A The regret bound of the Adaptive greedy algorithm

We present a finite-time bound on the cumulative regret defined in Equation (2).
Let $\mathcal{H}_{t-1}$ is the set of all possible histories (after deterministic initialization) of the game up to turn $t-1$:

$$\mathcal{H}_{t-1} = \left\{ h = \begin{bmatrix} b_{m_I+1} & b_{m_I+2} & \dots & b_{t-1} \\ i_{m_I+1} & i_{m_I+2} & \dots & i_{t-1} \end{bmatrix} : b_s \in \{0,1\},\ i_s \in M_s, \quad \forall s \in \{m_I+1, \dots, t-1\} \right\}. \tag{10}$$

If $b_s = 1$ we say that the algorithm explored at time s, if $b_s = 0$ we say that the algorithm exploited at time s, while i_s is the index of the arm that was played at time s.

Theorem 2.1 *Let us define the following quantities:*

- $g(p) = b + (a-b)p$,
- $f_{M(h,s)}(g(p))$ *is the PDF (or PMF) of the maximum of the estimated mean rewards at time* s *given that each arm has been pulled according to history* h *up to time* $s-1$*:*

$$f_{M(h,k)}(x) = \frac{1}{(m_t-1)!} \operatorname{perm} \left(\begin{bmatrix} F_1(x) & F_2(x) & \dots & F_{m_k}(x) \\ \vdots & \vdots & \vdots & \vdots \\ F_1(x) & F_2(x) & \dots & F_{m_k}(x) \\ f_1(x) & f_2(x) & \dots & f_{m_k}(x) \end{bmatrix} \right) \Bigg\} \ m_k - 1 \text{ rows} \ ,$$

  where $f_1(x), \cdots, f_{m_k}(x)$ *and* $F_1(x), \cdots, F_{m_k}(x)$ *are the PDFs (or PMFs) of the distributions of the average rewards,*

- $u_s(h, i_s)$ *is an upper bound on the probability that arm* i_s *is considered to be the best arm at time* s *given the history of pulls (according to* h*) up to time* $s-1$*:*

$$u_s(h, i_s) = \prod_{i:\mu_i > \mu_{i_s}} \left(\exp\left\{ -\frac{t_{i_s}(h,s)\Delta(i,i_s)^2}{2r} \right\} + \exp\left\{ -\frac{t_i(h,s)\Delta(i,i_s)^2}{2r} \right\} \right),$$

- $U_s(h, i_s)$ *is an upper bound on the probability that arm* i_s *was pulled at time* s *given the history of pulls (according to* h*) up to time* $s-1$*:*

$$U_s(h, i_s) = \int_0^1 \left(\frac{p}{m_s} \mathbb{1}_{\{b_s=1\}} + (1-p) u_s(h, i_s) \mathbb{1}_{\{b_s=0\}} \right) f_{M(h,s)}(g(p))\, dp,$$

- $u_t(h, j)$ *is an upper bound on the probability that arm* j *is considered to be the best arm at time* t *given the history of pulls (according to* h*) up to time* $t-1$*:*

$$u_t(h, j) = \prod_{i:\mu_i > \mu_j} \left(\exp\left\{ -\frac{t_j(h,t)\Delta(i,j)^2}{2r} \right\} + \exp\left\{ -\frac{t_i(h,t)\Delta(i,j)^2}{2r} \right\} \right),$$

- $U_t(h, j)$ *is an upper bound on the probability that arm* j *was pulled at time* t *given the history of pulls (according to* h*) up to time* $t-1$*:*

$$U_t(h, j) = \int_0^1 \left(\frac{p}{m_t} + (1-p) u_t(h, j) \right) f_{M(h,t)}(g(p))\, dp.$$

Then, an upper bound on the expected cumulative regret R_n *at round* n *is given by*

$$\mathbb{E}[R_n] \leq \sum_{j \in M_I} \Delta_{j,i_j^*} + \sum_{t=m_I+1}^{n} \sum_{j \in M_t} \Delta_{j,i_t^*} \sum_{h \in \mathcal{H}_{t-1}} \left(U_t(h,j) \prod_{s=m_I+1}^{t-1} U_s(h, i_s) \right).$$

First step: Decomposition of $\mathbb{E}[R_n]$.

$$\mathbb{E}[R_n] = \sum_{j \in M_I} \Delta_{j,i_j^*} + \sum_{t=m_I+1}^{n} \sum_{j \in M_t} \Delta_{j,i_t^*} \mathbb{P}\left(t \in I(j)\right), \tag{11}$$

where we can write $\mathbb{P}\left(t \in I(j)\right)$ as

$$\mathbb{P}\left(t \in I(j)\right) = \sum_{h \in \mathcal{H}_{t-1}} \mathbb{P}\left(t \in I(j) \;\middle|\; H_{t-1} = h\right) \mathbb{P}\left(H_{t-1} = h\right), \tag{12}$$

where H_{t-1} is a random variable that takes values in $\mathcal{H}_{t-1}$ defined as

$$\mathcal{H}_{t-1} = \left\{ h = \begin{bmatrix} b_{m_I+1} & b_{m_I+2} & \dots & b_{t-1} \\ i_{m_I+1} & i_{m_I+2} & \dots & i_{t-1} \end{bmatrix} : b_s \in \{0,1\},\ i_s \in M_s \quad \forall s \in \{m_I+1, \dots, t-1\} \right\}. \tag{13}$$

