}'\|_2$$

with probability at least $1/4$. The result follows.

Let’s define the set of optimistic model parameters:

$$\Theta_{t,p} = \{\theta \in \mathbb{R}^d \ s.t. \ \max_{x \in X_{t,p}} x^\top \theta \geq (x_{t,p}')^\top \theta\}.$$

Where $\mathcal{D}_t$ denotes the event that round $t$ is a doubling round. We now show how to bound the instantaneous regret $r_{t+p}$ during all rounds by using these results:

**Proof of Lemma 8.** Recall that $E'$ is the event defined by Equation 20 in Lemma 6 applied to the $\{\hat{\theta}_{t,p}\}_{t,p}$ sequence. Define $\{E_{t,p}'\}_{t,p}$ be the corresponding event family defined by Equation 19 in Lemma 6 applied to the $\{\hat{\theta}_{t,p}'\}_{t,p}$ sequence. It follows that $\mathbb{P}(E_{t,p}' \mid F_{t,p-1}) \geq 1 - \frac{\delta}{2(t(t-1)+p^2)}$. Notice that if $D_t^c \cap E'$ holds (meaning $\|\hat{\theta}_{t,p} - \theta\|_{V_{t,p}} \leq \delta_{t,p}$ and because $x_{t,p} = \arg\max_{x \in X_{t,p}} \theta_{t,p}' x$ then:

$$\hat{\theta}_{t,p}' x_{t,p} \geq \inf_{\theta \in \mathcal{C}(\theta, V_{t,p}, \gamma_{t,p}(\delta))} \max_{x \in X_{t,p}} \theta^\top x \equiv \hat{\theta}_{t,p}' x_{t,p}.$$

When $\hat{\theta}_{t,p}'$ is optimistic:

$$\theta^\top x_{t,p}' - \theta^\top x_{t,p}' \leq (\hat{\theta}_{t,p}' x_{t,p}') - \theta^\top x_{t,p}' \mid \hat{\theta}_{t,p}' \in \Theta_{t,p}.$$
Therefore:

\[
\begin{align*}
\mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] (\theta_*^T x_{t,p} - \theta^*_t x_{t,p}) &= \mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] (\theta^T x_{t,p} - \theta^T x_{t,p}) + \mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] (\theta^T x_{t,p} - \theta^T x_{t,p}) \\
&\leq \mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] (\theta^T x_{t,p} - \theta^T x_{t,p}) + \mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] \|\theta_t - \theta_*\|_{v,p} \|x_{t,p}\|_{v^{-1}} \\
&\leq \mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] \|\hat{\theta}_t - \theta_*\|_{v,p} \|x_{t,p}\|_{v^{-1}}^{i} \\
&\leq \mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] \|\hat{\theta}_t - \theta_*\|_{v,p} \|x_{t,p}\|_{v^{-1}}^{ii} \\
&\leq 2\gamma_t(\delta) \mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] \|x_{t,p}'\|_{v^{-1}} \|\hat{\theta}_t - \theta_*\|_{v,p} \|x_{t,p}\|_{v^{-1}}^{iv} \\
&\leq \frac{2\gamma_t(\delta)}{1 - \frac{\delta}{2(t(P^{-1}+p)^2)}} \mathbb{I}[\mathcal{D}_t^c \cap \mathcal{E}^c] \|x_{t,p}'\|_{v^{-1}} \|\hat{\theta}_t - \theta_*\|_{v,p} \|x_{t,p}\|_{v^{-1}}^{v}
\end{align*}
\]

Inequality (i) follows by Equation 23, (ii) by the definition of \( \hat{x}_t \), (iii) is a consequence of Cauchy Schwartz, (iv) follows by the definition of \( \mathcal{E}^c \) and \( \mathcal{E}''_t \), and (v) follows because \( |x_{t,p}'|_{v^{-1}} \) is nonnegative and because by Lemma 6, it follows that \( \mathbb{P} (\hat{\theta}_t \in \Theta, \mathcal{E}''_{t-1}, X_{t-1}) \geq \frac{1}{2} - \frac{\delta}{2(t(P^{-1}+p))^2} \). The result follows.

\[\square\]

### A.2.4 Ancillary Lemmas

In the proof of Theorem 2 we will also make use of the following supporting result:

**Lemma 10.** Similar to Lemma 8, let \( \hat{\theta}'_{t,p} \) copy of \( \hat{\theta}_{t,p} \), equally distributed to \( \hat{\theta}_{t,p} \) and independent conditionally on \( X_{t-1} \). We call \( x_{t,p}' \), \( \hat{\theta}'_{t,p} \), to the resulting argmax action \( \text{argmax}_{x \in \mathcal{X}_{t,p}} \langle x, \hat{\theta}'_{t,p} \rangle \). With probability at least 1 – \( \delta \):

\[
\sum_{t=1}^{T} \sum_{p=1}^{P} \mathbb{E}[|x_{t,p}'|_{\mathbb{V}^{-1}} | X_{t-1}, F_{t-1}] \leq \sum_{t=1}^{T} \sum_{p=1}^{P} |x_{t,p}'|_{\mathbb{V}^{-1}} + \frac{2L}{\sqrt{\lambda}} \sqrt{TP\log(1/\delta)}
\]

**Proof.** Define the martingale difference sequence \( Z_{t,p} = |x_{t,p}'|_{\mathbb{V}^{-1}} - \mathbb{E}[|x_{t,p}'|_{\mathbb{V}^{-1}} | F_{t-1}] \) (the indexing is lexicographic over pairs \((t,p)\)). It is easy to see that \( |Z_{t,p}| \leq \frac{2L}{\sqrt{\lambda}} \), since for all valid \( x \in \mathcal{X}_{t,p} \), \( |x|_{\mathbb{V}^{-1}} \leq \frac{L}{\sqrt{\lambda}} \) for all \( t, p \) pairs. Consequently a simple use of Hoeffding bound yields the result.

\[\square\]

### A.3 Proof of Theorem 2

We proceed to prove the general version of Theorem 2:

**Theorem 5.** Let Assumptions 1, 2 and 3 hold and \( \mathcal{D}_t \) denote the event that round \( t \) is a doubling round (see Condition 1). Then the regret of both Algorithms 3 and 4 satisfy

\[
R(T, P) \leq O \left( LSP \cdot \sum_{t=1}^{T} \mathbb{1}[\mathcal{D}_t] + 4\gamma_T(\delta) \left( \sqrt{dTP \left( 1 + \frac{L^2}{\lambda} \right) \ln \left( \frac{d\lambda + TP\lambda}{d\lambda} \right) + \frac{2L}{\sqrt{\lambda}} \sqrt{TP\log(1/\delta)} + cTP \right) \right)
\]

(24)
with probability at least $1 - 3\delta$, whenever $\delta \leq \frac{1}{6}$. Here

$$\gamma_T(\delta) = \sqrt{2(\sqrt{T\delta} + \sqrt{(T - 1)\epsilon})} \left( \sqrt{d} + 2\sqrt{\log \left( \frac{TP - 1 + P}{\delta} \right) + 1} \right).$$

**Proof of Theorem 2.** The regret in terms of $f(\cdot)$ can be linearized at the cost of an additive $2\epsilon \sqrt{dTP}$ as in the Proof of Theorem 1. After this we can decompose the regret for Algorithm 3, splitting on the event each round over the event $D_t$,

$$\mathcal{R}(T, P) = \sum_{t=1}^{T} \left( \sum_{p=1}^{P} \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \right) = \sum_{t=1}^{T} \mathbb{I}[D_t] \left( \sum_{p=1}^{P} \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \right) + \sum_{t=1}^{T} \mathbb{I}[D^c_t] \left( \sum_{p=1}^{P} \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \right)$$

