Anytime Model Selection in Linear Bandits

Parnian Kassraie[1] **Nicolas Emmenegger**[1] **Andreas Krause**[1] **Aldo Pacchiano**[2,3]
[1]ETH Zurich [2]Broad Institute of MIT and Harvard [3]Boston University
{pkassraie, nicolaem, krausea}@ethz.ch pacchian@bu.edu

Abstract

Model selection in the context of bandit optimization is a challenging problem, as it requires balancing exploration and exploitation not only for action selection, but also for model selection. One natural approach is to rely on online learning algorithms that treat different models as experts. Existing methods, however, scale poorly ($\mathrm{poly}M$) with the number of models M in terms of their regret. Our key insight is that, for model selection in linear bandits, we can emulate full-information feedback to the online learner with a favorable bias-variance trade-off. This allows us to develop ALExp, which has an exponentially improved ($\log M$) dependence on M for its regret. ALExp has anytime guarantees on its regret, and neither requires knowledge of the horizon n, nor relies on an initial purely exploratory stage. Our approach utilizes a novel time-uniform analysis of the Lasso, establishing a new connection between online learning and high-dimensional statistics.

1 Introduction

When solving bandit problems or performing Bayesian optimization, we need to commit to a reward model *a priori*, based on which we estimate the reward function and build a policy for selecting the next action. In practice, there are many ways to model the reward by considering different feature maps or hypothesis spaces, e.g., for optimizing gene knockouts [Gonzalez et al., 2015, Pacchiano et al., 2022] or parameter estimation in biological dynamical systems [Ulmasov et al., 2016, Imani et al., 2019]. It is not known a priori which model is going to yield the most sample efficient bandit algorithm, and we can only select the right model as we gather empirical evidence. This leads us to ask, can we perform adaptive model selection, while simultaneously optimizing for reward?

In an idealized setting with no sampling limits, given a model class of size M, we could initialize M bandit algorithms (a.k.a agents) in parallel, each using one of the available reward models. Then, as the algorithms run, at every step we can select the most promising agent, according to the cumulative rewards that are obtained so far. Model selection can then be cast into an online optimization problem on a M-dimensional probability simplex, where the probability of selecting an agent is dependent on its cumulative reward, and the optimizer seeks to find the distribution with the best return in hindsight. This approach is impractical for large model classes, since at every step, it requires drawing M different samples in parallel from the environment so that the reward for each agent is realized.

In realistic applications of bandit optimization, we can only draw *one* sample at a time, and so we need to design an algorithm which allocates more samples to the agents that are deemed more promising. Prior work [e.g., Maillard and Munos, 2011, Agarwal et al., 2017] propose to run a single meta algorithm which interacts with the environment by first selecting an agent, and then selecting an action according to the suggestion of that agent. The online model selection problem can still be emulated in this setup, however this time the optimizer receives partial feedback, coming from only one agent. Consequently, many agents need to be queried, and the overall regret scales with $\mathrm{poly}M$, again restricting this approach to small model classes. In fact, addressing the limited scope of such algorithms, Agarwal et al. [2017] raise an open problem on the feasibility of obtaining a $\log M$ dependency for the regret.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

We show that this rate is achievable, in the particular case of linearly parametrizable rewards. We develop a technique to "hallucinate" the reward for every agent that was not selected, and run the online optimizer with emulated full-information feedback. This allows the optimizer to assess the quality of the agents, without ever having queried them. As a result, our algorithm ALEXP, satisfies a regret rate $O(\max\{\sqrt{n \log^3 M}; n^{3/4}\sqrt{\log M}\})$, with high probability, simultaneously for all $n \geq 1$ (Theorem 1). Our key idea, leading to the $\log M$ dependency, is to employ the Lasso as a low-variance online regression oracle, and estimate the reward for the agents that were not chosen. This trick is made possible through our novel time-uniform analysis of online Lasso regression (Theorem 3). Consequently, ALEXP is horizon-independent, and explores adaptively without requiring an initial exploration stage. Empirically we find that ALEXP consistently outperforms prior work across a range of environments.

Table 1: Overview of literature on online model selection for bandit optimization

	MS Technique	Regret	MS Guarantee	adaptive exploration	anytime
Sparse Linear Bandits	Lasso	$\log M$	✗	✗	✗
MS for Black-Box Bandits	OMD with bandit feedback	$\text{poly}M$	✓	✓	✗
MS for Linear Bandits (Ours)	EXP4 with full-information	$\log M$	✓	✓	✓

2 Related Work

Online Model selection (MS) for bandits considers combining a number of agents in a master algorithm, with the goal of performing as well as the best agent [Maillard and Munos, 2011, Agarwal et al., 2017, Pacchiano et al., 2020, Luo et al., 2022]. This literature operates on black-box model classes of size M and uses variants of Online Mirror Descent (OMD) to sequentially select the agents. The optimizer operates on importance-weighted estimates of the agents' rewards, which has high variance ($\text{poly}M$) and is non-zero only for the selected agent. Effectively, the optimizer receives partial (a.k.a. bandit) feedback and agents are at risk of starvation, since at every step only the selected agent gets the new data point. These works assume knowledge of the horizon, however, we suspect that this may be lifted with a finer analysis of OMD.

Sparse linear bandits use sparsity-inducing methods, often Lasso [Tibshirani, 1996], for estimating the reward in presence of many features, as an alternative to model selection. Early results are in the data-rich regime where the stopping time is known and larger than the number of models, and some suffer from $\text{poly}M$ regret dependency [e.g., Abbasi-Yadkori et al., 2012]. Recent efforts often consider the contextual case, where at every step only a finite stochastic subset of the action domain is presented to the agent, and that the distribution of these points is i.i.d. and sufficiently diverse [Li et al., 2022, Bastani and Bayati, 2020, Kim and Paik, 2019, Oh et al., 2021, Cella and Pontil, 2021]. We do not rely on such assumptions. Most sparse bandit algorithms either start with a purely exploratory phase [Kim and Paik, 2019, Li et al., 2022, Hao et al., 2020, Jang et al., 2022], or rely on a priori scheduled exploration [Bastani and Bayati, 2020]. The exploration budget is set according to the horizon n. Therefore, such algorithms inherently require the knowledge of n and can be made anytime only via the doubling trick [Auer et al., 1995]. Table 2 presents an in-depth overview.

ALEXP inherits the best of both worlds (Table 1): its regret enjoys the $\log M$ dependency of sparse linear bandits even on compact domains, and it has adaptive probabilistic exploration with anytime guarantees. In contrast to prior literature, we perform model selection with an online optimizer (EXP4), which hallucinates full-information feedback using a low-variance Lasso estimator instead of importance-weighted estimates. Moreover, our anytime approach lifts the horizon dependence and the exploration requirement of sparse linear bandits.

Our work is inspired by and contributes to the field of Learning with Expert Advice, which analyzes incorporating the advice of M oblivious (non-adaptive) experts, with bandit or full-information feedback [Haussler et al., 1998, Auer et al., 2002b, McMahan and Streeter, 2009]. The idea of employing an online optimizer for learning stems from this literature, and has been used in various applications of online learning [Foster et al., 2017, Singla et al., 2018, Muthukumar et al., 2019, Karimi et al., 2021, Liu et al., 2022]. In particular, we are inspired by Foster and Rakhlin [2020] and Moradipari et al. [2022], who apply EXP4 to least squares estimates, for arm selection in K-armed

contextual bandits. However, their algorithms are not anytime, and due to the choice of estimator, the corresponding regret scales with $\mathcal{O}(\min(\sqrt{dM}, \sqrt{K \log M}))$.

3 Problem Setting

We consider a bandit problem where a learner interacts with the environment in rounds. At step t the learner selects an action $x_t \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^{d_0}$ is a compact domain and observes a noisy reward $y_t = r(x_t) + \varepsilon_t$ such that ε_t is an i.i.d. zero-mean sub-Gaussian variable with parameter σ^2. We assume the reward function $r : \mathcal{X} \to \mathbb{R}$ is linearly parametrizable by some unknown feature map, and that the model class $\{\phi_j : \mathbb{R}^{d_0} \to \mathbb{R}^d, j = 1, \ldots, M\}$ contains the set of plausible feature maps. We consider the setting where M can be very large, and while the set $\{\phi_j\}$ may include misspecified feature maps, it contains at least one feature map that represents the reward function, i.e., there exists $j^\star \in [M]$ such that $r(\cdot) = \theta_{j^\star}^\top \phi_{j^\star}(\cdot)$. We assume the image of ϕ_j spans $\mathbb{R}^d$, and no two feature maps are linearly dependent, i.e. for any $j \neq j' \in [M]$, there exists no $\alpha \in \mathbb{R}$ such that $\phi_j(\cdot) = \alpha \phi_{j'}(\cdot)$. This assumption, which is satisfied by design in practice, ensures that the features are not ill-posed and we can explore in all relevant directions. We assume that the concatenated feature map $\phi(x) := (\phi_1(x), \ldots, \phi_M(x))$ is normalized $\|\phi(x)\| \leq 1$ for all $x \in \mathcal{X}$ and that $\|\theta_{j^\star}\| \leq B$, which implies $|r(x)| \leq B$ for all $x \in \mathcal{X}$.

We will model this problem in the language of model selection where a meta algorithm aims to optimize the unknown reward function by relying on a number of base learners. In order to interact with the environment the meta algorithm selects an agent that in turn selects an action. In our setting we thinking of each of these M feature maps as controlled by a base agent running its own algorithm. Base agent j uses the feature map ϕ_j for modeling the reward. At step t of the bandit problem, each agent j is given access to the full history $H_{t-1} := \{(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\}$, and uses it to locally estimate the reward as $\hat{r}_{t,j}(\cdot) = \hat{\theta}_{t-1,j}^\top \phi_j(\cdot)$, where $\hat{\theta}_{t-1,j} \in \mathbb{R}^d$ is the estimated coefficients vector. The agent then uses this estimate to develop its action selection policy $p_{t,j} \in \mathcal{M}(\mathcal{X})$. Here, $\mathcal{M}$ denotes the space of probability measures defined on $\mathcal{X}$. The condition on existence of $j^\star$ will ensure that there is at least one agent which is using a correct model for the reward, and thereby can solve the bandit problem if executed in isolation. We refer to agent $j^\star$ as the oracle agent.

Our goal is to find a sample-efficient strategy for iterating over the agents, such that their suggested actions maximize the cumulative reward $\sum_{t=1}^n r(x_t)$ achieved over any horizon $n \geq 1$. This is equivalent to minimizing the cumulative regret $R(n) = \sum_{t=1}^n r(x^\star) - r(x_t)$, where $x^\star$ is a global maximizer of the reward function. We neither know $x^\star$, nor assume knowledge of it.

4 Method

As warm-up, consider an example with deterministic agents, i.e., where $p_{t,j}$ is a Dirac measure on a specific action $x_{t,j}$. Suppose it was practically feasible to draw the action suggested by every agent and observe the corresponding reward vector $r_t = (r_{t,j})_{j=1}^M$. In this case, model selection becomes a full-information online optimization problem, and we can design a minimax optimal algorithm as follows. We assign a probability distribution $q_t = (q_{t,j})_{j=1}^M$ to the models, and update it such that the overall average return $\sum_{t=1}^n q_t^\top r_t$ is competitive to the best agent's average return $\max_j \sum_{t=1}^n r_{t,j}$. At every step, we update $q_{t+1,j} \propto \exp(\sum_{s=1}^t r_{s,j})$, since such exponential weighting is known to lead to an optimal solution for this classic online learning problem [Cesa-Bianchi and Lugosi, 2006]. In our setting however, the agents are stochastic, and we do not have access to r_t vector.

We propose the Anytime EXPonential weighting algorithm based on Lasso reward estimates (ALEXP), summarized in Algorithm 1. At step t we first sample an agent j_t, and then sample an action x_t according to the agent's policy p_{t,j_t}. Let Δ_M denote the M-dimensional probability simplex. We maintain a probability distribution $q_t \in \Delta_M$ over the agents, and update it sequentially as we accumulate evidence on the performance of each agent. Ideally, we would have adjusted $q_{t,j}$ according to the average return of model j, that is, $\mathbb{E}_{x \sim p_{t,j}} r(x)$. However, since r is unknown, we estimate the average reward with some $\hat{r}_{t,j}$. We then update q_t for the next step via,

$$q_{t+1,j} = \frac{\exp(\eta_t \sum_{s=1}^t \hat{r}_{s,j})}{\sum_{i=1}^M \exp(\eta_t \sum_{s=1}^t \hat{r}_{s,i})}$$

for all $j = 1, \ldots, M$, where η_t is the learning rate, and controls the sensitivity of the updates. This rule allows us to imitate the full-information example that we mentioned above. By utilizing $\hat{r}_{t,j}$ and hallucinating feedback from all agents, we can reduce the probability of selecting a badly performing agent, without ever having sampled them (c.f. Fig. 4). It remains to design the estimator $\hat{r}_{t,j}$. We concatenate all feature maps, and, knowing that many features are redundant, use a sparsity inducing estimator over the resulting coefficients vector. Mainly, let $\theta = (\theta_1, \ldots, \theta_M) \in \mathbb{R}^{Md}$ be the concatenated coefficients vector. We then solve

$$\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^{Md}} L(\theta, H_t, \lambda_t) = \arg\min_{\theta \in \mathbb{R}^{Md}} \frac{1}{t} \|y_t - \Phi_t \theta\|_2^2 + 2\lambda_t \sum_{j=1}^{M} \|\theta_j\|_2 \tag{1}$$

where $\Phi_t = [\phi(x_s)]_{s \le t} \in \mathbb{R}^{t \times Md}$ is the feature matrix, $y_t \in \mathbb{R}^t$ is the concatenated reward vector, and λ_t is an adaptive regularization parameter. Problem (1) is the online variant of the group Lasso [Lounici et al., 2011]. The second term is the loss is the mixed $(2,1)$-norm of θ, which can be seen as the ℓ_1-norm of the vector $(\|\theta_1\|, \ldots, \|\theta_M\|) \in \mathbb{R}^M$. This norm induces sparsity at the group level, and therefore, the sub-vectors $\hat{\theta}_{t,j} \in \mathbb{R}^d$ that correspond to redundant feature maps are expected to be 0, i.e. the null vector. We then estimate the average return of each model by simply taking an expectation $\hat{r}_{t,j} = \mathbb{E}_{x \sim p_{t+1,j}}[\hat{\theta}_t^\top \phi(x)]$. This quantity is the average return of the agent's policy $p_{t+1,j}$, according to our Lasso estimator. In Section 5.2 we explain why the particular choice of Lasso is crucial for obtaining a $\log M$ rate for the regret.

For action selection, with probability $1 - \gamma_t$, we sample agent j with probability $q_{t,j}$ and draw $x_t \sim p_{t,j}$ as per suggestion of the agent. With probability γ_t, we choose the action according to some exploratory distribution $\mu \in \mathcal{M}(\mathcal{X})$ which aims to sample informative actions. This can be any design where $\mathrm{supp}(\mu) = \mathcal{X}$. We mix $p_{t,j}$ with μ, to collect sufficiently diverse data for model selection. We are not restricting the agents' policy, and therefore can not rely on them to explore adequately. In Theorem 1, we choose a decreasing sequence (γ_t) of 1 the probabilities of exploration at step t 1, since less exploration will be required as data accumulates. To conclude, the action selection policy of ALEXP is formally described as the mixture

$$p_t(x) = \gamma_t \mu(x) + (1 - \gamma_t) \sum_{j=1}^{M} q_{t,j}\, p_{t,j}(x).$$

5 Main Results

For the regret guarantee, we consider specific choices of base agents and exploratory distribution. Our analysis may be extended to include other policies, since ALEXP can be wrapped around any bandit agent that is described by some $p_{t,j}$, and allows for random exploration with any distribution μ.

Base Agents. We assume that the oracle agent has either a UCB [Abbasi-Yadkori et al., 2011] or a GREEDY [Auer et al., 2002a] policy, and all other agents are free to choose any arbitrary policy. Similar treatment can be applied to cases where the oracle uses other (sublinear) polices for solving linear or contextual bandits [e.g., Thompson, 1933, Kaufmann et al., 2012, Agarwal et al., 2014]. In either case, agent j at t calculates a ridge estimate of the coefficients vector based on the history H_t:

$$\hat{\theta}_{t,j^\star} := \arg\min_{\theta \in \mathbb{R}^d} \|y_t - \Phi_{t,j^\star}\theta\|_2^2 + \tilde{\lambda}\|\theta\|_2^2 = \left(\Phi_{t,j^\star}^\top \Phi_{t,j^\star} + \tilde{\lambda} I\right)^{-1} \Phi_{t,j^\star}^\top y_t.$$

Here $\Phi_{t,j^\star} \in \mathbb{R}^{t \times d}$ is the feature matrix, where each row is $\phi_{j^\star}(x_s)$ and $\tilde{\lambda}$ is the regularization constant. Then at step t, a GREEDY oracle suggests the action which maximizes the reward estimate $\hat{\theta}_{t,j^\star}^\top \phi_{j^\star}(x)$, and a UCB oracle queries $\arg\max_x u_{t,j^\star}(x)$ where $u_{t,j^\star}(\cdot)$ is the upper confidence bound that this agent calculates for $r(x)$. Proposition 21 shows that the sequence $(u_{t,j^\star})_{t \ge 1}$ is in fact an anytime valid upper bound for r over the entire domain.

Exploratory policy. Performance of ALEXP depends on the quality of the samples that μ suggests. The eigenvalues of the covariance matrix $\Psi(\mu; \phi) := \mathbb{E}_{x \sim \mu}\, \phi(x)\phi(x)^\top$ reflect how diverse the data is, and thus are a good indicator for data quality. van de Geer and Bühlmann [2009] present a survey on the notions of diversity defined based on these eigenvalues. Let $\lambda_{\min}(A)$ denote the minimum eigenvalue of a matrix A. Similar to Hao et al. [2020], we assume that μ is the maximizer of the problem below and present our regret bounds in terms of

$$C_{\min} = C_{\min}(\mathcal{X}; \phi) := \max_{\mu \in \mathcal{M}(\mathcal{X})} \lambda_{\min}(\Psi(\mu; \phi)); \tag{2}$$

Algorithm 1 ALEXP

Inputs: γ; $(\alpha_t; \beta_t; \varepsilon_t)$ for $t \geq 1$
Let $q_1 \leftarrow \text{Unif}(M)$ and initialize base agents $(p_{1;1}; \ldots; p_{1;M})$.
for $t = 1$ do
  Draw $\varepsilon_t \sim \text{Bernoulli}(\varepsilon_t)$. . Decide to explore or exploit
  if $\varepsilon_t = 1$ then
    Choose action x_t randomly according to π. . Explore
  else
    Draw $j_t \sim q_t$. . Select an agent
    Draw $x_t \sim p_{t;j_t}$. . Select the action suggested by the agent
  end if
  Observe $y_t = r(x_t) + \varepsilon_t$. . Receive reward
  $H_t = H_{t-1} \cup \{(x_t; y_t)\}$. . Append history
  $\hat{\beta}_t \leftarrow \arg\min L(\beta; H_t; \alpha_t)$. . Update the parameter estimate
  Report H_t to all agents, and get updated policies $p_{t+1;1}; \ldots; p_{t+1;M}$. . Update agents
  Update estimated average return of every agent

$$\hat{r}_{t;j} \leftarrow \mathbb{E}_{x \sim p_{t+1;j}}[\hat{\beta}_t^\top \phi(x)]; \quad j = 1; \ldots; M$$

  Update agent probabilities

$$q_{t+1;j} \leftarrow \frac{\exp(\eta_t \sum_{s=1}^{t} \hat{r}_{s;j})}{\sum_{i=1}^{M} \exp(\eta_t \sum_{s=1}^{t} \hat{r}_{s;i})}$$

end for

which is greater than zero under the conditions specified in our problem setting. Prior works in the sparse bandit literature all require a similar or stronger assumption of this kind, and Table 2 gives an overview. Alternatively, one can work with an arbitrary π e.g., $\text{Unif}(X)$, as long as $\phi_{\min}(\mathbb{E}[\phi(X)\phi(X)^\top])$ is bounded away from zero. Appendix C.1 reviews some configurations of $(\pi; X; \phi)$ that lead to a non-zero minimum eigenvalue, and Corollary 12 bounds the regret of ALEXP with uniform exploration.