$\mathcal{H}_{t-1}$ is the set of all possible histories (after deterministic initialization) of the game up to turn $t-1$. If $b_s = 1$ we say that the algorithm explored at time s, if $b_s = 0$ we say that the algorithm exploited at time s, while i_s is the index of the arm that was played at time s. The set $\mathcal{H}_{t-1}$ has $\prod_{s=m_I+1}^{t-1}(2m_s)$ elements. Note also that, by design of the algorithm, if an arm j is new at time s,

$$\mathbb{P}\left(H_{t-1} = \begin{bmatrix} b_{m_I+1} & \dots & b_s = 0 & \dots & b_{t-1} \\ i_{m_I+1} & \dots & i_s = j & \dots & i_{t-1} \end{bmatrix}\right) = 0,$$

because the algorithm does not allow exploitation of a new arm. In the following steps we study (and find an upper bound when needed) each term in (12).

Second step: Upper bound for $\mathbb{P}\left(H_{t-1} = h\right)$.

Let us define h_k, with $k > m_I,\ k \in \mathbb{N}$, the first $k - m_I$ columns of h (so h and h_{t-1} are the same).
For each h, we indicate how many times arm j has been pulled up to time k with

$$t_j(h,k) = \mathbb{1}_{\{j \in M_I\}} + \sum_{s=m_I+1}^{k} \mathbb{1}_{\{i_s \in I(j)\}},$$

and, similarly to the definition of $\widehat{X}_j$ given in (1), we denote the mean estimated reward for arm j, given history of pulled arms h, with

$$\widehat{X}_j(h,k) = \frac{1}{t_j(h,k-1)} \sum_{s \in I(j)}^{t_j(h,k-1)} X_j(s). \tag{14}$$

For each h, the probability of exploration at time k is a random variable $E(h,k)$ with distribution given by

$$\mathbb{P}\left(E(h,k) = p\right) = \mathbb{P}\left(1 - \frac{\max_{j \in M_k} \widehat{X}_j(h,k) - a}{b-a} = p\right), \tag{15}$$

Let us define $g(p) = b + (a-b)p$, then we can rewrite (15) as

$$\mathbb{P}\left(E(h,k) = p\right) = \mathbb{P}\left(\max_{j \in M_k} \widehat{X}_j(h,k) = g(p)\right). \tag{16}$$

We will give a formula for 16 in the next step of the proof.
We can compute $\mathbb{P}\left(H_{t-1} = h\right)$ recursively using the fact that $\mathbb{P}\left(H_{t-1} = h\right)$ is equal to

$$\mathbb{P}\left(H_{t-1} = h \,\middle|\, H_{t-2} = h_{t-2}\right) \mathbb{P}\left(H_{t-2} = h_{t-2} \,\middle|\, H_{t-3} = h_{t-3}\right) \cdots \mathbb{P}\left(H_{m_I+2} = h_{m_I+2} \,\middle|\, H_{m_I+1} = h_{m_I+1}\right) \mathbb{P}\left(H_{m_I+1} = h_{m_I+1}\right). \tag{17}$$

h_{m_I+1} has only one column: $\begin{bmatrix} b_{m_I+1} \\ i_{m_I+1} \end{bmatrix}$, where $b_{m_I+1} \in \{0,1\}$ and $i_{m_I+1} \in M_{m_I+1}$.
We can write $\mathbb{P}\left(H_{m_I+1} = h_{m_I+1}\right)$ as

$$\int_0^1 \left(\frac{p}{m_{I+1}} \mathbb{1}_{\{b_{m_I+1}=1\}} + (1-p) \mathbb{P}\left(\widehat{X}_{i_{m_I+1}}(h, m_I+1) > \widehat{X}_i(h, m_I+1) \ \ \forall i \neq i_{m_I+1}\right) \mathbb{1}_{\{b_{m_I+1}=0\}} \right) \mathbb{P}\left(E(h, m_I+1) = p\right) \, \mathrm{d}p \tag{18}$$

Similarly, we can compute each term in (17). For each $s \in \{m_I+2, \cdots, t-1\}$, we have that $\mathbb{P}\left(H_s = h \,\middle|\, H_{s-1} = h_{s-1}\right)$ is given by

$$\int_0^1 \left(\frac{p}{m_s}\mathbb{1}_{\{b_s=1\}} + (1-p)\mathbb{P}\left(\widehat{X}_{i_s}(h,s) > \widehat{X}_i(h,s) \;\; \forall i \neq i_s\right)\mathbb{1}_{\{b_s=0\}}\right)\mathbb{P}\left(E(h,s)=p\right)\, \mathrm{d}p \tag{19}$$

