Supplementary Material for Paper ID: 867

A Theoretical Preliminaries

We require the following two versions of the Chernoff-Hoeffeding inequality in our proof.

Lemma A.1 (Chernoff-Hoeffeding inequality)**.** *Suppose $X_1, \ldots, X_T$ are independent random variables taking values in the interval $[0,1]$, and let $X = \sum_{t \in [T]} X_t$ and $\overline{X} = \frac{\sum_{t \in [T]} X_t}{T}$. Then for any $\varepsilon \geq 0$ the following holds:*

$$a) \ \ \mathbb{P}\{X - E[X] \geq \varepsilon\} \leq e^{\frac{-2\varepsilon^2}{T}} ,$$

$$b) \ \ \mathbb{P}\{\overline{X} - E[\overline{X}] \geq \varepsilon\} \leq e^{-2\varepsilon^2 T} \ .$$

B Omitted Proofs

B.1 Proof of Theorem 1

For $B \in \mathbb{N}$, let $\varepsilon = \sqrt{\frac{2}{pB}\log(16pMB)}$. Also, let $L = \min_{t \in \mathbb{N}}\{2\sqrt{\frac{1}{t}\log 16pMt} \leq p\}$. Note that L is a *finite* constant dependent on p and M, and that for all $B \geq L$

$$\varepsilon \leq \sqrt{p/2} \ . \tag{4}$$

In this proof, i indexes the set $[M]$, and x indexes the set $\{0,1\}$. Recall $p_{i,x} = \mathbb{P}\{X_i = x\}$ and $p_i = p_{i,1}$. Also note that `OBS-ALG` plays the arm a_0 for B rounds. For $i \in [M]$, let $X_i(t)$ be the value of X_i sampled in round $t \in [B]$. For all i, x, let

$$\widehat{p}_{i,x} = \frac{\sum_{t \in [B]} \mathbb{1}\{X_i(t) = x\}}{B} \ , \quad \text{and}$$

$$\widehat{\mu}_{i,x} = \frac{\sum_{t \in [B]} Y_t \cdot \mathbb{1}\{X_i(t) = x\}}{\sum_{t \in [B]} \mathbb{1}\{X_i(t) = x\}} \ ,$$

where Y_t is value of Y sampled in round t. Notice that $\widehat{\mu}_{i,x}$ is the empirical estimate of $\mu_{i,x}$ computed by `OBS-ALG` at the end of B rounds. Similarly the empirical estimate of μ_0, denoted $\widehat{\mu}_0$, is computed by `OBS-ALG` at the end of B rounds as follows:

$$\widehat{\mu}_0 = \frac{\sum_{t \in [B]} Y_t}{B} \ .$$

Finally, also let $\widehat{p}_i = \widehat{p}_{i,1}$. The proof of the theorem is completed using the following lemma.

Lemma B.1. *At the end of B rounds played by `OBS-ALG` the following hold:*

1. $\mathbb{P}\{|\widehat{\mu}_0 - \mu_0| \geq \varepsilon\} \leq 2e^{-2\varepsilon^2 B} \leq 4e^{-\varepsilon^2 pB}$,
2. *For any fixed (i,x)* $\quad \mathbb{P}\Big\{\widehat{p}_{i,x}B \leq \frac{pB}{2}\Big\} \leq 2e^{-\varepsilon^2 pB}$,
3. *For any fixed (i,x)* $\quad \mathbb{P}\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon\} \leq 4e^{-\varepsilon^2 pB}$.

Proof. 1) Part 1 directly follows from Lemma A.1.

2) Observe that $E[\widehat{p}_i] = p_i$, and hence from Lemma A.1, for an $i \in [M]$ at the end of B rounds we have

$$\mathbb{P}\left\{|(\widehat{p}_i - p_i)B| \geq \varepsilon B\sqrt{\frac{p}{2}}\right\} \leq 2e^{-\varepsilon^2 pB} \ . \tag{5}$$

Since $\varepsilon \leq \sqrt{p/2}$ (from Equation 4), $\varepsilon B\sqrt{\frac{p}{2}} \leq \frac{pB}{2}$. This implies

$$\frac{pB}{2} \leq pB - \varepsilon B\sqrt{\frac{p}{2}} \ . \tag{6}$$

Hence from Equations 5 and 6, for a fixed (i,x) the following holds:

$$\mathbb{P}\Big\{\widehat{p}_{i,x}B \leq \frac{pB}{2}\Big\} \leq 2e^{-\varepsilon^2 pB} \ .$$

3) Notice that $\widehat{p}_{i,x}B$ is the number of times X_i was sampled as x in B rounds. In particular, part 2 of Lemma B.1 bounds the probability that the number of times X_i was sampled as x is small. We use this to prove part 3. First observe that from Lemma A.1 we have

$$\mathbb{P}\left\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon \middle| \widehat{p}_{i,x}B > \frac{pB}{2}\right\} \leq 2e^{-\varepsilon^2 pB} \ . \tag{7}$$

In particular, Equation 7 bounds the error probability of estimating $\widehat{\mu}_{i,x}$ conditioned on the event that X_i has been sampled as x sufficiently many times. Next by law of total probability, for any fixed (i, x),

$$\begin{aligned}\mathbb{P}\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon\} &= \mathbb{P}\left\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon \middle| \widehat{p}_{i,x}B > \frac{pB}{2}\right\} \cdot \mathbb{P}\left\{\widehat{p}_{i,x}B > \frac{pB}{2}\right\} \\ &\quad + \mathbb{P}\left\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon \middle| \widehat{p}_{i,x}B \leq \frac{pB}{2}\right\} \cdot \mathbb{P}\left\{\widehat{p}_{i,x}B \leq \frac{pB}{2}\right\} \\ \mathbb{P}\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon\} &\leq \mathbb{P}\left\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon \middle| \widehat{p}_{i,x}B > \frac{pB}{2}\right\} + \mathbb{P}\left\{\widehat{p}_{i,x}B \leq \frac{pB}{2}\right\} \ .\end{aligned}$$

Hence, from Equation 7 and part 2 of Lemma B.1 we have

$$\mathbb{P}\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon\} \leq 4e^{-\varepsilon^2 pB} \ .$$

□

Let U_0 be the event that $|\widehat{\mu}_0 - \mu_0| \leq \varepsilon$, and for any i, x let $U_{i,x}$ be the event $|\widehat{\mu}_{i,x} - \mu_{i,x}| \leq \varepsilon$. Also let $U = (\cap_{i,x} U_{i,x}) \cap U_0$, $\overline{U}$ denote the compliment of U. Then applying union bound on the events in part 1 and 3 in Lemma B.1, we have that

$$\mathbb{P}\{\overline{U}\} \leq (2M + 1) \cdot 4e^{-\varepsilon^2 pB}$$

Hence, we have that

$$\mathbb{P}\{U\} \geq 1 - (8M + 4)e^{-\varepsilon^2 pB} \geq 1 - 16Me^{-\varepsilon^2 pB}.$$

Let $a^* = \arg\max_{a \in \mathcal{A}}(\mu_a)$. Note that if event $\overline{U}$ holds then the simple regret of `OBS-ALG` , $r_{\texttt{OBS-ALG}}\ (B) \leq 1$. On the other hand, if the event U holds, and a_B is the arm output by the algorithm, then $r_{\texttt{OBS-ALG}}\ (B) = \mu_{a^*} - \mu_{a_B} \leq 2\varepsilon$. Setting $\delta = 16Me^{-\varepsilon^2 pB}$, and substituting the value of ε, we have $\delta = \frac{1}{16Mp^2B^2}$. Hence, the expected simple regret is at most:

$$\delta + \sqrt{\frac{8}{pB}\log(16pMB)} = \frac{1}{16Mp^2B^2} + \sqrt{\frac{8}{pB}\log(16pMB)} = O\left(\sqrt{\frac{1}{pB}\log(pMB)}\right) \ . \tag{8}$$

B.2 Proof of Theorem 2

For convenience, we denote $m(\mathbf{p})$ and $m(\widehat{\mathbf{p}})$ as m and $\widehat{m}$ respectively. Throughout the proof we assume that B is such that: a) $B \geq \max(\gamma m, pM)$ and b) $B \geq \max(\frac{16}{p^2}\log\frac{2MB}{\gamma m}, \frac{16}{p^2}\log 2pMB)$. Note that the two constraints hold for sufficiently large B. To begin with observe that if $\gamma = \theta(\frac{1}{p \cdot m(\mathbf{p})})$ then $O\left(\sqrt{\frac{1}{pB}\log(pMB)}\right) = O\left(\sqrt{\frac{\gamma m}{B}\log\frac{MB}{\gamma m}}\right)$. Hence, it is sufficient to show that if $\gamma \leq \frac{1}{5p \cdot m(\mathbf{p})}$ then the expected simple regret of γ-`NB-ALG` is $O\left(\sqrt{\frac{\gamma m}{B}\log\frac{MB}{\gamma m}}\right)$ and if $\gamma \geq \frac{5}{p \cdot m(\mathbf{p})}$ then the expected simple regret of γ-`NB-ALG` is $O\left(\sqrt{\frac{1}{pB}\log(pMB)}\right)$. Theorem 2 is proved using Lemmas B.2 and B.3.

