A Technical Proofs and

We gather in this section the proofs omitted in the core text.

A.1 Recovering stationary linear bandits, rotting and rising rested bandits, and contextual bandits

Example 1 (Stationary linear bandits) *Consider a linear bandit model, defined by an action set $\mathcal{A} \subset \mathcal{B}_d$ and $\theta^* \in \mathcal{B}_d$. This is equivalent to a LBM with the same $\mathcal{A}$ and θ^*, and memory matrix A such that $A(a_1, \dots, a_m) = I_d$ for any $a_1, \dots, a_m \in \mathcal{A}^m$, i.e., when $m = 0$ or $\gamma = 0$.*

Example 2 (Rotting and rising rested bandits) *In rotting [Levine et al., 2017, Seznec et al., 2019] or rising [Metelli et al., 2022] rested bandits, the expected reward of an arm k at time step t is fully determined by the number $n_k(t)$ of times arm k has been played before time t. Formally, each arm is equipped with a function μ_k such that the expected reward at time t is given by $\mu_k(n_k(t))$. In particular, requiring all the μ_k to be nonincreasing corresponds to the rotting bandits model, and requiring all the μ_k to be nondecreasing corresponds to the rested rising bandits model. Now, let $d = K$, $\mathcal{A} = (e_k)_{1 \le k \le K}$, $\theta^* = (1/\sqrt{K}, \dots, 1/\sqrt{K})$, and $m \to \infty$[2]. By the definition of A, see (2), and the orthogonality of the actions, it is easy to check that the expected reward of playing action e_k at time step t is given by $(1 + n_k(t))^\gamma / \sqrt{K}$. When $\gamma \le 0$, this is a nonincreasing function of $n_k(t)$, and we recover rotting rested bandits. Conversely, when $\gamma \ge 0$, we recover rising rested bandits. We note however that the class of decreasing (respectively increasing) functions we can consider is restricted to the set of monomials of the form $n \mapsto (1+n)^\gamma / \sqrt{K}$, for $\gamma \le 0$ (respectively $\gamma \ge 0$). Extending it to generic polynomials is clearly possible, although it requires more computations in the model selection phase, see Remark 4 and Section 3.3.*

After presenting Example 2, we explain the motivation for considering a finite memory m. Although rotting and rising bandits require infinite memory, we argue on both practical and theoretical grounds that in our setting a finite value of m is preferable. First, in many applications it is reasonable to assume that the effect of past actions will vanish at some point. For example, listening to a song now does not affect how much we will enjoy the same song in a distant enough future. Second, permanent effects may trivialize the problem on the theoretical side: consider $m \to \infty$ and $\gamma \le -1/2$, then for any sequence of actions $(a_t)_{t \ge 1}$ we have

$$\sum_{t=1}^T \langle a_t, A_{t-1}\theta^* \rangle \le \sum_{t=1}^T \left\| A_{t-1} a_t \right\|_2 \le \sqrt{T \sum_{t=1}^T \left\| A_{t-1} a_t \right\|_2^2} \le \sqrt{2dT \log(1 + T/d)} \coloneqq B_T \,,$$

where we have used the elliptical potential lemma [Lattimore and Szepesvári, 2020, Lemma 19.4]. Hence, as soon as $\gamma \le -1/2$, we have $\mathrm{OPT} \le B_T$, and the trivial strategy consistently playing 0 enjoys a small regret B_T. Conversely, consider $\gamma \ge 0$. The strategy consistently playing θ^* achieves, after t rounds, an instantaneous *reward of $(1+t)^\gamma$, which is diverging for $\gamma \ge 1$. This is not realistic in most application and, incidentally, violates the concave payoffs assumption [Metelli et al., 2022, Assumption 3.2]. Therefore, although considering $m = +\infty$ may look attractive at first sight, it actually fails to adequately model song satiation, and restricts the range of relevant γ from $\mathbb{R}$ to $(-1/2, 1)$. Instead, focusing on finite memory m yields more interesting problems, although it prevents a full generalization of rotting bandits with finitely many arms. We note however that when $m < \infty$, the spirit of rotting (resp., rising) bandits is still preserved, as playing an action does decrease (resp., increase) its efficiency for the next pulls (within the time window).*

We conclude this exposition by highlighting that LBMs may also be generalized to contextual bandits [Lattimore and Szepesvári, 2020].

Remark 3 (Contextual bandits) *In contextual bandits, at each time step t the learner is provided a context c_t (e.g., data about a user). The learner then picks an action $a_t \in \mathcal{A}$ (based on c_t), and receives a reward whose expectation depends linearly on the vector $\psi(c_t, a_t) \in \mathbb{R}^d$, where ψ is a known feature map. Note that it is equivalent to have the learner playing actions $a_t \in \mathbb{R}^d$ that belong to a subset $\mathcal{A}_t = \{\psi(c_t, a) \in \mathbb{R}^d \colon a \in \mathcal{A}\}$. The analysis developed in Section 3 still holds true when $\mathcal{A}_t$ depends on t, and can thus be generalized to contextual bandits with memory.*

[2] In the next paragraph, however, we explain why a bounded memory m is preferable within our model.

A.2 Proof of Proposition 1

Proposition 1 *The oracle greedy strategy, which plays* $a_t^{\text{greedy}} = \arg\max_{a \in \mathcal{A}} \langle a, A_{t-1}\theta^* \rangle$ *at time step* t*, can suffer linear regret, both in rotting or rising scenarios.*

Proof We build two instances of LBM, one rotting, one rising, in which the oracle greedy strategy suffers linear regret. We highlight that the other strategy exhibited, which performs better than oracle greedy, may not be optimal.

Rotting instance. Let $\mathcal{A} = \mathcal{B}_d$, $\theta^* = e_1$, $m = d - 1$, and A such that

$$A(a_1, \dots, a_m) = \left(I_d + \sum_{s=1}^{m} a_s a_s^\top \right)^{-\gamma} ,$$

for some $\gamma > 0$ to be specified later. Oracle greedy, which plays at each time step $a_t^{\text{greedy}} = \arg\max_{a \in \mathcal{A}} \langle a, A_{t-1}\theta^* \rangle$, constantly plays e_1. After the first m pulls, it collects a reward of $1/d^\gamma$ at every time step. On the other side, the strategy that plays cyclically the block $e_1 \dots e_d$ collects a reward of 1 every $d = m + 1$ time steps, i.e., an average reward of $1/d$ per step. Hence, up to the transitive first m pulls, the cumulative reward of oracle greedy after T rounds is T/d^γ, and that of the cyclic policy is T/d. The regret of oracle greedy is thus at least

$$T \left(\frac{1}{d} - \frac{1}{d^\gamma} \right) ,$$

which is linear for $\gamma > 1$.