For this choice of agents and exploratory distribution, Theorem 1 presents an informal regret bound. Here, we have used the $\tilde{O}$ notation, and only included the fastest growing terms. The inequality is made exact in Theorem 14, up to constant multiplicative factors.

Theorem 1 (Cumulative Regret of ALEXP, Informal). Let $\gamma \in (0; 1]$ and set β to be the maximizer of (2). Choose learning rate $\eta_t = O(C_{\min} t^{-1/2}/C(M; \gamma; d))$, exploration probability $\varepsilon_t = O(t^{-1/4})$ and Lasso regularization parameter $\alpha_t = O(C_{\min} t^{-1/2} C(M; \gamma; d))$, where

$$C(M; \gamma; d) = O\left(\sqrt{1 + \log(M/\gamma) + (\log\log d)_+} + \sqrt{d(\log(M/\gamma) + (\log\log d)_+)}\right).$$

Then ALEXP satisfies the cumulative regret

$$R(n) = O\left(n^{3/4} B + n^{3/4} C(M; \gamma; d) + C_{\min}^{-1} \sqrt{n} C(M; \gamma; d) \log M + C_{\min}^{-1/2} n^{5/8} \sqrt{d \log(n) + \log(1/\gamma)}\right)$$

simultaneously for all $n \geq 1$, with probability greater than $1 - \gamma$.

In this bound, the first term is the regret incurred at exploratory steps (when $\varepsilon_t = 1$), the second term is due to the estimation error of Lasso (i.e., $\|\hat{\beta}_t\|$), and the third term is the regret of the exponential weights sub-algorithm. The fourth term, is the regret bound for the oracle agent when run within the ALEXP framework. It does not depend on the agent's policy (greedy or optimistic), and is worse than the minimax optimal rate of $\sqrt{d \log n}$. This is because the oracle is suggesting actions based on the history H_t, which includes uninformative action-reward pairs queried by other, potentially misspecified, agents. In Corollary 12, we provide a regret bound independent of $C_{\min}$ for orthogonal feature maps, and show that the same $O(\max\{\sqrt{n \log^3 M}; n^{3/4}\sqrt{\log M}\})$ rate may be achieved even with the simple choice $\pi = \text{Unif}(X)$.

5.1 Proof Sketch

The regret is caused by two sources: selecting a sub-optimal agent, and an agent selecting a sub-optimal action. Accordingly, for any $j = 1, \dots, M$, we decompose the regret as

$$R(n) = \sum_{t=1}^{n} r(x^\star) - r(x_t) = \sum_{t=1}^{n} r(x^\star) - r_{t,j} + \sum_{t=1}^{n} r_{t,j} - r(x_t) \,. \tag{3}$$

The first term shows the cumulative regret of agent j, when run within ALEXP. The second term evaluates the received reward against the cumulative average reward of model j. We bound each term separately.

Virtual Regret. The first term $R_j(n) := \sum_{t=1}^{n} r(x^\star) - r_{t,j}$ compares the suggestion of agent j against the optimal action. We call it the virtual regret since the sequence $(x_{t,j})_{t=1}$ of the actions suggested by model j are not necessarily selected by the meta algorithm, unless $j = j^\star$. This regret is merely a technical tool, and not actually realized when running ALEXP. The virtual regret of the oracle agent may still be analyzed using standard techniques for linear bandits, e.g., Abbasi-Yadkori et al. [2011], however we need to adapt it to take into account a subtle difference: The confidence sequence of model $j^\star$ is constructed according to the true sequence of actions $(x_t)_{t=1}$, while its virtual regret is calculated based on the virtual sequence $(x_{t,j^\star})_{t=1}$, which the model suggests. The two sequences only match at the steps when model $j^\star$ is selected. Adapting the analysis of Abbasi-Yadkori et al. [2011] to this subtlety, we obtain in Lemma 15 that with probability greater than $1-\delta$, simultaneously for all $n \geq 1$,

$$R_{j^\star}(n) = O\left(n^{5/8} C_{\min}^{1/2} \sqrt{d \log(n) + \log(1/\delta)}\right) \,.$$

Model Selection Regret. The second term in (3) is the model selection regret $R_t(n; j) := \sum_{t=1}^{n} r_{t,j} - r(x_t)$, which evaluates the chosen actions by ALEXP against the suggestion of the jth agent. Our analysis relies on a careful decomposition of $R(n; j)$,

$$R(n;j) = \sum_{t=1}^{n} \Big[\underbrace{r_{t,j} - \hat{r}_{t,j}}_{\text{(I)}} + \underbrace{\hat{r}_{t,j} - \sum_{i=1}^{M} q_{t,i}\, \hat{r}_{t,i}}_{\text{(II)}} + \underbrace{\sum_{i=1}^{M} q_{t,i}\, \hat{r}_{t,i} - \sum_{i=1}^{M} q_{t,i}\, r_{t,i}}_{\text{(III)}} + \underbrace{\sum_{i=1}^{M} q_{t,i}\, r_{t,i} - r(x_t)}_{\text{(IV)}} \Big] \,.$$

We bound away each term in a modular manner, until we are left with the regret of the standard exponential weights algorithm. The terms (I) and (III) are controlled by the bias of the Lasso estimator, and are $O(n^{3/4} C(M; \delta; d))$ (Lemma 19). The last term (IV) is zero in expectation, and reflects the deviation of $r(x_t)$ from its mean. We observe that the summands form a bounded Martingale difference sequence, and their sum grows as $O(\sqrt{n})$ (Lemma 18). Term (II) is the regret of our online optimizer, which depends on the variance of the Lasso estimator. We bound this term with $O(\sqrt{n C(M; \delta; d) \log M})$, by first conducting a standard anytime analysis of exponential weights (Lemma 17), and then incorporating the anytime Lasso variance bound (Lemma 20). We highlight that neither of the above steps require assumptions about the base agents. Combining these steps, Lemma 16 establishes the formal bound on the model selection regret.

Anytime Lasso. We develop novel confidence intervals for Lasso with history-dependent data, which are uniformly valid over an unbounded time horizon. This result may be of independent interest in applications of Lasso for online learning or sequential decision making. Here, we use these confidence intervals to bound the bias and variance terms that appear in our treatment of the model selection regret. The width of the Lasso confidence intervals depends on the quality of feature matrix $_t$, often quantified by the restricted eigenvalue property [Bickel et al., 2009, van de Geer and Bühlmann, 2009, Javanmard and Montanari, 2014]:

Definition 2. For the feature matrix $_t \in R^{t \times dM}$ we define $(\ _t; s)$ for any $1 \leq s \leq M$ as

$$(\ _t; s) := \inf_{(J,b)} \sqrt{\frac{1}{t}}\, \frac{\| \ _t b \|_2}{\sqrt{\sum_{j \in J} \|b_j\|_2^2}}$$
$$\text{s.t. } b \in R^{dM} \setminus \{0\}, \quad \sum_{j \notin J} \|b_j\|_2 \leq 3 \sum_{j \in J} \|b_j\|_2, \ J \subseteq \{1, \dots, M\}, \ |J| \leq s.$$

Our analysis is in terms of this quantity, and Lemma 8 explains the connection between $\lambda_{\min}(V_t)$ and $C_{\min}$ as defined in (2), particularly that $\lambda_{\min}(V_t) > 2)$ is also positive with a high probability.

Theorem 3 (Anytime Lasso Confidence Sequences). Consider the data model $y_t = \langle \beta, x_t \rangle + \eta_t$ for all $t \geq 1$, where η_t is i.i.d. zero-mean sub-Gaussian noise, and $(x_t)_{t \geq 1}$ is $(\mathcal{F}_t)_{t \geq 1}$-predictable, where $\mathcal{F}_t := (x_1, \ldots, x_t, \eta_1, \ldots, \eta_{t-1})$. Then the solution of (1) guarantees

$$\mathbb{P}\left(\forall t \geq 1 : \|\hat{\beta}_t - \beta\|_2 \leq \frac{4\sqrt{10}\, \gamma_t}{\lambda_{\min}(V_t) > 2)} \right) \geq 1 - \delta$$

if the regularization parameter satisfies for all $t \geq 1$

$$\gamma_t \geq \frac{2}{\sqrt{t}} \sqrt{1 + \frac{12}{t}} \left(\log(2M/\delta) + (\log\log t)_+ \right) + \frac{5}{\sqrt{2}} \sqrt{d \left(\log(2M/\delta) + (\log\log t)_+ \right)}.$$

Our confidence bound enjoys the same rate as Lasso with fixed (offline) data, up to log log t factors. We prove this theorem by constructing a self-normalized martingale sequence based on the ℓ_2-norm of the empirical process error $\sum_{s=1}^{t} x_s \eta_s$. We then apply the "stitched" time-uniform boundary of Howard et al. [2021]. Appendix B elaborates on this technique. Previous work on sparse linear bandits also include analysis of Lasso in an online setting, where x_t is $\mathcal{F}_t$ measurable. Cella and Pontil [2021] imitate offline analysis and then apply a union bound over the time steps, which multiplies the width by $\log n$ and requires knowledge of the horizon. Bastani and Bayati [2020] also rely on knowledge of σ and employ a scalar-valued Bernstein inequality on ℓ_1-norm of the empirical process error, which inflates the width of the confidence sets by a factor of $\sqrt{d}$. We work directly on the ℓ_2-norm, and use a curved boundary for sub-Gamma martingale sequences, which according to Howard et al. [2021] is uniformly tighter than a Bernstein bound, especially for t small.

5.2 Discussion

In light of Theorem 1 and Theorem 3, we discuss some properties of ASLE.

Sparse EXP4. Our approach presents a new connection between online learning and high-dimensional statistics. The rule for updating the probabilities in ASLE.EXP is inspired by the exponential weighting for Exploration and Exploitation with Experts (EXP4) algorithm, which was proposed by Auer et al. [2002b] and has been extensively studied in the adversarial bandit and learning with expert advice literature [e.g., McMahan and Streeter, 2009, Beygelzimer et al., 2011]. EXP4 classically uses importance-weighted (IW) or ordinary least squares (LS) estimators to estimate the return, both of which are unbiased but high-variance choices [Bubeck et al., 2012]. In particular, in our linearly parametrized setting, the variance of IW and LS scales with M, which will lead to a $\mathrm{poly}(M)$ regret. However, it is known that introducing bias can be useful if it reduces the variance [Zimmert and Lattimore, 2022]. For instance, EXP3-IX [Kocák et al., 2014] and EXP4-IX [Neu, 2015] construct a biased IW estimator. Equivalently, others craft regularizers for the reward of the online optimizer, seeking to improve the bias-variance balance [e.g., Abernethy et al., 2008, Bartlett et al., 2008, Abernethy and Rakhlin, 2009, Bubeck et al., 2017, Lee et al., 2020, Zimmert and Lattimore, 2022]. A key technical observation in this work is that our online Lasso estimator helps EXP4 to achieve sublinear regret which depends logarithmically on M. This is due to the fact that while the estimator itself is M-dimensional, its bias squared and variance scale with $d \log M$. To the best of our knowledge, this work is first to instantiate the EXP4 algorithm with a sparse low-variance estimator.

Adaptive and Anytime. To estimate the reward, prior work on sparse bandits commonly emulate the Lasso analysis on offline data or on a martingale sequence with a known length [Hao et al., 2020, Bastani and Bayati, 2020]. These works require a long enough sequence of exploratory samples, and knowledge of the horizon to plan this sequence. ASLE.EXP removes both of these constraints, and presents a fully adaptive algorithm. Crucially, we employ the elegant martingale bounds of Howard et al. [2020] to present the first time-uniform analysis of the Lasso with history-dependent data (Theorem 3). This way we can use all the data points and explore with a probability which vanishes at a $O(t^{-1/4})$ rate. Our anytime confidence bound for Lasso, together with the horizon-independent analysis of the exponential weights algorithm, also allow ASLE.EXP to be stopped at any time with valid guarantees.

Rate Optimality. For $M = n$, we obtain a $O(\sqrt{n \log^3 M})$ regret, which matches the rate conjectured by Agarwal et al. [2017]. However, if M is comparable to or smaller, our regret scales with

Figure 1: ALEXP can model-select in both orthogonal and correlated classes ($M = 55$)

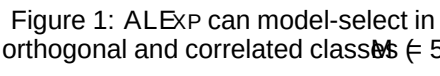
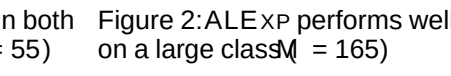


Figure 2: ALEXP performs well on a large class ($M = 165$)



$O(n^{3/4}\sqrt{\log M})$, and while it is still sublinear and scales logarithmically with M, the dependency on n is sub-optimal. This may be due to the conservative nature of our model selection analysis, during which we do not make assumptions about the dynamics of the base agents. Therefore, to ensure sufficiently diverse data for successful model selection, we need to occasionally choose exploratory actions with a vanishing probability of ε_t. We conjecture that this is avoidable, if we make more assumptions about the agents, e.g., that a sufficient number of agents can achieve sublinear regret if executed in isolation. Banerjee et al. [2023] show that the data collected by sublinear algorithms organically satisfies a minimum eigenvalue lowerbound, which may also be sufficient for model selection. We leave this as an open question for future work.

6 Experiments

Experiment Setup. We create a synthetic dataset based on our data model (Section 3), and choose the domain to be 1-dimensional, $\mathcal{X} = [-1, 1]$. As a natural choice of features, we consider the set of degree-p Legendre polynomials, since they form an orthonormal basis for $L^2(\mathcal{X})$ if p grows unboundedly. We construct each feature map, by choosing s different polynomials from this set, and therefore obtaining $M = \binom{p+1}{s}$ different models. More formally, we let $\phi_j(x) = (P_{j_1}(x), \ldots, P_{j_s}(x)) \in \mathbb{R}^s$ where $\{j_1, \ldots, j_s\} \subseteq \{0, \ldots, p\}$ and P_{j_0} denotes a degree-j_0 Legendre polynomial. To construct the reward function, we randomly sample $j^\star$ from $[M]$, and draw $\theta^\star_{j^\star}$ from an i.i.d. standard gaussian distribution. We then normalize $\theta^\star$ so that $\|\theta^\star_{j^\star}\| = 1$. When sampling from the reward, we add Gaussian noise with standard deviation $\sigma = 0.01$. Figure 5 in the appendix shows how the random reward functions may look. For all experiments we set $n = 100$, and plot the cumulative regret $R(n)$ averaged over 20 different random seeds, the shaded areas in all figures show the standard error across these runs.

Algorithms. We perform experiments on two LinUCB algorithms, one with oracle knowledge of $j^\star$, and a naive one which takes into account all M feature maps. We run Explore-then-Commit (ETC) by Hao et al. [2020], which explores for a horizon of n_0 steps, performs Lasso once, and then selects actions greedily for the remaining steps. As another baseline, we introduce Explore-then-Stop (ETS) that explores for n_0 steps, performs model selection using the sparsity pattern of the Lasso estimator. For the remaining steps, the policy switches to LinUCB, calculated based on the selected features. Performance of ETC and ETS depends highly on n_0, so we tune this hyperparameter separately for each experiment. We also run CORRAL as proposed by Agarwal et al. [2017], with LinUCB agents similar to ALEXP. We tune the hyper-parameters of CORRAL as well. To initialize ALEXP we set the rates of ε_t, γ_t and η_t according to Theorem 1, and perform a light hyper-parameter tuning to choose the scaling constants. We have included the details and results of our hyper-parameter tuning in Appendix F.1. To solve (1), we use CELER, a fast solver for the group Lasso [Massias et al., 2018]. Every time LinUCB policy is used, we set the exploration coefficient $\alpha_t = 2$, and every time exploration is required, we sample according to $\pi = \mathrm{Unif}(\mathcal{X})$. Appendix F includes the pseudo-code for all baseline algorithms.[1]

Easy vs. Hard Cases. We construct an easy problem instance, where $s = 2$, $p = 10$, and thus $M = 55$. Models are lightly correlated since each two model can have at most one Legendre polynomial in common. We also generate an instance with highly correlated feature maps where $s = 8$ and $p = 10$, which will be a harder problem, since out of the total $M = 55$ models, there are 36 models which have at least 6 Legendre polynomials in common with the oracle model $j^\star$. Figure 1 shows that not only ALEXP is not affected by the correlations between the models, but also it achieves

[1] The PYTHON code for reproducing the experiments is accessible on github.com/lasgroup/ALEXP.

Figure 3: ALEXP is hardly affected by increasing the number of models (y-axis have various scales)

Figure 4: ALEXP can rule out models without ever having queried them ($M = 165$)

a performance competitive to the oracle in both cases, implying that our exponential weights technique for model selection is robust to choice of features. ETC and ETS rely on Lasso for model selection, which performs poorly in the case of correlated features. CORRAL uses log-barrier-OMD with an importance-weighted estimator, which has a significantly high variance. The curve for CORRAL in Figures 1 and 2 is cropped since the regret values get very large. Figure 6 shows the complete results. We construct another hard instance (Fig. 2), where the model class is large ($\sigma = 10$; $M = 165$). ALEXP continues to outperform all baselines with a significant gap. It is clear in the regret curves how explore-then-commit style algorithms are inherently horizon-dependent, and may exhibit linear regret, if stopped at an arbitrary time. This is not an issue with the other algorithms.

Scaling with M. Figure 3 shows how well the algorithms scale as M grows. For this experiment we set $s = 2$ and change $p \in \{9, \ldots, 13\}$. While increasing M hardly affects ALEXP and OracleUCB, other baselines become less and less sample efficient.

