Supplementary material to ”Preselection Bandits”

A. Proofs of Theorems 4.1 and 4.2

For the proofs of Theorem 4.1 and Theorem 4.2 we need the following result on the Kullback-Leibler divergence of categorical probability distributions, which is Lemma 3 in Chen & Wang (2018). Throughout the proofs we let $\gamma \in (0, \infty)$ be some arbitrary degree of preciseness.

Lemma A.1. *Let $P \sim Cat(p_1, \ldots, p_m)$, i.e. $P(i) = p_i$ for $i = 1, \ldots, m$ and $\sum_{i=1}^m p_i = 1$, as well as $Q \sim Cat(q_1, \ldots, q_m)$, such that $q_i = p_i + \varepsilon_i$ and $|\varepsilon_i| < 1$ for any $i = 1, \ldots, m$. Then,*

$$\mathrm{KL}\big(P, Q\big) \leq \sum_{i=1}^m \frac{\varepsilon_i^2}{q_i}.$$

Moreover, we will need the following auxiliary result for all lower bound results.

Lemma A.2. *For any $\delta \in (0, 1)$ and any $\gamma \in (0, \infty)$ it holds that*

$$1 - (1 - \delta)^{1/\gamma} \geq \min\{1, 1/\gamma\}\, \delta.$$

Proof. First, consider the case $\gamma \in (0, 1]$. Then the assertion follows immediately as the left-hand side of the inequality is monotonically decreasing with γ and for $\gamma = 1$ the inequality is valid.

Next, let us consider the case $\gamma \in (1, \infty)$. The assertion is equivalent to showing that $f(x) = 1 - x\delta - (1 - \delta)^x$ is non-negative for $x \in (0, 1)$. The first and second derivatives are respectively

$$f'(x) = -\delta - \log(1 - \delta)(1 - \delta)^x,$$
$$f''(x) = -\log(1 - \delta)^2 (1 - \delta)^x.$$

By straightforward computations it can be shown that f has a global maximum on $(0, 1)$ at $x_{max} = \frac{\log\left(\frac{-\delta}{\log(1-\delta)}\right)}{\log(1-\delta)}$ and f is strictly increasing on $(0, x_{max})$ and strictly decreasing on $(x_{max}, 1)$. As $\lim\limits_{x\to 0} f(x) = \lim\limits_{x \to 1} f(x) = 0$, we can conclude the lemma. □

Proof of Theorem 4.1. We will use a similar proof technique as in Chen & Wang (2018). Let φ be some arbitrary algorithm suggesting the l-sized subsets (preselections) $(S_t^\varphi)_{t\in[T]} \subset \mathbb{A}_l$. For a set $S \in \mathbb{A}_l$ we write $\theta_S = (\theta_S(1), \ldots, \theta_S(n))$ to denote the score parameter of the PL-model with components given by

$$\theta_S(i) := \begin{cases} 1, & i \in S, \\ 1 - \varepsilon, & i \notin S, \end{cases}$$

where $\varepsilon \in (0, 1/2)$ is some hardness parameter specified below. Note that for any $S \in \mathbb{A}_l$ the score parameter θ_S is an element of the parameter space Θ. For sake of convenience, we will write $\mathbb{P}_S$ and $\mathbb{E}_S$ to express the law and expectation associated with the parameter θ_S, i.e., $\mathbb{P}_S = \mathbb{P}_{\theta_S}$. Recall the decomposition in (5) such that we have $\theta_S(i) = v_S(i)^\gamma$ for some suitable $v_S(i)$'s respectively and we define v_S in the same spirit as θ_S.

First, for any $S, \tilde{S} \in \mathbb{A}_l$ with $S \neq \tilde{S}$ it holds that

$$\begin{aligned} &\mathrm{U}(S; v_S, \gamma) - \mathrm{U}(\tilde{S}; v_S, \gamma) \\ &\quad \geq 1 - \frac{(l-1) + (1-\varepsilon)^{\frac{(1+\gamma)}{\gamma}}}{l - \varepsilon} \\ &\quad = \frac{(1-\varepsilon)(1 - (1-\varepsilon)^{\frac{1}{\gamma}})}{l - \varepsilon} \\ &\quad > \frac{\min\{1, 1/\gamma\}\, \varepsilon}{2\, l}, \end{aligned} \tag{A.1}$$

where we used for the last step $1 - \varepsilon \geq 1/2$ and $l - \varepsilon < l$ as well as Lemma A.2. For $i \in [n]$ let $N_i(t) = \sum_{s=1}^t 1_{\{i \in S_s^\varphi\}}$ denote the number of times an arm i is part of a preselection till time instance t suggested by some algorithm φ. In particular, write $N_i = N_i(T)$, then (A.1) implies

$$\begin{aligned} &\mathbb{E}_S \sum_{t=1}^T \mathrm{U}(S, \theta_S) - \mathrm{U}(S_t^\varphi, \theta_S) \\ &\qquad \geq \frac{\min\{1, 1/\gamma\}\, \varepsilon}{2\, l} \sum_{i \notin S} \mathbb{E}_S N_i. \end{aligned} \tag{A.2}$$

We can bound the expected regret from below as follows

$$\begin{aligned} &\sup_{\theta \in \Theta} \mathbb{E}_\theta \mathcal{R}(T) \\ &\quad \geq \sup_{S \in \mathbb{A}_l} \mathbb{E}_S\, \mathcal{R}(T) \\ &\quad = \sup_{S \in \mathbb{A}_l} \mathbb{E}_S \sum_{t=1}^T \mathrm{U}(S, \theta_S) - \mathrm{U}(S_t^\varphi, \theta_S) \\ &\quad \geq \frac{1}{\binom{n}{l}} \sum_{S \in \mathbb{A}_l} \mathbb{E}_S \sum_{t=1}^T \mathrm{U}(S, \theta_S) - \mathrm{U}(S_t^\varphi, \theta_S) \end{aligned}$$

$$\begin{aligned}&\geq \frac{1}{\left(\genfrac{}{}{0pt}{}{n}{l}\right.} \sum_{S\in\mathbb{A}_l} \sum_{i\notin S} \frac{\min\{1,1/\gamma\}\,\varepsilon}{2\,l}\, \mathbb{E}_S N_i \\ &= \frac{\min\{1,1/\gamma\}\,\varepsilon}{2} \Big(T - \frac{1}{l\left(\genfrac{}{}{0pt}{}{n}{l}\right.} \sum_{S\in\mathbb{A}_l} \sum_{i\in S} \mathbb{E}_S N_i\Big),\end{aligned}$$

where we used for the last inequality (A.2) and for the last equality that $T\,l = \sum_{i=1}^n \mathbb{E}_S N_i = \sum_{i\in S} \mathbb{E}_S N_i + \sum_{i\notin S} \mathbb{E}_S N_i$. Now, using Formulas (5) – (7) in Chen & Wang (2018) and Hölder's resp. Jensen's inequality as in Section 3.4 of Chen & Wang (2018) one obtains

$$\begin{aligned}&\sup_{\theta\in\Theta} \mathbb{E}_\theta \mathcal{R}(T) \\ &\quad\geq \frac{\min\{1,1/\gamma\}\,\varepsilon\, T}{2} \\ &\quad\times \Big(\frac{2}{3} - \sup_{S'\in\mathbb{A}_{l-1}} \sqrt{\sum_{i\in S'} \frac{\mathrm{KL}\big(\mathbb{P}_{S'},\mathbb{P}_{S'\cup\{i\}}}{2(n-l+1)}}\Big).\end{aligned}$$

The Kullback-Leibler divergence in the latter display can be dealt with by the following lemma which is proved below.

Lemma A.3. *For each $S' \in \mathbb{A}_{l-1}$ and $i\in S'$ the following bound is true*

$$\mathrm{KL}\big(\mathbb{P}_{S'},\mathbb{P}_{S'\cup\{i\}} \;\leq \frac{22\,\varepsilon^2\,\mathbb{E}_{S'}N_i}{l}.$$

With Lemma A.3 we have that for any $S' \in \mathbb{A}_{l-1}$

$$\sqrt{\sum_{i\in S'} \frac{\mathrm{KL}\big(\mathbb{P}_{S'},\mathbb{P}_{S'\cup\{i\}}}{2(n-l+1)}} \leq \sqrt{\frac{11\,\varepsilon^2\,T}{n}},$$

since $\sum_{i\in S'} \mathbb{E}_{S'}N_i \leq Tl$. Thus, choosing $\varepsilon = \min(C\sqrt{n/T}, 1/2)$ for some appropriate small constant $C > 0$, independent of T, n and l, we obtain the assertion. □

Proof of Lemma A.3. Let $\tilde{S} \in \mathbb{A}_l$ be arbitrary. Then $\mathbb{P}_{S'}(\cdot|\tilde{S})$ denotes the (categorical) probability distribution on the set $\tilde{S}$ parameterized by $\theta_{S'}$, i.e.,

$$\mathbb{P}_{S'}(j|\tilde{S}) = \begin{array}{ll} \frac{\theta_{S'}(j)}{\sum_{k\in\tilde{S}}\theta_{S'}(k)}, & j\in\tilde{S}, \\ 0, & \text{else}.\end{array}$$

If $i\notin\tilde{S}$ then $\mathrm{KL}\big(\mathbb{P}_{S'}(\cdot|\tilde{S}),\mathbb{P}_{S'\cup\{i\}}(\cdot|\tilde{S}) \;= 0$, as both distributions coincide in this case. Thus, we have the following bound

$$\begin{aligned}&\mathrm{KL}\big(\mathbb{P}_{S'},\mathbb{P}_{S'\cup\{i\}} \\ &\quad\leq \mathrm{KL}\big(\mathbb{P}_{S'}(\cdot|\tilde{S},\, i\in\tilde{S}),\mathbb{P}_{S'\cup\{i\}}(\cdot|\tilde{S},\, i\in\tilde{S}) \qquad \text{(A.3)} \\ &\quad\times \mathbb{E}_{S'}N_i,\end{aligned}$$

as $i\in\tilde{S}$ happens $\mathbb{E}_{S'}N_i$ times in expectation. We proceed by bounding the Kullback-Leibler-divergence on the right-hand side of (A.3). Define $J_+ = |\tilde{S}\cap S'|$, and $J_- = |\tilde{S}\cap (S')^{\complement}|$. Since $\tilde{S}\in\mathbb{A}_l$ it holds that $J_+ + J_- = l$. With this, the categorical probabilities for $j\in\tilde{S}$ are given by

$$\begin{aligned}p_j &:= \mathbb{P}_{S'}(j|\tilde{S},\, i\in\tilde{S}) = \frac{\theta_{S'}(j)}{J_+ + (1-\varepsilon)J_-}, \\ q_j &:= \mathbb{P}_{S'\cup\{i\}}(j|\tilde{S},\, i\in\tilde{S}) = \frac{\theta_{S'}(j)}{J_+ + 1 + (1-\varepsilon)(J_- - 1)}.\end{aligned}$$

