# Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems*

**Yasin Abbasi-Yadkori**
abbasiya@cs.ualberta.ca
Dept. of Computing Science
University of Alberta

**Dávid Pál**
dpal@cs.ualberta.ca
Dept. of Computing Science
University of Alberta

**Csaba Szepesvári**
szepesva@cs.ualberta.ca
Dept. of Computing Science
University of Alberta

## Abstract

The analysis of online least squares estimation is at the heart of many stochastic sequential decision-making problems. We employ tools from the self-normalized processes to provide a simple and self-contained proof of a tail bound of a vector-valued martingale. We use the bound to construct new tighter confidence sets for the least squares estimate.

We apply the confidence sets to several online decision problems, such as the multi-armed and the linearly parametrized bandit problems. The confidence sets are potentially applicable to other problems such as sleeping bandits, generalized linear bandits, and other linear control problems.

We improve the regret bound of the Upper Confidence Bound (UCB) algorithm of Auer et al. (2002) and show that its regret is with high-probability a problem dependent constant. In the case of linear bandits (Dani et al., 2008), we improve the problem dependent bound in the dimension and number of time steps. Furthermore, as opposed to the previous result, we prove that our bound holds for small sample sizes, and at the same time the worst case bound is improved by a logarithmic factor and the constant is improved.

## 1 Introduction

The least squares method forms a cornerstone of statistics and machine learning. It is used as the main component of many stochastic sequential decision problems, such as multi-armed bandit, linear bandits, and other linear control problems. However, the analysis of least squares in these online settings is non-trivial because of the correlations between data points. Fortunately, there is a connection between online least squares estimation and the area of self-normalized processes. Study of self-normalized processes has a long history that goes back to Student and is treated in detail in recent book by de la Peña et al. (2009). Using these tools we provide a proof of a bound on the deviation for vector-valued martingales. A less general version of the bound can be found already in de la Peña et al. (2004, 2009). Additionally our proof, based on the method of mixtures, is new, simpler and self-contained. The bound improves the previous bound of Rusmevichientong and Tsitsiklis (2010) and it is applicable to virtually any online least squares problem.

The bound that we derive, gives immediately rise to tight confidence sets for the online least squares estimate that can replace the confidence sets in existing algorithms. In particular, the confidence sets can be used in the UCB algorithm for the multi-armed bandit problem, the ConfidenceBall algorithm of Dani et al. (2008) for the linear bandit problem, and LinRel algorithm of Auer (2003) for the associative reinforcement learning problem. We show that this leads to improved performance of these algorithms. Our hope is that the new confidence sets can be used to improve the performance of other similar linear decision problems.

The multi-armed bandit problem, introduced by Robbins (1952), is a game between the learner and the environment. At each time step, the learner chooses one of $K$ actions and receives a reward which is generated independently at random from a fixed distribution associated with the chosen arm. The objective of the learner is to maximize his total reward. The performance of the learner is evaluated by the regret, which is defined as the difference between his total reward and the total reward of the best action. Lai and Robbins (1985) prove a $(\sum_{i \neq i_*} 1/D(p_j, p_{i_*}) - o(1)) \log T$ lower

*Submitted to the 24th Annual Conference on Learning Theory (COLT 2011)

bound on the expected regret of any algorithm, where $T$ is the number of time steps, $p_{i_*}$ and $p_i$ are the reward distributions of the optimal arm and arm $i$ respectively, and $D$ is the KL-divergence.

Auer et al. (2002) designed the UCB algorithm and proved a finite-time logarithmic bound on its regret. He used Hoeffding's inequality to construct confidence intervals and obtained a $O((K \log T)/\Delta)$ bound on the expected regret, where $\Delta$ is the difference between the expected rewards of the best and the second best action. We modify UCB so that it uses our new confidence sets and we show a stronger result. Namely, we show that with probability $1-\delta$, the regret of the modified algorithm is $O(K \log(1/\delta)/\Delta)$. Seemingly, this result contradicts the lower bound of Lai and Robbins (1985), however our algorithm depends on $\delta$ which it receives as an input. The expected regret of the modified algorithm with $\delta = 1/T$ matches the regret of the original algorithm.

In the linear bandit problem, the learner chooses repeatedly actions from a fixed subset of $\mathbb{R}^d$ and receives a random reward, expectation of which is a linear function of the action. Dani et al. (2008) proposed the ConfidenceBall algorithm and showed that its regret is at most $O(d \log(T) \sqrt{T \log(T/\delta)})$ with probability at most $1-\delta$. We modify their algorithm so that it uses our new confidence sets and we show that its regret is at most $O(d \log(T)\sqrt{T} + \sqrt{dT \log(T/\delta)})$. Additionally, constants in our bound are smaller, and our bound holds for all $T \geq 1$, as opposed the previous one which holds only for sufficiently large $T$. Dani et al. (2008) prove also a problem dependent regret bound. Namely, they show that the regret of their algorithm is $O(\frac{d^2}{\Delta} \log^2 T \log(T/\delta))$ where $\Delta$ is the "gap" as defined in (Dani et al., 2008). For our modified algorithm we prove an improved $O(\frac{\log(1/\delta)}{\Delta}(\log T + d \log\log T)^2)$ bound.

### 1.1 Notation

We use $\|\cdot\|$ to denote the 2-norm. For a positive definite matrix $A \in \mathbb{R}^{d\times d}$, the weighted 2-norm is defined by $\|x\|_A^2 = x^\top A x$, where $x \in \mathbb{R}^d$. The inner product is denoted by $\langle \cdot, \cdot \rangle$ and the weighted inner-product $x^\top A y = \langle x, y \rangle_A$. We use $\lambda_{\min}(A)$ to denote the minimum eigenvalue of the positive definite matrix $A$. We use $A \succ 0$ to denote that $A$ is positive definite, while we use $A \succeq 0$ to denote that it is positive semidefinite. The same notation is used to denote the Loewner partial order of matrices. We shall use $\mathbf{e}_i$ to denote the $i^{\text{th}}$ unit vector, i.e., for all $j \neq i$, $\mathbf{e}_{ij} = 0$ and $\mathbf{e}_{ii} = 1$.

## 2 Vector-Valued Martingale Tail Inequalities

Let $(\mathcal{F}_k; k \geq 0)$ be a filtration, $(m_k; k \geq 0)$ be an $\mathbb{R}^d$-valued stochastic process adapted to $(\mathcal{F}_k)$, $(\eta_k; k \geq 1)$ be a real-valued martingale difference process adapted to $(\mathcal{F}_k)$. Assume that $\eta_k$ is conditionally sub-Gaussian in the sense that there exists some $R > 0$ such that for any $\gamma \in \mathbb{R}$, $k \geq 1$,

$$\mathbb{E}[\exp(\gamma \eta_k) \,|\, \mathcal{F}_{k-1}] \leq \exp\left(\frac{\gamma^2 R^2}{2}\right) \quad \text{a.s.} \tag{1}$$

Consider the martingale

$$S_t = \sum_{k=1}^{t} \eta_k m_{k-1} \tag{2}$$

and the matrix-valued processes

$$V_t = \sum_{k=1}^{t} m_{k-1} m_{k-1}^\top, \qquad \overline{V}_t = V + V_t, \quad t \geq 0, \tag{3}$$

where $V$ is an $\mathcal{F}_0$-measurable, positive definite matrix. In particular, assume that with probability one, the eigenvalues of $V$ are larger than $\lambda_0 > 0$ and that $\|m_k\| \leq L$ holds a.s. for any $k \geq 0$.

