I. System Model

A. The Causal Graph Model with Soft Intervention

Without intervention, the causal structure is represented by a DAG, $(\mathcal{V}, \boldsymbol{B})$, where $\mathcal{V}$ is the set of N nodes and $\boldsymbol{B}$ is the observational edge-weight matrix. We consider soft interventions, defined as

$$\boldsymbol{a} = (a_1, \ldots, a_N)^T \in \{0,1\}^N \equiv \mathcal{A}, \tag{1}$$

where a_i represents whether node i is intervened (1) or not (0). Different from hard interventions, soft interventions do not cut off causal relationships between the intervened node and its parents, but change the upcoming edges to the node.

Further, we denote the interventional weight matrix by $\boldsymbol{B}'$ so that the post-intervention matrix $\boldsymbol{B_a}$ can be constructed as

$$[\boldsymbol{B_a}]_i = \mathbf{1}(a_i = 1)\,[\boldsymbol{B}']_i + \mathbf{1}(a_i = 0)\,[\boldsymbol{B}]_i\,, \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function and $[\cdot]_i$ represents the i-th column of a matrix. Denote the set of parents of node i by $\mathrm{pa}(i, a_i)$ and the set of ancestors by $\mathrm{an}(i, a_i)$. The i-th column of the post-intervention weight matrix determines $\mathrm{pa}(i, a_i)$ and how these parents influence node i.

With intervention, the vector of stochastic values associated with the nodes is represented by $X = (X_1, \ldots, X_N)^T$. The causal relationship among nodes is described by a linear structure equation model (LinSEM),

$$X = (\boldsymbol{B_a})^T X + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon}$ is a vector of Gaussian noise/exogenous variables, independent of X. We assume $\boldsymbol{\epsilon}$ has independent elements with known mean vector $\boldsymbol{\nu}$ and unknown covariance.

B. The Causal Bandit Model

In the MAB framework, an agent performs a sequence of actions to maximize cumulative reward over a finite horizon T. We consider node N as the reward node in the causal graph model, which generates stochastic rewards in each time step.

To compute the expected reward under intervention $\boldsymbol{a}$, we recognize that in LinSEM, there exists a causal flow between every ancestor-descendant pair. Thus each X_i can be written as a linear combination of exogenous variables in $\boldsymbol{\epsilon}$, weighted by the causal flow. Define the flow-weight matrix as

$$\boldsymbol{C_a} \doteq (\boldsymbol{I} - \boldsymbol{B_a})^{-1}\,, \tag{4}$$

where the (i,j)-th entry represents the net flow weight from node i to j. In this way, we rewrite (3) as

$$X = (\boldsymbol{I} - \boldsymbol{B_a})^{-T}\boldsymbol{\epsilon} = (\boldsymbol{C_a})^T\boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{I}$ denotes the identity matrix. The expectation of X under intervention $\boldsymbol{a} \in \mathcal{A}$ is formulated as

$$\boldsymbol{\mu_a} \doteq \mathbb{E}\left[(\boldsymbol{I} - \boldsymbol{B_a})^{-T}\boldsymbol{\epsilon}\right] = (\boldsymbol{I} - \boldsymbol{B_a})^{-T}\boldsymbol{\nu}. \tag{6}$$

Thus with the knowledge of post-intervention weight matrices, the optimal intervention can be obtained as

$$\boldsymbol{a}^* \doteq \operatorname*{arg\,max}_{\boldsymbol{a} \in \mathcal{A}}\ [\boldsymbol{\mu_a}]_N\,. \tag{7}$$

In each time step, the agent selects an intervention $\boldsymbol{a}^t$, observes X^t and collects reward X_N^t. The randomness of observation comes from the exogenous variables $\boldsymbol{\epsilon}^t$, which is independent of the intervention. The objective is to maximize the expected cumulative reward, $\sum_t [\boldsymbol{\mu}_{\boldsymbol{a}}^t]_N$.

