Submitted to the Annals of Applied Statistics

BAYESIAN INVERSE REINFORCEMENT LEARNING FOR COLLECTIVE ANIMAL MOVEMENT

BY TORYN L. J. SCHAFER[1], CHRISTOPHER K. WIKLE[1] AND MEVIN B. HOOTEN[2,3]

[1]*University of Missouri*

[2]*U. S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit*

[3]*Colorado State University*

Agent-based methods allow for defining simple rules that generate complex group behaviors. The governing rules of such models are typically set *a priori* and parameters are tuned from observed behavior trajectories. Instead of making simplifying assumptions across all anticipated scenarios, inverse reinforcement learning provides inference on the short-term (local) rules governing long term behavior policies by using properties of a Markov decision process. We use the computationally efficient linearly-solvable Markov decision process to learn the local rules governing collective movement for a simulation of the self propelled-particle (SPP) model and a data application for a captive guppy population. The estimation of the behavioral decision costs is done in a Bayesian framework with basis function smoothing. We recover the true costs in the SPP simulation and find the guppies value collective movement more than targeted movement toward shelter.

1. Introduction. Understanding individual animal decision-making processes in social groups is challenging. Traditionally, agent-based models (ABMs) of individual interactions are used as building blocks for complex group dynamics (Vicsek et al., 1995; Couzin et al., 2002; Scharf et al., 2016; McDermott, Wikle and Millspaugh, 2017; Scharf et al., 2018). ABMs attempt to recreate what is observed in nature by defining a mechanistic model *a priori*. While the simple individual-based rules lead to complex group dynamics, ABMs suffer from automatic behavior after reaching some equilibrium, challenges to incorporate interactions with habitat, and no notion of memory (Ried, Müller and Briegel, 2019). The goal of inverse modeling is to instead learn the underlying local rules from observations of sequential behavior decisions (Lee et al., 2017; Kangasrääsiö and Kaski, 2018; Yamaguchi et al., 2018).

Parameters of ABMs in practice need to be tuned or learned by supervised learning (Ried, Müller and Briegel, 2019; Wikle and Hooten, 2016; Hooten, Wikle and Schwob, 2020). A recent alternative to supervised learning is reinforcement learning (RL). RL is goal-oriented learning from continuous interaction between an agent and its environment (Sutton and Barto, 1998). That is, RL methods learn parameters controlling global behavior by trial and error experiments within the defined environment and local rules. The agents learn preferences by paying costs to (or receiving rewards from) the environment and choose optimal behavior by minimizing the cumulative expected future costs (also referred to as "costs-to-go"). Similar to difficulty in tuning ABMs, defining the cost function to produce desired long term behavior is challenging (Ng and Russell, 2000; Finn, Levine and Abbeel, 2016; Arora and Doshi, 2018).

In systems where observations of behavior trajectories can be observed, inverse reinforcement learning (IRL) methods aim to learn the state costs or costs-to-go that governed the observed agents' decisions. Ng and Russell (2000) introduced the first IRL algorithms including dynamic programming, which solves a system of equations based on the state transition probabilities and a grid search method for exploring potential state costs that may have

Keywords and phrases: agent-based model, inverse optimal control, Markov decision process, variational approximation.

generated observed trajectory samples. As surveyed by Arora and Doshi (2018), many more methods have since been developed or adapted to address problems of meaningful size and non-identifiability of the costs. The methods can be broadly categorized as maximum margin optimization (Ratliff, Bagnell and Zinkevich, 2006), entropy optimization (Ziebart et al., 2008), Bayesian IRL (Ramachandran and Amir, 2007; Choi and Kim, 2011; Jin et al., 2017; Šošić, Zoubir and Koeppl, 2017), and deep learning IRL (Wulfmeier, Ondruska and Posner, 2015), with the majority of the methods being applied to Markov decision processes (MDPs). The benefit of Bayesian frameworks to address the non-identifiability of the IRL problem is that they provide a distribution of costs that can generate the observed expert behavior.

Many of the aforementioned methods parameterize the likelihood by the immediate state costs, because the state cost function is a concise description of the task (Ng and Russell, 2000; Ramachandran and Amir, 2007). A computational challenge associated with parametrizing the likelihood by the costs is the necessity to solve the forward MDP each iteration. This is especially challenging in multiagent MDPs, which describe collective animal movement based on calculations of distance between agents that control an agent's state or perception of the environment. An alternative class of MDP, the linearly-solvable MDP (LMDP) introduced by Todorov (2009), is linear in its solution for the optimal policy and thus, less computationally costly for forward modeling. The LMDP is defined by a set of passive dynamics that describe an agent's state transitions in the absence of state costs or environmental feedback and then the optimal state transitions minimize costs-to-go. Moreover, IRL for LMDPs does not require the forward solution for each iteration as there is a linear relationship between the costs-to-go and immediate costs. Therefore, inference about immediate state costs can be obtained by transformation of the estimated costs-to-go. As a special case, Dvijotham and Todorov (2010) showed that maximum entropy IRL is the solution to a LMDP with uniform passive dynamics. Kohjima, Matsubayashi and Sawada (2017) proposed a Bayesian IRL method for learning state values for LMDPs using variational approximation.

As argued by Ried, Müller and Briegel (2019), an MDP (or LMDP) for collective animal movement is a better model for the system than traditional self-propelled particle models (Vicsek et al., 1995). The MDP incorporates the internal processes of an animal by modeling the behavior as perception (state space), planning (state values), and action (see Hooten, Scharf and Morales (2019) for related individual-level models). Furthermore, the behavior is governed by feedback from the environment (which includes other agents) rather than assuming automatic interaction rules. Few applied examples of IRL for collective animal movement exist in the literature. Exceptions include the application of maximum entropy IRL to flocking pigeons of Pinsler et al. (2018) and Bayesian policy estimation of the self-propelled particle (SPP) and Ising models (Šošić et al., 2017).

We present the first application of IRL for collective animal movement using Bayesian learning of state costs-to-go for an LMDP. As an extension of Kohjima, Matsubayashi and Sawada (2017), we reduce the dimension of the state space with basis function approximation, compare variational approximation to MCMC sampling, and consider the multiagent LMDP. We first demonstrate the modeling framework for a simulation of the Vicsek et al. (1995) SPP model to illustrate the mechanisms of the LMDP framework in Section 3. In Section 4, we use the new methodology to estimate state costs-to-go for collective movement of guppies (*Poecilia reticulata*) in a tank to infer trade offs between targeted motion and group cohesion. Finally, we discuss the findings and direction for future work in Section 5.