For $j = i$ it holds that ${}^{(p_j-q_j)^2}/{}_{q_j} \leq {}^{8\varepsilon^2}/{}_{l^3}$. We show this exemplary for the case, where $j = i$ and $j\in\tilde{S}\cap S'$, while the case $j = i$ and $j\notin\tilde{S}\cap S'$, can be dealt with similarly. It holds that $J_+ + (1-\varepsilon)J_- = l - \varepsilon J_-$ and $J_+ + 1 + (1-\varepsilon)(J_- - 1) = l + \varepsilon(1 - J_-)$, so that

$$p_j - q_j = \frac{\varepsilon}{\big[l-\varepsilon J_-\big]\big[l+\varepsilon(1-J_-)\big]}$$

and with this

$$\frac{(p_j-q_j)^2}{q_j} = \frac{\varepsilon^2}{\big[l-\varepsilon J_-\big]^2\big[l+\varepsilon(1-J_-)\big]} \leq \frac{8\varepsilon^2}{l^3},$$

as the terms inside the squared brackets are respectively greater than $l/2$, since $\varepsilon\in(0,1/2)$ and $|J_+|, |J_-| \leq l$. If $j = i$, then ${}^{(p_j-q_j)^2}/{}_{q_j} \leq {}^{20\varepsilon^2}/{}_{l}$. Indeed, we have

$$p_j - q_j = \frac{\varepsilon\big(1-l-\varepsilon(1-J_-)}{\big[l-\varepsilon J_-\big]\big[l+\varepsilon(1-J_-)\big]},$$

so that

$$\frac{(p_j-q_j)^2}{q_j} = \frac{\varepsilon^2\big(1-l-\varepsilon(1-J_-)\;^2}{\big[l-\varepsilon J_-\big]^2\big[l+\varepsilon(1-J_-)\big]} \leq \frac{20\varepsilon^2}{l},$$

since $\big(1-l-\varepsilon(J_+-J_-)\;^2 \leq 2l^2 + 2\varepsilon^2 l^2 \leq {}^{5l^2}/{}_2$. Note that $|p_j - q_j| < 1$ for each case, so that by using Lemma A.1 and $l\geq 2$ we obtain for Equation (A.3) that

$$\begin{aligned}&\mathrm{KL}\big(\mathbb{P}_{S'},\mathbb{P}_{S'\cup\{i\}} \\ &\quad\leq \mathbb{E}_{S'}N_i\cdot\Big(\frac{(l-1)8\varepsilon^2}{l^3} + \frac{20\varepsilon^2}{l}\Big) \leq \mathbb{E}_{S'}N_i\cdot\frac{22\varepsilon^2}{l},\end{aligned}$$

which completes the proof. □

Proof of Theorem 4.2 (i). Let φ be some arbitrary algorithm suggesting the subsets $(S_t^{\varphi})_{t\in[T]} \subset \mathbb{A}_{full}$. In the following we define two problem instances characterized by score parameters $\theta^{(1)}, \theta^{(2)} \in \Theta$ such that

$$\inf_{\varphi} \Big\{\mathbb{E}^{\varphi}_{\theta^{(1)}}\big(\mathcal{R}(T) \;+ \mathbb{E}^{\varphi}_{\theta^{(2)}}\big(\mathcal{R}(T) \;\Big\} \geq \check{C}\sqrt{T}, \qquad \text{(A.4)}$$

where the infimum is taken over all terminating algorithms φ for the flexible Pre-bandit problem and $\check{C} > 0$ is a constant similar to C as in the assertion. The proof will be then complete due to

$$\inf_{\varphi} \sup_{\theta\in\Theta} \mathbb{E}_{\theta}^{\varphi}(\mathcal{R}(T)) \\ \geq \frac{1}{2} \inf_{\varphi} \quad \mathbb{E}_{\theta^{(1)}}^{\varphi}\big(\mathcal{R}(T) \ + \mathbb{E}_{\theta^{(2)}}^{\varphi}\big(\mathcal{R}(T) \quad .$$

Thus, we proceed by showing (A.4).

The observation at t under the PL model assumption for the algorithm φ for an instance with score parameter θ is a random sample of $P_{S_t^\varphi,\theta} = P_{S_t^\varphi}$, where

$$P_{S_t^\varphi,\theta}(i) := \begin{array}{ll} \frac{\theta_i}{_{j\in S_t^\varphi}\theta_j}, & i \in S_t^\varphi, \\ 0, & \text{else.} \end{array} \tag{A.5}$$

The probability distribution with respect to φ and θ is denoted by $\mathbb{P}_\theta^\varphi = \mathbb{P}_\theta$ and the corresponding expectation by $\mathbb{E}_\theta^\varphi = \mathbb{E}_\theta$. The regret of φ for a PL model with parameter θ over the time horizon T is

$$\begin{aligned} &\mathbb{E}_\theta^\varphi\big(\mathcal{R}(T) \\ &= \sum_{t=1}^{T} \mathbb{E}_\theta^\varphi\big(\mathrm{U}(S^*) - \mathrm{U}(S_t^\varphi)\big) \\ &= \sum_{S\in\mathbb{A}_{full}} \big(\mathrm{U}(S^*) - \mathrm{U}(S) \ \mathbb{E}_\theta^\varphi(N_S(T)), \end{aligned} \tag{A.6}$$

where $N_S(t) = \sum_{s=1}^{t} 1_{\{S_s^\varphi = S\}}$ denotes the number of times the subset $S \in \mathbb{A}_{full}$ was suggested by φ till time $t \in [T]$. Note that we suppressed here the dependency of S^* on θ in the notation for sake of brevity.

Next, define

$$\begin{aligned} \theta^{(1)} &:= \ \ 1, 1-\varepsilon, \theta_{min}, \ldots, \theta_{min} \ \ , \\ \theta^{(2)} &:= \ \ 1-\varepsilon, 1, \theta_{min}, \ldots, \theta_{min} \ \ , \end{aligned} \tag{A.7}$$

where $\varepsilon \in (0, 1-\theta_{min})$ is a hardness parameter of the instances, which will be specified below. Note that both score parameters are elements of Θ and only differ in two of the n components. It is easy to see that for any $S \in \mathbb{A}_{full}\backslash\{1\}$ and $S' \in \mathbb{A}_{full}\backslash\{2\}$ one has that

$$\begin{aligned} \mathrm{U}(\{1\},\theta^{(1)}) - \mathrm{U}(S,\theta^{(1)}) \geq \min\{1, 1/\gamma\}\,\varepsilon, \\ \mathrm{U}(\{2\},\theta^{(2)}) - \mathrm{U}(S',\theta^{(2)}) \geq \min\{1, 1/\gamma\}\,\varepsilon. \end{aligned} \tag{A.8}$$

Indeed, recall the decomposition of θ in (5) and obtain

$$\begin{aligned} \mathrm{U}(\{1\},\theta^{(1)}) - \mathrm{U}(S,\theta^{(1)}) &\geq 1 - \frac{(1-\varepsilon)^{\frac{1+\gamma}{\gamma}}}{1-\varepsilon} \\ &= 1 - (1-\varepsilon)^{\frac{1}{\gamma}} \\ &\geq \min\{1, 1/\gamma\}\,\varepsilon. \end{aligned}$$

The inequality $\mathrm{U}(\{2\},\theta^{(2)}) - \mathrm{U}(S',\theta^{(2)}) \geq \min\{1, 1/\gamma\}\,\varepsilon$ can be shown similarly. Clearly, the optimal subset to suggest for the problem instance characterized by $\theta^{(1)}$ is $\{1\}$, while $\{2\}$ is optimal for the other scenario associated with $\theta^{(2)}$. Suggesting other subsets respectively results in an at least linear regret in the hardness parameter ε. By means of representation (A.6) and (A.8) it follows that for $i = 1, 2$

$$\mathbb{E}_{\theta^{(i)}}^{\varphi}\big(\mathcal{R}(T) \\ > \mathbb{P}_{\theta^{(i)}}^{\varphi}\big(N_{\{1\}}(T) \leq T/2 \quad \frac{\min\{1,1/\gamma\}\,\varepsilon T}{2}.$$

The inequalities are intuitive: if the optimal set $\{1\}$ for the parameter $\theta^{(1)}$ is suggested at most $T/2$ times, then one obtains a regret of at least ε for the suggested sets in the remaining cases, which occur at least $T/2$ times. Similarly, if the suboptimal set $\{1\}$ for the problem instance with $\theta^{(2)}$ is suggested at least $T/2$ times, then one obtains a regret of at least ε in all these timesteps. The latter display implies

$$\begin{aligned} &\mathbb{E}_{\theta^{(1)}}^{\varphi}\big(\mathcal{R}(T) \ + \mathbb{E}_{\theta^{(2)}}^{\varphi}\big(\mathcal{R}(T) \\ &> \frac{\min\{1,1/\gamma\}\,\varepsilon T}{2} \ \mathbb{P}_{\theta^{(1)}}^{\varphi}\big(N_{\{1\}}(T) \leq T/2 \\ &\qquad\qquad + \mathbb{P}_{\theta^{(2)}}^{\varphi}\big(N_{\{1\}}(T) > T/2 \\ &\geq \frac{\min\{1,1/\gamma\}\,\varepsilon T}{2} \exp \ \ - \mathrm{KL}\big(\mathbb{P}_{\theta^{(1)}}^{\varphi}, \mathbb{P}_{\theta^{(2)}} \ \ ^{\varphi} \ , \end{aligned}$$

where we used in the last line a version of Pinkser's inequality, see Theorem 14.2 in Lattimore & Szepesvári (2020). We proceed by analyzing the Kullback-Leibler distance in the latter display by means of Lemma A.1 and the following decomposition of the Kullback-Leibler divergence for the family of probability distributions $(\mathbb{P}_\theta^\varphi)_{\theta\in\Theta}$ which can be shown analogously to Lemma 15.1 in Lattimore & Szepesvári (2020).

Lemma A.4. *Let* $\theta, \theta' \in \Theta$, *then*

$$\mathrm{KL}\big(\mathbb{P}_\theta^\varphi, \mathbb{P}_{\theta'}^\varphi \ = \sum_{S\in\mathbb{A}_{full}} \mathbb{E}_\theta(N_S(T))\,\mathrm{KL}\big(P_{S,\theta}, P_{S,\theta'} \ .$$

Note that by definition of the score parameters in (A.7) it holds that $\mathrm{KL}\big(P_{S,\theta^{(1)}}, P_{S,\theta^{(2)}} \ = 0$ for any subset $S \in \mathbb{A}_{full}$ which does not contain $\{1\}$ and $\{2\}$, as both distributions are the same for such subsets. For the remaining subsets S', which are of order $\mathcal{O}(2^{n-2})$ many, Lemma A.1 yields $\mathrm{KL}\big(P_{S',\theta^{(1)}}, P_{S',\theta^{(2)}} \ \leq 2\theta_{min}^{-1}\varepsilon^2$ (cf. the proof of Lemma A.3). We distinguish two cases in the following.

Case 1: $T > 2^n - 1$.

As $\sum_{S\in\mathbb{A}_{full}} \mathbb{E}_\theta(N_S(T)) = T$ for any $\theta \in \Theta$ it is true that $\mathbb{E}_\theta(N_S(T)) \leq {}^{T}/_{2^n-1}$ for each $S \in \mathbb{A}_{full}$ by the pigeonhole principle. Thus, by means of Lemma A.4 obtain $\mathrm{KL}\big(\mathbb{P}_{\theta^{(1)}}, \mathbb{P}_{\theta^{(2)}} \ \leq \widetilde{C}\,T\varepsilon^2$, where $\widetilde{C} > 0$ is some constant independent of n and T. Hence,

$$\mathbb{E}_{\theta^{(1)}}^{\varphi}\big(\mathcal{R}(T) \ + \mathbb{E}_{\theta^{(2)}}^{\varphi}\big(\mathcal{R}(T)$$

$$\geq \frac{\min\{1,1/\gamma\}\,\varepsilon T}{2}\exp\quad -\widetilde{C}T\varepsilon^2\quad .$$

Case 2: $T \leq 2^n-1$.

In this case, note that there are at least 2^n-1-T many zero summands in $\sum_{S\in\mathbb{A}_{full}}\mathbb{E}_\theta(N_S(T))$ as the sum equals T. Therefore, similar to the case before obtain by means of Lemma A.4 that $\mathrm{KL}\big(\mathbb{P}_{\theta^{(1)}},\mathbb{P}_{\theta^{(2)}}\quad\leq \widetilde{C}T\varepsilon^2$ for some constant $\widetilde{C}>0$ independent of n and T. Consequently,

$$\begin{aligned}&\mathbb{E}^{\varphi}_{\theta^{(1)}}\big(\mathcal{R}(T)\quad+\mathbb{E}^{\varphi}_{\theta^{(2)}}\big(\mathcal{R}(T)\\&\quad\geq \frac{\min\{1,1/\gamma\}\,\varepsilon T}{2}\exp\quad -\widetilde{C}T\varepsilon^2\quad .\end{aligned}$$

By choosing in both cases $\varepsilon=\min(\bar{C}\sqrt{1/T},\,1-\theta_{min})$ for some appropriate constant $\bar{C}>0$ we obtain the assertion with some constants $C,C'>0$ which are independent of T,l and n. □

Proof of Theorem 4.2 (ii). For the gap-dependent lower bound we will make use of the following result, which is Lemma 1 in Kaufmann et al. (2016).

Lemma A.5. *Let ν and ν' be two MAB models with n arms and ν_i resp. ν_i' denotes the reward distribution for arm $i\in[n]$ respectively. Let A_t denote the arm played at round t and R_t be the corresponding observed reward. Moreover, let $\mathcal{F}_t=\sigma(A_1,R_1,\ldots,A_t,R_t)$ be the sigma algebra generated by the observations till time instance t. Suppose that ν_i and ν_i' are mutually absolutely continuous for each $i\in[n]$, then it holds that*

$$\sum_{i\in[n]}\mathbb{E}_\nu[N_i(T)]\mathrm{KL}\big(\nu_i,\nu_i'\quad\geq d(\mathbb{E}_\nu(\mathcal{E}),\mathbb{E}_{\nu'}(\mathcal{E}))$$

for any $\mathcal{F}_T$-measurable random variable $\mathcal{E}$. Here, $d(x,y)=x\log(x/y)+(1-x)\log((1-x)/(1-y))$ and $N_i(t)=\sum_{s=1}^t 1_{i_s^\varphi=i}$ is the number of times an algorithm φ plays arm i till time instance t.