The following standard inequality plays a crucial role in the following developments:

**Lemma 1.** *Consider $(\eta_t)$, $(m_t)$ as defined above and let $\tau$ be a stopping time with respect to the filtration $(\mathcal{F}_t)$. Let $\lambda \in \mathbb{R}^d$ be arbitrary and consider*

$$P_t^\lambda = \exp\left(\sum_{k=1}^{t} \left[\frac{\eta_k \langle \lambda, m_{k-1}\rangle}{R} - \frac{1}{2}\langle \lambda, m_{k-1}\rangle^2\right]\right).$$

*Then $P_\tau$ is almost surely well-defined and*

$$\mathbb{E}\left[P_\tau^\lambda\right] \leq 1.$$

*Proof.* The proof is standard (and is given only for the sake of completeness). We claim that $P_t = P_t^\lambda$ is a supermartingale. Let

$$D_k = \exp\left(\frac{\eta_k \langle \lambda, m_{k-1}\rangle}{R} - \frac{1}{2}\langle \lambda, m_{k-1}\rangle^2\right).$$

Observe that by (1), we have $\mathbb{E}[D_k | \mathcal{F}_{k-1}] \le 1$. Clearly, $D_k$ is $\mathcal{F}_k$-adapted, as is $P_k$. Further,

$$\mathbb{E}[P_t|\mathcal{F}_{t-1}] = \mathbb{E}[D_1 \cdots D_{t-1} D_t | \mathcal{F}_{t-1}] = D_1 \cdots D_{t-1}\, \mathbb{E}[D_t | \mathcal{F}_{t-1}] \le P_{t-1},$$

showing that $(P_t)$ is indeed a supermartingale.

Now, this immediately leads to the desired result when $\tau = t$ for some deterministic time $t$. This is based on the fact that the mean of any supermartingale can be bounded by the mean of its first element. In the case of $(P_t)$, for example, we have $\mathbb{E}[P_t] = \mathbb{E}[\mathbb{E}[P_t|\mathcal{F}_{t-1}]] \le \mathbb{E}[P_{t-1}] \le \ldots \le \mathbb{E}[P_0] = \mathbb{E}[D_0] = 1$.

Now, in order to consider the general case, let $S_t = P_{\tau \wedge t}$.[1] It is well known that $(S_t)$ is still a supermartingale with $\mathbb{E}[S_t] \le \mathbb{E}[S_0] = \mathbb{E}[P_0] = 1$. Further, since $P_t$ was nonnegative, so is $S_t$. Hence, by the convergence theorem for nonnegative supermartingales, almost surely, $\lim_{t\to\infty} S_t$ exists, i.e., $P_\tau$ is almost surely well-defined. Further, $\mathbb{E}[P_\tau] = \mathbb{E}[\liminf_{t\to\infty} S_t] \le \liminf_{t\to\infty} \mathbb{E}[S_t] \le 1$ by Fatou's Lemma. □

Before stating our main results, we give some recent results, which can essentially be extracted from the paper by Rusmevichientong and Tsitsiklis (2010).

**Theorem 2.** *Consider the processes $(S_t)$, $(\overline{V}_t)$ as defined above and let*

$$\kappa = \sqrt{3 + 2\log((L^2 + \mathrm{trace}(V))/\lambda_0)}. \tag{4}$$

*Then, for any $0 < \delta < 1$, $t \ge 2$, with probability at least $1 - \delta$,*

$$\|S_t\|_{\overline{V}_t^{-1}} \le 2\,\kappa^2 R \sqrt{\log t}\,\sqrt{d\,\log(t) + \log(1/\delta)}\,. \tag{5}$$

We now show how to strengthen the previous result using the method of mixtures, originally used by Robbins and Siegmund (1970) to evaluate boundary crossing probabilities for Brownian motion.

**Theorem 3** (Self-normalized bound for vector-valued martingales)**.** *Let $(\eta_t)$, $(m_t)$, $(S_t)$, $(\overline{V}_t)$, and $(\mathcal{F}_t)$ be as before and let $\tau$ be a stopping time with respect to the filtration $(\mathcal{F}_t)$. Assume that $V$ is deterministic. Then, for any $0 < \delta < 1$, with probability $1 - \delta$,*

$$\|S_\tau\|^2_{\overline{V}_\tau^{-1}} \le 2R^2 \log\left(\frac{\det(\overline{V}_\tau)^{1/2}\det(V)^{-1/2}}{\delta}\right). \tag{6}$$

*Proof.* Without loss of generality, assume that $R = 1$ (by appropriately scaling $S_t$, this can always be achieved). Let

$$M_t(\lambda) = \exp\left(\langle \lambda, S_t\rangle - \tfrac{1}{2}\,\|\lambda\|^2_{V_t}\right).$$

Notice that by Lemma 1, the mean of $M_\tau(\lambda)$ is not larger than one.

Let $\Lambda$ be a Gaussian random variable which is independent of all the other random variables and whose covariance is $V^{-1}$. Define

$$M_t = \mathbb{E}[M_t(\Lambda)|\mathcal{F}_\infty].$$

Clearly, we still have $\mathbb{E}[M_\tau] = \mathbb{E}[\,\mathbb{E}[\,M_\tau(\Lambda)\,|\,\Lambda\,]\,] \le 1$.

Let us calculate $M_t$: Let $f$ denote the density of $\Lambda$ and for a positive definite matrix $P$ let $c(P) = \sqrt{(2\pi)^d/\det(P)} = \int \exp(-\frac{1}{2}x^\top P x)dx$. Then,

$$\begin{aligned}
M_t &= \int_{\mathbb{R}^d} \exp\left(\langle \lambda, S_t\rangle - \tfrac{1}{2}\,\|\lambda\|^2_{V_t}\right) f(\lambda)\, d\lambda \\
&= \int_{\mathbb{R}^d} \exp\left(-\tfrac{1}{2}\,\left\|\lambda - V_t^{-1}S_t\right\|^2_{V_t} + \tfrac{1}{2}\,\|S_t\|^2_{V_t^{-1}}\right) f(\lambda)\, d\lambda \\
&= \frac{1}{c(V)}\exp\left(\tfrac{1}{2}\,\|S_t\|^2_{V_t^{-1}}\right)\int_{\mathbb{R}^d}\exp\left(-\tfrac{1}{2}\left\{\left\|\lambda - V_t^{-1}S_t\right\|^2_{V_t} + \|\lambda\|^2_V\right\}\right) d\lambda.
\end{aligned}$$

[1] $\tau \wedge t$ is a shorthand notation for $\min(\tau, t)$.

