MERMADE: K-SHOT ROBUST ADAPTIVE MECHANISM DESIGN VIA MODEL-BASED META-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanism design (MD) studies how rules and rewards shape the behavior of intelligent agents, e.g., in auctions or the economy. Simulations with AI agents are powerful tools for MD, but real-world agents may behave and learn differently than simulated agents under a given mechanism. Also, the mechanism designer may not fully observe an agent's learning strategy or rewards, and executing a mechanism may be costly, e.g., enforcing a tax might require extra labor. Hence, it is key to design robust adaptive mechanisms that generalize well to agents with unseen (learning) behavior, are few-shot adaptable, and are cost-efficient. Here, we introduce MERMADE, a model-based meta-learning framework to learn mechanisms that can quickly adapt when facing out-of-distribution agents with different learning strategies and reward functions. First, we show that meta-learning allows adapting to the theoretically known and appropriate Stackelberg equilibrium in a simple matrix game at meta-test time, with few interactions with the agent. Second, with bandit agents, we show empirically that our approach yields strong meta-test time performance against agents with various unseen explore-exploit behaviors. Finally, we outperform baselines that separately use either meta-learning or agent behavior modeling to learn a cost-effective mechanism that is K-shot adaptable with only partial information about the agents.

1 INTRODUCTION

Mechanism design studies how to design mathematical games, i.e., reward functions and environment rules, that are experienced by a set of intelligent agents, including real people. It has applications across many domains, e.g., maximizing revenue in auctions (Milgrom & Milgrom, 2004), optimizing social welfare with economic policy (Zheng et al., 2022), or optimizing skill acquisition in personalized education (Maghsudi et al., 2021). For instance, a government may want to incentivize consumers to use environmentally-friendly ("clean") products, but also may want to limit market distortion. Hence, it may adjust prices by taxing "dirty" or subsidizing "clean" products, but needs to understand how people (strategically) change their consumption behavior. Here, existing models of human behavior that assume rationality or use simplified models of bounded rationality often do not suffice, and interacting with the agents is required to *learn* their behavior. However, executing mechanisms is not "free", e.g., a tax or subsidy policy may require effort to apply it fairly and to measure its impact on both consumers and producers.

To mitigate the need for costly real-world interactions, we can often use simulations, which we can couple with deep RL methods. In particular, simulations can use artificial agents to emulate real-world agents, and can be run safely and as often as needed. Hence, we can use deep RL to design mechanisms that are effective even in the face of complex agent behaviors and *sequential* dynamic environments. However, it also faces several challenges. When deployed, mechanisms can typically interact only a few times and rarely be repeated under identical circumstances; in contrast to simulations, we cannot reset the real world. Agents may learn and adapt to the mechanism, but the mechanism designer (or *planner*) typically does not know an agent's rewards or its learning strategy.

To address these challenges, here we study the problem of designing a learning-based method that yields mechanisms that 1) perform well even when agents learn, 2) can be quickly adapted, 3) are robust to distribution shifts, and 4) are effective despite having only *partial information*.

Contributions. We propose MERMADE, a deep RL approach to mechanism design that 1) learns a world model (i.e., is *model-based*) and 2) uses gradient-based meta-learning to learn a mechanism that can be quickly adapted to perform well on unseen test agents. In MERMADE, the mechanism designer intervenes *at a cost* to incentivize agents to (learn to) achieve the designer's objective. We assume that the agent behaves in a first-order strategic manner and the mechanism designer in a second-order strategic manner. Here, the agents optimize their experienced rewards, but do not account for their influence on the mechanism designer's actions. In contrast, the mechanism designer intervenes explicitly as to influence the agent's actions. In order to focus on the challenges of mechanism design, we study a planner interacting *with only one agent at a time*.

We first analyze the one-shot adaptation performance of a meta-learned planner in a matrix game setting, under both perfect and noisy observations for the agent and the planner. We show that meta-training reliably finds solutions that one-shot adapt well, and characterize how the planner's out-of-distribution performance depends on its observable information about the agent.

We next develop and empirically verify these insights with more adaptive agents and propose MERMADE which finds well-performing mechanisms in the sequential bandit setting. Here, RL based mechanisms perform well against out-of-distribution bandit learners, with test-time performance and robustness depending on the level of exploration of the agents and their pessimism in the face of uncertainty, confirming and extending the analysis and conclusions from the single-round setting.

2 Related Work

Bilevel optimization. Learning a mechanism with agents who also learn is a bilevel optimization problem, which is NP-hard (Ben-Ayed & Blair, 1990; Sinha et al., 2017). Possible solution techniques include branch-and-bound and trust regions (Colson et al., 2007). In particular, solving bilevel optimization using joint learning of the mechanism and the agents can be unstable, as the agents continuously adapt their behavior to changes in the mechanism. This can be stabilized using curriculum learning (Zheng et al., 2020), but generally bilevel problems remain challenging, especially with nonlinear objectives or constraints.