II. Proof of Proposition 1

Proposition 1. With the signal model defined in Section I-A and MMSE estimation, the following inequalities hold, $\forall \boldsymbol{a}$,

$$\begin{aligned} I(R_i^t(a_i); X_j) \le\ & I\Big(\sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li}\epsilon_l - \epsilon_i;\ \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{lj}\,\epsilon_l\Big) \\ &+ I\Big(\sum_{k \in \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t - \boldsymbol{C_a}\big]_{ki}\epsilon_k;\ \sum_{k \in \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{kj}\,\epsilon_k\Big), \end{aligned} \tag{8}$$

$$\begin{aligned} I\Big(\sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li}\epsilon_l - \epsilon_i;\ \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{lj}\,\epsilon_l\Big) \le \\ \log\Big(1 + \frac{\sigma_{E1}}{\sigma_{E2}}\big|\big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji}\big|\Big) - \frac{1}{2}\log(1 - \rho_E^2), \end{aligned} \tag{9}$$

where σ_{E1}^2, σ_{E2}^2 are the variance of the following variables

$$E(X_j, i) = \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{lj}\,\epsilon_l, \tag{10}$$

$$E(R_i^t(a_i), \backslash j) = \sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li}\epsilon_l - \epsilon_i - \big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji}E(X_j, i), \tag{11}$$

and ρ_E stands for the correlation between them.

Proof: Focus on node i under intervention $\boldsymbol{a}$, with observation up to step t. The true causal relationship can be represented as

$$X_i = \sum_{k \in \mathrm{pa}(i,a_i)} [\boldsymbol{B_a}]_{ki}\,X_k + \epsilon_i = \sum_{k \in \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{ki}\,\epsilon_k + \epsilon_i. \tag{12}$$

And the estimate can be decomposed similarly as

$$\hat{X}_i^t(a_i) = \sum_{k \in \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{ki}\epsilon_k + \sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li}\epsilon_l. \tag{13}$$

Since we do not assume any causal knowledge, all other nodes are considered as covariates. The residual is represented as

$$\begin{aligned} R_i^t(a_i) = -\epsilon_i + \\ \underbrace{\sum_{k \in \mathrm{an}(i,a_i)} \Big(\big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{ki} - [\boldsymbol{C_a}]_{ki}\Big)\epsilon_k}_{\text{intrinsic error}} + \underbrace{\sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li}\epsilon_l}_{\text{causal error}}\,. \end{aligned} \tag{14}$$

Any non-ancestor node l should not be considered as a covariate and consequently, ϵ_l should not appear in the residual. Therefore we define the last term in (14) as the causal error. As a source of the causal error, we consider a specific non-ancestor node j and decompose its value as

$$X_j = \sum_{k \in \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{kj}\, \epsilon_k + \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{lj}\, \epsilon_l. \tag{15}$$

Next, based on the data processing inequality [1] and the fact that exogenous variables are independent, we can bound the mutual information as

$$\begin{aligned} &I(R_i^t(a_i); X_j) \le \\ &\quad I\Big(\sum_{k \in \mathrm{an}(i,a_i)} \Big(\big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{ki} - [\boldsymbol{C_a}]_{ki} \Big) \epsilon_k ; \sum_{k \in \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{kj}\, \epsilon_k \Big) + \\ &\quad I\Big(\sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li} \epsilon_l - \epsilon_i ; \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{lj}\, \epsilon_l \Big) \end{aligned} \tag{16}$$

where the first and second parts are the mutual information induced by the intrinsic and causal error, respectively. Due to the existence of intrinsic error, we may not be able to detect the causal error and thus misjudge causal directions. To illustrate this point, let us look deeper into the causal error induced mutual information. To simplify notation, denote the non-causal part in X_j with respect to node i by

$$E(X_j, i) \doteq \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{lj}\, \epsilon_l, \tag{17}$$

which is Gaussian with variance denoted by σ_{E1}^2. The non-causal edge used in the estimation, $j \to i$, passes this causal error to the residual, weighted by $\big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji}$. We denote the causal error in the residual induced by other nodes by

$$E(R_i^t(a_i), \backslash j) \doteq \sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li} \epsilon_l - \epsilon_i - \big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji} E(X_j, i), \tag{18}$$

which is Gaussian as well, with variance denoted by σ_{E2}^2. Thus we can rewrite and bound the causal error induced mutual information as

