A Proof of Theorem 1

First, let's review the lower bound of the linear bandit setting. The linear bandit setting is almost identical to ours except that the θ_t's do not vary across rounds, and are equal to the same (unknown) θ, *i.e.*, $\forall t \in [T]\ \theta_t = \theta$.

Lemma 4 ([20]). *For any $T_0 \geq \sqrt{d}/2$ and let $D = \left\{x \in \Re^d : \|x\| \leq 1\right\}$, then there exists a $\theta \in \left\{\pm\sqrt{d/4T_0}\right\}^d$, such that the worst case regret of any algorithm for linear bandits with unknown parameter θ is $\Omega(d\sqrt{T_0})$.*

Going back to the non-stationary environment, suppose nature divides the whole time horizon into $\lceil T/H \rceil$ blocks of equal length H rounds (the last block can possibly have less than H rounds), and each block is a decoupled linear bandit instance so that the knowledge of previous blocks cannot help the decision within the current block. Following Lemma 4, we restrict the sequence of θ_t's are drawn from the set $\left\{\pm\sqrt{d/4H}\right\}^d$. Moreover, θ_t's remain fixed within a block, and can vary across different blocks, *i.e.*,

$$\forall i \in \left[\left\lceil \frac{T}{H} \right\rceil\right] \forall t_1, t_2 \in [(i-1)H + 1, i \cdot H \wedge T] \quad \theta_{t_1} = \theta_{t_2}. \tag{27}$$

We argue that even if the learner knows this additional information, it still incur a regret $\Omega(d^{2/3}B_T^{1/3}T^{2/3})$. Note that different blocks are completely decoupled, and information is thus not passed across blocks. Therefore, the regret of each block is $\Omega\left(d\sqrt{H}\right)$, and the total regret is at least

$$\left(\left\lceil \frac{T}{H} \right\rceil - 1\right)\Omega\left(d\sqrt{H}\right) = \Omega\left(dTH^{-\frac{1}{2}}\right). \tag{28}$$

Intuitively, if H, the number of length of each block, is smaller, the worst case regret lower bound becomes larger. But too small a block length can result in a violation of the variation budget. So we work on the total variation of θ_t's to see how small can H be. The total variation of the θ_t's can be seen as the total variation across consecutive blocks as θ_t remains unchanged within a single block. Observe that for any pair of $\theta, \theta' \in \left\{\pm\sqrt{d/4H}\right\}^d$, the ℓ_2 difference between θ and θ' is upper bounded as

$$\sqrt{\sum_{i=1}^{d} \frac{4d}{4H}} = \frac{d}{\sqrt{H}} \tag{29}$$

and there are at most $\lfloor T/H \rfloor$ changes across the whole time horizon, the total variation is at most

$$B = \frac{T}{H} \cdot \frac{d}{\sqrt{H}} = dTH^{-\frac{3}{2}}. \tag{30}$$

By definition, we require that $B \leq B_T$, and this indicates that

$$H \geq (dT)^{\frac{2}{3}} B_T^{-\frac{2}{3}}. \tag{31}$$

Taking $H = \left\lceil (dT)^{\frac{2}{3}} B_T^{-\frac{2}{3}} \right\rceil$, the worst case regret is

$$\Omega\left(dT\left((dT)^{\frac{2}{3}} B_T^{-\frac{2}{3}}\right)^{-\frac{1}{2}}\right) = \Omega\left(d^{\frac{2}{3}} B_T^{\frac{1}{3}} T^{\frac{2}{3}}\right). \tag{32}$$

B Proof of Lemma 1

In the proof, we denote $B(1)$ as the unit Euclidean ball, and $\lambda_{\max}(M)$ as the maximum eigenvalue of a square matrix M. By folklore, we know that $\lambda_{\max}(M) = \max_{z \in B(1)} z^\top M z$. In addition, recall the definition that $V_{t-1}^{-1} = \lambda I + \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top$ We prove the Lemma as follows:

$$\left\| V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top (\theta_s - \theta_t) \right\| = \left\| V_{t-1}^{-1} \sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top \left[\sum_{p=s}^{t-1} (\theta_p - \theta_{p+1})\right] \right\|$$

