Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 56-2 (2023) 10120–10125

Stackelberg Meta-Learning Based Control for Guided Cooperative LQG Systems

Yuhan Zhao * **Quanyan Zhu** *

* *New York University, Brooklyn, NY 11201 USA*
(e-mail: {yhzhao,qz494}@nyu.edu)

Abstract: Guided cooperation allows intelligent agents with heterogeneous capabilities to work together by following a leader-follower type of interaction. However, the associated control problem becomes challenging when the leader agent does not have complete information about follower agents. There is a need for learning and adaptation of cooperation plans. To this end, we develop a meta-learning-based Stackelberg game-theoretic framework to address the challenges in the guided cooperative control for linear systems. We first formulate the guided cooperation between agents as a dynamic Stackelberg game and use the feedback Stackelberg equilibrium as the agent-wise cooperation strategy. We further leverage meta-learning to address the incomplete information of follower agents, where the leader agent learns a meta-response model from a prescribed set of followers offline and adapts to a new coming cooperation task with a small amount of learning data. We use a case study in robot teaming to corroborate the effectiveness of our framework. Comparison with other learning approaches also shows that our learned cooperation strategy provides better transferability for different cooperation tasks.

Keywords: Learning for control, Dynamic games, Linear systems, Intelligent robotics, Data-driven control, Adaptive control of multi-agent systems

1. INTRODUCTION

Cooperative control aims to address the collaboration between multiple intelligent agents and has become indispensable in modern system design to accomplish complex tasks (e.g., Wang et al. (2017); Chen and Zhu (2019); Liao et al. (2021); Papalia et al. (2022)). Particularly, guided cooperation is gaining increasing attention and popularity as we witness the rapid advances in the development of Artificial Intelligence (AI) aided technology in control systems. Guided cooperation allows intelligent agents with heterogeneous capabilities to work together and has a leader-follower or mentor-apprentice structure of interactions. A more resourceful agent (leader) can utilize its resources (e.g., sensing and computational resources) to provide strategic guidance to a less resourceful agent (follower) so that all agents can fully utilize their advantages to achieve the cooperation task objective. Guided cooperation is also broadly used in many applications such as multi-agent teaming (e.g., Choi et al. (2017); Hu et al. (2020)), human-robot interaction (e.g., Van Zoelen et al. (2020)), and collective transportation and manufacturing (e.g., Du et al. (2019); Fu et al. (2022)).

As a prevalent approach, game theory has been widely adopted for studying multi-agent interactions and cooperative control (see Marden et al. (2009)). Most game-theoretic literature leverages Nash games to develop cooperation plans, where all agents are homogeneous in interactions and decision-making (e.g., Mylvaganam et al. (2017); Yao et al. (2020); Zhu et al. (2020)). However, it fails to capture the discussed asymmetric structure in guided cooperation. Stackelberg games (e.g., Başar and Olsder (1998)) provide a suitable framework for quantifying the heterogeneous capabilities and leader-follower type of interactions in guided cooperation. The corresponding Stackelberg equilibrium solution can be used as an agent-wise optimal control strategy for guided cooperation. Some recent works have investigated asymmetric interactions and cooperation in multi-agent systems based on Stackelberg games. For example, Fisac et al. (2019) have used feedback Stackelberg games to develop driving strategies for autonomous vehicles to proactively assist the human driver to drive more safely and efficiently. Zhao et al. (2022) have studied the cooperation strategies for guided multi-robot rearrangement tasks based on stochastic Stackelberg games.

Although Stackelberg game-theoretic approaches capture the asymmetric interactions between the prescribed heterogeneous agents (agent-level heterogeneity), it is insufficient to address the following rising control challenges for guided cooperation. First, there is incomplete information about the agents. A leader (agent) may not know the exact model of the follower (agent), demanding learning-based approaches for game-theoretic control. Second, a leader often needs to work with different types of followers for heterogeneous tasks (also known as task-level heterogeneity). As the number of followers increases, designing distinct cooperation plans (using Stackelberg games) for hetero-

⋆ This work is partially supported by grants ECCS-1847056, CNS-2027884, and BCS-2122060 from National Science Foundation (NSF), grant 20-19829 from DOE-NE, and grant W911NF-19-1-0041 from Army Research Office (ARO).

10.1016/j.ifacol.2023.10.884

geneous followers demands a tractable and fast adaptive approach.

