Optimal Control of Complex Systems through Variational Inference with a Discrete Event Decision Process

Fan Yang
University at Buffalo
Buffalo, New York
fyang24@buffalo.edu

Bo Liu
Auburn University
Auburn, Alabama
boliu@auburn.edu

Wen Dong
University at Buffalo
Buffalo, New York
wendong@buffalo.edu

ABSTRACT

Complex social systems are composed of interconnected individuals whose interactions result in group behaviors. Optimal control of a real-world complex system has many applications, including road traffic management, epidemic prevention, and information dissemination. However, such real-world complex system control is difficult to achieve because of high-dimensional and non-linear system dynamics, and the exploding state and action spaces for the decision maker. Prior methods can be divided into two categories: simulation-based and analytical approaches. Existing simulation approaches have high-variance in Monte Carlo integration, and the analytical approaches suffer from modeling inaccuracy. We adopted simulation modeling in specifying the complex dynamics of a complex system, and developed analytical solutions for searching optimal strategies in a complex network with high-dimensional state-action space. To capture the complex system dynamics, we formulate the complex social network decision making problem as a discrete event decision process. To address the curse of dimensionality and search in high-dimensional state action spaces in complex systems, we reduce control of a complex system to variational inference and parameter learning, introduce Bethe entropy approximation, and develop an expectation propagation algorithm. Our proposed algorithm leads to higher system expected rewards, faster convergence, and lower variance of value function in a real-world transportation scenario than state-of-the-art analytical and sampling approaches.

ACM Reference Format:
Fan Yang, Bo Liu and Wen Dong. 2019. Optimal Control of Complex Systems through Variational Inference with a Discrete Event Decision Process. *In Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), May 13-17, 2019, Montréal, Canada.* IFAAMAS, 9 pages.

1 INTRODUCTION

A complex social system is a collective system composed of a large number of interconnected entities that as whole exhibit properties and behaviors resulted from the interaction of its individual parts [27]. Achieving optimal control of a real-world complex social system is important and valuable. For example, in an urban transportation system, a central authority wants to reduce the overall delay and the total travel time by controlling traffic signals using traffic data collected from the Internet of Things [2]. In an epidemic system, an optimal strategy is pursued to minimize the expected discounted losses resulting from the epidemic process over an infinite horizon, with specific aims such as varying the birth and death rates [18] or early detection of epidemic outbreaks [13]. In the sphere of public opinion, components such as information consumers, news media, social websites, and governments aim to maximize their own utility functions with optimal strategies. A specific objective in the public opinion environment is tackling fake news to create a clearer, more trusted information environment [1]. These scenarios exemplify the enormous potential applications of a versatile optimal control framework for general complex systems.

However, key characteristics of complex systems make establishing such a framework very challenging. The system typically has a large number of interacting units and thus a high-dimensional state space. State transition dynamics in complex systems are already non-linear, time-variant, and high-dimensional, and then these frameworks must account for additional control variables. Prior research in decision making for complex social systems has generally gone in one of two directions: a simulation or analytical approach [21]. Simulation approaches specify system dynamics through a simulation model and develop sampling-based algorithms to reproduce the dynamic flow [29, 41, 46, 47]. These approaches can capture the microscopic dynamics of a complex system with high fidelity, but have a high variance and are time-consuming [20]. Analytical approaches instead formulate the decision-making problem as a constrained optimization problem with an analytical model through specifying the macroscopic state transitions directly, deriving analytical solutions for optimizing strategies [6, 36]. These approaches can provide a robust solution with less variance, but are applicable only to scenarios with small state spaces [16], or cases with low resolution intervention [9], due to modeling costs and errors [20]. The above research points to a new direction in research opportunities that combines the ability to capture system dynamics more precisely from simulation approaches with the benefit of having less variance and being more robust from analytical approaches. In this paper, we adopted simulation modeling in specifying the dynamics of a complex system, and developed analytical solutions for searching optimal strategies in a complex network with high-dimensional state-action space specified through simulation modeling.

We formulate the problem of decision making in a complex system as a discrete event decision process (DEDP), which identifies the decision making process as a Markov decision process (MDP) and introduces a discrete event model — a kind of simulation model [17] — to specify the system dynamics. A discrete event model defines a Markov jump process with probability measure on a sequence of elementary events that specify how the system

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

components interact and change the system states. These elementary events individually effect only minimal changes to the system, but in sequence together are powerful enough to induce non-linear, time-variant, high-dimensional behavior. Comparing with an MDP which specifies the system transition dynamics of a complex system analytically through a Markov model, a DEDP describes the dynamics more accurately through a discrete event model that captures the dynamics using a simulation process over the microscopic component-interaction events. We will demonstrate this merit through benchmarking with an analytical approach based on a MDP in the domain of transportation optimal control.

To solve a DEDP analytically, we derived a duality theorem that recasts optimal control to variational inference and parameter learning, which is an extension of the current equivalence results between optimal control and probabilistic inference [24, 38] in Markov decision process research. With this duality, we can include a number of existing probabilistic-inference and parameter-learning techniques, and integrate signal processing and decision making into a holistic framework. When exact inference becomes intractable, which is often the case in complex systems due to the formidable state space, our duality theorem implies the possibility of introducing recent approximate inference techniques to infer complex dynamics. The method in this paper is an expectation propagation algorithm, part of a family of approximate inference algorithms with local marginal projection. We will demonstrate that our approach is more robust and has less variance in comparison with other simulation approaches in the domain of transportation optimal control.

This research makes several important contributions. First, we formulate a DEDP — a general framework for modeling complex system decision-making problems — by combining MDP and simulation modeling. Second, we reduce the problem of optimal control to variational inference and parameter learning, and develop an approximate solver to find optimal control in complex systems through Bethe entropy approximation and an expectation propagation algorithm. Finally, we demonstrate that our proposed algorithm can achieve higher system expected rewards, faster convergence, and lower variance of value function within a real-world transportation scenario than even state-of-the-art analytical and sampling approaches.

2 BACKGROUND

In this section, we review the complex social system, the discrete event model, the Markov decision process, and the variational inference framework for a probabilistic graphical model.

2.1 Complex Social System

A complex social system is a collective system composed of a large number of interconnected entities that as whole exhibit properties and behaviors resulted from the interaction of its individual parts. A complex system tends to have four attributes: Diversity, interactivity, interdependency, and adaptivity [27]. Diversity means the system contains a large number of entities with various attributes and characteristics. Interactivity means the diverse entities interact with each other in an interaction structure, such as a fixed network or an ephemeral contact. Interdependency means the state change of one entity is dependent on others through the interactions. Adaptivity means the entities can adapt to different environments automatically.

In this paper, we temporarily exclude the attribute of adaptivity, the study of which will be future work. We focus on studying the optimal control of a complex social system with the attributes of diversity, interactivity, and interdependency, which leads to a system containing a large number of diverse components, the interactions of which lead to the states change. Examples of complex social systems include the transportation system where the traffic congestions are formed and dissipated through the interaction and movement of individual vehicles, the epidemic system where the disease is spread through the interaction of different people, and the public opinion system where people's minds are influenced and shaped by the dissemination of news through social media.

2.2 Discrete Event Model

A discrete event model defines a discrete event process, also called a Markov jump process. It is used to specify complex system dynamics with a sequence of stochastic events that each changes the state only minimally, but when combined in a sequence induce complex system evolutions. Specifically, a discrete event model describes the temporal evolution of a system with M species $\mathcal{X} = \{X^{(1)}, X^{(2)}, \cdots, X^{(M)}\}$ driven by V mutually independent events parameterized by rate coefficients $\mathbf{c} = (c_1, \ldots, c_V)$. At any specific time t, the populations of the species are $x_t = (x_t^{(1)}, \ldots, x_t^{(M)})$.

