Zhijie Zhang, Wei Chen, Xiaoming Sun, and Jialin Zhang. Online influence maximization with node-level feedback using standard offline oracles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9153–9161, 2022.

Appendix

Appendix A. Conversion of Function f_X

A.1. Conversion to g_X such that $\lim_{x\to+\infty} g_X(x) = +\infty$

In this section, we firstly prove that any monotone increasing function f_X that satisfies Assumptions 1 and 2 can be converted to a function g_X such that the conversion does not impact the propagation of BGLM, i.e., $f_X(x) = g_X(x)$ for $x \in [0, |\boldsymbol{Pa}(X)|]$, $\lim_{x\to+\infty} g_X(x) = +\infty$, g_X is twice differentiable and Assumptions 1 and 2 still hold.

On one hand, if for all $x \geq 2|\boldsymbol{Pa}(X)|$, $f''_X(x) \geq 0$, then $f_X(x) \geq f_X(2|\boldsymbol{Pa}(X)|) + f'_X(2|\boldsymbol{Pa}(X)|)(x - 2|\boldsymbol{Pa}(X)|)$, which already satisfies $\lim_{x\to+\infty} f_X(x) = +\infty$. In this case, no conversion is needed (let $g_X \equiv f_X$). On another hand, we can find a $x^* \geq 2|\boldsymbol{Pa}(X)|$ such that $f''_X(x^*) < 0$.

We define the conversion as

$$g_X(x) = \begin{cases} f_X(x) & x \leq x^* \\ f_X(x^*) + \frac{f'_X(x^*)^2}{f''_X(x^*)} \ln\left(-\frac{f'_X(x^*)}{f''_X(x^*)}\right) - \frac{f'_X(x^*)^2}{f''_X(x^*)} \ln\left(x - x^* - \frac{f'_X(x^*)}{f''_X(x^*)}\right) & x > x^* \end{cases}.$$

During the propagation of the BGLM, the input of f_X is $\boldsymbol{Pa}(X) \cdot \boldsymbol{\theta}^*_X$, which is in the range $[0, |\boldsymbol{Pa}(X)|] \subseteq [0, x^*]$. Hence, when we replace f_X by g_X in the BGLM, the propagation is not impacted.

Moreover, we can compute that

$$g'_X(x) = \begin{cases} f'_X(x) & x \leq x^* \\ -\frac{f'_X(x^*)^2}{f''_X(x^*)\left(x - x^* - \frac{f'_X(x^*)}{f''_X(x^*)}\right)} & x > x^* \end{cases},$$

and

$$g''_X(x) = \begin{cases} f''_X(x) & x \leq x^* \\ \frac{f'_X(x^*)^2}{f''_X(x^*)\left(x - x^* - \frac{f'_X(x^*)}{f''_X(x^*)}\right)^2} & x > x^* \end{cases}.$$

Therefore, we have $\lim_{x\to x^{*+}} g_X(x) = f_X(x^*)$ and $\lim_{x\to x^{*-}} g_X(x) = f_X(x^*)$. Hence, g_X is continuous. Moreover, $\lim_{x\to x^{*+}} g'_X(x) = f'_X(x^*) = \lim_{x\to x^{*-}} g'_X(x)$ and $\lim_{x\to x^{*+}} g''_X(x) = f''_X(x^*) = \lim_{x\to x^{*-}} g''_X(x)$, so $g_X(x)$ is twice differentiable and g''_X is continuous.

Now we only need to verify Assumptions 1 and 2. Firstly, when $x > x^*$, we have $g'_X(x) < g'_X(x^*) = f'_X(x^*) \leq L^{(1)}_{f_X}$ and $g''_X(x) < g''_X(x^*) = f''_X(x^*) \leq L^{(2)}_{f_X}$, so Assumption 1 holds. Secondly, $\max_{\boldsymbol{v}\in[0,1]^{|\boldsymbol{Pa}(X)|}, \|\boldsymbol{\theta}-\boldsymbol{\theta}^*_X\|\leq 1} \boldsymbol{v} \cdot \boldsymbol{\theta} \leq 2|\boldsymbol{Pa}(X)| \leq x^*$, so the conversion does not impact the value of κ. Until now, we complete the conversion.

A.2. Conversion to h_X such that $\lim_{x\to-\infty} h_X(x) = -\infty$ and $\lim_{x\to+\infty} h_X(x) = +\infty$

Then we prove that the monotone increasing function g_X that satisfies Assumptions 1 and 2 can be converted to a function h_X such that the conversion does not impact the propagation of BGLM, i.e., $g_X(x) = h_X(x)$ for $x \in [0, |\boldsymbol{Pa}(X)|]$, $\lim_{x\to-\infty} h_X(x) = -\infty$, $\lim_{x\to+\infty} h_X(x) = +\infty$, h_X is twice differentiable and Assumptions 1 and 2 still hold.

On one hand, if for all $x \leq -|\boldsymbol{Pa}(X)|$, $f''_X(x) \leq 0$, then $f_X(x) \leq f_X(-|\boldsymbol{Pa}(X)|) - f'_X(-|\boldsymbol{Pa}(X)|)(-x - |\boldsymbol{Pa}(X)|)$, which already satisfies $\lim_{x \to -\infty} f_X(x) = -\infty$. In this case, no conversion is needed (let $h_X \equiv g_X$). On another hand, we can find a $x^* \leq -|\boldsymbol{Pa}(X)|$ such that $f''_X(x^*) > 0$.

We define the conversion as

$$h_X(x) = \begin{cases} g_X(x) & x \geq x^* \\ g_X(x^*) - \frac{g'_X(x^*)^2}{g''_X(x^*)} \ln\left(-\frac{g'_X(x^*)}{g''_X(x^*)}\right) + \frac{g'_X(x^*)^2}{g''_X(x^*)} \ln\left(-x + x^* + \frac{g'_X(x^*)}{g''_X(x^*)}\right) & x < x^* \end{cases}.$$

During the propagation of the BGLM, the input of g_X is $\boldsymbol{Pa}(X) \cdot \boldsymbol{\theta}^*_X$, which is in the range $[0, |\boldsymbol{Pa}(X)|] \subseteq [0, x^*]$. Hence, when we replace g_X by h_X in the BGLM, the propagation is not impacted.

Moreover, we can compute that

$$h'_X(x) = \begin{cases} g'_X(x) & x \geq x^* \\ \frac{g'_X(x^*)^2}{g''_X(x^*)\left(-x+x^*+\frac{g'_X(x^*)}{g''_X(x^*)}\right)} & x < x^* \end{cases},$$

and

$$h''_X(x) = \begin{cases} g''_X(x) & x \geq x^* \\ \frac{g'_X(x^*)^2}{g''_X(x^*)\left(x-x^*-\frac{g'_X(x^*)}{g''_X(x^*)}\right)^2} & x < x^* \end{cases}.$$

Therefore, we have $\lim_{x \to x^{*+}} h_X(x) = g_X(x^*)$ and $\lim_{x \to x^{*-}} h_X(x) = g_X(x^*)$. Hence, h_X is continuous. Moreover, $\lim_{x \to x^{*+}} h'_X(x) = g'_X(x^*) = \lim_{x \to x^{*-}} h'_X(x)$ and $\lim_{x \to x^{*+}} h''_X(x) = g''_X(x^*) = \lim_{x \to x^{*-}} h''_X(x)$, so $h_X(x)$ is twice differentiable and h''_X is continuous.

Now we only need to verify Assumptions 1 and 2. Firstly, when $x < x^*$, we have $h'_X(x) < h'_X(x^*) = g'_X(x^*) \leq L^{(1)}_{f_X}$ and $h''_X(x) < h''_X(x^*) = g''_X(x^*) \leq L^{(2)}_{f_X}$, so Assumption 1 holds. Secondly, $\min_{\boldsymbol{v} \in [0,1]^{|\boldsymbol{Pa}(X)|}, \|\boldsymbol{\theta} - \boldsymbol{\theta}^*_X\| \leq 1} \boldsymbol{v} \cdot \boldsymbol{\theta} \geq -|\boldsymbol{Pa}(X)| \geq x^*$, so the conversion does not impact the value of κ. Until now, we complete the conversion.

In conclusion we have found a conversion from f_X to h_X such that the conversion does not impact the propagation of BGLM, i.e., $h_X(x) = f_X(x)$ for $x \in [0, |\boldsymbol{Pa}(X)|]$, $\text{Range}(h_X) = \mathbb{R}$, h_X is twice differentiable and Assumptions 1 and 2 still hold.

Appendix B. Pseudocode of Algorithm 5

Here, we want to give a lemma to clarify why we can always find a solution for equation $\sum_{i=1}^t (X^{(i)} - f_X(\boldsymbol{V}_{i,X}^\intercal \boldsymbol{\theta}_X))\boldsymbol{V}_{i,X} = 0$ in Line 5 of Algorithm 5.

Lemma 10 *When $\lim_{x \to +\infty} f_X(x) = +\infty$, $\lim_{x \to -\infty} f_X(x) = -\infty$, and f_X is monotone increasing, equation $\sum_{i=1}^t (X^{(i)} - f_X(\boldsymbol{V}_{i,X}^\intercal \boldsymbol{\theta}_X))\boldsymbol{V}_{i,X} = 0$ has a solution.*

Proof We define $m_X(x)$ as

$$m_X(x) = \begin{cases} \int_0^x f_X(c)\mathrm{d}c & x \geq 0 \\ -\int_x^0 f_X(c)\mathrm{d}c & x < 0 \end{cases}.$$

Algorithm 5: BGLM-Estimate

```
Input: All observations $((\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_t, Y_t))$ until round $t$.
Output: $\{\hat{\boldsymbol{\theta}}_{t,X}, M_{t,X}\}_{X \in \boldsymbol{X} \cup \{Y\}}$
For each $X \in \boldsymbol{X} \cup \{Y\}$, $i \in [t]$, construct data pair $(\boldsymbol{V}_{i,X}, X^{(i)})$ with $\boldsymbol{V}_{i,X}$ the vector of ancestors of $X$ in round $i$, and $X^{(i)}$ the value of $X$ in round $i$ if $X \notin S_i$.
for $X \in \boldsymbol{X} \cup \{Y\}$ do
    Calculate the maximum-likelihood estimator $\hat{\boldsymbol{\theta}}_{t,X}$ by solving the equation $\sum_{i=1}^{t}(X^{(i)} - f_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X))\boldsymbol{V}_{i,X} = 0$.
    $M_{t,X} = \sum_{i=1}^{t} \boldsymbol{V}_{i,X}\boldsymbol{V}_{i,X}^{\intercal}$.
end for
```

Then we can compute

$$\sum_{i=1}^{t}(X^{(i)} - f_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X))\boldsymbol{V}_{i,X}$$

as

$$\nabla_{\boldsymbol{\theta}} \sum_{i=1}^{t} \left(X^{(i)}\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X - m_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X) \right).$$

Hence, we only need to prove that

$$H_X(\boldsymbol{\theta}_X) \triangleq \sum_{i=1}^{t} \left(X^{(i)}\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X - m_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X) \right)$$

is a concave function with respect to $\boldsymbol{\theta}_X$ and $\lim_{(\boldsymbol{\theta}_X)_j \to \infty} H_X(\boldsymbol{\theta}_X) = -\infty$ or $\frac{\partial H_X(\boldsymbol{\theta}_X)}{\partial(\boldsymbol{\theta}_X)_j} \equiv 0$ for all $j \in [|\boldsymbol{Pa}(X)|]$, which implies that H_X has a maximal point. Firstly, we know that

$$\frac{\partial^2 m_X(x)}{\partial x^2} = f_X'(x) > 0,$$

so m_X is a convex function. Therefore, for any vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^{|\boldsymbol{Pa}(X)|}$ and $\lambda \in [0,1]$, we have

$$\begin{aligned} m_X\left(\boldsymbol{V}_{i,X}^{\intercal}(\lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2)\right) &= m_X\left(\lambda\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_2)\right) \\ &\leq \lambda m_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_1) + (1-\lambda)m_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_2), \end{aligned}$$

so $m_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X)$ is also a convex function with respect to $\boldsymbol{\theta}_X$ and the Hessian matrix $\mathbf{H}[m_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X)]$ of $m_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X)$ with respect to $\boldsymbol{\theta}_X$ should be positive semidefinite. Now we can compute the Hessian matrix $\mathbf{H}[H_X(\boldsymbol{\theta}_X)]$ as

$$\mathbf{H}[H_X(\boldsymbol{\theta}_X)] = \sum_{i=1}^{t} \left(-\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{V}_{i,X} \cdot \mathbf{H}[m_X(\boldsymbol{V}_{i,X}^{\intercal}\boldsymbol{\theta}_X)] \right).$$

Hence, $\mathbf{H}[H_X(\boldsymbol{\theta}_X)]$ is negative semidefinite because multiplying a positive semidefinite matrix by a negative scalar preserves the semidefiniteness. Thus H_X is a concave function with respect to $\boldsymbol{\theta}_X$.

Now for any $j \in [\boldsymbol{Pa}(X)]$, we prove that $\lim_{(\boldsymbol{\theta}_X)_j \to +\infty} H_X(\boldsymbol{\theta}_X) = -\infty$ and $\lim_{(\boldsymbol{\theta}_X)_j \to -\infty} H_X(\boldsymbol{\theta}_X) = -\infty$ or $\frac{\partial H_X(\boldsymbol{\theta}_X)}{\partial (\boldsymbol{\theta}_X)_j} \equiv 0$. Firstly, we have

$$\begin{aligned}\frac{\partial H_X(\boldsymbol{\theta}_X)}{\partial (\boldsymbol{\theta}_X)_j} &= \sum_{i=1}^{t} \left(X^{(i)} (\boldsymbol{V}_{i,X})_j - (\boldsymbol{V}_{i,X})_j m'_X(\boldsymbol{V}_{i,X}^{\intercal} \boldsymbol{\theta}_X) \right) \\ &= \sum_{i=1}^{t} \left(X^{(i)} (\boldsymbol{V}_{i,X})_j - (\boldsymbol{V}_{i,X})_j f_X(\boldsymbol{V}_{i,X}^{\intercal} \boldsymbol{\theta}_X) \right).\end{aligned}$$

If $(\boldsymbol{V}_{i,X})_j = 0$ for all $i \in [t]$, we have $\frac{\partial H_X(\boldsymbol{\theta}_X)}{\partial (\boldsymbol{\theta}_X)_j} \equiv 0$. Otherwise, we have

$$\begin{aligned}\lim_{(\boldsymbol{\theta}_X)_j \to +\infty} \frac{\partial H_X(\boldsymbol{\theta}_X)}{\partial (\boldsymbol{\theta}_X)_j} &= \lim_{(\boldsymbol{\theta}_X)_j \to +\infty} \sum_{i=1}^{t} \left(X^{(i)} (\boldsymbol{V}_{i,X})_j - (\boldsymbol{V}_{i,X})_j f_X(\boldsymbol{V}_{i,X}^{\intercal} \boldsymbol{\theta}_X) \right) \\ &= \lim_{(\boldsymbol{\theta}_X)_j \to +\infty} \sum_{i=1}^{t} (\boldsymbol{V}_{i,X})_j \left(X^{(i)} - f_X(\boldsymbol{V}_{i,X}^{\intercal} \boldsymbol{\theta}_X) \right) \\ &= -\infty, \qquad\qquad (\lim_{(\boldsymbol{\theta}_X)_j \to +\infty} f_X(\boldsymbol{V}_{i,X}^{\intercal} \boldsymbol{\theta}_X) = +\infty)\end{aligned}$$

which indicates that $\lim_{(\boldsymbol{\theta}_X)_j \to +\infty} H_X(\boldsymbol{\theta}_X) = -\infty$. Also, we have

$$\begin{aligned}\lim_{(\boldsymbol{\theta}_X)_j \to -\infty} \frac{\partial H_X(\boldsymbol{\theta}_X)}{\partial (\boldsymbol{\theta}_X)_j} &= \lim_{(\boldsymbol{\theta}_X)_j \to -\infty} \sum_{i=1}^{t} \left(X^{(i)} (\boldsymbol{V}_{i,X})_j - (\boldsymbol{V}_{i,X})_j f_X(\boldsymbol{V}_{i,X}^{\intercal} \boldsymbol{\theta}_X) \right) \\ &= \lim_{(\boldsymbol{\theta}_X)_j \to -\infty} \sum_{i=1}^{t} (\boldsymbol{V}_{i,X})_j \left(X^{(i)} - f_X(\boldsymbol{V}_{i,X}^{\intercal} \boldsymbol{\theta}_X) \right) \\ &= +\infty, \qquad\qquad (\lim_{(\boldsymbol{\theta}_X)_j \to -\infty} f_X(\boldsymbol{V}_{i,X}^{\intercal} \boldsymbol{\theta}_X) = -\infty)\end{aligned}$$

which indicates that $\lim_{(\boldsymbol{\theta}_X)_j \to -\infty} H_X(\boldsymbol{\theta}_X) = -\infty$.

Until now, we have proved that $H_X(\boldsymbol{\theta}_X)$ has at least one global maximum, which indicates that the equation has at least one solution. ■

Appendix C. Proofs for Propositions in Section 5

In this section, we give proofs that are omitted in Section 5 of our main text.

C.1. Proof of Lemma 2

Lemma 11 *Let G be a BGLM with parameter $\boldsymbol{\theta}^*$ that satisfies Assumption 2. Recall that $\theta^*_{\min} = \min_{(X',X) \in E} \theta^*_{X',X}$. If $X_i \in \boldsymbol{Pa}(X_j)$, we have $\mathbb{E}[X_j | do(X_i = 1)] - \mathbb{E}[X_j | do(X_i = 0)] \geq \kappa \theta^*_{X_i,X_j} \geq \kappa \theta^*_{\min}$; if X_i is not an ancestor of X_j, we have $\mathbb{E}[X_j | do(X_i = 1)] = \mathbb{E}[X_j | do(X_i = 0)]$.*

Proof At first, we define an equivalent threshold model form of the BGLM as follows. For each node X, we randomly sample a threshold γ_X uniformly from $[0, 1]$, i.e., $\gamma_X \sim \mathcal{U}[0, 1]$.

Then if $f_X(\boldsymbol{Pa}(X)\cdot\boldsymbol{\theta}^*_X)+\varepsilon_X \geq \gamma_X$, X is activated, i.e., X is set to 1; otherwise, X is not activated, i.e., X is set to 0. Therefore, if we ignore $\boldsymbol{\varepsilon}$, the BGLM model belongs to the family of general threshold models (Kempe et al., 2003). For convenience, we denote the vector of all $\gamma_X, X \in \boldsymbol{X}\cup\{Y\}\backslash\{X_1\}$ by $\boldsymbol{\gamma}$. The vector of fixing all entries in $\boldsymbol{\gamma}$ except γ_X is denoted by $\boldsymbol{\gamma}_{-X}$.

