A Causal Bandit Approach to Learning Good Atomic Interventions in Presence of Unobserved Confounders (Supplementary material)

Aurghya Maiti[1] **Vineet Nair**[2] **Gaurav Sinha**[1]

[1]Adobe Research, Bangalore, India
[2]Technion Israel Institute of Technology, Haifa, Israel

A PRELIMINARY LEMMAS

We state some standard concentration bounds that are used in our proofs. Their proofs can be found in the citations provided.

Lemma A.1 (Chernoff Bounds, Section 4.2 in Mitzenmacher and Upfal [2005])
Let Z be any random variable. Then for any $t > 0$,

1. $\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \min_{\lambda > 0} \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}]e^{-\lambda t}$
2. $\mathbb{P}(Z \leq \mathbb{E}[Z] - t) \leq \min_{\lambda > 0} \mathbb{E}[e^{\lambda(\mathbb{E}[Z] - Z)}]e^{-\lambda t}$

Lemma A.2 (Hoeffding's Lemma, Lemma 2.6 in Massart and Picard [2007])
Let Z be a bounded random variable with $Z \in [a, b]$. Then,

$$\mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

for all $\lambda \in \mathbb{R}$.

Lemma A.3 (Chernoff-Hoeffeding inequality, Chernoff [1952], Hoeffding [1963])
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 $\overline{X} = \frac{1}{T}(\sum_{t \in [T]} X_t)$. Then for any $\varepsilon \geq 0$ the following holds:

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

B EXAMPLE OF CBN WITH $m(\mathcal{C}) \ll N$

Consider a CBN $\mathcal{C} = (\mathcal{G}, \mathbb{P})$ with N intervenable nodes and in-degree at most $k-1$, and let k be such that $2^k \ll N$. Further, let $\mathbb{P}$ be such that for at most 2^k nodes, chosen in the reverse topological order, the conditional probability of a node being 1 given its parents is Bernoulli with parameter $1/2^{k+1}$, and for the remaining nodes the conditional probability of a node being 1 given its parents is Bernoulli with parameter $1/2$. Now, using the definition of $m(\mathcal{C})$ provided in Section 3, it is easy to see that $m(\mathcal{C}) \leq 2^k \ll N$.

C ESTIMATION OF REWARD FROM OBSERVATION

In Algorithm C.1 below we explain our strategy (derived from Bhattacharyya et al. [2020]) for estimating the reward of the interventional arms $a_{i,x}$ using $T/2$ observational samples collected by playing the observational arm a_0. This is followed by details on each of the steps involved. Recall, N is the number of intervenable nodes.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Algorithm C.1 Estimating Rewards from Observational Samples

INPUT: His containing the $T/2$ observational samples collected by playing arm a_0, and $\mathcal{G}$

1: For each $i \in [N]$, reduce the input ADMG $\mathcal{G}$ to ADMG $\mathcal{H}_i$ as outlined in Algorithm C.2.

2: Next, for each $i \in [N]$ and $x \in \{0,1\}$, construct the Bayes net $D_{i,x}$ which simulates the causal effect of intervention $do(X_i = x)$ on the reduced graph $\mathcal{H}_i$.

3: Using Algorithm C.3 on the input samples, estimate the distributions of all $D_{i,x}$. Then, using learned $D_{i,x}$, generate samples to estimate marginal of Y and return them as estimated rewards.

Step 1 **:** This step executes Algorithm C.2 based on the reduction algorithm from Bhattacharyya et al. [2020].

Algorithm C.2 Reducing $\mathcal{G}$ to $\mathcal{H}_i$

INPUT: ADMG $\mathcal{G}$ and index $i \in [N]$.

Let $\mathbf{W} = Y \cup X_i \cup \mathbf{Pa}^c(X_i)$, and $\mathcal{G}'_i$ be the graph obtained by considering $\mathbf{V}\backslash\mathbf{W}$ as hidden variables. Let $\mathbf{V}_i$ denote the nodes in $\mathcal{G}'_i$

Projection Algorithm: It reduces $\mathcal{G}'_i$ to $\mathcal{H}_i$ as follows:

1. Add all observable variables in $\mathcal{G}'_i$ to $\mathcal{H}_i$.
2. For every pair of observable variable $V^i_j, V^i_k \in \mathbf{V}_i$, add a directed edge from V^i_j to V^i_k in $\mathcal{H}_i$, if (a) there exists a directed edge from V^i_j to V^i_k in $\mathcal{G}'_i$, or if (b) there exists a directed path from V^i_j to V^i_k in $\mathcal{G}'_i$ which contains only unobservable variables.
3. For every pair of observable variable $V^i_j, V^i_k \in \mathbf{V}_i$, add a bi-directed edge between V^i_j and V^i_k in $\mathcal{H}_i$, if (a) there exists an unobserved variable U with two directed paths in $\mathcal{G}'_i$ going from U to V^i_j and U to V^i_k and containing only unobservable variables.

Return $\mathcal{H}_i$.

Algorithm C.3 Estimating distributions of $D_{i,x}$

INPUT: ADMG $\mathcal{H}_i$ and $x \in \{0,1\}$.

for every $V_j \in S_1$ **do**
  for every assignment $V_j = v$ and $\mathbf{Z_j} = \mathbf{z}$ where $\mathbf{Z_j}$ are effective parents of V_j in $\mathcal{H}_i$ **do**
    $N_j \leftarrow$ the number of samples with $\mathbf{Z_j} = \mathbf{z}$
    $N_{j,v} \leftarrow$ the number of samples with $\mathbf{Z_j} = \mathbf{z}$ and $V_j = v$
    $\widehat{D}_{i,x}(V_j = v|\mathbf{Z_i} = \mathbf{z}) \leftarrow \frac{N_{j,v}+1}{N_j+2}$
for every $V_j \in \mathbf{V_i}\backslash\mathbf{S_1}$ **do**
  for every $V_j = v$ and $\mathbf{Z_j}\backslash X_i = \mathbf{z}$, where $\mathbf{Z_j}$ are effective parents of V_j in $\mathcal{H}_i$ **do**
    if $X \in \mathbf{Z_i}$ **then**
      $N_j \leftarrow$ the number of samples with $\mathbf{Z_j}\backslash X_i = \mathbf{z}$ and $X_i = x$
      $N_{j,v} \leftarrow$ the number of samples with $V_j = v$, $\mathbf{Z_j}\backslash X_i = \mathbf{z}$ and $X_i = x$
      if $N_j \geq t$ **then**
        $\widehat{D}_{i,x}(V_j = v|\mathbf{Z_i} = \mathbf{z}) \leftarrow \frac{N_{j,v}+1}{N_j+2}$
      else
        $\widehat{D}_{i,x}(V_j = v|\mathbf{Z_j} - \{X_i\} = \mathbf{z}, X_i = x) \leftarrow \frac{1}{2}$
    else
      $N_j \leftarrow$ the number of samples with $\mathbf{Z_j} = \mathbf{z}$
      $N_{j,v} \leftarrow$ the number of samples with $V_j = v$ and $\mathbf{Z_j} = \mathbf{z}$
      if $N_j \geq t$ **then**
        $\widehat{D}_{i,x}(V_j = v|\mathbf{Z_i} = \mathbf{z}) \leftarrow \frac{N_{j,v}+1}{N_j+2}$
      else
        $\widehat{D}_{i,x}(V_j = v|\mathbf{Z_j} = \mathbf{z}) \leftarrow \frac{1}{2}$
Return $\widehat{D}_{i,x}$.

Step 2 **:** Construction of $D_{i,x}$ is done using the method described in Section 4.1 of Bhattacharyya et al. [2020]. Without loss of generality let $\mathbf{S_1}$ be the c-component containing X_i. To construct $D_{i,x}$, we start with $\mathcal{H}_i$. Then, for each $V \notin \mathbf{S_1}$ such that X_i is in the set $\mathbf{Z_i}$ of "effective parents" (Section 4, Bhattacharyya et al. [2020]) of V, we create a clone of X_i and fix its value to x (i.e. the clone has no parents). Then we remove all the outgoing edges from the original X_i. Note that, for any assignment $\boldsymbol{v}$ of all variables except X_i in $\mathcal{H}_i$, the causal effect $\mathbb{P}_{\mathcal{H}_i}(\boldsymbol{v}|do(X_i = x)) = \sum_x \mathbb{P}_{D_{i,x}}(\boldsymbol{v}, X_i = x)$.

Step 3 **:** In this step, we estimate the distributions of all $D_{i,x}$ using the $T/2$ samples that were provided as input. Details are described in Algorithm C.3. Using this estimated distribution, we get $O(T)$ samples and compute an empirical estimate $\widehat{\mu}_{i,x}$ of the reward $\mu_{i,x} = \mathbb{P}_{\mathcal{G}}(Y = 1|do(X_i = x))$. This follows from the construction of $D_{i,x}$ in Step 2 which implies,

$$\mu_{i,x} = \mathbb{P}_{\mathcal{G}}(Y = 1|do(X_i = x)) = \mathbb{P}_{\mathcal{H}_i}(Y = 1|do(X_i = x)) = \sum_{x,\boldsymbol{v}'} \mathbb{P}_{D_{i,x}}(Y = 1, \boldsymbol{v}', X_i = x)$$

where $\boldsymbol{v}'$ is an assignment of nodes in $D_{i,x}$ other than X_i and Y.

D PROOF OF THEOREM 3.1

For the sake of analysis, we assume without loss of generality that $q_1, q_2, \ldots, q_N$ are arranged such that their corresponding c-component sizes $k_1, k_2, \ldots, k_N$ satisfy the following relation: $(q_1)^{k_1} \leq (q_2)^{k_2} \leq \ldots \leq (q_N)^{k_N}$. Also, let $q = \min_{i\{q_i>0\}} q_i$ (if $q_i = 0$ for all $i \in [N]$ then $q = \frac{1}{N+1}$), $k = \max_i k_i$, and $p^{i,x}_{\mathbf{z}} = \mathbb{P}(X_i = x, \mathbf{Pa}^c(X_i) = \mathbf{z})$. We remark that $p^{i,x}_{\mathbf{z}}$ is different from $p^{i,x}_{\mathbf{z}}$ used in Section 5 to denote $\mathbb{P}(X_i = x, \mathbf{Pa}(X_i) = \mathbf{z})$; note that $\mathbf{Pa}(X_i) \subseteq \mathbf{Pa}^c(X_i)$. Let d be the maximum indegree of any node in S_i for $i \in [N]$. Finally, let Z_i be the size of the domain from which $\mathbf{Pa}^c(X_i)$ takes values, and note that $Z_i \leq 2^{k_i d + k_i}$ and let $Z = \max_i Z_i$. Note that, by our assumption Z is $O(1)$. Also, in this section, let $m(\mathcal{C})$ be denoted by m.

