Contents of Appendix

A Extended Literature Review

The sparse linear bandit literature considers linear reward functions of the form $\boldsymbol{\theta}^\top \boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^p$, 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 $\boldsymbol{\phi}_j$ 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 $M = \mathcal{O}(p)$, which is the case when $d \ll p$.

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 $\hat{\boldsymbol{\theta}}_{n_0}^\top \boldsymbol{\phi}(\boldsymbol{x})$. Explore-then-($\epsilon$)Greedy takes a similar approach, however it does not settle on $\hat{\boldsymbol{\theta}}_{n_0}$, rather it continues to update the parameter estimate and select $\boldsymbol{x}_t = \arg\max \hat{\boldsymbol{\theta}}_t^\top \boldsymbol{\phi}(\boldsymbol{x})$. 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 τ in Table 2, following the notation of Hao et al. [2020]. Note that τ 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 $R(n) = \mathcal{O}(\sqrt{dMn})$, 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. $y_t = \boldsymbol{x}^\top(\boldsymbol{\theta} + \boldsymbol{\varepsilon}_t)$. 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 x_i and does less "damage " [Chapter 29.3 Lattimore and Szepesvári, 2020]. In this setting, Carpentier and Munos [2012] present a $\mathcal{O}(d\sqrt{n})$ regret bound.

Recent work considers contextual linear bandits, where at every step $\mathcal{A}_t$, 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 (Definition 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 $\lambda_{\min}(\Phi_t)$ 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. $\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})$, as oppose to linear rewards, i.e. $\boldsymbol{\theta}^\top \boldsymbol{x}$.

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 specific horizon [Bastani and Bayati, 2020]. Oh et al. [2021], which analyzes 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 τ shows existence of other problem-dependent terms which are not constants, and varies across different rows. The regret bounds are simplified 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.	∞	✗	✓	✓	UCB	EXP4 on Sqrd error	$\mathcal{A}$ is compact	$\sqrt{dMn}$, w.h.p.
Foster et al.	K	✓	✓	✗	UCB	EXP4 on Sqrd error	$\lambda_{\min}(\Sigma) \geq c_\lambda$	$(Mn)^{3/4}K^{1/4} + \sqrt{KdMn}$ w.h.p
Carpentier and Munos	∞	✓	✗	✗	UCB	Hard Thresh.	$\mathcal{A}$ is a ball param. noise	$d\sqrt{n}$, w.h.p.
Bastani and Bayati	K	✗	✗	✗	Explore then Greedy	Lasso	$\kappa(\Sigma) > c_K$	$\tau K d^2(\log n + \log M)^2$ w.h.p.
Kim and Paik	K	✗	✗	✗	Explore then ϵ-Greedy	Lasso	$\kappa(\Sigma) > c_K$	$\tau d\sqrt{n}\log(Mn)$, w.h.p.
Oh et al.	K	✓	✓	✗	Greedy	Lasso	$\kappa(\Sigma) > c_\kappa$ + other assums.	$\tau d\sqrt{n\log(Mn)}$ in expectation
Li et al.	K	✓	✗	✗	ETC	Lasso	$\lambda_{\min}(\hat{\Sigma}) > c_\lambda$	$\tau(n^2 d)^{1/3}\sqrt{\log Mn}$ in expectation
Hao et al.	∞	✓	✗	✗	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.	∞	✓	✗	✗	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)	∞	✓	✓	✓	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. Definition 22).

Lemma 4 (Empirical Process is sub-Gamma)**.** *For $t \geq 1$, suppose ξ_t are a sequence conditionally standard sub-Gaussians adapted to the filtration $\mathcal{F}_t = \sigma(\xi_1, \ldots, \xi_t)$. Let $v_t \in \mathbb{R}$, and $Z_t := \xi_t^2 - 1$. Define 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 definition [c.f. Definition 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 $t \geq 1$:

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

We show the above holds in equality by proving that the left hand side itself, is a supermartingale w.r.t. $\mathcal{F}_t$. We define, $M_t(\lambda) \coloneqq \exp\{\lambda S_t - \lambda^2 V_t/2(1-c\lambda)\}$, therefore,

$$\begin{aligned}\mathbb{E}\left[M_t|\mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[\exp\left(\lambda S_{t-1} - \frac{\lambda^2}{2(1-c\lambda)}V_{t-1} + \lambda Z_t v_t - \frac{2\lambda^2 v_t^2}{(1-c\lambda)}\right)|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[M_{t-1}|\mathcal{F}_{t-1}\right]\mathbb{E}\left[\exp\left(\lambda Z_t v_t\right)|\mathcal{F}_{t-1}\right]\exp\left(-\frac{2\lambda^2 v_t^2}{1-c\lambda}\right) \\ &= M_{t-1}\mathbb{E}\left[\exp\left(\lambda Z_t v_t\right)|\mathcal{F}_{t-1}\right]\exp\left(-\frac{2\lambda^2 v_t^2}{1-c\lambda}\right).\end{aligned}$$

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

$$\mathbb{E}\left[\exp\left(\lambda v_t Z_t\right)|\mathcal{F}_{t-1}\right] \leq \exp\left(2\lambda^2 v_t^2\right) \leq \exp\left(\frac{2\lambda^2 v_t^2}{1-c\lambda}\right),$$

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

$$\mathbb{E}\left[M_t|\mathcal{F}_{t-1}\right] \leq M_{t-1}\exp\left(\frac{2\lambda^2 v_t^2}{1-c\lambda}\right)\exp\left(-\frac{2\lambda^2 v_t^2}{1-c\lambda}\right) = M_{t-1}.$$

for $\lambda \in [0, 1/c)$, concluding the proof. □

We now construct a self-normalizing martingale sequence based on 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*

$$A_j = \left\{\forall\, t \geq 1 :\ \left\|(\Phi_t^\top \boldsymbol{\varepsilon}_t)_j\right\|_2/t \leq \lambda_t/2\right\}.$$

Then, for any $0 \leq \delta < 1$*, the event* $A = \cap_{j=1}^M A_j$ *happens with probability* $1-\delta$*, if for all* $t \geq 1$*,*

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

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 ε_i are zero-mean sub-gaussian variables, as driven in Lemma 3.1 [Lounici et al., 2011], it holds that

$$A_j^c = \left\{\exists t :\ \frac{1}{t^2}\boldsymbol{\varepsilon}_t^T \Phi_{t,j}\Phi_{t,j}^\top \boldsymbol{\varepsilon}_t \geq \frac{\lambda^2}{4}\right\} = \left\{\exists t :\ \frac{\sum_{i=1}^t v_i(\xi_i^2 - 1)}{\sqrt{2}\|\boldsymbol{v}_t\|} \geq \alpha_{t,j}\right\}$$

where ξ_i are sub-gaussian variables with variance proxy 1, scalar v_i denotes the i-th eigenvalue of $\Phi_{t,j}\Phi_{t,j}^\top/t$ with the concatenated vector $\boldsymbol{v}_t = (v_1, \ldots, v_t)$, and

$$\alpha_{t,j} = \frac{t^2\lambda^2/(4\sigma^2) - \mathrm{tr}(\Phi_{t,j}^\top\Phi_{t,j})}{\sqrt{2}\left\|\Phi_{t,j}^\top\Phi_{t,j}\right\|_{\mathrm{Fr}}}.$$

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

$$\sqrt{2}\|\boldsymbol{v}_t\|_2\alpha_{t,j} \geq \frac{5}{2}\sqrt{\max\left\{4\|\boldsymbol{v}_t\|_2^2, 1\right\}\omega_{\delta/M}(\|\boldsymbol{v}_t\|_2)} + 12\omega_{\delta/M}(\|\boldsymbol{v}_t\|_2)\max_{t\geq 1} v_t \tag{4}$$

Recall that w.l.o.g. feature maps are bounded everywhere $\|\phi_j(\cdot)\|_2 \leq 1$, and $\mathrm{rank}(\Phi_j) \leq d$ which allows for the following matrix inequalities,

$$\mathrm{tr}(\Phi_{t,j}^\top\Phi_{t,j}) = \mathrm{tr}(\Phi_{t,j}\Phi_{t,j}^\top) = \sum_{i=1}^t \phi_j^\top(x_i)\phi_j(x_i) \leq t$$

$$\left\|\Phi_{t,j}\Phi_{t,j}^\top\right\| \le \operatorname{tr}(\Phi_{t,j}\Phi_{t,j}^\top) \le t$$
$$\left\|\Phi_{t,j}\Phi_{t,j}^\top\right\| \le \left\|\Phi_{t,j}\Phi_{t,j}^\top\right\|_{\mathrm{Fr}} \le \sqrt{d}\left\|\Phi_{t,j}\Phi_{t,j}^\top\right\| \le t\sqrt{d}$$

Therefore,

$$\|\boldsymbol{v}_t\| = \left\|\Phi_{t,j}\Phi_{t,j}^\top\right\|_{\mathrm{Fr}}/t \le \sqrt{d}, \quad \max_{t\ge 1} v_t = \max_{t\ge 1}\left\|\Phi_{t,j}\Phi_{t,j}^\top\right\|/t \le 1.$$

For Eq. (4) to hold, is suffices that for all $t \ge 1$,

$$\lambda \ge \frac{2\sigma}{\sqrt{t}}\sqrt{1 + \frac{5}{2\sqrt{2}}\sqrt{4d\left(\log(2M/\delta) + (\log\log d)_+\right)} + \frac{12}{\sqrt{2}}\left(\log(2M/\delta) + (\log\log d)_+\right)}.$$

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

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

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

$$\frac{1}{t}\left\|\Phi_t(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta})\right\|_2^2 \le \frac{2}{t}\boldsymbol{\varepsilon}_t^T\Phi_t(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}) + 2\lambda_t\sum_{j=1}^M\left(\|\boldsymbol{\theta}_j\|_2 - \left\|\hat{\boldsymbol{\theta}}_{t,j}\right\|_2\right).$$