2. IRL Methodology.

2.1. *LMDP.* We focus on the discrete state space LMDP defined by the tuple $(S, \bar{\mathbf{P}}, \gamma, R)$ where $S = \{1, ..., J\}$ is a finite set of states, $\gamma \in [0, 1]$ is a discount factor, $R : S \to \mathbb{R}$ is a state

cost function, and $\bar{\mathbf{P}}$ is a $J \times J$ transition probability matrix with elements $\bar{p}_{ij}$ for $i = 1, ..., J$ and $j = 1, ..., J$ corresponding to the transition from state i to state j under no control (e.g., passive dynamics). We denote an observation from the set of states as $\mathbf{s} \in \{1, ..., J\}$ and the state cost at state i as r_i for $i = 1, ..., J$ (see Appendix A for a notational reference).

The policy (e.g., how to choose the next state) of an LMDP is defined by continuous controls, $\mathbf{u} = \{u_{ij} \in \mathbb{R}; \forall i, j = 1, ..., J\}$, such that the controlled dynamics are expressed as:

$$p(s_t = j | s_{t-1} = i) = p_{ij}(u_{ij}) \equiv \bar{p}_{ij} \text{exp}(u_{ij}), \tag{1}$$

and the controls are defined to be 0 when the passive transition probability is 0 (i.e., if $\bar{p}_{ij} = 0$, then $u_{ij} = 0$). The controls, u_{ij}, are interpretable as the cost the agent is willing to pay to go against the passive dynamics (Todorov, 2009). For a given policy, the joint costs of the state and control, $l(i, \mathbf{u})$, are:

$$l(i, \mathbf{u}) = r_i + KL(\mathbf{p}_i(\mathbf{u}) || \bar{\mathbf{p}}_i), \tag{2}$$

where r_i is the immediate state cost for states $i = 1, ..., J$ and $KL(\cdot)$ is the Kullback-Leibler (KL) divergence between the controlled transition probability, $\mathbf{p}_i(\mathbf{u}) = (p_{i1}(u_{i1}), ..., p_{iJ}(u_{iJ}))'$, and passive transition probabilities, $\bar{\mathbf{p}}_i = (\bar{p}_{i1}, ..., \bar{p}_{iJ})'$. The KL divergence penalty requires the agent to "pay" a larger price for behavior that deviates from the passive dynamics (Todorov, 2007).

The state costs-to-go, v_i, for $i = 1, ..., J$, are the discounted sum of future expected costs incurred from beginning in state i:

$$v_i = l(i, \mathbf{u}) + E[\gamma \sum_{t=1}^{T} l(j, \mathbf{u})], \tag{3}$$

where the expectation is with respect to the controlled transitions (1). The value of T determines whether the problem has finite- or infinite-horizon (e.g., $T < \infty$ or $T = \infty$). A finite-horizon LMDP can be modeled as an infinite-horizon LMDP by assuming the agent remains in the final observed state and incurs no future costs (Todorov, 2007). Costs-to-go can also be interpreted as relative time to goal completion where a smaller cost-to-go indicates that the agent can reach a desirable state more quickly by transitioning to that state than transitioning to a state with a higher cost-to-go. Based on the definition, there is a recursive relationship between the cost-to-go functions such that (Sutton and Barto, 1998; Todorov, 2009):

$$v_i = l(i, \mathbf{u}) + E[\gamma v_j]. \tag{4}$$

The forward problem of the LMDP is an optimization problem for the set of controls that minimize the cost-to-go and can be expressed by the Bellman optimality equation (e.g., Bellman, 1957) for the state costs-to-go, v_i, for $i = 1, ..., J$:

$$v_i = \min_{\mathbf{u}} \left(l(i, \mathbf{u}) + \gamma \sum_{\forall j \in \mathcal{S}} p_{ij}(u_{ij}) v_j \right), \tag{5}$$

where the summation is over the reachable states $j \in S$ as determined by the policy $p_{ij}(u_{ij})$ for all $j \in S$ (i.e., the expectation in (3) is now expressed as the sum over the discrete distribution defined by (1)). The computational advantage of the LMDP for RL is the Bellman optimality can be solved analytically using the method of Lagrange multipliers for the optimal transition probabilities (Todorov, 2009):

$$p^*(s_t = j | s_{t-1} = i) = \frac{\bar{p}_{ij} \text{exp}(-\gamma v_j)}{\sum_{k=1}^{J} \bar{p}_{ik} \text{exp}(-\gamma v_k)}. \tag{6}$$

By substituting equation (6) into the Bellman optimality (5) and exponentiating, the optimal costs-to-go are a solution to an eigenvector problem that is obtained using a power iteration method (Todorov, 2009), which we demonstrate in Section 3.

2.2. *Inverse Reinforcement Learning (IRL).* Assume we observe a collection of sequences of optimal behavioral state trajectories, $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_N\}$, and $\mathcal{D}_n = \{s_{n0}, ..., s_{nT}\}$, where s_{nt} is the observed state for individual n, for $n = 1, ..., N$, and time point t, for $t = 0, 1, ..., T$. Then, the observed state transitions are summarized into frequencies, $y_{ij} = \sum_{n=1}^{N}\sum_{t=1}^{T} I(s_{nt} = j | s_{n(t-1)} = i)$. We assume that each individual operates according to an LMDP with identical parameters $(S, \bar{\mathbf{P}}, \gamma, R)$, but that the state costs, R, and therefore, costs-to-go, $\mathbf{v}$, are unknown. The likelihood of $\mathcal{D}$ is:

$$
\begin{aligned}
P(\mathcal{D}|\bar{\mathbf{P}}, \mathbf{v}) &= \prod_{n=1}^{N}\prod_{t=1}^{T}\prod_{i=1}^{J}\prod_{j=1}^{J} p^*(s_{nt} = j | s_{n(t-1)} = i), \\
&= \prod_{i=1}^{J}\prod_{j=1}^{J}\left(\frac{\bar{p}_{ij}\exp(-\gamma v_j)}{\sum_k \bar{p}_{ik}\exp(-\gamma v_k)}\right)^{y_{ij}},
\end{aligned} \tag{7}
$$

for all individuals $n = 1, ..., N$, times points $t = 1, ..., T$, transitions from state $i \in S$ to state $j \in S$ and the second equality is based on the optimal transitions (6). We express the costs-to-go vector, $\mathbf{v} = (v_1, ..., v_J)'$, as a linear combination of features in the $J \times n_b$ matrix $\mathbf{X}$ with unknown weights $\boldsymbol{\beta}$ (e.g., $\mathbf{v} = \mathbf{X}\boldsymbol{\beta}$). We estimate the weights in a Bayesian framework by assuming the following hierarchical prior:

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \mathrm{N}\left(\mathbf{0}, \frac{1}{\tau}\mathbf{I}_{n_b}\right), \\
\tau &\sim \mathrm{Gamma}(0.1, 0.1),
\end{aligned} \tag{8}
$$

where $\mathbf{0}$ is an n_b-dimensional vector of zeroes and the parameters are estimated using MCMC sampling and variational approximation with the statistical platform STAN using the R package rstan (Carpenter et al., 2017; Stan Development Team, 2020). For the MCMC sampling, we used the Hamiltonian Monte Carlo with no-U-turn sampler (e.g., Hoffman and Gelman, 2014), which is the default algorithm in STAN. For variational inference, STAN assumes a Gaussian approximating distribution on a transformation of the parameters to a continuous domain (Kucukelbir et al., 2015). We provide brief definitions of the algorithms and the STAN code in an online supplement. Note that the costs-to-go are only estimable up to a constant due to the exponential in (7) and therefore all resulting mean costs-to-go functions are shifted to have a minimum value of 0, which typically corresponds to a terminal state or a state in which an agent incurs no cost indefinitely (Todorov, 2009).

3. SPP LMDP. We illustrate the LMDP for collective movement using the Vicsek et al. (1995) SPP model. We consider the dynamics of the SPP model for agent $n = 1, ..., N$ as defined by Šošić et al. (2017) for the direction θ_{nt} and location (x_{nt}, y_{nt}) as:

$$
\begin{aligned}
\theta_{n(t+1)} &= \langle\theta_{nt}\rangle_\rho + \epsilon_{nt}, \quad \epsilon_{nt} \sim \mathrm{N}(0, \sigma^2), \\
x_{n(t+1)} &= x_{nt} + v_{nt} \cdot \cos(\theta_{nt}), \\
y_{n(t+1)} &= y_{nt} + v_{nt} \cdot \sin(\theta_{nt}),
\end{aligned} \tag{9}
$$

where the agent heads in the mean direction, $\langle\theta_{nt}\rangle_\rho$, of other agents including itself within radius ρ with a speed of v_{nt}. The local misalignment of an agent is the difference between the mean neighborhood direction and the agent's direction, $\langle\theta_{nt}\rangle_\rho - \theta_{nt}$. Šošić et al. (2017) formulated the SPP model as an MDP with 13 discrete actions corresponding to turning angles, $\phi \in [-60°, -50°, ..., 60°]$ and a discrete state space defined by a grid of local misalignment values. The local misalignment grid was defined by $J = 36$ equally sized bins of 10 degrees

with centers $\mathbf{s} = (\pm 180°, -170°, ..., 170°)'$. An agent of the SPP MDP chooses a turning angle, ϕ_{nt}, given the observation of local misalignment bin, s_{nt}. The distribution of the next direction given the turning angle is $\theta_{n(t+1)}|\phi_{nt} \sim \mathrm{N}(\theta_{nt} + \phi_{nt}, \sigma^2)$. The optimal policy corresponds to choosing the turning angle that minimizes the current local misalignment. In our simulation, we assume a constant velocity of 1, fixed interaction radius $\rho = 0.1$, and turning angle standard deviation of 10 degrees (i.e., $\sigma = 10°$) and embed the MDP of Šošić et al. (2017) into the LMDP framework as outlined by Todorov (2007). All angular differences were calculated with the two argument arc-tangent function.

We defined the state cost function, R, as:

$$r_i = \begin{cases} 0 & \text{if } |s_i| \leq 5° \\ 2.5 & \text{if } |s_i| \leq 15° \\ 4 & \text{if } |s_i| \leq 25° \\ 5 & \text{otherwise} \end{cases}, \tag{10}$$

where s_i is the center of the local misalignment bin $i = 1, ..., J$. The costs were chosen based on the results of Šošić et al. (2017) and were constrained in magnitude such that $\exp(-r_i)$ was not numerically 0 (Todorov, 2009).

We assumed agents synchronously chose their next state. The turning angle was therefore equivalent to the change in state (i.e., the difference in states was the difference in directions); this implied a continuous transition distribution for the next state given the turning angle, $s_{n(t+1)}|\phi_{nt} \sim \mathrm{N}(s_{nt} + \phi_{nt}, \sigma^2)$, which can be discretized to provide a transition probability function over the discrete grid defined by $\mathbf{s}$. The LMDP passive state transition probabilities were constructed by summing over the conditional transition probabilities given the discrete turning angles, $\phi \in [-60°, -50°, ..., 60°]$, of the MDP. Therefore the passive dynamics between the discrete grid cell centers s_i and s_j are a discretization of a mixture of normal distribution functions:

$$\bar{p}(s_j|s_i) \propto \sum_{\phi \in [-60°, -50°, ..., 60°]} \Phi\left(\frac{s_j - s_i - \phi + 5°}{10°}\right) - \Phi\left(\frac{s_j - s_i - \phi - 5°}{10°}\right), \tag{11}$$

where Φ is the standard normal cumulative distribution function and the discretization length $5°$ was determined by the half-length of the state grid cells. The passive dynamics were then normalized to have row sums equal to one (i.e., $\sum_{j \in S} \bar{p}(s_j|s_i) = 1$). Lastly, as stated in Section 2.1, the true costs-to-go can be calculated as the solution an eigenproblem. The SPP LMDP setup defines an infinite-horizon problem without an absorbing state (i.e., $\bar{p}_{ii} \neq 1$ for any $i = 1, ..., J$) so we choose to consider the average cost LMDP defined by the following system of equations:

$$\mathbf{z} = \frac{1}{\lambda} \mathrm{diag}(\exp(-\mathbf{r}))\bar{\mathbf{P}}\mathbf{z}, \tag{12}$$

where $\mathbf{z} = \exp(-\mathbf{v})$ is a J-dimensional vector referred to as the desirability function, $\mathrm{diag}(\exp(-\mathbf{r}))$ is a $J \times J$ diagonal matrix with the state costs, $\mathbf{r} = (r_1, ..., r_J)$, on the main diagonal, $\bar{\mathbf{P}}$ is the $J \times J$ passive transition probability matrix, λ is the principal eigenvalue of $[\mathrm{diag}(\exp(-\mathbf{r}))\bar{\mathbf{P}}]$ and $-\log(\lambda)$ corresponds to the average cost of each time step (see supplementary information in Todorov, 2009). The scaling by the largest eigenvalue allows for numerical stability in estimation. The system of equations is solved by initializing the vector $\mathbf{z}$ to all ones, $\mathbf{z} = \mathbf{1}$, and repeatedly multiplying by $[\frac{1}{\lambda}\mathrm{diag}(\exp(-\mathbf{r}))\bar{\mathbf{P}}]$ until convergence. This method is referred to as Z-iteration in the LMDP literature (Todorov, 2009). When applied to the SPP example here, the true cost-to-go function is symmetric about $0°$ with larger relative differences between states near $0°$ than states with local misalignment greater in absolute value than $25°$ (Figure 1). Because the states with local misalignment values greater in

absolute value than 25° have the same immediate cost (10), the differences are related to the average number of time steps it takes an agent to be able to turn toward the group as defined by the passive dynamics; the passive dynamics do not allow an agent to turn more than 90° in one step.

We simulated from the calculated optimal policy with 200 agents for 100 time points and calculated the state transition frequencies using the following algorithm:

1. Initialize $(x_{n0}, y_{n0}, \theta_{n0})$ and calculate local misalignment to determine grid cell s_{n0} for $n = 1, ..., 200$.
2. Repeat the following for $t = 1, ..., 100$ synchronously for $n = 1, ..., 200$:
   a) Sample next local misalignment from $p^*(\cdot|s_{n(t-1)} = i)$.
   b) Calculate turning angle, ϕ_{nt} as difference between $\theta_{n(t-1)}$ and 2a.
   c) Update location (x_{nt}, y_{nt}) according to (9), $\theta_{nt} = \theta_{n(t-1)} + \phi_{nt}$, and local misalignment s_{nt}.

We estimated the costs-to-go with full MCMC sampling and variational approximation for comparison. Additionally, we estimated the costs-to-go for each state separately, (e.g., $\mathbf{X} = \mathrm{I}$), and with Gaussian basis functions with centers on every other grid cell to reduce the state dimension by a factor of 2.

From Figure 1, it is evident that all modeling scenarios estimated the relative true costs-to-go and the estimates from MCMC sampling capture more uncertainty than those from variational approximation. It appears the uncertainty of the estimates increases with an increase in local misalignment and the difference is more apparent for the variational approximation. This pattern generally reflects the amount of data; there were more transitions to states with smaller local misalignment. Furthermore, there were no transitions to states in grid cells centered on $-170^\circ, -150^\circ, 170^\circ, \pm 180^\circ$ misalignment.

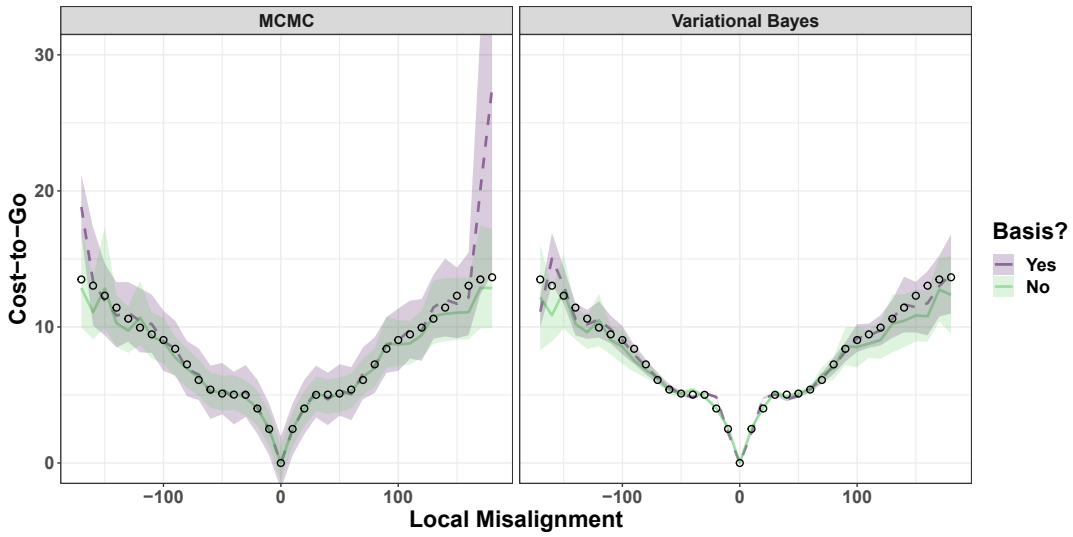

FIG 1. *Estimated cost-to-go function for a Vicsek et al. (1995) SPP model using LMDP IRL for Bayesian MCMC sampling and variational approximation under known passive dynamics. The models either used Gaussian basis functions (dashed lines) or independent state parameters (solid lines). The shaded regions correspond to the 95% C.I. The black open circles is the true cost-to-go function calculated from equation (12). The mean and true cost-to-go functions were shifted to have minimum 0.*

The LMDP framework for IRL allows for efficient estimation of the state costs r_i from the estimation of the cost-to-go by rearranging (12):

$$r_i = \log(\lambda) + v_i + \log\left(\sum_j \bar{p}_{ij}\exp(-v_j)\right), \tag{13}$$

for $i = 1, ..., J$. Figure 2 shows that the estimated costs from the mean cost-to-go functions in Figure 1 generally match the arbitrary state costs defined in (10).

