A Variational Perturbative Approach to Planning in Graph-based Markov Decision Processes

Dominik Linzner[1] **and Heinz Koeppl**[1, 2]
[1]Department of Electrical Engineering and Information Technology
[2]Department of Biology
Technische Universität Darmstadt
{dominik.linzner, heinz.koeppl}@bcs.tu-darmstadt.de

Abstract

Coordinating multiple interacting agents to achieve a common goal is a difficult task with huge applicability. This problem remains hard to solve, even when limiting interactions to be mediated via a static interaction-graph. We present a novel approximate solution method for multi-agent Markov decision problems on graphs, based on variational perturbation theory. We adopt the strategy of planning via inference, which has been explored in various prior works. We employ a non-trivial extension of a novel high-order variational method that allows for approximate inference in large networks and has been shown to surpass the accuracy of existing variational methods. To compare our method to two state-of-the-art methods for multi-agent planning on graphs, we apply the method different standard GMDP problems. We show that in cases, where the goal is encoded as a non-local cost function, our method performs well, while state-of-the-art methods approach the performance of random guess. In a final experiment, we demonstrate that our method brings significant improvement for synchronization tasks.

Introduction

Understanding and designing the behavior of multiple agents interacting through large networks in order to achieve a common goal is a task studied across many fields, such as artificial intelligence (Sigaud and Buffet 2013), electrical engineering (Tousi, Hosseinian, and Menhaj 2010), but also economics and biological sciences (Castellano, Fortunato, and Loreto 2009) and epidemics (Venkatramanan et al. 2018). Finding optimal policies, e.g., for the distribution of information across a social or communication network, for effective intervention in molecular networks or for vaccinations in order to prevent spreading of diseases are actively discussed problems. In many of these applications, there exists no unique natural time-scale. In such cases, it is appropriate to reason in continuous-time. The setting of multiple agents on a graph in continuous-time has been previously explored (Kan and Shelton 2008).

For a Markov decision process (MDP), an optimal policy can be computed in time scaling polynomially in the size of the state and action space using dynamic programming (Puterman 2005). However, in many realistic scenarios, these spaces are high dimensional, e.g., in multi-agent settings (Boutilier, Dean, and Hanks 1996), where the size of the state and action space of the underlying global MDP in general scales exponentially in the number of agents. Solving such problems exactly is infeasible for large-scale systems. For this reason, various simplifying assumptions on the structure of MDPs have been proposed. Assuming a factorized state space, a local representation of the transition model and the reward function, decomposing according to a graph-structure, so-called factored MDPs (FMDPs) (Guestrin, Koller, and Parr 2001; Boutilier 1996) have been defined. For this model, various approximate solution schemes have been developed (Guestrin, Koller, and Parr 2001; Guestrin et al. 2003).

Graph-based MDPs (GMDPs), as proposed in (Sabbadin, Peyrard, and Forsell 2012), present a subclass of FMDPs, where additionally, agent-wise policies are assumed. We note, that this renders GMDPs equivalent to mMDPs (Boutilier, Dean, and Hanks 1996), interacting and communicating over a graph-structure. GMDPs can be solved approximately using approximate linear programming (Sabbadin, Peyrard, and Forsell 2012), approximate policy iteration (Sabbadin, Peyrard, and Forsell 2012) or approximate value iteration using mean field or cluster variational methods (Cheng and Chen 2013). Additional simplifying assumptions, such as transition-independence of agents (TI-Dec-MDP) can be made (Sigaud and Buffet 2013), however reducing the descriptive power of the model. We will thus not compare to such models.

In this work, we propose a novel method for approximate inference and planning for GMDPs inspired by advances in statistical physics. We emphasize that in planning problems (Fleming and Soner 2006), system dynamics are known, given a policy. Thus, we do not encounter problems as in reinforcement learning, e.g., as the *exploration-exploitation* dilemma (Puterman 2005). We employ a scheme based on variational perturbation theory (Tanaka 1999; Paquet, Winther, and Opper 2009; Opper, Paquet, and Winther 2013; Linzner and Koeppl 2018), which was originally introduced in (Plefka 1982).

The manuscript is organized as follows: In Section 2, we briefly summarize the connection between variational inference and planning. Here, the main result is that maximization of expected reward can be coined as maximization of a variational lower bound (Toussaint and Storkey 2006; Furmston and Barber 2010; Kappen, Gómez, and Opper 2012). In Section 3 and 4, we develop an expectation-maximization algorithm to iteratively improve the policy for

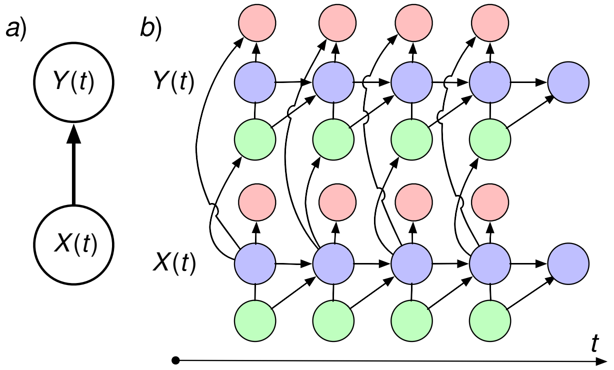

Figure 1: a) A minimal example of a GMDP. The state of agent Y is modulated by its parent X. b) The same GMDP unrolled in time as directed graphical model. Agent X affects agent Y's state (blue) by influencing agent Y's choice over actions (green) defined by Y's policy. The rewards (red) of agent Y are determined by Y's and X's state. It is also possible to incorporate direct modulation of the transition models by the states of adjacent agents (not displayed for readability).

each agent individually. Lastly, we perform simulated experiments on several standard planning task and show realistic cases, where current state-of-the-art methods perform similar to random guess, while our method performs well (Section 5). An implementation of our method is available via Git[1].

Background

Continuous-time MDPs on Graphs. A MDP models an agent picking actions according to a policy, depending on its current state. Its objective is to minimize its reward, while being subject to some, possibly hostile, environment. Herein, we define a homogeneous continuous-time MDP by a tupel $(\mathcal{S}, \mathcal{A}, \mathcal{W}, R)$. It defines a two-component Markov process $\{S(t), A(t)\}$ through a transition intensity matrix $\mathcal{W} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ over a countable state space $\mathcal{S}$ and a countable action space $\mathcal{A}$ together with a policy $\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$. Each state-action pair is mapped to a reward via the *reward function* $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_-$. In this work we only consider negative rewards, which poses no restriction as any bounded reward function can be trivially shifted into the negative half-space. For the sake of conciseness, we will often adopt shorthand notations of the type $p_{t-t'}(s' \mid s, a) \equiv p(S(t) = s' \mid S(t') = s, A(t') = a)$, with $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$. Given a sequence of actions, the evolution of the MDP can be understood as a usual continuous-time Markov chain (CTMC) with the (infinitesimal) transition probability

$$p_h(s' \mid s, a) = \delta_{s,s'} + h\, \mathcal{W}(s' \mid s, a) + o(h),$$

for some time-step h with $\lim_{h \to 0} o(h)/h = 0$, and $\delta_{s,s'}$ the indicator function. We note, that any intensity matrix $\mathcal{W}$ fullfils $\mathcal{W}(s \mid s) = -\sum_{s' \neq s} \mathcal{W}(s' \mid s)$. A multi-agent MDP (mMDP) can be understood as an N-component MDP over state- and action-spaces $\mathcal{S} = \bigtimes_{n=1}^N \mathcal{X}_n$, $\mathcal{A} = \bigtimes_{n=1}^N \mathcal{A}_n$, with $\bigtimes$ denoting the Cartesian product, evolving jointly as an MDP. We state explicitly that single component states and actions are entries of the states and actions of the global MDP, i.e. $s = (x_1, \dots, x_N)$ for $s \in \mathcal{S}$ with $x_n \in \mathcal{X}_n$ and $a = (a_1, \dots, a_N)$ for $a \in \mathcal{A}$ with $a_n \in \mathcal{A}_n$ for all $n \in \{1, \dots, N\}$. In this multi-agent setting, each component, referred to as an individual agent, has no direct access to the global state of the system, but can only observe the states of a subset of agents, which we will call its *parent-set*. In the following analysis, we want to restrict ourselves to mMDPs on graphs (GMDPs).