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

$$\boldsymbol{\varepsilon}_t^T\Phi_t(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}) \le \sum_{j=1}^M\|(\Phi_t^T\boldsymbol{\varepsilon}_t)_j\|_2\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2 \le \frac{t\lambda}{2}\sum_{j=1}^M\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2$$

then adding $\lambda_t\sum_{j=1}^M\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2$ to both sides, applying the triangle inequality, and recalling from Section 3 that $\boldsymbol{\theta}_j = 0$ for $j \ne j^\star$ gives

$$\begin{aligned}\frac{1}{t}\left\|\Phi_t(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta})\right\|_2^2 + \lambda_t\sum_{j=1}^M\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2 &\le 2\lambda_t\sum_{j=1}^M\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2 + 2\lambda_t\sum_{j=1}^M\left(\|\boldsymbol{\theta}_j\|_2 - \left\|\hat{\boldsymbol{\theta}}_{t,j}\right\|_2\right)\\ &\le 4\lambda_t\left\|\hat{\boldsymbol{\theta}}_{t,j^\star} - \boldsymbol{\theta}_{j^\star}\right\|_2.\end{aligned}$$

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

$$\frac{1}{t}\left\|\Phi_t(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta})\right\|_2^2 \le 4\lambda_t\left\|\hat{\boldsymbol{\theta}}_{t,j^\star} - \boldsymbol{\theta}_{j^\star}\right\|_2 \tag{5}$$

$$\sum_{\substack{j=1\\ j\ne j^\star}}^M\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2 \le 3\left\|\hat{\boldsymbol{\theta}}_{t,j^\star} - \boldsymbol{\theta}_{j^\star}\right\|_2 \tag{6}$$

Step 2. Consider a sequence $(c_1, \dots, c_k, \dots)$, where $c_1 \ge \cdots \ge c_k \dots$, then

$$c_k \le \frac{1}{k}(kc_k + \sum_{i>k} c_i) \le \sum_{i\ge 1}\frac{c_i}{k}. \tag{7}$$

Define $J_1 = \{j^\star\}$ and $J_2 = \{j^\star, j'\}$ where

$$j' = \operatorname*{arg\,max}_{\substack{j\in[M]\\ j\ne j^\star}}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2.$$

For any $J \subset [M]$ the complementing set is denoted as $J^c = [M] \setminus J$. For simplicity let $c_j = \left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2$, and let $\pi(k)$ 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,

$$\begin{aligned}\sum_{j\in J_2^c}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2^2 = \sum_{\substack{k>1\\ \pi(k)\in J_1^c}} c_k^2 &\overset{(7)}{\leq} \sum_{\substack{k>1\\ \pi(k)\in J_1^c}} \frac{\left(\sum_{i\in J_1^c} c_i\right)^2}{k^2} \\ &\leq \Big(\sum_{i\in J_1^c} c_i\Big)^2 \sum_{\substack{k>1\\ \pi(k)\in J_1^c}} \frac{1}{k^2} \overset{(6)}{\leq} 9c_{j^\star}^2 \\ &\leq 9(c_{j^\star}^2 + c_{j'}^2) = 9\sum_{j\in J_2}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2^2,\end{aligned}$$

which, in turn, gives

$$\left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\right\|_2 = \sqrt{\sum_{j=1}^{M}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2^2} \leq \sqrt{10\sum_{j\in J_2}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2^2}. \tag{8}$$

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

$$\sum_{j\in J_2^c}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2 \leq 3\sum_{j\in J_2}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2.$$

From the theorem assumptions and Definition 2, we know that there exists $0 < \kappa(\Phi_t, 2)$, therefore by Definition 2, the feature matrix Φ_t satisfies,

$$\begin{aligned}\sum_{j\in J_2}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2^2 &\leq \frac{1}{t\kappa^2(\Phi_t,2)}\left\|\Phi_t(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta})\right\|_2^2 \\ &\overset{(5)}{\leq} \frac{1}{\kappa^2(\Phi_t,2)}4\lambda_t\left\|\hat{\boldsymbol{\theta}}_{t,j^\star} - \boldsymbol{\theta}_{j^\star}\right\|_2 \leq \frac{1}{\kappa^2(\Phi_t,2)}4\lambda_t\sqrt{\sum_{j\in J_2}\left\|\hat{\boldsymbol{\theta}}_{t,j} - \boldsymbol{\theta}_j\right\|_2^2}.\end{aligned}$$

From here, by applying (8) we get,

$$\left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\right\|_2 \leq \frac{4\sqrt{10}\lambda_t}{\kappa^2(\Phi_t,2)}.$$

If λ_t are chosen according to Lemma 5, event 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 $\Phi_t\Phi_t^\top$, as it is later used in our regret analysis. In particular, we show that the feature matrix Φ_t satisfies the restricted eigenvalue condition (Definition 2) required for valid Lasso confidence set (Theorem 3), and calculate a lower bound on $\kappa(\Phi_t, 2)$. 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 $\kappa(\Phi_t, s) = \inf_{b\in\Xi_s}\|\Phi_t\boldsymbol{b}\|_2/\sqrt{t}$ where

$$\Xi_s := \Big\{\boldsymbol{b}\in\mathbb{R}^d\backslash\{0\}\Big|\textstyle\sum_{j\notin J}\|\boldsymbol{b}_j\|_2 \leq 3\sum_{j\in J}\|\boldsymbol{b}_j\|_2,\ \sqrt{\sum_{j\in J}\|\boldsymbol{b}_j\|_2^2} \leq 1 \text{ s.t. } J\subset\{1,\dots,M\},\ |J|\leq s.\Big\}. \tag{9}$$

For simplicity in notation, we further define

$$\tilde{\kappa}(A, s) := \min_{b\in\Xi_s} b^\top Ab. \tag{10}$$

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

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

Lemma 6. *Suppose $\Phi_{\pi,t}$ has t' rows. Then,*

$$\kappa^2(\Phi_t, s) \geq \frac{t'}{t}\kappa^2(\Phi_{\pi,t}, s)$$

Proof of Lemma 6. Let $\Psi^{(t)} = \boldsymbol{\phi}(\boldsymbol{x}_t)\boldsymbol{\phi}^\top(\boldsymbol{x}_t) \in \mathbb{R}^{dM\times dM}$ for all $t = 1, \dots, n$. Note that $\Psi^{(t)}$ is positive semi-definite by construction. We have,

$$\|\Phi_t b\|_2^2 = \sum_{s=1}^{t} b^\top \Psi^{(s)} b = \sum_{s\in T_\pi}^{t} b^\top \Psi^{(s)} b + \sum_{s\notin T_\pi}^{t} b^\top \Psi^{(s)} b$$

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

$$\begin{aligned}
\kappa^2(\Phi_t, s) &= \frac{1}{t}\min_{b\in\Xi_s} \|\Phi_t \boldsymbol{b}\|_2^2 \\
&= \min_{b\in\Xi_s} \frac{1}{t}\sum_{s\in T_\pi} b^\top \Psi^{(s)} b + \frac{1}{t}\sum_{s\notin T_\pi} b^\top \Psi^{(s)} b \\
&\geq \min_{b\in\Xi_s} \frac{1}{t}\sum_{s\in T_\pi} b^\top \Psi^{(s)} b
\end{aligned}$$

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

$$\kappa^2(\Phi_t, s) \geq \min_{b\in\Xi_s} b^\top \left(\frac{1}{t}\sum_{s\in T_\pi}^{t} \Psi^{(s)}\right) b = \frac{|T_\pi|}{t}\kappa^2(\Phi_{\pi,t}, s)$$

□

While the number of rows of $\Phi_{\pi,t}$ is a random variable, we continue to condition on the event that $\Phi_{\pi,t}$ has t' 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 π be the solution to (2), and $s \in \mathbb{N}$. Suppose $\Phi_{\pi,t}$ has t' rows. Then for all $\delta > 0$,*

$$\mathbb{P}\left(\forall t' : \kappa^2(\Phi_{\pi,t}, s) \geq \tilde{\kappa}(\Sigma, s) - \frac{80s}{\sqrt{t'}}\sqrt{\log(2Md/\delta) + (\log\log 4t')_+}\right) \geq 1 - \delta$$

where $\Sigma = \Sigma(\pi, \boldsymbol{\phi}) := \mathbb{E}_{\boldsymbol{x}\sim\pi}\boldsymbol{\phi}(\boldsymbol{x})\boldsymbol{\phi}^\top(\boldsymbol{x})$ and $\tilde{\kappa}$ is defined in (10).

Step III. Remains to combine the two above lemmas and incorporate a high probability bound on t', showing that it is close to $\sum_{s=1}^{t}\gamma_s$.

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

$$\mathbb{P}\left(\forall t \geq 1 : \ \kappa^2(\Phi_t, 2) \geq C_1\tilde{\kappa}(\Sigma, 2)t^{-1/4} - C_2 t^{-5/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\right) \geq 1 - \delta$$

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

$$\mathbb{P}\left(\forall t \geq 1 : \ \kappa^2(\Phi_t, 2) \geq C_1 C_{\min} t^{-1/4} - C_2 t^{-5/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\right) \geq 1 - \delta$$

The regret analysis of Hao et al. [2020] also relies on connecting $\kappa(\Phi_t, s)$ 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 $\kappa^2(\Phi_t, s) \geq C_1 C_{\min}$ with high probability, if $t \geq n_0$ and roughly $n_0 = \mathcal{O}(\sqrt[3]{n^2 \log M})$. We highlight that Lemma 8, presents a lower-bound which holds for all $t \geq 1$, 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 t. 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 $n \gg M$. It is unclear to us if this rate can be improved without assuming knowledge of n, or that $n \geq n_0$.

