2024 IEEE Conference on Control Technology and Applications (CCTA)
August 21-23, 2024. Newcastle upon Tyne, UK

Dynamic Intervention in Gene Regulatory Networks: A Partially Observed Zero-Sum Markov Game

Seyed Hamid Hosseini, and Mahdi Imani

2024 IEEE Conference on Control Technology and Applications (CCTA) | 979-8-3503-7094-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CCTA60707.2024.10666558

Abstract— **Gene Regulatory Networks (GRNs) are pivotal in governing diverse cellular processes, such as stress response, DNA repair, and mechanisms associated with complex diseases like cancer. The interventions in GRNs aim to restore the system state to its normal condition by altering gene activities over time. Unlike most intervention approaches that rely on the direct observability of the system state and assume no response of the cell against intervention, this paper models the fight between intervention and cell dynamic response using a partially observed zero-sum Markov game with binary state variables. The paper derives a stochastic intervention policy under partial state observability of genes. The optimal Nash equilibrium intervention policy is first obtained for the underlying system. To overcome the challenges of partial state observability, the paper employs the optimal minimum mean-square error (MMSE) state estimator to estimate the system state, given all available information. The proposed intervention policy utilizes the optimal Nash intervention policy associated with the optimal MMSE state estimator. The performance of the proposed method is examined using numerical experiments on the melanoma regulatory network observed through gene-expression data.**

I. Introduction

Recent advancements in genomics technology have allowed for a better understanding of complex biological systems, particularly gene regulatory networks (GRNs) [1]–[5]. These networks are composed of interacting genes that control ecosystem functioning and cellular processes such as DNA repair, stress response, and complex diseases like cancer [6]. A critical goal in genomics is to develop effective intervention strategies to alter the undesirable behavior of GRNs, particularly those associated with chronic diseases [7], [8].

Several intervention strategies have been developed for GRNs [9]–[12]. These include dynamic perturbations, which provide time-dependent interventions [13], and structural interventions, which make a single-time change in gene interactions to properly shift their dynamics [14], [15]. The majority of existing intervention methods consider cells as isolated and non-responsive entities [16]. However, cells are highly dynamic and intelligent, often fighting back against interventions or therapies through internal stimuli.

The impact of dynamic cell responses on the performance of existing dynamic intervention methods is studied in [17]. These intervention methods rely on the stationarity of the intervention process, wherein cells do not respond to therapies. Lack of consideration of cell responses often leads to early success of existing methods in shifting the system's undesirable behaviors, followed by the recurrence of unhealthy conditions at a later time. This comes from the deterministic nature of these policies, which allows cells, as intelligent and dynamic entities, to find ways to fight against interventions through their internal stimuli. In practice, direct access to cell responses during the intervention is not achievable; rather, noisy gene-expression data, which provides partial knowledge about the state of the genes, should be utilized for making intervention decisions.

This paper models the intervention process in GRNs using a partially observed two-player zero-sum Markov game with binary state variables. The underlying model represents a two-player zero-sum game in which the two players are the cell and the intervention, each with opposing objectives [18]. This model accounts for realistic conditions regarding no access to cell responses and partial access to the system state. A recursive Bayesian approach is derived to capture the posterior distribution of the state based on the available gene-expression data. The optimal minimum mean square error (MMSE) state estimator is used to recursively estimate the genes' state, given the state posterior distribution. The proposed intervention method utilizes the optimal Nash intervention policy corresponding to the estimated state, as the true state is unknwon, during the intervention process. This process resembles a state-feedback controller scheme, where the optimal solution of the zero-sum game serves as a controller, and the optimal MMSE state estimator acts as an state estimator.

This paper demonstrates the empirical convergence of the proposed policy to the optimal Nash policy associated with the true state. Analytical results are used to investigate the shortcomings of existing dynamic intervention methods. Meanwhile, we measure the distance of the proposed intervention policy, which is a stochastic policy, with the optimal Nash policy. We show that the expected error of state estimation can be computed and used as a confidence measure to assess the deviation of the proposed intervention policy from the optimal Nash policy. The numerical results using the well-known melanoma regulatory network demonstrate the superiority of the performance of the proposed method compared to several existing intervention policies.

II. Background

This paper models gene regulatory networks using the partially observed Boolean dynamical systems (POBDS) [19]–[24]. The POBDS is a generalization of Boolean network models [25], [26], where the genes' activity is modeled

S. H. Hosseini and M. Imani are with the Department of Electrical and Computer Engineering at Northeastern University. Emails: hosseini.ha@northeastern.edu, m.imani@northeastern.edu

through the binary state process, and the uncertainty in the gene-expression data is modeled through the measurement process.

The state vector of a GRN, consisting d genes at time step k, can be expressed as $\mathbf{x}_k \in \{0,1\}^d$, where $\mathbf{x}_k(i) = 0$ and $\mathbf{x}_k(i) = 1$ represent the inactivation and activation of the ith gene at time step k, respectively. The genes' state gets updated according to the following model:

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) \oplus \mathbf{a}_{k-1} \oplus \mathbf{u}_{k-1} \oplus \mathbf{n}_k, \quad k = 1, 2, \ldots, \quad (1)$$

where $\{\mathbf{a}_k; k = 0, 1, ...\}$ denotes a set of external interventions or therapies, while $\{\mathbf{u}_k; k = 0, 1, ...\}$ represents internal cell stimuli. The variable $\mathbf{n}_k \in \{0,1\}^d$ characterizes process noise at the time step k. The symbol "$\oplus$" indicates element-wise modulo-2 addition, and $\mathbf{f}$ stands for the *network function*. If $\mathbf{n}_k(j) = 0$, the state of the jth gene at time k is determined by the network function, while $\mathbf{n}_k(j) = 1$ alters the jth gene value predicted by the network function. We assume the noise process $\mathbf{n}_k$ comprises independent components coming from a Bernoulli distribution with a parameter $0 \le p \le 0.5$. This parameter p determines the level of "stochasticity" within the Boolean state process. Larger p values model more chaotic systems, whereas smaller p values indicate nearly deterministic systems/processes.

The measurement process, in a most general form, can be expressed as:

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k, \mathbf{v}_k), \ k = 1, 2, \ldots, \quad (2)$$

where $\mathbf{h}(.)$ is the observation function and $\mathbf{v}_k$ is the measurement noise. The observation function varies depending on the types of genomics data. Assuming the measurements are from cDNA microarrays [27] or live-cell imaging-based assays [28], the measurement process can be expressed using the following Gaussian model:

$$\mathbf{y}_k(j) = m + \delta \mathbf{x}_k(j) + \mathbf{v}_k(j), \quad k = 1, 2, \ldots, \quad (3)$$

for $j = 1, \ldots, d$, where $\mathbf{v}_k(j) \sim \mathcal{N}(0, \sigma^2)$ represents a sample from Gaussian noise with a mean of zero and a variance of σ^2. In this scenario, m denotes the baseline expression for the inactivated genes, and δ represents the magnitude of the differential expression. The differential expression parameter determines how much genes are expressed differently in the measurements in inactivated and activated states.