We begin by proving Lemmas D.1, D.2, and D.3 which would be used to prove Theorem 3.1. The following lemma bounds the probability of making a bad estimate of q_i for any $i \in [N]$, at the end of $T/2$ rounds.

Lemma D.1
Let $F = \mathbb{1}\{$At the end of $T/2$ rounds, there exists i such that $|\widehat{q}_i - q_i| \geq \frac{1}{4}(1 - 2^{-1/k})q\}$. Then $\mathbb{P}(F = 1) \leq 4NZe^{-\frac{1}{16}(1-2^{-1/k})^2 q^2 T}$.

Proof. Let $F_{i,x} = \mathbb{1}\{$*At the end of $T/2$ rounds there exists* $\mathbf{z}$ *such that* $|\widehat{p}^{i,x}_{\mathbf{z}} - p^{i,x}_{\mathbf{z}}| \geq \frac{1}{4}(1 - 2^{-1/k})q\}$. From Lemma A.3, it follows that,

$$\mathbb{P}\Big(|\widehat{p}^{i,x}_{\mathbf{z}} - p^{i,x}_{\mathbf{z}}| \geq \frac{1}{4}(1 - 2^{-1/k})q\Big) \leq 2e^{-2\frac{1}{16}(1-2^{-1/k})^2 q^2 \frac{T}{2}}$$

By union bound,

$$\mathbb{P}(F_{i,x} = 1) \leq 2Z_i e^{-\frac{1}{16}(1-2^{-1/k})^2 q^2 T}$$

By definition $q_i = \min_{x,\mathbf{z}} p^{i,x}_{\mathbf{z}}$ and $\widehat{q}_i = \min_{x,\mathbf{z}} \widehat{p}^{i,x}_{\mathbf{z}}$. Hence,

$$\mathbb{P}\Big(|\widehat{q}_i - q_i| \geq \frac{1}{4}(1 - 2^{-1/k})q\Big) \leq 2P(F_{i,x} = 1) \leq 4Z_i e^{-\frac{1}{16}(1-2^{-1/k})^2 q^2 T}$$

Taking union bound, we get $\mathbb{P}(F = 1) \leq 4NZe^{-\frac{1}{16}(1-2^{-1/k})^2 q^2 T}$. □

The next lemma shows that with high probability the estimate of m at Step 6 of `SRM-ALG` is good.

Lemma D.2
Let F be as defined in Lemma D.1 and let $J = \mathbb{1}\{$At the end of $T/2$ rounds the following holds $\widehat{m} \leq 2m\}$. Then $F = 0$ implies $J = 1$, and in particular, $\mathbb{P}(J = 1) \geq 1 - 4NZe^{-\frac{1}{16}(1-2^{-1/k})^2 q^2 T}$.

Proof. Note that if $q_i = 0$ for all $i \in [N]$, then our proposition is trivially true. $F = 0$ implies after $T/2$ rounds for all $i \in [N]$, $|\widehat{q}_i - q_i| \leq \frac{1}{4}(1 - 2^{-1/k})q$. Now from definition of m we know that there is an $l \leq m$ such that for $i > l$,

$(q_i)^{k_i} \geq (\frac{1}{m})$. Hence, for $i > l$, since $q \leq q_i$ by definition

$$(\widehat{q}_i)^{k_i} \geq \Big(q_i - \frac{1}{4}(1-2^{-1/k})q\Big)^{k_i} \geq \Big(q_i - (1-2^{-1/k})q_i\Big)^{k_i} \geq \frac{1}{2^{k_i/k}m} \geq \frac{1}{2m}$$

Since, $l \leq m$, we have $|\{j \mid \widehat{q}_j^{k_j} < \frac{1}{2m}\}| \leq 2m$. This implies $\widehat{m} \leq 2m$.

$\square$

The next lemma provides the confidence bound on the estimate of $\mu_{i,x}$ computed by Algorithm C.1 for each i, x .

Lemma D.3
For an action $a_{i,x} \in \mathcal{A}$, at the end of $T/2$ rounds $\mathbb{P}(|\widehat{\mu}_{i,x} - \mu_{i,x}| > \epsilon) \leq \exp\big(-\epsilon^2 \frac{q_i^{k_i} T}{K_{\mathcal{G}}}\big)$, where $K_{\mathcal{G}} \geq 1$ is a constant dependent on the structure of $\mathcal{G}$ but independent of $\mathbb{P}$.

Proof. Using **Theorem 2.5** and **Theorem A.1** in Bhattacharyya et al. [2020], it can be inferred that the learner can estimate $\widehat{\mu}_{i,x}$, such that $|\widehat{\mu}_{i,x} - \mu_{i,x}| \leq \epsilon$, with probability $1-\delta_i$, using $O\big(2^{2u_i^2} \log 2^{2u_i^2} \log \frac{1}{\delta_i} / (q_i^{k_i}\epsilon^2)\big)$ samples, where $u_i = 1 + k_i(d+1)$. Hence using samples $T = K' \frac{2^{2.2u_i^2}}{q_i^{k_i}\epsilon^2} \log \frac{1}{\delta_i}$, where K' is a constant independent of the problem instance, we get, $P(|\widehat{\mu}_{i,x} - \mu_{i,x}| \leq \epsilon) \geq 1 - \delta_i$. Writing δ_i in terms of T and ϵ, and using $K_{\mathcal{G}} = \max\{1, K' 2^{2.2u_i^2}\}$,

$$\mathbb{P}(|\widehat{\mu}_{i,x} - \mu_{i,x}| > \epsilon) \leq \exp\bigg(-\frac{T}{K'}\frac{q_i^{k_i}\epsilon^2}{2^{2.2u_i^2}}\bigg) \leq \exp\bigg(-\epsilon^2 \frac{q_i^{k_i} T}{K_{\mathcal{G}}}\bigg)$$

Also by A.3, for a_0, by,

$$\mathbb{P}(|\widehat{\mu}_0 - \mu_0| \geq \epsilon) \leq \exp\bigg(-2\epsilon^2 \frac{T}{2}\bigg).$$

$\square$

Now we are ready to prove the theorem using the above Lemmas, and let $K = 2^{k-1}K_{\mathcal{G}}$. Let $L_1 = \min_{t \in \mathbb{N}}(4NZe^{-\frac{1}{16}(1-2^{-1/k})^2 q^2 t} \leq \sqrt{\frac{144Km}{t}\log\frac{Nt}{m}})$ and $L_2 = \min_{t\in\mathbb{N}} \frac{6}{N^3}(\frac{m}{t})^4 \leq \sqrt{\frac{16Km}{t}\log\frac{Nt}{m}}$ and we assume throughout the proof that $T \geq \max\{L_1, L_2\}$. Consider $a_{i,x} \in \mathcal{Q}$. By Lemma A.3, and Lemma D.2,

$$\mathbb{P}\big(|\widehat{\mu}_{i,x} - \mu_{i,x}| \geq \epsilon \mid F = 0\big) \leq 2\exp\bigg(-\epsilon^2 \frac{2T}{4\widehat{m}}\bigg) \leq 2\exp\bigg(-\epsilon^2 \frac{T}{4m}\bigg) \leq 2\exp\bigg(-\epsilon^2 \frac{T}{4Km}\bigg)$$

If $a_{i,x} \notin \mathcal{Q}$, and $q_i^{k_i} \geq \frac{1}{m}$, then given $F = 0$ we get,

$$\mathbb{P}\big(|\widehat{\mu}_{i,x} - \mu_{i,x}| > \epsilon \mid F = 0\big) \leq \exp\bigg(-\epsilon^2 \frac{q_i^{k_i} T}{K_{\mathcal{G}}}\bigg) \leq \exp\bigg(-\epsilon^2 \frac{T}{4Km}\bigg)$$

If $a_{i,x} \notin \mathcal{Q}$, and $q_i^{k_i} < \frac{1}{m}$, then given $F = 0$ from Lemma D.1, $q_i^{k_i} \geq (\widehat{q}_i - \frac{1}{4}(1-2^{-1/k})q)^{k_i} \geq ((\frac{1}{\widehat{m}})^{1/k_i} - \frac{1}{4}(\frac{1}{m})^{1/k_i}))^{k_i} \geq ((\frac{1}{2m})^{1/k_i} - \frac{1}{4}(\frac{1}{m})^{1/k_i}))^{k_i} \geq \frac{1}{2^{k+1}m}$ we get,

$$\mathbb{P}\big(|\widehat{\mu}_{i,x} - \mu_{i,x}| > \epsilon \mid F = 0\big) \leq \exp\bigg(-\epsilon^2 \frac{q_i^{k_i} T}{K_{\mathcal{G}}}\bigg) \leq \exp\bigg(-\epsilon^2 \frac{T}{2^{k+1}K_{\mathcal{G}}m}\bigg) \leq \exp\bigg(-\epsilon^2 \frac{T}{4Km}\bigg)$$

$$\begin{aligned}\mathbb{P}\{\textit{There exists an action } a \textit{ such that } |\widehat{\mu}_a - \mu_a| > \epsilon \mid F = 0\} &\leq (4N+2)\exp\bigg(-\epsilon^2 \frac{T}{4Km}\bigg) \\ &\leq 6N \exp\bigg(-\epsilon^2 \frac{T}{4Km}\bigg)\end{aligned}$$

Substituting $\epsilon = \sqrt{\frac{16Km}{T}\log\frac{NT}{m}}$, we get,

$$E[r_T \mid F=0] \le 2\sqrt{\frac{16Km}{T}\log\frac{NT}{m}} + \frac{6}{N^3}\left(\frac{m}{T}\right)^4 \le \sqrt{\frac{144Km}{T}\log\frac{NT}{m}}$$

Finally, the expected simple regret of Algorithm 1 is as follows:

$$\begin{aligned} E[r_T] &= E[r_T|F=0]\mathbb{P}(F=0) + E[r_T|F=1]\mathbb{P}(F=1) \\ &\le E[r_T|F=0] + \mathbb{P}(F=1) \\ &\le \sqrt{\frac{144Km}{T}\log\frac{NT}{m}} + 4NZe^{-\frac{1}{16}(1-2^{-1/k})^2q^2T} \end{aligned}$$

Since $T \ge \max(L_1, L_2)$ the simple regret is $\mathcal{O}\left(\sqrt{\frac{m}{T}\log\frac{NT}{m}}\right)$.