Now we prove the first part of this lemma: $\mathbb{E}[X_j|do(X_i=1)]-\mathbb{E}[X_j|do(X_i=0)] \geq \kappa\theta^*_{X_i,X_j} \geq \kappa\theta^*_{\min}$ if $X_i \in \boldsymbol{Pa}(X_j)$. By the definition of our equivalent threshold model, we know that after fixing all the thresholds γ_X's and noises ε_X's, the propagation result is completely determined merely by the intervention. Therefore, we have

$$\begin{aligned}\mathbb{E}[X_j|do(X_i=1)] &= \mathbb{E}_{\boldsymbol{\gamma}\in(\mathcal{U}[0,1])^n,\boldsymbol{\varepsilon}}[X_j|do(X_i=1)]\\ &= \mathbb{E}_{\boldsymbol{\gamma}_{-X_j}\in(\mathcal{U}[0,1])^{n-1},\boldsymbol{\varepsilon}}\left[\Pr_{\gamma_{X_j}\sim\mathcal{U}[0,1]}\left\{X_j=1|do(X_i=1),\boldsymbol{\gamma}_{-X_j},\varepsilon\right\}\right],\end{aligned}$$

and

$$\mathbb{E}[X_j|do(X_i=0)] = \mathbb{E}_{\boldsymbol{\gamma}_{-X_j}\in(\mathcal{U}[0,1])^{n-1},\boldsymbol{\varepsilon}}\left[\Pr_{\gamma_{X_j}\sim\mathcal{U}[0,1]}\left\{X_j=1|do(X_i=0),\boldsymbol{\gamma}_{-X_j},\varepsilon\right\}\right].$$

Hence, in order to prove $\mathbb{E}[X_j|do(X_i=1)]-\mathbb{E}[X_j|do(X_i=0)] \geq \kappa\theta^*_{X_i,X_j} \geq \kappa\theta^*_{\min}$, we only need to prove

$$\Pr_{\gamma_{X_j}\sim\mathcal{U}[0,1]}\left\{X_j=1|do(X_i=1),\boldsymbol{\gamma}_{-X_j},\varepsilon\right\} - \Pr_{\gamma_{X_j}\sim\mathcal{U}[0,1]}\left\{X_j=0|do(X_i=0),\boldsymbol{\gamma}_{-X_j},\varepsilon\right\} \geq \kappa\theta^*_{\min}.$$

When $\boldsymbol{\gamma}_{-X_j}$ and $\boldsymbol{\varepsilon}$ are fixed, all the nodes in $\boldsymbol{X}\cup\{Y\}\backslash(\{X_j\}\cup\{\boldsymbol{Des}(X_j)\})$ are already fixed given an arbitrarily fixed intervention. Here, $\boldsymbol{Des}(X_j)$ is used to represent the descendants of X_j. Suppose under $do(X_i=1), \boldsymbol{\gamma}_{-X_j}$ and $\boldsymbol{\varepsilon}$, the value vector of parents of X_j is $\boldsymbol{pa}_1(X_j)$; under $do(X_i=0), \boldsymbol{\gamma}_{-X_j}$ and $\boldsymbol{\varepsilon}$, the value vector of parents of X_j is $\boldsymbol{pa}_0(X_j)$. By induction along the topological order, nodes in $\boldsymbol{X}\cup\{Y\}\backslash(\{X_j\}\cup\{\boldsymbol{Des}(X_j)\})$ that is activated under $do(X_i=0), \boldsymbol{\gamma}_{-X_j}$ and $\boldsymbol{\varepsilon}$ must be also activated under $do(X_i=1), \boldsymbol{\gamma}_{-X_j}$ and $\boldsymbol{\varepsilon}$. Therefore, entries in $\boldsymbol{pa}_1(X_j)-\boldsymbol{pa}_0(X_j)$ are all non-negative and the entry in $\boldsymbol{pa}_1(X_j)-\boldsymbol{pa}_0(X_j)$ for the value of X_j is 1. From this observation, we can deduce that

$$\begin{aligned}f_{X_j}(\boldsymbol{pa}_1(X_j)\cdot\boldsymbol{\theta}^*_{X_j}) - f_{X_j}(\boldsymbol{pa}_0(X_j)\cdot\boldsymbol{\theta}^*_{X_j}) &\geq \kappa\left(\boldsymbol{pa}_1(X_j)\cdot\boldsymbol{\theta}^*_{X_j} - \boldsymbol{pa}_0(X_j)\cdot\boldsymbol{\theta}^*_{X_j}\right)\\ &\geq \kappa\theta^*_{X_i,X_j}.\end{aligned}$$

Hence, we have

$$\begin{aligned}&\Pr_{\gamma_{X_j}\sim\mathcal{U}[0,1]}\left\{X_j=1|do(X_i=1),\boldsymbol{\gamma}_{-X_j},\varepsilon\right\} - \Pr_{\gamma_{X_j}\sim\mathcal{U}[0,1]}\left\{X_j=0|do(X_i=0),\boldsymbol{\gamma}_{-X_j},\varepsilon\right\}\\ &= \Pr_{\gamma_{X_j}\sim\mathcal{U}[0,1]}\left\{f_{X_j}(\boldsymbol{pa}_1(X_j)\cdot\boldsymbol{\theta}^*_{X_j}) \geq \gamma_{X_j}+\varepsilon_{X_j}|\varepsilon_{X_j}\right\}\\ &\quad - \Pr_{\gamma_{X_j}\sim\mathcal{U}[0,1]}\left\{f_{X_j}(\boldsymbol{pa}_0(X_j)\cdot\boldsymbol{\theta}^*_{X_j}) \geq \gamma_{X_j}+\varepsilon_{X_j}|\varepsilon_{X_j}\right\}\\ &= \left(f_{X_j}(\boldsymbol{pa}_1(X_j)\cdot\boldsymbol{\theta}^*_{X_j})-\varepsilon_{X_j}\right) - \left(f_{X_j}(\boldsymbol{pa}_0(X_j)\cdot\boldsymbol{\theta}^*_{X_j})-\varepsilon_{X_j}\right)\\ &\geq \kappa\theta^*_{X_i,X_j} \geq \kappa\theta^*_{\min},\end{aligned}$$

which is what we want. Until now, the first part of Lemma 2 has been proved.

Then we prove the second part of this lemma: $\mathbb{E}[X_j|do(X_i=1)] = \mathbb{E}[X_j|do(X_i=0)]$ if X_j is not a descendant of X_i. In this situation, we know from the graph structure that $(X_j \perp\!\!\!\perp X_i)_{G_{\overline{\{X_i\}}}}$, where $G_{\overline{\{X_i\}}}$ is the graph obtained by deleting from G all arrows pointing to X_i. According to the third law of *do*-calculus (Pearl, 2012), we deduce that

$$\begin{aligned}\mathbb{E}[X_j|do(X_i=1)] &= \Pr\{X_j=1|do(X_i=1)\} = \Pr\{X_j=1|\}\\ &= \Pr\{X_j=1|do(X_i=0)\} = \mathbb{E}[X_j|do(X_i=0)].\end{aligned}$$

Now Lemma 2 is completely proved. ■

Corollary 12 (An Extension of Lemma 2) *Suppose G is a BGLM with parameter $\boldsymbol{\theta}^*$ that satisfying Assumption 2 and $do(\boldsymbol{S}=\boldsymbol{s})$ is an intervention such that $X_i, X_j \notin \boldsymbol{S}$. If $X_i \in \boldsymbol{Pa}(X_j)$, we have $\mathbb{E}[X_j|do(X_i=1), do(\boldsymbol{S}=\boldsymbol{s})] - \mathbb{E}[X_j|do(X_i=0), do(\boldsymbol{S}=\boldsymbol{s})] \geq \kappa\theta^*_{X_i,X_j} \geq \kappa\theta^*_{\min}$; if X_i is not an ancestor of X_j, we have $\mathbb{E}[X_j|do(X_i=1), do(\boldsymbol{S}=\boldsymbol{s})] = \mathbb{E}[X_j|do(X_i=0), do(\boldsymbol{S}=\boldsymbol{s})]$.*

Proof According to Pearl (2012), $\Pr\{X_j|do(X_i), do(\boldsymbol{S})\}$ is equivalent to $\Pr\{X_j|do(X_i)\}$ in a new model G' such that all in-edges of $\boldsymbol{S}$ are deleted and all nodes in $\boldsymbol{S}$ are fixed by $\boldsymbol{s}$. We know that Lemma 2 holds in G', so this corollary holds in G. ■

C.2. Proof of Lemma 3

Lemma 13 (Positive Rate of BGLM-Order) *Suppose Assumption 2 holds for BGLM G. In the initialization phase of Algorithm 1, Algorithm 2 finds a consistent ancestor-descendant relationship for G with probability no less than $1-2\binom{n-1}{2}\exp\left(-\frac{c_0c_1^2T^{1/10}}{2}\right)$ when $\theta^*_{\min} \geq 2c_1\kappa^{-1}T^{-1/5}$.*

Proof We first assume that for every pair of nodes if $X_i \in \boldsymbol{Pa}(X_j)$, Algorithm 2 puts X_j as a descendant of X_i in the ancestor-descendant relationship; if X_j is not a descendant of X_i, Algorithm 2 do not put X_j as an descendant of X_i in the ancestor-descendant relationship. This event is denoted by $\mathcal{E}$ for simplicity. We prove that when event $\mathcal{E}$ does occur, the ancestor-descendant relationship we find is absolutely consistent with the true graph structure of G. Otherwise, suppose there is a mistake in the ancestor-descendant relationship such that X_i is an ancestor of X_j but not put in $\widehat{\boldsymbol{Anc}}(X_j)$. We denote a directed path from X_i to X_j by $X_i \to X_{k_1} \to X_{k_2} \to \cdots \to X_{k_p} \to X_j$. Therefore, X_{k_1} must be put in $\widehat{\boldsymbol{Anc}}(X_i)$, X_{k_2} must be put in $\widehat{\boldsymbol{Anc}}(X_{k_1})$, $\ldots$, X_j must be put in $\widehat{\boldsymbol{Anc}}(X_{k_p})$. In conclusion, X_j should be put in $\widehat{\boldsymbol{Anc}}(X_i)$, which is a contradiction. Hence, there is no mistake in the ancestor-descendant relationship given event $\mathcal{E}$.

Now we only prove that using Algorithm 2, with probability no less than

$$1-2\binom{n-1}{2}\exp\left(-\frac{c_0c_1^2T^{1/10}}{2}\right),$$

event $\mathcal{E}$ defined in the paragraph above occurs. For a pair of nodes $X_i, X_j \in \boldsymbol{X}\backslash\{X_1\}$, if $X_i \in \boldsymbol{Pa}(X_j)$, we know from Lemma 2 that $\mathbb{E}[X_j|do(X_i = 1)] - \mathbb{E}[X_j|do(X_i = 0)] \geq \kappa\theta^*_{\text{min}}$. We denote the difference between random variable X_j given $do(X_i = 1)$ and random variable X_j given $do(X_i = 0)$ by Z. In $\sum_{k=1}^{c_0T^{1/2}}\left(X_j^{\left(2ic_0T^{1/2}+k\right)} - X_j^{\left((2i+1)c_0T^{1/2}+k\right)}\right)$, each term $X_j^{\left(2ic_0T^{1/2}+k\right)} - X_j^{\left((2i+1)c_0T^{1/2}+k\right)}$ is an i.i.d. sample of Z. We denote $X_j^{\left(2ic_0T^{1/2}+k\right)} - X_j^{\left((2i+1)c_0T^{1/2}+k\right)}$ by Z_k. We know that $Z_k \in [-1, 1]$ and $\mathbb{E}[Z_k] \geq \kappa\theta^*_{\text{min}}$, so according to Hoeffding's inequality (Hoeffding, 1994), we have

$$\begin{aligned}
&\Pr\left\{\sum_{k=1}^{c_0T^{1/2}}\left(X_j^{\left(2ic_0T^{1/2}+k\right)} - X_j^{\left((2i+1)c_0T^{1/2}+k\right)}\right) > c_0c_1T^{3/10}\right\}\\
&= \Pr\left\{\sum_{k=1}^{c_0T^{1/2}} Z_k > c_0c_1T^{3/10}\right\}\\
&= 1 - \Pr\left\{\sum_{k=1}^{c_0T^{1/2}} Z_k \leq c_0c_1T^{3/10}\right\}\\
&\geq 1 - \exp\left(-\frac{2\left(c_0T^{1/2}\kappa\theta^*_{\text{min}} - c_0c_1T^{3/10}\right)^2}{4c_0T^{1/2}}\right) = 1 - \exp\left(-\frac{c_0\left(T^{1/4}\kappa\theta^*_{\text{min}} - c_1T^{1/20}\right)^2}{2}\right)\\
&\geq 1 - \exp\left(-\frac{c_1^2c_0T^{1/10}}{2}\right). \qquad\qquad \text{(because } T \geq 32\left(\tfrac{c_1}{\kappa\theta^*_{\text{min}}}\right)^5\text{)}
\end{aligned}$$

Similarly, if X_j is not a descendant of X_i, we do not put X_i in $\widehat{\boldsymbol{Anc}}(X_j)$ in the ancestor-descendant relationship if and only if $\sum_{k=1}^{c_0T^{1/2}}\left(X_j^{\left(2ic_0T^{1/2}+k\right)} - X_j^{\left((2i+1)c_0T^{1/2}+k\right)}\right) \leq c_0c_1T^{3/10}$. Now we still have $Z_k \in [-1, 1]$ but $\mathbb{E}[Z_k] = 0$. Therefore, according to Hoeffding's inequality (Hoeffding, 1994), we have

$$\begin{aligned}
&\Pr\left\{\sum_{k=1}^{c_0T^{1/2}}\left(X_j^{\left(2ic_0T^{1/2}+k\right)} - X_j^{\left((2i+1)c_0T^{1/2}+k\right)}\right) \leq c_0c_1T^{3/10}\right\}\\
&= 1 - \Pr\left\{\sum_{k=1}^{c_0T^{1/2}} Z_k > c_0c_1T^{3/10}\right\}\\
&> 1 - \exp\left(-\frac{2\left(c_0c_1T^{3/10}\right)^2}{4c_0T^{1/2}}\right) = 1 - \exp\left(-\frac{c_1^2c_0T^{1/10}}{2}\right).
\end{aligned}$$

Hence, by union bound (Boole's inequality (Bonferroni, 1936)), the probability of $\mathcal{E}$ is no less than $1 - 2\binom{n-1}{2}\exp\left(-\frac{c_1^2c_0T^{1/10}}{2}\right)$. This is because when $X_i, X_j \in \boldsymbol{X}\backslash\{X_1\}$, there are $2\binom{n-1}{2}$ possible choices of them that are tested by Algorithm 2. When $\mathcal{E}$ happens, Algorithm 2 gets the ancestor-descendant relationship correct, so Lemma 3 is proved. ■

C.3. Proof of Theorem 4

In the following proofs on a BGLM G, when $X' \in \boldsymbol{Anc}(X)$ but $X' \notin \boldsymbol{Pa}(X)$, we add an edge $X' \to X$ with weight $\theta_{X',X} = 0$ into G and this does not impact the propagation results of G. Let $D = \max_{X \in \boldsymbol{X} \cup Y} |\boldsymbol{Pa}(X)|$ represent the maximum in-degree. After applying this transformation, $D = n$ and $\boldsymbol{Anc}(X) = \boldsymbol{Pa}(X)$ for all $X \in \boldsymbol{X} \cup Y$ in this subsection. This transformation effectively converts the ancestor-descendant relationship into an ancestor-descendant graph.

Before the proof of this theorem, we introduce several lemmas at first. The first component is based on the result of maximum-likelihood estimation (MLE). It gives a theoretical measurement for the accuracy of estimated $\hat{\boldsymbol{\theta}}$ computed by MLE. One who is interested could find the proof of this lemma in Appendix C.2 of Feng and Chen (2022).

Lemma 14 (Lemma 1 in Feng and Chen (2023)) *Suppose that Assumptions 1 and 2 hold. Moreover, given $\delta \in (0,1)$, assume that*

$$\lambda_{\min}(M_{t,X}) \geq \frac{512|\boldsymbol{Pa}(X)|\left(L_{f_X}^{(2)}\right)^2}{\kappa^4}\left(|\boldsymbol{Pa}(X)|^2 + \ln\frac{1}{\delta}\right). \tag{8}$$

Then with probability at least $1-3\delta$, the maximum-likelihood estimator satisfies , for any $\boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{Pa}(X)|}$,

$$\left|\boldsymbol{v}^\intercal(\hat{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*_X)\right| \leq \frac{3}{\kappa}\sqrt{\log(1/\delta)}\,\|\boldsymbol{v}\|_{M_{t,X}^{-1}},$$

where the probability is taken from the randomness of all data collected from round 1 to round t.

The second component is called the group observation modulated (GOM) bounded smoothness property (Li et al., 2020). It shows that a small change in parameters $\boldsymbol{\theta}$ leads to a small change in the reward. Under our BGLM setting, this lemma is proved in Appendix C.3 of Feng and Chen (2022).

Lemma 15 (Lemma 2 in Feng and Chen (2023)) *For any two weight vectors $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 \in \Theta$ for a BGLM G, the difference of their expected reward for any intervened set $\boldsymbol{S}$ can be bounded as*

$$\left|\sigma(\boldsymbol{S},\boldsymbol{\theta}^1) - \sigma(\boldsymbol{S},\boldsymbol{\theta}^2)\right| \leq \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\gamma}}\left[\sum_{X \in \boldsymbol{X}_{\boldsymbol{S},Y}} \left|\boldsymbol{V}_X^\intercal(\boldsymbol{\theta}^1_X - \boldsymbol{\theta}^2_X)\right| L_{f_X}^{(1)}\right], \tag{9}$$

where $\boldsymbol{X}_{\boldsymbol{S},Y}$ is the set of nodes in paths from $\boldsymbol{S}$ to Y excluding $\boldsymbol{S}$, and $\boldsymbol{V}_X$ is the propagation result of the parents of X under parameter $\boldsymbol{\theta}^2$. The expectation is taken over the randomness of the thresholds $\boldsymbol{\gamma}$ and the noises $\boldsymbol{\varepsilon}$.

Thirdly, we propose a lemma in order to bound the sum of $\|\boldsymbol{V}_{t,X}\|_{M_{t-1,X}^{-1}}$ at first. This lemma is proved in Appendix C.4 of Feng and Chen (2022).

Lemma 16 (Lemma 9 in Feng and Chen (2022)) *Let $\{\boldsymbol{W}_t\}_{t=1}^\infty$ be a sequence in $\mathbb{R}^d$ satisfying $\|\boldsymbol{W}_t\| \leq \sqrt{d}$. Define $\boldsymbol{W}_0 = \boldsymbol{0}$ and $M_t = \sum_{i=0}^t \boldsymbol{W}_i\boldsymbol{W}_i^\intercal$. Suppose there is an integer t_1 such that $\lambda_{\min}(M_{t_1+1}) \geq 1$, then for all $t_2 > 0$,*

$$\sum_{t=t_1}^{t_1+t_2} \|\boldsymbol{W}_t\|_{M_{t-1}^{-1}} \leq \sqrt{2t_2 d\log(t_2 d + t_1)}.$$

At last, in order to show that $\lambda_{\min}(M_{T_1,X}) \geq R$ after the initialization phase of Algorithm 1 and thus satisfy the condition of Lemma 14, we introduce Lemma 17. This lemma is improved upon Lemma 7 in Feng and Chen (2022) and enables us to use Lecué and Mendelson's inequality (Nie, 2022) in our later theoretical regret analysis.

Let *Sphere*(d) denote the sphere of the d-dimensional unit ball.