III. Proposed Intervention Policy

A. Intervention as Two-Player Zero-Sum Game

We model the fight between the cell and intervention as a two-player zero-sum game. This scenario is characterized by a tuple denoted by $\langle \mathcal{X}, \mathcal{A}, \mathcal{U}, R^a, \mathcal{P} \rangle$, where $\mathcal{X} = \{0,1\}^d$ represents the *state space*, $\mathcal{A}$ corresponds to the *intervention space*, $\mathcal{U}$ is the *cell stimuli space*, R^a is the *intervention reward function*, and $\mathcal{P}$ stands for *state transition probability function*. The $p(\mathbf{x}' \mid \mathbf{x}, \mathbf{a}, \mathbf{u})$ indicates the probability of transitioning from state $\mathbf{x}$ to state $\mathbf{x}'$ given the external intervention $\mathbf{a}$ and the internal cell input $\mathbf{u}$. Additionally, $R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}')$ represents the immediate intervention reward gained after moving from state $\mathbf{x}$ to state $\mathbf{x}'$, upon performing the intervention $\mathbf{a}$ and the internal cell input $\mathbf{u}$.

The intervention aims at deviating the system from unhealthy conditions (e.g., slowing down cell proliferation), whereas the cell aims at enhancing uncontrolled cell proliferation and keeping the system in unhealthy conditions. Consequently, the cell's reward function R^u is the negative of the intervention reward function, expressed as $R^u(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') = -R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}')$. This creates a give-and-take scenario, where what benefits the intervention is against the cell's interests. This study focuses on stationary Markov Nash equilibria within GRNs, as modeled through the infinite-horizon discounted Markov game framework. Let $\pi^a(\mathbf{a} \mid \mathbf{x})$ be an intervention policy, which determines the probability of taking interventions $\mathbf{a} \in \mathcal{A}$ at a given state $\mathbf{x} \in \mathcal{X}$. Similarly, $\pi^u(\mathbf{u} \mid \mathbf{x})$ represents the cell's policy, indicating the probability of the cell input $\mathbf{u} \in \mathcal{U}$ at state $\mathbf{x} \in \mathcal{X}$. We define the state value function for the intervention under the joint stochastic policy (π^a, π^u) as:

$$V^a_{\pi^a,\pi^u}(\mathbf{x}) = \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t R^a(\mathbf{x}_t, \mathbf{a}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) \mid \mathbf{a}_{0:\infty} \sim \pi^a, \mathbf{u}_{0:\infty} \sim \pi^u, \mathbf{x}_0 = \mathbf{x}\Bigg], \quad (4)$$

for $\mathbf{x} \in \mathcal{X}$, with $0 < \gamma < 1$ representing a discount factor signifying the relative importance of rewards at early stages compared to future ones. According to (4), the state values of the cell and the intervention are intertwined. In reality, the solution to a Markov game differs from a Markov Decision Process (MDP) due to the fact that the optimal performance of each agent is influenced not solely by its individual policy but also by the decisions of both cell and intervention within the game.

The solution of a Markov game yields a Nash equilibrium policy denoted by $\pi^* = (\pi^a_*, \pi^u_*)$. This policy, for any combination of joint strategies $\pi = (\pi^a, \pi^u)$ and any state $\mathbf{x} \in \mathcal{X}$, satisfies the following condition:

$$V^a_{\pi^a_*,\pi^u_*}(\mathbf{x}) \ge V^a_{\pi^a,\pi^u_*}(\mathbf{x}), \text{ and } V^a_{\pi^a_*,\pi^u_*}(\mathbf{x}) \le V^a_{\pi^a_*,\pi^u}(\mathbf{x}). \quad (5)$$

where the optimal Nash equilibrium policy corresponds to a case where neither the cell nor the intervention finds any incentive to deviate from their individual strategies. Employing the min-max theorem in the matrix form of zero-sum games results in the following expression of the optimal Nash equilibrium policy [29]:

$$(\pi^a_*, \pi^u_*) = \operatorname*{argmin}_{\pi^u} \operatorname*{argmax}_{\pi^a} V^a_{\pi^a,\pi^u}(\mathbf{x}), \quad \text{for} \quad \mathbf{x} \in \mathcal{X}, \quad (6)$$

where the space of π^a contains 2^d simplexes of size $\mathcal{A}$ and π^u space is 2^d simplex of size $\mathcal{U}$. A systematic approach to calculating the Nash policy is provided in the next section.

B. Optimal MMSE State Estimator

Since the true system state (e.g., $\mathbf{x}_k$) is unknown, directly performing the optimal Nash equilibrium policy in (6) according to the observed gene-expression data is impossible. Thus, this paper builds the intervention policy according to the best estimate of the system state. Let $\mathbf{a}_{0:k-1} =$

$(\mathbf{a}_0, ..., \mathbf{a}_{k-1})$ be the sequence of performed interventions, and $\mathbf{y}_{1:k} = (\mathbf{y}_1, ..., \mathbf{y}_k)$ be the observed gene-expression data up to time step k. The sequence of cell stimuli $\mathbf{u}_{0:k-1} = (\mathbf{u}_0, ..., \mathbf{u}_{k-1})$ is not directly observable. If the true system state $\mathbf{x}_k$ at time step k was known, then the optimal policy in (6) could be implemented as $\mathbf{a}_k \sim \pi^a_*(. \mid \mathbf{x}_k)$, which guarantees the best intervention outcome. Our objective is to find the state estimator $\hat{\mathbf{x}}_{k|k}$ of the true state $\mathbf{x}_k$ by minimizing a criterion measuring the difference between the estimated and true unobserved state. The following theorem characterizes the exact optimal minimum mean-square error (MMSE) solution [19], [30]:

Theorem 1: *Given $\mathbf{a}_{0:k-1}$ and $\mathbf{y}_{1:k}$ be the set of performed interventions and observed measurements up to time step k, the optimal MMSE state estimator at time step k can be obtained as:*

$$\hat{\mathbf{x}}^{\mathrm{MS}}_{k|k} = \overline{\mathbb{E}\left[\mathbf{x}_k \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k}\right]}, \tag{7}$$

where $\overline{\mathbf{v}}$ is a nonlinear operator mapping the vector elements greater than $1/2$ to 1 and others to 0. The expected error of the optimal estimator can be computed as:

$$C^{MS}_{k|k} = \frac{d}{2} - \sum_{i=1}^{d} \left| \mathbb{E}\left[\mathbf{x}_k(i) \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k}\right] - \frac{1}{2} \right|. \tag{8}$$

The error takes in $0 \le C^{\mathrm{MS}}_{k|k} \le d/2$, where small values close to 0 represent accurate state estimator, whereas expected errors close to $d/2$ indicate less accurate estimation. See [19], for the proof of the Theorem.