Rising instance. Let $m \geq 1$, $d = 2$, $\mathcal{A} = \mathcal{B}_2$, $\theta^* = (\varepsilon, 1)$ where $\varepsilon > 0$ is to be specified later, and A such that

$$A(a_1, \dots, a_m) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \sum_{s=1}^{m} a_s a_s^\top .$$

Oracle greedy constantly plays e_1 collecting a reward of $(m+1)\theta_1^*$ from round $m+1$ onward. On the other side, the strategy that plays constantly e_2 collects a reward of $m\theta_2^*$ from round $m + 1$ onward. Hence, the regret of oracle greedy from round $m + 1$ onward is at least $(T - m)[m - (m + 1)\varepsilon]$, which is linear for $\varepsilon < m/(m + 1)$. □

A.3 Proof of Proposition 2

Proposition 2 *For any* $m, L \geq 1$*, let* $\widetilde{\boldsymbol{a}}$ *be the block of* $m + L$ *actions defined in* (5) *and* $(\widetilde{r}_t)_{t=1}^T$ *be the expected rewards collected when playing cyclically* $\widetilde{\boldsymbol{a}}$*. We have*

$$\text{OPT} - \sum_{t=1}^{T} \widetilde{r}_t \leq \frac{2mR}{m + L} T . \tag{6}$$

Proof Recall that the optimal sequence is denoted $(a_t^*)_{t=1}^T$ and collects rewards $(r_t^*)_{t=1}^T$. Let $L > 0$; by definition, there exists a block of actions of length L in $(a_t^*)_{t=1}^T$ with average expected reward higher that OPT/T. Let t^* be the first index of this block, we thus have $(1/L) \sum_{t=t^*}^{t^*+L-1} r_t^* \geq \text{OPT}/T$. However, this average expected reward is realized only using the initial matrix A_{t^*-1}, generated from $a_{t^*-1}^*, \dots, a_{t^*-m}^*$. Let $\boldsymbol{a}^* = a_{t^*-m}^*, \dots, a_{t^*+L-1}^*$ of length $m + L$. Note that, by definition, we have that $\widetilde{r}(\widetilde{\boldsymbol{a}}) \geq \widetilde{r}(\boldsymbol{a}^*) = \sum_{t=t^*}^{t^*+L-1} r_t^* \geq L \, \text{OPT}/T$. Furthermore, by (8), when playing cyclically $\widetilde{\boldsymbol{a}}$ one obtains at least a reward of $-R$ in each one of the first m pulls of the block. Collecting all the pieces, we obtain

$$\begin{aligned} \sum_{t=1}^{T} \widetilde{r}_t &\geq \frac{T}{m + L} \Big(- mR + \widetilde{r}(\widetilde{\boldsymbol{a}}) \Big) \\ &\geq \frac{T}{m + L} \Big(- mR + \widetilde{r}(\boldsymbol{a}^*) \Big) \\ &\geq \frac{T}{m + L} \left(-mR + L \, \frac{\text{OPT}}{T} \right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{L}{m+L}\mathrm{OPT} - \frac{mR}{m+L}\,T \\
&\geq \frac{L}{m+L}\mathrm{OPT} + \frac{m}{m+L}\mathrm{OPT} - \frac{mR}{m+L}\,T - \frac{mR}{m+L}\,T \qquad (13)\\
&= \mathrm{OPT} - \frac{2mR}{m+L}\,T\,,
\end{aligned}
$$

where (13) derives from $\mathrm{OPT} \leq RT$. □

A.4 Proof of Proposition 4

We prove the (stronger) high probability version of Proposition 4.

Proposition 5 *Let $\lambda \geq 1$, $\delta \in (0,1)$, and $\boldsymbol{a}_\tau$ be the blocks of actions in $\mathbb{R}^{d(m+L)}$ associated to the $\boldsymbol{b}_\tau$ defined in (9). Then, with probability at least $1-\delta$ we have*

$$
\sum_{\tau=1}^{T/(m+L)} \widetilde{r}(\widetilde{\boldsymbol{a}}) - \widetilde{r}(\boldsymbol{a}_\tau) \leq 4L(m+1)^{\gamma^+} \sqrt{Td\,\ln\left(1+\frac{T(m+1)^{2\gamma^+}}{d(m+L)\lambda}\right)} \\
\cdot \left(\sqrt{\lambda L} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d(m+L)\,\ln\left(1+\frac{T(m+1)^{2\gamma^+}}{d(m+L)\lambda}\right)}\right).
$$

Proof The proof essentially follows that of [Abbasi-Yadkori et al., 2011, Theorem 3]. The main difference is that our version of OFUL operates at the block level. This implies a smaller time horizon, but also and increased dimension and an instantaneous regret $\langle \widetilde{\boldsymbol{b}}, \boldsymbol{\theta}^* \rangle - \langle \boldsymbol{b}_\tau, \boldsymbol{\theta}^* \rangle$ upper bounded by $2L(m+1)^{\gamma^+}$ instead of 1. We detail the main steps of the proof for completeness. Recall that running OFUL in our case amounts to compute at every block time step τ

$$
\widehat{\boldsymbol{\theta}}_\tau = \boldsymbol{V}_\tau^{-1}\left(\sum_{\tau'=1}^{\tau} \boldsymbol{y}_{\tau'}\,\boldsymbol{b}_{\tau'}\right),
$$

where

$$
\boldsymbol{V}_\tau = \sum_{\tau'=1}^{\tau} \boldsymbol{b}_{\tau'}\boldsymbol{b}_{\tau'}^\top + \lambda I_{d(m+L)}\,, \qquad \text{and} \qquad \boldsymbol{y}_\tau = \sum_{i=m+1}^{m+L} y_{\tau,i}\,,
$$

since we associate with a block of actions the sum of rewards obtained after time step m. Note that by the determinant-trace inequality, see e.g., [Abbasi-Yadkori et al., 2011, Lemma 10], with actions $\boldsymbol{b}_\tau$ that satisfy $\|\boldsymbol{b}_\tau\|_2^2 \leq m + L(m+1)^{2\gamma^+}$ we have

$$
\frac{|\boldsymbol{V}_\tau|}{|\lambda I_{d(m+L)}|} \leq \left(1+\frac{\tau(m+L(m+1)^{2\gamma^+})}{d(m+L)\lambda}\right)^{d(m+L)} \leq \left(1+\frac{\tau(m+1)^{2\gamma^+}}{d\lambda}\right)^{d(m+L)}. \qquad (14)
$$

The action played at block time step τ is the block $\boldsymbol{a}_\tau \in \mathcal{B}_d^{m+L}$ associated with

$$
\boldsymbol{b}_\tau = \operatorname*{arg\,max}_{\boldsymbol{b}\in\boldsymbol{\mathcal{B}}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\mathcal{C}}_{\tau-1}} \langle \boldsymbol{b}, \boldsymbol{\theta} \rangle\,, \qquad (15)
$$

where

$$
\boldsymbol{\mathcal{C}}_\tau = \left\{\boldsymbol{\theta} \in \mathbb{R}^{d(m+L)} \colon \left\|\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}\right\|_{\boldsymbol{V}_\tau} \leq \boldsymbol{\beta}_\tau(\delta)\right\},
$$

with

$$
\boldsymbol{\beta}_\tau(\delta) = \sqrt{2\ln\left(\frac{1}{\delta}\right) + d(m+L)\,\ln\left(1+\frac{\tau(m+1)^{2\gamma^+}}{d\lambda}\right)} + \sqrt{\lambda L}\,. \qquad (16)
$$