To address the challenges, we leverage meta-learning to enable learning a customizable plan from a prescribed set of tasks and fast adaptation to a new task with a small amount of learning data (see Hospedales et al. (2021)). Meta-learning has been used in many areas to seek adaptive cooperation plans, such as multi-agent systems by Jia et al. (2022), Internet of Things by Yue et al. (2022), and human robot-interaction by Gao et al. (2019). Some recent works have also focused on meta-learning-based control. For example, Harrison et al. (2018) have developed a meta-learning-based approach to stabilize the unseen dynamical system. Richards et al. (2021) have adopted meta-learning to learn a control policy and adapt to unknown environment noise for UAVs to achieve better trajectory tracking. Meta-learning also provides a suitable learning mechanism for Stackelberg game-theoretic approaches to address guided cooperation. The leader can learn a meta-knowledge of cooperative control strategies from experience. When a new cooperation task is initiated, the leader can transfer the meta-knowledge to fit the new task only using a small amount of interactive data.

In this work, we establish a Stackelberg meta-learning framework to enable guided cooperative control in linear systems and evaluate the framework using an application of robot teaming. Specifically, a leader guides different types of followers from different starting positions to reach the target destination and forms a team. The framework captures the guided interactions as a dynamic Stackelberg game and uses associated feedback Stackelberg equilibrium (FSE) as the agent-wise optimal control strategy for cooperation. When guiding heterogeneous followers to target destinations, the leader leverages meta-learning to learn a meta-response model for all foreseeable followers and adapts to a customized model for cooperative control when working with a specific follower. We use numerical experiments to corroborate that the proposed Stackelberg meta-learning framework not only enables promising guided control for diverse followers but also achieves a cost-efficient solution compared with other learning approaches and shows better transferability in the learned cooperation strategy compared with individual learning schemes.

2. PROBLEM FORMULATION

2.1 Stackelberg Games for Cooperative Control

We consider that a leader L (she) cooperates with a follower F (he) to complete a task driven by linear-Gaussian dynamics

$$x_{t+1} = Ax_t + B^L u_t^L + B^F u_t^F + w_t, \tag{1}$$

where $x_t \in \mathbb{R}^n$ and $u_t^L \in \mathbb{R}^{r^L}$ represents the system state, and the leader's control input at time t, $A \in \mathbb{R}^{n\times n}$ and $B^L \in \mathbb{R}^{n\times r^L}$ are state transition matrix and the leader's control input matrix, respectively. Likewise, $u_t^F \in R^{r^F}$ and $B^F \in \mathbb{R}^{n\times r^F}$ are the follower's control and the control input matrix, respectively. Here, $w_t \in \mathbb{R}^n$ are i.i.d. process noise with Gaussian distribution $\mathcal{N}(0, \mathbf{\Sigma})$.

Followers have heterogeneous characteristics, distinguished by their type $\theta \in \Theta$. The leader works with one follower at a time to achieve the control task. We assume that the leader does not know the exact type of the follower except for a type distribution p over Θ, where $p(\theta)$ represents the probability that the leader cooperates with a follower with type θ.

The cooperative interactions are strategic. The leader's goal is to minimize the guidance cost J_θ^L over time horizon T by finding an optimal control trajectory $\mathbf{u}^{L*} := \{u_t^{L*}\}_{t=0}^{T-1}$. The less resourceful followers are assumed to be myopic and only minimize one-step cost J_θ^F after observing the current state x_t and the leader's control u_t. This asymmetric interaction in the cooperation can be captured by a dynamic Stackelberg game $\mathcal{G}_\theta$ as follows:

$$\begin{aligned} \min_{\mathbf{u}^L} \quad & J_\theta^L(\mathbf{u}^L) \\ & := \mathbb{E}\left[\sum_{t=0}^{T} x_t^\mathsf{T} Q^L x_t + {u_t^L}^\mathsf{T} R^L u_t^L + x_t^\mathsf{T} Q_f^L x_T\right], \quad (2) \\ \text{s.t.} \quad & x_{t+1} = Ax_t + B^L u_t^L + B_\theta^F u_{\theta,t}^{F*}(x_t, u_t^L) + w_t, \quad (3) \\ & \qquad t = 0, \dots, T-1, \\ & u_{\theta,t}^{F*}(x_t, u_t^L) = \arg\min_{u^F} \quad J_\theta^F(u^F; x_t, u_t^L) \quad (4) \\ & := \mathbb{E}[x_{t+1}^\mathsf{T} Q_\theta^F x_{t+1} + {u^F}^\mathsf{T} R_\theta^F u^F], \ t = 0, \dots, T-1. \end{aligned}$$

Here, $Q^L \succeq 0, Q_f^L \succeq 0, R^L \succ 0, Q_\theta^F \succeq 0, R_\theta^F \succ 0$ are the leader and the follower's cost matrices with proper dimensions. $B_\theta^F \in \mathbb{R}^{n\times r^F}$ is the type-specific control input matrix. The follower's problem is captured by (4), where x_{t+1} in (4) represents the follower's one-step prediction and evolves according to the dynamics (1) after the follower observes x_t and u_t^L. The leader anticipates the follower's response $u_{\theta,t}^{F*}(x_t, u_t^L)$ and uses it for long-term planning. We adopt the FSE $\langle \mathbf{u}^{L*}(\mathbf{x}), \mathbf{u}_\theta^{F*}(\mathbf{x})\rangle$ of the game $\mathcal{G}_\theta$ as the agent-wise cooperation plan for the leader and the follower. Note that the equilibrium strategies are the functions of the state. The leader and follower can generate controls based on the observed state.