Lemma 17 *For any* $\boldsymbol{v} = (v_1, v_2, \dots, v_{|\boldsymbol{Pa}(X)|}) \in \mathit{Sphere}(|\boldsymbol{Pa}(X)|)$ *and any* $X \in \boldsymbol{X} \cup \{Y\}$ *in a BGLM that satisfies Assumption 3, we have*

$$\Pr_{\boldsymbol{\varepsilon},\boldsymbol{X},Y}\left\{|\boldsymbol{Pa}(X)\cdot\boldsymbol{v}| \geq \frac{1}{\sqrt{4D^2-3}}\right\} \geq \zeta,$$

where $\boldsymbol{Pa}(X)$ *is the random vector generated by the natural Bayesian propagation in BGLM* G *with no interventions (except for setting* X_1 *to 1).*

Proof The lemma is similarly proved as Lemma 7 in Feng and Chen (2022) using the idea of Pigeonhole principle. Let $\boldsymbol{Pa}(X) = (X_{i_1} = X_1, X_{i_2}, X_{i_3}, \dots, X_{i_{|\boldsymbol{Pa}(X)|}})$ as the random vector and $\boldsymbol{pa}(X) = (x_1 = 1, x_{i_1}, x_{i_2}, x_{i_3}, \dots, x_{i_{|\boldsymbol{Pa}(X)|}})$ as a possible valuation of $\boldsymbol{Pa}(X)$. Without loss of generality, we suppose that $|v_2| \geq |v_3| \geq \dots \geq \left|v_{|\boldsymbol{Pa}(X)|}\right|$. For simplicity, we denote $D_0 = \sqrt{D-1} + \frac{1}{2\sqrt{D-1}}$. If $|v_1| \geq \frac{D_0}{\sqrt{D_0^2+1}}$, we can deduce that

$$\begin{aligned}
|\boldsymbol{pa}(X)\cdot\boldsymbol{v}| &\geq |v_1| - |v_2| - |v_3| - \cdots - \left|v_{|\boldsymbol{Pa}(X)|}\right| \\
&\geq \frac{D_0}{\sqrt{D_0^2+1}} - \sqrt{(D-1)\left(|v_2|^2 + |v_3|^2 + \cdots + \left|v_{|\boldsymbol{Pa}(X)|}\right|^2\right)} && (10) \\
&\geq \frac{D_0}{\sqrt{D_0^2+1}} - \sqrt{(D-1)\left(1 - \frac{D_0^2}{D_0^2+1}\right)} && (11) \\
&= \frac{1}{2\sqrt{(D_0^2+1)(D-1)}} = \frac{1}{\sqrt{4D^2-3}},
\end{aligned}$$

where Inequality (10) is by the Cauchy-Schwarz inequality and the fact that $|\boldsymbol{Pa}(X)| \leq D$, and Inequality (11) uses the fact that $\boldsymbol{v} \in \mathit{Sphere}(|\boldsymbol{Pa}(X)|)$. Thus, when $|v_1| \geq \frac{D_0}{\sqrt{D_0^2+1}}$, the event $|\boldsymbol{Pa}(X)\cdot\boldsymbol{v}| \geq \frac{1}{\sqrt{4D^2-3}}$ holds deterministically. Otherwise, when $|v_1| < \frac{D_0}{\sqrt{D_0^2+1}}$, we use the fact that $|v_2|$ is the largest among $|v_2|, |v_3|, \dots$ and deduce that

$$|v_2| \geq \frac{1}{\sqrt{n-1}}\sqrt{|v_2|^2 + |v_3|^2 + \cdots} \geq \frac{\sqrt{1 - \left(\frac{D_0}{\sqrt{D_0^2+1}}\right)^2}}{\sqrt{n-1}} = \frac{2}{\sqrt{4D^2-3}}. \tag{12}$$

Therefore, using the fact that

$$\begin{aligned}
&\Pr_{\boldsymbol{\varepsilon},\boldsymbol{X},Y}\{X_{i_1} = 1, X_{i_2} = x_{i_2}, X_{i_3} = x_{i_3}, \dots\} \\
&= \Pr_{\boldsymbol{\varepsilon},\boldsymbol{X},Y}\{X_{i_2} = x_{i_2} | X_{i_1} = 1, X_{i_3} = x_{i_3}, \dots\} \cdot \Pr_{\boldsymbol{\varepsilon},\boldsymbol{X},Y}\{(X_{i_1} = 1, X_{i_3} = x_{i_3}, \dots\} \\
&\geq \zeta \Pr_{\boldsymbol{\varepsilon},\boldsymbol{X},Y}\{X_{i_1} = 1, X_{i_3} = x_{i_3}, \dots\}
\end{aligned}$$

and $\sum_{x_{i_3},x_{i_4},\dots} \Pr_{\boldsymbol{\varepsilon},\boldsymbol{X},Y}\left\{X_{i_1}=1, X_{i_3}=x_{i_3},\dots\right\}=1$, we have

$$
\begin{aligned}
&\Pr_{\boldsymbol{\varepsilon},\boldsymbol{X},Y}\left\{|\boldsymbol{Pa}(X)\cdot\boldsymbol{v}|\geq\frac{1}{\sqrt{4D^2-3}}\right\}\\
&=\sum_{x_{i_3},x_{i_4},\dots}\Pr\{X_{i_1}=1,X_{i_2}=1,X_{i_3}=x_{i_3},\dots\}\cdot\mathbb{I}\left\{|(1,1,x_{i_3},x_{i_4},\dots)\cdot(v_1,v_2,v_3,\dots)|\geq\frac{1}{\sqrt{4D^2-3}}\right\}\\
&+\sum_{x_{i_3},x_{i_4},\dots}\Pr\{X_{i_1}=1,X_{i_2}=0,X_{i_3}=x_{i_3},\dots\}\cdot\mathbb{I}\left\{|(1,0,x_{i_3},x_{i_4},\dots)\cdot(v_1,v_2,v_3,\dots)|\geq\frac{1}{\sqrt{4D^2-3}}\right\}\\
&\geq\sum_{x_{i_3},x_{i_4},\dots}\zeta\Pr\{X_{i_1}=1,X_{i_3}=x_{i_3},X_{i_4}=x_{i_4}\dots\}\cdot\mathbb{I}\left\{|(1,1,x_{i_3},x_{i_4},\dots)\cdot(v_1,v_2,v_3,\dots)|\geq\frac{1}{\sqrt{4D^2-3}}\right\}\\
&+\sum_{x_{i_3},x_{i_4},\dots}\zeta\Pr\{X_{i_1}=1,X_{i_3}=x_{i_3},X_{i_4}=x_{i_4},\dots\}\cdot\mathbb{I}\left\{|(1,0,x_{i_3},x_{i_4},\dots)\cdot(v_1,v_2,v_3,\dots)|\geq\frac{1}{\sqrt{4D^2-3}}\right\}\\
&=\zeta\sum_{x_{i_3},x_{i_4},\dots}\Pr\{X_{i_1}=1,X_{i_3}=x_{i_3},X_{i_4}=x_{i_4},\dots\}\left(\mathbb{I}\left\{|(1,1,x_{i_3},x_{i_4},\dots)\cdot(v_1,v_2,v_3,\dots)|\geq\frac{1}{\sqrt{4D^2-3}}\right\}\right.\\
&\left.+\mathbb{I}\left\{|(1,0,x_{i_3},x_{i_4},\dots)\cdot(v_1,v_2,v_3,\dots)|\geq\frac{1}{\sqrt{4D^2-3}}\right\}\right)\\
&\geq\zeta\sum_{x_{i_3},x_{i_4},\dots}\Pr\{X_{i_1}=1,X_{i_3}=x_{i_3},X_{i_4}=x_{i_4},\dots\} \qquad (13)\\
&=\zeta,
\end{aligned}
$$

which is exactly what we want to prove. Inequality (13) holds because otherwise, at least for some $x_{i_3},x_{i_4},\dots$, both indicators on the left-hand side of the inequality have to be 0, which implies that

$$|(1,1,x_{i_3},x_{i_4},\dots)\cdot(v_1,v_2,v_3,\dots)-(1,0,x_{i_3},x_{i_4},\dots)\cdot(v_1,v_2,v_3,\dots)|=|v_2|<\frac{2}{\sqrt{4D^2-3}}, \qquad (14)$$

but this contradicts to Inequality (12). ■

Having these four lemmas above together with Lemma 3 proved in Appendix C.2, we are finally able to prove the regret bound of BGLM-OFU-Unknown algorithm (Theorem 4) as below.

Theorem 4 (Regret Bound of BGLM-OFU-Unknown) *Under Assumptions 1, 2 and 3, the regret of BGLM-OFU-Unknown (Algorithms 1, 2 and 5) is bounded as*

$$R(T)=O\left(\frac{1}{\kappa}n^{\frac{3}{2}}L_{\max}^{(1)}\sqrt{T}\log T\right), \qquad (5)$$

where $L_{\max}^{(1)}=\max_{X\in\boldsymbol{X}\cup\{Y\}}L_{f_X}^{(1)}$ *and the terms of* $o(\sqrt{T}\ln T)$ *are omitted, and the big* O *notation holds for* $T\geq 32\left(\frac{c_1}{\kappa\theta^*_{\min}}\right)^5$.

Proof We only consider the case of $T\geq 32\left(\frac{c_1}{\kappa\theta^*_{\min}}\right)^5$ in this proof because the big O notation is asymptotic.

Let H_t be the history of the first t rounds and R_t be the regret in the t^{th} round. Because the reward node Y is in interval $[0,1]$, we can deduce that for any $t \le T_1$, $R_t \le 1$. Now we consider the case of $t > T_1$. According to Lemma 3, with probability at least $1 - 2\binom{n-1}{2}\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right)$, Algorithm 2 returns a correct ancestor-descendant relationship, i.e., $\widehat{\boldsymbol{Anc}}(X) = \boldsymbol{Anc}(X)$ for $X \in \boldsymbol{X} \cup \{Y\}$. Next we bound the regret conditioned on the correct ancestor-descendant relationship. When $t > T_1$, we have

$$\mathbb{E}[R_t | H_{t-1}] = \mathbb{E}[\sigma(\mathbf{S}^{\text{opt}}, \boldsymbol{\theta}^*) - \sigma(\mathbf{S}_t, \boldsymbol{\theta}^*) | H_{t-1}], \tag{15}$$

where the expectation is taken over the randomness of $\mathbf{S}_t$. Then for $T_1 < t \le T$, we define $\xi_{t-1,X}$ for $X \in \mathbf{X} \cup \{Y\}$ as $\xi_{t-1,X} = \left\{\left|\boldsymbol{v}^T(\hat{\boldsymbol{\theta}}_{t-1,X} - \boldsymbol{\theta}^*_X)\right| \le \rho \cdot \|\boldsymbol{v}\|_{M^{-1}_{t-1,X}}, \forall \boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{Pa}(X)|}\right\}$. According to the definition of Algorithm 1, we can deduce that $\lambda_{\min}(M_{t-1,X}) \ge \lambda_{\min}(M_{T_1,X})$. By Lecué and Mendelson's inequality (Nie, 2022; Feng and Chen, 2022) (conditions of this inequality satisfied according to Lemma 17), we have

$$\Pr\left\{\lambda_{\min}(M_{T_1,X}) < R\right\} \le \Pr\left\{\lambda_{\min}(M_{T_1,X} - M_{T_0,X}) < R\right\} \le \exp\left(-\frac{(T_1 - T_0)\zeta^2}{c}\right)$$

where c, ζ are constants. Then we can define $\xi_{t-1} = \wedge_{X \in \mathbf{X} \cup \{Y\}} \xi_{t-1,X}$ and let $\overline{\xi_{t-1}}$ be its complement. By Lemma 14, we have

$$\Pr\left\{\overline{\xi_{t-1}}\right\} \le \left(3\delta + \exp\left(-\frac{(T_1 - T_0)\zeta^2}{c}\right) + 3\delta \exp\left(-\frac{(T_1 - T_0)\zeta^2}{c}\right)\right) n \triangleq p_{\text{error}}.$$

Because under ξ_{t-1}, for any $X \in \mathbf{X} \cup \{Y\}$ and $\boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{Pa}(X)|}$, we have $\left|\boldsymbol{v}^T(\hat{\boldsymbol{\theta}}_{t-1,X} - \boldsymbol{\theta}^*_X)\right| \le \rho \cdot \|\boldsymbol{v}\|_{M^{-1}_{t-1,X}}$. Therefore, by the definition of $\tilde{\boldsymbol{\theta}}_t$, we have $\sigma(\mathbf{S}_t, \tilde{\boldsymbol{\theta}}_t) \ge \sigma(\mathbf{S}^{\text{opt}}, \boldsymbol{\theta}^*)$ because $\boldsymbol{\theta}^*$ is in our confidence ellipsoid. Hence,

$$\begin{aligned}
\mathbb{E}[R_t] &\le \Pr\left\{\xi_{t-1}\right\} \cdot \mathbb{E}[\sigma(\mathbf{S}^{\text{opt}}, \boldsymbol{\theta}^*) - \sigma(\mathbf{S}_t, \boldsymbol{\theta}^*)] + \Pr(\overline{\xi_{t-1}}) \\
&\le \mathbb{E}[\sigma(\mathbf{S}^{\text{opt}}, \boldsymbol{\theta}^*) - \sigma(\mathbf{S}_t, \boldsymbol{\theta}^*)] + p_{\text{error}} \\
&\le \mathbb{E}[\sigma(\mathbf{S}_t, \tilde{\boldsymbol{\theta}}_t) - \sigma(\mathbf{S}_t, \boldsymbol{\theta}^*)] + p_{\text{error}}.
\end{aligned}$$

Then we need to bound $\sigma(\mathbf{S}_t, \tilde{\boldsymbol{\theta}}_t) - \sigma(\mathbf{S}_t, \boldsymbol{\theta}^*)$ carefully.

Therefore, according to Lemma 14 and Lemma 15, we can deduce that

$$\begin{aligned}
\mathbb{E}[R_t] &\le \mathbb{E}\left[\sum_{X \in \mathbf{X}_{\mathbf{S}_t,Y}} \left|\boldsymbol{V}_{t,X}(\tilde{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*_X)\right| L^{(1)}_{f_X}\right] + p_{\text{error}} \\
&\le \mathbb{E}\left[\sum_{X \in \mathbf{X}_{\mathbf{S}_t,Y}} \|\boldsymbol{V}_{t,X}\|_{M^{-1}_{t-1,X}} \left\|\tilde{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*_X\right\|_{M_{t-1,X}} L^{(1)}_{f_X}\right] + p_{\text{error}} \\
&\le 2\rho \cdot \mathbb{E}\left[\sum_{X \in \mathbf{X}_{\mathbf{S}_t,Y}} \|\boldsymbol{V}_{t,X}\|_{M^{-1}_{t-1,X}} L^{(1)}_{f_X}\right] + p_{\text{error}}.
\end{aligned}$$

The last inequality holds because

$$\left\|\tilde{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*_X\right\|_{M_{t-1,X}} \leq \left\|\tilde{\boldsymbol{\theta}}_{t,X} - \hat{\boldsymbol{\theta}}_{t-1,X}\right\|_{M_{t-1,X}} + \left\|\hat{\boldsymbol{\theta}}_{t-1,X} - \boldsymbol{\theta}^*_X\right\|_{M_{t-1,X}} \leq 2\rho.$$

Therefore, conditioned on the correct ancestor-descendant relationship, the total regret can be bounded as

$$R(T) \leq 2\rho \cdot \mathbb{E}\left[\sum_{t=T_0+1}^{T} \sum_{X \in \mathbf{X}_{\mathbf{S}_t,Y}} \|\boldsymbol{V}_{t,X}\|_{M^{-1}_{t-1,X}} L^{(1)}_{f_X}\right] + p_{\text{error}}(T - T_1) + T_1.$$

For convenience, we define $\boldsymbol{W}_{t,X}$ as a vector such that if $X \in S_t$, $\boldsymbol{W}_{t,X} = \mathbf{0}^{|\boldsymbol{Pa}(X)|}$; if $X \notin S_t$, $\boldsymbol{W}_{t,X} = \boldsymbol{V}_{t,X}$. Using Lemma 16, we can get the result:

$$\begin{aligned}
R(T) \leq & \left(2\rho\mathbb{E}\left[\sum_{t=T_0+1}^{T} \sum_{X \in \mathbf{X}_{\mathbf{S}_t,Y}} \|\boldsymbol{V}_{t,X}\|_{M^{-1}_{t-1,X}} L^{(1)}_{f_X}\right] + p_{\text{error}}(T - T_1) + T_1\right) \\
& \cdot \left(1 - 2\binom{n-1}{2}\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right)\right) + 2\binom{n-1}{2}\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right) T \\
\leq & \ 2\rho\mathbb{E}\left[\sum_{t=T_0+1}^{T} \sum_{X \in \boldsymbol{X} \cup \{Y\}} \|\boldsymbol{W}_{t,X}\|_{M^{-1}_{t-1,X}} L^{(1)}_{f_X}\right] + p_{\text{error}}(T - T_1) + T_1 \\
& + 2\binom{n-1}{2}\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right) T \\
\leq & \ 2\rho \cdot \max_{X \in \mathbf{X} \cup \{Y\}} \left(L^{(1)}_{f_X}\right) \mathbb{E}\left[\sum_{X \in \boldsymbol{X} \cup \{Y\}} \sqrt{2(T - T_0)|\boldsymbol{Pa}(X)| \log\left((T - T_0)|\boldsymbol{Pa}(X)| + T_0\right)}\right] \\
& + p_{\text{error}}(T - T_1) + T_1 + 2\binom{n-1}{2}\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right) T \\
= & \ O\left(\frac{1}{\kappa} n^{\frac{3}{2}} \sqrt{T} L^{(1)}_{\max} \ln T\right) = \tilde{O}\left(\frac{1}{\kappa} n^{\frac{3}{2}} \sqrt{T} L^{(1)}_{\max}\right)
\end{aligned}$$

because $\rho = \frac{3}{\kappa}\sqrt{\log(1/\delta)}$, $\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right) T = o(\sqrt{T})$ and $p_{\text{error}} T = o(\sqrt{T})$. ■

Appendix D. A BLM CCB Algorithm with Minimum Weight Gap Based on Linear Regression

As BLM is a special case of BGLM, the initialization phase in BGLM-OFU-Unknown to determine the ancestor-descendant relationship can also be used on BLMs. Feng and Chen (2023) propose a CCB algorithm for BLMs using linear regression instead of MLE to remove the requirement of Assumption 3. Furthermore, BLM takes the identity function as f_X's, so Assumptions 1 and 2 is neither required. The specific algorithm BLM-LR-Unknown-SG

Algorithm 6: BLM-LR-Unknown-SG for BLM and Linear Model CCB Problem

```
Input: Graph $G = (\boldsymbol{X} \cup \{Y\}, E)$, action set $\mathcal{A}$, positive constants $c_0$ and $c_1$ for initialization phase such that $c_0\sqrt{T} \in \mathbb{N}^+$.
/* Initialization Phase: */
Initialize $T_0 \leftarrow 2(n-1)c_0T^{1/2}$.
Do each intervention among $do(X_2 = 1), do(X_2 = 0), \ldots, do(X_n = 1), do(X_n = 0)$ for $c_0T^{1/2}$ times in order and observe the feedback $(\mathbf{X}_t, Y_t)$ for $1 \le t \le T_0$.
Determine a feasible ancestor-descendant relationship $\widehat{\boldsymbol{Anc}}(X)$'s for $X \in \boldsymbol{X} \cup \{Y\}$ by BGLM-Ancestors($(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{T_0}, Y_{T_0}), c_1$) (see Algorithm 2).
/* Parameters Initialization: */
Initialize $M_{T_0,X} \leftarrow \mathbf{I} \in \mathbb{R}^{|\widehat{\boldsymbol{Anc}}(X)|\times|\widehat{\boldsymbol{Anc}}(X)|}$, $\boldsymbol{b}_{T_0,X} \leftarrow \mathbf{0}^{|\widehat{\boldsymbol{Anc}}(X)|}$ for all $X \in \boldsymbol{X} \cup \{Y\}$, $\hat{\boldsymbol{\theta}}_{T_0,X} \leftarrow 0 \in \mathbb{R}^{|\widehat{\boldsymbol{Anc}}(X)|}$ for all $X \in \boldsymbol{X} \cup \{Y\}$, $\delta \leftarrow \frac{1}{n\sqrt{T}}$ and $\rho_t \leftarrow \sqrt{n\log(1+tn) + 2\log\frac{1}{\delta}} + \sqrt{n}$ for $t = 0, 1, 2, \ldots, T$.
/* Iterative Phase: */
for $t = T_0 + 1, T_0 + 2, \ldots, T$ do
    Compute the confidence ellipsoid $\mathcal{C}_{t,X} = \{\boldsymbol{\theta}'_X \in [0,1]^{|\widehat{\boldsymbol{Anc}}(X)|} : \left\|\boldsymbol{\theta}'_X - \hat{\boldsymbol{\theta}}_{t-1,X}\right\|_{M_{t-1,X}} \le \rho_{t-1}\}$ for any node $X \in \boldsymbol{X} \cup \{Y\}$.
    $(\boldsymbol{S}_t, \boldsymbol{s}_t, \tilde{\boldsymbol{\theta}}_t) = \operatorname{argmax}_{do(\boldsymbol{S}=\boldsymbol{s})\in\mathcal{A}, \boldsymbol{\theta}'_{t,X}\in\mathcal{C}_{t,X}} \mathbb{E}[Y|do(\boldsymbol{S} = \boldsymbol{s})]$.
    Intervene all the nodes in $\boldsymbol{S}_t$ to $\boldsymbol{s}_t$ and observe the feedback $(\boldsymbol{X}_t, Y_t)$.
    for $X \in \boldsymbol{X} \cup \{Y\}$ do
        Construct data pair $(\boldsymbol{V}_{t,X}, X^{(t)})$ with $\boldsymbol{V}_{t,X}$ the vector of ancestors of $X$ in round $t$, and $X^{(t)}$ the value of $X$ in round $t$ if $X \notin S_t$.
        $M_{t,X} = M_{t-1,X} + \boldsymbol{V}_{t,X}\boldsymbol{V}_{t,X}^\intercal$, $\boldsymbol{b}_{t,X} = \boldsymbol{b}_{t-1,X} + X^{(t)}\boldsymbol{V}_{t,X}$, $\hat{\boldsymbol{\theta}}_{t,X} = M_{t,X}^{-1}\boldsymbol{b}_{t,X}$.
    end for
end for
```

(BLM-LR-Unknown Algorithm with Safety Gap (Minimum Weight Gap)) is demonstrated in Algorithm 6.