Applying [Abbasi-Yadkori et al., 2011, Theorem 2] to $\boldsymbol{\theta}^* \in \mathbb{R}^{d(m+L)}$ which satisfies $\|\boldsymbol{\theta}^*\|_2 \leq \sqrt{L}$ we have that $\boldsymbol{\theta}^* \in \boldsymbol{\mathcal{C}}_\tau$ for every τ with probability at least $1-\delta$. Denoting by $\widetilde{\boldsymbol{\theta}}_\tau$ the model that

maximizes (15), we thus have that with probability at least $1-\delta$, the inequality $\langle \widetilde{\boldsymbol{b}}, \boldsymbol{\theta}^* \rangle \leq \langle \boldsymbol{b}_\tau, \widetilde{\boldsymbol{\theta}}_\tau \rangle$ holds for every τ, and consequently

$$
\begin{aligned}
&\sum_{\tau=1}^{T/(m+L)} \langle \widetilde{\boldsymbol{b}}, \boldsymbol{\theta}^* \rangle - \langle \boldsymbol{b}_\tau, \boldsymbol{\theta}^* \rangle \\
&\qquad \leq \sum_{\tau=1}^{T/(m+L)} \min\left\{ 2L(m+1)^{\gamma^+} \,,\, \langle \boldsymbol{b}_\tau, \widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^* \rangle \right\} \\
&\qquad \leq \sum_{\tau=1}^{T/(m+L)} \min\left\{ 2L(m+1)^{\gamma^+} \,,\, \left\| \widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^* \right\|_{\boldsymbol{V}_{\tau-1}} \|\boldsymbol{b}_\tau\|_{\boldsymbol{V}_{\tau-1}^{-1}} \right\} \\
&\qquad \leq \sum_{\tau=1}^{T/(m+L)} \min\left\{ 2L(m+1)^{\gamma^+} \,,\, 2\boldsymbol{\beta}_\tau(\delta) \|\boldsymbol{b}_\tau\|_{\boldsymbol{V}_{\tau-1}^{-1}} \right\} \\
&\qquad \leq 2L(m+1)^{\gamma^+} \boldsymbol{\beta}_{T/(m+L)}(\delta) \sum_{\tau=1}^{T/(m+L)} \min\left\{ 1 \,,\, \|\boldsymbol{b}_\tau\|_{\boldsymbol{V}_{\tau-1}^{-1}} \right\} \\
&\qquad \leq 2L(m+1)^{\gamma^+} \boldsymbol{\beta}_{T/(m+L)}(\delta) \sqrt{\frac{T}{m+L} \sum_{\tau=1}^{T/(m+L)} \min\left\{ 1 \,,\, \|\boldsymbol{b}_\tau\|^2_{\boldsymbol{V}_{\tau-1}^{-1}} \right\}} \\
&\qquad \leq 2\sqrt{2}L(m+1)^{\gamma^+} \boldsymbol{\beta}_{T/(m+L)}(\delta) \sqrt{\frac{T}{m+L} \ln \frac{|\boldsymbol{V}_{T/(m+L)}|}{|\lambda I_{d(m+L)}|}} \\
&\qquad \leq 4L(m+1)^{\gamma^+} \sqrt{Td \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d(m+L)\lambda}\right)} \\
&\qquad\qquad \cdot \left(\sqrt{\lambda L} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d(m+L) \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d(m+L)\lambda}\right)} \right) ,
\end{aligned}
$$

where we have used [Abbasi-Yadkori et al., 2011, Lemma 11], as well as (14) and (16). Note that in the stationary case, i.e., when $m = 0$ and $L = 1$, we exactly recover [Abbasi-Yadkori et al., 2011, Theorem 3]. Proposition 4 is obtained by setting $\lambda \in [1, d]$, $L \geq m$, and $\delta = 1/T$. □

A.5 Proof of Proposition 3

Proof Let $d = m + 1$, $\mathcal{A} = \{0_d\} \cup (e_k)_{k \leq d}$, $\theta^* = (1/\sqrt{d}, \ldots, 1/\sqrt{d})$, and $\gamma \leq 0$. For simplicity, we note the basis modulo d, i.e., $e_{k+d} = e_k$ for any $k \in \mathbb{N}$. Note that for any $a_1, \ldots, a_{m+1} \in \mathcal{A}$ we have $\left|\langle a_{m+1}, A_m \theta^* \rangle\right| \leq \|a_{m+1}\|_1 \, \|A_m \theta^*\|_\infty \leq 1/\sqrt{d}$, such that one can take $R = 1/\sqrt{d}$. Observe now that the strategy which plays cyclically $e_1, \ldots, e_d$ collects a reward of $1/\sqrt{d}$ at each time step, which is optimal, such that $\mathrm{OPT} = T/\sqrt{d}$. Further, it is easy to check that block $\widetilde{\boldsymbol{a}}$, composed of m pulls of 0_d followed by $e_1, \ldots, e_L$ satisfies $\widetilde{r}(\widetilde{\boldsymbol{a}}) = L/\sqrt{d}$, which is optimal for similar reasons. Playing cyclically $\widetilde{\boldsymbol{a}}$, one gets a reward of $L/\sqrt{d}$ every $m + L$ pulls. In other terms, we have

$$
\mathrm{OPT} - \sum_{t=1}^{T} \widetilde{r}_t = \frac{T}{\sqrt{d}} - \frac{L}{m+L} \frac{T}{\sqrt{d}} = \frac{m}{m+L} \frac{T}{\sqrt{d}} = \frac{mR}{m+L} T \,.
$$

□

A.6 Proof of Theorem 1

We prove the high probability version of Theorem 1, obtained by setting $\lambda \in [1, d]$, and $\delta = 1/T$.

Theorem 2 *Let $\lambda \geq 1$, $\delta \in (0,1)$, and $\boldsymbol{a}_\tau$ be the blocks of actions in $\mathbb{R}^{d(m+L)}$ defined in (11). Then, with probability at least $1-\delta$ we have*

$$\sum_{\tau=1}^{T/(m+L)} \widetilde{r}(\widetilde{\boldsymbol{a}}) - \widetilde{r}(\boldsymbol{a}_\tau) \leq 4L(m+1)^{\gamma^+} \sqrt{Td \, \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d\lambda}\right)} \cdot \left(\sqrt{\lambda} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d \, \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d(m+L)\lambda}\right)}\right).$$

Let $m \geq 1$, $T \geq m^2 d^2 + 1$, and set $L = \lceil \sqrt{m/d}\, T^{1/4} \rceil - m$. Let r_t be the rewards collected when playing $\boldsymbol{a}_\tau$ as defined in (11). Then, with probability at least $1-\delta$ we have

$$\mathrm{OPT} - \sum_{t=1}^{T} r_t \leq 4\sqrt{d}\,(m+1)^{\frac{1}{2}+\gamma^+} \, T^{3/4} \left[1 + 2\sqrt{\ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d\lambda}\right)} \cdot \left(\sqrt{\frac{\lambda}{d}} + \sqrt{\frac{\ln(1/\delta)}{d} + \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d\lambda}\right)}\right)\right].$$

Proof The proof is along the lines of OFUL's analysis. The main difficulty is that we cannot use the elliptical potential lemma, see e.g., [Lattimore and Szepesvári, 2020, Lemma 19.4] due to the delay accumulated by V_τ, which is computed every $m+L$ round only. Let

$$\beta_\tau(\delta) = \sqrt{2\ln\left(\frac{1}{\delta}\right) + d \, \ln\left(1 + \frac{\tau(m+1)^{2\gamma^+}}{d\lambda}\right)} + \sqrt{\lambda}\,. \tag{17}$$