A discrete event process initially in state x_0 at time $t = 0$ can be simulated by: (1) Sampling the event $v \in \{\emptyset, 1, \cdots, V\}$ according to categorical distribution $v \sim (1 - h_0, h_1, \ldots, h_V)$, where $h_v(x, c_v) = c_v \prod_{m=1}^{M} g_v^{(m)}(x_t^{(m)})$ is the rate of event v, which equals to the rate coefficients c_v times a total of $\prod_{m=1}^{M} g_v^{(m)}(x^{(m)})$ different ways for the individuals to react, and $h_0(x, c) = \sum_{v=1}^{V} h_v(x, c_v)$ the rate of all events. The formulation of $h_v(x, c_v)$ comes from the formulations of the stochastic kinetic model and stochastic petri net [43, 44]. (2) Updating the network state deterministically $x \leftarrow x + \Delta_v$, where Δ_v represents how an event v changes the system states, until the termination condition is satisfied. In a social system, each event involves only a few state and action variables. This generative process thus assigns a probabilistic measure to a sample path induced by a sequence of events $v_0, \ldots, v_T$ happening between times $0, 1, \cdots, T$, where δ is an indicator function.

$$P(x_{0:T}, v_{0:T}) = p(x_0) \prod_{t=0}^{T} p(v_t|x_t) \delta_{x_{t+1} = x_t + \Delta_{v_t}},$$

$$\text{where } p(v_t|x_t) = \begin{cases} 1 - h_0(x_t, c), & v_t = \emptyset \\ h_k(x_t, c_k), & v_t = k, \end{cases}$$

The discrete event model is widely used by social scientists to specify social system dynamics [3] where the system state transitions are induced by interactions of individual components. Recent research [5, 26, 45] has also applied the model to infer the hidden state of social systems, but this approach has not been explored in social network intervention and decision making.

2.3 Markov Decision Process

A Markov decision process [33] is a framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. Formally, an MDP is defined as a tuple MDP$\langle S, A, P, R, \gamma \rangle$, where S represents the state space and $s_t \in S$ the state at time t, A the action space and a_t the action taken at time t, P the transition kernel of states such as $P(s_{t+1}|s_t, a_t)$, R the reward function such as $R(s_t, a_t)$ [6] or $R(s_t)$ [42] that evaluates the immediate reward at each step, and $\gamma \in [0, 1)$ the discount factor. Let us further define a policy π as a mapping from a state s_t to an action $a_t = \mu(s_t)$ or a distribution of it parameterized by θ — that is, $\pi = p(a_t|s_t; \theta)$. The probability measure of a length $-$ T MDP trajectory is $p(\xi_T) = p(s_0) \prod_{t=0}^{T-1} p(a_t|s_t; \theta) p(s_{t+1}|s_t, a_t)$, where $\xi_T = (s_{0:T}, a_{0:T})$. Solving an MDP involves finding the optimal policy π or its associated parameter θ to maximize the expected future reward — $\arg\max_\theta \mathbb{E}_\xi(\sum_t \gamma^t R_t; \theta)$.

The graphical representation of an MDP is shown in Figure 1, where we assume that the full system state s_t can be represented as a collection of component state variables $s_t = (s_t^{(1)}, ..., s_t^{(M)})$, so that the state space S is a Cartesian product of the domains of component state $s_t^{(m)}$: $S = S^{(1)} \times S^{(2)} \times \cdots \times S^{(M)}$. Similarly, the action variable a_t can be represented as a collection of action variables $a_t = (a_t^{(1)}, ..., a_t^{(D)})$, and the action space $A = A^{(1)} \times A^{(2)} \times \cdots \times A^{(D)}$. Here M is not necessarily equal to D because s_t represents the state of each component of the system while a_t represents the decisions taken by the system as a whole. For example, in the problem of optimizing the traffic signals in a transportation system where s_t represents the locations of each vehicle and a_t represents the status of each traffic light, the number of vehicles M may not necessarily equal to the number of traffic lights D. Usually in complex social systems, the number of individual components M is much greater than the system decision points D.

Prior research in solving a Markov decision process for a complex social system could be generally categorized into simulation or analytical approaches. A simulation approach reproduces the dynamic flow through sampling-based method. It describes the state transition dynamics with a high-fidelity simulation tool such as MATSIM [11], which simulates the microscopic interactions of the components and how these interactions leads to macroscopic state changes. Given current state s_t and action a_t at time t, a simulation approach uses a simulation tool to generate the next state s_{t+1}.

An analytical approach develops analytical solutions to solve a constrained optimization problem. Instead of describing the dynamics with a simulation tool, an analytical approach specifies the transition kernel analytically with probability density functions that describe the macroscopic state changes directly. Given current state s_t and action a_t, it computes the probability distribution of the next state s_{t+1} according to the state transition kernel $p(s_{t+1} \mid s_t, a_t)$. However, approximations are required to make the computation tractable. For an MDP containing M binary state variables and D binary action variables, the state space is 2^M, the action space is 2^D, the policy kernel is a $2^M \times 2^D$ matrix, and the state transition kernel (fixed action) is a $2^M \times 2^M$ matrix. Since M is usually much larger than D in complex social systems, the complexity bottleneck is usually the transition kernel with size $2^M \times 2^M$, the complexity of which grows exponentially with the number of state variables. Certain factorizations and approximations must be applied to lower the dimensionality of the transition kernel.

Usually analytical approaches solve complex social system MDPs approximately by enforcing certain independence constraints [31]. For example, Cheng [4] assumed that a state variable is only dependent on its neighboring variables. Sabbadin, Peyrard and Sabbadin [28, 30] exploited a mean field approximation to compute and update the local policies. Weiwei approximated the state transition kernel with differential equations [22]. These assumptions introduces additional approximations that results in modeling errors. In the next section, we propose a discrete event decision process which reduces the complexity of the transition probabilities, and which does not introduce additional independence assumptions.

Two specific approaches of solving an MDP are optimal control [32] and reinforcement learning [34]. Optimal control problems consist of finding the optimal decision sequence or the time-variant state-action mapping that maximizes the expected future reward, given the dynamics and reward function. Reinforcement-learning problems target the optimal stationary policy that maximizes the expected future reward while not assuming knowledge of the dynamics or the reward function. In this paper, we address the problem of optimizing a stationary policy to maximize the expected future reward, assuming known dynamics and reward function.

2.4 Variational Inference

A challenge in evaluating and improving a policy in a complex system is that the state space grows exponentially with the number of state variables, which makes probabilistic inference and parameter learning intractable. For example, in a system with M binary components, the size of state space S will be 2^M, let alone the exploding transition kernel. One way to resolve this issue is applying variation inference to optimize a tractable lower bound of the log expected future reward through conjugate duality. Variational inference is a classical framework in the probabilistic graphical model community [40]. It exploits the conjugate duality between log-partition function and the entropy function for exponential family distributions. Specifically, it solves the variational problem $\log \int \exp \langle \theta, \phi(x) \rangle \, dx = \sup_{q(x)} \left\{ \int q(x) \langle \theta, \phi(x) \rangle \, dx + H(q) \right\}$, where θ and $\phi(x)$ are respectively canonical parameters and sufficient statistics of an exponential family distribution, q is an auxiliary distribution and $H(q) = -\int dx q(x) \log q(x)$ is the entropy of q. For a tree-structured graphical model (here we use the simplified notation of a series of x_t), the Markov property admits a factorization of q into the product and division of local marginals $q(x) = \prod_t q_t(x_{t-1,t}) / \prod_t q_t(x_t)$. Substituting the factored forms of q into the variational target, we get an equivalent constrained optimization problem involving local marginals and consistency constraints among those marginals, and a fixed-point algorithm involving forward statistics α_t and backward statistics β_t. There are two primary forms of approximation of the original variational problem: the Bethe entropy problem and a structured mean field. The Bethe entropy problem is typically solved by loopy belief propagation or an expectation propagation algorithm.