Using independence of the arms and Proposition 3, for each $s \in \{m_I+1, \cdots, t-1\}$ we can write

$$\begin{aligned}
& \mathbb{P}\left(\widehat{X}_{i_s}(h,s) > \widehat{X}_i(h,s) \;\; \forall i \neq i_s\right) && (20)\\
\leq \;& \mathbb{P}\left(\widehat{X}_{i_s}(h,s) > \widehat{X}_i(h,s) \;\; \forall i : \mu_i > \mu_{i_s}\right) && (21)\\
\leq \;& \prod_{i:\mu_i>\mu_{i_s}} \mathbb{P}\left(\widehat{X}_{i_s}(h,s) > \widehat{X}_i(h,s)\right) && (22)\\
\leq \;& \prod_{i:\mu_i>\mu_{i_s}} \left[\mathbb{P}\left(\widehat{X}_{i_s}(h,s) > \mu_{i_s} + \frac{\Delta(i,i_s)}{2}\right) + \mathbb{P}\left(\widehat{X}_i(h,s) < \mu_i - \frac{\Delta(i,i_s)}{2}\right)\right] && (23)
\end{aligned}$$

and then bound each term by using Hoeffding's inequality[1]:

$$\mathbb{P}\left(\widehat{X}_{i_s}(h,s) > \mu_{i_s} + \frac{\Delta(i,i_s)}{2}\right) \leq \exp\left\{-\frac{t_{i_s}(h,s)\Delta(i,i_s)^2}{2r}\right\} \tag{24}$$

and

$$\mathbb{P}\left(\widehat{X}_i(h,s) < \mu_i - \frac{\Delta(i,i_s)}{2}\right) \leq \exp\left\{-\frac{t_i(h,s)\Delta(i,i_s)^2}{2r}\right\}. \tag{25}$$

Let us define

$$u_s(h,i_s) = \prod_{i:\mu_i>\mu_{i_s}} \left(\exp\left\{-\frac{t_{i_s}(h,s)\Delta(i,i_s)^2}{2r}\right\} + \exp\left\{-\frac{t_i(h,s)\Delta(i,i_s)^2}{2r}\right\}\right), \tag{26}$$

then, $\mathbb{P}\left(H_s = h \,\middle|\, H_{s-1} = h_{s-1}\right) \leq U_s(h,i_s)$, where

$$U_s(h,i_s) = \int_0^1 \left(\frac{p}{m_s}\mathbb{1}_{\{b_s=1\}} + (1-p)u_s(h,i_s)\mathbb{1}_{\{b_s=0\}}\right)\mathbb{P}\left(E(h,s)=p\right)\, \mathrm{d}p, \tag{27}$$

and from (17)

$$\mathbb{P}\left(H_{t-1} = h\right) \leq \prod_{s=m_I+1}^{t-1} U_s(h,i_s). \tag{28}$$

Third step: Formula for $\mathbb{P}\left(E(h,k)=p\right)$.

We can determine $\mathbb{P}\left(E(h,k)=p\right) = \mathbb{P}\left(\max_{j\in M_k} \widehat{X}_j(h,k) = g(p)\right)$ by using a result from Vaughan and Venables [1972] that describes the PDF of the maximum of random variables coming from different distributions. Note that each $\widehat{X}_j(h,k)$ has a different distribution[2] that depends also on s_j. Given a square matrix A, let $\mathrm{perm}(A)$ be the permanent[3] of A. Then, the PDF of $\max_{j\in M_t} \widehat{X}_j(h,k)$ is given by

$$f_{M(h,k)}(x) = \frac{1}{(m_t-1)!}\,\mathrm{perm}\left(\begin{bmatrix} F_1(x) & F_2(x) & \dots & F_{m_k}(x) \\ \vdots & \vdots & \vdots & \vdots \\ F_1(x) & F_2(x) & \dots & F_{m_k}(x) \\ f_1(x) & f_2(x) & \dots & f_{m_k}(x) \end{bmatrix}\right) \;\Bigg\}\; m_k - 1 \text{ rows} \tag{29}$$

[1] **Hoeffding's bound:** Let $X_1, \cdots, X_n$ be r.v. bounded in $[a_i, b_i]$ $\forall i$. Let $\widehat{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[\widehat{X}]$. Then, $\mathbb{P}\left(\widehat{X} - \mu \geq \varepsilon\right) \leq \exp\left\{-\frac{2n^2\varepsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}\right\}$. In our case, $\varepsilon = \frac{\Delta_{i_s,i}}{2}$, $n = t_s$ or t_{i_s}, $b-a = r$.

[2] For example, if X_i has a Bernoulli distribution with parameter μ_i, $\widehat{X}_{i,2}$ assumes values in $\{0,1,2\}$, with probabilities $(1-\mu_i)^2, (1-\mu_i)\mu_i, \mu_i^2$, while $\widehat{X}_{i,3}$ assumes values in $\{0,1,2,3\}$, with probabilities $(1-\mu_i)^3, (1-\mu_i)^2\mu_i, (1-\mu_i)\mu_i^2, \mu_i^3$.

