Kullback–Leibler Control in Boolean Control Networks

Mitsuru Toyoda, *Member, IEEE*, and Yuhu Wu, *Member, IEEE*

***Abstract*—This article addresses the Kullback–Leibler (KL) control problem in Boolean control networks. In the considered problem, an extended stage cost function depending on the control inputs is introduced; in contrast to a stage cost of the conventional KL control problems in the Markov decision process cannot take into consideration the control inputs. An associated Bellman equation and a matrix-based iteration algorithm are presented. The theoretical analysis shows that the proposed KL control results in an approximated form of conventional dynamic programming (DP). Furthermore, the convergence analysis is presented, with the weight parameter converging to zero and diverging to infinity. In practical application examples, a comparison of the conventional DP and proposed KL control is illustrated.**

***Index Terms*—Boolean control networks (BCNs), convergence analysis, gene regulatory networks, Kullback–Leibler (KL) control, optimal control, semi-tensor product (STP) of matrices.**

I. Introduction

LOGICAL dynamic systems have been employed for modeling complicated dynamical behavior, such as biological systems [1], combustion engines [2], and transportation systems [3]. The mathematically tractable structure of Boolean control networks (BCNs) [4] has resulted in their widespread use. After the development of the semi-tensor product (STP) technique [5], [6], Boolean networks have been analyzed in terms of their matrix-based expressions. With respect to BCNs, theoretical concepts similar to those of continuous-valued systems and related analyses have been developed: stabilization problem [7], estimation problem [8], [9], decoupling problem [10], robustness analysis [11], [12], conservation law [13], Lyapunov function [14], [15], and reinforcement learning [16]. To satisfy various control objectives, several control structure types have been reported. A common closed-loop structure has been frequently used for the stabilization: stabilization under random switching [17] and pinning control purpose [18], [19]. Event-based [20], [21] and sample-value [22], [23] control schemes, which have been explored in the continuous-valued systems analysis, have been developed for BCNs. In contrast to these settings, the time-varying feedback law is exploited in the optimal control formulation [16], [24], as discussed in this article.

In practical systems, as there exist some preferred and undesired states [25], control schemes that avoid forbidden states in biological systems [26] or obstacles in robotic systems [27] have been discussed. In these reports, feedback raw is provided for making the transition probabilities to the forbidden state equal zero; this technique is similar to dynamic programming (DP) [28] with a modified stage cost having a value of the infinity, which is discussed in this article. However, the aforementioned schemes are specialized to make the transition probabilities zero, and they cannot quantitatively evaluate the probabilistic transitions to preferred and undesired states. The development of a control scheme quantitatively addressing the transition probabilities while minimizing the given objective function with optimal control is still an open and challenging issue.

Based on the background summarized above, in this study, the Kullback–Leibler (KL) divergence, which evaluates the similarity of two given probability distributions, is used, and the KL control problem formulation [29] is applied to BCNs with an extended stage cost function taking into consideration the control input.

In the conventional Markov decision process (MDP), it is assumed that the KL control can control the transition probabilities directly. This assumption holds in various trajectory planning problems, such as the maze game (see [30]) the trajectory planning of robots (see [27]), wherein the control input of the problem is the transition of the state itself. In BCNs, the state transition depends on the system structure, and the relationship between an applied input and the resulting state transition probability is complicated; therefore, a theoretical analysis and a practical code implementation for the BCNs result in difficult jobs. To the best of the authors' knowledge, optimal control problems with an extended stage cost function taking into consideration the control input and the corresponding Bellman equation have not been explored and is still an open issue. In addition, although the existing KL control is successively implemented in a matrix-based form, which is computationally efficient in recent programming languages, a matrix-based expression of the aforementioned formulation, including the extended stage cost function, is required to be developed.

The contribution of this article is broadly summarized as follows.

1) The optimal control problem with the KL divergence and extended stage cost function depending on the control input is addressed; in contrast, the existing KL control (see [27], [29]) has a stage cost that is dependent only

Manuscript received 24 March 2023; revised 8 June 2023; accepted 30 June 2023. This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP19K15019, and in part by the National Natural Science Foundation of China under Grant 62173062 and Grant 61773090. This article was recommended by Associate Editor Z. Li. *(Corresponding author: Mitsuru Toyoda.)*

Mitsuru Toyoda is with the Department of Mechanical Systems Engineering, Tokyo Metropolitan University, Tokyo 191-0065, Japan (e-mail: toyoda@tmu.ac.jp).

Yuhu Wu is with the Key Laboratory of Intelligent Control and Optimization for Industrial Equipment of Ministry of Education and the School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: wuyuhu@dlut.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2023.3292819.

Digital Object Identifier 10.1109/TCYB.2023.3292819

on the state. That is, the considered problem has a high modeling capacity. This article presents the corresponding Bellman equation and the matrix-based iteration of the proposed algorithm.

2) The theoretical analysis provides the new insight that the solution of the aforementioned KL control problem results in the conventional DP modified by replacing the max and arg max operations with an extended log-sum-exp-based approximation function and extended softmax function, respectively. Furthermore, the theoretical support for the following two limiting cases is given as follows.
   a) If the weight parameter in the KL control scheme converges to zero, the value function and optimal control input obtained using the KL control converge to those of the conventional DP.
   b) If the weight parameter in the KL control scheme diverges to $+\infty$, the value function and optimal control input obtained using the KL control converge to those of the desired transition probabilities given for the KL divergence.

II. Preliminaries

The notations used in this article are summarized as follows.

1) For an $m \times n$-valued matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, its (i, j) element is denoted by $[\boldsymbol{A}]_{i,j}$ or $A_{i,j}$. The $m \times n$ zero matrix is denoted by $\boldsymbol{O}_{m\times n}$. The $n \times n$ identity is denoted by $\boldsymbol{I}_n$, and its ith column vector is denoted by $\boldsymbol{\delta}_n^i$. Its set is denoted by $\Delta^n = \{\boldsymbol{\delta}_n^i, i = 1, \ldots, n\}$.
2) A matrix $\mathrm{Diag}(\boldsymbol{a}) \in \mathbb{R}^{n\times n}$ calculated with a vector $\boldsymbol{a} \in \mathbb{R}^n$ is a diagonal matrix, the (i, i) element of which is a_i.
3) A matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, with all its column vectors belonging to Δ^m is called a logical matrix. $\mathbb{L}^{m\times n}$ is the set of all $m \times n$-valued logical matrices.
4) The Boolean domain, which comprises T (True) and F (False), is denoted by $\mathcal{D} = \{T = 1, F = 0\}$. T and F are identified with $\boldsymbol{\delta}_2^1$ and $\boldsymbol{\delta}_2^2$, respectively.
5) The STP [6] of matrices $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ is defined as $\boldsymbol{A} \ltimes \boldsymbol{B} = (\boldsymbol{A} \otimes \boldsymbol{I}_{s/n})(\boldsymbol{B} \otimes \boldsymbol{I}_{s/p})$, where $\otimes$ is the Kronecker product, and s is the least common multiple of n and p. This article omits the "$\ltimes$" in $\boldsymbol{A} \ltimes \boldsymbol{B}$ for notational simplicity.
6) A logical matrix $[\boldsymbol{\delta}_m^{i_1}, \ldots, \boldsymbol{\delta}_m^{i_n}] \in \mathbb{L}^{m\times n}$ is simply denoted by $\boldsymbol{\delta}_m[i_1, \ldots, i_n]$.
7) The n-dimensional unit simplex is denoted by $\mathbb{S}^n = \{\boldsymbol{x} \in \mathbb{R}^n | \sum_{i=1}^n x_i = 1, x_1, \ldots, x_n \geq 0\}$.
8) For a vector $\boldsymbol{a} \in \mathbb{R}^n$, $\min(\boldsymbol{a})$ and $\max(\boldsymbol{a})$ are the minimum and maximum elements of $\boldsymbol{a}$, respectively. The minimizing and maximizing arguments are defined as $\arg\min(\boldsymbol{a}) = \arg\min_{\boldsymbol{\delta}\in\mathbb{S}^n} \boldsymbol{\delta}^\top \boldsymbol{a}$ and $\arg\max(\boldsymbol{a}) = \arg\max_{\boldsymbol{\delta}\in\mathbb{S}^n} \boldsymbol{\delta}^\top \boldsymbol{a}$, respectively.
9) The inequality $\boldsymbol{a} \leq \boldsymbol{b}$ for vectors $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^n$ indicates that $a_i \leq b_i$ for each $i = 1, \ldots, n$.
10) In this article, the argument $x \in \mathbb{R} \cup \{-\infty\}$ of the exp operation is considered, and $\exp(-\infty) = 0$ is defined. In addition, $0 \times +\infty = 0$ and $0 \log 0 = \lim_{x\searrow 0} x \log x = 0$ are set formally, which makes the function $x \log x$ continuous and convex on $[0, +\infty)$.
11) In this article, mathematical operations for matrices and vectors are defined as elementwise operations, for example, $[\exp(\boldsymbol{A})]_{i,j} = \exp(A_{i,j})$.
12) $\odot$ and $\oslash$ are the elementwise product and division of matrices, respectively.
13) $P(A)$ is the probability of an event A, and $P(A|B)$ is the conditional probability of A under a condition B. $\mathbb{E}[X]$ is the expectation of the stochastic variable X, and $\mathbb{E}[X|B]$ is the conditional expectation of X under B.

Furthermore, the max operation approximation performed by the log-sum-exp function is summarized here. The log-sum-exp function $\mathtt{LSE}(\boldsymbol{x}, \mu) = \mu \log(\boldsymbol{1}_n^\top \exp(\boldsymbol{x}/\mu))$ for $\boldsymbol{x} \in \mathbb{R}^n$ and $\mu > 0$ (see [31, Example 10.45]) has been used in various research fields. Based on the definition of the log-sum-exp function, the extended log-sum-exp function $\mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}):(\mathbb{R} \cup \{-\infty\})^n \times \mathbb{R}_{>0} \times \mathbb{S}^n \to \mathbb{R}$ is defined as

$$\mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) = \mu \log\Big(\boldsymbol{p}^\top \exp(\boldsymbol{x}/\mu)\Big).$$

Furthermore, the extended softmax function $\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p}) : (\mathbb{R} \cup \{-\infty\})^n \times \mathbb{R}_{>0} \times \mathbb{S}^n \to \mathbb{S}^n$ is defined by

$$\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p}) = \frac{\boldsymbol{p} \odot \exp(\boldsymbol{x}/\mu)}{\boldsymbol{p}^\top \exp(\boldsymbol{x}/\mu)} = \frac{\partial\, \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p})}{\partial \boldsymbol{x}}$$

where a simple calculation provides the second equality, thus indicating that $\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})$ is the partial derivative of $\mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p})$ with respect to the variable $\boldsymbol{x}$ on $\mathbb{R}^n$.

Theorem 1: The extended log-sum-exp function $\mathtt{eLSE}$ satisfies the following inequalities.

1) For an arbitrary $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$

$$\begin{aligned}\mathtt{eLSE}\big(\boldsymbol{x}', \mu, \boldsymbol{p}\big) \geq{}& \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) \\ &+ \frac{\partial\, \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p})}{\partial \boldsymbol{x}^\top}\big(\boldsymbol{x}' - \boldsymbol{x}\big) \qquad (1)\end{aligned}$$

and

$$\begin{aligned}\mathtt{eLSE}\big(\boldsymbol{x}', \mu, \boldsymbol{p}\big) \leq{}& \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) \\ &+ \frac{\partial\, \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p})}{\partial \boldsymbol{x}^\top}\big(\boldsymbol{x}' - \boldsymbol{x}\big) \\ &+ \frac{1}{2\mu}\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2. \qquad (2)\end{aligned}$$

2) If $\min(\boldsymbol{p}) > 0$, for an arbitrary $\boldsymbol{x} \in \mathbb{R}^n$

$$\begin{aligned}\mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) &\leq \max(\boldsymbol{x}) \\ &\leq \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) + \mu \log(1/\min(\boldsymbol{p})). \qquad (3)\end{aligned}$$

3) For an arbitrary $\boldsymbol{x} \in \mathbb{R}^n$

$$\mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) - \frac{1}{2\mu}\|\boldsymbol{x}\|_2^2 \leq \boldsymbol{p}^\top \boldsymbol{x} \leq \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}). \qquad (4)$$

Proof: Item 1): The inequalities are obtained by evaluating the Hessian of $\mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p})$ with respect to $\boldsymbol{x}$ directly. That is

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{x}^\top}\frac{\partial\, \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p})}{\partial \boldsymbol{x}} &= \frac{\partial\, \mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})}{\partial \boldsymbol{x}^\top} \\ &= \mu^{-1}\big[\mathrm{Diag}(\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})) \\ &\quad -\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})\mathtt{esoftmax}^\top \\ &\quad (\boldsymbol{x}, \mu, \boldsymbol{p})\big]. \qquad (5)\end{aligned}$$

For the arbitrary vector $\boldsymbol{\xi} \in \mathbb{R}^n$, the quadratic form is calculated as

$$\begin{aligned}&\boldsymbol{\xi}^\top\big[\mathrm{Diag}(\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})) \\ &\quad -\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})\mathtt{esoftmax}^\top(\boldsymbol{x}, \mu, \boldsymbol{p})\big]\boldsymbol{\xi}\end{aligned}$$

$$= \left\| \left[\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})\right]^{\frac{1}{2}} \odot \boldsymbol{\xi} \right\|_2^2 \\ -|\mathtt{esoftmax}^\top(\boldsymbol{x}, \mu, \boldsymbol{p})\boldsymbol{\xi}|^2. \tag{6}$$

The following Cauchy–Schwarz inequality holds:

$$\begin{aligned} &|\mathtt{esoftmax}^\top(\boldsymbol{x}, \mu, \boldsymbol{p})\boldsymbol{\xi}| \\ &= |\left[\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})\right]^{\frac{1}{2},\top}\left[\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})\right]^{\frac{1}{2}} \odot \boldsymbol{\xi}| \\ &\le \|\left[\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})\right]^{\frac{1}{2}}\|_2 \cdot \|\left[\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})\right]^{\frac{1}{2}} \odot \boldsymbol{\xi}\|_2 \\ &= \|\left[\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})\right]^{\frac{1}{2}} \odot \boldsymbol{\xi}\|_2 \end{aligned} \tag{7}$$

where the equality $\|[\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})]^{1/2}\|_2 = \sqrt{\mathbf{1}_n^\top([\mathtt{esoftmax}(\boldsymbol{x}, \mu, \boldsymbol{p})]^{1/2})^2} = 1$ is used. On applying (7) to the quadratic form (6), the quadratic form becomes non-negative, which indicates that the Hessian is positive semi-definite. Therefore, the convexity of $\mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p})$ is demonstrated, and the resulting convex inequality (1) follows. As the second term $(*)(*)^\top$ of the Hessian (5) is positive semi-definite, the largest eigenvalue of the Hessian (5) is upper bounded by the largest eigenvalue of the first diagonal matrix, which means that $\mu^{-1}\max(\boldsymbol{p} \odot \exp(\boldsymbol{x}/\mu)/\boldsymbol{p}^\top \exp(\boldsymbol{x}/\mu)) \le \mu^{-1}$; the equivalence of the largest eigenvalue and Lipschitz smoothness (see [32, Sec. 2.1]) results in the inequality of (2) of Item 1).

Item 2): The flow is similar to that of the conventional log-sum-exp function ([31, Example 10.45]). The first inequality is given by

$$\begin{aligned} \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) &= \mu \log\left(\boldsymbol{p}^\top \exp(\boldsymbol{x}/\mu)\right) \\ &\le \mu \log\left(\boldsymbol{p}^\top \exp(\max(\boldsymbol{x})/\mu)\mathbf{1}_n\right) = \max(\boldsymbol{x}) \end{aligned}$$

where the last equality used $\boldsymbol{p}^\top\mathbf{1}_n = 1$ on recalling $\boldsymbol{p} \in \mathbb{S}^n$. On using $i^* \in \arg\max_{i=1,\ldots,n} x_i$, the following inequality results in the second inequality:

$$\begin{aligned} \max(\boldsymbol{x}) &= \mu \log\left(p_{i^*}^{-1} \cdot p_{i^*} \exp(\max(\boldsymbol{x})/\mu)\right) \\ &\le \mu \log\left(p_{i^*}^{-1}\boldsymbol{p}^\top \exp(\boldsymbol{x}/\mu)\right) \\ &= \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) + \mu \log\left(p_{i^*}^{-1}\right) \\ &\le \mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p}) + \mu \log(1/\min(\boldsymbol{p})). \end{aligned}$$

Item 3): Equation (2) of Item 1) is applied with $\boldsymbol{x} = \mathbf{0}_n$. The substitution of

$$\begin{cases} \mathtt{eLSE}(\mathbf{0}_n, \mu, \boldsymbol{p}) = \mu \log\left(\boldsymbol{p}^\top \exp(\mathbf{0}_n/\mu)\right) \\ \qquad = \mu \log\left(\boldsymbol{p}^\top \mathbf{1}_n\right) = 0, \\ \frac{\partial\, \mathtt{eLSE}(\mathbf{0}_n, \mu, \boldsymbol{p})}{\partial \boldsymbol{x}} = \mathtt{esoftmax}(\mathbf{0}_n, \mu, \boldsymbol{p}) \\ \qquad = \dfrac{\boldsymbol{p} \odot \exp(\mathbf{0}_n/\mu)}{\boldsymbol{p}^\top \exp(\mathbf{0}_n/\mu)} = \boldsymbol{p} \end{cases}$$

into (2) results in the first inequality of (4). On recalling $\boldsymbol{p} \in \mathbb{S}^n$ and the convexity of exp, the Jensen inequality indicates that $\boldsymbol{p}^\top \exp(\boldsymbol{x}/\mu) \ge \exp(\boldsymbol{p}^\top\boldsymbol{x}/\mu)$, which results in the second inequality of (4). ■

Remark 1: The conventional log-sum-exp function $\mathtt{LSE}(\boldsymbol{x}, \mu) = \mu \log(\mathbf{1}_n^\top \exp(\boldsymbol{x}/\mu))$ is a special case of the extended log-sum-exp function $\mathtt{eLSE}(\boldsymbol{x}, \mu, \boldsymbol{p})$ with $\boldsymbol{p} = \mathbf{1}_n/n \in \mathbb{S}^n$, where n is the dimension of $\boldsymbol{x}$.