In the following, we will adapt the proof of Theorem 3 in (Saha & Gopalan, 2019b) to our case, which boils down to incorporating our (different) notion of regret into their proof.

To make use of Lemma A.5 we embed the flexible Pre-Bandit problem into a classical MAB problem by considering each subset $S\in\mathbb{A}_{full}$ as an arm. Moreover, we define the score parameters

$$\begin{aligned}&\theta^{(1)}=(1,1-\Delta,\ldots,1-\Delta),\\&\theta^{(i)}=\big(1,1-\Delta,\ldots,1-\Delta,1+\varepsilon,1-\Delta,\ldots,1-\Delta\quad,\\&\qquad i=2,\ldots,n,\end{aligned}\tag{A.9}$$

where $\Delta\in(0,1-\theta_{min})$ and $\varepsilon>0$ and the i-th component of $\theta^{(i)}$ is $1+\varepsilon$. For $\theta\in\Theta$ and $S\in\mathbb{A}_{full}$ let $P_{S,\theta}$ denote the categorical distribution as in (A.5). Using Lemma A.5 with $\nu_S=P_{S,\theta^{(1)}}$ and $\nu_S'=P_{S,\theta^{(i)}}$ for $i=1$ for any $S\in\mathbb{A}_{full}$ as the reward distributions of the arms and the $\mathcal{F}_T$-measurable random variable $\mathcal{E}=N_{\{i\}}(T)/T$, one has that

$$\begin{aligned}&\sum_{S\in\mathbb{A}_{full}}\mathbb{E}_{\theta^{(1)}}[N_S(T)]\,\mathrm{KL}\big(P_{S,\theta^{(1)}},P_{S,\theta^{(i)}}\\&\qquad=\sum_{S\in\mathbb{A}_{full}}\mathbb{E}_{\theta^{(1)}}[N_S(T)]\,\mathrm{KL}\big(\nu_S,\nu_S'\qquad\qquad(\text{A.10})\\&\qquad\geq d(\mathbb{E}_{\theta^{(1)}}[N_{\{i\}}(T)/T],\mathbb{E}_{\theta^{(i)}}[N_{\{i\}}(T)/T]).\end{aligned}$$

Now, since $d(x,y)\geq(1-x)\log(1/(1-y))-\log(2)$ derive that

$$\begin{aligned}&d(\mathbb{E}_{\theta^{(1)}}[N_{\{i\}}(T)/T],\mathbb{E}_{\theta^{(i)}}[N_{\{i\}}(T)/T])\\&\quad\geq\quad 1-\frac{\mathbb{E}_{\theta^{(1)}}[N_{\{i\}}]}{T}\quad\log\quad\frac{T}{T-\mathbb{E}_{\theta^{(i)}}[N_{\{i\}}]}\quad-\log(2).\end{aligned}$$

As we assume that φ is a no-regret algorithm, we have that $\mathbb{E}_{\theta^{(1)}}[N_{\{i\}}]=o(T^\alpha)$ and $T-\mathbb{E}_{\theta^{(i)}}[N_{\{i\}}]=\mathbb{E}_{\theta^{(i)}}[\sum_{S\in\mathbb{A}_{full},S\neq\{i\}}N_{\{i\}}]=o(T^\alpha)$ for some $\alpha\in(0,1]$. Hence, by dividing the latter display by $\log(T)$ and by considering $T\to\infty$ one obtains

$$\begin{aligned}&\lim_{T\to\infty}\frac{d(\mathbb{E}_{\theta^{(1)}}[N_{\{i\}}(T)/T],\mathbb{E}_{\theta^{(i)}}[N_{\{i\}}(T)/T])}{\log(T)}\\&\geq\lim_{T\to\infty}\frac{1}{\log(T)}\quad 1-o(T^{\alpha-1})\quad\log\quad\frac{T}{o(T^\alpha)}\quad-\frac{\log(2)}{\log(T)}\\&\geq(1-\alpha).\end{aligned}$$

Hence, dividing (A.10) by $\log(T)$ and considering the limit case obtain

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{\log(T)}\sum_{S\in\mathbb{A}_{full}}\mathbb{E}_{\theta^{(1)}}[N_S(T)]\,\mathrm{KL}\big(P_{S,\theta^{(1)}},P_{S,\theta^{(i)}}\\&\qquad\qquad\geq(1-\alpha).\end{aligned}\tag{A.11}$$

The Kullback-Leibler divergence in (A.11) can be bounded by the following lemma, which first statement can be shown by following the lines of display (2) in Saha & Gopalan (2019b), while the second statement is straightforward from the choice of the score parameters in (A.9).

Lemma A.6. *For each $i=1$ it holds that*

$$\mathrm{KL}\big(P_{S,\theta^{(1)}},P_{S,\theta^{(i)}}\quad\leq\frac{(\Delta+\varepsilon)^2}{(1-\Delta)|S|(1+\varepsilon)}.$$

Moreover, if $i\notin S$ or if $|S|=1$, then

$$\mathrm{KL}\big(P_{S,\theta^{(1)}},P_{S,\theta^{(i)}}\quad=0.$$

Using Lemma A.6 we can derive from (A.11) by multiplying with ${}^{(1-\Delta)^2}/_{(\Delta+\varepsilon)}$ that

$$\lim_{T\to\infty} \frac{1}{\log(T)} \sum_{\substack{S\in\mathbb{A}_{full}\setminus\{i\},\\ i\in S}} \frac{\mathbb{E}_{\theta^{(1)}}[N_S(T)](1-\Delta)(\Delta+\varepsilon)}{|S|(1+\varepsilon)} \geq \frac{(1-\Delta)^2}{(\Delta+\varepsilon)}\,(1-\alpha).$$

Summing over $i \in \{2,\dots,n\}$ and taking the limit $\varepsilon \to 0$ in the latter display leads to

$$\lim_{T\to\infty} \frac{1}{\log(T)} \sum_{i=2}^{n} \sum_{\substack{S\in\mathbb{A}_{full}\setminus\{i\},\\ i\in S}} \mathbb{E}_{\theta^{(1)}}[N_S(T)]\,\frac{(1-\Delta)\Delta}{|S|} \geq \frac{(1-\Delta)^2}{\Delta}\,(n-1)\,(1-\alpha). \tag{A.12}$$

Next, we bound the cumulative regret in (9) for any algorithm φ for the flexible Pre-Bandit problem from below. For this purpose recall the decomposition in (5) and denote the ith component of $\theta^{(1)}$ by $\theta_i^{(1)}$ and let $v_i^{(1)} = (\theta_i^{(1)})^{1/\gamma}$. Hence, we get

$$\begin{aligned}
&\mathbb{E}_{\theta^{(1)}}\big(\mathcal{R}(T)\\
&= \sum_{t=1}^{T} \mathbb{E}_{\theta^{(1)}}\big(\mathrm{U}(S^*) - \mathrm{U}(S_t^{\varphi})\big)\\
&= \sum_{t=1}^{T} \mathbb{E}_{\theta^{(1)}} \quad v_1^{(1)} - \frac{\sum_{i\in S_t^{\varphi}} \big(v_i^{(1)} \;\; {}^{1+\gamma}}{\sum_{i\in S_t^{\varphi}} \big(v_i^{(1)} \;\; {}^{\gamma}}\\
&= \mathbb{E}_{\theta^{(1)}} \quad \sum_{t=1}^{T} \sum_{S\in\mathbb{A}_{full}} 1_{S_t^{\varphi}=S}\\
&\qquad\qquad \frac{\sum_{i=2}^{n} 1_{i\in S}\,\big(v_i^{(1)} \;\; {}^{\gamma}\big(v_1^{(1)} - v_i^{(1)}\big)}{\sum_{i=1}^{n} 1_{i\in S}\big(v_i^{(1)} \;\; {}^{\gamma}}\\
&\geq \min\{1, 1/\gamma\}\\
&\quad \times \mathbb{E}_{\theta^{(1)}} \quad \sum_{t=1}^{T} \sum_{S\in\mathbb{A}_{full}} 1_{S_t^{\varphi}=S} \sum_{i=2}^{n} \frac{1_{i\in S}\,(1-\Delta)\Delta}{|S|}\\
&= \min\{1, 1/\gamma\}\\
&\quad \times \sum_{i=2}^{n} \sum_{S\in\mathbb{A}_{full}} \mathbb{E}_{\theta^{(1)}} \quad \sum_{t=1}^{T} 1_{S_t^{\varphi}=S} \quad 1_{i\in S}\,\frac{(1-\Delta)\Delta}{|S|}\\
&= \min\{1, 1/\gamma\}\\
&\quad \times \sum_{i=2}^{n} \sum_{S\in\mathbb{A}_{full},\, i\in S} \mathbb{E}_{\theta^{(1)}}(N_S[T])\,\frac{(1-\Delta)\Delta}{|S|},
\end{aligned}$$

where we used Lemma A.2 for the inequality together with $\sum_{i=1}^{n} 1_{i\in S}\big(v_i^{(1)} \;\; {}^{\gamma} \leq |S|$. With this obtain from (A.12) that if φ is a no-regret algorithm, then

$$\lim_{T\to\infty} \frac{1}{\log(T)} \mathbb{E}_{\theta^{(1)}}\big(\mathcal{R}(T) \geq \frac{\min\{1, 1/\gamma\}\cdot(1-\alpha)(1-\Delta)^2}{\Delta}\,(n-1),$$

which concludes the proof as Δ corresponds to $\min_{i\notin S^*} \theta_{max} - \theta_i$ for $\theta = \theta^{(1)}$ and $(1-\alpha)(1-\Delta)^2$ is some constant independent of T and n. □

B. Proof of Theorem 5.1

We start by introducing the notation for the rest of the proof and recalling the main terms of the TRCB algorithm. Thereafter we give an outline of the proof, before deriving the details.

B.1. Notation and relevant terms

Throughout $(S_t)_{t=1,\dots,T}$ denotes the suggested subsets (the preselections) of the TRCB algorithm at each time instance respectively and $(i_t)_{t=1,\dots,T}$ the corresponding decisions of the selector, i.e., $i_t \in S_t$. Furthermore, let $\gamma \in (0,\infty)$ be some arbitrary degree of preciseness. Next, we clarify the notation as well as recall the main terms emerging in the TRCB algorithm. We define

$$w_{i,j}(t) := \begin{matrix} \sum_{s=1}^{t-1} 1_{\{i_s=i,\,\{i,j\}\in S_s\}}, & t>1,\\ 0, & t=1, \end{matrix} \tag{B.13}$$

to denote the number of times i has been picked by the selector till time instance t, when i and j were both part of the preselection, while $\overline{w}_{i,j}(t) := w_{i,j}(t) + w_{j,i}(t)$ is the number of times either i or j was picked till time instance t, when both were part of the preselection. The relative scores in (4) are estimated in time instance t by

$$\hat{O}_{i,j}(t) := \begin{matrix} \frac{\overline{w}_{i,j}(t)}{w_{j,i}(t)} - 1, & w_{j,i}(t) = 0,\\ \theta_{min}, & \text{else}, \end{matrix} \qquad i,j \in [n]. \tag{B.14}$$

The arm with the most picks till time instance t is

$$J := J(t) = \underset{i\in[n]}{\arg\max}\, \#\{w_{i,j}(t) \geq w_{j,i}(t) \mid j = i\}. \tag{B.15}$$

Note that in the following we will suppress its dependency on the time instance t in the notation. The (thresholded) random value inside the confidence region of $\hat{O}_{i,J}(t)$ is

$$\hat{O}_{i,J}^{\mathrm{TRCB}}(t) = \theta_{min}^{-1} \wedge \quad \big(\hat{O}_{i,J} + C_{shrink}\,\beta_i(t) \quad \vee \theta_{min} \quad ,$$

for $i = J$ and $\hat{O}_{i,J}^{\mathrm{TRCB}}(t) = 1$ for $i = J$, where

$$\beta_i(t) \sim \mathrm{Unif}[-c_{i,J}(t), c_{i,J}(t)],$$

$$c_{i,J}(t) = \sqrt{\frac{32\,\log(l\,t^{3/2})}{\theta_{min}^4\,\overline{w}_{i,J}(t)}},$$

and $C_{shrink} \in (0, 1/2)$ is some finite constant. Note that the β_i's are mutually independent. Recall the definition of regret for any time instance $t \in [T]$ in (9). Due to (10) we will consider the following scaled regret per time

$$\tilde{r}(t) := \widetilde{\mathrm{U}}(S^*; O_J, \gamma) - \widetilde{\mathrm{U}}(S_t; O_J, \gamma) = v_J^{-1}\, r(S_t). \tag{B.16}$$

Finally, let $\mathcal{F}_t$ denote the σ-algebra generated by $S_1, i_1, \ldots, S_{t-1}, i_{t-1}$ in time instance t, with $\mathcal{F}_1$ being the trivial σ-algebra. Note that $J(t)$ as well as $\overline{w}_{i,J}(t)$ resp. $c_{i,J}$ are $\mathcal{F}_t$-measurable for any $t \in [T]$.