[3] The permanent of a square matrix A is defined like the determinant, except that all signs are positive.

where $f_1(x), \cdots, f_{m_k}(x)$ and $F_1(x), \cdots, F_{m_k}(x)$ are the PDFs (or PMFs) of the cumulative distributions of the average rewards $\widehat{X}_j(h,k)$ of arms $j \in M_k$ (if unknown, they are approximated by a Normal r.v. by CLT). Thus,

$$\mathbb{P}\left(E(h,k)=p\right) = f_{M(h,k)}(g(p)). \tag{30}$$

Fourth step: Formula for $\mathbb{P}\left(t \in I(j) \;\middle|\; H_{t-1}=h\right)$.

We have that

$$\mathbb{P}\left(t \in I(j) \;\middle|\; H_{t-1}=h\right) = \int_0^1 \left[p\frac{1}{m_t} + (1-p)\mathbb{P}\left(\widehat{X}_j(h,t) > \widehat{X}_i(h,t) \;\; \forall i \neq j\right)\right] f_{M(h,t)}(g(p))\mathrm{d}p \tag{31}$$

Similarly to Step 2, $\mathbb{P}\left(\widehat{X}_j(h,t) > \widehat{X}_i(h,t) \;\; \forall i \neq j\right)$ has upper bound

$$u_t(h,j) = \prod_{i:\mu_i>\mu_j} \left(\exp\left\{-\frac{t_j(h,t)\Delta(i,j)^2}{2r}\right\} + \exp\left\{-\frac{t_i(h,t)\Delta(i,j)^2}{2r}\right\}\right), \tag{32}$$

and (31) has upper bound $U_t(h,j)$, where

$$U_t(h,j) = \int_0^1 \left(\frac{p}{m_t} + (1-p)u_t(h,j)\right) f_{M(h,t)}(g(p))\ \mathrm{d}p. \tag{33}$$

Note that $U_t(h,j)$ is different from $U_s(h,i_s)$ defined in (27) that have values of b_s available.

Fifth step: Bringing together all the bounds of the previous steps.

From (12) we have that

$$\mathbb{P}\left(t \in I(j)\right) = \sum_{h \in \mathcal{H}_{t-1}} \mathbb{P}\left(t \in I(j) \;\middle|\; H_{t-1}=h\right)\mathbb{P}\left(H_{t-1}=h\right) \leq \sum_{h \in \mathcal{H}_{t-1}} U_t(h,j) \prod_{s=m_I+1}^{t-1} U_s(h,i_s), \tag{34}$$

and from (11) in conclusion:

$$\mathbb{E}[R_n] \leq \sum_{j \in M_I} \Delta_{j,i_j^*} + \sum_{t=m_I+1}^{n} \sum_{j \in M_t} \Delta_{j,i_t^*} \sum_{h \in \mathcal{H}_{t-1}} \left(U_t(h,j) \prod_{s=m_I+1}^{t-1} U_s(h,i_s)\right).$$

B The regret bound of the UCB mortal algorithm

Theorem 2.3 *Let $\bigcup_{z=1}^{E_j} L_j^z$ be a partition of L_j into epochs with different best available arm, s_j^z and l_j^z be the first and last step of epoch L_j^z, and for each epoch let $u_{j,z}$ be defined as*

$$u_{j,z} = \max_{t \in \{s_j^z, \cdots, l_j^z\}} \left\lceil \frac{8\psi(j,t)\log(t-s_j)}{\Delta_{j,z}^2} \right\rceil, \tag{35}$$

where

$$\Delta_{j,i_t^*} = \Delta_{j,z} \ \text{ for } t \in L_j^z. \tag{36}$$

Then, the bound on the mean regret $\mathbb{E}[R_n]$ at time n is given by

$$\begin{aligned}\mathbb{E}[R_n] \quad &\leq \quad \sum_{j \in M_I} \Delta_{j,i_j^*} \\ &+ \quad \sum_{j \in M} \sum_{z=1}^{E_j} \Delta_{j,z} \min \left(l_j^z - s_j^z \ , \ u_{j,z} + \sum_{\substack{t \in L_j^z \\ t > m_I}} (t - s_{i_t^*})(t - s_j - u_{j,z} + 1) \left[(t-s_j)^{-\frac{4}{r^2}\psi(j,t)} + (t - s_{i_t^*})^{-\frac{4}{r^2}\psi(i_t^*,t)} \right] \right).\end{aligned}$$

First step: Decomposition of $\mathbb{E}[R_n]$.

Let us partition the set of steps L_j during which arm j is available into E_j epochs L_j^z, such that

- $\bigcup_{z=1}^{E_j} L_j^z = L_j$,
- $L_j^{z_1} \cap L_j^{z_2} = \emptyset$ if $z_1 \neq z_2$,
- $i_t^* \neq i_s^*$ if $t \in L_j^{z_1}$ and $s \in L_j^{z_2}$ (i.e., if different epochs have different best arm available).