2.2 Meta Response and Meta-learning Objectives

Model-based methods such as dynamic programming can find the FSE of $\mathcal{G}_\theta$ if the leader knows the follower's exact decision-making model (i.e., J_θ^F). However, this information may not be known to the leader. Therefore, the leader needs learning-based approaches to first learn the follower's behavior model and then compute the FSE to find the cooperation plan.

The follower's behavior model can be estimated in various ways, including neural networks. From (4), we can obtain

$$u_{\theta,t}^{F*}(x_t, u_t^L) = -({B_\theta^F}^\mathsf{T} Q_\theta^F B_\theta^F + R_\theta^F)^{-1} {B_\theta^F}^\mathsf{T} Q_\theta^F (Ax_t + B^L u_t^L) \tag{5}$$

for $t = 0, \dots, T-1$. It shows the follower's optimal response has a linear structure in x_t and u_t^L under a quadratic cost. We leverage this linear structure and use a matrix parameter $M \in \mathbb{R}^{r^F\times n}$ to estimate the follower's response with $r_\theta : \mathbb{R}^n \times \mathbb{R}^{r^L} \to \mathbb{R}^{r^F}$, defined by

$$r_\theta(x_t, u_t^L; M) = M(Ax_t + B^L u_t^L). \tag{6}$$

We substitute the follower's problem (4) with (6). Then, $\mathcal{G}_\theta$ becomes a new Stackelberg game, denoted as $\widetilde{\mathcal{G}}_\theta(M)$, where the leader does not know the follower's model but assumes that the follower uses (6) as his response parameterized M. Then, the leader uses the FSE $\langle \tilde{\mathbf{u}}^{L^*}(M), \mathbf{r}^*_\theta(M)\rangle$ of $\widetilde{\mathcal{G}}_\theta(M)$, where $\mathbf{r}^*_\theta(M) := \{r_\theta(x_t, \tilde{u}^{L*}_t; M)\}_{t=0}^{T-1}$, to approximate the one of $\mathcal{G}_\theta$. The leader's optimal guidance cost in $\widetilde{\mathcal{G}}_\theta(M)$ is denoted as $\widetilde{J}^{L*}(M)$.

The leader faces different game problems when cooperating with heterogeneous followers. It can be cost-prohibitive for the leader to estimate each follower's response model and compute FSEs. Meta-learning provides a learning mechanism to learn from a set of followers and fast adapt to a specific individual follower to achieve cooperative control. Specifically, the leader learns a meta-response from encountered followers as the meta-knowledge of followers' behavior. When working with a new follower, the leader only needs small learning data to adapt the meta-response to the follower-specific one and uses the adapted response model to compute cooperative strategies, i.e., the FSE.

With a slight abuse of notation, we use M as the meta parameter and use (6) as the meta-response model for all $\theta \in \Theta$. We refer to the guided cooperation between the leader and the follower with type θ as task $\mathcal{T}_\theta$ to align with the meta-learning context. A meta-response model should approximate the follower's real behavior (optimal response) and reduce the leader's guidance cost. The latter objective can be quantified by the leader's optimal cost function $\widetilde{J}^{L*}(M)$. The former can be achieved by minimizing the response data fitting cost. Let $\mathcal{D}_\theta = \{\hat{x}_i, \hat{u}^L_i, \hat{u}^{F*}_i\}_{i=1}^D$ be a best-response data set of D samples collected from the follower with type θ. The data fitting cost is given by

$$Q_\theta(M) = \frac{1}{N}\sum_{i=1}^N \left\| M(A\hat{x}_i + B^L\hat{u}^L_i) - \hat{u}^{F*}_i \right\|_2^2 .$$

We define the meta-learning objective for the task $\mathcal{T}_\theta$ as

$$L_\theta(M) = \widetilde{J}^{L*}_\theta(M) + \gamma Q_\theta(M), \tag{7}$$

where $\gamma > 0$ is the weighting parameter.

Interpretation on γ The weighting parameter γ represents how the leader values the follower's response data in learning the follower's behavior model. When $\gamma = 0$, the leader shows zero interest in the follower's real behavior. She only seeks a unilaterally optimal model, which helps minimize her guidance cost. The learned response model can differ significantly from the follower's real response. When $\gamma \to \infty$, the leader aims to learn the follower's real response as precisely as possible. The approximation accuracy becomes the exclusive objective in meta-learning. Therefore, γ provides flexibility in balancing different meta-learning criteria.