The first term can be bounded using Assumptions 2 and 3 along with the Cauchy-Schwarz inequality which gives

$$\left( \sum_{p=1}^{P} \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \right) \leq 2LS P \sum_{t=1}^{T} \mathbb{I}[D_t].$$

This holds for any sequence of $\mathbf{x}_{t,p}$ chosen by the doubling round routine DR. We turn our attention to the second term.

$$\sum_{t=1}^{T} \mathbb{I}[D^c_t] \left( \sum_{p=1}^{P} \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \right) = \sum_{t=1}^{T} \left( \sum_{p=1}^{P} \mathbb{I}[D^c_t \cap \mathcal{E}'] \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \right) + \sum_{t=1}^{T} \left( \sum_{p=1}^{P} \mathbb{I}[D^c_t \cap \mathcal{E}^c'] \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \right)$$

Notice that:

$$\sum_{t=1}^{T} \sum_{p=1}^{P} \mathbb{I}[D^c_t \cap \mathcal{E}'] \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \leq \sum_{t=1}^{T} \sum_{p=1}^{P} 2 \mathbb{I}[D^c_t \cap \mathcal{E}^c'] \leq 2LSP \mathbb{I}[\mathcal{E}']$$

And therefore we can forget this term when we condition on $\mathcal{E}'$, an event that occurs with probability at least $1 - 2\delta$ (recall $\mathcal{E}'$ as the event from Lemma 6). It remains to bound the first term of Equation 25.

$$\sum_{t=1}^{T} \sum_{p=1}^{P} \mathbb{I}[D^c_t \cap \mathcal{E}'] \langle \mathbf{x}_{t,p}^* - \mathbf{x}_{t,p}, \theta^* \rangle \leq \sum_{t=1}^{T} \sum_{p=1}^{P} 8 \left( \gamma_{t,p}(\delta) \mathbb{I}[D^c_t \cap \mathcal{E}'] \mathbb{E} \left[ \| \mathbf{x}_{t,p}^* \| \mathbf{V}^{-1}_{t-p} | \mathcal{F}_{t-p} \right] + \gamma_{t,p}(\delta) \mathbb{I}[D^c_t \cap \mathcal{E}'] \| \mathbf{x}_{t,p} \| \mathbf{V}^{-1}_{t}\right)$$

$$\leq \left( \frac{8}{3\delta} + 1 \right) \gamma_T(\delta) \left( \sum_{t=1}^{P} \sum_{p=1}^{P} \| \mathbf{x}_{t,p} \| \mathbf{V}^{-1}_{t-p} \right) + \frac{2L}{\sqrt{\lambda}} \sqrt{TP \log(1/\delta)}$$

Inequality (i) holds by Lemma 8 and the assumption that $\delta \leq \frac{1}{6}$. Inequality (ii) holds because all terms $\| \mathbf{x}_{t,p}^* \| \mathbf{V}^{-1}_{t-p}$ and $\| \mathbf{x}_{t,p} \| \mathbf{V}^{-1}_{t}$ are nonnegative. Inequality (iii) holds with probability at least $1 - \delta$ and is a consequence of lexicographic monotonicity (in $t, p$) of $\gamma_{t,p}(\delta)$ and Lemma 10. The last two inequalities are a simple consequence of the determinant lemma (Lemma 11).
A.4 Auxiliary Results

Here we summarize the self-normalized vector martingale inequality used to establish the confidence ball for the least-squares estimator in a well-specified linear model,

\[ r_{t,p} = x_{t,p}^T \theta^* + \xi_{t,p}. \tag{26} \]

Here \( \xi_{t,p} \) is an i.i.d. noise process.

**Theorem 6.** ([Theorem 1 in Abbasi-Yadkori et al. [2011]]) For all \( t \in \mathbb{N} \):

\[ \| \hat{\theta}_t - \theta^* \|_{V_{t,1}} \leq \sqrt{\beta_t(\delta)} \]

with probability at least \( 1 - \delta \). Moreover, on this event by definition,

\[ \theta^* \in C(\hat{\theta}_t, V_{t,1}, \beta_t(\delta)). \]

We can now prove a generalization of this result which applies to the analysis of the lazy LinUCB algorithm in a well-specified model.

**Theorem 7.** Let \( N \subseteq [T] \) be the set of rounds which are not doubling rounds (see Condition 1). Then,

\[ \forall t \in N, \quad \| \hat{\theta}_t - \theta^* \|_{V_{t,p}} \leq \sqrt{2} \sqrt{\beta_t(\delta)} \]

with probability at least \( 1 - \delta \). Moreover, on this event by definition,

\[ \theta^* \in C(\hat{\theta}_t, V_{t,p}, 2\beta_t(\delta)). \]

**Proof.** Let \( N \subseteq \mathbb{N} \) be the set of rounds which are not doubling rounds. Then if \( t \in N \),

\[ \| \hat{\theta}_t - \theta^* \|_{V_{t,p}} \leq \sqrt{2} \| \hat{\theta}_t - \theta^* \|_{V_{t,1}} \]

from the definition in Condition 1. Hence,

\[ P[\| \hat{\theta}_t - \theta^* \|_{V_{t,p}} \geq \sqrt{2} \sqrt{\beta_t(\delta)}, t \in N] \leq P[\| \hat{\theta}_t - \theta^* \|_{V_{t,1}} \geq \sqrt{\beta_t(\delta)}, t \in N] \leq \]

\[ P[\| \hat{\theta}_t - \theta^* \|_{V_{t,1}} \geq \sqrt{\beta_t(\delta)}, \forall t \in N] \leq \delta. \]

where the final inequality follows by Theorem 6.

Define \( \mathcal{E} \) to the \( 1 - \delta \) probability event defined in Theorem 7:

\[ \mathcal{E} := \{ \forall t \in N, \quad \| \hat{\theta}_t - \theta^* \|_{V_{t,p}} \leq \sqrt{2} \sqrt{\beta_t(\delta)} \} \]

We are now in a position to prove generalizations of these results which provide valid confidence sets for the linear regression estimator in misspecified models. In summary, the confidence sets are modified with a growing, additive correction to accommodate the bias arising from the misspecification.

**Theorem 8.** If the rewards are generated from a model satisfying Assumption 3, then for all \( t \in \mathbb{N} \):

\[ \| \hat{\theta}_t - \theta^* \|_{V_{t,1}} \leq \sqrt{\beta_t(\delta)} + \sqrt{(t - 1)P\epsilon} \]

with probability at least \( 1 - \delta \). Moreover, on this event by definition,

\[ \theta^* \in C(\hat{\theta}_t, V_{t,1}, \beta_t(\delta), \epsilon). \]

**Proof.** The argument uses a bias-variance decomposition. First, define the linearized reward \( \tilde{r}_{a,b} = x_{a,b}^T \theta^* + \xi_{a,b} \) and linear estimator using these rewards as \( \hat{\theta} = V_{t,1}^{-1}(\sum_{a=1}^{t-1} \sum_{b=1}^{P} x_{a,b} \tilde{r}_{a,b}) \). By definition, \( r_{a,b} - \tilde{r}_{a,b} = \epsilon_{a,b} \) (which all satisfy \( |\epsilon_{a,b}| \leq \epsilon \) uniformly for all \( x_{a,b} \) by assumption). Then,

\[ \hat{\theta}_t - \theta^* = \hat{\theta} - \theta^* = \hat{\theta} - \theta^* + V_{t,1}^{-1}(\sum_{a=1}^{t-1} \sum_{b=1}^{P} x_{a,b} \epsilon_{a,b}) \implies \]

\[ \| \hat{\theta}_t - \theta^* \|_{V_{t,1}} \leq \| \hat{\theta}_t - \theta^* \|_{V_{t,1}} + \| \sum_{a=1}^{t-1} \sum_{b=1}^{P} x_{a,b} \epsilon_{a,b} \|_{V_{t,1}} \]

\[ \| \hat{\theta}_t - \theta^* \|_{V_{t,1}} \leq \| \hat{\theta}_t - \theta^* \|_{V_{t,1}} + \sqrt{2} \sqrt{\beta_t(\delta)} + \sqrt{(t - 1)P\epsilon} \]