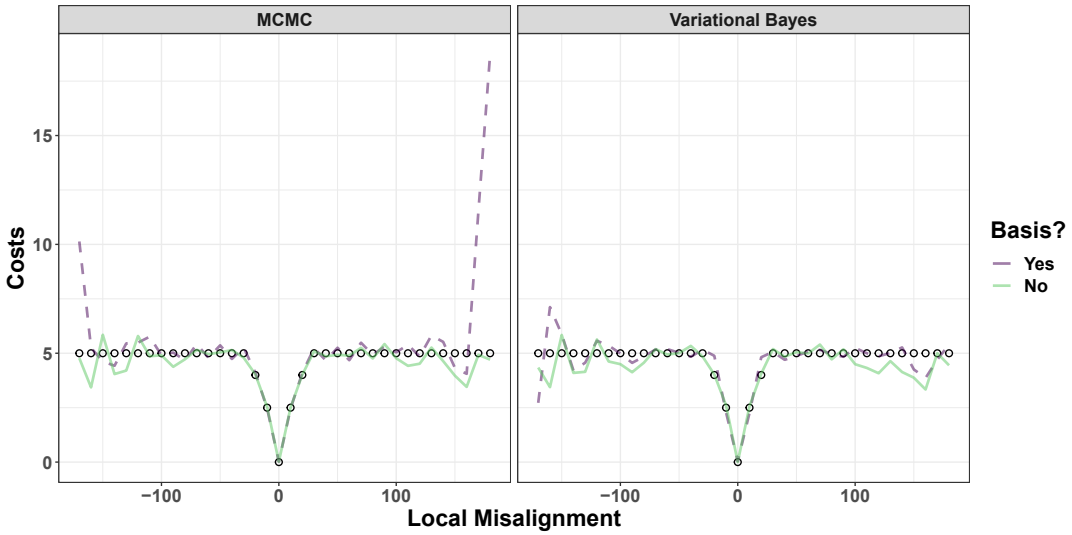

FIG 2. *Estimated state costs for a Vicsek et al. (1995) SPP model using LMDP IRL for Bayesian MCMC sampling and variational approximation under known passive dynamics. The models either used Gaussian basis functions (dashed lines) or independent state parameters (solid lines). The black open circles are the true state costs from* (10).

For the MCMC estimation with bisquare basis functions, there is an increase in cost-to-go and uncertainty at the boundaries. The obvious spike at $180°$ is the cost-to-go for the state defined by local misalignment less than $-175°$ and greater than $175°$. The Gaussian basis functions were not defined on a circle, but rather the continuous real line and could be contributing to the lack of smoothness near the boundary. Additionally, some flexibility is lost in estimation by reducing the dimensionality of the state space with the basis functions.

4. Guppy Application. We used the data available from Bode et al. (2012) on an experiment involving a captive population of guppies (*Poecilia reticulata*). Groups of 10 same sex guppies were filmed from above in a square tank with one corner containing gravel and shade, which is defined by a point. The shaded corner provided shelter and is hypothesized to be attractive to the guppies. The guppies were released in the tank in the opposite corner. The data consist of movement trajectories truncated to the time points when all individuals were moving until one guppy reached the shaded target area. There were 26 experiments with 14 experiments consisting of all males and 12 of all females. We used trajectories from all experiments to estimate the cost-to-go functions.

We defined an LMDP for the guppy trajectories with a discrete state space of local misalignment and target misalignment. Local misalignment was defined as in Section 3 and target misalignment was defined as the difference between the current heading and the direction to

the target point. We rescaled all pixel locations to the unit square and calculated the local misalignment between an individual and all other individuals. The assumption of interaction with all other agents is reasonable as the movement was bounded and there were no visual obstructions outside the target area (Bode et al., 2012). The two misalignment states were discretized using the same $J = 36$ bins of length 10° as in the previous section resulting in a discretized grid of 36×36 states. We assume a fixed discount factor of 1 (i.e., all future costs/rewards are not discounted). Across the 26 experiments, there were 7,816 unique state transitions. Estimation of parameters in the guppy application was done by variational approximation due to the size of the state space.

For the first set of estimated costs-to-go, we assumed the passive dynamics to be discrete uniform, (e.g., $\bar{p}_{ij} \propto 1$ for all $i, j = 1, ..., J$). The features, $\mathbf{X}$, considered were the identity matrix and 819 multiresolution bisquare basis functions generated uniformly within the gridded state space by the R package FRK (Zammit-Mangion, 2020) referred to as "Identity" and "Bisquare" respectively in Figure 3) . The results shown in Figure 3 show a similar pattern among feature matrices with the bisquare basis functions providing more smoothing across the state space. In general, the results suggest the guppies perceived less cost for aligning with other guppies as the low costs-to-go in yellow are concentrated around 0° and there is more flexibility in target alignment as the low costs-to-go have more spread along the target misalignment axis. When comparing the two feature matrices, there is more contrast between the estimated cost-to-go functions that is likely attributable to the dimension reduction and creation of basis functions in the continuous domain rather than circular.

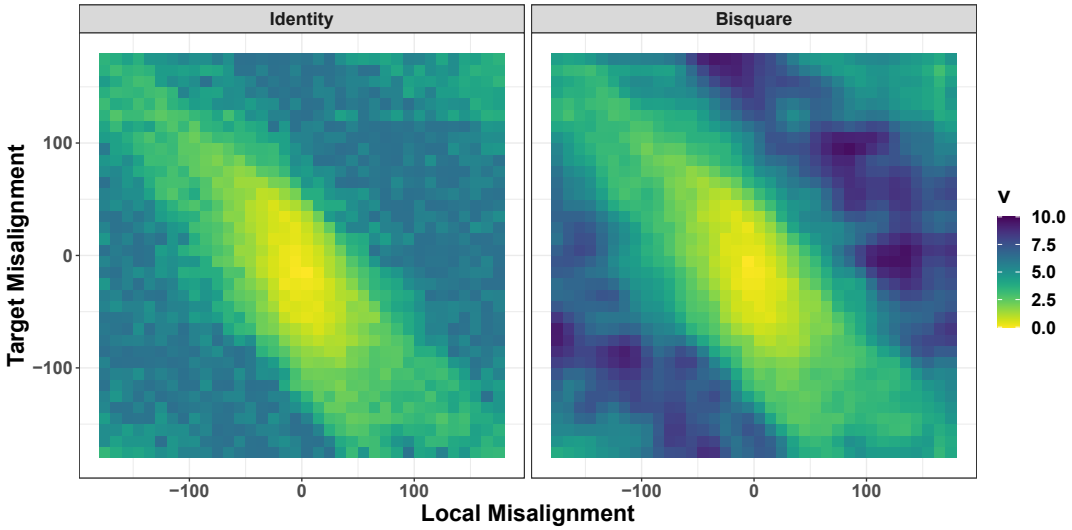

FIG 3. *Variational posterior mean costs-to-go for the guppy experiments for a gridded state space of target and local misalignment across two sets of features: full (identity matrix) and bisquare basis functions. The passive dynamics are assumed to be discrete uniform and the mean estimated costs have been shifted to have a minimum of 0. The yellow indicates states with lower costs-to-go and therefore states to which the guppies choose to transition.*