B.2. Outline of the proof

We introduce in the following the core lemmas to prove the main result, which will be gradually verified in the next subsection. For $t \in [T]$ define

$$A_t := \{\exists i \in S_t \cup S^* : \, |\hat{O}_{i,J}^{\mathrm{TRCB}}(t) - O_{i,J}| > c_{i,J}(t)\}. \tag{B.17}$$

Thus, A_t is the event on which the estimates for the relative scores for arms in the chosen preselection and the optimal preselection with respect to the currently most winning arm J are not close enough to their actual relative score, where the length of the confidence region $c_{i,J}(t)$ determines how closeness is to be understood in this case.
As a consequence, one wishes that the probability that A_t happens is sufficiently small. The following lemma establishes this requirement.

Lemma B.1. *It holds that*

$$\mathbb{E}_\theta\big(1_{\{A_t\}}\,|\mathcal{F}_t \; = \mathcal{O} \quad \overline{\frac{\log(t)}{t}} \;\;,$$

where the constant in the $\mathcal{O}$-term is independent of T, l and n. In particular, for any $i \in S_t \cup S^*$,

$$\mathbb{E}_\theta\Big[\mathbb{E}_\theta\big(|\hat{O}_{i,J}^{\mathrm{TRCB}}(t) - O_{i,J}|\,1_{A_t^\complement}\,|\mathcal{F}_t \;\Big] \le \mathbb{E}_\theta \;\, c_{i,J}(t) \;\, = \mathbb{E}_\theta\Big[\sqrt{\frac{32\,\log(l\,t^{3/2})}{\theta_{min}^4\,\overline{w}_{i,J}(t)}}\Big].$$

Next, we investigate the deviation between the scaled regret per time (cf. (B.16)) and its empirical counterpart. For this purpose, note that

$$\begin{aligned}\tilde{r}(t) &= \widetilde{\mathrm{U}}(S^*; O_J, \gamma) - \widetilde{\mathrm{U}}(S_t; O_J, \gamma) \\ &\le \;\; \widetilde{\mathrm{U}}(S^*; O_J, \gamma) - \widetilde{\mathrm{U}}(S^*; \hat{O}_J^{\mathrm{TRCB}}, \gamma) \\ &\quad + \;\; \widetilde{\mathrm{U}}(S_t; \hat{O}_J^{\mathrm{TRCB}}, \gamma) - \widetilde{\mathrm{U}}(S_t; O_J, \gamma) \;\;,\end{aligned} \tag{B.18}$$

since $\widetilde{\mathrm{U}}(S^*; \hat{O}_J^{\mathrm{TRCB}}) - \widetilde{\mathrm{U}}(S_t; \hat{O}_J^{\mathrm{TRCB}}) \le 0$, by the definition of S_t in line 11 of the TRCB algorithm. Here, we abbreviated $\hat{O}_J^{\mathrm{TRCB}} = (\hat{O}_{1,J}^{\mathrm{TRCB}}, \ldots, \hat{O}_{n,J}^{\mathrm{TRCB}})$.

The following lemma gives a bound on the ratio between the two terms in squared brackets on the right-hand side of the latter display.

Lemma B.2. *Conditioned on $\mathcal{F}_t$ there exist constants $C_1, C_2 > 0$ depending if at all on θ_{min} and γ (but independent of T, l and n) such that on $A_t^\complement$ it holds with probability at least $1 - \frac{C_1}{\sqrt{t}}$ that*

$$\frac{\widetilde{\mathrm{U}}(S^*; O_J, \gamma) - \widetilde{\mathrm{U}}(S^*; \hat{O}_J^{\mathrm{TRCB}}, \gamma)}{\widetilde{\mathrm{U}}(S_t; \hat{O}_J^{\mathrm{TRCB}}, \gamma) - \widetilde{\mathrm{U}}(S_t; O_J, \gamma)} \le C_2.$$

Moreover, C_2 is of the form $\mathrm{const} \cdot \theta_{min}^{-2(3+\gamma)}$. In particular, with probability at least $1 - \frac{C_1}{\sqrt{t}}$

$$\begin{aligned}&\mathbb{E}_\theta\big(\tilde{r}(t) 1_{A_t^\complement}\,|\mathcal{F}_t \\ &\quad \le (C_2 + 1)\mathbb{E}_\theta\big(\big|\widetilde{\mathrm{U}}(S_t; \hat{O}_J^{\mathrm{TRCB}}) - \widetilde{\mathrm{U}}(S_t; O_J)\big|\,1_{A_t^\complement}\,|\mathcal{F}_t \;\; .\end{aligned}$$

The next pillar of the proof is to transfer the high concentration of $\hat{O}_J^{\mathrm{TRCB}}$ around O_J to a high concentration of the corresponding utilities $\widetilde{\mathrm{U}}$ by exploiting its Lipschitz smoothness.

Lemma B.3. *For any $t \in [T]$*

$$\begin{aligned}&\big|\widetilde{\mathrm{U}}(S_t; \hat{O}_J^{\mathrm{TRCB}}, \gamma) - \widetilde{\mathrm{U}}(S_t; O_J, \gamma)\big| \\ &\quad \le \frac{\max\{\theta_{min}^{(\gamma-1)/(\gamma)}, \theta_{min}^{(1-\gamma)/(\gamma)}\}}{\gamma} \sum_{i \in S_t} |\hat{O}_{i,J}^{\mathrm{TRCB}}(t) - O_{i,J}|.\end{aligned}$$

Finally, an upper bound on the expected length of the confidence regions over time (that is basically $(\overline{w}_{i,J}(t))^{-1/2}$) has to be verified.

Lemma B.4. *The following statement is valid,*

$$\sum_{t \in T} \mathbb{E}_\theta\big(\sum_{i \in S_t} {}^1\!/\!\sqrt{\overline{w}_{i,J}(t)} \;\; \le 4\sqrt{Tn}.$$

Conclusion: Proof of Theorem 5.1 Given these core lemmas, we are now in the position to verify Theorem 5.1.

Let $\theta \in \Theta$ and $T \in \mathbb{N}$ with $T > n$, then since $r(S_t) \le \tilde{r}(t)$, for any $t \in [T]$, we have

$$\mathbb{E}_\theta[\mathcal{R}(T)] \le \sum_{t=1}^{T} \mathbb{E}_\theta\big(\mathbb{E}(\tilde{r}(t)|\mathcal{F}_t) \;\; ,$$

where we used the tower property of the conditional expected value. Note that $\tilde{r} \le 1/\theta_{min}$ such that by applying

Lemma B.2, Lemma B.1 and then Lemma B.3, one can derive that

$$
\begin{aligned}
&\mathbb{E}_\theta[\mathcal{R}(T)] \\
&\leq \sum_{t=1}^{T} \ \mathbb{E}_\theta\big(\mathbb{E}(\tilde{r}(t)1_{A_t}|\mathcal{F}_t) \ + \mathbb{E}_\theta \sum_{i\in S_t} \big(\mathbb{E}(\tilde{r}(t)1_{A_t^{\complement}}|\mathcal{F}_t) \\
&\leq \sum_{t=1}^{T} \ \mathbb{E}_\theta\big(\mathbb{E}(\tilde{r}(t)1_{A_t}|\mathcal{F}_t) \ + C_0 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \\
&\qquad + C_1 \mathbb{E}_\theta\big(\sum_{i\in S_t} \mathbb{E}(|\hat{O}^{\mathrm{TRCB}}_{i,J}(t) - O_{i,J}| \, 1_{A_t^{\complement}} |\mathcal{F}_t) \\
&\leq C_2 \sum_{t=1}^{T} \ \overline{\frac{\log(t)}{t}} \\
&\qquad + C_1 \sum_{t=1}^{T} \mathbb{E}_\theta\big(\sum_{i\in S_t} \mathbb{E}(|\hat{O}^{\mathrm{TRCB}}_{i,J}(t) - O_{i,J}| \, 1_{A_t^{\complement}} |\mathcal{F}_t) \\
&\leq C_2 \sum_{t=1}^{T} \ \overline{\frac{\log(t)}{t}} + C_3 \sum_{t=1}^{T} \mathbb{E}_\theta\big(\sum_{i\in S_t} \sqrt{\log(l\cdot t)/\overline{w}_{i,J}(t)} \ ,
\end{aligned}
$$

where $C_i > 0$, for $i \in \{0,1,2,3\}$, are constants depending if at all only on θ_{min} and γ, but are independent of T, l and n. Next, since $\sum_{t=1}^{T} t^{-1/2} \leq 2\sqrt{T}$ and $\log(l \cdot t) \leq 2\log(T)$, due to $l \leq n < T$, we can further estimate the right-hand side of the latter display to obtain

$$
\begin{aligned}
\mathbb{E}_\theta[\mathcal{R}(T)] &\leq C_4 \sqrt{T \, \log(T)} \\
&\qquad + C_5 \sqrt{\log(T)} \sum_{t=1}^{T} \mathbb{E}_\theta \sum_{i\in S_t} \sqrt{1/\overline{w}_{i,J}(t)} \\
&\leq C_4 \sqrt{T \, \log(T)} + C_6 \sqrt{\log(T) \, T \, n}
\end{aligned}
$$

where we used Lemma B.4 for the second last inequality. Here, the constants $C_4, C_5, C_6 > 0$ are as before depending (if at all) on θ_{min} and γ, but are independent of T, l and n. In particular, we have C_4 is of the form $\mathrm{const} \cdot \theta_{min}^{-1}$, while C_6 is of the form

$$
\mathrm{const} \cdot \frac{\max\{\theta_{min}^{(\gamma-1)/(\gamma)}, \theta_{min}^{(1-\gamma)/(\gamma)}\}}{\gamma} \cdot \theta_{min}^{-2(3+\gamma)}.
$$

This concludes the proof.

B.3. Proofs of the core lemmas in Subsection B.2

We start with the proof of Lemma B.1. For this we need the following result, which is Lemma 1 in Saha & Gopalan (2019a).

