REPLICABLE LINEAR BANDITS WITH UCB BASED EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study replicable algorithms for stochastic linear bandits with optimism-based exploration. A bandit algorithm is ρ-replicable if two executions using shared internal randomness but independent reward realizations produce the same action sequence with probability at least $1-\rho$. For linear bandits with infinitely many actions, existing replicable algorithms are elimination-based and rely on discretizing the action space, leading to suboptimal dependence on the dimension d and on ρ. We first introduce `RepRidge`, a replicable ridge regression estimator that satisfies a self-normalized confidence bound whose radius is inflated by a factor of $(1+d/\rho)$. We believe this estimator is of independent interest. Building on it, we design `RepLinUCB`, a replicable optimistic algorithm that combines determinant-triggered batching with `RepRidge`, and show that its regret is $\widetilde{\mathcal{O}}(d^3\sqrt{T}/\rho)$. This improves the state-of-the-art by a factor of d/ρ and gives the first linear bandit algorithm with optimal dependence on ρ for large action sets. Because our approach does not rely on action-space discretization, it also applies to time-varying action sets, which existing results could not handle. Finally, we extend our approach to generalized linear bandits via a replicable penalized GLM estimator, `RepGLM`, and a corresponding optimistic algorithm, `RepGLMUCB`. In synthetic experiments, `RepLinUCB` achieves substantially lower regret than existing elimination-based algorithm.

1 INTRODUCTION

Replicable learning studies the design of randomized learning algorithms whose outputs remain stable across repeated executions on fresh data drawn from the same underlying distribution. This line of work was initiated by Impagliazzo et al. (2022), motivated by the broader reproducibility crisis in science and machine learning, where even small changes in sampled data can lead to substantially different conclusions (National Academies of Sciences, Engineering, and Medicine, 2019). Replicability asks for more than accurate performance: it requires that the algorithm itself reach essentially the same conclusion across independent datasets from the same source. This is particularly important in scientific and high-stakes pipelines, where downstream conclusions should not be overly sensitive to sample-level randomness. Replicability is closely connected to other central notions in learning theory, including differential privacy, stability, and adaptive generalization (Bun et al., 2023), and achieving it can entail nontrivial computational tradeoffs (Kalavasis et al., 2024). The study of replicability now spans statistical clustering (Esfandiari et al., 2023b), high-dimensional estimation and hypothesis testing (Hopkins et al., 2024), active learning (Hira et al., 2026), online learning (Ahmadi et al., 2025), and reinforcement learning (Karbasi et al., 2023b;a).

In a bandit problem, a learner repeatedly selects actions, while observing a noisy reward for the selected action, and seeks to minimize cumulative regret with respect to the best action. A bandit algorithm is ρ-replicable if, when run twice with the same internal randomness but independent reward realizations, the two executions select the same action at every round $t \in [T]$ with probability at least $1-\rho$ (Definition 2.1). Esfandiari et al. (2023a) introduced this notion and designed batched elimination algorithms for stochastic multi-armed bandits (MABs) and linear bandits, built on the replicable mean estimator of Impagliazzo et al. (2022). For MABs, Komiyama et al. (2026) showed that the additional regret due to replicability can be made of lower order in T, and Ahmadi et al. (2025) showed that a regret of $\Omega(\sqrt{T}/\rho)$ is unavoidable. Bollini et al. (2026) recently showed that optimism is compatible with replicability. Their algorithm selects an arm by maximizing an upper confidence bound (UCB) computed from replicable mean estimates, commits to that arm until its pull count doubles, and thereby limits the number of estimator calls to $\mathcal{O}(K \log T)$. Replicable bandits have also been studied under the weaker notion of list replicability (Chen et al., 2025) and in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

applications such as digital health interventions, where replicability supports consistent downstream statistical analysis across repeated trials (Zhang et al., 2025).

Large action spaces. Much less is known when the action set is large or infinite, which is the regime where linear bandits are most relevant (Dani et al., 2008; Abbasi-yadkori et al., 2011; Chu et al., 2011). For linear bandits with K actions, the guarantee of Esfandiari et al. (2023a) scales as K^2/ρ^2 and becomes vacuous as K grows. For infinite action sets, their approach discretizes the action space with a $1/T$-net, computes a G-optimal design and its core set of $\widetilde{\mathcal{O}}(d)$ actions, and runs batched elimination. At each phase, the value of every core-set action is estimated with a separate replicable mean estimator, and these estimates are combined into a least-squares parameter estimate. The resulting regret is $\widetilde{\mathcal{O}}(d^4\sqrt{T}/\rho^2)$, which is suboptimal in two ways. The replicability cost is paid once per core-set action rather than once per parameter estimate, and the dependence on ρ is quadratic, whereas the lower bound scales as $1/\rho$. The construction also requires a fixed action set across all time steps. This raises the question that motivates this paper:

Can optimistic exploration achieve the optimal dependence on ρ in linear bandits with large or infinite action sets, without discretizing the action space?

Why the finite-armed template does not transfer. The batched optimistic scheme for MABs is tied to finitely many arms in two ways. First, batches are scheduled by per-arm pull counts. This is what bounds the number of replicable estimator calls, and hence the union-bound cost of replicability, by $\mathcal{O}(K \log T)$. With infinitely many actions no such counts exist, and discretizing to recover them reintroduces a dependence on the size of the discretization. Second, replicability is obtained by rounding scalar mean estimates of individual arms. In a linear bandit, the natural object to make replicable is a d-dimensional parameter estimate whose uncertainty is described by a data-dependent confidence ellipsoid, and rounding it coordinate by coordinate in the original parameter space does not respect that geometry. We overcome these challenges and answer the question affirmatively. Our main contributions are as follows.

1. **Replicable ridge regression.** We develop `RepRidge` (Algorithm 2), a replicable ridge regression estimator. Given $\delta \in (0,1)$ and $\rho \in (3\delta, 1)$, we show that `RepRidge` returns a ρ replicable estimator $\widetilde{\theta}_n$ satisfying $\|\widetilde{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta)\Big(1 + \frac{d}{\rho - 2\delta}\Big)$, with probability at least $1-\delta$. Here $\beta_n(\delta)$ is the standard ridge confidence radius and $V_n = \lambda I + \sum_{i=1}^n x_i x_i^\top$ is the regularized Gram matrix (see Theorem 3.1). This is a replicable analogue of the self-normalized confidence bound (Abbasi-yadkori et al., 2011), showing that replicability costs a multiplicative inflation of the confidence radius, and we expect it to be useful wherever confidence sets and replicability are needed simultaneously. As a first consequence, we show that under the coverage assumption of Esfandiari et al. (2023a), this implies a uniform prediction error with a factor-d improvement in sample complexity (see Proposition 3.1).
2. **Replicable linear bandits.** We develop `RepLinUCB` (Algorithm 3) a replicable optimistic algorithm for linear bandits. We combine determinant-triggered batching with our estimator `RepRidge` at batch starts and show it is ρ-replicable, and achieves regret $\widetilde{\mathcal{O}}\big(\frac{d^3}{\rho}\sqrt{T}\big)$ (see Theorem 4.1). This improves the $\widetilde{\mathcal{O}}(d^4\sqrt{T}/\rho^2)$ bound of Esfandiari et al. (2023a) by a factor of d/ρ, and it matches the $\Omega(\sqrt{T}/\rho)$ lower bound (Ahmadi et al., 2025) in its dependence on both ρ and T. To the best of our knowledge, it is the first linear bandit algorithm with optimal dependence on ρ when the number of actions is large or infinite. The algorithm requires no discretization, elimination, or design-based core set, and it allows the action sets $(\mathcal{A}_t)_{t=1}^T$ to vary over time, chosen by an oblivious adversary (unlike existing algorithms that only work for fixed action set $\mathcal{A}$).
3. **Replicable generalized linear bandits.** We next extend our approach to generalized linear bandits (Filippi et al., 2010). We introduce `RepGLM` (Algorithm 4), a replicable penalized quasi-maximum-likelihood estimator, and show that it satisfies an analogous inflated confidence bound. Using it within the same batching scheme, we obtain `RepGLMUCB` (Algorithm 5), which is ρ-replicable with regret $\widetilde{\mathcal{O}}\big(\frac{k_\mu}{c_\mu}\, d^3\sqrt{T}/\rho\big)$. To the best of our knowledge, this is the first replicable algorithm for generalized linear bandits.
4. **Experiments.** On synthetic linear bandit instances, `RepLinUCB` attains $5.8\times$ lower regret than the elimination-based replicable algorithm of Esfandiari et al. (2023a) while meeting the same replicability requirement, and it outperforms it at every replicability level we tested, $\rho \in [0.05, 0.5]$.

Table 1: Regret guarantees for ρ-replicable stochastic bandits. $\widetilde{\mathcal{O}}$ hides polylogarithmic factors.

Work	Setting	Approach	Regret
Esfandiari et al. (2023a)	Linear, K actions	batched elimination	$\widetilde{\mathcal{O}}\big(\frac{K^2}{\rho^2}\sqrt{dT}\big)$
Esfandiari et al. (2023a)	Linear, infinite	net + G-optimal design + elimination	$\widetilde{\mathcal{O}}\big(\frac{d^4}{\rho^2}\sqrt{T}\big)$
Ours (Thm. 4.1)	Linear, infinite	batched optimism + `RepRidge`	$\widetilde{\mathcal{O}}\big(\frac{d^3}{\rho}\sqrt{T}\big)$
Ours (Thm. 5.1)	GLM, infinite	batched optimism + `RepGLM`	$\widetilde{\mathcal{O}}\big(\frac{k_\mu}{c_\mu}\frac{d^3}{\rho}\sqrt{T}\big)$
Komiyama et al. (2026)	lower bound	–	$\Omega\big(\sqrt{T}/\rho\big)$

2 Preliminaries

We consider an online decision-making problem over a horizon of T rounds. At each round $t \in [T]$, the learner interacts with an environment, selects an action a_t, and observes a stochastic reward $r_t \equiv r_t(a_t)$. The precise form of the action space depends on the model; in the stochastic multi-armed bandit setting, the learner chooses one of K arms, while in the linear bandit setting, the learner selects an action from a feasible set $A_t \subset \mathbb{R}^d$. In both cases, the learner's decision at round t is based only on the information available prior to that round. Suppose a_t is the action chosen by the learner at round t, and $a_t^\star$ denotes an oracle action maximizing the expected reward at round t over the feasible action set, i.e., $a_t^\star = \mathrm{argmax}_{a \in X_t} \mathbb{E}[r_t(a)]$, where the expectation is taken over the randomness of the rewards r_t. Then we define the cumulative regret as $\mathrm{Reg}(T) := \sum_{t=1}^T \mathbb{E}[r_t(a_t^\star)] - \mathbb{E}[r_t(a_t)]$.

Multi-Armed Bandits. In the stochastic multi-armed bandit setting, there are K arms, indexed by $a \in [K]$. Each arm a is associated with an unknown mean reward $\mu_a \in \mathbb{R}$. At each round $t \in [T]$, the learner selects an arm $a_t \in [K]$ and observes a reward $r_t(a_t)$ satisfying $r_t(a_t) = \mu_{a_t} + \eta_t$, where η_t is a zero-mean R sub-gaussian noise term. Let $\mu^\star := \max_{a \in [K]} \mu_a$ denote the optimal mean reward, and let $a^\star \in \arg\max_{a \in [K]} \mu_a$ be an optimal arm. The gap of any arm a is defined as $\Delta_a := \mu^\star - \mu_a$. Suppose $N_a(T) := \sum_{t=1}^T \mathbf{1}\{a_t = a\}$ is the number of times arm a is pulled up to round T, then the cumulative regret over T rounds is $\mathrm{Reg}(T) := \sum_{t=1}^T (\mu^\star - \mu_{a_t}) = \sum_{a=1}^K \Delta_a N_a(T)$.

Linear Bandits. In the linear bandit setup, there is an unknown parameter $\theta^\star \in \mathbb{R}^d$. At each round $t \in [T]$, the learner is given a decision set $A_t \subset \mathbb{R}^d$, selects action $a_t \in A_t$, and observes a reward $r_t(a_t)$ satisfying $r_t(a_t) = \langle a_t, \theta^\star \rangle + \eta_t$, where η_t is zero-mean R sub-gaussian noise. Let $a_t^\star \in \arg\max_{a \in A_t} \langle a, \theta^\star \rangle$ be the optimal action at round t. The cumulative regret over T rounds is $\mathrm{Reg}(T) := \sum_{t=1}^T \left(\langle a_t^\star, \theta^\star \rangle - \langle a_t, \theta^\star \rangle \right)$.

Replicability. A bandit algorithm is replicable if, when it is executed twice with the same internal random seed but independent reward realizations, the two runs produce exactly the same sequence of chosen arms with high probability. We define it more formally as follows.

Definition 2.1 (Replicable Bandit). *Let $\rho \in [0, 1]$ and consider two executions of the same algorithm on the same bandit instance over T rounds, using the same realization of the algorithm's internal randomness, but with independent reward realizations from the environment. Let $(a_1^{(1)}, \ldots, a_T^{(1)})$ and $(a_1^{(2)}, \ldots, a_T^{(2)})$ denote the resulting action sequences. We say that the algorithm is ρ-replicable if*

$$\mathbb{P}\left((a_1^{(1)}, \ldots, a_T^{(1)}) = (a_1^{(2)}, \ldots, a_T^{(2)}) \right) \geq 1 - \rho.$$

2.1 Warm-Up: Replicable Batched UCB for Multi-Armed Bandits

Bollini et al. (2026) showed that the UCB rule can be made replicable by combining batching with replicable mean estimation. We recall their construction in our notation as a warm-up for the linear case. The main insight is that recomputing empirical means after every new reward can cause two runs with independent reward realizations to drift apart quickly, even when the same internal randomness is used. The construction therefore replaces empirical means with the replicable estimator RepMean (Impagliazzo et al., 2022), and updates the optimistic index only at the end of each batch.

Proposition 2.1 (Replicable Mean Estimation). *Let $\tau, \delta_1, \rho_1 \in (0, 1)$. There exists a ρ_1-replicable mean estimation algorithm* RepMean *s.t., given samples from a distribution with mean μ, it draws at most $\frac{C_{\mathrm{ME}} \log(1/\delta_1)}{\tau^2(\rho_1 - \delta_1)^2}$ samples and returns $\widehat{\mu}$ satisfying $|\widehat{\mu} - \mu| \leq \tau$ with probability at least $1 - \delta_1$.*

Algorithm 1 Replicable UCB (RepUCB) (Similar to DEBORA in Bollini et al. (2026))

Require: K arms, horizon T, RepMean parameters (δ_1, ρ_1), radius $\tau(n) := \sqrt{\frac{C_{\mathrm{ME}} \log(1/\delta_1)}{n(\rho_1 - \delta_1)^2}}$

Initialize empty sample sets $\mathcal{S}_a \leftarrow \emptyset$ and counts $N_a \leftarrow 0$ for all $a \in [K]$, $t \leftarrow K + 1$
for $a = 1, 2, \ldots, K$ **do**
  Pull arm a once; observe reward $r_1(a)$ and set $\mathcal{S}_a \leftarrow \mathcal{S}_a \cup \{r_1(a)\}$ ▷ Round-robin initialization
  $N_a \leftarrow 1$, $r_a \leftarrow \tau(1)$, $\widehat{\mu}_a \leftarrow \mathrm{RepMean}(\delta_1, \tau(1), \rho_1)$ run on $\mathcal{S}_a$, and $U_a \leftarrow \widehat{\mu}_a + 2r_a$
while $t \leq T$ **do**
  Choose $a_t \in \arg\max_{a \in [K]} U_a$ (break ties by smallest index) ▷ Optimistic arm selection
  $B \leftarrow \min\{N_{a_t}, T - t + 1\}$ ▷ Batch length equals current sample count
  for $j = 1, 2, \ldots, B$ **do**
    Play arm a_t; observe reward $r_t(a_t)$ ▷ Repeat selected arm for B rounds
    $\mathcal{S}_{a_t} \leftarrow \mathcal{S}_{a_t} \cup \{r_t(a_t)\}$
  $N_{a_t} \leftarrow N_{a_t} + B$, $t \leftarrow t + B$ ▷ Update sample count and time index
  $\widehat{\mu}_{a_t} \leftarrow \mathrm{RepMean}(\delta_1, \tau(N_{a_t}), \rho_1)$ run on $\mathcal{S}_{a_t}$ ▷ Recompute only for selected arm
  $r_{a_t} \leftarrow \tau(N_{a_t})$, $U_{a_t} \leftarrow \widehat{\mu}_{a_t} + 2r_{a_t}$ ▷ Updated radius and optimistic estimate after batch

For each arm $a \in [K]$, RepUCB (Algorithm 1) maintains a sample set $\mathcal{S}_a$, a pull count N_a, a replicable mean estimate $\widehat{\mu}_a$, and an optimistic index $U_a = \widehat{\mu}_a + 2\tau(N_a)$. After pulling each arm once, the algorithm repeatedly selects $a_t \in \arg\max_{a \in [K]} U_a$, plays it for a batch of $B = \min\{N_{a_t}, T - t + 1\}$ rounds, so that its pull count doubles, and then recomputes $\widehat{\mu}_{a_t}$ with RepMean and updates U_{a_t}. Since each arm count can double at most $\lceil \log_2 T \rceil$ times, there are at most $M := K(1 + \lceil \log_2 T \rceil)$ calls to RepMean. The replicability budget is split across these calls by setting $\rho_1 := \rho/M$ and $\delta_1 := 1/(2MT)$, and the radius $\tau(n) := \sqrt{C_{\mathrm{ME}} \log(1/\delta_1)/(n(\rho_1 - \delta_1)^2)}$ is the accuracy at which the sample requirement of Proposition 2.1 equals n. By a union bound, all RepMean calls return identical outputs in the two runs with probability at least $1 - \rho$; on this event, the indices, and hence the selected arms and batch lengths, coincide at every batch start, and since the index is frozen within each batch, the two action sequences are identical.

With these choices, RepUCB is ρ-replicable and satisfies $\mathbb{E}[\mathrm{Reg}_{\mathrm{MAB}}(T)] = \mathcal{O}\big(\frac{K^2 \log^2 T}{\rho^2} \sum_{a:\Delta_a > 0} \big(\Delta_a + \frac{\log(KT \log T)}{\Delta_a}\big)\big)$; for completeness, we state and prove this in our notation in Appendix A (Theorem A.1). This matches, up to logarithmic factors, the guarantees of Bollini et al. (2026) and of the elimination-based algorithm of Esfandiari et al. (2023a, Theorem 4), while Komiyama et al. (2026) further separate the $\log T$ term from the K^2/ρ^2 term.

From multi-armed to linear bandits. Two features of this construction carry over to our linear algorithms. First, the number of batches, and hence of replicable estimator calls, is bounded, so the replicability budget can be split across calls by a union bound. Second, the optimistic index is frozen within each batch, so replicability of the entire action sequence reduces to replicability of the estimates computed at batch starts. The remaining ingredients do not carry over when the action set is infinite. Batches can no longer be scheduled by per-arm pull counts; rather than discretizing the action space to recover such counts, as Esfandiari et al. (2023a) do via a $1/T$-net and a G-optimal core set, we trigger batches by the growth of $\det(V_t)$, which yields $\mathcal{O}(d \log T)$ batches regardless of the number of actions. Next, we replace scalar replicable estimates of individual discretized arms in Esfandiari et al. (2023a) by a single vector-level estimate. To do this, we develop RepRidge in Section 3 that rounds the whitened ridge estimate $V_n^{1/2}\widehat{\theta}_n$ once per batch, so the replicability cost is paid once per batch rather than once per core-set action. Finally, whereas the samples of a fixed arm are i.i.d., in a linear bandit the design is chosen adaptively; our analysis in Section 4 handles this by induction over batches, using that the two runs share the same design matrix at a batch start whenever their action sequences agree up to that point.

3 Replicable Ridge Regression

Before introducing our algorithm for linear bandits, we first develop a replicable ridge regression estimator, which serves as the core estimation primitive underlying our approach. In the next section, we invoke this estimator at the start of a batch to obtain parameter estimates that are both statistically accurate and replicable across repeated executions with shared internal randomness. Beyond its role in our linear bandit algorithm, this estimator, along with our guarantees may also be of independent

Algorithm 2 Replicable Ridge Regression via Randomized Grid Rounding (RepRidge)

Require: Regularization $\lambda > 0$, confidence $\delta \in (0,1)$, target replicability $\rho \in (2\delta, 1)$, $\{(x_i, y_i)\}_{i=1}^n$.

1: $V_n \leftarrow \lambda I + \sum_{i=1}^n x_i x_i^\top$, $\widehat{\theta}_n \leftarrow V_n^{-1} \sum_{i=1}^n x_i y_i$ ▷ Ridge Regression Estimate

2: $\beta_n(\delta) \leftarrow \sigma\sqrt{2\log\left(\frac{\det(V_n)^{1/2}}{\det(\lambda I)^{1/2}\,\delta}\right)} + \sqrt{\lambda}\,S$ ▷ Ridge Confidence Radius

3: $z_n \leftarrow V_n^{1/2}\widehat{\theta}_n$ ▷ Whitening

4: Set $\alpha \leftarrow \frac{2\,\beta_n(\delta)\,\sqrt{d}}{\rho - 2\delta}$ and sample $u \sim \mathrm{Unif}([0,\alpha)^d)$ ▷ Grid width chosen to guarantee ρ-replicability

5: Define $Q_{\alpha,u} : \mathbb{R}^d \to \mathbb{R}^d$ coordinatewise by ▷ Randomized Grid Rounding

$$\big(Q_{\alpha,u}(z)\big)_j = \alpha\left\lfloor \frac{z_j - u_j}{\alpha} \right\rfloor + u_j + \frac{\alpha}{2}, \qquad j = 1, \ldots, d$$

6: $\widetilde{\theta}_n \leftarrow V_n^{-1/2}\, Q_{\alpha,u}(z_n)$ ▷ Randomized rounding and map back to parameter space

7: **return** $\widetilde{\theta}_n$

interest in statistical learning where confidence guarantees and replicability must both be satisfied. We consider a fixed-design linear regression problem. Let $d, n \geq 1$, and $\lambda > 0$. Given inputs $x_1, \ldots, x_n \in \mathbb{R}^d$, the response satisfies the following assumption.

Assumption 3.1. *There is an unknown parameter $\theta^\star \in \mathbb{R}^d$ with $\|\theta^\star\|_2 \leq S$. Observations follow the linear model $y_i = x_i^\top \theta^\star + \varepsilon_i, i \in [n]$, where (ε_i) is conditionally σ-sub-Gaussian with respect to the filtration $(\mathcal{F}_i)$ generated by $\{x_1, \ldots, x_i, y_1, \ldots, y_i\}$.*

The objective is to estimate the regression parameter θ^* using input-output samples $(x_i, y_i)_{i=1}^n$ satisfying Assumption 3.1. The ridge estimator and its confidence radius for $\delta \in (0,1)$ are given by

$$\widehat{\theta}_n := V_n^{-1} \sum_{i=1}^n x_i y_i, \quad \beta_n(\delta) := \sigma\sqrt{2\log\left(\frac{\det(V_n)^{1/2}}{\det(\lambda I)^{1/2}\,\delta}\right)} + \sqrt{\lambda}\,S,$$

where $V_n := \lambda I + \sum_{i=1}^n x_i x_i^\top$, and standard concentration results guarantee that $\|\theta_* - \hat{\theta}_n\|_{V_n} \leq \beta_n(\delta)$ with probability at least $1 - \delta$ (see (Abbasi-yadkori et al., 2011)).