Meta-learning and distribution shift. In recent years, gradient-based meta-learning has proven effective in learning initializations for complex policy models that generalize well to unseen tasks (Finn et al., 2017a; Nagabandi et al., 2018). Luketina et al. (2022) showed that context-conditioned meta-gradients are effective for adapting in environments with controlled sources of non-stationarity, but they do not account for non-stationarity from interactions between strategic agents that learn. Prior works in imitation learning (Argall et al., 2009) and inverse RL (Abbeel & Ng, 2004) assume access to expert demonstrations with a fixed policy that the (RL) agent wants to emulate. In contrast, our mechanism designer aims to learn a policy that can strategically alter the behavior of such demonstrators (our agents), who are themselves learning during an episode of the demonstration. Recently, Boutilier et al. (2020) studied meta-learning for bandit policies, while Guo et al. (2021) introduced the inverse bandit setup for learning from low-regret demonstrators. However, these works do not consider shifts in the bandit learning algorithm between training and test time.

Modeling agents. A key challenge in multi-agent learning is that each agent experiences a non-stationary environment if other agents are learning. As such, agents can benefit from having a *world model*, e.g., to know what the policy of value function of the other agents are. World models can stabilize multi-agent RL (Lowe et al., 2017a) and enable higher-order learning methods (Foerster et al., 2018), and can be seen as a form of model-based RL. However, this may require a large amount of observational data or prior knowledge, which may be hard to acquire.

Adaptive mechanism design. Pardoe et al. (2006) found that a form of meta-learning that adapts the learning process itself can design English auctions (sequential bidding) that perform better with adaptive bidders who are loss-averse, and is still effective when the distribution of bidder behaviors (slowly) shifts. Baumann et al. (2020) studied gradient-based adaptive meta-learning, assuming access to the value function of the agents; having knowledge of or modeling other agents is important, e.g., to stabilize multi-agent learning (Lowe et al., 2017b; Baumann et al., 2020). Our work expands on this theme by explicitly modeling agents that learn, considering shifts in the *learning algorithm*

of the agents, and using deep RL with gradient-based meta-learning. The combination of these techniques enable learning mechanisms that generalize well across more complex tasks. However, we note that modeling other agents requires observational data about the agents.

3 K-SHOT ROBUST ADAPTIVE MECHANISM DESIGN

Overview. We model a *planner* who aims to incentivize an *agent* to (learn to) execute the planner's preferred action. To do so, the planner can *intervene* and change the agent's rewards at a cost. Without interventions, the agent may learn to prefer an action different than the planner's.

For example, consider consumers who can use either environmentally "clean" or "dirty" goods. Indifferent at first, consumers may gradually learn to prefer dirty goods if those are consistently cheaper than clean ones, whereas the government may want them to prefer clean goods. Here, the agent's reward is the negative of the cost of consumption, for instance, and an intervention changes the price of goods through taxes or subsidies. If we have access to a simulation, the planner can compute an optimal intervention. However, the simulation might be inaccurate and real-world agents might behave differently. As an example of such test-time distribution shift, simulated agents may be fast to change their consumption preferences, while real agents may be slow. A "good" mechanism (trained in a simulation) could quickly be fine-tuned to intervene more in the latter case and adapt quickly if such behavior is observed during deployment.

In particular, we focus on *learning* mechanisms that need to be adapted quickly on a single test-time trajectory (e.g., taxes and subsidies are deployed in the real world), and that are effective when the agent's learning algorithm differs from that seen during train-time.

We now formalize this setting. In this work, we focus on agents in a stateless environment for ease of exposition. For all variables and their meaning, see Tables 2 and 3 in the Appendix.

The agent. The agents are characterized by their action space A and a base reward function $r : A \to \mathbb{R}$. We call it the base reward because the agent experiences an *intervened* reward

$$\tilde{r}_t\left(a_t\right) = r\left(a_t\right) + r'_t\left(a_t\right), \tag{1}$$

where the intervention r'_t is provided externally (by the planner) for the agent action a_t. We index time as $t = 1, \ldots, T$. At each time step t, the agent's policy π_t computes a distribution over its actions based on the observations for the agent up to timestep t and executes $a_t \sim \pi_t$. We assume that the planner has a *preferred action* a^* that the agent should execute, whereas the agent's optimal policy can prefer a different action than a^* without intervention. Finally, at time t, the agent learns using an update rule $f : (\pi_t, a_t, \tilde{r}_t) \mapsto \pi_{t+1}$ to maximize the agent's intervened rewards, e.g., under UCB (Lai et al., 1985), f updates the confidence bounds for the action selected at time t.

The planner. The standard mechanism design problem studies which rewards and environment dynamics maximize the planner's objective, e.g., maximize social welfare and incentivize truthful behavior (Myerson, 1981). A standard assumption is that the agents are rational and that they do not change their behavior between training and testing time. Moreover, agents may have a private state (referred to as its *type*) which the planner cannot see.

In this work, from the planner's point of view, the *world (environment)* consists of the agent who maximizes $\tilde{r}$. Although the agent faces a stateless problem, *the planner faces a stateful problem with partial observability*. The full state $s \in S$ includes the planner's internal state h^p_t (e.g., the planner's belief about the value of the private agent information), and all information about the agent, including its past actions, reward function, and policy model; often, the latter two are private.