To assess the sensitivity to the assumed passive dynamics, we estimated the costs-to-go under a set of passive dynamics corresponding to an independent, normal random walk on the gridded state space with standard deviation 90°. The standard deviation was chosen to be large enough to ensure all non-zero transition probabilities used all of the observed data. The variational posterior mean and standard deviation for the costs-to-go are shown in Figure 4.

Comparing to the previously estimated states, the variational posterior mean cost-to-go functions are similar. The variational posterior standard deviations reflect the pattern of observed frequencies with states more frequently observed having smaller uncertainty.

In Figures 3 and 4, the diagonal pattern can be attributed to the corners appearing far when plotted in the 2-D plane, but are close together in circular space so they have similar costs-to-go.

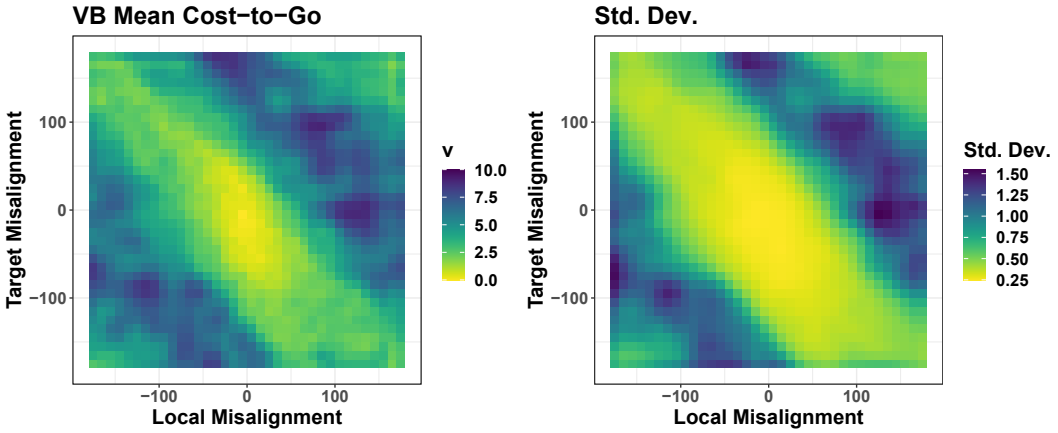

FIG 4. *Variational posterior mean costs-to-go (left panel) and standard deviations (right panel) for the guppy experiments for a gridded state space of target and local misalignment with passive dynamics assumed to be a normal random walk and bisquare basis functions. The mean estimated costs-to-go have been shifted to have a minimum of 0. The yellow indicates states with lower costs-to-go and therefore states to which the guppies choose to transition.*

The marginal costs-to-go based on the estimates in Figure 4 are shown in Figure 5, where the costs-to-go are calculated as the mean across all values of the other state variable and shifted to have a minimum of 0. There is evidence of collective alignment as shown in the local misalignment costs-to-go function due to the minimum cost occurring at 0° with gradual increase as misalignment increases in absolute value. Furthermore, the guppies appear to perceive local misalignments from -15° to 45° as equally optimal, which can be contrasted with the sharp dip in cost-to-go for 0° local misalignment in the SPP simulation Figure 1. The dip in the target misalignment cost-to-go function corresponds to the grid cells defined by -55° to -45° and -45° to -35°, suggesting it is less costly to approach the upper corner with the target 55° to 35° to the right. From inspection of the observed data shown in Figure 6, it appears many of the guppies moved across the tank to the left first, which would require a right turn to decrease the target misalignment. A symmetry constraint could be applied to the costs-to-go by considering the absolute target alignment if it were assumed to be equally costly to approach from the right or left.

5. Discussion. Collective motion from generative local interaction rules limit possible behavior, but the (L)MDP framework extends the definition of the agent to include perception and internal processes (Ried, Müller and Briegel, 2019). By estimating the state costs-to-go or value functions, system specific local rules can be estimated.

Our analysis of the captive guppy populations confirms previous works that find evidence of social interactions between individuals (Bode et al., 2012; Russell, Hanks and Haran,

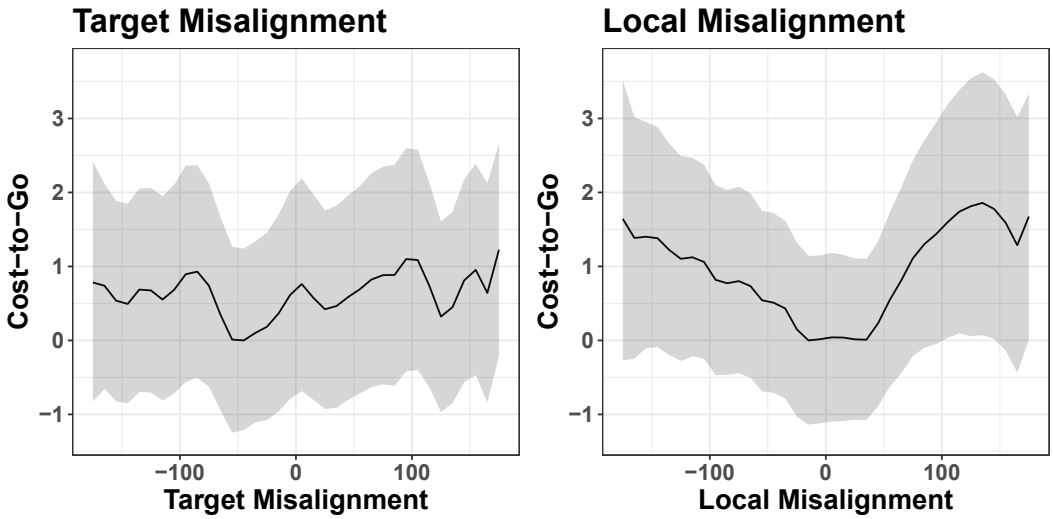

FIG 5. *Marginal costs-to-go of target and local misalignment for the guppy experiments for a gridded state space of target and local misalignment with passive dynamics assumed to be a normal random walk and bisquare basis functions. The mean estimated costs have been shifted to have a minimum of 0. The lower costs-to-go increase the probability of transitioning into that state.*