2.3 Meta-Learning as Bilevel Optimization Problems

The leader uses meta-learning to gain meta-knowledge of followers' behavior and trains an adapted response model for the new coming follower for cooperation. We split the data $\mathcal{D}_\theta = \mathcal{D}^{train}_\theta \cup \mathcal{D}^{test}_\theta$ following typical learning settings and formulate the meta-learning problem as a bilevel optimization problem (see Rajeswaran et al. (2019)):

$$\min_M \quad \mathbb{E}_{\theta\sim p}[L_\theta(Z^*_\theta(M); \mathcal{D}^{test}_\theta)] \tag{8}$$

with

$$Z^*_\theta(M) = \arg\min_Z L_\theta(Z; \mathcal{D}^{train}_\theta) + \lambda \|Z - M\|_F^2 , \tag{9}$$

where $\lambda > 0$ is the weighting parameter. The inner-level problem (9) learns a task-specific optimizer on the training data $\mathcal{D}^{train}_\theta$. The outer-level problem (8) improves the generalized performance of the meta parameter on all sampled tasks with data $\mathcal{D}^{test}_\theta$.

3. STACKELBERG META-LEARNING

3.1 Parametric Optimal Control

We take the linear meta-response (6) into the dynamics (3). Then the Stackelberg game $\widetilde{\mathcal{G}}_\theta(M)$ becomes a single-agent linear-quadratic-Gaussian (LQG) control problem:

$$\begin{aligned} \min_{\mathbf{u}^L} \quad & \widetilde{J}^L(\mathbf{u}^L) = \mathbb{E}\left[\sum_{t=0}^T x_t^\top Q^L x_t + {u^L_t}^\top R^L u^L_t + x_t^\top Q^L_f x_T\right], \\ \text{s.t.} \quad & x_{t+1} = \widetilde{A}x_t + \widetilde{B}^L u^L_t + w_t, \quad , t = 0, \dots, T-1, \end{aligned} \tag{10}$$

where $\widetilde{A} := A + B^F_\theta MA$ and $\widetilde{B}^L := B^L + B^F_\theta MB^L$. Given a meta parameter M, we can evaluate the leader's optimal guidance cost $\widetilde{J}^{L*}(M)$ and the feedback control law $\tilde{\mathbf{u}}^{L*}(M)$ by solving the discrete Riccati equation

$$\begin{aligned} P_t = Q^L + & \widetilde{A}^\top P_{t+1}\widetilde{A} \\ & - \widetilde{A}^\top P_{t+1}\widetilde{B}(R^L + \widetilde{B}^\top P_{t+1}\widetilde{B})^{-1}\widetilde{B}^\top P_{t+1}\widetilde{A} \end{aligned} \tag{11}$$

for $t = 0, \dots, T-1$ with $P_T = Q^L_f$. The feedback control $u^{L*}_t = -K_t x_t$ where $K_t := (R^L + \widetilde{B}^\top P_{t+1}\widetilde{B})^{-1}\widetilde{B}^\top P_{t+1}\widetilde{A}$. The optimal guidance cost $\widetilde{J}^{L*}(M) = x_0^\top P_0 x_0 + \text{res}_0$, where $\text{res}_t = \sum_{j=t+1}^T \text{tr}(\mathbf{\Sigma} P_{t+1})$ and $\text{res}_T = 0$.

3.2 Meta-Response Training

Solving the inner-level problem (9) requires optimizing the parameter M over the parameterized cost $\widetilde{J}^{L*}(M)$. We have the following proposition to characterize the property of $\widetilde{J}^{L*}(M)$.

Proposition 1. With the parametrization of $\widetilde{A}$ and $\widetilde{B}$ in (10), the parameterized cost $\widetilde{J}^{L*}(M)$ is a rational polynomial of entries of M.

Proof. See Appendix A in report (Zhao and Zhu, 2022).

Therefore, $\widetilde{J}^{L*}(M)$ is continuously differentiable in the entries of M, and we can develop gradient methods to solve the inner-level problem. To evaluate $\frac{\partial \widetilde{J}^{L*}}{\partial M}$, we note that the matrix P_t, $t = 0, \dots, T-1$, is also parameterized by M. Thus, we leverage the Riccati equation (11) to evaluate $\frac{\partial P_t}{\partial M}$ backward from $t = T-1, \dots 0$ with $\frac{\partial P_T}{\partial M} = 0$. $\frac{\partial\, \text{res}_t}{\partial M}$ can be evaluated similarly. The convergence of gradient methods on the inner-level problem (9) is guaranteed because the objective is continuously differentiable in M and is lower bounded by 0 (see Bertsekas (1997)). The weight λ can be used to convexify the inner-level problem and help search for local minimizers.