By [Abbasi-Yadkori et al., 2011, Theorem 2], we have with probability at least $1-\delta$ that $\theta^* \in \mathcal{C}_\tau$ for every τ. It follows directly that $\boldsymbol{\theta}^* \in \boldsymbol{\mathcal{D}}_\tau$ for any τ, such that $\langle \widetilde{\boldsymbol{b}}, \boldsymbol{\theta}^* \rangle \leq \langle \boldsymbol{b}_\tau, \widetilde{\boldsymbol{\theta}}_\tau \rangle$, where $\widetilde{\boldsymbol{\theta}}_\tau = (0_d, \dots, 0_d, \widetilde{\theta}_\tau, \dots, \widetilde{\theta}_\tau)$ with $\widetilde{\theta}_\tau \in \mathbb{R}^d$ that maximizes (11) over $\mathcal{C}_{\tau-1}$. It can be shown that the regret is upper bounded by $\sum_\tau \sum_{i=m+1}^{m+L} \langle b_{\tau,i}, \widetilde{\theta}_\tau - \theta^* \rangle$. Following the standard analysis, one could then use

$$\left\langle b_{\tau,i}, \widetilde{\theta}_\tau - \theta^* \right\rangle \leq \|b_{\tau,i}\|_{V_{\tau-1}^{-1}} \left\|\widetilde{\theta}_t - \theta^*\right\|_{V_{\tau-1}}.$$

While the confidence set gives $\left\|\widetilde{\theta}_t - \theta^*\right\|_{V_{\tau-1}} \leq 2\beta_{\tau-1}(\delta)$, the quantity $\sum_{i=m+1}^{m+L} \|b_{\tau,i}\|_{V_{\tau-1}^{-1}}$ is much more complex to bound. Indeed, the elliptical potential lemma allows to bound $\sum_t \|a_t\|^2_{V_{t-1}^{-1}}$ when $V_t = \sum_{s \leq t} a_s a_s^\top + \lambda I_d$. However, recall that in our case we have $V_\tau = \sum_{\tau'=1}^{\tau} \sum_{i=m+1}^{m+L} b_{\tau',i} b_{\tau',i}^\top + \lambda I_d$, which is only computed every $m+L$ rounds. As a consequence, there exists a "delay" between $V_{\tau-1}$ and the action $b_{\tau,i}$ for $i \geq m+2$, preventing from using the lemma. Therefore, we propose to use instead

$$\left\langle b_{\tau,i}, \widetilde{\theta}_\tau - \theta^* \right\rangle \leq \|b_{\tau,i}\|_{V_{\tau,i-1}^{-1}} \left\|\widetilde{\theta}_t - \theta^*\right\|_{V_{\tau,i-1}}, \quad \text{where} \quad V_{\tau,i} = V_{\tau-1} + \sum_{j=m+1}^{i} b_{\tau,j} b_{\tau,j}^\top. \tag{18}$$

By doing so, the elliptical potential lemma applies. On the other hand, one has to control $\left\|\widetilde{\theta}_t - \theta^*\right\|_{V_{\tau,i-1}}$, which is not anymore bounded by $2\beta_{\tau-1}(\delta)$ since the subscript matrix is $V_{\tau,i-1}$ instead of $V_{\tau-1}$. Still, one can show that for any $i \leq m+L$ we have

$$\begin{aligned}
\left\|\widetilde{\theta}_t - \theta^*\right\|^2_{V_{\tau,i-1}} &= \mathrm{Tr}\left(V_{\tau,i-1} \left(\widetilde{\theta}_t - \theta^*\right)\left(\widetilde{\theta}_t - \theta^*\right)^\top\right) \\
&= \mathrm{Tr}\left(\left(V_{\tau-1} + \sum_{j=m+1}^{i-1} b_{\tau,j} b_{\tau,j}^\top\right)\left(\widetilde{\theta}_t - \theta^*\right)\left(\widetilde{\theta}_t - \theta^*\right)^\top\right) \\
&= \mathrm{Tr}\left(\left(I_d + \sum_{j=m+1}^{i-1} \left(V_{\tau-1}^{-1/2} b_{\tau,j}\right)\left(V_{\tau-1}^{-1/2} b_{\tau,j}\right)^\top\right) V_{\tau-1}^{1/2}\left(\widetilde{\theta}_t - \theta^*\right)\left(\widetilde{\theta}_t - \theta^*\right)^\top V_{\tau-1}^{1/2}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| I_d + \sum_{j=m+1}^{i-1} \left(V_{\tau-1}^{-1/2} b_{\tau,j}\right)\left(V_{\tau-1}^{-1/2} b_{\tau,j}\right)^\top \right\|_* \operatorname{Tr}\left(V_{\tau-1}^{1/2}\left(\widetilde{\theta}_t - \theta^*\right)\left(\widetilde{\theta}_t - \theta^*\right)^\top V_{\tau-1}^{1/2}\right) \\
&\leq \left(1 + \sum_{j=m+1}^{i-1} \left\| V_{\tau-1}^{-1/2} b_{\tau,j} \right\|_2^2\right) \left\| \widetilde{\theta}_t - \theta^* \right\|_{V_{\tau-1}}^2 \\
&\leq \left(1 + (L-1)(m+1)^{2\gamma^+}\right) \left\| \widetilde{\theta}_t - \theta^* \right\|_{V_{\tau-1}}^2 \\
&\leq L(m+1)^{2\gamma^+} \left\| \widetilde{\theta}_t - \theta^* \right\|_{V_{\tau-1}}^2 . \qquad (19)
\end{aligned}
$$

Recalling also that $\langle \widetilde{\boldsymbol{b}}, \boldsymbol{\theta}^* \rangle - \langle \boldsymbol{b}_\tau, \boldsymbol{\theta}^* \rangle \leq 2L(m+1)^{\gamma^+}$, we have with probability at least $1-\delta$

$$
\begin{aligned}
&\sum_{\tau=1}^{T/(m+L)} \langle \widetilde{\boldsymbol{b}}, \boldsymbol{\theta}^* \rangle - \langle \boldsymbol{b}_\tau, \boldsymbol{\theta}^* \rangle \\
&\qquad \leq \sum_{\tau=1}^{T/(m+L)} \min\left\{ 2L(m+1)^{\gamma^+} , \langle \boldsymbol{b}_\tau, \widetilde{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^* \rangle \right\} \\
&\qquad = \sum_{\tau=1}^{T/(m+L)} \min\left\{ 2L(m+1)^{\gamma^+} , \sum_{i=m+1}^{m+L} \langle b_{\tau,i}, \widetilde{\theta}_\tau - \theta^* \rangle \right\} \\
&\qquad \leq \sum_{\tau=1}^{T/(m+L)} \min\left\{ 2L(m+1)^{\gamma^+} , \sum_{i=m+1}^{m+L} \|b_{\tau,i}\|_{V_{\tau,i-1}^{-1}} \left\| \widetilde{\theta}_t - \theta^* \right\|_{V_{\tau,i-1}} \right\} \\
&\qquad \leq \sum_{\tau=1}^{T/(m+L)} \min\left\{ 2L(m+1)^{\gamma^+} , 2\sqrt{L}(m+1)^{\gamma^+} \beta_{\tau-1}(\delta) \sum_{i=m+1}^{m+L} \|b_{\tau,i}\|_{V_{\tau,i-1}^{-1}} \right\} \\
&\qquad \leq 2L(m+1)^{\gamma^+} \beta_{T/(m+L)}(\delta) \sum_{\tau=1}^{T/(m+L)} \sum_{i=m+1}^{m+L} \min\left\{ 1 , \|b_{\tau,i}\|_{V_{\tau,i-1}^{-1}} \right\} \\
&\qquad \leq 2L(m+1)^{\gamma^+} \beta_{T/(m+L)}(\delta) \sqrt{\frac{TL}{m+L} \sum_{\tau=1}^{T/(m+L)} \sum_{i=m+1}^{m+L} \min\left\{ 1 , \|b_{\tau,i}\|_{V_{\tau,i-1}^{-1}}^2 \right\}} \\
&\qquad \leq 2\sqrt{2}L(m+1)^{\gamma^+} \beta_{T/(m+L)}(\delta) \sqrt{T \ln \frac{|V_{T/(m+L)}|}{|\lambda I_d|}} \\
&\qquad \leq 4L(m+1)^{\gamma^+} \sqrt{Td \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d\lambda}\right)} \\
&\qquad\qquad \cdot \left(\sqrt{\lambda} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d(m+L)\lambda}\right)} \right) , \qquad (20)
\end{aligned}
$$