Lemma B.2. *Let $\widehat{p}_{i,1} = \widehat{p}_i$ and $F = \mathbb{1}\{$At the end of $B/2$ rounds there is an $i \in [M]$ such that $|\widehat{p}_i - p_i| \geq \frac{p}{4}\}$. Then $\mathbb{P}\{F = 1\} \leq 2Me^{-\frac{p^2}{16}B}$.*

Proof. Let $F_i = \mathbb{1}\{$At the end of $B/2$ rounds $|\widehat{p}_i - p_i| \geq \frac{p}{4}\}$. Then from Lemma A.1,

$$\mathbb{P}\{F_i = 1\} \leq 2e^{-\frac{p^2}{16}B} \ .$$

Taking union bound over $F_i = 1$ for $i \in [M]$, we have $\mathbb{P}\{F = 1\} \leq 2Me^{-\frac{p^2}{16}B}$. □

The following lemma is similar to Lemma 8 in Lattimore et al. (2016).

Lemma B.3. *Let F be as in Lemma B.2, and let $I = \mathbb{1}\{$At the end of $B/2$ rounds $\frac{2m(\mathbf{p})}{5} \leq m(\widehat{\mathbf{p}}) \leq 2m(\mathbf{p})\}$. Then $F = 0$ implies $I = 1$, and in particular, $\mathbb{P}\{I = 1\} \geq 1 - 2Me^{-\frac{p^2}{16}B}$.*

Proof. We are interested in the quantity $\min_{x\in\{0,1\}} p_{i,x}$ for each $i \in [M]$. Without loss of generality, let us assume $\min_{x\in\{0,1\}} p_{i,x} = p_{i,1} = p_i$ for each $i \in [M]$, and also $p_1 \leq p_2 \leq \ldots \leq p_M \leq \frac{1}{2}$. Note that $F = 0$ implies after $B/2$ rounds for all $i \in [M]$ $|\widehat{p}_i - p_i| \leq \frac{p}{4}$. Now, from the definition of $m(\mathbf{p})$ we know that there is an $\ell \leq m$ such that the following is true: for $i > \ell$, $p_i \geq \frac{1}{m}$. Further, we can also conclude that $p \leq \frac{1}{m-1}$ (otherwise $m(\mathbf{p}) = m - 1$). Hence, $\widehat{p}_i \geq p_i - \frac{p}{4} \geq \frac{1}{m} - \frac{1}{4(m-1)}$. Hence for $i > \ell$, $\widehat{p}_i \geq \frac{3m-4}{4m(m-1)} \geq \frac{1}{2m}$ (since $m \geq 2$). Since $\ell \leq m$, we have $|\{j \mid \widehat{p}_j < \frac{1}{2m}\}| \leq 2m$. This implies $\widehat{m} \leq 2m$. To prove the other inequality, observe that for each $i \leq m$, we have $p_i \leq \frac{1}{m-1}$ (otherwise, $m(\mathbf{p}) \leq m - 1$). Then, $\widehat{p}_i \leq p_i + \frac{p}{4} \leq \frac{1}{m-1} + \frac{1}{4(m-1)} \leq \frac{5}{4(m-1)} \leq \frac{5}{2m}$. Hence for $i \leq m$, $\widehat{p}_i \leq \frac{5}{2m}$. This implies $\widehat{m} \geq \frac{2m}{5}$. □

From Lemmas B.2 and B.3 it follows that if $F = 0$ then at the end of $B/2$ rounds the following holds:

$$\frac{p}{2} \leq \widehat{p} \leq \frac{3p}{2} \quad \text{and} \quad \frac{2m}{5} \leq \widehat{m} \leq 2m$$

This implies that if $F = 0$ then at then end of $B/2$ rounds the following holds:

$$\frac{p \cdot m}{5} \leq \widehat{p} \cdot \widehat{m} \leq 5p \cdot m \tag{9}$$

Case a ($\gamma < \frac{1}{5p\cdot m}$): We condition on $F = 0$. Hence, from the argument above it follows that Equation 9 holds. Hence, $\gamma < \frac{1}{5p\cdot m} \leq \frac{1}{\widehat{p}\cdot\widehat{m}}$. This implies at step 6 in γ-`NB-ALG` , $\widehat{p}\cdot\widehat{m} < \frac{1}{\gamma}$, and γ-`NB-ALG` executes steps 11-14. That is γ-`NB-ALG` makes $\frac{B}{4\gamma}$ interventions in the remaining rounds. The algorithm constructs set $A = \{a_{i,x} \mid \widehat{p}_{i,x} \leq \frac{1}{\widehat{m}}\}$. Now for arms in A, $\widehat{\mu}_{i,x}$ is computed as in step 14 of γ-`NB-ALG`, i.e for $a_{i,x} \in A$

$$\widehat{\mu}_{i,x} = \frac{2\gamma|A|}{B} \sum_{t=B/2+1}^{B/2\gamma} Y_t \cdot \mathbb{1}\{a_t = a_{i,x}\} \ .$$

Notice that $|A| \leq \widehat{m}$ (from the definition of $m(\widehat{\mathbf{p}})$). Hence

$$\frac{B}{2\gamma \cdot |A|} \geq \frac{B}{2\gamma \cdot \widehat{m}} \geq \frac{B}{4\gamma \cdot m} \qquad \text{(from Lemma B.3)}.$$

Thus from Lemma A.1 for each arm $a_{i,x} \in A$ and any $\varepsilon > 0$

$$\mathbb{P}\Big\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon \Big| F = 0\Big\} \leq 2e^{-\varepsilon^2 \frac{B}{2\gamma m}} \tag{10}$$

Also for arms not in A, $\widehat{\mu}_{i,x}$ is computed as in step 3 of γ-`NB-ALG`, i.e. for $a_{i,x} \notin A$

$$\widehat{\mu}_{i,x} = \frac{\sum_{t=1}^{B/2} Y_t \cdot \mathbb{1}\{X_i = x\}}{\sum_{t=1}^{B/2} \mathbb{1}\{X_i = x\}} \ .$$

Moreover, if $a_{i,x} \notin A$ then $\widehat{p}_{i,x} \geq \frac{1}{\widehat{m}} \geq \frac{1}{2m}$. Since $\widehat{p}_{i,x} = \frac{2}{B}\sum_{t=1}^{B/2} \mathbb{1}\{X_i = x\}$, this implies if $a_{i,x} \notin A$ then $\sum_{t=1}^{B/2} \mathbb{1}\{X_i = x\} \geq \frac{B}{4m}$. Hence from Lemma A.1, for each arm $a_{i,x} \notin A$ and any $\varepsilon > 0$,

$$\mathbb{P}\Big\{|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \varepsilon \Big| F = 0\Big\} \leq 2e^{-\varepsilon^2 \frac{B}{2m}} \leq 2e^{-\varepsilon^2 \frac{B}{2\gamma m}} \tag{11}$$

The last inequality holds since $\gamma \geq 1$. Using Equations 10 and 11 we have for any arm $a \in \mathcal{A}$,

$$\mathbb{P}\Big\{|\widehat{\mu}_a - \mu_a| \geq \varepsilon \Big| F = 0\Big\} \leq 2e^{-\varepsilon^2 \frac{B}{2\gamma m}} \ .$$

Hence, applying union bound we have

$$\mathbb{P}\Big\{\text{there is an } a \in \mathcal{A} \text{ such that } |\widehat{\mu}_a - \mu_a| \geq \varepsilon \Big| F = 0\Big\} \leq (4M+2)e^{-\varepsilon^2 \frac{B}{2\gamma m}} \leq 8Me^{-\varepsilon^2 \frac{B}{2\gamma m}} \ .$$

Substituting $\varepsilon = \sqrt{\frac{8\gamma m}{B} \log \frac{MB}{\gamma m}}$ we have

$$E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)|F=0] \leq \sqrt{\frac{8\gamma m}{B} \log \frac{MB}{\gamma m}} + \frac{8}{M^3}\left(\frac{\gamma m}{B}\right)^4 \leq \sqrt{\frac{32\gamma m}{B} \log \frac{MB}{\gamma m}} \ . \tag{12}$$

To get the last inequality, we use that $\frac{8}{M^3}\left(\frac{\gamma m}{B}\right)^4 \leq \sqrt{\frac{8\gamma m}{B} \log \frac{MB}{\gamma m}}$, as $M \geq 1$ and $B \geq \gamma m$. Finally, we use Equation 12 and Lemma B.2 to bound the expected simple regret of γ-`NB-ALG` in this case as follows:

$$\begin{aligned}
E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)] &= E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)|Y=0]Pr\{Y=0\} + E[r(B)|Y=1]Pr\{Y=1\} \\
&\leq E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)|Y=0] + Pr\{Y=1\} \\
&\leq \sqrt{\frac{32\gamma m}{B} \log \frac{MB}{\gamma m}} + 2Me^{-\frac{p^2}{16}B} \\
&= O\left(\sqrt{\frac{\gamma m}{B} \log \frac{MB}{\gamma m}}\right) \ .
\end{aligned}$$

In last but one line of the above equation, we use that B satisfies $B \geq \frac{4}{p^2} \log \frac{2MB}{\gamma m}$ and $B \geq \gamma m$ implying $2Me^{-\frac{p^2}{16}B}$ is at most $\sqrt{\frac{32\gamma m}{B} \log \frac{MB}{\gamma m}}$.