Learning Dynamics of ALEXP. Figure 4 gives some insight into the dynamics of ALEXP when $M = 165$. In particular, it shows how ALEXP can rule out sub-optimal agents without ever having queried them. Figure (a) shows the distribution q_t at $t = 20$ which is roughly equal to the optimal a_0 for ETC in this configuration. The oracle model $j^\star$ is annotated with a star, and has the highest probability of selection. We observe that, already at this time step, more than 80% of the agents are practically ruled out, due to small probability of selection. However, according to Figure (b), which shows M_t the total number of visited models, less than 10% of the models are queried at $t = 20$. This is the key practical benefit of ALEXP compared to black-box algorithms such as CORRAL. Lastly, Figure (c) shows how $q_{t,j^\star}$ the probability of selecting the oracle agent changes with time. While this probability is higher than that of the other agents, Figure (c) shows that it is not exceeding 0.25, therefore there is always a probability of greater than 0.75 that we sample another agent, making ALEXP robust to hard problem instances where many agents perform efficiently. We conclude that ALEXP seems to rapidly recognize the better performing agents, and select among them with high probability.

7 Conclusion

We proposed ALEXP, an algorithm for simultaneous online model selection and bandit optimization. As a first, our approach leads to anytime valid guarantees for model selection and bandit regret, and does not rely on a priori determined exploration schedule. Further, we showed how the Lasso can be used together with the exponential weights algorithm to construct a low-variance online learner. This new connection between high-dimensional statistics and online learning opens up avenues for future research on high-dimensional online learning. We established empirically that ALEXP has favorable exploration–exploitation dynamics, and outperforms existing baselines. We tackled the open problem of Agarwal et al. [2017], and showed that $\log M$ dependency for regret is achievable for linearly parametrizable rewards. This problem remains open for more general, non-parametric reward classes.

Acknowledgments

We thank Jonas Rothfuss and Miles Wang-Henderson for their valuable suggestions regarding the writing. We thank Scott Sussex and Felix Schur for their thorough feedback. This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and Innovation Program Grant agreement no. 815943. Nicolas Emmenegger was supported by the Zeno Karl Schindler Foundation and the Swiss Study Foundation. Aldo Pacchiano would like to thank the support of the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard. This work was supported in part by funding from the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard.

References

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

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-con dence-set conversions and application to sparse stochastic bandits. *Arti cial Intelligence and Statistics*. PMLR, 2012.

Jacob Abernethy and Alexander Rakhlin. Beating the adaptive bandit with high probability. *2009 Information Theory and Applications Workshop*. IEEE, 2009.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. An ef cient algorithm for bandit linear optimization. In *Conference on Learning Theory*, 2008.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. *International Conference on Machine Learning*. PMLR, 2014.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*. PMLR, 2017.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Foundations of Computer Science*. IEEE, 1995.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 2002b.

Debangshu Banerjee, Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Exploration in linear bandits with rich action sets and its implications for inference. *International Conference on Arti cial Intelligence and Statistics*. PMLR, 2023.

Peter L Bartlett, Varsha Dani, Thomas P Hayes, Sham M Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. *Conference on Learning Theory*, 2008.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 2020.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. *International Conference on Arti cial Intelligence and Statistics*, 2011.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 2009.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 2012.

Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017.

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

Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Arti cial Intelligence and Statistics*. PMLR, 2012.

Leonardo Cella and Massimiliano Pontil. Multi-task and meta-learning with sparse linear bandits. In *Conference on Uncertainty in Arti cial Intelligence*, 2021.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and ef cient contextual bandits with regression oracles. In *International Conference on Machine Learning*. PMLR, 2020.

Dylan J Foster, Satyen Kale, Mehryar Mohri, and Karthik Sridharan. Parameter-free online learning via model selection. *Advances in Neural Information Processing Systems*, 2017.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. In *Conference on Learning Theory*, 2011.

Apostolos Giannopoulos. Notes on isotropic convex bodies, October 2003.

Javier Gonzalez, Joseph Longworth, David C James, and Neil D Lawrence. Bayesian optimization for synthetic gene design. *arXiv preprint*, 2015.

Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 2020.

David Haussler, Jyrki Kivinen, and Manfred K Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 1998.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 2020.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic con dence sequences. *The Annals of Statistics*, 2021.

Mahdi Imani, Seyede Fatemeh Ghoreishi, Douglas Allaire, and Ulisses M Braga-Neto. Mfbo-ssm: Multi- delity bayesian optimization for fast inference in state-space models. *Proceedings of the AAAI Conference on Arti cial Intelligence*, 2019.

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. Popart: Ef cient sparse regression and experimental design for optimal sparse linear bandits. *Advances in Neural Information Processing Systems*, 2022.

Adel Javanmard and Andrea Montanari. Con dence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 2014.

Mohammad Reza Karimi, Nezihe Merve Gürel, Bojan Karlaš, Johannes Rausch, Ce Zhang, and Andreas Krause. Online active model selection for pre-trained classi ers. In *International Conference on Arti cial Intelligence and Statistics*. PMLR, 2021.

Parnian Kassraie, Jonas Rothfuss, and Andreas Krause. Meta-learning hypothesis spaces for sequential decision-making. In *International Conference on Machine Learning*. PMLR, 2022.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper con dence bounds for bandit problems. In *Arti cial intelligence and statistics*. PMLR, 2012.

Gi-Soo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Ef cient learning by implicit exploration in bandit problems with side observations. Advances in Neural Information Processing Systems, 2014.

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

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. Advances in Neural Information Processing Systems, 2020.

Wenjie Li, Adarsh Barik, and Jean Honorio. A simple uni ed framework for high dimensional bandit problems. In International Conference on Machine Learning. PMLR, 2022.

Xuefeng Liu, Fangfang Xia, Rick L Stevens, and Yuxin Chen. Cost-effective online contextual model selection. arXiv preprint, 2022.

Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. The annals of statistics, 2011.

Haipeng Luo, Mengxiao Zhang, Peng Zhao, and Zhi-Hua Zhou. Corralling a larger band of bandits: A case study on switching regret for linear bandits. Conference on Learning Theory. PMLR, 2022.

Odalric-Ambrym Maillard and Rémi Munos. Adaptive bandits: Towards the best history-dependent strategy. In International Conference on Arti cial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, 2011.

Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. Celer: a fast solver for the lasso with dual extrapolation. In International Conference on Machine Learning, 2018.

H. Brendan McMahan and Matthew Streeter. Tighter bounds for multi-armed bandits with expert advice. In Proceedings of the 22nd Annual Conference on Learning Theory, 2009.

Ahmadreza Moradipari, Berkay Turan, Yasin Abbasi-Yadkori, Mahnoosh Alizadeh, and Mohammad Ghavamzadeh. Feature and parameter selection in stochastic linear bandits. International Conference on Machine Learning. PMLR, 2022.

Vidya Muthukumar, Mitas Ray, Anant Sahai, and Peter Bartlett. Best of many worlds: Robust model selection for online supervised learning. Proceedings of the Twenty-Second International Conference on Arti cial Intelligence and Statistics, 2019.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. Advances in Neural Information Processing Systems, 2015.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandits. International Conference on Machine Learning. PMLR, 2021.

Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. Advances in Neural Information Processing Systems, 2020.

Aldo Pacchiano, Drausin Wulsin, Robert A Barton, and Luis Voloch. Neural design for genetic perturbation experiments. arXiv preprint, 2022.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. Advances in Neural Information Processing Systems, 2007.

Felix Schur, Parnian Kassraie, Jonas Rothfuss, and Andreas Krause. Lifelong bandit optimization: No prior and no regret. In Conference on Uncertainty in Arti cial Intelligence, 2023.

Adish Singla, Hamed Hassani, and Andreas Krause. Learning to interact with learning agents. In AAAI Conference on Arti cial Intelligence, 2018.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 1933.

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

Doniyor Ulmasov, Caroline Baroukh, Benoit Chachuat, Marc Peter Deisenroth, and Ruth Misener. Bayesian optimization with dimension scheduling: Application to biological systems. In Computer Aided Chemical Engineering. Elsevier, 2016.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics, 2009.

Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge university press, 2018.

Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. Cambridge university press, 2019.

Julian Zimmert and Tor Lattimore. Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In Conference on Learning Theory. PMLR, 2022.

Contents of Appendix

A Extended Literature Review 14

B Time Uniform Lasso Analysis 15

C Results on Exploration 18

C.1 ALExp with Uniform Exploration 20

C.2 Proof of Results on Exploration 20

D Proof of Regret Bound 23

D.1 Proof of Model Selection Regret 24

D.2 Proof of Virtual Regret 30

E Time-Uniform Concentration Inequalities 32

F Experiment Details 34

F.1 Hyper-Parameter Tuning Results 34

A Extended Literature Review

The sparse linear bandit literature considers linear reward functions of the form $x^\top \theta$, where $x \in \mathbb{R}^p$, however a sub-vector of size d is sufficient to span the reward function. This can be formulated as model selection among $M = \binom{p}{d}$ different linear parametrizations, where each is a d-dimensional feature map. We present the bounds in terms of d and M for coherence with the rest of the text, assuming that $\log M = O(p)$, which is the case when $d \asymp p$.

Table 2 compares recent work on sparse linear bandits based on a number of important factors. In this table, the ETC algorithms follow the general format of exploring, performing parameter estimation once at $t = n_0$, and then repeatedly suggesting the same action which maximizes $x^\top \hat{\theta}_{n_0}$. Explore-then-$\varepsilon$-Greedy takes a similar approach, however it does not settle on $\hat{\theta}_{n_0}$, rather it continues to update the parameter estimate and select $x_t = \arg\max_{x \in \mathcal{A}_t} x^\top \hat{\theta}_t$. The UCB algorithms iteratively update upper confidence bound, and choose actions which maximize them. The regret bounds in Table 2 are simplified to the terms with largest rate of growth, the reader should check the corresponding papers for rigorous results. Some of the mentioned bounds depend on problem-dependent parameters (e.g. c_K), which may not be treated as absolute constants and have complicated forms. To indicate such parameters we use $\asymp$ in Table 2, following the notation of Hao et al. [2020]. Note that $\asymp$ varies across the rows of the table, and is just an indicator for existence of other terms.

Abbasi-Yadkori et al. [2012] use the SEQSEW online regression oracle [Gerchinovitz, 2011] for estimating the parameter vector, together with a UCB policy. The regression oracle is an exponential weights algorithm, which runs on the squared error loss. This subroutine, and thereby the algorithm proposed by Abbasi-Yadkori et al. [2012] are not computationally efficient, and this is believed to be unavoidable. This work considers the data-rich regime and shows $R(n) = O(\sqrt{dMn})$, matching the lower bound of Theorem 24.3 in Lattimore and Szepesvári [2020].

Carpentier and Munos [2012] assume that the action set is a Euclidean ball, and that the noise is directly added to the parameter vector, i.e. $y_t = x^\top(\theta + \varepsilon_t)$. Roughly put, linear bandits with parameter noise are "easier" to solve than stochastic linear bandits with reward noise, since the noise is scaled proportionally to the features, and does less "damage" [Chapter 29.3 Lattimore and Szepesvári, 2020]. In this setting, Carpentier and Munos [2012] prescribe $\tilde{O}(d\sqrt{n})$ regret bound.

Recent work considers contextual linear bandits, where at every step a stochastic finite subset of size K from $\mathcal{A}$, is presented to the agent. It is commonly assumed that members of $\mathcal{A}_t$ are i.i.d., and the sampling distribution is diverse and time-independent. The diversity assumption is often in the

form of a restricted eigenvalue condition (De nition 2) on the covariance of the context distribution [e.g. in, Kim and Paik, 2019, Bastani and Bayati, 2020]. Li et al. [2022] require a stronger condition which directly assumes that $\lambda_{\min}(\Sigma_t)$ the minimum eigenvalue of the empirical covariance matrix is lower bounded. This is generally not true, but may hold with high probability. Hao et al. [2020] assume that the action set spans $\mathbb{R}^{dM}$. We believe that this assumption is the weakest in the literature, and conjecture that it is necessary for model selection. If not met, the agent can not explore in all relevant directions, and may not identify the relevant features. Our diversity assumption is similar to Hao et al. [2020], adapted to our problem setting. Mainly, we consider reward functions which are linearly parametrizable, i.e. $\beta^\top \phi(x)$, as oppose to linear rewards, i.e. $\beta^\top x$.

A key distinguishing factor between ALEXP and existing work on sparse linear bandit is that ALEXP is horizon-independent and does not rely on a forced exploration schedule. As shown on Table 2, majority of prior work relies either on an initial exploration stage, the length of which is determined according to n [e.g., Carpentier and Munos, 2012, Kim and Paik, 2019, Li et al., 2022, Hao et al., 2020, Jang et al., 2022], or on a hand crafted schedule, which is again designed for a speci c horizon [Bastani and Bayati, 2020]. Oh et al. [2021], which analyze K-armed contextual bandits, does not require explicit exploration, and instead imposes restrictive assumptions on the diversity of context distribution, e.g. relaxed symmetry and balanced covariance. Regardless, the regret bounds hold in expectation, and are not time-uniform.

Table 2: Overview of recent work on high-dimensional Bandits. Parameter shows existence of other problem-dependent terms which are not constants, and varies across different rows. The regret bounds are simpli ed and are not rigorous.

	$\lvert\mathcal{A}_t\rvert$	data-poor	adap. exp.	any-time	action selection policy	MS algo	context or action assumpt.	Regret
Abbasi-Yadkori et al.	1	✗	✓	✓	UCB	EXP4 on Sqrd error	$\mathcal{A}$ is compact	$\sqrt{dMn}$, w.h.p.
Foster et al.	K	✓	✓	✗	UCB	EXP4 on Sqrd error	$\min_{\pi}\Psi(\pi) > c$	$(Mn)^{3/4}K^{1/4} + \sqrt{KdMn}$ w.h.p
Carpentier and Munos	1	✓	✗	✗	UCB	Hard Thresh.	$\mathcal{A}$ is a ball param. noise	$d\sqrt{n}$, w.h.p.
Bastani and Bayati	K	✗	✗	✗	Explore then Greedy	Lasso	$\Psi(\Sigma) > c_K$	$Kd^2(\log n + \log M)^2$ w.h.p.
Kim and Paik	K	✗	✗	✗	Explore then ε-Greedy	Lasso	$\Psi(\Sigma) > c_K$	$d\sqrt{n}\log(Mn)$, w.h.p.
Oh et al.	K	✓	✓	✗	Greedy	Lasso	$\Psi(\Sigma) > c$ + other assums.	$d\sqrt{n\log(Mn)}$ in expectation
Li et al.	K	✓	✗	✗	ETC	Lasso	$\min_{\pi}\Psi(\hat{\Sigma}) > c$	$(n^2d)^{1/3}\sqrt{\log Mn}$ in expectation
Hao et al.	1	✓	✗	✗	ETC	Lasso	$\mathcal{A}$ spans $\mathbb{R}^{dM}$ + is compact	$(ndC_{\min}^{-1})^{2/3}(\log M)^{1/3}$ w.h.p.
Jang et al.	1	✓	✗	✗	ETC	Hard Thresh.	$\mathcal{A} \subset [-1,1]^{Md}$ + spans $\mathbb{R}^{Md}$	$(nd)^{2/3}(C_{\min}^{-1}\log M)^{1/3}$ w.h.p.
ALEXP (Ours)	1	✓	✓	✓	Greedy or UCB	EXP4 on reward est.	$\mathrm{Im}(\phi_j)$ spans $\mathbb{R}^d$ $\mathcal{A}$ is compact	$\sqrt{n\log M}(n^{1/4} + C_{\min}^{-1}\log M)$ w.h.p

B Time Uniform Lasso Analysis

We start by showing that the sum of squared sub-gaussian variables is a sub-Gamma process (c.f. De nition 22).

Lemma 4 (Empirical Process is sub-Gamma). For $t \geq 1$, suppose η_t are a sequence conditionally standard sub-Gaussians adapted to the ltration $\mathcal{F}_t = \sigma(\eta_1, \ldots, \eta_t)$. Let $v_t \in \mathbb{R}$, and $Z_t := \eta_t^2 - 1$. De ne the processes $S_t := \sum_{i=1}^{t} Z_i v_i$ and $V_t := 4\sum_{i=1}^{t} v_i^2$. Then $(S_t)_{t=0}^{\infty}$ is sub-Gamma with variance process $(V_t)_{t=0}^{\infty}$ and scale parameter $c = 4\max_{i \geq 1} v_i$.

Proof of Lemma 4. By de nition [c.f. De nition 1, Howard et al., 2021], S_t is sub-Gamma if for each $\lambda \in [0, 1/c)$, there exists a supermartingale $(M_t(\lambda))_{t=0}^{\infty}$ w.r.t. $\mathcal{F}_t$, such that $\mathbb{E}\, M_0 = 1$ and for all $t \geq 1$:

$$\exp\left\{\lambda S_t - \frac{\lambda^2}{2(1 - c\lambda)} V_t\right\} \leq M_t(\lambda) \quad \text{a.s.}$$

We show the above holds in equality by proving that the left hand side itself, is a supermartingale w.r.t. F_t. We define, $M_t(\ ) := \exp\{ S_t \quad {}^2V_t/2(1 \quad c\ )\}$, therefore,

$$\begin{aligned} E[M_t|F_{t\ 1}] &= E\left[\exp\left\{ S_{t\ 1} \quad \frac{{}^2}{2(1 \quad c\ )}V_{t\ 1} + Z_t v_t \quad \frac{2\ {}^2v_t^2}{(1 \quad c\ )} \right\} |F_{t\ 1}\right] \\ &= E[M_{t\ 1}|F_{t\ 1}]E\left[\exp\left(Z_t v_t\right)|F_{t\ 1}\right]\exp\left\{ \frac{2\ {}^2v_t^2}{1 \quad c} \right\} \\ &= M_{t\ 1}E\left[\exp\left(Z_t v_t\right)|F_{t\ 1}\right]\exp\left\{ \frac{2\ {}^2v_t^2}{1 \quad c} \right\} : \end{aligned}$$

Note that Z_t is $F_{t\ 1}$-measurable, conditionally centered and conditionally sub-exponential with parameters $(\ ;\ ) = (2\ ;4)$ (c.f. Vershynin [2018, Lemma 2.7.6] and Wainwright [2019, Example 2.8]). Therefore, for $< 1=c$,

$$E\left[\exp\left(v_t Z_t\right)|F_{t\ 1}\right] \quad \exp\left\{ 2\ {}^2v_t^2 \right\} \quad \exp\left\{ \frac{2\ {}^2v_t^2}{1 \quad c} \right\} ;$$

where the last inequality holds due to the fact that $0 \quad 1 \quad c < 1$. Therefore,

$$E[M_t|F_{t\ 1}] \quad M_{t\ 1}\exp\left\{ \frac{2\ {}^2v_t^2}{1 \quad c} \right\} \quad \exp\left\{ \frac{2\ {}^2v_t^2}{1 \quad c} \right\} = M_{t\ 1}:$$

for $\ 2 [0; 1=c)$, concluding the proof. □

We now construct a self-normalizing martingale sequence based on the ℓ_2-norm of the empirical process error term, and recognize that it is a sub-gamma process. We then employ our curved Bernstein bound Lemma 25 to control the boundary. This step will allow us to "ignore" the empirical process error term later in the lasso analysis.

