Supplementary Material for "Combinatorial Partial Monitoring Game with Linear Feedback and Its Application".

A. Full proof for Theorems 4.1 and 4.2

If the reader will recall, we have the following problem-specific constants in the main text: the size of the global observer set $|\sigma|$, parameter $L > 0$ from the continuity assumption, error bound $\beta_\sigma = \max \|(M_\sigma^\intercal M_\sigma)^{-1} \sum_{i=1}^{|\sigma|} M_{x_i}^\intercal M_{x_i}(\boldsymbol{\nu}_i - \boldsymbol{\nu}_0)\|_2$, where the max is taken from $\boldsymbol{\nu}_0, \boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_{|\sigma|} \in [0,1]^n$, and the maximum difference in the expected reward $R_{\max} = \max_{x_1,x_2 \in X, \boldsymbol{\nu} \in [0,1]^n} |\overline{r}(x_1, \boldsymbol{\nu}) - \overline{r}(x_2, \boldsymbol{\nu})|$.

For technical reasons, we also defined $\phi(\boldsymbol{\nu}) = \max(\min(\boldsymbol{\nu}, \vec{\mathbf{1}}), \vec{\mathbf{0}})$ to adjust $\boldsymbol{\nu}$ to the nearest vector in $[0,1]^n$, and $\overline{r}(x, \boldsymbol{\nu}) = \overline{r}(x, \phi(\boldsymbol{\nu})), \forall \boldsymbol{\nu} \in \mathbb{R}^n \setminus [0,1]^n$ to preserve the Lipschitz continuity throughout $\mathbb{R}^n$.

To make our proof clearer, we define $v(t)$ as the state of any variable v by the end of time step t. Our analysis is based on the snapshot of all variables just before the statement $t \leftarrow t+1$ (Line 14 and 30). One batch processing in exploration phase is called *one round*, and then n_σ is increased by 1. Denote $\hat{\boldsymbol{\nu}}^{(j)}$ as the estimated mean of outcomes after j rounds of exploration. For example, at time t, the estimated mean of outcomes is $\hat{\boldsymbol{\nu}}(t)$ and the exploration counter is $n_\sigma(t)$, so we have $\hat{\boldsymbol{\nu}}^{(n_\sigma(t))} = \hat{\boldsymbol{\nu}}(t)$. And for time step $t+1$, the player will use the previous knowledge of $\hat{\boldsymbol{\nu}}(t)$ to get $\hat{x}(t+1) = \operatorname{argmax}_{x \in \mathcal{X}} \overline{r}(x, \hat{\boldsymbol{\nu}}(t))$ and $\hat{x}^-(t+1) = \operatorname{argmax}_{x \in \mathcal{X} \setminus \{\hat{x}(t+1)\}} \overline{r}(x, \hat{\boldsymbol{\nu}}(t))$.

In the following analysis, the frequency function is set to $f_{\mathcal{X}}(t) = \ln t + 2 \ln |\mathcal{X}|$. Note that by using $f_{\mathcal{X}}(t)$, we can construct the confidence interval $\sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma}}$ to eliminate failures with high probability. Define $N(t) = |\mathcal{X}|^2 t$, which will be frequently used in our analysis, then $\exp\{-f_{\mathcal{X}}(t)\} = N(t)^{-1}$. Let $\alpha = \frac{8L^2\beta_\sigma^2}{a^2}$, where $a > 0$ is a parameter to be tuned later.

The symbols used in the proof are listed in Table 1, and to facilitate the understanding of our proof, each lemma is briefly summarized in Table 2. Below we give an outline of the complete proof, which consists of three main parts:

- First, we introduce some general insights, present the preliminaries, and prove basic properties of our model and the Algorithm GCB, which is shared by the proofs of both distribution-independent and distribution-dependent regret bounds. Next, we obtain the concentration property of the empirical mean of outcomes via Azuma-Hoeffding Inequality (Fact A.2) in Lemma A.3. Lemma A.13 shows that our algorithm bounds the number of exploration rounds by $O(T^{2/3} \log T)$, which implies that our algorithm will not play exploration for too long. In Lemma A.14, we prove that when the gap between the estimated optimal action $\hat{x}$ and the second optimal $\hat{x}^-$ is large (i.e. the first condition in Line 8), with low probability the estimated optimal action is sub-optimal. This means that our global confidence bound will exclude sub-optimal actions effectively.
- In Section A.1, we prove the distribution-independent regret bound of $O(T^{\frac{2}{3}} \log T)$ (Theorem 4.1 in the main text). In Lemma A.15, we show that the gap between the estimated optimal action $\hat{x}$ and the real one x^* decays exponentially with the number of exploration rounds. Thus, the penalty in exploitation phase can be derived in Lemma A.16. Then, we use Lemmas A.13 and A.16 to prove Theorem 4.1 in the main text. Hence the distribution-independent bound $O(T^{\frac{2}{3}} \log T)$ is achieved.
- In Section A.2, we prove the distribution-dependent bound of $O(\log T)$ related to the predetermined distribution p, assuming that the optimal action x^* is unique (Theorem 4.2 in the main text). First, we show in Lemma A.17 that, when the algorithm plays $\Omega\left(\frac{\ln t + 2\ln|\mathcal{X}|}{\Delta_{\min}^2}\right)$ rounds of exploration, the probability of the estimated optimal action $\hat{x}$ being sub-optimal is low. Then, in Lemma A.18, we combine the results of Lemmas A.14 and A.17 and show that with a low probability the algorithm exploits with a sub-optimal action. Thus, Lemma A.18 is enough to bound the regret of exploitation. Next, we bound the regret of exploration by bounding the number of exploration rounds to $O\left(\frac{\ln T + 2\ln|\mathcal{X}|}{\Delta_{\min}^2}\right)$ in Lemma A.22. This is done by showing that whenever the algorithm has conducted $\Theta\left(\frac{\ln t + 2\ln|\mathcal{X}|}{\Delta_{\min}^2}\right)$ rounds of exploration, with high probability it switches to exploitation (Lemma A.19), and then aggregating multiple switches between exploration and exploitation in the proof of Lemma A.22. Finally, we combine Lemmas A.18 and A.22 to prove Theorem 4.2 in the main text.

Fact A.1. *The following probability laws will be used in the analysis.*

Symbols in the main text	*Definition*
$v(t)$	state any variable v by the end of time step t
$\mathbf{v} \in [0,1]^n$	outcomes of the environment
$\tilde{\mathbf{v}} \in [0,1]^n$	estimation of outcomes through inversion
$\boldsymbol{\nu}$	mean of outcomes
$\hat{\boldsymbol{\nu}}$	empirical mean of outcomes
$x \in \mathcal{X}$	action x in action set $\mathcal{X}$
$r(x, \mathbf{v}), \overline{r}(x, \boldsymbol{\nu}) \in \mathbb{R}$	reward function taking x and $\mathbf{v}$, and expected reward function taking x and $\boldsymbol{\nu}$
$M_x \in \mathbb{R}^{m_x \times n}$	transformation matrix of action x where m_x depends on x
$\mathbf{y}(t) \in \mathbb{R}^{m_{x(t)}}$	feedback vector under the choice of $x(t)$
$\vec{\mathbf{y}}$	vector that stacks feedbacks from different times
$\Delta_x, \Delta_{\max}, \Delta_{\min}$	reward gap of action x, of the maximum, of the (positive) minimum
$\sigma \subset \mathcal{X}$	global observer set of actions
L	Lipschitz constant
β_σ	distribution-independent error bound from σ
$R_{\max}$	distribution-independent largest gap of expected reward
$f_{\mathcal{X}}(t)$	frequency function
x^*	real optimal action
$\hat{x}(t)$	estimated optimal action at time t
$\hat{x}^-(t)$	estimated second optimal action at time t
Symbols in the proof	*Definition*
n_σ	exploration counter
$\mu(t), \eta(t)$	threshold functions
$\mathcal{X}^{\text{Good}}, \mathcal{X}^{\text{Bad}} \subset \mathcal{X}$	good action set, and bad action set
$\mathcal{F}^{\text{Good}}, \mathcal{F}^{\text{Bad}}$	event of choosing good action set, and bad action set
$\mathcal{L}_{CI}, \mathcal{L}_{CI}^c$	event of occurring gap being larger than confidence interval and its complement
$\mathcal{E}^{\text{Explore}}, \mathcal{E}^{\text{InExplore}}, \mathcal{E}^{\text{FinExplore}}$	event of doing exploration, being in the middle of it, and being at its end
$\mathcal{E}^{\text{Exploit}}$	event of exploitation
$\delta_{x_i,x_j}, \hat{\delta}_{x_i,x_j}(t) \in \mathbb{R}$	reward gap of action x_i and x_j, and its estimated value at time t
$\mathcal{G}_k$	the event indicating the first occurrence of k rounds of exploration

Table 1. List of symbols in the proof.

	Succinct interpretation of the results	Dependence
Lemma A.3	Estimate of outcomes concentrates around the mean.	Fact A.2
Lemma A.7	Difference of real and estimated gap is bounded.	
Lemma A.8	Estimated error of outcomes is small compared to the confidence interval.	Lemma A.3
Lemma A.13	The counter of exploration is bounded within $O(T^{2/3}\log T)$.	
Lemma A.14	Finding a bad action to fail confidence interval occurs rarely.	Lemma A.8
Lemma A.15	Incurring a large penalty for current optimal action is rare.	Lemma A.3
Lemma A.16	The penalty in the exploitation phase is bounded.	Lemmas A.14, A.15
Theorem 4.1	Distribution-independent bound: $O(T^{2/3}\log T)$.	Lemmas A.13, A.16
Lemma A.17	With enough exploration, finding a bad action is rare.	Lemma A.3
Lemma A.18	Finding a bad action and exploiting it become rare as time elapses.	Lemmas A.14, A.17
Lemma A.19	With enough exploration, finding a good action but yet exploring becomes rare.	Lemmas A.3, A.7, A.13
Lemma A.20	Once the algorithm performs enough exploration, it switches to exploitation.	Lemmas A.17, A.19
Lemma A.22	Exploration rounds are bounded.	Lemma A.20
Theorem 4.2	Distribution-dependent bound: $O(\log T)$.	Lemmas A.18, A.22

Table 2. List of lemmas and their dependencies in the proof.