Elementary calculation shows that if $P \succeq 0$, $Q \succ 0$,

$$\|x - a\|_P^2 + \|x\|_Q^2 = \left\|x - (P+Q)^{-1}Pa\right\|_{P+Q}^2 + \|a\|_P^2 - \|Pa\|_{(P+Q)^{-1}}^2 .$$

Therefore,

$$\begin{aligned}
\left\|\lambda - V_t^{-1}S_t\right\|_{V_t}^2 + \|\lambda\|_V^2 &= \left\|\lambda - (V+V_t)^{-1}S_t\right\|_{V+V_t}^2 + \left\|V_t^{-1}S_t\right\|_{V_t}^2 - \|S_t\|_{(V+V_t)^{-1}}^2 \\
&= \left\|\lambda - (V+V_t)^{-1}S_t\right\|_{V+V_t}^2 + \|S_t\|_{V_t^{-1}}^2 - \|S_t\|_{(V+V_t)^{-1}}^2 ,
\end{aligned}$$

which gives

$$\begin{aligned}
M_t &= \frac{1}{c(V)} \exp\left(\tfrac{1}{2}\ \|S_t\|_{(V+V_t)^{-1}}^2\right) \int_{\mathbb{R}^d} \exp\left(-\tfrac{1}{2}\ \left\|\lambda - (V+V_t)^{-1}S_t\right\|_{V+V_t}^2\right) d\lambda \\
&= \frac{c(V+V_t)}{c(V)} \exp\left(\tfrac{1}{2}\ \|S_t\|_{(V+V_t)^{-1}}^2\right) = \left(\frac{\det(V)}{\det(V+V_t)}\right)^{1/2} \exp\left(\tfrac{1}{2}\ \|S_t\|_{(V+V_t)^{-1}}^2\right) .
\end{aligned}$$

Now, from $\mathbb{E}[M_\tau] \le 1$, we obtain

$$\begin{aligned}
\mathbb{P}\left(\|S_\tau\|_{(V+V_\tau)^{-1}}^2 > 2\log\left(\frac{\det(V+V_\tau)^{1/2}}{\det(V)^{1/2}}\ \frac{1}{\delta}\right)\right) &= \mathbb{P}\left(\frac{\exp\left(\tfrac{1}{2}\ \|S_\tau\|_{(V+V_\tau)^{-1}}^2\right)}{\delta^{-1}\left(\det(V+V_\tau)\big/\det(V)\right)^{\frac{1}{2}}} > 1\right) \\
&\le \mathbb{E}\left[\frac{\exp\left(\tfrac{1}{2}\ \|S_\tau\|_{(V+V_\tau)^{-1}}^2\right)}{\delta^{-1}\left(\det(V+V_\tau)\big/\det(V)\right)^{\frac{1}{2}}}\right] \\
&= \mathbb{E}[M_\tau]\,\delta \le \delta,
\end{aligned}$$

thus finishing the proof. ∎

*Corollary* 1 (Uniform Bound). Under the same assumptions as in the previous theorem, for any $0 < \delta < 1$, with probability $1 - \delta$,

$$\forall t \ge 0, \qquad \|S_t\|_{\overline{V}_t^{-1}}^2 \le 2R^2 \log\left(\frac{\det(\overline{V}_t)^{1/2}\det(V)^{-1/2}}{\delta}\right). \tag{7}$$

*Proof.* We will use a stopping time construction, which goes back at least to Freedman (1975). Define the bad event

$$B_t(\delta) = \left\{\omega \in \Omega\ :\ \|S_t\|_{\bar{V}_t^{-1}}^2 > 2R^2\log\left(\frac{\det(\bar{V}_t)^{1/2}\det(V)^{-1/2}}{\delta}\right)\right\} \tag{8}$$

We are interested in bounding the probability that $\bigcup_{t\ge 0} B_t(\delta)$ happens. Define $\tau(\omega) = \min\{t \ge 0\ :\ \omega \in B_t(\delta)\}$, with the convention that $\min\emptyset = \infty$. Then, $\tau$ is a stopping time. Further,

$$\bigcup_{t\ge 0} B_t(\delta) = \{\omega\ :\ \tau(\omega) < \infty\}.$$

Thus, by Theorem 3

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{t\ge 0} B_t(\delta)\right) &= \mathbb{P}(\tau < \infty) \\
&= \mathbb{P}\left(\|S_\tau\|_{\bar{V}_\tau^{-1}}^2 > 2R^2\log\left(\frac{\det(\bar{V}_\tau)^{1/2}\det(V)^{-1/2}}{\delta}\right),\ \tau < \infty\right) \\
&\le \mathbb{P}\left(\|S_\tau\|_{\bar{V}_\tau^{-1}}^2 > 2R^2\log\left(\frac{\det(\bar{V}_\tau)^{1/2}\det(V)^{-1/2}}{\delta}\right)\right) \\
&\le \delta .
\end{aligned}$$

∎

Let us now turn our attention to understanding the determinant term on the right-hand side of (6).

**Lemma 4.** *We have that*

$$\log \frac{\det(\overline{V}_t)}{\det V} \le \sum_{k=1}^{t} \|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}} .$$

*Further, we have that*

$$\sum_{k=1}^{t} \left( \|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}} \wedge 1 \right) \le 2(\log\det(\overline{V}_t) - \log\det V) \le 2(d\log((\mathrm{trace}(V) + tL^2)/d) - \log\det V).$$

*Finally, if $\lambda_0 \ge \max(1, L^2)$ then*

$$\sum_{k=1}^{t} \|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}} \le 2\log\frac{\det(\overline{V}_t)}{\det(V)}.$$

*Proof.* Elementary algebra gives

$$\begin{aligned} \det(\overline{V}_t) &= \det(\overline{V}_{t-1} + m_{t-1}m_{t-1}^\top) = \det(\overline{V}_{t-1})\det(I + \overline{V}_{t-1}^{-1/2} m_{t-1}(\overline{V}_{t-1}^{-1/2} m_{t-1})^\top) \\ &= \det(\overline{V}_{t-1})\,(1 + \|m_{t-1}\|^2_{\overline{V}^{-1}_{t-1}}) = \det(V)\prod_{k=1}^{t}\left(1 + \|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}}\right), \end{aligned} \tag{9}$$

where we used that all the eigenvalues of a matrix of the form $I + xx^\top$ are one except one eigenvalue, which is $1 + \|x\|^2$ and which corresponds to the eigenvector $x$. Using $\log(1+t) \le t$, we can bound $\log\det(\overline{V}_t)$ by

$$\log\det(\overline{V}_t) \le \log\det V + \sum_{k=1}^{t} \|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}} .$$

Combining $x \le 2\log(1+x)$, which holds when $x \in [0,1]$, and (9), we get

$$\sum_{k=1}^{t}\left(\|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}} \wedge 1\right) \le 2\sum_{k=1}^{t}\log\left(1 + \|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}}\right) = 2(\log\det(\overline{V}_t) - \log\det V).$$

The trace of $\overline{V}_t$ is bounded by $\mathrm{trace}(V) + tL^2$, assuming $\|m_k\| \le L$. Hence, $\det(\overline{V}_t) = \prod_{i=1}^{d}\lambda_i \le \left(\frac{\mathrm{trace}(V)+tL^2}{d}\right)^d$ and therefore,

$$\log\det(\overline{V}_t) \le d\log((\mathrm{trace}(V) + tL^2)/d),$$

finishing the proof of the second inequality. The sum $\sum_{k=1}^{t}\|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}}$ can itself be upper bounded as a function of $\log\det(\overline{V}_t)$ provided that $\lambda_0$ is large enough. Notice $\|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}} \le \lambda^{-1}_{\min}(\overline{V}_{k-1})\,\|m_{k-1}\|^2 \le L^2/\lambda_0$. Hence, we get that if $\lambda_0 \ge \max(1, L^2)$,

$$\log\frac{\det(\overline{V}_t)}{\det V} \le \sum_{k=1}^{t}\|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}} \le 2\log\frac{\det(\overline{V}_t)}{\det(V)}.$$

□

Most of this argument can be extracted from the paper of Dani et al. (2008). However, the idea goes back at least to Lai et al. (1979), Lai and Wei (1982) (a similar argument is used around Theorem 11.7 in the book by Cesa-Bianchi and Lugosi (2006)). Note that Lemmas B.9–B.11 of Rusmevichientong and Tsitsiklis (2010) also give a bound on $\sum_{k=1}^{t}\|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}}$, with an essentially identical argument. Alternatively, one can use the bounding technique of Auer (2003) (see the proof of Lemma 13 there on pages 412–413) to derive a bound like $\sum_{k=1}^{t}\|m_{k-1}\|^2_{\overline{V}^{-1}_{k-1}} \le Cd\log t$ for a suitable chosen constant $C > 0$.