The following theorem shows the regret bound of BLM-LR-Unknown-SG. It is not surprising that this algorithm could also work on linear models with continuous variables as Appendix F in Feng and Chen (2022). The dominant term in the expected regret does not increase compared to BLM-LR in Feng and Chen (2023).

Theorem 18 (Regret Bound of BLM-LR-Unknown-SG) *The regret of BLM-LR-Unknown-SG running on BLM or linear model is bounded as*

$$R(T) = O\left(n^{\frac{5}{2}}\sqrt{T}\log T\right),$$

where the terms of $o(\sqrt{T}\ln T)$ are omitted, and the big O notation holds for $T \ge 32\left(\frac{c_1}{\kappa\theta^*_{\min}}\right)^5$.

Proof In the following proof on G, when $X' \in \boldsymbol{Anc}(X)$ but $X' \notin \boldsymbol{Pa}(X)$, we add an edge $X' \to X$ with weight $\theta_{X',X} = 0$ into G and this does not impact the propagation results of G. After doing this transformation, $D = n$ and $\boldsymbol{Anc}(X) = \boldsymbol{Pa}(X)$ for all $X \in \boldsymbol{X} \cup \{Y\}$.

According to Lemma 3, with probability at least $1 - 2\binom{n-1}{2}\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right)$, Algorithm 2 returns a correct ancestor-descendant relationship, i.e., $\boldsymbol{Anc}(X) = \widehat{\boldsymbol{Anc}}(X)$ for $X \in \boldsymbol{X} \cup \{Y\}$. Moreover, by Lemma 11 in Feng and Chen (2022), with probability at most $n\delta$, event $\left\{\exists T_0 < t \leq T, x \in \boldsymbol{X} \cup \{Y\} : \left\|\boldsymbol{\theta}_X^{*'} - \hat{\boldsymbol{\theta}}_{t,X}\right\| > \rho_t\right\}$ occurs. Now we bound the expected regret conditioned on the absence of this event and finding a correct ancestor-descendant relationship. For $T_0 < t \leq T$, according to Theorem 1 in Li et al. (2020) and Theorem 15, we can deduce that

$$\begin{aligned}
\mathbb{E}\left[R_t\right] &= \mathbb{E}\left[\sigma'(\boldsymbol{S}^{\text{opt}}, \boldsymbol{\theta}^{*'}) - \sigma'(\boldsymbol{S}_t, \boldsymbol{\theta}^{*'})\right] \\
&\leq \mathbb{E}\left[\sigma'(\boldsymbol{S}_t, \tilde{\boldsymbol{\theta}}_t) - \sigma'(\boldsymbol{S}_t, \boldsymbol{\theta}^{*'})\right] \\
&\leq \mathbb{E}\left[\sum_{X \in \boldsymbol{X}_{\boldsymbol{S}_t, Y}} \left|\boldsymbol{V}_{t,X}^{\intercal}(\tilde{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}_X^{*'})\right|\right] \\
&\leq \mathbb{E}\left[\sum_{X \in \boldsymbol{X}_{\boldsymbol{S}_t, Y}} \left\|\boldsymbol{V}_{t,X}\right\|_{M_{t-1,X}^{-1}} \left\|\tilde{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}_X^{*'}\right\|_{M_{t-1,X}}\right] \\
&\leq \mathbb{E}\left[\sum_{X \in \boldsymbol{X}_{\boldsymbol{S}_t, Y}} 2\rho_{t-1} \left\|\boldsymbol{V}_{t,X}\right\|_{M_{t-1,X}^{-1}}\right],
\end{aligned}$$

since $\tilde{\boldsymbol{\theta}}_{t,X}, \boldsymbol{\theta}_X^{*}$ are both in the confidence set. Thus, we have

$$\begin{aligned}
R(T) = \mathbb{E}\left[\sum_{t=1}^{T} R_t\right] &\leq \mathbb{E}\left[\sum_{t=T_0+1}^{T} R_t\right] + T_0 \\
&\leq 2\rho_T \cdot \mathbb{E}\left[\sum_{t=T_0+1}^{T} \sum_{X \in \boldsymbol{X}_{\boldsymbol{S}_t, Y}} \left\|\boldsymbol{V}_{t,X}\right\|_{M_{t-1,X}^{-1}}\right] + T_0.
\end{aligned}$$

For convenience, we define $\boldsymbol{W}_{t,X}$ as a vector such that if $X \in S_t$, $\boldsymbol{W}_{t,X} = \boldsymbol{0}^{|\boldsymbol{Pa}(X)|}$; if $X \notin S_t$, $\boldsymbol{W}_{t,X} = \boldsymbol{V}_{t,X}$. According to Cauchy-Schwarz inequality, we have

$$\begin{aligned}
R(T) &\leq 2\rho_T \cdot \mathbb{E}\left[\sum_{t=T_0+1}^{T} \sum_{X \in \boldsymbol{X} \cup \{Y\}} \left\|\boldsymbol{W}_{t,X}\right\|_{M_{t-1,X}^{-1}}\right] + T_0 \\
&\leq 2\rho_T \cdot \mathbb{E}\left[\sqrt{T} \cdot \sum_{X \in \boldsymbol{X} \cup \{Y\}} \sqrt{\sum_{t=T_0+1}^{T} \left\|\boldsymbol{W}_{t,X}\right\|_{M_{t-1,X}^{-1}}^2}\right] + T_0 \\
&\leq 2\rho_T \cdot \mathbb{E}\left[\sqrt{T} \cdot \sum_{X \in \boldsymbol{X} \cup \{Y\}} \sqrt{\sum_{t=1}^{T} \left\|\boldsymbol{W}_{t,X}\right\|_{M_{t-1,X}^{-1}}^2}\right] + 2(n-1)c_0 T^{1/2}.
\end{aligned}$$

Note that $M_{t,X} = M_{t-1,X} + \boldsymbol{W}_{t,X}\boldsymbol{W}_{t,X}^{\intercal}$ and therefore,

$$\det\left(M_{t,X}\right) = \det(M_{t-1,X})\left(1 + \left\|\boldsymbol{W}_{t,X}\right\|_{M_{t-1,X}^{-1}}^2\right),$$

we have

$$\begin{aligned}\sum_{t=1}^{T}\|\boldsymbol{W}_{t,X}\|^2_{M^{-1}_{t-1,X}} &\leq \sum_{t=1}^{T}\frac{n}{\log(n+1)}\cdot\log\left(1+\|\boldsymbol{W}_{t,X}\|^2_{M^{-1}_{t-1,X}}\right)\\ &\leq \frac{n}{\log(n+1)}\cdot\log\frac{\det(M_{T,X})}{\det(\mathbf{I})}\\ &\leq \frac{n|\boldsymbol{Pa}(X)|}{\log(n+1)}\cdot\log\frac{\operatorname{tr}(M_{T,X})}{|\boldsymbol{Pa}(X)|}\\ &\leq \frac{n|\boldsymbol{Pa}(X)|}{\log(n+1)}\cdot\log\left(1+\sum_{t=1}^{T}\frac{\|\boldsymbol{W}_{t,X}\|_2^2}{|\boldsymbol{Pa}(X)|}\right)\\ &\leq \frac{nD}{\log(n+1)}\log(1+T).\end{aligned}$$

Therefore, the final conditional regret $R(T)$ is bounded by

$$R(T)\leq 2\rho_T n\sqrt{T\frac{nD}{\log(n+1)}\log(1+T)}+2(n-1)c_0T^{1/2},$$

because $\rho_T=\sqrt{D\log(1+TD)+2\log\frac{1}{\delta}}+\sqrt{D}$. When

$$\left\{\exists t\in(T_0,T], x\in\boldsymbol{X}\cup\{Y\}:\left\|\boldsymbol{\theta}^{*\prime}_X-\hat{\boldsymbol{\theta}}_{t,X}\right\|>\rho_t\right\}$$

does occur or Algorithm 2 finds an incorrect order, the regret is no more than T. Therefore, the total regret is no more than

$$\begin{aligned}&\left(2\rho_T n\sqrt{T\frac{nD}{\log(n+1)}\log(1+T)}+2(n-1)c_0T^{1/2}\right)\left(1-n\delta-2\binom{n-1}{2}\exp\left(-\frac{c_0c_1^2T^{1/10}}{2}\right)\right)\\ &\quad+T\left(n\delta+2\binom{n-1}{2}\exp\left(-\frac{c_0c_1^2T^{1/10}}{2}\right)\right)\\ &\leq 2\rho_T n\sqrt{T\frac{nD}{\log(n+1)}\log(1+T)}+o(\sqrt{T}\ln T)\\ &=O\left(n^{\frac{5}{2}}\sqrt{T}\log T\right),\end{aligned}$$

which is exactly what we want.

Replacing Lemma 11 in Feng and Chen (2022) by Lemma 12 in Feng and Chen (2022), the above proof for BLMs is still feasible for the regret on linear models without any other modification. ■

Remark 19 *According to the transformation in Section 5.1 of Feng and Chen (2023), this algorithm also works for some BLMs with hidden variables. Using that transformation, running BLM-LR-Unknown-SG on G is equivalent to running on a Markovian BLM or linear model G', where parameter $\boldsymbol{\theta}^*$ is also transformed to a new set of parameters $\boldsymbol{\theta}^{*\prime}$. Here, we disallow the graph structure where a hidden node has two paths to X_i and X_i's descendant X_j and the paths contain only hidden nodes except the end points X_i and X_j.*

Appendix E. Proofs for Propositions in Section 6

E.1. Proof of Lemma 6

Lemma 20 *In Algorithm 3, if the constants c_0 and c_1 satisfy that $c_0 \geq \max\{\frac{1}{c_1^2}, \frac{1}{(1-c_1)^2}\}$, with probability at least $1-(n-1)(n-2)\frac{1}{T^{1/3}}$, after the initialization phase we have*

1). If X' is a true parent of X in G with weight $\theta^*_{X',X} \geq T^{-1/3}$, the edge $X' \rightarrow X$ will be identified and added to the estimated graph G'.

2). If X' is not an ancestor of X in G, $X' \rightarrow X$ will not be added into G'.

Proof First, for each node X_j and its parent X_i with weight $\theta^*_{X_i,X_j} \geq T^{-1/3}$, by Lemma 2, we can have

$$\mathbb{E}[X_j \mid do(X_i = 1)] - \mathbb{E}[X_j \mid do(X_i = 0)] \geq \theta^*_{X_i,X_j}$$

Then each element $X_j^{\left(c_0(2i)T^{2/3}+k\right)} - X_j^{\left(c_0(2i+1)T^{2/3}+k\right)}$ is an i.i.d sample of $Z = X_j \mid_{do(X_i=1)} - X_j \mid_{do(X_i=0)}$ with $\mathbb{E}[Z] \geq \theta^*_{X_i,X_j} \geq T^{-1/3}$. By the Hoeffding's inequality, if we choose $c_1 < 1$ and $c_0(1-c_1)^2 > \frac{1}{3}$, we have

$$\begin{aligned}
&\Pr\left\{\sum_{k=1}^{c_0T^{2/3}}\left(X_j^{\left(c_0(2i)T^{2/3}+k\right)} - X_j^{\left(c_0(2i+1)T^{2/3}+k\right)}\right) > c_0c_1T^{1/3}\log(T^2)\right\}\\
&\geq 1 - \exp\left(-\frac{2\log(T^2)\left(c_0T^{2/3}\mathbb{E}[Z] - c_0c_1T^{1/3}\right)^2}{4c_0T^{2/3}}\right)\\
&\geq 1 - \exp\left(-\frac{2\log(T^2)\left(c_0T^{1/3} - c_0c_1T^{1/3}\right)^2}{4c_0T^{2/3}}\right)\\
&\geq 1 - \exp\left(-\frac{c_0(1-c_1)^2\log(T^2)}{2}\right)\\
&\geq 1 - T^{-c_0(1-c_1)^2}\\
&\geq 1 - \frac{1}{T}.
\end{aligned}$$

Taking the union bound for all X and X', with probability at least $1-\binom{n-1}{2}\frac{1}{T^2}$, the edge $X' \rightarrow X$ with $\theta^*_{X',X}$ will be identified and added to the estimated graph G'. Also, assume X_i is not an ancestor of X_j, then

$$\mathbb{E}[X_j \mid do(X_i = 1)] - \mathbb{E}[X_i \mid do(X_i = 0)] = 0.$$

Thus the element $X_j^{\left(c_0(2i)T^{2/3}+k\right)} - X_j^{\left(c_0(2i+1)T^{2/3}+k\right)}$ is an i.i.d sample of $Z' = X_j \mid_{do(X_i=1)} -X_j \mid_{do(X_i=0)}$ with $\mathbb{E}[Z'] = 0$. Thus by Hoeffding's inequality,

$$
\begin{aligned}
&\Pr\left\{\sum_{k=1}^{c_0T^{2/3}}\left(X_j^{\left(c_0(2i)T^{2/3}+k\right)} - X_j^{\left(c_0(2i+1)T^{2/3}+k\right)}\right) > c_0c_1T^{1/3}\log(T^2)\right\}\\
&\leq \exp\left(-\frac{2\log(T^2)\left(c_0T^{2/3}\mathbb{E}[Z] - c_0c_1T^{1/3}\right)^2}{4c_0T^{2/3}}\right)\\
&\leq \exp\left(-c_0c_1^2\log T\right)\\
&\leq T^{-c_0c_1^2}\\
&\leq \frac{1}{T}.
\end{aligned}
$$

and then with probability at least $1 - \binom{n-1}{2}\frac{1}{T}$, we will not add the edge $X' \to X$ in the graph G. Combining these two facts, we complete the proof. ■

E.2. Proof of Lemma 7

For each node X, consider the estimated possible parent $\boldsymbol{Pa}'(X)$, then our observation $\boldsymbol{V}_{t,X} \in \{0,1\}^{\boldsymbol{Pa}'(X)}$ are the values of $\boldsymbol{Pa}'(X)$. Since we have θ' that

$$\mathbb{E}[X_t \mid \boldsymbol{V}_{t,X}] = \boldsymbol{\theta}_{t,X}^T\boldsymbol{V}_{t,X}. \tag{16}$$

Thus applying Lemma 1 in Li et al. (2020), we can have

$$|\boldsymbol{\theta}'_X - \boldsymbol{\theta}'_{t,X}|_{M_{t,X}} \leq \sqrt{n\log(1+tn) + 2\log(1/\delta)} + \sqrt{n}. \tag{17}$$

E.3. Proof of Lemma 8

Note that M represents the model with true graph G and true weights $\boldsymbol{\theta}$, and M' represents the model with estimated graph G' and estimated weights M', then difference

$$|\theta'_{X_i,X} - \theta_{X_i,X}| \leq nr \tag{18}$$

Now we construct a auxillary model M'', which has graph G' and weights θ on it. The parent of X in model M $\boldsymbol{Pa}''(X)$ is equivalent to $\boldsymbol{Pa}'(X)$. Then we prove the following two claims:

Claim 1 $|\mathbb{E}_M[Y \mid do(S = \boldsymbol{1})] - \mathbb{E}_{M''}[Y \mid do(S = \boldsymbol{1})]| \leq n^2r.$

Proof Let the topological order be $X_1, X_2, \ldots, X_n$. First, $\mathbb{E}_M[X_1 \mid do(S)] - \mathbb{E}_{M''}[X_1 \mid do(S)] = 0 \leq nr$ because X_1 is always 1. Assume $X_{q+1} \notin S$ $\mathbb{E}_M[X_i \mid do(S)] - \mathbb{E}_{M''}[X_i \mid do(S)] \leq qnr$ for all $i \leq q$, then if $X_{q+1} \in S$, $\mathbb{E}_M[X_{q+1} \mid do(S)] - \mathbb{E}_{M''}[X_{q+1} \mid do(S)] = 0 \leq$

$(q+1)nr$ holds trivially. Thus now we assume $X_{q+1} \notin S$.

$$\begin{aligned}
&\mathbb{E}_M[X_{q+1} \mid do(S)] - \mathbb{E}_{M''}[X_{q+1} \mid do(S)] \\
&= \mathbb{E}_M\left[\sum_{X_i \in \boldsymbol{Pa}(X_{q+1})} \theta_{X_i,X_{q+1}} X_i \middle| do(S)\right] - \mathbb{E}_{M''}\left[\sum_{X_i \in \boldsymbol{Pa}''(X_{q+1})} \theta_{X_i,X_{q+1}} X_i \middle| do(S)\right] \\
&= \sum_{X_i \in \boldsymbol{Pa}''(X_{q+1})} \theta_{X_i,X_{q+1}}(\mathbb{E}_M[X_i \mid do(S)] - \mathbb{E}_{M''}[X_i \mid do(S)]) + \\
&\qquad \sum_{X_i \in \boldsymbol{Pa}(X_{q+1}) \setminus \boldsymbol{Pa}''(X_{q+1})} \theta_{X_i,X_{q+1}} \mathbb{E}_M[X_i \mid do(S)] \\
&\leq \sum_{X_i \in \boldsymbol{Pa}(X_{q+1})} \theta_{X_i,X_{q+1}} qnr + rn \\
&\leq (q+1)nr
\end{aligned}$$

where the first equality follows the definition of linear model, the second equality is because $\theta'_{X',X} = 0$ if X' is not a true parent of X in G. The third inequality is derived by induction, and the last inequality is because $\|\theta_{X',X_{q+1}}\|_1 \leq 1$. ■

Claim 2 $|\mathbb{E}_{M'}[Y \mid do(S = \boldsymbol{1})] - \mathbb{E}_{M''}[Y \mid do(S = \boldsymbol{1})]| \leq n^3 r$.