- *Law of Conditional Probability:* $\Pr[A \wedge B] = \Pr[A \mid B] \cdot \Pr[B]$.
- *Law of Total Probability: if* $\{B_n : n = 1, 2, \cdots\}$ *is a set of disjoint events whose union is the entire sample space, then*

$$\Pr[A] = \sum_n \Pr[A \wedge B_n].$$

Fact A.2 (Azuma-Hoeffding Inequality in Euclidean Space (Theorem 1.8 of (Hayes, 2003))). *Let* $\mathbf{X} = (X_0, \cdots, X_n)$ *be a very-weak martingale, which is defined for every* i, $\mathbb{E}[X_i \mid X_{i-1}] = X_{i-1}$, *and it takes values in Euclidean Space, such that for every* i, $X_i \in \mathbb{R}^d$. *Suppose* $X_0 = 0$, *and for* $i = 1, \cdots, n$, $\|X_i - X_{i-1}\|_2 \leq 1$. *Then, for every* $\epsilon > 0$,

$$\Pr[\|X_n\|_2 \geq \epsilon] < 2e^{1-\frac{(\epsilon-1)^2}{2n}} < 2e^2 e^{-\frac{\epsilon^2}{2n}}. \tag{1}$$

We can use the preceding fact to obtain the concentration property of outcomes during exploration.

Lemma A.3 (Concentration during exploration). *After the exploration round* $i = 1, 2, \cdots, j$ *at* $t_1, t_2, \cdots, t_j$ *respectively, we use the inverse to get* $\tilde{\mathbf{v}}_i = \mathbf{I}(M_\sigma, \vec{\mathbf{y}}(t_i)) = M_\sigma^+ \vec{\mathbf{y}}_i$ *and their mean is* $\hat{\boldsymbol{\nu}}^{(j)} = \frac{1}{j}\sum_{i=1}^{j} \tilde{\mathbf{v}}_i$. *Then,* $\forall \gamma > 0$*:*

$$\Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \geq \gamma\right] \leq 2e^2 \exp\{-\frac{\gamma^2 j}{2\beta_\sigma^2}\}. \tag{2}$$

Proof. For each i, let X_i be the sequence sum satisfying $X_i = \sum_{\ell=1}^{i} \frac{\boldsymbol{\nu} - \tilde{\mathbf{v}}_\ell}{\beta_\sigma}$, where $\mathbb{E}\left[\frac{\boldsymbol{\nu} - \tilde{\mathbf{v}}_i}{\beta_\sigma}\right] = 0$, and $\|\frac{\boldsymbol{\nu} - \tilde{\mathbf{v}}_i}{\beta_\sigma}\|_2 \leq 1$. So $X_i - X_{i-1} = \frac{\boldsymbol{\nu} - \tilde{\mathbf{v}}_i}{\beta_\sigma}$ implies $\|X_i - X_{i-1}\|_2 \leq 1$. And we know that $\tilde{\mathbf{v}}_i$ is independent of the previous inverse $\tilde{\mathbf{v}}_1, \cdots, \tilde{\mathbf{v}}_{i-1}$, so it holds that

$$\begin{aligned}
\mathbb{E}[X_i \mid X_{i-1}] - X_{i-1} =& \mathbb{E}[X_i - X_{i-1} \mid X_{i-1}] && (3)\\
=& \mathbb{E}\left[\frac{\boldsymbol{\nu} - \tilde{\mathbf{v}}_i}{\beta_\sigma} \,\middle|\, X_{i-1}\right] && (4)\\
=& \mathbb{E}\left[\frac{\boldsymbol{\nu} - \tilde{\mathbf{v}}_i}{\beta_\sigma}\right] = 0. && (5)
\end{aligned}$$

Therefore, $\mathbf{X} = (X_0, \cdots, X_n)$ satisfies the definition of a very-weak martingale. Apply Fact A.2, and it will achieve the bound $\forall \epsilon > 0, \Pr[\|X_j\|_2 \geq \epsilon] < 2e^2 e^{-\frac{\epsilon^2}{2j}}$. Let $\gamma = \epsilon\frac{\beta_\sigma}{j}$, as $\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)} = \frac{\beta_\sigma}{j} X_j$, we will get:

$$\forall \gamma > 0,\ \Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \geq \gamma\right] < 2e^2 \exp\{-\frac{\gamma^2 j}{2\beta_\sigma^2}\}. \tag{6}$$

□

Under a predetermined outcome distribution p with mean outcome vector $\boldsymbol{\nu}$ and $x^* = \operatorname{argmax}_{x \in \mathcal{X}} \overline{r}(x, \boldsymbol{\nu})$, in the main text we define the gap:

$$\begin{aligned}
\Delta_x =& \overline{r}(x^*, \boldsymbol{\nu}) - \overline{r}(x, \boldsymbol{\nu}), && (7)\\
\Delta_{\max} =& \max\{\Delta_x : x \in \mathcal{X}\}, && (8)\\
\Delta_{\min} =& \min\{\Delta_x : x \in \mathcal{X}, \Delta_x > 0\}. && (9)
\end{aligned}$$

Definition A.4 (Good actions / bad actions). *Based on the distance to the optimal action, define good actions and bad actions as:*

$$\begin{aligned}
\mathcal{X}^{\textit{Good}} \triangleq& \{x : \forall x \in \mathcal{X}, \Delta_x = 0\} && (10)\\
\mathcal{X}^{\textit{Bad}} \triangleq& \{x : \forall x \in \mathcal{X}, \Delta_x > 0\}. && (11)
\end{aligned}$$

Therefore, $\mathcal{X} = \mathcal{X}^{\textit{Good}} \cup \mathcal{X}^{\textit{Bad}}$. *Moreover,* $x^* \in \mathcal{X}^{\textit{Good}}$. *(*$x^*$ *is unique if and only if* $|\mathcal{X}^{\textit{Good}}| = 1$.*)*

Definition A.5 (Events of finding a good action / bad action)**.** *Define $\hat{x}(t) \triangleq \operatorname{argmax}_{x\in\mathcal{X}} \overline{r}(x, \hat{\boldsymbol{\nu}}(t-1))$ as the current optimal action at time t. Let $\mathcal{F}^{Bad}(t)$ be the event that fails to choose the optimal action at time t. Formally, $\mathcal{F}^{Bad}(t)$ and its complement event are:*

$$\mathcal{F}^{Bad}(t) \triangleq \left\{\hat{x}(t) \in \mathcal{X}^{Bad}\right\} \tag{12}$$

$$\mathcal{F}^{Good}(t) \triangleq \left\{\hat{x}(t) \in \mathcal{X}^{Good}\right\}. \tag{13}$$

To build the connection with the exploration round j, we define the time-invariant event $\mathcal{F}^{Bad}_{(j)}$ as the event in which the algorithm fails to choose the optimal action after j rounds of exploration:

$$\mathcal{F}^{Bad}_{(j)} \triangleq \left\{\hat{x}^{(j)} \in \mathcal{X}^{Bad}\right\} \tag{14}$$

$$\mathcal{F}^{Good}_{(j)} \triangleq \left\{\hat{x}^{(j)} \in \mathcal{X}^{Good}\right\}, \tag{15}$$

where $\hat{x}^{(j)} = \operatorname{argmax}_{x\in\mathcal{X}} \overline{r}(x, \hat{\boldsymbol{\nu}}^{(j)})$.

By definition, it is always true that $\mathcal{F}^{\text{Bad}}_{(n_\sigma(t-1))} = \mathcal{F}^{\text{Bad}}(t)$ and $\mathcal{F}^{\text{Good}}_{(n_\sigma(t-1))} = \mathcal{F}^{\text{Good}}(t)$.

Definition A.6 (Estimated gap and real gap)**.** *For any pair of action $x_i, x_j \in \mathcal{X}$, defined the gap of estimated reward between x_i, x_j as*

$$\hat{\delta}_{x_i,x_j}(t) \triangleq \overline{r}(x_i, \hat{\boldsymbol{\nu}}(t-1)) - \overline{r}(x_j, \hat{\boldsymbol{\nu}}(t-1)),$$

and the gap of real reward between them as

$$\delta_{x_i,x_j} \triangleq \overline{r}(x_i, \boldsymbol{\nu}) - \overline{r}(x_j, \boldsymbol{\nu}).$$

Lemma A.7 (Bound of the gap)**.** *For any pair of action $x_i, x_j \in \mathcal{X}$, we establish the inequality over time t as:*

$$|\hat{\delta}_{x_i,x_j}(t) - \delta_{x_i,x_j}| \leq 2L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2. \tag{16}$$

Proof.

$$\begin{aligned}
|\hat{\delta}_{x_i,x_j}(t) - \delta_{x_i,x_j}| =& |(\overline{r}(x_i, \hat{\boldsymbol{\nu}}(t-1)) - \overline{r}(x_i, \boldsymbol{\nu})) - (\overline{r}(x_j, \hat{\boldsymbol{\nu}}(t-1)) - \overline{r}(x_j, \boldsymbol{\nu}))| && (17)\\
\leq& |\overline{r}(x_i, \hat{\boldsymbol{\nu}}(t-1)) - \overline{r}(x_i, \boldsymbol{\nu})| + |\overline{r}(x_j, \hat{\boldsymbol{\nu}}(t-1)) - \overline{r}(x_j, \boldsymbol{\nu})| && (18)\\
\leq& L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2 + L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2 && (19)\\
=& 2L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2. && (20)
\end{aligned}$$