*Remark* 5. By combining Corollary 1 and Lemma 4, we get a simple worst case bound that holds with probability $1-\delta$:

$$\forall t \geq 0, \quad \|S_t\|^2_{\overline{V}_t^{-1}} \leq d\, R^2 \log\left(\frac{\mathrm{trace}(V) + tL^2}{d\,\delta}\right). \tag{10}$$

Still, the new bound is considerably better than the previous one given by Theorem 2. Note that the $\log(t)$ factor cannot be removed, as shown by Problem 3, page 203 in the book by de la Peña et al. (2009).

## 3 Optional Skipping

Consider the case when $d = 1$, $m_k = \varepsilon_k \in \{0, 1\}$, i.e., the case of an optional skipping process. Then, using again $V = I = 1$, $\overline{V}_t = 1 + \sum_{k=1}^t \varepsilon_{k-1} \stackrel{\text{def}}{=} 1 + N_t$ and thus the expression studied becomes

$$\|S_t\|_{\overline{V}_t^{-1}} = \frac{|\sum_{k=1}^t \varepsilon_{k-1}\eta_k|}{\sqrt{1+N_t}}.$$

We also have

$$\log\det(\overline{V}_t) = \sum_{k=1}^t \log\left(1 + \frac{\varepsilon_{k-1}}{1+N_k}\right) \leq \sum_{k=1}^t \frac{\varepsilon_{k-1}}{1+N_k} = \sum_{k=1}^{N_t+1} \frac{1}{k} \leq 1 + \int_1^{N_t+1} x^{-1}\,dx = 1 + \log(1+N_t).$$

Thus, we get, with probability $1-\delta$

$$\forall s \geq 0, \quad \left|\sum_{k=1}^s \varepsilon_{k-1}\eta_k\right| \leq \sqrt{(1+N_s)\,\left(1 + 2\log\left(\frac{(1+N_s)^{1/2}}{\delta}\right)\right)}. \tag{11}$$

If we apply Doob's optional skipping and Hoeffding-Azuma, with a union bound (see, e.g., the paper of Bubeck et al. (2008)), we would get, for any $0 < \delta < 1$, $t \geq 2$, with probability $1-\delta$,

$$\forall s \in \{0, \dots, t\}, \qquad \left|\sum_{k=1}^s \varepsilon_{k-1}\eta_k\right| \leq \sqrt{2N_s\,\log\left(\frac{2t}{\delta}\right)}. \tag{12}$$

The major difference between these bounds is that (12) depends explicitly on $t$, while (11) does not. This has the positive effect that one need not recompute the bound if $N_t$ does not grow, which helps e.g. in the paper of Bubeck et al. (2008) to improve the computational complexity of the HOO algorithm. Also, the coefficient of the leading term in (11) under the square root is 1, whereas in (12) it is 2.

Instead of a union bound, it is possible to use a "peeling device" to replace the conservative $\log t$ factor in the above bound by essentially $\log\log t$. This is done e.g. in Garivier and Moulines (2008) in their Theorem 22.[2] From their derivations, the following one sided, uniform bound can be extracted (see Remark 24, page 19): For any $0 < \delta < 1$, $t \geq 2$, with probability $1-\delta$,

$$\forall s \in \{0, \dots, t\}, \qquad \sum_{k=1}^s \varepsilon_{k-1}\eta_k \leq \sqrt{\frac{4\,N_s}{1.99}\,\log\left(\frac{6\log t}{\delta}\right)}. \tag{13}$$

As noted by Garivier and Moulines (2008), due to the law of iterated logarithm, the scaling of the right-hand side as a function of $t$ cannot be improved in the worst-case. However, this leaves open the possibility of deriving a maximal inequality which depends on $t$ only through $N_t$.

## 4 The Multi-Armed Bandit Problem

Now we turn our attention to the multi-armed bandit problem. Let $\mu_i$ denote the expected reward of action $i$ and $\Delta_i = \mu_* - \mu_i$, where $\mu_*$ is the expected reward of the optimal action. We assume that if we choose action $I_t$ in round $t$, we obtain reward $\mu_{I_t} + \eta_t$. Let $N_{i,t}$ denote the number of times that we have played action $i$ up to time $t$, and $\bar{X}_{i,t}$ denote the average of the rewards received by action $i$ up to time $t$. From (11) with $\delta/K$ instead of $\delta$ and a union bound over the actions, we have the following confidence intervals that hold with probability at least $1-\delta$:

$$\forall i \in \{1, \dots, K\},\ \forall s \in \{1, 2, \dots\}, \qquad \left|\bar{X}_{i,s} - \mu_i\right| \leq c_{i,s}\ , \tag{14}$$

[2] They give their theorem as ratios, which they should not, since their inequality then fails to hold for $N_t = 0$. However, this is easy to remedy by reformulating their result as we do it here.

where

$$c_{i,s} = \sqrt{\frac{1+N_{i,s}}{N_{i,s}^2}\left(1+2\log\left(\frac{K(1+N_{i,s})^{1/2}}{\delta}\right)\right)}.$$

Modify the UCB Algorithm of Auer et al. (2002) to use the confidence intervals (14) and change the action selection rule accordingly. Hence, at time $t$, we choose the action

$$I_t = \operatorname*{argmax}_i \bar{X}_{i,t} + c_{i,t}. \tag{15}$$

We call this algorithm UCB($\delta$).