The optimal state estimator yields the exact MMSE optimality and the best estimate of the system state given all available information (i.e., $\mathbf{a}_{0:k-1}, \mathbf{y}_{1:k}$). One can select the intervention at time step k by replacing the true system state $\mathbf{x}_k$ with the optimal MMSE state estimator as:

$$\mathbf{a}_k \sim \pi^a_*\left(. \mid \hat{\mathbf{x}}^{\mathrm{MS}}_{k|k}\right), \tag{9}$$

where π^a_* is the optimal Nash intervention policy for the underlying system computed in (6).

The proposed intervention in (9) requires the computation of the state estimator as a new intervention is performed and a new measurement is observed. Let $p(\mathbf{x}_k \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k})$ be the state posterior distribution given the information up to time step k. The expectation in (7) can be expressed using the posterior distribution of state as:

$$\mathbb{E}[\mathbf{x}_k \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k}] = \sum_{i=1}^{2^d} \mathbf{x}^i \, p(\mathbf{x}_k = \mathbf{x}^i \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k}), \tag{10}$$

where $\{\mathbf{x}^1 = [0, \ldots, 0]^T, \ldots, \mathbf{x}^{2^d} = [1, \ldots, 1]^T\}$ are all the possible system states. Upon performing a new intervention $\mathbf{a}_k$ and observing a new measurement $\mathbf{y}_{k+1}$, the optimal MMSE state estimator computations require recursive computation of the posterior of the system state at time step $k+1$. This can be expressed as:

$$\begin{aligned} &p(\mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k+1}) \\ &= \frac{p(\mathbf{y}_{k+1}, \mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k})}{\sum_{l=1}^{2^d} p(\mathbf{y}_{k+1}, \mathbf{x}_{k+1} = \mathbf{x}^l \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k})} \\ &= \frac{p(\mathbf{y}_{k+1} \mid \mathbf{x}_{k+1} = \mathbf{x}^i)\, p(\mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k})}{\sum_{l=1}^{2^d} p(\mathbf{y}_{k+1} \mid \mathbf{x}_{k+1} = \mathbf{x}^l)\, p(\mathbf{x}_{k+1} = \mathbf{x}^l \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k})}, \end{aligned} \tag{11}$$

where in the last expression, the first term in the numerator represents the measurement model, and the second term can be further expanded as:

$$\begin{aligned} &p(\mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k}) \\ &= \sum_{j=1}^{2^d} p(\mathbf{x}_{k+1} = \mathbf{x}^i, \mathbf{x}_k = \mathbf{x}^j \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k}) \\ &= \sum_{j=1}^{2^d} p(\mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k}, \mathbf{x}_k = \mathbf{x}^j) p(\mathbf{x}_k = \mathbf{x}^j \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k}). \end{aligned} \tag{12}$$

The second part of the last expression in (12) is the jth element of the state posterior distribution at time step k, and the first term can be expanded through the marginalization over the unobserved/unmeasured cell action as:

$$\begin{aligned} p(\mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k}, \mathbf{x}_k = \mathbf{x}^j) &= \sum_{\mathbf{u} \in \mathcal{U}} p(\mathbf{x}_{k+1} = \mathbf{x}^i, \mathbf{u}_k = \mathbf{u} \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k}, \mathbf{x}_k = \mathbf{x}^j) \\ &= \sum_{\mathbf{u} \in \mathcal{U}} p(\mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k}, \mathbf{x}_k = \mathbf{x}^j, \mathbf{u}_k = \mathbf{u}) \\ &\qquad \times p(\mathbf{u}_k = \mathbf{u} \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k}, \mathbf{x}_k = \mathbf{x}^j) \\ &= \sum_{\mathbf{u} \in \mathcal{U}} p(\mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_k, \mathbf{x}_k = \mathbf{x}^j, \mathbf{u}_k = \mathbf{u}) \\ &\qquad \times \underbrace{p(\mathbf{u}_k = \mathbf{u} \mid \mathbf{x}_k = \mathbf{x}^j)}_{\pi^u_*(\mathbf{u} \mid \mathbf{x}^j)}, \end{aligned} \tag{13}$$

where the first term of the last expression is obtained according to Markov properties of the state process, and the second term $p(\mathbf{u}_k = \mathbf{u} \mid \mathbf{x}_k = \mathbf{x}^j)$ represents the probability that the cell action/response would be $\mathbf{u}_k = \mathbf{u}$ if the system state is $\mathbf{x}_k = \mathbf{x}^j$. Note that assuming the cell follows the optimal Nash equilibrium policy, the last term is replaced by $\pi^u_*(\mathbf{u} \mid \mathbf{x}^j)$. This assumption is valid for an intelligent cell, as it yields the best results for the cell when the intervention is close to the optimal Nash intervention policy.

Replacing (12) and (13) into (11) leads to the recursive posterior update of the system state required for computation of the optimal MMSE state estimator. The details of the computation of the proposed method are expressed in the next section.

IV. Matrix-Form Computation of Proposed Intervention Policy

This section outlines a matrix-form procedure for efficient computation of the proposed intervention policy, involving offline Nash policy computation and online state estimation.

A. Offline Step

We describe a dynamic programming technique for computing the optimal Nash equilibrium policy for a two-player zero-sum game. Let $\mathbf{V} = [\mathbf{V}(1), ..., \mathbf{V}(2^d)]^T$ be the state value vector associated with a given cell and intervention policy. We define the intervention joint state-action value function, known as Q-function, associated with the state vector $\mathbf{V}$ as:

$$Q^a_{\mathbf{V}}(\mathbf{x}, \mathbf{a}, \mathbf{u}) = \mathbb{E}_{\mathbf{x}'|\mathbf{x},\mathbf{a},\mathbf{u}}\left[R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') + \gamma \mathbf{V}(\mathbf{x}')\right], \quad (14)$$

for $\mathbf{x} \in \mathcal{X}, \mathbf{a} \in \mathcal{A}$ and $\mathbf{u} \in \mathcal{U}$, where $Q^a_{\mathbf{V}}(\mathbf{x}, ., .)$ can be seen as a matrix in $\mathbb{R}^{|\mathcal{A}|\times|\mathcal{U}|}$, and the expectation is with respect to the next system state. The Q-value identifies the anticipated intervention rewards when the joint actions $(\mathbf{a}, \mathbf{u})$ are chosen at state $\mathbf{x}$, and subsequently, the policy corresponding to the state value function $\mathbf{V}$ is followed.

For cell and intervention actions $(\mathbf{a}, \mathbf{u})$, we define the corresponding *transition matrix* as:

$$\begin{aligned}(M(\mathbf{a}, \mathbf{u}))_{ij} &= P\left(\mathbf{x}_k = \mathbf{x}^j \mid \mathbf{x}_{k-1} = \mathbf{x}^i, \mathbf{a}_{k-1} = \mathbf{a}, \mathbf{u}_{k-1} = \mathbf{u}\right) \\ &= p^{\|\mathbf{f}(\mathbf{x}^i) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j\|_1} (1-p)^{d - \|\mathbf{f}(\mathbf{x}^i) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j\|_1},\end{aligned} \quad (15)$$

for $i, j = 1, \ldots, 2^d$, $\mathbf{a} \in \mathcal{A}$, and $\mathbf{u} \in \mathcal{U}$, where $\|.\|_1$ represents the L-1 norm of a vector. In the absence of noise, $\mathbf{f}(\mathbf{x}^i) \oplus \mathbf{a} \oplus \mathbf{u}$ would represent the genes' state in the next time step. Therefore, $\|\mathbf{f}(\mathbf{x}^i) \oplus \mathbf{a} \oplus \mathbf{u} \oplus \mathbf{x}^j\|_1$ counts the number of flips in genes activities caused by noise when the system transitions from state $\mathbf{x}^i$ to state $\mathbf{x}^j$.