More formally, the planner can be modeled as a POMDP $(S, o^p, A^p, r^p, \gamma, \mathcal{P})$. The observation function o^p determines what part of a world state s is visible to the planner, A^p is its action space of interventions, r^p is its reward, γ is a discounting factor, and $\mathcal{P}$ are the environment dynamics, e.g., as caused by the agent's actions. At time t, the planner samples an action $\boldsymbol{a}^p_t \sim \pi^p\left(\boldsymbol{a}^p_t | o^p_{t-1}, h^p_{t-1}\right)$ which determines its intervention on each possible agent action a, i.e. $\boldsymbol{a}^p_t = \left[r'_1, \ldots, r'_{|A|}\right]$.

K-shot robust adaptive mechanism design. To model distribution shift at test time, we follow the meta-learning terminology (Finn et al., 2017b) and view each distinct agent as a *task* τ^i. The

planner has access to a *meta-train set of agents* $\tau^i \in \mathcal{T}_{\text{train}}; i = 1, \ldots, n_{\text{train}}$ and is evaluated on a *meta-test set of agents* $\tau^i \in \mathcal{T}_{\text{test}}; i = 1, \ldots, n_{\text{test}}$. *We emphasize that during a task, both the mechanism and agent may learn and adapt, both at meta-train and meta-test time.*

Here, we focus on two key challenges: K-shot adaptation and distribution shift. First, the planner gets only K episodes for fine-tuning for each meta-test task (but can train indefinitely for each meta-train task). Second, the planner faces two types of distribution shift: 1) across tasks and 2) intra-task non-stationarity. The meta-train and meta-test tasks may differ (significantly) in their temporal distribution of actions, e.g., due to different agent updates f or the agent rewards r_t being centered around different values (e.g., average price levels are higher in the real world vs in the simulation). Within a task, the agents learning is affected by the planner's interventions that change its reward $\tilde{r}$. This gives rise to non-stationarity in the agent's environment, as its learning objective may shift over time. These forms of distribution shift distinguish our robust adaptive mechanism design setting from most prior work in meta-learning, which often assume stationarity within a task and also assume similar task distributions at meta-train and meta-test times.

Objectives. The planner's objective is to maximize how often meta-test-time agents choose a^* during learning and have them converge to a policy that always chooses a^*. To do so, the planner aims to maximize the cost-adjusted test-time return $J^p_{\text{test}}\left(\pi^p, \pi^i\right) = \mathbb{E}_{\mathcal{T}_{\text{test}}}\left[\sum_{t=1}^{T} \gamma^{t-1}(r^p_t - \alpha c_t)\right]$, where the agent executes its (optimal) policy $\pi^i\left[\pi^p\right]$ in response to π^p:

$$\arg\max_{\pi^p} \mathbb{E}_{\tau^i \in \mathcal{T}_{\text{test}}} \mathbb{E}_{\pi^p} \mathbb{E}_{\pi^i[\pi^p]} \left[\sum_{t=1}^{T} \gamma^{t-1}(r^p_t - \alpha c_t)\right], \quad r^p_t = \mathbf{1}\left[a_t = a^*\right], \quad \alpha > 0, \tag{2}$$

where the planner incurs a cost c_t if it intervenes. A simple cost function is $c_t = \mathbf{1}\left[r'_t \neq 0\right]$, i.e., the cost is constant across non-trivial interventions, where $\alpha > 0$ is a constant. Note that if intervention were free ($c_t = 0$), a trivial solution is to always add a large $r'\left(a^*\right) \gg 0$ for its preferred action a^*, such that it always yields the highest reward. Hence, we focus on learning non-trivial strategies when intervention is costly, which forces the planner to strategically alter the agent's learning behavior.

During an episode of T time steps, each agent i starts with a uniformly initialized action probability distribution π^i_0 and optimizes π^i_t subject to interventions π^p to maximize its return: $\mathbb{E}_{\pi^i}\mathbb{E}_{\pi^p}\left[\sum_{t=1}^{T} \tilde{r}^i_t\left(a^i_t, a^p_t\right)\right]$. Here, we assume that T and γ are sufficiently large so the agent converges to its optimal policy under $\tilde{r}$, using its learning algorithm f. That is, we assume that the objective in Equation (2) is sufficient to describe the planner's objective of ensuring the agent converges to preferring a^* at some $t < T$.

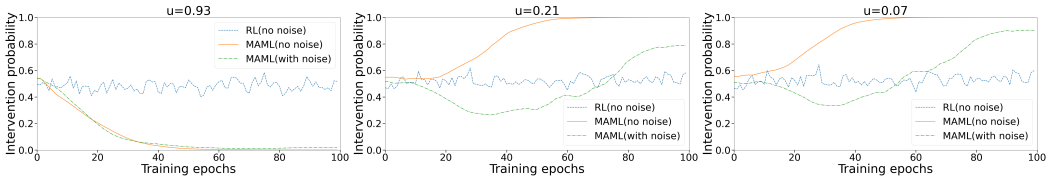
In the K-shot adaptation setting, at meta-test time, the planner gets K episodes to interact with any agent, each episode of length T steps. The planner has a fixed mechanism during an episode and it can update its mechanism at the end of an episode. The agent is reset across episodes, and within each episode, the agent follows its own learning strategy in response to the planner's interventions. On the $K+1^{\text{th}}$ episode, the planner evaluates its K-shot adapted mechanism on the agent. Note this assumes that the planner has a separate copy of the meta-test time agent for evaluation.