Case b ($\gamma \geq \frac{5}{p \cdot m(\mathbf{p})}$): Again we condition on $F = 0$, and hence Equation 9 holds. Hence, $\gamma \geq \frac{5}{p \cdot m(\mathbf{p})} \geq \frac{1}{\widehat{p} \cdot m(\widehat{\mathbf{p}})}$. This implies at step 6 in γ-`NB-ALG` , $\widehat{p} \cdot m(\widehat{\mathbf{p}}) \geq \frac{1}{\gamma}$, and γ-`NB-ALG` executes steps 7-9. That is it plays the arm a_0 for B rounds. Thus, from the analysis of Theorem 1 we have that (see Equation 8)

$$E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)|Y=0] \leq \sqrt{\frac{1}{pB}} + \sqrt{\frac{8}{pB} \log(16pMB)} \ . \tag{13}$$

We use Equation 13 and Lemma B.2 to bound the expected simple regret of γ-`NB-ALG` in this case as follows:

$$\begin{aligned}
E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)] &= E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)|Y=0]Pr\{Y=0\} + E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)|Y=1]Pr\{Y=1\} \\
&\leq E[r_{\gamma\text{-}\mathtt{NB\text{-}ALG}}(B)|Y=0] + Pr\{Y=1\} \\
&\leq \sqrt{\frac{1}{pB}} + \sqrt{\frac{8}{pB} \log(16pMB)} + 2Me^{-\frac{p^2}{16}B} \\
&= O\left(\sqrt{\frac{1}{pB} \log(16pMB)}\right)
\end{aligned}$$

Again in the last but one line of the above equation, we use that $\frac{4 \log MB}{p^2 B} \leq 1$ and hence $2Me^{-\frac{p^2}{16}B}$ is at most $\sqrt{\frac{8}{pB} \log(16pMB)}$.

B.3 Proof of Theorem 3

The proof of Theorem 3 requires the the following lemmas.

Lemma B.4. *For any $T \in \mathbb{N}$, at the end of T rounds the following hold:*

1. $\mathbb{P}\Big\{|\widehat{\mu}_0(T) - \mu_0| \geq \frac{d_0}{4}\Big\} \leq \frac{2}{T^{\frac{d_0^2}{8}}}$,
2. *Let* $\widehat{p}_{i,x} = \frac{\sum_{t=1}^{T} \mathbb{1}\{a_t = a_0 \quad and \quad X_i = x\}}{N_T^0}$. *Then* $\mathbb{P}\{\widehat{p}_{i,x} \geq \frac{p}{2}\} \geq 1 - \frac{1}{T^{\frac{p^2}{2}}}$,

3. $\mathbb{P}\Big\{\Big|\frac{\widehat{\mu}_{i,x}(T)}{\gamma} - \frac{\mu_{i,x}}{\gamma}\Big| \geq \frac{d_0}{4}\Big\} \leq \frac{2}{T^{\frac{d_0^2 p\gamma^2}{16}}} + \frac{1}{T^{\frac{p^2}{2}}}$.

Proof. 1. Since $\beta \geq 1$, at the end of T rounds arm a_0 is pulled by CRM-NB-ALG at least $(\ln T)^2$ times. Hence, $N_T^0 \geq (\ln T)$, and from Lemma A.1 we have

$$\mathbb{P}\Big\{|\widehat{\mu}_0(T) - \mu_0| \geq \frac{d_0}{4}\Big\} \leq 2e^{-\frac{d_0^2}{8}\ln T} = \frac{2}{T^{\frac{d_0^2}{8}}} \tag{14}$$

2. Observe that for any (i,x), $p_{i,x} \geq p$, and $\mathbb{E}[\widehat{p}_{i,x}] = p_{i,x}$. Using Lemma A.1 and that $N_T^0 \geq \ln T$, for a fixed (i,x) we have

$$\mathbb{P}\{\widehat{p}_{i,x} \geq p_{i,x} - \frac{p}{2} \geq \frac{p}{2}\} \geq 1 - e^{-\frac{p^2}{2}\ln T} = 1 - \frac{1}{T^{\frac{p^2}{2}}} \quad .$$

3. Recall that the effective number of arm pulls of arm $a_{i,x}$ at the end of T rounds is

$$E_T^{i,x} = N_T^{i,x} + \sum_{t=1}^{T} \mathbb{1}\{a_t = a_0 \quad \text{and} \quad X_i = x\} \quad .$$

Hence, $E_T^{i,x} = N_T^{i,x} + \widehat{p}_{i,x}N_T^0$, where $\widehat{p}_{i,x}$ is as defined in part two of this lemma. Hence for any i,x at the end of T rounds if $\widehat{p}_{i,x} \geq \frac{p}{2}$ then $E_T^{i,x} \geq \frac{pN_T^0}{2}$. Further, as $N_T^0 \geq \ln T$, it follows that at the end of T rounds if $\widehat{p}_{i,x} \geq \frac{p}{2}$ then $E_T^{i,x} \geq \frac{p\ln T}{2}$. Hence, from the definition of $\widehat{\mu}_{i,x}(T)$ and Lemma A.1, at the end of T rounds we have for any fixed i,x:

$$\mathbb{P}\Big\{\Big|\frac{\widehat{\mu}_{i,x}(T)}{\gamma} - \frac{\mu_{i,x}}{\gamma}\Big| \geq \frac{d_0}{4}\Big|\widehat{p}_{i,x} \geq \frac{p}{2}\Big\} \leq 2e^{-\frac{\gamma^2 d_0^2}{16}p\ln T} = \frac{2}{T^{\frac{p\gamma^2 d_0^2}{16}}} \ . \tag{15}$$

Finally by law of total probability,

$$\begin{aligned}
\mathbb{P}\Big\{\Big|\frac{\widehat{\mu}_{i,x}(T)}{\gamma} - \frac{\mu_{i,x}}{\gamma}\Big| \geq \frac{d_0}{4}\Big\} &= \mathbb{P}\Big\{\Big|\frac{\widehat{\mu}_{i,x}(T)}{\gamma} - \frac{\mu_{i,x}}{\gamma}\Big| \geq \frac{d_0}{4}\Big|\widehat{p}_{i,x} \geq \frac{p}{2}\Big\}\mathbb{P}\{\widehat{p}_{i,x} \geq \frac{p}{2}\} \\
&\quad + \mathbb{P}\Big\{\Big|\frac{\widehat{\mu}_{i,x}(T)}{\gamma} - \frac{\mu_{i,x}}{\gamma}\Big| \geq \frac{d_0}{4}\Big|\widehat{p}_{i,x} \leq \frac{p}{2}\Big\}\mathbb{P}\{\widehat{p}_{i,x} \leq \frac{p}{2}\} \\
&\leq \mathbb{P}\Big\{\Big|\frac{\widehat{\mu}_{i,x}(T)}{\gamma} - \frac{\mu_{i,x}}{\gamma}\Big| \geq \frac{d_0}{4}\Big|\widehat{p}_{i,x} \geq \frac{p}{2}\Big\} + \mathbb{P}\{\widehat{p}_{i,x} \leq \frac{p}{2}\} \\
&\leq \frac{2}{T^{\frac{p\gamma^2 d_0^2}{16}}} + \frac{1}{T^{\frac{p^2}{2}}} \ .
\end{aligned}$$

The last line in the above inequality follows from Equation 15 and part 2 of this lemma. □

Lemma B.5. *Let* $L = \arg\min_{t\in\mathbb{N}}\{\frac{t^{\frac{p^2 d_0^2}{16}}}{\ln t} \geq 15M\}$*, and suppose* CRM-NB-ALG *pulls arms for* T *rounds, where* $T \geq \max(L, e^{\frac{50}{d_0^2}})$*, and let* $a^* \neq a_0$*. Then at the end of* T *rounds* $\frac{8}{9d_0^2} \leq E[\beta^2] \leq \frac{50}{d_0^2}$*. (Note that* $\max(L, e^{\frac{50}{d_0^2}})$ *is a finite constant dependent on instance constants* p, d_0*, and* M*.)*

Proof. Recall that β is set as in steps 11-14 in CRM-NB-ALG . We begin by making the following easy to see observations.