$$= \left\| V_{t-1}^{-1} \sum_{p=1\vee(t-w)}^{t-1} \sum_{s=1\vee(t-w)}^{p} X_s X_s^\top (\theta_p - \theta_{p+1}) \right\| \tag{33}$$

$$\leq \sum_{p=1\vee(t-w)}^{t-1} \left\| V_{t-1}^{-1} \left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right) (\theta_p - \theta_{p+1}) \right\| \tag{34}$$

$$\leq \sum_{p=1\vee(t-w)}^{t-1} \lambda_{\max} \left(V_{t-1}^{-1} \left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right) \right) \|\theta_p - \theta_{p+1}\| \tag{35}$$

$$\leq \sum_{p=1\vee(t-w)}^{t-1} \|\theta_p - \theta_{p+1}\| . \tag{36}$$

Equality (33) is by the observation that both sides of the equation is summing over the terms $X_s X_s^\top(\theta_p - \theta_{p+1})$ with indexes (s, p) ranging over $\{(s, p) : 1 \vee (t-w) \leq s \leq p \leq t-1\}$. Inequality (34) is by the triangle inequality.

To proceed with the remaining steps, we argue that, for any index subset $S \subseteq \{1 \vee (t-w), \ldots, t-1\}$, the matrix $V_{t-1}^{-1} \left(\sum_{s\in S} X_s X_s^\top \right)$ is positive semi-definite (PSD). Now, let's denote $A = \sum_{s\in S} X_s X_s^\top$. Evidently, matrix A is PSD, while matrix V_{t-1}^{-1} is positive definite, and both matrices A, V_{t-1}^{-1} are symmetric. Matrices $V_{t-1}^{-1}A$ and $V_{t-1}^{-1/2} A V_{t-1}^{-1/2}$ have the same sets of eigenvalues, since these matrices have the same characteristics polynomial (with the variable denoted as η below):

$$\begin{aligned} \det(\eta I - V_{t-1}^{-1}A) &= \det(V_{t-1}^{-1/2}) \det(\eta V_{t-1}^{1/2} - V_{t-1}^{-1/2} A) \\ &= \det(\eta V_{t-1}^{1/2} - V_{t-1}^{-1/2} A) \det(V_{t-1}^{-1/2}) = \det(\eta I - V_{t-1}^{-1/2} A V_{t-1}^{-1/2}). \end{aligned}$$

Evidently, $V_{t-1}^{-1/2} A V_{t-1}^{-1/2}$ is PSD, since for any $y \in \mathbb{R}^d$ we clearly have $y^\top V_{t-1}^{-1/2} A V_{t-1}^{-1/2} y = \|A^{1/2} V_{t-1}^{-1/2} y\|^2 \geq 0$ (Matrices $A^{1/2}, V_{t-1}^{-1/2}$ are symmetric). Altogether, we have shown that $V_{t-1}^{-1} \left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right)$ is PSD.

Inequality (35) is by the fact that, for any matrix $M \in \mathbb{R}^{d\times d}$ with $\lambda_{\max}(M) \geq 0$ and any vector $y \in \mathbb{R}^d$, we have $\|My\| \leq \lambda_{\max}(M) \|y\|$. Without loss of generality, assume $y \neq 0$. Now, it is evident that

$$\|My\| = \left\| M \frac{y}{\|y\|} \right\| \cdot \|y\| \leq \left\| \max_{z\in B(1)} Mz \right\| \cdot \|y\| = |\lambda_{\max}(M)| \cdot \|y\| = \lambda_{\max}(M) \|y\| .$$

Applying the above claim with $M = V_{t-1}^{-1} \left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right)$, which is PSD, and $y = \theta_p - \theta_{p+1}$ demonstrates inequality (35).

Finally, inequality (36) is by the inequality $\lambda_{\max} \left(V_{t-1}^{-1} \left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right) \right) \leq 1$. Indeed,

$$\begin{aligned} &\lambda_{\max} \left(V_{t-1}^{-1} \left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right) \right) = \max_{z\in B(1)} z^\top V_{t-1}^{-1} \left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right) z \\ &\leq \max_{z\in B(1)} \left\{ z^\top V_{t-1}^{-1} \left(\sum_{s=1\vee(t-w)}^{p} X_s X_s^\top \right) z + z^\top V_{t-1}^{-1} \left(\sum_{s=p+1}^{t-1} X_s X_s^\top \right) z + \lambda z^\top V_{t-1}^{-1} z \right\} \qquad (37) \\ &= \max_{z\in B(1)} z^\top V_{t-1}^{-1} V_{t-1} z = 1, \end{aligned}$$

where inequality (37) is by the property that both matrices $V_{t-1}^{-1} \left(\sum_{s=p+1}^{t-1} X_s X_s^\top \right), V_{t-1}^{-1}$ are PSD, as we establish previously. Altogether, the Lemma is proved.