To obtain a replicable estimator, we modify the ridge estimate only after first passing to the geometry induced by the confidence ellipsoid. The key observation is that the natural error metric for ridge regression is the V_n-norm, so it is convenient to work in the whitened coordinates $z_n := V_n^{1/2}\widehat{\theta}_n$, where the standard confidence ellipsoid for $\widehat{\theta}_n$ becomes an Euclidean ball. Our construction in RepRidge therefore applies randomized midpoint rounding in the whitened space. Specifically, Algorithm 2 first computes the ridge estimator $\widehat{\theta}_n$ and its confidence radius $\beta_n(\delta)$, then maps $\widehat{\theta}_n$ to the whitened variable $z_n = V_n^{1/2}\widehat{\theta}_n$ in Line 3. It then rounds z_n using the shared randomly shifted grid $Q_{\alpha,u}$ in Line 5, where the grid width α is chosen as a function of $\beta_n(\delta)$, d, and the target replicability level ρ. Finally, the rounded point is mapped back to parameter space in Line 6 to produce the replicable estimated ridge parameter $\widetilde{\theta}_n$.

This construction is designed so that if two executions yield whitened estimates that are sufficiently close, then with high probability they fall into the same shifted grid cell and hence return the same rounded estimator, while the rounding error remains small enough to preserve a useful confidence guarantee. The following theorem establishes these two main properties and a complete proof is given in Appendix B. Self-normalized ridge regression guarantees are widely used beyond bandits, for example, in system identification of linear systems (Simchowitz et al., 2019; Mania et al., 2022), and statistical estimation (Whitehouse et al., 2026). This result states that, in the replicable setting, confidence radius inflates by a d/ρ factor. Hence, we expect this result would serve as a valuable building block in constructing other novel replicable estimation procedures.

Theorem 3.1 (Confidence Ellipsoid and Replicability of RepRidge (Algorithm 2)). *Suppose Assumption 3.1 holds, fix $\delta \in (0,1)$ and $\rho \in (3\delta, 1)$, and let $\widetilde{\theta}_n$ be the output of RepRidge (Algorithm 2) for fixed $\{(x_i, y_i)\}_{i=1}^n$. Then with probability at least $1 - \delta$,*

$$\|\widetilde{\theta}_n - \theta^\star\|_{V_n} \leq \beta_n(\delta)\left(1 + \frac{d}{\rho - 2\delta}\right).$$

Moreover, the estimator is ρ-replicable: if `RepRidge` *(Algorithm 2) is run twice on the same covariate sequence with the same shared random shift u but with independent response noise, producing outputs $\widetilde{\theta}_n^{(1)}$ and $\widetilde{\theta}_n^{(2)}$, then $\mathbb{P}\left(\widetilde{\theta}_n^{(1)} = \widetilde{\theta}_n^{(2)}\right) \geq 1 - \rho$.*

3.1 Comparison with the Replicable Least Squares Estimator

Our approach differs substantially from the Replicable Least Squares Estimator (LSE) (see Lemma 9 of Esfandiari et al. (2023a)). Their construction estimates the value of each action in a finite core set using a reproducible scalar mean-estimation subroutine, and then combines these arm-wise estimates to form a least-squares parameter estimate. The analysis therefore proceeds through a reduction to multiple one-dimensional reproducible estimation problems, followed by a reconstruction step that yields a uniform prediction guarantee over the action set. In contrast, our estimator operates directly on the full ridge regression solution. We first pass to the whitened coordinates $V_n^{1/2}\hat{\theta}_n$, apply randomized grid rounding once at the vector level, and then map the rounded point back to parameter space. Replicability is obtained by showing that two nearby whitened ridge estimates fall in the same shifted grid cell with high probability, while accuracy follows by combining the ridge confidence bound with a deterministic rounding error bound. As a result, our analysis yields a direct V_n-norm confidence ellipsoid, without requiring a core-set.

As an application, under the same core-set coverage assumption (c.f. Assumption 3.2) as in Esfandiari et al. (2023a), our ellipsoidal guarantee can be converted into a uniform prediction guarantee over the full action set (c.f. Proposition 3.1). The resulting estimator attains the same uniform error bound as their Replicable Least Squares Estimator (Esfandiari et al., 2023a, Lemma 9), but with sample complexity better by a factor of d (see Appendix C for proof and a comparison).

Assumption 3.2 (Core-set design). *Let $\mathcal{A} \subseteq \mathbb{R}^d$ be an action set with $\|a\|_2 \leq 1$ for all $a \in \mathcal{A}$. The responses follow $Y(a) = \langle a, \theta^\star \rangle + \eta$, where $\|\theta^\star\|_2 \leq 1$ and the noise variables are independent, mean-zero, and 1-sub-Gaussian. There exists a 4-approximate G-optimal design π supported on a finite core set $C \subseteq \mathcal{A}$ with $|C| = \mathcal{O}(d \log\log d)$ such that $V(\pi) = \sum_{a \in C} \pi(a) a a^\top \succ 0$, $\sup_{a \in \mathcal{A}} \|a\|^2_{V(\pi)^{-1}} \leq 4d$.*

Proposition 3.1. *Suppose Assumption 3.2 holds. Then there exists an algorithm that uses $M = \widetilde{O}\left(\frac{d^3(d+\log(1/\delta))}{\rho^2 \epsilon^2}\right)$ total samples and outputs a ρ replicable estimator $\widetilde{\theta} \in \mathbb{R}^d$ satisfying $\sup_{a \in \mathcal{A}} |\langle a, \widetilde{\theta} - \theta^\star \rangle| \leq \epsilon$, with probability at least $1 - \delta$.*

4 Replicable Linear Bandit

In the linear bandit setting, our objective is to design an optimistic algorithm whose action sequence remains stable across repeated executions, while retaining the statistical efficiency of standard linear UCB methods. We assume that the rewards r_t satisfy the following assumption.

Assumption 4.1. *There is an unknown parameter $\theta^\star \in \mathbb{R}^d$ with $\|\theta^\star\|_2 \leq S$. At each round $t \in [T]$, after the learner selects an action $a_t \in A_t \subset \mathbb{R}^d$, it observes a reward $r_t = a_t^\top \theta^\star + \varepsilon_t$, where (ε_t) is conditionally σ-sub-Gaussian with respect to the filtration $(\mathcal{F}_t)$ generated by $\{A_1, a_1, r_1, \ldots, A_t, a_t, r_t\}$, i.e. for every t and $\gamma \in \mathbb{R}$, $\mathbb{E}[\exp(\gamma \varepsilon_t) \mid \mathcal{F}_{t-1}, A_t, a_t] \leq \exp\left(\frac{\sigma^2 \gamma^2}{2}\right)$.*

The main difficulty of establishing replicability of a LinUCB algorithm (Abbasi-yadkori et al., 2011) is that recomputing the ridge estimate after every observation makes the action highly sensitive to the realized reward, causing two runs with independent rewards to diverge quickly even when they use the same internal algorithmic randomness. To address this, we combine determinant-triggered batching (Gu et al., 2024) with the replicable ridge estimator `RepRidge`. Specifically, the policy is updated only at batch starts determined by a determinant-growth rule, and at each such update we replace the usual ridge estimate by the replicable estimate returned by `RepRidge`. The resulting procedure is therefore a batched optimistic linear bandit algorithm that uses a single replicable confidence estimate at the start of each batch and reuses it throughout the batch.

More formally, Algorithm 3 implements the preceding idea by updating the policy only at determinant-triggered batch starts. Whenever a new batch begins, the algorithm invokes the replicable ridge estimator `RepRidge` in Line 5 to obtain the batch-start parameter estimate $\widetilde{\theta}_b$, computes the ridge confidence radius β_b in Line 6, and forms the corresponding inflated radius $\widetilde{\beta}_b$. Throughout the batch, the same pair $(\widetilde{\theta}_b, \widetilde{\beta}_b)$ is reused, and actions are selected by maximizing the batch-start UCB score

Algorithm 3 Replicable Batched Linear UCB (RepLinUCB)

Require: Regularization $\lambda > 0$, horizon T, batch budget $B \in \mathbb{N}$, determinant growth factor $q > 1$, Per-batch budgets $(\delta_b, \rho_b)_{b=0}^{B-1}$ with $\delta_b \in (0,1)$, $\rho_b \in (0,1)$, and $\rho_b > 2\delta_b$

1: Initialize $V_1 \leftarrow \lambda I_d$, $b_1 \leftarrow 0$, $b \leftarrow 0$ and batch-start time $t_b \leftarrow 1$
2: **for** $t = 1$ to T **do**
3: Observe compact action set $A_t \subset \mathbb{R}^d$
4: **if** $t = t_b$ **then**
5: $\widetilde{\theta}_b \leftarrow \texttt{RepRidge}\left(\lambda, \delta_b, \rho_b; \{(a_s, r_s)\}_{s=1}^{t_b-1}\right)$ ▷ Replicable batch parameter
6: $\beta_b \leftarrow \beta_{t_b}(\delta_b)$ ▷ Ridge confidence radius at batch start
7: $\widetilde{\beta}_b \leftarrow \beta_b\left(1 + \frac{d}{\rho_b - 2\delta_b}\right)$ ▷ Inflated batch confidence radius
8: Choose $a_t \in \arg\max_{a \in A_t}\left\{\langle a, \widetilde{\theta}_b\rangle + \widetilde{\beta}_b \|a\|_{V_{t_b}^{-1}}\right\}$ (break ties with fixed rule) ▷ UCB rule
9: Observe r_t and update $V_{t+1} \leftarrow V_t + a_t a_t^\top$ and $b_{t+1} \leftarrow b_t + a_t r_t$ ▷ Update statistics
10: **if** $b \le B - 2$ **and** $\det(V_{t+1}) > q \det(V_{t_b})$ **then**
11: Start a new batch: set $b \leftarrow b + 1$ and $t_b \leftarrow t + 1$ ▷ Determinant trigger

over the current action set in Line 8. After observing the reward, the sufficient statistics are updated in Line 9, but the policy changes only when the determinant trigger in Line 11 fires, at which point the next batch begins. The next theorem gives the main guarantee for RepLinUCB (Algorithm 3). It shows that combining determinant-triggered batching with the replicable ridge estimator RepRidge yields a linear bandit algorithm that is both ρ-replicable and achieves a sub-linear regret.

Theorem 4.1 (Replicability and Regret Bound for RepLinUCB (Algorithm 3)). *Suppose Assumption 4.1 holds and the action sets $(A_t)_{t=1}^T$ are chosen by an oblivious adversary. Further suppose RepLinUCB (Algorithm 3) is run with $B = \lceil d \log\left(1 + \frac{TL^2}{\lambda d}\right)\rceil$, $q = \left(1 + \frac{TL^2}{\lambda d}\right)^{\frac{d}{B}}$, $\delta_b = \frac{\delta}{B}$, $\rho_b = \frac{\rho}{B}$, for all $b \in \{0, \dots, B-1\}$. Then for $\rho \in (3\delta, 1)$, RepLinUCB (Algorithm 3) is ρ-replicable, i.e., if the algorithm is run twice on action sets $(A_t)_{t=1}^T$ with shared internal randomness and i.i.d. reward sequences, producing action sequences $(a_t^{(1)})_{t=1}^T$ and $(a_t^{(2)})_{t=1}^T$, then $\mathbb{P}\left(\forall t \in [T] : a_t^{(1)} = a_t^{(2)}\right) \ge 1 - \rho$. Moreover, there exists a constant $C > 0$ such that, with probability at least $1 - \delta$, the regret of satisfies*

$$\mathrm{Reg}(T) \le C\left(\sigma\sqrt{d\log\left(1 + \tfrac{TL^2}{\lambda d}\right) + \log\left(\tfrac{B}{\delta}\right)} + \sqrt{\lambda}\, S\right)\left(1 + \frac{dB}{\rho}\right)\sqrt{T\, d \log\left(1 + \tfrac{TL^2}{\lambda d}\right)}.$$

Remark 4.1 (Regret Improvement). With the choice $B = \lceil d\log\left(1 + \frac{TL^2}{\lambda d}\right)\rceil$, and σ, λ, and S treated as constants, Theorem 4.1 yields the final regret bound $\mathrm{Reg}(T) = \widetilde{\mathcal{O}}(\frac{d^3}{\rho}\sqrt{T})$. This improves on the existing regret bound of $\widetilde{\mathcal{O}}(d^4\sqrt{T}/\rho^2)$ (Esfandiari et al., 2023a, Theorem 10), by a factor of $\mathcal{O}\left(\frac{d}{\rho}\right)$. Existing lower bounds imply a regret of $\Omega(\sqrt{T}/\rho)$ (Ahmadi et al., 2025; Komiyama et al., 2026), and hence our bound is optimal in ρ, in addition to T. To the best of our knowledge, this is the first linear-bandit algorithm with an *optimal* dependence on ρ for large number of arms. Although we have improved the d dependence from d^4 to d^3, it is not known what the optimal dependence on d is for replicable linear bandits. In fact even for the K-armed case, the optimal dependence on K is not known and we pose this as an open problem in Section 7.

Remark 4.2 (Comparison with Prior Work). Our approach differs substantially from the infinite-action algorithm of Esfandiari et al. (2023a, Theorem 10). Their method constructs a deterministic $1/T$-net of the action space, computes a G-optimal design and its associated core set, and then proceeds by batched elimination using replicable LSE subroutine on the core-set arms. In contrast, RepLinUCB is a direct optimistic algorithm: it does not discretize the action space, does not rely on elimination, and does not require a core-set construction. Instead, at each batch start it computes a replicable ridge estimate via RepRidge and, forms an inflated confidence radius. Then it selects actions throughout the batch by maximizing a batch-start UCB score over the current action set.

Remark 4.3 (Analysis). The proof of Theorem 4.1 is more delicate than that of Theorem 3.1, since RepRidge is proved under a fixed design, whereas in the linear bandit setting the actions are chosen adaptively from the observed history. We show that replicability can nevertheless be established inductively over batches: if the action sequence is identical up to the start of a batch, then the design is shared across runs, so conditioning on this common design allows us to apply the guarantee of

RepRidge to obtain a replicable batch-start parameter estimate. We then show that, because the same batch-start estimate and confidence radius are used throughout the batch, all actions selected within that batch are also replicable across runs. Repeating this argument across batches yields the overall ρ-replicability guarantee. A rigorous proof is provided in Appendix D.

Remark 4.4 (Action Set). Comparing Esfandiari et al. (2023a, Theorem 10) and our result, note that their guarantee assumes a single fixed action set $\mathcal{A}$, whereas in our setting the action set A_t may vary with t, provided the sequence $(A_t)_{t=1}^T$ is chosen by an oblivious adversary, and is therefore more general. However, in the non-replicable linear bandit analysis of Abbasi-yadkori et al. (2011), a $\mathcal{O}(\sqrt{T})$ regret guarantee holds even against an adaptive adversary. The oblivious-adversary assumption is needed in our analysis for the replicability argument: once we show that two runs agree up to the start of a batch, they must face the same action sets within that batch so that the UCB maximization in Line 8 produces the same actions across runs. Extending our results to the adaptive adversarial setting would require novel techniques and is left for future work.

5 Replicable Generalized Linear Bandits

We now extend our framework to the generalized linear bandit setting of Filippi et al. (2010), where rewards depend on actions through a known nonlinear inverse link function.

Assumption 5.1 (Generalized linear bandit). *There is an unknown $\theta^\star \in \Theta \subset \mathbb{R}^d$ with Θ a known closed convex set and $\|\theta^\star\|_2 \le S$. At each round $t \in [T]$, the learner observes an action set $A_t \subset \mathbb{R}^d$ with $\|a\|_2 \le L$ for all $a \in A_t$, selects $a_t \in A_t$, and observes $r_t = \mu(a_t^\top \theta^\star) + \varepsilon_t$, where (ε_t) is conditionally σ-sub-Gaussian. The inverse link μ is continuously differentiable and k_μ-Lipschitz, with $c_\mu := \inf_{\theta\in\Theta, t, a\in A_t} \mu'(a^\top\theta) > 0$, and mean rewards lie in $[0,1]$.*

The main difficulty in lifting our linear construction to this setting is that the penalized GLM estimator is defined implicitly through a strictly concave optimization problem rather than a closed-form ridge expression. Recomputing it after every new reward would again make the chosen action highly sensitive to the realized noise, so two runs with independent rewards would quickly diverge. We address this through the same approach used in the linear case: *(i)* a fixed-design replicable GLM estimator, and *(ii)* a determinant-triggered batching scheme that freezes the policy between updates.

Replicable GLM estimation. Given covariates $\{(x_i, y_i)\}_{i=1}^n$, our primitive RepGLM (Algorithm 4) first computes the penalized quasi-MLE $\widehat{\theta}_n = \arg\max_{\theta\in\Theta}\{\sum_{i=1}^n (y_i x_i^\top\theta - b(x_i^\top\theta)) - \frac{\lambda}{2}\|\theta\|_2^2\}$ where $b' = \mu$, then whitens via $z_n = V_n^{1/2}\widehat{\theta}_n$ with $V_n = \lambda I + \sum_i x_i x_i^\top$, applies randomized midpoint rounding on a shared shifted grid of width $\alpha = 2\beta_n(\delta)\sqrt{d}/(\rho - 2\delta)$, and maps back to parameter space. The whitening step is what enables this: the natural statistical metric for the penalized GLM estimator is the V_n-norm, and after whitening, V_n-closeness becomes Euclidean closeness, so the same shifted-grid argument used in the linear case applies. Theorem E.1 shows that RepGLM is ρ-replicable and satisfies $\|\widetilde{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta)(1 + d/(\rho - 2\delta))$ with probability at least $1-\delta$, where $\beta_n(\delta)$ is the standard GLM confidence radius inflated by $1/\underline{c}_\mu$.

Replicable batched GLM-UCB. Our bandit algorithm RepGLMUCB (Algorithm 5) follows the optimistic principle of Filippi et al. (2010), but constructs its upper confidence bound only at the start of each batch using the replicable estimator above. Within a batch, it reuses the same pair $(\widetilde{\theta}_b, \widetilde{\beta}_b)$ to select $a_t \in \arg\max_{a\in A_t}\left\{\mu(\langle a, \widetilde{\theta}_b\rangle) + \widetilde{\beta}_b\|a\|_{V_{t_b}^{-1}}\right\}$, and a new batch is triggered whenever the determinant of the design matrix has grown by a factor of q. Freezing the policy between determinant triggers ensures that two parallel runs make identical decisions throughout a batch whenever their batch-start estimators coincide, and the determinant rule guarantees that there are at most $B = O(d\log(TL^2/(\lambda d)))$ batches. This yields our main guarantee for the GLM setting.

Theorem 5.1 (Informal; see Theorem E.5). *With the parameter choices $B = \lceil d\log(1 + TL^2/(\lambda d))\rceil$, $q = (1 + TL^2/(\lambda d))^{d/B}$, and $\delta_b = \delta/B$, $\rho_b = \rho/B$ for $\rho \in (3\delta, 1)$, RepGLMUCB is ρ-replicable, and with probability at least $1-\delta$ its regret satisfies* $\mathrm{Reg}_{\mathrm{GLB}}(T) = \widetilde{O}\left(\frac{k_\mu}{\underline{c}_\mu}\left(\sigma\sqrt{d} + \sqrt{\lambda}\,S\right)\left(1 + \frac{dB}{\rho}\right)\sqrt{Td}\right)$.

The proof combines a batch-start concentration argument for the replicable GLM estimator with a stale-norm-to-current-norm comparison enabled by the determinant trigger, and a union-bound replicability argument over batches. For a complete description, see Appendix E.

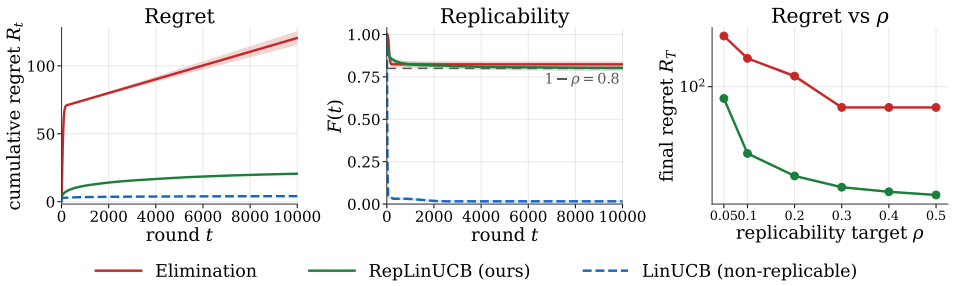


Figure 1: **Left:** cumulative regret. **Centre:** replicability $F(t)$, with the requirement $1-\rho=0.8$ marked; LinUCB falls below it. **Right:** final regret as a function of ρ: `RepLinUCB` beats Elimination across all choices. Shaded regions are 95% intervals over 30 held-out instance seeds.

6 Experiments

We compare `RepLinUCB` against the elimination-based replicable linear bandit algorithm of Esfandiari et al. (2023a) ("Elimination"), with non-replicable LinUCB as a reference.

Setup. Each instance draws $K=100$ actions with i.i.d. standard normal entries, normalised to the unit sphere in $\mathbb{R}^d$ with $d=5$, and a unit-norm $\theta^\star$. Rewards are $r_t=\langle a_t,\theta^\star\rangle+\eta_t$ with η_t sub-Gaussian, $\sigma=10^{-3}$, and the horizon is $T=10^4$. We report pseudo-regret R_T and the empirical replicability $F(t)=\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\tau_i>t]$, where τ_i is the first round at which the two runs of pair i play different actions; two runs share internal randomness and draw independent reward noise. An algorithm meets the requirement when $F(T)\geq 1-\rho$. We use $N=1800$ pairs (30 instances, 60 pairs each) and set $\rho=0.2$ unless stated otherwise.

Hyperparameters. All three algorithms are tuned independently on 12 instance seeds and evaluated on 30 disjoint seeds. For RepLinUCB we sweep the ridge parameter, the rounding-grid width, the batch count and the confidence-radius scale (168 configurations); for Elimination the elimination width, the rounding width of the arm-wise estimate and the per-batch sample budget (150 configurations); for LinUCB the ridge parameter, batch count and radius scale (24 configurations). For the two replicable algorithms we select the configuration minimising mean regret subject to $F(T)\geq 1-\rho$; LinUCB is selected on regret alone. Selection is repeated at each ρ.

Results. Figure 1 reports the outcome. `RepLinUCB` attains $R_T=20.6\pm0.2$ against 120.5 ± 2.6 for Elimination, a $5.8\times$ improvement, while both satisfy the replicability requirement ($F(T)=0.80$ and 0.82). LinUCB attains lower regret (4.1 ± 0.03) but is non-replicable ($F(T)=0.02$). The regret of RepLinUCB is sublinear throughout, whereas Elimination incurs most of its regret in an initial exploration phase and still grows linearly. The ordering is also stable across $\rho\in[0.05,0.5]$.