Lemma 5 (Empirical Process is dominated by regularization) Let

$$A_j = \ 8t \quad 1 : \ \left({}^{>}_t "_t\right)_j {}_2 = t \quad {}_t = 2 \ :$$

Then, for any $0 < 1$, the event $A = \ \bigcap_{j=1}^{M} A_j$ happens with probability $1 \quad$, if for all $t \quad 1$,

$$_t \quad \frac{2}{\sqrt{t}}\left(1 + \frac{5}{2}\sqrt{d\left(\log(2M=\ ) + (\log\log\ d)_+\right)} + \frac{12}{2}\left(\log(2M=\ ) + (\log\log\ d)_+\right)\right):$$

Proof of Lemma 5. This proof includes a treatment of the empirical process similar to Lemma B.1 in Kassraie et al. [2022], but adapts it to our time-uniform setting. Since $_i$ are zero-mean sub-gaussian variables, as driven in Lemma 3.1 [Lounici et al., 2011], it holds that

$$A_j^c = \ 9t : \frac{1}{t^2}"_t^T \ {}_{t;j} \ {}^{>}_{t;j} "_t \quad \frac{{}^2}{4} = \ 9t : \frac{\left(\sum_{i=1}^{t} v_i\left({}_i^2 \quad 1\right)\right)}{\sqrt{2}kv_t k \ {}_{t;j}}$$

where $_i$ are sub-gaussian variables with variance proxy 1, v_i denotes the i-th eigenvalue of ${}_{t;j} \ {}^{>}_{t;j} = t$ with the concatenated vector $v_t = (v_1; : : : ; v_t)$, and

$${}_{t;j} = \frac{t^2 \ {}^2=(4 \ {}^2)}{\sqrt{2}\ {}^{>}_{t;j} \ {}_{t;j}}\ {}_{Fr}\left({}^{>}_{t;j} \ {}_{t;j}\right):$$

We can apply Lemma 25 to control the probability of event A_j^c by tuning . Mainly, for A_j^c to happen with probability less than $=M$, Lemma 25 states that the following must hold for all t,

$$\sqrt{2}kv_t k_2 \ {}_{t;j} \quad \frac{5}{2}\sqrt{\max\left\{ 4kv_t k_2^2; 1 \right\}} \ ! \ {}_{=M}\left(kv_t k_2\right) + 12 ! \ {}_{=M}\left(kv_t k_2\right)\max_{t\ 1} v_t \qquad (4)$$

Recall that w.l.o.g. feature maps are bounded everywhere, $k_j(x)k_2 \quad 1$, and $\text{rank}(\ _j) \quad d$ which allows for the following matrix inequalities,

$$\text{tr}\left({}^{>}_{t;j} \ {}_{t;j}\right) = \text{tr}\left(\ {}_{t;j} \ {}^{>}_{t;j}\right) = \sum_{i=1}^{t} {}^{>}_j(x_i) \ {}_j(x_i) \quad t$$

$$\Sigma_{t;j}\,\hat{\Sigma}_{t;j} \qquad \mathrm{tr}(\Sigma_{t;j}\,\hat{\Sigma}_{t;j}) \qquad t$$

$$\Sigma_{t;j}\,\hat{\Sigma}_{t;j} \qquad \Sigma_{t;j}\,\hat{\Sigma}_{t;j}\,_{Fr} \qquad p\,\bar{d} \qquad \Sigma_{t;j}\,\hat{\Sigma}_{t;j} \qquad t\,p\,\bar{d}$$

Therefore,

$$\|v_t\| = \Sigma_{t;j}\,\hat{\Sigma}_{t;j}\,_{Fr} = t\,p\,\bar{d}; \quad \max_{t\ 1} v_t = \max_{t\ 1} \Sigma_{t;j}\,\hat{\Sigma}_{t;j} = t \qquad 1:$$

For Eq. (4) to hold, is suf ces that for all 1,

$$\frac{2}{p\bar{t}}\left(1 + \frac{5}{2}\sqrt{\frac{p}{2}}\right)\sqrt{4d\,(\log(2M/\ ) + (\log\log\ d)_+)} + \frac{12}{p\bar{2}}\,(\log(2M/\ ) + (\log\log\ d)_+):$$

Therefore, if $_t$ are chosen to satisfy the above inequality, each A_j happens with probability less than $/M$. Then by applying union bound, $\bigcap_{j=1}^{M} A_j^c$ happens with probability less than $\square$

Proof of Theorem 3 The theorem statement requires that the regularization parameter is chosen such that condition of Lemma 5 is met, and therefore event happens with probability 1 . Throughout this proof, which adapts the analysis of Theorem 3.1. Lounici et al. [2011] to the time-uniform setting, we condition on happening, and later incorporate the probability.

Step 1. Let $\hat{}_t$ be a minimizer of L and $$ be the true coef cients vector, then $L(\hat{}_t; H_t; \ _t) \ L(\ ; H_t; \ _t)$. Writing out the loss and re-ordering the inequality we obtain,

$$\frac{1}{t}\ _t(\hat{}_t \qquad)\ _2^2 \quad \frac{2}{t}\|\varepsilon_t^T\ _t(\hat{}_t \qquad)\| + 2\ _t\sum_{j=1}^{M} k_j k_2 \qquad \hat{}_{t;j}\ _2 :$$

which is often referred to as the Basic inequality [Bühlmann and Van De Geer, 2011]. By Cauchy-Schwarz and conditioned on event ,

$$\|\varepsilon_t^T\ _t(\hat{}_t \qquad)\| \sum_{j=1}^{M} (\ _t^T\varepsilon_t)_j\ _2\ \hat{}_{t;j} \qquad _j\ _2 \quad \frac{t}{2}\sum_{j=1}^{M} \hat{}_{t;j} \qquad _j\ _2$$

then adding $_t\sum_{j=1}^{M} \hat{}_{t;j} \qquad _j\ _2$ to both sides, applying the triangle inequality, and recalling from Section 3 that $_j = 0$ for $j \in j^?$ gives

$$\frac{1}{t}\ _t(\hat{}_t \qquad)\ _2^2 + \ _t\sum_{j=1}^{M} \hat{}_{t;j} \qquad _j\ _2 \quad 2\ _t\sum_{j=1}^{M} \hat{}_{t;j} \qquad _j\ _2 + 2\ _t\sum_{j=1}^{M} k_j k_2 \qquad \hat{}_{t;j}\ _2$$

$$4\ _t\ \hat{}_{t;j^?} \qquad _{j^?}\ _2:$$

Since each term on the left hand side is positive, then each is also individually smaller than the right hand side, and we obtain,

$$\frac{1}{t}\ _t(\hat{}_t \qquad)\ _2^2 \quad 4\ _t\ \hat{}_{t;j^?} \qquad _{j^?}\ _2 \tag{5}$$

$$\sum_{\substack{j=1 \\ j \in j^?}}^{M} \hat{}_{t;j} \qquad _j\ _2 \quad 3\ \hat{}_{t;j^?} \qquad _{j^?}\ _2 \tag{6}$$

Step 2. Consider a sequence $(c_1; : : : ; c_k; : : :)$, where $c_1 \qquad c_k \ : : :$, then

$$c_k \quad \frac{1}{k}(kc_k + \sum_{i>k} c_i) \qquad \sum_{i \ 1} \frac{c_i}{k}: \tag{7}$$

De ne $J_1 = f j^?g$ and $J_2 = f j^?; j^0g$ where

$$j^0 = \arg\max_{\substack{j \in [M] \\ j \in j^?}} \hat{}_{t;j} \qquad _j\ _2:$$

For any $J \subseteq [M]$ the complementing set is denoted as $J^c = [M] \setminus J$. For simplicity let $c_j = \hat{\mu}_{t,j}^\top \theta_j$, and let (k) denote the index of the k-th largest element of $\{c_j : j \in J_1^c\}$. By definition of J_2^c we have,

$$\sum_{j \in J_2^c} \hat{\mu}_{t,j}^\top \theta_j \Big\|_2^2 = \sum_{k > 1,\, (k) \in J_1^c} c_{(k)}^2 \overset{(7)}{\leq} \sum_{k > 1,\, (k) \in J_1^c} \frac{\left(\sum_{i \in J_1^c} c_i\right)^2}{k^2}$$
$$\sum_{i \in J_1^c} c_i \Big\|_2^2 \sum_{k > 1,\, (k) \in J_1^c} \frac{1}{k^2} \overset{(6)}{\leq} 9 c_{j_?}^2$$
$$9(c_{j_?}^2 + c_{j_0}^2) = 9 \sum_{j \in J_2} \hat{\mu}_{t,j}^\top \theta_j \Big\|_2^2;$$

which, in turn, gives

$$\hat{\mu}_t^\top \theta_2 \Big\|_2 = \sqrt{\sum_{j=1}^M \hat{\mu}_{t,j}^\top \theta_j \Big\|_2^2} \leq \sqrt{10 \sum_{j \in J_2} \hat{\mu}_{t,j}^\top \theta_j \Big\|_2^2}. \tag{8}$$

Step 3. On the other hand, due to (6), and by definition of J_2 it also holds that

$$\sum_{j \in J_2^c} \hat{\mu}_{t,j}^\top \theta_j \Big\|_2^2 \leq 3 \sum_{j \in J_2} \hat{\mu}_{t,j}^\top \theta_j \Big\|_2^2.$$

From the theorem assumptions and Definition 2, we know that there exists $\phi(\Psi_t, 2)$, therefore by Definition 2, the feature matrix Ψ_t satisfies,

$$\sum_{j \in J_2} \hat{\mu}_{t,j}^\top \theta_j \Big\|_2^2 \frac{1}{t\, \phi^2(\Psi_t, 2)} \Big\| \Psi_t (\hat{\mu}_t^\top \theta_2) \Big\|_2^2$$
$$\overset{(5)}{\leq} \frac{1}{\phi^2(\Psi_t, 2)} 4 \Big\| \Psi_t \hat{\mu}_{t,j_?}^\top \theta_{j_?} \Big\|_2^2 \leq \frac{1}{\phi^2(\Psi_t, 2)} 4 \sqrt{\sum_{j \in J_2} \hat{\mu}_{t,j}^\top \theta_j \Big\|_2^2}.$$

From here, by applying (8) we get,

$$\hat{\mu}_t^\top \theta_2 \Big\|_2 \leq \frac{4\sqrt{10}}{\phi^2(\Psi_t, 2)}.$$

If Ψ_t are chosen according to Lemma 5, event $\mathcal{A}$ and, in turn, the inequality above hold with probability greater than $1 - \delta$. □

C Results on Exploration

In this section we present lower-bounds on the eigenvalues of the covariance matrix $\Psi_t^\top \Psi_t$, as it is later used in our regret analysis. In particular, we show that the feature matrix Ψ_t satisfies the restricted eigenvalue condition (Definition 2) required for valid Lasso confidence set (Theorem 3), and calculate a lower bound on $\phi(\Psi_t, 2)$. The lower bound is later used by Lemma 19 and Lemma 20 to develop the model selection regret. We show this bound in three steps.

Equivalent to Definition 2, we write $\phi(\Psi_t, s) = \inf_{b \in \mathcal{C}_s} \|\Psi_t b\|_2 / \sqrt{t}$ where

$$\mathcal{C}_s := \left\{ b \in \mathbb{R}^d \setminus \{0\} : \sum_{j \in J} \|b_j\|_2 \leq 3 \sum_{j \in J} \|b_j\|_2;\ \sqrt{\sum_{j \in J} \|b_j\|_2^2} = 1 \text{ s.t. } J \subseteq \{1, \ldots, M\},\ |J| \leq s \right\}. \tag{9}$$

For simplicity in notation, we further define

$$\tilde{\phi}(A, s) := \min_{b \in \mathcal{C}_s} b^\top A b. \tag{10}$$

since $\tilde{\phi}(\frac{\Psi_t^\top \Psi_t}{t}, s) = \phi^2(\Psi_t, s)$.

Step 1. Consider the exploratory steps at which $\eta_t = 1$. Let $\Psi_{t,t'}$ be a sub-matrix of Ψ_t where only rows from exploratory steps are included. Note that $\Psi_{t,t'} \in \mathbb{R}^{t' \times dM}$ is a random matrix, where the number of rows t' are also random. We show that $\phi^2(\Psi_t, s)$ is lower bounded by $\phi^2(\Psi_{t,t'}, s)$.

Lemma 6. Suppose $_{;t}$ has t^0 rows. Then,

$$^2(\ _t;s) \quad \frac{t^0}{t}\ ^2(\ _{;t};s)$$

Proof of Lemma 6. Let $^{(t)} = \ (x_t)\ ^>(x_t) \ 2\ R^{dM\ \ dM}$ for all $t = 1;:::;n$. Note that $^{(t)}$ is positive semi-de nite by construction. We have,

$$k\ _t bk_2^2 = \sum_{s=1}^{t} b^> \ ^{(s)}b = \sum_{s2T} b^> \ ^{(s)}b + \sum_{s2T} b^> \ ^{(s)}b$$

where the set T contains the indices of the exploratory steps at which the action is selected according to . Therefore,

$$\begin{aligned} ^2(\ _t;s) &= \frac{1}{t}\min_{b2\ s} k\ _t bk_2^2 \\ &= \min_{b2\ s} \frac{1}{t}\sum_{s2T} b^> \ ^{(s)}b + \frac{1}{t}\sum_{s2T} b^> \ ^{(s)}b \\ &\quad \min_{b2\ s} \frac{1}{t}\sum_{s2T} b^> \ ^{(s)}b \end{aligned}$$

where the last inequality holds due to $^{(s)}$ being PSD. Then we have,

$$^2(\ _t;s) \quad \min_{b2\ s} b^> \ \frac{1}{t}\sum_{s2T}^{t} \ ^{(s)} \ b = \frac{jT\ j}{t}\ ^2(\ _{;t};s)$$

□

While the number of rows of $_{;t}$ is a random variable, we continue to condition on the event that $_{;t}$ has t^0 rows, and investigate the distribution of its restricted eigenvalues.

Step II . The restricted eigenvalues of the exploratory submatrix are well bounded away from zero.

Lemma 7. Let be the solution to (2), and $s\ 2\ N$. Suppose $_{;t}$ has t^0 rows. Then for all > 0,

$$P\ \ 8t^0 : \ ^2(\ _{;t};s) \quad \tilde{}(\ ;s) \ \frac{80s}{\sqrt{t^0}}\sqrt{\log(2Md=\ ) + (\log\log 4t^0)_+} \quad 1$$

where $= (\ \ ;\ \ ) := E_x \ \ (x)\ ^>(x)$ and $\tilde{}$ is de ned in (10).

Step III . Remains to combine the two above lemmas and incorporate a high probability bound on t^0, showing that it is close to $\sum_{s=1}^{t}\ _s$.

Lemma 8. There exist absolute constants $C_1; C_2$ which satisfy,

$$P\ \ 8t \quad 1 : \ ^2(\ _t;2) \quad C_1\tilde{}(\ ;2)t\ ^{1=4} \quad C_2t\ ^{5=8}\sqrt{\log(Md=\ ) + (\log\log\ t)_+} \quad 1$$

if $_t = O(t\ ^{1=4})$. Let be the solution to (2), then it further holds that

$$P\ \ 8t \quad 1 : \ ^2(\ _t;2) \quad C_1C_{min}\ t\ ^{1=4} \quad C_2t\ ^{5=8}\sqrt{\log(Md=\ ) + (\log\log\ t)_+} \quad 1$$

The regret analysis of Hao et al. [2020] also relies on connecting $(n_s;s)$ to C_{min}, and for this, they use Theorem 2.4 of Javanmard and Montanari [2014]. This theorem states that there exists a problem-dependent constant C_1 for which $^2(\ _t;s) \quad C_1C_{min}$ with high probability, if $\quad n_0$ and roughly $n_0 = O(\ ^3\ n^2\log M)$. We highlight that Lemma 8, presents a lower-bound which holds for all $t \quad 1$, however this comes at the cost of getting a looser lower bound than the result of Javanmard and Montanari [2014] for the larger time steps. In fact, due to the sub-optimal dependency of Lemma 8 on t, we later obtain sub-optimal dependency on the horizon for the case when M . It is unclear to us if this rate can be improved without assuming knowledge of that $n \quad n_0$.

For the last lemma in this section we show that the empirical sub-matrices $_{t;j}$ are also bounded away from zero. This will be required later to prove Lemma 15.

Lemma 9 (Base Model $\lambda_{\min}$ Bound). Assume θ is the maximizer of Eq. (2). Then, with probability greater than $1-\delta$, simultaneously for all $j = 1, \ldots, M$ and $t \geq 1$,

$$\lambda_{\min}\left(\sum_{s<t} x_{t,j} x_{t,j}^\top\right) \geq C_1 C_{\min} t^{3/4} - C_2 t^{3/8} \sqrt{\log(M d/\delta) + (\log\log t)_+}$$

if $\epsilon_t = O(t^{-1/4})$.

C.1 ALEXP with Uniform Exploration

We presented our main regret bound (Theorem 1) in terms of $C_{\min}$, which only depends on properties of the feature maps and the action domain. We give a lower-bound on $C_{\min}$ for a toy scenario which corresponds to the problem of linear feature selection over convex action sets.

Proposition 10 (Invertible Features). Suppose $\phi(x) := Ax : \mathbb{R}^d \to \mathbb{R}^d$ is an invertible linear map, and $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex body. Then,

$$C_{\min} \geq \frac{\lambda_{\min}(A)}{\lambda^2_{\max}(T)} > 0$$

where T is the transformation which maps $\mathcal{X}$ to an isotropic body.

The lower-bound of Proposition 10 is achieved by simply exploring via $\pi = \mathrm{Unif}(\mathcal{X})$. Inspired by Schur et al. [2023, Lemma E.13], we show that even for non-convex action domains and orthogonal feature maps, the uniform exploration yields a constant lower-bound on restricted eigenvalues.

Proposition 11 (Orthonormal Features). Suppose $\phi_j : \mathcal{X} \to \mathbb{R}$ are chosen from an orthogonal basis of $L^2(\mathcal{X})$, and satisfy $\|\phi_j\|_{L^2(\mathcal{X})}/\mathrm{Vol}(\mathcal{X}) \geq 1$. Then there exist absolute constants C_1 and C_2 for which the exploration distribution $\pi = \mathrm{Unif}(\mathcal{X})$ satisfies

$$P\left[\forall t \geq 1 : \lambda_2\left(\sum_{s<t} x_s x_s^\top\right) \geq C_1 t^{1/4} - C_2 t^{5/8} \sqrt{\log(M d/\delta) + (\log\log t)_+}\right] \geq 1 - \delta.$$

The $d = 1$ condition is met without loss of generality, by splitting the higher dimensional feature maps and introducing more base features, which will increase M. Moreover, the orthonormality condition is met by orthogonalizing and re-scaling the feature maps. Basis functions such as Legendre polynomials and Fourier features [Rahimi et al., 2007] satisfy these conditions.