For GMDPs, the parent configuration can be summarized via a directed graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ encoding the relationship among the agents $\mathcal{V} \equiv \{V_1, \dots, V_N\}$, in this context also referred to as nodes. These are connected via an edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. The parent-set is then defined as $\mathrm{pa}(n) \equiv \{m \mid (m, n) \in \mathcal{E}\}$. Conversely, we define the *child-set* $\mathrm{ch}(n) \equiv \{m \mid (n, m) \in \mathcal{E}\}$. The n'th agents process $\{S_n(t), A_n(t)\}$ then depend only on its current state $x_n \in \mathcal{X}_n$, its action $a_n \in \mathcal{A}_n$ and of all his parents $U_n(t) = u_n$ taking values in $\mathcal{U}_n \equiv \bigtimes_{m \in \mathrm{pa}(n)} \mathcal{X}_m$. We display a sketch of a GMDP in Fig. 1. We note, that cycles in a graphical model as in Fig. 1(a) are unproblematic, as the corresponding temporally unrolled model, as displayed in Fig. 1(b), would be acyclic. For a GMDP, the global marginal transition matrix $p_h(s' \mid s, a)$ then factorizes over agents

$$p_h(s' \mid s, a) = \prod_{n=1}^N p_h(y_n \mid x_n, u_n, a_n),$$

into local conditional transition probabilities. We define local transition rates $w_n^u : \mathcal{X}_n \times \mathcal{X}_n \times \mathcal{A}_n \to \mathbb{R}$ and policies $\pi_n^u : \mathcal{A}_n \times \mathcal{X}_n \to [0, 1]$ for each parent configuration $u \in \mathcal{U}_n$. In the following we write compactly $w_n^u(y \mid x, a) \equiv w_n^u(y_n \mid x_n, a_n)$ and $\pi_n^u(a \mid x) \equiv \pi_n^u(a_n \mid x_n)$. Subsequently, we can express the local conditional transition probabilities as

$$p_h(y_n \mid x_n, u_n, a_n) = \delta_{x,y} + h w_n^u(y_n \mid x_n, a) + o(h). \tag{1}$$

We consider the problem of planning in continuous time over a countable state space.

Definition 1. *Consider a MDP $(\mathcal{S}, \mathcal{A}, \mathcal{W}, R)$ with initial state $s_0 \in \mathcal{S}$ and a policy π. Then, we can define the (discounted) infinite horizon value function in continuous time as*

$$V_p^\pi(s_0) = \mathsf{E}_p\left[\int_0^\infty \mathrm{d}t\, \gamma^t R(S(t), A(t)) \mid S(0) = s_0, \pi\right],$$

with E_p being the expectation with respect to the MDPs path measure p.

We can now cast the planning problem as: for a given initial state s_0, find a policy π^*, such that

$$\pi^* = \arg\max_\pi \{V_p^\pi(s_0)\}. \tag{2}$$

[1] https://git.rwth-aachen.de/bcs/vpt-planning

A common solution strategy for these kinds of problems is to solve the Bellman equation (Puterman 2005). Instead of trying to optimize a Bellman equation, we want to take advantage of the close relationship of planning and inference (Dayan and Hinton 1997; Furmston and Barber 2010; Toussaint and Storkey 2006; Kappen, Gómez, and Opper 2012; Levine and Koltun 2013). In the following, we restrict ourselves to *finite horizon* MDPs, for which the process evolution terminates at time T, and later extend to the *infinite horizon* problem.

Finite Horizon Planning via Inference. In order to establish the connection between inference and planning we can, following (Dayan and Hinton 1997; Toussaint and Storkey 2006; Levine and Koltun 2013) or similarly (Kappen, Gómez, and Opper 2012; Furmston and Barber 2010), define a boolean auxiliary process $Z(t)$ taking values in $\{0,1\}$, with emission probability $p(Z(t)=1 \mid S(t)=s, A(t)=a) = \exp\{R(s,a)\}$. We define the finite horizon trajectories $S_{[0,T]} \equiv \{S(\xi) \mid 0 \le \xi \le T\}$ and $A_{[0,T]} \equiv \{A(\xi) \mid 0 \le \xi \le T\}$, we can express the reward-optimal posterior process for a given policy π according to Definition 1 as $p(S_{[0,T]}, A_{[0,T]} \mid Z_{[0,T]}=1, \pi, s_0)$, with $Z_{[0,T]}=1$ meaning that $Z(\xi)=1$ for $0 \le \xi \le T$. We consider the Kullback–Leibler (KL) divergence between the posterior $p(S_{[0,T]}, A_{[0,T]} \mid Z_{[0,T]}=1, \pi, s_0)$ and a variational measure $q(S_{[0,T]}, A_{[0,T]} \mid \pi, s_0)$ induced by a time-inhomogeneous MDP with the same policy as p (in supplementary B, we show that the KL-divergences between two continuous-time MDPs with different policies diverges). We arrive at a lower bound for the marginal log-likelihood in the finite horizon case

$$\ln p(Z_{[0,T]}=1 \mid \pi, s_0) \ge \mathcal{F}[q,\pi] + V_q^\pi(s_0), \qquad (3)$$

$$\mathcal{F}[q,\pi] \equiv \\ - D_{KL}[q(S_{[0,T]}, A_{[0,T]} \mid \pi, s_0) || p(S_{[0,T]}, A_{[0,T]} \mid \pi, s_0)],$$

with the variational lower bound $\mathcal{F}[q,\pi]$. The full derivation and structure of (3) can be found in supplementary A and B. When performing exact inference, meaning that $q(S_{[0,T]}, A_{[0,T]} \mid \pi, s_0) = p(S_{[0,T]}, A_{[0,T]} \mid \pi, s_0)$, lower bound and log-likelihood coincide and the maximization of the value function as in Definition 1 corresponds to a maximization of the log-likelihood w.r.t the policy

$$\arg\max_\pi \{V_q^\pi(s_0)\} = \arg\max_\pi \{\ln p(Z_{[0,T]}=1 \mid \pi, s_0)\},$$

establishing the connection between planning and inference. When performing approximate inference, we can iteratively maximize the lower bound with respect to q and thereby approximate the log-likelihood, following a maximization with respect to π. This is the expectation-maximization algorithm, which has been previously applied to policy optimization (Toussaint and Storkey 2006; Levine and Koltun 2013).

Infinite Horizon Planning via Inference. The same framework as above can be used in order to solve (discounted) infinite horizon problems. Following (Toussaint and Storkey 2006), this can be achieved by introducing a prior over horizons $p(T)$. As a derivation in continuous-time is missing in literature, we provide it in supplementary C. By choosing $p(T) = \ln(\gamma)\gamma^T$, one recovers exponential discounting with discount factor γ.

Variational Perturbation Theory for GMDPs

Calculating a variational lower bound exactly is in general intractable for interacting systems. This is often circumvented by assuming a factorized proposal distribution $q(x) = \prod_i q_i(x_i)$, which corresponds to the *naive mean-field* approximation. Variational perturbation theory (VPT) offers a different approach. Here, the similarity measure (the KL-divergence) itself is approximated via a series expansion (Tanaka 1999). A prominent example of this approach is Plefka's expansion (Plefka 1982; Bachschmid et al. 2016). The central assumption is that variables are only weakly coupled, i.e. the interaction of variables is scaled in some small perturbation parameter ε. In this case, the objective is to find an expansion of the KL-divergence in orders of the interaction parameter ε: $\mathcal{F}[q,\pi] = \mathcal{F}^{(0)}[q,\pi] + \varepsilon\mathcal{F}^{(1)}[q,\pi] + \ldots$. This approximate variational lower bound is then maximized with respect to q. We note that $\mathcal{F}[q,\pi]$, like in the case of cluster variational methods (Yedidia, Freeman, and Weiss 2000) (CVMs), no longer provides a lower bound but only an approximation. However, in contrast to CVMs (which can be used construct similar approximate KL–divergences (Vázquez, Ferraro, and Ricci-Tersenghi 2017)), variational perturbation theory yields a controlled approximation in the perturbation parameter ε.

Weak Coupling Expansion. In the following, we want to briefly recapitulate and extend the weak coupling expansion for the lower bound in (3), as derived in (Linzner and Koeppl 2018) in the context of factorized CTMCs, to (discounted) infinite horizon GMDPs. For this we notice, that the lower bound $\mathcal{F}[q,\pi]$ decomposes over time

$$\mathcal{F}[q,\pi] = \lim_{h\to 0} \frac{1}{h} \int_0^\infty \mathrm{d}t \, f_t^h[q,\pi],$$

$$f_t^h[q,\pi] = \sum_{s,s',a} \pi(a \mid s) q(s;t) q_h(s' \mid s,a) \ln \frac{p_h(s' \mid s,a)}{q_h(s' \mid s,a)},$$

where we introduced the shorthands for the marginals $q(s;t) \equiv q(S(t)=s)$ and the infinitesimal transition matrix $q_h(s' \mid s,a) \equiv q(S(t+h)=s' \mid S(t)=s, A(t)=a)$ of the variational process q, for notational convenience.