III. Problem Formulation

This article is focused on BCNs with a dth state update represented as follows:

$$\boldsymbol{x}_{k+1,d} = \boldsymbol{f}_d(\boldsymbol{x}_k, \boldsymbol{u}_k),\ \ d = 1, \ldots, n_{\boldsymbol{x}}. \tag{8}$$

The state and control variables are expressed in the STPs as $\boldsymbol{x}_k = \boldsymbol{x}_{k,1} \ltimes \cdots \ltimes \boldsymbol{x}_{k,n_x} \in \Delta^{2^{n_{\boldsymbol{x}}}}$ and $\boldsymbol{u}_k = \boldsymbol{u}_{k,1} \ltimes \cdots \ltimes \boldsymbol{u}_{k,n_u} \in \Delta^{2^{n_{\boldsymbol{u}}}}$, respectively. Although conventional studies on BCNs have addressed the deterministic control input $\boldsymbol{u}_k \in \Delta^{2^{n_{\boldsymbol{u}}}}$, this study takes into consideration the randomized control input and the corresponding conditional probabilities of $\boldsymbol{u}_k$ under a given state $\boldsymbol{x}_k$ at the kth step as follows:

$$c_{k,i,l} = P\left(\boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_{\boldsymbol{x}}}}^l | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_{\boldsymbol{x}}}}^i\right), \\ i = 1, \ldots, 2^{n_{\boldsymbol{x}}},\ \ l = 1, \ldots, 2^{n_{\boldsymbol{u}}}.$$

An initial state $\boldsymbol{x}_0 \in \Delta^{2^{n_{\boldsymbol{x}}}}$ is deterministically given, and the design problem of $c_{k,i,l}$ is addressed. It should be noted that $\boldsymbol{c}_{k,i} = [c_{k,i,1}, \ldots, c_{k,i,2^{n_{\boldsymbol{u}}}}]^\top \in \mathbb{S}^{2^{n_{\boldsymbol{u}}}}$, which indicates that $\boldsymbol{c}_{k,i}$ should be a point on the unit simplex. Herein, $c_{k,i,l}\,(k = 0, \ldots, N-1,\ i = 1, \ldots, 2^{n_{\boldsymbol{x}}},\ l = 1, \ldots, 2^{n_{\boldsymbol{u}}})$ is referred to as a selection probability in analogy with the probabilistic BCNs, using a similar concept to that of random state switching [33]. Then, the structure matrix $\boldsymbol{M} \in \mathbb{L}^{2^{n_{\boldsymbol{x}}} \times 2^{n_{\boldsymbol{u}}+n_{\boldsymbol{x}}}}$ of the BCNs (8) uniquely exists and satisfies the state equation $\boldsymbol{x}_{k+1} = \boldsymbol{M}\boldsymbol{u}_k\boldsymbol{x}_k$ [6]. Although $\boldsymbol{x}_k$ depends on the design and stochastic behavior of the randomized control, it is simply denoted by $\boldsymbol{x}_k$ for notational simplicity herein. Next, a desired transition probability $\overline{P}(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k, \boldsymbol{u}_k)$ is given, and the difference between the desired and actual transition probabilities is introduced as the KL divergence

$$\mathrm{KL}\left(P(\cdot|\boldsymbol{x}_k)\middle|\overline{P}(\cdot|\boldsymbol{x}_k)\right) = \sum_{\substack{\boldsymbol{x}' \in \Delta^{2^{n_{\boldsymbol{x}}}}, \\ \overline{P}(\boldsymbol{x}'|\boldsymbol{x}_k) \neq 0}} P\left(\boldsymbol{x}'|\boldsymbol{x}_k\right) \log\left(\frac{P\left(\boldsymbol{x}'|\boldsymbol{x}_k\right)}{\overline{P}\left(\boldsymbol{x}'|\boldsymbol{x}_k\right)}\right). \tag{9}$$

An objective function is the sum of the objective function of the conventional optimal control problem and the KL divergence

$$\min_{\substack{\boldsymbol{c}_{k,i} \in \mathcal{C}_{k,i}, \\ k=0,\ldots,N-1, \\ i=1,\ldots,2^{n_{\boldsymbol{x}}}}} \mathbb{E}\left[\sum_{k=0}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + h(\boldsymbol{x}_N) \right. \\ \left. + \mu \sum_{k=0}^{N-1} \mathrm{KL}\left(P(\cdot|\boldsymbol{x}_k)\middle|\overline{P}(\cdot|\boldsymbol{x}_k)\right)\right] \tag{10} \\ \text{subject to the BCNs (8) with } \boldsymbol{x}_0 = \boldsymbol{x}_{\text{init}}.$$

g_k and h are bounded, and the KL divergence $P(\cdot|\boldsymbol{x}_k)$ depends on the selection probabilities $\boldsymbol{c}_{k,i}\,(i = 1, \ldots, 2^{n_{\boldsymbol{u}}})$ of $\boldsymbol{u}_k$, the feasible sets of which are $\mathcal{C}_{k,i}$. Compared with the conventional KL control [29], the KL divergence is introduced with a weight coefficient $\mu > 0$, and the stage cost function $g_k(\boldsymbol{x}_k, \boldsymbol{u}_k)$ is also extended to include the cost of the control input $\boldsymbol{u}_k$, instead of the conventional form $g_k(\boldsymbol{x}_k)$. In the case of $\mu = 0$ in Problem (10), which means that Problem (10) ignores the KL divergence, the problem can be solved using the conventional DP (see [28]).

As observed in Example 2, the KL divergence quantitatively evaluates the similarity between P and $\overline{P}$, and it has a broader modeling capability compared with conventional forbidden-state-based techniques [12], [26]. If it is preferred that the system be fixed at a target point or in a given set of states similar to the stabilization problems (see [15], [18], [22]), the desired transition probabilities to the target points are set as large values. Throughout this article, the following simplified notations of the transition probabilities are used:

$$\begin{cases} p_{k,i,j} = P\left(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j} | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}\right) \\ \overline{p}_{k,i,j} = \overline{P}\left(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j} | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}\right). \end{cases}$$

Furthermore, the following notations are introduced in this article:

$$\begin{aligned} &\mathtt{inv}(j|i) = \left\{ l = 1, \dots, 2^{n_u} | \boldsymbol{\delta}_{2^{n_x}}^{j} = \boldsymbol{M}\boldsymbol{\delta}_{2^{n_u}}^{l}\boldsymbol{\delta}_{2^{n_x}}^{i} \right\} \\ &\mathtt{inv}^*(j|i,k) \in \arg\min_{l \in \mathtt{inv}(j|i)} g_k\left(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_u}}^{l}\right) \\ &\mathtt{inv}^\dagger(j|i,k) = \mathtt{inv}(j|i) \setminus \mathtt{inv}^*(j|i,k). \end{aligned} \tag{11}$$

In summary, $\mathtt{inv}(j|i)$ is the set of all the indices of the control inputs $\boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^{l}$ driving $\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}$ to $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j}$. An index in $\mathtt{inv}(j|i)$ with the minimum stage cost $g_k(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_u}}^{l})$ is selected as $\mathtt{inv}^*(j|i,k)$; the trivial nonoptimal inputs are in $\mathtt{inv}^\dagger(j|i,k)$ and set as $c_{k,i,l} = 0$ if $l \in \mathtt{inv}^\dagger(j|i,k)$ for j satisfying $\boldsymbol{\delta}_{2^{n_x}}^{j} = \boldsymbol{M}\boldsymbol{\delta}_{2^{n_u}}^{l}\boldsymbol{\delta}_{2^{n_x}}^{i}$. Thus, $\mathtt{inv}(j|i) \neq \emptyset$, which means $\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}$ can be driven to $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j}$, $\boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^{l}$ where $l = \mathtt{inv}^*(j|i,k)$ is used. The optimality of $\mathtt{inv}^*(j|i,k)$ [equivalently, the nonoptimality of $\mathtt{inv}^\dagger(j|i,k)$] can be easily confirmed and omitted herein owing to space limitations. From the definition of $\mathtt{inv}(j|i)$, the union of $\mathtt{inv}(j|i)$ with respect to $j = 1, \dots, 2^{n_x}$ is $\{1, \dots, 2^{n_u}\}$, indicating that

$$\begin{aligned} \bigcup_{j=1}^{2^{n_x}} \mathtt{inv}(j|i) &= \bigcup_{j=1}^{2^{n_x}} \left(\mathtt{inv}^*(j|i,k) \cup \mathtt{inv}^\dagger(j|i,k)\right) \\ &= \left\{1, \dots, 2^{n_u}\right\} \end{aligned} \tag{12}$$

and $\mathtt{inv}(j|i) \cap \mathtt{inv}(j'|i) = \emptyset$ if $j \neq j'$.

Example 1: The following BCNs with a 2-D state and a single input are considered:

$$\boldsymbol{x}_{k+1,1} = \boldsymbol{x}_{k,1} \wedge \boldsymbol{u}_k, \ \boldsymbol{x}_{k+1,2} = \boldsymbol{x}_{k,2} \wedge \neg\boldsymbol{u}_k. \tag{13}$$

The aforementioned BCNs have a structure matrix $\boldsymbol{M} = \delta_4[2, 2, 4, 4, 3, 4, 3, 4]$. On using the state variable $\boldsymbol{x}_k = \boldsymbol{x}_{k,1}\boldsymbol{x}_{k,2} \in \Delta^4$ and the control input $\boldsymbol{u}_k \in \Delta^2$, the state equation is expressed as $\boldsymbol{x}_{k+1} = \boldsymbol{M}\boldsymbol{u}_k\boldsymbol{x}_k$. In the example of $\boldsymbol{x}_k = \boldsymbol{\delta}_4^1$, equations $\boldsymbol{M}\boldsymbol{\delta}_2^1\boldsymbol{\delta}_4^1 = \boldsymbol{\delta}_4^2$ and $\boldsymbol{M}\boldsymbol{\delta}_2^2\boldsymbol{\delta}_4^1 = \boldsymbol{\delta}_4^3$ result in $\mathtt{inv}(2|1) = \mathtt{inv}^*(2|1,k) = 1$ and $\mathtt{inv}(3|1) = \mathtt{inv}^*(3|1,k) = 2$, respectively, without relation to the stage cost g. In the case of $\boldsymbol{x}_k = \boldsymbol{\delta}_4^4$, because $\boldsymbol{M}\boldsymbol{\delta}_2^1\boldsymbol{\delta}_4^4 = \boldsymbol{M}\boldsymbol{\delta}_2^2\boldsymbol{\delta}_4^4 = \boldsymbol{\delta}_4^4$, there are multiple inputs resulting in $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_4^4$. For such cases, $\mathtt{inv}(4|4) = \{1, 2\}$, and $\mathtt{inv}^*(4|4,k)$ is a control input minimizing the stage cost. An inequality $g_k(\boldsymbol{\delta}_4^4, \boldsymbol{\delta}_2^1) < g_k(\boldsymbol{\delta}_4^4, \boldsymbol{\delta}_2^2)$ results in $\mathtt{inv}^*(4|4,k) = 1$ and $\mathtt{inv}^\dagger(4|4,k) = 2$, as illustrated in Example 3.

In Problem (10), the following assumption is made.

Assumption 1: In Problem (10):

1) The desired transition probability $\overline{\boldsymbol{p}}_{k,i} = [\overline{p}_{k,i,1}, \dots, \overline{p}_{k,i,2^{n_x}}]^\top \in \mathcal{S}^{2^{n_x}}$ ($k = 0, \dots, N-1$, $i = 1, \dots, 2^{n_x}$) is given as follows:

$$\begin{aligned} \overline{p}_{k,i,j} &= \overline{P}\left(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j} | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}\right) \\ &= \begin{cases} (\text{constant}) \in [0,1], & (\mathtt{inv}(j|i) \neq \emptyset) \\ 0, & (\mathtt{inv}(j|i) = \emptyset) \end{cases} \end{aligned}$$

2) The selection probability $\boldsymbol{c}_{k,i} \in \mathbb{S}^{2^{n_u}}$ ($k = 0, \dots, N-1$, $i = 1, \dots, 2^{n_x}$) satisfies

$$c_{k,i,l} \begin{cases} = 0 & \left(l \in \mathtt{inv}^\dagger(j|i,k)\right), \\ = \overline{p}_{k,i,j} & \left(l = \mathtt{inv}^*(j|i,k) \text{ with } \overline{p}_{k,i,j} \in \{0,1\}\right), \\ \in [0,1] & \left(l = \mathtt{inv}^*(j|i,k) \text{ with } \overline{p}_{k,i,j} \in (0,1)\right) \end{cases}$$

for $l = 1, \dots, 2^{n_u}$, where j is an index satisfying $\boldsymbol{\delta}_{2^{n_x}}^{j} = \boldsymbol{M}\boldsymbol{\delta}_{2^{n_u}}^{l}\boldsymbol{\delta}_{2^{n_x}}^{i}$.

Equation (11) indicates that $\mathtt{inv}^*(j|i,k)$ and $\mathtt{inv}^\dagger(j|i,k)$ are the disjoint separation of $\mathtt{inv}(j|i)$, which means that $\mathtt{inv}^*(j|i,k) \cup \mathtt{inv}^\dagger(j|i,k) = \mathtt{inv}(j|i)$ and $\mathtt{inv}^*(j|i,k) \cap \mathtt{inv}^\dagger(j|i,k) = \emptyset$; therefore, the aforementioned three cases of $c_{k,i,l}$ are independent.

Remark 2: The first line of the definition of $c_{k,i,l}$, which means $c_{k,i,l} = 0\,(l \in \mathtt{inv}^\dagger(j|i,k))$, is considered. The condition implies the exclusion of a trivial nonoptimal control input, which results in $g_k(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_u}}^{l}) > \arg\min_{l'=1,\dots,2^{n_u}} g_k(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_u}}^{l'})$ with the same next state $\boldsymbol{M}\boldsymbol{\delta}_{2^{n_u}}^{l}\boldsymbol{\delta}_{2^{n_x}}^{i} = \boldsymbol{M}\boldsymbol{\delta}_{2^{n_u}}^{l'}\boldsymbol{\delta}_{2^{n_x}}^{i}$, the meaning of which is clear.

Remark 3: The second and third lines of the definition of $c_{k,i,l}$, which indicate that the case of $l = \mathtt{inv}^*(j|i,k)$, are considered. These two cases are classified using the value of $\overline{p}_{k,i,j}$ as follows.

1) $c_{k,i,l} = \overline{p}_{k,i,j}\,(l = \mathtt{inv}^*(j|i,k)$ with $\overline{p}_{k,i,j} \in \{0,1\})$:

   a) If $\overline{p}_{k,i,j} = 0$, then $c_{k,i,l} = \overline{p}_{k,i,j} = 0$; to avoid the zero-division issue, a transition probability corresponding to $\overline{p}_{k,i,j} = 0$, which means that $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j}$ is a forbidden state, is excluded in the definition of the KL divergence (9). Instead, the feasible set $\mathcal{C}_{k,i}$ takes into consideration the forbidden state.

   b) If $\overline{p}_{k,i,j} = 1$, then $c_{k,i,l} = \overline{p}_{k,i,j} = 1$, which is a redundant condition. Here, an index j_{fixed} is set subject to satisfying $\overline{p}_{k,i,j} = 0$ for an arbitrary $j \neq j_{\text{fixed}}$. $\sum_{j=1,\dots,2^{n_x}} \overline{p}_{k,i,j} = 1$ results in $\overline{p}_{k,i,j_{\text{fixed}}} = 1$. In addition, the index l satisfying $l = \mathtt{inv}^*(j_{\text{fixed}}|i,k)$ results in $c_{k,i,l} = 1$ on using the aforementioned 1-a) case because $c_{k,i,l} = \overline{p}_{k,i,j} = 0$ for arbitrary $j \neq j_{\text{fixed}}$ and $\sum_{l=1,\dots,2^{n_u}} c_{k,i,l} = 1$.