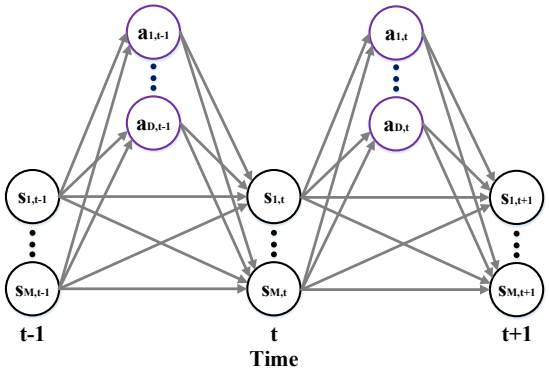

Figure 1: Graph representation of an MDP

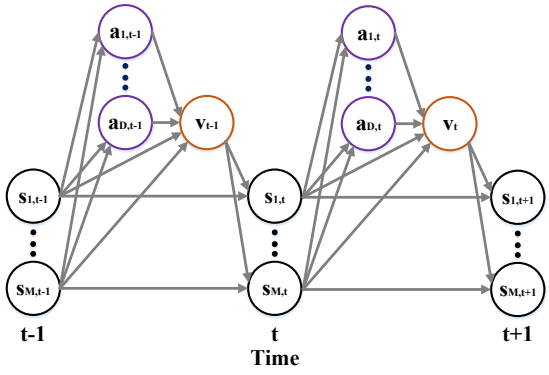

Figure 2: Graph representation of a DEDP

3 METHODOLOGY

In this section, we develop a DEDP and present a duality theorem that extends the equivalence of optimal control with probabilistic inference and parameter learning. We also develop an expectation propagation algorithm as an approximate solver to be applied in real-world complex systems.

3.1 Discrete Event Decision Process

The primary challenges in modeling a real-world complex system are the exploding state space and complex system dynamics. Our solution is to model the complex system decision-making process as a DEDP.

The graphical representation of a DEDP is shown in Figure 2. Formally, a DEDP is defined as a tuple DEDP$\langle S, A, \mathcal{V}, C, P, R, \gamma\rangle$, where S is the state space and $s_t = \left(s_t^{(1)}, ..., s_t^{(M)}\right) \in S$ a M-dimensional vector representing the state of each component at time t, A is the action space and $a_t = \left(a_t^{(1)}, ..., a_t^{(D)}\right) \in A$ a D-dimensional vector representing the action taken by the system at time t. As before, M is not necessarily equal to D, and M is usually much larger than D in complex social systems. Both $s_t^{(m)}$ and $a_t^{(d)}$ could take real or categorical values depending on the applications.

$\mathcal{V} = \{\emptyset, 1, ..., V\}$ is the set of events and $v_t \in \mathcal{V}$ a scalar following categorical distributions indicating the event taken at time t and changing the state by Δ_{v_t}. C is the function mapping actions to event rate coefficients which takes a D-dimensional vector as input, and outputs a V-dimensional vector $\boldsymbol{c} = (c_1, ..., c_V) = C(a_t)$, and P is the transition kernel of states induced by events $P(s_{t+1}, v_t|s_t, a_t) = p(v_t \mid s_t, a_t)\delta_{s_{t+1}=s_t+\Delta v_t}$, where $p(v_t = v \mid s_t, a_t)$ represents the probability of an event v happened at time t:

$$p(v_t = v \mid s_t, a_t) = \begin{cases} h_v(s_t, c_v) & \text{if } v \neq \emptyset \\ 1 - \sum_{v=1}^{V} h_v(s_t, c_v) & \text{if } v = \emptyset \end{cases}$$

Following the definitions in the discrete event model, $h_v(s_t, c_v) = c_v \cdot \prod_{m=1}^{M} g_v^{(m)}(s_t^{(m)})$ is the probability for event v to happen, which equals to the rate coefficients c_v times a total of $\prod_{m=1}^{M} g_v^{(m)}(s_t^{(m)})$ ways that the components can interact to trigger an event.

The immediate reward function R is a function of system states, defined as the summation of reward evaluated at each component $R(s_t) = \sum_{m=1}^{M} R_t^{(m)}(s_t^{(m)})$, and $\gamma \in [0, 1)$ is the discount factor.

We further define a policy π as a mapping from a state s_t to an action $a_t = \mu(s_t; \theta)$ or a distribution of it parameterized by θ — that is, $\pi = p(a_t|s_t; \theta)$. The parameterized policy can take any form, such as a lookup table where θ represents values of each entries in the table, a Gaussian distribution where θ represents the mean and variance, or a neural network where θ represents the network weights. Solving a DEDP involves finding the optimal policy π or its associated parameter θ to maximize the expected future reward — $\arg\max_\theta \mathbb{E}_\xi(\sum_t \gamma^t R_t; \theta)$. The probability measure of a length$-T$ DEDP trajectory with a stochastic policy is as follows, where δ is an indicator function and $\xi_T = (s_{0:T}, a_{0:T}, v_{0:T})$:

$$p(\xi_T) = p(s_0) \prod_{t=0}^{T-1} \left(p(a_t \mid s_t) p(v_t \mid s_t, a_t) \delta_{s_{t+1}=s_t+\Delta_{v_t}}\right)$$

The probability measure of that with a deterministic policy is this:

$$\begin{aligned} p(\xi_T) &= p(s_0) \prod_{t=0}^{T-1} \left(\delta_{a_t=\mu(s_t)} p(v_t \mid s_t, a_t) \delta_{s_{t+1}=s_t+\Delta_{v_t}}\right) \\ &= p(s_0) \prod_{t=0}^{T-1} \left(p(v_t \mid s_t, \mu(s_t)) \delta_{s_{t+1}=s_t+\Delta_{v_t}}\right) \\ &:= p(s_0) \prod_{t=0}^{T-1} \left(p(v_t \mid s_t) \delta_{s_{t+1}=s_t+\Delta_{v_t}}\right) \end{aligned}$$

A DEDP makes a tractable representation of complex system control problem by representing the non-linear and high-dimensional state transition kernel with microscopic events. A vanilla MDP is an intractable representation because the state-action space grows exponentially with the number of state-action variables. In comparison, the description length of a DEDP grows linearly in the number of events. As such, a DEDP greatly reduces the complexity of specifying a complex system control problem through introducing an auxiliary variable (event), and can potentially describe complex and high-fidelity dynamics of a complex system.

A DEDP could be reduced to an MDP if marginalizing out the events

$$\begin{aligned} &\sum_{v_{0:T}} p(s_{0:T}, a_{0:T}, v_{0:T}) \\ &= \sum_{v_{0:T}} p(s_0) \prod_{t=0}^{T-1} \left(p(a_t \mid s_t) p(v_t \mid s_t, a_t) \delta_{s_{t+1}=s_t+\Delta_{v_t}}\right) \\ &= p(s_0) \prod_{t=0}^{T-1} \left(p(a_t \mid s_t) \sum_{v_t} p(s_{t+1}, v_t \mid s_t, a_t)\right) \\ &= p(s_0) \prod_{t=0}^{T-1} \left(p(a_t \mid s_t) p(s_{t+1} \mid s_t, a_t)\right) \end{aligned}$$

Thus, the only difference between a DEDP and an MDP is that a DEDP introduces events to describe the state transition dynamics. Compared with the aforementioned models introducing independence constraints to make MDPs tractable, a DEDP does not make

any independence assumptions. It defines a simulation process that describes how components interact and trigger an event, and how the aggregation of events leads to system state changes. In this way, a DEDP captures the microscopic dynamics in a macroscopic system, leading to more accurate dynamics modeling.

3.2 Duality Theorem on Value Function

To solve a DEDP with a high-dimensional state and action space, we derive the convex conjugate duality between the log expected future reward function and the entropy function of a distribution over finite-length DEDP trajectories, and the corresponding duality in the parameter space between the log discounted trajectory-weighted reward and the distribution of finite-length DEDP trajectories. As a result, we can reduce the policy evaluation problem to a variational inference problem that involves the entropy function and can be solved by various probabilistic inference techniques.

Specifically, in a complex system decision process DEDP$\langle S, A, \mathcal{V}, C, P, R, \gamma\rangle$, let T be a discrete time, m the component index, ξ_T the length-T trajectory of a DEDP starting from initial state s_0, and $V^\pi = E\left(\sum_{t=0}^{\infty}\gamma^t R_t;\pi\right)$ the expected future reward (value function), where $\gamma\in[0,1)$ is a discount factor. Define $q(T,m,\xi_T)$ as a proposal joint probability distribution over finite length$-T$ DEDP trajectories, $r(T,m,\xi_T;\pi)=\gamma^T P(\xi_T;\pi)R_T^{(m)}(s_T^{(m)})$ as the discounted trajectory-weighted reward component where $P(\xi_T;\pi)$ is the probability distribution of a trajectory with policy π, and $R_T^{(m)}(s_T^{(m)})$ as the reward evaluated at component m at time T with state $s_T^{(m)}$. We thus have the following duality theorem.