Observation B.1. *1. If* $a^* \neq a_0$ *then* $d_0 = \frac{\mu_{a^*}}{\gamma} - \mu_0$*.*

2. Let $\widehat{\mu}^* = \max_{i,x}(\widehat{\mu}_{i,x}(T))$ *(as computed in step 11 of* CRM-NB-ALG *). If* $|\widehat{\mu}_0(T) - \mu_0| \leq \frac{d_0}{4}$ *and* $|\frac{\widehat{\mu}_{i,x}(T)}{\gamma} - \frac{\mu_{i,x}}{\gamma}| \leq \frac{d_0}{4}$ *for all* (i,x) *then* $\frac{d_0}{2} \leq \frac{\widehat{\mu}^*}{\gamma} - \widehat{\mu}_0(T) \leq \frac{3d_0}{2}$*, and* $\frac{32}{9d_0^2} \leq \beta^2 \leq \frac{32}{d_0^2}$*. Notice that since* $T \geq e^{\frac{50}{d_0^2}}$*,* $\frac{32}{d_0^2} \leq \ln T$*.*

Let U_0 be the event that $|\widehat{\mu}_0 - \mu_0| \leq \frac{d_0}{4}$, and for any i,x let $U_{i,x}$ be the event $|\frac{\widehat{\mu}_{i,x}}{\gamma} - \frac{\mu_{i,x}}{\gamma}| \leq \frac{d_0}{4}$. Also let $U = (\cap_{i,x}U_{i,x}) \cap U_0$, and let $\overline{U}_0$, $\overline{U}_{i,x}$, and $\overline{U}$ denote the compliment of the events $U_0, U_{i,x}$, and $\overline{U}$ respectively. From parts 1 and 3 of Lemma B.4, we have

$$\mathbb{P}\Big\{\overline{U}_0\Big\} \leq \frac{2}{T^{\frac{d_0^2\ln T}{8}}} \ , \text{ and}$$

$$\text{for a fixed } (i,x) \quad \mathbb{P}\Big\{\overline{U}_{i,x}\Big\} \le \frac{2}{T^{\frac{p\gamma^2 d_0^2}{16}}} + \frac{1}{T^{\frac{p^2}{2}}} \ .$$

Hence applying union bound,

$$\begin{aligned}\mathbb{P}\{\overline{U}\} &\le 4M\left(\frac{1}{T^{\frac{p\gamma^2 d_0^2}{16}}} + \frac{1}{T^{\frac{p^2}{2}}}\right) + \frac{2}{T^{\frac{d_0^2}{8}}} \\ &\le 4M\left(\frac{1}{T^{\frac{p^2 d_0^2}{16}}} + \frac{1}{T^{\frac{p^2 d_0^2}{16}}}\right) + \frac{2M}{T^{\frac{p^2 d_0^2}{16}}} \qquad \text{as } \ \gamma \ge 1, p \le 1, d_0 \le 1 \\ &\le \frac{10M}{T^{\frac{p^2 d_0^2}{16}}} = \delta \ .\end{aligned}$$

We will use the above arguments to first show that $E[\beta^2] \ge \frac{8}{d_0^2}$. From part 2 of Observation B.1 we have that the event U implies $\beta^2 \ge \frac{32}{9d_0^2}$. Since $\mathbb{P}\{U\} \ge 1-\delta$,

$$E[\beta^2] \ge \frac{32}{9d_0^2}(1-\delta) = \frac{32}{9d_0^2} - \frac{32\delta}{9d_0^2}$$

Since T satisfies $\frac{T^{\frac{p^2 d_0^2}{16}}}{\ln T} \ge 15M$, this implies $\frac{32\delta}{9d_0^2} \le \frac{24}{9d_0^2}$, and hence $E[\beta^2] \ge \frac{8}{9d_0^2}$. Similarly, from part 2 of Observation B.1 we have that the event U implies $\beta^2 \le \frac{32}{d_0^2}$. Here, we use that if U does not hold then $\beta^2 \le \ln T$. Hence

$$E[\beta^2] \le \frac{32}{d_0^2}(1-\delta) + \delta \ln T \le \frac{32}{d_0^2} + \delta \ln T \ .$$

Since T satisfies $\frac{T^{\frac{p^2 d_0^2}{16}}}{\ln T} \ge 15M$, we have $\delta \ln T \le \frac{18}{d_0^2}$, and hence $E[\beta^2] \le \frac{50}{d_0^2}$. □

Lemma B.6. *Suppose the algorithm pulls the arms for T rounds and if $a^* \ne a_{i,x}$. Then*

$$E[N_T^{i,x}|T] \le \max\left(0, \frac{8\ln T}{d_{i,x}^2} + 1 - p_{i,x}E[N_t^0]\right) + \frac{\pi^2}{3} \ .$$

Further if $a^* \ne a_0$ then

$$E[N_{0,t}|T] \le \max\Big(E[\beta^2]\ln T, \ \frac{8\ln T}{d_0^2} + 1\Big) + \frac{\pi^2}{3}, \ .$$

Proof. For ease of notation we denote $E[N_T^{i,x}|T]$ as $E[N_T^{i,x}]$. Observe that

$$N_T^{i,x} = \sum_{t\in T} \mathbb{1}\{a(t) = a_{i,x}\} \ . \tag{16}$$

Since $E_T^{i,x} = N_T^{i,x} + \sum_{t\in[T]} \mathbb{1}\{a(t) = a_0 \ \text{ and } \ X_i = x\}$, if $E_T^{i,x} = \ell$ then $N_T^{i,x} = \max(0, \ell - \sum_{t\in[T]} \mathbb{1}\{a(t) = a_0 \ \text{ and } \ X_i = x\})$. We use this to rewrite Equation 16 as follows

$$N_T^{i,x} \le \max(0, \ell - \sum_{t\in[T]} \mathbb{1}\{a(t) = a_0 \ \text{ and } \ X_i = x\}) + \sum_{t\in T} \mathbb{1}\{a(t) = a_{i,x}, E_t^{i,x} \ge \ell\} \ . \tag{17}$$

We require the following observation which is easy to prove.

Observation B.2. $\sum_{t\in[T]} E[\mathbb{1}\{a(t) = a_0 \ \ \textit{and} \ \ X_i = x\}] = p_{i,x}E[N_T^0]$.

Proof. Observe that

$$E[\sum_{t\in[T]} \mathbb{1}\{a(t) = a_0 \ \text{ and } \ X_i = x\}] = \sum_{t\in[T]} E[\mathbb{1}\{a(t) = a_0 \ \text{ and } \ X_i = x\}] = \sum_{t\in[T]} \mathbb{P}\{\mathbb{1}\{a(t) = a_0 \ \text{ and } \ X_i = x\}\}$$

Also observe that

$$\mathbb{P}\{\mathbb{1}\{a(t) = a_0 \ \text{ and } \ X_i = x\} = \mathbb{P}\{\mathbb{1}\{X_i = x\} \mid a(t) = a_0\}\} \cdot \mathbb{P}\{a(t) = a_0\} = p_{i,x}\mathbb{P}\{a(t) = a_0\} \ .$$

□

We continue by taking expectation on both sides of Equation 17 and use Observation B.2,

$$E[N_T^{i,x}] \leq \max\left(0, \ell - p_{i,x}E[N_t^0]\right) + \sum_{t\in[\ell+1,T]} \mathbb{P}\{a(t) = (i,x), E_t^{i,x} \geq \ell\} \ . \tag{18}$$

Now we bound $\sum_{t\in T} \mathbb{P}\{a(t) = a_{i,x}, E_t^{i,x} \geq \ell\}$, and assuming $a^* \neq a_0$. The proof for $a^* = a_0$ is similar. Before proceeding we make a note of few notations. We use $E_T^{a^*}$ to denote the effective number of pulls of a^* at the end of T rounds. Also, for better clarity in the arguments below, we use $\widehat{\mu}_{i,x}(E_T^{i,x}, T)$ (instead of $\widehat{\mu}_{i,x}(T)$) and $\widehat{\mu}_0(N_T^0, T)$ (instead of $\widehat{\mu}_0(T)$) to denote the empirical estimates of $\mu_{i,x}$ and μ_0 computed by `CRM-NB-ALG` at the end of T rounds using $E_T^{i,x}$ and N_T^0 samples respectively.