E PROOF OF THEOREM 4.1

Throughout this proof we assume the following terminology: a) a node is a root node if it has not parents, b) a node is a leaf node if it has no children. Consider an n-ary tree $\mathcal{T} \in \mathsf{T}$ on N intervenable nodes. Note that since $\mathcal{T}$ is a tree, each node X_i for $i \in [N]$ has at most one parent. In addition $\mathcal{T}$ has one special node Y, called the outcome. There is a directed from every leaf node in $\mathcal{T}$ to Y, and let $L_{\mathcal{T}}$ be the set of all leaf nodes. We use $\mathbf{V}$ to denote the set of nodes in $\mathcal{T}$, that is, $\mathbf{V} = \{X_1, \ldots, X_N, Y\}$. Without loss of generality, we assume that $X_1, \ldots, X_N$ is in the reverse topological order, that is, X_1 is a leaf node, X_N is a root node, X_{N-1} is either a root node or a child of X_N, and so on. Let $\mathcal{T}_M$ be the sub-graph of $\mathcal{T}$ defined by the nodes $X_1, \ldots, X_M$. An edge belongs to $\mathcal{T}_M$ if both its endpoints belong to $\{X_1, \ldots, X_M\}$. Further, let h be the maximum number of nodes in a (directed) path from a root node to Y. Now we define distributions $\mathbb{P}_0, \ldots, \mathbb{P}_M$ all compatible with $\mathcal{T}$ such that the optimal arm in the CBN $\mathcal{C}_i = (\mathcal{T}, \mathbb{P}_i)$ is $a_{i,1}$ for $i \in [M]$, and for $\mathcal{C}_0 = (\mathcal{T}, \mathbb{P}_0)$ every arm is an optimal arm.

Defining $\mathbb{P}_0$: For X_i not belonging to $\mathcal{T}_M$ let $\mathbb{P}_0(X_i = 1|.) = 0.5$, and for X_i belonging to $\mathcal{T}_M$ and for an appropriately chosen α let

$$\begin{aligned} &\mathbb{P}_0(X_i = 1) = \alpha && \text{If } X_i \text{ is a root node,} \\ &\mathbb{P}_0(X_i = 1 \mid \mathbf{Pa}(X_i) = 0) = \alpha && \text{If } X_i \text{ is not a root node,} \\ &\mathbb{P}_0(X_i = 1 \mid \mathbf{Pa}(X_i) = 1) = 1-\alpha && \text{If } X_i \text{ is not a root node,} \\ &\mathbb{P}_0(Y = 1 \mid .) = 0.5 \quad \mathbb{P}_0(Y = 0 \mid .) = 0.5 \end{aligned}$$

The value of α is appropriately chosen later to achieve the desired lower bound. Note that in the above equations if X_i is not a root node then $\mathbf{Pa}(X_i)$ is a singleton set. Also, $\mathbb{P}_0(Y = 1|.)$ denotes the probability of $Y = 1$ conditioned on any value of its parents. Next, we define $\mathbb{P}_i$ for $i \in [N]$.

Defining $\mathbb{P}_i$: Let L_i be the set of leaf nodes that are reachable from X_i, that is there is a directed path from X_i to every leaf node in L_i. Note that if X_i is a leaf then $L_i = \{X_i\}$. We use $L_i = \mathbf{1}$ and $L_i = \mathbf{0}$ to denote all nodes in L_i evaluated to 1 and 0 respectively. Also, let $L_{\mathcal{T}}^M$ be the set of all leaves in $\mathcal{T}_M$ and $L_i' = L_{\mathcal{T}}^M \setminus L_i$. Then

$$\mathbb{P}_i(Y|L_i = \mathbf{1}, L_i' = \mathbf{0}) = 0.5 + \epsilon\ .$$

The value of ϵ is appropriately chosen later to achieve the desired lower bound. The distributions of X_i given its parents corresponding to $\mathbb{P}_i$ is the same as those defined for $\mathbb{P}_0$.

We set $\alpha = \min\{(2h|L_{\mathcal{T}}| + 2^{h+1})^{-1}, (2^h|L_{\mathcal{T}}|M)^{-1}\}$ and hence $\alpha < \frac{1}{M}$. Using this it is easy to see that $m(\mathcal{C}_i) = M$ for $i \in [0, M]$ and $M > 4$. Additionally, in $\mathcal{C}_i$ arm $a_{i,1}$ is the optimal arm for $i \in [1, M]$ and the reward for every arm in $\mathcal{C}_0$ is 0.5. We will denote a^* as the optimal arm for every $\mathcal{C}_i$, and note that $a^* = a_{i,1}$ for $\mathcal{C}_i$, where $i \in [M]$. First, in Lemma E.1, we lower bound the regret of returning a sub-optimal arm in $\mathcal{C}_i$ at the end of T rounds. Further, in Lemma E.2, we show that any algorithm would have a non-trivial probability of returning a sub-optimal arm in at least one of the constructed CBNs. Finally, we would use Lemmas E.1 and E.2 to lower bound the expected regret of any algorithm. Let $\mathrm{rew}_i(a_{j,x})$ denote the expected reward of action $do(X_j = x)$ under the distribution $\mathbb{P}_i$. We deviate from the usual notation of μ in this case, because the reward now depends on the arm and the corresponding distribution. We require the following sets in Lemmas E.1 and E.2: $V_1 = L_i \setminus L_j$, $V_2 = L_i \cap L_j$, $V_3 = L_j \setminus L_i$, $V_4 = L_{\mathcal{T}}^M \setminus (L_i \cup L_j)$, and $V_5 = V \setminus L_{\mathcal{T}}^M$.

Lemma E.1
For every $i \in [1, M]$, $j \in [1, N]$, $x \in \{0, 1\}$, and $(j, x) \neq (i, 1)$ the following holds: rew$_i(a_{i,1})$ − rew$_i(a_{j,x}) \geq 0.5\epsilon$.

Proof. For any $i, j \in [M]$, we have

$$\text{rew}_i(a_{i,1}) = 0.5 + \mathbb{P}_i(V_4 = \mathbf{0}, V_1 = \mathbf{1}, V_2 = \mathbf{1}, V_3 = \mathbf{0} \mid do(X_i = 1))(\epsilon) \tag{E.1}$$
$$\text{rew}_i(a_{j,1}) = 0.5 + \mathbb{P}_i(V_4 = \mathbf{0}, V_1 = \mathbf{1}, V_2 = \mathbf{1}, V_3 = \mathbf{0} \mid do(X_j = 1))(\epsilon) \tag{E.2}$$

Subtracting Equation E.2 from Equation E.1 we have

$$\begin{aligned}
&\text{rew}_i(a_{i,1}) - \text{rew}_i(a_{j,1}) \\
&= \mathbb{P}_i(V_4 = \mathbf{0})\big[\mathbb{P}_i(V_1 = \mathbf{1}, V_2 = \mathbf{1}, V_3 = \mathbf{0} \mid do(X_i = 1)) - \mathbb{P}_i(V_1 = \mathbf{1}, V_2 = \mathbf{1}, V_3 = \mathbf{0} \mid do(X_j = 1))\big]\epsilon \\
&= \mathbb{P}_i(V_4 = \mathbf{0})\big[\mathbb{P}_i(V_3 = \mathbf{0})\mathbb{P}_i(V_1 = \mathbf{1}, V_2 = \mathbf{1} \mid do(X_i = 1)) - \mathbb{P}_i(V_1 = \mathbf{1})P(V_2 = \mathbf{1}, V_3 = \mathbf{0} \mid do(X_j = 1))\big]\epsilon \\
&\underset{(i)}{\geq} (1-\alpha)^{h|V_4|}\big[(1-\alpha)^{h(|L_i|+|V_3|)} - (2^h\alpha)\big]\epsilon \\
&\geq ((1-\alpha)^{h|L_{\mathcal{T}}|} - 2^h\alpha)\epsilon \\
&\geq ((1 - h|L_{\mathcal{T}}|\alpha) - 2^h\alpha)\epsilon \\
&\geq 0.5\epsilon
\end{aligned}$$

(i) in the above equations follows from the definitions of h and $\mathbb{P}_i$. Similarly, it can be shown that $\text{rew}_i(a_{i,1}) - \text{rew}_i(a_{j,0}) \geq 0.5\epsilon$ for $j \in [N]$, and $\text{rew}_i(a_{i,1}) - \text{rew}_i(a_{j,1}) \geq 0.5\epsilon$ for $j \in [M+1, N]$. Also $\text{rew}_i(a_{i,1}) - \text{rew}_i(a_0) \geq 0.5\epsilon$. ☐

Let ALG be an algorithm that outputs arm a_T at the end of T rounds. We choose $\epsilon = \min\{\frac{1}{4}, \sqrt{\frac{M}{18T}}\}$. Note that corresponding to every $\mathcal{C}_i$ for $i \in [0, M]$, ALG and $\mathbb{P}_i$ together define a probability measure on all the sampled values of the nodes of $\mathcal{T}$ over T rounds. Denote $\mathbb{D}_i$ as this measure and $\mathbb{E}_i$ as the expectation over $\mathbb{D}_i$ for $i \in [0, M]$. Let $\mathcal{G}_t$ be the sampled values of the nodes of $\mathcal{T}$ at time t and let $\mathbf{G}_t = \{\mathcal{G}_1, \ldots, \mathcal{G}_t\}$. Also, for $i \in [0, M]$ let $\mathbb{D}_i(.|\mathbf{G}_{t-1}) = \mathbb{P}_i^t(.)$; here $\mathbb{D}_i(.|\mathbf{G}_{t-1})$ denotes the probability of the sampled values of the nodes of $\mathcal{G}$ conditioned on its history till time $t-1$. Observe that conditioned on history $\mathbf{G}_{t-1}$, ALG determines an arm, say a_t, to pull at time t (either deterministically or in a randomized way), and for $j, j' \in [1, N]$ if $a_t = a_{j,x}$ then $\mathbb{P}_i^t(X_{j'} = x | do(X_j = x)) = \mathbb{P}_i(X_{j'} = x | do(X_j = x))$.

Lemma E.2
For any algorithm ALG there exists an $i \in [M]$ such that $\mathbb{D}_i(a_T \neq a_{i,1}) \geq \frac{\frac{M}{4e} - 1}{M}$.