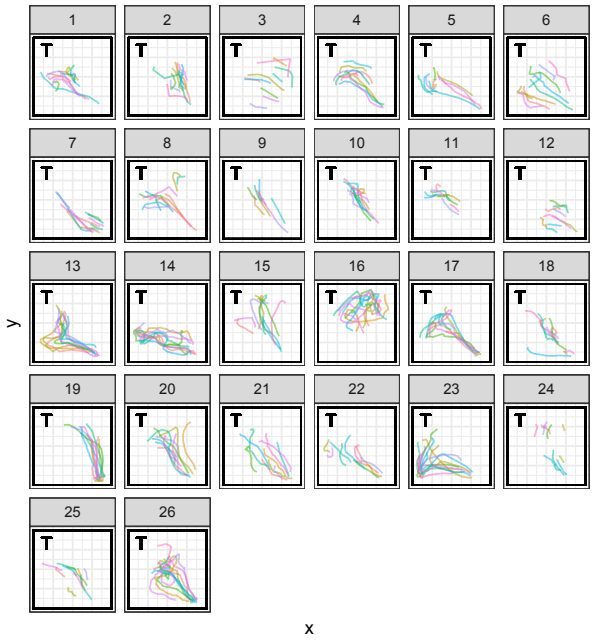

FIG 6. *Trajectories of all 26 experiments of groups of 10 guppies in a tank. The target is located at the point marked "T." The different colors represent the different individuals.*

2016; McDermott, Wikle and Millspaugh, 2017). However, instead of defining a set of behavioral rules *a priori*, we estimated the decision-making mechanisms. Our results suggested the captive guppies value collective movement more than targeted movement toward shelter. Furthermore, the behavioral mechanisms determined by the cost-to-go functions were non-linear and non-symmetric.

Inference is constrained to relative differences in costs-to-go. This is similar to the estimation of relative selection probabilities in animal resource selection modeling (Hooten et al., 2017, 2020) and therefore IRL can still provide useful inference. However, the SPP simulation demonstrated the ability to recover the magnitude of the true state costs despite the inability to estimate the true magnitude of the cost-to-go function.

It may be possible to improve the inference for the guppy data by relaxing the assumptions, estimating passive dynamics, and expanding the state space to include other features. We tested sensitivity of inference to choice of passive dynamics with two simple models. We did not detect a substantial difference, but for full quantification of uncertainty, joint estimation of passive dynamics could be considered. In future work, estimation of the passive dynamics parameters such as the random walk variance may be helpful. Additionally, the state space could include features based on physical distance to assess hypotheses about zonal collective movement which is a primary feature of collective movement ABMs (e.g., Couzin et al., 2002).

In the SPP simulation and guppy application, we assumed a discount factor of 1 which may be realistic for trajectories from such a short time frame. For observations spanning longer periods of time, it would be more realistic to assume there is some "forgetting" of past states which would correspond to a discount factor less than 1. Additionally, the discount factor can also be interpreted as the degree to which agents behave optimally (Choi and Kim, 2014). It might be expected that observations from animals in the wild are subject to more stochasticity than experimental settings and therefore do not always behave optimally.

APPENDIX A: LMDP NOTATION

The following is a table of LMDP notation used throughout the manuscript in order of appearance:

Symbol	Definition
S	Discrete state space with values $\{1, ..., J\}$ and observations are denoted as s
$\bar{\mathbf{P}}$	$J \times J$ passive transition probability matrix
$\bar{p}_{ij}$	An element of $\bar{\mathbf{P}}$; passive transition probability from state i to state j
γ	Discount factor in $[0, 1]$
R	State cost function with values denoted r_i for $i \in S$
$\mathbf{u}$	Continous controls which define the policy (1)
u_{ij}	An element of $\mathbf{u}$
$p_{ij}(u_{ij})$	Controlled transitions or policy defined by continuous controls and passive dynamics (1)
$p^*(s_t = j \vert s_i = i)$	Same as $p_{ij}(u_{ij})$
$l(\cdot, \mathbf{u})$	State and control cost function; it is the sum of the state cost R and KL divergence between passive and controlled transition probabilities (2)
$\mathbf{v}$	Cost-to-go function or the expected discounted future state control costs (3) with values denoted by v_i for $i \in S$

Acknowledgements. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1443129. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the National Science Foundation. CKW was supported by NSF Grant DMS-1811745; MBH was supported by NSF Grant DMS-1614392. Any use of

trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. government.

SUPPLEMENTARY MATERIAL

STAN Algorithms and Code.
(doi: COMPLETED BY THE TYPESETTER; .pdf). Definitions of the HMC, NUTS, and variational approximation algorithms and STAN model code.

REFERENCES

ARORA, S. and DOSHI, P. (2018). A survey of inverse reinforcement learning: Challenges, methods and progress. *arXiv preprint arXiv:1806.06877*.

BELLMAN, R. (1957). *Dynamic Programming*, 1 ed. Princeton University Press, Princeton, NJ, USA.

BODE, N. W., FRANKS, D. W., WOOD, A. J., PIERCY, J. J., CROFT, D. P. and CODLING, E. A. (2012). Distinguishing social from nonsocial navigation in moving animal groups. *The American Naturalist* **179** 621–632.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* **76**.

CHOI, J. and KIM, K.-E. (2011). Map inference for Bayesian inverse reinforcement learning. In *Advances in Neural Information Processing Systems* 1989–1997.