**Theorem 6.** *With probability at least $1-\delta$, the total regret of the UCB($\delta$) algorithm with the action selection rule* (15) *is constant and is bounded by*

$$R(T) \le \sum_{i:\Delta_i>0}\left(3\Delta_i + \frac{16}{\Delta_i}\log\frac{2K}{\Delta_i\delta}\right).$$

*where $i_*$ is the index of the optimal action.*

*Proof.* Suppose the confidence intervals do not fail. If we play action $i$, the upper estimate of the action is above $\mu^*$. Hence,

$$c_{i,s} \ge \frac{\Delta_i}{2}.$$

Substituting $c_{i,s}$ and squaring gives

$$\frac{N_{i,s}^2-1}{N_{i,s}+1} \le \frac{N_{i,s}^2}{N_{i,s}+1} \le \frac{4}{\Delta_i^2}\left(1+2\log\frac{K(1+N_{i,s})^{1/2}}{\delta}\right).$$

By using Lemma 8 of Antos et al. (2010), we get that

$$N_{i,s} \le 3 + \frac{16}{\Delta_i^2}\log\frac{2K}{\Delta_i\delta}.$$

Thus, using $R(T) = \sum_{i\neq i_*}\Delta_i N_{i,T}$, we get that with probability at least $1-\delta$, the total regret is bounded by

$$R(T) \le \sum_{i:\Delta_i>0}\left(3\Delta_i + \frac{16}{\Delta_i}\log\frac{2K}{\Delta_i\delta}\right).$$

□

*Remark* 7. Lai and Robbins (1985) prove that for any suboptimal arm $j$,

$$\mathbb{E}[N_{i,t}] \ge \frac{\log t}{D(p_j, p_*)},$$

where, $p_*$ and $p_j$ are the reward density of the optimal arm and arm $j$ respectively, and $D$ is the KD-divergence. This lower bound does not contradict Theorem 6, as Theorem 6 only states a high probability upper bound for the regret. Note that UCB($\delta$) takes delta as its input. Because with probability $\delta$, the regret in time $t$ can be $t$, on expectation, the algorithm might have a regret of $t\delta$. Now if we select $\delta = 1/t$, then we get $O(\log t)$ upper bound on the expected regret.

## 5 Application to Least Squares Estimation and Linear Bandit Problem

In this section we first apply Theorem 3 to derive confidence intervals for least-squares estimation, where the covariate process is an arbitrary process and then use these confidence intervals to improve the regret bound of Dani et al. (2008) for the linear bandit problem. In particular, our assumption on the data is as follows:

**Assumption A1** Let $(\mathcal{F}_i)$ be a filtration, $(x_1, y_1)$, …, $(x_t, y_t)$ be a sequence of random variables over $\mathbb{R}^d \times \mathbb{R}$ such that $x_i$ is $\mathcal{F}_i$-measurable, and $y_i$ is $\mathcal{F}_{i+1}$-measurable ($i = 1, 2, \ldots$). Assume that there exists $\theta_* \in \mathbb{R}^d$ such that $\mathbb{E}[y_i|\mathcal{F}_i] = x_i^\top\theta_*$, i.e., $\varepsilon_i = y_i - x_i^\top\theta_*$ is a martingale difference sequence ($\mathbb{E}[\varepsilon_i|\mathcal{F}_i] = 0$, $i = 1, 2, \ldots$) and that $\varepsilon_i$ is sub-Gaussian: There exists $R > 0$ such that for any $\gamma \in \mathbb{R}$,

$$\mathbb{E}[\exp(\gamma\varepsilon_i)|\mathcal{F}_{i-1}] \le \exp(\gamma^2R^2/2).$$

We shall call the random variables $x_i$ covariates and the random variables $y_i$ the responses. Note that the assumption allows any sequential generation of the covariates.

Let $\hat{\theta}_t$ be the $\ell^2$-regularized least-squares estimate of $\theta_*$ with regularization parameter $\lambda > 0$:

$$\hat{\theta}_t = (X^\top X + \lambda I)^{-1} X^\top Y, \qquad \hat{\theta}_0 = 0, \tag{16}$$

where $X$ is the matrix whose rows are $x_1^\top, \ldots, x_{t-1}^\top$ and $Y = (y_1, \ldots, y_{t-1})^\top$. We further let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{t-1})^\top$.

We are interested in deriving a confidence bound on the error of predicting the mean response $x^\top \theta_*$ at an arbitrarily chosen random covariate $x$ using the least-squares predictor $x^\top \hat{\theta}_t$. Using

$$\begin{aligned}
\hat{\theta}_t &= (X^\top X + \lambda I)^{-1} X^\top (X\theta_* + \varepsilon) \\
&= (X^\top X + \lambda I)^{-1} X^\top \varepsilon + (X^\top X + \lambda I)^{-1} (X^\top X + \lambda I)\theta_* - \lambda (X^\top X + \lambda I)^{-1} \theta_* \\
&= (X^\top X + \lambda I)^{-1} X^\top \varepsilon + \theta_* - \lambda (X^\top X + \lambda I)^{-1} \theta_* \,,
\end{aligned}$$

we get

$$\begin{aligned}
x^\top \hat{\theta}_t - x^\top \theta_* &= x^\top (X^\top X + \lambda I)^{-1} X^\top \varepsilon - \lambda x^\top (X^\top X + \lambda I)^{-1} \theta_* \\
&= \langle x, X^\top \varepsilon \rangle_{V_t^{-1}} - \lambda \langle x, \theta_* \rangle_{V_t^{-1}},
\end{aligned}$$

where $V_t = X^\top X + \lambda I$. Note that $V_t$ is positive definite (thanks to $\lambda > 0$) and hence so is $V_t^{-1}$, so the above inner product is well-defined. Using the Cauchy-Schwartz inequality, we get

$$\begin{aligned}
|x^\top \hat{\theta}_t - x^\top \theta_*| &\leq \|x\|_{V_t^{-1}} \left( \left\| X^\top \varepsilon \right\|_{V_t^{-1}} + \lambda \, \|\theta_*\|_{V_t^{-1}} \right) \\
&\leq \|x\|_{V_t^{-1}} \left( \left\| X^\top \varepsilon \right\|_{V_t^{-1}} + \lambda^{1/2} \, \|\theta_*\| \right),
\end{aligned}$$

where we used that $\|\theta_*\|_{V_t^{-1}}^2 \leq 1/\lambda_{\min}(V_t) \, \|\theta_*\|^2 \leq 1/\lambda \, \|\theta_*\|^2$. Fix any $0 < \delta < 1$. By Corollary 1, with probability at least $1 - \delta$,

$$\forall t \geq 1, \quad \left\| X^\top \varepsilon \right\|_{V_t^{-1}} \leq R \sqrt{2 \log \left( \frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)}.$$

Therefore, on the event where this inequality holds, one also has

$$|x^\top \hat{\theta}_t - x^\top \theta_*| \leq \|x\|_{V_t^{-1}} \left( R \sqrt{2 \log \left( \frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} \, \|\theta_*\| \right).$$

Similarly, we can derive a worst-case bound. The result is summarized in the following statement:

**Theorem 8.** *Let $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ satisfy the linear model Assumption A1 with some $R > 0$, $\theta_* \in \mathbb{R}^d$ and let $(\mathcal{F}_t)$ be the associated filtration. Assume that w.p.1 the covariates satisfy $\|x_i\| \leq L$, $i = 1, \ldots, n$ and $\|\theta_*\| \leq S$. Consider the $\ell^2$-regularized least-squares parameter estimate $\hat{\theta}_n$ with regularization coefficient $\lambda > 0$ (cf. (16)). Let $x$ be an arbitary, $\mathbb{R}^d$-valued random variable. Let $V_t = \lambda I + \sum_{i=1}^{t-1} x_i x_i^\top$ be the regularized design matrix underlying the covariates. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\forall t \geq 1, \quad |x^\top \hat{\theta}_t - x^\top \theta_*| \leq \|x\|_{V_t^{-1}} \left( R \sqrt{2 \log \left( \frac{\det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} \, S \right). \tag{17}$$

*Similarly, with probability $1 - \delta$,*

$$\forall t \geq 1, \quad |x^\top \hat{\theta}_t - x^\top \theta_*| \leq \|x\|_{V_t^{-1}} \left( R \sqrt{d \log \left( \frac{1 + \frac{tL}{\lambda}}{\delta} \right)} + \lambda^{1/2} \, S \right). \tag{18}$$

*Remark* 9. We see that $\lambda \to \infty$ increases the second term (the "bias term") in the parenthesis of the estimate. In fact, $\lambda \to \infty$ for $n$ fixed gives $\lambda^{1/2} \, \|x\|_{V_t^{-1}} \to \text{const}$ (as it should be). Decreasing $\lambda$, on the other hand increases $\|x\|_{V_t^{-1}}$ and the log term, while it decreases the bias term $\lambda^{1/2} S$.