where we have used (17), (18), and (19). Similarly to Proposition 5, note that in the stationary case, i.e., when $m = 0$ and $L = 1$, we exactly recover [Abbasi-Yadkori et al., 2011, Theorem 3]. The first claim of Theorem 1 is obtained by setting $\lambda \in [1, d]$, and $\delta = 1/T$.

Let R_T denote the right-hand side of (20). Combining this bound with the arguments of Proposition 2, we have with probability $1-\delta$

$$
\begin{aligned}
\sum_{t=1}^{T} r_t &\geq \sum_{\tau=1}^{T/(m+L)} \widetilde{r}(\boldsymbol{a}_\tau) - \frac{m(m+1)^{\gamma^+}}{m+L} T \qquad (21) \\
&= \sum_{\tau=1}^{T/(m+L)} \langle \boldsymbol{b}_\tau, \boldsymbol{\theta}^* \rangle - \frac{m(m+1)^{\gamma^+}}{m+L} T
\end{aligned}
$$

$$\geq \sum_{\tau=1}^{T/(m+L)} \langle \widetilde{\boldsymbol{b}}, \boldsymbol{\theta}^* \rangle - R_T - \frac{m(m+1)^{\gamma^+}}{m+L} T \tag{22}$$

$$= \sum_{\tau=1}^{T/(m+L)} \widetilde{r}(\widetilde{\boldsymbol{a}}) - R_T - \frac{m(m+1)^{\gamma^+}}{m+L} T$$

$$\geq \sum_{t=1}^{T} \widetilde{r}_t - R_T - \frac{2m(m+1)^{\gamma^+}}{m+L} T \tag{23}$$

$$\geq \mathrm{OPT} - R_T - \frac{4m(m+1)^{\gamma^+}}{m+L} T \tag{24}$$

$$\geq \mathrm{OPT} - 4(m+1)^{\gamma^+} \left[\frac{mT}{m+L} + (m+L)\sqrt{Td \, \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d\lambda}\right)} \right.$$
$$\left. \cdot \left(\sqrt{\lambda} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d(m+L)\lambda}\right)} \right) \right],$$

where (21) and (23) come from the fact that any instantaneous reward is bounded by $(m+1)^{\gamma^+}$, see (8), (22) from (20), and (24) from Proposition 2.

Now, assume that $m \geq 1$, $T \geq d^2 m^2 + 1$, and let $L = \lceil \sqrt{m/d}\, T^{1/4} \rceil - m$. By the condition on T, we have $\sqrt{m/d}\, T^{1/4} > m \geq 1$, such that $L \geq 1$ and

$$\sqrt{\frac{m}{d}}\, T^{1/4} \leq \left\lceil \sqrt{\frac{m}{d}}\, T^{1/4} \right\rceil = L + m \leq \sqrt{\frac{m}{d}}\, T^{1/4} + 1 \leq 2\sqrt{\frac{m}{d}}\, T^{1/4}.$$

Substituting in the above bound, we have with probability $1 - \delta$

$$\mathrm{OPT} - \sum_{t=1}^{T} r_t \leq 4\sqrt{d}\,(m+1)^{\frac{1}{2}+\gamma^+}\, T^{3/4} \left[1 + 2\sqrt{\ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d\lambda}\right)} \right.$$
$$\left. \cdot \left(\sqrt{\frac{\lambda}{d}} + \sqrt{\frac{\ln(1/\delta)}{d} + \ln\left(1 + \frac{T(m+1)^{2\gamma^+}}{d\lambda}\right)} \right) \right].$$

The second claim of Theorem 1 is obtained by setting $\lambda \in [1, d]$, and $\delta = 1/T$. □

Remark 4 (Generic matrix mapping A) *Note that our analysis naturally extends to any matrix mapping A, as long as it is known. The term $(m+1)^{\gamma^+}$ in Theorem 1 is then replaced with $\sup_{a_1 \dots a_m} \|A(a_1, \dots, a_m)\|_*$. We highlight however that having access to such knowledge is unlikely in practice. This is why we focus on the simpler parametric family (2), which encompasses many rotting and rising scenarios while allowing us to learn simultaneously m and γ, as shown in the next section. It is of course possible to extend the family of monomials (2) to a family of polynomials, but this requires tracking more parameters (namely, the different coefficients of the polynomial), thus degrading the final regret bound.*

Remark 5 (Solving LBM with a general Reinforcement Learning (RL) approach) *Our setting may be seen as an MDP with a d-dimensional continuous space of actions, a (md)-dimensional continuous state space (for the past m actions), a deterministic transition function parameterized by an unknown scalar γ, and a stochastic reward function with a linear dependence on an additional d-dimensional latent parameter θ^*. The optimal policy in this MDP is generally nonstationary, and we are not aware of RL algorithms whose regret can be bounded without relying on more specific assumptions on the MDP. By exploiting the structure of the MDP, and restricting to cyclic policies, we show instead that the original problem can be solved using stationary bandit techniques.*

A.7 Computational complexity of LBM

As described in Algorithm 1, our approach consists of two steps: updating the confidence region C_τ, i.e., $\widehat{\theta}_\tau$ and β_τ according to (10) and (17), and computing the block $\boldsymbol{a}_\tau$ that maximizes the UCB index.

The first step is performed by online Ridge regression, and has a computational cost of $\mathcal{O}(Ld^2)$. We note here the advantage of our refined algorithm over the naive concatenated approach, whose Ridge regression update has cost $\mathcal{O}(L^2d^2)$. The maximization of the UCB indices, performed through gradient ascent has time complexity per iteration of $\mathcal{O}\big((m+L)d^2\big)$. Hence, the overall complexity of an epoch of Algorithm 1 is $\mathcal{O}\big((m+L)d^2 \cdot n_{\text{it}}\big)$, where n_{it} is the number of iterations performed by gradient ascent. Recall that the epochs of Algorithm 1 correspond to blocks of $m+L$ actions, such that the actual per-round complexity is $\mathcal{O}(d^2 \cdot n_{\text{it}})$.