For a weak coupling expansion, we decompose the node-wise transition probability into an uncoupled part, given by averaging over parents $p_h(y_n \mid x_n, a_n) = \mathsf{E}[p_h(y_n \mid x_n, u_n, a_n) \mid x_n]$ and a deviation around it, defined as $g(y_n, x_n, u_n, a_n) \equiv p_h(y_n \mid x_n, u_n, a_n) - p_h(y_n \mid x_n, a_n)$. Following standard mean-field procedure, we extract a scale parameter $g(y_n, x_n, u_n, a_n) = \varepsilon\tilde{g}(y_n, x_n, u_n, a_n)$, with $\tilde{g}(y_n, x_n, u_n, a_n)$ having the same magnitude as the uncoupled part. This allows to rewrite the transition matrix

$$p_h(y_n \mid x_n, u_n, a_n) = p_h(y_n \mid x_n, a_n) + \varepsilon \tilde{g}(y_n, x_n, u_n, a_n). \tag{4}$$

We emphasize, that this procedure is generic and can be performed for any transition probability. This motivates the weak-coupling expansion on which the results in this manuscript are build upon, for which we define the shorthand $q(y_n, x_n, u_n, a_n; t) \equiv q(S_n(t+h) = y_n, S_n(t) = x_n, U_n(t) = u_n, A_n(t) = a_n)$.

Theorem 1 (Weak coupling expansion for GMDPs)**.** *The time point wise lower bound $f_t^h[q, \pi]$ admits an expansion in ε, as given in* (4)*, into node-wise terms $f_{t,n}[q, \pi]$*

$$f_t^h[q, \pi] = \sum_{n=1}^{N} f_{t,n}^h[q, \pi_n] + o(\varepsilon),$$
$$f_{t,n}^h[q, \pi_n] = \sum_{x_n, y_n, u_n, a_n} \pi_n^{u_n}(a_n \mid x_n) q_t(y_n, x_n, u_n) \times \ln \frac{p_h(y_n \mid x_n, u_n, a_n)}{q_h(y_n \mid x_n, u_n, a_n)}.$$

The proof of this theorem is along the same lines as in (Linzner and Koeppl 2018).

Weak Coupling Expansion for GMDPs in Continuous Time. In order to derive the approximate variational lower bound in continuous time for a GMDP, we define variational marginal rates

$$\tau_n^{u_n}(x_n, y_n, a_n; t) \equiv \lim_{h \to 0} \frac{q(y_n, x_n, a_n, u_n; t)}{h} \quad \text{for } x_n \neq y_n$$

and $\tau_n^{u_n}(x_n, x_n, a_n; t) = -\sum_{y \neq x} \tau_n^{u_n}(x_n, y_n, a_n; t)$ but will from now on use the redefinition $x \equiv x_n, y \equiv y_n, a \equiv a_n$, $u \equiv u_n$ for these objects, in order to avoid clutter. We further make use a mean-field assumption $q(y_n, x_n, u_n; t) = q_h(y_n \mid x_n, u_n; t) q_n(x; t) q_n^u(t)$, with the shorthand $q_n^u(t) \equiv \prod_{j \in \mathrm{par(n)}} q_n(u_j, t)$, assuming factorization of the marginals. We emphasize, that in contrast to naive mean-field (Opper and Sanguinetti 2008; Cohn et al. 2010), we only have to assume a factorization of these marginals, but keep the dependency on the parents in the rates $\tau_n^u(x, y, a; t)$. Together with the normalization constraint, this defines an expansion of the proposal transition probability in time-steps of h: $q(y_n, x_n, u_n, a_n; t) = \delta_{x,y} q_n(x; t) q_n^u(t) \pi_n^u(a \mid x) + h\tau_n^u(x, y, a; t) + o(h)$. The proposal transition probability defines an inhomogeneous master equation

$$\dot{q}_n(x; t) = \sum_{y \neq x, u, a} [\tau_n^u(y, x, a; t) - \tau_n^u(x, y, a; t)]. \tag{5}$$

In order for q to describe a probability distribution, this constraint has to be enforced at all times.

Proposition 1. *The variational lower bound of a GMDP has an expansion into agent-wise terms in the perturbation parameter ε*

$$\mathcal{F}[q, \pi] = \mathcal{F}_{\mathrm{VPT}}[q, \pi] + o(\varepsilon)$$
$$\mathcal{F}_{\mathrm{VPT}}[q, \pi] = \sum_{n=1}^{N} \int_0^\infty \mathrm{d}t \, d_\gamma(t) \left\{ H_n(t) + E_n(t) \right\}, \tag{6}$$
$$H_n(t) = \sum_{y, x \neq y, u, a} \tau_n^u(y, x, a; t) \ln \left\{ \frac{\tau_n^u(y, x, a; t)}{q_n(x; t) q_n^u(t)} - 1 \right\},$$
$$E_n(t) = \sum_{y, x \neq y, u, a} \{ q_n(x; t) q_n^u(t) \pi_n^u(a \mid x) w_n^u(y, x|a) + \tau_n^u(y, x, a; t) \ln [w_n^u(y, x|a) \pi_n^u(a \mid x)] \},$$

with the discounting function $d_\gamma(t) \equiv 1 - \int_0^t \mathrm{d}T \, p(T)$.

Algorithm 1 Stationary points of Euler–Lagrange equation

```
Input: Initial trajectories $q_n(x;t) \forall t \in [0,T]$ obeying normalization, boundary conditions $q(x;0)$ and $\rho(x;T)$, reward function $R(s,a)$.
repeat
    for all $n \in \{1,\dots,N\}$ do
        Update $\rho_n(x;t)$ by backward propagation (9).
        Update $q_n(x;t)$ by forward propagation using (8) given $\rho_n(x;t)$.
    end for
until Convergence (6)
Output: Set of $q_n(x;t)$ and $\rho_n(x;t)$.
```

Proof. We proof our proposition by inserting the marginals into the expansion of Theorem 1. We insert the expression of the conditional transition matrix (1). Subsequently, we perform $h \to 0$. We arrive at the approximate lower bound of a GMDP. The discounting function follows from Fubini's theorem. For a detailed derivation, see supplementary D. □

By minimizing this functional, while fulfilling continuity, we can derive approximate dynamic equations corresponding to the stationary solutions of the Lagrangian

$$\mathcal{L}[q, \pi, \eta] = \mathcal{F}_{\mathrm{VPT}}[q, \pi] + \mathcal{C}[q, \eta] + V_q^\pi(s_0), \tag{7}$$

with $\mathcal{C}[q, \eta]$ being the constrain enforcing (5) (see supplementary E) and Lagrange multipliers $\eta_n(t)$.

Approximate Inference

We finally derive approximate dynamics of the GMDP as stationary points of the Lagrangian, satisfying the Euler–Lagrange equation. These are the key equations that enable us to perform scalable approximate inference for large GMDPs.

Proposition 2. *We define the agent-wise expectation $\mathsf{E}_n^\pi[f(x)] \equiv \sum_{u,a} \pi_n^u(a \mid x) q_n^u(t) f(a, u, x)$. The stationary points of the Lagrangian* (7) *are given by the set of ordinary differential equations for every component $n \in \{1, \dots, N\}$*

$$\dot{q}_n(t) = q_n(t) \Omega_n(t) \tag{8}$$

$$\dot{\rho}_n(t) = \{ \Omega_n(t) + \Theta_n(t) + \Psi_n(t) \} \rho_n(t) \tag{9}$$

with

$$\Omega_n(x,y;t) \equiv \mathsf{E}_n^{\pi}[w_n^u(x,y \mid a)]\frac{\rho_n(y;t)}{\rho_n(x;t)}$$

$$\Theta_n(x,y;t) \equiv \delta_{x,y}\left(\mathsf{E}_n^{\pi}[R_n^u(x,a)] + \ln\rho_n(x;t)\frac{\partial_t d_\gamma(t)}{d_\gamma(t)}\right)$$

with $\Psi_n(t)$ *as given in the supplementary and* $R(s,a) = \sum_{n=1}^N R_n^u(x,a)$. *We note, that for exponential discounting* $\frac{\partial_t d_\gamma(t)}{d_\gamma(t)} = \ln\gamma$.