$$\begin{aligned}
\sum_{t\in[\ell+1,T]} \mathbb{P}\left\{a(t) = (i,x), E_t^{i,x} \geq \ell\right\} &= \sum_{t\in[\ell,T-1]} \mathbb{P}\left\{\frac{\widehat{\mu}_{a^*}(E_t^{a^*},t)}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 E_t^{a^*}}} \leq \frac{\widehat{\mu}_{i,x}(E_t^{i,x},t)}{\gamma} + \sqrt{\frac{2\ln(t)}{\gamma^2 E_t^{i,x}}}, \quad E_t^{i,x} \geq \ell\right\} \\
&\leq \sum_{t\in[0,T-1]} \mathbb{P}\left\{\min_{s\in[0,t]} \frac{\widehat{\mu}_{a^*}(s,t)}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 s}} \leq \max_{s_j\in[\ell-1,t]} \frac{\widehat{\mu}_{i,x}(s_j,t)}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 s_j}}\right\} \\
&\leq \sum_{t\in T}\sum_{s\in[0,t-1]}\sum_{s_j\in[\ell-1,t]} \mathbb{P}\left\{\frac{\widehat{\mu}_{a^*}(s,t)}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 s}} \leq \frac{\widehat{\mu}_{i,x}(s_j,t)}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 s_j}}\right\}
\end{aligned}$$

If $\frac{\widehat{\mu}_{a^*}(s,t)}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 s}} \leq \frac{\widehat{\mu}_{i,x}(s_j,t)}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 s_j}}$ is true then at least one of the following events is true

$$\frac{\widehat{\mu}_{a^*}(s,t)}{\gamma} \leq \frac{\mu_{a^*}}{\gamma} - \sqrt{\frac{2\ln t}{\gamma^2 s}} \ , \tag{19a}$$

$$\frac{\widehat{\mu}_{i,x}(s_j,t)}{\gamma} \geq \frac{\mu_{i,x}}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 s_j}} \ , \tag{19b}$$

$$\frac{\mu_{a^*}}{\gamma} \leq \frac{\mu_{i,x}}{\gamma} + 2\sqrt{\frac{2\ln t}{\gamma^2 s_j}} \ . \tag{19c}$$

The probability of the events in Equations 19a and 19b can be bounded using Lemma A.1,

$$\mathbb{P}\left\{\frac{\widehat{\mu}_{a^*}(s,t)}{\gamma} \leq \frac{\mu_{a^*}}{\gamma} - \sqrt{\frac{2\ln t}{\gamma^2 s}}\right\} \leq t^{-4} \ ,$$

$$\mathbb{P}\left\{\frac{\widehat{\mu}_{i,x}(s_j,t)}{\gamma} \geq \frac{\mu_{i,x}}{\gamma} + \sqrt{\frac{2\ln t}{\gamma^2 s_j}}\right\} \leq t^{-4} \ .$$

Also if $\ell \geq \lceil \frac{8\ln T}{d_{i,x}^2} \rceil$ then the event in Equation 19c is false, i.e. $\frac{\mu_{a^*}}{\gamma} > \frac{\mu_{i,x}}{\gamma} + 2\sqrt{\frac{2\ln t}{\gamma^2 s_j}}$ (as $\gamma \geq 1$). Thus we set $\ell = \frac{8\ln T}{d_{i,x}^2} + 1 \geq \lceil \frac{8\ln T}{d_{i,x}^2} \rceil$, which implies

$$\sum_{t\in T} \mathbb{P}\{a(t) = a_{i,x}, E_t^{i,x} \geq \ell\} \leq \sum_{t\in[T]}\sum_{s\in[T]}\sum_{s_j\in[\ell,T]} 2t^{-4} \leq \frac{\pi^2}{3} \tag{20}$$

If $a^* = a_0$ then using the exact arguments as above we can show that Equation 20 still holds. Hence, using Equations 18 and 20 we have if $a^* \neq a_{i,x}$ then

$$E[N_T^{i,x}] \leq \max\left(0, \frac{8\ln T}{d_{i,x}^2} + 1 - p_{i,x}E[N_t^0]\right) + \frac{\pi^2}{3} \ .$$

The arguments used to bound $E[N_T^0|T]$ (denoted $E[N_T^0]$ for convenience), when $a^* \neq a_0$ is similar. In this case the equation corresponding to Equation 18 is

$$E[N_T^0] \leq \max\Big(E[\beta^2]\ln T, \ \ell\Big) + \sum_{t\in[\ell+1,T]} \mathbb{P}\{a(t) = a_0, N_t^0 \geq \ell\} \ . \tag{21}$$

Also the same arguments as above can be used to show that for $\ell = \frac{8 \ln T}{d_0^2} + 1$,

$$\sum_{t \in T} \mathbb{P}\{a(t) = a_0, N_t^0 \geq \ell\} \leq \frac{\pi^2}{3} \ . \tag{22}$$

Finally using Equations 21 and 22, we have

$$E[N_T^0] \leq \max\left(E[\beta^2] \ln T, \ \frac{8 \ln T}{d_0^2} + 1\right) + \frac{\pi^2}{3} \ .$$

□

Lemma B.7. *If $a^* = a_0$ and suppose the algorithm pulls the arms for T rounds then*

$$E[N_T^0 | T] \geq T - \left(2M(1 + \frac{\pi^2}{3}) \sum_{i,x} \frac{8 \ln T}{d_{i,x}^2}\right) \ .$$

Proof. For convenience, we denote $E[N_T^{i,x}|T]$ and $E[N_T^0|T]$ as $E[N_T^{i,x}]$ and $E[N_T^0]$ respectively. At the end of T rounds we have

$$N_T^0 + \sum_{i,x} N_T^{i,x} = T \ .$$

Taking expectation on both sides of the above equation and rearranging the terms we have,

$$E[N_T^0] = T - \sum_{i,x} E[N_T^{i,x}] \ .$$

Now we use Lemma B.6 to conclude that

$$E[N_T^0] \geq T - \left(2M(1 + \frac{\pi^2}{3}) \sum_{i,x} \frac{8 \log T}{d_{i,x}^2}\right) \ .$$

□

Before we bound the regret of the algorithm we make the following observation regarding T, which is the number of rounds CRM-NB-ALG pulls the arms before exhausting the budget B:

$$\frac{B}{\gamma} \leq T \leq B \quad \Rightarrow \quad \frac{B}{\gamma} \leq E_T[T] \leq B \ . \tag{23}$$

Now are ready to bound the expected cumulative regret of CRM-NB-ALG for the two cases:

Case a ($a^* = a_0$): In this case we bound the expected cumulative regret of CRM-NB-ALG for B satisfying

$$\frac{B}{\gamma} \geq \frac{1}{p_{i,x}}(1 + \frac{8 \ln B}{d_{i,x}^2}) + \left(2M(1 + \frac{\pi^2}{3}) \sum_{i,x} \frac{8 \ln B}{d_{i,x}^2}\right) \ . \tag{24}$$

Observe that the constraint on B in Equation 24 is satisfied for any large B. We begin by making the following observation which shows that in this case the expected number of pulls of a sub-optimal arm is bounded by a constant for any large B. Observe that the constraint on B in Observation B.3 is satisfied for any large B.

Observation B.3. *Let $a^* = a_0$, and T be the number of rounds CRM-NB-ALG pulls the arms before the budget B is exhausted, where B satisfies the constraint in Equation 24. Then $E_T[N_T^{i,x}] \leq \frac{\pi^2}{3}$.*

Proof. From Lemmas B.6 and B.7 for any T satisfying

$$T \geq \frac{1}{p_{i,x}}(1 + \frac{8 \ln T}{d_{i,x}^2}) + \left(2M(1 + \frac{\pi^2}{3}) \sum_{i,x} \frac{8 \ln T}{d_{i,x}^2}\right) \tag{25}$$

we have $E[N_T^{i,x}|T] \leq \frac{\pi^2}{3}$. Notice that the constraint on T in Equation 25 is the same as the constraint on $\frac{B}{\gamma}$ in Equation 24. Moreover, observe that if $\frac{B}{\gamma}$ satisfies the constraint in Equation 24 then $T \geq \frac{B}{\gamma}$ satisfies Equation 25 with probability 1. Hence, $E_T[N_T^{i,x}|T] \leq \frac{\pi^2}{3}$. □

Next observe that in this case G_B (see Equation 2) is $B\mu_0$, i.e the optimal solution is to play arm a_0 in all the rounds. We require the following observation which lower bounds $E_T[T]$ in terms of B, which is the total number of rounds played by the optimal solution.