7 Conclusion

We studied replicable algorithms for stochastic linear and generalized linear bandits. In the linear setting, we introduced `RepRidge`, a replicable estimator for ridge regression, and used it as the core estimation primitive in `RepLinUCB`. This yields a replicable optimistic algorithm for stochastic linear bandits whose regret improves the previous best known bound by a factor of d/ρ. We extend our approach to generalized linear bandits via a replicable penalized GLM estimator, `RepGLM`, and a corresponding optimistic algorithm, `RepGLMUCB`. Several directions remain open. In stochastic multi-armed bandits, the best known lower bound scales as $\sqrt{T}/\rho$, leaving a gap from the upper bound; an important next step is therefore to close this gap, either by developing replicable algorithms with improved regret guarantees or by strengthening lower bounds. In stochastic linear bandits, it remains open whether one can further improve the dependence on the dimension d, or establish matching lower bounds that clarify the optimal price of replicability in this setting. Another natural direction is to extend replicable optimistic methods beyond linear models to more general non-linear contextual and bandit settings (Deb et al., 2024a; Foster & Rakhlin, 2020; Foster & Krishnamurthy, 2021). Recently Bollini et al. (2026) started the study of constrained MABs under replicability, and it would be interesting to understand whether replicable algorithms can be extended to other linear and non-linear constrained bandit settings as well (Agrawal & Devanur, 2016; Kazerouni et al., 2017; Moradipari et al., 2021; Deb et al., 2025a;b; 2024b). Finally, our replicable ridge estimator may be useful beyond bandits in other adaptive statistical and online decision-making problems.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, support qualitative and thematic data analysis, interpret results, and propose or refine hypotheses. We did not use generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, or clean and reformat data sets. Assistance with translation is not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures and to polish the writing of the paper. We have reviewed all AI-assisted work: we checked LLM-generated research ideas for potential plagiarism through a manual literature survey, and LLM-generated code was verified and tested for correctness by one author. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided our code to generate our results as a zip file with all the instructions. The hyper-parameters are provided in the code.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/f3144cefe89a60d6a1afaf7859c5076b-Paper.pdf.

Saba Ahmadi, Siddharth Bhandari, and Avrim Blum. Replicable online learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=TGuWVMssSa.

Matteo Bollini, Gianmarco Genalti, Francesco Emanuele Stradi, Matteo Castiglioni, and Alberto Marchesi. Replicable constrained bandits, 2026. URL https://arxiv.org/abs/2602.14580.

Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, pp. 520–527, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399135. doi: 10.1145/3564246.3585246. URL https://doi.org/10.1145/3564246.3585246.

Michael Chen, A. Pavan, N. V. Vinodchandran, Ruosong Wang, and Lin Yang. Regret-optimal list replicable bandit learning: Matching upper and lower bounds. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0T49QbSOho.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic Linear Optimization Under Bandit Feedback. In *Annual Conference on Computational Learning Theory*, pp. 355–366, 2008. URL https://mlanthology.org/colt/2008/dani2008colt-stochastic/.

Rohan Deb, Yikun Ban, Shiliang Zuo, Jingrui He, and Arindam Banerjee. Contextual bandits with online neural regression. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=5ep85sakT3.

Rohan Deb, Aadirupa Saha, and Arindam Banerjee. Think before you duel: Understanding complexities of preference learning under constrained resources. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4546–4554. PMLR, 02–04 May 2024b. URL https://proceedings.mlr.press/v238/deb24a.html.

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Conservative contextual bandits: Beyond linear representations. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=SThJXvucjQ.

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Thompson sampling for constrained bandits. *Reinforcement Learning Journal*, 6:2820–2843, 2025b.

Hossein Esfandiari, Alkis Kalavasis, Amin Karbasi, Andreas Krause, Vahab Mirrokni, and Grigoris Velegkas. Replicable bandits. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=gcD2UtCGMc2.

Hossein Esfandiari, Amin Karbasi, Vahab Mirrokni, Grigoris Velegkas, and Felix Zhou. Replicable clustering. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 39277–39320. Curran Associates, Inc., 2023b. doi: 10.52202/075280-1707. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7bc3fe234454107149fa9d44faacaa64-Paper-Conference.pdf.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18907–18919. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9d684c589d67031a627ad33d59db65e5-Paper.pdf.

Quanquan Gu, Amin Karbasi, Khashayar Khosravi, Vahab Mirrokni, and Dongruo Zhou. Batched neural bandits. *ACM / IMS J. Data Sci.*, 1(1), January 2024. doi: 10.1145/3592474. URL https://doi.org/10.1145/3592474.

Rupkatha Hira, Dominik Kau, and Jessica Sorrell. The cost of replicability in active learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=ZsqJu9eITd.

Max Hopkins, Russell Impagliazzo, Daniel Kane, Sihan Liu, and Christopher Ye. Replicability in high dimensional statistics. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1–8, 2024. doi: 10.1109/FOCS61266.2024.00103.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, pp. 818–831, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392648. doi: 10.1145/3519935.3519973. URL https://doi.org/10.1145/3519935.3519973.

Alkis Kalavasis, Amin Karbasi, Grigoris Velegkas, and Felix Zhou. On the computational landscape of replicable learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105887–105927. Curran Associates, Inc., 2024. doi: 10.52202/079017-3359. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bf331c87e29f473b610336f00fe1cb51-Paper-Conference.pdf.

Amin Karbasi, Grigoris Velegkas, Lin Yang, and Felix Zhou. Replicability in reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74702–74735. Curran Associates, Inc., 2023a. doi: 10.52202/075280-3265. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ec4d2e436794d1bf55ca83f5ebb31887-Paper-Conference.pdf.

Amin Karbasi, Grigoris Velegkas, Lin Yang, and Felix Zhou. Replicability in reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74702–74735. Curran Associates, Inc., 2023b. doi: 10.52202/075280-3265. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ec4d2e436794d1bf55ca83f5ebb31887-Paper-Conference.pdf.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/bdc4626aa1d1df8e14d80d345b2a442d-Paper.pdf.

Junpei Komiyama, Shinji Ito, Yuichi Yoshida, and Souta Koshino. Replicability is asymptotically free in multi-armed bandits. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=E8rmbq8BYP. J2C Certification.

Horia Mania, Michael I. Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *Journal of Machine Learning Research*, 23(32):1–30, 2022. URL http://jmlr.org/papers/v23/20-807.html.

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69: 3755–3767, 2021. doi: 10.1109/TSP.2021.3089822.

National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC, 2019. doi: 10.17226/25303.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2714–2802. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/simchowitz19a.html.

Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. Time-uniform self-normalized concentration for vector-valued processes. *The Annals of Applied Probability*, 36(3):1972 – 2013, 2026. doi: 10.1214/25-AAP2246. URL https://doi.org/10.1214/25-AAP2246.

Kelly W. Zhang, Nowell Closser, Anna L. Trella, and Susan A. Murphy. Replicable Bandits for Digital Health Interventions. *Statistical Science*, 40(4):671 – 691, 2025. doi: 10.1214/25-STS1017. URL https://doi.org/10.1214/25-STS1017.

A Replicable Multi Armed Bandit Proof

Theorem A.1 (Replicability and Regret of RepUCB (Algorithm 1), cf. Bollini et al. (2026)). *Consider RepUCB (Algorithm 1) and let* $M := K\big(1 + \lceil \log_2 T \rceil\big)$, $\rho_1 := \frac{\rho}{M}$, *and* $\delta_1 := \frac{1}{2MT}$, *and assume* $\delta_1 \le \rho_1/2$. *Define* $\tau(n) := \sqrt{\frac{C_{\mathrm{ME}} \log(1/\delta_1)}{n(\rho_1 - \delta_1)^2}}$ *for all* $n \ge 1$. *Then RepUCB (Algorithm 1) is* ρ*-replicable, and there exists a constant* $C > 0$ *such that the regret is bounded as follows*

$$\mathbb{E}[\mathrm{Reg}_{\mathrm{MAB}}(T)] \le C \cdot \frac{K^2(1 + \lceil \log_2 T \rceil)^2}{\rho^2} \sum_{a:\Delta_a > 0} \left(\Delta_a + \frac{\log\big(2KT(1 + \lceil \log_2 T \rceil)\big)}{\Delta_a} \right) + \frac{1}{2}.$$

A.1 Proof of Theorem A.1

Recall that in this setup, arms $a \in [K]$ produce i.i.d. rewards in $[0,1]$ with mean μ_a, $\mu^\star := \max_{a \in [K]} \mu_a$, and the gap $\Delta_a := \mu^\star - \mu_a$. For horizon T, regret in the MAB case is defined as

$$\mathrm{Reg}_{\mathrm{MAB}}(T) := \sum_{t=1}^{T} (\mu^\star - \mu_{a_t}) = \sum_{a=1}^{K} \Delta_a \, N_a(T),$$

where $N_a(T)$ is the number of pulls of arm a up to time T. The replicability proof of RepUCB (Algorithm 1) critically uses the replicable mean estimation procedure from Impagliazzo et al. (2022) and we state it as an oracle as follows (shorter version in Proposition 2.1).

RepMean **(Replicable Mean Estimation).** There exists a universal constant $C_{\mathrm{ME}} > 0$ such that for any $\delta_1 \in (0,1)$, $\rho_1 \in (0,1)$, and $\tau > 0$ with $\delta_1 < \rho_1$, there is a mean estimator $\mathrm{RepMean}(\delta_1, \tau, \rho_1)$ satisfying:

1. **Accuracy.** If it is run on n i.i.d. samples from a $[0,1]$-valued distribution with mean μ and
$$n \ge \frac{C_{\mathrm{ME}} \log(1/\delta_1)}{\tau^2(\rho_1 - \delta_1)^2},$$
then its output $\widehat{\mu}$ satisfies $\mathbb{P}(|\widehat{\mu} - \mu| \le \tau) \ge 1 - \delta_1$.
2. **Replicability.** If it is run twice with the same internal randomness on two independent size-n sample sets from the same distribution, the two outputs are identical with probability at least $1 - \rho_1$.

Lemma A.2 (Replicability of RepUCB). *Let* $M := K\big(1 + \lceil \log_2 T \rceil\big)$ *and suppose Algorithm 1 is run with* $\rho_1 := \frac{\rho}{M}, \delta_1 := \frac{1}{2MT}$, *where* $\delta_1 < \rho_1$. *Then RepUCB is* ρ*-replicable. More precisely, let* $\{a_t^{(1)}\}_{t=1}^T$ *and* $\{a_t^{(2)}\}_{t=1}^T$ *denote the action sequences produced by two executions of Algorithm 1 using the same internal randomness and two independent reward realizations from the same bandit instance. Then*

$$\mathbb{P}\Big(\forall t \in [T],\ a_t^{(1)} = a_t^{(2)}\Big) \ge 1 - \rho.$$

Proof of Lemma A.2. Consider two executions of Algorithm 1 on the same bandit instance, using the same internal randomness and two independent reward realizations. Let $\{a_t^{(1)}\}_{t=1}^T$ and $\{a_t^{(2)}\}_{t=1}^T$ denote the corresponding action sequences.

Step 1. Bound the total number of RepMean calls. During initialization, the algorithm pulls each arm once and computes one RepMean estimate for each arm. Hence initialization makes exactly K calls to RepMean.

After initialization, RepMean is called only in Line 12, that is, only after a batch is completed, and only for the arm selected for that batch. Fix an arm $a \in [K]$. Initially, $N_a = 1$. Whenever arm a is selected in the main loop, the batch length is

$$B = \min\{N_a,\ T - t + 1\}.$$

If the horizon truncation is inactive, then $B = N_a$, and after the batch the count is updated to

$$N_a \leftarrow N_a + B = 2N_a.$$

Therefore, apart from a possible final truncated batch, each selection of arm a doubles its count. Starting from $N_a = 1$, this doubling can occur at most $\lceil \log_2 T \rceil$ times before the count exceeds T. Thus arm a can trigger at most $\lceil \log_2 T \rceil$ post-initialization calls to RepMean.

Summing over all K arms, the total number of post-initialization calls is at most $K\lceil \log_2 T \rceil$. Hence the total number of RepMean calls made by the algorithm is at most

$$M := K + K\lceil \log_2 T \rceil = K\big(1 + \lceil \log_2 T \rceil\big).$$

Step 2. Union bound over per-call RepMean replicability. Let $\mathcal{G}$ denote the event that every RepMean call made in the two executions returns the same output in both runs. Since the total number of calls is at most M, and since each call is ρ_1-replicable by the replicability guarantee in Proposition 2.1, the probability that a given call produces different outputs in the two executions is at most ρ_1.

Therefore, by the union bound, $\mathbb{P}(\mathcal{G}^c) \le M\rho_1$. With the choice $\rho_1 = \rho/M$, it follows that

$$\mathbb{P}(\mathcal{G}) \ge 1 - M\rho_1 = 1 - \rho.$$

Step 3: on $\mathcal{G}$, the action sequences are identical. We prove that on the event $\mathcal{G}$, $a_t^{(1)} = a_t^{(2)}$ for all $t \in [T]$. First, during initialization, both executions pull arms $1, 2, \ldots, K$ in the same deterministic order. Hence

$$a_t^{(1)} = a_t^{(2)} \qquad \text{for all } t = 1, \ldots, K.$$

Moreover, for each arm $a \in [K]$, the initialization call to RepMean agrees across the two executions on $\mathcal{G}$. Therefore the initial estimates $\widehat{\mu}_a$, radii r_a, and indices U_a are identical in the two runs after initialization.

Now suppose inductively that the two executions agree up to some decision time t, in the sense that they have produced the same actions in all previous rounds. Then for each arm $a \in [K]$, the number of times arm a has been played is the same in both executions, so the counts N_a are identical across the two runs. Since $r_a = \tau(N_a)$ depends only on N_a, the radii are also identical.

Further, the only times an estimate $\widehat{\mu}_a$ can change are the initialization call and the update call in Line 12 after a batch of arm a. Every such RepMean call agrees across the two executions on $\mathcal{G}$. Hence all estimates $\widehat{\mu}_a$ are identical across the two runs, and therefore all optimistic indices $U_a = \widehat{\mu}_a + 2r_a$ are identical as well.

Consequently, at decision time t, both executions solve the same maximization problem

$$a_t \in \arg\max_{a \in [K]} U_a,$$

with the same deterministic tie-breaking rule. Thus both executions choose the same arm, say $a^\star$. Since the counts match, both executions also compute the same batch length

$$B = \min\{N_{a^\star},\, T - t + 1\}.$$

Therefore both executions play arm $a^\star$ for exactly the next B rounds, so the action sequences continue to agree throughout the batch. After the batch, both executions update the same arm $a^\star$, with the same new count $N_{a^\star}$, and invoke the corresponding RepMean call on that arm. On $\mathcal{G}$, this call also agrees across the two runs, so the updated estimate, radius, and optimistic index for arm $a^\star$ remain identical across the two executions.

This closes the induction. Hence, on $\mathcal{G}$, $\forall t \in [T], a_t^{(1)} = a_t^{(2)}$. Combining the previous steps, we obtain

$$\mathbb{P}\Big(\forall t \in [T],\ a_t^{(1)} = a_t^{(2)}\Big) \ge \mathbb{P}(\mathcal{G}) \ge 1 - \rho.$$

Therefore Algorithm 1 is ρ-replicable. □

Lemma A.3 (Regret guarantee of RepUCB)**.** *Let $L = \lceil \log_2 T \rceil$, $M = K(1+L)$, $\rho_1 = \frac{\rho}{M}$, and $\delta_1 = \frac{1}{2MT}$. Assume $\delta_1 < \rho_1$ and $\delta_1 \le \rho_1/2$. Then there exists a universal constant $C > 0$ such that*

$$\mathbb{E}[\mathrm{Reg}(T)] \le C \cdot \frac{K^2(1+L)^2}{\rho^2} \sum_{a:\Delta_a > 0} \left(\Delta_a + \frac{\log(2KT(1+L))}{\Delta_a}\right) + \frac{1}{2}.$$

Proof. We prove the regret bound in four steps.

Step 1: global accuracy event. Let $\mathcal{E}$ be the event that every call to RepMean made by Algorithm 1 is accurate, namely, whenever RepMean is called on arm a with current sample size $n = N_a$, the returned estimate $\widehat{\mu}_a$ satisfies

$$|\widehat{\mu}_a - \mu_a| \le \tau(n).$$

By Step 1 in the proof of Lemma A.2, the total number of calls to RepMean is at most $M = K(1+L)$. For any such call, the definition

$$\tau(n) = \sqrt{\frac{C_{\mathrm{ME}} \log(1/\delta_1)}{n(\rho_1 - \delta_1)^2}}$$

implies

$$n = \frac{C_{\mathrm{ME}} \log(1/\delta_1)}{\tau(n)^2(\rho_1 - \delta_1)^2}.$$

Hence the sample-size requirement in the accuracy guarantee of RepMean is satisfied with equality. Therefore each call fails with probability at most δ_1. By a union bound,

$$\mathbb{P}(\mathcal{E}^c) \le M\delta_1 = M \cdot \frac{1}{2MT} = \frac{1}{2T}.$$

Step 2: selecting a suboptimal arm forces its radius to be large. Consider any decision time in the main loop where the algorithm selects a suboptimal arm a with $\Delta_a > 0$. Let $n := N_a$ denote the count of arm a at that moment. The index used for arm a is

$$U_a = \widehat{\mu}_a + 2\tau(n),$$

where $\widehat{\mu}_a$ was computed the last time arm a was updated, which is exactly when it had n samples. On the event $\mathcal{E}$,

$$\widehat{\mu}_a \le \mu_a + \tau(n),$$

and thus

$$U_a \le \mu_a + 3\tau(n).$$

Let $a^\star$ be any optimal arm, so $\mu_{a^\star} = \mu^\star$. Its current index is

$$U_{a^\star} = \widehat{\mu}_{a^\star} + 2\tau(N_{a^\star}).$$

Again on $\mathcal{E}$,

$$\widehat{\mu}_{a^\star} \ge \mu^\star - \tau(N_{a^\star}),$$

which gives

$$U_{a^\star} \ge \mu^\star - \tau(N_{a^\star}) + 2\tau(N_{a^\star}) = \mu^\star + \tau(N_{a^\star}) \ge \mu^\star.$$

Since arm a is chosen by the argmax rule, we have $U_a \ge U_{a^\star}$. Therefore $\mu_a + 3\tau(n) \ge \mu^\star$, and hence

$$3\tau(n) \ge \mu^\star - \mu_a = \Delta_a.$$

Substituting the definition of $\tau(n)$ and rearranging yields

$$n \le \frac{9C_{\mathrm{ME}} \log(1/\delta_1)}{(\rho_1 - \delta_1)^2 \Delta_a^2} =: H_a.$$

Step 3: doubling batches bound total pulls of a suboptimal arm. Fix a suboptimal arm a. If it is never selected after initialization, then $N_a(T) = 1$. Otherwise, consider the last time it is selected, and let n be its count immediately before that selection. Since arm a is selected at that time, Step 2 implies that on $\mathcal{E}$, $n \le H_a$.

The batch length is $B = \min\{n, \text{remaining rounds}\} \le n$. Therefore, after that final batch,

$$N_a(T) = n + B \le 2n \le 2H_a.$$

Since initialization already contributes one pull, it is valid to write the slightly looser bound

$$N_a(T) \le 1 + 2H_a.$$

Thus on $\mathcal{E}$,

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{a:\Delta_a>0} \Delta_a N_a(T) \\
&\le \sum_{a:\Delta_a>0} \Delta_a (1 + 2H_a) \\
&= \sum_{a:\Delta_a>0} \left(\Delta_a + 2\Delta_a \cdot \frac{9C_{\mathrm{ME}} \log(1/\delta_1)}{(\rho_1 - \delta_1)^2 \Delta_a^2} \right) \\
&= \sum_{a:\Delta_a>0} \left(\Delta_a + \frac{18C_{\mathrm{ME}} \log(1/\delta_1)}{(\rho_1 - \delta_1)^2 \Delta_a} \right).
\end{aligned}$$

Step 4: expected regret. Decompose

$$\mathbb{E}[\mathrm{Reg}(T)] = \mathbb{E}[\mathrm{Reg}(T)\mathbf{1}\{\mathcal{E}\}] + \mathbb{E}[\mathrm{Reg}(T)\mathbf{1}\{\mathcal{E}^c\}].$$

On $\mathcal{E}$, Step 3 yields

$$\mathbb{E}[\mathrm{Reg}(T)\mathbf{1}\{\mathcal{E}\}] \le \sum_{a:\Delta_a>0} \left(\Delta_a + \frac{18C_{\mathrm{ME}} \log(1/\delta_1)}{(\rho_1 - \delta_1)^2 \Delta_a} \right).$$

On $\mathcal{E}^c$, always $\mathrm{Reg}(T) \le T$, so Step 1 implies

$$\mathbb{E}[\mathrm{Reg}(T)\mathbf{1}\{\mathcal{E}^c\}] \le T\mathbb{P}(\mathcal{E}^c) \le T \cdot \frac{1}{2T} = \frac{1}{2}.$$

Therefore

$$\mathbb{E}[\mathrm{Reg}(T)] \le \sum_{a:\Delta_a>0} \Delta_a + \frac{18C_{\mathrm{ME}} \log(1/\delta_1)}{(\rho_1 - \delta_1)^2} \sum_{a:\Delta_a>0} \frac{1}{\Delta_a} + \frac{1}{2}.$$

Now use the assumption $\delta_1 \le \rho_1/2$, which implies

$$\rho_1 - \delta_1 \ge \frac{\rho_1}{2}, \qquad \text{hence} \qquad \frac{1}{(\rho_1 - \delta_1)^2} \le \frac{4}{\rho_1^2}.$$

Also, since $\delta_1 = \frac{1}{2MT}$,

$$\log(1/\delta_1) = \log(2MT) = \log(2KT(1+L)).$$

Using $\rho_1 = \rho/M$, we obtain

$$\frac{\log(1/\delta_1)}{(\rho_1 - \delta_1)^2} \le 4\log(2KT(1+L)) \cdot \frac{M^2}{\rho^2} = 4\log(2KT(1+L)) \cdot \frac{K^2(1+L)^2}{\rho^2}.$$

Substituting this into the previous display gives

$$\mathbb{E}[\mathrm{Reg}(T)] \le \sum_{a:\Delta_a>0} \Delta_a + 72C_{\mathrm{ME}} \cdot \frac{K^2(1+L)^2}{\rho^2} \sum_{a:\Delta_a>0} \frac{\log(2KT(1+L))}{\Delta_a} + \frac{1}{2}.$$

Since $\log(2KT(1+L)) \ge 1$ and $\frac{K^2(1+L)^2}{\rho^2} \ge 1$, the first term can be absorbed into the same scale. Therefore, for a universal constant $C > 0$,

$$\mathbb{E}[\mathrm{Reg}(T)] \le C \cdot \frac{K^2(1+L)^2}{\rho^2} \sum_{a:\Delta_a>0} \left(\Delta_a + \frac{\log(2KT(1+L))}{\Delta_a} \right) + \frac{1}{2}.$$

□

B Replicable Ridge Regression Proof

Theorem 3.1 *Fix $\delta \in (0,1)$ and $\rho \in (2\delta, 1)$, and let $\widetilde{\theta}_n$ be the output of* RepRidge *(Algorithm 2) for fixed $\{(x_i, y_i)\}_{i=1}^n$. Then with probability at least $1-\delta$,*

$$\|\widetilde{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta)\left(1 + \frac{d}{\rho - 2\delta}\right).$$

Moreover, the estimator is ρ-replicable: if RepRidge *(Algorithm 2) is run twice on the same covariate sequence with the same shared random shift u but with independent response noise, producing outputs $\widetilde{\theta}_n^{(1)}$ and $\widetilde{\theta}_n^{(2)}$, then $\mathbb{P}\left(\widetilde{\theta}_n^{(1)} = \widetilde{\theta}_n^{(2)}\right) \ge 1 - \rho$.*

B.1 Proof of Theorem 3.1

Let $d \ge 1$ and $n \ge 1$ and recall that we have a fixed covariate sequence $x_1, \dots, x_n \in \mathbb{R}^d$,. Fix $\lambda > 0$ and define the Gram matrix

$$V_n := \lambda I + \sum_{i=1}^n x_i x_i^\top.$$

Following Assumption 3.1, there is an unknown parameter $\theta^\star \in \mathbb{R}^d$ with $\|\theta^\star\|_2 \le S$ and Observations follow the linear model $y_i = x_i^\top \theta^\star + \varepsilon_i, i = 1, \dots, n$, where (ε_i) is conditionally σ-sub-Gaussian with respect to the filtration $(\mathcal{F}_i)$ generated by $\{x_1, \dots, x_i, y_1, \dots, y_i\}$, i.e. for every i and every $\gamma \in \mathbb{R}$,

$$\mathbb{E}[\exp(\gamma \varepsilon_i) \mid \mathcal{F}_{i-1}] \le \exp\left(\frac{\sigma^2 \gamma^2}{2}\right).$$

The ridge estimator is given by

$$\hat{\theta}_n := V_n^{-1} \sum_{i=1}^n x_i y_i.$$

We next define the standard ridge confidence radius. For $\delta \in (0,1)$, define

$$\beta_n(\delta) := \sigma \sqrt{2 \log\left(\frac{\det(V_n)^{1/2}}{\det(\lambda I)^{1/2}\, \delta}\right)} + \sqrt{\lambda}\, S.$$

A standard self-normalized martingale inequality implies (Abbasi-yadkori et al., 2011)

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta)\right) \ge 1 - \delta.$$

Lemma B.1 (High-probability confidence bound for RepRidge**).** *Suppose Assumption 3.1 holds and let $\tilde{\theta}_n$ be the output of* RepRidge *(Algorithm 2). Then with probability at least $1-\delta$,*

$$\|\tilde{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta) + \frac{\alpha}{2}\sqrt{d} = \beta_n(\delta)\left(1 + \frac{d}{\rho - 2\delta}\right).$$

Proof of Lemma B.1. Let $\mathcal{E} := \{\|\hat{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta)\}$. Under Assumption 3.1, the standard self-normalized confidence bound for ridge regression implies

$$\mathbb{P}(\mathcal{E}) \ge 1 - \delta.$$

We now prove the claim in three steps.