Lemma B.5. *It holds that for any $r \in \mathbb{N}, i, j \in [n]$ and $\varepsilon > 0$ that*

$$
\begin{aligned}
&\mathbb{P} \ \ \frac{w_{i,j}(t)}{\overline{w}_{i,j}(t)} - \frac{\theta_i}{\theta_i + \theta_j} \ \geq \varepsilon, \, \overline{w}_{i,j}(t) = r \\
&\leq \mathbb{P} \ \ \frac{w_{i,j}(t)}{\overline{w}_{i,j}(t)} - \frac{\theta_i}{\theta_i + \theta_j} \ \geq \varepsilon, \, \overline{w}_{i,j}(t) \geq r \\
&\leq 2\exp(-2\, r\, \varepsilon^2).
\end{aligned}
$$

Proof of Lemma B.1. Define the function $\phi(x) = x^{-1} - 1$, then note that $\phi\big(\frac{w_{j,i}(t)}{\overline{w}_{i,j}(t)} \ = \hat{O}_{i,j}(t)$ and $\phi\big(\frac{\theta_j}{\theta_i+\theta_j} \ = O_{i,j}$. Further, by the mean value theorem there exists for any pair of arms (i,j) some $\tilde{z}_{i,j}$ between $\frac{w_{j,i}(t)}{\overline{w}_{i,j}(t)}$ and $\frac{\theta_j}{\theta_i+\theta_j}$ such that

$$
\begin{aligned}
\hat{O}_{i,j}(t) - O_{i,j} &= \phi \ \ \frac{w_{j,i}(t)}{\overline{w}_{i,j}(t)} \ \ - \phi \ \ \frac{\theta_j}{\theta_i + \theta_j} \\
&= \phi'(\tilde{z}_{i,j}) \ \ \frac{w_{j,i}(t)}{\overline{w}_{i,j}(t)} - \frac{\theta_j}{\theta_i + \theta_j} \\
&= -\frac{1}{\tilde{z}_{i,j}^2} \ \ \frac{w_{j,i}(t)}{\overline{w}_{i,j}(t)} - \frac{\theta_j}{\theta_i + \theta_j} \ \ .
\end{aligned}
$$

Note that

$$
\begin{aligned}
\tilde{z}_{i,j} &\geq \min\big(w_{j,i}(t)/\overline{w}_{i,j}(t), \theta_j/\theta_i+\theta_j\big) \\
&\geq \min\big(w_{j,i}(t)/\overline{w}_{i,j}(t), \theta_{min}/2\big)
\end{aligned}
$$

and in particular if $j = J$ then

$$
\tilde{z}_{i,J} \geq \min\big(1/2, \theta_{min}/2\big) = \theta_{min}/2,
$$

as $\overline{w}_{i,J} \leq 2 w_{J,i}$ by definition of J and $\theta_{min} < 1$. Let us write $E_{i,J}(t) = \ \hat{O}_{i,J}(t) - O_{i,J}$ for sake of brevity, then we get with the deviation above for $\varepsilon > 0$ for any $t \in [2, T] \cap \mathbb{N}$ that

$$
\begin{aligned}
&\mathbb{P} \ \ \Big\{E_{i,J}(t) \geq \varepsilon/\sqrt{\overline{w}_{i,J}(t)}\Big\} \\
&\quad \leq \sum_{r=1}^{t-1} \mathbb{P} \qquad \frac{w_{J,i}(t)}{\overline{w}_{i,J}(t)} - \frac{\theta_J}{\theta_i + \theta_J} \ \ \geq \frac{\theta_{min}^2 \, \varepsilon}{4\sqrt{\overline{w}_{i,J}(t)}} \\
&\qquad\qquad\qquad \cap \{\overline{w}_{i,J}(t) = r\} \\
&\quad = \sum_{r=1}^{t-1} \mathbb{P} \qquad \frac{w_{J,i}(t)}{\overline{w}_{i,J}(t)} - \frac{\theta_J}{\theta_i + \theta_J} \ \ \geq \frac{\theta_{min}^2 \, \varepsilon}{4\sqrt{r}} \\
&\qquad\qquad\qquad \cap \{\overline{w}_{i,J}(t) = r\} \\
&\quad \leq 2(t-1)\exp \ \ - \frac{\theta_{min}^4 \varepsilon^2}{8} \ \ ,
\end{aligned}
$$

where Lemma B.5 was used in the last step. Setting $\varepsilon = \sqrt{8\ \log(l\, t^{3/2})/\theta_{min}^4}$ in the last display, we obtain in combination with the law of total expectation that conditioned on $\mathcal{F}_t$ that

$$
\begin{aligned}
&\mathbb{P}\big(A_t \\
&\quad \leq \sum_{i\in S_t \cup S^*} \int_{[-c_{i,J}(t), c_{i,J}(t)]} (2c_{i,J}(t))^{-1} \\
&\qquad\quad \times \mathbb{P} \ \ \big\{E_{i,J}(t) \geq c_{i,J}(t) - C_{shrink}\, y\big\} \ \ dy
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{i \in S_t \cup S^*} \mathbb{P} \quad E_{i,J}(t) \geq (1 - C_{shrink})\, c_{i,J}(t) \\
&\leq 4\, l\, (t-1) \exp\big(- {}^{\theta_{min}^4 \varepsilon^2}\!/_8 \;\; = \mathcal{O}(\sqrt{\log(t)}/\sqrt{t}),
\end{aligned}
$$

where we used that thresholding of the relative scores only makes the probability of the event smaller for the first inequality and in the second last step that, firstly, $C_{shrink} \leq 1/2$ in combination with $1/2\, c_{i,J}(t) = {}^{\varepsilon}\!/\sqrt{\overline{w}_{i,J}(t)}$ and secondly, that $|S_t \cup S^*| \leq 2l$. The constant in the $\mathcal{O}$-term is independent of l, T and n. This concludes the lemma. □

Proof of Lemma B.2. Let us write $S_2^*(O) = \sum_{i \in S^*} O_{i,J}^{\frac{1+\gamma}{\gamma}}$ and $S_1^*(O) = \sum_{i \in S^*} O_{i,J}$. In the same spirit define $S_2^*(\hat{O})$, $S_1^*(\hat{O})$, $S_2^t(O)$, $S_1^t(O)$, $S_2^t(\hat{O})$ and $S_1^t(\hat{O})$, where $\hat{O}$ is short for $\hat{O}_J^{\mathrm{TRCB}}$. Then,

$$
\begin{aligned}
\frac{\widetilde{\mathrm{U}}(S^*; O_J, \gamma) - \widetilde{\mathrm{U}}(S^*; \hat{O}_J^{\mathrm{TRCB}}, \gamma)}{\widetilde{\mathrm{U}}(S_t; \hat{O}_J^{\mathrm{TRCB}}, \gamma) - \widetilde{\mathrm{U}}(S_t; O_J, \gamma)} &= \frac{\frac{S_2^*(O)}{S_1^*(O)} - \frac{S_2^*(\hat{O})}{S_1^*(\hat{O})}}{\frac{S_2^t(\hat{O})}{S_1^t(\hat{O})} - \frac{S_2^t(O)}{S_1^t(O)}} \\
&= \frac{\frac{[S_2^*(O) - S_2^*(\hat{O})]}{S_1^*(O)} + \frac{S_2^*(\hat{O})[S_1^*(\hat{O}) - S_1^*(O)]}{S_1^*(O) S_1^*(\hat{O})}}{\frac{[S_2^t(\hat{O}) - S_2^t(O)]}{S_1^t(\hat{O})} + \frac{S_2^t(O)[S_1^t(O) - S_1^t(\hat{O})]}{S_1^t(\hat{O}) S_1^t(O)}}.
\end{aligned} \tag{B.19}
$$

It holds that

$$
\begin{aligned}
{}^{\theta_{min}}\!/_{l} &\leq \frac{1}{S_1^*(O)} \leq {}^{1}\!/_{\theta_{min} l}, \\
{}^{\theta_{min}^{3+\gamma}}\!/_{l} &\leq \frac{S_2^*(\hat{O})}{S_1^*(O) S_1^*(\hat{O})} \leq {}^{1}\!/_{\theta_{min}^{3+\gamma} l}, \\
{}^{\theta_{min}}\!/_{l} &\leq \frac{1}{S_1^t(\hat{O})} \leq {}^{1}\!/_{\theta_{min} l}, \\
{}^{\theta_{min}^{3+\gamma}}\!/_{l} &\leq \frac{S_2^t(O)}{S_1^t(\hat{O}) S_1^t(O)} \leq {}^{1}\!/_{\theta_{min}^{3+\gamma} l}.
\end{aligned}
$$

Hence, all of the latter terms can be bounded from below resp. above by $\tilde{C}_j/l$ for some suitable constants C_j which depend if at all on θ_{min}. Following the lines of proof of Lemma B.1, it can be shown that there exists a constant $C_1 > 0$ (depending on θ_{min} and γ) such that the ratios of the terms in the squared brackets in (B.19) are bounded by some constant $C_2 > 0$ on the event $A_t^{\complement}$, with probability at least $1 - \frac{C_1}{\sqrt{t}}$. Hence, the whole term in (B.19) can be bounded with probability at least $1 - \frac{C_1}{\sqrt{t}}$ by some constant $C_3 > 0$ which if at all depends only on θ_{min}. This yields the first part of the lemma. The second part is just a consequence of the first part together with (B.18). □

Proof of Lemma B.3. Define the function $\phi(x_1, \ldots, x_l) = {}^{l}_{i=1} x_i^{(1+\gamma)/\gamma} / \;\; {}^{l}_{i=1} x_i$ for $x_1, \ldots, x_l \in [A, B]$ for $0 < A < B$. Then, we have that for $i = 1, \ldots, l$

$$
\frac{\partial \phi(x_1, \ldots, x_l)}{\partial x_i} = \frac{\frac{1+\gamma}{\gamma} x_i^{1/\gamma} \sum_j x_j - \sum_j x_j^{(1+\gamma)/\gamma}}{(\sum_j x_j)^2},
$$

It can be easily checked that

$$
\sup_i \sup_{x_i \in [A,B]} \frac{\partial \phi(x_1, \ldots, x_l)}{\partial x_i} \leq \begin{cases} \frac{B^{\frac{1-\gamma}{\gamma}}}{\gamma}, & \gamma \leq 1, \\ \frac{A^{\frac{1-\gamma}{\gamma}}}{\gamma}, & \gamma > 1. \end{cases},
$$

Without loss of generality assume that $S_t = \{1, \ldots, l\}$, then by setting $x_i = O_{i,J}$ and $y_i = O_{i,J}^{\mathrm{TRCB}}(t)$ and noting that $\phi(x_1, \ldots, x_l) = \widetilde{\mathrm{U}}(S_t; O_J)$ as well as $\phi(y_1, \ldots, y_l) = \widetilde{\mathrm{U}}(S_t; \hat{O}_J^{\mathrm{TRCB}})$, we obtain with the mean value theorem that

$$
\begin{aligned}
\widetilde{\mathrm{U}}(S_t; \hat{O}_J^{\mathrm{TRCB}}, \gamma) - \widetilde{\mathrm{U}}(S_t; O_J, \gamma) \\
\leq C \sum_{i \in S_t} |\hat{O}_{i,J}^{\mathrm{TRCB}}(t) - O_{i,J}|,
\end{aligned}
$$

by choosing $C = \max\{\theta_{min}^{\frac{\gamma-1}{\gamma}}/\gamma, \theta_{min}^{\frac{1-\gamma}{\gamma}}/\gamma, \}$, since $\theta_{min} \leq O_{i,J} \leq 1/\theta_{min}$ and $\theta_{min} \leq O_{i,J}^{\mathrm{TRCB}}(t) \leq 1/\theta_{min}$. □

Proof of Lemma B.4. Since $\sum_{t=1}^{T} t^{-1/2} \leq 2\sqrt{T}$ one has $\sum_{\overline{w}_{i,J}(t)=1}^{\overline{w}_{i,J}(T)} {}^{1}\!/\sqrt{\overline{w}_{i,J}(t)} \leq 2\sqrt{\overline{w}_{i,J}(T)}$. Due to $\sum_{i \in [n]} \mathbb{E}\overline{w}_{i,J}(T) \leq T$ it follows by Jensen's inequality that $\sum_{t=1}^{T} \mathbb{E} \sum_{i \in S_t} \sqrt{\frac{1}{\overline{w}_{i,J}(t)}} \leq 4\sqrt{T\, n}$. □

C. Proof of Theorem 5.2

We start by introducing the notation for the rest of the proof and recalling the main terms of the CBR algorithm. Thereafter we give an outline of the proof, before deriving the technical details.

We break the proof down into two core lemmas, for which we first clarify the notation. We assume that without loss of generality $|S^*| = 1$, i.e., there is only one best arm, as this makes the learning problem only more difficult. Indeed, having several arms with the same highest score extends the opportunities to identify one of these highest score arms. To ease the notation we denote the score of the highest scored arm with θ_{max}, which is 1 by definition of Θ and its index by i_{max}.

C.1. Notation and relevant terms

We define the estimate for the pairwise winning probability $q_{i,j}$ (cf. (3)) by

$$
\hat{q}_{i,j} = \hat{q}_{i,j}(t) = \begin{matrix} \frac{w_{i,j}(t)}{w_{i,j}(t) + w_{j,i}(t)}, & i, j \in [n], i = j, \\ 0, & i = j, \end{matrix}
$$

where $w_{i,j}$ are as in (B.13) and with the convention that ${}^{x}/_{0} = 0$. With $J(t) = J$ we again denote the arm (within the active set) with the most picks till time instance t as in (B.15). With $\Delta_i = \theta_{max} - \theta_i$ we define the gap between the score of the ith arm and the overall best arm. The lengths of the confidence intervals are

$$c_{i,j}^{\mathrm{CBR}}(t) = c_{i,j} = \begin{array}{ll} \sqrt{\frac{2\log(n\,t^{3/2})}{\overline{w}_{i,j}(t)}}, & i,j \in [n], i = j, \\ 0, & i = j, \end{array}$$

thereby implicitly setting $\overline{w}_{ii}(t) = \infty$ for any $i \in [n]$.