2) $c_{k,i,l} \in [0,1]\,(l = \mathtt{inv}^*(j|i,k)$ with $\overline{p}_{k,i,j} \in (0,1))$. As discussed subsequently, if neither $(l \in \mathtt{inv}^\dagger(j|i,k))$ nor $(l = \mathtt{inv}^*(j|i,k)$ with $\overline{p}_{k,i,j} \in \{0,1\})$ is satisfied, an optimal solution $c^*_{k,i,l}$ always lies in $c^*_{k,i,l} \in (0,1)$.

Example 2: The BCNs having a 2-D state and a single input of Example 1, the structure matrix of which is given by $\boldsymbol{M} = \delta_4[2, 2, 4, 4, 3, 4, 3, 4]$, are considered. Equations $\boldsymbol{M}\boldsymbol{\delta}_2^1\boldsymbol{\delta}_4^1 = \boldsymbol{\delta}_4^2$ and $\boldsymbol{M}\boldsymbol{\delta}_2^2\boldsymbol{\delta}_4^1 = \boldsymbol{\delta}_4^3$ (or equivalently $\mathtt{inv}(2|1) = \mathtt{inv}^*(2|1,k) = 1$ and $\mathtt{inv}(3|1) = \mathtt{inv}^*(3|1,k) = 2$, respectively) result in arbitrary design parameters $\overline{p}_{k,1,2} \in [0,1]$ and $\overline{p}_{k,1,3} \in [0,1]$, subject to $\overline{p}_{k,1,2} + \overline{p}_{k,1,3} = 1$, while $\overline{p}_{k,1,1} = \overline{p}_{k,1,4} = 0$ is in accordance with Assumption 1. That is, the state $\boldsymbol{x}_k$ cannot arrive at either $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_4^1$ or $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_4^4$.

Here, Problem (10) with a simple setting of $N = 1, g_0 = 0, h = 0, \mu = 1$, and $\boldsymbol{x}_0 = \boldsymbol{\delta}_4^1$ is considered, that is, the minimization problem of $\mathrm{KL}\left(P(\cdot|\boldsymbol{x}_0)\middle|\overline{P}(\cdot|\boldsymbol{x}_0)\right)$ is considered.

Case 1): The desired transition probabilities are given by

$$\overline{p}_{0,1,2} = \overline{p}_{0,1,3} = 0.5, \ \overline{p}_{0,1,1} = \overline{p}_{0,1,4} = 0.$$

The case of $\boldsymbol{x}_k = \boldsymbol{\delta}_4^1$ results in

$$\begin{cases} c_{k,1,1} = P\left(\boldsymbol{u}_k = \boldsymbol{\delta}_2^1 \middle| \boldsymbol{x}_k = \boldsymbol{\delta}_4^1\right) \\ \qquad = P\left(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_4^2 \middle| \boldsymbol{x}_k = \boldsymbol{\delta}_4^1\right) = p_{k,1,2} \\ c_{k,1,2} = P\left(\boldsymbol{u}_k = \boldsymbol{\delta}_2^2 \middle| \boldsymbol{x}_k = \boldsymbol{\delta}_4^1\right) \\ \qquad = P\left(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_4^3 \middle| \boldsymbol{x}_k = \boldsymbol{\delta}_4^1\right) = p_{k,1,3}. \end{cases} \tag{14}$$

Because the initial value $\boldsymbol{x}_0 = \boldsymbol{\delta}_4^1$ is deterministic, the objective function no longer requires the expectation operation and is given by

$$\begin{aligned} \mathrm{KL}\left(P(\cdot|\boldsymbol{x}_0)\middle|\overline{P}(\cdot|\boldsymbol{x}_0)\right) &= p_{0,1,2}\log\frac{p_{0,1,2}}{\overline{p}_{0,1,2}} + p_{0,1,3}\log\frac{p_{0,1,3}}{\overline{p}_{0,1,3}} \\ &= c_{0,1,1}\log\frac{c_{0,1,1}}{0.5} + c_{0,1,2}\log\frac{c_{0,1,2}}{0.5}. \end{aligned}$$

The optimal solution is $c_{0,1,1}^* = c_{0,1,2}^* = 0.5$, and the optimal objective function value is zero (the detailed derivation is omitted herein because the problem is a special case of Problem (26), introduced subsequently). In the conventional optimal problems, the optimal control is obtained as according to a deterministic law, which means that $(c_{0,1,1}, c_{0,1,2}) = (0, 1)$ or $(1, 0)$. On recalling that $0\log 0$ is formally defined by $\lim_{x\searrow 0} x\log x = 0$ in this article, both the aforementioned deterministic laws result in $\mathrm{KL}\left(P(\cdot|\boldsymbol{x}_0)\middle|\overline{P}(\cdot|\boldsymbol{x}_0)\right) = \log 2$. This example suggests that the randomized control input is required to be taken into consideration to minimize the objective function while including the KL divergence.

Case 2): The following desired transition probabilities are given:

$$\overline{p}_{0,1,2} = 1,\ \overline{p}_{0,1,3} = 0,\ \overline{p}_{0,1,1} = \overline{p}_{0,1,4} = 0.$$

This means that the transition from $\boldsymbol{x}_0 = \boldsymbol{\delta}_4^1$ to $\boldsymbol{x}_1 = \boldsymbol{\delta}_4^3$ is prohibited. The definition of $\mathcal{C}_{0,1}$ in Assumption 1 results in

$$c_{0,1,\mathtt{inv}^*(2|1,0)} = c_{0,1,1} = 1,\ c_{0,1,\mathtt{inv}^*(3|1,0)} = c_{0,1,2} = 0.$$

Eventually, the feasible set $\mathcal{C}_{0,1}$ degenerates to a point $(c_{0,1,1}, c_{0,1,2}) = (1, 0)$, which is a trivial optimal solution. The transition probability, which is as desired, results in

$$\mathrm{KL}\left(P(\cdot|\boldsymbol{x}_0)\middle|\overline{P}(\cdot|\boldsymbol{x}_0)\right) = p_{0,1,2}\log\frac{p_{0,1,2}}{\overline{p}_{0,1,2}} = \log 1 = 0.$$

IV. Main Results

A. Reformulation as Optimal Trajectory Planning Problem

This section reformulates the optimal control problem (Problem (10) with Assumption 1) as an optimal trajectory planning problem because the latter has a more tractable structure. If $\mathtt{inv}(j|i) \neq \emptyset$ and $l = \mathtt{inv}^*(j|i,k)$, which means that there exists a $\boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^l$ driving $\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i$ to $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^j$, the transition probability from $\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i$ to $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^j$ is $c_{k,i,l}$

$$\begin{aligned} p_{k,i,j} &= P\left(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^j | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i\right) \\ &= \begin{cases} c_{k,i,l} = P\left(\boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^l | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i\right), & (\mathtt{inv}(j|i) \neq \emptyset) \\ 0, & (\mathtt{inv}(j|i) = \emptyset). \end{cases} \end{aligned} \tag{15}$$

An optimal selection probability $c_{k,i,l}^*$ of $c_{k,i,l}$ can be obtained from an optimal transition probability $p_{k,i,j}^*$ of $p_{k,i,j}$. Therefore, the optimal control Problem (10) is the design problem of $p_{k,i,j}$. Here, the consideration of Assumption 1 results in

$$\begin{aligned} \overline{p}_{k,i,j} &= \overline{P}\left(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^j | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i\right) \\ &= \begin{cases} (\text{constant}) \in [0, 1], & (\mathtt{inv}(j|i) \neq \emptyset) \\ 0, & (\mathtt{inv}(j|i) = \emptyset). \end{cases} \end{aligned} \tag{16}$$

There are two settings resulting in $\overline{p}_{k,i,j} = 0$: 1) $\mathtt{inv}(j|i) = \emptyset$, that is, the state $\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i$ cannot move to $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^j$ and 2) $\mathtt{inv}(j|i) \neq \emptyset$ and $\mathtt{inv}^*(j|i,k) = l$. However, a designer deliberately sets $\overline{p}_{k,i,j} = 0$, that is, $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^j$ is a forbidden state. $\overline{p}_{k,i,j} = 0$ and Assumption 1 result in $p_{k,i,j} = 0$; therefore, the transition probability $p_{k,i,j}$ the desired value of which is set such that $\overline{p}_{k,i,j} \neq 0$ is required to be designed, and the number of such variables is

$$\left|\{j = 1, \ldots, 2^{n_x} | \overline{p}_{k,i,j} \neq 0\}\right| \leq 2^{n_u} \tag{17}$$

because the number of possible next states $\boldsymbol{x}_{k+1}$, which means that $\mathtt{inv}^*(j|i,k) = l$ in (16), is less than that of $\boldsymbol{u}_k \in \Delta^{2^{n_u}}$. The feasible set of the variable $\boldsymbol{p}_{k,i}$ is denoted as follows:

$$\mathcal{P}_{k,i} = \left[\boldsymbol{p}_{k,i} \in \mathbb{S}^{2^{n_x}},\ p_{k,i,j}\begin{cases} = \overline{p}_{k,i,j} & (\overline{p}_{k,i,j} \in \{0, 1\}) \\ \in [0, 1] & (\overline{p}_{k,i,j} \in (0, 1)). \end{cases}\right]. \tag{18}$$

The stage cost function $g_k(\boldsymbol{x}_k, \boldsymbol{u}_k)$ is reformulated in the form of $w_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1})$ as follows:

$$w_k\left(\boldsymbol{\delta}_{2^{n_x}}^i, \boldsymbol{\delta}_{2^{n_x}}^j\right) = \begin{cases} g_k\left(\boldsymbol{\delta}_{2^{n_x}}^i, \boldsymbol{\delta}_{2^{n_u}}^{\mathtt{inv}^*(j|i,k)}\right), & (\overline{p}_{k,i,j} > 0) \\ +\infty, & (\overline{p}_{k,i,j} = 0). \end{cases} \tag{19}$$

The equivalence of the expectation of $g_k(\boldsymbol{x}_k, \boldsymbol{u}_k)$ and $w_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1})$ is presented here. The expectation of g_k is calculated as follows:

$$\mathbb{E}\left[g_k(\boldsymbol{x}_k, \boldsymbol{u}_k)\right] = \sum_{i=1}^{2^{n_x}}\sum_{l=1}^{2^{n_u}} P\begin{pmatrix}\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i, \\ \boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^l\end{pmatrix} g_k\left(\boldsymbol{\delta}_{2^{n_x}}^i, \boldsymbol{\delta}_{2^{n_u}}^l\right).$$

$P(\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i, \boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^l)$ is expanded as

$$P\begin{pmatrix}\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i, \\ \boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^l\end{pmatrix} = P\left(\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i\right) P\left(\boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^l | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i\right)$$

and $P(\boldsymbol{u}_k = \boldsymbol{\delta}_{2^{n_u}}^l | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i) = c_{k,i,l}$. Therefore

$$\mathbb{E}\left[g_k(\boldsymbol{x}_k, \boldsymbol{u}_k)\right] = \sum_{i=1}^{2^{n_x}}\sum_{l=1}^{2^{n_u}} P\left(\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i\right) c_{k,i,l}\, g_k\left(\boldsymbol{\delta}_{2^{n_x}}^i, \boldsymbol{\delta}_{2^{n_u}}^l\right).$$

On recalling (12), which indicates that $\{1, \ldots, 2^{n_u}\} = \bigcup_{j=1}^{2^{n_x}} \mathtt{inv}(j|i)$ and $\mathtt{inv}(j|i) \cap \mathtt{inv}(j'|i) = \emptyset$ if $j \neq j'$, the sum with respect to $l = 1, \ldots, 2^{n_u}$, as mentioned above, can be expressed as that with respect to $j = 1, \ldots, 2^{n_x}$, as follows:

$$\begin{aligned} &\mathbb{E}\left[g_k(\boldsymbol{x}_k, \boldsymbol{u}_k)\right] \\ &= \sum_{i=1}^{2^{n_x}}\sum_{j=1}^{2^{n_x}}\sum_{l' \in \mathtt{inv}(j|i)} P\left(\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^i\right) c_{k,i,l'}\, g_k\left(\boldsymbol{\delta}_{2^{n_x}}^i, \boldsymbol{\delta}_{2^{n_u}}^{l'}\right). \end{aligned}$$

In the sum with respect to $l' \in \mathtt{inv}(j|i) = \mathtt{inv}^*(j|i,k) \cup \mathtt{inv}^\dagger(j|i,k)$ above, the definition of the feasible set $\mathcal{C}_{k,i}$ in Assumption 1 indicates that $c_{k,i,l'} = 0$ if $l' \neq \mathtt{inv}^*(j|i,k)$, or

$l' \in \mathtt{inv}^{\dagger}(j|i,k)$ equivalently. Therefore, the sum is simplified as follows:

$$\begin{aligned}
&\mathbb{E}\big[g_k(\boldsymbol{x}_k, \boldsymbol{u}_k)\big] \\
&= \sum_{i=1}^{2^{n_x}} \sum_{j=1}^{2^{n_x}} P\big(\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}\big) c_{k,i,\mathtt{inv}^*(j|i,k)} g_k\Big(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_u}}^{\mathtt{inv}^*(j|i,k)}\Big) \\
&= \sum_{i=1}^{2^{n_x}} \sum_{j=1}^{2^{n_x}} P\big(\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}\big) p_{k,i,j} w_k\Big(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_x}}^{j}\Big) \\
&= \sum_{i=1}^{2^{n_x}} \sum_{j=1}^{2^{n_x}} P\Big(\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j}\Big) w_k\Big(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_x}}^{j}\Big) \\
&= \mathbb{E}\big[w_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1})\big]
\end{aligned}$$

where the second equality uses the relationship between $p_{k,i,j}$ and $c_{k,i,l}$ in (15) and the definition of w_k in (19).

Consequently, the optimal control problem (Problem (10) with Assumption 1) is reformulated in the following optimal trajectory planning problem:

$$\begin{aligned}
\min_{\substack{\boldsymbol{p}_{k,i} \in \mathcal{P}_{2^{n_x}}, \\ k=0,\ldots,N-1, \\ i=1,\ldots,2^{n_x}}} \; & \mathbb{E}\Bigg[\sum_{k=0}^{N-1} w(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) + h(\boldsymbol{x}_N) \\
& \qquad + \mu \sum_{k=0}^{N-1} \mathrm{KL}\big(P(\cdot|\boldsymbol{x}_k) \big| \overline{P}(\cdot|\boldsymbol{x}_k)\big)\Bigg] \qquad (20)\\
\text{subject to } & P\Big(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j} | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}\Big) = p_{k,i,j} \\
& \quad (k = 0, \ldots, N-1, \;\; i,j = 1, \ldots, 2^{n_x}), \\
& \boldsymbol{x}_0 = \boldsymbol{x}_{\mathrm{init}}.
\end{aligned}$$

The optimal trajectory planning problem of the state $\boldsymbol{x}_k$ is more tractable because the state variable is directly controlled (see the maze game in [30, Sec. 8.1]).

Example 3: The BCNs (13) in Example 1 is considered again. The structure matrix is $\boldsymbol{M} = \boldsymbol{\delta}_4[2,2,4,4,3,4,3,4]$. $\boldsymbol{x}_k = \boldsymbol{\delta}_4^1$ results in (14) and indicates that the selection probabilities of the control are equivalent to the transition probabilities. Because $\mathtt{inv}(1|1) = \mathtt{inv}(4|1) = \emptyset$, which means that there is no control input $\boldsymbol{u}_k$ driving $\boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^1$ to either $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^1$ or $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^4$, the following transition probabilities are set as zero, based on the definition of $\mathcal{P}_{k,i}$ in (18)

$$\overline{p}_{k,1,1} = p_{k,1,1} = \overline{p}_{k,1,4} = p_{k,1,4} = 0.$$

Similarly, equations

$$\begin{cases} \mathtt{inv}(1|2) = \mathtt{inv}(3|2) = \emptyset \\ \mathtt{inv}(1|3) = \mathtt{inv}(2|3) = \emptyset \\ \mathtt{inv}(1|4) = \mathtt{inv}(2|4) = \mathtt{inv}(3|4) = \emptyset \end{cases}$$

result in

$$\begin{cases} \overline{p}_{k,2,1} = p_{k,2,1} = \overline{p}_{k,2,3} = p_{k,2,3} = 0 \\ \overline{p}_{k,3,1} = p_{k,3,1} = \overline{p}_{k,3,2} = p_{k,3,2} = 0 \\ \overline{p}_{k,4,1} = p_{k,4,1} = \overline{p}_{k,4,2} = p_{k,4,2} = \overline{p}_{k,4,3} = p_{k,4,3} = 0 \end{cases}$$

respectively. In the case of $\boldsymbol{x}_k = \boldsymbol{\delta}_4^4$, because $\mathtt{inv}(4|4) = \{1,2\}$, there are multiple inputs resulting in $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_4^4$. For such cases, a control input minimizing the stage cost in Problem (20) is selected as $\mathtt{inv}^*(4|4,k)$. A time-invariant stage cost $[\boldsymbol{G}_k]_{i,l} = g_k(\boldsymbol{\delta}_4^i, \boldsymbol{\delta}_2^l)$ given by

$$\boldsymbol{G}_k = \begin{bmatrix} 0.1 & 0.2 & 0.4 & 0.4 \\ 0.1 & 0.3 & 0.5 & 0.7 \end{bmatrix}^{\top}$$

for each $k = 0, \ldots, N-1$ is considered. An inequality $g_k(\boldsymbol{\delta}_4^4, \boldsymbol{\delta}_2^1) = 0.4 < g_k(\boldsymbol{\delta}_4^4, \boldsymbol{\delta}_2^2) = 0.7$ results in $\mathtt{inv}^*(4|4,k) = 1$ and $\mathtt{inv}^{\dagger}(4|4,k) = 2$, and the resulting problem (20) is addressed. Here, the desired time-invariant transition probability $\overline{p}_{k,i,j} = [\overline{\boldsymbol{P}}_k]_{i,j}$ for $k = 0, \ldots, N-1$ is set as follows:

$$\overline{\boldsymbol{P}}_k = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

It should be noted that the transition probability matrices used in this article are similar to those used in the conventional MDP. This means that they are the transposed matrices of the STP-based transition matrices for the probabilistic BCNs [26]. The control variable $p_{k,i,j} = [\boldsymbol{P}_k]_{i,j}$ is given as follows:

$$\begin{aligned}
\boldsymbol{P}_k &= \begin{bmatrix} 0 & p_{k,1,2} & p_{k,1,3} & 0 \\ 0 & p_{k,2,2} & 0 & p_{k,2,4} \\ 0 & 0 & p_{k,3,3} & p_{k,3,4} \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 0 & c_{k,1,1} & c_{k,1,2} & 0 \\ 0 & c_{k,2,1} & 0 & c_{k,2,2} \\ 0 & 0 & c_{k,3,2} & c_{k,3,1} \\ 0 & 0 & 0 & 1 \end{bmatrix}.
\end{aligned}$$

The second equality shown above claims that the original control variables, which are the selection probabilities of the control input, are recovered using the control variable $\boldsymbol{P}_k$.