Proof. We use $KL(\mathbb{D}_0, \mathbb{D}_i)$ to denote the KL divergence between $\mathbb{D}_0$ and $\mathbb{D}_i$ for any $i \in [M]$. Let $N_T^{(i,1)}$ be the number of times ALG plays the arm $a_{i,1}$ at the end of T rounds. Also, let $\mathcal{B} = \{a_{i,1} \mid i \leq M \textit{ and } \mathbb{E}_0[N_T^{(i,1)}] \leq 2T/M\}$. Observe that $|\mathcal{B}| \geq M/2$, as otherwise the sum of the expected number of arm pulls of arms not in $\mathcal{B}$ would be greater than T. First, using Lemma 2.6 from Tsybakov [2008], we have,

$$\mathbb{D}_0(a_T = a_{i,1}) + \mathbb{D}_i(a_T \neq a_{i,1}) \geq \frac{1}{2} \cdot \exp\left(-KL(\mathbb{D}_0, \mathbb{D}_i)\right)$$

Rearranging and summing the above equation over arms in $\mathcal{B}$, and observing that $\sum_{a_{i,1} \in \mathcal{B}} \mathbb{D}_0(a_T = a_{i,1}) \leq 1$ we have

$$\sum_{a_{i,1} \in \mathcal{B}} \mathbb{D}_i(a_T \neq a_{i,1}) \geq \frac{1}{2} \cdot \sum_{a_{i,1} \in \mathcal{B}} \exp(-KL(\mathbb{D}_0, \mathbb{D}_i)) - 1 \tag{E.3}$$

Now we bound $\exp(-KL(\mathbb{D}_0, \mathbb{D}_i))$ for every i such that $a_{i,1} \in \mathcal{B}$. Using the chain rule for product distributions (see Auer et al. [1995] and Chapter 2 in Slivkins [2019]) the KL divergence of $\mathbb{D}_0$ and $\mathbb{D}_i$ for any $i \in [M]$ can be written as

$$KL(\mathbb{D}_0, \mathbb{D}_i) = \sum_{t=1}^{T} KL(\mathbb{D}_0(\mathcal{G}_t|\mathbf{G}_{t-1}), \mathbb{D}_i(\mathcal{G}_t|\mathbf{G}_{t-1}) = \sum_{t=1}^{T} KL(\mathbb{P}_0^t(\mathcal{G}_t), \mathbb{P}_i^t(\mathcal{G}_t)) \tag{E.4}$$

Each term on the right hand side of the above summation can be computed as follows:

$$\begin{aligned}
KL(\mathbb{P}_0^t, \mathbb{P}_i^t) &= \sum_{\mathbf{v}} \mathbb{P}_0^t(V = \mathbf{v}) \log \frac{\mathbb{P}_0^t(\mathbf{V} = \mathbf{v})}{\mathbb{P}_i^t(\mathbf{V} = \mathbf{v})} \\
&\underset{(i)}{=} \sum_{x, \mathbf{v_5}} \mathbb{P}_0^t(Y = x, L_i = \mathbf{1}, L_i' = \mathbf{0}, V_5 = \mathbf{v_5}) \log \frac{\mathbb{P}_0^t(Y = x | L_i = \mathbf{1}, L_i' = \mathbf{0}, V_5 = \mathbf{v_5})}{\mathbb{P}_i^t(Y = x | L_i = \mathbf{1}, L_i' = \mathbf{0}, V_5 = \mathbf{v_5})} \\
&\underset{(ii)}{=} 0.5 \cdot \mathbb{P}_0^t(L_i = \mathbf{1}, L_i' = \mathbf{0}) \Big[\log \frac{0.5}{0.5 + \epsilon} + \log \frac{0.5}{0.5 - \epsilon} \Big] \\
&\underset{(iii)}{\leq} 0.5 \Big(\mathbb{P}_0^t\{do(X_i = 1)\} + 2^h |L_{\mathcal{T}}| \alpha \Big) \log \frac{0.25}{0.25 - \epsilon^2} \\
&= -0.5 \Big(\mathbb{P}_0^t\{do(X_i = 1)\} + 2^h |L_{\mathcal{T}}| \alpha \Big) \log(1 - 4\epsilon^2) \\
&= 0.5 \Big(\mathbb{P}_0^t\{do(X_i = 1)\} + 2^h |L_{\mathcal{T}}| \alpha \Big) \Big(4\epsilon^2 + \frac{(4\epsilon^2)^2}{2} + \frac{(4\epsilon^2)^3}{3} + \dots \Big) \\
&\leq 6 \Big(\mathbb{P}_0^t\{do(X_i = 1)\} + 2^h |L_{\mathcal{T}}| \alpha \Big) \epsilon^2 \, .
\end{aligned} \tag{E.5}$$

In the above equations: (i) follows by observing that for every other evaluation of $\mathbf{V}$ the distributions $\mathbb{P}_0^t$ and $\mathbb{P}_i^t$ are same hence the corresponding terms in KL divergence amount to zero, (ii) follows from the definitions of $\mathbb{P}_0^t$ and $\mathbb{P}_i^t$, and (iii) follows by observing that

$$\mathbb{P}_0^t(L_i = \mathbf{1}, L_i' = \mathbf{0}) \leq \mathbb{P}_0^t\{do(X_i = 1)\} + 2^h |L_{\mathcal{T}}| \alpha \, .$$

Using Equations E.4 and E.5, we have for every $a_{i,1} \in \mathcal{B}$,

$$KL(\mathbb{D}_0, \mathbb{D}_i) \leq \sum_{t=1}^{T} 6 \big(\mathbb{E}_0[N_T^{(i,1)}] + 2^h |L_{\mathcal{T}}| \alpha T \big) \epsilon^2 \underset{(i)}{\leq} \frac{18T}{M} \epsilon^2 \leq 1 \, , \tag{E.6}$$

where (i) follows from the definition of $\mathcal{B}$. Finally, using Equations E.3 and E.6, and $|\mathcal{B}| \geq M/2$, we have

$$\begin{aligned}
\sum_{a_{i,1} \in \mathcal{B}} \mathbb{D}_i(a_T \neq a_{i,1}) &\geq \frac{1}{2} \sum_{a_{i,1} \in \mathcal{B}} \exp(-KL(\mathbb{D}_0, \mathbb{D}_i)) - 1 \\
&\geq \frac{|\mathcal{B}|}{2e} - 1 \\
&\geq \frac{M}{4e} - 1 \, .
\end{aligned}$$

Therefore as $|\mathcal{B}| \leq M$, by averaging argument there exists an $i \in [M]$ such that

$$\mathbb{D}_i(a_T^* \neq a_{i,1}) \geq \frac{\frac{M}{4e} - 1}{M} \, .$$

$\square$

From Lemmas E.1 and E.2 for any algorithm ALG, if $\epsilon < \frac{1}{4}$ then the expected simple regret of ALG can be upper bounded as follows

$$r_{\texttt{ALG}}(T) \geq \mathbb{D}_i(a_T^* \neq a_{i,1}) \frac{1}{2} \epsilon \geq \frac{\frac{M}{4e} - 1}{M} \cdot \big(\frac{1}{2} \epsilon\big) \geq \frac{\frac{M}{4e} - 1}{2M} \sqrt{\frac{M}{18T}} \, . \tag{E.7}$$

On the contrary, if $\epsilon \geq \frac{1}{4}$ then $M \geq T$, so $\sqrt{M/T} = \Omega(1)$ and regret $r_{\texttt{ALG}}(T) \geq \Omega(1)$. Hence, for any algorithm there exists an $i \in [0, M]$ such that the expected simple regret of the algorithm on $\mathcal{C}_i$ is $\Omega\left(\sqrt{\frac{m(\mathcal{C}_i)}{T}}\right)$.

F PROOF OF THEOREM 4.2

We begin by constructing the causal graph $\mathcal{G}$ on $N+1$ nodes $\{X_1, \ldots, X_N, Y\}$, where $N \geq 3$. In $\mathcal{G}$, X_N is the parent of $X_1, \ldots, X_{N-1}$ and there is a directed edge form each node to the outcome node Y. The strategy remains the same as in the proof of Theorem 4.1; Now given $q_1, q_2, \ldots, q_N$, compatible with the graph $\mathcal{G}$, we will construct $\mathbb{P}_0, \ldots, \mathbb{P}_N$ such that on at least one CBN $\mathcal{C}_i = (\mathcal{G}, \mathbb{P}_i)$ the expected simple regret of any algorithm is tight. Also, without loss of generality, assume that $q_1 \leq q_2 \leq \cdots \leq q_N$.

Defining $\mathbb{P}_0$: For all the nodes in the graph $\mathcal{G}$, we define the distribution $\mathbb{P}_0$ as follows:

$$\begin{aligned}
&\mathbb{P}_0(X_N = 1) = q_N \\
&\mathbb{P}_0(X_i = 1 | X_N = 0) = \frac{q_i}{1 - q_N} \\
&\mathbb{P}_0(X_i = 1 | X_N = 1) = \frac{1}{2} \\
&\mathbb{P}_0(Y = 1 | .) = 0.5
\end{aligned}$$

$\mathbb{P}_0(Y = 1|.)$ denotes the probability of $Y = 1$ conditioned on any value of the parents. Also, note that since $q_1, \ldots, q_N$ are compatible with the given graph $\mathcal{G}$, we have, for any $i \neq N$, $q_i = \min_{x_i, x_N} \mathbb{P}_0(X_i = x_i, X_N = x_N) \leq \mathbb{P}_0(X_i = 1, X_N = 1) = q_N/2$. In addition, $\mathbb{P}_0(X_i = 1|X_N = 0) = q_i/(1 - q_N) \leq 2q_i$. Let $M = m(\mathcal{C}_i)$ for all $i \in [N]$ and $M' = M - 1$.

Case a: $M \geq 12$.

Defining $\mathbb{P}_i$: For $i = N$, define $\mathbb{P}_N(Y = 1|X_N = 1) = 0.5 + \epsilon$, and for $i \neq N$, $\mathbb{P}_i(Y = 1|X_i = 1, X_N = 0) = 0.5 + \epsilon$. The remaining conditional distributions are same as those of $\mathbb{P}_0$.

Now, it is easy to see that the optimal action for $\mathbb{P}_i$ is $a_{i,1}$. As in proof of Theorem 4.1, let $\text{rew}_i(a_{j,x})$ denote the expected reward of action $do(X_j = x)$ under the distribution $\mathbb{P}_i$.

Lemma F.1
For every $i \in [M']$, $j \in [N]$, $x \in \{0, 1\}$, and $(j, x) \neq (i, 1)$ the following holds: $\text{rew}_i(a_{i,1}) - \text{rew}_i(a_{j,x}) \geq 0.1\epsilon$.

Proof. For $i = N$, the regret for choosing a sub-optimal arm a is $\text{rew}_N(a_{N,1}) - \text{rew}_N(a) \geq (1 - q_N)\epsilon \geq 0.5\epsilon$. For $i \neq N$, the regret for choosing a sub-optimal arm $a_{j,x}$, where $j \neq N$ is as follows:

$$\begin{aligned}
\text{rew}_i(a_{i,1}) - \text{rew}_i(a_{j,x}) &\geq (1 - q_N)\epsilon - q_i\epsilon \\
&\geq \left(1 - \frac{3q_N}{2}\right)\epsilon \\
&\geq 0.25\epsilon
\end{aligned}$$

For $j = N$, the regret is as follows:

$$\begin{aligned}
&\text{rew}_i(a_{i,1}) - \text{rew}_i(a_{N,0}) = (1 - q_1)\epsilon - \mathbb{P}_i(X_i = 1|X_N = 0)\epsilon \geq (0.5 - 2q_i)\epsilon \\
&\text{rew}_i(a_{i,1}) - \text{rew}_i(a_{N,1}) = (1 - q_1)\epsilon \geq 0.5\epsilon
\end{aligned}$$

Hence, if $q_i \leq 1/M' \leq \frac{1}{5}$, the regret of pulling a sub-optimal arm is 0.1ϵ.