Proof First, $\mathbb{E}_M[X_1 \mid do(S)] - \mathbb{E}_{M''}[X_1 \mid do(S)] = 0 \leq n^2 r$ Then similarly, assume $\mathbb{E}_M[X_i \mid do(S)] - \mathbb{E}_{M''}[X_i \mid do(S)] \leq qn^2 r$ for all $i \leq q$ and $X_{q+1} \notin S$. Then

$$\begin{aligned}
&\mathbb{E}_{M'}[X_{q+1} \mid do(S)] - \mathbb{E}_{M''}[X_{q+1} \mid do(S)] \\
&= \mathbb{E}_{M'}\left[\sum_{X_i \in \boldsymbol{Pa}'(X_{q+1})} \theta'_{X_i,X_{q+1}} X_i \middle| do(S)\right] - \mathbb{E}_{M''}\left[\sum_{X_i \in \boldsymbol{Pa}''(X_{q+1})} \theta_{X_i,X_{q+1}} X_i \middle| do(S)\right] \\
&= \sum_{X_i \in \boldsymbol{Pa}''(X_{q+1})} \theta'_{X_i,X_{q+1}} \mathbb{E}_{M'}[X_i \mid do(S)] - \theta_{X_i,X_{q+1}} \mathbb{E}_{M''}[X_i \mid do(S)] \\
&= \sum_{X_i \in \boldsymbol{Pa}''(X_{q+1})} (\theta'_{X_i,X_{q+1}} - \theta_{X_i,X_{q+1}}) \mathbb{E}_{M'}[X_i \mid do(S)] + \\
&\qquad \sum_{X_i \in \boldsymbol{Pa}''(X_{q+1})} \theta_{X_i,X_{q+1}} (\mathbb{E}_{M'}[X_i \mid do(S)] - \mathbb{E}_{M''}[X_i \mid do(S)]) \\
&= n^2 r + n^2 qr \\
&\leq (q+1)n^2 r.
\end{aligned}$$

where the first equality follows the definition, the second equality is because $\boldsymbol{Pa}'(X) = \boldsymbol{Pa}''(X)$ for any node X. The fourth inequality derived from induction , inequality (18) and $X_i \in [0,1]$. By induction, we complete the proof. ■

Now we prove the Lemma 8:

Proof Combining Claim 1 and Claim 2, we have

$$\mathbb{E}_M[Y \mid do(S)] - \mathbb{E}_{M'}[Y \mid do(S)] \leq n^2(n+1)r. \tag{19}$$

■

E.4. Proof of Theorem 9

Proof Denote the original model and estimated model as M and M' The initialization phase will lead to regret at most $T_0 = 16(n-1)T^{2/3}$. At Iterative phase, denote the optimal action to be $do(\boldsymbol{S}^* = \mathbf{1})$, by Lemma 6 and the guarantee of BLM-LR, with probability at least $1-(n-1)(n-2)\frac{1}{T}$

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_M[Y \mid do(\boldsymbol{S}^* = \mathbf{1})] - \mathbb{E}_M[Y \mid do(\boldsymbol{S}_t = \mathbf{1})] \\
&= \sum_{t=1}^{T}((\mathbb{E}_M[Y \mid do(\boldsymbol{S}^* = \mathbf{1})] - \mathbb{E}_{M'}[Y \mid do(\boldsymbol{S}^* = \mathbf{1})]) \\
&\quad + (\mathbb{E}_{M'}[Y \mid do(\boldsymbol{S}^* = \mathbf{1})] - \mathbb{E}_{M'}[Y \mid do(\boldsymbol{S}_t = \mathbf{1})])) \\
&\leq T_0 + \sum_{t=T_0+1}^{T} n^2(n+1)T^{-1/3} + \sum_{t=T_0+1}^{T} (\mathbb{E}_{M'}[Y \mid do(\boldsymbol{S}^* = \mathbf{1})] - \mathbb{E}_{M'}[Y \mid do(\boldsymbol{S}_t = \mathbf{1})]) \\
&\leq T_0 + n^2(n+1)T^{2/3} + cn^2\sqrt{nT\log T} \\
&= O((n^3T^{2/3} + n^3\sqrt{T})\log T) \\
&= O(n^3T^{2/3}\log T),
\end{aligned}$$

where the first inequality is derived from Lemma 8, and the second inequality is the guarantee of BLM-LR in Theorem 3 of Feng and Chen (2023).

Thus the total regret will be bounded by

$$\begin{aligned}
R(T) &\leq \frac{(n-1)(n-2)}{T} \cdot T + O((n^3T^{2/3})\log T) \\
&= O((n^3T^{2/3})\log T).
\end{aligned}$$

The first inequality is because our regret have an upper bound T. ■

E.5. Proof of Theorem 1

Proof Consider the causal bandit instances $\mathcal{T}_i$ with parallel graph ($E = \{X_i \to Y, 1 \leq i \leq n\}$.) and $\boldsymbol{A} = \{do(), do(X = x), do(\boldsymbol{X} = \boldsymbol{x})\}$ for all node X, $x \in \{0,1\}$, $\boldsymbol{x} \in \{0,1\}^n$ be all observation, atomic intervention and actions that intervene all nodes.

For $\mathcal{T}_1$, we assume X_i are independent with each other and $P(X_i = 1) = P(X_i = 0) = 0.5$. Define

$$P(Y=1) = \begin{cases} 0.5 + \Delta & \text{if } X_1 = X_2 = \cdots = X_n = 0 \\ 0.5 & \text{otherwise} \end{cases}$$

Then for $\mathcal{T}_i, 2 \le i \le 2^n$, consider the binary representation of $i-1$ as $\overline{b_1 b_2 \dots b_n}$. Then assume X_i are independent with each other and $P(X_i = 1) = 0.5$, and define

$$P(Y=1) = \begin{cases} 0.5 + \Delta & \text{if } X_1 = X_2 = \cdots = X_n = 0 \\ 0.5 + 2\Delta & \text{if } X_j = b_j \text{ for all } 1 \le j \le n \\ 0.5 & \text{otherwise} \end{cases}$$

Now in $\mathcal{T}_i$, $do\left(\boldsymbol{X} = \overline{b_1 b_2 \dots b_n}\right)$ is the best action, and other actions will lead to at least Δ regret.

Denote $T_a(t)$ for action $a \in \boldsymbol{A}$ as the number of times taking a until time t. To simplify the notation, we denote a_i as $do(\boldsymbol{X} = \boldsymbol{x})$, where $\boldsymbol{x}$ is the binary representation of $i-1$, $\{b_1, b_2, \dots, b_n\}$. Then for instances $\mathcal{T}_1$ and $\mathcal{T}_i$, we have

$$\mathbb{E}_{\mathcal{T}_1}[R(t)] \ge \mathbb{P}_{\mathcal{T}_1}(T_{a_1}(t) \le t/2)\frac{t\Delta}{2}, \quad \mathbb{E}_{\mathcal{T}_i}[R(t)] \ge \mathbb{P}_{\mathcal{T}_i}(T_{a_1}(t) > t/2)\frac{t\Delta}{2}.$$

Thus

$$\begin{aligned} \mathbb{E}_{\mathcal{T}_1}[R(t)] + \mathbb{E}_{\mathcal{T}_i}[R(t)] &> \frac{t\Delta}{2}(\mathbb{P}_{\mathcal{T}_1}(T_{a_1}(t) \le t/2) + \mathbb{P}_{\mathcal{T}_i}(T_{a_1}(t) > t/2)) \\ &\ge \frac{t\Delta}{4}\exp\left(-\mathrm{KL}(\mathbb{P}_{\mathcal{T}_1}, \mathbb{P}_{\mathcal{T}_i})\right). \end{aligned}$$

Now we need to bound $\mathrm{KL}(\mathbb{P}_{\mathcal{T}_i}, \mathbb{P}_{\mathcal{T}_1})$.

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_{\mathcal{T}_1}, \mathbb{P}_{\mathcal{T}_i}) &\le \sum_{a \in \boldsymbol{A}} \mathbb{E}_{\mathcal{T}_1}[T_a(t)] \mathrm{KL}(\mathbb{P}_{\mathcal{T}_1}(\boldsymbol{X}, Y \mid a) \| \mathbb{P}_{\mathcal{T}_i}(\boldsymbol{X}, Y \mid a)) & (20) \\ &= \sum_{a \in \boldsymbol{A}} \mathbb{E}_{\mathcal{T}_1}[T_a(t)] \mathrm{KL}(\mathbb{P}_{\mathcal{T}_1}(Y \mid a) \| \mathbb{P}_{\mathcal{T}_i}(Y \mid a)) & (21) \\ &\le \mathbb{E}_{\mathcal{T}_1}[T_{a_i}(t)] \cdot \mathrm{KL}(0.5 \| 0.5 + 2\Delta) + \sum_{a = do(X_i = x), do()} \mathbb{E}_{\mathcal{T}_1}[T_a(t)] \cdot \mathrm{KL}(0.5 \| 0.5 + \frac{\Delta}{2^{n-2}}) & (22) \\ &\le \mathbb{E}_{\mathcal{T}_1}[T_{a_i}(n)] \cdot 2\Delta^2 + t \cdot \frac{\Delta^2}{2^{2n-3}}, & (23) \end{aligned}$$

where (22) is because for $a = do(X_i = x)$ or $a = do()$, $P(Y \mid do(a)) \ge 0.5$ in $\mathcal{T}_1$, and $P(Y \mid do(a)) \le 0.5 + \frac{2\Delta}{2^{n-1}} = 0.5 + \frac{\Delta}{2^{n-2}}$ in $\mathcal{T}_i$. Now we choose

$$i = \operatorname*{argmin}_{j > 1} \mathbb{E}_{\mathcal{T}_1}[T_{a_j}(t)], \tag{24}$$

then we have

$$\mathbb{E}_{\mathcal{T}_1}[T_{a_i}(t)] \le \frac{T}{2^n - 1}. \tag{25}$$

Then by (23), choosing $\Delta = \sqrt{\frac{2^n - 1}{3t}}$, we have

$$\mathrm{KL}(\mathbb{P}_{\mathcal{T}_1}, \mathbb{P}_{\mathcal{T}_i}) \le \frac{2t\Delta^2}{2^n - 1} + \frac{t\Delta^2}{2^{2n-3}} \le t\Delta^2 \cdot \frac{3}{2^n - 1} = 1 \tag{26}$$

Thus

$$\begin{aligned}\mathbb{E}_{\mathcal{T}_1}[R(t)] + \mathbb{E}_{\mathcal{T}_i}[R(t)] &\geq \frac{t\Delta}{4}\exp\left(-\mathrm{KL}(\mathbb{P}_{\mathcal{T}_1}, \mathbb{P}_{\mathcal{T}_i})\right)\\ &\geq \frac{t\Delta}{4e}\\ &\geq \frac{\sqrt{(2^n-1)t}}{4\sqrt{3}e}\\ &\geq \frac{\sqrt{2^n t}}{8e}.\end{aligned}$$

Then $\max\{\mathbb{E}_{\mathcal{T}_1}[R(t)], \mathbb{E}_{\mathcal{T}_i}[R(t)]\} \geq \frac{\sqrt{2^n t}}{16e}$. We complete the proof when $t \geq \frac{16(2^n-1)}{3}$.

Now suppose $t \leq \frac{16(2^n-1)}{3}$, choose $\Delta = \frac{1}{4}$, then based on (23) and (25), we have

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\mathcal{T}_1}, \mathbb{P}_{\mathcal{T}_i}) &\leq \frac{t}{8(2^n-1)} + \frac{t}{2^{2n+1}}\\ &\leq \frac{2}{3} + \frac{16}{3}\cdot\frac{2^n-1}{2^{2n+1}}\\ &\leq 1.\end{aligned}$$

Then we have

$$\begin{aligned}\mathbb{E}_{\mathcal{T}_1}[R(t)] + \mathbb{E}_{\mathcal{T}_i}[R(t)] &\geq \frac{t\Delta}{4}\exp\left(-\mathrm{KL}(\mathbb{P}_{\mathcal{T}_1}, \mathbb{P}_{\mathcal{T}_i})\right)\\ &\geq \frac{t\Delta}{4e}\\ &\geq \frac{t}{16e},\end{aligned}$$

and $\max\{\mathbb{E}_{\mathcal{T}_1}[R(t)], \mathbb{E}_{\mathcal{T}_i}[R(t)]\} \geq \frac{t}{32e}$. ■

Appendix F. An Explanation of Weight Gap Assumption

The weight gap assumption states that the parameter $\theta_{\min}$ is larger than a term relative to T. In Lemma 2, the parameter $\theta_{\min}$ represents the minimum difference between $\mathbb{E}[X_j \mid do(X_i = 1)]$ and $\mathbb{E}[X_j \mid do(X_i = 0)]$, where X_i and X_j form a causal edge. Intuitively, this assumption suggests that the causal relationship represented by each edge is sufficiently significant, making it a stronger version of the causal faithfulness assumption. If the causal relationship is too weak to be observed, it may indicate the presence of intermediate factors not accounted for in practice. In such cases, one could address the issue by collecting and observing additional intermediate factors.

Furthermore, it is important to note that the weight gap assumption on $\theta^*_{\min}$ depends on T. Therefore, if the weight gap assumption is not satisfied and the intermediate factors are unobservable, the user has two options. The first is to increase the number of rounds until $\theta^*_{\min} \geq 2c_1\kappa^{-1}T^{-1/5}$. Alternatively, BGLM-OFU-Unknown can guarantee an $O(\sqrt{T})$

regret bound for $T \geq 32\left(\frac{c_1}{\kappa\theta^*_{\min}}\right)^5$. The second option is to use BLM-LR-Unknown if T cannot be increased. In this case, a theoretical regret bound of $O(T^{\frac{2}{3}})$ can be achieved.

Therefore, our results account for both scenarios, whether the weight gap assumption is satisfied or not.

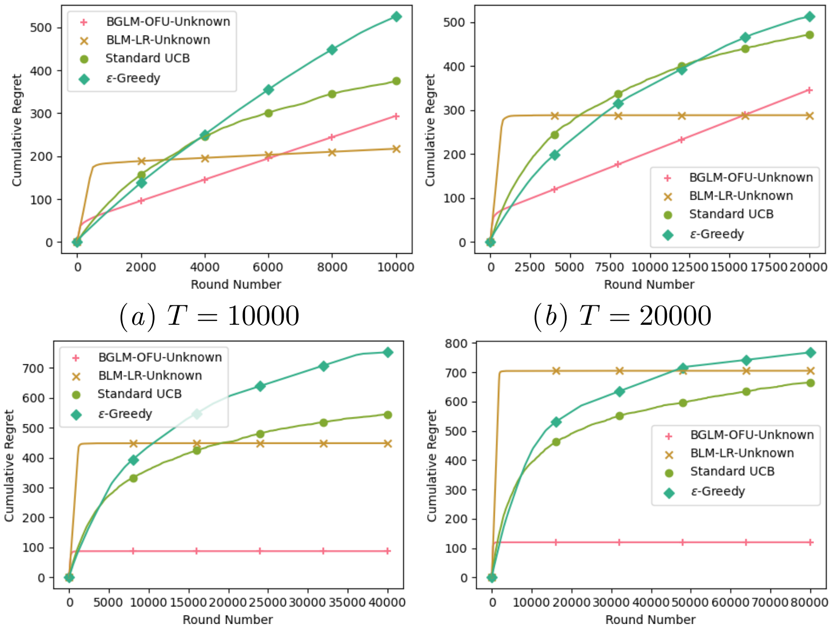
Appendix G. Experiments

G.1. Experiment Results

We conduct our experiments on a parallel BLM consisting of 7 nodes, $X_1, \dots, X_6$, and Y, with X_1 being the unique always-1 node. To simplify the analysis, we apply Algorithms 1 and 3 solely to identify the edges between $X_2, \dots, X_6$ and Y. As per the definition of our algorithms, if a node $X_i, 2 \leq i \leq 6$ is not a parent of Y, it will never be selected for interventions. We set $\mathcal{A}$ to be all interventions budgeted by 2 nodes. The parameters are set as follows:

$$\theta^*_{X_1,X_2} = \theta^*_{X_1,X_3} = 0.3, \theta^*_{X_1,X_4} = \theta^*_{X_1,X_5} = \theta^*_{X_1,X_6} = 0.2,$$
$$\theta^*_{X_2,Y} = \theta^*_{X_3,Y} = 0.3, \theta^*_{X_4,Y} = \theta^*_{X_5,Y} = \theta^*_{X_6,Y} = 0.13.$$

We run BGLM-OFU-Unknown and BLM-LR-Unknown on this BLM and compare them to the standard Upper Confidence Bound (UCB) algorithm and the ϵ-greedy algorithm ($\epsilon = 0.02$) as baseline methods. Additional implementation details can be found in the Appendix G.2. Due to computational resource constraints, we run these 4 algorithms on this BLM for $T = 10000, 20000, 40000, 80000$, each executed 50 times, and compute the average regrets as follows.



(a) $T = 10000$ (b) $T = 20000$

(c) $T = 40000$ (d) $T = 80000$

We can observe from the results that when T is small, BGLM-OFU-Unknown struggles to accurately learn the graph structure, leading to a significant regret. In contrast, BLM-LR-Unknown performs well under these conditions. However, when T is sufficiently large, BGLM-OFU-Unknown is able to consistently identify the correct graph structure, resulting in superior performance compared to all other algorithms.

G.2. Experiment Settings

Due to the limited number of rounds, we adjust ρ_t and ρ to be $\frac{1}{10}$ of our original parameter settings for BGLM-OFU-Unknown and BLM-LR-Unknown. Both algorithms have constants c_0 and c_1 set to 0.1. We employ the pair-oracle implementation as described in Appendix H.1 of Feng and Chen (2022). When BGLM degenerates to BLM, we remove the second initialization phase (line 8 of Algorithm 1) of BGLM-OFU-Unknown by setting $T_1 = T_0$. This is because the second-order derivative of a linear function is 0, making $L_{f_X}^{(2)}$ and R in BGLM-OFU-Unknown arbitrarily small; thus, the minimum eigenvalues of $M_{t,X}$'s should satisfy Lemma 14's condition after T_0 rounds. Additionally, for completeness, we provide the specific BLM used to test our algorithms in Fig. 1.