□

Lemma A.8 (Small error in estimation)**.** *Given time t, for $f_{\mathcal{X}}(t) = \ln t + 2\ln|\mathcal{X}|$, $\alpha = \frac{8L^2\beta_\sigma^2}{a^2}$, and $a > 0$,*

$$\forall\gamma > 0,\ \Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2 \geq \gamma\sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}}\right] \leq \frac{2e^2}{|\mathcal{X}|^2} N(t)^{1-\frac{4\gamma^2L^2}{a^2}}. \tag{21}$$

Proof. As the time of exploration equals to the counter $n_\sigma(t-1)$ and $\hat{\boldsymbol{\nu}}(t-1) = \hat{\boldsymbol{\nu}}^{(n_\sigma(t-1))}$, we have:

$$\Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2 \geq \gamma \cdot \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}}\right] \tag{22}$$

$$= \sum_{j=1}^{t-1} \Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(n_\sigma(t-1))}\|_2 \geq \gamma \cdot \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}} \ \wedge \ n_\sigma(t-1) = j\right] \tag{23}$$

$$= \sum_{j=1}^{t-1} \Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \geq \gamma \cdot \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{j}} \ \wedge \ n_\sigma(t-1) = j\right] \tag{24}$$

$$\leq \sum_{j=1}^{t-1} \Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \geq \gamma \cdot \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{j}}\right] \tag{25}$$

$$\leq \sum_{j=1}^{t-1} 2e^2 \exp\left\{-\left(\gamma \cdot \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{j}}\right)^2 \frac{j}{2\beta_\sigma^2}\right\} \qquad \{\text{Lemma A.3}\} \tag{26}$$

$$\leq \sum_{j=1}^{t-1} 2e^2 \exp\left\{-\frac{\gamma^2 \alpha f_{\mathcal{X}}(t)}{2\beta_\sigma^2}\right\} \tag{27}$$

$$\leq \sum_{j=1}^{t-1} 2e^2 N(t)^{-\frac{\gamma^2\alpha}{2\beta_\sigma^2}}. \tag{28}$$

As $\alpha = \frac{8L^2\beta_\sigma^2}{a^2}$ and $a > 0$, the probability is:

$$\Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2 \geq \gamma\sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}}\right] \leq 2e^2(t-1) \cdot N(t)^{-\frac{\gamma^2\alpha}{2\beta_\sigma^2}} \leq \frac{2e^2}{|\mathcal{X}|^2} N(t)^{1-\frac{4\gamma^2L^2}{a^2}}. \tag{29}$$

□

Definition A.9 (Events of exploration or exploitation)**.** *In Algorithm GCB, for any time t, we can define three events, namely the beginning of exploration $\mathcal{E}^{\textit{Explore}}(t)$, in the process of exploration $\mathcal{E}^{\textit{InExplore}}(t)$ and exploitation $\mathcal{E}^{\textit{Exploit}}(t)$. They are mutually exclusive, and $\mathcal{E}^{\textit{Explore}}(t) \vee \mathcal{E}^{\textit{InExplore}}(t) \vee \mathcal{E}^{\textit{Exploit}}(t)$ is always true. Formally, it is:*

$$\mathcal{E}^{\textit{Explore}}(t) \triangleq \{state(t) = \textit{begin\_exploration}\} \tag{30}$$

$$\mathcal{E}^{\textit{InExplore}}(t) \triangleq \{state(t) = \textit{in\_exploration}\} \tag{31}$$

$$\mathcal{E}^{\textit{Exploit}}(t) \triangleq \{state(t) = \textit{exploitation}\}. \tag{32}$$

Definition A.10 (Events related to confidence interval)**.** *In Line 8 of Algorithm GCB, we can define the event for the first condition where the gap of estimated optimal action and other actions is larger than confidence interval as $\mathcal{L}_{CI}(t)$ at time t, i.e.,*

$$\mathcal{L}_{CI}(t) = \left\{\forall x \in \mathcal{X} \setminus \{\hat{x}(t)\}, \hat{\delta}_{\hat{x}(t),x}(t) > \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}}\right\}. \tag{33}$$

And its complement event is:

$$\mathcal{L}_{CI}^{\mathrm{c}}(t) = \left\{\exists x \in \mathcal{X} \setminus \{\hat{x}(t)\}, \hat{\delta}_{\hat{x}(t),x}(t) \leq \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}}\right\}. \tag{34}$$

Remark 1. In Algorithm GCB, we know that the first condition of Line 8 is true, if and only if $\mathcal{L}_{CI}(t) = \left\{\forall x \in \mathcal{X} \setminus \{\hat{x}(t)\}, \hat{\delta}_{\hat{x}(t),x}(t) > \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}}\right\}$ occurs. Thus, we use the equivalent event in the following proof to make it clearer.

Definition A.11. *For simplicity, suppose $\alpha = \frac{8L^2\beta_\sigma^2}{a^2}$, constant $a > 0$ and $\theta > 0$, then we can define two threshold functions:*

$$\eta(t) = t^{\frac{2}{3}} f_{\mathcal{X}}(t) \tag{35}$$

$$\mu(t) = (1+\theta a)^2 \frac{\alpha f_{\mathcal{X}}(t)}{\Delta_{\min}^2}. \tag{36}$$

Note that $\eta(t)$ and $\mu(t)$ are values, not random variables.

Proposition A.12. *If $t > T_0 = \frac{(1+\theta a)^3 \alpha^{\frac{3}{2}}}{\Delta_{\min}^3}$, then $\mu(t) < \eta(t)$. (It can be verified by the definition.)*

Lemma A.13 (Exploration Ceiling)**.** *Let $\alpha \geq \frac{8L^2\beta_\sigma^2}{a^2}$ and $a > 0$. For any time t, if the exploration counter $n_\sigma(t-1) > \eta(t)$, the algorithm will play exploitation surely, i.e.,*

$$\Pr\left[\mathcal{E}^{\textit{Explore}}(t) \;\middle|\; n_\sigma(t-1) > \eta(t)\right] = 0. \tag{37}$$

Proof. If $n_\sigma(t-1) > \eta(t)$, then Line 8 of Algorithm GCB will be true because of its second condition. According to the algorithm, it will not go to exploration phase, so we know that

$$\Pr\left[\mathcal{E}^{\text{Explore}}(t) \;\middle|\; n_\sigma(t-1) > \eta(t)\right] = 0, \tag{38}$$

which restricts $n_\sigma(t-1)$ to no larger than $\lfloor \eta(t) \rfloor + 1$ at any time t. □

Lemma A.14 (Low failure probability of the confidence interval)**.** *Let $f_{\mathcal{X}}(t) = \ln t + 2\ln|\mathcal{X}|$, $\alpha \geq \frac{8L^2\beta_\sigma^2}{a^2}$ and $0 < a \leq \frac{1}{\sqrt{3}}$. For any time t, the probability that both choosing bad action and the gap is larger than confidence interval satisfies:*

$$\Pr\left[\mathcal{L}_{CI}(t) \;\wedge\; \mathcal{F}^{\textit{Bad}}(t)\right] \leq \frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2}. \tag{39}$$

Proof. The definition of $\mathcal{F}^{\text{Bad}}(t) = \{\hat{x}(t) \in \mathcal{X}^{\text{Bad}}\}$ implies $\exists x^* \in \mathcal{X} \setminus \{\hat{x}(t)\}$. Their gap is

$$\begin{aligned}
\hat{\delta}_{\hat{x}(t),x^*}(t) =& \overline{r}(\hat{x}(t), \hat{\boldsymbol{\nu}}(t-1)) - \overline{r}(x^*, \hat{\boldsymbol{\nu}}(t-1)) && (40)\\
\leq& \overline{r}(\hat{x}(t), \hat{\boldsymbol{\nu}}(t-1)) - \overline{r}(x^*, \hat{\boldsymbol{\nu}}(t-1)) + \overline{r}(x^*, \boldsymbol{\nu}) - \overline{r}(\hat{x}(t), \boldsymbol{\nu}) && \{\text{Definition of } x^*\} \quad (41)\\
\leq& |\overline{r}(\hat{x}(t), \hat{\boldsymbol{\nu}}(t-1)) - \overline{r}(\hat{x}(t), \boldsymbol{\nu})| + |\overline{r}(x^*, \boldsymbol{\nu}) - \overline{r}(x^*, \hat{\boldsymbol{\nu}}(t-1))| && (42)\\
\leq& 2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2. && (43)
\end{aligned}$$

Thus, we can write the probability as:

$$\begin{aligned}
& \Pr\left[\mathcal{L}_{CI}(t) \;\wedge\; \mathcal{F}^{\text{Bad}}(t)\right] && (44)\\
=& \Pr\left[\forall x \in X \setminus \{\hat{x}(t)\}, \hat{\delta}_{\hat{x}(t),x} \geq \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}} \;\wedge\; \mathcal{F}^{\text{Bad}}(t)\right] && (45)\\
\leq& \Pr\left[\hat{\delta}_{\hat{x}(t),x^*} \geq \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}} \;\wedge\; \mathcal{F}^{\text{Bad}}(t)\right] && (46)\\
\leq& \Pr\left[\hat{\delta}_{\hat{x}(t),x^*} \geq \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}}\right] && (47)\\
\leq& \Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(n_\sigma(t-1))}\|_2 \geq \frac{1}{2L}\sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}}\right] && (48)\\
\leq& \frac{2e^2}{|\mathcal{X}|^2} N(t)^{1-\frac{1}{a^2}} && \{\text{Lemma A.8 with } \gamma = \frac{1}{2L}\} \quad (49)\\
\leq& \frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2}. && \{0 < a \leq \frac{1}{\sqrt{3}}\} \quad (50)
\end{aligned}$$