$\square$

Let $\texttt{ALG}$ be an algorithm that outputs arm a_T at the end of T rounds. We choose $\epsilon = \min\{\frac{1}{4}, \sqrt{\frac{M'}{24T}}\}$. For $i \in [N]$, denote $\mathbb{D}_i$ as the measure on all the sampled values of the nodes of $\mathcal{G}$ over T rounds and $\mathbb{E}_i$ as the expectation over $\mathbb{D}_i$. Let $\mathcal{G}_t$ be the sampled values of the nodes of $\mathcal{G}$ at time t and let $\mathbf{G}_t = \{\mathcal{G}_1, \ldots, \mathcal{G}_t\}$. Also, for $i \in [0, M']$ let $\mathbb{D}_i(.|\mathbf{G}_{t-1}) = \mathbb{P}_i^t(.)$. Note that $\texttt{ALG}$ determines the arm a_t conditioned on $\mathbf{G}_{t-1}$ (either in a deterministic or randomized way). Also for $j, j' \in [1, N]$, if $a_t = a_{j,x}$ and $j' \neq j$, then $\mathbb{P}_i^t(X_{j'} = x|do(X_j) = x) = \mathbb{P}_i(X_{j'} = x|do(X_j = x))$.

Lemma F.2
For any algorithm $\texttt{ALG}$, there exists an $i \in [M']$, such that $\mathbb{D}_i(a_T \neq a_{i,1}) \geq \frac{\frac{M'}{4e} - 1}{M'}$.

Proof. We use $KL(\mathbb{D}_0, \mathbb{D}_i)$ to denote the KL divergence between $\mathbb{D}_0$ and $\mathbb{D}_i$ for any $i \in [N]$. Let $N_T^{(i,1)}$ be the number of times `ALG` plays the arm $a_{i,1}$ at the end of T rounds. Also, let $\mathcal{B} = \{a_{i,1} \mid i \le M' \textit{ and } \mathbb{E}_0[N_T^{(i,1)}] \le 2T/M'\}$. Observe that $|\mathcal{B}| \ge M'/2$, as otherwise the sum of the expected number of arm pulls of arms not in $\mathcal{B}$ would be greater than T. First, using Lemma 2.6 from Tsybakov [2008], we have,

$$\mathbb{D}_0(a_T = a_{i,1}) + \mathbb{D}_i(a_T \neq a_{i,1}) \ge \frac{1}{2} \cdot \exp\left(-KL(\mathbb{D}_0, \mathbb{D}_i)\right)$$

Rearranging and summing the above equation over arms in $\mathcal{B}$, and observing that $\sum_{a_{i,1} \in \mathcal{B}} \mathbb{D}_0(a_T = a_{i,1}) \le 1$ we have

$$\sum_{a_{i,1} \in \mathcal{B}} \mathbb{D}_i(a_T \neq a_{i,1}) \ge \frac{1}{2} \cdot \sum_{a_{i,1} \in \mathcal{B}} \exp(-KL(\mathbb{D}_0, \mathbb{D}_i)) - 1 \tag{F.8}$$

Now we bound $\exp(-KL(\mathbb{D}_0, \mathbb{D}_i))$ for every i such that $a_{i,1} \in \mathcal{B}$. Using the chain rule for product distributions (see Auer et al. [1995] and Chapter 2 in Slivkins [2019]) the KL divergence of $\mathbb{D}_0$ and $\mathbb{D}_i$ for any $i \in [M]$ can be written as

$$KL(\mathbb{D}_0, \mathbb{D}_i) = \sum_{t=1}^{T} KL(\mathbb{D}_0(\mathcal{G}_t|\mathbf{G}_{t-1}), \mathbb{D}_i(\mathcal{G}_t|\mathbf{G}_{t-1}) = \sum_{t=1}^{T} KL(\mathbb{P}_0^t(\mathcal{G}_t), \mathbb{P}_i^t(\mathcal{G}_t)) \tag{F.9}$$

Now each term in the summation can be written as, for $i \neq N$,

$$\begin{aligned}
&KL(\mathbb{P}_0^t, \mathbb{P}_i^t) \\
&= \sum_{\mathbf{v}} \mathbb{P}_0^t(\mathbf{v}) \log \frac{\mathbb{P}_0^t(\mathbf{v})}{\mathbb{P}_i^t(\mathbf{v})} \\
&= \sum_{y} \mathbb{P}_0^t(Y = y|X_N = 0, X_i = 1)\mathbb{P}_0^t(X_N = 0, X_i = 1) \log \frac{\mathbb{P}_0^t(Y = y|X_N = 0, X_i = 1)}{\mathbb{P}_i^t(Y = y|X_N = 0, X_i = 1)} \\
&= 0.5\mathbb{P}_0^t(X_N = 0, X_i = 1)\left[\log \frac{0.5}{0.5+\epsilon} + \log \frac{0.5}{0.5-\epsilon}\right] \\
&\le 6\mathbb{P}_0^t(X_N = 0, X_i = 1)\epsilon^2
\end{aligned} \tag{F.10}$$

For $i = N$,

$$\begin{aligned}
KL(\mathbb{P}_0^t, \mathbb{P}_i^t) &= \sum_{\mathbf{v}} \mathbb{P}_0^t(\mathbf{v}) \log \frac{\mathbb{P}_0^t(\mathbf{v})}{\mathbb{P}_i^t(\mathbf{v})} \\
&= \sum_{y} \mathbb{P}_0^t(Y = y|X_N = 1)\mathbb{P}_0^t(X_N = 1) \log \frac{\mathbb{P}_0^t(Y = y|X_1 = 1)}{\mathbb{P}_i^t(Y = y|X_N = 1)} \\
&= 0.5\mathbb{P}_0^t(X_N = 1)\left[\log \frac{0.5}{0.5+\epsilon} + \log \frac{0.5}{0.5-\epsilon}\right] \\
&\le 6\mathbb{P}_0^t(X_N = 1)\epsilon^2
\end{aligned} \tag{F.11}$$

Using Equation F.10 and F.11 in equation F.9, we get when $q_i \le \frac{1}{M'}$

$$\begin{aligned}
KL(\mathbb{D}_0, \mathbb{D}_i) &\le 6\left[\mathbb{E}_0[N_T^{(i,1)}] + \frac{2}{M'}T\right]\epsilon^2 \\
&\le \frac{24T}{M'}\epsilon^2 \\
&\le 1
\end{aligned}$$

Now putting the value of $KL(\mathbb{D}_0, \mathbb{D}_i)$ in Equation F.8 we get the following,

$$\begin{aligned}\sum_{a_{i,1}\in\mathcal{B}} \mathbb{D}_i(a_T \neq a_{i,1}) &\geq \frac{1}{2}\sum_{a_{i,1}\in\mathcal{B}} \exp(-KL(\mathbb{D}_0, \mathbb{D}_i)) - 1\\ &\geq \frac{|\mathcal{B}|}{2e} - 1\\ &\geq \frac{M'}{4e} - 1\,.\end{aligned}$$

Therefore as $|\mathcal{B}| \leq M'$, by averaging argument there exists an $i \in [M']$ such that

$$\mathbb{D}_i(a_T^* \neq a_{i,1}) \geq \frac{\frac{M'}{4e} - 1}{M'}\,.$$

From Lemmas F.1 and F.2 for any algorithm $\mathtt{ALG}$, if $\epsilon < \frac{1}{4}$ then the expected simple regret of $\mathtt{ALG}$ can be upper bounded as follows

$$r_{\mathtt{ALG}}(T) \geq \mathbb{D}_i(a_T^* \neq a_{i,1}) \cdot (0.1\epsilon) \geq \frac{\frac{M'}{4e} - 1}{M'} \cdot (0.1\epsilon) \geq \frac{\frac{M'}{4e} - 1}{10M'}\sqrt{\frac{M'}{24T}}\,. \tag{F.12}$$

Otherwise, if $\epsilon \geq \frac{1}{4}$, $M' \geq T$, so $\sqrt{M'/T} = \Omega(1)$ and regret $r_{\mathtt{ALG}}(T) \geq \Omega(1)$.

Hence, it is proved that regret is lower bounded by $\Omega\big(\sqrt{\frac{M}{T}}\big)$.

$\square$

Case b: $M < 12$. Define N distributions $\mathbb{P}_1, \ldots, \mathbb{P}_N$ as follows. We choose $\epsilon = \sqrt{\frac{1}{45T}}$. The rest of conditional distributions remain same as $\mathbb{P}_0$. For all $i \in [N]$,

$$\mathbb{P}_i(Y = 1|X_i = 1) = 0.5 + \epsilon$$

Now, the optimal arm for action $\mathbb{P}_i$ is $a_{i,1}$, and the regret of pulling a sub-optimal arm in place of the optimal arm $a_{i,1}$ is $(1 - q_i)\epsilon \geq 0.5 \cdot \epsilon$. Each term in the summation of Equation F.9 can be written as

$$\begin{aligned}KL(\mathbb{P}_0^t, \mathbb{P}_i^t) &= \sum_{\mathbf{v}} \mathbb{P}_0(\mathbf{v}) \log\frac{\mathbb{P}_0^t(\mathbf{v})}{\mathbb{P}_i^t(\mathbf{v})}\\ &= \sum_{y} \mathbb{P}_0^t(Y = y|X_i = 1)\mathbb{P}_0^t(X_i = 1) \log\frac{\mathbb{P}_0^t(Y = y|X_i = 1)}{\mathbb{P}_i^t(Y = y|X_i = 1)}\\ &= 0.5\mathbb{P}_0^t(X_i = 1)\left[\log\frac{0.5}{0.5+\epsilon} + \log\frac{0.5}{0.5-\epsilon}\right]\\ &\leq 6\mathbb{P}_0^t(X_i = 1)\epsilon^2\end{aligned}$$

Since $\mathbb{P}_0(X_i = 1|.) \leq 0.5$.

$$KL(\mathbb{D}_0, \mathbb{D}_i) \leq 6\left[\mathbb{E}_0[N_T^{(i,1)}] + \frac{T}{2}\right]\epsilon^2 \tag{F.13}$$

Note that $\mathbb{E}_0[N_T^{(i,1)}] \leq T$

$$KL(\mathbb{D}_0, \mathbb{D}_i) \leq 9T\epsilon^2 \leq 0.2 \tag{F.14}$$

Now putting the value of $KL(\mathbb{D}_0, \mathbb{D}_i)$ in Equation F.8 we get the following,

$$\begin{aligned}\sum_{i\in[N]} \mathbb{D}_i(a_T \neq a_{i,1}) &\geq \frac{1}{2}\sum_{i\in[N]} \exp(-KL(\mathbb{D}_0, \mathbb{D}_i)) - 1\\ &\geq \frac{N}{2e^{0.2}} - 1\,.\end{aligned}$$

Hence any algorithm $\texttt{ALG}$ there exists an i such that the regret incurred by it is

$$r_{\texttt{ALG}}(T) \geq 0.5\mathbb{D}_i(a_T \neq a_{i,1})\epsilon \geq \frac{\frac{N}{2e^{0.2}} - 1}{N}\sqrt{\frac{1}{45T}} \tag{F.15}$$

Finally, from Equations F.12 and F.15 it follows that the expected simple regret of any algorithm is $\Omega\big(\sqrt{\frac{M}{T}}\big)$, where M depends on $\mathbf{q}$ and k_i for $i \in [N]$.