In addition, the reformulation of the stage cost function $g_k(\boldsymbol{x}_k, \boldsymbol{u}_k)$ as $w_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1})$ is examined here. The stage cost $[\boldsymbol{W}_k]_{i,j} = w_k(\boldsymbol{\delta}_4^i, \boldsymbol{\delta}_4^j)$ for each $k = 0, \ldots, N-1$, is given by

$$\begin{aligned}
\boldsymbol{W}_k &= \begin{bmatrix} +\infty & g_k\big(\boldsymbol{\delta}_4^1, \boldsymbol{\delta}_2^1\big) & g_k\big(\boldsymbol{\delta}_4^1, \boldsymbol{\delta}_2^2\big) & +\infty \\ +\infty & g_k\big(\boldsymbol{\delta}_4^2, \boldsymbol{\delta}_2^1\big) & +\infty & g_k\big(\boldsymbol{\delta}_4^2, \boldsymbol{\delta}_2^2\big) \\ +\infty & +\infty & g_k\big(\boldsymbol{\delta}_4^3, \boldsymbol{\delta}_2^2\big) & g_k\big(\boldsymbol{\delta}_4^3, \boldsymbol{\delta}_2^1\big) \\ +\infty & +\infty & +\infty & g_k\big(\boldsymbol{\delta}_4^4, \boldsymbol{\delta}_2^1\big) \end{bmatrix} \\
&= \begin{bmatrix} +\infty & 0.1 & 0.1 & +\infty \\ +\infty & 0.2 & +\infty & 0.3 \\ +\infty & +\infty & 0.5 & 0.4 \\ +\infty & +\infty & +\infty & 0.4 \end{bmatrix}.
\end{aligned}$$

In the equation above, the path from $\boldsymbol{x}_k = \boldsymbol{\delta}_4^4$ to $\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_4^4$ uses the control input $\boldsymbol{u}_k = \boldsymbol{\delta}_2^1 = \boldsymbol{\delta}_2^{\mathtt{inv}^*(4|4,k)}$, and the corresponding cost is $[\boldsymbol{W}_k]_{4,4} = w_k(\boldsymbol{\delta}_4^4, \boldsymbol{\delta}_4^4) = g_k(\boldsymbol{\delta}_4^4, \boldsymbol{\delta}_2^1) = 0.4$ instead of $g_k(\boldsymbol{\delta}_4^4, \boldsymbol{\delta}_2^2) = 0.7$.

For $s = 0, \ldots, N-1$, the following optimal trajectory planning subproblem derived from the optimal control problem (10) with Assumption 1 is considered

$$\begin{aligned}
&v_s^{\mathrm{KL},\mu}(\boldsymbol{x}_{\mathrm{init}}) = \\
&\min_{\substack{\boldsymbol{p}_{k,i} \in \mathcal{P}_{k,i}, \\ k=s,\ldots,N-1, \\ i=1,\ldots,2^{n_x}}} \; \mathbb{E}\Bigg[\sum_{k=s}^{N-1} w_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) + h(\boldsymbol{x}_N) \\
&\qquad\qquad + \mu \sum_{k=s}^{N-1} \mathrm{KL}\big(P(\cdot|\boldsymbol{x}_k) \big| \overline{P}(\cdot|\boldsymbol{x}_k)\big) \Bigg| \boldsymbol{x}_s = \boldsymbol{x}_{\mathrm{init}}\Bigg]
\end{aligned}$$

$$\begin{aligned} \text{subject to}\quad & P\left(\boldsymbol{x}_{k+1} = \boldsymbol{\delta}_{2^{n_x}}^{j} | \boldsymbol{x}_k = \boldsymbol{\delta}_{2^{n_x}}^{i}\right) = p_{k,i,j}\\ & \left(k = s, \dots, N-1,\ \ i,j = 1, \dots, 2^{n_x}\right)\\ & \boldsymbol{x}_s = \boldsymbol{x}_{\text{init}}. \end{aligned} \tag{21}$$

At $s = N$, $v_N^{\mathrm{KL},\mu}(\boldsymbol{\delta}_{2^{n_x}}^{i}) = h(\boldsymbol{\delta}_{2^{n_x}}^{i})$, $i = 1, \dots, 2^{n_x}$. The following theorem provides the Bellman equation of the optimal trajectory planning problem (20).

Theorem 2: The optimal trajectory planning problem (20) derived from the optimal control problem (10) with Assumption 1 is considered. In Problem (20), the Bellman equation on the value function $v_s^{\mathrm{KL},\mu}$ $(s = 0, \dots, N-1)$ in (21) is given as

$$\begin{aligned} v_s^{\mathrm{KL},\mu}\left(\boldsymbol{\delta}_{2^{n_x}}^{i}\right) = \min_{\boldsymbol{p}_{s,i} \in \mathcal{P}_{s,i}} \Bigg(& \sum_{j=1}^{2^{n_x}} p_{s,i,j}\left[w_s\left(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_x}}^{j}\right) + v_{s+1}^{\mathrm{KL},\mu}\left(\boldsymbol{\delta}_{2^{n_x}}^{j}\right)\right]\\ & + \mu \sum_{\substack{j=1,\dots,2^{n_x}\\ \bar{p}_{s,i,j} \neq 0}} p_{s,i,j} \log \frac{p_{s,i,j}}{\bar{p}_{s,i,j}} \Bigg). \end{aligned} \tag{22}$$

Proof: In $v_s^{\mathrm{KL},\mu}(\boldsymbol{\delta}_{2^{n_x}}^{i})$, as defined in (21), the sum of the stage cost from $s+1$ and the terminal cost is rearranged as follows:

$$\begin{aligned} & \mathbb{E}\Bigg[\sum_{k=s+1}^{N-1} w_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) + h(\boldsymbol{x}_N)\\ & \qquad + \mu \sum_{k=s+1}^{N-1} \mathrm{KL}\left(P(\cdot|\boldsymbol{x}_k) | \bar{P}(\cdot|\boldsymbol{x}_k)\right) \Bigg| \boldsymbol{x}_s = \boldsymbol{\delta}_{2^{n_x}}^{i}\Bigg]\\ & = \sum_{j=1}^{2^{n_x}} p_{s,i,j} \mathbb{E}\Bigg[\sum_{k=s+1}^{N-1} w_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) + h(\boldsymbol{x}_N)\\ & \qquad + \mu \sum_{k=s+1}^{N-1} \mathrm{KL}\left(P(\cdot|\boldsymbol{x}_k) | \bar{P}(\cdot|\boldsymbol{x}_k)\right) \Bigg| \boldsymbol{x}_{s+1} = \boldsymbol{\delta}_{2^{n_x}}^{j}\Bigg]. \end{aligned} \tag{23}$$

On the right-hand side of the equation above, the minimum value and minimizer of the sum of the two expectations are $v_{s+1}^{\mathrm{KL},\mu}(\boldsymbol{\delta}_{2^{n_x}}^{j})$ and the corresponding optimal $\boldsymbol{p}_{k,i}$ $(k = s+1, \dots, N-1,\ i = 1, \dots, 2^{n_x})$, respectively. The stage cost function at the sth step is expanded as follows:

$$\begin{aligned} \mathbb{E}\left[w_s(\boldsymbol{x}_s, \boldsymbol{x}_{s+1}) | \boldsymbol{x}_s = \boldsymbol{\delta}_{2^{n_x}}^{i}\right] &= \sum_{j=1}^{2^{n_x}} p_{s,i,j} w_s\left(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_x}}^{j}\right),\\ \mathbb{E}\left[\mathrm{KL}\left(P(\cdot|\boldsymbol{x}_s) | \bar{P}(\cdot|\boldsymbol{x}_s)\right) | \boldsymbol{x}_s = \boldsymbol{\delta}_{2^{n_x}}^{i}\right] &= \sum_{\substack{j=1,\dots,2^{n_x}\\ \bar{p}_{s,i,j} \neq 0}} p_{s,i,j} \log \frac{p_{s,i,j}}{\bar{p}_{s,i,j}}. \end{aligned} \tag{24}$$

Consequently, the combination of (23) and (24) results in the claim of the theorem. ■

Remark 4: In Theorem 2, especially at $s = 0$, the value of $v_0^{\mathrm{KL},\mu}(\boldsymbol{x}_{\text{init}})$ is the optimal objective function value of Problem (20) and the minimizer of $v_0^{\mathrm{KL},\mu}(\boldsymbol{x}_{\text{init}})$ is an optimal solution of Problem (20).

The following theorem provides the vectorized expression of the Bellman equation (22) using the vectorized value and objective functions

$$\begin{aligned} \boldsymbol{v}_k^{\mathrm{KL},\mu} &= \left[v_{k,1}^{\mathrm{KL},\mu}, \dots, v_{k,2^{n_x}}^{\mathrm{KL},\mu}\right]^{\top}\\ &= \left[v_k^{\mathrm{KL},\mu}\left(\boldsymbol{\delta}_{2^{n_x}}^{1}\right), \dots, v_k^{\mathrm{KL},\mu}\left(\boldsymbol{\delta}_{2^{n_x}}^{2^{n_x}}\right)\right]^{\top} \in \mathbb{R}^{2^{n_x}}\\ \boldsymbol{w}_{k,i} &= \left[w_{k,i,1}, \dots, w_{k,i,2^{n_x}}\right]^{\top}\\ &= \left[w_k\left(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_x}}^{1}\right), \dots, w_k\left(\boldsymbol{\delta}_{2^{n_x}}^{i}, \boldsymbol{\delta}_{2^{n_x}}^{2^{n_x}}\right)\right]^{\top} \in (\mathbb{R} \cup \{+\infty\})^{2^{n_x}}. \end{aligned} \tag{25}$$

Theorem 3: The optimal trajectory planning problem (20) derived from the optimal control problem (10) with Assumption 1 is considered. The value function $\boldsymbol{v}_k^{\mathrm{KL},\mu}$ $(k = 0, \dots, N-1,\ i = 1, \dots, 2^{n_x})$ satisfies the following iterative equation:

$$v_{k,i}^{\mathrm{KL},\mu} = -\mu \log\left(\bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)\right).$$

The optimal transition probability is given by

$$\boldsymbol{p}_{k,i}^{*,\mu} = \frac{\bar{\boldsymbol{p}}_{k,i} \odot \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)}{\bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)}.$$

Proof: The basic flow of the proof is similar to that presented in the original article of the KL control [29]. For the variable $p_{k,i,j}$, a variable transform with $d_{k,i,j} \in \mathbb{R}$ is introduced as $p_{k,i,j} = \bar{p}_{k,i,j} d_{k,i,j}$. The case of $\bar{p}_{k,i,j} = p_{k,i,j} \in \{0, 1\}$ formally defines $d_{k,i,j} = \bar{p}_{k,i,j}$. For the transformed control variable vector $\boldsymbol{d}_{k,i} = [d_{k,i,1}, \dots, d_{k,i,2^{n_x}}]^{\top}$, the feasible set is denoted as

$$\mathcal{D}_{k,i} = \left[\begin{matrix} \boldsymbol{d}_{k,i} \in \mathbb{R}^{2^{n_x}},\\ \bar{\boldsymbol{p}}_{k,i}^{\top} \boldsymbol{d}_{k,i} = 1, \end{matrix}\ d_{k,i,j} \begin{cases} = \bar{p}_{k,i,j} & (\bar{p}_{k,i,j} \in \{0, 1\})\\ \in \left[0, 1/\bar{p}_{k,i,j}\right] & (\bar{p}_{k,i,j} \in (0, 1)) \end{cases}\right].$$

Problem (20) results in the equivalent design problem of $\boldsymbol{d}_{k,i} \in \mathcal{D}_{k,i}$ $(k = 0, \dots, N-1,\ i = 1, \dots, 2^{n_x})$. If there exists j such that $\bar{p}_{k,i,j} = p_{k,i,j} = d_{k,i,j} \in \{0, 1\}$, this variable degenerates to a point, and it need not to be considered (especially if there exists $\bar{p}_{k,i,j} = p_{k,i,j} = d_{k,i,j} = 1$, and $\mathcal{D}_{k,i}$ is a single feasible point). The optimal solution is this feasible point and evidently satisfies the Bellman equation presented subsequently; this case is trivial and is excluded from the remainder of this proof. Thus, $\mathcal{D}_{k,i}$ is convex and has interior feasible points. On using Theorem 2 and vectorized $\boldsymbol{v}_k^{\mathrm{KL},\mu}$, $\boldsymbol{w}_{k,i}$ in (25), the Bellman equation of $\boldsymbol{\delta}_{2^{n_x}}^{i} \in \Delta^{2^{n_x}}$ at each $k = 0, \dots, N-1$, is given by

$$v_{k,i}^{\mathrm{KL},\mu} = \min_{\boldsymbol{d}_{k,i} \in \mathcal{D}_{k,i}} \sum_{j=1}^{2^{n_x}} \bar{p}_{k,i,j} d_{k,i,j}\left(w_{k,i,j} + \mu \log d_{k,i,j} + v_{k+1,j}^{\mathrm{KL},\mu}\right).$$

On vectorizing the equation above, the following subproblem is obtained:

$$v_{k,i}^{\mathrm{KL},\mu} = \min_{\boldsymbol{d}_{k,i} \in \mathcal{D}_{k,i}} \left(\bar{\boldsymbol{p}}_{k,i} \odot \boldsymbol{d}_{k,i}\right)^{\top}\left(\boldsymbol{w}_{k,i} + \mu \log \boldsymbol{d}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right). \tag{26}$$

The Hessian of the objective function $J_{k,i}(\boldsymbol{d}_{k,i}) = (\bar{\boldsymbol{p}}_{k,i} \odot \boldsymbol{d}_{k,i})^{\top}(\boldsymbol{w}_{k,i} + \mu \log \boldsymbol{d}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu})$ is given as

$$\frac{\partial^2 J_{k,i}(\boldsymbol{d}_{k,i})}{\partial d_{k,i,j'} \partial d_{k,i,j}} = \begin{cases} \mu \bar{p}_{k,i,j}/d_{k,i,j} > 0, & (j = j', d_{k,i,j} > 0)\\ 0, & (j \neq j') \end{cases}$$

which implies that $J_{k,i}(\boldsymbol{d}_{k,i})$ is convex. Here, the following points are introduced:

$$\begin{cases} \boldsymbol{d}_{k,i}^{*,\mu} = \dfrac{\exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)}{\bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)} \\ \boldsymbol{p}_{k,i}^{*,\mu} = \bar{\boldsymbol{p}}_{k,i} \odot \boldsymbol{d}_{k,i}^{*,\mu} = \dfrac{\bar{\boldsymbol{p}}_{k,i} \odot \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)}{\bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)} \\ \lambda_i^{*,\mu} = \mu \log\left(\bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)\right) - \mu. \end{cases} \tag{27}$$

The point $(\boldsymbol{d}_{k,i}^{*,\mu}, \lambda_i^{*,\mu})$ is the Karush–Kuhn–Tucker (KKT) point associated with a Lagrange function with Lagrange multipliers $\lambda_i \, (i = 1, \ldots, 2^{n_x})$ for the equality constraint, which is introduced as follows:

$$\mathcal{L}_i(\boldsymbol{d}_{k,i}, \lambda_i) = \left(\bar{\boldsymbol{p}}_{k,i} \odot \boldsymbol{d}_{k,i}\right)^{\top}\left[\boldsymbol{w}_{k,i} + \mu \log \boldsymbol{d}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right] + \lambda_i\left(\bar{\boldsymbol{p}}_{k,i}^{\top}\boldsymbol{d}_{k,i} - 1\right).$$

The optimality condition is

$$\begin{cases} \frac{\partial \mathcal{L}_i(\boldsymbol{d}_{k,i}, \lambda_i)}{\partial d_{k,i,j}} = \bar{p}_{k,i,j}\left(w_{k,i,j} + \mu \log d_{k,i,j} + v_{k+1,j}^{\mathrm{KL},\mu} + \lambda_i + \mu\right) = 0, \\ \bar{\boldsymbol{p}}_{k,i}^{\top}\boldsymbol{d}_{k,i} = 1. \end{cases}$$

Then, 1) the objective function and equality constraint set are both convex and 2) the feasible set has an interior feasible point, indicating that the Slater condition is satisfied. Therefore, a point satisfying the optimality condition is an optimal solution (see [34, Sec. 5.5.3]). Because $\boldsymbol{p}_{k,i}^{*,\mu}$ of the KKT point (27) indicates $0 < p_{k,i,j}^{*,\mu} = \bar{p}_{k,i,j} d_{k,i,j}^{*,\mu} < 1$ provided $\bar{p}_{k,i,j} \notin \{0, 1\}$, the inequality constraint $0 \leq d_{k,i,j} \leq 1/\bar{p}_{k,i,j}$ is inactive; therefore, the Lagrange multiplier for the inequality constraint is omitted here. It should be noted that, on using $w_{k,i,j} = +\infty$ in (19) and $\exp(-\infty) = 0$, a case of a feasible and trivial optimal point $p_{k,i,j}^{*,\mu} = d_{k,i,j}^{*,\mu} = \bar{p}_{k,i,j} \in \{0, 1\}$ is also covered with the expressions of $\boldsymbol{d}_{k,i}^{*,\mu}$ and $\boldsymbol{p}_{k,i}^{*,\mu}$ at the KKT point (27).