By invoking Proposition 11, instead of Lemma 8 in the proof of Theorem 1, we obtain the regret of ALEXP with uniform exploration.

Corollary 12 (ALEXP with Uniform Exploration). Let $\delta \in (0, 1]$. Suppose $\phi_j : \mathcal{X} \to \mathbb{R}$ are chosen from an orthogonal basis of $L^2(\mathcal{X})$, and satisfy $\|\phi_j\|_{L^2(\mathcal{X})}/\mathrm{Vol}(\mathcal{X}) \geq 1$. Assume the oracle agent employs a UCB or a Greedy policy, as laid out in Section 5. Choose $\epsilon = O(1/\sqrt{t}\, C(M, \delta, d))$ and $\epsilon_t = O(t^{-1/4})$ and $\gamma_t = O(C(M, \delta, d)/\sqrt{t})$, then ALEXP with uniform exploration $\pi = \mathrm{Unif}(\mathcal{X})$ attains the regret

$$\begin{aligned} R(n) = O\Big(& B n^{3/4} + \sqrt{n}\, C(M, \delta, d) \log M + B^2 \sqrt{n} + B \sqrt{n\,((\log\log n B^2)_+ + \log(1/\delta))} \\ & + (n^{3/4} + \log n) C(M, \delta, d) + n^{5/8} \sqrt{d \log n + \log(1/\delta) + B^2} \Big) \end{aligned}$$

with probability greater than $1-\delta$, simultaneously for all $n \geq 1$. Here,

$$C(M, \delta, d) = O\left(\sqrt{1 + \sqrt{d\,(\log(M/\delta) + (\log\log d)_+)} + (\log(M/\delta) + (\log\log d)_+)}\right).$$

C.2 Proof of Results on Exploration

As an intermediate step, we consider the restricted eigenvalue property of the empirical covariance matrix. Given t_0 samples, the empirical estimate of Σ is

$$\hat{\Sigma}_{t_0} := \frac{1}{t_0} \sum_{s=1}^{t_0} \phi(x_s)^\top \phi(x_s) \tag{11}$$

where x_s are sampled according to π. We show that every entry of $\hat{\Sigma}_{t_0}$ is close to the corresponding entry in Σ, and later use it in the proofs of eigenvalue lemmas.

Lemma 13 (Anytime Bound for The Entries of Empirical Covariance Matrix)Let $\hat{\Sigma}_{t^0}$ be the empirical covariance matrix corresponding to $(\phi(\cdot)\,;\,\cdot)$ given t^0 samples. Then,

$$P\left[\exists t^0 :\ d_1\left(\Sigma\,;\hat{\Sigma}_{t^0}\right) \geq \frac{5}{\sqrt{t^0}}\sqrt{(\log\log 4t^0)_+ + \log(2Md/\delta)}\right] \leq \delta$$

where $d_1(A,B) := \max_{i,j}|A_{i,j} - B_{i,j}|$.

Proof of Lemma 13. We show the element-wise convergence of $\hat{\Sigma}_t$ for the (i,j) entry where $i,j = 1,\ldots,dM$. Consider the random sequence $X_s := \phi_i(x_s)\phi_j(x_s) - \Sigma_{i,j}$. We show that $X_1,\ldots,X_n$ satisfies conditions of Lemma 26. We first observe that

$$E[X_s|X_{1:s-1}] = EX_s = \Sigma(i,j) - E_x\left[\phi_i(x)\phi_j(x)\right] = 0$$

since by definition $\Sigma_{i,j} = E_x\left[\phi_i(x)^\top\phi_j(x)\right]$. Moreover, we have normalized features $\|\phi(\cdot)\| \leq 1$, therefore, each entry $\phi_i(\cdot)\phi_j(\cdot)$ is also bounded, yielding $|X_s| \leq 2$. Then Lemma 26 implies that for all $\tilde{\delta} > 0$,

$$P\left[\exists t^0 :\ \frac{1}{t^0}\sum_{s=1}^{t^0} X_s \geq \frac{5}{\sqrt{t^0}}\sqrt{(\log\log 4t^0)_+ + \log(2/\tilde{\delta})}\right] \leq \tilde{\delta}.$$

Setting $\tilde{\delta} = \delta/(dM)$ and taking a union bound over all indices concludes the proof. □

We are now ready to present the proofs to the lemmas in Appendix C.

Proof of Lemma 7. By (11) we have $\hat{\Sigma}_{t^0} = \frac{\Phi_{\cdot,t}\Phi_{\cdot,t}^\top}{t^0}$, and thereby

$$\sigma^2(\Phi_{\cdot,t}\,;s) = \min_{b\in\mathbb{R}^{dM}} b^\top\hat{\Sigma}_{t^0}b = \tilde{\sigma}(\hat{\Sigma}_{t^0}\,;s).$$

Inspired by Lemma 10.1 in van de Geer and Bühlmann [2009], we show that element-wise closeness of matrices Σ and $\hat{\Sigma}_{t^0}$ (c.f. Lemma 13) implies closeness in $\tilde{\sigma}$:

$$\begin{aligned}\sigma^2(\Phi_{\cdot,t}\,;s) - \tilde{\sigma}(\Sigma\,;s) &= \tilde{\sigma}(\hat{\Sigma}_{t^0};s) - \tilde{\sigma}(\Sigma\,;s)\\ &= \tilde{\sigma}\left(\hat{\Sigma}_{t^0} - \Sigma\,;s\right)\\ &\leq \min_{b\in\mathbb{R}^{dM}} d_1\left(\Sigma\,;\hat{\Sigma}_{t^0}\right)\|b\|_1^2\end{aligned}$$

where the last line holds due to Hölder's. Moreover, since $b\in\mathbb{R}^{dM}_s$, for any $J \subseteq [dM]$ where $|J| \leq s$ it additionally holds that $\|b_J\|_2 \leq 1$ and

$$\|b\|_1 \leq (1+3)\|b_J\|_1 \leq 4\sqrt{s}\|b_J\|_2 \leq 4\sqrt{s}$$

which gives,

$$\sigma^2(\Phi_{\cdot,t}\,;s) - \tilde{\sigma}(\Sigma\,;s) \leq 16s\,d_1\left(\Sigma\,;\hat{\Sigma}_{t^0}\right).$$

Therefore by Lemma 13,

$$\sigma^2(\Phi_{\cdot,t}\,;s) - \tilde{\sigma}(\Sigma\,;s) \leq \frac{80s}{\sqrt{t^0}}\sqrt{(\log\log 4t^0)_+ + \log(2Md/\delta)} \tag{12}$$

with probability greater than $1-\delta$, simultaneously for all $t^0 \geq 1$. □

Proof of Lemma 8. In Lemma 6 we showed that

$$\sigma^2(\Phi_t\,;s) \leq \frac{t^0}{t}\,\sigma^2(\Phi_{\cdot,t}\,;s)$$

where t^0 indicates the number of rows in the exploratory sub-matrix of Φ. Recall that $t^0 = \sum_{s=1}^{t}\eta_s$ where η_s are i.i.d Bernoulli random variables with success probability ϵ. Due to Lemma 24,

$$P\left(\forall t \geq 1 :\ |t^0 - \epsilon t| \leq \epsilon t\right) \geq 1 - \delta/2 \tag{13}$$

where

$$\gamma_t := \frac{5}{2}\sqrt{\frac{(\log\log t)_+ + \log(8/\delta)}{t}}, \qquad \Gamma_t := \sum_{s=1}^{t} \gamma_s$$

Due to Lemma 7, with probability greater than $1 - \delta/2$ the following holds for all $t \geq 1$

$$\sigma^2(\mu_t, z) \leq \frac{t'}{t}\sigma(\mu', z) + \frac{160}{t}\sqrt{t'(\log\log 4t')_+ + \log(4M/\delta)}$$
$$\frac{t' - t}{t}\sigma(\mu', z) \leq 160\sqrt{\frac{t' + t}{t^2}\left((\log\log(4\mu_t + \Gamma_t))_+ + \log(4M/\delta)\right)}$$

where the second inequality holds with probability $1 - \delta/2$, by incorporating (13) and taking a union bound. For the rest of the proof and to keep the calculations simple, we ignore the values of the absolute constants. We use the notation $g(t) = o(f(t))$ to show that $f(t)$ grows much faster than $g(t)$. More formally, if for every constant c there exists t_0, where $g(t) \leq c|f(t)|$ for all $t \geq t_0$. Since $\gamma_s = O(s^{-1/4})$ there exists C such that $\Gamma_t = Ct^{3/4}$, then it is straightforward to observe that there exists absolute constants C_i's which satisfy,

$$\sigma^2(\mu_t, z) \leq C_1 t^{-1/4}\sigma(\mu', z) + \frac{5t^{3/2}\sigma(\mu', z)}{2}\sqrt{\frac{(\log\log t)_+ + \log(8/\delta)}{2}}$$
$$\leq C_2 t^{5/8}\sqrt{\log(M/\delta) + (\log\log t)_+} + o\left(t^{5/8}\sqrt{\log(M/\delta) + (\log\log t)_+}\right)$$
$$\leq C_1 t^{-1/4}\sigma(\mu', z) + C_3 t^{5/8}\sqrt{\log(M/\delta) + (\log\log t)_+}$$

The last inequality holds since $t^{3/2}\sqrt{\log\log t} = o(t^{5/8}\sqrt{\log\log t})$. The above chain of inequalities imply that there exist absolute constants C_1, C_2, for which

$$\mathbb{P}\left(\forall t \geq 1 : \sigma^2(\mu_t, z) \leq C_1\sigma(\mu', z)t^{-1/4} + C_2 t^{5/8}\sqrt{\log(M/\delta) + (\log\log t)_+}\right) \geq 1 - \delta.$$

If δ is chosen according to (2), then $\sigma(\mu', z) \leq C_{\min}$ yielding the lemma's second argument. □

Proof of Lemma 9. Fix $j \in \{1, \ldots, M\}$, and construct the set

$$\mathcal{B}_{1:j} = \left\{b \in \mathbb{R}^d \setminus \{0\} \;\middle|\; b = (b_1, \ldots, b_M);\ \text{s.t.}\ b_j \in \mathbb{R}^d; \|b_j\|_2 \leq 1 \text{ and } \forall j' \neq j : b_{j'} = 0\right\}.$$

Note that $\mathcal{B}_{1:j} \subseteq \mathcal{B}_s$. Therefore,

$$\inf_{b \in \mathcal{B}_{1:j}} \|\Gamma_t b\|_2 \geq \inf_{b \in \mathcal{B}_s} \|\Gamma_t b\|_2 = \sqrt{t}\,\sigma(\mu_t, s).$$

Moreover, by construction of $\mathcal{B}_{1:j}$ we have for all $b \in \mathcal{B}_{1:j}$ that $\Gamma_t b = \Gamma_{t:j} b_j$, therefore,

$$\inf_{b \in \mathcal{B}_{1:j}} \|\Gamma_t b\|_2^2 = \inf_{\substack{b_j \in \mathbb{R}^d \\ \|b_j\|_2^2 \leq 1}} \|\Gamma_{t:j} b_j\|_2^2 = \lambda_{\min}\left(\Gamma_{t:j}^\top \Gamma_{t:j}\right).$$

From the above equations we conclude that $\lambda_{\min}(\Gamma_{t:j}^\top \Gamma_{t:j}) \geq t\,\sigma^2(\mu_t, s)$, for all $j = 1, \ldots, M$. Therefore, using Lemma 8 we obtain that there exists C_1, C_2 such that

$$\mathbb{P}\left(\forall t \geq 1, j = 1, \ldots, M : \lambda_{\min}(\Gamma_{t:j}^\top \Gamma_{t:j}) \geq C_1 C_{\min} t^{3/4} - C_2 t^{3/8}\sqrt{\log(M/\delta) + (\log\log t)_+}\right) \geq 1 - \delta$$

□

Proof of Proposition 10. Since X is a convex body, then there exists an invertible matrix T, such that $T(X)$ is an isotropic body [e.g. Proposition 1.1.1., Giannopoulos, 2003]. Then by definition, $X \sim \mathrm{Unif}(T(X))$ is an isotropic distribution and $\mathrm{Cov}(X) = I_d$ [e.g., c.f. Chapter 3.3.5 Vershynin, 2018]. Since ϕ is linear and invertible, it may be written as $\phi(x) = Ax$, where A is an invertible matrix. Therefore,

$$\Sigma(\phi, \mu) = \mathrm{Cov}(\phi(X)) = A^\top \mathrm{Cov}(X) A = A^\top \mathrm{Cov}\left(T^{-1}X\right) A = A^\top (T^{-1})^2 A.$$

As for the minimum eigenvalue, suppose $v \in \mathbb{R}^d$ and $\|v\| = 1$, then

$$C_{\min} \geq \lambda_{\min}(\Sigma(\phi, \mu)) = v^\top A^\top (T^{-1})^2 A v \geq \|Av\|_2\,\lambda_{\min}(T^{-2}) = \frac{\|Av\|_2}{\lambda^2_{\max}(T)} \geq \frac{\lambda_{\min}(A)}{\lambda^2_{\max}(T)}.$$

□

Proof of Proposition 11. By the assumption of the proposition, for all $i \in M$

$$[(\ ;\ )]_{i,i} = E_x\ \ {}_i^2(x) = \frac{1}{\mathrm{Vol}(X)} \int_X {}_i^2(x)d\ (x)\ \ 1$$

and for all $i \neq j$,

$$[(\ ;\ )]_{i,j} = E_x\ \ {}_i(x)\ {}_j(x) = \frac{1}{\mathrm{Vol}(X)} \int_X {}_i(x)\ {}_j(x)d\ (x) = 0$$

We use $= (\ ;\ )$. For any $b \in R^{Md}$ where $kbk\ \ 1$,

$$\begin{aligned} b^{>}\ b &= \sum_{i,j\ 2[M]} b_j^{>}\ {}_{i,j}\ b_i = \sum_{i 2[M]} b_i^{>}\ {}_{i,i}\ b_i + \sum_{i,j\ 2[M],i \neq j} b_j^{>}\ {}_{i,j}\ b_i \\ &= \sum_{i 2[M]} b_i^{>}\ {}_{i,i}\ b_i \\ &\ \ \sum_{i 2[M]} kb_i k_2^2\ \ 1: \end{aligned}$$

Which implies,

$$\tilde{}(\ ;s) = \min_{b2\ s} b^{>}\ b\ \ 1:$$

By Lemma 8, there exist absolute constants C_1 and C_2 for which,

$$P\ \ 8t\ \ 1 :\ \ {}^2(\ {}_t;2)\ \ \tilde{}(\ ;2)C_1 t^{\ 1=4}\ \ C_2 t^{\ 5=8}\sqrt{\log(Md=\ ) + (\log\log\ t)_+}\ \ \ 1\ \ :$$

concluding the proof.

□

D Proof of Regret Bound

Theorem 14 (Anytime Regret, Formal) Let $\ \ 2\ (0;1]$ and $\ $ be the maximizer of (2). Assume the oracle agent employs a UCB or a Greedy policy, as laid out in Section 5. Suppose $O(C_{min}\ t^{\ 1=2}C(M;\ ;d))$ and ${}_t = O(t^{\ 1=4})$ and ${}_t = O(C(M;\ ;d)t^{\ 1=2})$, then exists absolute constants $C_1;::::;C_6$ for which ALEXP attains the regret

$$\begin{aligned} R(n)\ \ & C_1 B n^{3=4} + C_2 \sqrt{n} C_{min}^{\ 1} C(M;\ ;d)\log M \\ & + C_3 B^{\ 2} C_{min}\sqrt{n} + C_4 B\sqrt{n\ ((\log\log nB^{\ 2})_+ + \log(1=\ ))} \\ & + C_5\ \ 1 + C_{min}^{\ 1} n^{\ 3=8}\sqrt{\log(Md=\ ) + (\log\log\ n)_+} \\ & \ \ B n^{\ 1=4} + (n^{3=4} + \frac{\log n}{C_{min}})C(M;\ ;d) + \sqrt{\frac{n^{5=8}}{C_{min}}}\sqrt{d\log n + \log(1=\ ) + B^{\ 2}} \end{aligned}$$

with probability greater than $1\ \ $, simultaneously for all $n\ \ 1$. Here,

$$C(M;\ ;d) = C_6\ \ \sqrt{1 + \sqrt{d(\log(M=\ ) + (\log\log\ d)_+)} + (\log(\ M=\ ) + (\log\log\ d)_+)}:$$

Our main regret bound is an immediate corollary of Lemma 15 and Lemma 16, considering the regret decomposition of (3).

Lemma 15 (Virtual Regret of the Oracle) Let $\ \ 2\ (0;1]$ and $\tilde{}> 0$. Assume the oracle agent employs a UCB or a Greedy policy, as laid out in Section 5, $_t = O(t^{1=4})$, there exists an absolute constant C_1 for which with probability greater than $1\ \ $, simultaneously for all $n\ \ 1$,

$$R_{j?}(n) = \sqrt{\frac{C_1 n^{5=8}}{C_{min}}}\ \ 1 + n^{\ 3=8} C_{min}^{\ 1}\sqrt{\log(Md=\ ) + (\log\log\ n)_+}\sqrt{\ {}^2 d\log\frac{n}{d} + 1\ \ + 2\ {}^2\log(1=\ ) + \tilde{}B^{\ 2}}$$

Lemma 16 (Any-Time Model-Selection Regret, Formal). Let $\in (0;1]$ and be the maximizer of (2). Suppose $_t = O(C_{min} = \bar{t}C(M; ;d))$ and $_t = O(t^{1=4})$ and $_t = O(C(M; ;d)=\bar{t})$, then exists absolute constants C_1 for which ALEXP attains the model selection regret

$$R(n;j) \quad C_1 B n^{3=4} + C_2 \sqrt{\bar{n}} C_{min}^{1} C(M; ;d) \log M + C_3 B^2 C_{min} \sqrt{\bar{n}} + C_4 B \sqrt{n((\log\log n B^2)_+ + \log(1=))} + C_5 \quad B n^{1=4} + (n^{3=4} + \frac{\log n}{C_{min}}) C(M; ;d) \quad 1 + C_{min}^{1} n^{3=8} \sqrt{\log(M d=) + (\log\log n)_+}$$

with probability greater than 1 , simultaneously for all n 1. Here,

$$C(M; ;d) = C_6 \sqrt{1 + \sqrt{d(\log(M=) + (\log\log d)_+) + (\log(M=) + (\log\log d)_+)}}:$$

D.1 Proof of Model Selection Regret

Our technique for bounding the model selection regret relies on a classic horizon-independent analysis of the exponential weights algorithm, presented in Lemma 17.