Since during the same epoch the best arm available does not change, let us define

$$\Delta_{j,i_t^*} = \Delta_{j,z} \ \text{ for } t \in L_j^z, \tag{37}$$

and $s_j^z = \min L_j^z$, $l_j^z = \max L_j^z$ the first and last step of epoch L_j^z.
Then, using the second formulation of the cumulative regret given in (3) we have that

$$R_n = \sum_{j \in M_I} \Delta_{j,i_j^*} + \sum_{j \in M} \sum_{\substack{t \in L_j \\ t > m_I}} \Delta_{j,i_t^*} \mathbb{1}\{t \in I(j)\} \tag{38}$$

$$= \sum_{j \in M_I} \Delta_{j,i_j^*} + \sum_{j \in M} \sum_{z=1}^{E_j} \Delta_{j,z} \sum_{\substack{t \in L_j^z \\ t > m_I}} \mathbb{1}\{t \in I(j)\} \tag{39}$$

Let us call

$$T_j^z(l_j^z) = \sum_{\substack{t \in L_j^z \\ t > m_I}} \mathbb{1}\{t \in I(j)\}$$

the total number of times we choose arm j in epoch z during the game (after initialization). Then, by taking the expectation of (39) we get

$$\mathbb{E}[R_n] = \sum_{j \in M_I} \Delta_{j,i_j^*} + \sum_{j \in M} \sum_{z=1}^{E_j} \Delta_{j,z} \, \mathbb{E}\left[T_j^z(l_j^z)\right]. \tag{40}$$

Therefore, finding an upper bound for the expected value of (38) can be accomplished by bounding the expected value of $T_j^z(l_j^z)$.

Second step: Decomposition of $T_j^z(l_j^z)$.

Recall that with $T_j(t-1)$ we indicate the number of times we played arm j before turn t starts. For any integer $u_{j,z}$, we can write

$$\begin{aligned} T_j^z(l_j^z) &= u_{j,z} + \sum_{\substack{t\in L_j^z \\ t>m_I}} \mathbb{1}\{t \in I(j), T_j(t-1) \geq u_{j,z}\} \\ &= u_{j,z} + \sum_{\substack{t\in L_j^z \\ t>m_I}} \mathbb{1}\left\{\widehat{X}_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}} > \widehat{X}_{i_t^*} + \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{T_{i_t^*}(t-1)}}, T_j(t-1) \geq u_{j,z}\right\} \\ &\leq u_{j,z} + \sum_{\substack{t\in L_j^z \\ t>m_I}} \sum_{k_j=u_{j,z}}^{t-s_j} \sum_{k_{i_t^*}=1}^{t-s_{i_t^*}} \mathbb{1}\left\{\widehat{X}_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{k_j}} > \widehat{X}_{i_t^*} + \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{k_{i_t^*}}}\right\}. \end{aligned}$$

Therefore we can find an upper bound for the expectation of $T_j^z(l_j^z)$ by finding an upper bound for the probability of the event

$$A = \left\{\widehat{X}_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{k_j}} > \widehat{X}_{i_t^*} + \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{k_{i_t^*}}}\right\}.$$

Third step: Upper bound for $\mathbb{E}[T_j^z(l_j^z)]$.

Using Proposition 1 and Proposition 2 we have that, by choosing $u_{j,z} = \max_{t\in\{s_j^z,\cdots,l_j^z\}} \left\lceil \frac{8\psi(j,t)\log(t-s_j^z)}{\Delta_{j,z}^2} \right\rceil$,

$$A \subset \left(\left\{\widehat{X}_{i_t^*} < \mu_{i_t^*} - \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{k_{i_t^*}}}\right\} \cup \left\{\widehat{X}_j > \mu_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{k_j}}\right\}\right). \tag{41}$$

Using Hoeffding's[4] bound we have that

$$\begin{aligned} &\mathbb{P}\left(\widehat{X}_{i_t^*} < \mu_{i_t^*} - \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{T_{i_t^*}(t-1)}}\right) \leq \exp\left\{-\frac{2k_{i_t^*}^2\psi(i_t^*,t)^2\frac{2\log(t-s_{i_t^*})}{k_{i_t^*}}}{k_{i_t^*}r^2}\right\} = (t-s_{i_t^*})^{-\frac{4}{r^2}\psi(i_t^*,t)} \\ &\mathbb{P}\left(\widehat{X}_j > \mu_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}}\right) \leq \exp\left\{-\frac{2k_j^2\psi(j,t)^2\frac{2\log(t-s_j)}{k_j}}{k_j r^2}\right\} = (t-s_j)^{-\frac{4}{r^2}\psi(j,t)}. \end{aligned}$$

Using the inclusion in (41) in combination with Hoeffding's bounds, we have that

$$\begin{aligned} \mathbb{E}\left[T_j^z(l_j^z)\right] &\leq u_{j,z} + \sum_{\substack{t\in L_j^z \\ t>m_I}} \sum_{k_j=u_{j,z}}^{l_j} \sum_{k_{i_t^*}=1}^{t-s_{i_t^*}} \mathbb{P}\left\{\widehat{X}_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{k_j}} > \widehat{X}_{i_t^*} + \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{k_{i_t^*}}}\right\} \\ &\leq u_{j,z} + \sum_{\substack{t\in L_j^z \\ t>m_I}} \sum_{k_j=u_{j,z}}^{t-s_j} \sum_{k_{i_t^*}=1}^{t-s_{i_t^*}} \left[(t-s_{i_t^*})^{-\frac{4}{r^2}\psi(i_t^*,t)} + (t-s_j)^{-\frac{4}{r^2}\psi(j,t)}\right] \\ &= u_{j,z} + \sum_{\substack{t\in L_j^z \\ t>m_I}} (t-s_{i_t^*})(t-s_j-u_{j,z}+1)\left[(t-s_j)^{-\frac{4}{r^2}\psi(j,t)} + (t-s_{i_t^*})^{-\frac{4}{r^2}\psi(i_t^*,t)}\right]. \end{aligned} \tag{42}$$