Observation B.4. *Let $a^* = a_0$, and T be the number of rounds `CRM-NB-ALG` pulls the arms before the budget B is exhausted, where B satisfies the constraint in Equation 24. Then $E_T[T] \geq B - 1 - \frac{2M\pi^2(\gamma-1)}{3}$.*

Proof. Let c_{a_t} denote the cost of arm a_t pulled at time $t \leq T$. That is $c_{a_t} = \gamma$ if $a_t = a_{i,x}$ and $c_{a_t} = 1$ if $a_t = a_0$. Then the following is always true, as `CRM-NB-ALG` pulls arms till the budget is the exhausted:

$$B - 1 \leq \sum_{t\in[T]} c_{a_t} \ . \tag{26}$$

Taking expectation over T and the sequence of arm pulls $\{a_t\}$ made by `CRM-NB-ALG` , on both sides of the above equation, we have

$$\begin{aligned}
B - 1 &\leq E_{T,\{a_t\}}\Big[\sum_{t\in[T]} c_{a_t}\Big] \\
&\leq E_T\Big[E_{\{a_t\}}[\sum_{t\in[T]} c_{a_t}]\Big] \\
&\leq E_T\Big[\sum_{t\in[T]}\Big(\mathbb{P}\{a_t = a_0\} + \gamma(\sum_{i,x}\mathbb{P}\{a_t = a_{i,x}\}\Big)\Big] \\
&\leq E_T\Big[T + \sum_{t\in[T]}(\gamma-1)(\sum_{i,x}\mathbb{P}\{a_t = a_{i,x}\})\Big] \\
&\leq E_T[T] + E_T\Big[\sum_{i,x}(\gamma-1)(\sum_{t\in[T]}\mathbb{P}\{a_t = a_{i,x}\})\Big] \\
&\leq E_T[T] + E_T\Big[\sum_{i,x}(\gamma-1)E[N_T^{i,x}|T]\Big] \ .
\end{aligned}$$

The third line in the above set of equations follows by using $\mathbb{P}\{a_t = a_0\} = 1 - \sum_{i,x}\mathbb{P}\{a_t = a_{i,x}\}$. Finally from Observation B.3, we have $E[N_T^{i,x}] \leq \frac{\pi^2}{3}$. Substituting this in the last line of the above equation, we have $E_T[T] \geq B - 1 - \frac{2M\pi^2(\gamma-1)}{3}$. □

Finally we bound the expected cumulative regret of `CRM-NB-ALG` when $a^* = a_0$ as follows:

$$\begin{aligned}
E[R_{\texttt{CRM-NB-ALG}}(B)] \ &\leq G_B - E_{T,\{a_t\}}\left[\sum_{t\in[T]}\mu_{a_t}\right] \\
&\leq B\mu_0 - E_T\left[\sum_{t=1}^{T} E_{\{a_t\}}[\mu_{a_t}]\right] \\
&\leq E_T\left[B\mu_0 - \sum_{t=1}^{T} E_{\{a_t\}}[\mu_{a_t}]\right] \\
&\leq E_T\left[B\mu_0 - \sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\mu_a\mathbb{P}\{a_t = a\}\right]
\end{aligned}$$

$$\begin{aligned}
&\leq E_T\left[(B-T)\mu_0 + \sum_{t=1}^{T}(\mu_0 - \sum_{a\in\mathcal{A}}\mu_a\mathbb{P}\{a_t = a\})\right] \\
&\leq E_T\left[(B-T)\mu_0\right] + E_T\left[\sum_{t=1}^{T}\sum_{\Delta_a>0}\Delta_a\mathbb{P}\{a_t = a\})\right] \ .
\end{aligned}$$

Thus, from Observations B.3 and B.4, we have

$$E[R_{\texttt{CRM-NB-ALG}}(B)] \leq 1 + \frac{2M\pi^2(\gamma-1)}{3} + \sum_{\Delta_a>0}\Delta_a\frac{\pi^2}{3} \ .$$

Observe that the expected regret of `CRM-NB-ALG` is bounded by a constant for large B and hence $O(1)$.

Case b ($a^* \neq a_0$): In this case we bound the expected cumulative regret of `CRM-NB-ALG` for B satisfying $B \geq \max(L, e^{\frac{50}{d_0^2}})$, where L is as in Lemma B.5. Observe that the constraint is satisfied for any large B. Let T be the number of rounds `CRM-NB-ALG` pulls the arms before exhausting the budget B. Then from Equation 25, we have $T \geq \max(L, e^{\frac{50}{d_0^2}})$. Hence, from Lemmas B.5 and B.6, and as $T \leq B$ (from Equation 23), we have for $a^* \neq a_{i,x}$

$$E_T\left[E[N_T^{i,x}|T]\right] \leq \max\left(0, 1 + 8\ln B\left(\frac{1}{d_{i,x}^2} - \frac{p_{i,x}}{9d_0^2}\right)\right) + \frac{\pi^2}{3} \ , \tag{27}$$

$$\text{and} \quad E_T\left[E[N_T^0|T]\right] \leq \frac{50\ln B}{d_0^2} + \frac{\pi^2}{3} \quad . \tag{28}$$

Also observe that in this case G_B is at most $\frac{B\mu_{a^*}}{\gamma}$. Below we bound the expected cumulative regret of `CRM-NB-ALG` when $a^* \neq a_0$

$$\begin{aligned}
E[R_{\texttt{CRM-NB-ALG}}(B)] \ &\leq \frac{B\mu_{a^*}}{\gamma} - E_{T,\{a_t\}}\left[\sum_{t\in[T]}\mu_{a_t}\right] \\
&\leq \frac{B\mu_{a^*}}{\gamma} - E_T\left[\sum_{t=1}^{T}E_{\{a_t\}}[\mu_{a_t}]\right] \\
&\leq E_T\left[\frac{B\mu_{a^*}}{\gamma} - \sum_{t=1}^{T}E_{\{a_t\}}[\mu_{a_t}]\right] \\
&\leq E_T\left[\frac{B\mu_{a^*}}{\gamma} - \sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\mu_a\mathbb{P}\{a_t = a|T\}\right] \\
&\leq E_T\left[\Big(\frac{B}{\gamma} - T\Big)\mu_{a^*} + \sum_{t=1}^{T}(\mu_{a^*} - \sum_{a\in\mathcal{A}}\mu_a\mathbb{P}\{a_t = a|T\})\right] \\
&\leq E_T\left[\Big(\frac{B}{\gamma} - T\Big)\mu_{a^*}\right] + E_T\left[\sum_{t=1}^{T}\sum_{\Delta_a>0}\Delta_a\mathbb{P}\{a_t = a|T\})\right] \ .
\end{aligned}$$

Now observe that as $T \geq \frac{B}{\gamma}$, $E_T[(\frac{B}{\gamma} - T)\mu_{a^*}] \leq 0$. Also note that $E_T[\sum_{t=1}^{T}\mathbb{P}\{a_t = a|T\}] = E_T[N_T^a|T]$. Using this and Equations 27 and 28, we have our result as follows:

$$\begin{aligned}
E[R_{\texttt{CRM-NB-ALG}}(B)] \ &\leq \Delta_0 E_T\left[E[N_T^0|T]\right] + \sum_{\Delta_{i,x}>0}\Delta_{i,x}E_T\left[E[N_T^{i,x}|T]\right] \\
&\leq \Delta_0\Big(\frac{50\ln B}{d_0^2} + \frac{\pi^2}{3}\Big) + \sum_{\Delta_{i,x}>0}\Delta_{i,x}\max\left(0, 1 + 8\ln B\left(\frac{1}{d_{i,x}^2} - \frac{p_{i,x}}{9d_0^2}\right)\right) + \frac{\pi^2}{3} \ .
\end{aligned}$$

Hence, we have that the expected cumulative regret of CRM-NB-ALG is:

$$E[R_{\texttt{CRM-NB-ALG}}(B)] \leq \begin{cases} 1 + \frac{2M\pi^2(\gamma-1)}{3} + \sum_{\Delta_a>0} \Delta_a \frac{\pi^2}{3} & \text{when } a^* = a_0 \\ \Delta_0\left(\frac{50 \ln B}{d_0^2} + \frac{\pi^2}{3}\right) + \sum_{\Delta_{i,x}>0} \Delta_{i,x} \max\left(0, 1 + 8 \ln B \left(\frac{1}{d_{i,x}^2} - \frac{p_{i,x}}{9d_0^2}\right)\right) + \frac{\pi^2}{3} & \text{when } a^* \neq a_0 \end{cases}$$

B.4 Proof of Theorem 4

Throughout this proof a and $\mathbf{y}$ indexes the sets $\mathcal{A}$ and S^n respectively. Let δ, L_1, $L_{2,a}$ and L_a for all a, be as in the theorem statement. Let $a^* = \arg\max_a(\mu_a)$. As is standard in MAB literature, we assume without loss of generality that a^* is unique. Further, let $\Delta_a = \mu_{a^*} - \mu_a$. The regret upper bound is proved using Lemmas B.8 and B.9.