Step 1: Bound the rounding error in the V_n-norm. Let $z_n := V_n^{1/2}\hat{\theta}_n$. By the definition of RepRidge (Algorithm 2),

$$V_n^{1/2}\tilde{\theta}_n = Q_{\alpha,u}(z_n).$$

Therefore

$$\begin{aligned}\|\tilde{\theta}_n - \hat{\theta}_n\|_{V_n} &= \|V_n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n)\|_2 \\ &= \|Q_{\alpha,u}(z_n) - z_n\|_2.\end{aligned}$$

Now fix any coordinate $j \in [d]$. By construction, $Q_{\alpha,u}(z_n)_j$ is the midpoint of the unique interval of the form $[u_j + k\alpha,\, u_j + (k+1)\alpha)$ that contains $(z_n)_j$. Since this interval has length α, its midpoint is at distance at most $\alpha/2$ from every point in the interval. Hence

$$|Q_{\alpha,u}(z_n)_j - (z_n)_j| \le \frac{\alpha}{2}.$$

Summing over coordinates gives

$$\begin{aligned}\|Q_{\alpha,u}(z_n) - z_n\|_2^2 &= \sum_{j=1}^{d} \big(Q_{\alpha,u}(z_n)_j - (z_n)_j\big)^2 \\ &\le \sum_{j=1}^{d} \left(\frac{\alpha}{2}\right)^2 \\ &= d\left(\frac{\alpha}{2}\right)^2.\end{aligned}$$

Taking square roots yields

$$\|Q_{\alpha,u}(z_n) - z_n\|_2 \le \frac{\alpha}{2}\sqrt{d}.$$

Consequently,

$$\|\tilde{\theta}_n - \hat{\theta}_n\|_{V_n} \le \frac{\alpha}{2}\sqrt{d}.$$

Step 2: Combine the statistical error and the rounding error. On the event $\mathcal{E}$, the triangle inequality gives

$$\begin{aligned}\|\tilde{\theta}_n - \theta^\star\|_{V_n} &\le \|\tilde{\theta}_n - \hat{\theta}_n\|_{V_n} + \|\hat{\theta}_n - \theta^\star\|_{V_n} \\ &\le \frac{\alpha}{2}\sqrt{d} + \beta_n(\delta).\end{aligned}$$

Thus, on $\mathcal{E}$,

$$\|\tilde{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta) + \frac{\alpha}{2}\sqrt{d}.$$

Step 3: Substitute the choice of α. By the definition of the grid width used in `RepRidge`, $\alpha = \frac{2\beta_n(\delta)\sqrt{d}}{\rho - 2\delta}$. Therefore

$$\begin{aligned}\frac{\alpha}{2}\sqrt{d} &= \frac{1}{2} \cdot \frac{2\beta_n(\delta)\sqrt{d}}{\rho - 2\delta} \cdot \sqrt{d} \\ &= \beta_n(\delta)\frac{d}{\rho - 2\delta}.\end{aligned}$$

Substituting this into the bound from Step 2 yields

$$\begin{aligned}\|\tilde{\theta}_n - \theta^\star\|_{V_n} &\le \beta_n(\delta) + \frac{\alpha}{2}\sqrt{d} \\ &= \beta_n(\delta)\left(1 + \frac{d}{\rho - 2\delta}\right).\end{aligned}$$

Since $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$, the stated bound holds with probability at least $1 - \delta$. □

Lemma B.2 (ρ-replicability of `RepRidge`). *Consider two independent noise sequences $(\varepsilon_i^{(1)})_{i=1}^n$ and $(\varepsilon_i^{(2)})_{i=1}^n$, generating two response sequences $y_i^{(k)} = x_i^\top \theta^\star + \varepsilon_i^{(k)}, k \in \{1, 2\}$, with the same fixed covariates $x_1, \ldots, x_n$. Fix $\rho \in (2\delta, 1)$ and run `RepRidge` (Algorithm 2) twice using the same shared internal randomness u, producing the estimated parameter outputs $\tilde{\theta}_n^{(1)}$ and $\tilde{\theta}_n^{(2)}$. Then*

$$\mathbb{P}\left(\tilde{\theta}_n^{(1)} = \tilde{\theta}_n^{(2)}\right) \ge 1 - \rho.$$

Proof of Lemma B.2. For $k \in \{1,2\}$, let $\hat{\theta}_n^{(k)}$ denote the ridge estimator computed from the response sequence $(y_i^{(k)})_{i=1}^n$. Since the covariates $x_1, \dots, x_n$ are the same in the two runs, the matrix V_n, the radius $\beta_n(\delta)$, and the grid width α are identical across the two executions. Both runs use the same shared internal randomness u.

Define

$$z^{(k)} := V_n^{1/2}\hat{\theta}_n^{(k)}, \qquad \tilde{\theta}_n^{(k)} := V_n^{-1/2} Q_{\alpha,u}(z^{(k)}), \qquad k \in \{1,2\}.$$

Since $V_n^{-1/2}$ is invertible, we have

$$\tilde{\theta}_n^{(1)} = \tilde{\theta}_n^{(2)} \quad \Longleftrightarrow \quad Q_{\alpha,u}(z^{(1)}) = Q_{\alpha,u}(z^{(2)}).$$

We now prove the claim in three steps.

Step 1: Bound the distance between the two whitened ridge estimators. For $k \in \{1,2\}$, define the confidence event

$$\mathcal{E}_k := \left\{ \|\hat{\theta}_n^{(k)} - \theta^\star\|_{V_n} \le \beta_n(\delta) \right\}.$$

Under Assumption 3.1, the standard self-normalized confidence bound for ridge regression gives

$$\mathbb{P}(\mathcal{E}_k) \ge 1 - \delta, \qquad k \in \{1,2\}.$$

Therefore, by a union bound,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - 2\delta.$$

On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\begin{aligned} \|z^{(1)} - z^{(2)}\|_2 &= \|V_n^{1/2}(\hat{\theta}_n^{(1)} - \hat{\theta}_n^{(2)})\|_2 \\ &\le \|V_n^{1/2}(\hat{\theta}_n^{(1)} - \theta^\star)\|_2 + \|V_n^{1/2}(\hat{\theta}_n^{(2)} - \theta^\star)\|_2 \\ &= \|\hat{\theta}_n^{(1)} - \theta^\star\|_{V_n} + \|\hat{\theta}_n^{(2)} - \theta^\star\|_{V_n} \\ &\le 2\beta_n(\delta). \end{aligned}$$

Step 2: Bound the probability that the two rounded vectors differ. Condition on fixed values of $z^{(1)}$ and $z^{(2)}$. If $Q_{\alpha,u}(z^{(1)}) \ne Q_{\alpha,u}(z^{(2)})$, then there exists at least one coordinate $j \in [d]$ such that

$$Q_{\alpha,u}(z^{(1)})_j \ne Q_{\alpha,u}(z^{(2)})_j.$$

Hence, by a union bound over coordinates,

$$\mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \ne Q_{\alpha,u}(z^{(2)}) \mid z^{(1)}, z^{(2)}\Big) \le \sum_{j=1}^d \mathbb{P}\Big(Q_{\alpha,u}(z^{(1)})_j \ne Q_{\alpha,u}(z^{(2)})_j \mid z^{(1)}, z^{(2)}\Big).$$

Now fix a coordinate j. As a function of the shared random shift $u_j \sim \mathrm{Unif}([0,\alpha))$, the one-dimensional rounded outputs differ only if there is a grid boundary between $z_j^{(1)}$ and $z_j^{(2)}$. The grid boundaries are the points $u_j + k\alpha$ for $k \in \mathbb{Z}$. Over one period of length α, the set of shifts u_j for which such a boundary lies between $z_j^{(1)}$ and $z_j^{(2)}$ has Lebesgue measure at most $|z_j^{(1)} - z_j^{(2)}|$. Therefore

$$\mathbb{P}\Big(Q_{\alpha,u}(z^{(1)})_j \ne Q_{\alpha,u}(z^{(2)})_j \mid z^{(1)}, z^{(2)}\Big) \le \frac{|z_j^{(1)} - z_j^{(2)}|}{\alpha}.$$

Summing over j yields

$$\begin{aligned} \mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \ne Q_{\alpha,u}(z^{(2)}) \mid z^{(1)}, z^{(2)}\Big) &\le \sum_{j=1}^d \frac{|z_j^{(1)} - z_j^{(2)}|}{\alpha} \\ &= \frac{\|z^{(1)} - z^{(2)}\|_1}{\alpha} \\ &\le \frac{\sqrt{d}\,\|z^{(1)} - z^{(2)}\|_2}{\alpha}. \end{aligned}$$

On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, Step 1 implies $\|z^{(1)} - z^{(2)}\|_2 \le 2\beta_n(\delta)$, and hence

$$\mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \ne Q_{\alpha,u}(z^{(2)}) \mid \mathcal{E}_1 \cap \mathcal{E}_2\Big) \le \frac{2\beta_n(\delta)\sqrt{d}}{\alpha}.$$

Step 3: Conclude ρ-replicability. Using the equivalence

$$\tilde{\theta}_n^{(1)} \neq \tilde{\theta}_n^{(2)} \iff Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)}),$$

we obtain

$$\begin{aligned}\mathbb{P}\Big(\tilde{\theta}_n^{(1)} \neq \tilde{\theta}_n^{(2)}\Big) &= \mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)})\Big) \\ &\leq \mathbb{P}((\mathcal{E}_1 \cap \mathcal{E}_2)^c) + \mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)}) \mid \mathcal{E}_1 \cap \mathcal{E}_2\Big) \\ &\leq 2\delta + \frac{2\beta_n(\delta)\sqrt{d}}{\alpha}.\end{aligned}$$

By the definition of the grid width in RepRidge,

$$\alpha = \frac{2\beta_n(\delta)\sqrt{d}}{\rho - 2\delta}.$$

Substituting this into the previous display gives

$$\mathbb{P}\Big(\tilde{\theta}_n^{(1)} \neq \tilde{\theta}_n^{(2)}\Big) \leq 2\delta + \rho - 2\delta = \rho.$$

Therefore

$$\mathbb{P}\Big(\tilde{\theta}_n^{(1)} = \tilde{\theta}_n^{(2)}\Big) \geq 1 - \rho.$$

This proves that RepRidge is ρ-replicable. □

C Core-Set Uniform Error Guarantee for RepRidge

Proposition 3.1. Suppose Assumption 3.2 holds. Then there exists an algorithm that uses $M = \widetilde{O}\left(\frac{d^3(d+\log(1/\delta))}{\rho^2\epsilon^2}\right)$ total samples and outputs a ρ replicable estimator $\widetilde{\theta} \in \mathbb{R}^d$ satisfying

$$\sup_{a\in\mathcal{A}} |\langle a, \widetilde{\theta} - \theta^\star\rangle| \leq \epsilon, \text{ with probability at least } 1-\delta.$$

Remark C.1 (Comparison with Lemma 9 of Esfandiari et al. (2023a))**.** Lemma 9 of Esfandiari et al. (2023a) requires pulling every arm in the core set, whose size is $|C| = \widetilde{\mathcal{O}}(d)$, and pays an arm-wise reproducible mean-estimation cost for each of these core actions. This leads to a total sample complexity of $\widetilde{\Omega}(d^5/(\rho^2\epsilon^2))$. Under the same core-set coverage condition, Proposition 3.1 obtains the same type of uniform prediction guarantee using $\widetilde{\mathcal{O}}(d^4/(\rho^2\epsilon^2))$ samples. Thus, up to logarithmic factors, the vector-level RepRidge construction improves the sample complexity by a factor of d.

C.1 Proof of Proposition 3.1

Proof. Fix a design scale $N \geq 1$. For each action $a \in C$, pull action a

$$n_a = \lceil N\pi(a) \rceil$$

times. Let $x_1, \ldots, x_M$ denote the resulting deterministic sequence of pulled actions. Thus each $x_t \in C \subseteq \mathcal{A}$, and

$$M = \sum_{a\in C} n_a.$$

Define the unregularized empirical design matrix

$$G_N = \sum_{t=1}^{M} x_t x_t^\top = \sum_{a\in C} n_a a a^\top,$$

and the regularized design matrix

$$V_N = I + G_N = I + \sum_{a \in C} n_a a a^\top.$$

Since $n_a = \lceil N\pi(a) \rceil \geq N\pi(a)$ for every $a \in C$, we have

$$G_N = \sum_{a \in C} n_a a a^\top \succeq N \sum_{a \in C} \pi(a) a a^\top = N V(\pi).$$

Since $V_N = I + G_N \succeq G_N$, it follows that

$$V_N \succeq N V(\pi).$$

Both matrices are positive definite, so inversion reverses the semidefinite order:

$$V_N^{-1} \preceq \frac{1}{N} V(\pi)^{-1}.$$

Therefore, for every $a \in \mathcal{A}$,

$$\|a\|^2_{V_N^{-1}} = a^\top V_N^{-1} a \leq \frac{1}{N} a^\top V(\pi)^{-1} a = \frac{1}{N} \|a\|^2_{V(\pi)^{-1}}.$$

Using Assumption 3.2,

$$\sup_{a \in \mathcal{A}} \|a\|^2_{V_N^{-1}} \leq \frac{4d}{N}.$$

Hence,

$$\sup_{a \in \mathcal{A}} \|a\|_{V_N^{-1}} \leq 2\sqrt{\frac{d}{N}}. \tag{1}$$

Now run `RepRidge` on the fixed design sequence $x_1, \ldots, x_M$ with regularization parameter $\lambda = 1$. The observations satisfy

$$y_t = \langle x_t, \theta^\star \rangle + \eta_t, \qquad t = 1, \ldots, M,$$

where the noise variables are independent, mean-zero, and 1-sub-Gaussian. Let

$$\delta_0 = \min\left\{\delta, \frac{\rho}{4}\right\}.$$

By Theorem 3.1, with probability at least $1 - \delta_0$,

$$\|\widetilde{\theta} - \theta^\star\|_{V_N} \leq \beta_N(\delta_0) \left(1 + \frac{d}{\rho - 2\delta_0}\right), \tag{2}$$

where, since $\lambda = 1$,

$$\beta_N(\delta_0) = \sqrt{2 \log\left(\frac{\det(V_N)^{1/2}}{\delta_0}\right)} + 1.$$

Equivalently,

$$\beta_N(\delta_0) = \sqrt{\log\det(V_N) + 2\log\left(\frac{1}{\delta_0}\right)} + 1.$$

We next upper bound $\log\det(V_N)$. Since $V_N = I + G_N$ and $G_N \succeq 0$, the trace-determinant inequality gives

$$\det(I + G_N) \leq \left(1 + \frac{\mathrm{tr}(G_N)}{d}\right)^d.$$

Moreover,

$$\mathrm{tr}(G_N) = \mathrm{tr}\left(\sum_{t=1}^{M} x_t x_t^\top\right) = \sum_{t=1}^{M} \|x_t\|_2^2.$$

Since $x_t \in \mathcal{A}$ and $\|x_t\|_2 \le 1$, we have

$$\mathrm{tr}(G_N) \le M.$$

Therefore,

$$\log\det(V_N) = \log\det(I + G_N) \le d\log\left(1 + \frac{M}{d}\right). \tag{3}$$

We now control the uniform prediction error. Fix any $a \in \mathcal{A}$. By Cauchy-Schwarz in the V_N-geometry,

$$\begin{aligned}|\langle a, \widetilde{\theta} - \theta^\star\rangle| &= |\langle V_N^{-1/2} a, V_N^{1/2}(\widetilde{\theta} - \theta^\star)\rangle| \\ &\le \|a\|_{V_N^{-1}} \|\widetilde{\theta} - \theta^\star\|_{V_N}.\end{aligned}$$

Taking the supremum over $a \in \mathcal{A}$ and using (1) and (2), on the probability-$(1-\delta_0)$ event,

$$\sup_{a\in\mathcal{A}} |\langle a, \widetilde{\theta} - \theta^\star\rangle| \le 2\sqrt{\frac{d}{N}}\left(1 + \frac{d}{\rho - 2\delta_0}\right)\beta_N(\delta_0).$$

Since $\delta_0 \le \rho/4$, we have $\rho - 2\delta_0 \ge \rho/2$. Hence,

$$1 + \frac{d}{\rho - 2\delta_0} \le 1 + \frac{2d}{\rho} \le \frac{3d}{\rho},$$

where the final inequality uses $d \ge 1$ and $\rho \in (0, 1)$. Therefore,

$$\sup_{a\in\mathcal{A}} |\langle a, \widetilde{\theta} - \theta^\star\rangle| \le \frac{6d}{\rho}\sqrt{\frac{d}{N}}\beta_N(\delta_0). \tag{4}$$

It remains to choose N large enough. Since

$$n_a = \lceil N\pi(a)\rceil \le N\pi(a) + 1,$$

we have

$$M = \sum_{a\in C} n_a \le \sum_{a\in C}(N\pi(a) + 1) = N + |C|.$$

By Assumption 3.2, $|C| = \mathcal{O}(d\log\log d)$. Thus,

$$M \le N + \mathcal{O}(d\log\log d). \tag{5}$$

Combining (3) and (5), there is a universal constant $c > 0$ such that

$$\beta_N(\delta_0) \le \sqrt{d\log\left(1 + \frac{N + cd\log\log d}{d}\right) + 2\log\left(\frac{1}{\delta_0}\right)} + 1.$$

Define

$$B_N = d\log\left(1 + \frac{N + cd\log\log d}{d}\right) + 2\log\left(\frac{1}{\delta_0}\right).$$

Then $\beta_N(\delta_0) \le \sqrt{B_N} + 1$. Since $(\sqrt{x} + 1)^2 \le 2x + 2$ for all $x \ge 0$,

$$\beta_N(\delta_0)^2 \le 2B_N + 2.$$

Squaring (4), we obtain

$$\left(\frac{6d}{\rho}\sqrt{\frac{d}{N}}\beta_N(\delta_0)\right)^2 = \frac{36d^3}{\rho^2 N}\beta_N(\delta_0)^2 \le \frac{72d^3}{\rho^2 N}(B_N + 1).$$

Thus, if

$$N \geq \frac{72d^3}{\rho^2\epsilon^2}(B_N + 1), \tag{6}$$

then

$$\sup_{a\in\mathcal{A}} |\langle a, \widetilde{\theta} - \theta^\star\rangle| \leq \epsilon.$$

The event on which this holds has probability at least $1 - \delta_0$, and since $\delta_0 \leq \delta$, it also has probability at least $1 - \delta$.

We next verify replicability. The design sequence $x_1, \ldots, x_M$ is deterministic because C, π, and the counts $n_a = \lceil N\pi(a)\rceil$ are fixed before observing any rewards. Consider two executions using the same design sequence and the same internal randomness in `RepRidge`, but independent reward noise. By Theorem 3.1,

$$\mathbb{P}\left(\widetilde{\theta}^{(1)} = \widetilde{\theta}^{(2)}\right) \geq 1 - \rho.$$

Therefore, the estimator is ρ-replicable. Finally, (6) is solved, up to logarithmic factors, by

$$N = \widetilde{\mathcal{O}}\left(\frac{d^3(d + \log(1/\delta))}{\rho^2\epsilon^2}\right).$$

Using $M \leq N + \mathcal{O}(d \log\log d)$, the total number of samples also satisfies

$$M = \widetilde{\mathcal{O}}\left(\frac{d^3(d + \log(1/\delta))}{\rho^2\epsilon^2}\right).$$

This proves the proposition. □

D Replicable Linear Bandit Proof

Theorem 4.1. Suppose `RepLinUCB` (Algorithm 3) is run with $B = \left\lceil d \log\left(1 + \frac{TL^2}{\lambda d}\right)\right\rceil$, $q = \left(1 + \frac{TL^2}{\lambda d}\right)^{\frac{d}{B}}$, $\delta_b = \frac{\delta}{B}$, $\rho_b = \frac{\rho}{B}$, for all $b \in \{0, \ldots, B-1\}$, and assume $\rho > 3\delta$. Then `RepLinUCB` (Algorithm 3) is ρ-replicable, i.e., if the algorithm is run twice on the action sets $(A_t)_{t=1}^T$ with shared internal randomness and independent noise realizations, producing action sequences $(a_t^{(1)})_{t=1}^T$ and $(a_t^{(2)})_{t=1}^T$, then $\mathbb{P}\left(\forall t \in [T] :\ a_t^{(1)} = a_t^{(2)}\right) \geq 1 - \rho$. Moreover, there exists a constant $C > 0$ such that, with probability at least $1 - \delta$, the regret satisfies

$$\mathrm{Reg}(T) \leq C\left(\sigma\sqrt{d\log\left(1 + \frac{TL^2}{\lambda d}\right) + \log\left(\frac{B}{\delta}\right)} + \sqrt{\lambda}\,S\right)\left(1 + \frac{dB}{\rho}\right)\sqrt{T\,d\log\left(1 + \frac{TL^2}{\lambda d}\right)}.$$

D.1 Proof of Theorem 4.1

Recall the sequence of action sets $(X_t)_{t=1}^T$ is fixed in advance by an oblivious adversary.