[4] **Hoeffding's bound:** Let $X_1,\cdots,X_n$ be r.v. bounded in $[a_i,b_i]$ $\forall i$. Let $\widehat{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[\widehat{X}]$. Then, $\mathbb{P}\left(\widehat{X} - \mu \geq \varepsilon\right) \leq \exp\left\{-\frac{2n^2\varepsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}\right\}$.

In our case, n is k_j or $k_{i_t^*}$, $b_i - a_i$ is r, μ is μ_j or $\mu_{i_t^*}$, and ε is $\psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}}$ or $\psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{T_{i_t^*}(t-1)}}$.

Of course, we also have that the expected number of times the algorithm chooses arm j during epoch L_j^z is also bounded by the length of the epoch itself $l_j^z - s_j^z$ (this bound is useful in case the epoch is very short). Combining this with (42) we have that

$$\mathbb{E}\left[T_j^z(l_j^z)\right] \leq \min\left(l_j^z - s_j^z \ , \ u_{j,z} + \sum_{\substack{t \in L_j^z \\ t > m_I}} (t - s_{i_t^*})(t - s_j - u_{j,z} + 1)\left[(t-s_j)^{-\frac{4}{r^2}\psi(j,t)} + (t - s_{i_t^*})^{-\frac{4}{r^2}\psi(i_t^*,t)}\right]\right). \quad (43)$$

Fourth step: Get upper bound for $\mathbb{E}[R_n]$.

Combining (43) with (40) we get that the bound on the cumulative regret is given by

$$\begin{aligned}\mathbb{E}[R_n] \ &\leq \ \sum_{j \in M_I} \Delta_{j,i_j^*} \\ &+ \ \sum_{j \in M} \sum_{z=1}^{E_j} \Delta_{j,z} \min\left(l_j^z - s_j^z \ , \ u_{j,z} + \sum_{\substack{t \in L_j^z \\ t > m_I}} (t - s_{i_t^*})(t - s_j - u_{j,z} + 1)\left[(t-s_j)^{-\frac{4}{r^2}\psi(j,t)} + (t - s_{i_t^*})^{-\frac{4}{r^2}\psi(i_t^*,t)}\right]\right).\end{aligned}$$

Notice that if $\psi(j,t) = 1$, $s_j = 0$ and $l_j > n \ \forall j, t$, you can recover the bound of the standard UCB algorithm used in the stochastic case. (Note that you should use $P > 2$ instead of 2 when r is not 1 to create the UCB.)

The results in Proposition 1 and 2 are similar to arguments used in Auer et al. [2002] for the proof of the regret bound for the UCB algorithm (here we have additional weighting of the upper confidence bound).

Proposition 1. *The event*

$$A = \left\{ \widehat{X}_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}} > \widehat{X}_{i_t^*} + \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{T_{i_t^*}(t-1)}} \right\}$$

is included in $B \cup C \cup D$, where

$$\begin{aligned} B &= \left\{ \widehat{X}_{i_t^*} < \mu_{i_t^*} - \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{T_{i_t^*}(t-1)}} \right\} \\ C &= \left\{ \widehat{X}_j > \mu_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}} \right\} \\ D &= \left\{ \mu_{i_t^*} - \mu_j < 2\psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}} \right\} \end{aligned}$$

The inclusion $A \subset (B \cup C \cup D)$ intuitively means that if the algorithm is choosing to play suboptimal arm j at turn t, then it is underestimating the best arm available (event B), or it is overestimating arm j (event C), or it has not pulled enough times arm j to distinguish its performance from the one of arm i_t^* (event D).

For the sake of contradiction let us assume there exists $\omega \in A$ such that $\omega \in (B \cup C \cup D)^{\mathcal{C}}$. Then, for that ω, none of the inequalities that define the events B, C, and D would hold, i.e. (using, in order, the inequality in B, then the one in D, then the one in C):

$$\begin{aligned} \widehat{X}_{i_t^*} &\geq \mu_{i_t^*} - \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{T_{i_t^*}(t-1)}} \\ &\geq \mu_j + 2\psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}} - \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{T_{i_t^*}(t-1)}} \\ &\geq \widehat{X}_j + \psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}} - \psi(i_t^*,t)\sqrt{\frac{2\log(t-s_{i_t^*})}{T_{i_t^*}(t-1)}}, \end{aligned}$$

which contradicts $\omega \in A$.

The result in Proposition 2 is similar to the one used in Auer et al. [2002] for the proof of the regret bound for the UCB algorithm.