Lemma 17 (Anytime Exponential Weights Guarantee). Assume $_t \hat{r}_{t;j}$ 1 for all 1 j M and t 1. If the sequence $(_t)_{t \ 1}$ is non-increasing, then for all n 1,

$$\sum_{t=1}^{n} \hat{r}_{t;k} \quad \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t;j} \hat{r}_{t;j} \quad \frac{\log M}{n} + \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t;j} \hat{r}_{t;j}^2$$

for any arm k 2 [M].

Proof of Lemma 17. Define $\hat{R}_{t;i} := \sum_{s=1}^{t} \hat{r}_{s;i}$ to be the expected cumulative reward of agent i after t steps. We rewrite for a fixed k

$$\sum_{t=1}^{n} \hat{r}_{t;k} \quad \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t;j} \hat{r}_{t;j} = \sum_{t=1}^{n} \hat{r}_{t;k} \quad \sum_{t=1}^{n} E_{j \ q_t}[\hat{r}_{t;j}]: \tag{14}$$

We focus on a single term in the second sum. For any t we have

$$\begin{aligned} E_{j \ q_t}[\hat{r}_{t;j}] &= \log(\exp(\ E_{j \ q_t}[\frac{t}{t}\hat{r}_{t;j}])) = \log(\exp(\ E_{j \ q_t}[\ _t\hat{r}_{t;j}])^{1=\ _t}) \\ &= \frac{1}{t}\log(\exp(\ E_{j \ q_t}[\ _t\hat{r}_{t;j}])) \\ &= \frac{1}{t}\log(E_{i \ q_t}\exp(\ E_{j \ q_t}[\ _t\hat{r}_{t;j}])) \end{aligned} \tag{15}$$

The inner expectation is over j, while the outer one is over i and therefore has no effect. Moreover,

$$\begin{aligned} \frac{1}{t}\log E_{i \ q_t}\exp(\ _t E_{j \ q_t}[\hat{r}_{t;j}] + \ _t\hat{r}_{t;i}) &= \frac{1}{t}\log(\exp(\ _t E_{j \ q_t}[\hat{r}_{t;j}])E_{i \ q_t}\exp(\ _t\hat{r}_{t;i})) \\ &= \frac{1}{t}\log E_{i \ q_t}\exp(\ _t E_{j \ q_t}[\hat{r}_{t;j}]) \\ &\quad + \frac{1}{t}\log E_{i \ q_t}\exp(\ _t\hat{r}_{t;i}) \end{aligned} \tag{16}$$

where again, the expectation can be reintroduced to get the last line. Combining (15) and (16),

$$E_{j \ q_t}[\hat{r}_{t;j}] = \frac{1}{t}\log E_{i \ q_t}\exp(\ _t E_{j \ q_t}[\hat{r}_{t;j}] + \ _t\hat{r}_{t;i}) \quad \frac{1}{t}\log E_{i \ q_t}\exp(\ _t\hat{r}_{t;i}) \tag{17}$$

This transformation is at the core of many exponential weight proofs [Bubeck et al., 2012, Lattimore and Szepesvári, 2020]. We first bound the first term in (17):

$$\log E_{i \ q_t}\exp(\ _t E_{j \ q_t}[\hat{r}_{t;j}] + \ _t\hat{r}_{t;i}) = \log\ E_{i \ q_t}\exp(\ _t\hat{r}_{t;i}) \quad \ _t E_{j \ q_t}\hat{r}_{t;j}$$

$$\overset{(I)}{=} \mathbb{E}_{i \sim q_t} \exp(\eta_t \hat{r}_{t,i}) - 1 - \eta_t \mathbb{E}_{j \sim q_t} \hat{r}_{t,j}$$
$$= \mathbb{E}_{i \sim q_t} [\exp(\eta_t \hat{r}_{t,i}) - 1 - \eta_t \hat{r}_{t,i}]$$
$$\overset{(II)}{\leq} \mathbb{E}_{i \sim q_t} \eta_t^2 \hat{r}_{t,i}^2 \tag{18}$$

where in (I) we use the fact that $\log(z) \leq z - 1$ and in (II) we use the fact that for $x \leq 1$, we have $\exp(x) \leq 1 + x + x^2$, and hence $\exp(x) - 1 - x \leq x^2$. For the second term in (17), we will mirror the potential argument in Bubeck et al. [2012], but with a slightly different potential function. We expand the definition of J_t:

$$\frac{1}{\eta_t} \log \mathbb{E}_{i \sim q_t} \exp(\eta_t \hat{r}_{t,i}) = \frac{1}{\eta_t} \log \frac{\sum_{i=1}^M \exp(\eta_t \hat{R}_{t,i})}{\sum_{i=1}^M \exp(\eta_t \hat{R}_{t-1,i})}$$
$$= \frac{1}{\eta_t} \log \frac{1}{M} \sum_{i=1}^M \exp(\eta_t \hat{R}_{t,i}) + \frac{1}{\eta_t} \log \frac{1}{M} \sum_{i=1}^M \exp(\eta_t \hat{R}_{t-1,i})$$
$$= J_t(\eta_t) - J_{t-1}(\eta_t); \tag{19}$$

where we define $J_t(\eta) = \frac{1}{\eta} \log \frac{1}{M} \sum_{i=1}^M \exp(\eta \hat{R}_{t,i})$. We also define $F_t(\eta) = \frac{1}{\eta} \log \frac{1}{M} \sum_{i=1}^M \exp(\eta \hat{R}_{t,i})$. We observe the relation $J_t(\eta) = F_t(-\eta)$. From this, it follows that for any η, we have $J_t'(\eta) = -F_t'(-\eta) \geq 0$, by the argument in Bubeck et al. [2012, Theorem 3.1] that shows $F_t'(\eta) \leq 0$ for any η.

Putting together the pieces Now, we can bound (17) by inputing (18) and (19):

$$\mathbb{E}_{j \sim q_t} [\hat{r}_{t,j}] - \mathbb{E}_{i \sim q_t} \eta_t \hat{r}_{t,i}^2 + J_t(\eta_t) - J_{t-1}(\eta_t)$$

With this, we rewrite (14) as

$$\sum_{t=1}^n \hat{r}_{t,k} - \sum_{t=1}^n \mathbb{E}_{j \sim q_t} [\hat{r}_{t,j}] = \sum_{t=1}^n \hat{r}_{t,k} + \sum_{t=1}^n \mathbb{E}_{i \sim q_t} \eta_t \hat{r}_{t,i}^2 + \sum_{t=1}^n J_t(\eta_t) - J_{t-1}(\eta_t) \tag{20}$$

Potential manipulation We can do an Abel transformation on the sum of potentials in (20), namely obtaining

$$\sum_{t=1}^n J_t(\eta_t) - J_{t-1}(\eta_t) = \sum_{t=1}^{n-1} (J_t(\eta_t) - J_t(\eta_{t+1})) + J_n(\eta_n);$$

where we used that $J_0(\eta) = 0$. We know $J_t'(\eta) \leq 0$ and so J_t is decreasing and since $\eta_{t+1} \geq \eta_t$, we have $J_t(\eta_{t+1}) \leq J_t(\eta_t)$ or $(J_t(\eta_t) - J_t(\eta_{t+1})) \leq 0$, so that for any fixed k

$$\sum_{t=1}^n J_t(\eta_t) - J_{t-1}(\eta_t) \leq \frac{\log(M)}{\eta_n} - \frac{1}{\eta_n} \log \left(\sum_{i=1}^M \exp(\eta_n \hat{R}_{n,i}) \right)$$
$$\overset{(\star)}{\leq} \frac{\log(M)}{\eta_n} - \frac{1}{\eta_n} \log \exp(\eta_n \hat{R}_{n,k})$$
$$= \frac{\log(M)}{\eta_n} - \sum_{t=1}^n \hat{r}_{t,k} \tag{21}$$

where $(\star)$ follows because exp is positive and log is decreasing (notice that we drop $M - 1$ terms from the sum). Plugging (21) into (20), we obtain

$$\sum_{t=1}^n \hat{r}_{t,k} - \sum_{t=1}^n \mathbb{E}_{j \sim q_t} [\hat{r}_{t,j}] \leq \sum_{t=1}^n \hat{r}_{t,k} + \sum_{t=1}^n \mathbb{E}_{i \sim q_t} \eta_t \hat{r}_{t,i}^2 + \sum_{t=1}^n J_t(\eta_t) - J_{t-1}(\eta_t)$$
$$\leq \sum_{t=1}^n \hat{r}_{t,k} + \sum_{t=1}^n \mathbb{E}_{i \sim q_t} \eta_t \hat{r}_{t,i}^2 + \frac{\log(M)}{\eta_n} - \sum_{t=1}^n \hat{r}_{t,k}$$
$$\leq \sum_{t=1}^n \mathbb{E}_{i \sim q_t} \eta_t \hat{r}_{t,i}^2 + \frac{\log(M)}{\eta_n}$$

$$= \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j}^2 + \frac{\log(M)}{n}.$$

□

We expressed in Section 5.2, that the model selection regret of EXP, is closely tied to the bias and variance of the reward estimates $\hat{r}_{t,j}$. The following lemma formalizes this claim.

Lemma 18. (Anytime Generic regret bound) If γ_t is picked such that $\gamma_t \geq \hat{r}_{t,j}$ for all $1 \leq j \leq M$ and $1 \leq t$ almost surely, then Algorithm 1 satisfies with probability greater than $1 - 2/3$, that simultaneously for all $n \geq 1$

$$R(n;i) \leq 2B \sum_{t=1}^{n} \gamma_t + \frac{\log M}{n} + \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j}^2 + \sum_{t=1}^{n} (\gamma_t + \hat{r}_{t,i}) + \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t,j} (\gamma_t + \hat{r}_{t,j}) + 10B \sqrt{\frac{\log\log(nB^2)_+ + \log(12/\delta)}{n}}$$

where $\gamma_{t,i} = |r_{t,i} - \hat{r}_{t,i}|$.

Proof of Lemma 18. Let ϵ_t denote the Bernoulli random variable that is equal to 1 if at step t we select actions according to p and 0 otherwise. At each step t with $\epsilon_t = 1$ ALEXP accumulates a regret of at most $2B$, since $r_t \in [-B, B]$ and $r(\cdot) \in [-1, 1]$. We can decompose the regret as,

$$R(n;i) \leq \sum_{t=1}^{n} 2B\epsilon_t + (r_{t,i} - r_t)(1 - \epsilon_t)$$

For the first term, by Lemma 24, we have

$$2B \sum_{t=1}^{n} \epsilon_t \leq 2B \sum_{t=1}^{n} \epsilon_t + \frac{5}{2} \sqrt{\frac{\log\log(n)_+ + \log(4/\delta_1)}{n}},$$

simultaneously for all $n \geq 1$, with probability $1 - \delta_1$. Let $\hat{r}_t := \sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j}$. We may re-write the second term of the regret as follows,

$$\sum_{t=1}^{n} (1 - \epsilon_t)(r_{t,i} - r_t) \leq \sum_{t=1}^{n} (1 - \epsilon_t)\left[(r_{t,i} - \hat{r}_{t,i}) + (\hat{r}_{t,i} - \hat{r}_t) + (\hat{r}_t - r_t)\right]$$
$$\leq \sum_{t=1}^{n} \gamma_{t,i} + (1 - \epsilon_t)\left[(\hat{r}_{t,i} - \hat{r}_t) + (\hat{r}_t - r_t)\right]$$

We bound the second term on the right hand side, using Lemma 17

$$\sum_{t=1}^{n} (1 - \epsilon_t)(\hat{r}_{t,i} - \hat{r}_t) \leq \sum_{t=1}^{n} (\hat{r}_{t,i} - \hat{r}_t) \leq \frac{\log M}{n} + \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j}^2.$$

As for the third term,

$$(1 - \epsilon_t)\left(\sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j} - r_t\right) = (1 - \epsilon_t)\left[\sum_{j=1}^{M} q_{t,j} (\hat{r}_{t,j} - r_{t,j} + r_{t,j}) - r_t\right]$$
$$\leq \sum_{j=1}^{M} q_{t,j} \gamma_{t,j} + (1 - \epsilon_t)\left(r_t - \sum_{j=1}^{M} q_{t,j} r_{t,j}\right).$$

It remains to bound the deviation term. For t that satisfy $\epsilon_t = 0$, the action/model is selected according to $q_{t,j}$, therefore the conditional expectation of r_t can be written as

$$E_{t-1} r_t = \sum_{j=1}^{M} q_{t,j} r_{t,j}$$

The sequence $X_t := r_t - E_{t-1} r_t$ is a martingale difference sequence adapted to the history H_t, since for every $t \geq 1$,

$$E_{t-1} X_t = E[r_t - E_{t-1} r_t | H_{t-1}] = 0.$$

Since $r_t \in [0, B]$, then $X_t \in [-B, B]$ almost surely, which allows for an application of anytime Azuma-Hoeffding (Lemma 26):

$$P\left(\exists n : \sum_{t=1}^{n} \left(r_t - \sum_{j=1}^{M} q_{t,j} r_{t,j}\right) \geq \frac{5B}{2} \sqrt{n\left((\log\log nB^2)_+ + \log(2/\delta_2)\right)}\right) \leq \delta_2$$

which, in turn, leads us to

$$\sum_{t=1}^{n} (1 - q_t)\left(r_t - \sum_{j=1}^{M} q_{t,j} r_{t,j}\right) \overset{\text{a.s.}}{\leq} \sum_{t=1}^{n} \left(r_t - \sum_{j=1}^{M} q_{t,j} r_{t,j}\right) \overset{\text{w.h.p.}}{\leq} \frac{5B}{2} \sqrt{n\left((\log\log nB^2)_+ + \log(2/\delta_2)\right)}$$

simultaneously for all $n \geq 1$. We set $\delta_1 = \delta_2 = \delta/3$, take a union bound and put the terms together obtaining,

$$R(n,i) \leq 2B \sum_{t=1}^{n} \eta_t + \frac{\log M}{\eta_n} + \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j}^2 + \sum_{t=1}^{n} (\hat{\ell}_{t,i} + \sum_{j=1}^{M} q_{t,j} \hat{\ell}_{t,j}) + \frac{5B}{2} \sqrt{n\left((\log\log nB^2)_+ + \log(6/\delta)\right)} + 5B \sqrt{n\left((\log\log n)_+ + \log(12/\delta)\right)}$$

We upper bound the sum of last two terms to conclude the proof. □

The next two lemmas bound the bias and variance terms which appear in Lemma 18.

Lemma 19 (Anytime Bound on the Bias Term) *If the regularization parameter of Lasso is chosen at every step as*

$$\gamma_t = \frac{2}{\sqrt{t}} \sqrt{1 + \sqrt{\frac{5}{2}} \sqrt{d\left(\log(2M/\delta) + (\log\log d)_+\right)} + \frac{12}{\sqrt{2}} \left(\log(2M/\delta) + (\log\log d)_+\right)}$$

and $\eta_t = O(t^{-1/4})$, then with probability greater than $1 - \delta$, simultaneously for all $n \geq 1$,

$$\sum_{t=1}^{n} |\hat{\ell}_{t,i} - r_{t,i}| \leq n^{3/4} C_{\min}^{-1} C(M, \delta, d) \left(1 + n^{3/8} C_{\min}^{-1} \sqrt{\log(Md/\delta) + (\log\log n)_+}\right)$$

where

$$C(M, \delta, d) := C \sqrt{1 + \sqrt{d\left(\log(M/\delta) + (\log\log d)_+\right)} + \left(\log(M/\delta) + (\log\log d)_+\right)}$$

and C is an absolute constant.

Proof of Lemma 19. By the definition of the expected reward and its estimate,

$$\begin{aligned}
\sum_{t=1}^{n} |\hat{\ell}_{t,i} - r_{t,i}| &= \sum_{t=1}^{n} \int_{\mathcal{X}} (r(x) - \hat{r}_t(x)) \, dp_{t+1,i}(x) \\
&\leq \sum_{t=1}^{n} \int_{\mathcal{X}} |r(x) - \hat{r}_t(x)| \, dp_{t+1,i}(x) \\
&\overset{\text{C.S.}}{\leq} \sum_{t=1}^{n} \int_{\mathcal{X}} \|\hat{\theta}_t - \theta^*\|_2 \|x\|_2 \, dp_{t+1,i}(x) \\
&\overset{\text{bdd.}}{\leq} \sum_{t=1}^{n} \|\hat{\theta}_t - \theta^*\|_2 \int_{\mathcal{X}} dp_{t+1,i}(x) = \sum_{t=1}^{n} \|\hat{\theta}_t - \theta^*\|_2 .
\end{aligned}$$

We highlight that the Cauchy-Schwarz step may be refined. By further assuming that β_j is bounded away from zero (also called the beta-min condition [Bühlmann and Van De Geer, 2011]) one can show that $\hat{\beta}_t$ is a 2-sparse vector. This will then allow one to only rely on boundedness of $\|\beta_j\|_2$ rather than $\|\beta_j\|_1$ to derive the last inequality, and relax our assumption of $\|\beta_j\|_1 \leq 1$ to $\|\beta_j\|_2 \leq 1$ for all $j \in [M]$. From Theorem 3, with probability greater than $1 - \alpha/2$ simultaneously for all $n \geq 1$,

$$\sum_{t=1}^{n} |\hat{r}_{t,i} - r_{t,i}| \leq \sum_{t=1}^{n} \frac{4\sqrt{10}\, t}{\phi^2(\gamma_t, 2)} = \bar{C}(M, \gamma, d) \sum_{t=1}^{n} \frac{1}{\phi^2(\gamma_t, 2)\sqrt{t}}$$

where,

$$\bar{C}(M, \gamma, d) := 8\left(1 + \sqrt{\frac{5}{2}}\sqrt{d(\log(4M/\alpha) + (\log\log d)_+)} + \sqrt{\frac{12}{2}}(\log(4M/\alpha) + (\log\log d)_+)\right).$$

From Lemma 8, there exist absolute constants C_1, C_2 for which,

$$P\left(\forall t \geq 1 : \phi^2(\gamma_t, 2) \geq C_1 C_{\min} t^{-1/4} - C_2 t^{-5/8}\sqrt{\log(Md/\alpha) + (\log\log t)_+}\right) \geq 1 - \alpha.$$

Using Taylor approximation we observe that $\frac{1}{1 - x^{-1/4}} = 1 + x^{-1/4} + o(x^{-1/4}) = O(1 + x^{-1/4})$. Therefore, these exists absolute constants C_3, C_4, for which with probability greater than $1 - \alpha$ for all $t \geq 1$

$$\begin{aligned}
\sum_{t=1}^{n} \frac{\bar{C}(M, \gamma, d)}{\phi^2(\gamma_t, 2)\sqrt{t}} &\leq \sum_{t=1}^{n} \frac{\bar{C}(M, \gamma, d)}{\sqrt{t}} \frac{1}{C_1 C_{\min} t^{-1/4} - C_2 t^{-5/8}\sqrt{\log(Md/\alpha) + (\log\log t)_+}} \\
&\leq \sum_{t=1}^{n} \frac{\bar{C}(M, \gamma, d)}{\sqrt{t}} \frac{C_3}{C_{\min} t^{-1/4}}\left(1 + \frac{C_2 t^{-5/8}\sqrt{\log(Md/\alpha) + (\log\log t)_+}}{C_1 C_{\min} t^{-1/4}}\right) \\
&\leq \sum_{t=1}^{n} \frac{C_3 \bar{C}(M, \gamma, d) t^{1/4}}{C_{\min}}\left(1 + C_4 \frac{t^{-3/8}}{C_{\min}}\sqrt{\log(Md/\alpha) + (\log\log t)_+}\right) \\
&= \frac{C_3 \bar{C}(M, \gamma, d) n^{3/4}}{C_{\min}}\left(1 + C_4 \frac{n^{-3/8}}{C_{\min}}\sqrt{\log(Md/\alpha) + (\log\log n)_+}\right).
\end{aligned}$$

□

Lemma 20 (Anytime Bound on Variance Term) *Suppose γ_t is chosen according to Lemma 19, $\gamma_t = O(t^{-1/4})$ and $\gamma_t = O(C_{\min} t^{-1/2}/\bar{C}(M, \gamma, d))$. Then with probability greater than $1 - \alpha$, the following holds simultaneously for all $n \geq 1$ and $t \geq 1$*

$$\hat{r}_{t,j} \leq \frac{4\sqrt{10}\, t}{\phi^2(\gamma_t, 2)} + B, \qquad \forall j \in [M]$$

$$\sum_{t=1}^{n}\sum_{j=1}^{M} q_{t,j}\hat{r}_{t,j}^2 \leq C_1 B^2 C_{\min}\sqrt{n} + C_2 B n^{1/4}\left(1 + \frac{n^{-3/8}}{C_{\min}}\sqrt{\log(Md/\alpha) + (\log\log n)_+}\right) + \bar{C}(M, \gamma, d)\frac{\log n}{C_{\min}}\left(1 + \frac{n^{-3/8}}{C_{\min}}\sqrt{\log(Md/\alpha) + (\log\log n)_+}\right)$$

where C_i are absolute constants, and $\bar{C}(M, \gamma, d)$ is as defined in Lemma 19, up to constant factors.