Proof. Differentiating $\mathcal{L}$ with respect to $q_n(x;t)$, its time-derivative $\dot{q}_n(x;t)$, $\tau_n^u(x,y,a;t)$ and the Lagrange multiplier $\eta_n(x;t)$ yield a closed set of coupled ODEs for the posterior process of the marginal distributions $q_n(x;t)$ and transformed Lagrange multipliers $\rho_n(x;t) \equiv \exp(\eta_n(x;t)/d_\gamma(t))$, eliminating $\tau_n^u(x,y,a;t)$. For more details, we refer the reader to supplementary E. □

Although, the restriction on the reward function to decompose into local terms is not necessary, we will assume it for readability. The coupled set of ODEs can be solved iteratively as a fixed-point procedure in the same manner as in previous works (Opper and Sanguinetti 2008) in a forward-backward procedure (see Algorithm 1). Because we only need to solve $2N$ ODEs to approximate the dynamics of an N-agent system, we recover a linear complexity in the number of agents, rendering our method scalable.

We require boundary conditions for the evolution interval in order to determine a unique solution to the set of equations in Proposition 2. We thus set $q_n(x;0) = \delta_{x,x_0}$ to the desired initial state x_0 and $\rho_n(x;t) = 1$ for free evolution of the system. We note that while we do not consider time-dependent reward in general, our method is capable of doing so. We use this in the following control setting: in control scenarios, a deterministic *goal* state of the system is often desired (Kappen, Gómez, and Opper 2012). In this case, we can put infinite reward on the goal state x_T at the boundary T. We then recover the terminal condition $\rho_n(x;t) = \delta_{x,x_T}$. By setting the reward-dependent terms in Proposition 2 to zero, we can evaluate the prior dynamics of the system given a policy. We will use this to evaluate Definition 1 approximately.

Expectation-Maximization for GMDPs. By examining the approximate lower bound of the value function, one notices that it decomposes over local agent-wise value functions, conditioned on its parents.

Remark. The marginal log-likelihood of a GMDP has an approximate agent-wise decomposition

$$\ln p(Z_{[0,T]} = 1 \mid \pi) \geq \sum_{n=1}^{N} \mathcal{F}_{\mathrm{VPT}}^n[q,\pi] + V_q^{\pi}(s_0) + o(\varepsilon), \tag{10}$$

where the $\mathcal{F}_{\mathrm{VPT}}^n[q,\pi]$ are given by Proposition 1.

Because of this, the global marginal log-likelihood can be maximized by locally maximizing local lower bounds of the individual agents with respect to local policies π_n.

Algorithm 2 Expectation-Maximization for Planning

```
Input: Initial trajectories $q_n(x;t) \forall t \in [0,T]$ obeying normalization, boundary conditions $q(x;0)$ and $\rho(x;T)$, reward function $R(s,a)$, initial policy $\pi^{(0)}$.
Set $i = 0$
repeat
    Solve Euler-Lagrange equations given $\pi^{(i)}$ using Algorithm 1.
    for all $n \in \{1,\dots,N\}$ do
        Maximize (10) with respect to $\pi_n$'s.
        Set maximizer $\pi_n^{(i+1)} = \pi_n^*$.
    end for
    $i \to i+1$
until Convergence of (10)
Output: Optimal policy $\pi^*$.
```

Given the dynamic equations from Proposition 2, we now devise a strategy for scalable planning for GMDPs. For this we notice, that the solutions of these equations maximize the lower bound, thereby providing an approximation to the marginal log-likelihood. Because of (10), we can maximize this object as well with respect to the policies π_n for each agent individually. Thus the complexity of our optimization scales linearly in the number of components. Given this maximizer, we again evaluate the dynamic equations. We do this repeatedly until convergence, thereby implementing an expectation-maximization (EM) algorithm. This strategy is summarized in Algorithm 2. We note that the resulting policy is probabilistic, but a MAP-deterministic policy can be constructed.

Experiments

We evaluate the performance of our method on real-world problem settings against two existing state-of-the-art methods for GMDPs on different network topologies. One method is based on policy iteration in mean-field approximation (API) (Sabbadin, Peyrard, and Forsell 2012), the other on approximate linear programming (ALP) (Guestrin, Koller, and Parr 2001). Both algorithms have been developed and implemented in the GMDPtoolbox (Cros et al. 2017). For small problems, we compare the performance of all algorithms to the exact solution. To ensure a correct evaluation, we first construct the GMDP problem and then transform it to the corresponding MDP problem by a built-in function in the GMDPtoolbox, in order to recover the exact solution. For small problems, we finally perform exact policy evaluation using this MDP.

As competing methods are implemented in discrete-time, we have to pass them an equivalent discrete-time version of the continuous-time problem via uniformization (Kan and Shelton 2008). For this we generate transformed rewards and

Table 1: Results of disease control problem. We give the relative deviation $d_r[\%]$ of the values returned by different methods from the exact optimal values.

(μ,ν)	π_{VPT}	π_{ALP}	π_{API}	π_{RND}
$(0.3, 0.3)$	**0**	61	**0**	200
$(0.6, 0.3)$	**0**	60	**0**	292
$(0.9, 0.3)$	**0**	59	**0**	270
$(0.3, 0.6)$	**0**	61	**0**	281
$(0.6, 0.6)$	**0**	60	**0**	354
$(0.9, 0.6)$	**0**	59	**0**	399
$(0.3, 0.9)$	**0**	61	**0**	390
$(0.6, 0.9)$	**0**	60	**0**	344
$(0.9, 0.9)$	**0**	59	**0**	352

Table 2: Results of forest management problem. We give the relative deviation $d_r[\%]$ of the values returned by different methods from the exact optimal values.

(μ,ν)	π_{VPT}	π_{ALP}	π_{API}	π_{RND}
$(0.3, 0.3)$	**0**	1	1	12
$(0.6, 0.3)$	**0**	8	1	12
$(0.9, 0.3)$	**0**	12	1	11
$(0.3, 0.6)$	**1**	27	12	23
$(0.6, 0.6)$	**1**	26	11	22
$(0.9, 0.6)$	**1**	26	12	21
$(0.3, 0.9)$	**9**	58	25	48
$(0.6, 0.9)$	**9**	58	27	50
$(0.9, 0.9)$	**10**	58	38	52

transition matrices

$$\tilde{R}^u_n(x,a) \equiv \frac{w^u_n(x \mid x,a) - \ln\gamma}{\kappa - \ln\gamma} R^u_n(x,a),$$

$$p_{1/\kappa}(y_n \mid a_n, u_n, x_n) \equiv \begin{cases} w^u_n(y \mid x,a), & x \neq y \\ \kappa + w^u_n(x \mid x,a), & \text{else} \end{cases}$$

for some $\kappa \geq |w^u_n(x \mid x,a)|$.

GMDPs have previously been applied to a variety of problems as in agriculture, forest management (Peyrard et al. 2007; Sabbadin, Peyrard, and Forsell 2012), socio-physics (Castellano, Fortunato, and Loreto 2009; Yang et al. 2018) and caching networks (Rezaei, Manoochehri, and Khalaj 2018), to name a few. In the following we want to benchmark our method on those problem sets. We want compare to the exact solution, thus the network considered is a small regular 2×3 grid, with nearest-neighbour bi-directional couplings, unless specified otherwise. In the end, we demonstrate scalability on a larger 5×5 grid in a synchronization task experiment. We denote the policies returned by the different methods with π_{ALP} for ALP, π_{API} for API, π_{RND} for a random policy and π_{VPT} for our method VPT. For all experiments, we set the discount factor to $\gamma = 0.9$ and the atomic reward $r = 1$. As a metric for performance, we calculate the relative deviation $d_r(\pi) \equiv (V^{\pi^*} - V^{\pi})/V^{\pi^*}$ (with π^* being the exact optimal policy) in percent for the crop and forest planning problem, and the 95% interval of total deviation for the opinion dynamics model.

Disease Control. First, we apply our method to the task of disease control, originally posed for crop fields (Sabbadin, Peyrard, and Forsell 2012). Each crop is in either of two states – susceptible or infected ($\mathcal{X} = \{1, 2\}$). The rate $\alpha(u) = 1 + \frac{1}{2}(1 - (1-\mu)^{|u|})$, with which a susceptible crop is infected, is proportional to the number of its infected neighbours, which we denote by $|u|$. The recovery rate is assumed to be constant ν. The planner has to decide between two local actions for each crop – either to harvest or to leave it fallow and treat it ($\mathcal{A} = \{1, 2\}$). Below, we summarize the transition model:

	$a = 1$		$a = 2$	
w^u_n	$x = 1$	$= 2$	$x = 1$	$= 2$
$y = 1$	$-\alpha(u)$	$\alpha(u)$	0	0
$y = 2$	0	0	ν	$-\nu$

The reward model is:

R^u_n	$x = 1$	$= 2$
$a = 1$	0	0
$a = 2$	r	$r/2$

In Table 1, we display the results of this experiment for different parameters. We find that API and VPT perform equally well in this problem.