CHOI, J. and KIM, K.-E. (2014). Hierarchical Bayesian inverse reinforcement learning. *IEEE transactions on cybernetics* **45** 793–805.

COUZIN, I. D., KRAUSE, J., JAMES, R., RUXTON, G. D. and FRANKS, N. R. (2002). Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology* **218** 1–11.

DVIJOTHAM, K. and TODOROV, E. (2010). Inverse Optimal Control with Linearly-Solvable MDPs. In *ICML* 335-342.

FINN, C., LEVINE, S. and ABBEEL, P. (2016). Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning* 49–58.

HOFFMAN, M. D. and GELMAN, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15** 1593-1623.

HOOTEN, M. B., SCHARF, H. R. and MORALES, J. M. (2019). Running on empty: Recharge dynamics from animal movement data. *Ecology Letters* **22** 377-389.

HOOTEN, M., WIKLE, C. and SCHWOB, M. (2020). Statistical Implementations of Agent-Based Demographic Models. *International Statistical Review*.

HOOTEN, M. B., JOHNSON, D. S., MCCLINTOCK, B. T. and MORALES, J. M. (2017). *Animal Movement: Statistical Models for Telemetry Data*. CRC Press.

HOOTEN, M. B., LU, X., GARLICK, M. J. and POWELL, J. A. (2020). Animal movement models with mechanistic selection functions. *Spatial Statistics* **37** 100406. Frontiers in Spatial and Spatio-temporal Research.

JIN, M., DAMIANOU, A., ABBEEL, P. and SPANOS, C. (2017). Inverse reinforcement learning via deep Gaussian process. In *Conference on Uncertainty in Artificial Intelligence*.

KANGASRÄÄSIÖ, A. and KASKI, S. (2018). Inverse reinforcement learning from summary data. *Machine Learning* **107** 1517–1535.

KOHJIMA, M., MATSUBAYASHI, T. and SAWADA, H. (2017). Generalized Inverse Reinforcement Learning with Linearly Solvable MDP. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 373–388. Springer.

KUCUKELBIR, A., RANGANATH, R., GELMAN, A. and BLEI, D. (2015). Automatic variational inference in Stan. In *Advances in Neural Information Processing Systems* 568–576.

LEE, K., RUCKER, M., SCHERER, W. T., BELING, P. A., GERBER, M. S. and KANG, H. (2017). Agent-based model construction using inverse reinforcement learning. In *2017 Winter Simulation Conference (WSC)* 1264–1275. IEEE.

MCDERMOTT, P. L., WIKLE, C. K. and MILLSPAUGH, J. (2017). Hierarchical nonlinear spatio-temporal agent-based models for collective animal movement. *Journal of Agricultural, Biological and Environmental Statistics* **22** 294–312.

NG, A. Y. and RUSSELL, S. J. (2000). Algorithms for Inverse Reinforcement Learning. In *ICML* 663-670.

PINSLER, R., MAAG, M., ARENZ, O. and NEUMANN, G. (2018). Inverse Reinforcement Learning of Bird Flocking Behavior. *ICRA Swarms Workshop*.

RAMACHANDRAN, D. and AMIR, E. (2007). Bayesian Inverse Reinforcement Learning. In *IJCAI* **7** 2586–2591.

RATLIFF, N. D., BAGNELL, J. A. and ZINKEVICH, M. A. (2006). Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning* 729–736.

RIED, K., MÜLLER, T. and BRIEGEL, H. J. (2019). Modelling collective motion based on the principle of agency: General framework and the case of marching locusts. *PLOS ONE* **14**.

RUSSELL, J. C., HANKS, E. M. and HARAN, M. (2016). Dynamic models of animal movement with spatial point process interactions. *Journal of Agricultural, Biological and Environmental Statistics* **21** 22–40.

SCHARF, H. R., HOOTEN, M. B., FOSDICK, B. K., JOHNSON, D. S., LONDON, J. M. and DURBAN, J. W. (2016). Dynamic social networks based on movement. *Ann. Appl. Stat.* **10** 2182–2202.

SCHARF, H. R., HOOTEN, M. B., JOHNSON, D. S. and DURBAN, J. W. (2018). Process convolution approaches for modeling interacting trajectories. *Environmetrics* **29** e2487. e2487 env.2487.

SUTTON, R. S. and BARTO, A. G. (1998). *Introduction to reinforcement learning* **2**. MIT press Cambridge.

STAN DEVELOPMENT TEAM (2020). RStan: the R interface to Stan. R package version 2.19.3.

TODOROV, E. (2007). Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems* 1369–1376.

TODOROV, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences* **106** 11478–11483.

VICSEK, T., CZIRÓK, A., BEN-JACOB, E., COHEN, I. and SHOCHET, O. (1995). Novel type of phase transition in a system of self-driven particles. *Physical Review Letters* **75** 1226.

ŠOŠIĆ, A., ZOUBIR, A. M. and KOEPPL, H. (2017). A Bayesian approach to policy recognition and state representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40** 1295–1308.

ŠOŠIĆ, A., KHUDABUKHSH, W. R., ZOUBIR, A. M. and KOEPPL, H. (2017). Inverse Reinforcement Learning in Swarm Systems. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* 1413–1421.

WIKLE, C. K. and HOOTEN, M. B. (2016). Hierarchical agent-based spatio-temporal dynamic models for discrete-valued data. In *Handbook of Discrete-Valued Time Series* (R. A. Davis, S. H. Holan, R. Lund and N. Ravishanker, eds.) 349 – 366. Chapman & Hall/CRC Press, Boca Raton, FL.

WULFMEIER, M., ONDRUSKA, P. and POSNER, I. (2015). Deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*.

YAMAGUCHI, S., NAOKI, H., IKEDA, M., TSUKADA, Y., NAKANO, S., MORI, I. and ISHII, S. (2018). Identification of animal behavioral strategies by inverse reinforcement learning. *PLoS Computational Biology* **14** e1006122.

ZAMMIT-MANGION, A. (2020). FRK: Fixed Rank Kriging. R package version 0.2.2.1.

ZIEBART, B. D., MAAS, A., BAGNELL, J. A. and DEY, A. K. (2008). Maximum Entropy Inverse Reinforcement Learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence*. *AAAI'08* **3** 1433–1438. AAAI Press.