26
The first term can be bounded by $\sqrt{\beta_t(\delta)}$ with probability $1 - \delta$ exactly by using Theorem 6. Using the projection bound in Zanette et al. [2020, Lemma 8] it follows that,
\begin{equation}
\| \sum_{a=1}^{t-1} \sum_{p=1}^{P} x_{a,b} \epsilon_{a,b} \|_{Y_i}^{-1} \leq \sqrt{(t-1)P\epsilon}
\end{equation}

since $|\epsilon_{a,b}| \leq \epsilon$ uniformly for all $a,b$.

The analogue for the lazy confidence set follows similarly,

**Theorem 9.** Let $N \subseteq [T]$ be the set of rounds which are not doubling rounds (see Condition 1). Then, If the rewards are generated from a model satisfying Assumption 3, for all $t \in N$:
\begin{equation}
\| \hat{\theta}_t - \theta^* \|_{V_{i,p}} \leq \sqrt{2(\sqrt{\beta_t(\delta)} + \sqrt{(t-1)P\epsilon})}
\end{equation}

with probability at least $1 - \delta$. Moreover, on this event by definition,
\begin{equation}
\theta^* \in C(\hat{\theta}_t, V_{t,1}, 2\beta_t(\delta), 2\epsilon).
\end{equation}

**Proof.** First, if $t \in N$,
\begin{equation}
\| \hat{\theta}_t - \theta^* \|_{V_{i,p}} \leq \sqrt{2\| \hat{\theta}_t - \theta^* \|_{V_{i,1}}}
\end{equation}

from the definition in Condition 1. The remainder of the argument follows as in the previous result,
\begin{equation}
\mathbb{P}[\| \hat{\theta}_t - \theta^* \|_{V_{i,p}} \geq \sqrt{2(\sqrt{\beta_t(\delta)} + \sqrt{(t-1)P\epsilon})}, t \in N] \leq \mathbb{P}[\| \hat{\theta}_t - \theta^* \|_{V_{i,1}} \geq \sqrt{\beta_t(\delta)} + \sqrt{(t-1)P\epsilon}, t \in N] \leq \mathbb{P}[\| \hat{\theta}_t - \theta^* \|_{V_{i,1}} \geq \sqrt{\beta_t(\delta)} \forall t \in N] \leq \delta.
\end{equation}

where the final inequality follows by Theorem 6, since $\tilde{\theta} = V_{t,1}^{-1}(\sum_{a=1}^{t-1} \sum_{b=1}^{P} x_{a,b} \epsilon_{a,b})$ is the estimator utilizing the linearized rewards. \hfill \Box

Next we recall the elliptical potential lemma which control the volumetric growth of the space spanned by sequence of covariance matrices.

**Lemma 11.** [Lemma 19.4 in Lattimore and Szepesvári [2020]] Let $V_0 \in \mathbb{R}^{d \times d}$ be a positive-definite matrix, $x_1, \ldots, x_n \in \mathbb{R}^d$ be a sequence of vectors with $\|x_i\| \leq L$ for all $i \in [n]$, and $V_n = V_0 + \sum_{s \leq n} x_s x_s^\top$. Then,
\begin{equation}
\sum_{s=1}^{n} \log(1 + \|x_s\|_{V_{s-1}}^2) = \log \left( \frac{\det V_n}{\det V_0} \right)
\end{equation}

\begin{equation}
\log \left( \frac{\det V_n}{\det V_0} \right) \leq d \log \left( \frac{\text{tr}(V_0) + nL^2}{d} \right) - \log(\det V_0)
\end{equation}

Finally, we state a prove a simple fact from linear algebra.

**Lemma 12.** If $A \succeq B > 0$,
\begin{equation}
\forall x \neq 0 \in \mathbb{R}^d, \quad \frac{x^\top Ax}{x^\top Bx} \leq \frac{\det(A)}{\det(B)}.
\end{equation}

**Proof of Lemma 12.** To this end, let $x = B^{-1/2}y$ for $y \in \mathbb{R}^d$. Then note that,
\begin{equation}
\sup_{x \neq 0} \frac{x^\top Ax}{x^\top Bx} = \|B^{-1/2}AB^{-1/2}\|_2
\end{equation}

by the definition of the operator norm. Similarly, we can rewrite $\frac{\det(A)}{\det(B)} = \det(B^{-1/2}AB^{-1/2})$. The claim then follows because all the eigenvalues of $B^{-1/2}AB^{-1/2}$ are $\geq 1$ (note $\inf_{x,\|x\|_2=1} x^\top B^{-1/2}AB^{-1/2}x = 1 + x^\top B^{-1/2}(A - B)B^{-1/2}x \geq 1$ since $A \succeq B$). \hfill \Box
Proof of Corollary 4. As a consequence of Theorem 1 we have for any choice of doubling round routine that the regret of Algorithm 1 and Algorithm 2 obeys,

$$\mathbb{E}[\mathcal{E}] R(T, P) \leq \mathbb{E}[\mathcal{E}] \left( LSP \sum_{t=1}^{T} \mathbb{I}[D_t] + \sqrt{TP} \max(\sqrt{\beta_T(\delta)} + \sqrt{TP} e, LS) \cdot \sqrt{d\log \left( 1 + \frac{TPL^2}{\lambda} \right) + eTP} \right)$$

where $\sqrt{\beta_T(\delta)} \leq \mathcal{R} \sqrt{d\log \left( 1 + \frac{TPL^2}{\lambda} \right) + \sqrt{T}\alpha}$. This inequality holds on the event $\mathcal{E}$ which occurs with probability at least $1 - \delta$ for both Algorithm 1 and Algorithm 2. Now we introduce an additional event $\mathcal{G} = \{ \sum_{t=1}^{T} \mathbb{I}[D_t] \geq \lceil \frac{10^5 L^2\pi_{\max}^2}{\pi_{\min}^2} \log(\frac{4d}{\delta}) \rceil \}$. Then we can consider two cases,

- First,

$$\mathbb{I}[\mathcal{G}^c](A + B) \leq \mathbb{I}[\mathcal{G}^c] \left( LSP \frac{10^5 L^2\pi_{\max}^2}{P} \frac{\log(\frac{4d}{\delta})}{\pi_{\min}^2} + B \right)$$

simply by definition of the event.