On using the optimality condition $w_{k,i,j} + \mu \log d_{k,i,j}^{*,\mu} + v_{k+1,j}^{\mathrm{KL},\mu} + \lambda_i^{*,\mu} + \mu = 0$ for $\bar{p}_{k,i,j} \in (0, 1)$ and $\bar{\boldsymbol{p}}_{k,i}^{\top}\boldsymbol{d}_{k,i}^{*,\mu} = 1$, the value function is rearranged as follows:

$$\begin{aligned} v_{k,i}^{\mathrm{KL},\mu} &= \min_{\boldsymbol{d}_{k,i} \in \mathcal{D}_{k,i}} \left(\bar{\boldsymbol{p}}_{k,i} \odot \boldsymbol{d}_{k,i}\right)^{\top}\left(\boldsymbol{w}_{k,i} + \mu \log \boldsymbol{d}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right) \\ &= -\lambda_i^{*,\mu} - \mu = -\mu \log\left(\bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)\right) \end{aligned}$$

and all the equations of the theorem have been obtained. ■

For efficient code implementation using the transformed variable [29], a variable $\boldsymbol{z}_k^{\mu}$ is introduced as follows:

$$\boldsymbol{z}_k^{\mu} = \exp\left(-\mu^{-1}\boldsymbol{v}_k^{\mathrm{KL},\mu}\right). \tag{28}$$

The following theorem provides an iterative equation in the matrix form.

Theorem 4: The same condition of Theorem 3 for the optimal trajectory planning subproblem (21) derived from the optimal control problem (10) with Assumption 1 is considered. The variable $\boldsymbol{z}_k^{\mu}$ in (28) satisfies the following matrix equation:

$$\begin{aligned} \boldsymbol{z}_k^{\mu} &= \boldsymbol{\Phi}_k^{\mu}\boldsymbol{z}_{k+1}^{\mu} \\ \boldsymbol{P}_k^{*,\mu} &= \left(\mathrm{Diag}\left(\mathbf{1}_{2^{n_x}} \oslash \boldsymbol{z}_k^{\mu}\right)\boldsymbol{\Phi}_k^{\mu}\right) \odot \left[\boldsymbol{z}_{k+1}^{\mu} \cdots \boldsymbol{z}_{k+1}^{\mu}\right]^{\top} \end{aligned}$$

where $\boldsymbol{\Phi}_k^{\mu} = \bar{\boldsymbol{P}}_k \odot \exp(-\mu^{-1}\boldsymbol{W}_k)$, $\bar{\boldsymbol{P}}_k = [\bar{\boldsymbol{p}}_{k,1} \cdots \bar{\boldsymbol{p}}_{k,2^{n_x}}]^{\top}$, $\boldsymbol{P}_k^{*,\mu} = [\boldsymbol{p}_{k,1}^{*,\mu} \cdots \boldsymbol{p}_{k,2^{n_x}}^{*,\mu}]^{\top}$, and $\boldsymbol{W}_k = [\boldsymbol{w}_{k,1} \cdots \boldsymbol{w}_{k,2^{n_x}}]^{\top}$.

Algorithm 1 DP for Problem (20) With $\mu = 0$

```
1: $\boldsymbol{v}_N^{\mathrm{DP}} = \boldsymbol{h}$                                      ▷ Initialization
2: for $k = N-1, \ldots, 0$ do                          ▷ Backward calculation on $k$
3:     for $i = 1, \ldots, 2^{n_x}$ do
4:         $v_{k,i}^{\mathrm{DP}} = \min(\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{DP}})$
5:         $\boldsymbol{p}_{k,i}^{\mathrm{DP}} \in \arg\min(\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{DP}})$
```

Proof: The Bellman equation given in Theorem 3 is reformulated as

$$-\mu^{-1}v_{k,i}^{\mathrm{KL},\mu} = \log\left(\bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)\right).$$

Taking the exponential of both the sides results in the following equation:

$$\begin{aligned} z_{k,i}^{\mu} &= \exp\left(-\mu^{-1}v_{k,i}^{\mathrm{KL},\mu}\right) = \bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right) \\ &= \left[\bar{\boldsymbol{p}}_{k,i} \odot \exp\left(-\mu^{-1}\boldsymbol{w}_{k,i}\right)\right]^{\top} \exp\left(-\mu^{-1}\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right) \\ &= \left[\bar{\boldsymbol{p}}_{k,i} \odot \exp\left(-\mu^{-1}\boldsymbol{w}_{k,i}\right)\right]^{\top} \boldsymbol{z}_{k+1}^{\mu}. \end{aligned}$$

On vertically stacking the equation above, a vectorized equation of $\boldsymbol{z}_k^{\mu}$ is obtained. In addition, the optimal transition probability $\boldsymbol{p}_{k,i}^{*,\mu}$ is rearranged as follows:

$$\begin{aligned} \boldsymbol{p}_{k,i}^{*,\mu} &= \frac{\bar{\boldsymbol{p}}_{k,i} \odot \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)}{\bar{\boldsymbol{p}}_{k,i}^{\top} \exp\left(-\mu^{-1}\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right]\right)} \\ &= \left(1/z_{k,i}^{\mu}\right)\bar{\boldsymbol{p}}_{k,i} \odot \exp\left(-\mu^{-1}\boldsymbol{w}_{k,i}\right) \odot \boldsymbol{z}_{k+1}^{\mu}. \end{aligned}$$

On transposing and vertically stacking the equation above, the matrix form of $\boldsymbol{P}_{k,i}^{*,\mu}$ in the theorem is obtained. ■

The algorithms of the DP for the optimal trajectory planning problem (20) with $\mu = 0$ (see [28]) and the matrix-valued KL control in Theorem 4 are presented as Algorithms 1 and 2, respectively.

Remark 5: The computation complexity of both the algorithms is given by $\mathcal{O}(N \cdot 2^{n_x} \cdot \min(2^{n_u}, 2^{n_x}))$ if appropriate implementation is considered. More precisely, in the fourth and fifth lines of Algorithm 1, because $\boldsymbol{w}_{k,i}$ has elements having a value of $+\infty$ and the number of the indices j satisfying $w_{k,i,j} < +\infty$ is not more than 2^{n_u} [see (17), (19)], the min and arg min operations need not take into consideration these elements. In Algorithm 2, the ith row vector of $\bar{\boldsymbol{P}}_k$, which is $\bar{\boldsymbol{p}}_{k,i}$, is a vector with nonzero elements not more than 2^{n_u} [see (17), (18)], that is, $\bar{\boldsymbol{P}}_k$ is a sparse matrix if $2^{n_u} \ll 2^{n_x}$ (for further details of sparse implementation, please refer to the Appendix). If the theoretical computation time is the same for the two algorithms, a unified comparison of their computation time cannot be obtained, and the practical computation time depends on the implementation.

It should be noted that the conventional DP (Algorithm 1) and KL control (Algorithm 2) can be used to solve Problem (20) with $\mu = 0$ and $\mu \in (0, +\infty)$, respectively; the target problem formulations of these two algorithms are independent and not overlapped; however, the obtained Algorithm 2 is consistent with the conventional Algorithm 1 in terms of computation time.

It should be noted that a very small value of μ causes $\exp(-\mu^{-1}[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}])$ and $\boldsymbol{z}_k^{\mu}$ to become almost zero and causes the overflow of $\mathrm{Diag}(\mathbf{1}_{2^{n_x}} \oslash \boldsymbol{z}_k^{\mu})$.

Algorithm 2 KL Control for Problem (20) With $\mu \in (0, +\infty)$

```
1: $\boldsymbol{v}_N^{\mathrm{KL},\mu} = \boldsymbol{h}, \boldsymbol{z}_N^{\mu} = \exp(-\mu^{-1}\boldsymbol{v}_N^{\mathrm{KL},\mu})$        ▷ Initialization
2: for $k = N-1, \ldots, 0$ do        ▷ Backward calculation on $k$
3:     $\boldsymbol{\Phi}_k^{\mu} = \overline{\boldsymbol{P}}_k \odot \exp(-\mu^{-1}\boldsymbol{W}_k)$
4:     $\boldsymbol{z}_k^{\mu} = \boldsymbol{\Phi}_k^{\mu}\boldsymbol{z}_{k+1}^{\mu}$
5:     $\boldsymbol{P}_k^{*,\mu} = \left(\mathrm{Diag}(\mathbf{1}_{2^{nx}} \oslash \boldsymbol{z}_k^{\mu})\boldsymbol{\Phi}_k^{\mu}\right) \odot \left[\boldsymbol{z}_{k+1}^{\mu} \cdots \boldsymbol{z}_{k+1}^{\mu}\right]^{\top}$
```

Remark 6: As the value function $v_{k,i}^{\mathrm{KL},\mu}$ and the optimal transition probability $\boldsymbol{p}_{k,i}^{*,\mu}$, given in Theorem 3, are interpreted as

$$\begin{aligned} v_{k,i}^{\mathrm{KL},\mu} &= -\mathtt{eLSE}\left(-\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right], \mu, \overline{\boldsymbol{p}}_{k,i}\right) \\ \boldsymbol{p}_{k,i}^{*,\mu} &= \mathtt{esoftmax}\left(-\left[\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\right], \mu, \overline{\boldsymbol{p}}_{k,i}\right) \end{aligned}$$

respectively. In other words, Algorithm 2 is the same as Algorithm 1 obtained on replacing the $\min(\cdot)$ and $\arg\min(\cdot)$ operations of the fourth and fifth lines at the kth iteration step with $-\mathtt{eLSE}(-(\cdot), \mu, \overline{\boldsymbol{p}}_{k,i})$ and $\mathtt{esoftmax}(-(\cdot), \mu, \overline{\boldsymbol{p}}_{k,i})$, respectively. To the best of the authors' knowledge, the form of Algorithm 2 has not been studied in the context of BCNs and related MDPs. It should be noted that the interpretation above is naturally derived from the theoretical analysis and not introduced as a heuristic algorithm. The subsequent analysis of $\mu \searrow 0$ implies that the KL control is a generalization of the conventional DP for $\mu = 0$.

Hereafter, the main focus of the remainder of this section is the convergence behavior of the KL control with respect to the weight parameter μ. The convergence behavior of the aforementioned approximation is summarized in the subsequent theorems. As observed in the arg min operation in Algorithm 1, an optimal solution of the conventional problem formulation, which is the case of $\mu = 0$ of Problem (20), is not necessarily a point but can be a simplex, which means that $\arg\min(\boldsymbol{a}) = \arg\min_{\boldsymbol{\delta} \in \mathbb{S}^n} \boldsymbol{\delta}^{\top}\boldsymbol{a} = \{\boldsymbol{\delta} \in \mathcal{S}^n | \delta_i = 0\,(a_i > \min(\boldsymbol{a}), i = 1, \ldots, n)\}$. Therefore, the convergence of a sequence $\boldsymbol{b}_k \in \mathcal{S}^n$ to $\arg\min(\boldsymbol{a}) \subset \mathcal{S}^n$, which is denoted as $\boldsymbol{b}_k \to \arg\min(\boldsymbol{a})$, is defined by $b_{k,i} \to 0$ for each i such that $a_i > \min(\boldsymbol{a})$.

Theorem 5: Algorithms 1 and 2 for the optimal trajectory planning problem (20) derived from the optimal control problem (10) with Assumption 1 are considered. In the limit of $\mu \searrow 0$,

1) $\boldsymbol{v}_k^{\mathrm{KL},\mu} \to \boldsymbol{v}_k^{\mathrm{DP}} (k = 0, \ldots, N)$.
2) $\boldsymbol{p}_{k,i}^{*,\mu} (k = 0, \ldots, N,\ i = 1, \ldots, 2^{n_x}) \to \arg\min(\boldsymbol{w}_{k,i} + \boldsymbol{v}_{k+1}^{\mathrm{DP}})$.

The proof of the theorem above can be referred to in the Appendix.

In contrast to the limit of $\mu \searrow 0$, the diverging $\mu \to +\infty$ emphasizes the minimization of the KL divergence. The objective function value obtained with the desired transition probabilities $\overline{\boldsymbol{p}}_{k,i}\,(k = 0, \ldots, N,\ i = 1, \ldots, 2^{n_x})$ is introduced as

$$\overline{v}_{k,i} = \begin{cases} \overline{\boldsymbol{p}}_{k,i}^{\top}\left(\boldsymbol{w}_{k,i} + \overline{\boldsymbol{v}}_{k+1}\right), & (k = 0, \ldots, N-1) \\ h_i, & (k = N). \end{cases}$$

The limiting behavior of $\mu \to +\infty$ is summarized in the following theorem.

Theorem 6: Algorithm 2 for for the optimal trajectory planning problem (20) derived from the optimal control problem (10) with Assumption 1 is considered. In the limit of $\mu \to +\infty$.

1) $\boldsymbol{v}_k^{\mathrm{KL},\mu} \to \overline{\boldsymbol{v}}_k\,(k = 0, \ldots, N)$.
2) $\boldsymbol{p}_{k,i}^{*,\mu} \to \overline{\boldsymbol{p}}_{k,i}\,(k = 0, \ldots, N,\ i = 1, \ldots, 2^{n_x})$.

The proof of the theorem above can be referred to in the Appendix.

Example 4: Here, the control problem of Example 3 for the BCNs (8) is considered again. The control period is set as $N = 3$, and the terminal cost h is set as

$$\boldsymbol{h} = \left[h\left(\delta_4^1\right), h\left(\delta_4^2\right), h\left(\delta_4^3\right), h\left(\delta_4^4\right)\right]^{\top} = [0.4, 0.7, 0, 0.3]^{\top}.$$

The value function and the optimal solution obtained using the conventional DP for $\mu = 0$ (Algorithm 1) and the KL control for $\mu \in (0, +\infty)$ (Algorithm 2) are compared. The matrix-valued value function and the control variable of the DP are defined as $\boldsymbol{V}^{\mathrm{DP}} = [\boldsymbol{v}_0^{\mathrm{DP}}, \ldots, \boldsymbol{v}_N^{\mathrm{DP}}]^{\top} \in \mathbb{R}^{(N+1)\times 2^{nx}}$ and $\boldsymbol{P}_k^{\mathrm{DP}} = [\boldsymbol{p}_{k,1}^{\mathrm{DP}}, \ldots, \boldsymbol{p}_{k,2^{nx}}^{\mathrm{DP}}]^{\top} \in \mathbb{R}^{2^{nx}\times 2^{nx}}$, respectively; those of the KL control, which are denoted as $\boldsymbol{V}^{\mathrm{KL},\mu}$ and $\boldsymbol{P}_k^{*,\mu}$, respectively, are similarly defined. As shown in the proof of Theorem 5 in the Appendix, the difference between $\boldsymbol{v}_k^{\mathrm{DP}}$ and $\boldsymbol{v}_k^{\mathrm{KL},\mu}$ increases backwards in k; therefore, this example provides the values of $\boldsymbol{P}_k$ at $k = 0$ because of the space limitations. In the case of $N = 3$, $\boldsymbol{V}^{\mathrm{DP}}$ and $\boldsymbol{P}_0^{\mathrm{DP}}$ are given as follows:

$$\left(\boldsymbol{V}^{\mathrm{DP}}, \boldsymbol{P}_0^{\mathrm{DP}}\right) = \begin{bmatrix} 0.9 & 1 & 1.5 & 1.5 \\ 0.6 & 0.8 & 1 & 1.1 \\ 0.1 & 0.6 & 0.5 & 0.7 \\ 0.4 & 0.7 & 0 & 0.3 \end{bmatrix}, \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \multicolumn{2}{c}{\boldsymbol{p}_{0,3,[3,4]}^{\mathrm{DP},\top}} \\ 0 & 0 & 0 & 1 \end{array}\right]$$

where $\boldsymbol{p}_{0,3,[3,4]}^{\mathrm{DP}}$ is an arbitrary vector in $\mathbb{S}^2$. The value function and optimal transition probabilities of the KL control in the cases of $\mu = 0.01$ and $\mu = 1$ are presented as follows:

$$\begin{aligned} &\left(\boldsymbol{V}^{\mathrm{KL},0.01}, \boldsymbol{P}_0^{*,0.01}\right) \\ &= \begin{bmatrix} 0.921 & 1.021 & 1.505 & 1.5 \\ 0.614 & 0.814 & 1.014 & 1.1 \\ 0.107 & 0.607 & 0.507 & 0.7 \\ 0.4 & 0.7 & 0 & 0.3 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &\left(\boldsymbol{V}^{\mathrm{KL},1}, \boldsymbol{P}_0^{*,1}\right) \\ &= \begin{bmatrix} 1.131 & 1.278 & 1.548 & 1.5 \\ 0.764 & 0.969 & 1.098 & 1.1 \\ 0.390 & 0.739 & 0.595 & 0.7 \\ 0.4 & 0.7 & 0 & 0.3 \end{bmatrix}, \\ &\quad \begin{bmatrix} 0 & 0.532 & 0.468 & 0 \\ 0 & 0.558 & 0 & 0.442 \\ 0 & 0 & 0.476 & 0.524 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

Furthermore, the differences $\boldsymbol{v}_0^{\mathrm{KL},\mu} - \boldsymbol{v}_0^{\mathrm{DP}}$ and $\overline{\boldsymbol{v}}_0 - \boldsymbol{v}_0^{\mathrm{KL},\mu}$ are depicted in Fig. 1. From the value above and Fig. 1, convergence is observed.