C Proof of Theorem 2

Fixed any $\delta \in [0,1]$, we have that for any $t \in [T]$ and any $x \in D_t$,

$$
\begin{aligned}
\left|x^\top(\hat{\theta}_t - \theta_t)\right| &= \left|x^\top\left(V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top(\theta_s - \theta_t)\right) + x^\top V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s - \lambda\theta_t\right)\right| \\
&\leq \left|x^\top\left(V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top(\theta_s - \theta_t)\right)\right| + \left|x^\top V_{t-1}^{-1}\left(\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s - \lambda\theta_t\right)\right| \qquad (38)\\
&\leq \|x\| \cdot \left\|V_{t-1}^{-1}\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top(\theta_s - \theta_t)\right\| + \|x\|_{V_{t-1}^{-1}}\left\|\sum_{s=1\vee(t-w)}^{t-1}\eta_s X_s - \lambda\theta_t\right\|_{V_{t-1}^{-1}} \qquad (39)\\
&\leq L\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s - \theta_{s+1}\| + \|x\|_{V_{t-1}^{-1}}\left[R\sqrt{d\ln\left(\frac{1+wL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S\right].,\qquad (40)
\end{aligned}
$$

where inequality (38) uses triangular inequality, inequality (39) follows from Cauchy-Schwarz inequality, and inequality (40) are consequences of Lemmas 1, 2.

D Proof of Theorem 3

In the proof, we choose λ so that $\beta \geq 1$, for example by choosing $\lambda \geq 1/S^2$. By virtue of UCB, the regret in any round $t \in [T]$ is

$$
\begin{aligned}
\langle x_t^* - X_t, \theta_t\rangle &\leq L\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s - \theta_{s+1}\| + \langle X_t, \hat{\theta}_t\rangle + \beta\|X_t\|_{V_{t-1}^{-1}} - \langle X_t, \theta_t\rangle \qquad (41)\\
&\leq 2L\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s - \theta_{s+1}\| + 2\beta\|X_t\|_{V_{t-1}^{-1}}. \qquad (42)
\end{aligned}
$$

Inequality (41) is by an application of our `SW-UCB` algorithm established in equation (10). Inequality (42) is by an application of inequality (40), which bounds the difference $|\langle X_t, \hat{\theta}_t - \theta_t\rangle|$ from above. By the evident fact that $\langle X_t, \hat{\theta}_t - \theta_t\rangle \leq 2$, we have

$$
\langle x_t^* - X_t, \theta_t\rangle \leq 2L\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s - \theta_{s+1}\| + 2\beta\left(\|X_t\|_{V_{t-1}^{-1}} \wedge 1\right). \qquad (43)
$$

Summing equation (43) over $1 \leq t \leq T$, the regret of the `SW-UCB` algorithm is upper bounded as

$$
\begin{aligned}
\mathbf{E}\left[\text{Regret}_T\left(\texttt{SW-UCB algorithm}\right)\right] &= \sum_{t\in[T]}\langle x_t^* - X_t, \theta_t\rangle \\
&\leq 2L\left[\sum_{t=1}^{T}\sum_{s=1\vee(t-w)}^{t-1}\|\theta_s - \theta_{s+1}\|\right] + 2\beta\sum_{t=1}^{T}\left(\|X_t\|_{V_{t-1}^{-1}} \wedge 1\right)\\
&= 2L\left[\sum_{s=1}^{T}\sum_{t=s+1}^{(s+w)\wedge T}\|\theta_s - \theta_{s+1}\|\right] + 2\beta\sum_{t=1}^{T}\left(\|X_t\|_{V_{t-1}^{-1}} \wedge 1\right)\\
&\leq 2LwB_t + 2\beta\sum_{t=1}^{T}\left(\|X_t\|_{V_{t-1}^{-1}} \wedge 1\right). \qquad (44)
\end{aligned}
$$