We use empirical value to approximate the expectation in the outer-level problem (8) and obtain

$$\min_{M} \quad \frac{1}{|\mathcal{T}_{batch}|} \sum_{\theta \sim p} L_\theta(Z^*(M); \mathcal{D}_\theta^{test}). \tag{12}$$

Here, $\theta \sim p$ represents the empirical task distribution of sampled batch tasks $\mathcal{T}_{batch} := \{\mathcal{T}_\theta\}$ from p. Following similar computations, we use gradient methods to solve the outer-level problem and find a meta-response model. The iteration follows

$$M_{k+1} \leftarrow M_k - \frac{\beta}{|\mathcal{T}_{batch}|} \sum_{\mathcal{T}_\theta \in \mathcal{T}_{batch}} \frac{\partial}{\partial M} L_\theta(Z_\theta^*; \mathcal{D}_\theta^{test}), \tag{13}$$

where $\beta > 0$ is the meta-learning step. We summarize the Stackelberg meta-learning algorithm for cooperative control in Alg. 1, which outputs a meta-response model.

Algorithm 1 Stackelberg Meta-learning algorithm.

Require: Step α, β; weight γ, λ; type distribution $p(\theta)$;
Require: Initial mete parameter M_0, initial state x_0;
$k \leftarrow 0$;
while $k <$ MAX_ITER **do**
  Sample a batch of tasks $\mathcal{T}_{batch} := \{\mathcal{T}_\theta\} \sim p$;
  ▷ Inner-level problem gradient evaluation
  for all task $\mathcal{T}_\theta \in \mathcal{T}_{batch}$ **do**
    iter $\leftarrow 0$; $Z_\theta \leftarrow M_k$;
    while True **do**
      $\frac{\partial P_t}{\partial M}, \frac{\partial \mathrm{res}_t}{\partial M} \leftarrow$ based on (11) and Z_θ $\forall t$;
      $\frac{\partial \widetilde{J}^{L*}}{\partial M}\big|_{Z_\theta} \leftarrow \frac{\partial}{\partial M} x_0^\top P_0 x_0 + \frac{\partial \mathrm{res}_0}{\partial M}$;
      $\widetilde{\mathbf{x}}(Z_\theta), \widetilde{\mathbf{u}}^{L*}(Z_\theta) \leftarrow$ simulate trajectory;
      Randomly sample N_1 data;
      Sample N_2 data around $\widetilde{\mathbf{x}}(Z_\theta), \widetilde{\mathbf{u}}^{L*}(Z_\theta)$;
      $\mathcal{D}_\theta^{train} \leftarrow$ all samples with $N = N_1 + N_2$;
      $g \leftarrow \frac{\partial}{\partial M} L_\theta(Z_\theta; \mathcal{D}_\theta^{train}) + 2\lambda(Z_\theta - M_k)$;
      $Z_\theta \leftarrow Z_\theta - \alpha g$;
      if iter $>$ MAX_GD or $\|g\| < \epsilon$ **then**
        $Z_\theta^* \leftarrow Z_\theta$; break;
      end if
      iter $\leftarrow$ iter $+1$;
    end while
  end for
  ▷ Outer-level problem gradient evaluation
  for all task $\mathcal{T}_\theta \in \mathcal{T}_{batch}$ **do**
    $\frac{\partial P_t}{\partial M}, \frac{\partial \mathrm{res}_t}{\partial M} \leftarrow$ based on (11) and Z_θ^* $\forall t$;
    Sample $\mathcal{D}_\theta^{test}$ (same sampling rule as $\mathcal{D}_\theta^{train}$);
    Compute $\frac{\partial}{\partial M} L_\theta(Z_\theta^*; \mathcal{D}_\theta^{test})$;
  end for
  Update M_{k+1} by (13);
  $k \leftarrow k + 1$;
end while
return $M_{meta} \leftarrow M_k$;

Sampling Follower's Response Data We note that the inner-level problem (9) shows that the meta parameter is updated within a small neighborhood of the original one due to the regularization term. Hence, the updated leader's trajectory will likely stay near the previous one. The samples near the trajectory can better help the leader refine the follower's response model near the trajectory and hence make a better update. This sampling technique is more useful when the leader uses a nonlinear response mode such as neural networks to estimate the follower's behavior. We set $\kappa := N_2/N_1$ to control the sample ratio in $\mathcal{D}_\theta^{train}$ and $\mathcal{D}_\theta^{test}$.