A.8 Proof of Corollary 1

Lemma 1 *Suppose that a block-based bandit algorithm (in our case the bandit combiner) produces a sequence of T_{bc} blocks $\boldsymbol{a}_\tau$, with possibly different cardinalities $|\boldsymbol{a}_\tau|$, such that*

$$\sum_{\tau=1}^{T_{\text{bc}}} \frac{\widetilde{r}(\widetilde{\boldsymbol{a}})}{|\widetilde{\boldsymbol{a}}|} - \sum_{\tau=1}^{T_{\text{bc}}} \frac{\widetilde{r}(\boldsymbol{a}_\tau)}{|\boldsymbol{a}_\tau|} \leq F(T_{\text{bc}})\,,$$

for some sublinear function F. Then, we have

$$\frac{\min_\tau |\boldsymbol{a}_\tau|}{\max_\tau |\boldsymbol{a}_\tau|} \left(\widetilde{r}(\widetilde{\boldsymbol{a}})\, \frac{\sum_\tau |\boldsymbol{a}_\tau|}{|\widetilde{\boldsymbol{a}}|}\right) - \sum_{\tau=1}^{T_{\text{bc}}} \widetilde{r}(\boldsymbol{a}_\tau) \;\leq\; \min_\tau |\boldsymbol{a}_\tau|\, F(T_{\text{bc}})\,.$$

In particular, if all blocks have the same cardinality the last bound is just the block regret bound scaled by $|\boldsymbol{a}_\tau|$.

Proof We have

$$\begin{aligned}
\sum_{\tau=1}^{T_{\text{bc}}} \widetilde{r}(\boldsymbol{a}_\tau) &\geq \min_\tau |\boldsymbol{a}_\tau| \sum_{\tau=1}^{T_{\text{bc}}} \frac{\widetilde{r}(\boldsymbol{a}_\tau)}{|\boldsymbol{a}_\tau|} \\
&\geq \min_\tau |\boldsymbol{a}_\tau| \left(\sum_{\tau=1}^{T_{\text{bc}}} \frac{\widetilde{r}(\widetilde{\boldsymbol{a}})}{|\widetilde{\boldsymbol{a}}|} - F(T_{\text{bc}})\right) \\
&= \frac{\min_\tau |\boldsymbol{a}_\tau|}{\max_\tau |\boldsymbol{a}_\tau|}\, \frac{\widetilde{r}(\widetilde{\boldsymbol{a}})}{|\widetilde{\boldsymbol{a}}|}\, \max_\tau |\boldsymbol{a}_\tau|\, T_{\text{bc}} - \min_\tau |\boldsymbol{a}_\tau|\, F(T_{\text{bc}}) \\
&\geq \frac{\min_\tau |\boldsymbol{a}_\tau|}{\max_\tau |\boldsymbol{a}_\tau|} \left(\widetilde{r}(\widetilde{\boldsymbol{a}})\, \frac{\sum_\tau |\boldsymbol{a}_\tau|}{|\widetilde{\boldsymbol{a}}|}\right) - \min_\tau |\boldsymbol{a}_\tau|\, F(T_{\text{bc}})\,.
\end{aligned}$$

□

Corollary 1 *Consider an instance of LBM with unknown parameters $(m_\star, \gamma_\star)$. Assume a bandit combiner is run on $N \leq d\sqrt{m_\star}$ instances of OFUL-memory (Algorithm 2), each using a different pair of parameters (m_i, γ_i) from a set $\mathcal{S} = \big\{(m_1,\gamma_1),\dots,(m_N,\gamma_N)\big\}$ such that $(m_\star,\gamma_\star) \in \mathcal{S}$. Let $M = (\max_j m_j)/(\min_j m_j)$. Then, for all $T \geq (m_\star+1)^{2\gamma_\star^+}/m_\star d^4$, the expected rewards $\big(r_t^{\text{bc}}\big)_{t=1}^T$ of the bandit combiner satisfy*

$$\frac{\text{OPT}}{\sqrt{M}} - \mathbb{E}\left[\sum_{t=1}^{T} r_t^{\text{bc}}\right] \;=\; \widetilde{\mathcal{O}}\Big(M\, d\, (m_\star+1)^{1+\frac{3}{2}\gamma_\star^+}\, T^{3/4}\Big)\,.$$

Proof Let $m_\star$ be the true memory size, and $L_\star = L(m_\star)$ the corresponding (partial) block length. Throughout the proof, $\widetilde{\boldsymbol{a}}$ denotes the block defined in (5) with length $m_\star + L_\star$. First observe that only one of the OFUL-memory instances we test is well-specified, i.e., has the true parameters $(m_\star, \gamma_\star)$. We can thus rewrite the regret bound for the Bandit Combiner [Cutkosky et al., 2020, Corollary 2], generalized to rewards bounded in $[-R, R]$ as follows

$$\text{Regret}_{\text{bc}} = \widetilde{\mathcal{O}}\left(C_\star T_{\text{bc}}^{\alpha_\star} + C_\star^{\frac{1}{\alpha_\star}} T_{\text{bc}} \eta_\star^{\frac{1-\alpha_\star}{\alpha_\star}} + R^2 T_{\text{bc}} \eta_\star + \sum_{j \neq \star} \frac{1}{\eta_j}\right)\,, \tag{25}$$

where $T_{\text{bc}} = T/(m_\star + L_\star)$ is the bandit combiner horizon, $C_\star$ and $\alpha_\star$ are the constants in the regret bound of the well-specified instance (see below how we determine them), and the η_j are free parameters to be tuned. We now derive $C_\star$ and $\alpha_\star$. To that end, we must establish the regret bound of the well-specified instance, and identify $C_\star$ and $\alpha_\star$ such that this bound is equal to $C_\star T_{\text{bc}}^{\alpha_\star}$, where $C_\star$ may contain logarithmic factors. For the well-specified instance, the first claim of Theorem 2 gives that, with probability at least $1-\delta$, we have

$$\sum_{\tau=1}^{T/(m_\star+L_\star)} \widetilde{r}(\widetilde{\boldsymbol{a}}) - \widetilde{r}(\boldsymbol{a}_\tau) \leq 4(m_\star + L_\star)(m_\star+1)^{\gamma_\star^+} \sqrt{Td\,\ln\left(1 + \frac{T(m_\star+1)^{2\gamma_\star^+}}{d\lambda}\right)} \left(\sqrt{\lambda} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d\,\ln\left(1 + \frac{T(m_\star+1)^{2\gamma_\star^+}}{d(m_\star+L_\star)\lambda}\right)}\right)$$

$$\sum_{\tau=1}^{T/(m_\star+L_\star)} \frac{\widetilde{r}(\widetilde{\boldsymbol{a}})}{|\widetilde{\boldsymbol{a}}|} - \frac{\widetilde{r}(\boldsymbol{a}_\tau)}{|\boldsymbol{a}_\tau|} \leq T^{1/2}\, 4(m_\star+1)^{\gamma_\star^+} \sqrt{d\ln\left(1 + \frac{T(m_\star+1)^{2\gamma_\star^+}}{d\lambda}\right)} \left(\sqrt{\lambda} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d\,\ln\left(1 + \frac{T(m_\star+1)^{2\gamma_\star^+}}{d(m_\star+L_\star)\lambda}\right)}\right), \tag{26}$$

where we have used that $|\boldsymbol{a}_\tau| = |\widetilde{\boldsymbol{a}}| = m_\star + L_\star$ for every τ. Note that the right-hand side of (26) is expressed in terms of T, which is not the correct horizon, $T/(m_\star + L_\star)$. However, recall that we have

$$\begin{aligned}
m_\star + L_\star &\leq 2\sqrt{\frac{m_\star}{d}}\, T^{1/4} \\
(m_\star + L_\star)^4 &\leq \left(\frac{4m_\star}{d}\right)^2 T \\
T^3 &\leq \left(\frac{4m_\star}{d}\right)^2 \left(\frac{T}{m_\star + L_\star}\right)^4 \\
T^{1/2} &\leq \left(\frac{4m_\star}{d}\right)^{1/3} \left(\frac{T}{m_\star + L_\star}\right)^{2/3},
\end{aligned}$$

such that by substituting in (26) and identifying we have $\alpha_\star = 2/3$, and

$$C_\star = 4\left(\frac{4m_\star}{d}\right)^{1/3} (m_\star+1)^{\gamma_\star^+} \sqrt{d\ln\left(1 + \frac{T_{\text{bc}}(m_\star+L_\star)(m_\star+1)^{2\gamma_\star^+}}{d\lambda}\right)} \left(\sqrt{\lambda} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d\,\ln\left(1 + \frac{T_{\text{bc}}(m_\star+1)^{2\gamma_\star^+}}{d\lambda}\right)}\right).$$