Forest Management. We consider the forest management problem as in (Sabbadin, Peyrard, and Forsell 2012). Here, each node has multiple states dependent of each trees age and whether it is damaged by wind or not. A tree can either age or become damaged over time. In a simplified scenario, we are going to assume, that a tree can either be grown – or not – or damaged ($\mathcal{X} = \{1, 2, 3\}$). As trees can shield one-another against wind-damage, this rate $\alpha(u) = 1 + \frac{1}{2}(1 - (1-\mu)^{-|u|})$ depends on the number of grown trees $|u|$. The planner has, again, two actions – either to harvest and cut down the tree or to leave it ($\mathcal{A} = \{1, 2\}$). The transition model is summarized below:

	$a = 1$			$a = 2$		
w^u_n	$x = 1$	$= 2$	$= 3$	$x = 1$	$= 2$	$= 3$
$y = 1$	$-\nu$	ν	0	0	0	0
$y = 2$	0	$-\alpha(u)$	$\alpha(u)$	1	-1	0
$y = 3$	0	0	0	1	0	-1

As yield depends on having neighbours for various reasons, the reward function in (Sabbadin, Peyrard, and Forsell 2012) has a non-local form. We consider reward functions as:

R^u_n	$x = 1$	$= 2$	$= 3$
$a = 1$	0	0	0
$a = 2$	0	$r - \lvert u\rvert$	$\frac{r - \lvert u\rvert}{2}$

Table 3: Results of voter model for an ensemble of 20 random reward functions. We give the 95% interval of the deviation of the achieved values returned by different methods from the exact optimal value.

(μ, ν)	π_{VPT}	π_{ALP}	π_{API}	π_{RND}
$(0.1, 0.0)$	**0.0**	**0.0**	**0.0**	1.9
$(0.2, 0.0)$	**0.4**	**0.4**	**0.4**	5.6
$(0.0, 0.1)$	**0.8**	5.5	5.2	4.7
$(0.1, 0.1)$	**1.8**	5.6	5.7	6.4
$(0.2, 0.1)$	**0.4**	1.2	1.8	5.6
$(0.0, 0.2)$	**0.0**	3.5	4.2	4.7
$(0.1, 0.2)$	**0.1**	4.4	8.8	9.1
$(0.2, 0.2)$	**0.0**	3.0	4.7	10.4

The results of this experiment are displayed in Table 2, where we give the relative deviation in percent between the optimal and the policies returned from the different methods. We find that for all parameters, our method performs significantly better than other methods.

Opinion Dynamics. In this experiment we test the performance of our method on the seminal Ising model, which has, among others, applications in socio-physics (Castellano, Fortunato, and Loreto 2009) to model opinion dynamics, swarming (Šošic et al. 2017), or as a benchmark for multi-agent reinforcement learning (Yang et al. 2018). In the Ising model, each node is in either of two states $\mathcal{X} = \{-1, 1\}$ and the reward function takes the form

$$R(s, a) = \sum_{n=1}^{N} x_n \left\{ J_n + \sum_{k \in \text{par}(n)} J_{n,k} x_k \right\}. \qquad (11)$$

In the following, we want to consider random reward functions, where couplings are drawn from gaussians $J_n \sim \mathcal{N}(0, \mu)$ and $J_{n,k} \sim \mathcal{N}(0, \nu)$. Further, we model the transition rates according to opinion dynamics (voter model) (Castellano, Fortunato, and Loreto 2009) $\alpha(u) = \frac{1}{2}[1 + \tanh(|u|)]$ and $\beta(u) = \frac{1}{2}[1 - \tanh(|u|)]$, with $|u|$, being the sum of the sequence u, see below:

	$a = 1$		$a = 2$	
w_n^u	$x = -1$	$= 1$	$x = -1$	$= 1$
$y = -1$	$-\alpha(u)$	$\alpha(u)$	$-\beta(u)$	$\beta(u)$
$y = 1$	$\beta(u)$	$-\beta(u)$	$\alpha(u)$	$-\alpha(u)$

The results for an ensemble 20 random reward functions displayed in Table 3. Again, we find that our method performs best in all tested parameter regimes, while in some cases RND achieves a higher value than API and ALP.

Synchronization of Agents. In a final experiment, we want to compare the performance of methods in a synchronization task. We consider a regular grid of 5 × 5 agents. We encode a synchronization goal by reward function as in (11) with $J_n = 0$ and $J_{n,k} = -1$. The reward function takes the from of an order parameter $\mathcal{R}(s) = \sum_{i,j \in \text{par}(i)} \delta_{x_i \neq x_j}$, which measures anti-parallel alignment between neighbouring agents. Each agents transition model is local:

	$a = 1$		$a = 2$	
w_n^u	$x = -1$	$= 1$	$x = -1$	$= 1$
$y = -1$	-0.9	0.9	-0.1	0.1
$y = 1$	0.1	-0.1	0.9	-0.9

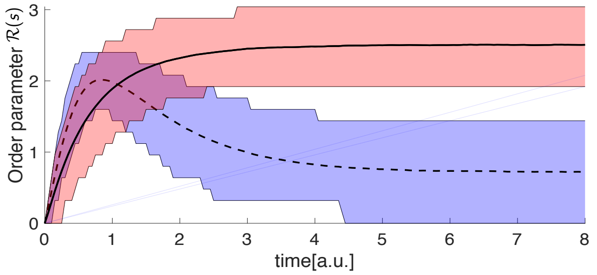

Figure 2: Results of the synchronization task. We track the mean order parameter over time under the VPT (red) and MF (blue-dashed) policy. Areas denote 90% percent of variance.

We display $\mathcal{R}(s)$ over time for different methods in Figure 2 (LP returned the same policy as MF). For evaluation, we simulated each trained model using Gillespie sampling.

Conclusion

We proposed a new method to conduct planning on large scale GMDPs based on variational perturbation theory. We compare our method to state-of-the-art methods for planning in GMDPs and show, that for non-local reward functions state-of-the-art methods approach the performance of random guess, while our method performs well. In the future, we want to use this planning method as the basis for a new reinforcement algorithm for multiple agents on a graph.

Acknowledgements

We thank the anonymous reviewers for helpful comments on the previous version of this manuscript. Dominik Linzner acknowledges funding by the European Union's Horizon 2020 research and innovation programme (iPC–Pediatric Cure, No. 826121) and (PrECISE, No. 668858). Heinz Koeppl acknowledges support by the European Research Council (ERC) within the CONSYN project, No. 773196, and by the Hessian research priority programme LOEWE within the project CompuGene.

References

Bachschmid, L.; Battistin, C.; Opper, M.; and Roudi, Y. 2016. Variational perturbation and extended Plefka approaches to dynamics on random networks: The case of the kinetic Ising model. *Journal of Physics A: Mathematical and Theoretical* 49(43):434003–33.

Boutilier, C.; Dean, T.; and Hanks, S. 1996. Planning under uncertainty: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research* 11:1–94.

Boutilier, C. 1996. Planning, learning and coordination in multiagent decision processes. *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge* 195–210.

Castellano, C.; Fortunato, S.; and Loreto, V. 2009. Statistical physics of social dynamics. *Reviews of Modern Physics* 81(2):1–58.

Cheng, Q., and Chen, F. 2013. Variational Planning for Graph-based MDPs. *Advances in Neural Information Processing Systems*.

Cohn, I.; El-Hay, T.; Friedman, N.; and Kupferman, R. 2010. Mean field variational approximation for continuous-time Bayesian networks. *Journal Of Machine Learning Research* 11:2745–2783.

Cros, M. J.; Aubertot, J. N.; Peyrard, N.; and Sabbadin, R. 2017. GMDPtoolbox: A Matlab library for designing spatial management policies. Application to the long-term collective management of an airborne disease. *PLoS ONE* 12(10):e0186014.

Dayan, P., and Hinton, G. E. 1997. Using EM for reinforcement learning. *Neural Computation* 278(1):271–278.

Fleming, W. H., and Soner, H. M. 2006. *Controlled Markov processes and viscosity solutions*. Springer.

Furmston, T., and Barber, D. 2010. Variational Methods for Reinforcement Learning. *International conference on artificial intelligence and statistics* 241–248.

Guestrin, C.; Koller, D.; Parr, R.; and Venkataraman, S. 2003. Efficient solution algorithms for factored MDPs. *J. Artificial Intelligence Res.* 19:399–468.

Guestrin, C.; Koller, D.; and Parr, R. 2001. Multiagent Planning with Factored MDPs. *Advances in Neural Information Processing Systems* 1523–1530.

Kan, K. F., and Shelton, C. R. 2008. Solving Structured Continuous-Time Markov Decision Processes. *AAAI*.

Kappen, H. J.; Gómez, V.; and Opper, M. 2012. Optimal control as a graphical model inference problem. *Machine Learning* 87(2):159–182.