***Theorem* 1.**

$$\log V^\pi(r)=\sup_q\left(\sum_{T,m,\xi_T} q(T,\,m,\,\xi_T)\log\, r(T,\,m,\,\xi_T;\pi)+\mathrm{H}(q)\right)$$

In the above, equality is satisfied when $\sum_T\sum_m\sum_{\xi_T} r(T,m,\xi_T;\pi)<\infty$ and $q(T,m,\xi_T)\propto\gamma^T P(\xi_T)R_T^{(m)}(s_T^{(m)})$. As such, $\log V^\pi(r)$ is the convex conjugate of $\mathrm{H}(q(T,m,\xi_T))$. $\mathrm{H}(q(T,m,\xi_T))$ is a convex function of $q(T,m,\xi_T)$, so by property of the convex conjugate, $\mathrm{H}(q(T,m,\xi_T))$ is also a conjugate of $\log V^\pi(r)$. The proof for this theorem is shown in the Appendix.

Theorem 1 provides a tight lower bound of the log expected reward and, more importantly, defines a variational problem in terms analogous to well-known variational inference formulations in the graphic model community [40], where a number of variational inference methods can be introduced such as exact inference methods, sampling-based approximate solutions, and variational inference. Theorem 1 extends the equivalence between optimal control and probability inference in recent literatures to a general variational functional problem. Specifically, it gets rid of the probability likelihood interpretation of the value function in [38, 39], and the prior assumption that value function is in the multiplication form of local-scope value functions [24].

3.3 Expectation Propagation for Optimal Control

Theorem 1 implies a generalized policy-iteration paradigm around the duality form: solving the variational problem as policy evaluation and optimizing the target over parameter θ with a known mixture of finite-length trajectories as policy improvement. In the following, we develop the policy evaluation and improvement algorithm with a deterministic policy $a_t=\mu(s_t;\theta)$, the derivation for which is given in the Appendix. The stochastic policy case $\pi=p(a_t|s_t;\theta)$ will lead to a similar result, which is not presented here due to the limit of space.

In policy evaluation, the Markov property of a DEDP admits factorizations $P(\xi_T;\pi)=p(s_0)\prod_{t=1}^{T}\left(p(v_t\mid s_t;\theta)\delta_{s_{t+1}=s_t+\Delta_{v_t}}\right)$ and $q(T,m,\xi_T)=q(T,m)q(s_0)\prod_{t=1}^{T}q(s_{t-1,t},v_{t-1}|T)/\prod_{t=1}^{T-1}q(s_t|T)$, where $q(T,m)=\gamma^T(1-\gamma)/M$ is the length prior distribution to match the discount factor, and $q(s_t|T)$ and $q(s_{t-1,t},v_{t-1}|T)$ are locally consistent one-slice and two-slice marginals. To cope with the exploding state space, we apply the Bethe entropy approximation. Specifically, we relax the formidable searching state space of s_t into an amenable space through the mean field approximation $q(s_t|T,m)=\prod_{\hat{m}=1}^{M}q(s_t^{(\hat{m})}|T,m)$, where $q(s_t^{(\hat{m})}|T,m)$ is the one-slice marginal involving only component $\hat{m}$. Applying the factorization and approximation, we get the following Bethe entropy problem (let $\sum_{s_{t-1,t},v_{t-1}\backslash s_t^{(\hat{m})}}$ represent the summation over all value combinations of s_{t-1},s_t,v_{t-1} except a fixed $s_t^{(\hat{m})}$).

$$\begin{aligned}&\max \text{ over } q(s_t^{(\hat{m})}|T,\,m),\, q(s_{t-1,t},\,v_t|T,\,m)\;\forall T,\,t\le T,\,m\\&\sum_{T,m}\sum_{t=1}^{T-1}\sum_{\hat{m}}\sum_{s_t^{(\hat{m})}}\sum_{T,m} q(T,\,m,\,s_t^{(\hat{m})})\log q(s_t^{(\hat{m})}|T,\,m)\\&-\textstyle\sum_{T,m}\sum_{t=1}^{T-1}\sum_{s_{t-1,t},v_{t-1}} q(T,\,m,\,s_{t-1,t},\,v_{t-1})\log\left(\frac{q(s_{t-1,t},v_{t-1}|T,m)}{p(s_t,v_{t-1}|s_{t-1};\theta)}\right)\\&-\textstyle\sum_{T,m,s_{T-1,T},v_{T-1}} q(T,\,m,\,s_{T-1,T},\,v_{T-1})\log\left(\frac{q(s_{T-1,T},v_{T-1}|T,m)}{p(s_T,v_{T-1}|s_{T-1};\theta)R_T^{(m)}}\right)\end{aligned}\tag{1}$$

$$\text{subject to: }\begin{array}{ll}\sum_{s_{t-1,t},v_{t-1}\backslash s_{t-1}^{(\hat{m})}} q(s_{t-1,t},\,v_{t-1}|T,\,m) &= q(s_{t-1}^{(\hat{m})}|T,\,m),\\ \sum_{s_{t-1,t},v_{t-1}\backslash s_t^{(\hat{m})}} q(s_{t-1,t},\,v_{t-1}|T,\,m) &= q(s_t^{(\hat{m})}|T,\,m)\end{array}$$

We solve this with the method of Lagrange multipliers, which leads to a forward-backward algorithm that updates the forward messages $\alpha_{t|T,m}^{(\hat{m})}(s_t^{(\hat{m})})$ and backward messages $\beta_{t|T,m}^{(\hat{m})}(s_t^{(\hat{m})})$ marginally according to the average effects of all other components — that is, a projected marginal kernel $p(s_t^{(\hat{m})},a_t|s_{t-1}^{(\hat{m})};\theta)$. Therefore, the algorithm achieves linear complexity over the number of components for each T and m, and quadratic time complexity over time horizon H to compute all messages for all $T\le H$. To further lower down the time complexity, we define $\beta_t^{(\hat{m})}(s_t^{(\hat{m})})=\sum_m\sum_{T=t}^{\infty}q(T,m)\beta_{t|T,m}^{(\hat{m})}(s_t^{(\hat{m})})$ by gathering together backward messages sharing t and $\alpha_t^{(\hat{m})}(s_t^{(\hat{m})})=\alpha_{t|T,m}^{(\hat{m})}(s_t^{(\hat{m})})$ by noting that $\alpha_{t|T,m}^{(\hat{m})}(s_t^{(\hat{m})})$ doesn't depend on T,m. This leads to the following forward-backward algorithm, which is linear in time horizon:

$$\alpha_t^{(\hat{m})}(s_t^{(\hat{m})})\propto\sum_{s_{t-1}^{(\hat{m})},v_{t-1}}\alpha_{t-1}^{(\hat{m})}(s_{t-1}^{(\hat{m})})\cdot p(s_t^{(\hat{m})},v_{t-1}|s_{t-1}^{(\hat{m})};\theta)\tag{2}$$

$$\begin{aligned}\beta_t^{(\hat{m})}(s_t^{(\hat{m})})=&\sum_m q(t,m)\beta_{t|t,m}^{(\hat{m})}(s_t^{(\hat{m})})\\&+\sum_{s_{t+1}^{(\hat{m})},v_t}p(s_{t+1}^{(\hat{m})},v_t|s_t^{(\hat{m})};\theta)\beta_{t+1}^{(\hat{m})}(s_{t+1}^{(\hat{m})})\end{aligned}\tag{3}$$