Setting $\eta_j = T_{\text{bc}}^{-2/3}$, and substituting in (25) with $R = (m_\star+1)^{\gamma_\star^+}$, we have that with high probability

$$\sum_{\tau=1}^{T_{\text{bc}}} \frac{\widetilde{r}(\widetilde{\boldsymbol{a}})}{|\widetilde{\boldsymbol{a}}|} - \frac{\widetilde{r}(\boldsymbol{a}_\tau^{\text{bc}})}{|\boldsymbol{a}_\tau^{\text{bc}}|} = \widetilde{\mathcal{O}}\Big(\big(C_\star^{3/2} + N\big)\, T_{\text{bc}}^{2/3} + \ (m_\star+1)^{2\gamma_\star^+}\, T_{\text{bc}}^{1/3}\Big).$$

Now, recall that $T_{\text{bc}} = \mathcal{O}\big(\sqrt{d/m_\star}\, T^{3/4}\big)$, and that $C_\star = \widetilde{\mathcal{O}}\big((m_\star+1)^{\frac{1}{3}+\gamma_\star^+}\, d^{2/3}\big)$. Hence, $N \leq d\sqrt{m_\star}$ implies $N = \mathcal{O}\big(C_j^{3/2}\big)$, and $(m_\star+1)^{\gamma_\star^+} \leq d^2\sqrt{m_\star T}$ implies $(m_\star+1)^{\gamma_\star^+}\, T_{\text{bc}}^{1/3} = \mathcal{O}\big(C_\star^{3/2}\, T_{\text{bc}}^{2/3}\big)$. Setting $\lambda \in [1, d]$, $\delta = 1/T$, we obtain

$$\mathbb{E}\left[\sum_{\tau=1}^{T_{\text{bc}}} \frac{\widetilde{r}(\widetilde{\boldsymbol{a}})}{|\widetilde{\boldsymbol{a}}|} - \frac{\widetilde{r}(\boldsymbol{a}_\tau^{\text{bc}})}{|\boldsymbol{a}_\tau^{\text{bc}}|}\right] = \widetilde{\mathcal{O}}\Big(d\,\sqrt{m_\star}\,(m_\star+1)^{\frac{3}{2}\gamma_\star^+}\, T_{\text{bc}}^{2/3}\Big). \tag{27}$$

Let m_τ be the memory size associated to the bandit played at block time step τ by Algorithm 2. Let $m_{\min} = \min_j m_j$ and $m_{\max} = \max_j m_j$. Finally, let $L_{\min}$ and $L_{\max}$ the (partial) block length associated with $m_{\min}$ and $m_{\max}$. We have

$$\sum_{t=1}^{T} r_t^{\text{bc}} \geq \sum_{\tau=1}^{T_{\text{bc}}} \left(\widetilde{r}(\boldsymbol{a}_\tau^{\text{bc}}) - m_\tau \, (m_\star + 1)^{\gamma_\star^+} \right) \geq \sum_{\tau=1}^{T_{\text{bc}}} \widetilde{r}(\boldsymbol{a}_\tau^{\text{bc}}) - m_{\max} \, (m_\star + 1)^{\gamma_\star^+} \, T_{\text{bc}} \, ,$$

such that by Lemma 1 and (27) we obtain

$$\mathbb{E}\left[\frac{\min_\tau |\boldsymbol{a}_\tau|}{\max_\tau |\boldsymbol{a}_\tau|} \left(\widetilde{r}(\widetilde{\boldsymbol{a}}) \, \frac{\sum_\tau |\boldsymbol{a}_\tau|}{|\widetilde{\boldsymbol{a}}|} \right) - \sum_{t=1}^{T} r_t^{\text{bc}} \right] \leq m_{\max} \, (m_\star + 1)^{\gamma_\star^+} \, T_{\text{bc}} + \min_\tau |\boldsymbol{a}_\tau| \, \widetilde{\mathcal{O}}\Big(d \sqrt{m_\star} \, (m_\star + 1)^{\frac{3}{2}\gamma_\star^+} \, T_{\text{bc}}^{2/3} \Big) \, ,$$

$$\mathbb{E}\left[\frac{m_{\min} + L_{\min}}{m_{\max} + L_{\max}} \left(\frac{L_\star \, \text{OPT}}{T} \, \frac{T}{m_\star + L_\star} \right) - \sum_{t=1}^{T} r_t^{\text{bc}} \right] \leq \frac{m_{\max} \, (m_\star + 1)^{\gamma_\star^+} \, T}{m_{\min} + L_{\min}} + (m_{\min} + L_{\min})^{1/3} \, \widetilde{\mathcal{O}}\Big(d \sqrt{m_\star} \, (m_\star + 1)^{\frac{3}{2}\gamma_\star^+} \, T^{2/3} \Big) \, ,$$

$$\mathbb{E}\left[\sqrt{\frac{m_{\min}}{m_{\max}}} \, \text{OPT} - \sum_{t=1}^{T} r_t^{\text{bc}} \right] \leq \frac{m_{\max}}{m_{\min}} \sqrt{d \, m_\star} \, (m_\star + 1)^{\gamma_\star^+} \, T^{3/4} + \widetilde{\mathcal{O}}\Big(d \, m_\star \, (m_\star + 1)^{\frac{3}{2}\gamma_\star^+} \, T^{3/4} \Big)$$

$$= \frac{m_{\max}}{m_{\min}} \, \widetilde{\mathcal{O}}\Big(d \, m_\star \, (m_\star + 1)^{\frac{3}{2}\gamma_\star^+} \, T^{3/4} \Big) \, ,$$

where we have used the fact that $m_{\min} + L_{\min} = \sqrt{m_{\min}/d} \, T^{1/4}$, and $m_{\max} + L_{\max} = \sqrt{m_{\max}/d} \, T^{1/4}$. Corollary 1 is obtained by setting $M = m_{\max}/m_{\min}$. □

B Bandit Combiner

In this section we provide more details on the algorithmic implementation of `Bandit Combiner`.