Proof of Lemma 20. We start by upper bounding $\hat{r}_{t,j}$. For all j and t it holds that:

$$\hat{r}_{t,j} = \int_{\mathcal{X}} |\hat{h}_t(\gamma, x) - h_{t+1,j}(x)|\, dp_{t+1,j}(x) \leq \int_{\mathcal{X}} \|\hat{\beta}_t\|_2 \|x\|_2\, dp_{t+1,j}(x) \leq \|\hat{\beta}_t\|_2 B + (\|\hat{\beta}_t\|_2 - \|\beta_j\|_2)B$$

since $\|x\|_2 \leq B$. To bound the last term, we only need to invoke Theorem 3, which, in turn, will simultaneously bound $\hat{r}_{t,j}$ for all $j = 1, \ldots, M$:

$$P\left(\forall t \geq 1, \forall j \in [M] : \hat{r}_{t,j} \leq \frac{4\sqrt{10}\, t}{c_{\gamma,t}^2} + B\right) \geq 1 - \alpha$$

Which implies for all $t \geq 1$,

$$\sum_{j=1}^{M} q_{t,j} \hat{\Delta}_{t,j}^2 \left(\frac{4\sqrt{10}\, t}{c_{t,t}^2} + B \right)^2 \sum_{j=1}^{M} q_{t,j} = \frac{160\, t^2}{c_{t,t}^4} + B^2 + \frac{8B\sqrt{10}\, t}{c_{t,t}^2}.$$

For the last term, similar to the proof of Lemma 19 we have,

$$\sum_{t=1}^{n} t \frac{8B\sqrt{10}\, t}{c_{t,t}^2} \leq C_1 B n^{1/4} \left(1 + \frac{n}{C_{\min}}\right)^{3/8} \sqrt{\log(M/\delta) + (\log\log n)_+}$$

for some absolute constant C_1. We treat the squared term similarly,

$$\begin{aligned}\sum_{t=1}^{n} t \frac{160\, t^2}{c_{t,t}^4} &\leq C(M,\gamma,d) \sum_{t=1}^{n} \frac{C_3}{t C_{\min}} \left(1 + C_4 \frac{t}{C_{\min}}\right)^{3/8} \sqrt{\log(M/\delta) + (\log\log t)_+} \\ &\leq C(M,\gamma,d) \frac{C_3 \log n}{C_{\min}} \left(1 + C_4 \frac{n}{C_{\min}}\right)^{3/8} \sqrt{\log(M/\delta) + (\log\log n)_+}\,.\end{aligned}$$

Note that the last inequality is not tight. This term will not be fastest growing term in the regret, so we have little motivation to bound it tightly. Therefore,

$$\begin{aligned}\sum_{t=1}^{n} t \sum_{j=1}^{M} q_{t,j} \hat{\Delta}_{t,j}^2 &\leq C_1 B^2 C_{\min} \sqrt{n} + C_2 B n^{1/4} \left(1 + \frac{n}{C_{\min}}\right)^{3/8} \sqrt{\log(M/\delta) + (\log\log t)_+} \\ &\quad + C(M,\gamma,d) \frac{\log n}{C_{\min}} \left(1 + C_4 \frac{n}{C_{\min}}\right)^{3/8} \sqrt{\log(M/\delta) + (\log\log n)_+}\end{aligned}$$

where C_i are absolute constants. □

Proof of Lemma 16 We start by conditioning on the event E that $\hat{\Delta}_t$ is picked such that $t\hat{\Delta}_{t,j} \leq 1$ for all $t \geq 1$ and $j = 1,\ldots,M$. Then by application of Lemma 18 we get with probability greater than $1 - 2\delta/3$,

$$\begin{aligned}R(n,\gamma) \leq 2B \Bigg(&\sum_{t=1}^{n} t + \frac{\log M}{n} + \sum_{t=1}^{n} \sum_{j=1}^{M} q_{t,j} \hat{\Delta}_{t,j}^2 + \sum_{t=1}^{n} (\hat{\Delta}_{t,i} + \sum_{j=1}^{M} q_{t,j} \hat{\Delta}_{t,j}) \\ &+ 10B \sqrt{\frac{\log(12/\delta)}{n\left((\log\log n B^2)_+ + \log(12/\delta)\right)}} \Bigg)\end{aligned}$$

We invoke Lemma 19 and Lemma 20 with $\delta = \delta/3$ take a union bound, to bound the variance and $\hat{\Delta}_{t,i}$ terms as well. These lemmas require one application of Theorem 3 to hold simultaneously and no additional union bound is required between them, since the randomness comes only from the confidence interval over $\hat{\Delta}_t$.

$$\begin{aligned}R(n,\gamma) \leq\ & C_1 B n^{3/4} + \frac{\log M}{n} \\ &+ C_2 B^2 C_{\min} \sqrt{n} + C_3 B n^{1/4} \left(1 + \frac{n}{C_{\min}}\right)^{3/8} \sqrt{\log(M/\delta) + (\log\log n)_+} \\ &+ C(M,\gamma,d) \frac{\log n}{C_{\min}} \left(1 + \frac{n}{C_{\min}}\right)^{3/8} \sqrt{\log(M/\delta) + (\log\log n)_+} \\ &+ n^{3/4} C(M,\gamma,d) \left(1 + \frac{n}{C_{\min}}\right)^{3/8} \sqrt{\log(M/\delta) + (\log\log n)_+} \\ &+ 10B \sqrt{\frac{\log(12/\delta)}{n\left((\log\log n B^2)_+ + \log(12/\delta)\right)}}\end{aligned}$$

with probability greater than $1 - \delta$, conditioned on event E. Assuming that event E happens with probability $1 - 2\delta/3$, let $B = B(\gamma, M, d, n, B, \ldots, C_{\min})$ denote the right-hand-side of the regret inequality above.

By the chain rule we may write,

$$P\,\mathrm{Reg}(n;i) \;\le\; B$$

$$P\;\; R(n;i) \;\le\; B \;\;\; 8t \;2\; [n];j \;2\; [M] : \;{}_t\hat{}_{t;j} \;\;\; 1 \;\; P(8t \;2\; [n];j \;2\; [M] : \;{}_t\hat{}_{t;j} \;\;\; 1)$$

$$P\;\; R(n;i) \;\le\; B \;\;\; 8t \;2\; [n];j \;2\; [M] : \;{}_t\hat{}_{t;j} \;\;\; 1 \;\; (1 \;\;\; 2\;)$$

$$(1 \;\;\;)(1 \;\;\; 2\;) \;\;\; 1 \;\;\; 3 :$$

It remains to verify that event E is met with probability $1 \;\;\; 2$. Recall that ${}_t = O(C_{\min} = \sqrt{t}C(M;\;;d))$, and that from Lemma 8 with probability $1 \;\;\; $,

$$\frac{C_{\min}}{4\sqrt{t}C(M;\;;d)} \;\;\; \frac{C_1 C_{\min}\, t^{1=4} \;\;\; C_2 t^{\;1=8}\sqrt{\log(M d=\;) + (\log\log t)_+}}{4\sqrt{10}C(M;\;;d)} \;\;\; \frac{^2(\;{}_t;2)}{4\sqrt{10}\;{}_t}$$

Therefore, from Lemma 20, there exists C such that ${}_t = C\; C_{\min}=B\sqrt{t}C(M;\;;d)$ satisfying,

$$P(8t \;\;\; 1;j \;2\; [M] : \;{}_t\hat{}_{t;j} \;\;\; 1) \;\;\; 1 \;\;\; 2$$

The proof is then nished by setting 3 (and updating the absolute constants). □

D.2 Proof of Virtual Regret

Proposition 21. For any xed $\tilde{} > 0$, there exists an absolute constant C_1 such that

$$P\;\; 8t \;\;\; 1 : \;\hat{}_{t;j\,?} \;\;\; {}_{j\,?} \;{}_2 \;\;\; !\,(t;\;;d) \;\;\; 1 \;\;\; :$$

where

$$!\,(t;\;;d) := C_1 \sqrt{\frac{^2 d\log\frac{t}{d} +1 \;\;\; +2\;\; ^2\log(1=\;)+\;\tilde{}B\;^2}{\tilde{}+\;C_{\min}\, t^{\,3=4}}}\;\; 1 + C_{\min}^{\;1}\, t^{\;\;3=8}\sqrt{\log(M d=\;) + (\log\log t)_+} \;\;:$$

Moreover, for $u_{t;j\,?}(\;) := \;\hat{}_{t;j\,?} \;{}_{j\,?}(\;) + \;!\,(t;\;;d)$,

$$P(8t \;\;\; 1;\; x \;2\; X \;:\; r(x) \;\;\; u_{t;j\,?}(x)) \;\;\; 1 \;\;\; :$$

Proof of Proposition 21. De ne for convenience $V_t = \;\sum_{t;j\,?} \;{}_{t;j\,?} + \;\tilde{}I$. We rst observe that

$$\hat{}_{t;j\,?} = V_t^{\;\;1}(\;{}_{t;j\,?})^{>}\, y_t$$

We can apply results from Abbasi-Yadkori et al. [2011] to get an anytime-valid con dence set. Their Theorem 2 asserts that with probability $1 \;\;\; $, for all $t \;\;\; 1$ we have[2]

$$\hat{}_{t;j\,?} \;\;\; {}_{j\,?} \;{}^2_{V_t} \;\;\; {}_t$$

where

$${}_t = 2\;\; ^2\log\;\frac{\det(V_t)^{1=2}}{\det(\;\tilde{}I\;)^{1=2}} \;\;\; + \;\tilde{}B\;^2$$

Clearly, $V_t \;\;\; {}_{\min}(V_t)I$, and therefore with high probability,

$$\hat{}_{t;j\,?} \;\;\; {}_{j\,?} \;{}_2 \;\;\; \sqrt{\frac{{}_t}{{}_{\min}(V_t)}}$$

uniformly over time. Our assumption is that ${}_j(x)k \;\;\; 1$, and hence, denoting by $_i$ the eigenvalues of V_t, the geometric-arithmetic mean inequality yields

$$\det(V_t) \;\; \prod_{i=1}^{d} {}_i \;\;\; \frac{1}{d}\mathrm{trace}(V_t) \;\;^d :$$

[2] Their theorem statement is slightly different, but they prove the stronger version we state below.

Given that

$$\mathrm{trace}(V_t) = \sum_{i=1}^{d}\sum_{s=1}^{t} \left(\phi_j(x)\right)_i^2 + \tilde{d} \qquad t + \tilde{d}$$

we can conclude that

$$\sum_{t} \quad 2\ ^2 \log\left(\frac{(t=d+\tilde{})^{d=2}}{\tilde{}^{d=2}}\right) \quad + \tilde{B}^{\ 2} = d\ ^2 \log\left(\frac{t}{\tilde{d}} + 1\right) \quad + 2\ ^2 \log(1=\ ) + \tilde{B}^{\ 2}$$

We note that

$$\lambda_{\min}(V_t) = \lambda_{\min}\left(\sum_{t;j}^{>} \phi_{t;j}\right) + \tilde{}:$$

Then due to Lemma 9, there exist absolute constants C_1 and C_2 such that for all $t \quad 1$,

$$\lambda_{\min}(V_t) \quad \tilde{} + C_1 C_{\min}\, t^{3=4} \quad C_2 t^{3=8} \sqrt{\log(M d=\ ) + (\log\log t)_+}$$

therefore, there exist C_3 and C_4 such that

$$\begin{aligned}
&\sqrt{\frac{1}{\lambda_{\min}(V_t)}} \quad \sqrt{\frac{1}{\tilde{} + C_1 C_{\min}\, t^{3=4} \quad C_2 t^{3=8}\sqrt{\log(M d=\ ) + (\log\log t)_+}}} \\
&\sqrt{\frac{C_3}{\tilde{} + C_1 C_{\min}\, t^{3=4}}} \quad 1 + \frac{t^{3=8}\sqrt{\log(M d=\ ) + (\log\log t)_+}}{\tilde{} + C_1 C_{\min}\, t^{3=4}}\ ! \\
&\sqrt{\frac{C_4}{\tilde{} + C_1 C_{\min}\, t^{3=4}}} \quad 1 + C_{\min}^{1}\, t^{\ 3=8}\sqrt{\log(M d=\ ) + (\log\log t)_+}
\end{aligned}$$

with high probability for all $t \quad 1$. Setting

$$!\,(t;\ ;d) = C_5 \sqrt{\frac{^2 d \log\left(\frac{t}{\tilde{d}}+1\right) + 2\ ^2 \log(1=\ ) + \tilde{B}^{\ 2}}{\tilde{} + C_{\min}\, t^{3=4}}} \quad 1 + C_{\min}^{1}\, t^{\ 3=8}\sqrt{\log(M d=\ ) + (\log\log t)_+}$$

where C_5 is an absolute constant concludes the parametric con dence bound. The upper con dence bound then simply follows: for any $x \ 2\ X$

$$r(x) \quad \hat{}^{>}_{t;j\ ?}\ \phi_{j\ ?}(x) = \left\langle \phi_{j\ ?}\ \hat{}_{t;j\ ?};\ \phi_{j\ ?}(x)\right\rangle \qquad \phi_{j\ ?}\ \hat{}_{t;j\ ?}\ _2 k\ _{j\ ?}(x)k_2 \quad !\,(t;\ ;d)$$

where the last inequality holds with high probability simultaneously for all t 1. □

Proof of Lemma 15 Using Proposition 21 and the Cauchy-Schwarz inequality we obtain,

$$\begin{aligned}
R_{j\ ?}(n) &= \sum_{t=1}^{n} r(x^?) \quad r(x_{t;j}) \\
&= \sum_{t=1}^{n} r(x^?) \quad \hat{r}_t(x^?) + \hat{r}_t(x^?) \quad \hat{r}_t(x_{t;j}) + \hat{r}_t(x_{t;j}) \quad r(x_{t;j}) \\
&\sum_{t=1}^{n} \quad \phi_j \quad \hat{}_{t;j}\ _2 \ k\ _j(x^?)k_2 + k\ _j(x_{t;j})k_2 \ + \hat{r}_t(x^?) \quad \hat{r}_t(x_{t;j}) \\
&\sum_{t=1}^{n} \quad !\,(t;\ ;d) \ k\ _j(x^?)k_2 + k\ _j(x_{t;j})k_2 \ + \hat{r}_t(x^?) \quad \hat{r}_t(x_{t;j})
\end{aligned}$$

with probability 1 . If the agent selects actions greedily, then $\hat{r}_t(x^?) \quad \hat{r}_t(x_{t;j})$, and

$$R_{j\ ?}(n) \quad \sum_{t=1}^{n} \quad !\,(t;\ ;d) \ k\ _j(x^?)k_2 + k\ _j(x_{t;j})k_2 \quad \sum_{t=1}^{n} \quad 2!\,(t;\ ;d)$$

since the feature map is normalized to satisfy $\|\phi_j(\cdot)\|_k$ 1. If the agent selects actions optimistically according to the upper con dence bound of Proposition 21, then

$$\hat{r}_t(x_{t;j}) + !\,(t;\ ;d)k\ _j(x_{t;j})k \quad \hat{r}_t(x^?) + !\,(t;\ ;d)k\ _j(x^?)k$$

which implies

$$\hat{r}_t(x^?) \quad \hat{r}_t(x_{t;j}) \quad !(t;;d)k_{\ j}(x_{t;j})k \quad !(t;;d)k_{\ j}(x^?)k$$

and therefore,

$$R_{j^?}(n) \quad \sum_{t=1}^{n} 2!(t;;d)k_{\ j}(x_{t;j})k_2 \quad \sum_{t=1}^{n} 2!(t;;d):$$

Then due to Proposition 21, with probability greater than , simultaneously for all $n \quad 1$,

$$\sum_{t=1}^{n} !(t;;d) \quad \sum_{t=1}^{n} C_1 \sqrt{\frac{^2 d \log \frac{t}{d} + 1 \quad + 2 \quad ^2 \log(1 = \ ) + \ B \ ^2}{\sim + C_{min} \ t^{3=4}}} \quad 1 + t^{\ 3=8} \sqrt{\frac{}{\log(\frac{Md}{\ }) + (\log\log \ t)_+}}$$

$$C_1 n^{5=8} \sqrt{\frac{^2 d \log \frac{n}{d} + 1 \quad + 2 \quad ^2 \log(1 = \ ) + \ B \ ^2}{C_{min}}} \quad 1 + C_{min}^{\ 1} n^{\frac{3}{8}} \sqrt{\frac{}{\log(\frac{Md}{\ }) + (\log\log \ n)_+}}$$

concluding the proof. □

E Time-Uniform Concentration Inequalities

We will make use of the elegant concentration results in Howard et al. [2021], which analyzes the boundary of sub-Gamma processes.