$$\begin{aligned} &I\Big(\sum_{l \notin \mathrm{an}(i,a_i)} \big[\hat{\boldsymbol{C}}_{\boldsymbol{a}}^t\big]_{li} \epsilon_l - \epsilon_i ; \sum_{l \notin \mathrm{an}(i,a_i)} [\boldsymbol{C_a}]_{lj}\, \epsilon_l \Big) \\ &= h\Big(\big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji} E(X_j, i) + E(R_i^t(a_i), \backslash j) \Big) + h\Big(E(X_j, i) \Big) \\ &\qquad - h\Big(E(R_i^t(a_i), \backslash j), E(X_j, i) \Big) \qquad (19) \\ &\le \log\Big(\big(|\big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji}| \sigma_{E1} + \sigma_{E2}\big) \sqrt{2\pi e} \Big) + \log\Big(\sigma_{E1} \sqrt{2\pi e} \Big) \\ &\qquad - \Big[\log(2\pi e) + \log\Big(\sigma_{E1} \sigma_{E2} \sqrt{1 - \rho_E^2} \Big) \Big] \qquad (20) \\ &= \log\Big(1 + \frac{\sigma_{E1}}{\sigma_{E2}} |\big[\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t\big]_{ji}| \Big) - \frac{1}{2} \log(1 - \rho_E^2), \qquad (21) \end{aligned}$$

where $h(\cdot)$ represents differential entropy and ρ_E stands for the correlation between $E(X_j, i)$ and $E(R_i^t(a_i), \backslash j)$. The inequality comes from the fact that the standard deviation is bounded above by the sum of each individual standard deviation. This inequality completes the proof. □

III. Proof of Theorem 1

We first state the following lemmas that are necessary for the proof of Theorem 1.

Lemma 1. With the signal model defined in Section I-A and MMSE estimation, the norm of weight error vectors have upper-bounded moments for $m \ge 2$:

$$\mathbb{E}\Big[\| \big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i \|_2^m \Big] \le m! \sqrt{\frac{4N}{N+2}} \left[\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i)) \frac{N+2}{4} \right]^{\frac{m}{2}}, \tag{22}$$

where $\lambda_{\max}(\cdot)$ stands for the maximum eigenvalue of a matrix.

Proof: According to the theory of linear least-square estimation, each weight error vector follows a multivariate normal distribution,

$$\big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i)), \tag{23}$$

where the covariance matrix can be decomposed as $\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i) = \boldsymbol{Q}_i \boldsymbol{\Lambda}_i \boldsymbol{Q}_i^T$. With the orthonormal matrix $\boldsymbol{Q}_i$, we have

$$\| \big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i \|_2 = \| \boldsymbol{Q}_i^T \big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i \|_2, \quad \boldsymbol{Q}_i^T \big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}_i). \tag{24}$$

Since the squared norm of $\boldsymbol{Q}_i^T \big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i$ is a weighted sum of independent squared Gaussian variables, it can be bounded as

$$\| \boldsymbol{Q}_i^T \big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i \|_2^2 \le \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i)) \| \boldsymbol{\Lambda}_i^{-1/2} \boldsymbol{Q}_i^T \big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i \|_2^2. \tag{25}$$

Notice that $\| \boldsymbol{\Lambda}_i^{-1/2} \boldsymbol{Q}_i^T [\Delta \boldsymbol{B}_{\boldsymbol{a}}^t]_i \|_2^2$ follows a Chi-squared distribution with N degree of freedom, which has well-defined moments. Thus we can upper bound the m-th moment as

$$\begin{aligned} \mathbb{E}\Big[\| \big[\Delta \boldsymbol{B}_{\boldsymbol{a}}^t\big]_i \|_2^m \Big] &\le \sqrt{ \mathbb{E}\Big[(\| [\Delta \boldsymbol{B}_{\boldsymbol{a}}^t]_i \|_2^2)^m \Big] } && (26) \\ &\le \sqrt{ \lambda_{\max}\left(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i) \right)^m 2^m \frac{\Gamma(m + N/2)}{\Gamma(N/2)} } && (27) \\ &\le m! \sqrt{\frac{4N}{N+2}} \left[\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i)) \frac{N+2}{4} \right]^{\frac{m}{2}}, && (28) \end{aligned}$$

where $\Gamma(\cdot)$ represents the Gamma function. The first two inequalities are based on Jensen's inequality and moments of Chi-squared distribution. To explain the last inequality, we first rewrite the quotient of Gamma functions as

$$2^m \frac{\Gamma(m + N/2)}{\Gamma(N/2)} = (m!)^2 \cdot \prod_{k=1}^{m} \frac{N + 2k - 2}{k^2}. \tag{29}$$