In policy improvement, we maximize the log expected future reward function $L(\theta) = \sum_{T,m,\xi_T} q(T, m, \xi_T; \theta^{\text{old}}) \log\left(\gamma^T P(\xi_T;\theta) R_T^{(m)}\right)$ over parameter θ with $q(T, m, \xi_T)$ inferred from θ^{old} via gradient ascent update $\theta^{\text{new}} = \theta^{\text{old}} + \epsilon \cdot \frac{\partial L}{\partial \theta}\Big|_{\theta^{\text{old}}}$, or more aggressively by setting θ^{new} so that $\frac{\partial L}{\partial \theta}|_{\theta^{\text{new}}} = 0$. This objective can be simplified by dropping irrelevant terms and keeping only those involving policy: $L(\theta) = \sum_{T,m,t,v_t,s_t} q(T, m, v_t, s_t; \theta^{\text{old}}) \cdot \log P(v_t|s_t;\theta) + \text{const}$. The gradient is obtained from chain rule and messages $\alpha_t(x_t)$, $\beta_t(x_t)$ through dynamic programming:

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{t,s_t} \frac{\prod_{\dot{m}} \alpha_t^{(\dot{m})}(s_t^{(\dot{m})}, v_t = v) \beta_t^{(\dot{m})}(s_t^{(\dot{m})}, v_t = v)}{c_v} \frac{\partial c_v}{\partial \theta}$$
$$- \sum_{t,s_t} \frac{\prod_{\dot{m}} \alpha_t^{(\dot{m})}(s_t^{(\dot{m})}, v_t = \emptyset) \beta_t^{(\dot{m})}(s_t^{(\dot{m})}, v_t = \emptyset) \cdot \prod_m g_v^m(s_t^{(m)})}{1 - \sum_{v=1}^{V} c_v \cdot \prod_m g_v^m(s_t^{(m)})} \frac{\partial c_v}{\partial \theta} \quad (4)$$

In summary, we give our optimal control algorithm of complex systems as Algorithm 1.

Algorithm 1 Optimal control of social systems

Input: The DEDP tuple DEDP$\langle S, A, \mathcal{V}, C, P, R, \gamma\rangle$, initial policy parameter θ
Output: Optimal policy parameter θ
Procedure: Iterate until convergence:

- policy evaluation

Iterate until convergence:
- Update forward messages according to Eq. (2)
- Update backward messages according to Eq. (3)

- policy improvement

Update parameter θ according to the gradient in Eq. (4).

3.4 Discussions

In the above we developed a DEDP for modeling complex system decision making problems, and derived a expectation propagation algorithm for solving the DEDP. our algorithm is also applicable on a Markov decision process with other simulation models. While we used a discounted expected total future reward in the previous derivation, our framework is also applicable to other types of expected future reward, such as a finite horizon future reward, where we use a different probability distribution of time $q(T) = \frac{1}{T}$.

4 EXPERIMENTS

In this experiment, we benchmark algorithm 1 against several decision-making algorithms for finding the optimal policy in a complex system.

Overview: The complex social system in this example is a transportation system. The goal is to optimize the policy such that each vehicle arriving at the correct facilities at correct time (being at work during work hours and at home during rest hours) and spending minimum time on roads. We formulate the transportation optimal control problem as a discrete event decision process DEDP$\langle S, A, \mathcal{V}, C, P, R, \gamma\rangle$. The state variables $s_t = (s_t^{(1)}, \cdots, s_t^{(M-1)}, t)$ represent the populations at $(M-1)$ locations and the current time t. All events are of the form $p \cdot m_1 \overset{c_{m_1 m_2}}{\rightarrow} p \cdot m_2$—an individual p moving from location m_1 to location m_2 with rate (probability per unit time) $c_{m_1 m_2}$—decreasing the population at m_1 by one and increasing the population at m_2 by one. We also introduce auxiliary event $\emptyset$ that doesn't change any system state, and set the rates of leaving facilities and selecting alternative downstream links as action variables. We implement the state transition $p(s_{t+1}, v_t \mid s_t, a_t)$ following the fundamental diagram of traffic flow [12] that simulate the movement of vehicles. The reward function $R(s_t) = \sum_m \beta_{t,\text{perf}}^{(m)} s_t^{(m)} + \beta_{t,\text{trav}}^{(m)} s_t^{(m)}$ emulates the Charypa-Nagel scoring function in transportation research [12] to reward performing the correct activities at facilities and penalize traveling on roads , where $\beta_{t,\text{trav}}^{(m)}$ and $\beta_{t,\text{perf}}^{(m)}$ are the score coefficients. We implement the deterministic policy as a function of states through a neural network $a_t = \mu(s_t) = \mathcal{NN}(s_t;\theta)$ and apply Algorithm 1 to find the optimal policy parameter θ.

Benchmark Model Description: The transition kernel based on an MDP in this general form $p(s_{t+1} \mid s_t, a_t)$ in a complex system is too complicated to be modeled exactly with an analytical form, due to the high-dimensional state-action space, and the complex system dynamics. As such, we benchmark with the following algorithms: (1) Analytical approaches based on an MDP that uses Taylor expansion to approximate the intractable transition dynamics with differential equations, which leads to a guided policy search (GPS) algorithm [25] with a known dynamics that uses iterative linear quadratic regulator (iLQR) for trajectory optimization and supervised learning for training a global policy $\mu(s_t)$, implemented as a five-layer neural network. Other aforementioned approximations [4, 28, 30] are not applicable because in their settings each component takes an action, resulting in local policies for each component. Whereas in our problem the action is taken by the system as a whole, and therefore no local policies. (2) Simulation approaches that reproduce the dynamic flow through sampling the state action trajectories from the current policy and the system transition dynamics, which leads to a policy gradient (PG) algorithm, the policy of which is implemented as a four-layer neural network; and (3) an actor-critic (AC) algorithm [23] that implements the policy $\mu(s_t)$ as an actor network with four layers and the state-action value function $Q(s_t, a_t)$ as a critic network with five layers.

Performance and Efficiency: We benchmark the algorithms using the SynthTown scenario (Fig. 3), which has one home facility, one work facility, 23 road links, and 50 individuals going to work facility in the morning (9 am) and home facility in the evening (5 pm). A training episode is one day. This scenario is small enough for studying the details of different algorithms.

In Figure 4, the value-epoch curve of VI (our algorithm) dominates those of the other algorithms almost everywhere. Table 1 indicates that VI requires the fewest training epochs to converge to the highest total rewards per episode. Figure 5 presents the average vehicle distribution of ten runs at different locations (Home (h), Work (w), and roads 1-23) using the learned policy with each algorithm. This figure implies that the learned policy of VI leads to the largest amount of vehicles being at work during work hours (9 am to 5 pm), and least amount of time the vehicles spending on roads.

In the SynthTown scenario, high rewards require the individuals to perform the correct activities (home, work, and so on) at the right time, and to spend less time on roads. VI achieves the best performance when evaluating a policy by considering the whole

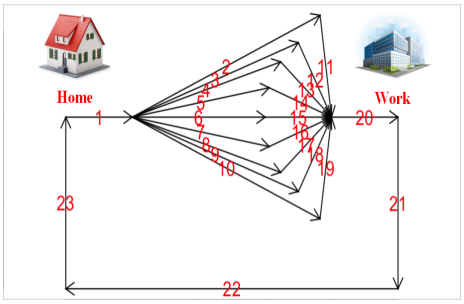

Figure 3: SynthTown road network

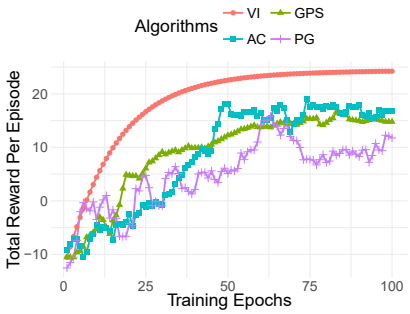

Figure 4: Training process on SynthTown

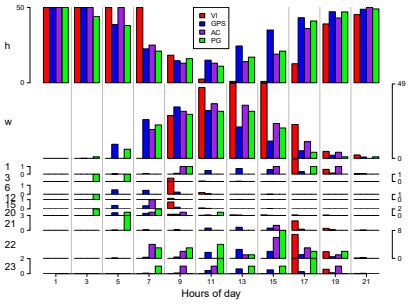

Figure 5: Average number of vehicles with trained policies

state-action space with VI approximation — the evolution of each state variables in the mean field of the other state variables. Modeling the complex system transition dynamics based on an MDP analytically with Taylor approximation will introduce modeling errors (GPS). Value estimation from Monte Carlo integration in high-dimensional state space has high variance (PG). A small perturbation of policy will result in significant change to the immediate reward and value in later steps, which makes it difficult to estimate the correct value from sampled trajectories, and as a result difficult to compute the correct value gradient (AC).