3.3 Response Adaptation

Using the meta parameter M_{meta} and the meta-response model from Alg. 1, the leader can fast adapt to a new coming follower using a small amount of data samples. Specifically, the leader samples a type-specific data set $\mathcal{D}'_\theta$ using M_{meta} when she starts cooperating with a follower with type θ. Then she customizes a response parameter M_θ^* from M_{meta} to adapt to the follower by solving

$$M_\theta^* = \arg\min_M L_\theta(M_{meta}; \mathcal{D}'_\theta) + \eta \|M - M_{meta}\|_F^2, \tag{14}$$

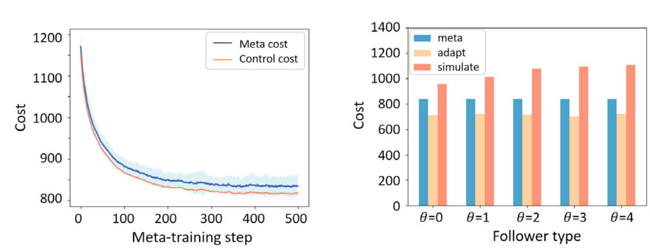
where $\eta \geq 0$ is the regularization weight. In practice, we can select $\eta = \lambda$.

4. EXPERIMENTS AND EVALUATIONS

In this section, we demonstrate our Stackelberg meta-learning framework using a case study in cooperative robot teaming, where a leader robot guides the follower robot to a target destination to form a team. Let $x^L = [p^L, v^L] \in \mathbb{R}^4, u^L \in \mathbb{R}^2$ ($x^F = [p^F, v^F] \in \mathbb{R}^4$, $u^F \in \mathbb{R}^2$) be the leader's (the follower's) position, velocity, and control input. The joint state $x := [x^L, x^F]$. We assume the leader and the follower have a double integrator dynamics, where $\ddot{p}^L = u^L$ and $\ddot{p}^F = u^F$. The corresponding discrete dynamical systems are obtained by setting a discretization time $\mathrm{d}t = 0.5$. We set the control time horizon $T = 10$ and the target destination $x^{\mathrm{d}} = 0$. $w_t \in \mathbb{R}^8$ are i.i.d. Gaussian noise $\sim \mathcal{N}(0, 0.5I)$. We consider five types of followers with a type distribution $p = [0.2,\ 0.3,\ 0.1,\ 0.2,\ 0.2]$.

4.1 Meta-learning Results

We set $\kappa = 2$ and use $N = 6$ response data in each iteration to perform meta-training. The hyperparameters are set by $\gamma = 5, \lambda = 100$. The training process is evaluated by the empirical meta-cost used in (12). We conduct 20 simulations with a randomly generated initial guess M_0 and plot the mean-variance training result in Fig. 1.

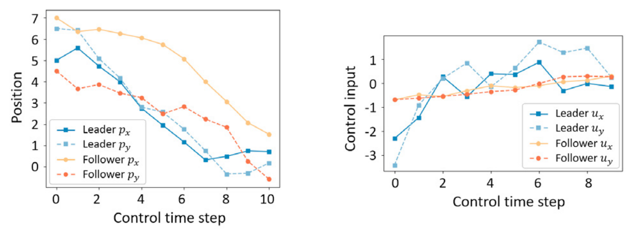


(a) Meta-training process. (b) Adaptation comparison.

Fig. 1. Meta training and adaptation results.

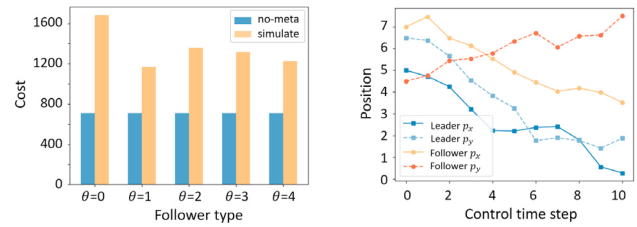
Fig. 1a shows that the meta-learning algorithm reduces the meta-cost and converges to a local minimum. The mean value of the leader's optimal guidance cost $\widetilde{J}^{L*}$ (orange line) is also reduced as the meta-training proceeds, which means that the meta-response model becomes more efficient for the leader to perform the guidance. The variance comes from different sampled response data in each simulation to train the meta model.

The adapted results for different types of followers are shown in Fig. 1b. The blue bar represents the leader's



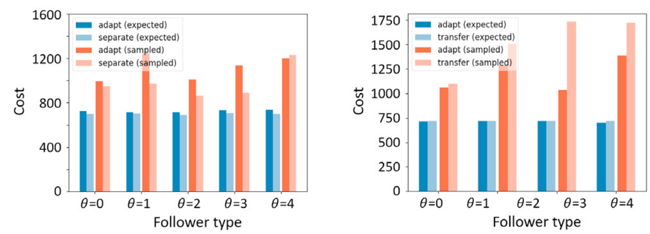
(a) Position trajectories. (b) Control trajectories.