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

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

$$\lambda_{\min}(\Phi_{t,j}^\top \Phi_{t,j}) \geq C_1 C_{\min} t^{3/4} - C_2 t^{3/8} \sqrt{\log(Md/\delta) + (\log\log t)_+}$$

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

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 $\boldsymbol{\phi}(\boldsymbol{x}) \coloneqq A\boldsymbol{x} : \mathbb{R}^d \to \mathbb{R}^d$ is an invertible linear map, and $\mathcal{X} \in \mathbb{R}^d$ 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 $\boldsymbol{\phi}_j : \mathcal{X} \to \mathbb{R}$ are chosen from an orthogonal basis of $L^2(\mathcal{X})$, and satisfy $\|\boldsymbol{\phi}_i\|_{L^2_\mu(\mathcal{X})}/Vol(\mathcal{X}) \geq 1$. Then there exist absolute constants C_1 and C_2 for which the exploration distribution $\pi = \mathrm{Unif}(\mathcal{X})$ satisfies*

$$\mathbb{P}\left(\forall t \geq 1 :\ \kappa^2(\Phi_t, 2) \geq C_1 t^{-1/4} - C_2 t^{-5/8} \sqrt{\log(Md/\delta) + (\log\log t)_+}\right) \geq 1 - \delta.$$

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 $\boldsymbol{\phi}_j : \mathcal{X} \to \mathbb{R}$ are chosen from an orthogonal basis of $L^2(\mathcal{X})$, and satisfy $\|\boldsymbol{\phi}_i\|_{L^2_\mu(\mathcal{X})}/Vol(\mathcal{X}) \geq 1$. Assume the oracle agent employs a UCB or a Greedy policy, as laid out in Section 5. Choose $\eta_t = \mathcal{O}(1/\sqrt{t}C(M, \delta, d))$ and $\gamma_t = \mathcal{O}(t^{-1/4})$ and $\lambda_t = \mathcal{O}(C(M, \delta, d)/\sqrt{t})$, then* ALExp *with uniform exploration $\pi = \mathrm{Unif}(\mathcal{X})$ attains the regret*

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

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

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

C.2 Proof of Results on Exploration

As an intermediate step, we consider the restricted eigenvalue property of the empirical covariance matrix. Given t' samples, the empirical estimate of Σ is

$$\hat{\Sigma}_{t'} \coloneqq \frac{1}{t'} \sum_{s=1}^{t'} \boldsymbol{\phi}(\boldsymbol{x}_s) \boldsymbol{\phi}^\top(\boldsymbol{x}_s) \tag{11}$$

where $\boldsymbol{x}_s$ are sampled according to π. We show that every entry of $\hat{\Sigma}_{t'}$ 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 $\hat{\Sigma}_{t'}$ be the empirical covariance matrix corresponding to $\Sigma(\pi, \boldsymbol{\phi})$ given t' samples. Then,*

$$\mathbb{P}\left(\exists t' :\ d_\infty(\Sigma, \hat{\Sigma}_{t'}) \geq \frac{5}{\sqrt{t'}}\sqrt{((\log\log 4t')_+ + \log(2Md/\delta))}\right) \leq \delta$$

where $d_\infty(A, B) \coloneqq \max_{i,j}|A_{i,j} - B_{i,j}|$.

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

$$\mathbb{E}[X_s|X_{1:s-1}] = \mathbb{E}X_s = \Sigma(i, j) - \mathbb{E}_{\boldsymbol{x}\sim\pi}\phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x}) = 0$$

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

$$\mathbb{P}\left(\exists t' :\ \frac{1}{t'}\sum_{s=1}^{t'} X_s \geq \frac{5}{\sqrt{t'}}\sqrt{\left((\log\log 4t')_+ + \log(2/\tilde{\delta})\right)}\right) \leq \tilde{\delta}.$$

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 $\hat{\Sigma}_{t'} = \frac{\Phi_{\pi,t}^\top\Phi_{\pi,t}}{t'}$, and thereby

$$\kappa^2(\Phi_{\pi,t}, s) = \min_{b\in\Xi_s} b^\top\hat{\Sigma}_{t'}b = \tilde{\kappa}(\hat{\Sigma}_{t'}, s).$$

Inspired by Lemma 10.1 in van de Geer and Bühlmann [2009], we show that element-wise closeness of matrices Σ and $\hat{\Sigma}_{t'}$ (c.f. Lemma 13) implies closeness in $\tilde{\kappa}$:

$$\begin{aligned}\left|\kappa^2(\Phi_{\pi,t}, s) - \tilde{\kappa}(\Sigma, s)\right| &= \left|\tilde{\kappa}(\hat{\Sigma}_{t'}, s) - \tilde{\kappa}(\Sigma, s)\right| \\ &= \left|\tilde{\kappa}\left(\hat{\Sigma}_{t'} - \Sigma, s\right)\right| \\ &\leq \min_{b\in\Xi_s} d_\infty(\Sigma, \hat{\Sigma}_{t'})\|b\|_1^2\end{aligned}$$

where the last line holds due to Hölder's. Moreover, since $b \in \Xi_s$, for any $J \subset [dM]$ 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,

$$\kappa^2(\Phi_{\pi,t}, s) \geq \tilde{\kappa}(\Sigma, s) - 16sd_\infty(\Sigma, \hat{\Sigma}_{t'}).$$

Therefore by Lemma 13,

$$\kappa^2(\Phi_{\pi,t}, s) \geq \tilde{\kappa}(\Sigma, s) - \frac{80s}{\sqrt{t'}}\sqrt{((\log\log 4t')_+ + \log(2Md/\delta))} \tag{12}$$

with probability greater than $1 - \delta$, simultaneously for all $t' \geq 1$. □

Proof of Lemma 8. In Lemma 6 we showed that

$$\kappa^2(\Phi_t, s) \geq \frac{t'}{t}\kappa^2(\Phi_{\pi,t}, s)$$

where t' indicates the number of rows in the exploratory sub-matrix of Φ_t. Recall that $t' = \sum_{s=1}^t \alpha_s$ where α_s are i.i.d Bernoulli random variables with success probability of γ_s. Due to Lemma 24,

$$\mathbb{P}\left(\forall t \geq 1 :\ |t' - \Gamma_t| \leq \Delta_t\right) \geq 1 - \delta/2 \tag{13}$$

where

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

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

$$\begin{aligned}\kappa^2(\Phi_t, 2) &\geq \frac{t'}{t}\tilde{\kappa}(\Sigma, 2) - \frac{160\sqrt{t'}}{t}\sqrt{(\log\log 4t')_+ + \log(4Md/\delta)} \\ &\geq \frac{\Gamma_t - \Delta_t}{t}\tilde{\kappa}(\Sigma, 2) - 160\sqrt{\frac{\Gamma_t + \Delta_t}{t^2}}\sqrt{\left(\log\log\left(4\Gamma_t + \Delta_t\right)\right)_+ + \log(4Md/\delta)}\end{aligned}$$

where the second inequality holds with probability $1-\delta$, 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 $\gamma_s = \mathcal{O}(s^{-1/4})$ there exists C such that $\Gamma_t = Ct^{3/4}$, then it is straightforward to observe that there exists absolute constants $\tilde{C}_i$ which satisfy,

$$\begin{aligned}\kappa^2(\Phi_t, 2) &\geq \tilde{C}_1 t^{-1/4}\tilde{\kappa}(\Sigma, 2) - \frac{5t^{-3/2}\tilde{\kappa}(\Sigma, 2)}{2}\sqrt{(\log\log t)_+ + \log(8/\delta)} \\ &\quad - \tilde{C}_2 t^{-5/8}\sqrt{\log(Md/\delta) + (\log\log t)_+} - o\left(t^{-5/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\right) \\ &\geq \tilde{C}_1 t^{-1/4}\tilde{\kappa}(\Sigma, 2) - \tilde{C}_3 t^{-5/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\end{aligned}$$

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

$$\mathbb{P}\left(\forall t \geq 1 :\ \kappa^2(\Phi_t, 2) \geq C_1\tilde{\kappa}(\Sigma, 2)t^{-1/4} - C_2 t^{-5/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\right) \geq 1-\delta.$$

If π is chosen according to (2), then $\tilde{\kappa}(\Sigma, 2) \geq C_{\min}$ yielding the lemma's second argument. □

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

$$\Xi_{1,j} = \left\{\boldsymbol{b} \in \mathbb{R}^d \setminus \{0\} \middle| \boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M),\ \text{s.t. } \boldsymbol{b}_j \in \mathbb{R}^d, \|\boldsymbol{b}_j\|_2 \leq 1 \text{ and } \forall j' \neq j :\ \boldsymbol{b}_{j'} = 0\right\}.$$

Note that $\Xi_{1,j} \subset \Xi_s$. Therefore,

$$\inf_{b\in\Xi_{1,j}} \|\Phi_t \boldsymbol{b}\|_2 \geq \inf_{b\in\Xi_s} \|\Phi_t \boldsymbol{b}\|_2 = \sqrt{t}\kappa(\Phi_t, s).$$

Moreover, by construction of $\Xi_{1,j}$ we have for all $\boldsymbol{b} \in \Xi_{1,j}$ that $\Phi_t \boldsymbol{b} = \Phi_{t,j}\boldsymbol{b}_j$, therefore,

$$\inf_{b\in\Xi_{1,j}} \|\Phi_t \boldsymbol{b}\|_2^2 = \inf_{\substack{\boldsymbol{b}_j \in \mathbb{R}^d \\ \|\boldsymbol{b}_j\|_2^2 \leq 1}} \|\Phi_{t,j}\boldsymbol{b}_j\|_2^2 = \lambda_{\min}(\Phi_{t,j}^\top \Phi_{t,j}).$$