Levine, S., and Koltun, V. 2013. Variational policy search via trajectory optimization. *Advances in Neural Information Processing Systems*.

Linzner, D., and Koeppl, H. 2018. Cluster Variational Approximations for Structure Learning of Continuous-Time Bayesian Networks from Incomplete Data. *Advances in Neural Information Processing Systems* 7891–7901.

Opper, M., and Sanguinetti, G. 2008. Variational inference for Markov jump processes. *Advances in Neural Information Processing Systems 20* 1105–1112.

Opper, M.; Paquet, U.; and Winther, O. 2013. Perturbative corrections for approximate inference in Gaussian latent variable models. *Journal of Machine Learning Research* 14:2857–2898.

Paquet, U.; Winther, O.; and Opper, M. 2009. Perturbation Corrections in Approximate Inference: Mixture Modelling Applications. *Journal of Machine Learning Research* 10:1263–1304.

Peyrard, N.; Sabbadin, R.; Lo-Pelzer, E.; and Aubertot, J. N. 2007. A Graph-based Markov Decision Process framework for Optimising Collective Management of Diseases in Agriculture: Application to Blackleg on Canola. *Modsim 2007: International Congress on Modelling and Simulation* 2175–2181.

Plefka, T. 1982. Convergence condition of the TAP equation for the infinite-range Ising spin glass model. *Journal of Physics A* 15:1971–1978.

Puterman, M. L. 2005. *Markov Decision Processes: Discrete stochastic dynamic programming*. Wiley-Interscience.

Rezaei, E.; Manoochehri, H. E.; and Khalaj, B. H. 2018. Multi-agent Learning for Cooperative Large-scale Caching Networks. *arXiv*.

Sabbadin, R.; Peyrard, N.; and Forsell, N. 2012. A framework and a mean-field algorithm for the local control of spatial processes. *International Journal of Approximate Reasoning* 53(1):66–86.

Sigaud, O., and Buffet, O. 2013. *Markov Decision Processes in Artificial Intelligence*. Wiley.

Šošic, A.; KhudaBukhsh, W. R.; Zoubir, A. M.; and Koeppl, H. 2017. Inverse reinforcement learning in swarm systems. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, volume 3, 1413–1420.

Tanaka, T. 1999. A Theory of Mean Field Approximation. *Advances in Neural Information Processing Systems*.

Tousi, M. R.; Hosseinian, S. H.; and Menhaj, M. B. 2010. A Multi-agent-based voltage control in power systems using distributed reinforcement learning. *Simulation* 87(7):581–599.

Toussaint, M., and Storkey, A. 2006. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *International conference on Machine learning*.

Vázquez, E. D.; Ferraro, G. D.; and Ricci-Tersenghi, F. 2017. A simple analytical description of the non-stationary dynamics in Ising spin systems. *Journal of Statistical Mechanics: Theory and Experiment* 2017(3):033303.

Venkatramanan, S.; Lewis, B.; Chen, J.; Higdon, D.; Vullikanti, A.; and Marathe, M. 2018. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics* 22:43–49.

Yang, Y.; Luo, R.; Li, M.; Zhou, M.; Zhang, W.; and Wang, J. 2018. Mean field multi-agent reinforcement learning. In *35th International Conference on Machine Learning, ICML 2018*, volume 12, 8869–8886.

Yedidia, J. S.; Freeman, W. T.; and Weiss, Y. 2000. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. *Advances in Neural Information Processing Systems* 13.

Appendix A – KL-Divergence between two MDPs

For any time-discretization, the KL-divergence takes the form

$$KL(q||p) = \sum_{X_1,\dots X_N, A_1,\dots A_{N-1}} q(A_1, \dots A_{N-1}, X_1, \dots X_N) \ln \left[\frac{q(A_1, \dots A_{N-1}, X_1, \dots X_N)}{p(A_1, \dots A_{N-1}, X_1, \dots X_N)} \right].$$

Making use of the Markov property of both distributions

$$p/q((A_1, X_1), \dots (A_N, X_N)) = p/q(A_0, X_0) \prod_{k=1}^{N} p/q((A_{k+1}, X_{k+1}) \mid (A_k, X_k)),$$

we arrive after some basic algebraic manipulations at

$$KL(q||p) = \sum_{k=1}^{N} \sum_{X_k A_k X_{k+1}} q(X_{k+1}, A_k, X_k) \ln \left[\frac{q(A_k \mid X_k) q(X_{k+1} \mid A_k, X_k)}{p(A_k \mid X_k) p(X_{k+1} \mid A_k, X_k)} \right],$$

– or in the notation of the main-paper

$$KL(q||p) = \sum_{t=1}^{N} \sum_{y,x,a} q(y, x \mid a; t) \pi_q(a \mid x) \ln \left(\frac{q(y \mid x, a; t) \pi_q(a \mid x)}{p(y \mid x, a; t) \pi_p(a \mid x)} \right),$$

where we identified the stationary policies $\pi_p(a \mid x) \equiv p(A_k \mid X_k)$ and $\pi_q(a \mid x) \equiv q(A_k \mid X_k)$.

Appendix B – KL-Divergence between two continuous-time MDPs

In order to perform the continuous-time limit, we have to define an expansion of the variational distribution q in some infinitesimal time-step h.

$$q(y, x, a; t) \equiv \delta_{x,y} q(x; t) \pi_q(a \mid x) + h \frac{\tau(x, y, a; t)}{q(x; t) \pi_q(a \mid x)} + o(h),$$

and $\tau(x, x, a; t) = -\sum_{y \neq x} \tau(x, y, a; t)$. Plugging in this definition and the expansion $p(y \mid x, a) = \delta_{x,y} + h w(y \mid x, a)$, we can write

$$KL(q||p) = \sum_{t} \sum_{y,x,a} \left[\delta_{x,y} q(x; t) + h \frac{w(x, y, a)}{\pi_q(a \mid x)} \right] \pi_q(a \mid x) \ln \left(\frac{\delta_{x,y} + h \frac{\tau(x,y,a;t)}{q(x;t)\pi_q(a|x)}}{\delta_{x,y} + h w(y \mid x, a)} \frac{\pi_q(a \mid x)}{\pi_p(a \mid x)} \right).$$

Finally, after some algebraic manipulations, using $\lim_{h \to 0} \ln(1 + hx) = hx$ and $\lim_{h \to 0} h \sum_{t=1}^{T} = \int_0^T \mathrm{d}t$

$$\begin{aligned} KL(q||p) = & \int_0^T \mathrm{dt} \sum_{x, y \neq x, a} \left\{ q(x; t) \pi_q(a \mid x) w(y \mid x, a) - \tau(x, y, a; t) \ln \left(w(y \mid x, a) \pi_p(a \mid x) \right) \right\} \\ & + \int_0^T \mathrm{dt} \sum_{x, y \neq x, a} \tau(x, y, a) \left\{ \ln \left(\tau(x, y, a; t) \right) - \ln q(x; t) - \ln \left(\frac{\pi_q(a \mid x)}{\pi_p(a \mid x)} \right) - 1 \right\} \\ & + \frac{1}{h} \int_0^T \mathrm{dt} \sum_{x, y \neq x, a} q(x; t) \pi_q(a \mid x) \ln \left(\frac{\pi_q(a \mid x)}{\pi_p(a \mid x)} \right). \end{aligned}$$

If now the policies $\pi_q \neq \pi_p$, then the last term in the KL-divergence becomes infinite in the limit $h \to 0$. Thus we have to enforce $\pi = \pi_q = \pi_p$ and arrive at

$$\begin{aligned} KL(q||p) = \int_0^T \mathrm{dt} & \underbrace{\sum_{x, y \neq x, a} \left\{ q(x; t) \pi(a \mid x) w(x, y \mid a) - \tau(x, y, a; t) \ln \left(w(x, y \mid a) \pi(a \mid x) \right) \right\}}_{\equiv E(t)} \\ & + \int_0^T \mathrm{dt} \underbrace{\sum_{x, y \neq x, a} \tau(x, y, a; t) \left\{ \ln \frac{\tau(x, y, a; t)}{q(x; t)} - 1 \right\}}_{\equiv H(t)} \end{aligned}$$