For varying the weight parameter μ, the optimal values without the KL divergence and the values of the KL divergence are illustrated in Fig. 2. The selected value of μ can balance the stage cost and KL divergence, which are indicated as x-axis and y-axis of Fig. 1, respectively.

V. Application Examples

A. Lac Operon Model

First, the lac operon model proposed in [35] (see [35] for further details of the model) is presented. The lac operon

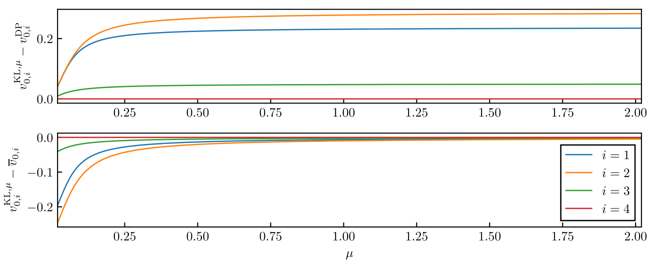

Fig. 1. Differences $v_{0,i}^{\mathrm{KL},\mu} - v_{0,i}^{\mathrm{DP}}$ and $\overline{v}_{0,i} - v_{0,i}^{\mathrm{KL},\mu}$ ($i = 1, 2, 3,$ and 4) with respect to the weight parameter $\mu > 0$.

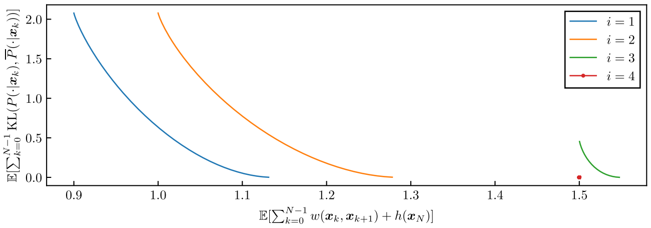

Fig. 2. Optimal values with varying weight parameter $\mu > 0$ ($v_{k,4}^{\mathrm{KL},\mu} = v_{k,4}^{\mathrm{DP}} = \overline{v}_{k,4} = 1.5$ with the zero value of the KL divergence degenerating to a point $(1.5, 0)$, as indicated in the plot above, because δ_4^1 cannot move to any other state).

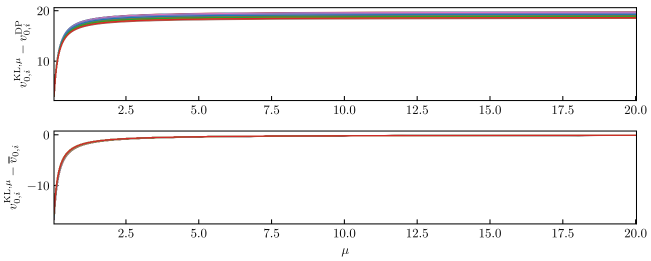

Fig. 3. Differences $v_{0,i}^{\mathrm{KL},\mu} - v_{0,i}^{\mathrm{DP}}$ and $\overline{v}_{0,i} - v_{0,i}^{\mathrm{KL},\mu}$ ($i = 1, \dots, 2^{10}$) with respect to the weight parameter $\mu > 0$.

model has the variables M (lac mRNA), B (β-galactosidase, LacZ), R (repressor protein, LacI), A (allolactose), L (lactose), P (transport protein, LacY, "lac permease"), and C (cAMP-CAP protein complex). A variable with a subscript "e" or "m" represents the extracellular and least medium concentrations, respectively; the other variables represent the intracellular concentrations. If a variable takes the values True or False, it indicates that its concentration is high or low, respectively. The state and control variables are given by $(x_1, \dots, x_{10}) = (M, B, R, A, L, P, C, R_{\mathrm{m}}, A_{\mathrm{m}}, L_{\mathrm{m}})$ and $(u_1, u_2, u_3) = (L_{\mathrm{e}}, L_{\mathrm{em}}, G_{\mathrm{e}})$, respectively. The Boolean update functions for the states are given as follows:

$$\begin{cases} f_M = C \land \neg R \land \neg R_{\mathrm{m}}, & f_P = M \\ f_B = M, & f_C = \neg G_{\mathrm{e}} \\ f_R = \neg A \land \neg A_{\mathrm{m}}, & f_{R_{\mathrm{m}}} = (\neg A \land \neg A_{\mathrm{m}}) \lor R \\ f_A = L \land B, & f_{A_{\mathrm{m}}} = L \lor L_{\mathrm{m}} \\ f_L = P \land L_{\mathrm{e}} \land \neg G_{\mathrm{e}}, & f_{L_{\mathrm{m}}} = ((L_{\mathrm{em}} \land P) \lor L_{\mathrm{e}}) \land \neg G_{\mathrm{e}} \end{cases}$$

In this numerical example, as in Example 3, the desired transition probabilities $\overline{p}_{k,i,j}$ were uniformly set to the reachable states of the next state for each $k = 0, \dots, N-1$, with $N = 100$, and the stage and terminal cost were randomly given. The differences $\boldsymbol{v}_0^{\mathrm{KL},\mu} - \boldsymbol{v}_0^{\mathrm{DP}}$ and $\overline{\boldsymbol{v}}_0 - \boldsymbol{v}_0^{\mathrm{KL},\mu}$ are depicted in Fig. 3, and the values of the cost and KL divergence are illustrated in Fig. 4.

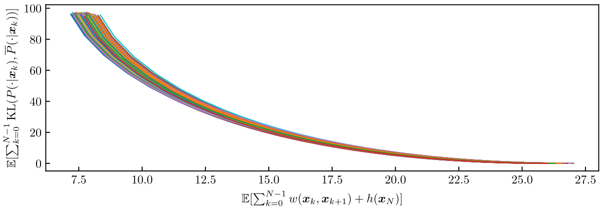

Fig. 4. Optimal values with varying weight parameter $\mu > 0$.

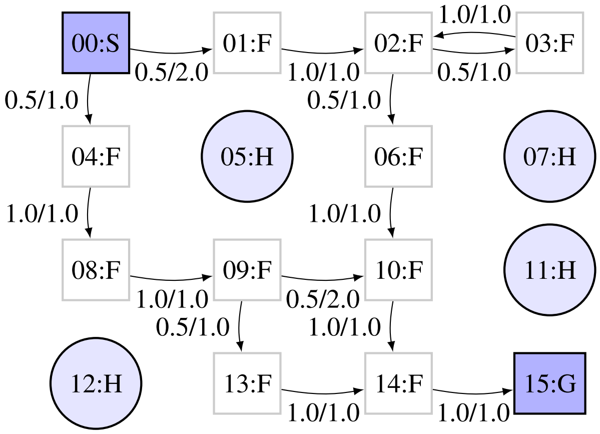

Fig. 5. Map of the frozen lake and associated desired transition probabilities and stage cost function. The values $*/*$ on the right-hand side of an edge represents the value of the time-invariant desired transition probability and the cost, respectively, indicating that $\overline{p}_{i,j}/w_{i,j}$.

B. Frozen Lake

As an example of a middle-scaled problem, a frozen lake [36] in Gymnasium is illustrated here. The detailed game setting is omitted here owing to space limitations but can be found in the Web documentation [36]. The frozen lake is not a Boolean model but a trajectory planning problem in the MDPs, and the proposed framework can be easily applied by setting the state space Δ^{16} rather than $\Delta^{2^{n_x}}$. Because deterministic systems are discussed in this article, the probabilistic slipping is ignored here.

In the map of the frozen lake (Fig. 5), a player starts at Start (S) and arrives at Goal (G) by moving on Frozen (F) ways. Because the player cannot move anymore after falling into Hall (H), the desired transition probabilities to the Halls are set as zero, which means that $\overline{p}_{k,i,05} = \overline{p}_{k,i,07} = \overline{p}_{k,i,11} = \overline{p}_{k,i,12} = 0$ for any k and i. In addition, an evidently time-consuming route, such as $04 \to 00$, is excluded by setting their desired transition probabilities to zero. The time-invariant desired transition probabilities and cost are given and indicated in a map (Fig. 5).

A situation with multiple people, such as an evacuation in the case of a disaster, is considered. A single optimal route, which the conventional DP provides, can result in congestion; therefore, the desired transition probabilities are selected to separate the flow of people, such as $\overline{p}_{00,01} = \overline{p}_{00,04} = 0.5$, although route $00 \to 04$ with cost $w_{00,04} = 1.0$ is more reasonable than $00 \to 01$ with $w_{00,01} = 2.0$. In addition, the buffer area 03 is utilized.

A long control period $N = 10$ is considered such that the player can reach the goal. The transition probabilities obtained are summarized in Table I. For a small value of μ, the minimization of the cost is emphasized, and the transition probability obtained of a nonoptimal route, such as $p_{00,01}$,

TABLE I
OBTAINED OPTIMAL TRANSITION PROBABILITIES

i	j	$p^{\mathrm{DP}}_{k,i}$	$p^{*,1}_{i,j}$	$p^{*,10}_{i,j}$	$\overline{p}_{i,j}$(desired)
00	01	0	0.225	0.453	0.5
00	04	1	0.775	0.547	0.5
02	03	0	0.068	0.417	0.5
02	06	1	0.932	0.583	0.5
09	10	0	0.269	0.475	0.5
09	13	1	0.731	0.525	0.5

takes a small value. In contrast, a large value of μ results in a small value of the KL divergence, indicating that the optimal transition probabilities are closed to the desired value.

Remark 7: The aforementioned numerical examples illustrate the advantages of the extended KL cost: 1) the extended KL cost can take into consideration the cost of the control (e.g., the transition costs indicated near the edges in Fig. 5), whereas the existing KL cost (e.g., [27] and [29]) does not take it into consideration and 2) using the weight parameter μ, the similarity to the desired transition probabilities and the transition costs can be balanced. Furthermore, the limiting cases $\mu \searrow 0$ and $\mu \to +\infty$ are theoretically supported by Theorems 5 and 6.

VI. CONCLUSION

This article addressed the optimal control problem with the stage cost function depending on the control input and the KL divergence. The introduced KL divergence can balance the objective function value and desired state transition probabilities. Furthermore, the convergence behavior of the KL control with respect to the weight parameter was presented.

In this work, the target system is a deterministic Boolean network, and thus, probabilistic Boolean networks (PBCNs) are not considered. The KL control problem for the PBCNs requires the consideration of innate stochastic behavior in the PBCNs and results in a more complicated discussion. It can be further investigated in future studies.

APPENDIX

A. Proof of Theorem 5

For notational simplicity, using a set $\mathcal{J}_{k,i} = \{j = 1, \dots, 2^{n_x},\ \overline{p}_{k,i,j} \neq 0\} = \{j_{k,i,1}, \dots, j_{k,i,|\mathcal{J}_{k,i}|}\}$, $\boldsymbol{D}_{k,i} = (\boldsymbol{\delta}_{2^{n_x}}[j_{k,i,1}, \dots, j_{k,i,|\mathcal{J}_{k,i}|}])^\top \in \mathbb{R}^{|\mathcal{J}_{k,i}| \times 2^{n_x}}$ is used. It should be noted that $\boldsymbol{D}_{k,i}\boldsymbol{w}_{k,i}$ comprises $w_{k,i,j}$ such that $\overline{p}_{k,i,j} > 0$, which means that $w_{k,i,j} = +\infty$ associated with $\overline{p}_{k,i,j} = 0$ is excluded. Similarly, $\boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}$ comprises $\overline{p}_{k,i,j} > 0$. On recalling that a term $\exp(-\infty) = 0$ in $\mathtt{eLSE}$ and the $+\infty/-\infty$ value in the max / min operations can be ignored, the following equations hold for an arbitrary vector $\boldsymbol{a} \in \mathbb{R}^{2^{n_x}}$:

$$\begin{cases} \mathtt{eLSE}\big(-\big[\boldsymbol{w}_{k,i} + \boldsymbol{a}\big], \mu, \overline{\boldsymbol{p}}_{k,i}\big) \\ \quad = \mathtt{eLSE}\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i} + \boldsymbol{a}\big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\big), \\ \min\big(\boldsymbol{w}_{k,i} + \boldsymbol{a}\big) = \min\big(\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i} + \boldsymbol{a}\big]\big) \\ \quad = -\max\big(-\big[\boldsymbol{w}_{k,i} + \boldsymbol{a}\big]\big) = -\max\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i} + \boldsymbol{a}\big]\big). \end{cases}$$

The modification using $\boldsymbol{D}_{k,i}$ above is introduced to exclude computations, including ∞ in fundamental inequalities given by Theorem 1.

1) Proof of Item 1): The difference $v^{\mathrm{KL},\mu}_{k,i} - v^{\mathrm{DP}}_{k,i}$ is separated into two parts using the equalities above

$$\begin{aligned} &v^{\mathrm{KL},\mu}_{k,i} - v^{\mathrm{DP}}_{k,i} \\ &\quad = -\mathtt{eLSE}\Big(-\Big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big], \mu, \overline{\boldsymbol{p}}_{k,i}\Big) - \min\big(\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{DP}}_{k+1}\big) \\ &\quad = \gamma^{(1)}_{k,i} + \gamma^{(2)}_{k,i} \end{aligned}$$

where

$$\begin{aligned} \gamma^{(1)}_{k,i} &= -\mathtt{eLSE}\Big(-\boldsymbol{D}_{k,i}\Big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\Big) \\ &\quad + \mathtt{eLSE}\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{DP}}_{k+1}\big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\big), \\ \gamma^{(2)}_{k,i} &= -\mathtt{eLSE}\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{DP}}_{k+1}\big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\big) \\ &\quad + \max\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{DP}}_{k+1}\big]\big). \end{aligned}$$

In this proof, the two terms $\gamma^{(1)}_{k,i}$ and $\gamma^{(2)}_{k,i}$ are lower- and upper-bounded using Theorem 1. More precisely, $\gamma^{(1)}_{k,i}$ is lower- and upper-bounded using (1) and (2), respectively, and $\gamma^{(2)}_{k,i}$ is lower- and upper-bounded using (3).