When $m \ge 2$, we have the following inequalities

$$\frac{N + 2k - 2}{k^2} = \frac{N - 2}{k^2} + \frac{2}{k} \le \frac{N+2}{4}, \quad \forall k \ge 2, \tag{30}$$

$$2^m \frac{\Gamma(m + N/2)}{\Gamma(N/2)} \le (m!)^2 \cdot N \cdot \left(\frac{N+2}{4} \right)^{m-1}, \tag{31}$$

which completes the proof. □

Lemma 2. With the signal model defined in Section I-A and MMSE estimation, the following inequality holds,

$$\mathbb{P}\Bigg\{\sigma_{\max}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t) \geq 2(N^2+2N)^{\frac{1}{4}} \cdot \sqrt{\ln\Big(\frac{2N}{\delta}\Big)\sum_{i=1}^{N}\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i))}\Bigg\} \leq \delta, \quad (32)$$

where $\sigma_{\max}(\cdot)$ represents the maximum singular-value.

Proof: For a specific intervention $\boldsymbol{a}$, consider the dilation [2] of the weight error matrix,

$$\mathcal{H}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t) \doteq \begin{bmatrix} \boldsymbol{0} & \Delta\boldsymbol{B}_{\boldsymbol{a}}^t \\ (\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)^T & \boldsymbol{0} \end{bmatrix}. \quad (33)$$

An important property is that the dilation preserves spectral information:

$$\lambda_{\max}(\mathcal{H}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)) = \sigma_{\max}(\mathcal{H}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)) = \sigma_{\max}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t), \quad (34)$$

Since the weight matrix consists of independently learned columns, we decompose the dilation as

$$\mathcal{H}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t) = \sum_{i=1}^{N}\mathcal{H}_i, \quad \mathcal{H}_i \doteq \begin{bmatrix} \boldsymbol{0} & (\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)_i \\ (\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)_i^T & \boldsymbol{0} \end{bmatrix}, \quad (35)$$

where the $N \times N$ matrix $(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)_i$ is composed of the i-th column of $\Delta\boldsymbol{B}_{\boldsymbol{a}}^t$ and zero everywhere else. Applying block matrix multiplication repeatedly, we have

$$\mathcal{H}_i^2 = \begin{bmatrix} (\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)_i\,(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)_i^T & \boldsymbol{0} \\ \boldsymbol{0} & (\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)_i^T\,(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)_i \end{bmatrix}, \quad (36)$$

$$\mathcal{H}_i^m = \begin{cases} \|[\Delta\boldsymbol{B}_{\boldsymbol{a}}^t]_i\|_2^{m-1}\cdot\mathcal{H}_i, & m \text{ is odd} \\ \|[\Delta\boldsymbol{B}_{\boldsymbol{a}}^t]_i\|_2^{m-2}\cdot\mathcal{H}_i^2, & m \text{ is even} \end{cases}. \quad (37)$$

We further observe that

$$\lambda_{\max}(\mathcal{H}_i) = \|\big[\Delta\boldsymbol{B}_{\boldsymbol{a}}^t\big]_i\|_2, \quad \lambda_{\max}(\mathcal{H}_i^2) = \|\big[\Delta\boldsymbol{B}_{\boldsymbol{a}}^t\big]_i\|_2^2, \quad (38)$$

which enables us to draw a general conclusion that

$$\mathcal{H}_i^m \preceq \|\big[\Delta\boldsymbol{B}_{\boldsymbol{a}}^t\big]_i\|_2^m \cdot \boldsymbol{I}, \quad (39)$$

where $\preceq$ represents the matrix semi-definite ordering. Further, taking expectation and applying Lemma 1, we have

$$\mathbb{E}[\mathcal{H}_i] = \boldsymbol{0} \quad (40)$$

$$\mathbb{E}[\mathcal{H}_i^m] \preceq m!\sqrt{\frac{4N}{N+2}}\left[\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i))\frac{N+2}{4}\right]^{\frac{m}{2}}\boldsymbol{I}, \; m \geq 2. \quad (41)$$

These properties enable us to apply the matrix Bernstein inequality for the sub-exponential case [3] and obtain

$$\mathbb{P}\{\sigma_{\max}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t) \geq \xi\} = \mathbb{P}\{\lambda_{\max}(\mathcal{H}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)) \geq \xi\} \quad (42)$$

$$= \mathbb{P}\left\{\lambda_{\max}\Big(\sum_{i=1}^{N}\mathcal{H}_i\Big) \geq \xi\right\} \quad (43)$$

$$\leq 2N\exp\left[-\frac{\xi^2}{4\sqrt{N(N+2)}\cdot\sum_{i=1}^{N}\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i))}\right], \quad (44)$$

which is valid when

$$\xi \leq \sqrt{\frac{4N}{\max_i\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i))}}\cdot\sum_{i=1}^{N}\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i)). \quad (45)$$