C.2. Outline of the proof

We define the following events

$$B_t = \{\exists i \in [n] \mid |\hat{q}_{i,J}(t) - q_{i,J}| > c_{i,J}(t)\},$$
$$R_t = \{J(t) = i_{max}\}, \quad E_t = \{|S_t| > 1\}.$$

Here, B_t is the event where an arm exists whose pairwise probability estimate for winning against J is not close enough to its actual parameter, where closeness is understood by means of the confidence length $c_{i,J}(t)$. R_t is the event when the most winning arm J is not the overall best arm and E_t is the event, where the offered subset at time instance t is not a singleton. All these events are ”bad” events and we will show that their probability of occurrence is sufficiently small.

We have the following key lemmas to prove the main result.

Lemma C.1. *There exist constants $C_1, C_2, C_3 > 0$ independent of T and n and depending if at all on the parameter space Θ, such that*

$$\sum_{t=1}^{T} \mathbb{P}(B_t) \leq C_1 \quad \textit{and}$$
$$\sum_{t=1}^{T} \mathbb{P}(R_t \cap B_t^{\complement}) \leq C_2 \log(T) \sum_{i\in[n]\setminus\{i_{max}\}} \frac{1}{\Delta_i^2} + C_3 n.$$

Lemma C.2. *There exist constants $C_1, C_2 > 0$ independent of T and n and depending if at all on the parameter space Θ, such that*

$$\sum_{t=1}^{T} \mathbb{P}(B_t^{\complement} \cap R_t^{\complement} \cap E_t) \leq C_1 \log(T) \sum_{i\in[n]\setminus\{i_{max}\}} \frac{1}{\Delta_i^2} + C_2 n.$$

Lemma C.3. *On the event $R_t^{\complement}$ it holds that*

$$r(S_t) \leq \frac{\max\{\theta_{min}^{(\gamma-1)/(\gamma)}, \theta_{min}^{(1-\gamma)/(\gamma)}\}}{\gamma} \sum_{i\in[n]\setminus S^*} \Delta_i.$$

Putting all together. Recalling the cumulative regret in (9), we obtain

$$\mathbb{E}\big(\mathcal{R}(T) = \sum_{t=1}^{T} \mathbb{E}\, r(S_t)$$
$$\leq \sum_{t=1}^{T} \mathbb{P}(B_t) + \sum_{t=1}^{T} \mathbb{P}(R_t \cap B_t^{\complement})$$
$$+ \sum_{t=1}^{T} \mathbb{E}\, r(S_t) 1_{B_t^{\complement} \cap R_t^{\complement}}$$
$$\leq \sum_{t=1}^{T} \mathbb{P}(B_t) + \sum_{t=1}^{T} \mathbb{P}(R_t \cap B_t^{\complement})$$
$$+ \sum_{t=1}^{T} \mathbb{E}\, r(S_t) 1_{B_t^{\complement} \cap R_t^{\complement} \cap E_t}$$
$$+ \sum_{t=1}^{T} \mathbb{E}\, r(S_t) 1_{B_t^{\complement} \cap R_t^{\complement} \cap E_t^{\complement}}$$
$$\leq C_0 n + C_1 \frac{\max\{\theta_{min}^{(\gamma-1)/(\gamma)}, \theta_{min}^{(1-\gamma)/(\gamma)}\}}{\gamma} \log(T)$$
$$\times \sum_{i\in[n]\setminus\{i_{max}\}} \frac{1}{\Delta_i},$$

where we used Lemma C.1 and Lemma C.2 to derive the constants $C_0, C_1 > 0$, which are both independent of T and n, while Lemma C.3 introduced the factor accompanying C_1. Furthermore, we used that on $R_t^{\complement} \cap E_t^{\complement}$ we have that S_t equals $\{i_{max}\} = S^*$ and thus $r(S_t) = 0$.

C.3. Proofs of the core lemmas in Subsection C.2

Proof of Lemma C.1. Using Lemma B.5 one obtains

$$\mathbb{P}(B_t)$$
$$\leq \sum_{i\in[n]} \sum_{r=1}^{t-1} \mathbb{P}\big(|\hat{q}_{i,J}(t) - q_{i,J}| > c_{i,J}(t),\ \overline{w}_{i,J}(t) = r$$
$$\leq 2n \sum_{r=1}^{t-1} \exp(-4\log(nt^{3/2})) \leq {}^{2}/_{t^5}.$$

By summing over t till T, we get $\sum_{t=1}^{T} {}^{2}/_{t^5} < 2\sum_{t=1}^{\infty} {}^{1}/_{t^2} = \pi^2/3$, which yields the first claim.

For the second claim, let A_t denote the set of active arms at time instance t, i.e.,

$$A_t = \quad i \in [n] \quad \sigma \quad \frac{\hat{q}_{i,J(s)}(s) + c_{i,J(s)}(s) - 1/2}{2c_{i,J(s)}(s)} \quad > 0,$$
$$\forall s \in [t]$$

It holds that conditioned on $B_t^{\complement}$ we have that $i_{max} \in A_t$ almost surely. Indeed,

$$\mathbb{P}(\{i_{max} \notin A_t\} \cap B_t^{\complement})$$
$$= \mathbb{P} \quad \sigma \quad \frac{\hat{q}_{i_{max},J}(t) + c_{i_{max},J}(t) - 1/2}{2c_{i_{max},J}(t)} \quad \leq 0,\ B_t^{\complement}$$
$$= \mathbb{P} \quad \hat{q}_{i_{max},J}(t) + c_{i_{max},J}(t) \leq 1/2,\ B_t^{\complement}$$

$$\leq \mathbb{P}\ \ q_{i_{max},J}(t) \leq 1/2) = 0,$$

where we used that $\sigma(x) \leq 0$ iff $x \leq 0$ and for the last inequality that $\hat{q}_{i_{max},J}(t) + c_{i_{max},J}(t) \geq q_{i_{max},J}(t)$ on $B_t^{\complement}$, while $q_{i_{max},J}(t) > 1/2$ holds by definition of i_{max}.
Next, consider the counting process $M_t^{i,i_{max}} := w_{i,i_{max}} - w_{i_{max},i}$ for some $i \in A_t \backslash \{i_{max}\}$ and define for sake of brevity the event $\tilde{S}_s^i = \{\{i, i_{max}\} \in S_s\}$. Note that $M_t^{i,i_{max}}$ can be written as

$$M_t^{i,i_{max}} = \sum_{s=1}^{t-1} 1_{\{\{i_s=i\}\cap\tilde{S}_s^i\}} - 1_{\{\{i_s=i_{max},\}\cap\tilde{S}_s^i\}}.$$

It holds that the event $\{\{i, i_{max}\} \in S_s\}$ has a strictly positive probability for any arm $i \in A_t\backslash\{i_{max}\}$ and any $s \in [t]$, as otherwise the arm would not be active anymore. Conditioned on some set S_s we have that

$$\begin{aligned}\mathbb{P}\big(\{i_s = i\}\ \ &- \mathbb{P}\big(\{i_s = i_{max}\}\\ &= \frac{\theta_i}{\sum_{j\in S_s}\theta_j} - \frac{\theta_{max}}{\sum_{j\in S_s}\theta_j} \leq -\frac{\Delta_i}{H'},\end{aligned}$$

where $H' = \sum_{i\in[n]}\theta_i$. Thus, we can find a constant $C > 0$, which depends only on Θ such that for each $s \in [t]$

$$\mathbb{P}\big(\{i_s = i\} \cap \tilde{S}_s^i\ \ - \mathbb{P}\big(\{i_s = i_{max}\} \cap \tilde{S}_s^i\ \ \leq -\Delta_i\, C.$$

Therefore, $\mathbb{E}M_t^{i,i_{max}} \leq -(t-1)\, C\, \Delta_i$ and by Lemma C.5 it follows that

$$\begin{aligned}&\mathbb{P}(w_{i,i_{max}} \geq w_{i_{max},i})\\ &= \mathbb{P}(M_t^{i,i_{max}} \geq 0) \leq \mathbb{P}(M_t^{i,i_{max}} \geq -2(t-1)\, C\, \Delta_i)\\ &\leq \exp\big(-\frac{C^2\,\Delta_i^2(t-1)}{8}\ \ .\end{aligned}$$

The event R_t is contained in the event that there exists an active arm i such that the winning count of i_{max} against i is smaller than the winning count of i against i_{max}, that is $M_t^{i,i_{max}} \geq 0$. Hence, using the union bound in combination with the latter display we obtain

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{P}(R_t \cap B_t^{\complement})\\ &\leq \sum_{t=1}^{T}\sum_{i\in[n]\backslash\{i_{max}\}} \exp\big(-\frac{C^2\,\Delta_i^2(t-1)}{8}\\ &= \sum_{i\in[n]\backslash\{i_{max}\}}\ \sum_{t=1}^{\lceil 8\ \log(T)/C^2\Delta_i^2\rceil} \exp\big(-\frac{C^2\,\Delta_i^2(t-1)}{8}\\ &\quad + \sum_{t\geq\lceil 8\ \log(T)/C^2\Delta_i^2\rceil}^{T} \exp\big(-\frac{C^2\,\Delta_i^2(t-1)}{8}\\ &\leq \frac{8\,\log(T)}{C^2}\sum_{i\in[n]\backslash\{i_{max}\}}\frac{1}{\Delta_i^2} + 2\,n\,T\exp(-\log(T)),\end{aligned}$$

from which we can conclude the lemma. □

Proof of Lemma C.2. For any $i = i_{max}$ we have that

$$\mathbb{E}(\overline{w}_{i,i_{max}}(t)) = \sum_{s=1}^{t-1}\mathbb{P}(i_s \in \{i, i_{max}\}, \{i, i_{max}\} \in S_s).$$

Now, similar as in the proof of Lemma C.1 before, we can find a constant $\tilde{C} > 0$ which depends if at all on Θ such that $\mathbb{P}(i_s \in \{i, i_{max}\}, \{i, i_{max}\} \in S_s) \geq \theta_{min}\tilde{C}$ for any active arm i and each $s \in [t]$. With this, we obtain that $\mathbb{E}(\overline{w}_{i,i_{max}}(t)) \geq (t-1)\theta_{min}\tilde{C}$. Using Lemma C.6 with $\overline{w}_{i,i_{max}}$ as the counting process one can derive that there exists a constant $C > 0$ depending on Θ such that

$$\mathbb{P}\ \ \overline{w}_{i,i_{max}}(t) \leq \frac{(t-1)C}{2}\ \ \leq \exp\big(-\frac{(t-1)C^2}{8}\ \ . \tag{C.20}$$

Next, write for short $\delta_{i,i_{max}} = \hat{q}_{i,i_{max}}(t) + c_{i,i_{max}}(t) - 1/2$ and note that

$$\begin{aligned}&\mathbb{P}(B_t^{\complement} \cap R_t^{\complement} \cap E_t)\\ &= \mathbb{P}(\exists i = i_{max} : \{i \in S_t\},\, B_t^{\complement} \cap R_t^{\complement})\\ &\leq \sum_{i\in[n]\backslash\{i_{max}\}}\mathbb{P}\ \ \sigma\ \ \frac{\delta_{i,i_{max}}}{2c_{i,i_{max}}(t)}\ \ \geq 0,\, B_t^{\complement} \cap R_t^{\complement}\\ &\leq \sum_{i\in[n]\backslash\{i_{max}\}}\mathbb{P}\ \ \delta_{i,i_{max}} \geq 0,\, B_t^{\complement} \cap R_t^{\complement}\\ &\leq \sum_{i\in[n]\backslash\{i_{max}\}}\mathbb{P}\ \ 2c_{i,i_{max}}(t) \geq 1/2 - q_{i,i_{max}}\\ &= \sum_{i\in[n]\backslash\{i_{max}\}}\mathbb{P}\ \ \overline{w}_{i,i_{max}}(t) \leq \frac{8\log(nt^{3/2})}{(1/2 - q_{i,i_{max}})^2}\\ &\leq \sum_{i\in[n]\backslash\{i_{max}\}}\mathbb{P}\ \ \overline{w}_{i,i_{max}}(t) \leq \frac{20\log(T)}{(1/2 - q_{i,i_{max}})^2}\ \ ,\end{aligned}$$

where we used that $J(t) = i_{max}$ on $R_t^{\complement}$ for the first inequality, $\sigma(x) \leq 0$ iff $x \leq 0$ for the second inequality, for the third inequality that $\hat{q}_{i,i_{max}}(t) - c_{i,i_{max}}(t) \leq q_{i,i_{max}}(t)$ on $B_t^{\complement}$, while the last inequality is due to $\log(nt^{3/2}) \leq 5/2\log(T)$, as $\max\{n,t\} \leq T$. One can find constants $C_i \in [1/4, 1/2]$ such that $1/2 - q_{i,i_{max}} = C_i\Delta_i$. Indeed, note that $1/2 - q_{i,i_{max}} = {}^{\Delta_i}/_{(2(\theta_i+\theta_{max}))}$ and it holds that

$$\frac{\Delta_i}{4} \leq \frac{\Delta_i}{2(\theta_i + \theta_{max})} \leq \frac{\Delta_i}{2}.$$