Let $R^a_{\mathbf{a},\mathbf{u}} = [R^a(\mathbf{x}^1, \mathbf{a}, \mathbf{u}), ..., R^a(\mathbf{x}^{2^d}, \mathbf{a}, \mathbf{u})]^T$ be the vector for the expected intervention reward, with ith element represented as:

$$\begin{aligned}R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}) &= \mathbb{E}_{\mathbf{x}'|\mathbf{x},\mathbf{a},\mathbf{u}}[R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}, \mathbf{x}')] \\ &= \sum_{j=1}^{2^d} R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}, \mathbf{x}^j) \\ &\quad \times P(\mathbf{x}_k = \mathbf{x}^j \mid \mathbf{x}_{k-1} = \mathbf{x}^i, \mathbf{a}_{k-1} = \mathbf{a}, \mathbf{u}_{k-1} = \mathbf{u}) \\ &= \sum_{j=1}^{2^d} R^a(\mathbf{x}^i, \mathbf{a}, \mathbf{u}, \mathbf{x}^j)\,(M(\mathbf{a}, \mathbf{u}))_{ij}.\end{aligned} \quad (16)$$

The Q-values in (14) can be expressed according to the controlled transition matrix $M(\mathbf{a}, \mathbf{u})$ and the vector-form expected reward function $R^a_{\mathbf{a},\mathbf{u}}$ as:

$$\begin{bmatrix} Q^a_{\mathbf{V}}(\mathbf{x}^1, \mathbf{a}, \mathbf{u}) \\ \vdots \\ Q^a_{\mathbf{V}}(\mathbf{x}^{2^d}, \mathbf{a}, \mathbf{u}) \end{bmatrix} = R^a_{\mathbf{a},\mathbf{u}} + \gamma M(\mathbf{a}, \mathbf{u})\,\mathbf{V}, \text{ for } \mathbf{a} \in \mathcal{A}, \mathbf{u} \in \mathcal{U}. \quad (17)$$

We define the Bellman operator $\mathcal{T}^*$ for any $\mathbf{x} \in \mathcal{X}$ as:

$$\begin{aligned}(\mathcal{T}^*[\mathbf{V}])(\mathbf{x}) &= \text{Value}[Q^a_{\mathbf{V}}(\mathbf{x}, ., .)] \\ &= \max_{\pi^a} \min_{\pi^u} \sum_{\mathbf{a}\in\mathcal{A}} \sum_{\mathbf{u}\in\mathcal{U}} \pi^a(\mathbf{a}|\mathbf{x})\,\pi^u(\mathbf{u}|\mathbf{x})\,Q^a_{\mathbf{V}}(\mathbf{x}, \mathbf{a}, \mathbf{u}),\end{aligned} \quad (18)$$

which should satisfy the condition $\sum_{\mathbf{a}\in\mathcal{A}} \pi^a(\mathbf{a}|\mathbf{x}) = \sum_{\mathbf{u}\in\mathcal{U}} \pi^u(\mathbf{u}|\mathbf{x}) = 1$. Linear programming techniques can be employed to calculate Value$[Q^a_{\mathbf{V}}(\mathbf{x}, ., .)]$ in (18).

Since the Bellman operator serves as a γ-contraction mapping for any arbitrary Markov game, we can start with an initial $\mathbf{V}_0$ and successively apply $\mathbf{V}_{t+1} = \mathcal{T}^*[\mathbf{V}_t]$ for $t = 0, 1, ...$ until a fixed vector is obtained. The fixed-point solution $\mathbf{V}^*$ corresponds to the optimal state value vector and the optimal Nash equilibrium policy (π^a_*, π^u_*). The optimal Nash policy corresponding to $\mathbf{V}^*$ can be obtained as:

$$\begin{aligned}&(\pi^a_*(.|\mathbf{x}), \pi^u_*(.|\mathbf{x})) \\ &= \operatorname*{argmax}_{\pi^a} \operatorname*{argmin}_{\pi^u} \sum_{\mathbf{a}\in\mathcal{A}} \sum_{\mathbf{u}\in\mathcal{U}} \pi^a(\mathbf{a}|\mathbf{x})\,\pi^u(\mathbf{u}|\mathbf{x})\,Q^a_{\mathbf{V}^*}(\mathbf{x}, \mathbf{a}, \mathbf{u}).\end{aligned} \quad (19)$$

for $\mathbf{x} \in \mathcal{X}$, where $Q^a_{\mathbf{V}^*}$ can be computed by using $\mathbf{V}^*$ in (17).

B. Online Step

This section provides a recursive computation method for the optimal MMSE state estimator. Consider a matrix $\mathcal{S} = [\mathbf{x}^1 = [0, \ldots, 0]^T, \ldots, \mathbf{x}^{2^d} = [1, \ldots, 1]^T]$ of size $d \times 2^d$, containing all possible system states. We define the vector form of the state posterior distribution at time step k, given the information available up to time step k, as:

$$\mathbf{\Pi}_{k|k}(i) = p(\mathbf{x}_k = \mathbf{x}^i \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k}), \quad \text{for} \quad i = 1, \ldots, 2^d. \quad (20)$$

According to (7), (10) and (20), the optimal MMSE state estimator can be computed as:

$$\hat{\mathbf{x}}^{\text{MS}}_{k|k} = \overline{\mathbb{E}\left[\mathbf{x}_k \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k}\right]} = \overline{\mathcal{S}\mathbf{\Pi}_{k|k}}. \quad (21)$$

We also define the predictive state posterior distribution vector as $\mathbf{\Pi}_{k+1|k} = p(\mathbf{x}_{k+1} = \mathbf{x}^i \mid \mathbf{a}_{0:k}, \mathbf{y}_{1:k})$. This vector specifies the probability distribution of state at time step $k+1$ given the information up to time step k. Using (13) and the controlled transition matrix in (15), the predictive posterior can be calculated as:

$$\mathbf{\Pi}_{k+1|k} = \sum_{\mathbf{u}\in\mathcal{U}} \left(M(\mathbf{a}_k, \mathbf{u})\mathbf{\Pi}_{k|k}\right) \circ p_{\mathbf{u}}, \quad (22)$$

where $p_{\mathbf{u}} = [\pi^u_*(\mathbf{u} \mid \mathbf{x}^1), ..., \pi^u_*(\mathbf{u} \mid \mathbf{x}^{2^d})]^T$ is a vector of size $2^d \times 1$ with the i'th element equal to the probability of cell action $\mathbf{u}$ at state $\mathbf{x}^i$ in the optimal Nash equilibrium policy, and $\circ$ is the component-wise multiplication of two vectors.