We maintain, for $t \geq 1$,

$$V_t := \lambda I + \sum_{s=1}^{t-1} x_s x_s^\top, \qquad b_t := \sum_{s=1}^{t-1} x_s r_s, \qquad \hat{\theta}_t := V_t^{-1} b_t.$$

We fix a realized batch grid $1 = t_0 < t_1 < \cdots < t_B = T + 1$ and per-batch budgets $\delta_b \in (0, 1)$ and $\rho_b \in (0, 1)$ for $b \in \{0, 1, \ldots, B-1\}$ satisfying $\rho_b > 2\delta_b$. At time t_b, let $n_b := t_b - 1$ and run replicable ridge on the past data $\{(x_s, r_s)\}_{s=1}^{n_b}$:

$$\tilde{\theta}_b \leftarrow \texttt{RepRidge}(\lambda, \delta_b, \rho_b; \{(x_s, r_s)\}_{s=1}^{n_b}).$$

Define, for $t \geq 1$ and $\delta \in (0, 1)$,

$$\beta_t(\delta) := \sigma \sqrt{2 \log\left(\frac{\det(V_t)^{1/2}}{\det(\lambda I)^{1/2}} \cdot \frac{1}{\delta} \right)} + \sqrt{\lambda}\, S,$$

and define the batch radius

$$\tilde{\beta}_b := \beta_{t_b}(\delta_b) \left(1 + \frac{d}{\rho_b - 2\delta_b} \right).$$

Then we can show that the estimated parameter $\tilde{\theta}_b$ by RepRidge satisfies a high probability concentration concentration around the true parameter θ^* as follows.

Lemma D.1 (Batch-start concentration for $\tilde{\theta}_b$). *Fix any realized batch grid $(t_b)_{b=0}^{B}$ and budgets $(\delta_b, \rho_b)_{b=0}^{B-1}$ with $\rho_b > 2\delta_b$ for all $b \in \{0, \dots, B-1\}$. For each batch $b \in \{0, \dots, B-1\}$, define the following event*

$$\mathcal{E}_b := \left\{ \|\hat{\theta}_{t_b} - \theta^\star\|_{V_{t_b}} \leq \beta_{t_b}(\delta_b) \right\}.$$

Then, for every $b \in \{0, \dots, B-1\}$, $\mathbb{P}(\mathcal{E}_b) \geq 1 - \delta_b$. Moreover, on $\mathcal{E}_b$,

$$\|\tilde{\theta}_b - \theta^\star\|_{V_{t_b}} \leq \tilde{\beta}_b := \beta_{t_b}(\delta_b) \left(1 + \frac{d}{\rho_b - 2\delta_b} \right).$$

Finally, with $\mathcal{E} := \bigcap_{b=0}^{B-1} \mathcal{E}_b$, we have $\mathbb{P}(\mathcal{E}) \geq 1 - \sum_{b=0}^{B-1} \delta_b$, and on $\mathcal{E}$ all of the above inequalities hold simultaneously for every batch $b \in \{0, \dots, B-1\}$.

Proof of Lemma D.1. Fix a batch $b \in \{0, \dots, B-1\}$. Let

$$\hat{\theta}_{t_b} = V_{t_b}^{-1} b_{t_b} \qquad \text{and} \qquad z_b := V_{t_b}^{1/2} \hat{\theta}_{t_b}.$$

By the definition of RepRidge at batch start t_b, the grid width is

$$\alpha_b := \frac{2\, \beta_{t_b}(\delta_b) \sqrt{d}}{\rho_b - 2\delta_b},$$

and if u_b denotes the shared random shift used at batch b, then

$$V_{t_b}^{1/2} \tilde{\theta}_b = Q_{\alpha_b, u_b}(z_b).$$

We prove the claim in four steps.

Step 1: Batch-start ridge confidence for $\hat{\theta}_{t_b}$. We first show that

$$\mathbb{P}(\mathcal{E}_b) \geq 1 - \delta_b.$$

Set $n_b := t_b - 1$. For each $s \leq n_b$, the action a_s is chosen using only information available before observing the noise at round s, so the played feature vector x_s is predictable with respect to the natural filtration. Under Assumption 3.1, the noise sequence is conditionally σ-sub-Gaussian, and therefore the standard self-normalized martingale inequality for ridge regression yields

$$\mathbb{P}\left(\|\hat{\theta}_{t_b} - \theta^\star\|_{V_{t_b}} \leq \sigma \sqrt{2 \log\left(\frac{\det(V_{t_b})^{1/2}}{\det(\lambda I)^{1/2}} \cdot \frac{1}{\delta_b} \right)} + \sqrt{\lambda}\, \|\theta^\star\|_2 \right) \geq 1 - \delta_b.$$

Using $\|\theta^\star\|_2 \leq S$, the quantity inside the probability is exactly $\beta_{t_b}(\delta_b)$. Hence

$$\mathbb{P}(\mathcal{E}_b) \geq 1 - \delta_b.$$

Step 2: Deterministic rounding error in the V_{t_b}-norm. We now bound $\|\tilde{\theta}_b - \hat{\theta}_{t_b}\|_{V_{t_b}}$. By construction,

$$\begin{aligned}\|\tilde{\theta}_b - \hat{\theta}_{t_b}\|_{V_{t_b}} &= \|V_{t_b}^{1/2}(\tilde{\theta}_b - \hat{\theta}_{t_b})\|_2 \\ &= \|Q_{\alpha_b,u_b}(z_b) - z_b\|_2.\end{aligned}$$

For each coordinate $j \in [d]$, the quantity $Q_{\alpha_b,u_b}(z_b)_j$ is the midpoint of the unique interval of length α_b in the shifted grid determined by u_b that contains $(z_b)_j$. Therefore

$$|Q_{\alpha_b,u_b}(z_b)_j - (z_b)_j| \le \frac{\alpha_b}{2}.$$

Summing over coordinates gives

$$\begin{aligned}\|Q_{\alpha_b,u_b}(z_b) - z_b\|_2^2 &= \sum_{j=1}^{d}\big(Q_{\alpha_b,u_b}(z_b)_j - (z_b)_j\big)^2 \\ &\le \sum_{j=1}^{d}\left(\frac{\alpha_b}{2}\right)^2 \\ &= d\left(\frac{\alpha_b}{2}\right)^2.\end{aligned}$$

Taking square roots yields

$$\|\tilde{\theta}_b - \hat{\theta}_{t_b}\|_{V_{t_b}} \le \frac{\alpha_b}{2}\sqrt{d}.$$

Step 3: Confidence bound for $\tilde{\theta}_b$. On the event $\mathcal{E}_b$, the triangle inequality gives

$$\begin{aligned}\|\tilde{\theta}_b - \theta^\star\|_{V_{t_b}} &\le \|\hat{\theta}_{t_b} - \theta^\star\|_{V_{t_b}} + \|\tilde{\theta}_b - \hat{\theta}_{t_b}\|_{V_{t_b}} \\ &\le \beta_{t_b}(\delta_b) + \frac{\alpha_b}{2}\sqrt{d}.\end{aligned}$$

Substituting the definition of α_b,

$$\begin{aligned}\beta_{t_b}(\delta_b) + \frac{\alpha_b}{2}\sqrt{d} &= \beta_{t_b}(\delta_b) + \frac{1}{2}\cdot\frac{2\,\beta_{t_b}(\delta_b)\sqrt{d}}{\rho_b - 2\delta_b}\cdot\sqrt{d} \\ &= \beta_{t_b}(\delta_b)\left(1 + \frac{d}{\rho_b - 2\delta_b}\right) \\ &= \tilde{\beta}_b.\end{aligned}$$

Hence, on $\mathcal{E}_b$,

$$\|\tilde{\theta}_b - \theta^\star\|_{V_{t_b}} \le \tilde{\beta}_b.$$

Step 4: Simultaneous validity over all batches. Define

$$\mathcal{E} := \bigcap_{b=0}^{B-1}\mathcal{E}_b.$$

By Step 1, for every batch b,

$$\mathbb{P}(\mathcal{E}_b^c) \le \delta_b.$$

Therefore, by a union bound,

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\left(\bigcup_{b=0}^{B-1}\mathcal{E}_b^c\right) \le \sum_{b=0}^{B-1}\delta_b.$$

Equivalently,

$$\mathbb{P}(\mathcal{E}) \ge 1 - \sum_{b=0}^{B-1}\delta_b.$$

On the event $\mathcal{E}$, we have $\mathcal{E}_b$ for every batch b, and therefore all the confidence and prediction inequalities established in Step 3 hold simultaneously for all $b \in \{0,\dots,B-1\}$. □

Using the concentration of the replicable parameter at the start of each batch $\tilde{\theta}_{t_b}$ we can show that the following bound on the regret holds.

Lemma D.2 (Regret bound for RepLinUCB on a realized determinant batch grid). *Run RepLinUCB (Algorithm 3) with parameters (λ, T, B, q) and budgets $(\delta_b, \rho_b)_{b=0}^{B-1}$. Let the realized batch grid be $1 = t_0 < t_1 < \cdots < t_{B'} = T + 1, B' \le B$. Define*

$$\widetilde{\beta}_{\max} := \max_{b \in \{0,\dots,B'-1\}} \widetilde{\beta}_b, \qquad M_T := \max\left\{\sqrt{q},\ \sqrt{\frac{\det(V_{T+1})}{q^{B-1}\det(\lambda I_d)}}\right\}.$$

Then with probability at least $1 - \sum_{b=0}^{B-1} \delta_b$,

$$\mathrm{Reg}(T) \le \max\{1,\ 2\widetilde{\beta}_{\max}\}\, M_T \sqrt{2TD_T}.$$

Proof. Let

$$\mathcal{E} := \bigcap_{b=0}^{B'-1} \mathcal{E}_b, \qquad \mathcal{E}_b := \left\{\|\hat{\theta}_{t_b} - \theta^\star\|_{V_{t_b}} \le \beta_{t_b}(\delta_b)\right\}.$$

By Lemma D.1, we have

$$\mathbb{P}(\mathcal{E}) \ge 1 - \sum_{b=0}^{B'-1} \delta_b \ge 1 - \sum_{b=0}^{B-1} \delta_b,$$

and on $\mathcal{E}$, for every realized batch $b \in \{0, \dots, B'-1\}$,

$$\|\tilde{\theta}_b - \theta^\star\|_{V_{t_b}} \le \widetilde{\beta}_b,$$

and equivalently, for every $x \in \mathbb{R}^d$,

$$|\langle x, \tilde{\theta}_b - \theta^\star\rangle| \le \widetilde{\beta}_b \|x\|_{V_{t_b}^{-1}}.$$

Fix the event $\mathcal{E}$ in what follows. We prove the regret bound in three steps.

Step 1: Reduce the regret to stale batch-start norms. For each round $t \in [T]$, let $b(t) \in \{0, \dots, B'-1\}$ denote the realized batch index such that $t \in [t_{b(t)}, t_{b(t)+1} - 1]$. Let $x_t^\star \in \arg\max_{x \in A_t} \langle x, \theta^\star\rangle$ denote an optimal action at round t. Since the algorithm uses the batch-start parameter $\tilde{\theta}_{b(t)}$ and radius $\widetilde{\beta}_{b(t)}$ throughout batch $b(t)$, the chosen action a_t satisfies

$$\langle a_t, \tilde{\theta}_{b(t)}\rangle + \widetilde{\beta}_{b(t)} \|a_t\|_{V_{t_{b(t)}}^{-1}} \ge \langle x_t^\star, \tilde{\theta}_{b(t)}\rangle + \widetilde{\beta}_{b(t)} \|x_t^\star\|_{V_{t_{b(t)}}^{-1}}.$$

Moreover, on $\mathcal{E}$,

$$\langle x_t^\star, \theta^\star\rangle \le \langle x_t^\star, \tilde{\theta}_{b(t)}\rangle + \widetilde{\beta}_{b(t)} \|x_t^\star\|_{V_{t_{b(t)}}^{-1}},$$
$$\langle a_t, \theta^\star\rangle \ge \langle a_t, \tilde{\theta}_{b(t)}\rangle - \widetilde{\beta}_{b(t)} \|a_t\|_{V_{t_{b(t)}}^{-1}}.$$

Subtracting the two displays gives

$$\langle x_t^\star, \theta^\star\rangle - \langle a_t, \theta^\star\rangle \le 2\widetilde{\beta}_{b(t)} \|a_t\|_{V_{t_{b(t)}}^{-1}}.$$

Also, since rewards lie in $[0, 1]$, the instantaneous regret is always at most 1. Therefore

$$\langle x_t^\star, \theta^\star\rangle - \langle a_t, \theta^\star\rangle \le \min\{1,\ 2\widetilde{\beta}_{b(t)} \|a_t\|_{V_{t_{b(t)}}^{-1}}\}.$$

Since $\widetilde{\beta}_{b(t)} \le \widetilde{\beta}_{\max}$, we obtain

$$\langle x_t^\star, \theta^\star\rangle - \langle a_t, \theta^\star\rangle \le \min\{1,\ 2\widetilde{\beta}_{\max} \|a_t\|_{V_{t_{b(t)}}^{-1}}\}.$$

Now for any $\kappa > 0$ and any $u \geq 0$,

$$\min\{1, \kappa u\} \leq \max\{1, \kappa\} \min\{1, u\}.$$

Applying this with $\kappa = 2\widetilde{\beta}_{\max}$ and $u = \|a_t\|_{V_{t_{b(t)}}^{-1}}$ yields

$$\langle x_t^\star, \theta^\star \rangle - \langle a_t, \theta^\star \rangle \leq \max\{1, 2\widetilde{\beta}_{\max}\} \min\{1, \|a_t\|_{V_{t_{b(t)}}^{-1}}\}.$$

Summing over t gives

$$\mathrm{Reg}(T) \leq \max\{1, 2\widetilde{\beta}_{\max}\} \sum_{t=1}^{T} \min\{1, \|a_t\|_{V_{t_{b(t)}}^{-1}}\}.$$

Step 2: Compare stale norms to current norms using determinant growth. We show that for every round $t \in [T]$,

$$\|a_t\|_{V_{t_{b(t)}}^{-1}} \leq M_T \, \|a_t\|_{V_t^{-1}}.$$

We split into two cases. If $B' < B$, then the batch budget is not exhausted. Hence for every realized batch $b \in \{0, \ldots, B' - 1\}$ and every $t \in [t_b, t_{b+1} - 1]$, the next batch has not yet been triggered by time t, so by the determinant batching rule,

$$\det(V_t) \leq q \, \det(V_{t_b}).$$

Also $V_t \succeq V_{t_b}$. Let

$$M := V_{t_b}^{-1/2} V_t V_{t_b}^{-1/2} \succeq I_d, \qquad u := V_{t_b}^{-1/2} a_t.$$

Then

$$\|a_t\|_{V_{t_b}^{-1}}^2 = \|u\|_2^2, \qquad \|a_t\|_{V_t^{-1}}^2 = u^\top M^{-1} u.$$

Diagonalizing M with eigenvalues $\lambda_1, \ldots, \lambda_d \geq 1$ and writing u in the eigenbasis gives

$$\|u\|_2^2 \leq \left(\max_i \lambda_i\right) u^\top M^{-1} u.$$

Since each $\lambda_i \geq 1$,

$$\max_i \lambda_i \leq \prod_{i=1}^{d} \lambda_i = \det(M) = \frac{\det(V_t)}{\det(V_{t_b})}.$$

Hence

$$\|a_t\|_{V_{t_b}^{-1}}^2 \leq \frac{\det(V_t)}{\det(V_{t_b})} \, \|a_t\|_{V_t^{-1}}^2 \leq q \, \|a_t\|_{V_t^{-1}}^2,$$

and so

$$\|a_t\|_{V_{t_b}^{-1}} \leq \sqrt{q} \, \|a_t\|_{V_t^{-1}} \leq M_T \, \|a_t\|_{V_t^{-1}}.$$

Now consider the case $B' = B$. For every batch $b \in \{0, \ldots, B - 2\}$, the trigger fired at t_{b+1}, so

$$\det(V_{t_{b+1}}) > q \, \det(V_{t_b}).$$

Multiplying these inequalities over $b = 0, \ldots, B - 2$ gives

$$\det(V_{t_{B-1}}) > q^{B-1} \det(V_{t_0}) = q^{B-1} \det(\lambda I_d).$$

For the non-final batches $b \in \{0, \ldots, B - 2\}$, exactly the same argument as above gives

$$\|a_t\|_{V_{t_b}^{-1}} \leq \sqrt{q} \, \|a_t\|_{V_t^{-1}} \leq M_T \, \|a_t\|_{V_t^{-1}} \qquad \text{for all } t \in [t_b, t_{b+1} - 1].$$

For the final batch $b = B-1$, if $t \in [t_{B-1}, T]$, then $V_t \preceq V_{T+1}$, and therefore

$$\frac{\det(V_t)}{\det(V_{t_{B-1}})} < \frac{\det(V_{T+1})}{q^{B-1}\det(\lambda I_d)}.$$

Applying the same determinant-ratio norm comparison as above with $A = V_t$ and $B = V_{t_{B-1}}$ yields

$$\|a_t\|_{V_{t_{B-1}}^{-1}} \le \|a_t\|_{V_t^{-1}} \sqrt{\frac{\det(V_t)}{\det(V_{t_{B-1}})}} \le \|a_t\|_{V_t^{-1}} \sqrt{\frac{\det(V_{T+1})}{q^{B-1}\det(\lambda I_d)}} \le M_T \|a_t\|_{V_t^{-1}}.$$

Thus in all cases,

$$\|a_t\|_{V_{t_{b(t)}}^{-1}} \le M_T \|a_t\|_{V_t^{-1}}.$$

Since $M_T \ge 1$, it follows that

$$\min\{1, \|a_t\|_{V_{t_{b(t)}}^{-1}}\} \le M_T \min\{1, \|a_t\|_{V_t^{-1}}\}.$$

Substituting into the bound from Step 1 gives

$$\mathrm{Reg}(T) \le \max\{1, 2\widetilde{\beta}_{\max}\} M_T \sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}\}.$$

Step 3: Apply the elliptical potential bound. Using $V_{t+1} = V_t + a_t a_t^\top$ and the matrix determinant lemma,

$$\det(V_{t+1}) = \det(V_t)\big(1 + \|a_t\|_{V_t^{-1}}^2\big).$$

Therefore

$$\log\left(\frac{\det(V_{T+1})}{\det(V_1)}\right) = \sum_{t=1}^{T} \log\left(1 + \|a_t\|_{V_t^{-1}}^2\right).$$

For every $u \ge 0$, we have $\log(1+u) \ge \frac{1}{2}\min\{u, 1\}$. Applying this with $u = \|a_t\|_{V_t^{-1}}^2$ yields

$$\sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}^2\} \le 2\log\left(\frac{\det(V_{T+1})}{\det(V_1)}\right) = 2D_T,$$

since $V_1 = \lambda I_d$. Finally, by Cauchy–Schwarz,

$$\begin{aligned}
\sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}\} &\le \sqrt{T \sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}\}^2} \\
&\le \sqrt{T \sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}^2\}} \\
&\le \sqrt{2TD_T}.
\end{aligned}$$

Combining this with the bound from Step 2 gives

$$\mathrm{Reg}(T) \le \max\{1, 2\widetilde{\beta}_{\max}\} M_T \sqrt{2TD_T}.$$

This proves the claim on the event $\mathcal{E}$, which holds with probability at least $1 - \sum_{b=0}^{B-1} \delta_b$. □

It remains to specialize Lemma D.2 to the parameter choices in Theorem 4.1. Since $\delta_b = \delta/B$ for all b, we have

$$\sum_{b=0}^{B-1} \delta_b = \delta,$$

so the high-probability event in Lemma D.2 holds with probability at least $1-\delta$.

We first control D_T. Since $V_{T+1} = \lambda I_d + \sum_{t=1}^{T} a_t a_t^\top$ and $\|a_t\|_2 \le L$ for all t, the standard determinant bound gives

$$\det(V_{T+1}) \le \det(\lambda I_d)\left(1+\frac{TL^2}{\lambda d}\right)^d.$$

Therefore

$$D_T = \log\left(\frac{\det(V_{T+1})}{\det(\lambda I_d)}\right) \le d\log\left(1+\frac{TL^2}{\lambda d}\right).$$

Next, by the choice

$$q = \left(1+\frac{TL^2}{\lambda d}\right)^{d/B},$$

we have

$$\begin{aligned} M_T &= \max\left\{\sqrt{q},\ \sqrt{\frac{\det(V_{T+1})}{q^{B-1}\det(\lambda I_d)}}\right\} \\ &\le \max\left\{\left(1+\frac{TL^2}{\lambda d}\right)^{\frac{d}{2B}},\ \sqrt{\frac{\left(1+\frac{TL^2}{\lambda d}\right)^d}{\left(1+\frac{TL^2}{\lambda d}\right)^{d(B-1)/B}}}\right\} \\ &= \left(1+\frac{TL^2}{\lambda d}\right)^{\frac{d}{2B}}. \end{aligned}$$

Since

$$B = \left\lceil d\log\left(1+\frac{TL^2}{\lambda d}\right)\right\rceil,$$

it follows that

$$\frac{d}{2B}\log\left(1+\frac{TL^2}{\lambda d}\right) \le \frac{1}{2},$$

and hence

$$M_T \le e^{1/2}.$$

We now bound $\widetilde{\beta}_{\max}$. For every realized batch $b \in \{0,\dots,B'-1\}$,

$$\widetilde{\beta}_b = \beta_b\left(1+\frac{d}{\rho_b - 2\delta_b}\right), \qquad \beta_b = \beta_{t_b}(\delta_b).$$

Since $t_b \le T+1$, $\delta_b = \delta/B$, and $\det(V_{t_b}) \le \det(V_{T+1})$, we have

$$\begin{aligned} \beta_b &= \sigma\sqrt{2\log\left(\frac{\det(V_{t_b})^{1/2}}{\det(\lambda I_d)^{1/2}}\cdot\frac{B}{\delta}\right)} + \sqrt{\lambda}\,S \\ &\le \sigma\sqrt{2\log\left(\left(1+\frac{TL^2}{\lambda d}\right)^{d/2}\frac{B}{\delta}\right)} + \sqrt{\lambda}\,S \\ &= \sigma\sqrt{d\log\left(1+\frac{TL^2}{\lambda d}\right) + 2\log\left(\frac{B}{\delta}\right)} + \sqrt{\lambda}\,S. \end{aligned}$$

Also, since $\rho_b = \rho/B$, $\delta_b = \delta/B$, and $\rho > 3\delta$,

$$\rho_b - 2\delta_b = \frac{\rho-2\delta}{B} \ge \frac{\rho}{3B},$$

so

$$1+\frac{d}{\rho_b-2\delta_b}\le 1+\frac{3dB}{\rho}\le 3\left(1+\frac{dB}{\rho}\right).$$

Combining the last two displays yields

$$\widetilde{\beta}_{\max}\le 3\left(\sigma\sqrt{d\log\left(1+\frac{TL^2}{\lambda d}\right)+2\log\left(\frac{B}{\delta}\right)}+\sqrt{\lambda}\,S\right)\left(1+\frac{dB}{\rho}\right).$$

Absorbing the factor of 2 inside the logarithm into a universal constant, we obtain

$$\widetilde{\beta}_{\max}\le C_1\left(\sigma\sqrt{d\log\left(1+\frac{TL^2}{\lambda d}\right)+\log\left(\frac{B}{\delta}\right)}+\sqrt{\lambda}\,S\right)\left(1+\frac{dB}{\rho}\right)$$

for a universal constant $C_1>0$.