Fig. 2. Trajectories for $\theta = 0$ follower after adaptation.



(a) Leader's guidance cost using unilateral learning. (b) Trajectories using unilateral learning for type $\theta = 2$ follower.

Fig. 3. Costs and trajectories for unilateral learning.



(a) Comparison with meta-learning. (b) Adapting $\theta = 4$ follower's model to the rest.

Fig. 4. Results for individual learning.

expected guidance cost $\widetilde{J}^{L*}$ using the meta-response model before the adaptation, serving as a baseline. The yellow bar shows the expected guidance cost $\widetilde{J}^{L*}$ using the adapted response model for different types of followers, respectively. As expected, the adapted model provides a lower guidance cost for the leader than the baseline.

Due to the process noise and the estimation error, the follower's real behavior can deviate from the leader's expectation. We simulate interactive trajectories to view the real cooperation performance, where the leader uses the adapted response model to design control strategies, and the follower uses his true model (4) to respond. The leader's simulated guidance costs for each type of follower are shown by the red bar in Fig. 1b. We observe that the simulated costs are higher but not significantly greater than the expected costs. It shows that the adapted response model and the resulting control strategies can provide satisfactory results in guidance tasks. For simplicity, we plot the position and control trajectories for the leader and the follower with type $\theta = 0$ in Fig. 2. The leader and the follower start from $[5, 6.5]$ and $[7, 4.5]$, respectively. From Fig. 2a, the leader can design effective cooperative strategies using the adapted model to guide the follower to approach the zero state. Their controls also approach 0 by the end of the guidance. The trajectory convergence direction shows that the guidance is effective.

4.2 Comparison with Unilateral Learning

The unilateral-learning approach refers to the leader learning a response model based solely on the guidance cost instead of the follower's real response. It is equivalent to set $\gamma = 0$ in (7). Since followers' response is not involved, the learned response models are the same for all followers. Therefore, the learning is fast and has an average training time of less than 1 min, which contrasts sharply with the meta-learning approach, with an average time of 32 min.

We evaluate the leader's expected (blue) and simulated (yellow) guidance cost using the model obtained from the unilateral-learning approach and compare them in Fig. 3a. The simulated cost significantly deviates from the expected one, indicating that the learned model is less effective in the guidance task. Besides, the expected and simulated costs are also greater than the counterparts in meta-learning (see Fig. 1b), showing the adapted models outperform the unilaterally learned ones. For simplicity, we show the simulated trajectory for the leader and the follower with type $\theta = 2$ in Fig 3b. The unilateral-learning approach fails to guide the follower to the origin. Instead, the follower moves in the opposite direction, resulting in a failure in the guidance task. Although unilateral learning saves considerable training time compared with the meta-learning approach, it can significantly sacrifice the model accuracy and the guidance performance.

4.3 Individual Learning and Transferability

The individual-learning approach refers to the leader learning separate response models for every follower and generating different guidance strategies. We evaluate the leader's expected (blue bars) and simulated (red bars) guidance cost in Fig. 4a. We also plot our meta-learning result with dark colors for comparison. It is not surprising that individual learning provides slightly smaller guidance costs compared with meta-learning because it trains designed models for different followers.

However, meta-learning can adapt the meta-response model to a specific follower and provide good guidance. Individual learning does not have such flexibility and transferability. To see this, we adapt the learned model of the follower $\theta = 4$ obtained by the individual-learning approach to other followers by following adaptation rule (14). We evaluate the leader's expected and simulated guidance cost in Fig. 4b. We can observe that meta-learning provides smaller expected and simulated guidance costs, showing that adapted models are more efficient for the leader in performing guidance control. The smaller simulated costs indicate that the meta-learning approach outperforms individual learning in real guidance tasks. Besides, individual learning requires considerable learning resources, especially when there are many followers. The meta-learning approach can leverage its better transferability in the learned response model, providing a more flexible and faster adaptation for different guidance tasks.

5. CONCLUSION

We have proposed a Stackelberg meta-learning framework for guided cooperative control in LQG systems. Our framework not only captures the leader-follower type of interactions in guided cooperation but also provides a learning mechanism to adapt to different guided control tasks. The case study in robot teaming application has demonstrated that the framework successfully provides effective and transferable guidance control strategies to accomplish different guidance tasks. As we have observed in the simulation, although a learned cooperation strategy can guide the follower toward the destination, it cannot perform as precisely as deterministic control. How to guarantee the control performance within an allowable range would be a valuable future research direction. For other future work, we would generalize our framework to more general control systems and investigate analytic properties such as optimality conditions and sample complexity.

REFERENCES

Başar, T. and Olsder, G.J. (1998). *Dynamic noncooperative game theory*. SIAM.

Bertsekas, D.P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3), 334–334.