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

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

□

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

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

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

$$C_{\min} \geq \lambda_{\min}\left(\Sigma(\pi, \boldsymbol{\phi})\right) \geq \boldsymbol{v}^\top A^\top (T^{-1})^2 A\boldsymbol{v} \geq \|A\boldsymbol{v}\|_2 \lambda_{\min}(T^{-2}) = \frac{\|A\boldsymbol{v}\|_2}{\lambda_{\max}^2(T)} \geq \frac{\lambda_{\min}(A)}{\lambda_{\max}^2(T)}.$$

□

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

$$[\Sigma(\pi, \boldsymbol{\phi})]_{i,i} = \mathbb{E}_{\boldsymbol{x}\sim\pi}\phi_i^2(\boldsymbol{x}) = \frac{1}{\mathrm{Vol}(\mathcal{X})}\int_{\mathcal{X}} \phi_i^2(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x}) \geq 1$$

and for all $i \neq j$,

$$[\Sigma(\pi, \boldsymbol{\phi})]_{i,j} = \mathbb{E}_{\boldsymbol{x}\sim\pi}\phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x}) = \frac{1}{\mathrm{Vol}(\mathcal{X})}\int_{\mathcal{X}} \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x}) = 0$$

We use $\Sigma = \Sigma(\pi, \boldsymbol{\phi})$. For any $\boldsymbol{b} \in \mathbb{R}^{Md}$ where $\|\boldsymbol{b}\| \leq 1$,

$$\begin{aligned}
\boldsymbol{b}^\top\Sigma\boldsymbol{b} = \sum_{i,j\in[M]} \boldsymbol{b}_j^\top\Sigma_{i,j}\boldsymbol{b}_i &= \sum_{i\in[M]} \boldsymbol{b}_i^\top\Sigma_{i,i}\boldsymbol{b}_i + \sum_{i,j\in[M],i\neq j} \boldsymbol{b}_j^\top\Sigma_{i,j}\boldsymbol{b}_i \\
&= \sum_{i\in[M]} \boldsymbol{b}_i^\top\Sigma_{i,i}\boldsymbol{b}_i \\
&\geq 1\sum_{i\in[M]} \|\boldsymbol{b}_i\|_2^2 \geq 1.
\end{aligned}$$

Which implies,

$$\tilde{\kappa}(\Sigma, s) = \min_{\boldsymbol{b}\in\Xi_s} \boldsymbol{b}^\top\Sigma\boldsymbol{b} \geq 1.$$

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

$$\mathbb{P}\left(\forall t \geq 1 :\ \kappa^2(\Phi_t, 2) \geq \tilde{\kappa}(\Sigma, 2)C_1 t^{-1/4} - C_2 t^{-5/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\right) \geq 1 - \delta.$$

concluding the proof.

□

D Proof of Regret Bound

Theorem 14 (Anytime Regret, Formal)**.** *Let $\delta \in (0, 1]$ and π be the maximizer of (2). Assume the oracle agent employs a UCB or a Greedy policy, as laid out in Section 5. Suppose $\eta_t = \mathcal{O}(C_{\min}t^{-1/2}C(M, \delta, d))$ and $\gamma_t = \mathcal{O}(t^{-1/4})$ and $\lambda_t = \mathcal{O}(C(M, \delta, d)t^{-1/2})$, then exists absolute constants $C_1, \ldots, C_6$ for which* ALExp *attains the regret*

$$\begin{aligned}
R(n) \leq\ & C_1 B n^{3/4} + C_2\sqrt{n}C_{\min}^{-1}C(M, \delta, d)\log M \\
&+ C_3 B^2 C_{\min}\sqrt{n} + C_4 B\sqrt{n\left((\log\log nB^2)_+ + \log(1/\delta)\right)} \\
&+ C_5\left(1 + C_{\min}^{-1}n^{-3/8}\sqrt{\log(Md/\delta) + (\log\log n)_+}\right) \\
&\quad\times\left[Bn^{1/4} + (n^{3/4} + \frac{\log n}{C_{\min}})C(M, \delta, d) + \frac{n^{5/8}}{\sqrt{C_{\min}}}\sqrt{d\log n + \log(1/\delta) + B^2}\right]
\end{aligned}$$

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

$$C(M, \delta, d) = C_6\sigma\sqrt{1 + \sqrt{d\left(\log(M/\delta) + (\log\log d)_+\right)} + (\log(M/\delta) + (\log\log d)_+)}.$$

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 $\delta \in (0, 1]$ and $\tilde{\lambda} > 0$. Assume the oracle agent employs a UCB or a Greedy policy, as laid out in Section 5. If $\gamma_t = \mathcal{O}(t^{1/4})$, there exists an absolute constant C_1 for which with probability greater than $1 - \delta$, simultaneously for all $n \geq 1$,*

$$\begin{aligned}
\tilde{R}_{j^\star}(n) = \frac{C_1 n^{5/8}}{\sqrt{C_{\min}}}&\left(1 + n^{-3/8}C_{\min}^{-1}\sqrt{\log(Md/\delta) + (\log\log n)_+}\right) \\
&\times\sqrt{\sigma^2 d\log\left(\tfrac{n}{\tilde{\lambda}d} + 1\right) + 2\sigma^2\log(1/\delta) + \tilde{\lambda}B^2}
\end{aligned}$$

Lemma 16 (Any-Time Model-Selection Regret, Formal)**.** *Let $\delta \in (0,1]$ and π be the maximizer of (2). Suppose $\eta_t = \mathcal{O}(C_{\min}/\sqrt{t}C(M,\delta,d))$ and $\gamma_t = \mathcal{O}(t^{-1/4})$ and $\lambda_t = \mathcal{O}(C(M,\delta,d)/\sqrt{t})$, then exists absolute constants C_i for which* ALExp *attains the model selection regret*

$$\begin{aligned}R(n,j) \le\ & C_1 B n^{3/4} + C_2\sqrt{n}C_{\min}^{-1}C(M,\delta,d)\log M \\ & + C_3 B^2 C_{\min}\sqrt{n} + C_4 B\sqrt{n\left((\log\log nB^2)_+ + \log(1/\delta)\right)} \\ & + C_5\left(Bn^{1/4} + (n^{3/4} + \frac{\log n}{C_{\min}})C(M,\delta,d)\right)\left(1 + C_{\min}^{-1}n^{-3/8}\sqrt{\log(Md/\delta) + (\log\log n)_+}\right)\end{aligned}$$

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

$$C(M,\delta,d) = C_6\sigma\sqrt{1 + \sqrt{d\left(\log(M/\delta) + (\log\log d)_+\right)} + \left(\log(M/\delta) + (\log\log d)_+\right)}.$$

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 $\eta_t\hat{r}_{t,j} \le 1$ for all $1 \le j \le M$ and $t \ge 1$. If the sequence $(\eta_t)_{t\ge1}$ is non-increasing, then for all $n \ge 1$,*

$$\sum_{t=1}^{n}\hat{r}_{t,k} - \sum_{t=1}^{n}\sum_{j=1}^{M}q_{t,j}\hat{r}_{t,j} \le \frac{\log M}{\eta_n} + \sum_{t=1}^{n}\eta_t\sum_{j=1}^{M}q_{t,j}\hat{r}_{t,j}^2$$

for any arm $k \in [M]$.

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

$$\sum_{t=1}^{n}\hat{r}_{t,k} - \sum_{t=1}^{n}\sum_{j=1}^{M}q_{t,j}\hat{r}_{t,j} = \sum_{t=1}^{n}\hat{r}_{t,k} - \sum_{t=1}^{n}\mathbb{E}_{j\sim q_t}[\hat{r}_{t,j}]. \tag{14}$$

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

$$\begin{aligned}-\mathbb{E}_{j\sim q_t}[\hat{r}_{t,j}] = \log(\exp(-\mathbb{E}_{j\sim q_t}[\frac{\eta_t}{\eta_t}\hat{r}_{t,j}])) &= \log(\exp(-\mathbb{E}_{j\sim q_t}[\eta_t\hat{r}_{t,j}])^{1/\eta_t}) \\ &= \frac{1}{\eta_t}\log(\exp(-\mathbb{E}_{j\sim q_t}[\eta_t\hat{r}_{t,j}])) \\ &= \frac{1}{\eta_t}\log(\mathbb{E}_{i\sim q_t}\exp(-\mathbb{E}_{j\sim q_t}[\eta_t\hat{r}_{t,j}])) \end{aligned} \tag{15}$$

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

$$\begin{aligned}\frac{1}{\eta_t}\log\mathbb{E}_{i\sim q_t}\exp(-\eta_t\mathbb{E}_{j\sim q_t}[\hat{r}_{t,j}] + \eta_t\hat{r}_{t,i}) &= \frac{1}{\eta_t}\log\left(\exp(-\eta_t\mathbb{E}_{j\sim q_t}[\hat{r}_{t,j}])\mathbb{E}_{i\sim q_t}\exp(\eta_t\hat{r}_{t,i})\right) \\ &= \frac{1}{\eta_t}\log\mathbb{E}_{i\sim q_t}\exp(-\eta_t\mathbb{E}_{j\sim q_t}[\hat{r}_{t,j}]) \\ &\quad + \frac{1}{\eta_t}\log\mathbb{E}_{i\sim q_t}\exp(\eta_t\hat{r}_{t,i}) \end{aligned} \tag{16}$$

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

$$-\mathbb{E}_{j\sim q_t}[\hat{r}_{t,j}] = \frac{1}{\eta_t}\log\mathbb{E}_{i\sim q_t}\exp(-\eta_t\mathbb{E}_{j\sim q_t}[\hat{r}_{t,j}] + \eta_t\hat{r}_{t,i}) - \frac{1}{\eta_t}\log\mathbb{E}_{i\sim q_t}\exp(\eta_t\hat{r}_{t,i}) \tag{17}$$

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):

$$\log\mathbb{E}_{i\sim q_t}\exp(-\eta_t\mathbb{E}_{j\sim q_t}[\hat{r}_{t,j}] + \eta_t\hat{r}_{t,i}) = \log\mathbb{E}_{i\sim q_t}\exp(\eta_t\hat{r}_{t,i}) - \eta_t\mathbb{E}_{j\sim q_t}\hat{r}_{t,j}$$

$$\begin{aligned}
&\overset{\text{(I)}}{\leq} \mathbb{E}_{i\sim q_t} \exp(\eta_t \hat{r}_{t,i}) - 1 - \eta_t \mathbb{E}_{j\sim q_t} \hat{r}_{t,j} \\
&= \mathbb{E}_{i\sim q_t} \left[\exp(\eta_t \hat{r}_{t,i}) - 1 - \eta_t \hat{r}_{t,i}\right] \\
&\overset{\text{(II)}}{\leq} \mathbb{E}_{i\sim q_t} \left[\eta_t^2 \hat{r}_{t,i}^2\right] \qquad (18)
\end{aligned}$$

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 q_t:

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

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 $J(\eta) = F(-\eta)$. From this, it follows that for any η, we have $J'(\eta) = -F'(-\eta) \leq 0$, by the argument in Bubeck et al. [2012, Theorem 3.1] that shows $F'(\eta) \geq 0$ for any η.