□

A.1. Distribution-independent bound

Lemma A.15. *For any $\epsilon > 0$, $\forall j = 1, 2, \cdots, t-1$, when the algorithm has played $n_\sigma(t-1) = j$ rounds' exploitation at time t, the probability of incurring penalty ϵ satisfies*

$$\Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ n_\sigma(t-1) = j\right] \leq 2e^2 \exp\left\{-\frac{j \cdot \epsilon^2}{8L^2\beta_\sigma^2}\right\}. \tag{51}$$

Proof. $\Delta_{\hat{x}(t)}$ is the real gap of reward between x^* and $\hat{x}(t)$:

$$\begin{aligned}
\Delta_{\hat{x}(t)} =& \delta_{x^*,\hat{x}(t)} && (52)\\
\leq& \delta_{x^*,\hat{x}(t)} + \hat{\delta}_{\hat{x}(t),x^*}(t) && \{\text{Definition of } \hat{x}(t)\} \quad (53)\\
\leq& \delta_{x^*,\hat{x}(t)} - \hat{\delta}_{x^*,\hat{x}(t)}(t) && (54)\\
\leq& |\overline{r}(x^*, \boldsymbol{\nu}) - \overline{r}(x^*, \hat{\boldsymbol{\nu}}(t-1))| + |\overline{r}(\hat{x}(t), \boldsymbol{\nu}) - \overline{r}(\hat{x}(t), \hat{\boldsymbol{\nu}}(t-1))| && (55)\\
\leq& 2L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2 = 2L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(n_\sigma(t-1))}\|_2 && (56)
\end{aligned}$$

When $n_\sigma(t-1) = j$, we can conclude that the probability of incurring a large penalty is:

$$\begin{aligned}
\Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ n_\sigma(t-1) = j\right] \leq& \Pr\left[2L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(n_\sigma(t-1))}\|_2 \geq \epsilon \ \wedge \ n_\sigma(t-1) = j\right] && (57)\\
\leq& \Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \geq \frac{\epsilon}{2L}\right] && (58)\\
\leq& 2e^2 \exp\left\{-\frac{j \cdot \epsilon^2}{8L^2\beta_\sigma^2}\right\}. && \{\text{Lemma A.3}\} \quad (59)
\end{aligned}$$

□

In Algorithm GCB, we know that the exploitation is penalized with respect to the regret only if it chooses a bad action and exploits it simultaneously, i.e., $\mathcal{F}^{\text{Bad}}(t)$ and $\mathcal{E}^{\text{Exploit}}(t)$ are both satisfied. When the algorithm chooses exploitation at time t, the regret at that time will be $\mathbb{E}\left[\Delta_{\hat{x}(t)} \cdot \mathbb{I}\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right]\right]$.

Lemma A.16 (Penalty of exploitation)**.** *$\forall \epsilon > 0$, Algorithm GCB with $f_{\mathcal{X}}(t) = \ln t + 2\ln|\mathcal{X}|$, $\alpha = \frac{8L^2\beta_\sigma^2}{a^2}$, $0 < a \leq \frac{1}{\sqrt{3}}$, and $\eta(t) = t^{\frac{2}{3}} f_{\mathcal{X}}(t)$, the penalty in the exploitation phase at time t will be in expectation:*

$$\mathbb{E}\left[\Delta_{\hat{x}(t)} \cdot \mathbb{I}\left[\mathcal{F}^{Bad}(t) \ \wedge \ \mathcal{E}^{Exploit}(t)\right]\right] \leq \epsilon + \Delta_{\max} \cdot \left(\frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2} + 2e^2 \exp\left\{-\frac{\eta(t) \cdot \epsilon^2}{8L^2\beta_\sigma^2}\right\}\right). \tag{60}$$

Proof. $\forall \epsilon > 0$, the expectation satisfies:

$$\begin{aligned}
&\mathbb{E}\left[\Delta_{\hat{x}(t)} \cdot \mathbb{I}\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right]\right] && (61)\\
=&\mathbb{E}\left[\Delta_{\hat{x}(t)} \ \middle| \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] \cdot \Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] && (62)\\
=&\mathbb{E}\left[\Delta_{\hat{x}(t)} \ \middle| \ \Delta_{\hat{x}(t)} < \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] \cdot \Pr\left[\Delta_{\hat{x}(t)} < \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right]\\
&+ \mathbb{E}\left[\Delta_{\hat{x}(t)} \ \middle| \ \Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] \cdot \Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] && (63)\\
\leq&\epsilon \cdot \Pr\left[\Delta_{\hat{x}(t)} < \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] + \Delta_{\max} \cdot \Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] && (64)\\
\leq&\epsilon + \Delta_{\max} \cdot \Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right]. && (65)
\end{aligned}$$

By definition, exploration event $\mathcal{E}^{\text{Exploit}}(t) = \{\mathcal{L}_{CI}(t) \vee n_\sigma(t) > \eta(t)\}$ happens when no other action is in the gap $\mathcal{L}_{CI}(t)$ or the counter $n_\sigma(t) > \eta(t)$. And we know that $n_\sigma(t)$ is no larger than $\lfloor \eta(t) \rfloor + 1$, because it is a hard constraint implied

by Lemma A.13. Therefore, the probability in the second term is the joint of these two events:

$$\begin{aligned}
&\Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] && (66)\\
=&\Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ (\mathcal{L}_{CI}(t) \ \vee \ n_\sigma(t-1) > \eta(t))\right] && (67)\\
=&\Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{L}_{CI}(t) \ \wedge \ n_\sigma(t-1) \leq \eta(t)\right]\\
&+ \Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) > \eta(t)\right] && (68)\\
\leq&\Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{L}_{CI}(t)\right] + \Pr\left[\Delta_{\hat{x}(t)} \geq \epsilon \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) = \lfloor \eta(t) \rfloor + 1\right] && (69)\\
\leq&\frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2} + 2e^2 \exp\left\{-\frac{\eta(t)\cdot\epsilon^2}{8L^2\beta_\sigma^2}\right\}. && \{\text{Lemma A.14 and A.15}\} \quad (70)
\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left[\Delta_{\hat{x}(t)} \cdot \mathbb{I}\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right]\right] \leq \epsilon + \Delta_{\max} \cdot \left(\frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2} + 2e^2 \exp\left\{-\frac{\eta(t)\cdot\epsilon^2}{8L^2\beta_\sigma^2}\right\}\right). \tag{71}$$

□

Theorem 4.1 (in the main text): (Distribution-independent bound)**.** Let $f_{\mathcal{X}}(t) = \ln t + 2\ln|\mathcal{X}|$, and $\alpha = 24L^2\beta_\sigma^2$. The distribution-independent regret bound of Algorithm GCB is:

$$R(T) \leq R_{\max}|\sigma| \cdot T^{\frac{2}{3}}(\ln T + 2\ln|\mathcal{X}|) + \frac{8}{3} L\beta_\sigma T^{\frac{2}{3}} + R_{\max}\left(|\sigma| + \frac{4e^2}{|\mathcal{X}|^4}\right). \tag{72}$$

Proof. From the algorithm, we know that it either plays actions in the exploration phase or in the exploitation phase. The exploration phase will take time $|\sigma|$ to finish, and its penalty is $\sum_{x\in\sigma}\Delta_x$. And the penalty of playing exploitation is $\Delta_{\hat{x}(t)}$ at each time step t.

$$R(T) = \sum_{\forall x\in\sigma} \Delta_x \mathbb{E}\left[n_\sigma(T)\right] + \sum_{t=1}^{T} \mathbb{E}\left[\Delta_{\hat{x}(t)} \cdot \mathbb{I}\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right]\right]. \tag{73}$$

From Lemma A.13, we can infer that if the exploration counter $n_\sigma(t) > \eta(t) = t^{\frac{2}{3}} f_{\mathcal{X}}(t)$, it will no longer play exploration. Therefore, the expected number of rounds of exploration satisfies $\mathbb{E}\left[n_\sigma(T)\right] \leq T^{\frac{2}{3}} f_{\mathcal{X}}(T) + 1$, so the regret for exploration is

$$\sum_{\forall x\in\sigma} \Delta_x \mathbb{E}\left[n_\sigma(T)\right] \leq \sum_{\forall x\in\sigma} \Delta_x \cdot \left(T^{\frac{2}{3}} f_{\mathcal{X}}(T) + 1\right). \tag{74}$$

Let $\epsilon = 4L\beta_\sigma t^{-\frac{1}{3}}$, then $\eta(t) = t^{\frac{2}{3}} f_{\mathcal{X}}(t)$ and $\frac{\eta(t)\epsilon^2}{8L^2\beta_\sigma^2} = 2f_{\mathcal{X}}(t)$. Therefore, we can apply Lemma A.16 to get the regret of exploitation part:

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\Delta_{\hat{x}(t)} \cdot \mathbb{I}\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right]\right] && (75)\\
\leq&\sum_{t=1}^{T}\left[\epsilon + \Delta_{\max}\cdot\left(\frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2} + 2e^2\exp\left\{-\frac{\eta(t)\cdot\epsilon^2}{8L^2\beta_\sigma^2}\right\}\right)\right] && (76)\\
\leq&\sum_{t=1}^{T}\left[4L\beta_\sigma t^{-\frac{1}{3}} + \Delta_{\max}\cdot\left(\frac{2e^2}{|\mathcal{X}|^2} + 2e^2\right) N(t)^{-2}\right] && (77)\\
=&\frac{8}{3} L\beta_\sigma T^{\frac{2}{3}} + \Delta_{\max}\cdot\left(\frac{2e^2}{|\mathcal{X}|^2} + 2e^2\right)\frac{1}{|\mathcal{X}|^4}. && (78)
\end{aligned}$$