Substituting the bounds on D_T, M_T, and $\widetilde{\beta}_{\max}$ into Lemma D.2, and absorbing universal constants, gives

$$\begin{aligned}\mathrm{Reg}(T)&\le \max\{1,2\widetilde{\beta}_{\max}\}\,M_T\sqrt{2TD_T}\\&\le C\left(\sigma\sqrt{d\log\left(1+\frac{TL^2}{\lambda d}\right)+\log\left(\frac{B}{\delta}\right)}+\sqrt{\lambda}\,S\right)\left(1+\frac{dB}{\rho}\right)\sqrt{T\,d\log\left(1+\frac{TL^2}{\lambda d}\right)},\end{aligned}$$

which is exactly the claimed regret bound.

Finally we show the replicability of the chosen actions using the following lemma.

Lemma D.3 (Trajectory ρ-replicability of `RepLinUCB`). *Fix any realized batch grid $(t_b)_{b=0}^{B'}$ with $B'\le B$, and budgets $(\rho_b)_{b=0}^{B-1}$. Consider two runs $k\in\{1,2\}$ of `RepLinUCB` (Algorithm 3) on the same obliviously fixed action-set sequence $(A_t)_{t=1}^{T}$, using the same shared internal randomness in every `RepRidge` call and independent noise sequences $(\xi_t^{(k)})_{t=1}^{T}$. Let $(a_t^{(k)})_{t=1}^{T}$ denote the resulting action sequences, and let $(\tilde{\theta}_b^{(k)})_{b=0}^{B'-1}$ denote the batch-start estimators produced in the two runs. Then*

$$\mathbb{P}\Big(\forall t\in[T]:\ a_t^{(1)}=a_t^{(2)}\Big)\ge 1-\sum_{b=0}^{B-1}\rho_b.$$

Proof. For each batch $b\in\{0,\dots,B'-1\}$, define the trajectory-match event up to the start of batch b by

$$\mathcal{M}_b:=\left\{\forall s\le t_b-1:\ a_s^{(1)}=a_s^{(2)}\right\}.$$

Since $t_0=1$, the event $\mathcal{M}_0$ holds trivially.

For each batch b, define the batch-estimator mismatch event

$$A_b:=\left\{\tilde{\theta}_b^{(1)}\neq\tilde{\theta}_b^{(2)}\right\}.$$

We prove the claim in three steps.

Step 1: If the batch-start estimators agree, then the entire batch trajectory agrees. Fix a batch $b\in\{0,\dots,B'-1\}$ and assume that $\mathcal{M}_b$ holds and also $\tilde{\theta}_b^{(1)}=\tilde{\theta}_b^{(2)}$. Since $\mathcal{M}_b$ holds, the two runs have selected the same actions up to time t_b-1. Therefore they have the same design matrix at batch start,

$$V_{t_b}^{(1)}=V_{t_b}^{(2)}=:V_{t_b}.$$

Because the batch-start confidence radius $\widetilde{\beta}_b$ is a deterministic function of V_{t_b} and the prescribed budget parameters, it is also identical across the two runs.

Now for any round $t \in \{t_b, \dots, t_{b+1} - 1\}$, both runs maximize the same function

$$a \mapsto \langle a, \tilde{\theta}_b \rangle + \widetilde{\beta}_b \|a\|_{V_{t_b}^{-1}}$$

over the same action set A_t, with the same deterministic tie-breaking rule. Hence

$$a_t^{(1)} = a_t^{(2)} \qquad \text{for all } t = t_b, \dots, t_{b+1} - 1.$$

Thus $\mathcal{M}_{b+1}$ holds. Equivalently,

$$\mathcal{M}_{b+1}^c \cap \mathcal{M}_b \subseteq A_b.$$

Step 2: Bound the probability that the batch-b `RepRidge` calls disagree on $\mathcal{M}_b$. Fix a batch $b \in \{0, \dots, B' - 1\}$ and condition on the event $\mathcal{M}_b$. Then the two runs have identical actions $a_1, \dots, a_{t_b - 1}$ entering batch b, and hence identical matrices

$$V_{t_b}^{(1)} = V_{t_b}^{(2)} =: V_{t_b}.$$

Therefore the corresponding ridge confidence radii and rounding widths are also identical:

$$\beta_{t_b}(\delta_b)^{(1)} = \beta_{t_b}(\delta_b)^{(2)} = \beta_{t_b}(\delta_b), \qquad \alpha_b^{(1)} = \alpha_b^{(2)} = \alpha_b,$$

where

$$\alpha_b = \frac{2\beta_{t_b}(\delta_b)\sqrt{d}}{\rho_b - 2\delta_b}.$$

Define

$$\hat{\theta}_{t_b}^{(k)} := V_{t_b}^{-1} b_{t_b}^{(k)}, \qquad z_b^{(k)} := V_{t_b}^{1/2} \hat{\theta}_{t_b}^{(k)}, \qquad k \in \{1, 2\}.$$

Also define the two batch-start confidence events

$$\mathcal{E}_b^{(k)} := \left\{ \|\hat{\theta}_{t_b}^{(k)} - \theta^\star\|_{V_{t_b}} \le \beta_{t_b}(\delta_b) \right\}, \qquad k \in \{1, 2\}.$$

By the standard self-normalized ridge confidence bound applied separately to the two runs,

$$\mathbb{P}(\mathcal{E}_b^{(k)}) \ge 1 - \delta_b, \qquad k \in \{1, 2\}.$$

Hence, by a union bound,

$$\mathbb{P}\big((\mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)})^c\big) \le 2\delta_b.$$

On the event $\mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}$, we have

$$\begin{aligned} \|z_b^{(1)} - z_b^{(2)}\|_2 &= \|V_{t_b}^{1/2}(\hat{\theta}_{t_b}^{(1)} - \hat{\theta}_{t_b}^{(2)})\|_2 \\ &\le \|V_{t_b}^{1/2}(\hat{\theta}_{t_b}^{(1)} - \theta^\star)\|_2 + \|V_{t_b}^{1/2}(\hat{\theta}_{t_b}^{(2)} - \theta^\star)\|_2 \\ &= \|\hat{\theta}_{t_b}^{(1)} - \theta^\star\|_{V_{t_b}} + \|\hat{\theta}_{t_b}^{(2)} - \theta^\star\|_{V_{t_b}} \\ &\le 2\beta_{t_b}(\delta_b). \end{aligned}$$

Now recall that

$$\tilde{\theta}_b^{(k)} = V_{t_b}^{-1/2} Q_{\alpha_b, u_b}(z_b^{(k)}), \qquad k \in \{1, 2\},$$

where the same shared random shift u_b is used in both runs. Since $V_{t_b}^{-1/2}$ is invertible,

$$A_b = \left\{ Q_{\alpha_b, u_b}(z_b^{(1)}) \ne Q_{\alpha_b, u_b}(z_b^{(2)}) \right\}.$$

We now bound the probability of A_b on $\mathcal{M}_b$. Decompose

$$\begin{aligned} \mathbb{P}(A_b \cap \mathcal{M}_b) &\le \mathbb{P}\big(\mathcal{M}_b \cap (\mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)})^c\big) + \mathbb{P}\big(A_b \cap \mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}\big) \\ &\le \mathbb{P}\big((\mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)})^c\big) + \mathbb{P}\big(A_b \cap \mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}\big). \end{aligned}$$

For the second term, condition on fixed values of $(z_b^{(1)}, z_b^{(2)})$. If

$$Q_{\alpha_b,u_b}(z_b^{(1)}) \neq Q_{\alpha_b,u_b}(z_b^{(2)}),$$

then there exists at least one coordinate $j \in [d]$ such that

$$Q_{\alpha_b,u_b}(z_b^{(1)})_j \neq Q_{\alpha_b,u_b}(z_b^{(2)})_j.$$

Hence, by a union bound over coordinates,

$$\mathbb{P}(A_b \mid z_b^{(1)}, z_b^{(2)}) \leq \sum_{j=1}^{d} \mathbb{P}\Big(Q_{\alpha_b,u_b}(z_b^{(1)})_j \neq Q_{\alpha_b,u_b}(z_b^{(2)})_j \mid z_b^{(1)}, z_b^{(2)}\Big).$$

Fix a coordinate j. As a function of the shared random shift $(u_b)_j \sim \mathrm{Unif}([0, \alpha_b))$, the one-dimensional rounded outputs differ only if there exists a grid boundary between $z_{b,j}^{(1)}$ and $z_{b,j}^{(2)}$. The grid boundaries are the points

$$(u_b)_j + m\alpha_b, \qquad m \in \mathbb{Z}.$$

As $(u_b)_j$ varies uniformly over one period of length α_b, the set of shifts for which such a boundary lies between $z_{b,j}^{(1)}$ and $z_{b,j}^{(2)}$ has Lebesgue measure at most $|z_{b,j}^{(1)} - z_{b,j}^{(2)}|$. Therefore

$$\mathbb{P}\Big(Q_{\alpha_b,u_b}(z_b^{(1)})_j \neq Q_{\alpha_b,u_b}(z_b^{(2)})_j \mid z_b^{(1)}, z_b^{(2)}\Big) \leq \frac{|z_{b,j}^{(1)} - z_{b,j}^{(2)}|}{\alpha_b}.$$

Summing over j gives

$$\begin{aligned}
\mathbb{P}(A_b \mid z_b^{(1)}, z_b^{(2)}) &\leq \sum_{j=1}^{d} \frac{|z_{b,j}^{(1)} - z_{b,j}^{(2)}|}{\alpha_b} \\
&= \frac{\|z_b^{(1)} - z_b^{(2)}\|_1}{\alpha_b} \\
&\leq \frac{\sqrt{d}\,\|z_b^{(1)} - z_b^{(2)}\|_2}{\alpha_b}.
\end{aligned}$$

On the event $\mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}$, this implies

$$\mathbb{P}(A_b \mid z_b^{(1)}, z_b^{(2)}) \leq \frac{2\beta_{t_b}(\delta_b)\sqrt{d}}{\alpha_b} = \rho_b - 2\delta_b.$$

Taking expectations yields

$$\begin{aligned}
\mathbb{P}\big(A_b \cap \mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}\big) &= \mathbb{E}\Big[\mathbf{1}_{\mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}} \mathbf{1}_{A_b}\Big] \\
&= \mathbb{E}\Big[\mathbf{1}_{\mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}}\, \mathbb{P}(A_b \mid z_b^{(1)}, z_b^{(2)})\Big] \\
&\leq \rho_b - 2\delta_b.
\end{aligned}$$

Combining the preceding bounds gives

$$\mathbb{P}(A_b \cap \mathcal{M}_b) \leq 2\delta_b + (\rho_b - 2\delta_b) = \rho_b.$$

Using Step 1, we conclude that

$$\mathbb{P}(\mathcal{M}_{b+1}^c \cap \mathcal{M}_b) \leq \mathbb{P}(A_b \cap \mathcal{M}_b) \leq \rho_b.$$

Step 3: Union bound over batches. Since $\mathcal{M}_0$ holds surely, any failure of full trajectory agreement must first occur at some batch transition. Therefore

$$\mathcal{M}_{B'}^c \subseteq \bigcup_{b=0}^{B'-1} (\mathcal{M}_{b+1}^c \cap \mathcal{M}_b).$$

By a union bound and Step 2,

$$\mathbb{P}(\mathcal{M}_{B'}^c) \le \sum_{b=0}^{B'-1} \mathbb{P}(\mathcal{M}_{b+1}^c \cap \mathcal{M}_b) \le \sum_{b=0}^{B'-1} \rho_b.$$

Since $\mathcal{M}_{B'}$ is exactly the event

$$\left\{\forall t \in [T] :\ a_t^{(1)} = a_t^{(2)}\right\},$$

we obtain

$$\mathbb{P}\Big(\forall t \in [T] :\ a_t^{(1)} = a_t^{(2)}\Big) \ge 1 - \sum_{b=0}^{B'-1} \rho_b.$$

Because $B' \le B$, this also implies

$$\mathbb{P}\Big(\forall t \in [T] :\ a_t^{(1)} = a_t^{(2)}\Big) \ge 1 - \sum_{b=0}^{B-1} \rho_b.$$

Finally, fix any batch $b \in \{0, \dots, B'-1\}$. First,

$$\mathcal{M}_b^c \subseteq \bigcup_{j=0}^{b-1} (\mathcal{M}_{j+1}^c \cap \mathcal{M}_j),$$

so by a union bound and Step 2,

$$\mathbb{P}(\mathcal{M}_b^c) \le \sum_{j=0}^{b-1} \rho_j.$$

Also,

$$\mathbb{P}(A_b) \le \mathbb{P}(A_b \cap \mathcal{M}_b) + \mathbb{P}(\mathcal{M}_b^c) \le \rho_b + \sum_{j=0}^{b-1} \rho_j = \sum_{j=0}^{b} \rho_j.$$

Therefore

$$\mathbb{P}\Big(\tilde{\theta}_b^{(1)} = \tilde{\theta}_b^{(2)}\Big) \ge 1 - \sum_{j=0}^{b} \rho_j.$$

In particular, if batch $B-1$ is realized, then

$$\mathbb{P}\Big(\tilde{\theta}_{B-1}^{(1)} = \tilde{\theta}_{B-1}^{(2)}\Big) \ge 1 - \sum_{b=0}^{B-1} \rho_b.$$

This completes the proof. □

Using $\rho_b = \rho/B$ immediately proves the replicability gurantee in Theorem 4.1

E Replicable Generalized Linear Bandits

We next extend the optimistic generalized linear bandit algorithm of Filippi et al. (2010) to a replicable setting. As in the linear case, the main difficulty is that recomputing the generalized linear estimator after every new reward makes the action highly sensitive to the realized noise, so two runs with shared internal randomness but independent rewards may quickly diverge. To control this sensitivity, we combine determinant-triggered batching with a replicable penalized generalized linear estimator. The resulting algorithm updates its policy only at batch starts, and within each batch it reuses a single replicable optimistic estimate throughout.

E.1 Replicable Generalized Linear Estimation

Before turning to the bandit problem, we first introduce the fixed-design estimation primitive that will be invoked at the start of each batch.

Let $d \geq 1$ and $n \geq 0$, let $\Theta \subset \mathbb{R}^d$ be a known closed convex set, and fix $\lambda > 0$. Given covariates $x_1, \ldots, x_n \in \mathbb{R}^d$, define

$$V_n := \lambda I + \sum_{i=1}^{n} x_i x_i^\top.$$

The response model is as follows.

Assumption E.1. *There is an unknown parameter $\theta^\star \in \Theta$ with $\|\theta^\star\|_2 \leq S$. Observations satisfy*

$$y_i = \mu(x_i^\top \theta^\star) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where (ε_i) is conditionally σ-sub-Gaussian with respect to the filtration $(\mathcal{F}_i)$ generated by $\{x_1, \ldots, x_i, y_1, \ldots, y_i\}$, that is, for every i and $\gamma \in \mathbb{R}$,

$$\mathbb{E}[\exp(\gamma \varepsilon_i) \mid \mathcal{F}_{i-1}] \leq \exp\left(\frac{\sigma^2 \gamma^2}{2}\right).$$

The inverse link function $\mu : \mathbb{R} \to \mathbb{R}$ is continuously differentiable and globally Lipschitz with constant k_μ. Moreover, there is a known constant $c_\mu > 0$ such that

$$\mu'(x_i^\top \theta) \geq c_\mu \qquad \text{for every } \theta \in \Theta \text{ and } i \in [n].$$

For $n = 0$, this derivative condition is vacuous.

Let $b : \mathbb{R} \to \mathbb{R}$ be any antiderivative of μ, so that $b' = \mu$. We define the penalized quasi-log-likelihood by

$$\mathcal{L}_n(\theta) := \sum_{i=1}^{n} \Big(y_i x_i^\top \theta - b(x_i^\top \theta) \Big) - \frac{\lambda}{2} \|\theta\|_2^2, \qquad \theta \in \Theta.$$

Since $b'' = \mu' \geq 0$ and $\lambda > 0$, the function $\mathcal{L}_n$ is strictly concave on Θ, and therefore it has a unique maximizer, which we denote by $\widehat{\theta}_n := \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$. We also define $\underline{c}_\mu := \min\{1, c_\mu\}$. For $n = 0$, all sums are empty and

$$\widehat{\theta}_0 = \arg\min_{\theta \in \Theta} \|\theta\|_2^2.$$

The estimator `RepGLM` below is the generalized linear analogue of `RepRidge`. It first computes the penalized generalized linear estimator $\widehat{\theta}_n$, passes to the whitened coordinates $z_n = V_n^{1/2} \widehat{\theta}_n$, and then performs randomized midpoint rounding on a shared shifted grid. The rounded point is then mapped back to parameter space.

The key point is that the natural statistical error metric for the penalized generalized linear estimator will again be the V_n-norm. After whitening, nearby estimators in V_n-norm become nearby in Euclidean distance, and the same shifted-grid argument used in the linear case can therefore be applied. The following theorem gives the fixed-design guarantee for `RepGLM`.

Theorem E.1 (Confidence and Replicability of `RepGLM` (Algorithm 4))**.** *Suppose Assumption E.1 holds, fix $\delta \in (0, 1)$ and $\rho \in (2\delta, 1)$, and let $\widetilde{\theta}_n$ be the output of `RepGLM` (Algorithm 4) for fixed covariates $\{x_i\}_{i=1}^n$. Then with probability at least $1 - \delta$,*

$$\|\widetilde{\theta}_n - \theta^\star\|_{V_n} \leq \beta_n(\delta) \left(1 + \frac{d}{\rho - 2\delta}\right).$$

Consequently, for every $x \in \mathbb{R}^d$, with probability at least $1 - \delta$,

$$|\mu(x^\top \widetilde{\theta}_n) - \mu(x^\top \theta^\star)| \leq k_\mu \, \beta_n(\delta) \left(1 + \frac{d}{\rho - 2\delta}\right) \|x\|_{V_n^{-1}}.$$

Moreover, the estimator is ρ-replicable: if `RepGLM` (Algorithm 4) is run twice on the same covariate sequence with the same shared random shift u but with independent response noise, producing outputs $\widetilde{\theta}_n^{(1)}$ and $\widetilde{\theta}_n^{(2)}$, then

$$\mathbb{P}\Big(\widetilde{\theta}_n^{(1)} = \widetilde{\theta}_n^{(2)}\Big) \geq 1 - \rho.$$

Algorithm 4 Replicable Penalized GLM Estimation via Randomized Grid Rounding (RepGLM)

Require: Regularization $\lambda > 0$, confidence $\delta \in (0,1)$, target replicability $\rho \in (2\delta, 1)$, data $\{(x_i, y_i)\}_{i=1}^n$, parameter set Θ

1: $V_n \leftarrow \lambda I + \sum_{i=1}^n x_i x_i^\top$

2: $\widehat{\theta}_n \leftarrow \arg\max_{\theta\in\Theta} \left\{ \sum_{i=1}^n (y_i x_i^\top \theta - b(x_i^\top \theta)) - \frac{\lambda}{2}\|\theta\|_2^2 \right\}$ ▷ Penalized GLM estimate

3: $\beta_n(\delta) \leftarrow \frac{1}{\underline{c}_\mu} \left(\sigma \sqrt{2 \log\left(\frac{\det(V_n)^{1/2}}{\det(\lambda I)^{1/2}\, \delta} \right)} + \sqrt{\lambda}\, S \right)$ ▷ Fixed-design confidence radius

4: $z_n \leftarrow V_n^{1/2} \widehat{\theta}_n$ ▷ Whitening

5: Set $\alpha \leftarrow \frac{2\,\beta_n(\delta)\sqrt{d}}{\rho - 2\delta}$, sample a fresh shared $v \sim \mathrm{Unif}([0,1)^d)$ independently of the data and previous calls, and set $u \leftarrow \alpha v$ ▷ Grid width chosen to guarantee ρ-replicability

6: Define $Q_{\alpha,u} : \mathbb{R}^d \to \mathbb{R}^d$ coordinatewise by ▷ Randomized Grid Rounding

$$\big(Q_{\alpha,u}(z)\big)_j = \alpha \left\lfloor \frac{z_j - u_j}{\alpha} \right\rfloor + u_j + \frac{\alpha}{2}, \qquad j = 1, \ldots, d$$

7: $\widetilde{\theta}_n \leftarrow V_n^{-1/2} Q_{\alpha,u}(z_n)$ ▷ Map back to parameter space

8: **return** $\widetilde{\theta}_n$

E.2 Proof of Theorem E.1

We begin with the following deterministic and statistical ingredients.

Lemma E.2 (Strong monotonicity of the penalized GLM score). *Define*

$$g_n(\theta) := \lambda\theta + \sum_{i=1}^n \mu(x_i^\top \theta) x_i, \qquad \theta \in \Theta.$$

Then for every $\theta, \theta' \in \Theta$,

$$\langle g_n(\theta) - g_n(\theta'), \theta - \theta' \rangle \geq \underline{c}_\mu \|\theta - \theta'\|_{V_n}^2.$$

Proof of Lemma E.2. Fix $\theta, \theta' \in \Theta$. Since Θ is convex, the segment

$$\theta_s := s\theta + (1-s)\theta', \qquad s \in [0,1],$$

lies in Θ. By the fundamental theorem of calculus,

$$g_n(\theta) - g_n(\theta') = \left(\int_0^1 \nabla g_n(\theta_s)\, ds \right) (\theta - \theta').$$

Moreover,

$$\nabla g_n(\vartheta) = \lambda I + \sum_{i=1}^n \mu'(x_i^\top \vartheta) x_i x_i^\top.$$

By Assumption E.1, for every $\vartheta \in \Theta$ and every $i \in [n]$,

$$\mu'(x_i^\top \vartheta) \geq c_\mu.$$

Hence

$$\begin{aligned} \nabla g_n(\vartheta) &\succeq \lambda I + c_\mu \sum_{i=1}^n x_i x_i^\top \\ &\succeq \underline{c}_\mu \lambda I + \underline{c}_\mu \sum_{i=1}^n x_i x_i^\top \\ &= \underline{c}_\mu V_n. \end{aligned}$$

Therefore

$$\begin{aligned}\langle g_n(\theta) - g_n(\theta'), \theta - \theta' \rangle &= \int_0^1 (\theta - \theta')^\top \nabla g_n(\theta_s)(\theta - \theta')\, ds \\ &\geq \underline{c}_\mu (\theta - \theta')^\top V_n (\theta - \theta') \\ &= \underline{c}_\mu \|\theta - \theta'\|_{V_n}^2.\end{aligned}$$

□

Lemma E.3 (High-probability confidence bound for `RepGLM`). *Suppose Assumption E.1 holds and let $\widetilde{\theta}_n$ be the output of `RepGLM` (Algorithm 4). Then with probability at least $1 - \delta$,*

$$\|\widetilde{\theta}_n - \theta^\star\|_{V_n} \leq \beta_n(\delta) \left(1 + \frac{d}{\rho - 2\delta}\right).$$

Proof of Lemma E.3. We prove the claim in four steps.