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

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

With this, we rewrite (14) as

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

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

$$\begin{aligned}
\sum_{t=1}^n J_t(\eta_t) - J_{t-1}(\eta_t) \leq J_n(\eta_n) &\leq \frac{\log(M)}{\eta_n} - \frac{1}{\eta_n} \log \left(\sum_{i=1}^M \exp(\eta_n \hat{R}_{n,i}) \right) \\
&\overset{(*)}{\leq} \frac{\log(M)}{\eta_n} - \frac{1}{\eta_n} \log \left(\exp(\eta_n \hat{R}_{n,k}) \right) \\
&= \frac{\log(M)}{\eta_n} - \sum_{t=1}^n \hat{r}_{t,k} \qquad (21)
\end{aligned}$$

where $(*)$ 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

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

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

□

We expressed in Section 5.2, that the model selection regret of ALEXP, 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 η_t is picked such that $\eta_t \hat{r}_{t,j} \leq 1$ for all $1 \leq j \leq M$ and $1 \leq t$ almost surely, then Algorithm 1 satisfies with probability greater than $1 - 2\delta/3$, that simultaneously for all $n \geq 1$*

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

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

Proof of Lemma 18. Let α_t denote the Bernoulli random variable that is equal to 1 if at step t we select actions according to π and 0 otherwise. At each step t with $\alpha_t = 1$ ALEXP accumulates a regret of at most $2B$, since $\|\boldsymbol{\theta}\|_\infty \leq B$ and $\|\boldsymbol{\phi}(\cdot)\| \leq 1$. We can decompose the regret as,

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

For the first term, by Lemma 24, we have

$$2B \sum_{t=1}^{n} \alpha_t \leq 2B \left(\sum_{t=1}^{n} \gamma_t + \frac{5}{2} \sqrt{n\left((\log\log n)_+ + \log(4/\delta_1)\right)} \right).$$

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

$$\begin{aligned}
\sum_{t=1}^{n} (1 - \alpha_t)\Big(r_{t,i} - r_t\Big) &\leq \sum_{t=1}^{n} (1 - \alpha_t)\Big[(r_{t,i} - \hat{r}_{t,i}) + (\hat{r}_{t,i} - \hat{r}_t) + (\hat{r}_t - r_t)\Big] \\
&\leq \sum_{t=1}^{n} \omega_{t,i} + (1 - \alpha_t)\Big[(\hat{r}_{t,i} - \hat{r}_t) + (\hat{r}_t - r_t)\Big]
\end{aligned}$$

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

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

As for the third term,

$$\begin{aligned}
(1 - \alpha_t)(\sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j} - r_t) &= (1 - \alpha_t)\Big[\sum_{j=1}^{M} q_{t,j}(\hat{r}_{t,j} - r_{t,j} + r_{t,j}) - r_t\Big] \\
&\leq \sum_{j=1}^{M} q_{t,j} \omega_{t,j} + (1 - \alpha_t)\left(r_t - \sum_{j=1}^{M} q_{t,j} r_{t,j}\right).
\end{aligned}$$

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

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

The sequence $X_t \coloneqq r_t - \mathbb{E}_{t-1} r_t$ is a martingale difference sequence adapted to the history H_t, since for every $t \geq 1$,

$$\mathbb{E}_{t-1} X_t = \mathbb{E}\left[r_t - \mathbb{E}_{t-1} r_t | H_{t-1}\right] = 0.$$

Since $r_t \leq B$, then $X_t \leq 2B$ almost surely, which allows for an application of anytime Azuma-Hoeffding (Lemma 26):

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

which, in turn, leads us to

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

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

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

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*

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

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

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

where

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

and C is an absolute constant.

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

$$\sum_{t=1}^{n}|\hat{r}_{t,i} - r_{t,i}| = \sum_{t=1}^{n}\left|\int_{\mathcal{X}}(r(\boldsymbol{x}) - \hat{r}_t(\boldsymbol{x}))\mathrm{d}p_{t+1,i}(\boldsymbol{x})\right|$$
$$\leq \sum_{t=1}^{n}\int_{\mathcal{X}}|r(\boldsymbol{x}) - \hat{r}_t(\boldsymbol{x})|\mathrm{d}p_{t+1,i}(\boldsymbol{x})$$
$$\overset{\text{C.S.}}{\leq} \sum_{t=1}^{n}\int_{\mathcal{X}}\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\right\|_2 \|\boldsymbol{\phi}(\boldsymbol{x})\|_2 \mathrm{d}p_{t+1,i}(\boldsymbol{x})$$
$$\overset{\text{bdd. }\phi}{\leq} \sum_{t=1}^{n}\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\right\|_2 \int_{\mathcal{X}}\mathrm{d}p_{t+1,i}(\boldsymbol{x}) = \sum_{t=1}^{n}\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\right\|_2.$$

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

$$\sum_{t=1}^{n} |\hat{r}_{t,i} - r_{t,i}| \leq \sum_{t=1}^{n} \frac{4\sqrt{10}\lambda_t}{\kappa^2(\Phi_t, 2)} = \tilde{C}(M, \delta, d) \sum_{t=1}^{n} \frac{1}{\kappa^2(\Phi_t, 2)\sqrt{t}}$$

where,

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

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

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

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

$$\begin{aligned}
\sum_{t=1}^{n} \frac{\tilde{C}(M, \delta, d)}{\kappa^2(\Phi_t, 2)\sqrt{t}} &\leq \sum_{t=1}^{n} \frac{\tilde{C}(M, \delta, d)}{\sqrt{t}} \frac{1}{C_1 C_{\min} t^{-1/4} - C_2 t^{-5/8} \sqrt{\log(Md/\delta) + (\log\log t)_+}} \\
&\leq \sum_{t=1}^{n} \frac{\tilde{C}(M, \delta, d)}{\sqrt{t}} \frac{\tilde{C}_3}{C_{\min} t^{-1/4}} \left(1 + \frac{C_2 t^{-5/8} \sqrt{\log(Md/\delta) + (\log\log t)_+}}{C_1 C_{\min} t^{-1/4}}\right) \\
&\leq \sum_{t=1}^{n} \frac{\tilde{C}_3 \tilde{C}(M, \delta, d) t^{-1/4}}{C_{\min}} \left(1 + \tilde{C}_4 \frac{t^{-3/8}}{C_{\min}} \sqrt{\log(Md/\delta) + (\log\log t)_+}\right) \\
&= \frac{C_3 \tilde{C}(M, \delta, d) n^{3/4}}{C_{\min}} \left(1 + C_4 \frac{n^{-3/8}}{C_{\min}} \sqrt{\log(Md/\delta) + (\log\log n)_+}\right).
\end{aligned}$$

□

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

$$\begin{aligned}
\hat{r}_{t,j} &\leq \frac{4\sqrt{10}\lambda_t}{\kappa^2(\Phi_t, 2)} + B, \qquad \forall j \in [M] \\
\sum_{t=1}^{n} \eta_t \sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j}^2 &\leq C_1 B^2 C_{\min} \sqrt{n} + C_2 B n^{1/4} \left(1 + \frac{n^{-3/8}}{C_{\min}} \sqrt{\log(Md/\delta) + (\log\log n)_+}\right) \\
&\quad + C(M, \delta, d) \frac{\log n}{C_{\min}} \left(1 + \frac{n^{-3/8}}{C_{\min}} \sqrt{\log(Md/\delta) + (\log\log n)_+}\right)
\end{aligned}$$

where C_i are absolute constants, and $C(M, \delta, d)$ 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:

$$\hat{r}_{t,j} = \int_{\mathcal{X}} \langle \hat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}(\boldsymbol{x}) \rangle \mathrm{d}p_{t+1,j}(\boldsymbol{x}) \leq \left\|\hat{\boldsymbol{\theta}}_t\right\| \int_{\mathcal{X}} \|\boldsymbol{\phi}(\boldsymbol{x})\| \mathrm{d}p_{t+1,j}(\boldsymbol{x}) \leq \left\|\hat{\boldsymbol{\theta}}_t\right\|_2 \leq B + \left\|(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta})\right\|_2$$

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

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

Which implies for all $t \geq 1$,

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

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

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

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

$$\begin{aligned}
\sum_{t=1}^{n} \eta_t \frac{160\lambda_t^2}{c_{\kappa,t}^4} &\leq C(M,\delta,d) \sum_{t=1}^{n} \frac{\bar{C}_3}{t C_{\min}} \left(1 + \bar{C}_4 \frac{t^{-3/8}}{C_{\min}} \sqrt{\log(Md/\delta) + (\log\log t)_+} \right) \\
&\leq C(M,\delta,d) \frac{C_3 \log n}{C_{\min}} \left(1 + C_4 \frac{n^{-3/8}}{C_{\min}} \sqrt{\log(Md/\delta) + (\log\log n)_+} \right).
\end{aligned}$$

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,

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

where C_i are absolute constants. □

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

$$\begin{aligned}
R(n,i) \leq 2B \sum_{t=1}^{n} \gamma_t + \frac{\log M}{\eta_n} + \sum_{t=1}^{n} \eta_t \sum_{j=1}^{M} q_{t,j} \hat{r}_{t,j}^2 + \sum_{t=1}^{n} (\omega_{t,i} + \sum_{j=1}^{M} q_{t,j} \omega_{t,j}) \\
+ 10B \sqrt{n \left((\log\log nB^2)_+ + \log(12/\delta) \right)}
\end{aligned}$$

We invoke Lemma 19 and Lemma 20 with $\delta \to \delta/3$ take a union bound, to bound the variance and $\omega_{t,i}$ 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 $\hat{\boldsymbol{\theta}}_t$.