Therefore, we will have

$$R(T) \leq \sum_{\forall x\in\sigma} \Delta_x \cdot T^{\frac{2}{3}} \cdot (\ln T + 2\ln|\mathcal{X}|) + \frac{8}{3} L\beta_\sigma T^{\frac{2}{3}} + \left(\sum_{\forall x\in\sigma}\Delta_x + \frac{4e^2}{|\mathcal{X}|^4}\Delta_{\max}\right). \tag{79}$$

As Δ_x and $\Delta_{\max}$ is bounded by $R_{\max}$ under any distribution, we conclude that:

$$R(T) \leq R_{\max}|\sigma| \cdot T^{\frac{2}{3}} \cdot (\ln T + 2\ln|\mathcal{X}|) + \frac{8}{3}L\beta_\sigma T^{\frac{2}{3}} + R_{\max}\left(|\sigma| + \frac{4e^2}{|\mathcal{X}|^4}\right). \tag{80}$$

□

A.2. Distribution-dependent Bound

Under a predetermined outcome distribution p, the minimum gap between optimal action and sub-optimal action is $\Delta_{\min}$. It follows that:

Lemma A.17 (Condition of choosing optimal action)**.** *Suppose we have played exploration round j, at time t. If $b \geq 1$, $\forall j \geq b \cdot \frac{8L^2\beta_\sigma^2}{\Delta_{\min}^2} f_{\mathcal{X}}(t)$, Algorithm GCB will choose the optimal action with high probability:*

$$\forall j \geq b \cdot \frac{8L^2\beta_\sigma^2}{\Delta_{\min}^2} f_{\mathcal{X}}(t), \quad \Pr\left[\mathcal{F}_{(j)}^{Bad}\right] \leq \frac{e^2}{t \cdot N(t)^{b-1}}. \tag{81}$$

Proof. According to the definition, $\mathcal{F}_{(j)}^{\text{Bad}}$ only occurs only if one sub-optimal action has the largest estimated reward.

$$\begin{aligned}
&\Pr\left[\mathcal{F}_{(j)}^{\text{Bad}}\right] && (82)\\
\leq &\Pr\left[\exists x^b \in \mathcal{X}^{\text{Bad}}, \forall x^g \in \mathcal{X}^{\text{Good}}, \overline{r}(x^g, \hat{\boldsymbol{\nu}}^{(j)}) \leq \overline{r}(x^b, \hat{\boldsymbol{\nu}}^{(j)})\right] && (83)\\
\leq &\Pr\left[\exists x^b \in \mathcal{X}^{\text{Bad}}, \exists x^g \in \mathcal{X}^{\text{Good}}, \overline{r}(x^g, \hat{\boldsymbol{\nu}}^{(j)}) \leq \overline{r}(x^b, \hat{\boldsymbol{\nu}}^{(j)})\right] && (84)\\
\leq &\sum_{\substack{x^b \in \mathcal{X}^{\text{Bad}}\\ x^g \in \mathcal{X}^{\text{Good}}}} \Pr\left[\overline{r}(x^g, \hat{\boldsymbol{\nu}}^{(j)}) - \overline{r}(x^b, \hat{\boldsymbol{\nu}}^{(j)}) \leq 0\right] && \{\text{Union bound}\} \quad (85)\\
\leq &\sum_{\substack{x^b \in \mathcal{X}^{\text{Bad}}\\ x^g \in \mathcal{X}^{\text{Good}}}} \left(\Pr\left[\overline{r}(x^g, \boldsymbol{\nu}) - \overline{r}(x^g, \hat{\boldsymbol{\nu}}^{(j)}) \geq \frac{\Delta_{\min}}{2}\right] + \Pr\left[\overline{r}(x^b, \boldsymbol{\nu}) - \overline{r}(x^b, \hat{\boldsymbol{\nu}}^{(j)}) < -\frac{\Delta_{\min}}{2}\right]\right) && (86)\\
\leq &\sum_{\substack{x^b \in \mathcal{X}^{\text{Bad}}\\ x^g \in \mathcal{X}^{\text{Good}}}} \left(\Pr\left[|\overline{r}(x^g, \boldsymbol{\nu}) - \overline{r}(x^g, \hat{\boldsymbol{\nu}}^{(j)})| \geq \frac{\Delta_{\min}}{2}\right] + \Pr\left[|\overline{r}(x^b, \boldsymbol{\nu}) - \overline{r}(x^b, \hat{\boldsymbol{\nu}}^{(j)})| > \frac{\Delta_{\min}}{2}\right]\right) && (87)\\
\leq &\sum_{\substack{x^b \in \mathcal{X}^{\text{Bad}}\\ x^g \in \mathcal{X}^{\text{Good}}}} 2\Pr\left[L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \geq \frac{\Delta_{\min}}{2}\right]. && (88)
\end{aligned}$$

Thus, by Lemma A.3, it is

$$\begin{aligned}
\Pr\left[\mathcal{F}_{(j)}^{\text{Bad}}\right] \leq &\sum_{x^b \in \mathcal{X}^{\text{Bad}}, x^g \in \mathcal{X}^{\text{Good}}} 2\Pr\left[L\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 > \frac{\Delta_{\min}}{2}\right] && (89)\\
\leq &\sum_{x^b \in \mathcal{X}^{\text{Bad}}, x^g \in \mathcal{X}^{\text{Good}}} 4e^2 \exp\{-\frac{j\Delta_{\min}^2}{8L^2\beta_\sigma^2}\} && (90)\\
\leq &4e^2|\mathcal{X}^{\text{Bad}}| \cdot |\mathcal{X}^{\text{Good}}| \cdot \exp\{-\frac{j\Delta_{\min}^2}{8L^2\beta_\sigma^2}\} && (91)\\
\leq &e^2|\mathcal{X}|^2 \cdot \exp\{-\frac{j\Delta_{\min}^2}{8L^2\beta_\sigma^2}\}. && \{|\mathcal{X}^{\text{Bad}}| + |\mathcal{X}^{\text{Good}}| = |\mathcal{X}|\} \quad (92)
\end{aligned}$$

Therefore, if $j \geq b \cdot \frac{8L^2\beta_\sigma^2}{\Delta_{\min}^2} f_{\mathcal{X}}(t)$, $b \geq 1$, we can conclude:

$$\Pr\left[\mathcal{F}_{(j)}^{\text{Bad}}\right] \leq \frac{e^2|\mathcal{X}|^2}{N(t)^b} = \frac{e^2}{tN(t)^{b-1}}. \tag{93}$$

□

Lemma A.18 (Exploit the Optimal Action)**.** *Let* $\alpha \geq \frac{8L^2\beta_\sigma^2}{a^2}$, $0 < a \leq \frac{1}{\sqrt{3}}$ *and* $\theta \geq \sqrt{3}$*. For any time* $t > T_0$*, the probability of* $\mathcal{F}^{Bad}(t)$ *and playing exploitation in Algorithm GCB is:*

$$\Pr\left[\mathcal{E}^{Exploit}(t) \ \wedge \ \mathcal{F}^{Bad}(t)\right] \leq 3e^2 N(t)^{-2}. \tag{94}$$

Proof. If $t > T_0$, and $\mathcal{E}^{\text{Exploit}}(t) = \{\mathcal{L}_{CI}(t) \ \vee \ n_\sigma(t-1) > \eta(t)\}$, we can write the probability of exploitation as:

$$\begin{aligned}
&\Pr\left[\mathcal{E}^{\text{Exploit}}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t)\right] && (95)\\
=&\Pr\left[\mathcal{E}^{\text{Exploit}}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) > \eta(t)\right] + \Pr\left[\mathcal{E}^{\text{Exploit}}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) \leq \eta(t)\right] && (96)\\
\leq&\Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) > \eta(t)\right] + \Pr\left[\mathcal{L}_{CI}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) \leq \eta(t)\right] && (97)\\
\leq&\Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) > \eta(t)\right] + \Pr\left[\mathcal{L}_{CI}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t)\right] && (98)\\
\leq&\Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) > \eta(t)\right] + \frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2}. && \{\text{Lemma A.14}\}\\
& && (99)
\end{aligned}$$

Since we know that $n_\sigma(t-1) > \eta(t)$, $0 < a \leq \frac{1}{\sqrt{3}}$ and $\theta \geq 1$, then

$$n_\sigma(t-1) > \eta(t) > \mu(t) = \frac{(1+\theta a)^2}{a^2} \cdot \frac{8L^2\beta_\sigma^2 f_{\mathcal{X}}(t)}{\Delta_{\min}^2} > 3 \cdot \frac{8L^2\beta_\sigma^2 f_{\mathcal{X}}(t)}{\Delta_{\min}^2}.$$

By Lemma A.17, the following inequality holds:

$$\begin{aligned}
&\Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) > \eta(t)\right] && (100)\\
\leq& \sum_{j=\eta(t)}^{t-1} \Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) = j\right] && (101)\\
=& \sum_{j=\eta(t)}^{t} \Pr\left[\mathcal{F}^{\text{Bad}}_{(n_\sigma(t-1))} \ \wedge \ n_\sigma(t-1) = j\right] && (102)\\
\leq& \sum_{j=\eta(t)}^{t-1} \Pr\left[\mathcal{F}^{\text{Bad}}_{(j)}\right] && (103)\\
\leq& \sum_{j=\eta(t)}^{t-1} \frac{e^2}{tN(t)^2} && \{\text{Lemma A.17 with } b=3\} \quad (104)\\
\leq& e^2 N(t)^{-2}. && (105)
\end{aligned}$$

Therefore, we can get:

$$\Pr\left[\mathcal{E}^{\text{Exploit}}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t)\right] \leq e^2 N(t)^{-2} + \frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2} \leq 3e^2 N(t)^{-2}. \tag{106}$$

□

Lemma A.19 (The exploration probability will drop)**.** *Suppose the instance has unique optimal action under distribution* p*, i.e.,* $|\mathcal{X}^{Good}| = 1$*. Let* $\alpha \geq \frac{8L^2\beta_\sigma^2}{a^2}$, $0 < a \leq \frac{1}{\sqrt{3}}$*. For any time* $t > T_0$*, when* $n_\sigma(t-1) \geq \mu(t) = (1+\theta a)^2 \frac{\alpha f_{\mathcal{X}}(t)}{\Delta_{\min}^2}$ *where* $\theta \geq \sqrt{3}$*, and the probability of* $\mathcal{F}^{Good}(t)$ *and exploration happening simultaneously is:*

$$\Pr\left[\mathcal{E}^{Explore}(t) \ \wedge \ \mathcal{F}^{Good}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(t)\right] \leq \frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2}. \tag{107}$$

Proof. By definition, the event that exploration happens at time t is $\mathcal{E}^{\text{Explore}}(t) = \{\mathcal{L}_{CI}^c(t) \ \wedge \ n_\sigma \leq \eta(t)\}$. When $t > T_0$, it is true that $\eta(t) > \mu(t)$.