An inequality $\boldsymbol{v}^{\mathrm{KL},\mu}_k \geq \boldsymbol{v}^{\mathrm{DP}}_k$ $(k = 0, \dots, N-1)$ is obtained by induction. First, at $k = N-1$, $\boldsymbol{v}^{\mathrm{KL},\mu}_N = \boldsymbol{v}^{\mathrm{DP}}_N = \boldsymbol{h} = [h(\boldsymbol{\delta}^1_{2^{n_x}}), \dots, h(\boldsymbol{\delta}^{2^{n_x}}_{2^{n_x}})]^\top$ and (3) result in

$$\gamma^{(1)}_{N-1,i} = 0,\ 0 \leq \gamma^{(2)}_{N-1,i} \leq \mu \log\big(1/\min\big(\boldsymbol{D}_{N-1,i}\overline{\boldsymbol{p}}_{N-1,i}\big)\big)$$

respectively. Therefore, $v^{\mathrm{KL},\mu}_{N-1,i} - v^{\mathrm{DP}}_{N-1,i} = \gamma^{(1)}_{N-1,i} + \gamma^{(2)}_{N-1,i}$ satisfies the following inequalities:

$$0 \leq v^{\mathrm{KL},\mu}_{N-1,i} - v^{\mathrm{DP}}_{N-1,i} \leq \mu \log\big(1/\min\big(\boldsymbol{D}_{N-1,i}\overline{\boldsymbol{p}}_{N-1,i}\big)\big).$$

Second, $\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} \geq \boldsymbol{v}^{\mathrm{DP}}_{k+1}$ is assumed. Equations (1) and (2) with the substitution $\boldsymbol{x}' = -\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{DP}}_{k+1}]$, $\boldsymbol{x} = -\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}]$, and $\boldsymbol{p} = \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}$ bounds $\gamma^{(1)}_{k,i}$ are obtained as follows:

$$\begin{aligned} &\mathtt{esoftmax}^\top\Big(-\boldsymbol{D}_{k,i}\Big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\Big) \\ &\quad \cdot \boldsymbol{D}_{k,i}\Big[\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1}\Big] \\ &\leq \gamma^{(1)}_{k,i} \\ &\leq \mathtt{esoftmax}^\top\Big(-\boldsymbol{D}_{k,i}\Big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\Big) \\ &\quad \cdot \boldsymbol{D}_{k,i}\Big[\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1}\Big] + \frac{1}{2\mu}\Big\|\boldsymbol{D}_{k,i}\Big[\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1}\Big]\Big\|_2^2. \end{aligned} \tag{29}$$

The bound of an inner product $\mathtt{esoftmax}^\top(-\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i})\boldsymbol{D}_{k,i}[\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1}]$ is considered. The $\mathtt{esoftmax}$ term always takes a value in $\mathcal{S}^{|\mathcal{J}_{k,i}|}$, indicating that it is a non-negative vector, and the induction hypothesis $\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} \geq \boldsymbol{v}^{\mathrm{DP}}_{k+1}$ indicates that $\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1} \geq \boldsymbol{0}_{2^{n_x}}$

$$\begin{aligned} 0 \leq &\ \mathtt{esoftmax}^\top\Big(-\boldsymbol{D}_{k,i}\Big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\Big) \\ &\quad \cdot \boldsymbol{D}_{k,i}\Big[\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1}\Big]. \end{aligned}$$

For the upper bound, on using the Hölder inequality $\boldsymbol{a}^\top\boldsymbol{b} \leq \|\boldsymbol{a}\|_1\|\boldsymbol{b}\|_\infty = (\boldsymbol{1}^\top\boldsymbol{a})\max(\boldsymbol{b})$ for non-negative vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ (see [37, Appendix B]), the following upper bound is given:

$$\begin{aligned} &\mathtt{esoftmax}^\top\Big(-\boldsymbol{D}_{k,i}\Big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\Big) \\ &\quad \cdot \boldsymbol{D}_{k,i}\Big[\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1}\Big] \\ &\leq \Big(\boldsymbol{1}^\top_{2^{n_x}}\mathtt{esoftmax}\Big(-\boldsymbol{D}_{k,i}\Big[\boldsymbol{w}_{k,i} + \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big], \mu, \boldsymbol{D}_{k,i}\overline{\boldsymbol{p}}_{k,i}\Big)\Big) \\ &\quad \cdot \max\Big(\boldsymbol{D}_{k,i}\Big[\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1}\Big]\Big) \\ &= \max\Big(\boldsymbol{D}_{k,i}\Big[\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1} - \boldsymbol{v}^{\mathrm{DP}}_{k+1}\Big]\Big). \end{aligned}$$

On applying the aforementioned upper and lower bounds of $\mathtt{esoftmax}^\top(-\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i}+\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i})\boldsymbol{D}_{k,i}[\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}]$, the bound of $\gamma_{k,i}^{(1)}$ in (29) is simplified as follows:

$$\begin{aligned}0 &\le \gamma_{k,i}^{(1)}\\ &\le \max\Big(\boldsymbol{D}_{k,i}\Big[\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\Big]\Big)+\frac{1}{2\mu}\Big\|\boldsymbol{D}_{k,i}\Big[\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\Big]\Big\|_2^2.\end{aligned} \tag{30}$$

For $\gamma_{k,i}^{(2)}$, (3) results in the following inequality:

$$0\le\gamma_{k,i}^{(2)}\le\mu\log\big(1/\min\big(\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\big). \tag{31}$$

Equations (30) and (31) show that $v_{k,i}^{\mathrm{KL},\mu}-v_{k,i}^{\mathrm{DP}}=\gamma_{k,i}^{(1)}+\gamma_{k,i}^{(2)}$ satisfies the following inequalities:

$$\begin{aligned}0\le v_{k,i}^{\mathrm{KL},\mu}-v_{k,i}^{\mathrm{DP}} &\le \max\Big(\boldsymbol{D}_{k,i}\Big[\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\Big]\Big)\\ &+\frac{1}{2\mu}\Big\|\boldsymbol{D}_{k,i}\Big[\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\Big]\Big\|_2^2+\mu\log\big(1/\min\big(\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\big).\end{aligned}$$

Therefore, the induction concludes that $\boldsymbol{v}_k^{\mathrm{KL},\mu}\ge\boldsymbol{v}_k^{\mathrm{DP}}$ $(k=0,\dots,N-1)$. On taking the maximum value of both the sides and using $C_1=\max_{i=1,\dots,2^{n_x},\,k=0,\dots,N-1}[1/\min(\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i})]$, the following equation is obtained:

$$\begin{aligned}&\max\Big(\boldsymbol{v}_k^{\mathrm{KL},\mu}-\boldsymbol{v}_k^{\mathrm{DP}}\Big)\\ &\le\max\Big(\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\Big)+\frac{1}{2\mu}\Big\|\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\Big\|_2^2+\mu C_1.\end{aligned} \tag{32}$$

The obtained inequality is a recursive equation of $\max(\boldsymbol{v}_k^{\mathrm{KL},\mu}-\boldsymbol{v}_k^{\mathrm{DP}})$; however, an explicit expression of the solution is difficult to obtain. Instead, using a sequence c_k that does not depend on μ, the remaining part is the proof for an inequality $\max(\boldsymbol{v}_k^{\mathrm{KL},\mu}-\boldsymbol{v}_k^{\mathrm{DP}})\le c_k\mu$. At N, the definition of $\boldsymbol{v}_N^{\mathrm{KL},\mu}=\boldsymbol{h}$ indicates $\boldsymbol{v}_N^{\mathrm{KL},\mu}=\boldsymbol{v}_N^{\mathrm{DP}}$ and $\max(\boldsymbol{v}_N^{\mathrm{KL},\mu}-\boldsymbol{v}_N^{\mathrm{DP}})=0=c_N\mu$ with $c_N=0$. At $k+1$, the inequality $\max(\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}})\le c_{k+1}\mu$ is assumed. Then, the ℓ_2 squared norm in (32) is bounded as

$$\begin{aligned}\|\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\|_2^2 &\le\Big[\max\Big(\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\Big)\Big]^2\|\boldsymbol{1}_{2^{n_x}}\|_2^2\\ &\le 2^{n_x}\Big[\max\Big(\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}-\boldsymbol{v}_{k+1}^{\mathrm{DP}}\Big)\Big]^2\le 2^{n_x}c_{k+1}^2\mu^2\end{aligned}$$

and (32) results in the following bound:

$$\max\Big(\boldsymbol{v}_k^{\mathrm{KL},\mu}-\boldsymbol{v}_k^{\mathrm{DP}}\Big)\le\Big(c_{k+1}+2^{n_x-1}c_{k+1}^2+C_1\Big)\mu$$

and the inequality $\max(\boldsymbol{v}_k^{\mathrm{KL},\mu}-\boldsymbol{v}_k^{\mathrm{DP}})\le c_k\mu$ is obtained by letting $c_k=c_{k+1}+2^{n_x-1}c_{k+1}^2+C_1$. The inequalities $\max(\boldsymbol{v}_k^{\mathrm{KL},\mu}-\boldsymbol{v}_k^{\mathrm{DP}})\le c_k\mu$ and $\boldsymbol{v}_k^{\mathrm{KL},\mu}\ge\boldsymbol{v}_k^{\mathrm{DP}}$ means that $\boldsymbol{v}_k^{\mathrm{KL},\mu}\to\boldsymbol{v}_k^{\mathrm{DP}}$ as $\mu\searrow 0$.

2) Proof of Item 2): The jth element of the optimal $\boldsymbol{p}_{k,i}^{*,\mu}\in\mathbb{S}^{2^{n_x}}$ is given as follows:

$$\begin{aligned}p_{k,i,j}^{*,\mu} &= \Big[\mathtt{esoftmax}\Big(-\Big(\boldsymbol{w}_{k,i}+\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\Big),\mu,\bar{\boldsymbol{p}}_{k,i}\Big)\Big]_j\\ &=\frac{\bar{p}_{k,i,j}\exp\Big(-\mu^{-1}\Big[w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}\Big]\Big)}{\sum_{j'=1}^{2^{n_x}}\bar{p}_{k,i,j'}\exp\Big(-\mu^{-1}\Big[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}\Big]\Big)}\\ &=\frac{\bar{p}_{k,i,j}}{\sum_{j'=1}^{2^{n_x}}\bar{p}_{k,i,j'}\exp\Big(\mu^{-1}\Big(\Big[w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}\Big]-\Big[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}\Big]\Big)\Big)}.\end{aligned}$$

The convergence of $\bar{p}_{k,i,j'}\exp(\mu^{-1}([w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}]-[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}]))$ is summarized as follows.

1) $w_{k,i,j}+v_{k+1,j}^{\mathrm{DP}}>w_{k,i,j'}+v_{k+1,j'}^{\mathrm{DP}}$: There exists $\varepsilon>0$ and $\underline{\mu}>0$ such that $[w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}]-[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}]>\varepsilon>0$ provided $0<\mu\le\underline{\mu}$ because of the convergence of $\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\to\boldsymbol{v}_{k+1}^{\mathrm{DP}}$. Therefore, $\bar{p}_{k,i,j'}\exp(\mu^{-1}([w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}]-[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}]))\ge\bar{p}_{k,i,j'}\exp(\mu^{-1}\varepsilon)$ if $0<\mu\le\underline{\mu}$ and diverges to $+\infty$ as $\mu\searrow 0$.
2) $w_{k,i,j}+v_{k+1,j}^{\mathrm{DP}}<w_{k,i,j'}+v_{k+1,j'}^{\mathrm{DP}}$: Similar to the discussion above, there exists $-\varepsilon<0$ and $\underline{\mu}>0$ such that $[w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}]-[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}]<-\varepsilon<0$ for $0<\mu\le\underline{\mu}$ because of the convergence of $\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\to\boldsymbol{v}_{k+1}^{\mathrm{DP}}$. Therefore, $0\le\bar{p}_{k,i,j'}\exp(\mu^{-1}([w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}]-[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}]))\le\bar{p}_{k,i,j'}\exp(-\mu^{-1}\varepsilon)$ if $0<\mu\le\underline{\mu}$ and converges to zero as $\mu\searrow 0$.
3) $w_{k,i,j}+v_{k+1,j}^{\mathrm{DP}}=w_{k,i,j'}+v_{k+1,j'}^{\mathrm{DP}}$: Inequalities $\boldsymbol{v}_k^{\mathrm{KL},\mu}\ge\boldsymbol{v}_k^{\mathrm{DP}}$ and $\max(\boldsymbol{v}_k^{\mathrm{KL},\mu}-\boldsymbol{v}_k^{\mathrm{DP}})\le c_k\mu$ implies that $[w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}]-[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}]=[v_{k+1,j}^{\mathrm{KL},\mu}-v_{k+1,j}^{\mathrm{DP}}]-[v_{k+1,j'}^{\mathrm{KL},\mu}-v_{k+1,j'}^{\mathrm{DP}}]\in[-c_{k+1}\mu,c_{k+1}\mu]$. Therefore, $\bar{p}_{k,i,j'}\exp(\mu^{-1}([w_{k,i,j}+v_{k+1,j}^{\mathrm{KL},\mu}]-[w_{k,i,j'}+v_{k+1,j'}^{\mathrm{KL},\mu}]))$ is bounded.

Item 1) concludes that $p_{k,i,j}^{*,\mu}\to 0$ if $w_{k,i,j}+v_{k+1,j+1}^{\mathrm{DP}}>\min(\boldsymbol{w}_{k,i}+\boldsymbol{v}_{k+1}^{\mathrm{DP}})$. The discussion above does not specify the convergence point, but the convergence of $\boldsymbol{p}_{k,i}^{*,\mu}\to\arg\min(\boldsymbol{w}_{k,i}+\boldsymbol{v}_{k+1}^{\mathrm{DP}})$ is established.

B. Proof of Theorem 6

1) Proof of Item 1): The flow of the proof is similar to that of Theorem 5, that is, the difference $\boldsymbol{v}_k^{\mathrm{KL},\mu}-\bar{\boldsymbol{v}}_k$ is lower- and upper-bounded. The difference $v_{k,i}^{\mathrm{KL},\mu}-\bar{v}_{k,i}$ is separated into two parts

$$\begin{aligned}v_{k,i}^{\mathrm{KL},\mu}-\bar{v}_{k,i} &= -\mathtt{eLSE}\Big(-\Big[\boldsymbol{w}_{k,i}+\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\Big],\mu,\bar{\boldsymbol{p}}_{k,i}\Big)\\ &\quad-\bar{\boldsymbol{p}}_{k,i}^\top\big(\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big)=\eta_{k,i}^{(1)}+\eta_{k,i}^{(2)}\end{aligned}$$

where

$$\begin{aligned}\eta_{k,i}^{(1)} &= -\mathtt{eLSE}\Big(-\boldsymbol{D}_{k,i}\Big[\boldsymbol{w}_{k,i}+\boldsymbol{v}_{k+1}^{\mathrm{KL},\mu}\Big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\Big)\\ &\quad+\mathtt{eLSE}\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\\ \eta_{k,i}^{(2)} &= -\mathtt{eLSE}\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\\ &\quad+\big(\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)^\top\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big]\big).\end{aligned}$$

In the equation above, $-\bar{\boldsymbol{p}}_{k,i}^\top(\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1})=(\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i})^\top(-\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}])$ is used. An inequality $\bar{\boldsymbol{v}}_k\ge\boldsymbol{v}_k^{\mathrm{KL},\mu}$ $(k=0,\dots,N-1)$ is obtained by induction. First, at $k=N-1$, $\boldsymbol{v}_N^{\mathrm{KL},\mu}=\bar{\boldsymbol{v}}_N=\boldsymbol{h}=[h(\delta_{2^{n_x}}^1),\dots,h(\delta_{2^{n_x}}^{2^{n_x}})]^\top$, and (4) results in

$$\eta_{N-1,i}^{(1)}=0,\ -\frac{1}{2\mu}\big\|\boldsymbol{D}_{N-1,i}\big[\boldsymbol{w}_{N-1,i}+\boldsymbol{h}\big]\big\|_2^2\le\eta_{N-1,i}^{(2)}\le 0$$

respectively. Therefore, $v_{N-1,i}^{\mathrm{KL},\mu}-\bar{v}_{N-1,i}=\eta_{N-1,i}^{(1)}+\eta_{N-1,i}^{(2)}$ satisfies the following inequalities:

$$-\frac{1}{2\mu}\big\|\boldsymbol{D}_{N-1,i}\big[\boldsymbol{w}_{N-1,i}+\boldsymbol{h}\big]\big\|_2^2\le v_{N-1,i}^{\mathrm{KL},\mu}-\bar{v}_{N-1,i}\le 0.$$

Second, $\bar{\boldsymbol{v}}_{k+1} \geq \boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}$ is assumed, and (1) and (2) with the substitutions $\boldsymbol{x}' = -\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i}+\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}]$, $\boldsymbol{x} = -\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}]$, and $\boldsymbol{p} = \boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}$ provide the following lower and upper bounds of $-\eta^{(1)}_{k,i}$:

$$\begin{aligned}
&\mathtt{esoftmax}^\top\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\\
&\quad\cdot\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\\
&\leq -\eta^{(1)}_{k,i}\\
&= \mathtt{eLSE}\Big(-\boldsymbol{D}_{k,i}\Big[\boldsymbol{w}_{k,i}+\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\Big)\\
&\quad -\mathtt{eLSE}\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\\
&\leq \mathtt{esoftmax}^\top\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\\
&\quad\cdot\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]+\frac{1}{2\mu}\Big\|\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\Big\|_2^2.
\end{aligned}\tag{33}$$

The bound of an inner product term $\mathtt{esoftmax}^\top(-\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i})\boldsymbol{D}_{k,i}[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}]$ is given using the Hölder inequality as follows:

$$\begin{aligned}
&0\leq \mathtt{esoftmax}^\top\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\\
&\quad\cdot\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\\
&\leq\Big(\boldsymbol{1}^\top_{2^{n_x}}\mathtt{esoftmax}\big(-\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big],\mu,\boldsymbol{D}_{k,i}\bar{\boldsymbol{p}}_{k,i}\big)\Big)\\
&\quad\cdot\max\Big(\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\Big)\\
&=\max\Big(\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\Big).
\end{aligned}$$

On using the inequalities above, the bound of $-\eta^{(1)}_{k,i}$ in (33) is simplified as follows:

$$\begin{aligned}
&0\leq -\eta^{(1)}_{k,i}\\
&\leq\max\Big(\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\Big)+\frac{1}{2\mu}\Big\|\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\Big\|_2^2.
\end{aligned}\tag{34}$$