4 Analysis in the Matrix Game Setting

We first study robust adaptive mechanism design with strategic agents using a simple 2-player game between a planner and an agent. The agent's actions are "cooperate" and "defect", while the planner can choose whether or not to intervene. Assuming the row player is the agent and the column player is the planner, the 2×2 payoff matrix is given by

$$\begin{array}{l} \\ \text{Cooperate} \\ \text{Defect} \end{array} \begin{array}{c} \begin{array}{cc} \text{No intervention} & \text{Intervene} \end{array} \\ \begin{pmatrix} u, 1 & u+1, 1-c \\ 1-u, 0 & -u, -c \end{pmatrix} \end{array}, \tag{3}$$

where $u \in (0, 1)$ and c is the cost of intervention ($c < 1$). The planner prefers cooperation: it gets 1 if the agent cooperates and 0 if the agent defects (minus the cost c if it intervenes). The agent's base payoff u is its type. Notice that an intervention incentivizes the agent to cooperate ($u + 1 > -u$).



(a) (b) (c)

Figure 1: **Single round game.** REINFORCE (RL) does not adapt to expected Stackelberg equilibrium during evaluation. MAML's adaptability suffers under observation noise.

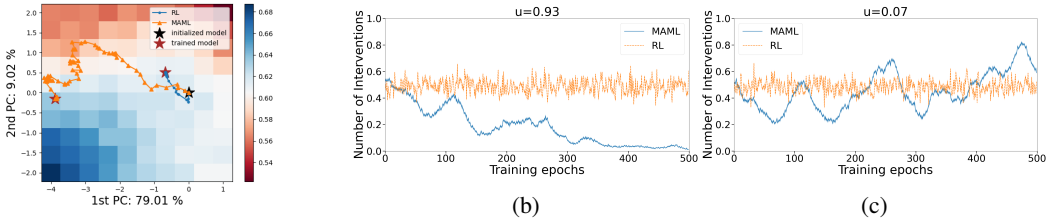
For mechanism design, it is natural to consider the Stackelberg setting where the planner is the leader and acts first (intervene or not), and the agent acts second (Von Stackelberg, 2010). Our goal is to learn a mechanism that can adapt to different agent types and find the Stackelberg equilibrium. We now analyze three scenarios with increasing difficulty:

1. First, we assume that the planner knows u. Here, there is a unique Stackelberg equilibrium at (Cooperate, No Intervention) when $u \geq \frac{1}{2}$, and at (Cooperate, Intervene) when $u < \frac{1}{2}$.
2. Second, the planner can observe a noisy version of u. In both these cases, the agent first observes the planner's action and plays the best response according to its payoff matrix.
3. Finally, we consider a repeated multi-stage game where the agent cannot observe the planner's action. Instead, we assume that the agent keeps a running average for the experienced payoffs for each of its actions. In a single-round setting this would correspond to the planner committing to a mixed action and then the agent choosing its best response. When $u \geq \frac{1}{2}$, the Stackelberg equilibrium occurs at (Cooperate, No Intervention). When $u < \frac{1}{2}$, at the Stackelberg equilibrium for this game the planner has a mixed action where it chooses to intervene for $\frac{2u+1}{2}$ fraction of times and the agent chooses to always cooperate.

With this equilibrium analysis in mind, we learn a neural network policy for the planner that predicts its probability of intervention and compare the behavior of the learned policy when trained using standard policy gradients (RL) versus meta-learning (MAML, (Finn et al., 2017b)). We set $c = 0.75$.

With perfect observability. Here, we study whether meta-learning finds a better initialization θ_{meta} for adaptation on unseen agents. In this setting, we assume that the planner observes an agent's exact payoff parameter u. It learns a stochastic policy $\pi_\theta^p(u)$ which determines its probability of intervening in a single-round game with an agent of type u. Given a set of training agents with different types $u \sim \mathcal{U}(0,1)$, for each u, the planner learns the optimal policy parameters $\theta^*(u) = \arg\max_\theta \mathbb{E}_{a^p \sim \pi_\theta^p(u)}[r^p(a^p)]$, where $r^p(a^p)$ is the planner's payoff for action a^p with agent u. The planner then learns θ_{meta} using the meta-learning algorithm in Appendix B and one-shot adapts on a meta-test set of agents with different us than at training. Note that we're studying the quality of the initialization, not the generalization performance of an already trained policy. Figure 1 shows the planner's meta-test time probability of intervening with 3 different agents from the test set, across training epochs. The planner and agent should be at different Stackelberg equilibria depending on the type u, as discussed above. We see that a planner trained from scratch on the test agents using standard policy gradients is unable to adapt to different agents in a single-shot adaptation setting. In contrast, with meta-learning, the planner learns a better policy that is one-shot adaptable to agents of different types and converges to the correct Stackelberg equilibrium at meta-test time.

With noisy observations for the planner. Here, we emulate a planner with partial observability of the agent, by letting the planner observe u with added i.i.d. Gaussian noise. The agent can see all payoffs and chooses the best response to achieve a Stackelberg equilibrium. Figure 1 shows that with noisy observations, the meta-learned planner policy requires more training time to be one-shot adaptable to the optimal intervention policy. This empirically indicates the increased difficulty of learning an adaptive intervention policy due to incomplete information about the agent, especially under limited adaptation time with unseen agents. It therefore motivates us to adopt a *model-based* approach for the planner to better estimate the agent type and learn an adaptive intervention policy.