On one hand, if $n_\sigma(t-1) > \eta(t)$, then by Lemma A.13, we know that

$$\begin{aligned}
&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ n_\sigma(t-1) > \eta(t)\right] && (108)\\
=&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \middle| \ n_\sigma(t-1) > \eta(t)\right] \cdot \Pr\left[n_\sigma(t-1) > \eta(t)\right] && (109)\\
=&0. && (110)
\end{aligned}$$

On the other hand, for $\mu(t) \le n_\sigma(t-1) \le \eta(t)$, whether to play exploration only depends on the event $\mathcal{L}_{CI}^c(t)$. If $\mathcal{F}^{\text{Good}}(t) = \left\{\hat{x}(t) \in \mathcal{X}^{\text{Good}}\right\}$ and with the assumption that $|\mathcal{X}^{\text{Good}}| = 1$, we know that $\mathcal{X}^{\text{Good}} \cap (\mathcal{X} \setminus \{\hat{x}(t)\}) = \varnothing$. So the gap at time t is, $\forall x \in \mathcal{X} \setminus \{\hat{x}(t)\}$,

$$\begin{aligned}
\hat{\delta}_{\hat{x}(t),x}(t) = \hat{\delta}_{x^*,x}(t) \ge& \delta_{x^*,x} - |\hat{\delta}_{x^*,x}(t) - \delta_{x^*,x}| && (111)\\
\ge& \Delta_{\min} - |\hat{\delta}_{x^*,x}(t) - \delta_{x^*,x}| && \{\Delta_{\min} \text{ is the minimum gap}\} \quad (112)\\
\ge& \Delta_{\min} - 2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2. && \{\text{Lemma A.7}\} \quad (113)
\end{aligned}$$

And we also know that if $n_\sigma(t-1) \ge \mu(t) = (1+\theta a)^2 \frac{\alpha f_{\mathcal{X}}(t)}{\Delta_{\min}^2}$,

$$\sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}} \le \frac{\Delta_{\min}}{1+\theta a}, \tag{114}$$

thus we can get

$$\begin{aligned}
&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ \mu(t) \le n_\sigma(t-1) \le \eta(t)\right] && (115)\\
=&\Pr\left[\exists x \in \mathcal{X} \setminus \{\hat{x}(t)\}, \hat{\delta}_{\hat{x}(t),x}(t) \le \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}} \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ \mu(t) \le n_\sigma(t-1) \le \eta(t)\right] && (116)\\
\le&\Pr\left[\Delta_{\min} - 2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2 \le \sqrt{\frac{\alpha f_{\mathcal{X}}(t)}{n_\sigma(t-1)}} \ \wedge \ \mu(t) \le n_\sigma(t-1) \le \eta(t)\right] && (117)\\
\le&\Pr\left[\Delta_{\min} - 2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}(t-1)\|_2 \le \frac{\Delta_{\min}}{1+\theta a} \ \wedge \ \mu(t) \le n_\sigma(t-1) \le \eta(t)\right] && (118)\\
=&\Pr\left[2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(n_\sigma(t-1))}\|_2 \ge \frac{\theta a}{1+\theta a}\Delta_{\min} \ \wedge \ \mu(t) \le n_\sigma(t-1) \le \eta(t)\right] && (119)\\
=&\Pr\left[2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(n_\sigma(t-1))}\|_2 \ge \frac{\theta a}{1+\theta a}\Delta_{\min} \ \wedge \ \mu(t) \le n_\sigma(t-1) \le \eta(t)\right] && (120)\\
=&\sum_{j=\mu(t)}^{\eta(t)} \Pr\left[2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(n_\sigma(t-1))}\|_2 \ge \frac{\theta a}{1+\theta a}\Delta_{\min} \ \wedge \ n_\sigma(t-1) = j\right] && (121)\\
\le&\sum_{j=\mu(t)}^{\eta(t)} \Pr\left[2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \ge \frac{\theta a}{1+\theta a}\Delta_{\min} \ \wedge \ n_\sigma(t-1) = j\right] && (122)\\
\le&\sum_{j=\mu(t)}^{\eta(t)} \Pr\left[2L \cdot \|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \ge \frac{\theta a}{1+\theta a}\Delta_{\min}\right]. && (123)
\end{aligned}$$

Let $\alpha \ge \frac{8L^2\beta_\sigma^2}{a^2}$, $0 < a \le \frac{1}{\sqrt{3}}$. For $j = \mu(t), \cdots, \eta(t)$ and $\mu(t) = (1+\theta a)^2 \frac{\alpha f_{\mathcal{X}}(t)}{\Delta_{\min}^2}$, recall Lemma A.3, then we have:

$$\begin{aligned}
&\Pr\left[\|\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}^{(j)}\|_2 \ge \frac{1}{2L} \cdot \frac{\theta a}{1+\theta a}\Delta_{\min}\right] && (124)\\
\le&2e^2 \exp\left\{-\frac{(\theta a)^2\Delta_{\min}^2}{(1+\theta a)^2}\frac{j}{8L^2{\beta_\sigma}^2}\right\} && \{\text{Lemma A.3}\} \quad (125)\\
\le&2e^2 \exp\left\{-\theta^2 f_{\mathcal{X}}(t)\right\} && \{j \ge \mu(t)\} \quad (126)\\
\le&2e^2 N(t)^{-\theta^2}. && (127)
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(t)\right] && (128)\\
=&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ n_\sigma(t-1) > \eta(t)\right] \\
&+ \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ \mu(t) \leq n_\sigma(t-1) \leq \eta(t)\right] && (129)\\
\leq&0 + \sum_{j=\mu(t)}^{\eta(t)} 2e^2 N(t)^{-\theta^2} && (130)\\
\leq&2e^2 t \cdot N(t)^{-\theta^2} && (131)\\
\leq&\frac{2e^2}{|\mathcal{X}|^2} N(t)^{1-\theta^2} && (132)\\
\leq&\frac{2e^2}{|\mathcal{X}|^2} N(t)^{-2}. && \{\text{Let } \theta \geq \sqrt{3}\} \quad (133)
\end{aligned}$$

□

When the instance has a unique optimal action x^* under distribution p, the following lemmata ensures that exploration will not continue endlessly, thus it will switch to exploitation gradually. For simplicity, we consider the case that the exploration round has already reached $\mu(T)$ at given time T.

Lemma A.20 (Switch to exploitation gradually)**.** *Suppose the instance has a unique optimal action x^* under distribution p. Given time T, if for time $i \leq T$ the exploration rounds $n_\sigma(i) = \mu(T)$ has already been satisfied, where $\mu(T) = (1+\theta a)^2 \frac{\alpha f_{\mathcal{X}}(T)}{\Delta_{\min}^2}$, $0 < a \leq \frac{1}{\sqrt{3}}$. $\theta \geq \sqrt{3}$. Then $\forall t, \max\{i+1, T_0\} \leq t \leq T$, the probability of playing exploration is:*

$$\Pr\left[\mathcal{E}^{\textit{Explore}}(t) \ \wedge \ n_\sigma(i) = \mu(T)\right] \leq 4e^2 N(t)^{-2}. \tag{134}$$

Proof. As $n_\sigma(i) = \mu(T)$, we know that

$$n_\sigma(i) = \mu(T) \Rightarrow n_\sigma(t-1) \geq n_\sigma(i) = \mu(T), \tag{135}$$

which implies that the event $n_\sigma(i) = \mu(T)$ is the subset of the event $n_\sigma(t-1) \geq \mu(T)$.

$$\begin{aligned}
&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ n_\sigma(i) = \mu(T)\right] && (136)\\
\leq&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(T)\right] && (137)\\
=&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(T)\right] \\
&+ \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(T)\right] && (138)\\
\leq&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(T)\right] && (139)\\
&+ \Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(T)\right]. && (140)
\end{aligned}$$

From Lemma A.19, the first part is

$$\begin{aligned}
&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(T)\right] && (141)\\
\leq&\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{F}^{\text{Good}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(t)\right] && (142)\\
\leq&\frac{2e^2}{|\mathcal{X}|^2} \cdot N(t)^{-2}. && (143)
\end{aligned}$$

For the second part, as $0 < a \leq \frac{1}{\sqrt{3}}$ and $\theta \geq \sqrt{3}$, we can get

$$\mu(T) = (1+\theta a)^2 \frac{\alpha f_{\mathcal{X}}(T)}{\Delta_{\min}^2} \geq (1+\theta a)^2 \frac{\alpha f_{\mathcal{X}}(t)}{\Delta_{\min}^2} \geq \frac{(1+\theta a)^2}{a^2} \cdot \frac{8L^2\beta_\sigma^2 f_{\mathcal{X}}(t)}{\Delta_{\min}^2} > 3 \cdot \frac{8L^2\beta_\sigma^2 f_{\mathcal{X}}(t)}{\Delta_{\min}^2}. \tag{144}$$