Chen, J. and Zhu, Q. (2019). Control of multilayer mobile autonomous systems in adversarial environments: A games-in-games approach. *IEEE Transactions on Control of Network Systems*, 7(3), 1056–1068.

Choi, T., Pavlic, T.P., and Richa, A.W. (2017). Automated synthesis of scalable algorithms for inferring nonlocal properties to assist in multi-robot teaming. In *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*, 1522–1527. IEEE.

Du, G., Xia, Y., Jiao, R.J., and Liu, X. (2019). Leader-follower joint optimization problems in product family design. *Journal of Intelligent Manufacturing*, 30(3), 1387–1405.

Fisac, J.F., Bronstein, E., Stefansson, E., Sadigh, D., Sastry, S.S., and Dragan, A.D. (2019). Hierarchical game-theoretic planning for autonomous vehicles. In *2019 International conference on robotics and automation (ICRA)*, 9590–9596. IEEE.

Fu, X., Wang, D., Hu, J., Wei, J., and Yan, C.B. (2022). Leader-follower based two-agv cooperative transportation system in 5g environment. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, 67–72. IEEE.

Gao, Y., Sibirtseva, E., Castellano, G., and Kragic, D. (2019). Fast adaptation with meta-reinforcement learning for trust modelling in human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 305–312. IEEE.

Harrison, J., Sharma, A., Calandra, R., and Pavone, M. (2018). Control adaptation via meta-learning dynamics. In *Workshop on Meta-Learning at NeurIPS*, volume 2018.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9), 5149–5169.

Hu, J., Bhowmick, P., Arvin, F., Lanzon, A., and Lennox, B. (2020). Cooperative control of heterogeneous connected vehicle platoons: An adaptive leader-following approach. *IEEE Robotics and Automation Letters*, 5(2), 977–984.

Jia, H., Zhao, Y., Zhai, Y., Ding, B., Wang, H., and Wu, Q. (2022). Crmrl: Collaborative relationship meta reinforcement learning for effectively adapting to type changes in multi-robotic system. *IEEE Robotics and Automation Letters*, 7(4), 11362–11369.

Liao, J., Liu, C., and Liu, H.H. (2021). Model predictive control for cooperative hunting in obstacle rich and dynamic environments. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 5089–5095. IEEE.

Marden, J.R., Arslan, G., and Shamma, J.S. (2009). Cooperative control and potential games. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(6), 1393–1407.

Mylvaganam, T., Sassano, M., and Astolfi, A. (2017). A differential game approach to multi-agent collision avoidance. *IEEE Transactions on Automatic Control*, 62(8), 4229–4235.

Papalia, A., Thumma, N., and Leonard, J. (2022). Prioritized planning for cooperative range-only localization in multi-robot networks. In *2022 International Conference on Robotics and Automation (ICRA)*, 10753–10759. IEEE.

Rajeswaran, A., Finn, C., Kakade, S.M., and Levine, S. (2019). Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32.

Richards, S.M., Azizan, N., Slotine, J.J., and Pavone, M. (2021). Adaptive-Control-Oriented Meta-Learning for Nonlinear Systems. In *Proceedings of Robotics: Science and Systems*. Virtual.

Van Zoelen, E.M., Barakova, E.I., and Rauterberg, M. (2020). Adaptive leader-follower behavior in human-robot collaboration. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1259–1265. IEEE.

Wang, Y., Garcia, E., Casbeer, D., and Zhang, F. (eds.) (2017). *Cooperative control of multi-agent systems: Theory and applications*. John Wiley & Sons.

Yao, K., Wang, J., Xu, Y., Xu, Y., Yang, Y., Zhang, Y., Jiang, H., and Yao, J. (2020). Self-organizing slot access for neighboring cooperation in uav swarms. *IEEE Transactions on Wireless Communications*, 19(4), 2800–2812.

Yue, S., Ren, J., Xin, J., Zhang, D., Zhang, Y., and Zhuang, W. (2022). Efficient federated meta-learning over multi-access wireless networks. *IEEE Journal on Selected Areas in Communications*, 40(5), 1556–1570.

Zhao, Y., Huang, B., Yu, J., and Zhu, Q. (2022). Stackelberg strategic guidance for heterogeneous robots collaboration. In *2022 International Conference on Robotics and Automation (ICRA)*, 4922–4928.

Zhao, Y. and Zhu, Q. (2022). Stackelberg meta-learning based control for guided cooperative lqg systems. *arXiv preprint arXiv:2211.06512*.

Zhu, F., Yang, Z., Lin, F., and Xin, Y. (2020). Decentralized cooperative control of multiple energy storage systems in urban railway based on multiagent deep reinforcement learning. *IEEE Transactions on Power Electronics*, 35(9), 9368–9379.