Dataset	SynthTown		Berlin	
Metrics	**TRPE**	**EC**	**TRPE**	**EC**
VI	24.21	75	11.52	200
GPS	14.79	100	-10.33	-
AC	16.76	100	-9.69	-
PG	11.77	150	-15.52	-

Table 1: Comparing algorithms in total reward per episode (TRPE) and epochs to converge (EC)

We also benchmark the performance of all algorithms using the Berlin scenario, which consists of a network of 1,530 locations and the trips of 9, 178 synthesized individuals [12]. Table 1 shows that VI outperformed in total rewards per episode, while the other algorithms did not even converge in a reasonable number of epochs.

In summary, VI outperformed guided policy search, policy gradient, and actor-critic algorithms in all scenarios. These benchmarking algorithms provide comparable results in a small dataset such as the SynthTown scenario, but became difficult to train when applied to a larger dataset such as Berlin.

5 RELATED WORKS

A number of prior works have explored the connection between decision making and probabilistic inference. The earliest such research is the Kalman duality proposed by Rudolf Kalman [14]. Subsequent works have expanded the duality in stochastic optimal control, MDP planning, and reinforcement learning. Todorov and Kappen expanded the connection between control and inference by restricting the immediate cost and identifying the stochastic optimal control problems as linearly-solvable MDPs [37] or KL control problems [15]. However, the passive dynamics in a linearly-solvable MDPs and the uncontrolled dynamics in a KL control problem become intractable in a social system scenario due to the complicated system dynamics. Toussaint, Hoffman, David, and colleagues broadened the connection by framing planning in an MDP as inference in a mixture of PGMs [38]. The exact [8] and variational inference [7] approaches in their framework encounter the transition dynamics intractability issue in a social network setting due to the complicated system dynamics. Their sampling-based alternatives [10] experience the high-variance problem in a social network due to the exploding state space. Levine, Ziebart, and colleagues widened the connection by establishing the equivalence between maximum entropy reinforcement learning, where the standard reward objective is augmented with an entropy term [48], and probabilistic inference in a PGM [19, 48]. They optimize a different objective function compared with our method. Our approach is an extension of Toussaint's formulation [38], but differs in that we establish the duality with a DEDP, and provide new insights into the duality from the perspective of convex conjugate duality. Compared with the existing exact inference [8] and approximate inference solutions [7, 10], our algorithm is more efficient with gathering the messages, and more scalable with the use of Bethe entropy approximation.

Other formulations of the decision-making problems which are similar to our framework are the Option network [35] and the multi-agent Markov decision process (MMDP) [31]. An option network extends a traditional Markov decision process with options — closed loop policies for taking actions over a period of time, with the goal of providing a temporal abstraction of the representation. In our framework, we introduce an auxiliary variable v with the goal of providing a more accurate and efficient modeling of the system dynamics. An MMDP represents sequential decision-making problems in cooperative multi-agent settings. There are two fundamental differences between our framework and MMDP. First, unlike MMDP where each agent takes an action at each time step, in a DEDP the actions are taken by the system and the dimensionality of states not necessarily equals to the dimensionality of actions. Second, unlike MMDP where each agent has its own information, a DEDP models the decision making of a large system where only one controller has all the information and makes decisions for the entire system.

6 CONCLUSIONS

In this paper, we have developed DEDP for modeling complex system decision-making processes. We reduce the optimal control of DEDP to variational inference and parameter learning around a variational duality theorem. Our framework is capable of optimally controlling real-world complex systems, as demonstrated by experiments in a transportation scenario.

7 APPENDIX

7.1 Derivation of Theorem 1

$$\begin{aligned}\log V^{\pi} &= \log\Big(\textstyle\sum_T \sum_m \sum_{\xi_T} q(T,m,\xi_T)\frac{r(T,m,\xi_T;\pi)}{q(T,m,\xi_T)}\Big)\\ &\geq \textstyle\sum_T \sum_m \sum_{\xi_T} q(T,m,\xi_T)\log\frac{r(T,m,\xi_T;\pi)}{q(T,m,\xi_T)}\\ &= \textstyle\sum_T \sum_m \sum_{\xi_T} q(T,m,\xi_T)\log r(T,m,\xi_T;\pi)+\mathrm{H}(q(T,m,\xi_T))\end{aligned}$$

7.2 Derivation of Eq. (1)

Rearranging the terms and applying the approximations described in the main text, the target becomes the following:

$$\begin{aligned}&\textstyle\sum_T \sum_m \sum_{\xi_T} q(T,m,\xi_T)\log\Big(\gamma^T P(\xi_T;\theta)R_T^{(m)}\Big)+\mathrm{H}(q(T,m,\xi_T))\\ &=\textstyle\sum_T \sum_m \sum_{\xi_T} q(T,m,\xi_T)\log\Big(\gamma^T \prod_{t=1}^{T-1} p(s_t,v_{t-1}|s_{t-1};\theta)\cdot p(s_T,v_{T-1}|s_{T-1};\theta)R_T^{(m)}\Big)\\ &-\textstyle\sum_T \sum_m \sum_{\xi_T} q(T,m,\xi_T)\log\left(q(T,m)\frac{\prod_{t=1}^{T} q(s_{t-1,t},v_{t-1,t}|T,m)}{\prod_{t=1}^{T-1} q(s_t|T,m)}\right)\\ &=\textstyle\sum_T \sum_m (q(T,m)\log\Big(\frac{\gamma^T}{q(T,m)}\Big)-\sum_{t=1}^{T-1}\sum_{s_{t-1,t},v_{t-1}} q(T,m,s_{t-1,t})\log\Big(\frac{q(s_{t-1,t},v_{t-1}|T,m)}{p(s_t,v_{t-1}|s_{t-1};\theta)}\Big)\\ &-\textstyle\sum_t \sum_m \sum_{s_{t-1,t},v_{t-1}} q(T,m,s_{t-1,t})\log\left(\frac{q(s_{t-1,t},v_{t-1}|T,m)}{p(s_t,v_{t-1}|s_{t-1};\theta)R_t^{(m)}}\right)\\ &+\textstyle\sum_{T,m}\sum_{t=1}^{T-1}\sum_{\hat{m}}\sum_{s_t^{(\hat{m})}} q(T,m,s_t^{(\hat{m})})\log q(s_t^{(\hat{m})}|T,m)\end{aligned}$$

7.3 Derivation to Solve Eq. (1)

We solve this maximization problem with the method of Lagrange multipliers.

$$\begin{aligned}L &= \textstyle\sum_{T,m} q(T,m)\log\Big(\frac{\gamma^T}{q(T,m)}\Big)\\ &-\textstyle\sum_{T,m}\sum_{t=1}^{T-1}\sum_{s_{t-1,t},v_{t-1}} q(T,m,s_{t-1,t},v_{t-1})\log\Big(\frac{q(s_{t-1,t},v_{t-1}|T,m)}{p(s_t,v_{t-1}|s_{t-1};\theta)}\Big)\\ &-\textstyle\sum_{T,m}\sum_{s_{T-1,T}} q(T,m,s_{T-1,T},v_{T-1})\log\left(\frac{q(s_{T-1,T},v_{T-1}|T,m)}{p(s_T,v_{T-1}|s_{T-1};\theta)R_T^{(m)}}\right)\\ &+\textstyle\sum_{T,m}\sum_{t=1}^{T-1}\sum_{\hat{m}}\sum_{s_t^{(\hat{m})}} q(T,m,s_t^{(\hat{m})})\log q(s_t^{(\hat{m})}|T,m)\\ &+\textstyle\sum_{t,\hat{m},s_{t-1}^{(\hat{m})},T,m}\alpha^{\hat{m}}_{t-1,s_{t-1}^{(\hat{m})},T,m}\Big(\sum_{s_{t-1,t},v_{t-1}\setminus s_{t-1}^{(\hat{m})}} q(s_{t-1,t},v_{t-1}|T,m)-q(s_{t-1}^{(\hat{m})}|T,m)\Big)\\ &+\textstyle\sum_{t,\hat{m},s_t^{(\hat{m})},T,m}\beta^{\hat{m}}_{t,s_t^{(\hat{m})},T,m}\Big(\sum_{s_{t-1,t},v_{t-1}\setminus s_t^{(\hat{m})}} q(s_{t-1,t},v_{t-1}|T,m)-q(s_t^{(\hat{m})}|T,m)\Big)\end{aligned}$$