Thus, by using Lemma A.17, it is

$$\Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) \geq \mu(T)\right] \tag{145}$$

$$= \sum_{j=\mu(T)}^{t-1} \Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ n_\sigma(t-1) = j\right] \tag{146}$$

$$= \sum_{j=\mu(T)}^{t-1} \Pr\left[\mathcal{F}^{\text{Bad}}_{(n_\sigma(t-1))} \ \wedge \ n_\sigma(t-1) = j\right] \tag{147}$$

$$= \sum_{j=\mu(T)}^{t-1} \Pr\left[\mathcal{F}^{\text{Bad}}_{(j)} \ \wedge \ n_\sigma(t-1) = j\right] \tag{148}$$

$$\leq \sum_{j=\mu(T)}^{t-1} \Pr\left[\mathcal{F}^{\text{Bad}}_{(j)}\right] \tag{149}$$

$$\leq t \cdot \frac{2e^2}{tN(t)^2} \qquad \{\text{Lemma A.17 with } b=3\} \tag{150}$$

$$\leq 2e^2 N(t)^{-2}. \tag{151}$$

Therefore, we can get

$$\Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ n_\sigma(i) = \mu(T)\right] \leq \frac{2e^2}{|\mathcal{X}|^2} N^{-2} + 2e^2 N^{-2} \leq 4e^2 N^{-2}. \tag{152}$$

□

For counter n_σ, the following definition characterizes its first occurrence to be k.

Definition A.21. *Given k, for any t, we define the event that $n_\sigma(t) = k$ and $n_\sigma(t-1) = k-1$ as $\mathcal{G}_k(t)$, i.e.,*

$$\mathcal{G}_k(t) = \{n_\sigma(t) = k \ \wedge \ n_\sigma(t-1) = k-1\}.$$

Lemma A.22 (Exploration Numbers)**.** *Let $\mu(T) = (1+\theta a)^2 \frac{\alpha f_{\mathcal{X}}(T)}{\Delta_{\min}^2}$, $0 < a \leq \frac{1}{\sqrt{3}}$ and $\theta \geq \sqrt{3}$. If under distribution p, there is a unique optimal action, i.e., $|\mathcal{X}^{Good}| = 1$, then the expected exploration round at time T ($T_0 \leq T$) is:*

$$\mathbb{E}\left[n_\sigma(T)\right] \leq \mu(T) + \frac{4e^2}{|\mathcal{X}|^4} \ln(T+1) + 1 + \sum_{t=1}^{T_0} \Pr\left[\mathcal{E}^{Explore}(t)\right]. \tag{153}$$

Proof. Note that it takes $|\sigma|$ time steps to play exploration and then to increase n_σ by 1. $\mathcal{E}^{\text{FinExplore}}(t)$ is the event that the algorithm finishes one round of exploration and updates n_σ at time t. Then, we have $\mathcal{E}^{\text{FinExplore}}(t) = \mathcal{E}^{\text{Explore}}(t - |\sigma| + 1)$ and $\forall t = 1, 2, \cdots, |\sigma| - 1$, $\Pr\left[\mathcal{E}^{\text{FinExplore}}(t)\right] = 0$, meaning that the event never happens for $t < |\sigma|$. By definition, we can get:

$$\mathbb{E}\left[n_\sigma(T)\right] = \sum_{t=1}^{T} \Pr\left[\mathcal{E}^{\text{FinExplore}}(t)\right] = \sum_{t=|\sigma|}^{T} \Pr\left[\mathcal{E}^{\text{FinExplore}}(t)\right] = \sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t)\right]. \tag{154}$$

Because the accumulation of exploration rounds is $n_\sigma(T)$, therefore its expected number can be:

$$\mathbb{E}\left[n_\sigma(T)\right] = \sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t)\right] \tag{155}$$

$$= \sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ n_\sigma(T) < \mu(T)\right] + \sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ n_\sigma(T) \geq \mu(T)\right]. \tag{156}$$

The following inequality ensures that the first part is not large:

$$\sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ n_\sigma(T) < \mu(T)\right] \tag{157}$$

$$\leq \sum_{t=1}^{T-|\sigma|+1} \Pr\left[n_\sigma(T) < \mu(T)\right] \cdot \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \middle| \ n_\sigma(T) < \mu(T)\right] \tag{158}$$

$$\leq \Pr\left[n_\sigma(T) < \mu(T)\right] \cdot \sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \middle| \ n_\sigma(T) < \mu(T)\right] \tag{159}$$

$$\leq \Pr\left[n_\sigma(T) < \mu(T)\right] \cdot \sum_{t=1}^{T} \Pr\left[\mathcal{E}^{\text{FinExplore}}(t) \ \middle| \ n_\sigma(T) < \mu(T)\right] \tag{160}$$

$$= \Pr\left[n_\sigma(T) < \mu(T)\right] \cdot \mathbb{E}\left[n_\sigma(T) \mid n_\sigma(T) < \mu(T)\right] \tag{161}$$

$$\leq \Pr\left[n_\sigma(T) < \mu(T)\right] \cdot \mu(T). \tag{162}$$

We know the counter n_σ could only increase by 1 at a time. For this reason, if the value of $n_\sigma(T)$ exceeds $\mu(T)$ at time T, this event must happen within $t = \mu(T), \cdots, T$. Thus, the occurrence of $\mu(T)$ is equivalent to the union of events $\left\{\bigvee_{i=\mu(T)}^{T} \mathcal{G}_{\mu(T)}(i)\right\}$. By definition, each event $\mathcal{G}_{\mu(T)}(i), \forall i = \mu(T), \cdots, T$, is mutually exclusive. Therefore, we have $\{n_\sigma(T) \geq \mu(T)\} = \left\{\bigvee_{i=\mu(T)}^{T} \mathcal{G}_{\mu(T)}(i)\right\}$, and the second part is:

$$\sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ n_\sigma(T) \geq \mu(T)\right] \tag{163}$$

$$= \sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \left(\bigvee_{i=\mu(T)}^{T} \mathcal{G}_{\mu(T)}(i)\right)\right] \tag{164}$$

$$= \sum_{t=1}^{T-|\sigma|+1} \Pr\left[\bigvee_{i=\mu(T)}^{T} \left(\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{G}_{\mu(T)}(i)\right)\right] \tag{165}$$

$$\leq \sum_{t=1}^{T-|\sigma|+1} \sum_{i=\mu(T)}^{T} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{G}_{\mu(T)}(i)\right] \qquad \{\text{Union bound}\} \tag{166}$$

$$\leq \sum_{i=\mu(T)}^{T} \sum_{t=1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{G}_{\mu(T)}(i)\right] \tag{167}$$

$$\leq \sum_{i=\mu(T)}^{T} \sum_{t=1}^{i} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{G}_{\mu(T)}(i)\right] + \sum_{i=\mu(T)}^{T} \sum_{t=i+1}^{T-|\sigma|+1} \Pr\left[\mathcal{E}^{\text{Explore}}(t) \ \wedge \ \mathcal{G}_{\mu(T)}(i)\right]. \tag{168}$$

Now we will prove that the first term is in $O(\mu(T))$:

$$\sum_{i=\mu(T)}^{T}\sum_{t=1}^{i}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \wedge\ \mathcal{G}_{\mu(T)}(i)\right] \tag{169}$$

$$=\sum_{i=\mu(T)}^{T}\sum_{t=1}^{i}\Pr\left[\mathcal{G}_{\mu(T)}(i)\right]\cdot\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \middle|\ \mathcal{G}_{\mu(T)}(i)\right] \tag{170}$$

$$=\sum_{i=\mu(T)}^{T}\Pr\left[\mathcal{G}_{\mu(T)}(i)\right]\cdot\sum_{t=1}^{i}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \middle|\ \mathcal{G}_{\mu(T)}(i)\right] \tag{171}$$

$$=\sum_{i=\mu(T)}^{T}\Pr\left[\mathcal{G}_{\mu(T)}(i)\right]\cdot\sum_{t=1}^{i+|\sigma|-1}\Pr\left[\mathcal{E}^{\text{FinExplore}}(t)\ \middle|\ \mathcal{G}_{\mu(T)}(i)\right] \tag{172}$$

$$=\sum_{i=\mu(T)}^{T}\Pr\left[\mathcal{G}_{\mu(T)}(i)\right]\cdot\mathbb{E}\left[n_\sigma(i+|\sigma|-1)\ \middle|\ \mathcal{G}_{\mu(T)}(i)\right] \tag{173}$$

$$\le\sum_{i=\mu(T)}^{T}\Pr\left[\mathcal{G}_{\mu(T)}(i)\right]\cdot\mathbb{E}\left[n_\sigma(i)+1\ \middle|\ \mathcal{G}_{\mu(T)}(i)\right] \qquad \{n_\sigma(i+|\sigma|-1)\le n_\sigma(i)+1\} \tag{174}$$

$$\le\Pr\left[n_\sigma(T)\ge\mu(T)\right]\cdot(\mu(T)+1)\,, \qquad \{\text{Mutually exclusive}\} \tag{175}$$