Hence, with these considerations one obtains

$$\sum_{t=1}^{T}\mathbb{P}(B_t^{\complement} \cap R_t^{\complement} \cap E_t)$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{T} \sum_{i\in[n]\setminus\{i_{max}\}} \mathbb{P} \ \overline{w}_{i,i_{max}}(t) \leq \frac{20\log(T)}{C_i^2\Delta_i^2} \\
&\leq \sum_{i\in[n]\setminus\{i_{max}\}} \frac{40\log(T)}{CC_i^2\Delta_i^2} \\
&+ \sum_{i\in[n]\setminus\{i_{max}\}} \sum_{t=\lceil \frac{40\log(T)}{CC_i^2\Delta_i^2}\rceil}^{T} \mathbb{P} \ \overline{w}_{i,i_{max}}(t) \leq \frac{20\log(T)}{C_i^2\Delta_i^2} \quad .
\end{aligned}
$$

Now, the summation over t on the right-hand side of the last display is such that ${}^{20\log(T)}/_{C_i^2\Delta_i^2} \leq {}^{(t-1)C}/_{2}$. Thus, we can use (C.20) to further estimate the last display by

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}(B_t^{\complement} \cap R_t^{\complement} \cap E_t) \\
&\qquad \leq \frac{40\log(T)}{C} \sum_{i\in[n]\setminus\{i_{max}\}} \frac{1}{C_i^2\Delta_i^2} + C_1 n T^{-C_2},
\end{aligned}
$$

for some constants $C_1, C_2 > 0$. From the latter display we can conclude the lemma. □

Proof of Lemma C.3. Note that

$$
\begin{aligned}
r(S_t) &= \mathrm{U}(S^*) - \mathrm{U}(S_t) \\
&= \frac{\sum_{i\in S_t}(v_{\max} - v_i)v_i^{\gamma}}{\sum_{i\in S_t} v_i^{\gamma}} \\
&= \frac{\sum_{i\in S_t}(\theta_{max}^{1/\gamma} - \theta_i^{1/\gamma})\theta_i}{\sum_{i\in S_t}\theta_i}.
\end{aligned}
$$

On the event $R_t^{\complement}$ it holds that $i_{max} \in S_t$, so that

$$
\sum_{i\in S_t} \theta_i \geq \theta_{max} = 1.
$$

With this, and the fact that $\theta_i \leq \theta_{max} = 1$, we can infer that

$$
r(S_t) \leq \sum_{i\in S_t} (\theta_{max}^{1/\gamma} - \theta_i^{1/\gamma}) \leq \sum_{i\in[n]\setminus S^*} (\theta_{max}^{1/\gamma} - \theta_i^{1/\gamma}).
$$

Considering the function $f(x) = x^{1/\gamma}$ defined for $x \in [\theta_{min}, \theta_{max}]$ the assertion follows easily by the mean-value theorem as in the proof of Lemma B.3. □

C.4. Technical results

In this subsection we collect the technical auxiliary results needed for the proofs of the core lemmas. These technical results could also be of independent interest.

The next two lemmas were of major importance for the proof of Lemma C.1.

Lemma C.4. *Let $M_t = \sum_{s=1}^{t} Z_s$, where $(Z_s)_{s=1,\dots,t}$ are random variables with values in $\{-1, 0, 1\}$, such that $\mathcal{F}_s$ is the canonical filtration generated by $\{Z_1, \dots, Z_{s-1}\}$ and Z_{s+1} is conditionally independent of $Z_{s+2}, \dots, Z_t$ given $\mathcal{F}_s$. We have that for any $z > 0$*

$$
\mathbb{P}(M_t - \mathbb{E}(M_t) > z) \leq \exp\big(- \frac{z^2}{8\,t} \ .
$$

Proof of Lemma C.4. The function $f(z_1, \dots, z_t) = z_1 + \dots + z_t$ is Lipschitz-continuous with Lipschitz constant $L = 2$ if $-1 \leq z_i \leq 1$ for each i. It is a well-known result that the sequence of random variables $(X_i)_{i=1,\dots,t}$ with $X_i = \mathbb{E}[f(Z_1, \dots, Z_t)|\mathcal{F}_i]$ is a martingale (the so-called Doob martingale) with bounded differences $|X_{i+1} - X_i| \leq 2L = 4$ (cf. Lemma 11 in Kocsis et al. (2006)). Consider the martingale difference sequence $\tilde{X}_i = X_i - \mathbb{E}X_i = X_i - \mathbb{E}M_t$ and note that $\tilde{X}_t = X_t - \mathbb{E}X_t = M_t - \mathbb{E}M_t$ and $\tilde{X}_0 = 0$ by setting $\mathcal{F}_0 = \{\varnothing, \Omega\}$. Thus, the Azuma-Hoeffding inequality implies for any $z > 0$ that

$$
\begin{aligned}
\mathbb{P}(M_t - \mathbb{E}(M_t) > z) &= \mathbb{P}(\tilde{X}_t - \tilde{X}_0 > z) \\
&\leq \exp(-z^2/(8\,t)).
\end{aligned}
$$

□

Lemma C.5. *Consider the setting of Lemma C.4 and assume that there exists Δ_t such that $\mathbb{E}(M_t) \leq \Delta_t/2$. Then,*

$$
\mathbb{P}(M_t \geq \Delta_t) \leq \exp\big(- \frac{\Delta_t^2}{32\,t} \ .
$$

Proof of Lemma C.5.

$$
\begin{aligned}
\mathbb{P}(M_t \geq \Delta_t) &= \mathbb{P}(M_t \geq \mathbb{E}(M_t) + \Delta_t - \mathbb{E}(M_t)) \\
&\leq \mathbb{P}(M_t \geq \mathbb{E}(M_t) + \Delta_t/2) \\
&\leq \exp(-\Delta_t^2/(32\,t)),
\end{aligned}
$$

where we used Lemma C.4 in the last step. □

For the proof of Lemma C.2 we use the following variant of Lemma 13 in Kocsis et al. (2006).

Lemma C.6. *Let $N_t = \sum_{s=1}^{t} Z_s$, where $(Z_s)_{s=1,\dots,t}$ are random variables with values in $\{0, 1\}$, such that $\mathcal{F}_s$ is the canonical filtration generated by $\{Z_1, \dots, Z_{s-1}\}$ and Z_{s+1} is conditionally independent of $Z_{s+2}, \dots, Z_t$ given $\mathcal{F}_s$. If $\mathbb{E}N_t \geq 2\Delta_t$, for some Δ_t then*

$$
\mathbb{P}(N_t \leq \Delta_t) \leq \exp\big(- {}^{\Delta_t^2}/_{2\,t} \ .
$$

Proof of Lemma C.6. By using $\mathbb{E}N_t \geq 2\Delta_t$, we have

$$
\begin{aligned}
\mathbb{P}(N_t \leq \Delta_t) &= \mathbb{P}(N_t \leq \mathbb{E}N_t + \Delta_t - \mathbb{E}N_t) \\
&\leq \mathbb{P}(N_t \leq \mathbb{E}N_t - \Delta_t) \\
&\leq \exp\big(- {}^{\Delta_t^2}/_{2\,t} \ ,
\end{aligned}
$$

where we used Lemma 12 of Kocsis et al. (2006) for the last inequality. □

D. Optimal subsets for restricted Pre-Bandits and an efficient algorithm for utility maximization

In this section, we show that the best arm is always element of the optimal preselection for the restricted Pre-Bandit case. Following this, we present a sophisticated algorithm (Algorithm 3) to avoid highly computational costs for determining the maximizing set in line 11 of Algorithm 1.

The following lemma, which can be verified by simple techniques of curve sketching, is the foundation for Algorithm 3 and the proof of Lemma D.2.

Lemma D.1. *Let $0 \leq a < b$ be real values, $(\theta_1, \ldots, \theta_n) \in [a,b]^n$ and $S \subseteq [n]$ be a nonempty subset. Further, define $f_\gamma : [a,b] \to \mathbb{R}^+$ by*

$$f_\gamma(\theta) = f_\gamma(\theta; S) = \frac{\theta^{1+\gamma} + \sum_{i\in S} \theta_i^{1+\gamma}}{\theta^\gamma + \sum_{i \in S} \theta_i^\gamma}.$$

The following statements are valid.

(i) For $\tilde{\theta} = {}_{i\in S}\, \theta_i^{1+\gamma} / {}_{i \in S}\, \theta_i^\gamma$ we have that $f_\gamma(\tilde{\theta}) = f_\gamma(0) = \tilde{\theta}$.

(ii) f_γ has a unique global minimum in $\bar{\theta}$, which is the (unique) real-valued solution of the following equation in x

$$x^{1+\gamma} + (1+\gamma)\Big(\sum_{i\in S} v_i^\gamma \; x - \gamma \sum_{i \in S} v_i^{1+\gamma} = 0.$$

It holds that f_γ is strictly decreasing in $[a, \bar{\theta}]$ and strictly increasing in $[\bar{\theta}, b]$. Moreover, $\bar{\theta} \leq \tilde{\theta}$.

Lemma D.2. *Let $\theta \in \Theta$ be such that $|\arg\max_{i\in[n]} \theta_i| = 1$ and let $J = \arg\max_{i \in [n]} \theta_i$. Then, for any $l \in \mathbb{N}$, one has $J \in S^*$, where each S^* is a maximizing subset as in (7) for $\mathbb{A} = \mathbb{A}_l$. Furthermore, if $|\arg\max_{i\in[n]} \theta_i| > 1$ then $\mathrm{U}(\{J\}) \geq \mathrm{U}(\{J\} \cup \{i\})$ for any $i \in [n]$, with an equality if and only if $\theta_i = \theta_J$. The same holds true for $\widetilde{\mathrm{U}}$.*

Proof of Lemma D.2. We prove the first assertion by contradiction. Hence, suppose that $J \notin S^*$. Let $\tilde{J} \in S^*$ be such that $\theta_{\tilde{J}} < \theta_J$ and define $\tilde{S} = S^* \backslash \{\tilde{J}\} \cup \{J\}$. Thus, by assumption it should hold that

$$\begin{aligned}\mathrm{U}(S^*) &= \frac{\theta_{\tilde{J}}^{1+\gamma} + \sum_{i \in S^*\backslash\{\tilde{J}\}} \theta_i^{1+\gamma}}{\theta_{\tilde{J}}^{\gamma} + \sum_{i\in S^*\backslash\{\tilde{J}\}} \theta_i^\gamma} = \frac{\sum_{i\in S^*} \theta_i^{1+\gamma}}{\sum_{i \in S^*} \theta_i^\gamma} \\ &> \frac{\sum_{i\in\tilde{S}} \theta_i^{1+\gamma}}{\sum_{i\in\tilde{S}} \theta_i^\gamma} = \frac{\theta_J^{1+\gamma} + \sum_{i\in S^*\backslash\{\tilde{J}\}} \theta_i^{1+\gamma}}{\theta_J^\gamma + \sum_{i \in S^*\backslash\{\tilde{J}\}} \theta_i^\gamma} \\ &= \mathrm{U}(\tilde{S}).\end{aligned}$$

In terms of Lemma D.1 this means that $f_\gamma(\theta_{\tilde{J}}, S^*\backslash\{\tilde{J}\}) > f_\gamma(\theta_J, S^*\backslash\{\tilde{J}\})$, but this is a contradiction due to (i) and (ii) of Lemma D.1, as $\theta_J > \tilde{\theta} = \frac{{}_{i\in S^*\backslash\{\tilde{J}\}}\, \theta_i^{1+\gamma}}{{}_{i\in S^*\backslash\{\tilde{J}\}}\, \theta_i^\gamma}$ and $\bar{\theta} \in [0, \tilde{\theta}]$. The second claim follows immediately by the strict monotonic behavior of f_γ and the claims for $\widetilde{\mathrm{U}}$ can be shown similarly. □

Algorithm 3 Utility-maximization

```
input  $n$ many paramters $\theta_1, \ldots, \theta_n$, preciseness parameter
       $\gamma$, preselection size $l$
initialization: $\tau \leftarrow Sort(\theta_1, \ldots, \theta_n)$ {determine permutation which sorts the scores in decreasing order}
$S \leftarrow \arg\max_{i\in\tau([n])} \theta_{\tau(i)}$ {select all high-score items}
if $|S| \geq l$ then
    return: randomly selected $l$ elements of $S$
else
    $A \leftarrow [n]\backslash[|S|]$ { set of active arms }
    repeat
        $\tilde{\theta} \leftarrow {}_{i\in S}\, \theta_i^{1+\gamma} / {}_{i\in S}\, \theta_i^\gamma$
        $A_{next} \leftarrow \arg\min_{i \in \{\min A, \max A\}} \{|\tilde{\theta} - f_\gamma(\theta_{\tau(i)}; S)|\}$
        {$f_\gamma$ as in Lemma D.1, break ties arbitrarily}
        $S \leftarrow S \cup \tau(A_{next})$
        $A \leftarrow A \backslash A_{next}$
    until $|S| == l$
    return: $S$
end if
```

Let $\theta_{(i)}$ denote the i-th order statistic for $(\theta_1, \ldots, \theta_n)$, i.e.,

$$\theta_{(1)} \leq \theta_{(2)} \leq \ldots \leq \theta_{(n)},$$

then Lemma D.1 implies that $f_\gamma\big(v; \{\theta_{(n)}\} \leq f_\gamma\big(\theta_{(n)}; \{\theta_{(n)}\}$ for any $v \in [0, \theta_{(n)}]$ and the smallest decrease of $f_\gamma\big(\cdot; \{\theta_{(n)}\}$ over the discrete set $\{\theta_{(1)}, \ldots, \theta_{(n-1)}\}$ is either for $\theta_{(n-1)}$ or for $\theta_{(1)}$.