(a) (b) (c)

Figure 2: **Multi-round game.** (a) Planner's optimization trajectory in the expected payoff landscape during training. Axes are PCA directions in the policy parameter space. (b)(c) MAML adapts (single shot) to Stackelberg equilibrium with a best response agent in a simplified form of MD.

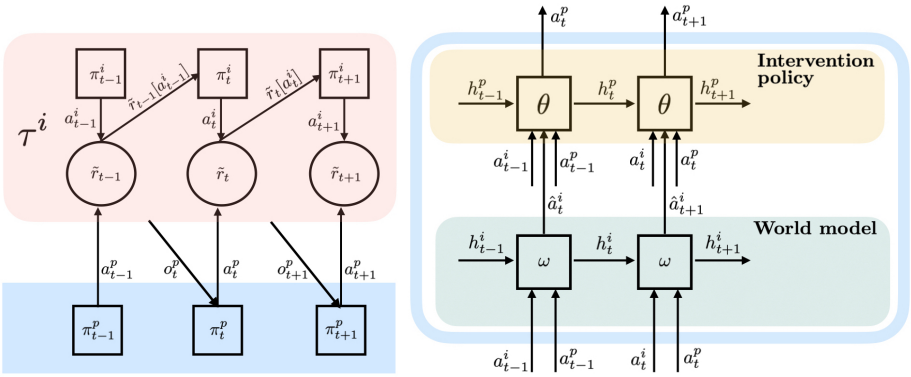


Figure 3: Overview of **MERMADE.** Left: Flow of planner and agent observables, rewards, and actions. Right: The planner's world model and intervention policy. Also see Algorithm 1.

Comparing Figure 1b and Figure 1c, we also see that when $u < \frac{1}{2}$, the difference in unintervened payoffs between the planner's preferred action (u) and the agent's preferred action ($1 - u$) also impacts the one-shot adaptability of the planner receiving noisy observations. This observation informs our analysis of the bandit setting in Section 6.

Multi-round repeated game with noisy rewards. In this setting, the planner and agent repeatedly play an iterated game over $T = 100$ steps. In each round, the planner observes the agent's type u with added i.i.d. Gaussian noise. The agent cannot observe the planner's actions, and plays a best response for its current estimate of the action payoffs. Whenever the agent selects an action, it receives a noisy observation of the true payoff and updates its estimate. Compared to the single-round setting, here the agent's best response behavior may change across rounds in the game depending on its observed payoffs, giving rise to non-stationarity in the planner's environment. The planner, in turn, has to learn to intervene so that the agent's best response is to cooperate.

Figure 2a compares the optimization trajectory followed using 1) standard policy gradients and 2) meta-learning for the planner. Starting from the same initialization, the meta-learned policy's parameters lie in a region of the payoff landscape with a higher expected value over the training agents. Moreover, Figure 2b and Figure 2c show the one-shot adaptability of the planner's policy for two different agent types at meta-test time. Meta-learning helps learn a better intervention strategy that is robust to the planner's observation noise as well as the agent's evolving best response strategy.

5 MERMADE

Motivated by our findings from Section 4, we now present MERMADE (Figure 3), consisting of:

Algorithm 1 MERMADE (Notations also in Table 3)

Initialize planner (θ_0, ω_0), and hidden states h_0^i, h_0^p.
for meta-train epoch $e = 1, \ldots, E_{\text{train}}$ **do**
  Update world model parameters $\omega = \omega_e$ using Equation (4).
  for agents (tasks) $i = 1, \ldots, n_{\text{train}}$ **do**
    Initialize agent: (μ^i, π_0^i), task specific planner policy parameter $\theta\left(\tau_0^i\right) = \theta_e$.
    for $k = 1, \ldots, K_{\text{train}}$ **do** ▷ Inner loop for K_{train} episodes.
      for time t = $1, \ldots, T$ **do** ▷ For each episode with T planner-agent interactions
        Predict $\hat{a}_t^i = \arg\max_{a_t^i} \hat{\pi}_\omega\left(a_t^i | a_{t-1}^i, a_{t-1}^p, h_{t-1}^i\right)$ using the world model.
        Intervention: $\tilde{\mu}^i = \mu^i + a_t^p, \quad a_t^p \sim \pi^p_{\theta\left(\tau_k^i\right)}\left(a_t^p | a_{t-1}^i, a_{t-1}^p, \hat{a}_t^i, h_{t-1}^p\right)$.
        Agent acts: $a_t^i \sim \pi_t^i$ and receives reward $r_t^i \sim \mathcal{N}\left(\tilde{\mu}^i, \sigma^2\right)$. $\pi_t^i \mapsto \pi_{t+1}^i$.
      Locally update $\theta\left(\tau_k^i\right) \mapsto \theta\left(\tau_{k+1}^i\right)$. ▷ Using REINFORCE.
    for $t = 1, \ldots, T$ **do** ▷ Rollout for meta-update; $\mathcal{D}_{\text{meta}}\left(\tau^i\right) = \{\}$
      Predict $\hat{a}_t^i = \arg\max_{a_t^i} \hat{\pi}_\omega\left(a_t^i | a_{t-1}^i, a_{t-1}^p, h_{t-1}^i\right)$ using the world model.
      Intervention: $\tilde{\mu}^i = \mu^i + a_t^p, \quad a_t^p \sim \pi^p_{\theta\left(\tau^i_{K_{\text{train}}}\right)}\left(a_t^p | a_{t-1}^i, a_{t-1}^p, \hat{a}_t^i, h_{t-1}^p\right)$.
      Agent acts: $a_t^i \sim \pi_t^i$, receives reward $r_t^i \sim \mathcal{N}\left(\tilde{\mu}^i, \sigma^2\right)$. Updates $\pi_t^i \mapsto \pi_{t+1}^i$.
      Collect $\mathcal{D}_{\text{meta}}\left(\tau^i\right) \cup \left\{a_t^i, a_t^p, \pi^p_{\theta\left(\tau^i_{K_{\text{train}}}\right)}\right\}$
  Meta-update $\theta_e \mapsto \theta_{e+1}$ using $\mathcal{D}_{\text{meta}} = \cup_{\tau^i} \mathcal{D}_{\text{meta}}\left(\tau^i\right)$. ▷ Using MAML.