Appendix C – Discounting

In order to incorporate discounting into the framework of planning via inference, one can introduce a prior over horizons $T \sim p(T \mid \gamma)$. In this case the KL-divergence between q and the reward optimal posterior (see main-text) $p(X_{[0,\infty]}, A_{[0,\infty]} \mid \pi, Z_{[0,\infty]} = 1)$ becomes

$$\begin{aligned}
&KL(q(X_{[0,T]}, A_{[0,T]} \mid \pi)p(T \mid \gamma)||p(T \mid \gamma)p(X_{[0,T]}, A_{[0,T]} \mid \pi, Z_{[0,T]} = 1)) = \\
&KL(q(X_{[0,T]}, A_{[0,T]} \mid \pi)p(T \mid \gamma)||p(T \mid \gamma)p(X_{[0,T]}, A_{[0,T]} \mid \pi)) \\
&+ \int_0^\infty \mathrm{d}T\, p(T \mid \gamma) \int_0^T \mathrm{d}t \sum_{s,a} q(s;t)\pi(a \mid s)R(s,a) - \ln p(Z_{[0,\infty]} = 1 \mid \pi).
\end{aligned}$$

By using Fubinis theorem, we can exchange the integration order $\int_0^\infty \mathrm{d}T\, p(T \mid \gamma) \int_0^T \mathrm{d}t = \int_0^\infty \mathrm{d}t \int_t^\infty \mathrm{d}T\, p(T \mid \gamma)$ and further noticing that $\int_t^\infty \mathrm{d}T\, p(T \mid \gamma) = 1 - \int_0^t \mathrm{d}T\, p(T \mid \gamma)$, we arrive at the discount factor from the main text $d_\gamma(t) \equiv 1 - \int_0^t \mathrm{d}T\, p(T \mid \gamma)$. We notice that for an exponential prior $p(T \mid \gamma) = \ln \gamma \gamma^T$, we get $d_\gamma(t) = \gamma^t$ – the standart exponential discount factor. We note, that the planning via inference framework allows naturally for non-exponential discounting, where in general a Bellmann equation can not be issued. Observing that $KL(q(X_{[0,T]}, A_{[0,T]} \mid \pi)p(T \mid \gamma)||p(T \mid \gamma)p(X_{[0,T]}, A_{[0,T]} \mid \pi, Z_{[0,T]} = 1)) \geq 0$ we arrive at a variational lower bound to the marginal likelihood in the discounted case

$$\ln p(Z_{[0,\infty]} = 1 \mid \pi) \geq \mathcal{F}[q, \pi] + V_q^\pi(s0),$$

where we inserted the definition

$$\begin{aligned}
V_q^\pi(s_0) &= \sum_{s,a} \int_0^\infty \mathrm{d}t\, \gamma^t q(s;t)\pi(a \mid s)R(s,a) \\
&= \mathsf{E}_q\left[\int_0^\infty \mathrm{d}t \gamma^t R(X(t), A(t)) \mid X(0) = s_0, \pi\right]
\end{aligned}$$

and

$$\mathcal{F}[q, \pi] \equiv -KL(q(X_{[0,T]}, A_{[0,T]} \mid \pi)p(T \mid \gamma)||p(T \mid \gamma)p(X_{[0,T]}, A_{[0,T]} \mid \pi)).$$

In a derivation, analogous to Appendix B above, we recover

$$\mathcal{F}[q, \pi] = -\int_0^\infty \mathrm{d}t\; d_\gamma(t) \left\{E(t) + H(t)\right\}.$$

Appendix D – Continuous-time variational lower-bound

In order to perform the continuous-time limit, we represent q by an expansion in h in set of marginals

$$q(y_n, x_n, u_n, a_n; t) = \delta_{x,y} q_n(x;t) q_n^u(t) \pi_n^u(a \mid x) + h \frac{\tau_n^u(x,y,a;t)}{q_n(x;t)q_n^u(t)\pi_n^u(a \mid x)} + o(h),$$

with $\tau_n^u(x,x,a,t) = -\sum_{y \neq x} \tau_n^u(x,y,a;t)$. By inserting $q's$ representation into $\mathcal{F}_{VPT}[q,\pi]$ we get

$$\begin{aligned}
\mathcal{F}_{VPT}[q,\pi] = \frac{1}{h}\int_0^\infty \mathrm{d}t d_\gamma(t) & \sum_{y \neq x, x, u, a} h\tau_n^u(x,y,a;t) \left[\ln h \frac{\tau_n^u(x,y,a;t)}{q_n(x;t)q_n^u(t)\pi_n^u(a \mid x)} - \ln h \pi_n^u(a \mid x) w_n^u(x, y \mid a)\right] \\
&+ \sum_{x,u} \left\{ q_n(x;t)q_n^u(t)\pi_n^u(a \mid x) - h\sum_{y \neq x} \tau_n^u(x,y,a;t) \right\} \\
&\times \left[\ln\left\{1 - h\frac{\sum_{y \neq x} \tau_n^u(x,y,a;t)}{q_n(x;t)q_n^u(t)\pi_n^u(a \mid x)}\right\} - \ln\left\{1 - h\sum_{y \neq x, a} \pi_n^u(a \mid x) w_n^u(y \mid x, a)\right\}\right]
\end{aligned}$$

where we also inserted $P(X_n(t) = y_n \mid X_n(t) = x_n, U_n(t) = u_n, A_n(t) = a_n) = \delta_{x,y} + w_n^u(x, y \mid a)\pi_n^u(a \mid x)h$. With the asymptotic identity $\ln(1 + hx) = hx$ we can simplify

$$\mathcal{F}_{VPT}[q, \pi] = \frac{1}{h}\int_0^\infty \mathrm{d}t d_\gamma(t) \sum_{y \neq x, x, u, a} h\tau_n^u(x, y, a; t)\left[\ln\frac{\tau_n^u(x, y, a; t)}{q_n(x; t)q_n^u(t)\pi_n^u(a \mid x)} - \ln \pi_n^u(a \mid x)w_n^u(x, y \mid a)\right]$$

$$+ \sum_{a,u,x}\left\{q_n(x; t)q_n^u(t)\pi_n^u(a \mid x) - h\sum_{y \neq x}\tau_n^u(x, y, a; t)\right\}$$

$$\times\left[h\sum_{y \neq x}\pi_n^u(a \mid x)w_n^u(x, y \mid a) - h\frac{\sum_{y \neq x}\tau_n^u(x, y, a; t)}{q_n(x; t)q_n^u(t)\pi_n^u(a \mid x)}\right]$$

which becomes in the continuous-time limit $h \to 0$

$$\mathcal{F}_{VPT}[q, \pi] = \sum_n \int_0^\infty \mathrm{d}t d_\gamma(t) \underbrace{\sum_{x, y \neq x, u} \tau_n^u(x, y, a; t)[1 - \ln \tau_n^u(x, y, a; t) + \ln(q_n^u(t)q_n(x; t))]}_{\equiv H_n(t)}$$

$$+ \sum_n \int_0^\infty \mathrm{d}t d_\gamma(t) \underbrace{\sum_{y \neq x, x, a, u} [q_n(x; t)q_n^u(t)w_n^u(x, y \mid a)\pi_n^u(a \mid x) + \tau_n^u(x, y, a; t)\ln w_n^u(x, y \mid a)\pi_n^u(a \mid x)].}_{\equiv E_n(t)}$$

The contribution of the likelihood term can be derived to be

$$\mathsf{E}[R(s, a)] = \sum_n \int_0^\infty \mathrm{d}t\, d_\gamma(t)\sum_{x,u} q_n(x; t)q_n^u(t)R_n^u(x, a)$$

Appendix E – Approximate GMDP dynamics

We are now going to derive the dynamics of GMDPs, defined by fulfilling the Euler–Lagrange equations

$$\partial_x\mathcal{L}[t, x, \dot{x}] - \partial_t[\partial_{\dot{x}}\mathcal{L}[t, x, \dot{x}]] = 0.$$

First lets consider the derivative with respect to $q_n(x; t)$:

$$\partial_{q_n(x;t)}H_n = d_\gamma(t)\sum_u\sum_{y \neq x}\frac{\tau_n^u(x, y; t)}{q_n(x; t)}, \quad \partial_{q_n(x;t)}E_j = d_\gamma(t)\mathsf{E}_n[\sum_a w_n^u(x, x \mid a)\pi_n^u(a \mid x)],$$

Further if node n has a child j

$$\partial_{q_n(x;t)}H_j = d_\gamma(t)\sum_{x,u \mid X_n(t) = x_n = x}\sum_{y \neq x}\frac{\tau_j^u(x, y, t)}{q_n(x; t)},$$

$$\partial_{q_n(x;t)}E_j = d_\gamma(t)\sum_x q_j(x; t)\mathsf{E}_n[\sum_a w_n^u(x, x \mid a)\pi_n^u(a \mid x) \mid X_n(t) = x].$$