We define *update vector* $T(\mathbf{y}_{k+1})$, given the observation vector $\mathbf{y}_{k+1}$, as:

$$(T(\mathbf{y}_{k+1}))_i = p(\mathbf{y}_{k+1} \mid \mathbf{x}_{k+1} = \mathbf{x}^i), \quad i = 1, \ldots, 2^d. \quad (23)$$

According to the measurement model described in (3), the i'th element of the update vector can be calculated as:

$$(T(\mathbf{y}_{k+1}))_i = \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\mathbf{y}_{k+1}(l) - m - \delta\mathbf{x}^i(l)\right)^2}{2\sigma^2}\right), \quad (24)$$

Finally, according to (11) and (24), the new posterior distribution, $\mathbf{\Pi}_{k+1|k+1}$, can be recursively computed as [31], [32]:

$$\mathbf{\Pi}_{k+1|k+1} = \frac{T(\mathbf{y}_{k+1}) \circ \left(\sum_{\mathbf{u}\in\mathcal{U}} \left(M(\mathbf{a}_k, \mathbf{u})\mathbf{\Pi}_{k|k}\right) \circ p_{\mathbf{u}}\right)}{\|T(\mathbf{y}_{k+1}) \circ \left(\sum_{\mathbf{u}\in\mathcal{U}} \left(M(\mathbf{a}_k, \mathbf{u})\mathbf{\Pi}_{k|k}\right) \circ p_{\mathbf{u}}\right)\|_1}. \quad (25)$$

Using the posterior distribution of state, the next intervention policy in (9) can be expressed as:

$$\mathbf{a}_{k+1} \sim \pi^a_*(. \mid \overline{\mathcal{S}\mathbf{\Pi}_{k+1|k+1}}). \tag{26}$$

The online process has a complexity of order $O(2^{2d} \times |\mathcal{U}|)$ due to the involvement of the controlled transition matrix and summation over cell actions. A major computational complexity occurs during the offline computation of the Nash intervention policy before the intervention begins.

V. Comparison with State-of-Art Methods

This section analyzes the performance of the proposed intervention policy within a well-known class of intervention policies [33]–[36]. These methods rely on the stationary assumption of the intervention process, where cells do not respond to interventions. This can be expressed using the cell space $\mathcal{U} = \{\}$, where the Markov game can be represented as an MDP with a single player, i.e., intervention. The intervention policy is deterministic in this case, as no competition with the cell is considered in deriving the intervention. This deterministic policy can be expressed as:

$$\mu^a(\mathbf{x}) = \underset{\mu}{\operatorname{argmax}} \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t R^a(\mathbf{x}_t, \mathbf{a}_t, \mathbf{u}_t = \mathbf{0}, \mathbf{x}_{t+1}) \mid \mathbf{x}_0 = \mathbf{x}, \mathbf{a}_{0:\infty} \sim \mu \Bigg], \tag{27}$$

where the maximization is over all deterministic action policies, i.e., $(\mathcal{A})^{2^d}$. The cell's aggressive response to the stationary and deterministic intervention policy in (27) can be expressed as:

$$\mu^u(\mathbf{x}) = \underset{\mu}{\operatorname{argmin}}\, \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t R^a(\mathbf{x}_t, \mathbf{a}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) \mid \mathbf{x}_0 = \mathbf{x}, \mathbf{a}_{0:\infty} \sim \mu_a, \mathbf{u}_{0:\infty} \sim \mu \Bigg]. \tag{28}$$

According to equation (5), one can observe that the stationary policy μ^a deviates from the optimal Nash policy π^a, which has caused the cell to change its policy in order to achieve higher accumulated rewards. The deviation from the optimal Nash intervention policy leads to the dominance of the cell. The difference between the stationary and the optimal Nash policies in terms of the expected discounted rewards, if the system starts at state $\mathbf{x} \in \mathcal{X}$, can be expressed as:

$$\mathbf{V}^*_{\mu_a,\mu_u}(\mathbf{x}) - \mathbf{V}^*_{\pi^a_*,\pi^u_*}(\mathbf{x}) \leq 0, \text{ for } \mathbf{x} \in \mathcal{X}, \tag{29}$$

where the inequality might become equality if and only if the optimal Nash policy is deterministic.

Let $p(\mathbf{x}_k = \mathbf{x}^i \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k})$ be the posterior distribution of the state at time step k, and $\hat{\mathbf{x}}^{\mathrm{MS}}_{k|k}$ be the optimal MMSE state estimator. The deviation of the optimal Nash policy from the proposed policy can be expressed using the Kullback–Leibler (KL) divergence as:

$$\mathrm{KL}\left(\pi^a_*(.|\mathbf{x}_k) \,||\, \pi^a_*(.|\hat{\mathbf{x}}^{\mathrm{MS}}_{k|k})\right). \tag{30}$$

The KL value becomes zero for the case where $\mathbf{x}_k = \hat{\mathbf{x}}^{\mathrm{MS}}_{k|k}$. One can see that the proposed intervention policy is stochastic, and its performance relies on the accuracy of the state estimation. The probability that the MMSE state estimator might not be the true state can be expressed using the posterior distribution of the state as $1 - P(\mathbf{x}_k = \hat{\mathbf{x}}^{\mathrm{MS}}_{k|k} \mid \mathbf{a}_{0:k-1}, \mathbf{y}_{1:k})$. Meanwhile, the expected error of the optimal MMSE state estimator in terms of mean square error can be computed using (8). Therefore, these two metrics can be used to measure the confidence in the MMSE state estimator and, consequently, the closeness of the proposed policy to the optimal Nash policy.

VI. Numerical Experiments

The performance of the proposed intervention policy is evaluated in this section using a well-known melanoma regulatory network [37], [38]. This network is associated with a dangerous type of skin cancer called melanoma, which grows and spreads on a molecular level. The relationships between the genes in this network are shown in Fig. 1. The network consists of the following 10 genes: WNT5A, pirin, S100P, RET1, MMP3, PHOC, MART1, HADHB, synuclein, and STC2. 10 genes in this network leads to $2^{10} = 1,024$ possible system states. The network function representing the system dynamics in (1) can be expressed as [37], [38]:

$$\mathbf{f}(\mathbf{x}_k) = [f_1(\mathbf{x}_k), f_2(\mathbf{x}_k), ..., f_{10}(\mathbf{x}_k)]^T = \begin{bmatrix} (\text{S100P} \wedge \text{MMP3} \wedge \neg\text{PHOC}) \vee (\neg\text{MMP3} \wedge \text{PHOC}) \\ (\neg\text{WNT5A} \wedge \neg\text{S100P} \wedge \text{MMP3}) \vee (\text{WNT5A} \wedge \neg\text{S100P} \wedge \neg\text{MMP3}) \\ \text{MART1} \\ (\neg\text{WNT5A} \wedge \text{pirin} \wedge \text{RET1}) \vee (\neg\text{pirin} \wedge \text{RET1}) \\ (\text{RET1} \wedge \text{synuclein}) \vee \neg\text{synuclein} \\ (\neg\text{RET1} \wedge \neg\text{MART1}) \vee (\text{RET1} \wedge \text{MART1} \wedge \text{STC2}) \\ \text{MART1} \\ (\text{WNT5A} \wedge \text{MMP3}) \vee (\neg\text{MMP3} \wedge \neg\text{synuclein}) \vee (\text{WNT5A} \wedge \neg\text{MMP3} \wedge \text{synuclein}) \\ (\neg\text{RET1} \wedge \neg\text{MART1} \wedge \neg\text{STC2}) \vee (\text{RET1} \wedge \neg\text{MART1} \wedge \text{STC2}) \vee \text{MART1} \\ \neg\text{S100P} \end{bmatrix}$$

where $\wedge$ and $\vee$ represent AND and OR operators, and $\neg$ indicates negation.