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

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

By the chain rule we may write,

$$\begin{aligned}
&\mathbb{P}\big(\mathrm{Reg}(n,i) \le \mathcal{B}\big) \\
&\quad \ge \mathbb{P}\left(R(n,i) \le \mathcal{B} \middle| \forall t \in [n], j \in [M] : \eta_t \hat{r}_{t,j} \le 1\right) \mathbb{P}\left(\forall t \in [n], j \in [M] : \eta_t \hat{r}_{t,j} \le 1\right) \\
&\quad \ge \mathbb{P}\left(R(n,i) \le \mathcal{B} \,\middle|\, \forall t \in [n], j \in [M] :\ \eta_t \hat{r}_{t,j} \le 1\right) (1-2\delta) \\
&\quad \ge (1-\delta)(1-2\delta) \ge 1 - 3\delta.
\end{aligned}$$

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

$$\frac{C_{\min}}{4\sqrt{t}C(M,\delta,d)} \le \frac{C_1 C_{\min} t^{1/4} - C_2 t^{-1/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}}{4\sqrt{10}C(M,\delta,d)} \le \frac{\kappa^2(\Phi_t, 2)}{4\sqrt{10}\lambda_t}$$

Therefore, from Lemma 20, there exists C_η such that $\eta_t = C_\eta C_{\min}/B\sqrt{t}C(M,\delta,d)$ satisfying,

$$\mathbb{P}\left(\forall t \ge 1, j \in [M] :\ \eta_t \hat{r}_{t,j} \le 1\right) \ge 1 - 2\delta$$

The proof is then finished by setting $\delta \leftarrow 3\delta$ (and updating the absolute constants). □

D.2 Proof of Virtual Regret

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

$$\mathbb{P}\left(\forall t \ge 1 :\ \left\|\hat{\boldsymbol{\beta}}_{t,j^\star} - \boldsymbol{\theta}_{j^\star}\right\|_2 \le \omega(t,\delta,d)\right) \ge 1 - \delta.$$

where

$$\omega(t,\delta,d) \coloneqq C_1 \sqrt{\frac{\sigma^2 d \log\left(\frac{t}{\tilde{\lambda}d}+1\right) + 2\sigma^2\log(1/\delta) + \tilde{\lambda}B^2}{\tilde{\lambda} + C_{\min}t^{3/4}}} \left(1 + C_{\min}^{-1} t^{-3/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\right).$$

Moreover, for $u_{t,j^\star}(\cdot) \coloneqq \hat{\boldsymbol{\beta}}_{t,j^\star}^\top \boldsymbol{\phi}_{j^\star}(\cdot) + \omega(t,\delta,d)$,

$$\mathbb{P}\left(\forall t \ge 1,\ \boldsymbol{x} \in \mathcal{X} : r(\boldsymbol{x}) \le u_{t,j^\star}(\boldsymbol{x})\right) \ge 1 - \delta.$$

Proof of Proposition 21. Define for convenience $V_t = \Phi_{t,j^\star}^\top \Phi_{t,j^\star} + \tilde{\lambda}\boldsymbol{I}$. We first observe that

$$\hat{\boldsymbol{\beta}}_{t,j^\star} = V_t^{-1}(\Phi_{t,j^\star})^\top \boldsymbol{y}_t$$

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

$$\left\|\hat{\boldsymbol{\beta}}_{t,j^\star} - \boldsymbol{\theta}_{j^\star}\right\|_{V_t}^2 \le \beta_t$$

where

$$\beta_t = 2\sigma^2 \log\left(\frac{\det(V_t)^{1/2}}{\det(\tilde{\lambda}\boldsymbol{I})^{1/2}\delta}\right) + \tilde{\lambda}B^2$$

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

$$\left\|\hat{\boldsymbol{\beta}}_{t,j^\star} - \boldsymbol{\theta}_{j^\star}\right\|_2 \le \sqrt{\frac{\beta_t}{\lambda_{\min}(V_t)}}$$

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

$$\det(V_t) \le \prod_{i=1}^d \nu_i \le \left(\frac{1}{d}\mathrm{trace}(V_t)\right)^d.$$

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

Given that

$$\mathrm{trace}(V_t) = \sum_{i=1}^{d}\sum_{s=1}^{t}(\boldsymbol{\phi}_{j^\star}(x))_i^2 + \tilde{\lambda}d \leq t + \tilde{\lambda}d$$

we can conclude that

$$\beta_t \leq 2\sigma^2 \log\left(\frac{(t/d+\tilde{\lambda})^{d/2}}{\tilde{\lambda}^{d/2}\delta}\right) + \tilde{\lambda}B^2 = d\sigma^2 \log\left(\frac{t}{\tilde{\lambda}d}+1\right) + 2\sigma^2\log(1/\delta) + \tilde{\lambda}B^2$$

We note that

$$\lambda_{\min}(V_t) = \lambda_{\min}(\Phi_{t,j}^\top \Phi_{t,j}) + \tilde{\lambda}.$$

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

$$\lambda_{\min}(V_t) \geq \tilde{\lambda} + C_1 C_{\min} t^{3/4} - C_2 t^{3/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}$$

therefore, there exists C_3 and C_4 such that

$$\begin{aligned}
\frac{1}{\sqrt{\lambda_{\min}(V_t)}} &\leq \frac{1}{\sqrt{\tilde{\lambda} + C_1 C_{\min} t^{3/4} - C_2 t^{3/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}}} \\
&\leq \frac{C_3}{\sqrt{\tilde{\lambda} + C_1 C_{\min} t^{3/4}}}\left(1 + \frac{t^{3/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}}{\tilde{\lambda} + C_1 C_{\min} t^{3/4}}\right) \\
&\leq \frac{C_4}{\sqrt{\tilde{\lambda} + C_1 C_{\min} t^{3/4}}}\left(1 + C_{\min}^{-1} t^{-3/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\right)
\end{aligned}$$

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

$$\omega(t,\delta,d) = C_5\sqrt{\frac{\sigma^2 d\log\left(\frac{t}{\tilde{\lambda}d}+1\right) + 2\sigma^2\log(1/\delta) + \tilde{\lambda}B^2}{\tilde{\lambda} + C_{\min}t^{3/4}}}\left(1 + C_{\min}^{-1}t^{-3/8}\sqrt{\log(Md/\delta) + (\log\log t)_+}\right)$$

where C_5 is an absolute constant concludes the parametric confidence bound. The upper confidence bound then simply follows: for any $\boldsymbol{x} \in \mathcal{X}$

$$r(\boldsymbol{x}) - \hat{\boldsymbol{\beta}}_{t,j^\star}^\top \boldsymbol{\phi}_{j^\star}(\boldsymbol{x}) = \langle \boldsymbol{\theta}_{j^\star} - \hat{\boldsymbol{\beta}}_{t,j^\star}, \boldsymbol{\phi}_{j^\star}(\boldsymbol{x})\rangle \leq \left\|\boldsymbol{\theta}_{j^\star} - \hat{\boldsymbol{\beta}}_{t,j^\star}\right\|_2 \|\boldsymbol{\phi}_{j^\star}(\boldsymbol{x})\|_2 \leq \omega(t,\delta,d)$$

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

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

$$\begin{aligned}
\tilde{R}_{j^\star}(n) &= \sum_{t=1}^{n} r(\boldsymbol{x}^\star) - r(\tilde{\boldsymbol{x}}_{t,j}) \\
&= \sum_{t=1}^{n} r(\boldsymbol{x}^\star) - \hat{r}_t(\boldsymbol{x}^\star) + \hat{r}_t(\boldsymbol{x}^\star) - \hat{r}_t(\tilde{\boldsymbol{x}}_{t,j}) + \hat{r}_t(\tilde{\boldsymbol{x}}_{t,j}) - r(\tilde{\boldsymbol{x}}_{t,j}) \\
&\leq \sum_{t=1}^{n} \left\|\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_{t,j}\right\|_2 \left(\|\boldsymbol{\phi}_j(\boldsymbol{x}^\star)\|_2 + \|\boldsymbol{\phi}_j(\tilde{\boldsymbol{x}}_{t,j})\|_2\right) + \hat{r}_t(\boldsymbol{x}^\star) - \hat{r}_t(\tilde{\boldsymbol{x}}_{t,j}) \\
&\leq \sum_{t=1}^{n} \omega(t,\delta,d)\left(\|\boldsymbol{\phi}_j(\boldsymbol{x}^\star)\|_2 + \|\boldsymbol{\phi}_j(\tilde{\boldsymbol{x}}_{t,j})\|_2\right) + \hat{r}_t(\boldsymbol{x}^\star) - \hat{r}_t(\tilde{\boldsymbol{x}}_{t,j})
\end{aligned}$$

with probability $1-\delta$. If the agent selects actions greedily, then $\hat{r}_t(\boldsymbol{x}^\star) \leq \hat{r}_t(\tilde{\boldsymbol{x}}_{t,j})$, and