With respect to the derivative $\dot{q}_n(x; t)$ we get

$$\partial_{\dot{q}_n(x;t)}\mathcal{L} = -w_n(x; t).$$

We derive with respect to the transitions

$$\partial_{\tau_n^u(x,y,a;t)}H_n = d_\gamma(t)\ln[q_n(x; t)q_n^u(t)] - \ln\tau_n^u(x, y, a; t), \quad \partial_{\tau_n^u(x,y,a;t)}E_n = d_\gamma(t)\ln w_n^u(x, y \mid a)\pi_n^u(a \mid x).$$

thus

$$\partial_{\tau_n^u(x,y,a;t)}\mathcal{L} = d_\gamma(t)\ln[q_n(x; t)q_n^u(t)] - d_\gamma(t)\ln\tau_n^u(x, y; t) + d_\gamma(t)\ln w_n^u(x, y \mid a)\pi_n^u(a \mid x) - \eta_n(x; t) + \eta_n(y; t).$$

The derivative with respect to the Lagrange-multipliers yields:

$$\partial_{\eta_n(x;t)}\mathcal{L} = -\left\{\dot{q}_n(x; t) - \left[\sum_{y \neq x, u}\tau_n^u(y, x; t) - \tau_n^u(x, y; t)\right]\right\}$$

And lastly derivatives of $\mathsf{E}[R(s,a)]]$

$$\partial_{q_n(x;t)}\mathsf{E}[R(s,a)] = d_\gamma(t)\sum_{u,a} q_n^u(t)\pi_n(a \mid x)R_n^u(x,a) + d_\gamma(t)\sum_{j\in\mathrm{child}(n)}\sum_{u,a} q_j(x;t)q_j^{u/n}(t)\pi_n(a\mid x)R_j^u(x,a)$$

These can then be combined as the following Euler-Lagrange equations:

$$\begin{aligned}
&\text{(I)}\quad 0 = d_\gamma(t)\sum_u\sum_{y\neq x}\frac{\tau_n^u(x,y,a;t)}{q_n(x;t)} + d_\gamma(t)\mathsf{E}_n[w_n^u(x,y\mid a)\pi_n^u(a\mid x)] + \dot{\eta}_n(x;t) + d_\gamma(t)\mathsf{E}_n[R_n^u(x,a)]\\
&\qquad + d_\gamma(t)\sum_{j\in\mathrm{child}(n)}\sum_{x,u\mid X_n(t)=x}\sum_{y\neq x}\frac{\tau_n^u(x,y,a;t)}{q_n(x;t)}\\
&\qquad + d_\gamma(t)\sum_x q_j(x;t)\left\{\mathsf{E}_j[w_j^u(x,x\mid a)\mid X_n(t)=x] + \mathsf{E}_j[R_j^u(x,a)\mid x_i = x]\right\}\\
&\text{(II)}\quad 0 = \ln[q_n(x;t)q_n^u(t)] - \ln\tau_n^u(x,y,a;t) + \ln w_n^u(x,y\mid a)\pi_n^u(a\mid x) - \eta_n(x;t)/d_\gamma(t) + \eta_n(y;t)/d_\gamma(t)\\
&\text{(III)}\quad \dot{q}_n(x;t) = \sum_{y\neq x,u,a}\left\{\tau_n^u(y,x,a;t) - \tau_n^u(x,y,a;t)\right\}.
\end{aligned}$$

Exponentiating (II) gives

$$\text{(II}^*\text{)}\quad \tau_n^u(x,y,a;t) = q_n(x;t)q_n^u(t)w_n^u(x,y,a\mid a)\pi_n^u(a\mid x)\rho_n(y;t)/\rho_n(x;t),$$

where $\rho_n(x;t)\equiv\exp(\eta_n(x;t)/d_\gamma(t))$. Assuming that w is irreducible, $\rho_n(x;t)$ and $q_n(x;t)$ are non-zero in $(0,T)$ and we can thus eliminate $\tau_n^u(x,y,a;t)$ in (I) and (II). Thus

$$\begin{aligned}
\text{(I}^*\text{)}\quad \dot{\rho}_n(x,t) &= \sum_{y\neq x,a}\mathsf{E}_n[w_n^u(x,y\mid a)\pi_n^u(a\mid x)]\rho_n(y;t)\\
&\quad + \left\{\mathsf{E}_n[w_n^u(x,x\mid a)\pi_n^u(a\mid x)] + \psi_n(x;t) + \ln\rho_n(x;t)\frac{\partial_t d_\gamma(t)}{d_\gamma(t)}\right\}\rho_n(x;t)\\
\text{(III}^*\text{)}\quad \dot{q}_n(x;t) &= \sum_{y\neq x,a}\{m_n(y)\mathsf{E}_n[w_n^u(y,x\mid a)\pi_n^u(a\mid y)]\rho_n(x;t)/\rho_n(y;t)\\
&\quad - q_n(x;t)\mathsf{E}_n[w_n^u(x,y\mid a)\pi_n^u(a\mid x)]\rho_n(y;t)/\rho_n(x;t)\},
\end{aligned}$$

where we used that

$$\frac{\partial_t\eta_i(x)}{d_\gamma(t)} = \frac{1}{\rho_i(x;t)}\partial_t\rho_i(x;t) + \ln\left(\rho_i(x;t)\right)\frac{\partial_t d_\gamma(t)}{d_\gamma(t)}$$

. We further summarized

$$\begin{aligned}
\psi_n(x;t) = \sum_{j\in\mathrm{child}(n)}\sum_x q_j(x;t)&\left\{\sum_{y\neq x,a}\frac{\rho_j(y;t)}{\rho_j(x;t)}\mathsf{E}_j[w_j^u(x,y\mid a)\pi_j^u(a\mid x)\mid X_n(t)=x]\right.\\
&\left. + \mathsf{E}_j[w_j^u(x,x\mid a)\pi_j^u(a\mid x)\mid X_n(t)=x] + \mathsf{E}_j[C_j(x,u)\mid x_i = x]\right\}.
\end{aligned}$$

Appendix F – Policy evaluation

We want to test how accurately we can approximate the true lower bound using Prop. 2. We define local state and action spaces $\mathcal{X}=\{0,1\}$ and $\mathcal{A}=\{0,1\}$. To test the performance under different types of policies, we consider policies

$$\pi_i^u(a\mid x) = \frac{1}{2} + \tanh\left(\beta x_i\sum_{j\in\mathrm{par(i)}}u_j\right), \tag{12}$$

that become increasingly deterministic with increasing β. In order to keep this experiment simple, we assume a deterministic transition model $w_n^u(y\mid x,a)=\delta_{x,(-1)^a}$ for $y\neq x$, independent on the parent configuration. Because in this experiment, the agents are coupled by their policies, a more deterministic policy corresponds to a stronger coupling between the agents, thus increasing the perturbation parameter ε. We test our method on two different tree topologies and random reward functions of type (10). We set the variances $\sigma_\alpha=\sigma_J=0.2$. One is is a tree-network sketched in the inset of Fig. 3 a), the other a bi-directional chain with periodic boundary conditions as sketched in Fig. 3 b). We find that for both networks, the accuracy of the approximation lowers for increasing β, however the accuracy is worse for the bi-directional chain.

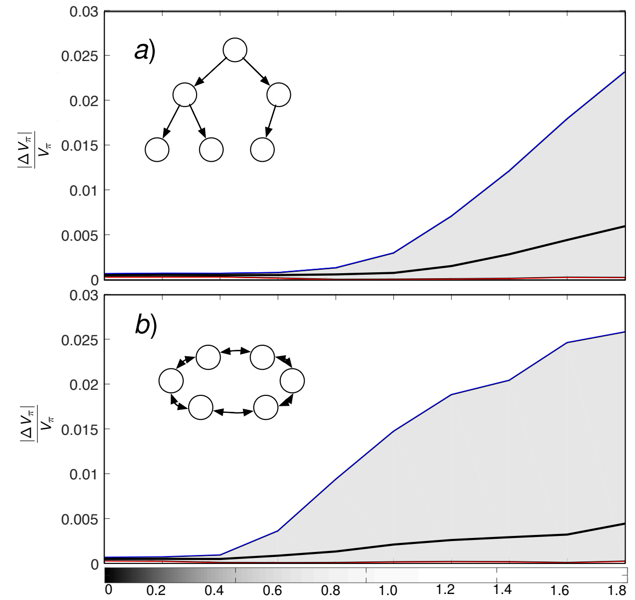

Figure 3: Relative deviation of the approximate expected reward from the exact one for different policies parametrized by β. For reference we plotted the scale of $\pi \propto 2\tanh(\beta)$ on the x-axis (colormap: white= 0 black= 1). The performance is evaluated using 50 random reward functions of type (10) for two different graph topologies (inset). We plotted the 5% (red), 50% (black) and 95% (blue) percentiles. We find that our method performs slightly better on trees a) than on bi-directed chains b).