- Second, by definition of the $V_{t,p}$ and $\mathcal{G}$ all rounds which are not doubling rounds (denoted by the set $N$) lead to randomly generated covariates being used to estimate the covariance. Let $r_1 \in N$ be the first doubling round to exceed the threshold $\lceil \frac{10^5 L^2\pi_{\max}^2}{\pi_{\min}^2} \log(\frac{4d}{\delta}) \rceil$. We claim (with high probability) there can be no more doubling rounds after round $r_1$. This follows since first by Lemma 13 on the event $\mathcal{G}$,

$$\| \sum_{a \in N} \sum_{p=1}^{P} x_{a,p}x_{a,p}^\top - (\Sigma_{\pi_{a,p}} + \mu_{a,p}\mu_{a,p}^\top) \|_2 \leq \begin{cases} \frac{1}{10^5 \pi_{\min}^2} \| N \|_P \quad \text{when } \| N \| > 1 \\ \frac{1}{10^5 \pi_{\min}^2} \| N \|_P \quad \text{when } \| N \| = 1 \end{cases}$$

with probability at least $1 - \delta$ (denote this further event $C$). This previous inequality follows by considering the two separate cases where $\| N \| > 1$ and $\| N \| = 1$ respectively. Thus in any round $t > r_1$ we must have with probability $1 - \delta$,

$$V_{t,1} \geq \sum_{a \in N} \sum_{p=1}^{P} (\Sigma_{\pi_{a,p}} + \mu_{a,p}\mu_{a,p}^\top) - \frac{\pi_{\min}^2}{10^5} \| N \|_P \geq \frac{9}{10^5} \| N \|_P \pi_{\min}^2 \mathbf{I} \quad \text{when } \| N \| > 1$$

$$V_{t,1} \geq P(\Sigma_{\pi_{t,p}} + \mu_{1,p}\mu_{1,p}^\top) - \frac{\ell^2 P}{10} \mathbf{I} \geq \frac{9}{10^5} P\ell^2 \mathbf{I} \quad \text{when } \| N \| = 1$$

using Definition 1. In summary for all $\| N \| > 1$ we then have that

$$V_{t,1} \geq \frac{9}{10^5} \| N \|_P \pi_{\min}^2 \mathbf{I}.$$ 

If, additionally it is the case that $\| N \| \geq \lceil \frac{10^5 \chi^2}{9 \pi_{\min}^2} \rceil$, let $r_2$ be the first round after which this occurs. Then it follows that for all $t > \max(r_1, r_2)$,

$$\mathbb{I}[\mathcal{C}] \left( \sum_{p=1}^{P} x_{t,p}x_{t,p}^\top \right) \leq \mathbb{I}[\mathcal{C}] P\chi^2 \mathbf{I} \leq \frac{9}{10^5} \| N \|_P \pi_{\min}^2 \mathbf{I} \leq V_{t,1},$$

so no $t > \max(r_1, r_2)$ can be a doubling round on this event. Concluding we have that,

$$\mathbb{I}[\mathcal{G}] \mathbb{I}[\mathcal{C}] A \leq LSP \cdot \max(\lceil \frac{10^5 L^2\pi_{\max}^2}{P} \frac{\log(\frac{4d}{\delta})}{\pi_{\min}^2} \rceil, \lceil \frac{10^5 \chi^2}{9 \pi_{\min}^2} \rceil).$$
Algorithm 5, labeled as \( E \) where both

\[
\mathbb{E}[x_i] = \mu_i, \quad \text{for all } i, \quad \mathbb{E}[\|x_i\|_2^2] = \sigma_i^2,
\]

Proof. Let Definition 1 and Assumption 2 hold and consider \( \mathbb{E}[x_i] = \mu_i \) i.i.d. copies of sets (with \( P \) elements) sampled from Algorithm 5, labeled as \( \{x_{i,p}\}_{i=1}^{N_p} \). Then,

\[
\mathbb{E}[\|x_{i,p}\|_2^2] = \sigma_i^2, \quad \text{for all } i, \quad \mathbb{E}[\|x_{i,p}\|_2] = \sigma_i.
\]

We now present the matrix concentration result we use, labeled as \( \{x_{i,p}\}_{i=1}^{N_p} \).

Lemma 13. Let Definition 1 and Assumption 2 hold and consider \( N \) i.i.d. copies of sets (with \( P \) elements) sampled from Algorithm 5, labeled as \( \{x_{i,p}\}_{i=1}^{N_p} \). Then,

\[
\mathbb{E}[\|x_{i,p}\|_2^2] = \sigma_i^2, \quad \text{for all } i, \quad \mathbb{E}[\|x_{i,p}\|_2] = \sigma_i.
\]

We now apply the matrix Bernstein inequality to control this first term [Tropp, 2012, Theorem 1.6.2]. Note that for all \( i \),

\[
\mathbb{E}[\|x_{i,p} - \mu_{i,p}\|_2^2] \leq 2L^2 \text{ and the matrix variance is bounded by } \mathbb{E}[\|\Sigma_{\pi_{i,p}}\|_2^2] \leq 2L^2 \sigma_{\pi_{i,p}}^2.
\]

Thus we obtain,

\[
\mathbb{E}[\|x_{i,p} - \mu_{i,p}\|_2^2] \leq 2L^2 \sigma_{\pi_{i,p}}^2.
\]

Together, we obtain,

\[
\mathbb{I}[\mathbb{E}[\|x_{i,p}\|_2^2] = \sigma_i^2, \quad \text{for all } i, \quad \mathbb{E}[\|x_{i,p}\|_2] = \sigma_i]
\]

Assembling and summing these two cases, then shows that,

\[
\mathbb{I}[\mathbb{C}((A + B) \leq (LSP \cdot \max(\frac{10^5 L^2 \sigma_{\max}^2}{\sigma_{\pi_{\min}}^4} \log(\frac{4d}{\delta}), \frac{10L^2}{9\ell_1^2}) + B).
\]

So it follows that,

\[
\mathbb{I}[\mathbb{E}[\|x_{i,p}\|_2^2] = \sigma_i^2, \quad \text{for all } i, \quad \mathbb{E}[\|x_{i,p}\|_2] = \sigma_i]
\]

Hiding logarithmic factors, this implies that,

\[
\mathbb{I}[\mathbb{E}[\|x_{i,p}\|_2^2] = \sigma_i^2, \quad \text{for all } i, \quad \mathbb{E}[\|x_{i,p}\|_2] = \sigma_i]
\]

where both \( \mathcal{E} \) and \( \mathcal{C} \) hold with probability \( 1 - \delta \). We can simplify the first term to,

\[
LSP \cdot \max(\frac{10^5 L^2 \sigma_{\max}^2}{\sigma_{\pi_{\min}}^4} \log(\frac{4d}{\delta}), \frac{10L^2}{9\ell_1^2}) \leq \tilde{O}(LSP \cdot \frac{10^5 L^2 \sigma_{\max}^2}{\sigma_{\pi_{\min}}^4} \log(\frac{4d}{\delta}) + 1 + \frac{\chi^2}{\ell_1^2}) \leq \tilde{O}(L(\frac{L^4}{\ell_1^2} + P \frac{\chi^2}{\ell_1^2}))
\]

with probability at least \( 1 - 2\delta \).

An identical argument establishes the result for the Thompson sampling algorithms save with Thompson sampling regret \( \mathcal{R}(T, P) \) used in place of the LinUCB regrets in the previous argument.

We now present the matrix concentration result we use, labeled as \( \{x_{i,p}\}_{i=1}^{N_p} \).
with probability at least $1 - \delta/2$. For the second term note by Jensen’s inequality that $\|\mu_i\|_2 = \|\mathbb{E}[x_i]\|_2 \leq \mathbb{E}[\|x_i\|_2] \leq L$ since $\|x_i\|_2 \leq L$. An identical calculation before to bound almost surely bound this term and its matrix variance we have that,

$$\frac{1}{NP} \sum_{i=1}^N \sum_{p=1}^P \mu_{i,p}(x_i - \mu_{i,p}) - \mu_{i,p}\mu_{i,p}^\top \leq 4 \left( L \sqrt{\frac{\sigma^2}{\max} \sqrt{\frac{\log(4d/\delta)}{NP}}} + \frac{L^2 \log(4d/\delta)}{NP} \right)$$

by the matrix Bernstein inequality with probability at least $1 - \delta/2$. Summing the terms and applying a union bound over the events on which they hold gives the result. \qed

C Proofs in Section 4

Here we include the proof of the main lower bound.