$$\tilde{R}_{j^\star}(n) \leq \sum_{t=1}^{n} \omega(t,\delta,d)\left(\|\boldsymbol{\phi}_j(\boldsymbol{x}^\star)\|_2 + \|\boldsymbol{\phi}_j(\tilde{\boldsymbol{x}}_{t,j})\|_2\right) \leq \sum_{t=1}^{n} 2\omega(t,\delta,d)$$

since the feature map is normalized to satisfy $\|\boldsymbol{\phi}_j(\cdot)\| \leq 1$. If the agent selects actions optimistically according to the upper confidence bound of Proposition 21, then

$$\hat{r}_t(\tilde{\boldsymbol{x}}_{t,j}) + \omega(t,\delta,d)\|\boldsymbol{\phi}_j(\tilde{\boldsymbol{x}}_{t,j})\| \geq \hat{r}_t(\boldsymbol{x}^\star) + \omega(t,\delta,d)\|\boldsymbol{\phi}_j(\boldsymbol{x}^\star)\|$$

which implies

$$\hat{r}_t(\boldsymbol{x}^\star) - \hat{r}_t(\tilde{\boldsymbol{x}}_{t,j}) \le \omega(t,\delta,d)\|\boldsymbol{\phi}_j(\tilde{\boldsymbol{x}}_{t,j})\| - \omega(t,\delta,d)\|\boldsymbol{\phi}_j(\boldsymbol{x}^\star)\|$$

and therefore,

$$\tilde{R}_{j^\star}(n) \le \sum_{t=1}^{n} 2\omega(t,\delta,d)\|\boldsymbol{\phi}_j(\tilde{\boldsymbol{x}}_{t,j})\|_2 \le \sum_{t=1}^{n} 2\omega(t,\delta,d).$$

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

$$\begin{aligned}\sum_{t=1}^{n}\omega(t,\delta,d) &\le \sum_{t=1}^{n} C_1\sqrt{\frac{\sigma^2 d\log\left(\frac{t}{\tilde{\lambda}d}+1\right)+2\sigma^2\log(1/\delta)+\tilde{\lambda}B^2}{\tilde{\lambda}+C_{\min}t^{3/4}}}\left(1+t^{-3/8}\sqrt{\log(\tfrac{Md}{\delta})+(\log\log t)_+}\right)\\ &\le \tilde{C}_1 n^{5/8}\sqrt{\frac{\sigma^2 d\log\left(\frac{n}{\tilde{\lambda}d}+1\right)+2\sigma^2\log(1/\delta)+\tilde{\lambda}B^2}{C_{\min}}}\left(1+C_{\min}^{-1}n^{\frac{-3}{8}}\sqrt{\log(\tfrac{Md}{\delta})+(\log\log n)_+}\right)\end{aligned}$$

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 $(S_t)_{t=0}^{\infty}$ and $(V_t)_{t=0}^{\infty}$ be real-valued processes adapted to $(\mathcal{F}_t)_{t=1}^{\infty}$ with $S_0 = V_0 = 0$ and V_t non-negative. We say that S_t is sub-Gamma if for $\lambda \in [0, 1/c)$, 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 $t \ge 1$:

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

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 $(S_t)_{t\ge0}$ be sub-Gamma with scale parameter c and variance process $(V_t)_{t\ge0}$. Define the boundary*

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

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

$$\mathbb{P}(\exists t:\ S_t \ge \mathcal{B}_\alpha(V_t)) \le \alpha.$$

Proof of Proposition 23. Suppose $\xi(\cdot)$ denotes the Riemann zeta function. Theorem 1 in Howard et al. [2021] states that if $(S_t)_{t\ge0}$ is a sub-Gamma process with variance process $(V_t)_{t\ge0}$ then the boundary

$$\mathcal{S}_\alpha(v') = k_1\sqrt{v'\left(s\log\log(\eta v') + \log\left(\frac{\zeta(s)}{\alpha\log^s\eta}\right)\right)} + ck_2\left(s\log\log(\eta v') + \log\left(\frac{\zeta(s)}{\alpha\log^s\eta}\right)\right).$$

satisfies,

$$\mathbb{P}(\exists t: S_t \ge \mathcal{S}_\alpha(\max(V_t,1))) \le \alpha$$

where

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

and $s, \eta \ge 1$. Choosing $s = 2$ and $\eta = e$, we obtain $\zeta(2) = \pi^2/6 \le 2$. Furthermore, we have $k_1 \le \frac{3}{2}$ and $k_2 \le \frac{3}{2}$. Then if $v' \ge 1$ (which we will enforce by the construction $v' = \max(1, v)$), we compute

$$s\log\log(\eta v') + \log\left(\frac{\zeta(s)}{\alpha\log^s\eta}\right) \le 2(\log\log ev')_+ + \log\left(\frac{2}{\alpha}\right).$$

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

$$\mathcal{S}_\alpha(v') \le \frac{5}{2}\sqrt{v'\left((\log\log ev')_+ + \log\left(\frac{2}{\alpha}\right)\right)} + 3c\left((\log\log ev')_+ + \log\left(\frac{2}{\alpha}\right)\right).$$

Now, since the boundary is given by $\mathcal{S}_\alpha(\max(v,1))$ and $v' = \max(v,1) \ge 1$ we deduce that

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

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 γ_s. Then for all $\delta > 0$,*

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

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

$$\mathbb{P}\left(\exists t : \ S_t \ge \mathcal{B}_\delta(V_t)\right) \le \delta,$$

where

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

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

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

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

Lemma 25 (Time-Uniform Bernstein)**.** *Let $(\xi_i)_{i=1}^\infty$ be a sequence of conditionally standard sub-gaussian variables, where each ξ_i is $\mathcal{F}_{i-1} = \sigma(\xi_1, \dots, \xi_i)$ measurable. Then, for $v_i \in \mathbb{R}$ and $\delta \in (0,1]$*

$$\mathbb{P}\left(\exists t : \ \sum_{i=1}^{t}(\xi_i^2 - 1)v_i \ge \frac{5}{2}\sqrt{\max\left\{1, 4\|\boldsymbol{v}_t\|_2^2\right\}\omega_\delta(\|\boldsymbol{v}_t\|_2)} + 12\omega_\delta(\|\boldsymbol{v}_t\|_2)\max_{i\ge 1} v_i\right) \le \delta$$

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

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

$$\mathbb{P}\left(\exists t : \ S_t \ge \mathcal{B}_\delta(V_t)\right) \le \delta,$$

where

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

□

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 $t > 1$ almost surely. Then for all $\delta > 0$,*

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

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

$$\mathbb{P}\left(\exists t:\ S_t \geq \mathcal{B}_\delta(V_t)\right) \leq \delta,$$

where

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

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,

$$\mathbb{P}\left(\exists t:\ S_t \geq \frac{5B}{2}\sqrt{t\left((\log\log etB^2)_+ + \log(2/\delta)\right)}\right) \leq \delta,$$

concluding the proof. □

F Experiment Details

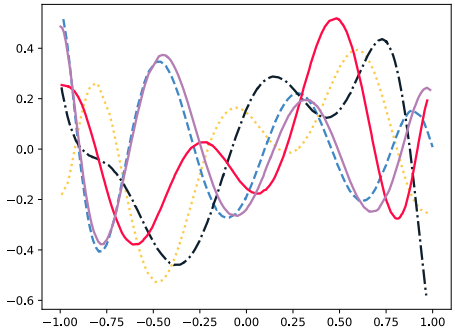

Figure 5: Examples of possible reward functions $r(\cdot)$ in our experiments.

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 $\phi_{j\star}$ (Oracle), and UCB with the concatenated feature map ϕ (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 $\pi = \text{Unif}(\mathcal{X})$. Figure 7 shows the results of our hyperparameter tuning experiment. To ensure that the curves are valid, we run each

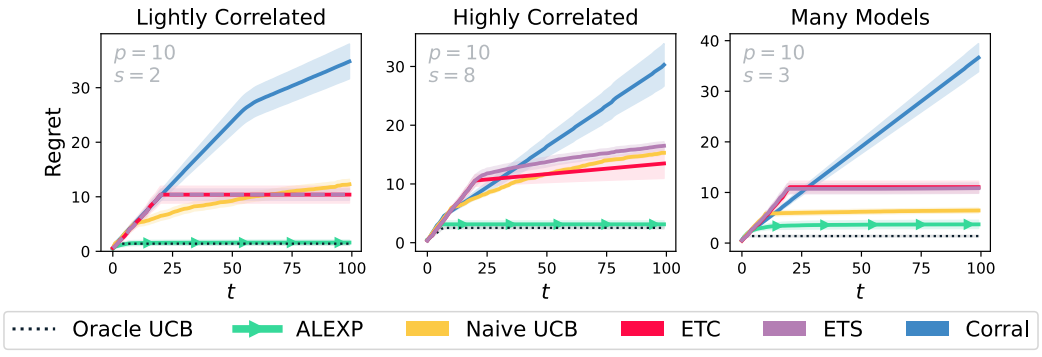

Figure 6: Bench-marking ALEXP and other baselines. Complete version of Fig. 1 and Fig. 2.