With this, Algorithm 3 successively builds a set S which will maximize the expected utility in (6) for a given score parameter $\theta = (\theta_1, \ldots, \theta_n)$. First, the scores are sorted in order to find the arms with the highest scores, as by Lemma D.2 these are always element of the maximizing subset. If more than $(l-1)$ elements have the same highest score, a randomly chosen l-sized set of these is returned, since the expected utility among all possible l-sized subsets of these is the same by Lemma D.1 or Lemma D.2.

Otherwise, an active index set A is initialized containing all indices for which it is not decided yet, if they are part of the maximizing set S eventually. As by Lemma D.2 the expected utility decreases from that point on by enlarging the set S, the algorithm determines the arm with the small-

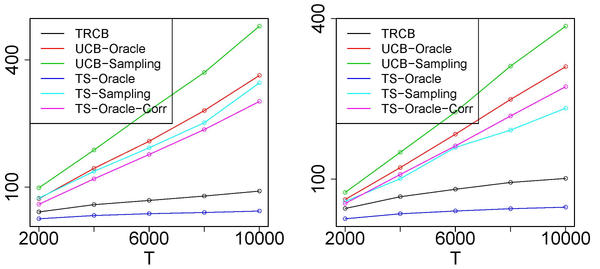

Figure 1. Mean cumulative regret for 1000 runs of randomly generated restricted PB instances for $(n, l) = (20, 4)$ (left) and $(n, l) = (30, 5)$ (right).

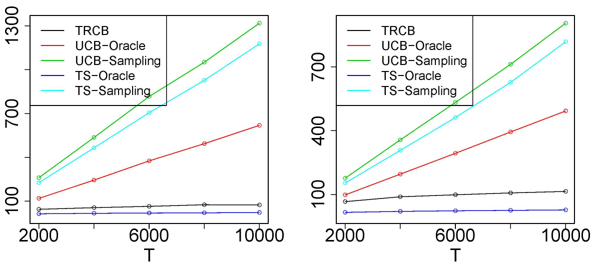

Figure 2. Mean cumulative regret for 1000 runs of randomly generated restricted PB instances for $(n, l) = (10, 3)$ (left) and $(n, l) = (20, 4)$ (right) and $\gamma = 1/20$.

est decrease for the expected utility, where ties are broken arbitrary by two possible candidates.

Since the expected utility of the currently set S is identical to $f_\gamma(0; S)$ only the arms with the smallest resp. highest score parameter in A have to be checked by the implication after Lemma D.2. It can be shown that the algorithm has worst complexity of $O(l\, n\, \log(n))$ if an efficient sorting algorithm is used in the initial step.

E. Further experiments for the Pre-Bandit problem

In this section, we provide further experiments on synthetic data for the two variants of the Pre-Bandit problem.

Restricted Pre-Bandit problem (larger number of arms) First, we present two additional scenarios of the simulation study in Section 6 for the restricted Pre-Bandit problem with larger numbers of arms n and different preselection sizes l. In particular, we investigate the performance of the following algorithms, which were also analyzed in Section 6, for the restricted Pre-Bandit problem:

- TRCB: The TRCB algorithm in Algorithm 1 with $C_{shrink} = 7 \cdot 10^{-5}$ and $\theta_{min} = 0.02$ (here as a parameter of the algorithm).
- UCB-Oracle: UCB-type algorithm of Agrawal et al. (2016) with knowledge of the best arm in advance and revenues are set to be the estimated score parameters (in short $r = \hat{\theta}$).
- UCB-Sampling: UCB-type algorithm of Agrawal et al. (2016) without knowledge of the best arm in advance (sampled with MNL probability among the three best) and $r = \hat{\theta}$.
- TS-Oracle: The Thompson sampling algorithm of Agrawal et al. (2017) (Algorithm 1) with knowledge of the best arm in advance and $r = \hat{\theta}$.
- TS-Sampling: The Thompson sampling algorithm of Agrawal et al. (2017) (Algorithm 1) without knowledge of the best arm in advance (sampled with MNL probability among the three best) and $r = \hat{\theta}$.
- TS-Oracle-Corr: Correlated Thompson sampling algorithm of Agrawal et al. (2017) (Algorithm 2) with knowledge of the best arm in advance and $r = \hat{\theta}$.

The left picture in Figure 1 provides the findings for the case $n = 20$ and $l = 4$, while the right picture illustrates our results for $n = 30$ and $l = 5$. Both scenarios are considered for the time horizons $T \in \{i \cdot 2000\}_{i=1}^{5}$ and the score parameters are drawn randomly from the n-simplex without any restrictions on θ_{min} and with $\gamma = 1$.

Table 1. Empirical standard deviations of the cumulative regret for the different time horizon steps for the scenarios $(n, l) = (20, 4)$ and $(n, l) = (30, 5)$.

	$(n, l) = (20, 4)$				
T	2000	4000	6000	8000	10000
TRCB	21.47	32.93	43.73	57.51	66.36
UCB-Oracle	43.90	79.27	119.19	165.42	202.54
UCB-Sampling	98.31	187.52	280.59	370.10	479.20
TS-Oracle	7.74	10.01	11.37	13.42	14.06
TS-Sampling	86.11	161.01	235.16	329.54	429.52
TS-Oracle-Corr	21.84	43.65	63.73	87.99	111.97
	$(n, l) = (30, 5)$				
TRCB	20.54	34.81	40.84	54.91	58.88
UCB-Oracle	34.45	65.40	102.52	143.31	172.84
UCB-Sampling	75.21	150.24	225.29	311.10	385.72
TS-Oracle	6.91	9.48	11.17	13.43	13.91
TS-Sampling	53.66	101.35	175.86	246.12	284.47
TS-Oracle-Corr	18.01	38.74	55.92	80.59	97.91

The findings are similarly as for the case $n = 10$ and $l = 3$, that is only the Thompson Sampling algorithm with knowledge of the best arm apriori (TS-Oracle) outperforms TRCB, while the other algorithms are outperformed by TRCB. Furthermore, we report the empirical standard deviations of the

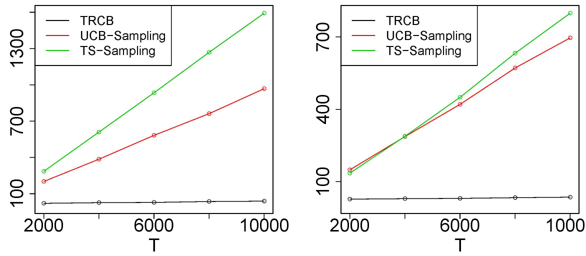

Figure 3. Mean cumulative regret for 1000 runs of randomly generated restricted PB instances for $(n, l) = (10, 3)$ (left) and $(n, l) = (20, 4)$ (right) and $\gamma = 20$.

considered algorithms for each time horizon in both scenarios in Table 1. Only TS-Oracle has a throughout smaller standard deviation than TRCB, while all the others have variations of a higher magnitude than TRCB.

Restricted Pre-Bandit problem (Varying degree of preciseness) Next, we consider two additional scenarios, in which we initially set $\gamma = 1/20$ such that the most preferred subsets consists throughout of the top-l arms and The results for $\gamma = 1/20$ are depicted in Figure 2 for the cases $(n, l) = (10, 3)$ and $(n, l) = (20, 4)$ for the algorithms described above. Note that TS-Oracle-Corr could not be compared as it sampled negative values for the score parameters, which lead to numerical issues regarding the evaluation of the utility function. Again the findings are in line with the observations we have made in the simulations before, i.e., only TS-Oracle is able to outperform our algorithm TRCB due to its advantage of knowing the best arm. In particular, this demonstrates that our algorithm performs well for scenarios where top-l subsets are the desired outcome for a user.

In addition, we consider the case $\gamma = 20$ such that the most preferred subsets are basically all subsets which contain the arm with the highest score (cf. Example 2 in the main paper). Figure 3 illustrates the results for the cases $(n, l) = (10, 3)$ and $(n, l) = (20, 4)$, where we do not included the algorithms which have prior knowledge of the best arm, as these naturally have throughout a regret of zero. This experiment indicates that the considered DAS algorithms depend too much on the assumption that the no-choice option corresponds to the highest scored arm as also remarked in Section 6.

Flexible Pre-Bandit problem In addition to the simulations in Section 6, we investigate the empirical regret growth over time for larger numbers of arms n for our CBR algorithm for the flexible Pre-Bandit problem. We consider two variants of the CBR-algorithm:

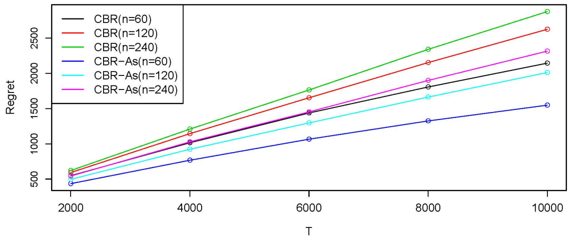

Figure 4. Mean cumulative regret of the variants of the CBR algorithm for 500 runs of randomly generated flexible Pre-Bandit instances for $n \in \{60, 120, 240\}$.

- CBR: The CBR algorithm with

$$\sigma(x) = (1 \wedge x) 1_{[0,\infty)}(x).$$

- CBR-As: The CBR algorithm with

$$\sigma(x) = \tfrac{1}{\pi} \arctan \quad \tfrac{x - 1/2}{(1-x)^{\rho} x^{\rho}} \quad + \tfrac{1}{2} \text{ and } \rho = 2.$$

Figure 4 illustrates the results of our simulations for both CBR algorithm variants over 500 repetitions, respectively, with $n \in \{60, 120, 240\}$, over the time horizons $T \in \{i \cdot 2000\}_{i=1}^{5}$ and the score parameters were drawn randomly from the unit interval and with $\gamma = 1$.

It is clearly visible that CBR-As outperforms CBR due to the more sophisticated choice of the S-curved function σ. Thus, it is reasonable to believe that the performance of CBR can be significantly improved by an appropriate choice of σ. Note that the Double Thompson Sampling considered in Section 6 was not competitive in these scenarios and is therefore omitted.

References

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. A near-optimal exploration-exploitation approach for assortment selection. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 599–600. ACM, 2016.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Thompson sampling for the mnl-bandit. In *Conference on Learning Theory*, pp. 76–78, 2017.

Chen, X. and Wang, Y. A note on a tight lower bound for capacitated mnl-bandit assortment selection models. *Operations Research Letters*, 46(5):534–537, 2018.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17 (1):1–42, 2016.

Kocsis, L., Szepesvári, C., and Willemson, J. Improved monte-carlo search. *Univ. Tartu, Tartu, Estonia, Tech. Rep. 1*, 2006.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press (to appear), 2020.

Saha, A. and Gopalan, A. Pac battling bandits in the plackett-luce model. In *Algorithmic Learning Theory*, pp. 700–737, 2019a.

Saha, A. and Gopalan, A. Combinatorial bandits with relative feedback. In *Advances in Neural Information Processing Systems*, pp. 983–993, 2019b.