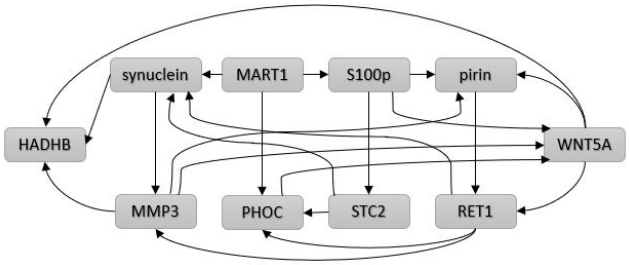

Fig. 1: The melanoma regulatory network containing 10 genes.

The increase in activation of WNT5A and pirin activation is shown to be associated with the metastasis state [37]. Thus, the intervention needs to decrease these genes activations, which can be expressed using the following intervention reward function:

$$R^a(\mathbf{x}, \mathbf{a}, \mathbf{u}, \mathbf{x}') = 2 - \mathbf{x}'(1) - \mathbf{x}'(2). \tag{31}$$

where the maximum reward is 2 occurs when both WNT5A and prin stay inactivated, and the minimum is 0 when both genes are in an activated state. The following intervention space is considered: $\mathcal{A} = \{\mathbf{a}^1, \mathbf{a}^2, \mathbf{a}^3, \mathbf{a}^4\}$, where $\mathbf{a}^1$ corresponds to no control, and $\mathbf{a}^2$, $\mathbf{a}^3$ and $\mathbf{a}^4$ correspond to an intervention over the pirin, RET1 and PHOC genes, respectively (e.g., $\mathbf{a}^2 = [0,1,0,0,0,0,0,0,0,0]^T$). The space for internal cell stimuli is considered as: $\mathcal{U} = \{\mathbf{u}^1, \mathbf{u}^2\}$, where $\mathbf{u}^1$ and $\mathbf{u}^2$ indicate stimuli altering the state value of the MMP3 and MART1 genes, respectively.

The results of the proposed method are compared with two state-of-the-art intervention policies: the stationary intervention policy [33], [39], which assumes cell as a non-responsive entity, and robust intervention policy [36], [40], which considers the cell responses as part of stochasticity in the intervention process. We also compare the proposed method with a random intervention policy as well as the baseline policy. The baseline represents the optimal Nash policy for the system under direct state observability, which corresponds to the best intervention results that can be achieved for the system under partial state observability. The parameters used for our experiments are as follows: $p = 0.05$, $\gamma = 0.95$, $\epsilon = 0.01$, $m = 30$, $\delta = 20$, $\sigma^2 = 10$, $\mathbf{\Pi}_{0|0}(i) = 1/2^{10}$, for $i = 1, ..., 2^{10}$. The results are averaged over 100 independent runs, and the standard error of the mean is presented for the average results.

Fig. 2 shows the average intervention reward over time for various intervention policies. As expected, the baseline policy yields the highest average reward, representing the optimal intervention results achievable under intelligent cell responses. The proposed policy outperforms other methods relying on partial state observability. For large time steps, the performance of the proposed policy and the baseline becomes close, due to the fading of the uniform initial state distribution leading to more accurate state estimation. In contrast, the stationary intervention policy and the random policy exhibit poor performance due to their inability to consider cell responses in deriving interventions. The robust intervention policy outperforms the stationary policy, as it considers the non-stationarity caused by cell responses in decision-making. However, the robust policy's performance still falls short of the proposed policy, partly from a lack of consideration of the cell's responses in the long term for deriving interventions.

Fig. 3 represents the average error of state estimation obtained by the MMSE state estimator for the trajectories obtained under the proposed intervention policy. The state estimation error is computed as $\sum_{i=1}^{10} |\mathbf{x}_k(i) - \hat{\mathbf{x}}_{k|k}^{\mathrm{MS}}(i)|$, which measures the number of genes with wrong estimated status. One can see that the error is initially large due to the uniform initial state distribution. However, as more data is observed, the estimation error becomes small and close to 0. It should be noted that the state estimation near zero ensures the proposed policy's closeness to the optimal Nash policy.

Fig. 4 represents the average Kullback-Leibler divergence between the optimal Nash policy (i.e., baseline policy) and different intervention policies. One can observe that the proposed policy has the smallest average KL values in comparison to the other two policies. In particular, the initial deviation is larger due to the large state estimation error, while the values become smaller as more data are observed. Interestingly, the stationary policy has the highest deviation among all methods (including the random policy), which results from the lack of consideration of the cell responses. The robust policy has less KL distance compared to others, but its values are still larger than the proposed policy. This is due to the deterministic nature of this policy, despite considering the non-stationarity coming from the cell responses.

The performance of the proposed policy with respect to the measurement noise (representing noise in gene expression data) is investigated in this part. Fig. 5 shows the average reward per step obtained by different policies under four different measurement noise levels. The proposed policy's results are close to the baseline policy under a small noise intensity (i.e., $\sigma^2 = 5$). This is due to a good state estimation performance and, consequently, the closeness of the

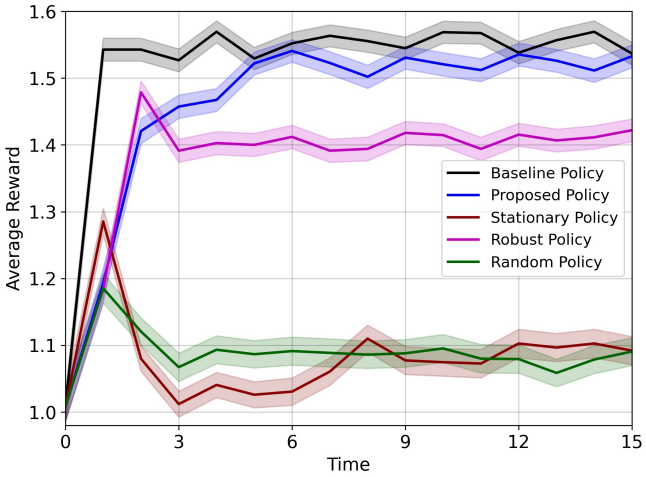

Fig. 2: The average reward obtained by different policies.