From the above result, we immediately obtain confidence bounds for $\theta_*$:

**Corollary 10.** *Under the condition of Theorem 8, with probability at least* $1-\delta$,

$$\forall t \geq 1, \quad \left\|\hat{\theta}_t - \theta_*\right\|_{V_t} \leq R\sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S.$$

*Also, with probability at least* $1-\delta$,

$$\forall t \geq 1, \quad \left\|\hat{\theta}_t - \theta_*\right\|_{V_t} \leq R\sqrt{d\log\left(\frac{1+\frac{tL}{\lambda}}{\delta}\right)} + \lambda^{1/2} S.$$

*Proof.* Plugging in $x = V_t(\hat{\theta}_t - \theta_*)$ into (17), we get

$$\left\|\hat{\theta}_t - \theta_*\right\|_{V_t}^2 \leq \left\|V_t(\hat{\theta}_t - \theta_*)\right\|_{V_t^{-1}} \left(R\sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2} S\right). \tag{19}$$

Now, $\left\|V_t(\hat{\theta}_t - \theta_*)\right\|_{V_t^{-1}}^2 = \left\|\hat{\theta}_t - \theta_*\right\|_{V_t}^2$ and therefore either $\left\|\hat{\theta}_t - \theta_*\right\|_{V_t} = 0$, in which case the conclusion holds, or we can divide both sides of (19) by $\left\|\hat{\theta}_t - \theta_*\right\|_{V_t}$ to obtain the desired result. □

*Remark* 11. In fact, the theorem and the corollary are equivalent. To see this note that $x^\top(\hat{\theta}_t - \theta_*) = (\hat{\theta}_t - \theta_*)^\top V_t^{1/2} V_t^{-1/2} x$, thus

$$\sup_{x \neq 0} \frac{|x^\top(\hat{\theta}_t - \theta_*)|}{\|x\|_{V_t^{-1}}} = \left\|\hat{\theta}_t - \theta_*\right\|_{V_t}.$$

*Remark* 12. The above bound could be compared with a similar bound of Dani et al. (2008) whose bound, under identical conditions, states that (with appropriate initialization) with probability $1-\delta$,

$$\text{for all } t \text{ large enough,} \qquad \left\|\hat{\theta}_t - \theta_*\right\|_{V_t} \leq R \max\left\{\sqrt{128\, d\log(t)\,\log\left(\frac{t^2}{\delta}\right)}, \frac{8}{3}\,\log\left(\frac{t^2}{\delta}\right)\right\}, \tag{20}$$

where large enough means that $t$ satisfies $0 < \delta < t^2 e^{-1/16}$. Denote by $\sqrt{\beta_t(\delta)}$ the right-hand side in the above bound. The restriction on $t$ comes from the fact that $\beta_t(\delta) \geq 2d(1 + 2\log(t))$ is needed in the proof of the last inequality of their Theorem 5.

On the other hand, Theorem 2 gives rise to the following result: For any *fixed* $t \geq 2$, for any $0 < \delta < 1$, with probability at least $1-\delta$,

$$\left\|\hat{\theta}_t - \theta_*\right\|_{V_t} \leq 2\,\kappa^2 R\sqrt{\log t}\,\sqrt{d\,\log(t) + \log(1/\delta)} + \lambda^{1/2} S,$$

where $\kappa$ is as in Theorem 2. To get a uniform bound one can use a union bound with $\delta_t = \delta/t^2$. Then $\sum_{t=2}^{\infty} \delta_t = \delta(\frac{\pi^2}{6} - 1) \leq \delta$. This thus gives that for any $0 < \delta < 1$, with probability at least $1-\delta$,

$$\text{for all} \quad t \geq 2, \quad \left\|\hat{\theta}_t - \theta_*\right\|_{V_t} \leq 2\,\kappa^2 R\sqrt{\log t}\,\sqrt{d\,\log(t) + \log(t^2/\delta)} + \lambda^{1/2} S,$$

This looks tighter than (20), but is still lagging beyond the result of Corollary 10.

### 5.1 The Linear Bandit Problem

We now turn our attention to the linear bandit problem. Assume the actions lie in $\mathcal{D} \subset \mathbb{R}^d$ and for any $x \in \mathcal{D}$, $\|x\|^2 \leq L$. Assume the reward of taking action $x \in \mathcal{D}$ has the form of

$$h_t(x) = \theta_*^\top x + \eta_t$$

and assume $\forall x \in \mathcal{D}$, $\theta_*^\top x \in [-1, 1]$. Define the regret by

$$R(T) = \sum_{t=1}^{T} (\theta_*^\top x_* - \theta_*^\top x_t),$$

where $x_*$ is the optimal action ($x_* = \operatorname{argmax}_{x \in \mathcal{D}} \theta_*^\top x$). Define the confidence set

$$\mathcal{C}_t(\delta) = \left\{\theta : (\theta - \hat{\theta}_t)^\top V_t(\theta - \hat{\theta}_t) \leq \beta_t(\delta)\right\}, \tag{21}$$

| |
|---|
| **Input**: Confidence $0 < \delta < 1$.<br>**for** $t := 1, 2, \ldots$ **do**<br>$(\tilde{\theta}_t, x_t) = \mathrm{argmax}_{(\theta,x)\in\mathcal{C}_t(\delta)\times\mathcal{D}}\, \theta^\top x$.<br>Play $x_t$ and observe reward $h_t(x_t)$.<br>Update $V_t$ and $C_t$.<br>**end for** |

Table 1: The Linear Bandit Algorithm

where

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

Consider the ConfidenceBall algorithm of Dani et al. (2008). We use the confidence intervals (21) and change the action selection rule accordingly. Hence, at time $t$, we define $\tilde{\theta}_t$ and $x_t$ by the following equation:

$$(\tilde{\theta}_t, x_t) = \underset{(\theta,x)\in\mathcal{C}_t(\delta)\times\mathcal{D}}{\mathrm{argmax}}\ \theta^\top x. \tag{22}$$

The algorithm is shown in Table 1.