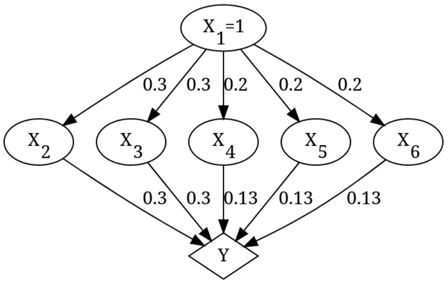


Figure 1: The BLM Employed for Evaluating Algorithms 1 and 3

For the standard UCB algorithm, we use the commonly adopted upper confidence bound $\sqrt{\frac{\ln t}{n_{i,t}}}$, where t is the current round number and $n_{i,t}$ is the number of times arm i has been played up to the t^{th} round (Slivkins et al., 2019). For the ϵ-greedy algorithm, we set $\epsilon = 0.02$, a typical implementation. We tested various settings for these two baselines, and our choices are near-optimal for BLMs. For both baselines, we treat each possible 2-node intervention set as an arm, resulting in a total of $\binom{7-2}{2} = 10$ arms. All experiments were executed using Python in a multithreaded environment on Arch Linux, utilizing 4 performance cores of an Intel Core™ i7-12700H Processor at 4.30GHz with 32GB DDR5 SDRAM. The total execution time amounts to 1687 seconds. Our Python implementation can be found in the supplementary material.

Appendix H. Pure Exploration of Causal Bandits without Graph Structure

Another performance measure for bandit algorithms is called sample complexity. In this setting, the agent aims to find an action with the maximum expected reward using as small number of rounds as possible. This setting is also called pure exploration. To be more specific, the agent is willing to find ε-optimal arm with probability at least $1-\delta$ by sampling as few rounds as possible for fixed parameter ε and δ. For pure exploration, we consider the general binary causal model with only null and atomic interventions, and study the gap-dependent bounds, meaning that the sample complexity depends on the reward gap between the optimal and suboptimal actions. Moreover, let a^* be one of the optimal actions. For each action $a = do(X_i = x)$, define $\mu_a = \mathbb{E}[Y \mid a]$ and the gap for action a to be

$$\Delta_a = \begin{cases} \mu_{a^*} - \max_{a \in \boldsymbol{A} \setminus \{a^*\}}\{\mu_a\}, & a = a^*; \\ \mu_{a^*} - \mu_a, & a \neq a^*. \end{cases} \tag{27}$$

Here, Δ_a can be 0.

According to the causal discovery literature (Pearl, 2009b), by passive observations alone one can obtain an essential graph of the causal graph, with some edge directions unidentified. We assume that the essential graph is known but the exact graph structure is unknown, which is also considered by Lu et al. (2021), with additional assumptions on the graph.

One naive solution for this problem is to first identify the graph structure and then to performed the pure exploration algorithm of causal bandits with known graph (Xiong and Chen, 2023). Define $c_e = |P(X \mid do(X' = 1)) - P(X' \mid do(X = 0))|$ for each edge $e = (X, X')$ and $c_X = \min_{e: X \to X'} \frac{1}{c_e^2}$. Then this naive solution admits a sample complexity about

$$\tilde{O}\left(\sum_{a \in S} \frac{1}{\max\{\Delta_a, \varepsilon/2\}^2} + \sum_{x \in X} \frac{1}{c_X^2}\right), \tag{28}$$

where S is a particular set defined following the previous work (Xiong and Chen, 2023) and the definition is provided in Appendix I. The first term is the sample complexity in Xiong and Chen (2023), while the second term is the cost for identifying the directions of all edges in the essential graph.

This naive solution separates the causal discovery phase and learning phase, so it cannot discover the directions adaptively. In Appendix I, we propose an adaptive algorithm to discover the edges' directions and learn the reward distribution in parallel, which can provide a lower sample complexity for some cases.

However, when the Δ_a and c_X is small, both the naive algorithm and our algorithms provided in Appendix I suffers $\Omega(\frac{n}{\varepsilon^2}\log(1/\delta))$ sample complexity. We claim that pure exploration for the general binary causal model is intrinsically hard due to unknown graph structure. To show this, we state a negative result for pure exploration of causal bandits on unknown graph structure with atomic intervention. It states that even if we have all observation distribution $P(\boldsymbol{X}, Y)$ as prior knowledge, we still cannot achieve better sample complexity result than the result in the classical pure exploration problem for the multi-armed bandit $O(\frac{n}{\varepsilon^2}\log(1/\delta))$.

Theorem 21 (Lower bound) *Consider causal bandits with only essential graph and atomic intervention, for any algorithm which can output ε-optimal action with probability at least $1-\delta$, there is a bandit instance with expected sample complexity $\Omega(\frac{n}{\varepsilon^2}\log(1/\delta))$ even if we have all observational distribution $P(\boldsymbol{X}, Y)$.*

Algorithm 7: Causal-PE-unknown$(G, A, \varepsilon, \delta)$

```
Initialize $t = 1$, $T_a(0) = 0, \hat{\mu}_a = 0$ for all arms $a \in A$, $\mathcal{A}_{known} = \emptyset$
for $t = 1, 2, \dots,$ do
    $a_h^{t-1} = \text{argmax}_{a \in A} \hat{\mu}_a^{t-1}$
    $a_l^{t-1} = \text{argmax}_{a \in A \setminus a_h^{t-1}} (U_a^{t-1})$
    if $U_{a_l^{t-1}} \le L_{a_h^{t-1}} + \varepsilon$ then
        Return $a_h^{t-1}$
    end if
    Perform $do()$ operation and observe $\boldsymbol{X}_t$ and $Y_t$. For $a = do()$, $T_a(t) = T_a(t-1) + 1$, $D_a(t) = D_a(t-1), r_{a,\emptyset}(t) = \frac{1}{T_a(t)} \sum_{j=1}^{t} Y_j,\ p_{a,\emptyset}(t) = 1$.
    for $a = do(X = x) \in \mathcal{A}_{known}$ do
        $T_{a,\boldsymbol{z}}(t) = T_{a,\boldsymbol{z}}(t-1) + \mathbb{I}\{X_t = x, \boldsymbol{P} = \boldsymbol{z}\}$, $T_a(t) = \min_{\boldsymbol{z}}\{T_{a,\boldsymbol{z}}(t)\}$, where $\boldsymbol{P} = \boldsymbol{Pa}(X)$. $D_a(t) = D_a(t-1)$.
        Update $r_{a,\boldsymbol{z}}(t) = \frac{1}{T_{a,\boldsymbol{z}}(t)} \sum_{j=1}^{t} \mathbb{I}\{X_j = x, \boldsymbol{P}_j = \boldsymbol{z}\} Y_j$.
        Update $p_{a,\boldsymbol{z}}(t) = \frac{1}{t} \sum_{j=1}^{t} \mathbb{I}\{\boldsymbol{P}_j = \boldsymbol{z}\}$.
        Estimate $\hat{\mu}_{O,a}(t) = \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) p_{a,\boldsymbol{z}}(t)$ and calculate $[L_{O,a}^t, U_{O,a}^t]$by (34) and (35).
    end for
    RECOVER-EDGE($a_h^{t-1}$).
    RECOVER-EDGE($a_l^{t-1}$).
    Update empirical mean $\hat{\mu}_{I,a}(t)$ using interventional dataand interventional confidence bound $[L_{I,a}^t, U_{I,a}^t]$
    Update confidence bound $[L_a^t, U_a^t]$ by (33), $\hat{\mu}_a = (L_a^t + U_a^t)/2$, for each arm $a$.
end for
```

Note that if we know distribution $P(\boldsymbol{X}, Y)$ and the exact graph structure, we can compute each intervention $P(Y \mid do(X = x))$ by do-calculus because the absence of hidden variables. So Theorem 21 shows the intrinsic hardness provided by unknown graph structure. The detailed proof can be found in Appendix I.

Appendix I. General Causal Bandits without Graph Structure

In this section, we only consider the atomic intervention, and provide an algorithm to solve causal bandits with the graph skeleton on binary model. We only consider the atomic intervention setting. An atomic intervention is $do(X = x)$, where X is a node of graph G and $x \in \{0, 1\}$.

I.1. General Causal Bandit Algorithms

We first provide the positive results, which provides an algorithm to improve the sample complexity comparing to applying the multi-armed bandit approach directly.

At each iteration we try to recover the edges' direction in parallel using sub-procedure "RECOVER-EDGE(a)" for $a \in A$. For action $a = do(X = x)$, this sub-procedure first performs two interventions $do(X = 1)$ and $do(X = 0)$, then chooses an undirected edge (X, X') corresponding to X (if exists), and then perform $do(X' = 1)$, $do(X' = 0)$. The

Algorithm 8: RECOVER-EDGE(a)

```
if $a = do()$ then
    Return.
else
    Assume $a = do(X = x)$. Sample action $do(X = 1), do(X = 0)$.
    $D_{a'}(t) = D_{a'}(t) + 1$ for $a' = do(X = 1)$ and $a' = do(X = 0)$.
    Estimate $P(X' = 1 \mid do(X = 1))$ and $P(X' = 1 \mid do(X = 0))$ using interventional data for neighbor $X'$, where the direction of $(X', X)$ is unknown.
    Update the confidence bound $[L_{X'|do(X=1)}, U_{X'|do(X=1)}]$ and $[L_{X'|do(X=0)}, U_{X'|do(X=0)}]$ by (31).
    if $[L_{X'|do(X=1)}, U_{X'|do(X=1)}] \cap [L_{X'|do(X=0)}, U_{X'|do(X=0)}] = \emptyset$ then
        recover $X \to X_i$.
    end if
    if $\exists X'$ such that $(X', X)$ is unknown then
        Choose one such $X'$ and perform $do(X' = 1)$ and $do(X' = 0)$.
        Estimate $P(X = 1 \mid do(X' = 0))$ and $P(X = 1 \mid do(X' = 1))$ using interventional data.
        Update the confidence bound $[L_{X|do(X'=1)}, U_{X|do(X'=1)}]$ and $[L_{X|do(X'=0)}, U_{X|do(X'=0)}]$ by (31).
        if $[L_{X|do(X'=1)}, U_{X|do(X'=1)}] \cap [L_{X|do(X'=0)}, U_{X|do(X'=0)}] = \emptyset$ then
            recover $X \to X_i$.
        end if
        $D_{a'}(t) = D_{a'}(t) + 1$ for $a' = do(X' = 1)$ and $a' = do(X' = 0)$.
    end if
end if
```

goal of these operations is to estimate the difference between $P(X = 1 \mid do(X' = 0))$ and $P(X = 1 \mid do(X' = 1))$, and also the difference between $P(X' = 1 \mid do(X = 0))$, $P(X' = 1 \mid do(X = 0))$. which decides whether $X' \to X$ or $X \to X'$. By this sub-procedure in parallel, the algorithm estimate the model and recover the edges' direction simultaneously and adaptively. To measure the difficulty for identified the direction of edges, for $e : X \to X'$ we define

$$c_e = P(X' = 1 \mid do(X = 1)) - P(X' = 1 \mid do(X = 0)) \tag{29}$$

$$c_a = c_X = \min_{e:X \to X'} c_e. \tag{30}$$

c_e measure the difficulty for distinguishing the direction for an edge, and $c_a = c_X$ represents the hardness for discovering all directions corresponding to X and its childs.

The main Algorithm 7 is followed from Xiong and Chen (2023). During the algorithm, we add "RECOVER-EDGE" sub-procedure to identify the directions of the unknown edges. This sub-procedure first perform intervention $do(X = 0)$ and $do(X = 1)$ on the node X. Then if there is an edge (X', X) which direction has not been identified, it chooses one such edge and perform $do(X' = 1)$ and $do(X' = 0)$. Then it constructs the confidence bound for all $P(X' = 1 \mid do(X = 1))$, $P(X' = 1 \mid do(X = 0))$, $P(X = 1 \mid do(X' = 1))$ and $P(X = 1 \mid do(X' = 0))$ based on Hoeffding's concentration bound. In fact, assume there are $D_a(t)$ samples for $a = do(X' = x), x \in \{0, 1\}$ until round t, then the confidence bound for X conditioning on $do(X' = x)$ is defined by

$$\begin{aligned}[L_{X|do(X'=x)}, U_{X|do(X'=x)}] = \Bigg[&\hat{P}(X = 1 \mid do(X' = x)) - \sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}, \\ &\hat{P}(X = 1 \mid do(X' = x)) + \sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}\Bigg],\end{aligned} \tag{31}$$

where n is the number of nodes, and $\hat{P}(X = 1 \mid do(X' = x))$ are the empirical mean of $P(X = 1 \mid do(X' = x))$ using all these $D_a(t)$ samples for $do(X' = x)$. Other confidence bounds define in this way similarly.

Moreover, at iteration t, Line 4-Line 6 first choose two actions a_h^{t-1} and a_l^{t-1} through LUCB1 algorithm. Then, we use $\mathcal{A}_{known}$ to represent all nodes actions $do(X = x)$ where all the edges corresponding to X are identified. In fact, if all the edges corresponding to X are identified, we can find the true parent set $\boldsymbol{Pa}(X)$. Then we can use do-calculus to estimate the causal effect:

$$\mathbb{E}[Y \mid do(X = x)] = \sum_z P(Y \mid X = x, Z = z)P(Z = z). \tag{32}$$

Line 9-14 enmurates all these actions, and calculate corresponding confidence bound. The confidence bound is calculated by

$$[L_a^t, U_a^t] = [L_{O,a}^t, U_{O,a}^t] \cap [L_{I,a}^t, U_{I,a}^t], \tag{33}$$

where the first term $[L_{O,a}^t, U_{O,a}^t] = (-\infty, \infty)$ for $a = do(X = x)$ if the parents of X are not sure at time t. In fact, if we do not discover all the edges corresponding to X, we cannot estimate the causal effect $\mathbb{E}[Y \mid do(X = x)]$ using do-calculus. For nodes which parent set is identified, we calculate

$$\begin{aligned}[L_{O,a}^t, U_{O,a}^t] &= [\hat{\mu}_{O,a}(t) - \beta_{O,a}(t), \hat{\mu}_{O,a}(t) + \beta_{O,a}(t)], \\ [L_{I,a}^t, U_{I,a}^t] &= [\hat{\mu}_{I,a}(t) - \beta_{I,a}(t), \hat{\mu}_{I,a}(t) + \beta_{I,a}(t)]\end{aligned} \tag{34}$$

The term $\hat{\mu}_{O,a}$ is calculated by estimating all terms at the right side of (32) empirically, and confidence radius is given by

$$\beta_{O,a}(t) = \sqrt{\frac{12}{T_a(t)} \log \frac{16n^2 Z_a t^3}{\delta}}, \beta_{I,a}(t) = 2\sqrt{\frac{1}{D_a(t)} \log \frac{2n \log(2t)}{\delta}} \tag{35}$$

Similar to Xiong and Chen (2023), we can prove it is a valid confidence radius, which means that the true effect $\mu_{O,a}$ will fall into the confidence bound $[L_{O,a}^t, U_{O,a}^t]$ with a high probability.

Line 15-16 try to recover the edge for action chosen by LUCB1 algorithm. At the end of this iteration, the algorithm updates all parameters and confidence bounds.

To represent the complexity result, we first provide the definition of gap-dependent threshold in Xiong and Chen (2023): For $a = do(X = x)$ and one possible configuration of the parent $\boldsymbol{z} \in \{0,1\}^{|\boldsymbol{Pa}(X)|}$, define $q_{a,\boldsymbol{z}} = P(X = x, \boldsymbol{Pa}(X) = \boldsymbol{z})$ and $q_a = \min_{\boldsymbol{z}}\{q_{a,\boldsymbol{z}}\}$. Then sort the arm set as $q_{a_1} \cdot \max\{\Delta_{a_1}, \varepsilon/2\}^2 \leq q_{a_2} \cdot \max\{\Delta_{a_2}, \varepsilon/2\}^2 \leq \ldots \leq q_{a_{|\boldsymbol{A}|}} \cdot \max\{\Delta_{a_{|\boldsymbol{A}|}}, \varepsilon/2\}^2$. Recall that $\Delta_a = \mu^* - \mu_a$ is the reward gap between the optimal reward and the reward of action a. Then H_r is defined by

$$H_r = \sum_{i=1}^{r} \frac{1}{\max\{\Delta_{a_i}, \varepsilon/2\}^2}. \tag{36}$$

Definition 22 (Gap-dependent observation threshold (Xiong and Chen, 2023)) *For a given causal graph G and its associated q_a's and Δ_a's, the* gap-dependent observation threshold *$m_{\varepsilon,\Delta}$ is defined as:*

$$m_{\varepsilon,\Delta} = \min\left\{\tau : \left|\left\{a \in \boldsymbol{A} \middle| q_a \max\left\{\Delta_a, \varepsilon/2\right\}^2 < \frac{1}{H_\tau}\right\}\right| \leq \tau\right\}.$$

Denote action set $S = \{a \in \boldsymbol{A} : q_a \max\{\Delta_a, \varepsilon/2\}^2 < \frac{1}{H_{m_{\varepsilon,\Delta}}}\}$ are all actions which q_a is relatively small, then $|S| \leq m_{\varepsilon,\Delta}$. Intuitively, action a with smaller q_a are harder to be estimated by observation: If we assume $q_a = q_{a,\boldsymbol{z}}$ for a fixed vector $\boldsymbol{z}$, then $P(X = x, \boldsymbol{Pa}(X) = \boldsymbol{z})$ is hard to observe and estimate by empirical estimation. Thus S contains all actions that are relatively hard to observe, so it is more efficient to estimate μ_a by intervention for $a \in S$. Based on this definition, we can provide the final sample complexity result:

Theorem 23 *Denote $H = \sum_{a\in S} \frac{1}{\max\{\Delta_a,\varepsilon/2\}^2} + \sum_{a\notin S} \min\{\frac{1}{\max\{\Delta_a,\varepsilon/2\}^2}, \frac{1}{c_a^2} + \sum_{e:X'\to X} \frac{1}{c_e^2}\}$. With probability $1 - 4\delta$, Algorithm 7 will return a ε-optimal arm with sample complexity bound at most*

$$T = O\left(H \log\left(\frac{nZH}{\delta}\right)\right),$$

where c_e, c_a is defined in (29) and (30).

The result can be explained in an intuitive way. The first term of H is the summation of all actions in S. As we discussed above, it is more efficient to estimate the μ_a with intervention for $a \in S$. Thus, this summation can be regarded as the sample complexity applying multi-armed bandit algorithm (e.g. LUCB1) directly. The second term is to estimate the actions by observation. For each action $a = do(X = x)$ with larger q_a, we can first identify the edge's direction corresponding to the node X, and then using do-calculus to estimate the reward. The term $\frac{1}{c_a^2} + \sum_{e:X'\to X} \frac{1}{c_e^2}$ represents the complextity to identify the directions, and the complexity for using do-calculus can be contained in the first term $\sum_{a\in S} \frac{1}{\max\{\Delta_a,\varepsilon/2\}^2}$ because of the definition of gap-dependent observation threshold. Also, the term $\min\{\frac{1}{\max\{\Delta_a,\varepsilon/2\}^2}, \frac{1}{c_a^2} + \sum_{e:X'\to X} \frac{1}{c_e^2}\}$ is because when we are discovering the edges' direction, if the reward can be estimated by intervention accurately, we turn to use interventional estimation and give up the causal discovery for this node. The detailed proof can be found in the Section I.2.

Even if these two mechanisms can reduce the sample complexity, at the worst case the complexity also degenerates to $O(n/\varepsilon^2)$, which is equal to the complexity for multi-armed bandit. We provide a lower bound to show that this problem cannot be avoided.