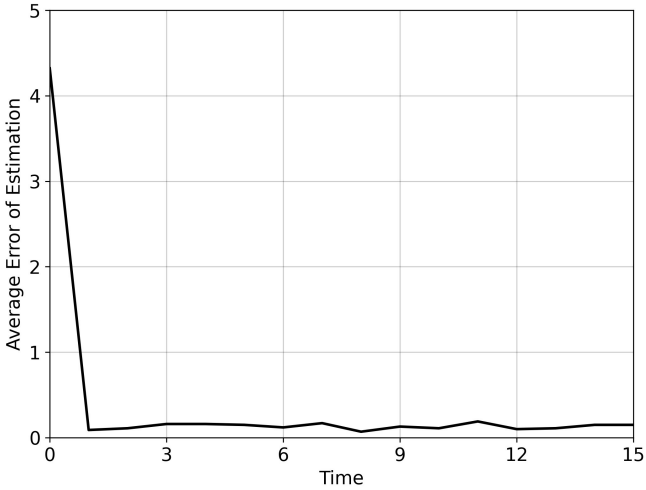

Fig. 3: The average state estimation error for trajectories obtained under the proposed policy.

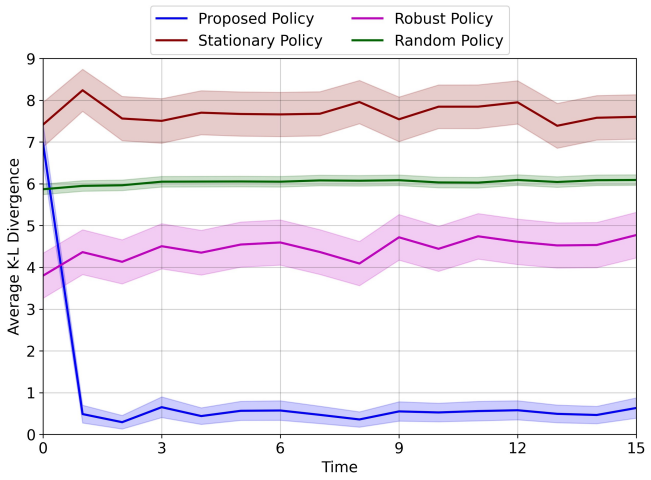

Fig. 4: The average KL-divergence between the optimal Nash policy (baseline) and other intervention policies.

proposed policy to the Nash policy. On the other hand, as measurement noise intensity increases, the performance of the proposed policy decreases. This is due to the fact that the true system state becomes indistinguishable under high measurement noise, which impacts both the state estimation and the intervention performances. One can also see that the proposed policy outperforms all competing methods for all noise values.

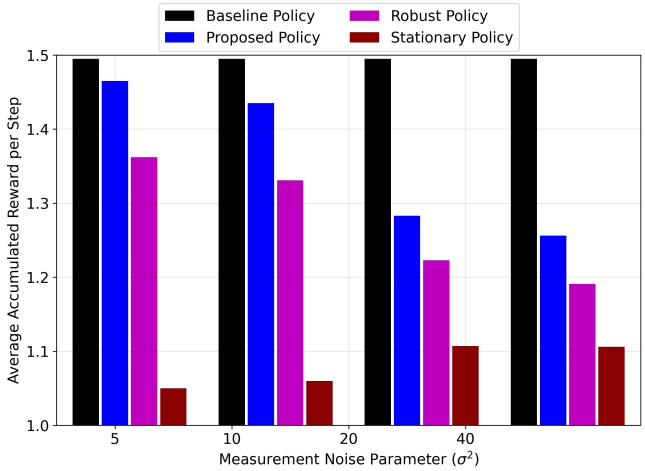

Fig. 5: The impact of the measurement noise on the performance of the proposed policy.

Finally, this section investigates the impact of the stochastic state process on the performance of the proposed method. We considered four process noise intensities, $p = 0.005, 0.05, 0.15, 0.3$, where larger values represent more chaotic systems. Table I represents the average reward obtained by all policies. It can be seen that the closest performance to the baseline policy is achieved by the proposed policy. The robust intervention policy yields the second-best results, and the stationary policy holds the worst results. One can observe that the increase in process noise has led to a reduction in average reward. In particular, for chaotic systems corresponding to $p = 0.3$, the results of all methods become similar, as the control or intervention, in this case, has less power to impact the system. However, for smaller noise intensity, the proposed method outperforms other competing methods.

TABLE I: Impact of process noise on the performance of the proposed policy.

	Process Noise (p)			
Policy	0.005	0.05	0.15	0.3
Baseline	1.54 ± 0.11	1.49 ± 0.13	1.40 ± 0.11	1.22 ± 0.06
Proposed	1.46 ± 0.14	1.42 ± 0.15	1.35 ± 0.11	1.21 ± 0.07
Robust	1.39 ± 0.09	1.37 ± 0.12	1.33 ± 0.10	1.21 ± 0.06
Stationary	1.03 ± 0.03	1.07 ± 0.07	1.11 ± 0.04	1.16 ± 0.03

VII. Conclusion

This paper introduces a stochastic intervention policy for gene regulatory networks (GRNs) with partial state observability. It models GRNs observed through gene-expression data using a two-player zero-sum game with binary state variables. A recursive approach is derived to compute the posterior distribution of genes' state, given the unmeasured cell stimuli and the available gene-expression data. The optimal minimum mean square error (MMSE) is obtained according to the posterior distribution of the state. The proposed intervention policy employs the optimal Nash policy associated with the MMSE state estimator at each time. This state-feedback control procedure ensures the stochasticity of the intervention policy, in contrast to most existing intervention policies that are deterministic. Our analytical and numerical results demonstrate a comparison between the proposed methods and existing approaches, as well as the empirical convergence of the proposed policy to the optimal Nash policy.

Acknowledgment

The authors acknowledge the support of the National Institute of Health award 1R21EB032480-01, National Science Foundation awards IIS-2311969 and IIS-2202395, ARMY Research Laboratory award W911NF2320179, ARMY Research Office award W911NF2110299, and Office of Naval Research award N00014-23-1-2850.

References

[1] N. Taou and M. Lones, "Optimising Boolean synthetic regulatory networks to control cell states," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 6, pp. 2649–2658, 2021.

[2] M. Alali and M. Imani, "Reinforcement learning data-acquiring for causal inference of regulatory networks," in *2023 American Control Conference (ACC)*, pp. 3957–3964, 2023.

[3] A. H. Abolmasoumi, M. Mohammadian, and L. Mili, "Robust Kalman filter state estimation for gene regulatory networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 2, pp. 1395–1405, 2023.

[4] A. Ravari, S. F. Ghoreishi, and M. Imani, "Optimal recursive expert-enabled inference in regulatory networks," *IEEE Control Systems Letters*, vol. 7, pp. 1027–1032, 2023.

[5] M. Alali and M. Imani, "Inference of regulatory networks through temporally sparse data," *Frontiers in control engineering*, vol. 3, 2022.