- a recurrent *world model* parameterized by ω that outputs a distribution over an agent i's actions at the next time step t: $\hat{\pi}_\omega\left(a_t^i | a_{t-1}^i, a_{t-1}^p, h_{t-1}^i\right)$, conditioned on the planner's intervention and the observed agent action at $t-1$. h_{t-1}^i is the hidden world model state.
- a recurrent *intervention policy* which outputs a distribution over interventions $a_t^p \sim \pi_\theta^p\left(a_t^p | a_{t-1}^i, a_{t-1}^p, \hat{a}_t^i, h_{t-1}^p\right)$, conditioned on its previous intervention, the observed agent action and the world model's predicted next agent action $\hat{a}_t^i = \max_a \hat{\pi}_\omega\left(a | a_{t-1}^i, a_{t-1}^p, h_{t-1}^i\right)$. h_{t-1}^p is the hidden state of the policy network.

We train this using gradient-based meta-learning and RL, see Algorithm 1. Here, the planner maximizes the meta-train objective J_{train}^p similar to the objective in Equation (2). The base RL algorithm is REINFORCE (Williams, 1992) and the meta-learning update uses MAML (Finn et al., 2017b). The agent optimizes its cumulative intervened reward, see Section 6 for details. The world model $\hat{\pi}_\omega$ trains by maximizing the log-likelihood of the observed a_t^i, using Adam (Kingma & Ba, 2014):

$$\arg\max_\omega \mathbb{E}_{a^p \sim \pi^p} \mathbb{E}_{a^i \sim \pi^i}\left[\sum_{t=1}^T \log \hat{\pi}_\omega\left(a_t^i | a_{t-1}^i, a_{t-1}^p, h_{t-1}^i\right)\right]. \tag{4}$$

Note that the planner's parameters θ are updated after each T-step episode, while the agent continuously learns during each episode. Also, the agent is reset in between episodes. At time 0, the world model makes a prediction based on zero initialization.We use a single world model for all agents. At meta-test time, only the intervention policy is updated by one-shot adaptation to a new agent.

6 Experimental Validation in the Bandit Setting

We now study the K-shot robust adaptive mechanism design problem, where the agent is a $|A|$-armed bandit instance with action set A having base reward $\boldsymbol{r} = \left[r_1, \ldots, r_{|A|}\right]$. At each time step t, the agent chooses an arm a and gets a reward sampled from $\mathcal{N}\left(r_a, \sigma^2\right)$. We assume $r_a \in (0, 1)\ \forall a$. The agent aims to maximize its cumulative reward over a horizon of T steps. The agent can only observe the reward for the chosen action, and hence faces a explore-exploit dilemma addressed by bandit algorithms like UCB (Lai et al., 1985). We assume there is a unique arm $\tilde{a}$ with the highest base reward: $\tilde{a} = \arg\max_a r_a$, i.e., the agent's preferred action without any intervention.

Costly interventions. To analyze the effect of the cost of intervention c_t on the planner's learnt policy, we assume that the planner decides among three different intervention levels $|r'| \in \{0, 0.5, 1\}$ such that $c_t = |r'|$. Across different bandit agent tasks τ^i with distinct base rewards $\boldsymbol{r}^i$ and reward gaps $\delta = \max_{a \in A} \boldsymbol{r}^i[a] - \boldsymbol{r}^i[a^*]$, the planner should learn to appropriately incentivize the agent while minimizing the total cost of intervening. We then define the experienced reward as:

$$\tilde{\boldsymbol{r}}_t[a^*] = \boldsymbol{r}^i[a^*] + r'_t; \quad \tilde{\boldsymbol{r}}_t[a] = \boldsymbol{r}^i[a] - r'_t, \quad \forall a \neq a^*, (a, a^* \in A). \tag{5}$$

Note that this ensures the agent always experiences an intervention, no matter which action it chooses. During each episode, the agent learns but the planner's policy is fixed; the planner can update its policy only at the end of each episode (Algorithm 1). Also, we assume that the planner can only observe the agent's actions a_t^i but not its base reward $\boldsymbol{r}^i$ or policy update function f^i. We measure the performance of the planner using Equation (2), with $\gamma = 1$.