Theorem 24 (Lower bound) *Consider causal bandits with only essential graph and atomic intervention, for any $(\varepsilon, \delta)-PAC$ algorithm, there is a bandit instance with expected sample complexity $\Omega(\frac{n}{\varepsilon^2}\log(1/\delta))$ even if we have all observational distribution $P(\boldsymbol{X}, Y)$.*

Theorem 24 states that even if we receive all observational distribution, which shows the intrinsic hardness for unknown graph. Indeed, the proof of lower bound shows that the unknown direction will lead to different interventional effects even when the observational distribution are the same, leading to a unavoidable hardness.

I.2. Proof of Theorem 23

First, fixed an action $a = do(X_i = x)$, $\boldsymbol{z} \in \{0,1\}^{|\boldsymbol{Pa}(X)|}$, then $T_{a,\boldsymbol{z}}(t) = \sum_{j=1}^t \mathbb{I}\{X_{j,i} = x, \boldsymbol{Pa}(X_i)_j = \boldsymbol{z}\}$ and the empirical mean $\hat{q}_{a,\boldsymbol{z}}(t) = T_{a,\boldsymbol{z}}(t)/t$. Then denote $2^{|\boldsymbol{Pa}(X)|} = Z_a$, if $q_{a,\boldsymbol{z}}(t) \geq \frac{6}{t}\log(2nZ_a/\delta)$, with probability at least $1 - \frac{\delta}{2nZ_a}$, we can have

$$|\hat{q}_{a,\boldsymbol{z}}(t) - q_{a,\boldsymbol{z}}(t)| < \sqrt{\frac{6q_{a,\boldsymbol{z}}(t)}{t}\log\left(\frac{2nZ_a}{\delta}\right)}$$

Hence

$$\hat{q}_a(t) = \min_{\boldsymbol{z}}\{\hat{q}_{a,\boldsymbol{z}}(t)\} \leq \min_{\boldsymbol{z}}\{q_{a,\boldsymbol{z}} + \sqrt{\frac{6q_{a,\boldsymbol{z}}}{t}\log\frac{2nZ_a}{\delta}}\} = q_a + \sqrt{\frac{6q_a}{t}\log\frac{2nZ_a}{\delta}}. \tag{37}$$

When $q_a \geq \frac{3}{t}\log\frac{2nZ_a}{\delta}$, $f(x) = x - \sqrt{\frac{6x}{t}\log\frac{2nZ_a}{\delta}}$ is a increasing function.

$$\hat{q}_a(t) \geq \min_{\boldsymbol{z}}\{q_{a,\boldsymbol{z}} - \sqrt{\frac{6q_{a,\boldsymbol{z}}}{t}\log\frac{2nZ_a}{\delta}}\} = q_a - \sqrt{\frac{6q_a}{t}\log\frac{2nZ_a}{\delta}}. \tag{38}$$

So define the event as

$$\mathcal{E}_1(t) = \left\{\forall a \in \boldsymbol{A} \text{ with } t \geq \frac{6}{q_a}\log\left(\frac{2nZ_a}{\delta}\right), |\hat{q}_a(t) - q_a| \leq \sqrt{\frac{6q_a}{t}\log\left(\frac{2nZ_a}{\delta}\right)}\right\}$$

then $\Pr\{\mathcal{E}_1^c(t)\} \leq \delta$, where $\mathcal{E}^c$ means the complement of the event $\mathcal{E}$.

Now we consider the concentration bound. First, by classical anytime confidence bound, with probability at least $1 - \frac{\delta}{2n}$, for any time $D_a(t) \geq 1$

$$|\hat{\mu}_{I,a}(t) - \mu_{I,a}| < 2\sqrt{\frac{1}{D_a(t)}\log\left(\frac{2n\log(2D_a(t))}{\delta}\right)} \leq 2\sqrt{\frac{1}{D_a(t)}\log\left(\frac{2n\log(2t)}{\delta}\right)}$$

Thus define the event as

$$\mathcal{E}_2 = \left\{\forall t, a, |\hat{\mu}_{I,a}(t) - \mu_{I,a}| < 2\sqrt{\frac{1}{D_a(t)}\log\left(\frac{2n\log(2t)}{\delta}\right)}\right\},$$

then $\Pr\{\mathcal{E}_2^c\} \leq \delta$.

Consider the observational confidence bound. First, if $a \notin \mathcal{A}_{known}$, $[L^t_{O,a}, U^t_{O,a}] = (-\infty, \infty)$ and then the $\hat{\mu}_{O,a}(t) \in [L^t_{O,a}, U^t_{O,a}]$. Now we consider that if $a = do(X = x) \in \mathcal{A}_{known}$ and the parent of X is $\boldsymbol{P}$. By Hoeffding's inequality, with probability at least $1 - \delta/16n^2 Z_a t^3$, for $a = do(X = x)$,

$$|r_{a,\boldsymbol{z}}(t) - P(Y = 1 \mid X = x, \boldsymbol{P} = \boldsymbol{z})| > \sqrt{\frac{1}{2T_{a,\boldsymbol{z}}(t)} \log \frac{16n^2 Z_a t^3}{\delta}} \tag{39}$$

Also, by Chernoff's inequality, since $q_a \le P(\boldsymbol{P} = \boldsymbol{z})$ for all $\boldsymbol{z} \in \{0,1\}^{|\boldsymbol{P}|}$, when $t \ge \frac{6}{q_a} \log\left(\frac{16n^2 Z_a t^3}{\delta}\right)$ with probability at least $1 - \delta/16n^2 Z_a t^3$ we will have

$$|p_{a,\boldsymbol{z}}(t) - P(\boldsymbol{P} = \boldsymbol{z})| > \sqrt{\frac{6P(\boldsymbol{P} = z)}{t} \log \frac{16n^2 Z_a t^3}{\delta}}, \tag{40}$$

then

$$\begin{aligned}
\hat{\mu}_{O,a} &= \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot p_{a,\boldsymbol{z}}(t) \\
&\le \sum_{\boldsymbol{z}} P(Y = 1 \mid X = x, \boldsymbol{P} = \boldsymbol{z}) p_{a,\boldsymbol{z}}(t) + \sum_{\boldsymbol{z}} p_{a,\boldsymbol{z}}(t) \sqrt{\frac{1}{2T_{a,\boldsymbol{z}}(t)} \log \frac{16n^2 Z_a t^3}{\delta}} \\
&\le \sum_{\boldsymbol{z}} P(Y = 1 \mid X = x, \boldsymbol{P} = \boldsymbol{z}) p_{a,\boldsymbol{z}}(t) + \sqrt{\frac{1}{2T_a(t)} \log \frac{16n^2 Z_a t^3}{\delta}} \\
&\le \sum_{\boldsymbol{z}} P(Y = 1 \mid X = x, \boldsymbol{P} = \boldsymbol{z}) P(\boldsymbol{P} = \boldsymbol{z}) + \sum_{\boldsymbol{z}} \sqrt{\frac{6P(\boldsymbol{P} = \boldsymbol{z})}{t} \log \frac{16n^2 Z_a t^3}{\delta}} + \\
&\sqrt{\frac{1}{2T_a(t)} \log \frac{16n^2 Z_a t^3}{\delta}} \\
&\le \mu_a + \sqrt{\frac{6Z}{t} \log \frac{16n^2 Z_a t^3}{\delta}} + \sqrt{\frac{1}{2T_a(t)} \log \frac{16n^2 Z_a t^3}{\delta}} \\
&\le \mu_a + \sqrt{\frac{6}{T_a(t)} \log \frac{16n^2 Z_a t^3}{\delta}} + \sqrt{\frac{1}{2T_a(t)} \log \frac{16n^2 Z_a t^3}{\delta}}, \\
&= \mu_a + \sqrt{\frac{8}{T_a(t)} \log \frac{16n^2 Z_a t^3}{\delta}}.
\end{aligned}$$

Also, if $t \le \frac{6}{q_a} \log \frac{16n^2 Z_a t^3}{\delta}$, first by Chernoff inequality, set $Q = \frac{6}{q_a} \log \frac{16n^2 Z_a t^3}{\delta}$, then with probability at least $1 - \delta/16n^2 Z_a t^3$, we have

$$\hat{q}_a(Q) \le 2q_a. \tag{41}$$

by $\mathcal{E}_1(Q)$.

$$T_a(t) \le T_a(Q) \le \hat{q}_a(Q) \cdot Q \le 2q_a \cdot Q = \frac{12}{q_a} \log \frac{16n^2 Z_a t^3}{\delta}.$$

Then $\sqrt{\frac{12}{T_a(t)}\log\frac{16n^2Z_at^3}{\delta}} \geq 1$ and the inequality

$$|\hat{\mu}_{O,a}(t) - \mu_{O,a}| \leq \sqrt{\frac{12}{T_a(t)}\log\frac{16n^2Z_at^3}{\delta}}$$

also holds. Thus we define the event

$$\mathcal{E}_3 = \left\{\forall a, t, |\hat{\mu}_{O,a}(t) - \mu_{O,a}| \leq \sqrt{\frac{12}{T_a(t)}\log\frac{16n^2Z_at^3}{\delta}}\right\}$$

then by taking the union bound of (39), (40) and (41),

$$\begin{aligned}\Pr\{\mathcal{E}_3^c\} &\leq \sum_{t=1}^{\infty}\sum_{a\in\boldsymbol{A}}\sum_{\boldsymbol{z}} 3\cdot\frac{\delta}{16n^2Z_at^3}\\ &\leq \sum_{t=1}^{\infty}\frac{\delta}{4t^3}\\ &\leq \delta.\end{aligned}$$

Now we consider how to bound our sample complexity based on events $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$. First, we provide the following lemma in Xiong and Chen (2023):

Lemma 25 (Lemma 6 in Xiong and Chen (2023)) *Under the event $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$, at round t, if we have*

$$\beta_{a_h^t}(t) \leq \frac{\max\{\Delta_{a_h^t}, \varepsilon/2\}}{4}, \beta_{a_l^t}(t) \leq \frac{\max\{\Delta_{a_l^t}, \varepsilon/2\}}{4},$$

where a_h^t, a_l^t are the actions performed by algorithm at round t. then the algorithm will stop at round $t+1$.

Now assume the algorithm does not terminate at $T_1 = 192H\log(nZT_1^3/\delta)$, where $Z = \max_a Z_a$. For $a \in S$, $D_a(t)$. Note that $H \geq H_{m_{\varepsilon,\Delta}}$. Thus at round T_1, for action a with $q_a \geq \frac{1}{H_{m_{\varepsilon,\Delta}}\cdot\max\{\Delta_a,\varepsilon/2\}^2} \geq \frac{192}{T_1}\log\frac{16nZ_aT_1^3}{\delta}$, if $a \in \mathcal{A}_{known}$, then under event $\mathcal{E}_1(T_1)$, we have

$$\hat{q}_a(T_1) \geq q_a - \sqrt{\frac{6q_a}{T_1}\log\frac{16nZ_aT_1^3}{\delta}} \geq \frac{q_a}{2}.$$

Then

$$\beta_a(T_1) \leq \beta_{O,a}(T_1) = \sqrt{\frac{12}{T_a(T_1)}\log\frac{16n^2Z_at^3}{\delta}} \leq \sqrt{\frac{12q_a}{2T_1}} \leq \frac{\max\{\Delta_a, \varepsilon/2\}^2}{4}.$$

Now we prove that if $D_a(t)$ is large for some a, then $a \in \mathcal{A}_{known}$.

Lemma 26 *With probability at least $1-\delta$, denote $C_a = \frac{1}{c_a^2} + \sum_{e:X'\to X}\frac{1}{c_e^2}$. If $D_a(t) \geq 32C_a\log(4n^2t^2/\delta)$, $a \in \mathcal{A}_{known}$.*

Proof If $D_a(t) \geq 8C_a \log t$, we have called sub-procedure RECOVER-EDGE(a) for $D_a(t)$ times. Then, for each edge $e : X \to X'$, we will perform intervention $do(X = 1)$, $do(X = 0)$ for at least $D_a(t)$ times and observe the empirical difference $|\hat{P}(X' \mid do(X = 1)) - \hat{P}(X' \mid do(X = 0))|$. By Hoeffding's inequality and union bound on all time t and the $\binom{n-1}{2}$ ordered-pair (X', X), with probability at least $1 - \delta$, for all $t \in [T]$ and all X', X we have

$$|\hat{P}(X' \mid do(X = 1)) - P(X' \mid do(X = 1))| \leq \sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}$$

$$|\hat{P}(X' \mid do(X = 0) - P(X' \mid do(X = 0))| \leq \sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}$$

Then for the confidence bounds

$$\begin{aligned}
&[L_{X'|do(X=1)}, U_{X'|do(X=1)}] \\
&= \left[\hat{P}(X' \mid do(X = 1)) - \sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}, \hat{P}(X' \mid do(X = 1)) + \sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}\right], \\
&[L_{X'|do(X=0)}, U_{X'|do(X=0)}] \\
&= \left[\hat{P}(X' \mid do(X = 0)) - \sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}, \hat{P}(X' \mid do(X = 0)) + \sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}\right],
\end{aligned}$$

the intersection

$$[L_{X'|do(X=1)}, U_{X'|do(X=1)}] \cap [L_{X'|do(X=0)}, U_{X'|do(X=0)}] = \emptyset,$$

since

$$\begin{aligned}
&|\hat{P}(X' \mid do(X = 1) - \hat{P}(X' \mid do(X = 0))| \\
\geq &|P(X' \mid do(X = 1) - P(X' \mid do(X = 0))| - |P(X' \mid do(X = 1) - \hat{P}(X' \mid do(X = 1))| \\
&\quad - |P(X' \mid do(X = 0) - \hat{P}(X' \mid do(X = 0))| \\
\geq &c_a - 2\sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}} \\
\geq &2\sqrt{\frac{2}{D_a(t)} \log \frac{4n^2t^2}{\delta}}.
\end{aligned}$$

where we use $D_a(t) \geq \frac{1}{c_a^2} \log \frac{4n^2t^2}{\delta}$. Then the edge's direction will be identified correctly.

Consider the edge $e : X' \to X$, then if we sample $do(X' = 1)$ and $do(X' = 0)$ for $\frac{1}{c_e^2} \log \frac{4n^2t^2}{\delta}$ times within sub-procedures RECOVER-EDGE(a), similarly we will identify the edge $X' \to X$. Then because the RECOVER-EDGE(a) will perform intervention $do(X' = 0)$ and $do(X' = 1)$ for the X' that the direction of (X', X) has not been discovered each time, after $\sum_{e:X' \to X} \frac{1}{c_e^2} \log \frac{4n^2t^2}{\delta}$. ■

Then we define

$$\mathcal{E}_4 = \{\text{Lemma } 26 \text{ holds}\}$$

Then $\Pr\{\mathcal{E}_4^c\} \leq \delta$. Also, under the event $\mathcal{E}_2$, the following lemma shows that if $D_a(t)$ is really large, we can estimate the μ_a accurately.

Lemma 27 *Under event $\mathcal{E}_2$, if $D_a(t) \geq \frac{64}{\max\{\Delta_a, \varepsilon/2\}^2} \log \frac{16n^2 Z_a t^3}{\delta}$, then*

$$\beta_a(T_1) \leq \frac{\max\{\Delta_a, \varepsilon/2\}^2}{4}.$$

Proof In fact,

$$\beta_a(t) \leq \beta_{I,a}(t) = 2\sqrt{\frac{1}{D_a(t)} \log\left(\frac{2n\log(2t)}{\delta}\right)} \leq 2\sqrt{\frac{1}{D_a(t)} \log\frac{16n^2 Z_a t^3}{\delta}} \leq \frac{\max\{\Delta_a, \varepsilon/2\}^2}{4}.$$

■

Now we turn to our main result. From the Lemma 25, at least one arm a with $\beta_a(t) \geq \frac{\max\{\Delta_a, \varepsilon/2\}}{4}$ will be performed an intervention at each round $t \geq T_1$. Under the event $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ and $\mathcal{E}_4$, these interventions will only performed in two types of action a:

- $q_a \leq \frac{1}{H_{m_{\varepsilon,\Delta}} \cdot \max\{\Delta_a, \varepsilon/2\}^2}$ and $D_a(t) \leq \frac{64}{\max\{\Delta_a, \varepsilon/2\}^2} \log \frac{16n^2 Z_a t^3}{\delta}$.
- $D_a(t) \leq \min\{MC_a \log(t), \frac{64}{\max\{\Delta_a, \varepsilon/2\}^2} \log \frac{16n^2 Z_a t^3}{\delta}\}$.

Note that $q_a \leq \frac{1}{H_{m_{\varepsilon,\Delta}} \cdot \max\{\Delta_a, \varepsilon/2\}^2}$ implies that $a \in S$, then after at most T_2 rounds, where

$$\begin{aligned}T_2 &= 64\left(\sum_{a \in S} \frac{1}{\max\{\Delta_a, \varepsilon/2\}^2} + \sum_{a \notin S} \min\left\{\frac{1}{\max\{\Delta_a, \varepsilon/2\}^2}, \frac{1}{c_a^2} + \sum_{e: X' \to X} \frac{1}{c_e^2}\right\}\right) \log \frac{16n^2 Z T_2^3}{\delta} \\ &= 64H \log \frac{16n^2 Z T_2^3}{\delta}\end{aligned}$$

the algorithm should terminates. The fist term is the summation of all actions in S, and the second term is for the second type of actions, where

$$D_a(t) \leq \min\{MC_a \log(t), \frac{64}{\max\{\Delta_a, \varepsilon/2\}^2} \log \frac{16n^2 Z_a t^3}{\delta}\}.$$

Denote $T = T_1 + T_2$, then

$$T = T_1 + T_2 \leq 256H \log \frac{16n^2 Z T^3}{\delta} \leq 768H \log \frac{16nZT}{\delta}$$

Then by the Lemma 28, with probability at least $1 - 4\delta$, the sample complexity has the upper bound

$$T = O\left(H \log\left(\frac{nZH}{\delta}\right)\right)$$

Replace δ to $\delta/4$, we derive the sample complexity in the Theorem 23. The correctness of algorithm can be derived by LUCB1 algorithm. We provide a short argument here. Because the stopping rule is $\hat{\mu}^t_{a^t_l} + \beta_{a^t_l}(t) \leq \hat{\mu}^t_{a^t_h} - \beta_{a^t_h}(t) + \varepsilon$, if $a^* \neq a^t_h$, we have

$$\mu_{a^t_h} + \varepsilon \geq \hat{\mu}_{a^t_h} - \beta_{a^t_h}(t) + \varepsilon \geq \hat{\mu}_{a^t_l} + \beta_{a^t_l}(t) \geq \hat{\mu}_{a^*} + \beta_{a^*}(t) \geq \mu_{a^*}.$$

Hence either $a^* = a^t_h$ or a^t_h is ε-optimal arm.