[6] E. R. Dougherty, R. Pal, X. Qian, M. L. Bittner, and A. Datta, "Stationary and structural control in gene regulatory networks: basic concepts," *International Journal of Systems Science*, vol. 41, no. 1, pp. 5–16, 2010.

[7] Q. Liu, Y. He, and J. Wang, "Optimal control for probabilistic Boolean networks using discrete-time Markov decision processes," *Physica A: Statistical Mechanics and its Applications*, vol. 503, pp. 1297–1307, 2018.

[8] R. Zandi, "Sparse coding for data augmentation of hyperspectral medical images," *San Jose State University*, 2021.

[9] M. Takizawa, K. Kobayashi, and Y. Yamashita, "Design of reduced-order and pinning controllers for probabilistic Boolean networks using reinforcement learning," *Applied Mathematics and Computation*, vol. 457, p. 128211, 2023.

[10] M. Imani and U. Braga-Neto, "Optimal control of gene regulatory networks with unknown cost function," in *2018 Annual American Control Conference (ACC)*, pp. 3939–3944, IEEE, 2018.

[11] A. Khan, G. Saha, and R. K. Pal, "Controlling the effects of external perturbations on a gene regulatory network using proportional-integral-derivative controller," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 3, pp. 1531–1544, 2022.

[12] M. Imani and S. F. Ghoreishi, "Optimal finite-horizon perturbation policy for inference of gene regulatory networks," *IEEE Intelligent Systems*, vol. 36, no. 1, pp. 54–63, 2020.

[13] M. Imani and U. Braga-Neto, "Point-based value iteration for partially-observed Boolean dynamical systems with finite observation space," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 4208–4213, IEEE, 2016.

[14] L. Van den Broeck, M. Gordon, D. Inzé, C. Williams, and R. Sozzani, "Gene regulatory network inference: connecting plant biology and mathematical modeling," *Frontiers in genetics*, vol. 11, p. 457, 2020.

[15] M. Imani, R. Dehghannasiri, U. M. Braga-Neto, and E. R. Dougherty, "Sequential experimental design for optimal structural intervention in gene regulatory networks based on the mean objective cost of uncertainty," *Cancer informatics*, vol. 17, p. 1176935118790247, 2018.

[16] J. Liang and J. Han, "Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks," *BMC systems biology*, vol. 6, no. 1, pp. 1–21, 2012.

[17] S. H. Hosseini and M. Imani, "Learning to fight against cell stimuli: A game theoretic perspective," in *2023 IEEE Conference on Artificial Intelligence (CAI)*, pp. 285–287, IEEE, 2023.

[18] S. H. Hosseini and M. Imani, "Modeling defensive response of cells to therapies: Equilibrium interventions for regulatory networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2024.

[19] M. Imani and U. M. Braga-Neto, "Maximum-likelihood adaptive filter for partially observed Boolean dynamical systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 359–371, 2017.

[20] L. D. McClenny, M. Imani, and U. Braga-Neto, "Boolfilter package vignette," 2017.

[21] S. H. Hosseini and M. Imani, "An optimal Bayesian intervention policy in response to unknown dynamic cell stimuli," *Information Sciences*, vol. 666, p. 120440, 2024.

[22] M. Imani and U. M. Braga-Neto, "Particle filters for partially-observed Boolean dynamical systems," *Automatica*, vol. 87, pp. 238–250, 2018.

[23] M. Alali and M. Imani, "Kernel-based particle filtering for scalable inference in partially observed boolean dynamical systems," in *IFAC-PapersOnLine, 20th IFAC Symposium on System Identification (SYSID 2024)*, Elsevier, 2024.

[24] A. Kazeminajafabadi and M. Imani, "Optimal joint defense and monitoring for networks security under uncertainty: A pomdp-based approach," *IET Information Security*, vol. 2024, no. 1, p. 7966713, 2024.

[25] A. Ravari, S. F. Ghoreishi, and M. Imani, "Structure-based inverse reinforcement learning for quantification of biological knowledge," in *IEEE Conference on Artificial Intelligence*, 2023.

[26] A. Ravari, S. F. Ghoreishi, and M. Imani, "Optimal inference of hidden Markov models through expert-acquired data," *IEEE Transactions on Artificial Intelligence*, 2024.

[27] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *Journal of Biomedical optics*, vol. 2, no. 4, pp. 364–374, 1997.

[28] J. Hua, C. Sima, M. Cypert, G. C. Gooden, S. Shack, L. Alla, E. A. Smith, J. M. Trent, E. R. Dougherty, and M. L. Bittner, "Dynamical analysis of drug efficacy and mechanism of action using GFP reporters," *Journal of Biological Systems*, vol. 20, no. 04, pp. 403–422, 2012.

[29] K. Zhang, Z. Yang, and T. Bacsar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

[30] M. Imani and U. Braga-Neto, "Gene regulatory network state estimation from arbitrary correlated measurements," *EURASIP Journal on Advances in Signal Processing*, vol. 2018, pp. 1–10, 2018.

[31] A. Kazeminajafabadi, S. F. Ghoreishi, and M. Imani, "Optimal detection for Bayesian attack graphs under uncertainty in monitoring and reimaging," *2023 American Control Conference (ACC)*, no. 1, 2024.

[32] A. Kazeminajafabadi and M. Imani, "Optimal monitoring and attack detection of networks modeled by Bayesian attack graphs," *Cybersecurity*, vol. 6, no. 1, p. 22, 2023.

[33] R. Pal, A. Datta, and E. R. Dougherty, "Optimal infinite-horizon control for probabilistic Boolean networks," *Signal Processing, IEEE Transactions on*, vol. 54, no. 6, pp. 2375–2387, 2006.

[34] A. Pezzotta and J. Briscoe, "Optimal control of gene regulatory networks for morphogen-driven tissue patterning," *Cell Systems*, vol. 14, no. 11, pp. 940–952, 2023.

[35] M. Imani and U. M. Braga-Neto, "Control of gene regulatory networks using Bayesian inverse reinforcement learning," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 4, pp. 1250–1261, 2018.

[36] H. Li, X. Yang, and S. Wang, "Robustness for stability and stabilization of boolean networks with stochastic function perturbations," *IEEE Transactions on Automatic Control*, vol. 66, no. 3, pp. 1231–1237, 2020.

[37] X. Qian and E. R. Dougherty, "Intervention in gene regulatory networks via phenotypically constrained control policies based on long-run behavior," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 1, pp. 123–136, 2011.

[38] A. T. Weeraratna, Y. Jiang, G. Hostetter, K. Rosenblatt, P. Duray, M. Bittner, and J. M. Trent, "Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma," *Cancer cell*, vol. 1, no. 3, pp. 279–288, 2002.

[39] M. Imani and U. Braga-Neto, "State-feedback control of partially-observed Boolean dynamical systems using RNA-seq time series data," in *2016 American Control Conference (ACC)*, pp. 227–232, IEEE, 2016.

[40] R. Pal, A. Datta, and E. R. Dougherty, "Robust intervention in probabilistic boolean networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1280–1294, 2008.