Step 1: A confidence bound for the penalized GLM estimator $\widehat{\theta}_n$. The gradient of the penalized quasi-log-likelihood is

$$\begin{aligned}\nabla \mathcal{L}_n(\theta) &= \sum_{i=1}^n \big(y_i - \mu(x_i^\top \theta)\big) x_i - \lambda\theta \\ &= \sum_{i=1}^n y_i x_i - g_n(\theta).\end{aligned}$$

Since $\widehat{\theta}_n$ maximizes the differentiable concave function $\mathcal{L}_n$ over the closed convex set Θ, the first-order optimality condition gives

$$\langle \nabla \mathcal{L}_n(\widehat{\theta}_n), \theta - \widehat{\theta}_n \rangle \leq 0 \qquad \text{for all } \theta \in \Theta.$$

Taking $\theta = \theta^\star \in \Theta$, we obtain

$$\left\langle \sum_{i=1}^n y_i x_i - g_n(\widehat{\theta}_n), \theta^\star - \widehat{\theta}_n \right\rangle \leq 0.$$

Equivalently,

$$\left\langle g_n(\widehat{\theta}_n) - \sum_{i=1}^n y_i x_i, \widehat{\theta}_n - \theta^\star \right\rangle \leq 0.$$

Using $y_i = \mu(x_i^\top \theta^\star) + \varepsilon_i$, we have

$$\begin{aligned}\sum_{i=1}^n y_i x_i &= \sum_{i=1}^n \mu(x_i^\top \theta^\star) x_i + \sum_{i=1}^n \varepsilon_i x_i \\ &= g_n(\theta^\star) - \lambda\theta^\star + \sum_{i=1}^n \varepsilon_i x_i.\end{aligned}$$

Substituting this into the previous display yields

$$\langle g_n(\widehat{\theta}_n) - g_n(\theta^\star), \widehat{\theta}_n - \theta^\star \rangle \leq \left\langle \sum_{i=1}^n \varepsilon_i x_i - \lambda\theta^\star, \widehat{\theta}_n - \theta^\star \right\rangle.$$

By Lemma E.2, the left-hand side is lower bounded by

$$\underline{c}_\mu \|\widehat{\theta}_n - \theta^\star\|_{V_n}^2.$$

Therefore

$$\underline{c}_\mu \|\widehat{\theta}_n - \theta^\star\|_{V_n}^2 \leq \left(\left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_{V_n^{-1}} + \|\lambda\theta^\star\|_{V_n^{-1}} \right) \|\widehat{\theta}_n - \theta^\star\|_{V_n}.$$

Since $V_n \succeq \lambda I$, we have

$$\|\lambda\theta^\star\|_{V_n^{-1}} \le \sqrt{\lambda}\|\theta^\star\|_2 \le \sqrt{\lambda}\, S.$$

Moreover, the standard self-normalized martingale inequality gives (Abbasi-yadkori et al., 2011)

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \varepsilon_i x_i\right\|_{V_n^{-1}} \le \sigma\sqrt{2\log\left(\frac{\det(V_n)^{1/2}}{\det(\lambda I)^{1/2}\,\delta}\right)}\right) \ge 1-\delta.$$

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

$$\underline{c}_\mu\|\widehat{\theta}_n - \theta^\star\|_{V_n} \le \sigma\sqrt{2\log\left(\frac{\det(V_n)^{1/2}}{\det(\lambda I)^{1/2}\,\delta}\right)} + \sqrt{\lambda}\, S.$$

By the definition of $\beta_n(\delta)$, this means that with probability at least $1-\delta$,

$$\|\widehat{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta).$$

Define the event

$$\mathcal{E} := \left\{\|\widehat{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta)\right\}.$$

Then $\mathbb{P}(\mathcal{E}) \ge 1-\delta$.

Step 2: Deterministic rounding error in the V_n-norm. Let $z_n := V_n^{1/2}\widehat{\theta}_n$. By the definition of `RepGLM` (Algorithm 4),

$$V_n^{1/2}\widetilde{\theta}_n = Q_{\alpha,u}(z_n).$$

Therefore

$$\begin{aligned}\|\widetilde{\theta}_n - \widehat{\theta}_n\|_{V_n} &= \|V_n^{1/2}(\widetilde{\theta}_n - \widehat{\theta}_n)\|_2 \\ &= \|Q_{\alpha,u}(z_n) - z_n\|_2.\end{aligned}$$

For every coordinate $j \in [d]$, the quantity $Q_{\alpha,u}(z_n)_j$ is the midpoint of the unique interval of length α in the shifted grid determined by u_j that contains $(z_n)_j$. Hence

$$|Q_{\alpha,u}(z_n)_j - (z_n)_j| \le \frac{\alpha}{2}.$$

Summing over j gives

$$\begin{aligned}\|Q_{\alpha,u}(z_n) - z_n\|_2^2 &= \sum_{j=1}^{d}\big(Q_{\alpha,u}(z_n)_j - (z_n)_j\big)^2 \\ &\le \sum_{j=1}^{d}\left(\frac{\alpha}{2}\right)^2 \\ &= d\left(\frac{\alpha}{2}\right)^2.\end{aligned}$$

Taking square roots yields

$$\|\widetilde{\theta}_n - \widehat{\theta}_n\|_{V_n} \le \frac{\alpha}{2}\sqrt{d}.$$

Step 3: Combine the statistical error and the rounding error. On the event $\mathcal{E}$, the triangle inequality gives

$$\begin{aligned}\|\widetilde{\theta}_n - \theta^\star\|_{V_n} &\le \|\widetilde{\theta}_n - \widehat{\theta}_n\|_{V_n} + \|\widehat{\theta}_n - \theta^\star\|_{V_n} \\ &\le \frac{\alpha}{2}\sqrt{d} + \beta_n(\delta).\end{aligned}$$

By the definition of the grid width in RepGLM,

$$\alpha = \frac{2\,\beta_n(\delta)\sqrt{d}}{\rho - 2\delta}.$$

Hence

$$\frac{\alpha}{2}\sqrt{d} = \beta_n(\delta)\frac{d}{\rho - 2\delta}.$$

Substituting this into the previous display gives

$$\|\widetilde{\theta}_n - \theta^\star\|_{V_n} \le \beta_n(\delta)\left(1 + \frac{d}{\rho - 2\delta}\right).$$

Since $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$, the claim follows.

Step 4: Reward-space consequence. Fix any $x \in \mathbb{R}^d$. By the Lipschitz property of μ,

$$\begin{aligned}|\mu(x^\top\widetilde{\theta}_n) - \mu(x^\top\theta^\star)| &\le k_\mu |x^\top(\widetilde{\theta}_n - \theta^\star)| \\ &\le k_\mu \|x\|_{V_n^{-1}} \|\widetilde{\theta}_n - \theta^\star\|_{V_n}.\end{aligned}$$

Substituting the bound from Step 3 completes the proof. □

Lemma E.4 (ρ-replicability of RepGLM). *Consider two independent noise sequences $(\varepsilon_i^{(1)})_{i=1}^n$ and $(\varepsilon_i^{(2)})_{i=1}^n$, generating two response sequences*

$$y_i^{(k)} = \mu(x_i^\top\theta^\star) + \varepsilon_i^{(k)}, \qquad k \in \{1,2\},$$

with the same fixed covariates $x_1, \dots, x_n$. Fix $\rho \in (2\delta, 1)$ and run RepGLM (Algorithm 4) twice using the same shared internal randomness u, producing outputs $\widetilde{\theta}_n^{(1)}$ and $\widetilde{\theta}_n^{(2)}$. Then

$$\mathbb{P}\left(\widetilde{\theta}_n^{(1)} = \widetilde{\theta}_n^{(2)}\right) \ge 1 - \rho.$$

Proof of Lemma E.4. For $k \in \{1,2\}$, let $\widehat{\theta}_n^{(k)}$ denote the penalized generalized linear estimator computed from the response sequence $(y_i^{(k)})_{i=1}^n$. Since the covariates are identical across the two runs, the matrix V_n, the radius $\beta_n(\delta)$, and the grid width α are also identical across the two executions. Both runs use the same shared internal randomness u.

Define

$$\begin{aligned}z^{(k)} &:= V_n^{1/2}\widehat{\theta}_n^{(k)}, \\ \widetilde{\theta}_n^{(k)} &:= V_n^{-1/2} Q_{\alpha,u}(z^{(k)}), \qquad k \in \{1,2\}.\end{aligned}$$

Since $V_n^{-1/2}$ is invertible, we have

$$\widetilde{\theta}_n^{(1)} = \widetilde{\theta}_n^{(2)} \iff Q_{\alpha,u}(z^{(1)}) = Q_{\alpha,u}(z^{(2)}).$$

We now prove the claim in three steps.

Step 1: Bound the distance between the two whitened penalized GLM estimators. For $k \in \{1,2\}$, define the confidence event

$$\mathcal{E}_k := \left\{\|\widehat{\theta}_n^{(k)} - \theta^\star\|_{V_n} \le \beta_n(\delta)\right\}.$$

By the argument in Step 1 of the proof of Lemma E.3, we have

$$\mathbb{P}(\mathcal{E}_k) \ge 1 - \delta, \qquad k \in \{1,2\}.$$

Hence, by a union bound,

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - 2\delta.$$

On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\begin{aligned}\|z^{(1)} - z^{(2)}\|_2 &= \|V_n^{1/2}(\widehat{\theta}_n^{(1)} - \widehat{\theta}_n^{(2)})\|_2 \\ &\le \|V_n^{1/2}(\widehat{\theta}_n^{(1)} - \theta^\star)\|_2 + \|V_n^{1/2}(\widehat{\theta}_n^{(2)} - \theta^\star)\|_2 \\ &= \|\widehat{\theta}_n^{(1)} - \theta^\star\|_{V_n} + \|\widehat{\theta}_n^{(2)} - \theta^\star\|_{V_n} \\ &\le 2\beta_n(\delta).\end{aligned}$$

Step 2: Bound the probability that the two rounded vectors differ. Condition on fixed values of $z^{(1)}$ and $z^{(2)}$. If $Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)})$, then there exists at least one coordinate $j \in [d]$ such that

$$Q_{\alpha,u}(z^{(1)})_j \neq Q_{\alpha,u}(z^{(2)})_j.$$

Hence, by a union bound over coordinates,

$$\begin{aligned}&\mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)}) \mid z^{(1)}, z^{(2)}\Big)\\ &\quad\leq \sum_{j=1}^{d} \mathbb{P}\Big(Q_{\alpha,u}(z^{(1)})_j \neq Q_{\alpha,u}(z^{(2)})_j \mid z^{(1)}, z^{(2)}\Big).\end{aligned}$$

Fix a coordinate j. As a function of the shared random shift $u_j \sim \mathrm{Unif}([0,\alpha))$, the one-dimensional rounded outputs differ only if there is a grid boundary between $z_j^{(1)}$ and $z_j^{(2)}$. The grid boundaries are the points $u_j + m\alpha$ for $m \in \mathbb{Z}$. Over one period of length α, the set of shifts for which such a boundary lies between $z_j^{(1)}$ and $z_j^{(2)}$ has Lebesgue measure at most $|z_j^{(1)} - z_j^{(2)}|$. Therefore

$$\mathbb{P}\Big(Q_{\alpha,u}(z^{(1)})_j \neq Q_{\alpha,u}(z^{(2)})_j \mid z^{(1)}, z^{(2)}\Big) \leq \frac{|z_j^{(1)} - z_j^{(2)}|}{\alpha}.$$

Summing over coordinates yields

$$\begin{aligned}&\mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)}) \mid z^{(1)}, z^{(2)}\Big)\\ &\quad\leq \sum_{j=1}^{d} \frac{|z_j^{(1)} - z_j^{(2)}|}{\alpha}\\ &\quad= \frac{\|z^{(1)} - z^{(2)}\|_1}{\alpha}\\ &\quad\leq \frac{\sqrt{d}\,\|z^{(1)} - z^{(2)}\|_2}{\alpha}.\end{aligned}$$

On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, Step 1 implies

$$\mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)}) \mid \mathcal{E}_1 \cap \mathcal{E}_2\Big) \leq \frac{2\beta_n(\delta)\sqrt{d}}{\alpha}.$$

Step 3: Conclude ρ-replicability. Using the equivalence

$$\widetilde{\theta}_n^{(1)} \neq \widetilde{\theta}_n^{(2)} \quad\Longleftrightarrow\quad Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)}),$$

we obtain

$$\begin{aligned}&\mathbb{P}\Big(\widetilde{\theta}_n^{(1)} \neq \widetilde{\theta}_n^{(2)}\Big)\\ &\quad= \mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)})\Big)\\ &\quad\leq \mathbb{P}((\mathcal{E}_1 \cap \mathcal{E}_2)^c)\\ &\qquad+ \mathbb{P}\Big(Q_{\alpha,u}(z^{(1)}) \neq Q_{\alpha,u}(z^{(2)}) \mid \mathcal{E}_1 \cap \mathcal{E}_2\Big)\\ &\quad\leq 2\delta + \frac{2\beta_n(\delta)\sqrt{d}}{\alpha}.\end{aligned}$$

By the definition of the grid width in `RepGLM`,

$$\alpha = \frac{2\beta_n(\delta)\sqrt{d}}{\rho - 2\delta}.$$

Substituting this into the previous display gives

$$\mathbb{P}\Big(\widetilde{\theta}_n^{(1)} \neq \widetilde{\theta}_n^{(2)}\Big) \leq 2\delta + \rho - 2\delta = \rho.$$

Therefore

$$\mathbb{P}\Big(\widetilde{\theta}_n^{(1)} = \widetilde{\theta}_n^{(2)}\Big) \geq 1 - \rho.$$

□

Algorithm 5 Replicable Batched GLM-UCB (RepGLMUCB)

Require: Regularization $\lambda > 0$, horizon T, batch budget $B \in \mathbb{N}$, determinant growth factor $q > 1$
Require: Per-batch budgets $(\delta_b, \rho_b)_{b=0}^{B-1}$ with $\delta_b \in (0,1)$, $\rho_b \in (0,1)$, and $\rho_b > 2\delta_b$

Initialize $V_1 \leftarrow \lambda I_d$
Initialize batch index $b \leftarrow 0$ and batch-start time $t_b \leftarrow 1$
for $t = 1$ to T **do**
  Observe compact action set $A_t \subset \mathbb{R}^d$
  if $t = t_b$ **then**
    $\widetilde{\theta}_b \leftarrow \texttt{RepGLM}\Big(\lambda, \delta_b, \rho_b; \{(a_s, r_s)\}_{s=1}^{t_b-1}, \Theta\Big)$ ▷ Replicable batch parameter
    $\beta_b \leftarrow \frac{1}{\underline{c}_\mu}\left(\sigma\sqrt{2\log\left(\frac{\det(V_{t_b})^{1/2}}{\det(\lambda I)^{1/2}\,\delta_b}\right)} + \sqrt{\lambda}\,S\right)$ ▷ GLM confidence radius at batch start
    $\widetilde{\beta}_b \leftarrow k_\mu \beta_b \left(1 + \frac{d}{\rho_b - 2\delta_b}\right)$ ▷ Inflated reward-space batch radius
  Choose

$$a_t \in \arg\max_{a \in A_t} \left\{ \mu(\langle a, \widetilde{\theta}_b \rangle) + \widetilde{\beta}_b \|a\|_{V_{t_b}^{-1}} \right\}$$

  using a fixed tie-breaking rule ▷ Batch-start GLM-UCB rule
  Observe r_t and update $V_{t+1} \leftarrow V_t + a_t a_t^\top$ ▷ Update design matrix
  if $b \le B - 2$ **and** $\det(V_{t+1}) > q \det(V_{t_b})$ **then**
    Start a new batch: set $b \leftarrow b + 1$ and $t_b \leftarrow t + 1$ ▷ Determinant trigger

Proof of Theorem E.1. The confidence and reward-confidence statements follow from Lemma E.3, and the replicability statement follows from Lemma E.4. □

E.3 Replicable Batched GLM-UCB

We now turn to the bandit problem. The setting is the generalized linear analogue of the linear bandit model of the previous section.

Assumption E.2. *There is an unknown parameter $\theta^\star \in \Theta \subset \mathbb{R}^d$, where Θ is a known closed convex set and $\|\theta^\star\|_2 \le S$. At each round $t \in [T]$, after the learner selects an action $a_t \in A_t \subset \mathbb{R}^d$, it observes a reward*

$$r_t = \mu(a_t^\top \theta^\star) + \varepsilon_t,$$

where (ε_t) is conditionally σ-sub-Gaussian with respect to the filtration $(\mathcal{F}_t)$ generated by $\{A_1, a_1, r_1, \ldots, A_t, a_t, r_t\}$, that is, for every t and $\gamma \in \mathbb{R}$,

$$\mathbb{E}[\exp(\gamma \varepsilon_t) \mid \mathcal{F}_{t-1}, A_t, a_t] \le \exp\left(\frac{\sigma^2 \gamma^2}{2}\right).$$

The inverse link function $\mu : \mathbb{R} \to \mathbb{R}$ is continuously differentiable and globally Lipschitz with constant k_μ. Moreover,

$$c_\mu := \inf_{\theta \in \Theta,\ t \in [T],\ a \in A_t} \mu'(a^\top \theta) > 0.$$

Finally, the action sets satisfy $\|a\|_2 \le L$ for all $t \in [T]$ and all $a \in A_t$, and the mean rewards are bounded:

$$\mu(a^\top \theta^\star) \in [0,1] \qquad \textit{for all feasible } a.$$

Every call to RepGLM uses the same known uniform lower bound c_μ from Assumption E.2.

As in the non-replicable GLM-UCB algorithm of Filippi et al. (2010), the optimistic score is constructed directly in the reward space rather than in the parameter space. The difference is that here we update the generalized linear estimator only at determinant-triggered batch starts, and at each such update we replace the usual generalized linear estimate by the replicable estimate returned by RepGLM. This yields the batched optimistic algorithm in RepGLMUCB.

Throughout a batch, the same pair $(\widetilde{\theta}_b, \widetilde{\beta}_b)$ is reused and actions are selected by maximizing the batch-start reward-space upper confidence bound in Line 9. The design matrix V_t is updated after every observation, but the policy itself only changes at determinant-triggered batch starts.

The following theorem gives the main guarantee for RepGLMUCB (Algorithm 5).

Theorem E.5 (Replicability and Regret Bound for RepGLMUCB (Algorithm 5))**.** *Suppose Assumption E.2 holds and the action sets $(A_t)_{t=1}^T$ are chosen by an oblivious adversary. Further suppose RepGLMUCB (Algorithm 5) is run with*

$$B = \left\lceil d \log\left(1 + \frac{TL^2}{\lambda d}\right)\right\rceil, \qquad q = \left(1 + \frac{TL^2}{\lambda d}\right)^{d/B}, \qquad \delta_b = \frac{\delta}{B}, \qquad \rho_b = \frac{\rho}{B},$$

for all $b \in \{0, \dots, B-1\}$. Then for $\rho \in (3\delta, 1)$, RepGLMUCB (Algorithm 5) is ρ-replicable, that is, if the algorithm is run twice on action sets $(A_t)_{t=1}^T$ with shared internal randomness and independent environment randomness, producing action sequences $(a_t^{(1)})_{t=1}^T$ and $(a_t^{(2)})_{t=1}^T$, then

$$\mathbb{P}\left(\forall t \in [T] : a_t^{(1)} = a_t^{(2)}\right) \geq 1 - \rho.$$

Moreover, there exists a constant $C > 0$ such that, with probability at least $1 - \delta$, the regret of RepGLMUCB satisfies

$$\mathrm{Reg}_{\mathrm{GLB}}(T) \leq C \max\left\{1,\ \frac{k_\mu}{\underline{c}_\mu}\left(\sigma\sqrt{d\log\left(1 + \tfrac{TL^2}{\lambda d}\right) + \log\left(\tfrac{B}{\delta}\right)} + \sqrt{\lambda}\, S\right)\left(1 + \tfrac{dB}{\rho}\right)\right\}\sqrt{T\, d \log\left(1 + \tfrac{TL^2}{\lambda d}\right)}.$$

E.4 Proof of Theorem E.5

We write

$$D_T := \log\left(\frac{\det(V_{T+1})}{\det(\lambda I_d)}\right).$$

The proof proceeds in three parts. We first establish batch-start concentration for the replicable generalized linear estimator, then derive a regret bound on an arbitrary realized determinant batch grid, and finally prove trajectory replicability.

Lemma E.6 (Batch-start concentration for $\widetilde{\theta}_b$). *Let $(t_b)_{b=0}^{B'}$ be the random batch grid generated by RepGLMUCB, with $1 = t_0 < \cdots < t_{B'} = T + 1$ and $B' \leq B$. Fix budgets $(\delta_b, \rho_b)_{b=0}^{B-1}$ with $\rho_b > 2\delta_b$ for every b. For each realized batch, let $\widehat{\theta}_{t_b}$ denote the penalized generalized linear estimator computed from $\{(a_s, r_s)\}_{s=1}^{t_b-1}$, and define*

$$\beta_b := \frac{1}{\underline{c}_\mu}\left(\sigma\sqrt{2\log\left(\frac{\det(V_{t_b})^{1/2}}{\det(\lambda I_d)^{1/2}\,\delta_b}\right)} + \sqrt{\lambda}\, S\right).$$

For every $b \in \{0, \dots, B-1\}$, let $\mathcal{E}_b$ be the event that either batch b is not realized or

$$\|\widehat{\theta}_{t_b} - \theta^\star\|_{V_{t_b}} \leq \beta_b.$$

Then $\mathbb{P}(\mathcal{E}_b) \geq 1 - \delta_b$ for every b. Moreover, on $\mathcal{E}_b$, whenever batch b is realized,

$$\|\widetilde{\theta}_b - \theta^\star\|_{V_{t_b}} \leq \beta_b\left(1 + \frac{d}{\rho_b - 2\delta_b}\right),$$

$$|\mu(a^\top \widetilde{\theta}_b) - \mu(a^\top \theta^\star)| \leq \widetilde{\beta}_b \|a\|_{V_{t_b}^{-1}} \qquad \text{for all } a \in \mathbb{R}^d,$$

where

$$\widetilde{\beta}_b := k_\mu \beta_b\left(1 + \frac{d}{\rho_b - 2\delta_b}\right).$$

Finally, with

$$\mathcal{E} := \bigcap_{b=0}^{B-1} \mathcal{E}_b,$$

we have

$$\mathbb{P}(\mathcal{E}) \geq 1 - \sum_{b=0}^{B-1} \delta_b,$$

and all the above inequalities hold simultaneously for every realized batch on $\mathcal{E}$.

Proof of Lemma E.6. Fix $b \in \{0, \dots, B-1\}$. For every $t \in \{1, \dots, T+1\}$, let $\widehat{\theta}_t$ be the penalized GLM estimator based on the first $t-1$ observations, and write

$$g_t(\theta) := \lambda\theta + \sum_{s=1}^{t-1} \mu(a_s^\top \theta) a_s.$$

The first-order optimality argument from Lemma E.3, applied pathwise to the observed data, gives

$$\langle g_t(\widehat{\theta}_t) - g_t(\theta^\star), \widehat{\theta}_t - \theta^\star \rangle \leq \left\langle \sum_{s=1}^{t-1} \varepsilon_s a_s - \lambda\theta^\star, \widehat{\theta}_t - \theta^\star \right\rangle.$$

By the uniform curvature assumption and Lemma E.2,

$$\underline{c}_\mu \|\widehat{\theta}_t - \theta^\star\|_{V_t}^2 \leq \left(\left\| \sum_{s=1}^{t-1} \varepsilon_s a_s \right\|_{V_t^{-1}} + \sqrt{\lambda}\, S \right) \|\widehat{\theta}_t - \theta^\star\|_{V_t}.$$

Thus, including the case of zero estimation error,

$$\|\widehat{\theta}_t - \theta^\star\|_{V_t} \leq \frac{1}{\underline{c}_\mu} \left(\left\| \sum_{s=1}^{t-1} \varepsilon_s a_s \right\|_{V_t^{-1}} + \sqrt{\lambda}\, S \right).$$

Each action is selected before its corresponding reward noise is observed. Therefore the time-uniform self-normalized martingale inequality (Abbasi-yadkori et al., 2011) gives, with probability at least $1 - \delta_b$, simultaneously for every $t \in \{1, \dots, T+1\}$,

$$\left\| \sum_{s=1}^{t-1} \varepsilon_s a_s \right\|_{V_t^{-1}} \leq \sigma \sqrt{2 \log\left(\frac{\det(V_t)^{1/2}}{\det(\lambda I_d)^{1/2}\, \delta_b} \right)}.$$

This event implies the desired confidence bound at the random time t_b whenever batch b is realized. Hence $\mathbb{P}(\mathcal{E}_b) \geq 1 - \delta_b$; no conditioning on the realized action sequence or batch grid is needed.