I.3. Proof of Lemma 25

For completeness, we provide the proof in Xiong and Chen (2023).
Proof If the optimal arm $a^* = a^t_h$,

$$\begin{aligned}
\hat{\mu}_{a^t_l} + \beta_{a^t_l}(t) &\leq \mu_{a^t_l} + 2\beta_{a^t_l}(t) \\
&\leq \mu_{a^t_l} + \frac{\max\{\Delta_{a^t_l}, \varepsilon/2\}}{2} \\
&\leq \mu_{a^t_h} - \Delta_{a^t_l} + \frac{\max\{\Delta_{a^t_l}, \varepsilon/2\}}{2} \\
&\leq \hat{\mu}_{a^t_h} + \beta_{a^*}(T_{a^*}(t)) - \Delta_{a^t_l} + \frac{\max\{\Delta_{a^t_l}, \varepsilon/2\}}{2} \\
&\leq \hat{\mu}_{a^t_h} - \beta_{a^*}(T_{a^*}(t)) + \frac{\max\{\Delta_{a^*}, \varepsilon/2\} + \max\{\Delta_{a^t_l}, \varepsilon/2\}}{2} - \Delta_{a^t_l} \\
&\leq \hat{\mu}_{a^t_h} - \beta_{a^*}(T_{a^*}(t)) + \frac{\Delta_{a^*} + \varepsilon/2 + \Delta_{a^t_l} + \varepsilon/2}{2} - \Delta_{a^t_l} \\
&\leq \hat{\mu}_{a^t_h} - \beta_{a^*}(T_{a^*}(t)) + \varepsilon.
\end{aligned}$$

If optimal arm $a^* \neq a^t_h$, and the algorithm doesn't stop at round $t+1$, then we prove $a^* \neq a^t_l$. Otherwise, assume $a^* = a^t_l$

$$\begin{aligned}
\hat{\mu}^t_{a^t_h} &\leq \mu^t_{a^t_h} + \frac{\max\{\Delta_{a^t_h}, \varepsilon/2\}}{4} && (42) \\
&= \mu^t_{a^t_l} - \Delta_{a^t_h} + \frac{\max\{\Delta_{a^t_h}, \varepsilon/2\}}{4} && (43) \\
&\leq \mu^t_{a^t_l} - \frac{3\Delta_{a^t_h}}{4} + \varepsilon/4 && (44) \\
&\leq \hat{\mu}^t_{a^t_l} + \frac{\max\{\Delta_{a^*}, \varepsilon/2\}}{4} - \frac{3\Delta_{a^t_h}}{4} + \varepsilon/4 && (45) \\
&\leq \hat{\mu}^t_{a^t_l} + \varepsilon/2 - \frac{\Delta_{a^t_h}}{2}. && (46)
\end{aligned}$$

From the definition of a^t_h, we know $\varepsilon > \Delta_{a^t_h} \geq \Delta_{a^*}, \beta_{a^t_h}(t) \leq \varepsilon/4, \beta_{a^t_l}(t) \leq \varepsilon/4$. Then $\hat{\mu}^t_{a^t_l} + \beta_{a^t_l}(t) + \beta_{a^t_h}(t) \leq \hat{\mu}_{a^t_l} + \varepsilon/2 \leq \hat{\mu}^t_{a^t_h} + \varepsilon$, which means the algorithm stops at round $t+1$.

Now we can assume $a^* \neq a^t_l, a^* \neq a^t_h$. Then

$$\mu_{a^t_l} + 2\beta_{a^t_l}(t) \geq \hat{\mu}_{a^t_l} + \beta_{a^t_l}(t) \geq \hat{\mu}_{a^*} + \beta_{a^*}(T_{a^*}(t)) \geq \mu_{a^*} = \mu_{a^t_l} + \Delta_{a^t_l}. \qquad (47)$$

Thus

$$\Delta_{a_l^t} \le 2\beta_{a_l^t}(t) \le \frac{\max\{\Delta_{a_l^t}, \varepsilon/2\}}{2}, \tag{48}$$

which leads to $\Delta_{a_l^t} \le \varepsilon/2, \beta_{a_l^t}(t) \le \varepsilon/8$. Since

Also,

$$\mu_{a_h^t} + \beta_{a_h^t}(t) \ge \hat{\mu}_{a_h^t} \ge \hat{\mu}_{a_l^t} \ge \mu_{a^*} - \beta_{a_l^t}(t) = \mu_{a_h^t} + \Delta_{a_h^t} - \beta_{a_l^t}(t), \tag{49}$$

which leads to

$$\frac{\max\{\Delta_{a_h^t}, \varepsilon/2\}}{4} \ge \Delta_{a_h^t} - \varepsilon/8, \tag{50}$$

and $\Delta_{a_h^t} \le \varepsilon/2, \beta_{a_h^t}(t) \le \varepsilon/8$. Hence $\hat{\mu}^t_{a_l^t} + \beta_{a_l^t}(t) + \beta_{a_h^t}(t) \le \hat{\mu}_{a_l^t} + \varepsilon/2 \le \hat{\mu}^t_{a_h^t} + \varepsilon$, which means the algorithm stops at round $t+1$. ■

I.4. Proof of Theorem 24

Proof We construct $n-1$ graphs with the same distribution $P(\boldsymbol{X}, Y)$ but different causal graph. Indeed, We construct the bandit instances $\{\xi_i\}_{2 \le i \le n}$ as follows. For instance ξ_2, the graph structure contains edge $X_1 \to Y, X_2 \to X_1, X_1 \to X_i (3 \le i \le n)$ and $X_2 \to X_i (3 \le i \le n)$. For instances $\xi_i (3 \le i \le n)$, we change $X_1 \to X_i$ to $X_i \to X_1$. The graph structure are shown in the Figure 2 and Figure 3.

The observational distribution for all instance is:

$$P(\boldsymbol{X}, Y) = p_1 p_2 \dots p_n, \tag{51}$$

where

$$p_1 = 0.5, \tag{52}$$

$$p_2 = \begin{cases} 0.5 + \varepsilon & x_2 = x_1 \\ 0.5 - \varepsilon & x_2 \ne x_1 \end{cases} \tag{53}$$

$$p_i = \begin{cases} 0.5 + 4\varepsilon & x_i = x_1 \\ 0.5 - 4\varepsilon & x_i \ne x_1 \end{cases}. \tag{54}$$

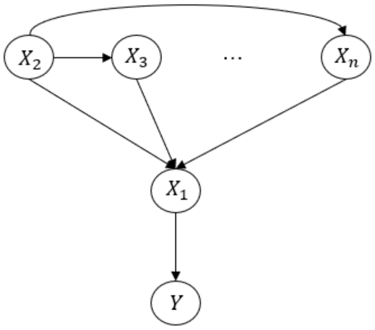


Figure 2: Causal Bandits Instance τ_2

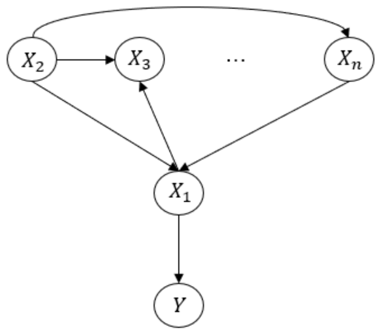


Figure 3: Causal Bandits Instance $\tau_i (i = 3)$

It is easy to check that $\sum_{\boldsymbol{x},y} P(\boldsymbol{X} = \boldsymbol{x}, Y = y) = 1$ and $P(X_i = 1) = 0.5$. The action set is $do(), do(X_i = 1), do(X_i = 0)$ where $2 \le i \le n$, which means the action set does not contain $do(X_1 = x)$ for $x = 0, 1$.

Now in ξ_2, we consider $P(Y = 1 \mid do(X_2 = 1))$. Actually, it is easy to show that $P(Y = 1 \mid do(X_2 = 1)) = P(X_1 = 1 \mid do(X_2 = 1)) = 0.5 + \varepsilon$. Similarly, $P(Y = 1 \mid do(X_2 = 0)) = 0.5 - \varepsilon$. For other actions, $P(Y = 1 \mid a) = P(X_1 = 1 \mid a) = 0.5$ since other actions a will not influence the value of X_1.

Now consider instance ξ_i for $3 \le i \le n$. For action $do()$ and $do(X_j = x)$ with $j \ne 2, i$, it will not influence the value of X_1 and then $P(Y = 1 \mid a) = 0.5$. Now consider action $a = do(X_2 = 1)$, we have

$$\begin{aligned} P(Y = 1 \mid do(X_2 = 1)) &= P(X_1 = 1 \mid do(X_2 = 1)) \\ &= P(X_1 = 1 \mid X_2 = 1) = 0.5 + \varepsilon. \end{aligned}$$

Similarly, $P(Y = 1 \mid do(X_2 = 0)) = 0.5 - \varepsilon$.

Now we calculate $P(Y = 1 \mid do(X_i = 1))$ in instance ξ_i. In fact, denote $q = 0.5 + 4\varepsilon$ and by do-calculus,

$$\begin{aligned} &P(X_1 = 1 \mid do(X_i = 1)) \\ &= \sum_{x=0,1} P(X_1 = 1 \mid X_i = 1, X_2 = x) P(X_2 = x) \\ &= 0.5(P(X_1 = 1 \mid X_i = 1, X_2 = 0) + P(X_1 = 1 \mid X_i = 1, X_2 = 1) \\ &= 0.5 \left(\frac{P(X_1 = 1, X_i = 1, X_2 = 0)}{P(X_i = 1, X_2 = 0)} + \frac{P(X_1 = 1, X_i = 1, X_2 = 1)}{P(X_i = 1, X_2 = 1)} \right) \\ &= 0.5 \left(\frac{(0.5 + 4\varepsilon)(0.5 - \varepsilon)}{(0.5 + 4\varepsilon)(0.5 - \varepsilon) + (0.5 - 4\varepsilon)(0.5 + \varepsilon)} + \frac{(0.5 + 4\varepsilon)(0.5 + \varepsilon)}{(0.5 + 4\varepsilon)(0.5 + \varepsilon) + (0.5 - 4\varepsilon)(0.5 - \varepsilon)} \right) \\ &= 0.5 \left(\frac{q(0.5 - \varepsilon)}{q(0.5 - \varepsilon) + (1 - q)(0.5 + \varepsilon)} + \frac{q(0.5 + \varepsilon)}{q(0.5 + \varepsilon) + (1 - q)(0.5 - \varepsilon)} \right) \\ &= 0.5 \left(\frac{q(0.5 - \varepsilon)}{0.5 - (2q - 1)\varepsilon} + \frac{q(0.5 + \varepsilon)}{0.5 + (2q - 1)\varepsilon} \right) \\ &= q \left(\frac{0.5^2 - (2q - 1)\varepsilon^2}{0.5^2 - (2q - 1)^2 \varepsilon^2} \right) \le q = 0.5 + 4\varepsilon. \end{aligned}$$

Also, we prove that

$$q\left(\frac{0.5^2-(2q-1)\varepsilon^2}{0.5^2-(2q-1)^2\varepsilon^2}\right)\geq 0.5+2\varepsilon.$$

Actually, this inequality is equal to

$$\begin{aligned}(0.5+4\varepsilon)(0.5^2-8\varepsilon^3)&\geq(0.5+2\varepsilon)(0.5^2-8\varepsilon^4)\\ \iff 1&\geq 56\varepsilon^3+8\varepsilon^2-32\varepsilon^4.\end{aligned}$$

When ε is small enough, this inequality holds. In summary, we have

$$P(X_1=1\mid do(X_i=1))\in[0.5+2\varepsilon,0.5+4\varepsilon].$$

Similarly, we can get

$$\begin{aligned}P(X_1=1\mid do(X_i=0))&=0.5(P(X_1=1\mid X_i=0,X_2=1)+P(X_1=1\mid X_i=0,X_2=0))\\ &=(1-q)\left(\frac{0.5^2-(1-2q)\varepsilon^2}{0.5^2-(1-2q)^2\varepsilon^2}\right)\in[0.5-4\varepsilon,0.5].\end{aligned}$$

Now in instance ξ_2, the output action should be $do(X_2=1)$, while in instance ξ_i, the output action should be $do(X_i=1)$.

Now by Pinkser's inequality, for an policy π, we have

$$2\delta\geq P_{\xi_2}(a^o=do(X_i=1))+P_{\xi_i}(a^o\neq do(X_i=1))\geq\exp(-\mathrm{KL}(\xi_2^\pi,\xi_i^\pi)).$$

Also, assume the stopping time as τ for the environment $\mathcal{E}$, the KL divergence can be rewritten as

$$\begin{aligned}\mathrm{KL}(\xi_2^\pi,\xi_i^\pi)&=\mathbb{E}_{A_t\sim\xi_2^\pi}\left[\sum_{t=1}^{\tau}\mathrm{KL}(P_{\xi_2}(\boldsymbol{X}_t,Y_t\mid A_t),P_{\xi_i}(\boldsymbol{X}_t,Y_t\mid A_t))\right] && (55)\\ &=\mathbb{E}_{\xi_2^\pi}\left[\sum_{t=1}^{\tau}P_{\xi_2}(\boldsymbol{X}_t,Y_t\mid A_t)\left(\log\frac{P_{\xi_2}(\boldsymbol{X}_t,Y_t\mid A_t)}{P_{\xi_i}(\boldsymbol{X}_t,Y_t\mid A_t)}\right)\right] && (56)\\ &=\mathbb{E}_{\xi_2^\pi}\left[\sum_{t=1}^{\tau}P_{\xi_2}(X_{t,i},X_{t,1}\mid A_t)\left(\log\frac{P_{\xi_2}(X_{t,i},X_{t,1}\mid A_t)}{P_{\xi_i}(X_{t,i},X_{t,1}\mid A_t)}\right)\right] && (57)\end{aligned}$$

where the last equation is derived as follows:

$$\frac{P_{\xi_2}(\boldsymbol{X}_t,Y_t\mid A_t)}{P_{\xi_i}(\boldsymbol{X}_t,Y_t\mid A_t)}=\frac{P_{\xi_2}(X_{t,i},X_{t,1}\mid A_t)\cdot P_{\xi_2}(\bar{\boldsymbol{X}}_{t,i},Y_t\mid X_{t,i},X_{t,1},A_t)}{P_{\xi_i}(X_{t,i},X_{t,1}\mid A_t)\cdot P_{\xi_i}(\bar{\boldsymbol{X}}_{t,i},Y_t\mid X_{t,i},X_{t,1},A_t)}$$

where $\bar{\boldsymbol{X}}_{t,i}=\boldsymbol{X}_t\setminus\{X_{t,i},X_{t,1}\}$. Now since $\bar{\boldsymbol{X}}_{t,i}$ is only decided by X_1,X_2 and X_2 is only decided by A_t, then

$$P_{\xi_2}(\bar{\boldsymbol{X}}_{t,i},Y_t\mid X_{t,i},X_{t,1},A_t)=P_{\xi_i}(\bar{\boldsymbol{X}}_{t,i},Y_t\mid X_{t,i},X_{t,1},A_t)$$

and then

$$\frac{P_{\xi_2}(\boldsymbol{X}_t, Y_t \mid A_t)}{P_{\xi_i}(\boldsymbol{X}_t, Y_t \mid A_t)} = \frac{P_{\xi_2}(X_{t,i}, X_{t,1} \mid A_t)}{P_{\xi_i}(X_{t,i}, X_{t,1} \mid A_t)}.$$

Note that only when $A_t = do(X_i = 1), do(X_i = 0)$, $P_{\xi_2}(X_{t,i}, X_{t,1} \mid A_t) \neq P_{\xi_i}(X_{t,i}, X_{t,1} \mid A_t)$. Then the equation (57) can be further calculated as

$$\begin{aligned}
(57) &= \sum_{x=0,1} \mathbb{E}_{\xi_2^\pi}\left[\sum_{t=1}^{\tau} \mathbb{I}\{A_t = do(X_i = x)\}\right] \cdot P_{\xi_2}(X_{t,i}, X_{t,1} \mid do(X_i = x)) \\
&\cdot \left(\log \frac{P_{\xi_2}(X_{t,i}, X_{t,1} \mid do(X_i = x))}{P_{\xi_i}(X_{t,i}, X_{t,1} \mid do(X_i = x))}\right) \\
&= \sum_{x=0,1} \mathbb{E}_{\xi_2^\pi}\left[\sum_{t=1}^{\tau} \mathbb{I}\{A_t = do(X_i = x)\}\right] \cdot P_{\xi_2}(X_{t,1} \mid do(X_i = x)) \\
&\cdot \left(\log \frac{P_{\xi_2}(X_{t,1} \mid do(X_i = x))}{P_{\xi_i}(X_{t,1} \mid do(X_i = x))}\right) \\
&\le \sum_{x=0,1} \mathbb{E}_{\xi_2^\pi}\left[\sum_{t=1}^{\tau} \mathbb{I}\{A_t = do(X_i = x)\}\right] \left(0.5 \cdot \left(\log \frac{0.5}{0.5 + 4\varepsilon} + \log \frac{0.5}{0.5 - 4\varepsilon}\right)\right) \\
&\le \sum_{x=0,1} \mathbb{E}_{\xi_2^\pi}\left[\sum_{t=1}^{\tau} \mathbb{I}\{A_t = do(X_i = x)\}\right] 96\varepsilon^2 \\
&= 96\varepsilon^2 \cdot \mathbb{E}_{\xi_2^\pi}[N(do(X_i = 1)) + N(do(X_i = 0))].
\end{aligned}$$

where the $\mathbb{E}_{\xi_2^\pi} N(a)$ represents that the number of times taking action a for policy π under the instance ξ_2. Now we have

$$\mathbb{E}_{\xi_2^\pi}[N(do(X_i = 1)) + N(do(X_i = 0))] \ge \frac{\mathrm{KL}(\xi_2^\pi, \xi_i^\pi)}{96\varepsilon^2} \ge \frac{1}{96\varepsilon^2} \log \frac{1}{2\delta}.$$

Hence the stopping time τ under policy π can be lower bounded by

$$\mathbb{E}_{\xi_2^\pi}[\tau] \ge \sum_{i=3}^{n} \mathbb{E}_{\xi_2^\pi}[N(do(X_i = 1)) + N(do(X_i = 0))] \ge \frac{n-2}{96\varepsilon^2} \log \frac{1}{2\delta} = O\left(\frac{n}{\varepsilon^2} \log \frac{1}{\delta}\right).$$

■

I.5. Technical Lemma

Lemma 28 *If $T = CH \log \frac{dT}{\delta}$ for some constant C and parameter d such that $d \ge e\delta$, then $T = O(H \log \frac{Hd}{\delta})$.*

Proof Let $f(x) = \frac{x}{\log(dx/\delta)}$, then for $x \ge 1$

$$f'(x) = \frac{\log(dx/\delta) - 1}{\log^2 dx/\delta} \ge 0$$

because $dx/\delta > e$. Then $f(x)$ is non-decreasing for $x \geq 1$.

To prove $T = O(H \log \frac{Hd}{\delta})$, we only need to show that $f(T) \leq f(C'H \log \frac{Hd}{\delta})$ for some constant C'. Since

$$\log \frac{C'Hd \log \frac{Hd}{\delta}}{\delta} = \log \frac{C'Hd}{\delta} + \log\log \frac{Hd}{\delta}$$

we only need to prove

$$f(C'H \log \frac{Hd}{\delta}) = \frac{C'H \log \frac{Hd}{\delta}}{\log \frac{C'Hd}{\delta} + \log\log \frac{Hd}{\delta}} \geq CH = f(T).$$

If we choose $C' \geq 2C + C \log C'$, then

$$\begin{aligned} CH\left(\log \frac{C'Hd}{\delta} + \log\log \frac{Hd}{\delta}\right) &\leq CH(\log \frac{C'Hd}{\delta} + \log \frac{Hd}{\delta}) \\ &\leq 2CH \log \frac{Hd}{\delta} + CH \log C' \\ &\leq (2C + C \log C')H \log \frac{Hd}{\delta} \\ &\leq C'H \log \frac{Hd}{\delta}. \end{aligned}$$

■