As mentioned in the main body of the paper,Our bandit combiner, see Algorithm 2 in Appendix B, builds upon the approach developed by Cutkosky et al. [2020] and works as follows. The meta-algorithm is fed with different bandit algorithms (in our case, instances of `O3M` with different choices of parameters m_j and γ_j) and at each round plays a block according to one of the algorithms. We relegate the explanation and details of this algorithmic solution to Appendix B. Each `O3M` instance comes with a *putative* regret bound $C_j T^{\alpha_j}$, which is the regret bound satisfied by the algorithm *should it be well-specified*, i.e., if the rewards are indeed generated through a memory matrix with memory m_j and exponent γ_j. Note that in order to be comparable across the different instances, the putative regrets apply to the average rewards. The values of C_j and α_j can be computed using Theorem 1, see the proof of Corollary 1 for details. The putative regrets are then used to successively discard the instances that are not well specified, and eventually identify the instance using parameters $(m_\star, \gamma_\star)$. Knowing C_j and α_j, we can compute for any j the target regret

$$R_j = C_j \, T_{\text{bc}}^{2/3} + \frac{5\sqrt{30}}{18} C_j^{3/2} \, T_{\text{bc}}^{2/3} + 1152 (m_j + 1)^{2\gamma_j^+} \, T^{1/3} \log(T_{\text{bc}}^3 N/\delta) + (N-1) T^{2/3} \, , \quad (28)$$

where T_{bc} is the number of blocks the Bandit Combiner is called on, see Appendix B for details. Here, we note how the presence of $(m_j + 1)^{2\gamma_j^+}$ is impacting differently the rising and rotting scenarios. Using [Cutkosky et al., 2020, Corollary 2], the regret of Algorithm 2 is finally given by $3R_{j_\star}$, where $j_\star$ is the index such that $(m_{j_\star}, \gamma_{j_\star}) = (m_\star, \gamma_\star)$.

In this section we show our adaptation of the numbers C_j and target regrets R_j for the Bandit Combiner algorithm Algorithm 2 which builds on Cutkosky et al. [2020]. For `O3M`(m_j, γ_j), $j = 1, \dots, N$, the numbers C_j and target regrets R_j are defined as

$$C_j = 4 \left(\frac{4 m_j}{d} \right)^{1/3} (m_j + 1)^{\gamma_j^+} \sqrt{d \ln \left(1 + \frac{T_{\text{bc}} (m_j + L_j)(m_j + 1)^{2\gamma_j^+}}{d\lambda} \right)} \quad (29)$$

Algorithm 2 `Bandit Combiner on O3M`

input : Instances $\texttt{O3M}(m_1, \gamma_1), \dots, \texttt{O3M}(m_N, \gamma_N)$, horizon T_{bc}
numbers $C_1, \dots, C_N > 0$, target regrets $R_1, \dots, R_N$.
Set $T(i) = 0, \mathcal{S}_i = 0, \Delta_i = 0$ for $i = 1, \dots, N$, and set $I_0 = \{1, \dots, N\}$

for $t = 1, \dots, T_{bc}$ **do**

  if *there is some* $i \in I_t$ *with* $T(i) = 0$ **then**

    $i_t = i$

  else

    For each $i \in I_t$, compute the UCB index:

$$\text{UCB}(i) = \min\left\{(m_i+1)^{2\gamma_i^+}, \frac{C_i}{\sqrt{T(i)}} + 4(m_i+1)^{2\gamma_i^+}\sqrt{\frac{2\log(T^3N/\delta)}{T(i)}}\right\} - \frac{R_i}{T_{\text{bc}}}$$

    Set $i_t = \arg\max_{i \in I_t} \frac{\mathcal{S}_i}{T(i)} + \text{UCB}(i)$

  Obtain from instance $\texttt{O3M}(m_{i_t}, \gamma_{i_t})$ a block of size $m_{i_t} + L_{i_t}$ and play it

  Return the total reward r_{i_t} collected in the last L_{i_t} time steps of the block to $\texttt{O3M}(m_{i_t}, \gamma_{i_t})$

  Compute the average reward $\widehat{r}_{i_t} = \frac{r_{i_t}}{L_{i_t}}$

  Update $\Delta_{i_t} \leftarrow \Delta_{i_t} + \mathcal{S}_{i_t}/T(i_t) - \widehat{r}_{i_t}$ (where we set $0/0 = 0$) and $\mathcal{S}_{i_t} \leftarrow \mathcal{S}_{i_t} + \widehat{r}_{i_t}$

  Update the number of plays $T(i_t) \leftarrow T(i_t) + 1$

  if $\Delta_{i_t} \geq C_{i_t} T(i_t)^{\gamma_{i_t}} + 12\,(m_{i_t}+1)^{2\gamma_{i_t}^+}\sqrt{2\log(T^3N/\delta)T(i_t)}$ **then**

    $I_t = I_{t-1} \setminus \{i_t\}$

  else

    $I_t = I_{t-1}$

$$\left(\sqrt{\lambda} + \sqrt{\ln\left(\frac{1}{\delta}\right) + d\,\ln\left(1 + \frac{T_{\text{bc}}(m_j+1)^{2\gamma_j^+}}{d\lambda}\right)}\right),$$

$$R_j = C_j T_{\text{bc}}^{\alpha_j} + \frac{(1-\alpha_j)^{\frac{1-\alpha_j}{\alpha_j}}(1+\alpha_j)^{\frac{1}{\alpha_j}}}{\alpha_j^{\frac{1-\alpha_j}{\alpha_j}}} C_j^{\frac{1}{\alpha_j}} T_{\text{bc}} \eta_j^{\frac{1-\alpha_j}{\alpha_j}} + 1152(m_j+1)^{2\gamma_j^+}\log(T_{\text{bc}}^3 N/\delta) T_{\text{bc}}\eta_j + \sum_{k \neq j}\frac{1}{\eta_k}.$$

Note that the form of the target regret R_j slightly differs from the one presented in [Cutkosky et al., 2020, Corollary 2] due to the different range of the rewards. The algorithm, which is an adaptation of Bandit Combiner in Cutkosky et al. [2020], is summarized in Algorithm 2.

C Additional Experiments

We provide an additional experiment comparing the regrets of `O3M` and `OM-Block`. In order to be able to plot the regret, we must know OPT which is hard to compute in general. Since in the rising scenario with an isotropic initialization OPT is oracle greedy, which is easy to compute, we present this experiment in a rising setting with $m = 1$ and $\gamma = 2$. We plot the regret of `O3M` and `OM-Block` against the number of time steps, measuring the performance at different time horizons and for different sizes of L (where L depends on T, see at the end of Section 3.2). Specifically, we instantiated `O3M` and `OM-Block` for increasing values of L, setting the horizon of each instance based on the equations in Theorem 1 and Proposition 4. Figure 3 shows how the dimension of $\widehat{\theta}$, which is d for `O3M` and $d \times L$ for `OM-Block`, has an actual impact on the performance since `O3M` outperforms `OM-Block`. The code is written in Python and it is publicly available at the following GitHub repository: Linear Bandits with Memory.

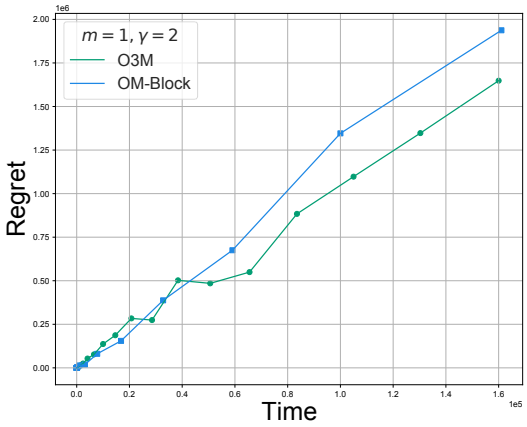

Figure 3: The regret of `O3M` and `OM-Block`. Each dot is a separate run where the value of L is tuned to the corresponding horizon.