Lastly, we denote the probability bound by δ and solve ξ from

$$\delta = 2N\exp\left[-\frac{\xi^2}{4\sqrt{N(N+2)}\cdot\sum_{i=1}^{N}\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i))}\right], \quad (46)$$

to arrive at the desired inequality. □

Theorem 1. With the signal model defined in Section I-A and MMSE estimation, the following inequality holds:

$$\mathbb{P}\Big\{\big|\big[\Delta\boldsymbol{\mu}_{\boldsymbol{a}}^t\big]_N\big| \geq U(\boldsymbol{X}^{1:t},\boldsymbol{a},\delta)\Big\} \leq \delta, \quad (47)$$

where $U(\boldsymbol{X}^{1:t},\boldsymbol{a},\delta)$ represents the error upper-bound at confidence level $1-\delta$,

$$U(\boldsymbol{X}^{1:t},\boldsymbol{a},\delta) = 2(N^2+2N)^{\frac{1}{4}}\left\|\Big[(\boldsymbol{I}-\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-1}\Big]_N\right\|_2 \cdot \|\boldsymbol{\mu}_{\boldsymbol{a}}\|_2\sqrt{\ln\Big(\frac{2N}{\delta}\Big)\sum_{i=1}^{N}\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{a}}^t(i))}.$$

Proof: First, recognize that the error of estimated reward is the last element of $\Delta\boldsymbol{\mu}_{\boldsymbol{a}}^t$. To understand the behavior of $\Delta\boldsymbol{\mu}_{\boldsymbol{a}}^t$, consider the following two linear systems,

$$(\boldsymbol{I}-\boldsymbol{B}_{\boldsymbol{a}})^T\boldsymbol{\mu}_{\boldsymbol{a}} = \boldsymbol{\nu}, \quad (48)$$

$$(\boldsymbol{I}-\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^T\hat{\boldsymbol{\mu}}_{\boldsymbol{a}}^t = \boldsymbol{\nu}, \quad (49)$$

where (48) and (49) represent the true and estimated systems, respectively. Based on the Woodbury matrix identity [4], we have

$$(\boldsymbol{I}-\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-1} = (\boldsymbol{I}-\boldsymbol{B}_{\boldsymbol{a}})^{-1} + (\boldsymbol{I}-\boldsymbol{B}_{\boldsymbol{a}})^{-1}\Delta\boldsymbol{B}_{\boldsymbol{a}}^t(\boldsymbol{I}-\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-1}. \quad (50)$$

Combining (49) with (50), we can express $\Delta\boldsymbol{\mu}_{\boldsymbol{a}}$ as

$$\Delta\boldsymbol{\mu}_{\boldsymbol{a}}^t = -\boldsymbol{\mu}_{\boldsymbol{a}} + (\boldsymbol{I}-\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-T}\boldsymbol{\nu} \quad (51)$$

$$= (\boldsymbol{I}-\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-T}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)^T\boldsymbol{\mu}_{\boldsymbol{a}}, \quad (52)$$

where (52) is a direct consequence of (48). Now, we can bound the error of estimated reward as

$$\big|\big[\Delta\boldsymbol{\mu}_{\boldsymbol{a}}^t\big]_N\big| = \left|\Big[(\boldsymbol{I}-\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-1}\Big]_N^T(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)^T\boldsymbol{\mu}_{\boldsymbol{a}}\right| \quad (53)$$

$$\leq \left\|\Big[(\boldsymbol{I}-\hat{\boldsymbol{B}}_{\boldsymbol{a}}^t)^{-1}\Big]_N\right\|_2\sigma_{\max}(\Delta\boldsymbol{B}_{\boldsymbol{a}}^t)\,\|\boldsymbol{\mu}_{\boldsymbol{a}}\|_2\,. \quad (54)$$

Based on that, we employ Lemma 2 to provide a high probability bound on the largest singular value, which completes the proof. □

References

[1] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[2] V. Paulsen, *Completely bounded maps and operator algebras*. Cambridge University Press, 2002, no. 78.

[3] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, pp. 389–434, 2012.

[4] N. J. Higham, *Accuracy and stability of numerical algorithms*. SIAM, 2002.