G PROOF OF THEOREM 5.1

Throughout the proof we use a^* to denote the optimal arm. First, we prove a few lemmas, and then use it to bound the expected cumulative regret of $\texttt{CRM-ALG}$. Recall the definitions of E_t, O_t and C_t^x from Section G. In this section we need to keep the context of which X_i the C_t^x corresponds to, therefore, we refer to it as $C_t^{i,x}$ instead. The following lemma shows that the expectation of $\widehat{\mu}_{i,x}$ as defined in Equation 5 is equal to $\mu_{i,x}$ for every i, x.

Lemma G.1
$\widehat{\mu}_{i,x}(t)$ is an unbiased estimator of $\mu_{i,x}$, that is $\mathbb{E}[\widehat{\mu}_{i,x}(t)] = \mu_{i,x}$. Moreover $\mathbb{P}(|\widehat{\mu}_{i,x}(t) - \mu_{i,x}| \geq \epsilon) \leq 2\exp(-2(N_t^{i,x} + C_t^{i,x})\epsilon^2)$.

Proof. We begin by restating the the definition of $\widehat{\mu}_{i,x}$ from Equation 5.

$$\widehat{\mu}_{i,x}(t) = \frac{\sum_{j \in S_t^{i,x}} \mathbb{1}\{Y_j = 1\} + \sum_{c \in [C_t^{i,x}]} Y_c^{i,x}}{N_t^{i,x} + C_t^{i,x}}$$

We note that in Equation 5, $Y_c^{i,x}$ is a random variable such that $\mathbb{E}[Y_c^{i,x}] = \mu_{i,x}$. Note that this holds because we partition the time steps where arm a_0 was pulled into odd and even instances O_t and E_t.Taking expectation on both sides of the above equation we have

$$\begin{aligned}
&\mathbb{E}[\widehat{\mu}_{i,x}(t)] \\
&= \mathbb{E}\Big[\frac{\sum_{j \in S_t^{i,x}} \mathbb{1}\{Y_j = 1\} + \sum_{c \in [C_t^{i,x}]} Y_c^{i,x}}{N_t^{i,x} + C_t^{i,x}}\Big] \\
&= \sum_{a=1}^{\infty}\sum_{b=0}^{\infty} \mathbb{E}\Big[\frac{\sum_{j \in S_t^{i,x}} \mathbb{1}\{Y_j = 1\} + \sum_{c \in [C_t^{i,x}]} Y_c^{i,x}}{N_t^{i,x} + C_t^{i,x}} \;\Big|\; N_t^{i,x} = a, C_t^{i,x} = b\Big]\mathbb{P}(N_t^{i,x} = a, C_t^{i,x} = b) \\
&= \sum_{a=1}^{\infty}\sum_{b=0}^{\infty} \Big(\frac{a\mu_{i,x} + b\mu_{i,x}}{a+b}\Big)\mathbb{P}(N_t^{i,x} = a, C_t^{i,x} = b) \\
&= \mu_{i,x}\sum_{a=1}^{\infty}\sum_{b=0}^{\infty} \mathbb{P}(N_t^{i,x} = a, C_t^{i,x} = b) \\
&= \mu_{i,x}
\end{aligned}$$

Next we prove the concentration inequality part of the lemma, which is similar to Chernoff-Hoeffding inequality (Lemma A.3) for our estimator.

$$
\begin{aligned}
&\mathbb{P}\bigg(\frac{\sum_{j\in S_t^{i,x}}\mathbb{1}\{Y_j=1\}+\sum_{c\in[C_t^{i,x}]}Y_c^{i,x}}{N_t^{i,x}+C_t^{i,x}}\geq \mu_{i,x}+\epsilon\bigg)\\
&=\mathbb{P}\bigg(\sum_{j\in S_t^{i,x}}\mathbb{1}\{Y_j=1\}+\sum_{c\in[C_t^{i,x}]}Y_c^{i,x}\geq (N_t^{i,x}+C_t^{i,x})\mu_{i,x}+(N_t^{i,x}+C_t^{i,x})\epsilon\bigg)\\
&\overset{(i)}{\leq}\min_{\lambda>0}E\bigg[\exp\Big(\lambda\big(\sum_{j\in S_t^{i,x}}(\mathbb{1}\{Y_j=1\}-\mu_{i,x})+\sum_{c\in[C_t^{i,x}]}(Y_c^{i,x}-\mu_{i,x})\big)\Big)\bigg]e^{-\lambda(N_t^{i,x}+C_t^{i,x})\epsilon}\\
&=\min_{\lambda>0}E\bigg[\prod_{j\in S_t^{i,x}}\exp\big(\lambda(\mathbb{1}\{Y_j=1\}-\mu_{i,x})\big)\prod_{c\in[C_t^{i,x}]}\exp\big(\lambda(Y_c^{i,x}-\mu_{i,x})\big)\bigg]e^{-\lambda(N_t^{i,x}+C_t^{i,x})\epsilon}\\
&\overset{(ii)}{=}\min_{\lambda>0}\prod_{j\in S_t^{i,x}}E\Big[\exp\big(\lambda(\mathbb{1}\{Y_j=1\}-\mu_{i,x})\big)\Big]\prod_{c\in[C_t^{i,x}]}E\Big[\exp\big(\lambda(Y_c^{i,x}-\mu_{i,x})\big)\Big]e^{-\lambda(N_t^{i,x}+C_t^{i,x})\epsilon}\\
&\overset{(iii)}{\leq}\min_{\lambda>0}\exp\bigg(\frac{N_t^{i,x}\lambda^2}{8}+\frac{C_t^{i,x}\lambda^2}{8}-\lambda(N_t^{i,x}+C_t^{i,x})\epsilon\bigg)\\
&\leq\exp\left(-2(N_t^{i,x}+C_t^{i,x})\epsilon^2\right)
\end{aligned}\tag{G.16}
$$

In the above equations, the inequality in (i) follows from Lemma A.1, the equality in (ii) follows from the fact that each term in the product are independent, and (iii) follows from Lemma A.2. Following the same steps as above we get the following two sided bound

$$
\mathbb{P}(|\widehat{\mu}_{i,x}(t)-\mu_{i,x}|\geq\epsilon)\leq 2\exp\left(-2(N_t^{i,x}+C_t^{i,x})\epsilon^2\right).\tag{G.17}
$$

□

Next we show that the estimates of μ_a at the end of T rounds is good with high probability.

Lemma G.2
Let $p=\min_{i,x,\mathbf{z}}\mathbb{P}(X_i=x,\mathbf{Pa}(X_i)=\mathbf{z})$. Then for sufficiently large $T\in\mathbb{N}$, at the end of T rounds the following hold:

1. $\mathbb{P}\Big(|\widehat{\mu}_0(T)-\mu_0|\geq\frac{\Delta_0}{4}\Big)\leq 2T^{-\frac{\Delta_0^2}{8}}$.
2. *Let $\widehat{p}_{\mathbf{z},T}^{\,i,x}=\frac{1}{|O_T|}\sum_{t\in O_T}\mathbb{1}\{X_i(t)=x,\mathbf{Pa}(X_i)(t)=\mathbf{z}\}$, and $\widehat{p}_T^{\,i,x}=\min_{\mathbf{z}}\widehat{p}_{\mathbf{z},T}^{\,i,x}$. Then $\mathbb{P}\big(\widehat{p}_T^{\,i,x}\geq\frac{p}{2}\big)\geq 1-Z_iT^{-\frac{p^2}{4}}$, where Z_i is the size of the domain from which $\mathbf{Pa}(X_i)$ takes values.*
3. $\mathbb{P}\Big(|\widehat{\mu}_{i,x}(T)-\mu_{i,x}|\geq\frac{\Delta_0}{4}\Big)\leq 2T^{-\frac{p\Delta_0^2}{32}}+Z_iT^{-\frac{p^2}{4}}$.

Proof. a) Since $\beta\geq 1$, at the end of T rounds arm a_0 is pulled by Algorithm 2 at least $(\ln T)$ times. Hence, $N_T^0\geq(\ln T)$, and by A.3,

$$
\mathbb{P}\Big(|\widehat{\mu}_0(T)-\mu_0|\geq\frac{\Delta_0}{4}\Big)\leq 2e^{-\frac{\Delta_0^2}{8}\ln T}=2T^{-\frac{\Delta_0^2}{8}}
$$

b) In this part we show, using union bound, that the estimation of $\widehat{p}_T^{\,i,x}$ being less that $p/2$ have low probability. Since, $|O_T|\geq N_T^0/2$, by Lemma A.3, we have,

$$
\mathbb{P}\bigg(\widehat{p}_{\mathbf{z},T}^{\,i,x}>p_{\mathbf{z}}^{i,x}-\frac{p}{2}\geq\frac{p}{2}\bigg)\geq 1-e^{-2\frac{p^2}{4}\frac{\ln T}{2}}=1-T^{-\frac{p^2}{4}}
$$