Lemma B.8. *Let T be the number of rounds C-UCB 2 has pulled the arms. Then for $T \geq L_1$ the following holds:*

1. *For all $\mathbf{y}$ such that $c_{\mathbf{y}} > 0$, $\mathbb{P}\big\{N_{\mathbf{y},T} \leq \frac{E[N_{\mathbf{y},T}]}{2}\big\} \leq e^{-\frac{E[N_{\mathbf{y},T}]^2}{2T}}$,*
2. *For all $\mathbf{y}$ such that $c_{\mathbf{y}} > 0$ and for any $\varepsilon_{\mathbf{y}} \geq 0$, $\mathbb{P}\{|\widehat{\mu}_{\mathbf{y}}(T) - \mu_{\mathbf{y}}| \geq \varepsilon_{\mathbf{y}}\} \leq 2e^{-c_{\mathbf{y}}^2 T\varepsilon_{\mathbf{y}}^2} + e^{-\frac{c_{\mathbf{y}}^2 T}{2}}$,*
3. *For all a, $\mathbb{P}\Big(|\widehat{\mu}_a(T) - \mu_a| \geq \sqrt{\frac{\log(k^n T^2/2)}{T}}\zeta_a\Big) \leq \frac{2}{T^2}$.*

Proof. 1. Part 1 of the lemma follows from Lemma A.1.

2. Using Lemma A.1 again, it follows that for all $\mathbf{y}$ such that $c_{\mathbf{y}} > 0$, and for all $\varepsilon_{\mathbf{y}} \geq 0$,

$$\mathbb{P}\Big\{|\widehat{\mu}_{\mathbf{y}}(T) - \mu_{\mathbf{y}}\Big| \geq \varepsilon_{\mathbf{y}} \mid N_{\mathbf{y},T} > \frac{E[N_{\mathbf{y},T}]}{2}\Big\} \leq 2e^{-E[N_{\mathbf{y},T}]\varepsilon_{\mathbf{y}}^2} . \tag{29}$$

Hence, for all $\mathbf{y}$ such that $c_{\mathbf{y}} > 0$, using the law of total probability we have

$$\begin{aligned} \mathbb{P}(|\widehat{\mu}_{\mathbf{y}}(T) - \mu_{\mathbf{y}}| \geq \varepsilon_{\mathbf{y}}) = \mathbb{P}\Big\{|\widehat{\mu}_{\mathbf{y}}(T) - \mu_{\mathbf{y}}| \geq \varepsilon_{\mathbf{y}}\Big| N_{\mathbf{y},T} > \frac{E[N_{\mathbf{y},T}]}{2}\Big\}\mathbb{P}\Big\{N_{\mathbf{y},T} > \frac{E[N_{\mathbf{y},T}]}{2}\Big\} + \\ \mathbb{P}\Big\{|\widehat{\mu}_{\mathbf{y}}(T) - \mu_{\mathbf{y}}| \geq \varepsilon_{\mathbf{y}}\Big| N_{\mathbf{y},T} \leq \frac{E[N_{\mathbf{y},T}]}{2}\Big\}\mathbb{P}\Big\{N_{\mathbf{y},T} \leq \frac{E[N_{\mathbf{y},T}]}{2}\Big\} \\ \leq \mathbb{P}\Big\{|\widehat{\mu}_{\mathbf{y}}(T) - \mu_{\mathbf{y}}| \geq \varepsilon_{\mathbf{y}}\Big| N_{\mathbf{y},T} > \frac{E[N_{\mathbf{y},T}]}{2}\Big\} + \mathbb{P}\Big\{N_{\mathbf{y},T} \leq \frac{E[N_{\mathbf{y}}, T]}{2}\Big\} \\ \leq 2e^{-E[N_{\mathbf{y},T}]\varepsilon_{\mathbf{y}}^2} + e^{-\frac{E[N_{\mathbf{y},T}]^2}{2T}} \\ \leq 2e^{-c_{\mathbf{y}} T\varepsilon_{\mathbf{y}}^2} + e^{-\frac{c_{\mathbf{y}}^2 T}{2}} \\ \leq 2e^{-c_{\mathbf{y}}^2 T\varepsilon_{\mathbf{y}}^2} + e^{-\frac{c_{\mathbf{y}}^2 T}{2}} . \end{aligned}$$

The second line in the above equations follows from Equation 29 and part one of this lemma. The last two inequalities follow by observing that $E[N_{\mathbf{y},T}] \geq c_{\mathbf{y}}T \geq c_{\mathbf{y}}^2 T$. This is true as for each $\mathbf{y}$, $c_{\mathbf{y}} = \min_a \mathbb{P}\{Pa(Y) = \mathbf{y} \mid do(a)\}$, and hence $0 < c_{\mathbf{y}} \leq 1$.

3. Let $\varepsilon_{\mathbf{y}} = \sqrt{\frac{\log(k^n T^2/2)}{c_{\mathbf{y}}^2 T}}$ if $c_{\mathbf{y}} > 0$, and $\varepsilon_{\mathbf{y}} = 0$ if $c_{\mathbf{y}} = 0$. Since the parent distributions have the same non-zero support, and as $\widehat{\mu}_a(T) = \sum_{\mathbf{y}} \widehat{\mu}_{\mathbf{y}}(T)\mathbb{P}\{Pa(Y) = \mathbf{y}|do(a)\}$, the event

$$|\widehat{\mu}_a(T) - \mu_a| \geq \sum_{\mathbf{y}, c_{\mathbf{y}}>0} \varepsilon_{\mathbf{y}}\mathbb{P}\{Pa(Y) = \mathbf{y}|do(a)\}$$

implies there is a $\mathbf{y}$ such that $c_{\mathbf{y}} > 0$ and $\{|\widehat{\mu}_{\mathbf{y}}(T) - \mu_{\mathbf{y}}| \geq \varepsilon_{\mathbf{y}}\}$. Hence, using part 2 of this lemma and applying union bound over all $\mathbf{y}$ such that $c_{\mathbf{y}} > 0$, we have for every a

$$\mathbb{P}\Big\{|\widehat{\mu}_a(T) - \mu_a| \geq \sum_{\mathbf{y}, c_{\mathbf{y}}>0} \varepsilon_{\mathbf{y}}\mathbb{P}\{Pa(Y) = \mathbf{y}|do(a)\}\Big\} \leq \sum_{\mathbf{y}, c_{\mathbf{y}}>0} \big(2e^{-c_{\mathbf{y}}^2 T\varepsilon_{\mathbf{y}}^2} + e^{-\frac{c_{\mathbf{y}}^2 T}{2}}\big) .$$

Substituting the values of $\varepsilon_{\mathbf{y}}$ and using $\zeta_a = \sum_{\mathbf{y},c_{\mathbf{y}}>0} \frac{\mathbb{P}\{Pa(Y)=\mathbf{y}|do(a)\}}{c_{\mathbf{y}}}$ in the above equation, we have

$$\begin{aligned}\mathbb{P}\Big\{|\widehat{\mu}_a(T)-\mu_a| \geq \sqrt{\frac{\log(k^nT^2/2)}{T}}\zeta_a\Big\} &\leq \frac{1}{T^2} + \sum_{\mathbf{y},c_{\mathbf{y}}>0} e^{-\frac{c_{\mathbf{y}}^2 T}{2}} \\ &\leq \frac{1}{T^2} + k^n e^{-\delta^2 T/2}\end{aligned}$$

where $\delta = \min_{c_{\mathbf{y}}>0} c_{\mathbf{y}}$. Since $T \geq L_1$, $T \geq \frac{2\log(k^nT^2)}{\delta^2}$. This implies $k^n e^{-\delta^2 T/2} \leq \frac{1}{T^2}$, and

$$\mathbb{P}\Big\{|\widehat{\mu}_a-\mu_a| \geq \sqrt{\frac{\log(k^nT^2/2)}{T}}\zeta_a\Big\} \leq \frac{2}{T^2}\ .$$

□

Lemma B.9. *Let $a \in A$ be a sub-optimal intervention. Then the expected number of times intervention a is made after $L_a = \max\{L_1, L_{2,a}\}$ rounds is at most $\frac{2\pi^2}{3}$.*