World model. The world model predicts the agent's next *action* (given the planner's prior observations) to characterize the agent's behavior. We do not train the planner's world model to estimate the base rewards, because bandit agents with distinct base rewards could still execute the same sequence of actions, depending on the agent's explore-exploit algorithm and its observations.

Challenges in the sequential setting. Compared to the simple game setting described in Section 4, mechanism design with sequential (bandit) learners gives rise to additional challenges:

- Bandit agents may follow different strategies for action selection to maximize their experienced reward. The agent's rate of exploration may be constant (e.g., ϵ-greedy) or it can reduce with time (e.g., UCB) within an episode, depending on its observations. This creates a highly non-stationary environment for the planner wherein its decision to intervene must adapt to different explore-exploit behaviors for the same agent within an episode. When the agent explores a larger action space, it further exacerbates the challenges in estimating the agent's behavior since the planner only has partial information about the agent.
- Bandit agents are sequential learners and feedback (a_t^i, $\tilde{r}_t^i$) can update the policy π^i differently at different steps t. The update may depend on how optimistic (e.g., UCB) or pessimistic (e.g., EXP3) the bandit agents are about their reward estimates. Hence, an intervention a^p may not equally incentivize the agent at different t. Since the planner's interventions have different costs, a strategic planner must decide *when* to intervene and *how much* ($|r'|$) depending on its observations of the agent's actions.

Results. In the following experiments, we use 15 bandit agents for training and 10 bandit agents for testing, each with different base rewards (both within and across train and test sets). $|A| = 10$. We consider two agent learning algorithms (UCB and ϵ-greedy). In each experiment, the train and test agents use the same algorithm, but with different tendencies for exploration vs exploitation, determined by their exploration coefficients: $\beta \in \{0.17, 0.27, 0.42, 0.5, 0.67\}$ for UCB (higher β gives more exploration) and $\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ for ϵ-greedy (higher ϵ gives more exploration). These constants were chosen such that they afford, on the average, the same number of exploratory actions when following either UCB or ϵ-greedy strategy without any intervention (see Appendix B).

In Table 1, we show the one-shot adapted planner's score on each test set over $T = 200$ time steps. We compare MERMADE against 1) model-free baselines (MF-RL using REINFORCE and MF-MAML using MAML), as well as 2) REINFORCE with world model (WM-RL) (see Appendix B for details). We also include a "No Intervention" baseline to show how agents behave by default.

Out-of-distribution performance. Table 1 shows the planner's score when evaluated on test agents having a *different exploration constant* than train agents. Using meta-learning for the mechanism (MF-MAML) and using a world model to predict the agent's behavior (WM-RL) both have advantages for training a robust and one-shot adaptable intervention policy. A world model is advantageous when 1) the test agent is more exploratory than the train set (e.g., $\epsilon = 0.1$ at training, $\epsilon = 0.4$ at test), or 2) the agent explores throughout an episode and is likely to often select actions other than the one with its current maximum mean reward estimate (e.g., $\epsilon = 0.5$ at training). Because we evaluate on $K = 1$, fine-tuning on only a single test-time episode, a trained world model provides a useful prior belief representation for the planner. Indeed, the MF-RL results show the

Train on UCB, $\beta = 0.17$	Test on $\beta = 0.17$	$\beta = 0.27$	$\beta = 0.42$	$\beta = 0.5$	$\beta = 0.67$
No intervention	3 (0)	5 (0)	8 (0)	10 (0)	12 (0)
MF-RL	119 (2)	109 (2)	98 (2)	90 (2)	77 (1)
MF-MAML	133 (2)	125 (3)	107 (1)	97 (1)	77 (0)
WM-RL	123 (7)	112 (6)	100 (4)	92 (2)	75 (1)
MERMADE (ours)	**154 (2)**	**141 (1)**	**115 (1)**	**103 (0)**	**80 (1)**
Train on ϵ**-greedy,** $\epsilon = 0.1$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
No intervention	3 (0)	4 (1)	7 (0)	9 (1)	11 (0)
MF-RL	115 (5)	94 (4)	54 (19)	39 (6)	22 (9)
MF-MAML	122 (4)	97 (3)	58 (5)	40 (2)	12 (1)
WM-RL	115 (4)	94 (5)	70 (1)	55 (3)	**38 (1)**
MERMADE (ours)	**134 (1)**	**108 (1)**	**85 (1)**	**57 (7)**	29 (1)
Train on UCB, $\beta = 0.67$	$\beta = 0.17$	$\beta = 0.27$	$\beta = 0.42$	$\beta = 0.5$	$\beta = 0.67$
No intervention	3 (0)	5 (0)	8 (0)	10 (0)	12 (0)
MF-RL	103 (3)	101 (3)	92 (2)	85 (1)	74 (1)
MF-MAML	124 (2)	116 (1)	102 (1)	94 (1)	80 (1)
WM-RL	100 (4)	89 (0)	85 (1)	85 (1)	74 (0)
MERMADE (ours)	**131 (2)**	**125 (2)**	**109 (1)**	**101 (1)**	**85 (1)**
Train on ϵ**-greedy,** $\epsilon = 0.5$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
No intervention	3 (0)	4 (1)	7 (0)	9 (1)	11 (0)
MF-RL	4 (5)	2 (3)	5 (0)	11 (5)	7 (1)
MF-MAML	2 (0)	4 (0)	6 (0)	8 (1)	11 (1)
WM-RL	102 (6)	79 (10)	68 (3)	47 (1)	30 (2)
MERMADE (ours)	87 (42)	**102 (3)**	**78 (6)**	**69 (1)**	**46 (2)**

Table 1: **Test-time planner scores across 3 random seeds.** The left column denotes the training algorithm (e.g., MERMADE) used and the training agent type (e.g., UCB with $\beta = 0.17$). The other columns show scores when testing on agents with the same algorithm, but different hyperparameters.

hidden state representation of the model-free planner might be unable to adapt to high environment non-stationarity without a trained next-agent-action world model.