Proposition 2. *When*

$$T_j(t-1) \geq \left\lceil \frac{8\psi(j,t)\log(t-s_j)}{\Delta_{j,i_t^*}^2} \right\rceil$$

event D in Preposition 1 can not happen.

In fact,

$$
\begin{aligned}
& \mu_{i_t^*} - \mu_j - 2\psi(j,t)\sqrt{\frac{2\log(t-s_j)}{T_j(t-1)}} \\
\geq \quad & \mu_{i_t^*} - \mu_j - 2\psi(j,t)\sqrt{\frac{2\log(t-s_j)}{\left\lceil \frac{8\psi(j,t)\log(t-s_j)}{\Delta_{j,i_t^*}^2} \right\rceil}} \\
\geq \quad & \mu_{i_t^*} - \mu_j - 2\psi(j,t)\sqrt{\frac{\log(t-s_j)\Delta_{j,i_t^*}^2}{4\psi(j,t)\log(t-s_j)}} \\
= \quad & \mu_{i_t^*} - \mu_j - \Delta_{j,i_t^*} = 0.
\end{aligned}
$$

C Useful results

The result in Proposition 3 is similar to the one used in Auer et al. [2002] for the proof of the regret bound for the ε-greedy algorithm.

Proposition 3. *Let $\mu_i > \mu_j$ and let us define the following events:*

$$\begin{aligned} A &= \left\{\widehat{X}_j > \widehat{X}_i\right\}, \\ B &= \left\{\widehat{X}_i < \mu_i - \frac{\Delta(i,j)}{2}\right\}, \\ C &= \left\{\widehat{X}_j > \mu_j + \frac{\Delta(i,j)}{2}\right\}. \end{aligned}$$

Then,

$$A \subset (B \cup C)\,. \tag{44}$$

Intuitively, the inclusion in (44) means that we play arm j when we underestimate the mean reward of the best arm, or when we overestimate that of arm j. Assume for the sake of contradiction that there exists an element $\omega \in A$ that does not belong to $B \cup C$. Then, we have that $\omega \in (B \cup C)^C$

$$\Rightarrow \quad \omega \quad \in \quad \left(\left\{\widehat{X}_i < \mu_i - \frac{\Delta(i,j)}{2}\right\} \cup \left\{\widehat{X}_j > \mu_j + \frac{\Delta(i,j)}{2}\right\}\right)^C \tag{45}$$

$$\Rightarrow \quad \omega \quad \in \quad \left\{\widehat{X}_i \geq \mu_i - \frac{\Delta(i,j)}{2}\right\} \cap \left\{\widehat{X}_j \leq \mu_j + \frac{\Delta(i,j)}{2}\right\}. \tag{46}$$

By definition we have $\mu_i - \frac{\Delta(i,j)}{2} = \mu_i - \frac{\mu_i - \mu_j}{2} = \frac{\mu_i + \mu_j}{2} = \mu_j + \frac{\Delta(i,j)}{2}$. From the inequalities given in (46) it follows that

$$\widehat{X}_i \geq \mu_i - \frac{\Delta(i,j)}{2} = \mu_j + \frac{\Delta(i,j)}{2} \geq \widehat{X}_j,$$

but this contradicts our assumption that $\omega \in A = \left\{\widehat{X}_j > \widehat{X}_i\right\}$.
Therefore, all elements of A belong to $B \cup C$.

D Numerical results

D.1 Dataset

The dataset can be found on the Yahoo Webscope program. It contains files recording 15 days of article recommendation history. Each record shows information about the displayed article id, user features, timestamp and the candidate pool of available articles at that time. The displayed article id shows the arm that recommenders pick each turn. User features were not used, since our algorithms look for articles generally liked by everyone. Timestamp tells the time that event happens; along with the candidate pool of available articles, we can scan through the records and find out each article's lifespan.

D.2 Evaluation methodology

A unique property of this dataset is that the displayed article is chosen uniformly at random from the candidate article pool. Therefore, one can use an unbiased offline evaluation method [Li et al., [2011]] to compare bandit algorithms in a reliable way. However, in the initialization phase, we applied a simpler and faster method (Algorithm 4), since initialization only plays 25 turns in a game and we care more about what happens later on.

In order to apply these evaluation methods, after parsing the original text log into structured data frame, we made an event stream generator out of it. The event stream generator has a member method "next_event()" that gives us the next record in the data frame. The fields in the record give information about the event. For example, in the initialization phase we checked the "article" field of the records to see if that article had been played before.

Algorithm 4: Initialization

event stream $Stream$
number of turns as initialization m
$i \leftarrow 0$
while $i < m$ **do**
  $Record \leftarrow Stream.next\_event()$
  if $Record.article$ was not seen before **then**
    update expectation of $Record.article$
    $i \leftarrow i + 1$
  end if
end while

D.3 Parameter tuning

AG-L filters out a portion of articles that expire soon. This portion is a tunable parameter. We tested different values with a smaller size dataset and finally used 0.1 as the threshold. In UCB-L's upper confidence bound, $\psi(j,t) = clog(l_j - t + 1)$ and c is a tunable parameter. After tuning, we set $c = 0.011$ for later experiments.