Taking derivative with respect to $q(s_{t-1,t},a_t|T,m)$ and $q(s_t^{(\hat{m})}|T,m)$, and setting it to zero, we then get

for t=1,…T−1

$$q(s_{t-1,t},v_{t-1}|T,m) \propto \exp\left(\frac{\sum_{\hat{m}}\alpha^{(\hat{m})}_{t-1,s_{t-1}^{(\hat{m})},T,m}}{q(T,m)}\right)\cdot p(s_t,v_{t-1}|s_{t-1};\theta)\exp\left(\frac{\sum_{\hat{m}}\beta^{(\hat{m})}_{t,s_t^{(\hat{m})},T,m}}{q(T,m)}\right)$$

for t=T

$$q(s_{t-1,t},v_{t-1}|T,m) \propto \exp\left(\frac{\sum_{\hat{m}}\alpha^{(\hat{m})}_{t-1,s_{t-1}^{(\hat{m})},T,m}}{q(T,m)}\right)\cdot p(s_t,v_{t-1}|s_{t-1};\theta)R_T^{(m)}\exp\left(\frac{\sum_{\hat{m}}\beta^{(\hat{m})}_{t,s_t^{(\hat{m})},T,m}}{q(T,m)}\right)$$

$$q(s_t^{(\hat{m})}|T,m) = \frac{1}{Z_T^{(m)}}\exp\left(\frac{\alpha^{(\hat{m})}_{t,s_t^{(\hat{m})},T,m}+\beta^{(\hat{m})}_{t,s_t^{(\hat{m})},T,m}}{q(T,m)}\right)$$

Marginalizing over $q(s_{t-1,t}|T,m)$, we have

$$\begin{aligned}&q(s_{t-1,t}^{(\hat{m})},v_{t-1}|T,m)\\ &=\textstyle\sum_{s_{t-1,t}^{(\hat{m}')},\hat{m}'\neq\hat{m}}\frac{1}{Z_t}\exp\left(\frac{\sum_{\hat{m}}\alpha^{(\hat{m})}_{t-1,s_{t-1}^{(\hat{m})},T,m}}{q(T,m)}\right)\cdot p(s_t,v_{t-1}|s_{t-1};\theta)\exp\left(\frac{\sum_{\hat{m}}\beta^{(\hat{m})}_{t,s_t^{(\hat{m})},T,m}}{q(T,m)}\right)\\ &:=\frac{1}{Z_t}\exp\left(\frac{\alpha^{(\hat{m})}_{t-1,s_t^{(\hat{m})},T,m}+\beta^{(\hat{m})}_{t,s_t^{(\hat{m})},T,m}}{q(T,m)}\right)\cdot p(s_t^{(\hat{m})},v_{t-1}|s_{t-1}^{(\hat{m})};\theta)\end{aligned}$$

We denote $\exp\left(\frac{\alpha^{(\hat{m})}_{t,s_t^{(\hat{m})},T,m}}{q(T,m)}\right)$ as $\alpha^{(\hat{m})}_{t|T,m}\left(s_t^{(\hat{m})}\right)$, $\exp\left(\frac{\beta^{(\hat{m})}_{t,s_t^{(\hat{m})},T,m}}{q(T,m)}\right)$ as $\beta^{(\hat{m})}_{t|T,m}\left(s_t^{(\hat{m})}\right)$.

We can compute α, β through a forward-backward iterative approach:

$$\begin{aligned}&\text{forward:} \quad q(s_t^{(\hat{m})}|T,m)=\textstyle\sum_{s_{t-1}^{(\hat{m})},v_{t-1}} q(s_{t-1,t}^{(\hat{m})},v_{t-1}|T,m)\\ &\Rightarrow \quad \alpha^{(\hat{m})}_{t|T,m}(s_t^{(\hat{m})})=\frac{Z_t^{(\hat{m})}}{Z_t}\textstyle\sum_{s_{t-1}^{(\hat{m})},v_{t-1}}\alpha^{(\hat{m})}_{t-1|T,m}(s_{t-1}^{(\hat{m})})\cdot p(s_t^{(\hat{m})},v_{t-1}|s_{t-1}^{(\hat{m})};\theta)\\ &\text{backward:} \quad q(s_{t-1}^{(\hat{m})}|T,m)=\textstyle\sum_{s_t^{(\hat{m})},v_{t-1}} q(s_{t-1,t}^{(\hat{m})},v_{t-1}|T,m)\\ &\Rightarrow \quad \beta^{(\hat{m})}_{t-1|T,m}(s_{t-1}^{(\hat{m})})=\frac{Z_t^{(\hat{m})}}{Z_t}\textstyle\sum_{s_t^{(\hat{m})},v_{t-1}} p(s_t^{(\hat{m})},v_{t-1}|s_{t-1}^{(\hat{m})};\theta)\cdot\beta^{(\hat{m})}_{t|T,m}(s_t^{(\hat{m})})\\ &\qquad\text{for } t=T,\hat{m}=m\\ &\qquad \beta^{(\hat{m})}_{t-1|T,m}(s_{t-1}^{(\hat{m})})=\frac{Z_t^{(\hat{m})}}{Z_t}\textstyle\sum_{s_t^{(\hat{m})},v_{t-1}} p(s_t^{(\hat{m})},v_{t-1}|s_{t-1}^{(\hat{m})};\theta)\cdot\beta^{(\hat{m})}_{t|T,m}(s_t^{(\hat{m})})R_T^{(m)}\end{aligned}$$

7.4 Derivation of Eq. (3)

We can further simplify our algorithm by gathering the messages. For notational simplicity, we can absorb the reward $R_T^{(m)}(s_T^{(m)})$ into the $\beta^{(m)}_{T|T,m}(s_T^{(m)})$ term, so that $\beta^{(m)}_{T|T,m}(s_T^{(m)})=R_T^{(m)}(s_T^{(m)})$ and $\beta^{(\hat{m})}_{T|T,m}(s_T^{(\hat{m})})=1$ for $\hat{m}\neq m$. We then define $\beta_t^{(\hat{m})}(s_t^{(\hat{m})})=\sum_m\sum_{T=t}^{\infty} q(T,m)\beta^{(\hat{m})}_{t|T,m}(s_t^{(\hat{m})})$ and have

$$\begin{aligned}&\beta_t^{(\hat{m})}(s_t^{(\hat{m})})\\ &=\textstyle\sum_m\sum_{T=t}^{\infty} q(T,m)\beta^{(\hat{m})}_{t|T,m}(s_t^{(\hat{m})})\\ &=\textstyle\sum_m q(t,m)\beta^{(\hat{m})}_{t|t,m}(s_t^{(\hat{m})})+\sum_{s_{t+1}^{(\hat{m})},v_t} p(s_{t+1}^{(\hat{m})},v_t|s_t^{(\hat{m})};\theta)\beta^{(\hat{m})}_{t+1}(s_{t+1}^{(\hat{m})})\end{aligned}$$

Observing that $\sum_{T=1}^{\infty}\sum_{t=1}^{T} \Longleftrightarrow \sum_{t=1}^{\infty}\sum_{T=t}^{\infty}$, this summation form of messages will be useful in the parameter-learning phase.