Since $\mathcal{G}_{\mu(T)}(i)=\{n_\sigma(i)=\mu(T)\ \wedge\ n_\sigma(i-1)=\mu(T)-1\}$, we can write the second term as:

$$\sum_{i=\mu(T)}^{T}\sum_{t=i+1}^{T-|\sigma|+1}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \wedge\ \mathcal{G}_{\mu(T)}(i)\right] \tag{176}$$

$$\le\sum_{i=\mu(T)}^{T}\sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \wedge\ \mathcal{G}_{\mu(T)}(i)\right]+\sum_{i=\mu(T)}^{T}\sum_{t=\max\{i+1,T_0+1\}}^{T-|\sigma|+1}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \wedge\ \mathcal{G}_{\mu(T)}(i)\right] \tag{177}$$

$$\le\sum_{i=\mu(T)}^{T}\sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \wedge\ \mathcal{G}_{\mu(T)}(i)\right]+\sum_{i=\mu(T)}^{T}\sum_{t=\max\{i+1,T_0+1\}}^{T-|\sigma|+1}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \wedge\ n_\sigma(i)=\mu(T)\right] \tag{178}$$

$$\le\sum_{t=1}^{T_0}\sum_{i=\mu(T)}^{T}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \wedge\ \mathcal{G}_{\mu(T)}(i)\right]+\sum_{i=\mu(T)}^{T}\sum_{t=\max\{i+1,T_0+1\}}^{T-|\sigma|+1}4e^2N(t)^{-2} \qquad \{\text{Lemma A.20}\} \tag{179}$$

$$\le\sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\ \wedge\ \left(\bigvee_{i=\mu(T)}^{T}\mathcal{G}_{\mu(T)}(i)\right)\right]+\frac{4e^2}{|\mathcal{X}|^4}\cdot\int_{i=\mu(T)}^{T}\int_{t=i}^{T}t^{-2}\cdot\mathrm{d}t\cdot\mathrm{d}i \qquad \{\text{Mutually exclusive}\} \tag{180}$$

$$\le\sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\right]+\frac{4e^2}{|\mathcal{X}|^4}\int_{i=\mu(T)}^{T}\frac{1}{i}di \tag{181}$$

$$\le\sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\right]+\frac{4e^2}{|\mathcal{X}|^4}\ln T. \tag{182}$$

Therefore, we can get

$$\mathbb{E}\left[n_\sigma(T)\right]\le\mu(T)+1+\frac{4e^2}{|\mathcal{X}|^4}\ln T+\sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\right]. \tag{183}$$

□

Theorem 4.2 (in the main text): (Distribution-dependent bound)**.** For Algorithm GCB, let $f_{\mathcal{X}}(t) = \ln t + 2\ln|\mathcal{X}|$, $\alpha = 24L^2\beta_\sigma^2$. If the instance has a unique optimal action under outcome distribution p and mean outcome vector $\boldsymbol{\nu}$, the distribution-dependent regret bound of Algorithm GCB is:

$$R(T) \leq \sum_{x\in\sigma} \Delta_x \cdot \left[\frac{96L^2\beta_\sigma^2}{\Delta_{\min}^2} \cdot (\ln T + 2\ln|\mathcal{X}|) + \frac{4e^2}{|\mathcal{X}|^4}\ln T + 1\right] + \Delta_{\max} \cdot \left(\frac{3e^2}{|\mathcal{X}|^4} + \frac{941L^3\beta_\sigma^3}{\Delta_{\min}^3}\right), \tag{184}$$

where $\sum_{\forall x\in\sigma}\Delta_x$, $\Delta_{\max}$ and $\Delta_{\min}$ are problem-specific constants under the distribution p.

Proof. If we penalize each time the algorithm plays a sub-optimal action by $\Delta_{\max}$, then the regret function is composed of exploration and exploitation:

$$R(T) \leq \sum_{x\in\sigma}\Delta_x \cdot \mathbb{E}\left[n_\sigma(T)\right] + \Delta_{\max}\cdot\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{I}\left[\mathcal{E}^{\text{Exploit}}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t)\right]\right] \tag{185}$$

$$\leq \sum_{x\in\sigma}\Delta_x \cdot \mathbb{E}\left[n_\sigma(T)\right] + \Delta_{\max}\cdot\sum_{t=1}^{T}\Pr\left[\mathcal{E}^{\text{Exploit}}(t) \ \wedge \ \mathcal{F}^{\text{Bad}}(t)\right]. \tag{186}$$

Suppose it has unique optimal action $|\mathcal{X}^{\text{Good}}| = 1$, from Lemma A.22 the expected rounds of exploration are:

$$\mathbb{E}\left[n_\sigma(T)\right] \leq (1+\theta a)^2\frac{\alpha f_{\mathcal{X}}(T)}{\Delta_{\min}^2} + 1 + \frac{4e^2}{|\mathcal{X}|^4}\ln(T+1) + \sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\right]. \tag{187}$$

The regret of exploitation phase can be inferred from Lemma A.18 that:

$$\Delta_{\max}\cdot\sum_{t=1}^{T}\Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right] \tag{188}$$

$$\leq \Delta_{\max}\cdot\left(\sum_{t=T_0+1}^{T}\frac{3e^2}{N(t)^2} + \sum_{t=1}^{T_0}\Pr\left[\mathcal{F}^{\text{Bad}}(t) \ \wedge \ \mathcal{E}^{\text{Exploit}}(t)\right]\right) \tag{189}$$

$$\leq \Delta_{\max}\cdot\left(\frac{3e^2}{|\mathcal{X}|^4}\sum_{t=T_0+1}^{T}\frac{1}{t^2} + \sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Exploit}}(t)\right]\right) \tag{190}$$

$$\leq \Delta_{\max}\cdot\left(\frac{3e^2}{|\mathcal{X}|^4} + \sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Exploit}}(t)\right]\right). \tag{191}$$

Since for $t = 1, 2, \cdots, T_0$, we perform either exploration or exploitation, the regret is no worse than $\Delta_{\max}\cdot T_0$, that is:

$$\sum_{x\in\sigma}\Delta_x\cdot\sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Explore}}(t)\right] + \Delta_{\max}\cdot\sum_{t=1}^{T_0}\Pr\left[\mathcal{E}^{\text{Exploit}}(t)\right] \leq \Delta_{\max}\cdot T_0. \tag{192}$$

Thus, for $f_{\mathcal{X}}(t) = \ln t + 2\ln|\mathcal{X}|$, $\alpha \geq \frac{8L^2\beta_\sigma^2}{a^2}$, $0 < a \leq \frac{1}{\sqrt{3}}$ and $\theta \geq \sqrt{3}$,

$$R(T) \leq \sum_{x\in\sigma}\Delta_x\cdot\left[\frac{(1+\theta a)^2\alpha}{\Delta_{\min}^2}\cdot(\ln T + 2\ln|\mathcal{X}|) + \frac{4e^2}{|\mathcal{X}|^4}\ln T + 1\right] + \Delta_{\max}\cdot\left(\frac{3e^2}{|\mathcal{X}|^4} + T_0\right), \tag{193}$$

where $T_0 = \frac{(1+\theta a)^3\alpha^{\frac{3}{2}}}{\Delta_{\min}^3}$.

Let $a = \frac{1}{\sqrt{3}}$, $\theta = \sqrt{3}$, and $\alpha = 24L^2\beta_\sigma^2$. As a conclusion, we will get:

$$R(T) \leq \sum_{x\in\sigma}\Delta_x\cdot\left[\frac{96L^2\beta_\sigma^2}{\Delta_{\min}^2}\cdot(\ln T + 2\ln|\mathcal{X}|) + \frac{4e^2}{|\mathcal{X}|^4}\ln T + 1\right] + \Delta_{\max}\cdot\left(\frac{3e^2}{|\mathcal{X}|^4} + \frac{941L^3\beta_\sigma^3}{\Delta_{\min}^3}\right). \tag{194}$$

□

B. An Example of M_σ and Global Observer Set Construction for $1 < s < N$ in the Crowdsourcing Application

In this section, we provide an example of constructing the stacked matrix M_σ and the global observer set in the crowdsourcing application when we require $1 < s < N$, where s is the number of matched worker-task pairs used for reporting the feedback. Recall that the feedback for a matching is the simple summation of these s matched worker-task pairs. This implies that for each matching $\mathbf{x}$, the transformation matrix $M_{\mathbf{x}}$ contains a single row with exactly s 1s and all other entries are 0, and $M_{\mathbf{x}} \cdot \mathbf{x} = s$.

As an illustration, consider the case that both N and M are divisible by $s+1$. Then we can construct a full-rank square matrix M_σ such that, after rearranging the columns of M_σ, it is a block diagonal matrix with each block B being an $(s+1)$-by-$(s+1)$ square matrix with 0 in the diagonal entries and 1 as off-diagonal entries. The following is an illustration of such a matrix for the case of $s+1 = N = M = 3$.

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

It is clear that this M_σ is full column rank. To recover the NM actions (matchings) corresponding to the NM rows, we map each block B to a matching that matches $s+1$ workers to $s+1$ tasks such that these matchings share no common edges. This can be done in the following way.

We partition N workers into $N/(s+1)$ groups of size $s+1$ each, and partition M tasks into $M/(s+1)$ groups of size $s+1$ each. Taking any group W of $s+1$ workers and any group U of $s+1$ tasks, we can find $s+1$ non-overlapping matchings between W and U by rotation: in the j-th matching, the i-th worker is matched with the $(i+j \mod s+1)$-th task. Since we have $NM/(s+1)^2$ worker-task group pairs, and each group pair generates $s+1$ non-overlapping matchings, in total we have $NM/(s+1)$ non-overlapping matchings, and we map these matches to the $NM/(s+1)$ blocks in the rearranged matrix M_σ. The above construction implies that we can find NM actions to form a global observer set, in which each action is a matching of $s+1$ workers to $s+1$ tasks, and each matching returns an aggregate performance feedback of s worker-task pairs in the matching. Thus the assumption on the existence of the global observer set holds and the set can be constructed easily.

The error bound β_σ for the above constructed M_σ is more complicated to analyze, but by our empirical evaluation using Matlab, we believe that it is also a low-degree polynomial in N and M.

References

Hayes, Thomas P. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 2003.