D.4 UCB score function

The original expression for the modified upper confidence bound in UCB-L is $X + \psi\sqrt{\frac{2log(t-s)}{T}}$. In the experiment, we used $X + \psi\sqrt{\frac{2log(t-s+1)}{T}}$ because sometimes an article is chosen the turn it becomes available and $t = s$, leading to an invalid value.

D.5 Timestamp vs Turn number

The original dataset records each action and reward along its corresponding timestamp, which we pre-scanned to get the articles' lifespan. But our algorithms count articles' lifespan in "turns." In this offline evaluation setting, a lot of events are discarded if they don't match the actions our algorithms pick, and there is no direct relation between an event's timestamp and turn number.

For AG-L, each time we rank articles by their remaining life, and filter out those with shorter remaining life. Since "timestamp" and "turn number" are positively correlated, ranking by "timestamp" and ranking by "turn number" would

almost give the same result ("almost" because sometimes consecutive events have equal timestamps, but later events should happen later in the game and are associated with bigger turn numbers). Then we can simulate AG-L using the timestamps.

In AG-L, if we use the average lifespan of previous arms as the estimation for current arms, we will be comparing arms by their birth time – all current arms are expected to have the same life length and the later an arm is born, the shorter remaining life it has.

However, UCB-L needs the exact turn number to compute the modified upper confidence bound. We can record the turn number when an article appears, but there's no way to know at which turn this article expires – We only know at what timestamp it disappears, but we can't transfer timestamps to turn numbers. Therefore, we only simulated UCB-L using life estimation.

For a short period in the early game we didn't have the life estimation because we hadn't yet seen an expired article. Also, sometimes our estimated life length $\hat{L}$ is too small, then $\hat{L}+s-t+1 \leq 0$, leading $\psi(j,t) = clog(l-t+1) = clog(\hat{L}+s-t+1)$ to an invalid value. In these cases we took $l-t+1 = \hat{L}+s-t+1 = 0$ and used only X as the upper confidence bound.

D.6 Contextual algorithm

Algorithm 5 is a similar adaptation of the LinUCB algorithm introduced by Li et al. [2010] to the mortal setting. Also in this case, the function $\psi(j,t)$ regulates the amplitude of the upper confidence bound above the estimated mean according to the remaining life of the arm. As before, new arms are initialized by using the average performance of past arms (i.e., if in the past a lot of bad arms appeared, new arms are considered more likely to be bad, and vice-versa if lots of good arms appeared in the past).

Algorithm 5: LinUCB-L algorithm

Input : number of rounds n, initial set of arms M_I, set M_t of available arms at time, rewards range $[a,b]$, dimension d (context space dimension + arms space dimension)

Initialization: For each $j \in M_I$, $A_j = I_d$, $b_j = 0_{d\times 1}$

for $t = 1$ **to** n **do**

  Get context x_t (or $x_{t,j}$ if each arm gets its context);

  for $j = 1$ **to** m_t **do**

    Set $\hat{\theta}_j = A_j^{-1} b_j$;

    Set $UCB_j = \hat{\theta}_j^T x_t + \psi(j,t)\sqrt{x_t^T A_j^{-1} x_t}$;

  end

  Play arm $j = \text{argmax}_i\, UCB_i$;

  Get reward $X_j(t)$;

  Update $A_j = A_j^{-1} + x_t x_t^T$;

  Update $b_j = b_j + r_t x_t$;

end

We have noticed that the contextual algorithm was not useful for the features made available in the Yahoo! Webscope Dataset, so for the experiments we used the non-contextual version presented in the main paper.

E Notation summary

- M_t as the set of all available arms at turn t;
- M_I the set of arms that are initialized;
- m_t: number of arms available at time t;
- n: total number of rounds;
- $X_j(t)$: random reward for playing arm j at time t;
- μ_*: mean reward of the optimal arm ($\mu_* = \max_{1 \leq j \leq m} \mu_j$);
- $\Delta(i, j)$: difference between the mean reward of arm i and arm j ($\Delta(i, j) = \mu_i - \mu_j$);
- $\hat{X_j}$: current estimate of μ_j;
- I_j: set of turns when arm j is played;
- $T_j(t-1)$: r.v. of the number of times arm j has been played before round t starts;
- $\mathcal{H}_{t-1}$: set of all possible histories h (after deterministic initialization) of the game up to turn $t-1$;
- $U_s(h, i_s)$: upper bound on the probability that arm i_s was pulled at time s given the history of pulls h up to time $s-1$;
- $u_s(h, i_s)$: upper bound on the probability that arm i_s is considered to be the best arm at time s given the history of pulls h up to time $s-1$;
- $f_{M(h,s)}(g(p))$: the PDF (or PMF) of the maximum of the estimated mean rewards at time s given that each arm has been pulled according to history h up to time $s-1$;
- $g(p)$: linear transformation $g(p) = b + (a-b)p$;
- $U_t(h, j)$: upper bound on the probability that arm j was pulled at time t given the history of pulls h up to time $t-1$;
- $u_t(h, j)$: upper bound on the probability that arm j is considered to be the best arm at time t given the history of pulls h up to time $t-1$;
- R_n: total regret at round n.