Proof of Theorem 3. The proof follows by first noting that in each of the instances claimed a single, fixed global context vector is used for all time and processors. Hence the parallel to sequential regret reduction established in Proposition 11 is applicable. Thus it suffices to establish the lower bounds for $R(TP, 1)$ in lieu of $R(T, P)$ for the instances claimed. The first term/result is an immediate consequence of Lemma 14. The second term/result we can obtain from Lemma 16. For the validity of the results, we inherit the constraints $d \geq [8 \log(m) L^2 / \epsilon^2]$ and $S \geq \|\theta^*\|_2 = \frac{\epsilon}{L} \sqrt{\frac{d-1}{8 \log(m)}} \implies \epsilon^2 \leq \frac{8(LS)^2 \log(m)}{d-1}$. If we take $d = [8 \log(m)/\epsilon^2]$ then the second constraint reduces too $\epsilon^2 \leq \frac{8(LS)^2 \log(m)}{8 \log(m)/\epsilon^2} \implies LS \geq 1$. \qed

C.1 Parallel to Sequential Reduction

Here we establish the reduction from parallel regret to sequential regret when considering parallel linear bandits where there is a single fixed action set/context set across all processors at a given time. So $X_{t,p} = X_t$ across all processors $p \in [P]$. Additionally, we assume as in the preamble that the reward of an action $x$ is determined by $r = f(x) + \epsilon$ so the law of the rewards is completely specified by a mean reward function $f$ and mean-zero noise distribution $\epsilon$.

We formalize the reduction from parallel to sequential bandits by first defining the canonical bandit environment. We consider the bandit instance to be indexed by the law of the rewards and the sequence of context sets.

Sequential In a model of purely sequential interaction we consider instances defined by two ingredients:

- the conditional distribution of the policy $\pi_t(\cdot | x_{i\leq t}, X_{i\leq t}, r_{i\leq t})$.
- the sequence of reward distributions $\nu_s \equiv \mathbb{P}_t(\cdot | x_{i\leq t}, X_{i\leq t}, r_{i\leq t}, f) \equiv \mathbb{P}_{x_t}(\cdot | f)$ for selected actions and the sequence of presented contexts $X_t$.

Parallel In a model of parallel interaction, there are two ingredients:

- the conditional distribution of policy $\psi_{t,p}(\cdot | x_{i,j \leq t,p}, X_{i\leq t}, r_{i,j \leq t,p})$.
- the sequence of reward distributions $\nu_p \equiv \mathbb{P}_{t,p}(\cdot | x_{i,j < t,p}, X_{i\leq t}, r_{i,j < t,p}) \equiv \mathbb{P}_{x_{t,p}}(\cdot | f)$ for selected actions and sequence of selected contexts $X_t$ (which for fixed $t$ are equal across all $p \in [P]$).

To formalize the reduction we make the following claim:

Proposition 10. If we consider the lexicographic ordering for $t, p \in [T, P]$, then for any sequence of parallel policy-environment interactions with law $(\psi_{t,p}(\cdot), \mathbb{P}_{t,p}(\cdot | f))$ and presented context sets $X_t$ (identical across $p \in [P]$), there exists a corresponding coupling to a purely sequential bandit environment $(\pi_m(\cdot), \mathbb{P}_m(\cdot | f))$ for $m \in [TP]$ and sequence of context sets $X_m = X_t$ for $m \in [TP, (t + 1)P]$ with an identical distribution.

Proof. We can construct the sequential environment inductively from the sequence of parallel interaction by coupling. To see this consider the first round of parallel interactions which are described by the measure,

$\Pi_{p=1}^P \psi_{1,p}(\cdot | x_{i,j < 1,p}, X_1, r_{i,j < 1,1}) \mathbb{P}_{x_{1,p}}(\cdot | f)$. 

30
By defining the measure over sequential policy-environment interactions,

$$\Pi_m^{\mathcal{P}} \pi_m(\cdot| x_{a<m}, r_{a<m}) \mathbf{P}_{X_{a<m}}(\cdot| f)$$

we can set, $$\pi_m(\cdot| x_{a<m}, r_{a<m}) = \psi_1, p(\cdot| x_{i,j<1}, p_1, r_{i,j<1,1})$$ by ignoring the conditioning on the further contexts and reward information in the sequential interaction to enforce equality of the policies. Since the policies are taken to be identical, by coupling the randomness between the sequential and parallel policies/environments, the sequence of selected actions will be identical (this uses the fact the context sets in the parallel block and sequential interaction can be taken equal). Inductively proceeding with the construction over the blocks of parallel interaction completes the argument.

With this claim in hand the reduction follows since the expected regret of an algorithm over an environment (in expectation) is determined by the law of the policy-environment interactions. Before beginning we first introduce the following notation for the expected parallel regret (further indexed by policy and environment,

$$R_{\psi, \nu_p}(T, P) = \sum_{t=1}^T \sum_{p=1}^P \max_{x \in \mathcal{X}} \mu(x) - \mathbb{E}_{\psi, \nu_p}[\mu(x)]$$

where $$\psi$$ denotes the parallel policy and $$\nu_p$$ the parallel environment indexed by the mean reward and sequence of context sets. Similarly, the sequential regret can be defined analogously as,

$$R_{\pi, \nu_s}(TP, 1) = \sum_{a=1}^{TP} \max_{x \in \mathcal{X}} \mu(x) - \mathbb{E}_{\pi, \nu_s}[\mu(x)]$$

where $$\pi$$ denotes the sequential policy and $$\nu_s$$ the sequential environment indexed by the mean reward and sequence of context sets which are fixed to be the same over consecutive blocks of length $$P$$.

**Proposition 11.** Consider both a parallel bandit policy/environment class and sequential bandit policy/environment class as defined in Proposition 10. Then,

$$\inf_{\psi} \sup_{\nu_p} R_{\psi, \nu_p}(T, P) \geq \inf_{\pi} \sup_{\nu_s} R_{\pi, \nu_s}(TP, 1)$$

where the infima over $$\psi$$ is taken place over the class of all parallel policies and the infima on the right hand side is taken over the class of all sequential policies.

**Proof.** By the the preceding claim in Proposition 10 every pair of parallel $$(\psi, \nu_p)$$ measures can be reproduced by a corresponding sequential measure $$(\pi, \nu_s)$$. Let the set of such induced sequential policies be $$\mathcal{P}$$. So the pointwise inequality,

$$R_{\psi, \nu_p}(T, P) = R_{\pi, \nu_s}(TP, 1)$$

holds by simply re-indexing the summation in lexicographic order since the expectations are identical. Hence, for a fixed $$\psi$$ (and induced $$\pi$$) the equality also holds after taming a supremum over the same indexing set on both sides (parameterized by $$f$$ and $$\mathcal{X}_t$$),

$$\sup_{\nu_p} R_{\psi, \nu_p}(T, P) = \sup_{\nu_s} R_{\pi, \nu_s}(TP, 1)$$

Now taming an infima over the policy class $$\psi$$ and equivalent induced sequential policy class shows,

$$\inf_{\psi} \sup_{\nu_p} R_{\psi, \nu_p}(T, P) = \inf_{\pi \in \mathcal{P}} \sup_{\nu_s} R_{\pi, \nu_s}(TP, 1) \geq \inf_{\pi} \sup_{\nu_1} R_{\pi, \nu_s}(TP, 1)$$

by relaxing the final infima to tamed place over the class of all sequential policies. 