Now using this we get,

$$
\mathbb{P}\bigg(\widehat{p}_T^{\,i,x}\leq\frac{p}{2}\bigg)=\mathbb{P}\bigg(\min_{\mathbf{z}}\widehat{p}_{\mathbf{z},T}^{\,i,x}\leq\frac{p}{2}\bigg)\leq\sum_{\mathbf{z}}\mathbb{P}\bigg(\widehat{p}_{\mathbf{z},T}^{\,i,x}\leq\frac{p}{2}\bigg)\leq Z_iT^{-\frac{p^2}{4}}\tag{G.18}
$$

c) Let the conditional probability distribution $\mathbb{P}(. \mid \widehat{p}_T^{\,i,x} > \frac{p}{2})$ be denoted by $\mathbb{P}_p$. Since $\beta \geq 1$, $N_T^0 \geq \ln T$. Further if $\widehat{p}_T^{\,i,x} > \frac{p}{2}$ then $C_T^{i,x} > \frac{p}{2}\frac{N_T^0}{2} \geq \frac{p}{4}\ln T$ (from the definition of $C_T^{i,x}$). Hence, from Lemma G.1 we have

$$\mathbb{P}_p\bigg(|\widehat{\mu}_{i,x}(T) - \mu_{i,x}| \geq \frac{\Delta_0}{4}\bigg) \leq 2\exp\bigg(-\frac{\Delta_0^2}{32}p\ln T\bigg) = 2T^{-\frac{p\Delta_0^2}{32}} \tag{G.19}$$

Finally by the law of total probability and using Equations G.18 and G.19

$$\begin{aligned}\mathbb{P}\bigg(|\widehat{\mu}_{i,x}(T) - \mu_{i,x}| \geq \frac{\Delta_0}{4}\bigg) &\leq \mathbb{P}_p\bigg(|\widehat{\mu}_{i,x}(T) - \mu_{i,x}| \geq \frac{\Delta_0}{4}\bigg) + \mathbb{P}\bigg(\widehat{p}_T^{\,i,x} \leq \frac{p}{2}\bigg) \\ &\leq 2T^{-\frac{p\Delta_0^2}{32}} + Z_i T^{-\frac{p^2}{4}}\end{aligned}$$

□

Next we show that β as set in `CRM-ALG` is bounded in expectation. Lemma G.3 and its proof is similar to Lemma 8.6 in Nair et al. [2021].

Lemma G.3
Let $L = \arg\min_{t\in\mathbb{N}}\left\{\frac{t^{\frac{p^2\Delta_0^2}{32}}}{\ln t} \geq 3N(Z+3)\right\}$, where $Z = \max_i Z_i$, and suppose `CRM-ALG` pulls arms for T rounds, where $T \geq \max(L, e^{\frac{50}{\Delta_0^2}})$, and let $a^* \neq a_0$. Then at the end of T rounds, $\frac{8}{9\Delta_0^2} \leq \mathbb{E}[\beta^2] \leq \frac{50}{\Delta_0^2}$.

Proof. Before proceeding to the proof of the lemma we make the following two observations.

Observation G.4
1. If $a^* \neq a_0$ then $\Delta_0 = \mu_{a^*} - \mu_0$
2. Let $\widehat{\mu}^* = \max_{i,x}(\widehat{\mu}_{i,x}(T))$. If $|\widehat{\mu}_0(T) - \mu_0| \leq \frac{\Delta_0}{4}$ and $|\widehat{\mu}_{i,x}(T) - \mu_{i,x}| \leq \frac{\Delta_0}{4}$ for all (i,x) then $\frac{\Delta_0}{2} \leq \widehat{\mu}_{a^*} - \widehat{\mu}_0(T) \leq \frac{3\Delta_0}{2}$, and $\frac{32}{9\Delta_0^2} \leq \beta^2 \leq \frac{32}{\Delta_0^2}$. Notice that since $T \geq e^{\frac{50}{\Delta_0^2}}$, $\frac{32}{\Delta_0^2} \leq \ln T$.

Let U_0 be the event that $|\widehat{\mu}_0(T) - \mu_0| \leq \frac{\Delta_0}{4}$, and for any i, x let $U_{i,x}$ be the event $|\widehat{\mu}_{i,x}(T) - \mu_{i,x}| \leq \frac{\Delta_0}{4}$. Also let $U = (\cap_{i,x} U_{i,x}) \cap U_0$. If $\overline{U}_0$, $\overline{U}_{i,x}$, and $\overline{U}$ denote the compliment of the events $U_0, U_{i,x}$, and U respectively, then

$$\mathbb{P}(\overline{U}_0) \leq 2T^{-\frac{\Delta_0^2}{8}}\text{ , and}$$

$$\text{for a fixed } (i,x) \quad \mathbb{P}(\overline{U}_{i,x}) \leq 2T^{-\frac{p\Delta_0^2}{32}} + Z_i T^{-\frac{p^2}{4}}\ .$$

Hence applying union bound,

$$\begin{aligned}\mathbb{P}(\overline{U}) &\leq 2N\left(\frac{2}{T^{\frac{p\Delta_0^2}{32}}} + \frac{Z}{T^{\frac{p^2}{4}}}\right) + \frac{2}{T^{\frac{\Delta_0^2}{8}}} \\ &\leq 2N\left(\frac{2}{T^{\frac{p^2\Delta_0^2}{32}}} + \frac{Z}{T^{\frac{p^2\Delta_0^2}{32}}}\right) + \frac{2N}{T^{\frac{p^2\Delta_0^2}{32}}} \qquad \text{as } p \leq 1, \Delta_0 \leq 1 \\ &\leq \frac{2N(Z+3)}{T^{\frac{p^2\Delta_0^2}{32}}} = \delta\end{aligned}$$

We will use the above arguments to first show that $\mathbb{E}[\beta^2] \geq \frac{8}{9\Delta_0^2}$. From part 2 of Observation we have that the event U implies $\beta^2 \geq \frac{32}{9\Delta_0^2}$. Since $\mathbb{P}\{U\} \geq 1 - \delta$,

$$\mathbb{E}[\beta^2] \geq \frac{32}{9\Delta_0^2}(1-\delta) = \frac{32}{9\Delta_0^2} - \frac{32\delta}{9\Delta_0^2}$$

Since T satisfies $\frac{T^{\frac{p^2\Delta_0^2}{32}}}{\ln T} \geq 3N(Z+3)$, this implies $\frac{32\delta}{9\Delta_0^2} \leq \frac{24}{9\Delta_0^2}$, and hence $\mathbb{E}[\beta^2] \geq \frac{8}{9\Delta_0^2}$. Similarly, from part 2 of Observation we have that the event U implies $\beta^2 \leq \frac{32}{\Delta_0^2}$. If U does not hold then $\beta^2 \leq \ln T$. Hence using the fact that T satisfies $\frac{T^{\frac{p^2\Delta_0^2}{32}}}{\ln T} \geq 3N(Z+3)$, and hence $\delta \ln T \leq \frac{18}{\Delta_0^2}$, we get,

$$\mathbb{E}[\beta^2] \leq \frac{32}{\Delta_0^2}(1-\delta) + \delta \ln T \leq \frac{32}{\Delta_0^2} + \delta \ln T \leq \frac{50}{\Delta_0^2} \, .$$

$\square$

Lemma G.5
Suppose $a^* \neq a_{i,x}$. Then at the end of T rounds the following holds:

$$\mathbb{E}[N_T^{i,x}] \leq \max\left(0, \frac{8\ln T}{\Delta_{i,x}^2} + 1 - \frac{1}{4} \cdot p_{i,x} \cdot \eta_T^{i,x} \cdot \mathbb{E}[N_T^0]\right) + \frac{\pi^2}{3} \, .$$

Further if $a^* \neq a_0$ then

$$\mathbb{E}[N_T^0] \leq \left(\mathbb{E}[\beta^2]\ln T + \frac{8\ln T}{\Delta_0^2} + 1\right) + \frac{\pi^2}{3} \, .$$

Proof. Let $F_T^{i,x} = N_T^{i,x} + C_T^{i,x}$. Then,

$$N_T^{i,x} = \sum_{t \in T} \mathbb{1}\{a_t = a_{i,x}\} \, . \tag{G.20}$$

$$N_T^{i,x} \leq \max(0, \ell - C_T^{i,x}) + \sum_{t\in T} \mathbb{1}\{a_t = a_{i,x}, F_t^{i,x} \geq \ell\} \tag{G.21}$$

Here, we make an observation regarding the expected value of $C_T^{i,x}$.

Observation G.6
$\mathbb{E}[C_T^{i,x}] = \mathbb{E}[\min_{\mathbf{z}} \widehat{p}_{\mathbf{z},T}^{i,x} \lceil N_T^0/2 \rceil] \geq \frac{1}{4} \cdot p_{i,x} \cdot \mathbb{E}[N_T^0] \cdot (1 - Z_i T^{-\frac{p_{i,x}^2}{2}}) = \frac{1}{4} \cdot p_{i,x} \cdot \eta_T^{i,x} \cdot \mathbb{E}[N_T^0]$

Proof. Note that the expectation of $\min_{\mathbf{z}} \widehat{p}_{\mathbf{z},T}^{i,x}$ is over the distribution of the CBN and that of N_T^0 over the distribution in the observation across all T rounds. Recall $p_{i,x} = \min_{\mathbf{z}} p_{\mathbf{z}}^{i,x}$. By Lemma A.3, we have,

$$\mathbb{P}\left(\widehat{p}_{\mathbf{z},T}^{i,x} > p_{\mathbf{z}}^{i,x} - \frac{p_{i,x}}{2} \geq \frac{p_{i,x}}{2}\right) \geq 1 - e^{-2\frac{p_{i,x}^2}{4}\frac{\ln T}{2}} = 1 - T^{-\frac{p_{i,x}^2}{4}}$$

Now using this we get,

$$\mathbb{P}\left(\widehat{p}_T^{i,x} \leq \frac{p_{i,x}}{2}\right) \leq \mathbb{P}\left(\min_{\mathbf{z}} \widehat{p}_{\mathbf{z},T}^{i,x} \leq \frac{p_{i,x}}{2}\right) \leq \sum_{\mathbf{z}} \mathbb{P}\left(\widehat{p}_{\mathbf{z},T}^{i,x} \leq \frac{p_{i,x}}{2}\right) \leq Z_i T^{-\frac{p_{i,x}^2}{4}}$$

We can now bound the expectation of $C_T^{i,x}$ for sufficiently large T as follows:

$$\begin{aligned}
\mathbb{E}[\min_{\mathbf{z}} \widehat{p}^{i,x}_{\mathbf{z},T} \lceil N_T^0/2 \rceil] &\geq \frac{1}{2}\mathbb{E}[\min_{\mathbf{z}} \widehat{p}^{i,x}_{\mathbf{z},T} N_T^0] \\
&= \frac{1}{2}\sum_{a=1}^{\infty} a \cdot \mathbb{E}[\min_{\mathbf{z}} \widehat{p}^{i,x}_{\mathbf{z},T} \mid N_T^0 = a]\mathbb{P}(N_T^0 = a) \\
&\geq \frac{1}{2}\sum_{a=1}^{\infty} a \cdot \frac{p_{i,x}}{2} \cdot \mathbb{P}\bigg(\min_{\mathbf{z}} \widehat{p}^{i,x}_{\mathbf{z},T} > \frac{p_{i,x}}{2} \mid N_T^0 = a \bigg)\mathbb{P}(N_T^0 = a) \\
&\geq \frac{1}{2}\sum_{a=1}^{\infty} a \cdot \frac{p_{i,x}}{2} \cdot \mathbb{P}\bigg(\min_{\mathbf{z}} \widehat{p}^{i,x}_{\mathbf{z},T} > \frac{p_{i,x}}{2} \mid N_T^0 = a \bigg)\mathbb{P}(N_T^0 = a) \\
&\geq \frac{p_{i,x}}{4}\mathbb{E}[N_T^0] \cdot \max\Big\{0, 1 - Z_i T^{-\frac{p_{i,x}^2}{4}}\Big\} \\
&= \frac{1}{4} \cdot p_{i,x} \cdot \eta_T^{i,x} \cdot \mathbb{E}[N_T^0] \qquad \text{(G.22)}
\end{aligned}$$