Proof. For ease of notation, we denote $\sqrt{\frac{\log(k^nt^2/2)}{t}}\zeta_a$ as $c_{a,t}$. Note that $c_{a,t}$ is the confidence radius of intervention a `C-UCB-2` maintains at the end of t rounds. Further, let $N'_{a,T}$ denote the number of times the algorithm performs intervention a from time L_a+1 to time $T \geq L_a$, and also let a_t denote the intervention performed at time t. Hence,

$$N'_{a,T} = \sum_{t=L_a+1}^{T} \mathbb{1}\Big\{a_t = a\Big\}\ . \tag{30}$$

Note that $a_t = a$ implies $\bar{\mu}_{a^*}(t-1) \leq \bar{\mu}_a(t-1)$ i.e. $\widehat{\mu}_{a^*}(t-1) + c_{a^*,t-1} \leq \widehat{\mu}_a(t-1) + c_{a,t-1}$. Hence from Equation 30, we have

$$N'_{a,T} \leq \sum_{t=L_a}^{T-1} \mathbb{1}\Big\{\widehat{\mu}_{a^*}(t) + c_{a^*,t} \leq \widehat{\mu}_a(t) + c_{a,t}\Big\}\ .$$

The event $\widehat{\mu}_{a^*,t} + c_{a^*,t} \leq \widehat{\mu}_{a,t} + c_{a,t}$ implies that at least one of the following events is true

$$\big\{\widehat{\mu}_{a^*}(t) \leq \mu_{a^*} - c_{a^*,t}\big\} \tag{31}$$
$$\big\{\widehat{\mu}_a(t) \geq \mu_a + c_{a,t}\big\} \tag{32}$$
$$\big\{\mu_{a^*} < \mu_a + 2c_{a,t}\big\} \tag{33}$$

Since $t \geq L_a \geq L_1$, using Lemma B.8 the probability of the events in Equations 31 and 32 can be bounded as:

$$\begin{aligned}\mathbb{P}\big\{\widehat{\mu}_{a^*}(t) \leq \mu_{a^*} - c_{1,t}\big\} &\leq 2t^{-2}\ , \\ \mathbb{P}\big\{\widehat{\mu}_a(t) \geq \mu_a + c_{a,t}\big\} &\leq 2t^{-2}\ .\end{aligned}$$

The event in equation 33 $\big\{\mu_{a^*} < \mu_a + 2c_{a,t}\big\}$ can be written as $\Big\{\mu_{a^*} - \mu_a - 2\sqrt{\frac{\log(k^nt^2/2)}{t}}\zeta_a < 0\Big\}$. Substituting $\Delta_a = \mu_{a^*} - \mu_a$ and since $t \geq L_a \geq L_{2,a}$, we have

$$\mathbb{P}\bigg(\Big\{\Delta_a - 2c_{a,t} < 0\Big\}\bigg) = 0\ . \tag{34}$$

Hence,

$$E[N'_{a,T}] \leq \sum_{t=L}^{T-1} \frac{4}{t^2} \leq \sum_{t=1}^{\infty} \frac{4}{t^2} \leq \frac{2\pi^2}{3}\ .$$

□

Now we bound the expected cumulative regret of C-UCB-2. From Equation 3 in Section 2, we have at the end of T rounds

$$\begin{aligned} E[R_{\texttt{C-UCB-2}}(T)] &= T\mu_{a^*} - \sum_{a\in\mathcal{A}} \mu_a E[N_{a,T}] \\ &= \sum_{a\in\mathcal{A}} \Delta_a E[N_{a,T}] \leq \sum_{a\in\mathcal{A}} \Delta_a (L_a + \frac{2\pi^2}{3}) \ . \end{aligned}$$

The inequality in the last line of the above equation follows from Lemma B.9.

C Rational for Simulation Parameters and Additional Experiments

C.1 Model and Parameter Choices in Section 6

Experiment 1: In this experiment the underlying causal graph and distributions $p_i = \mathbb{P}(X_i = 1)$ is the same as that in Lattimore et al. (2016). The expected reward of the best arm is set to 0.8 by choosing $\epsilon = 0.3$. Under these settings, Lattimore et al. (2016) demonstrated a faster exponential decay of simple regret compared to the non-causal algorithms. Since in this experiment we compare γ-NB-ALG to PB-ALG (adapted to the budgeted version), we choose the same causal graph and distribution.

Experiment 2: In this experiment 2 the underlying causal graph and distributions p_i are the same as in Experiment 1. If the reward distribution is the same as in experiment 1 the cumulative regret of CRM-NB-ALG even with $\gamma = 1.1$ converges very quickly to a small constant. This is attributed to the fact that the observation arm is closer to being optimal (i.e. d_0 is smaller). Even though this validates the better performance of our algorithm, for a better visual appeal we set the expected reward of the best arm to 1. Even with this reward distribution, the performance of CRM-NB-ALG is much better than F-KUBE.

Experiment 3: The causal graph used in this experiment is as shown in figure 2. Notice that this graph has a backdoor path from X_2 to Y and therefore algorithms such as PB-ALG , γ-NB-ALG and CRM-NB-ALG , which are for no-backdoor graphs, cannot be used. Our conditional probabilities for nodes ($P(\text{node}|Pa(\text{node}))$) are given in Table 1.

Conditional Variable	**Probability**
$X_1 = 0$	0.45
$X_1 = 1$	0.55
$X_2 = 0\|X_1 = 0$	0.55
$X_2 = 1\|X_1 = 0$	0.45
$X_2 = 0\|X_1 = 1$	0.45
$X_2 = 1\|X_1 = 1$	0.55
$W_1 = 0\|X_1 = 0$	0.46
$W_1 = 1\|X_1 = 0$	0.54
$W_1 = 0\|X_1 = 1$	0.54
$W_1 = 1\|X_1 = 1$	0.46
$W_2 = 0\|X_2 = 0$	0.52
$W_2 = 1\|X_2 = 0$	0.48
$W_2 = 0\|X_2 = 1$	0.48
$W_2 = 1\|X_2 = 1$	0.52

Table 1: Conditional Probability Distributions

The conditional distribution of the reward variable Y was chosen as $Y|w_1, w_2 = \theta_1 X_1 + \theta_2 X_2 + \epsilon$, where ϵ is distributed as $\mathcal{N}(0, 0.01)$. Here $\mathcal{N}(0, 0.01)$ denotes the normal distribution with mean 0 and standard deviation 0.01. Since we compare the performance of C-UCB-2 with C-UCB proposed by Lu et al. (2020), we choose our conditional distribution similar to that in Lu et al. (2020). The conditional probabilities in the above table are chosen to be close to each other in order to ensure that the expected rewards for all the arms are competitive and the algorithm takes longer to distinguish between them. The expected reward of the four arms $do(X_i = x), i \in [2], x \in \{0, 1\}$ are given in the Table 2.

Arm	Expected Reward
$do(X_1 = 0)$	0.2595
$do(X_1 = 1)$	0.2405
$do(X_2 = 0)$	0.244
$do(X_2 = 1)$	0.254

Table 2: Expected Reward of the Arms

C.2 Additional Experiment

In this section we show results for an additional experiment that validates our result in Theorem 1. The experiment 4 below plots the regret of OBS-ALG and PB-ALG (from Lattimore et al. (2016)) with respect to the minimum probability.

Experiment 4 **(OBS-ALG vs. PB-ALG)**: This experiment demonstrates the performance of OBS-ALG with respect to PB-ALG on the same graph and parameters as chosen in Experiment 1 of Section 6. In figure 4, we plot *simple regret* vs. *minimum probability*. Note that the minimum probability is $p = \min_{i,x}\{p_{i,x}\}$, as defined in Section 3.1. We fix the budget B to a moderate value 100 and cost of intervention γ to 1. Note that the performance of γ-NB-ALG is best for $\gamma = 1$. The plot shows an inverse relationship between simple regret of OBS-ALG and p as proved in Theorem 1, whereas the simple regret of PB-ALG does not depend on p. Recall that the expected simple regret of PB-ALG depends on $m(\mathbf{p})$ (and not on p) and for the p_i's used in Experiment 1, the quantity $m(\mathbf{p}) = 2$, does not change. Finally, also observe that after a threshold value of p, OBS-ALG starts performing much better than PB-ALG as can be seen from the plot.

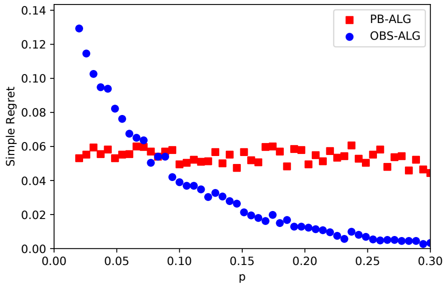

Figure 4: OBS-ALG vs PB-ALG