**C.2 Unit Ball Lower Bound**

Here we record the lower bound over a fixed action set for the unit ball for a sequential bandit instance. The proof is an immediate generalization of [Lattimore and Szepesvári, 2020, Theorem 24.2] with the scales restored. Throughout this section we assume that the rewards are generated as,

$$r_{t,p} = x_{t,p}^T \theta^* + \xi_{t,p}$$

where $$\xi_{t,p} \sim \mathcal{N}(0, R^2)$$. 

31
Lemma 14. Let the fixed action set $X = \{x \in \mathbb{R}^d : \|x\|_2 \leq L\}$ and parameter set $\Theta = \{\pm \Delta\}^d$ for $\Delta = \frac{R\sqrt{d}}{3\sqrt{2TL}}$. If $T \geq d \max(1, \frac{1}{3\sqrt{2dSNR}})$, then for any policy $\pi$, there is a vector $\theta^* \in \Theta$ such that:

$$R_{\pi, (x, \theta)}(T, 1) \geq \Omega \left( Rd\sqrt{T} \right).$$

Proof of Lemma 14. Let $A = \{x \in \mathbb{R}^d : \|x\|_2 \leq L\}$ and $\theta \in \mathbb{R}^d$ such that $\|\theta\|_2^2 = S^2$. Let $\Delta = \frac{R\sqrt{d}}{3\sqrt{2TL}}$ and $\theta \in \{\pm \Delta\}^d$ and for all $i \in [d]$ define $\tau_i = \min(T, \min(t : \sum_{t=1}^T x_{t,i}^2 \geq \frac{\alpha n}{d})$. We will set $\alpha = \ell^2$ at the end of the proof but in order to make the derivations clearer and easier to read we will keep this $\alpha$ explicit. Then for any policy $\pi$:

$$R_{\pi, (x, \theta)}(T, 1) = \Delta E_\theta \left[ \sum_{t=1}^T \sum_{i=1}^d \left( \frac{L}{\sqrt{d}} - x_{t,i} \text{sign}(\theta_i) \right) \right]$$

$$\geq \frac{\Delta \sqrt{d}}{2L} E_\theta \left[ \sum_{t=1}^T \sum_{i=1}^d \left( \frac{L}{\sqrt{d}} - x_{t,i} \text{sign}(\theta_i) \right)^2 \right]$$

$$\geq \frac{\Delta \sqrt{d}}{2L} \sum_{i=1}^d E_\theta \left[ \sum_{t=1}^{\tau_i} \left( \frac{L}{\sqrt{d}} - x_{t,i} \text{sign}(\theta_i) \right)^2 \right]$$

Where the first inequality uses that $\|x_i\|_2^2 \leq L^2$. Fix $i \in [d]$ and let $x \in \{-1, 1\}$, define $U_i(x) = \sum_{t=1}^{\tau_i} \left( \frac{L}{\sqrt{d}} - x_{t,i} x \right)^2$. Assume without loss of generality that $\theta_i > 0$. Let $P_\theta$ and $P_{\theta'}$ be the laws of $U_i(1)$ w.r.t. the bandit learner interaction measure induced by $\theta$ and $\theta'$ respectively. By a simple calculation we conclude that:

$$\text{KL}(P_\theta, P_{\theta'}) \leq 2 \frac{\Delta^2}{4R^2 E_\theta} \left[ \sum_{t=1}^{\tau_i} x_{t,i}^2 \right]$$

(37)

Also, observe that:

$$U_i(1) = \sum_{t=1}^{\tau_i} \left( \frac{L}{\sqrt{d}} - x_{t,i} \right)^2 \leq 2L^2 \sum_{t=1}^{\tau_i} \frac{1}{d} + 2 \sum_{t=1}^{\tau_i} x_{t,i}^2 \leq \left( \frac{2L^2T + 2\alpha T}{d} + 2L^2 \right)$$

Then:

$$E_\theta[U_i(1)] \geq E_{\theta'}[U_i(1)] - \left( \frac{2L^2T + 2\alpha T}{d} + 2L^2 \right) \text{KL}(P_\theta, P_{\theta'})$$

$$\geq E_{\theta'}[U_i(1)] - \left( \frac{2L^2T + 2\alpha T}{d} + 2L^2 \right) \frac{\Delta}{2R} \sqrt{E_\theta \left[ \sum_{t=1}^{\tau_i} x_{t,i}^2 \right]}$$

$$\geq E_{\theta'}[U_i(1)] - \left( \frac{2L^2T + 2\alpha T}{d} + 2L^2 \right) \frac{\Delta}{2R} \sqrt{\frac{T\alpha}{d} + L^2}$$

$$\geq E_{\theta'}[U_i(1)] - \left( \frac{2L^2T + 4\alpha T}{d} \right) \frac{\Delta}{2R} \sqrt{\frac{2T\alpha}{d}}$$

The last inequality follows by assuming $\alpha n \geq dL^2$, which holds because by assumption $d \geq L^2$ (recall $\alpha = L^2$). We can then conclude that:
\[ \mathbb{E}_\theta [U_i(1)] + \mathbb{E}_\theta [U_i(-1)] \geq \mathbb{E}_\theta [U_i(1) + U_i(-1)] - \left( \frac{2L^2T + 4\alpha T}{d} \right) \frac{\Delta}{2R} \sqrt{\frac{2T\alpha}{d}} \]

\[ = 2\mathbb{E}_\theta \left[ \frac{\tau_i L^2}{d} + \sum_{i=1}^n x^2_{i,i} \right] - \left( \frac{2L^2n + 4\alpha T}{d} \right) \frac{\Delta}{2R} \sqrt{\frac{2T\alpha}{d}} \]

\[ \geq \min \left( \frac{2\alpha T}{d}, \frac{2L^2 T}{d} \right) - \left( \frac{2L^2 T + 4\alpha T}{d} \right) \frac{\Delta}{2R} \sqrt{\frac{2T\alpha}{d}} \]

Therefore using the Randomization Hammer we conclude there must exist a parameter \( \theta \in \{ \pm \Delta \}^d \) such that \( R_{\pi,(x,\theta)}(T, 1) \) such that:

\[ R_{\pi,(x,\theta)}(T, 1) \geq d \frac{\Delta}{2L} \left( \min \left( \frac{2\alpha T}{d}, \frac{2L^2 T}{d} \right) - \left( \frac{2L^2 T + 4\alpha T}{d} \right) \frac{\Delta}{2R} \sqrt{\frac{2T\alpha}{d}} \right). \]

Let \( \alpha = L^2 \) and \( \Delta = \frac{R\sqrt{d}}{3\sqrt{2TL}}. \) In this case:

\[ R_{\pi,(x,\theta)}(T, 1) \geq \frac{Rd}{6\sqrt{2}} \sqrt{T}. \]

The additional constraint on \( T \) comes from the fact we must have that \( S \geq \| \theta^* \|_2 = \sqrt{d}\Delta = \frac{Rd}{3\sqrt{2TL}} \implies T \geq \frac{dR}{3\sqrt{2LS}} \]

\[ \frac{d}{3\sqrt{2\SNR}}. \]

### C.3 Misspecification Lower Bound

In this section record a scale-aware version of the lower bound for misspecified linear bandits from [Lattimore et al., 2020]. For future reference, recall the definition of misspecification stated in Assumption 3 specialized to a finite context set of size \( m \): Let \( \mathcal{X} \subset \mathbb{R}^d \) be a finite context set of size \( m \). A function \( f : \mathcal{X} \to \mathbb{R} \) is \( \epsilon \)-close to linear if there is a parameter \( \theta^* \) such that for all \( \mathbf{x} \in \mathcal{X}, \)

\[ |f(\mathbf{x}) - \mathbf{x}^\top \theta^*| \leq \epsilon. \]

Before stating our main result we will require the following supporting lemma from [Lattimore et al., 2020].