Algorithm 2 GetPosterior

Inputs: H_t, $\boldsymbol{\phi}$, $\tilde{\lambda}$
Let $K_t \leftarrow [\boldsymbol{\phi}^\top(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_j)]_{i,j\le t}$, and $V_t \leftarrow (K_t + \tilde{\lambda}^2 \boldsymbol{I})$, and $\boldsymbol{k}(\cdot) \leftarrow [\boldsymbol{\phi}^\top(\boldsymbol{x}_i)\boldsymbol{\phi}(\cdot)]_{i\le t}$

Calculate $\mu_t(\cdot) \leftarrow \boldsymbol{k}^T(\cdot)V_t^{-1}\boldsymbol{y}_t$

Calculate $\sigma_t(\cdot) \leftarrow \sqrt{\boldsymbol{\phi}^\top(\cdot)\boldsymbol{\phi}(\cdot) - \boldsymbol{k}^\top(\cdot)V_t^{-1}\boldsymbol{k}(\cdot)}$
Return: μ_t, σ_t

Algorithm 3 UCB

Inputs: $\tilde{\lambda}$, β_t, $\boldsymbol{\phi}$
for $t = 1, \dots, n$ **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 $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$. ▷ Receive reward
  $H_t \leftarrow H_{t-1} \cup \{(\boldsymbol{x}_t, y_t)\}$ ▷ Append history
  Update $\mu_t, \sigma_t \leftarrow \text{GetPosterior}(H_t, \boldsymbol{\phi}, \tilde{\lambda})$
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 $\tilde{\lambda} \in \{0.01, 0.1, 0.5\}$. 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 $\gamma_0 \in [10^{-4}, 10^{-1}]$ and $\eta_0 \in [10^0, 10^2]$. 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 $\lambda_1 \propto \sqrt{\log M / n_0}$ 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 $n_0 \in [2, 80]$ since the horizon is $n = 100$. The optimal value often happens around $n_0 = 20$.

CORRAL. We set the rates of the parameters as $\gamma = \mathcal{O}(1/n)$ and $\eta = \mathcal{O}(\sqrt{M/n})$ 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 $\beta_t = \mathcal{O}(\sqrt{d \log T})$ as shown in Proposition 21.

Algorithm 4 ETC [Hao et al., 2020]

Inputs: n_0, n, λ_1, π
Let $H_0 = \emptyset$
for $t = 1, \dots, n_0$ **do**
  Draw $\boldsymbol{x}_t \sim \pi$. ▷ Explore.
  Observe $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$. ▷ Receive reward
  $H_t \leftarrow H_{t-1} \cup \{(\boldsymbol{x}_t, y_t)\}$ ▷ Append history
end for
$\hat{\boldsymbol{\theta}}_{n_0} \leftarrow \mathcal{L}(\boldsymbol{\theta}, H_{n_0}, \lambda_1)$ ▷ Perform Lasso once
for $t = n_0 + 1, \dots, n$ **do**
  Choose $\boldsymbol{x}_t = \arg\max \hat{\boldsymbol{\theta}}_{n_0}^\top \boldsymbol{\phi}(\boldsymbol{x})$ ▷ Choose actions greedily
end for

Algorithm 5 ETS

Inputs: n_0, n, λ_1, $\tilde{\lambda}$, β_t, π
Let $H_0 = \emptyset$
for $t = 1, \dots, n_0$ **do**
  Draw $\boldsymbol{x}_t \sim \pi$. ▷ Explore
  Observe $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$. ▷ Receive reward
  $H_t \leftarrow H_{t-1} \cup \{(\boldsymbol{x}_t, y_t)\}$ ▷ Append history
end for
$\hat{\boldsymbol{\theta}}_{n_0} \leftarrow \mathcal{L}(\boldsymbol{\theta}, H_{n_0}, \lambda_1)$ ▷ Perform Lasso once
$\hat{J} \leftarrow \{j \,|\, \hat{\boldsymbol{\theta}}_{n_0,j} \neq \mathbf{0},\ j \in [M]\}$ ▷ Get sparsity pattern
$\boldsymbol{\phi}_{\hat{J}}(\cdot) \leftarrow [\boldsymbol{\phi}_j(\cdot)]_{j \in \hat{J}}$ ▷ Model-select acc. to $\hat{J}$
for $t = n_0 + 1, \dots, n$ **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 $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$
  $H_t \leftarrow H_{t-1} \cup \{(\boldsymbol{x}_t, y_t)\}$
  Update $\mu_t, \sigma_t \leftarrow \text{GetPosterior}(H_t, \boldsymbol{\phi}_{\hat{J}}, \tilde{\lambda})$
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 $j = 1, \dots, M$
Set $\boldsymbol{q}_1 = \bar{\boldsymbol{q}}_1 = \frac{\mathbf{1}}{M}$ and initialize base agents $(p_{1,1}, \dots, p_{1,M})$.
for $t = 1, \dots, n$ **do**
  Choose $j_t \sim \bar{\boldsymbol{q}}_t$. ▷ Sample Agent
  Draw $\boldsymbol{x}_t \sim p_{t,j_t}$. ▷ Play action according to agent j_t
  Observe $y_t = r(\boldsymbol{x}_t) + \varepsilon_t$.
  Calculate IW estimates $\hat{r}_{t,j} = \frac{y_t}{\bar{q}_{t,j}} \mathbb{I}\{j = j_t\}$ for all $j = 1, \dots, M$.
  Send $\hat{r}_{t,j} = \frac{y_t}{\bar{q}_{t,j}} \mathbb{I}\{j = j_t\}$ to agents and get updated policies $p_{t+1,j}$.
  $\boldsymbol{q}_{t+1} = \text{LOG-BARRIER-OMD}(\boldsymbol{q}_t, \hat{r}_{t,j_t} \boldsymbol{e}_{j_t}, \boldsymbol{\eta}_t)$ ▷ Update agent probabilities
  $\bar{\boldsymbol{q}}_{t+1} = (1-\gamma)\boldsymbol{q}_{t+1} + \gamma \frac{\mathbf{1}}{M}$ ▷ Mix with exploratory distribution
  for $j = 1, \dots, M$ **do** ▷ Update parameters
    if $\frac{1}{\bar{q}_{t+1,j}} > \rho_{t,j}$ **then** $\rho_{t+1,j} \leftarrow \frac{2}{\bar{q}_{t,j}}$, and $\eta_{t+1,j} \leftarrow \beta \eta_{t,j}$
    else $\rho_{t+1,j} \leftarrow \rho_{t,j}$ and $\eta_{t+1,j} \leftarrow \eta_{t,j}$
    end if
  end for
end for

Algorithm 7 LOG-BARRIER-OMD

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

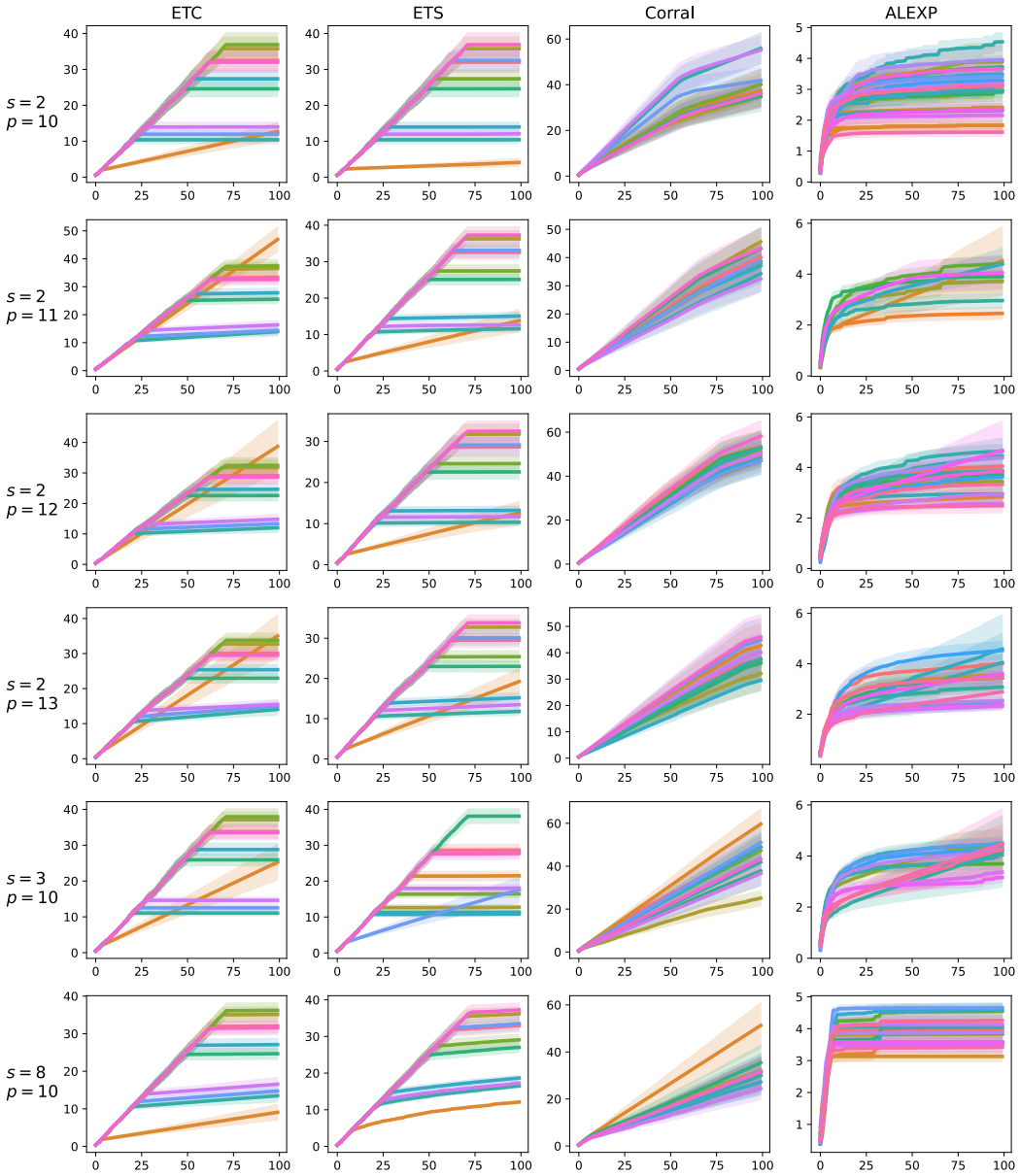

Figure 7: Results for different hyper-parameters across different problem instances. ALEXP is robust to the choice of hyper-paramters.