**Theorem 13.** *With probability at least* $1-\delta$*, the regret of the Linear Bandit Algorithm shown in Table 1 satisfies*

$$\forall T \geq 1, \quad R(T) \leq 4\sqrt{Td\log(\lambda + TL/d)}\left(\lambda^{1/2}S + R\sqrt{2\log 1/\delta + d\log(1 + TL/(\lambda d))}\right).$$

*Proof.* Lets decompose the instantaneous regret as follows:

$$\begin{aligned} r_t &= \theta_*^\top x_* - \theta_*^\top x_t \\ &\leq \tilde{\theta}_t^\top x_t - \theta_*^\top x_t \\ &= (\tilde{\theta}_t - \theta_*)^\top x_t \\ &= (\hat{\theta}_t - \theta_*)^\top x_t + (\tilde{\theta}_t - \hat{\theta}_t)^\top x_t \\ &\leq \sqrt{\beta_t(\delta)}\, \|x_t\|_{V_t^{-1}}\,, \end{aligned} \tag{23}$$

where the last step holds by Cauchy-Schwarz. Using (23) and the fact that $r_t \leq 2$, we get that

$$r_t \leq 2\min(\sqrt{\beta_t(\delta)}\,\|x_t\|^2_{V_t^{-1}}, 1) \leq 2\sqrt{\beta_t(\delta)}\min(\|x_t\|^2_{V_t^{-1}}, 1).$$

Thus, with probability at least $1-\delta$, $\forall T \geq 1$

$$\begin{aligned} R(T) &\leq \sqrt{T\sum_{t=1}^T r_t^2} \leq \sqrt{8\beta_T T\sum_{t=1}^T \min(w_t^2, 1)} \leq 4\sqrt{\beta_T T\log(\det(V_T))} \\ &\leq 4\sqrt{Td\log(\lambda + tL/d)}\left(\lambda^{1/2}S + R\sqrt{2\log 1/\delta + d\log(1 + tL/(\lambda d))}\right). \end{aligned}$$

where the last two steps follow from Lemma 4. □

### 5.2 Saving Computation

The action selection rule (22) is NP-hard in general (Dani et al., 2008). In this section, we show that we essentially need to solve this problem only $O(\log t)$ times up to time $t$ and hence saving computations. Algorithm 2 achieves this objective by changing its policy only when the volume of the confidence set is halved and still enjoyes almost the same regret bound as for Algorithm 1.

**Theorem 14.** *With probability at least* $1-\delta$, $\forall T \geq 1$*, the regret of the Linear Bandit Algorithm shown in Table 2 satisfies*

$$R(T) \leq 4\sqrt{2Td\log(\lambda + TL/d)}\left(\lambda^{1/2}S + R\sqrt{2\log 1/\delta + d\log(1 + TL/(\lambda d))}\right) + 4\sqrt{d\log(T/d)}.$$

**Input**: Confidence $0 < \delta < 1$.
$\tau = 1$ {This is the last timestep that we changed the action}
**for** $t := 1, 2, \ldots$ **do**
  **if** $\det(V_t) > 2\det(V_\tau)$ **then**
    $(\tilde{\theta}_t, x_t) = \operatorname{argmax}_{(\theta,x)\in\mathcal{C}_t(\delta)\times\mathcal{D}} \theta^\top x$.
    $\tau = t$.
  **end if**
  $x_t = x_\tau$.
  Play $x_t$ and observe reward $h_t(x_t)$.
**end for**

Table 2: The Linear Bandit Algorithm

First, we prove the following lemma:

**Lemma 15.** *Let $A$, $B$ and $C$ be positive semi-definite matrices such that $A = B + C$. Then, we have that*

$$\sup_{x\neq 0} \frac{x^\top A x}{x^\top B x} \leq \frac{\det(A)}{\det(B)}.$$

*Proof.* We consider first a simple case. Let $A = B + mm^\top$, $B$ positive definite. Let $x \neq 0$ be an arbitrary vector. Using the Cauchy-Schwartz inequality, we get

$$(x^\top m)^2 = (x^\top B^{1/2} B^{-1/2} m)^2 \leq \left\| B^{1/2} x \right\|^2 \left\| B^{-1/2} m \right\|^2 = \|x\|_B^2 \|m\|_{B^{-1}}^2 .$$

Thus,

$$x^\top (B + mm^\top) x \leq x^\top B x + \|x\|_B^2 \|m\|_{B^{-1}}^2 = (1 + \|m\|_{B^{-1}}^2) \|x\|_B^2$$

and so

$$\frac{x^\top A x}{x^\top B x} \leq 1 + \|m\|_{B^{-1}}^2 .$$

We also have that

$$\det(A) = \det(B + mm^\top) = \det(B)\det(I + B^{-1/2} m (B^{-1/2} m)^\top) = \det(B)(1 + \|m\|_{B^{-1}}^2),$$

thus finishing the proof of this case.

If $A = B + m_1 m_1^\top + \cdots + m_{t-1} m_{t-1}^\top$, then define $V_s = B + m_1 m_1^\top + \cdots + m_{s-1} m_{s-1}^\top$ and use

$$\frac{x^\top A x}{x^\top B x} = \frac{x^\top V_t x}{x^\top V_{t-1} x} \frac{x^\top V_{t-1} x}{x^\top V_{t-2} x} \cdots \frac{x^\top V_2 x}{x^\top B x}.$$

By the above argument, since all the terms are positive, we get

$$\frac{x^\top A x}{x^\top B x} \leq \frac{\det(V_t)}{\det(V_{t-1})} \frac{\det(V_{t-1})}{\det(V_{t-2})} \cdots \frac{\det(V_2)}{\det(B)} = \frac{\det(V_t)}{\det(B)} = \frac{\det(A)}{\det(B)}.$$

This finishes the proof of this case.

Now, if $C$ is a positive definite matrix, then the eigendecomposition of $C$ gives $C = U^\top \Lambda U$ , where $U$ is orthonormal and $\Lambda$ is positive diagonal matrix. This, in fact gives that $C$ can be written as the sum of at most $d$ rank-one matrices, finishing the proof for the general case.

□

*Proof of Theorem 14.* Let $\tau_t$ be the smallest timestep $\leq t$ such that $x_t = x_{\tau_t}$. By an argument similar to the one used in Theorem 13, we have

$$r_t \leq (\hat{\theta}_{\tau_t} - \theta_*)^\top x_t + (\tilde{\theta}_{\tau_t} - \hat{\theta}_{\tau_t})^\top x_t.$$

We also have that for all $\theta \in \mathcal{C}_{\tau_t}$ and $x$,

$$\begin{aligned}
\left|(\theta - \hat{\theta}_{\tau_t})^\top x\right| &\le \left\|V_t^{1/2}(\theta - \hat{\theta}_\tau)\right\| \sqrt{x^\top V_t^{-1} x} \\
&\le \left\|V_{\tau_t}^{1/2}(\theta - \hat{\theta}_{\tau_t})\right\| \sqrt{\frac{\det(V_t)}{\det(V_{\tau_t})}} \sqrt{x^\top V_t^{-1} x} \\
&\le \sqrt{2}\left\|V_{\tau_t}^{1/2}(\theta - \hat{\theta}_\tau)\right\| \sqrt{x^\top V_t^{-1} x} \\
&\le \sqrt{2\beta_{\tau_t}}\sqrt{x^\top V_t^{-1} x},
\end{aligned}$$

where the second step follows from Lemma 15, and the third step follows from the fact that at time $t$ we have $\det(V_t) < 2\det(V_{\tau_t})$. The rest of the argument is identical to that of Theorem 13. We conclude that with probability at least $1-\delta$, $\forall T \ge 1$,

$$R(T) \le 4\sqrt{2Td\log(\lambda + tL/d)}\left(\lambda^{1/2}S + R\sqrt{2\log 1/\delta + d\log(1 + tL/(\lambda d))}\right).$$

□

### 5.3 Problem Dependent Bound ($\Delta > 0$)

Let $\Delta$ be as defined in (Dani et al., 2008). In this section we assume that $\Delta > 0$. This includes the case when the action set is a polytope. First we state a matrix perturbation theorem from Stewart and Sun (1990) that will be used later.