What's left is to upper bound the quantity $2\beta\sum_{t\in[T]}\left(1\wedge\|X_t\|_{V_{t-1}^{-1}}\right)$. Following the trick introduced by the authors of [1], we apply Cauchy-Schwarz inequality to the term $\sum_{t\in[T]}\left(1\wedge\|X_t\|_{V_{t-1}^{-1}}\right)$.

$$\sum_{t\in[T]}\left(1\wedge\|X_t\|_{V_{t-1}^{-1}}\right)\leq\sqrt{T}\sqrt{\sum_{t\in[T]}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}}. \tag{45}$$

By dividing the whole time horizon into consecutive pieces of length w, we have

$$\sqrt{\sum_{t\in[T]}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}}\leq\sqrt{\sum_{i=0}^{\lceil T/w\rceil-1}\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}}. \tag{46}$$

While a similar quantity has been analyzed by Lemma 11 of [1], we note that due to the fact that V_t's are accumulated according to the sliding window principle, the key eq. (6) in Lemma 11's proof breaks, and thus the analysis of [1] cannot be applied here. To this end, we state a technical lemma based on a novel use of the Sherman-Morrison formula.

Lemma 5. *For any* $i\leq\lceil T/w\rceil-1$,

$$\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}\leq\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge\|X_t\|^2_{\overline{V}_{t-1}^{-1}},$$

where

$$\overline{V}_{t-1}=\sum_{s=i\cdot w+1}^{t-1}X_sX_s^\top+\lambda I. \tag{47}$$

Proof. Proof of Lemma 5. For a fixed $i\leq\lceil T/w\rceil-1$,

$$\begin{aligned}\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge\|X_t\|^2_{V_{t-1}^{-1}}&=\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge X_t^\top V_{t-1}^{-1}X_t\\&=\sum_{t=i\cdot w+1}^{(i+1)w}1\wedge X_t^\top\left(\sum_{s=1\vee(t-w)}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}X_t.\end{aligned} \tag{48}$$

Note that $i\cdot w+1\geq1$ and $i\cdot w+1\geq t-w\ \forall t\leq(i+1)w$, we have

$$i\cdot w+1\geq1\vee(t-w). \tag{49}$$

Consider any d-by-d positive definite matrix A and d-dimensional vector y, then by the Sherman-Morrison formula, the matrix

$$B=A^{-1}-\left(A+yy^\top\right)^{-1}=A^{-1}-A^{-1}+\frac{A^{-1}yy^\top A^{-1}}{1+y^\top A^{-1}y}=\frac{A^{-1}yy^\top A^{-1}}{1+y^\top A^{-1}y} \tag{50}$$

is positive semi-definite. Therefore, for a given t, we can iteratively apply this fact to obtain

$$\begin{aligned}&X_t^\top\left(\sum_{s=i\cdot w+1}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}X_t\\=&X_t^\top\left(\sum_{s=i\cdot w}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}X_t+X_t^\top\left(\left(\sum_{s=i\cdot w+1}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}-\left(\sum_{s=i\cdot w}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}\right)X_t\\=&X_t^\top\left(\sum_{s=i\cdot w}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}X_t+X_t^\top\left(\left(\sum_{s=i\cdot w+1}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}-\left(X_{i\cdot w}X_{i\cdot w}^\top+\sum_{s=i\cdot w+1}^{t-1}X_sX_s^\top+\lambda I\right)^{-1}\right)X_t\end{aligned}$$

$$\geq X_t^\top \left(\sum_{s=i\cdot w}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t$$

$$\vdots$$

$$\geq X_t^\top \left(\sum_{s=1\vee(t-w)}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t. \tag{51}$$