Definition 22 (Sub-Gamma process)**.** Let $(S_t)_{t=0}^{1}$ and $(V_t)_{t=0}^{1}$ be real-valued processes adapted to $(F_t)_{t=1}^{1}$ with $S_0 = V_0 = 0$ and V_t non-negative. We say that S_t is sub-Gamma if for $2 [0; 1=c)$, there exists a supermartingale $(M_t( ))_{t=0}^{1}$ w.r.t. F_t, such that $E\, M_0 = 1$ and for all $t \quad 1$:

$$\exp\{ S_t \quad \frac{^2}{2(1 \quad c \ )} V_t \} \quad M_t( ) \qquad a.s.$$

The following is a special case of Theorem 1 in Howard et al. [2021]. We have simplified it by making a few straightforward choices for the parameters used originally by Howard et al. [2021], which will yield an easier-to-use bound in our scenario.

Proposition 23 (Curved Boundary of Sub-Gamma Processes)**.** Let $(S_t)_{t \quad 0}$ be sub-Gamma with scale parameter c and variance process $(V_t)_{t \quad 0}$. Define the boundary

$$B \ (v) := \frac{5}{2} \sqrt{\max\{v; 1\} \quad (\log\log ev)_+ + \log \quad \frac{2}{\ }} \quad + 3c \quad (\log\log ev)_+ + \log \quad \frac{2}{\ } \quad ;$$

for $v > 0$, where $(x)_+ = \max(0; x)$. Then,

$$P(\exists t : S_t \quad B \quad (V_t)) \qquad :$$

Proof of Proposition 23. Suppose $( )$ denotes the Riemann zeta function. Theorem 1 in Howard et al. [2021] states that if $(S_t)_{t \quad 0}$ is a sub-Gamma process with variance process $(V_t)_{t \quad 0}$ then the boundary

$$S \ (v^0) = k_1 \quad v^0 \quad s \log\log(v^0) + \log \quad \frac{(s)}{\log^s} \quad + ck_2 \quad s \log\log(v^0) + \log \quad \frac{(s)}{\log^s} \quad :$$

satisfies,

$$P(\exists t : S_t \quad S \quad (\max(V_t; 1)))$$

where

$$k_1 := \frac{^{1=4} + \sqrt{}^{1=4}}{\sqrt{2}} \qquad \text{and} \qquad k_2 := (\sqrt{} + 1) =2$$

and $s; \quad 1$. Choosing $s = 2$ and $= e$, we obtain $(2) = \ ^2=6 \quad 2$. Furthermore, we have $k_1 \quad \frac{3}{2}$ and $k_2 \quad \frac{3}{2}$. Then if $v^0 \quad 1$ (which we will enforce by the construction $v^0 = \max(1; v)$), we compute

$$s \log\log(v^0) + \log \quad \frac{(s)}{\log^s} \quad 2(\log\log ev^0)_+ + \log \quad \frac{2}{\ } \quad :$$

Therefore, we can upper bound (using our bounds on k_1, k_2)

$$S(v^0) \le \frac{5}{2}\sqrt{v^0\left((\log\log ev^0)_+ + \log\frac{2}{\delta}\right)} + 3c\left((\log\log ev^0)_+ + \log\frac{2}{\delta}\right).$$

Now, since the boundary is given by $S(\max(v,1))$ and $v^0 = \max(v,1)$ we deduce that

$$B(v) := \frac{5}{2}\sqrt{\max\{v,1\}\left((\log\log ev)_+ + \log\frac{2}{\delta}\right)} + 3c\left((\log\log ev)_+ + \log\frac{2}{\delta}\right).$$

is an any-time valid boundary. □

Lemma 24 (Time-Uniform Two-sided Bernoulli) Let $X_1, \dots, X_s, \dots, X_t$ be a martingale sequence of Bernoulli random variables with conditional mean μ_s. Then for all $\delta > 0$,

$$P\left(\exists t : \sum_{s=1}^{t}(X_s - \mu_s) \ge \frac{5}{2}\sqrt{t\left((\log\log t)_+ + \log(4/\delta)\right)}\right) \le \delta;$$

Proof of Lemma 24 By Proposition 23, we know that S_t is sub-Gamma with variance process V_t and scale parameter c, then

$$P(\exists t : S_t \ge B(V_t)) \le \delta;$$

where

$$B(v) := \frac{5}{2}\sqrt{\max\{1, v\}\left((\log\log ev)_+ + \log(2/\delta)\right)} + 3c\left((\log\log ev)_+ + \log(2/\delta)\right).$$

By Howard et al. [2020], we know that if $(X_t)_{t=1}^{\infty}$ is a Bernoulli sequence, then $S_t = \sum_{s=1}^{t}(X_s - \mu_s)$ is sub-Gamma with variance process $V_t = t$ and scale parameter $c = 0$ (hence, sub-Gaussian). This implies,

$$P\left(\exists t : \sum_{s=1}^{t}(X_s - \mu_s) \ge \frac{5}{2}\sqrt{t\left((\log\log t)_+ + \log(2/\delta)\right)}\right) \le \delta;$$

The above arguments also holds for the sequence $Z_s = -X_s$. Then taking a union bound and adjusting $\delta = 2\delta$ concludes the proof. □

Lemma 25 (Time-Uniform Bernstein) Let $(\eta_i)_{i=1}^{\infty}$ be a sequence of conditionally standard sub-gaussian variables, where each η_i is $\mathcal{F}_{i-1} = \sigma(\eta_1, \dots, \eta_{i-1})$ measurable. Then, for $v_i \in \mathbb{R}$ and $\delta \in (0, 1]$

$$P\left(\exists t : \sum_{i=1}^{t}(\eta_i^2 - 1)v_i \ge \frac{5}{2}\sqrt{\max\left\{1, 4\|v_t\|_2^2\right\}\left(\iota + \frac{1}{2}\iota(\|v_t\|_2) + 12\iota(\|v_t\|_2)\max_{i-1} v_i\right)}\right) \le \delta$$

where, $v_t = (v_1, \dots, v_t) \in \mathbb{R}^t$ and $\iota(v) := \log\log(4ev^2)_+ + \log(2/\delta)$.

Proof of Lemma 25 From Lemma 4, $S_t = \sum_{i=1}^{t}(\eta_i^2 - 1)v_i$ is sub-Gamma with variance process $V_t = 4\sum_{i=1}^{t} v_i^2$ and $c = 4\max_{i-1} v_i$. By Proposition 23, we know that S_t is sub-Gamma with variance process V_t and scale parameter c, then

$$P(\exists t : S_t \ge B(V_t)) \le \delta;$$

where

$$B(v) := \frac{5}{2}\sqrt{\max\{1, v\}\left((\log\log ev)_+ + \log(2/\delta)\right)} + 3c\left((\log\log ev)_+ + \log(2/\delta)\right).$$

□

Lemma 26 (Time-Uniform Azuma-Hoeffding) Let $X_1, \dots, X_n$ be a martingale difference sequence such that $|X_t| \le B$ for all $t \ge 1$ almost surely. Then for all $\delta > 0$,

$$P\left(\exists t : \sum_{s=1}^{t} X_s \ge \frac{5B}{2}\sqrt{t\left((\log\log etB^2)_+ + \log(2/\delta)\right)}\right) \le \delta;$$

Proof of Lemma 26. By Proposition 23, we know that if S_t is sub-Gamma with variance process V_t and scale parameter c, then

$$\mathsf{P}\left(\forall t:\ S_t \quad B_\delta(V_t)\right) \quad \delta,$$

where

$$B_\delta(v) := \frac{5}{2}\sqrt{\max\{1, v\}\left((\log\log ev)_+ + \log(2/\delta)\right)} + 3c\left((\log\log ev)_+ + \log(2/\delta)\right).$$

By [Howard et al., 2020], we know that if $(X_t)_{t=1}^{\infty}$ is B-bounded martingale difference sequence, then $S_t = \sum_{s=1}^{t} X_s$ is sub-Gamma with variance process $V_t = tB^2$ and scale parameter $c = 0$. This implies,

$$\mathsf{P}\quad \forall t:\ S_t \quad \frac{5B}{2}\sqrt{t\left((\log\log etB^2)_+ + \log(2/\delta)\right)} \qquad \delta,$$

concluding the proof. □

F Experiment Details

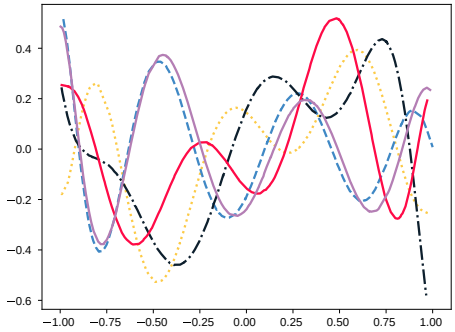


Figure 5: Examples of possible reward functions $r(\,)$ in our experiments.

F.1 Hyper-Parameter Tuning Results

We implement 6 algorithms in our experiments, ETC [Algorithm 4, Hao et al., 2020], ETS (Algorithm 5), CORRAL [Algorithm 6, Agarwal et al., 2017], ALEXP (Algorithm 1), and Lastly UCB (Algorithm 3) with the oracle feature map $_{j\star}$ (Oracle), and UCB with the concatenated feature map (Naive). The Python code is available on github.com/lasgroup/ALEXP. When algorithms require exploration, e.g., in the case of ETC or ALEXP, we simply set $\pi = \mathrm{Unif}(\mathcal{X})$. Figure 7 shows the results of our hyperparameter tuning experiment. To ensure that the curves are valid, we run each

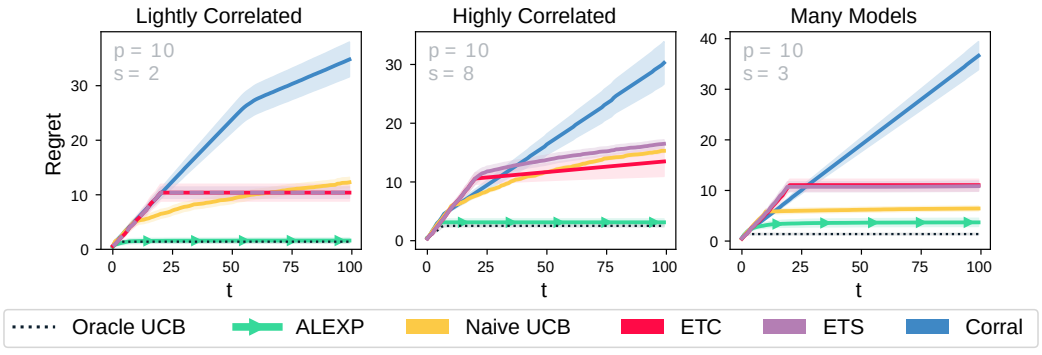


Figure 6: Bench-marking ALEXP and other baselines. Complete version of Fig. 1 and Fig. 2.

Algorithm 2 GetPosterior

Inputs: H_t, , $\tilde{\lambda}$
Let K_t [$^{>}(x_i)$ $(x_j)]_{i,j}$ $_t$, and V_t $(K_t + \tilde{\lambda}^2 I)$, and $k(\ )$ [$^{>}(x_i)$ $(\ )]_i$ $_t$

Calculate $\mu_t(\ )$ $k^T(\ )V_t$ $^1 y_t$

Calculate $\sigma_t(\ )$ $\sqrt{\ ^{>}(\ )\ (\ )\ k^{>}(\ )V_t\ ^1 k(\ )}$

Return: μ_t, σ_t

Algorithm 3 UCB

Inputs: $\tilde{\lambda}$, β_t,
for $t = 1, \ldots, n$ **do**
Choose $x_t \arg\max u_t$ $_1(x) = \mu_t$ $_1(x) + \beta_t \sigma_t$ $_1(x)$. ▷ Choose actions optimistically
Observe $y_t = r(x_t) + \varepsilon_t$. ▷ Receive reward
H_t H_t $_1$ [$f(x_t, y_t)g$ ▷ Append history
Update μ_t, σ_t GetPosterior$(H_t,$, $\tilde{\lambda})$
end for

configuration for 20 different random seeds, i.e. on different random environments. The shaded areas in Figure 7 show the standard error.

UCB. For all the experiments, we set the exploration coefficient of UCB to $\beta_t = 2$[3] and choose the regression regularizer from $\tilde{\lambda}$ 2 $f0.01, 0.1, 0.5g$. We use PyTorch [Paszke et al., 2017] for updating the upper confidence bounds, which requires more regularization for longer feature maps (e.g. when $s = 8$, $p = 2$), to be computationally stable.

Lasso. Every time we need to solve Eq. (1), we set λ_t according to the *rate* suggested by Theorem 3. To find a suitable constant scaling coefficient, we perform a hyper-parameter tuning experiment sampling 20 values in $[10^{-5}, 10^0]$. We choose $\lambda_0 = 0.009$, and scale λ_t with it across all experiments.

ALExp. We set the rates for γ_t and η_t as prescribed by Theorem 1. For the scaling constants, we perform a hyper-parameter tuning experiment log-uniformly sampling 20 different configurations from γ_0 2 $[10^{-4}, 10^{-1}]$ and η_0 2 $[10^0, 10^2]$. For each problem instance (i.e. as s and p change) we repeat this process. However we observe that the optimal hyper-parameters work well across all problem instances.

ETC/ETS. For these algorithms, we separately tune n_0 for each problem instance. We set λ_1 / $\sqrt{\log M/n_0}$ according to Theorem 4.2 of [Hao et al., 2020] and scale it with $\lambda_0 = 0.009$, as stated before. We uniformly sample 10 different values where n_0 2 $[2, 80]$ since the horizon is $n = 100$. The optimal value often happens around $n_0 = 20$.

Corral. We set the rates of the parameters as $\gamma = O(1/n)$ and $\eta = O(\sqrt{M/n})$ according to Agarwal et al. [2017, Theorem 5,]. Then similar to ALExp, we tune the scaling constants. The procedure for tuning the constants is identical to ALExp, as in we use the same search interval, and try 10 different configurations for γ and η.

[3] To achieve the $\sqrt{dT \log T}$ regret, one has to set $_t = O(\sqrt{d \log T})$ as shown in Proposition 21.

Algorithm 4 ETC [Hao et al., 2020]

Inputs: n_0, n, λ_1, π
Let $H_0 = ;$
for $t = 1, \ldots, n_0$ **do**
  Draw $\boldsymbol{x}_t \quad \pi$. ▷ Explore.
  Observe $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$. ▷ Receive reward
  $H_t \quad H_{t \ 1} \ [\ f(\boldsymbol{x}_t, y_t)g$ ▷ Append history
end for
$\hat{\ }_{n_0} \quad L(\ , H_{n_0}, \lambda_1)$ ▷ Perform Lasso once
for $t = n_0 + 1, \ldots, n$ **do**
  Choose $\boldsymbol{x}_t = \arg\max \ \hat{\ }_{n_0}^{>} \ (\boldsymbol{x})$ ▷ Choose actions greedily
end for

Algorithm 5 ETS

Inputs: n_0, n, λ_1, $\tilde{\lambda}$, β_t, π
Let $H_0 = ;$
for $t = 1, \ldots, n_0$ **do**
  Draw $\boldsymbol{x}_t \quad \pi$. ▷ Explore
  Observe $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$. ▷ Receive reward
  $H_t \quad H_{t \ 1} \ [\ f(\boldsymbol{x}_t, y_t)g$ ▷ Append history
end for
$\hat{\ }_{n_0} \quad L(\ , H_{n_0}, \lambda_1)$ ▷ Perform Lasso once
$\hat{J} \quad fj \ \hat{\ }_{n_0,j} \ 6= 0, \ j \ 2 \ [M]g$ ▷ Get sparsity pattern
$\ _{\hat{J}}( ) \quad [\ _j( )]_{j2\hat{J}}$ ▷ Model-select acc. to $\hat{J}$
for $t = n_0 + 1, \ldots, n$ **do**
  Choose $\boldsymbol{x}_t = \arg\max u_{t \ 1}(\boldsymbol{x}) = \mu_{t \ 1}(\boldsymbol{x}) + \beta_t \sigma_{t \ 1}(\boldsymbol{x})$ ▷ Choose actions optimistically
  Observe $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$
  $H_t \quad H_{t \ 1} \ [\ f(\boldsymbol{x}_t, y_t)g$
  Update $\mu_t, \sigma_t \quad \text{GetPosterior}(H_t, \ _{\hat{J}}, \tilde{\lambda})$
end for

Algorithm 6 CORRAL [Agarwal et al., 2017]

Inputs: n, γ, η
Initialize $\beta = e^{1/\ln n}$, $\eta_{1,j} = \eta$, $\rho_{1,j} = 2M$ for all $j = 1, \ldots, M$
Set $\boldsymbol{q}_1 = \bar{\boldsymbol{q}}_1 = \frac{\mathbf{1}}{M}$ and initialize base agents $(p_{1,1}, \ldots, p_{1,M})$.
for $t = 1, \ldots, n$ **do**
  Choose $j_t \quad \bar{\boldsymbol{q}}_t$. ▷ Sample Agent
  Draw $\boldsymbol{x}_t \quad p_{t,j_t}$. ▷ Play action according to agent j_t
  Observe $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$.
  Calculate IW estimates $\hat{r}_{t,j} = \frac{y_t}{\bar{q}_{t;j}} \mathsf{I}\{j = j_t\}$ for all $j = 1, \ldots, M$.
  Send $\hat{r}_{t,j} = \frac{y_t}{\bar{q}_{t;j}} \mathsf{I}\{j = j_t\}$ to agents and get updated policies $p_{t+1,j}$.
  $\boldsymbol{q}_{t+1} = \text{LOG-BARRIER-OMD}(\boldsymbol{q}_t, \hat{r}_{t,j_t} \boldsymbol{e}_{j_t}, \ _t)$ ▷ Update agent probabilities
  $\bar{\boldsymbol{q}}_{t+1} = (1 \quad \gamma)\boldsymbol{q}_{t+1} + \gamma \frac{\mathbf{1}}{M}$ ▷ Mix with exploratory distribution
  for $j = 1, \ldots, M$ **do** ▷ Update parameters
    if $\frac{1}{\bar{q}_{t+1;j}} > \rho_{t,j}$ **then** $\rho_{t+1,j} \quad \frac{2}{\bar{q}_{t;j}}$, and $\eta_{t+1,j} \quad \beta\eta_{t,j}$
    else $\rho_{t+1,j} \quad \rho_{t,j}$ and $\eta_{t+1,j} \quad \eta_{t,j}$
    end if
  end for
end for

Algorithm 7 LOG-BARRIER-OMD

Inputs: $\boldsymbol{q}_t$, $\grave{}_t$, $\ _t$
Find $\xi \ 2 \ [\min_j \ell_{t,j}, \max_j \ell_{t,j}]$ such that $\sum_{j=1}^{M} \ q_{t,j}^{\ 1} + \eta_{t,j}(\ell_{t,j} \quad \xi) \quad ^1 = 1$
Return: $\boldsymbol{q}_{t+1}$ where $q_{t+1,j}^{\ 1} = q_{t,j}^{\ 1} + \eta_{t,j}(\ell_{t,j} \quad \xi)$ for all $j \ 2 \ [M]$