7.5 Derivation of Eq. (4)

$$\begin{aligned}\frac{\partial L}{\partial\theta} &= \sum_{T,m,t}\sum_{s_t}\frac{q(T,m,v_t=v,s_t)}{c_v}\frac{\partial c_v}{\partial\theta}\\ &-\sum_{T,m,t}\sum_{s_t}\frac{q(T,m,v_t=\emptyset,s_t)\cdot g_v(s_t)}{1-\sum_{v=1}^{V} c_v g_v(s_t)}\frac{\partial c_v}{\partial\theta}\\ &=\sum_{t,s_t}\frac{\prod_{\hat{m}}\alpha_t^{(\hat{m})}(s_t^{(\hat{m})},v_t=v)\beta_t^{(\hat{m})}(s_t^{(\hat{m})},v_t=v)}{c_v}\frac{\partial c_v}{\partial\theta}\\ &-\sum_{t,s_t}\frac{\prod_{\hat{m}}\alpha_t^{(\hat{m})}(s_t^{(\hat{m})},v_t=\emptyset)\beta_t^{(\hat{m})}(s_t^{(\hat{m})},v_t=\emptyset)\cdot\prod_m g_v^m(s_t^{(m)})}{1-\sum_{v=1}^{V} c_v\cdot\prod_m g_v^m(s_t^{(m)})}\frac{\partial c_v}{\partial\theta}\end{aligned}$$

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.

[2] PG Balaji, X German, and D Srinivasan. 2010. Urban traffic signal control using reinforcement learning agents. *IET Intelligent Transport Systems* 4, 3 (2010), 177–188.

[3] Andrei Borshchev. 2013. *The big book of simulation modeling: multimethod modeling with AnyLogic 6.* AnyLogic North America Chicago.

[4] Qiang Cheng, Qiang Liu, Feng Chen, and Alexander T Ihler. 2013. Variational planning for graph-based MDPs. In *Advances in Neural Information Processing Systems*. 2976–2984.

[5] Le Fang, Fan Yang, Wen Dong, Tong Guan, and Chunming Qiao. 2017. Expectation Propagation with Stochastic Kinetic Model in Complex Interaction Systems. In *Advances in Neural Information Processing Systems*. 2026–2036.

[6] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. 2015. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*. 1954–1962.

[7] Thomas Furmston and David Barber. 2010. Variational methods for reinforcement learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 241–248.

[8] Thomas Furmston and David Barber. 2012. Efficient inference in markov control problems. *arXiv preprint arXiv:1202.3720* (2012).

[9] Miao He, Lei Zhao, and Warren B Powell. 2010. Optimal control of dosage decisions in controlled ovarian hyperstimulation. *Annals of Operations Research* 178, 1 (2010), 223–245.

[10] Matt Hoffman, Hendrik Kueck, Nando de Freitas, and Arnaud Doucet. 2009. New inference strategies for solving Markov decision processes using reversible jump MCMC. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 223–231.

[11] Andreas Horni, Kai Nagel, and Kay W Axhausen. 2016. The multi-agent transport simulation MATSim. *Ubiquity, London* 9 (2016).

[12] Andreas Horni, Kai Nagel, and Kay W Axhausen. 2016. The multi-agent transport simulation MATSim. *Ubiquity, London* 9 (2016).

[13] Masoumeh T Izadi and D Buckeridge. 2007. Optimizing anthrax outbreak detection methods using reinforcement learning. Citeseer.

[14] Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering* 82, 1 (1960), 35–45.

[15] Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. 2012. Optimal control as a graphical model inference problem. *Machine learning* 87, 2 (2012), 159–182.

[16] Lin Liao Dieter Fox Henry Kautz. 2004. Learning and inferring transportation routines. In *Proceedings: Nineteenth National Conference on Artificial Intelligence (AAAI-04): Sixteenth Innovative Applications of Artificial Intelligence Conference (IAAI-04)*. Aaai Press, 348.

[17] Averill M Law, W David Kelton, and W David Kelton. 1991. *Simulation modeling and analysis*. Vol. 2. McGraw-Hill New York.

[18] Claude Lefévre. 1981. Optimal control of a birth and death epidemic process. *Operations Research* 29, 5 (1981), 971–982.

[19] Sergey Levine. 2018. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. *arXiv preprint arXiv:1805.00909* (2018).

[20] Li Li, Yisheng Lv, and Fei-Yue Wang. 2016. Traffic signal timing via deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica* 3, 3 (2016), 247–254.

[21] Li Li, Ding Wen, and Danya Yao. 2014. A survey of traffic control with vehicular communications. *IEEE Transactions on Intelligent Transportation Systems* 15, 1 (2014), 425–432.

[22] Weiwei Li and Emanuel Todorov. 2004. Iterative linear quadratic regulator design for nonlinear biological movement systems.. In *ICINCO (1)*. 222–229.

[23] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[24] Qiang Liu and Alexander Ihler. 2012. Belief Propagation for Structured Decision Making. In *Proceedings of the Twenty-Eighth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-12)*. AUAI Press, Corvallis, Oregon, 523–532.

[25] William H Montgomery and Sergey Levine. 2016. Guided policy search via approximate mirror descent. In *Advances in Neural Information Processing Systems*. 4008–4016.

[26] Manfred Opper and Guido Sanguinetti. 2008. Variational inference for Markov jump processes. In *Advances in Neural Information Processing Systems*. 1105–1112.

[27] Scott E Page. 2015. What sociologists should know about complexity. *Annual Review of Sociology* 41 (2015), 21–41.

[28] Nathalie Peyrard and Régis Sabbadin. 2006. Mean field approximation of the policy iteration algorithm for graph-based Markov decision processes. *Frontiers in Artificial Intelligence and Applications* 141 (2006), 595.

[29] Victor M Preciado, Michael Zargham, Chinwendu Enyioha, Ali Jadbabaie, and George Pappas. 2013. Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*. IEEE, 7486–7491.

[30] Régis Sabbadin, Nathalie Peyrard, and Nicklas Forsell. 2012. A framework and a mean-field algorithm for the local control of spatial processes. *International Journal of Approximate Reasoning* 53, 1 (2012), 66–86.

[31] Olivier Sigaud and Olivier Buffet. 2013. *Markov decision processes in artificial intelligence*. John Wiley & Sons.

[32] Robert F Stengel. 1994. *Optimal control and estimation*. Courier Corporation.

[33] Richard S Sutton and Andrew G Barto. 2011. Reinforcement learning: An introduction. (2011).

[34] Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*. MIT press.

[35] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1-2 (1999), 181–211.

[36] Stelios Timotheou, Christos G Panayiotou, and Marios M Polycarpou. 2015. Distributed traffic signal control using the cell transmission model via the alternating direction method of multipliers. *IEEE Transactions on Intelligent Transportation Systems* 16, 2 (2015), 919–933.

[37] Emanuel Todorov. 2007. Linearly-solvable Markov decision problems. In *Advances in neural information processing systems*. 1369–1376.

[38] Marc Toussaint and Amos Storkey. 2006. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 945–952.

[39] Marc Toussaint, Amos Storkey, and Stefan Harmeling. 2010. Expectation-Maximization methods for solving (PO) MDPs and optimal control problems. *Inference and Learning in Dynamic Models. Cambridge University Press: Cambridge, England* (2010).

[40] Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1, 1–2 (2008), 1–305.

[41] Fei-Yue Wang. 2010. Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications. *IEEE Transactions on Intelligent Transportation Systems* 11, 3 (2010), 630–638.

[42] MA Wiering, J van Veenen, Jilles Vreeken, and Arne Koopman. 2004. Intelligent traffic light control. (2004).

[43] Darren J Wilkinson. 2011. *Stochastic modelling for systems biology*. CRC press.

[44] Zhen Xu, Wen Dong, and Sargur N Srihari. 2016. Using social dynamics to make individual predictions: variational inference with a stochastic kinetic model. In *Advances in Neural Information Processing Systems*. 2783–2791.

[45] Fan Yang and Wen Dong. 2017. Integrating simulation and signal processing with stochastic social kinetic model. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 193–203.

[46] Fan Yang and Wen Dong. 2018. Integrating simulation and signal processing in tracking complex social systems. *Computational and Mathematical Organization Theory* (2018), 1–22.

[47] Fan Yang, Alina Vereshchaka, and Wen Dong. 2018. Predicting and Optimizing City-Scale Road Traffic Dynamics Using Trajectories of Individual Vehicles. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 173–180.

[48] Brian D Ziebart. 2010. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Ph.D. Dissertation.