Plugging inequality (51) to (48), we have

$$\begin{aligned}\sum_{t=i\cdot w+1}^{(i+1)w} 1\wedge \|X_t\|^2_{V_{t-1}^{-1}} &\leq \sum_{t=i\cdot w+1}^{(i+1)w} 1\wedge X_t^\top \left(\sum_{s=i\cdot w+1}^{t-1} X_s X_s^\top + \lambda I\right)^{-1} X_t \\ &\leq \sum_{t=i\cdot w+1}^{(i+1)w} 1\wedge \|X_t\|^2_{\overline{V}_{t-1}^{-1}}, \end{aligned}\tag{52}$$

which concludes the proof. ☐

From Lemma 5 and eq. (46), we know that

$$\begin{aligned}2\beta \sum_{t\in[T]}\left(1\wedge \|X_t\|_{V_{t-1}^{-1}}\right) \leq & 2\beta\sqrt{T}\cdot\sqrt{\sum_{i=0}^{\lceil T/w\rceil-1}\sum_{t=i\cdot w+1}^{(i+1)w} 1\wedge\|X_t\|^2_{\overline{V}_{t-1}^{-1}}} \\ \leq & 2\beta\sqrt{T}\cdot\sqrt{\sum_{i=0}^{\lceil T/w\rceil-1} 2d\ln\left(\frac{d\lambda+wL^2}{d\lambda}\right)} \\ \leq & 2\beta T\sqrt{\frac{2d}{w}\ln\left(\frac{d\lambda+wL^2}{d\lambda}\right)}. \end{aligned}\tag{53}$$

Here, eq. (53) follows from Lemma 11 of [1].

Now putting these two parts to eq. (44), we have

$$\begin{aligned}\mathbf{E}\left[\text{Regret}_T\left(\texttt{SW-UCB}\text{ algorithm}\right)\right] \leq & 2LwB_T + 2\beta T\sqrt{\frac{2d}{w}\ln\left(\frac{d\lambda+wL^2}{d\lambda}\right)} + 2T\delta \\ = & 2LwB_T + \frac{2T}{\sqrt{w}}\left(R\sqrt{d\ln\left(\frac{1+wL^2/\lambda}{\delta}\right)}+\sqrt{\lambda}S\right)\sqrt{2d\ln\left(\frac{d\lambda+wL^2}{d\lambda}\right)} + 2T\delta. \end{aligned}\tag{54}$$

Now if B_T is known, we can take $w = O\left((dT)^{2/3}B_t^{-2/3}\right)$ and $\delta = 1/T$, we have

$$\mathbf{E}\left[\text{Regret}_T\left(\texttt{SW-UCB}\text{ algorithm}\right)\right] = \widetilde{O}\left(d^{\frac{2}{3}}B_T^{\frac{1}{3}}T^{\frac{2}{3}}\right);$$

while if B_T is not unknown taking $w = O\left((dT)^{2/3}\right)$ and $\delta = 1/T$, we have

$$\mathbf{E}\left[\text{Regret}_T\left(\texttt{SW-UCB}\text{ algorithm}\right)\right] = \widetilde{O}\left(d^{\frac{2}{3}}\left(B_T+1\right)T^{\frac{2}{3}}\right).$$

E Proof of Lemma 3

For any block i, the absolute sum of rewards can be written as

$$\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\langle X_t,\theta_t\rangle + \eta_t\right| \leq \sum_{t=(i-1)H+1}^{i\cdot H\wedge T}|\langle X_t,\theta_t\rangle| + \left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right| \leq H + \left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right|,$$

where we have iteratively applied the triangular inequality as well as the fact that $|\langle X_t, \theta_t \rangle| \leq 1$ for all t.

Now by property of the R-sub-Gaussian [22], we have the absolute value of the noise term η_t exceeds $2R\sqrt{\ln T}$ for a fixed t with probability at most $1/T^2$ *i.e.*,

$$\Pr\left(\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right| \geq 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right) \leq \frac{2H}{T^2}. \tag{55}$$

Applying a simple union bound, we have

$$\Pr\left(\exists i \in \left\lceil\frac{T}{H}\right\rceil : \left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right| \geq 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right) \leq \sum_{i=1}^{\lceil T/H\rceil}\Pr\left(\left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right| \geq 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right) \leq \frac{2}{T}. \tag{56}$$

Therefore, we have

$$\Pr\left(Q \geq H + 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right) \leq \Pr\left(\exists i \in \left\lceil\frac{T}{H}\right\rceil : \left|\sum_{t=(i-1)H+1}^{i\cdot H\wedge T}\eta_t\right| \geq 2R\sqrt{H\ln\frac{T}{\sqrt{H}}}\right) \leq \frac{2}{T}. \tag{57}$$

The statement then follows.