**Lemma 15.** For any \( \epsilon, L > 0 \), and \( d \in \{ m \} \) with \( d \geq \lceil 8\log(m)L^2/\epsilon^2 \rceil \), and an action set \( \mathcal{X} = \{ \mathbf{x}_i \}_{i=1}^m \subset \mathbb{R}^d \) satisfying \( \| \mathbf{x}_i \|_2 = L \) for all \( i \) and such that for all \( i, j \in [m] \) with \( i \neq j \), \( |\mathbf{x}_i^\top \mathbf{x}_j| \leq L^2 \sqrt{\frac{8\log(m)}{d-1}}. \)

**Proof.** This is simply a slightly modified version of in Lattimore et al. [2020, Lemma 3.1].

**Lemma 16.** Let \( \epsilon, L > 0 \) and \( m \in \mathbb{N} \). For any \( d \in \{ m \} \) with \( d \geq \lceil 8\log(m)L^2/\epsilon^2 \rceil \) consider a finite action set \( \mathcal{X} \subset \mathbb{R}^d \) of size \( m \) satisfying \( \| \mathbf{x}_i \| = L \) for all \( i \in [m] \). Then for any policy \( \pi \), there is a parameter \( \theta^* \), \( \epsilon \)-close to a function \( f : \mathcal{X} \to \mathbb{R}^m \) for which:

\[ R_{\pi,(x,\theta^*,f)}(T, 1) \geq \epsilon \sqrt{\frac{d-1}{8\log(m)}} \frac{\min(T, m-1)}{4}. \]

Moreover, this parameter \( \theta^* \) satisfies \( \| \theta^* \|_2 = \frac{\epsilon}{T} \sqrt{\frac{d-1}{8\log(m)}}. \)

**Proof.** Observe that by assumption since \( d \geq \lceil 8\log(m)L^2/\epsilon^2 \rceil \) we have that \( \epsilon \geq L \sqrt{\frac{8\log(m)}{d}}. \) By Lemma 15, we may choose \( \mathcal{X} = \{ \mathbf{x}_i \}_{i=1}^m \subset \mathbb{R}^d \) satisfying:

1. \( \| \mathbf{x}_i \|_2 = L \) for all \( i \in [m] \).
2. \( |\mathbf{x}_i^\top \mathbf{x}_j| \leq L^2 \sqrt{\frac{8\log(m)}{d-1}} \) for all \( i \neq j \).
We now consider a family of \( m \) bandit instances indexed by each of the arms \( x_i \in \mathcal{X} \). For any \( x_i \in \mathcal{X} \) define \( \theta_i^* = \delta x_i \) with \( \delta = \frac{\epsilon}{T^2} \sqrt{\frac{d-1}{8 \log(m)}} \) and define \( f_i \) as:

\[
f_i(x) = \begin{cases} L^2 \delta & \text{if } x = x_i \\ 0 & \text{otherwise} \end{cases}
\]

By definition \( f_i \) is \( \epsilon \)-close to linear since for all \( x \in \mathcal{X} \) with \( x \neq x_i \),

\[
\langle x, \theta_i^* \rangle \leq \delta L^2 \sqrt{\frac{8 \log(m)}{d-1}} = \epsilon \quad \text{and} \quad \langle x, \theta^* \rangle = L^2 \delta
\]

Denote by \( x^{(t)} \) the action played by algorithm \( \pi \) at time \( t \) and define \( \tau_i = \max \{ t \leq n : x^{(s)} \neq x_i, \forall s \leq t \} \). Then \( E[R_{\pi, (x, \theta^*, f)}(T, 1)] \geq L^2 \delta E_i[\tau_i] \). Where \( E_i \) denotes the expectation under the law of bandit problem \( \theta^* \) and algorithm \( \pi \). Observe that any algorithm \( \pi \) that queries arm \( x_j \) with \( j \neq i \) more than once before pulling arm \( x_i \) will have a larger \( E_i[\tau_i] \) than one that only queries each arm once before querying arm \( x_i \). This means that in order to lower bound \( E_i[\tau_i] \) we can restrict ourselves to algorithms \( \pi \) that do not repeat an arm pull before \( \tau_i \). In fact we can assume algorithm \( \pi \) behaves the same for all \( i \in [m] \). Observe that for such algorithms whenever facing problem \( \theta^* \) and \( t \leq \tau_i \), the law of the rewards is independent of \( x_i \) for all \( i \in [m] \). Let \( f_0 : \mathcal{X} \rightarrow \mathbb{R} \) denote the zero function such that \( f(x) = 0 \) for all \( x \in \mathcal{X} \) and let \( E_0 \) be the expectation under the law of the bandit problem induced by \( f_0 \) and algorithm \( \pi \). Let \( T_i \) be the first time that algorithm \( A \) encountered arm \( i \) when interacting with \( f_0 \). Observe that since the interactions of \( \pi \) with \( f_i \) before \( \tau_i + 1 \) and the interactions of \( A \) with \( f_0 \) before \( A \) pulls \( x_i \) (or the time runs up) are indistinguishable \( E_i[\tau_i] = E_0[\min(T, T_i - 1)] \). A simple averaging argument shows that

\[
\frac{1}{m} \sum_{i=1}^{m} R_{\pi, (x, \theta^*, f)}(T, 1) \geq \frac{L^2 \delta}{m} \sum_{i=1}^{m} E_i[\tau_i] = \frac{L^2 \delta}{m} \sum_{i=1}^{m} E_0[\min(T, T_i - 1)] = \frac{L^2 \delta}{m} E_0 \left[ \sum_{i=1}^{m} \min(T, T_i - 1) \right]
\]

Since \( \pi \) is assumed to interact with \( f_0 \) by never pulling the same arm twice, \( \{T_i - 1\}_{i=1}^{m} = \{i - 1\}_{i=1}^{m} \). Using this fact, we can write

\[
\frac{1}{m} \sum_{i=1}^{m} \min(T, T_i - 1) = \frac{\min(m, T)}{m} \sum_{i=1}^{m} i - 1 = \frac{\min(m, T)}{m} \sum_{i=1}^{m} i - 1 + 1(T \leq m - 1)T.
\]

In order to bound the expression above we analyze two cases, first when \( T \leq m \), and second when \( T > m \). In the first case

\[
\frac{1}{m} \sum_{i=1}^{m} i - 1 + 1(T \leq m - 1)T = \frac{1}{m} \left( \frac{T(T - 1)}{2} + T(m - T) \right)
\]

Let’s consider two sub-cases. If \( T - 1 \geq \frac{m}{2} \), then \( \frac{1}{m} \left( \frac{T(T - 1)}{2} + T(m - T) \right) \geq \frac{T}{2} \). If \( T - 1 < \frac{m}{2} \) then \( m - T > m - (\frac{m}{2} + 1) = \frac{m}{2} - 1 \). And therefore \( \frac{1}{m} \left( \frac{T(T - 1)}{2} + T(m - T) \right) \geq \frac{T}{2} - \frac{T}{m} > \frac{T}{2} - 1 \). It follows that whenever \( T \leq m \) (and \( T > 1 \)), then \( \frac{1}{m} \left( \frac{T(T - 1)}{2} + T(m - T) \right) \geq \frac{T}{2} \).