□

Taking expectation of Equation G.21, we get

$$\mathbb{E}[N_T^{i,x}] \leq \max\big\{0, \ell - \frac{p_{i,x}}{4} \cdot \eta_T^{i,x} \cdot \mathbb{E}[N_T^0]\big\} + \sum_{t\in[\ell+1,T]} \mathbb{P}(a_t = a_{i,x}, F_t^{i,x} \geq \ell) \qquad \text{(G.23)}$$

Now we bound $\sum_{t\in[l+1,T]} \mathbb{P}(a(t) = a_{i,x}, F_t^{i,x} \geq \ell)$, and assuming $a^* \neq a_0$. The proof for $a^* = a_0$ is similar. We use $F_T^{a^*}$ to denote the effective number of pulls of a^* at the end of T rounds. Also, for better clarity, we use $\widehat{\mu}_{i,x}(F_T^{i,x}, T)$ (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 Algorithm 2 at the end of T rounds.

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

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

$$\widehat{\mu}_{a^*}(s,t) \leq \mu_{a^*} - \sqrt{\frac{2\ln t}{s}}\ , \qquad \text{(G.24a)}$$

$$\widehat{\mu}_{i,x}(s_j,t) \geq \mu_{i,x} + \sqrt{\frac{2\ln t}{s_j}}\ , \qquad \text{(G.24b)}$$

$$\mu_{a^*} \leq \mu_{i,x} + 2\sqrt{\frac{2\ln t}{s_j}}\ . \qquad \text{(G.24c)}$$

The probability of the events in Equations G.24a and G.24b can be bounded using Chernoff-Hoeffding inequality

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

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

Also if $\ell \ge \lceil \frac{8\ln T}{\Delta_{i,x}^2} \rceil$ then the event in Equation G.24c is false, i.e. $\mu_{a^*} > \mu_{i,x} + 2\sqrt{\frac{2\ln t}{s_j}}$. Thus for $\ell = \frac{8\ln T}{\Delta_{i,x}^2} + 1 \ge \lceil \frac{8\ln T}{\Delta_{i,x}^2} \rceil$, which implies

$$\sum_{t\in[\ell+1,T]} \mathbb{P}\{a(t) = a_{i,x}, F_t^{i,x} \ge \ell\} \le \sum_{t\in[T]} \sum_{s\in[0,t-1]} \sum_{s_j\in[\ell-1,t]} 2t^{-4} \le \frac{\pi^2}{3} \tag{G.25}$$

If $a^* = a_0$ then using the exact arguments as above we can show that Equation G.25 still holds. Hence, using Equations G.23 and G.25 we have if $a^* \ne a_{i,x}$ then

$$\mathbb{E}[N_T^{i,x}] \le \max\left\{0, \frac{8\ln T}{\Delta_{i,x}^2} + 1 - \frac{p_{i,x}}{4} \cdot \eta_T^{i,x} \cdot \mathbb{E}[N_T^0]\right\} + \frac{\pi^2}{3} .$$

The arguments used to bound $\mathbb{E}[N_T^0]$, when $a^* \ne a_0$ is similar. In this case the equation corresponding to Equation G.23 is

$$\mathbb{E}[N_T^0] \le \mathbb{E}[\beta^2]\ln T + \ell + \sum_{t\in[\ell+1,T]} \mathbb{P}\{a(t) = a_0, N_t^0 \ge \ell\} . \tag{G.26}$$

Also the same arguments as above can be used to show that for $\ell = \frac{8\ln T}{\Delta_0^2} + 1$,

$$\sum_{t\in T} \mathbb{P}\{a(t) = a_0, N_t^0 \ge \ell\} \le \frac{\pi^2}{3} . \tag{G.27}$$

Finally using Equations G.26 and G.27, we have

$$\mathbb{E}[N_T^0] \le \left(\mathbb{E}[\beta^2]\ln T + \frac{8\ln T}{\Delta_0^2} + 1\right) + \frac{\pi^2}{3} .$$

Lemma G.7
If $a^* = a_0$ then at the end of T rounds the following is true:

$$\mathbb{E}[N_T^0] \ge T - \left(2N(1 + \frac{\pi^2}{3}) + \sum_{i,x} \frac{8\ln T}{\Delta_{i,x}^2}\right) .$$

Proof. 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,

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

Now we use Lemma G.5 to conclude that

$$\mathbb{E}[N_T^0] \ge T - \left(2N(1 + \frac{\pi^2}{3}) + \sum_{i,x} \frac{8\ln T}{\Delta_{i,x}^2}\right) .$$

$\square$

Now that we have bounds on $\mathbb{E}[N_T^0]$ and $\mathbb{E}[N_T^{i,x}]$, we can bound the regret as follows.

Case a ($a^* = a_0$): In this case we bound the expected cumulative regret of Algorithm 2. From Lemma G.5 and G.7 for any T satisfying both $T^{-\frac{p_{i,x}^2}{4}} > Z_i$ and

$$T \geq \frac{4}{p_{i,x} \cdot \eta_T^{i,x}}\left(1 + \frac{8 \ln T}{\Delta_{i,x}^2}\right) + \left(2N(1 + \frac{\pi^2}{3}) + \sum_{i,x} \frac{8 \ln T}{\Delta_{i,x}^2}\right) \tag{G.28}$$

we have $\mathbb{E}[N_T^{i,x}] \leq \frac{\pi^2}{3}$. Notice that Equation G.28 holds for sufficiently large T. Hence the cumulative regret caused by pulling sub-optimal arms $a_{i,x}$ is

$$\mathbb{E}[R(T)] \leq \sum_{\Delta_a > 0} \Delta_a \frac{\pi^2}{3} \tag{G.29}$$

Case b ($a^* \neq a_0$): In this case we bound the regret of pulling sub-optimal arms when $T \geq \max(L, e^{\frac{50}{\Delta_0^2}})$, where L is as defined in Lemma G.3. Note that this is satisfied for sufficiently large T. Hence from Lemma G.3 and Lemma G.5, we have for $a^* \neq a_{i,x}$ and for a_0

$$\mathbb{E}[N_T^{i,x}] \leq \max\left\{0, 1 + 8 \ln T\left(\frac{1}{\Delta_{i,x}^2} - \frac{p_{i,x} \cdot \eta_T^{i,x}}{36\Delta_0^2}\right)\right\} + \frac{\pi^2}{3} \tag{G.30}$$

$$\mathbb{E}[N_T^0] \leq \frac{58 \ln T}{\Delta_0^2} + 1 + \frac{\pi^2}{3} \tag{G.31}$$

Hence the cumulative regret can be written as

$$\mathbb{E}[R(T)] \leq \Delta_0\left(\frac{58 \ln T}{\Delta_0^2} + 1 + \frac{\pi^2}{3}\right) + \sum_{\Delta_{i,x} > 0} \Delta_{i,x}\left(\max\left\{0, 1 + 8 \ln T\left(\frac{1}{\Delta_{i,x}^2} - \frac{p_{i,x} \cdot \eta_T^{i,x}}{36\Delta_0^2}\right\}\right) + \frac{\pi^2}{3}\right) \tag{G.32}$$

□

H REMARKS ON EXPERIMENT INVOLVING ALGORITHM FROM Yabe et al. [2018]

We mention few issues faced while implementing `PROP-INF` using the details from Yabe et al. [2018] and how we resolved them: (a) In Step (3) of Algorithm 1 in Yabe et al. [2018] (which is a subroutine for `PROP-INF`), they iterate over all possible assignments to the parents of each node. Specifically, the algorithm would be exponential time in the in-degree of the reward node Y and therefore it runs efficiently only when Y has a small number of parents. `SRM-ALG` does not face this issue. To compare both algorithms we therefore created instances where in-degree of Y was small. (b) Another issue faced while implementing their algorithm is in an inequality condition specified in Equation 4 of Yabe et al. [2018]. We observe that this inequality is trivially satisfied unless the time period becomes very large (of the order of $\geq 10^{10}$) even for their experiments given in Section 5 of Yabe et al. [2018]. Since running the algorithms for such a long time period is not feasible, we run both algorithms till we see clear convergence of `SRM-ALG`. (c) A third problem we faced was in setting the time period range for our Experiments. They use $T \in \{C, 2C, \ldots, 9C\}$, but in Step 3 of Algorithm 1 and Step 4 of Algorithm 2 in Yabe et al. [2018], they estimate probabilities using $T/3C$ samples. This would leave them with at most 3 samples for such an estimation which would give noisy and unreliable estimates. Instead of using this set of values for T, we use equally spaced points in a time range where we see clear convergence of `SRM-ALG` (d) Finally, it is not discussed how the optimization problem giving $\widehat{\eta}$ in Step 12 of Algorithm 2 of Yabe et al. [2018] is solved, and they use a fixed value for $\widehat{\eta}$ in experiments. Since there is no technique proposed to solve the optimization problem, we use the same fixed $\widehat{\eta}$ as them.

References

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, 1995. doi: 10.1109/SFCS.1995.492488.

Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, Ashwin Maran, and Vinodchandran N. Variyam. Learning and sampling of atomic interventions from observations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 842–853. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/bhattacharyya20a.html`.

Herman Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *The Annals of Mathematical Statistics*, 23(4):493 – 507, 1952. doi: 10.1214/aoms/1177729330. URL `https://doi.org/10.1214/aoms/1177729330`.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459. URL `http://www.jstor.org/stable/2282952`.

P. Massart and J. Picard. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2007. ISBN 9783540485032. URL `https://books.google.co.in/books?id=ZI67BQAAQBAJ`.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, USA, 2005. ISBN 0521835402.

Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2017–2025. PMLR, 2021.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *CoRR*, abs/1904.07272, 2019. URL `http://arxiv.org/abs/1904.07272`.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. Causal bandits with propagating inference. In *International Conference on Machine Learning, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5508–5516. PMLR, 2018.