**Theorem 16** (Stewart and Sun (1990), Corollary 4.9)**.** *Let $A$ be a symmetric matrix with eigenvalues $\nu_1 \ge \nu_2 \ge \ldots \ge \nu_d$, $E$ be a symmetric matrix with eigenvalues $e_1 \ge e_2 \ge \ldots \ge e_d$, and $V = A + E$ denote a symmetric perturbation of $A$ such that the eigenvalues of $V$ are $\tilde{\nu}_1 \ge \tilde{\nu}_2 \ge \ldots \ge \tilde{\nu}_d$. Then, for $i = 1, \ldots, d$,*

$$\tilde{\nu}_i \in [\nu_i + e_d, \nu_i + e_1].$$

**Theorem 17.** *Assume that $\Delta > 0$ for the gap $\Delta$ defined in (Dani et al., 2008). Further assume that $\lambda \ge 1$ and $S \ge 1$. With probability at least $1-\delta$, $\forall T \ge 1$, the regret of the algorithm shown in Table 1 satisfies*

$$R(T) = \frac{16R^2\lambda S^2}{\Delta}\left(\log(LT) + (d-1)\log\frac{64R^2\lambda S^2 L}{\Delta^2} + 2(d-1)\log\left(d\log\frac{d\lambda + TL^2}{d} + 2\log(1/\delta)\right) + 2\log(1/\delta)\right)^2.$$

*Proof.* First we bound the regret in terms of $\log\det(V_T)$. We have that

$$R(T) = \sum_{t=1}^T r_t \le \sum_{t=1}^T \frac{r_t^2}{\Delta} \le \frac{16\beta_T}{\Delta}\log(\det(V_T)), \tag{24}$$

where the first inequality follows from the fact that either $r_t = 0$ or $\Delta < r_t$, and the second inequality can be extracted from the proof of Theorem 13. Let $b_t$ be the number of times we have played a sub-optimal action (an action $x_s$ for which $\theta_*^\top x_* - \theta_*^\top x_s \ge \Delta$) up to time $t$. Next we bound $\log\det(V_t)$ in terms of $b_t$. We bound the eigenvalues of $V_t$ by using Theorem 16.

Let $E_t = \sum_{s:x_s \ne x_*}^t x_s x_s^\top$ and $A_t = V_t - E_t = (t - b_t)x_* x_*^\top$. The only non-zero eigenvalue of $(t-b_t)x_* x_*^\top$ is $(t-b_t)L^*$, where $L^* = x_*^\top x_* \le L$. Let the eigenvalues of $V_t$ and $E_t$ be $\lambda_1 \ge \cdots \ge \lambda_d$ and $e_1 \ge \cdots \ge e_d$ respectively. By Theorem 16, we have that

$$\lambda_1 \in [(t-b_t)L^* + e_d, (t-b_t)L^* + e_1] \quad \text{and} \quad \forall i \in \{2,\ldots,d\},\, \lambda_i \in [e_d, e_1].$$

Thus,

$$\det(V_t) = \prod_i^d \lambda_i \le ((t-b_t)L^* + e_1)e_1^{d-1} \le ((t-b_t)L + e_1)e_1^{d-1}.$$

Therefore,

$$\log\det(V_t) \le \log((t-b_t)L + e_1) + (d-1)\log e_1.$$

Because $\mathrm{trace}(E) = \sum_{s:x_s \ne x_*}^t \mathrm{trace}(x_s x_s^\top) \le Lb_t$, we conclude that $e_1 \le Lb_t$. Thus,

$$\begin{aligned}
\log\det(V_t) &\le \log((t-b_t)L + Lb_t) + (d-1)\log(Lb_t) \\
&= \log(Lt) + (d-1)\log(Lb_t). \qquad (25)
\end{aligned}$$

With some calculations, we can show that

$$\beta_t \log\det V_t \le 4R^2\lambda S^2 (2\log(1/\delta) + \log\det V_t)^2 \le 4R^2\lambda S^2 \left(d\log\frac{d\lambda + tL^2}{d} + 2\log\frac{1}{\delta}\right)^2, \qquad (26)$$

where the second inequality follows from Lemma 4. Hence,

$$b_t \le \frac{16\beta_t}{\Delta^2}\log(\det(V_t)) \le \frac{64R^2\lambda S^2}{\Delta^2}\left(d\log\frac{d\lambda + tL^2}{d} + 2\log\frac{1}{\delta}\right)^2, \qquad (27)$$

where the first inequality follows from $R(t) \ge b_t\Delta$. Thus, with probability $1-\delta$, $\forall T \ge 1$,

$$\begin{aligned}
R(T) &\le \frac{16\beta_T}{\Delta}\log(\det(V_T)) \\
&\le \frac{64R^2\lambda S^2}{\Delta}(\log(\det(V_T)) + 2\log(1/\delta))^2 \\
&\le \frac{16R^2\lambda S^2}{\Delta}(\log(LT) + (d-1)\log(Lb_T) + 2\log(1/\delta))^2 \\
&\le \frac{16R^2\lambda S^2}{\Delta}\left(\log(LT) + (d-1)\log\frac{64R^2\lambda S^2 L}{\Delta^2} + 2(d-1)\log\left(d\log\frac{d\lambda + TL^2}{d} + 2\log(1/\delta)\right) + 2\log(1/\delta)\right)^2,
\end{aligned}$$

where the first step follows from (24), the second step follows from the first inequality in (26), the third step follows from (25), and the last step follows from the second inequality in (27). □

*Remark* 18. The problem dependent regret of (Dani et al., 2008) scales like $O(\frac{d^2}{\Delta}\log^3 T)$, while our bound scales like $O(\frac{1}{\Delta}(\log^2 T + d\log T + d^2\log\log T))$.

## References


A. Antos, V. Grover, and Cs. Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29-30):2712–2728, 2010.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2003. ISSN 1533-7928.

S. Bubeck, R. Munos, G. Stoltz, and Cs. Szepesvári. Online optimization in X-armed bandits. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS-21*, pages 201–208. MIT Press, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

V. Dani, T.P. Hayes, and S.M. Kakade. Stochastic linear optimization under bandit feedback. *COLT-2008*, pages 355–366, 2008.

V.H. de la Peña, M.J. Klass, and T.L. Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of Probability*, 32(3):1902–1933, 2004.

V.H. de la Peña, T.L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

D.A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

A Garivier and E Moulines. On upper-confidence bound policies for non-stationary bandit problems. Technical report, LTCI, Dec 2008. URL `http://arxiv.org/pdf/0805.3415`.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

T.L. Lai and C.Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

T.L. Lai, H. Robbins, and C.Z. Wei. Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Sciences*, 75(7):3034–3036, 1979.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

H. Robbins and D. Siegmund. Boundary crossing probabilities for the Wiener process and sample sums. *Annals of Math. Statistics*, 41:1410–1429, 1970.

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

G.W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.