Now let’s consider the case when \( T > m \). If this holds,

\[
\frac{1}{m} \sum_{i=1}^{m} \min(T, T_i - 1) \geq \frac{\min(m, T - 1)}{4}.
\]

Assembling these facts together we conclude that in all cases,

\[
\frac{1}{m} \sum_{i=1}^{m} \min(T, T_i - 1) \geq \frac{\min(T, m - 1)}{4}.
\]

The result follows by noting this implies there must exist one \( \theta^* \) such that

\[
R_{\pi, (x, \theta^*, f)}(T, 1) \geq \epsilon \sqrt{\frac{d-1}{8 \log(m)} \frac{\min(T, m - 1)}{4}}.
\]

The result follows.
D Experiment Details

In this section we provide all of the experimental details for training the bandit algorithms.

D.1 Hyperparameters

The synthetic hyperparameters were fixed to the theoretical values. For the randomly initialized neural network experiments, the grid for the regularizer \( \lambda \), norm bound \( S \), and noise subgaussianity \( R \) were: \{0.01, 0.1, 1.0, 10.0, 100.0\} which was selected post-hoc for each experiment. For the superconductor experiments, the grid for all 3 parameters were: \{0.1, 1.0, 10.0\}. For the transcription factor binding dataset, the parameters grid was \( \lambda = \{1.0, 10.0\}, \) \( R = \{0.01, 0.1, 1.0, 10.0, 100.0, 1000.0\}, \) and \( S = \{0.01, 0.1, 1.0, 10.0, 100.0, 1000.0\} \). For \( \epsilon \)-greedy, the parameter grid was set over \( \epsilon = \{0.01, 0.02, \ldots, 0.99\} \) across all relevant experiments.

D.2 Feature Engineering for Random Neural Network

Two feature sets were considered:

- Linear features \( x_i \) was encoded as a 14-length feature vector
- Linear + Quadratic features where interaction terms of \( x_i \) were included in 105 quadratic features alongside the original 14 linear features into a 119 features.

The resulting parallel regret plots are shown in Fig. 8.

Figure 7: TFBinding parallel regret with linear features. From left to right: \( P = 1 \), \( P = 10 \), and \( P = 100 \).

Figure 8: RandomNN with Linear features. From left to right: \( P = 1 \), \( P = 10 \), and \( P = 100 \).
D.3 Feature Engineering for Transcription Factor Binding

Three feature sets were considered:

- Linear features $x_i$ was one-hot encoded into a 32-length feature vector.

- 250 random ReLU features where a $250 \times 32$ random matrix $W$ is sampled such that $W_{ij} \sim \text{Normal}(0,1)$. Then, the feature map was evaluated as:

  $$\phi(x_i) = \frac{1}{\sqrt{250}} \text{ReLU}\left(\frac{Wx_i}{\sqrt{32}}\right)$$

- Linear + Quadratic features where $x_i$ was one-hot encode and all 32 linear features and 528 quadratic features were combined into a 560-length vector.

The off-line test $R^2$ (without added noise in the $y_i$) on a train set of $TP = 2500$ matching the number of total arm queries yields values 0.15, 0.26, 0.29 for linear, ReLU, and quadratic respectively. This matches with the number of features and level of expressivity of the model class. However, as shown in Figs. 9 and 10 the linear features perform the best followed by the ReLU features, and then the quadratic features. One can gain further insight by examining the off-line test $R^2$ for a smaller training size and see that the larger feature expansions accrue variance making the linear features perform the best. This matches our understanding that model fitting is less statistically efficient in an online setting.

![Figure 9: TFBinding best arm with ReLU features. From left to right: $P = 1$, $P = 10$, and $P = 100$.](image)

![Figure 10: TFBinding best arm with quadratic features. From left to right: $P = 1$, $P = 10$, and $P = 100$.](image)

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In Advances in Neural Information Processing Systems, pages 2312–2320, 2011.
Marc Abeille, Alessandro Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Alekhu Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

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.

Christof Angermueller, David Belanger, Andreea Gane, Zelda Mariet, David Dohan, Kevin Murphy, Lucy Colwell, and D Sculley. Population-based black-box optimization for biological sequence design. *arXiv preprint arXiv:2006.03227*, 2020.

Frances H Arnold. Design by directed evolution. *Accounts of chemical research*, 31(3):125–131, 1998.

Ruth E Baker, Jose-Maria Peña, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, 14(5):20170660, 2018.

Luis A Barrera, Anastasia Vedenko, Jesse V Kurland, Julia M Rogers, Stephen S Gisselbrecht, Elizabeth J Rossin, Jaie Woodard, Luca Mariani, Kian Hong Kock, Sachi Inukai, et al. Survey of variation in human transcription factors reveals prevalent dna binding changes. *Science*, 351(6280):1450–1454, 2016.

Mohsen Bayati, Nima Hamidi, Ramesh Johari, and Khashayar Khosravi. Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. *Advances in Neural Information Processing Systems*, 33, 2020.

David Belanger, Suhani Vora, Zelda Mariet, Ramya Deshpande, David Dohan, Christof Angermueller, Kevin Murphy, Olivier Chapelle, and Lucy Colwell. Biological sequences design using batched bayesian optimization. In *NeurIPS Workshop on ML for the Sciences*, 2019.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

David H Brookes and Jennifer Listgarten. Design by adaptive sampling. *arXiv preprint arXiv:1810.03714*, 2018.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Thomas Desautels, Andreas Krause, and Joel W Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:3873–3923, 2014.

Miroslav Dudík, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Robert E Hawkins, Stephen J Russell, and Greg Winter. Selection of phage antibodies by binding affinity: mimicking affinity maturation. *Journal of molecular biology*, 226(3):889–896, 1992.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. *Advances in Neural Information Processing Systems*, 26:854–862, 2013.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461, 2013.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Jeff Schneider, and Barnabás Póczos. Parallelised bayesian optimisation via thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 133–142, 2018.
Tarun Kathuria, Amit Deshpande, and Pushmeet Kohli. Batched gaussian process bandit optimization via determinantal point processes. In *Advances in Neural Information Processing Systems*, pages 4206–4214, 2016.

Nathan Korda, Balazs Szorenyi, and Shuai Li. Distributed clustering of linear bandits in peer to peer networks. In *International Conference on Machine Learning*, pages 1301–1309. PMLR, 2016.

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

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007.

Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, 2013.

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

Fathima Aidha Shaikh and Stephen G Withers. Teaching old enzymes new tricks: engineering and evolution of glycosidases and glycosyl transferases for improved glycoside synthesis. *Biochemistry and Cell Biology*, 86(2):169–177, 2008.

Sam Sinai and Eric Kelsic. A primer on model-guided exploration of fitness landscapes for biological sequence design. *arXiv preprint arXiv:2010.10614*, 2020.

Sam Sinai, Richard Wang, Alexander Whatley, Stewart Slocum, Elina Locane, and Eric Kelsic. Adalead: A simple and robust adaptive greedy algorithm for sequence design. *arXiv preprint*, 2020.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR, 2015.

Balazs Szorenyi, Róbert Busa-Fekete, István Hegedus, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *International Conference on Machine Learning*, pages 19–27. PMLR, 2013.

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.

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

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR, 2020.

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

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. *arXiv preprint arXiv:1904.06309*, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. *arXiv preprint arXiv:2003.00153*, 2020.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems*, pages 5197–5208, 2019.