Furthermore, (4) results in the following inequalities:

$$-\frac{1}{2\mu}\big\|\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big]\big\|_2^2\leq\eta^{(2)}_{k,i}\leq 0.\tag{35}$$

On combining (34) and (35), the following bounds of $v^{\mathrm{KL},\mu}_{k,i}-\bar{v}_{k,i}=\eta^{(1)}_{k,i}+\eta^{(2)}_{k,i}$ is obtained:

$$\begin{aligned}
0&\geq v^{\mathrm{KL},\mu}_{k,i}-\bar{v}_{k,i}\\
&\geq -\max\Big(\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\Big)-\frac{1}{2\mu}\|\boldsymbol{D}_{k,i}\Big[\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\|_2^2\\
&\quad-\frac{1}{2\mu}\big\|\boldsymbol{D}_{k,i}\big[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}\big]\big\|_2^2.
\end{aligned}\tag{36}$$

By induction, an inequality $\bar{\boldsymbol{v}}_k\geq\boldsymbol{v}^{\mathrm{KL},\mu}_k$ is obtained. By using $C_2=\max_{i=1,\ldots,2^{n_x},\,k=0,\ldots,N-1}(1/2)\|\boldsymbol{D}_{k,i}[\boldsymbol{w}_{k,i}+\bar{\boldsymbol{v}}_{k+1}]\|_2^2$, (36) results in

$$\begin{aligned}
\max\Big(\bar{\boldsymbol{v}}_k-\boldsymbol{v}^{\mathrm{KL},\mu}_k\Big)&\leq\max\Big(\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big)\\
&\quad+\frac{1}{2\mu}\Big\|\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big\|_2^2+\mu^{-1}C_2.
\end{aligned}\tag{37}$$

Now, an inequality $\max(\bar{\boldsymbol{v}}_k-\boldsymbol{v}^{\mathrm{KL},\mu}_k)\leq e_k\mu^{-1}$ with a sequence $e_k\geq 0\,(k=0,\ldots,N-1)$ is derived here. At $k+1$, the inequality $\max(\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1})\leq e_{k+1}\mu^{-1}$ with $e_{k+1}\geq 0$ is assumed. The ℓ_2 squared norm is bounded as

$$\begin{aligned}
&\Big\|\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\|_2^2\leq\Big[\max\Big(\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big)\Big]^2\Big\|\boldsymbol{1}_{2^{n_x}}\|_2^2\\
&\leq 2^{n_x}\Big[\max\Big(\bar{\boldsymbol{v}}_{k+1}-\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big)\Big]^2\leq 2^{n_x}e^2_{k+1}\mu^{-2}.
\end{aligned}$$

The equation above applies to (37) and shows that

$$\max\Big(\bar{\boldsymbol{v}}_k-\boldsymbol{v}^{\mathrm{KL},\mu}_k\Big)\leq e_{k+1}\mu^{-1}+2^{n_x-1}e^2_{k+1}\mu^{-3}+\mu^{-1}C_2.$$

Because the limit of $\mu\to+\infty$ is now considered, $\mu^{-3}\leq\mu^{-1}$ is assumed without a loss of generality. Therefore, $\max(\bar{\boldsymbol{v}}_k-\boldsymbol{v}^{\mathrm{KL},\mu}_k)\leq e_k\mu^{-1}$ with $e_k=e_{k+1}+2^{n_x-1}e^2_{k+1}+C_2$. The inequalities $\bar{\boldsymbol{v}}_k\geq\boldsymbol{v}^{\mathrm{KL},\mu}_k$ and $\max(\bar{\boldsymbol{v}}_k-\boldsymbol{v}^{\mathrm{KL},\mu}_k)\leq e_k\mu^{-1}$ result in $\boldsymbol{v}^{\mathrm{KL},\mu}_k\to\bar{\boldsymbol{v}}_k$ as $\mu\to+\infty$.

2) Proof of Item 2): The jth element of $\boldsymbol{p}^{*,\mu}_{k,i}$ is given as follows:

$$\begin{aligned}
p^{*,\mu}_{k,i,j}&=\Big[\mathtt{esoftmax}\Big(-\Big(\boldsymbol{w}_{k,i}+\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big),\mu,\bar{\boldsymbol{p}}_{k,i}\Big)\Big]_j\\
&=\frac{\bar{p}_{k,i,j}\exp\Big(-\mu^{-1}\Big[w_{k,i,j}+v^{\mathrm{KL},\mu}_{k+1,j}\Big]\Big)}{\bar{\boldsymbol{p}}^\top_{k,i}\exp\Big(-\mu^{-1}\Big[\boldsymbol{w}_{k,i}+\boldsymbol{v}^{\mathrm{KL},\mu}_{k+1}\Big]\Big)}.
\end{aligned}$$

In the exponential terms above

$$\begin{aligned}
&\exp\Big(-\mu^{-1}\Big[w_{k,i,j}+v^{\mathrm{KL},\mu}_{k+1,j}\Big]\Big)\\
&\quad=\exp\Big(-\mu^{-1}\big[w_{k,i,j}+\bar{v}_{k+1,j}\big]\Big)\cdot\exp\Big(\mu^{-1}\Big[\bar{v}_{k+1,j}-v^{\mathrm{KL},\mu}_{k+1,j}\Big]\Big).
\end{aligned}\tag{38}$$

The first term of the right side of (38), $\exp(-\mu^{-1}[w_{k,i,j}+\bar{v}_{k+1,j}])$, converges to 1 as $\mu\to+\infty$. The inequalities $0\leq\bar{v}_{k+1,j}-v^{\mathrm{KL},\mu}_{k+1,j}\leq e_{k+1}\mu^{-1}$ result in upper and lower bounds of the second term of the right side of (38) given by

$$\exp(0)\leq\exp\Big(\mu^{-1}\Big[\bar{v}_{k+1,j}-v^{\mathrm{KL},\mu}_{k+1,j}\Big]\Big)\leq\exp\Big(e_{k+1}\mu^{-2}\Big)$$

that is, $\exp(\mu^{-1}[\bar{v}_{k+1,j}-v^{\mathrm{KL},\mu}_{k+1,j}])$ also converges to 1. Therefore, $p^{*,\mu}_{k,i,j}\to\bar{p}_{k,i,j}/(\bar{\boldsymbol{p}}^\top_{k,i}\boldsymbol{1}_{2^{n_x}})=\bar{p}_{k,i,j}$ as $\mu\to+\infty$, which means that $\boldsymbol{p}^{*,\mu}_{k,i}\to\bar{\boldsymbol{p}}_{k,i}$.

C. Sparse Implementation of Algorithm 2

For practical implementation, the third line $\boldsymbol{\Phi}^\mu_k=\bar{\boldsymbol{P}}_k\odot\exp(-\mu^{-1}\boldsymbol{W}_k)$ can consume a significant amount of time because naive implementation results in a dense matrix $\exp(-\mu^{-1}\boldsymbol{W}_k)$. In particular, $\boldsymbol{W}_k$, defined in (19), has many $+\infty$ elements, and $\exp(-\mu^{-1}\boldsymbol{W}_k)$ is computed in the manner of the dense matrix computation. Instead, the following sparse matrix $\boldsymbol{W}'_k$ can be used:

$$W'_{k,i,j}=\begin{cases}g_k\Big(\boldsymbol{\delta}^i_{2^{n_x}},\boldsymbol{\delta}^{\mathtt{inv}^*(j|i,k)}_{2^{n_u}}\Big) & \big(\bar{p}_{k,i,j}>0\big),\\ 0 & \big(\bar{p}_{k,i,j}=0\big).\end{cases}$$

Because $\bar{p}_{k,i,j}\exp(-\mu^{-1}W'_{k,i,j})=\bar{p}_{k,i,j}\exp(-\mu^{-1}W_{k,i,j})=0$ if $\bar{p}_{k,i,j}=0$, the modified matrix $\boldsymbol{W}'_k$ of $\boldsymbol{W}_k$ does not affect the value of $\boldsymbol{\Phi}^\mu_k$, which means that $\bar{\boldsymbol{P}}_k\odot\exp(-\mu^{-1}\boldsymbol{W}'_k)=\bar{\boldsymbol{P}}_k\odot\exp(-\mu^{-1}\boldsymbol{W}_k)=\boldsymbol{\Phi}^\mu_k$. However, $\exp(-\mu^{-1}\boldsymbol{W}'_k)$ has many elements with the value of 1 and is still dense. Therefore, a common implementation named `expm1`, which computes $\mathtt{expm1}(x) = \exp(x) - 1$, can be used as $\boldsymbol{\Phi}_k^{\mu} = \overline{\boldsymbol{P}}_k + \overline{\boldsymbol{P}}_k \odot \mathtt{expm1}(-\mu^{-1}\boldsymbol{W}'_k)$; both $\overline{\boldsymbol{P}}_k$ and $\mathtt{expm1}(-\mu^{-1}\boldsymbol{W}'_k)$ are sparse, and $\boldsymbol{\Phi}_k^{\mu}$ is efficiently computed.

References

[1] B. P. Kramer, C. Fischer, and M. Fussenegger, "BioLogic gates enable logical transcription control in mammalian cells," *Biotechnol. Bioeng.*, vol. 87, no. 4, pp. 478–484, 2004.

[2] Y. Wu and T. Shen, "Policy iteration algorithm for optimal control of stochastic logical dynamical systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 2031–2036, May 2018.

[3] M. Kang, Y. Wu, and T. Shen, "Logical control approach to fuel efficiency optimization for commuting vehicles," *Int. J. Automot. Technol.*, vol. 18, no. 3, pp. 535–546, Jun. 2017.

[4] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *J. Theor. Biol.*, vol. 22, no. 3, pp. 437–467, 1969.

[5] D. Cheng and H. Qi, "A linear representation of dynamics of Boolean networks," *IEEE Trans. Autom. Control*, vol. 55, no. 10, pp. 2251–2258, Oct. 2010.

[6] D. Cheng, H. Qi, and Z. Li, *Analysis and Control of Boolean Networks: A Semi-Tensor Product Approach*. London, U.K.: Springer, 2011. [Online]. Available: https://link.springer.com/book/10.1007/978-0-85729-097-7

[7] G. Jia, M. Meng, J. Lam, and J.-E. Feng, "Further results for pinning stabilization of Boolean networks," *IEEE Trans. Control Netw.*, vol. 8, no. 2, pp. 897–905, Jun. 2021.

[8] H. Chen, Z. Wang, J. Liang, and M. Li, "State estimation for stochastic time-varying Boolean networks," *IEEE Trans. Autom. Control*, vol. 65, no. 12, pp. 5480–5487, Dec. 2020.

[9] Y. Chen, J. Yang, Z. Li, and X. Bu, "State estimation via designing controller and state estimation-based stabilization for Boolean control networks," *IEEE Trans. Cybern.*, early access, Nov. 16, 2022, doi: 10.1109/TCYB.2022.3219522.

[10] Y. Yu, J.-E. Feng, J. Pan, and D. Cheng, "Block decoupling of Boolean control networks," *IEEE Trans. Autom. Control*, vol. 64, no. 8, pp. 3129–3140, Aug. 2019.

[11] L. Lin, J. Cao, and L. Rutkowski, "Robust event-triggered control invariance of probabilistic Boolean control networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 1060–1065, Mar. 2020.

[12] Y. Guo, Q. Li, and W. Gui, "Optimal state estimation of Boolean control networks with stochastic disturbances," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1355–1359, Mar. 2020.

[13] Q. Chen, H. Li, and X. Yang, "Total-activity conservation analysis and design of Boolean networks," *IEEE Trans. Cybern.*, early access, Apr. 13, 2022, doi: 10.1109/TCYB.2022.3163608.

[14] H. Li and Y. Wang, "Lyapunov-based stability and construction of Lyapunov functions for Boolean networks," *SIAM J. Control Optim.*, vol. 55, no. 6, pp. 3437–3457, 2017.

[15] J. Liu, Y. Liu, Y. Guo, and W. Gui, "Sampled-data state-feedback stabilization of probabilistic Boolean control networks: A control Lyapunov function approach," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3928–3937, Sep. 2020.

[16] A. Acernese, A. Yerudkar, L. Glielmo, and C. D. Vecchio, "Reinforcement learning approach to feedback stabilization problem of probabilistic Boolean control networks," *IEEE Contr. Syst. Lett.*, vol. 5, pp. 337–342, 2021.

[17] Y. Wang, X. Ding, L. Lin, J. Lu, and J. Lou, "State-feedback set stabilization of Boolean networks with state-dependent random impulses," *IEEE Trans. Cybern.*, early access, Oct. 20, 2022, doi: 10.1109/TCYB.2022.3209982.

[18] L. Wang, M. Fang, Z.-G. Wu, and J. Lu, "Necessary and sufficient conditions on pinning stabilization for stochastic Boolean networks," *IEEE Trans. Cybern.*, vol. 50, no. 10, pp. 4444–4453, Oct. 2020.

[19] F. Li and Y. Tang, "Pinning controllability for a Boolean network with arbitrary disturbance inputs," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 3338–3347, Jun. 2021.

[20] J. Lu, J. Yang, J. Lou, and J. Qiu, "Event-triggered sampled feedback synchronization in an array of output-coupled Boolean control networks," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 2278–2283, Apr. 2021.

[21] M. Meng, G. Xiao, and D. Cheng, "Self-triggered scheduling for Boolean control networks," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 8911–8921, Sep. 2022.

[22] M. Xu, Y. Liu, J. Lou, Z.-G. Wu, and J. Zhong, "Set stabilization of probabilistic Boolean control networks: A sampled-data control approach," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3816–3823, Aug. 2020.

[23] B. Chen, J. Cao, G. Lu, and L. Rutkowski, "Stabilization of Markovian jump Boolean control networks via sampled-data control," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10290–10301, Oct. 2022.

[24] Y. Wu and T. Shen, "Dynamic programming algorithm for stochastic logical systems and its application to residual gas fraction control," in *Proc. ISCIE Int. Symp. Stochastic Syst. Theory Appl.*, 2015, pp. 136–141.

[25] W. Abou-Jaoudé, D. A. Ouattara, and M. Kaufman, "From structure to dynamics: Frequency tuning in the p53-Mdm2 network: I. Logical approach," *J. Theor. Biol.*, vol. 258, no. 4, pp. 561–577, 2009.

[26] Y. Liu, H. Chen, J. Lu, and B. Wu, "Controllability of probabilistic Boolean control networks based on transition probability matrices," *Automatica*, vol. 52, pp. 340–345, Feb. 2015.

[27] T. Matsubara, V. Gómez, and H. J. Kappen, "Latent Kullback Leibler control for continuous-state systems using probabilistic graphical models," in *Proc. 30th Conf. Uncertainty Artif. Intell.*, 2014, pp. 583–592.

[28] Y. Wu and T. Shen, "An algebraic expression of finite horizon optimal control algorithm for stochastic logical dynamical systems," *Syst. Control Lett.*, vol. 82, pp. 108–114, Aug. 2015.

[29] E. Todorov, "Linearly-solvable Markov decision problems," in *Advances in Neural Information Processing Systems*, vol. 19, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA, USA: MIT Press, 2006.

[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[31] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA, USA: Soc. Ind. Appl. Math., 2017.

[32] E. Hazan, "Introduction to online convex optimization," *Found. Trends Optim.*, vol. 2, nos. 3–4, pp. 157–325, 2016.

[33] A. Datta, E. R. Dougherty, M. L. Bittner, and R. Pal, "Intervention in context-sensitive probabilistic Boolean networks," *Bioinformatics*, vol. 21, no. 7, pp. 1211–1218, Nov. 2004.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[35] S. Chen, Y. Wu, M. Macauley, and X.-M. Sun, "Monostability and bistability of Boolean networks using semitensor products," *IEEE Trans. Control. Netw.*, vol. 6, no. 4, pp. 1379–1390, Dec. 2019.

[36] The Farama Foundation. "Frozen lake." Gymnasium Documentation. 2023. [Online]. Available: https://gymnasium.farama.org/environments/toy_text/frozen_lake/

[37] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2012.

Mitsuru Toyoda (Member, IEEE) received the Ph.D. degree in mechanical engineering from Sophia University, Tokyo, Japan, in March 2018.

Since April 2018, he has held a Project Assistant Professor position with the Institute of Statistical Mathematics, Tokyo. In April 2019, he joined the Department of Mechanical Systems Engineering, Tokyo Metropolitan University, Tokyo, where he is currently an Assistant Professor. His research interests are related to optimal control and stochastic system control theory.

Yuhu Wu (Member, IEEE) received the Ph.D. degree in mathematics from Harbin Institute of Technology, Harbin, China, in January 2012.

Since September 2012, he has held an Assistant Professor position with Harbin University of Science and Technology, Harbin. He held a Postdoctoral Research position with Sophia University, Tokyo, Japan, from April 2012 to September 2015. In October 2015, he joined the School of Control Science and Engineering, Dalian University of Technology, Dalian, China, where he is currently a Full Professor. His research interests are related to Boolean networks, nonlinear control theory and applications of control to automotive powertrain systems, and unmanned aerial vehicles.