At a realized batch start, the deterministic rounding error is

$$\|\widetilde{\theta}_b - \widehat{\theta}_{t_b}\|_{V_{t_b}} \leq \frac{\alpha_b \sqrt{d}}{2} = \frac{d\beta_b}{\rho_b - 2\delta_b}.$$

Consequently, on $\mathcal{E}_b$,

$$\|\widetilde{\theta}_b - \theta^\star\|_{V_{t_b}} \leq \beta_b \left(1 + \frac{d}{\rho_b - 2\delta_b} \right).$$

Global Lipschitzness of μ and Cauchy–Schwarz then give, simultaneously for every $a \in \mathbb{R}^d$,

$$\begin{aligned} |\mu(a^\top \widetilde{\theta}_b) - \mu(a^\top \theta^\star)| &\leq k_\mu \|a\|_{V_{t_b}^{-1}} \|\widetilde{\theta}_b - \theta^\star\|_{V_{t_b}} \\ &\leq \widetilde{\beta}_b \|a\|_{V_{t_b}^{-1}}. \end{aligned}$$

Finally, a union bound over the deterministic set of batch indices $b \in \{0, \dots, B-1\}$ yields

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{b=0}^{B-1} \mathbb{P}(\mathcal{E}_b^c) \leq \sum_{b=0}^{B-1} \delta_b.$$

□

Lemma E.7 (Regret bound for RepGLMUCB on a realized determinant batch grid). *Run RepGLMUCB (Algorithm 5) with parameters (λ, T, B, q) and budgets $(\delta_b, \rho_b)_{b=0}^{B-1}$. Let the realized batch grid be*

$$1 = t_0 < t_1 < \cdots < t_{B'} = T + 1, \qquad B' \le B.$$

Define

$$\widetilde{\beta}_{\max} := \max_{b \in \{0,\dots,B'-1\}} \widetilde{\beta}_b,$$

$$M_T := \max\left\{\sqrt{q},\ \sqrt{\frac{\det(V_{T+1})}{q^{B-1}\det(\lambda I_d)}}\right\}.$$

Then with probability at least $1 - \sum_{b=0}^{B-1} \delta_b$,

$$\mathrm{Reg}_{\mathrm{GLB}}(T) \le \max\{1, 2\widetilde{\beta}_{\max}\}\, M_T \sqrt{2TD_T}.$$

Proof of Lemma E.7. Let

$$\mathcal{E} := \bigcap_{b=0}^{B-1} \mathcal{E}_b,$$

where $\mathcal{E}_b$ is the event from Lemma E.6. By Lemma E.6,

$$\mathbb{P}(\mathcal{E}) \ge 1 - \sum_{b=0}^{B-1} \delta_b,$$

and on $\mathcal{E}$, for every realized batch $b \in \{0, \dots, B'-1\}$ and every $a \in \mathbb{R}^d$,

$$|\mu(a^\top \widetilde{\theta}_b) - \mu(a^\top \theta^\star)| \le \widetilde{\beta}_b \|a\|_{V_{t_b}^{-1}}.$$

Fix the event $\mathcal{E}$ in what follows. We prove the regret bound in three steps.

Step 1: Reduce the regret to stale batch-start norms. For each round $t \in [T]$, let $b(t) \in \{0, \dots, B'-1\}$ denote the realized batch index such that $t \in [t_{b(t)}, t_{b(t)+1} - 1]$. Let

$$a_t^\star \in \arg\max_{a \in A_t} \mu(a^\top \theta^\star)$$

be an optimal action at round t. Since the algorithm uses the batch-start parameter $\widetilde{\theta}_{b(t)}$ and radius $\widetilde{\beta}_{b(t)}$ throughout batch $b(t)$, the chosen action a_t satisfies

$$\begin{aligned}\mu(a_t^\top \widetilde{\theta}_{b(t)}) + \widetilde{\beta}_{b(t)} \|a_t\|_{V_{t_{b(t)}}^{-1}} \\ \ge \mu((a_t^\star)^\top \widetilde{\theta}_{b(t)}) + \widetilde{\beta}_{b(t)} \|a_t^\star\|_{V_{t_{b(t)}}^{-1}}.\end{aligned}$$

Moreover, on $\mathcal{E}$,

$$\mu((a_t^\star)^\top \theta^\star) \le \mu((a_t^\star)^\top \widetilde{\theta}_{b(t)}) + \widetilde{\beta}_{b(t)} \|a_t^\star\|_{V_{t_{b(t)}}^{-1}},$$

$$\mu(a_t^\top \theta^\star) \ge \mu(a_t^\top \widetilde{\theta}_{b(t)}) - \widetilde{\beta}_{b(t)} \|a_t\|_{V_{t_{b(t)}}^{-1}}.$$

Subtracting the two displays gives

$$\mu((a_t^\star)^\top \theta^\star) - \mu(a_t^\top \theta^\star) \le 2\widetilde{\beta}_{b(t)} \|a_t\|_{V_{t_{b(t)}}^{-1}}.$$

Since mean rewards lie in $[0, 1]$, the instantaneous regret is always at most 1. Therefore

$$\mu((a_t^\star)^\top \theta^\star) - \mu(a_t^\top \theta^\star) \le \min\left\{1,\ 2\widetilde{\beta}_{b(t)} \|a_t\|_{V_{t_{b(t)}}^{-1}}\right\}.$$

Since $\widetilde{\beta}_{b(t)} \le \widetilde{\beta}_{\max}$, we obtain

$$\mu((a_t^\star)^\top \theta^\star) - \mu(a_t^\top \theta^\star) \le \min\left\{1,\ 2\widetilde{\beta}_{\max}\|a_t\|_{V_{t_{b(t)}}^{-1}}\right\}.$$

Now for any $\kappa > 0$ and any $u \ge 0$,

$$\min\{1, \kappa u\} \le \max\{1, \kappa\} \min\{1, u\}.$$

Applying this with $\kappa = 2\widetilde{\beta}_{\max}$ and $u = \|a_t\|_{V_{t_{b(t)}}^{-1}}$ yields

$$\mu((a_t^\star)^\top \theta^\star) - \mu(a_t^\top \theta^\star) \le \max\{1, 2\widetilde{\beta}_{\max}\} \min\{1, \|a_t\|_{V_{t_{b(t)}}^{-1}}\}.$$

Summing over t gives

$$\mathrm{Reg}_{\mathrm{GLB}}(T) \le \max\{1, 2\widetilde{\beta}_{\max}\} \sum_{t=1}^{T} \min\{1, \|a_t\|_{V_{t_{b(t)}}^{-1}}\}.$$

Step 2: Compare stale norms to current norms using determinant growth. We show that for every round $t \in [T]$,

$$\|a_t\|_{V_{t_{b(t)}}^{-1}} \le M_T \|a_t\|_{V_t^{-1}}.$$

We split into two cases.

If $B' < B$, then the batch budget is not exhausted. Hence for every realized batch $b \in \{0, \dots, B'-1\}$ and every $t \in [t_b, t_{b+1} - 1]$, the next batch has not yet been triggered by time t, so by the determinant batching rule,

$$\det(V_t) \le q \det(V_{t_b}).$$

Also $V_t \succeq V_{t_b}$. Let

$$M := V_{t_b}^{-1/2} V_t V_{t_b}^{-1/2} \succeq I_d, \qquad u := V_{t_b}^{-1/2} a_t.$$

Then

$$\|a_t\|_{V_{t_b}^{-1}}^2 = \|u\|_2^2, \qquad \|a_t\|_{V_t^{-1}}^2 = u^\top M^{-1} u.$$

Diagonalizing M with eigenvalues $\lambda_1, \dots, \lambda_d \ge 1$ and writing u in the eigenbasis gives

$$\|u\|_2^2 \le \left(\max_i \lambda_i\right) u^\top M^{-1} u.$$

Since each $\lambda_i \ge 1$,

$$\max_i \lambda_i \le \prod_{i=1}^{d} \lambda_i = \det(M) = \frac{\det(V_t)}{\det(V_{t_b})}.$$

Hence

$$\|a_t\|_{V_{t_b}^{-1}}^2 \le \frac{\det(V_t)}{\det(V_{t_b})} \|a_t\|_{V_t^{-1}}^2 \le q \|a_t\|_{V_t^{-1}}^2,$$

and so

$$\|a_t\|_{V_{t_b}^{-1}} \le \sqrt{q}\, \|a_t\|_{V_t^{-1}} \le M_T \|a_t\|_{V_t^{-1}}.$$

Now consider the case $B' = B$. For every batch $b \in \{0, \dots, B-2\}$, the trigger fired at t_{b+1}, so

$$\det(V_{t_{b+1}}) > q \det(V_{t_b}).$$

Multiplying these inequalities over $b = 0, \dots, B-2$ gives

$$\det(V_{t_{B-1}}) \ge q^{B-1} \det(V_{t_0}) = q^{B-1} \det(\lambda I_d).$$

For the non-final batches $b \in \{0, \dots, B-2\}$, exactly the same argument as above gives

$$\|a_t\|_{V_{t_b}^{-1}} \le \sqrt{q}\,\|a_t\|_{V_t^{-1}} \le M_T \|a_t\|_{V_t^{-1}} \qquad \text{for all } t \in [t_b, t_{b+1}-1].$$

For the final batch $b = B-1$, if $t \in [t_{B-1}, T]$, then $V_t \preceq V_{T+1}$, and therefore

$$\frac{\det(V_t)}{\det(V_{t_{B-1}})} \le \frac{\det(V_{T+1})}{q^{B-1}\det(\lambda I_d)}.$$

Applying the same determinant-ratio norm comparison as above with $A = V_t$ and $B = V_{t_{B-1}}$ yields

$$\begin{aligned}
\|a_t\|_{V_{t_{B-1}}^{-1}} &\le \|a_t\|_{V_t^{-1}} \sqrt{\frac{\det(V_t)}{\det(V_{t_{B-1}})}} \\
&\le \|a_t\|_{V_t^{-1}} \sqrt{\frac{\det(V_{T+1})}{q^{B-1}\det(\lambda I_d)}} \\
&\le M_T \|a_t\|_{V_t^{-1}}.
\end{aligned}$$

Thus in all cases,

$$\|a_t\|_{V_{t_{b(t)}}^{-1}} \le M_T \|a_t\|_{V_t^{-1}}.$$

Since $M_T \ge 1$, it follows that

$$\min\{1, \|a_t\|_{V_{t_{b(t)}}^{-1}}\} \le M_T \min\{1, \|a_t\|_{V_t^{-1}}\}.$$

Substituting into the bound from Step 1 gives

$$\mathrm{Reg}_{\mathrm{GLB}}(T) \le \max\{1, 2\widetilde{\beta}_{\max}\}\, M_T \sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}\}.$$

Step 3: Apply the elliptical potential bound. Using $V_{t+1} = V_t + a_t a_t^\top$ and the matrix determinant lemma,

$$\det(V_{t+1}) = \det(V_t)\big(1 + \|a_t\|_{V_t^{-1}}^2\big).$$

Therefore

$$\log\left(\frac{\det(V_{T+1})}{\det(V_1)}\right) = \sum_{t=1}^{T} \log\Big(1 + \|a_t\|_{V_t^{-1}}^2\Big).$$

For every $u \ge 0$, we have $\log(1+u) \ge \frac{1}{2}\min\{u, 1\}$. Applying this with $u = \|a_t\|_{V_t^{-1}}^2$ yields

$$\sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}^2\} \le 2\log\left(\frac{\det(V_{T+1})}{\det(V_1)}\right) = 2D_T,$$

since $V_1 = \lambda I_d$. Finally, by Cauchy–Schwarz,

$$\begin{aligned}
\sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}\} &\le \sqrt{T \sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}\}^2} \\
&\le \sqrt{T \sum_{t=1}^{T} \min\{1, \|a_t\|_{V_t^{-1}}^2\}} \\
&\le \sqrt{2TD_T}.
\end{aligned}$$

Combining this with the bound from Step 2 gives

$$\mathrm{Reg}_{\mathrm{GLB}}(T) \le \max\{1, 2\widetilde{\beta}_{\max}\}\, M_T \sqrt{2TD_T}.$$

□

Lemma E.8 (Trajectory ρ-replicability of RepGLMUCB). *Fix a batch budget B and budgets $(\delta_b, \rho_b)_{b=0}^{B-1}$ with $\rho_b > 2\delta_b$. Consider two runs $k \in \{1, 2\}$ of RepGLMUCB (Algorithm 5) on the same obliviously fixed action-set sequence $(A_t)_{t=1}^T$, with independent environment randomness. The runs share vectors $v_b \sim \mathrm{Unif}([0,1)^d)$, independent across batches and independent of the environment randomness. At call b, run k uses $u_b^{(k)} = \alpha_b^{(k)} v_b$. Let $(a_t^{(k)})_{t=1}^T$ denote the resulting action sequences. Then*

$$\mathbb{P}\Big(\forall t \in [T] : a_t^{(1)} = a_t^{(2)}\Big) \geq 1 - \sum_{b=0}^{B-1} \rho_b.$$

Proof of Lemma E.8. For each run k and each $b \in \{0, \dots, B-1\}$, let $t_b^{(k)}$ be its batch-b start time, with $t_b^{(k)} = \infty$ if that batch is not realized. Define

$$\mathcal{M}_b := \Big\{ t_b^{(1)} = t_b^{(2)} \leq T, \quad a_s^{(1)} = a_s^{(2)} \text{ for every } s < t_b^{(1)} \Big\}.$$

Since $t_0^{(1)} = t_0^{(2)} = 1$, $\mathcal{M}_0$ holds surely. Let A_b be the event that both batch-b calls occur and their rounded estimators disagree.

Step 1: Agreement of the estimators preserves the batch trajectory. On $\mathcal{M}_b$, the two runs have a common batch start t_b, a common design matrix V_{t_b}, and identical confidence radii and grid widths. If their rounded estimators also agree, both runs maximize the same function

$$a \mapsto \mu(a^\top \widetilde{\theta}_b) + \widetilde{\beta}_b \|a\|_{V_{t_b}^{-1}}$$

over the same action set A_t, with the same deterministic tie-breaking rule. Their actions therefore agree throughout the batch. Their design updates and determinant triggers also agree, so the next batch starts at the same time in both runs, unless the horizon has already been reached.

Step 2: Bound disagreement on the matching-history event. For each run, let $\mathcal{E}_b^{(k)}$ be the event from Lemma E.6. That lemma applies separately to the two runs and gives

$$\mathbb{P}\Big((\mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)})^c\Big) \leq 2\delta_b.$$

On $\mathcal{M}_b$, write

$$z_b^{(k)} := V_{t_b}^{1/2} \widehat{\theta}_{t_b}^{(k)}, \qquad \alpha_b := \frac{2\beta_b \sqrt{d}}{\rho_b - 2\delta_b}.$$

On $\mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}$,

$$\begin{aligned} \|z_b^{(1)} - z_b^{(2)}\|_2 &\leq \|\widehat{\theta}_{t_b}^{(1)} - \theta^\star\|_{V_{t_b}} + \|\widehat{\theta}_{t_b}^{(2)} - \theta^\star\|_{V_{t_b}} \\ &\leq 2\beta_b. \end{aligned}$$

Let $\mathcal{H}_b$ contain both runs' histories immediately before their respective batch-b rounding operations, including all earlier rounding randomness. For a run that does not reach batch b, use its history through time T. These histories, the events $\mathcal{M}_b$ and $\mathcal{E}_b^{(k)}$, and the unrounded estimators and grid widths are determined without using v_b. Conditional on $\mathcal{H}_b$, v_b remains uniform and independent. On $\mathcal{M}_b$, both runs consequently use the same shift $u_b = \alpha_b v_b$, uniform on $[0, \alpha_b)^d$. The shifted-grid argument therefore yields, on $\mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}$,

$$\begin{aligned} \mathbb{P}(A_b \mid \mathcal{H}_b) &\leq \frac{\|z_b^{(1)} - z_b^{(2)}\|_1}{\alpha_b} \\ &\leq \frac{\sqrt{d}\, \|z_b^{(1)} - z_b^{(2)}\|_2}{\alpha_b} \\ &\leq \rho_b - 2\delta_b. \end{aligned}$$

Taking expectations conditional on these histories gives

$$\begin{aligned}\mathbb{P}\Big(A_b \cap \mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}\Big) &= \mathbb{E}\Big[\mathbf{1}_{\mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}} \mathbb{P}(A_b \mid \mathcal{H}_b)\Big] \\ &\le \rho_b - 2\delta_b.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{P}(A_b \cap \mathcal{M}_b) &\le \mathbb{P}\Big((\mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)})^c\Big) \\ &\quad + \mathbb{P}\Big(A_b \cap \mathcal{M}_b \cap \mathcal{E}_b^{(1)} \cap \mathcal{E}_b^{(2)}\Big) \\ &\le 2\delta_b + (\rho_b - 2\delta_b) = \rho_b.\end{aligned}$$

Step 3: Union bound over batches. If an action disagreement occurs, consider its first round. All earlier actions, design matrices, and determinant triggers agree. Thus the batch containing that round starts at a common time in both runs, and $\mathcal{M}_b$ holds. By Step 1, its rounded estimators must disagree. Hence

$$\Big\{\exists t \in [T] : a_t^{(1)} \neq a_t^{(2)}\Big\} \subseteq \bigcup_{b=0}^{B-1} (A_b \cap \mathcal{M}_b).$$

A union bound and Step 2 give

$$\mathbb{P}\Big(\exists t \in [T] : a_t^{(1)} \neq a_t^{(2)}\Big) \le \sum_{b=0}^{B-1} \rho_b,$$

which proves the claim. □

Proof of Theorem E.5. Using $\rho_b = \rho/B$, Lemma E.8 immediately gives the replicability guarantee:

$$\mathbb{P}\Big(\forall t \in [T] : a_t^{(1)} = a_t^{(2)}\Big) \ge 1 - \sum_{b=0}^{B-1} \rho_b = 1 - \rho.$$

It remains to prove the regret bound. Since $\delta_b = \delta/B$ for all b, we have

$$\sum_{b=0}^{B-1} \delta_b = \delta,$$

and therefore Lemma E.7 applies with probability at least $1 - \delta$.

We first control D_T. Since

$$V_{T+1} = \lambda I_d + \sum_{t=1}^{T} a_t a_t^\top$$

and $\|a_t\|_2 \le L$ for all t, the standard determinant bound gives

$$\det(V_{T+1}) \le \det(\lambda I_d) \left(1 + \frac{TL^2}{\lambda d}\right)^d.$$

Therefore

$$D_T = \log\left(\frac{\det(V_{T+1})}{\det(\lambda I_d)}\right) \le d \log\left(1 + \frac{TL^2}{\lambda d}\right).$$

Next, by the choice

$$q = \left(1 + \frac{TL^2}{\lambda d}\right)^{d/B},$$

we have

$$\begin{aligned} M_T &= \max\left\{\sqrt{q},\ \sqrt{\frac{\det(V_{T+1})}{q^{B-1}\det(\lambda I_d)}}\right\} \\ &\le \max\left\{\left(1+\frac{TL^2}{\lambda d}\right)^{d/(2B)},\ \sqrt{\frac{\left(1+\frac{TL^2}{\lambda d}\right)^{d}}{\left(1+\frac{TL^2}{\lambda d}\right)^{d(B-1)/B}}}\right\} \\ &= \left(1+\frac{TL^2}{\lambda d}\right)^{d/(2B)}. \end{aligned}$$

Since

$$B = \left\lceil d\log\left(1+\frac{TL^2}{\lambda d}\right)\right\rceil,$$

it follows that

$$\frac{d}{2B}\log\left(1+\frac{TL^2}{\lambda d}\right) \le \frac{1}{2},$$

and hence

$$M_T \le e^{1/2}.$$

We now bound $\widetilde{\beta}_{\max}$. For every realized batch $b \in \{0,\dots,B'-1\}$,

$$\begin{aligned} \widetilde{\beta}_b &= k_\mu \beta_b \left(1+\frac{d}{\rho_b - 2\delta_b}\right), \\ \beta_b &= \frac{1}{\underline{c}_\mu}\left(\sigma\sqrt{2\log\left(\frac{\det(V_{t_b})^{1/2}}{\det(\lambda I_d)^{1/2}\,\delta_b}\right)} + \sqrt{\lambda}\,S\right). \end{aligned}$$

Since $t_b \le T+1$, $\delta_b = \delta/B$, and $\det(V_{t_b}) \le \det(V_{T+1})$, we have

$$\begin{aligned} \beta_b &\le \frac{1}{\underline{c}_\mu}\left(\sigma\sqrt{2\log\left(\left(1+\frac{TL^2}{\lambda d}\right)^{d/2}\frac{B}{\delta}\right)} + \sqrt{\lambda}\,S\right) \\ &= \frac{1}{\underline{c}_\mu}\left(\sigma\sqrt{d\log\left(1+\frac{TL^2}{\lambda d}\right) + 2\log\left(\frac{B}{\delta}\right)} + \sqrt{\lambda}\,S\right). \end{aligned}$$

Also, since $\rho_b = \rho/B$, $\delta_b = \delta/B$, and $\rho > 3\delta$,

$$\rho_b - 2\delta_b = \frac{\rho - 2\delta}{B} \ge \frac{\rho}{3B},$$

so

$$1+\frac{d}{\rho_b - 2\delta_b} \le 1 + \frac{3dB}{\rho} \le 3\left(1+\frac{dB}{\rho}\right).$$

Combining the last two displays yields

$$\widetilde{\beta}_{\max} \le \frac{3k_\mu}{\underline{c}_\mu}\left(\sigma\sqrt{d\log\left(1+\frac{TL^2}{\lambda d}\right) + 2\log\left(\frac{B}{\delta}\right)} + \sqrt{\lambda}\,S\right)\left(1+\frac{dB}{\rho}\right).$$

Absorbing the factor of 2 inside the logarithm into a universal constant, we obtain

$$\widetilde{\beta}_{\max} \le C_1\frac{k_\mu}{\underline{c}_\mu}\left(\sigma\sqrt{d\log\left(1+\frac{TL^2}{\lambda d}\right) + \log\left(\frac{B}{\delta}\right)} + \sqrt{\lambda}\,S\right)\left(1+\frac{dB}{\rho}\right)$$

for a universal constant $C_1 > 0$.

Substituting the bounds on D_T, M_T, and $\widetilde{\beta}_{\max}$ into Lemma E.7, and absorbing universal constants, gives

$$\begin{aligned}\mathrm{Reg}_{\mathrm{GLB}}(T) &\le \max\{1, 2\widetilde{\beta}_{\max}\}\, M_T \sqrt{2TD_T} \\ &\le C \max\left\{1,\ \frac{k_\mu}{c_\mu}\left(\sigma\sqrt{d\log\left(1+\frac{TL^2}{\lambda d}\right)+\log\left(\frac{B}{\delta}\right)}+\sqrt{\lambda}\,S\right)\left(1+\frac{dB}{\rho}\right)\right\}\sqrt{T\,d\log\left(1+\frac{TL^2}{\lambda d}\right)},\end{aligned}$$

which is exactly the claimed regret bound. □