Compared to an ϵ-greedy agent, the UCB agent explores mostly at the start of an episode, for all β. Hence, with UCB agents, the planner learns an effective one-shot adaptable intervention policy using meta-learning (MF-MAML) only (even without a world model), as the agents cause less distribution shift across different c. It further emphasizes the effectiveness of meta-learning for K-shot robust adaptive mechanism design: unlike MF-MAML, neither the world model nor the intervention policy is meta-learned in WM-RL. Moreover, it also shows that for the same amount of distribution shift (characterized in Appendix B), the relative benefit of a world model or meta-learning the planner's policy depends on the nature of the agent's exploration strategy (which is unknown to the planner).

In all, these results show that MERMADE combines the best of both techniques, so that the planner obtains a higher score across agents with different learning algorithms and explore-exploit behaviors.

Agent exploration vs intervention cost. In order to intervene effectively, the planner should learn *when* to intervene and *how much* to incentivize the agent while minimizing its incurred cost. This is a challenging learning problem for the planner not just during meta-training, but more so during one-shot adaptation at meta-test time. Bandit algorithms like EXP3 (Auer et al., 2002) use pessimism in the face of uncertainty, and encourage continued exploration. This increases the non-stationarity for the planner. In order to effectively incentivize such agents to prefer a^*, the planner needs to accurately predict the agent's policy from its observations; otherwise it can incur a high cost for intervening ineffectively *and* lowering its score, and learn to stop intervening. Indeed, our results when training on $\epsilon = 0.5$-greedy agents show that the MF-RL and MF-MAML planner stop intervening. In contrast, in that setting, MERMADE learns an effective intervention policy that outperforms all baselines, even under distribution shift between meta-train and meta-test agents.

7 Discussion and Future Work

We have proposed a framework for K-shot robust adaptive mechanism design with strategic agents that can adapt their behavior. In this work, we focused on a singple planner and agent; future work could extend this to mechanism design with multiple learning agents who may coordinate, compete, or a combination thereof. Moreover, it is interesting to extend MERMADE to agents that adapt adversarially to the mechanism, a challenging non-stationary problem for the planner.

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Tobias Baumann, Thore Graepel, and John Shawe-Taylor. Adaptive mechanism design: Learning to promote cooperation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Omar Ben-Ayed and Charles E Blair. Computational difficulties of bilevel linear programming. *Operations Research*, 38(3):556–560, 1990.

Craig Boutilier, Chih-wei Hsu, Branislav Kveton, Martin Mladenov, Csaba Szepesvari, and Manzil Zaheer. Differentiable meta-learning of bandit policies. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2122–2134. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/171ae1bbb81475eb96287dd78565b38b-Paper.pdf.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of operations research*, 153(1):235–256, 2007.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017a.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv:1703.03400 [cs]*, July 2017b. URL http://arxiv.org/abs/1703.03400.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with Opponent-Learning Awareness. *arXiv:1709.04326 [cs]*, September 2018. URL http://arxiv.org/abs/1709.04326.

Wenshuo Guo, Kumar Krishna Agrawal, Aditya Grover, Vidya Muthukumar, and Ashwin Pananjady. Learning from an exploring demonstrator: Optimal reward estimation for bandits, 2021. URL https://arxiv.org/abs/2106.14866.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6379–6390. Curran Associates, Inc., 2017a.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017b.

Jelena Luketina, Sebastian Flennerhag, Yannick Schroecker, David Abel, Tom Zahavy, and Satinder Singh. Meta-gradients in non-stationary environments. In *ICLR Workshop on Agent Learning in Open-Endedness*, 2022.

Setareh Maghsudi, Andrew Lan, Jie Xu, and Mihaela van Der Schaar. Personalized education in the artificial intelligence era: what to expect next. *IEEE Signal Processing Magazine*, 38(3):37–50, 2021.

Paul Milgrom and Paul Robert Milgrom. *Putting auction theory to work*. Cambridge University Press, 2004.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2018.

David Pardoe, Peter Stone, Maytal Saar-Tsechansky, and Kerem Tomak. Adaptive mechanism design: A metalearning approach. In *Proceedings of the 8th International Conference on Electronic Commerce: The New e-Commerce: Innovations for Conquering Current Barriers, Obstacles and Limitations to Conducting Successful Business on the Internet*, ICEC '06, pp. 92–102, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933921. doi: 10.1145/1151454.1151480. URL https://doi.org/10.1145/1151454.1151480.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22 (2):276–295, 2017.

Richard S Sutton and Andrew G Barto. Introduction to reinforcement learning, 1998.

Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C. Parkes, and Richard Socher. The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies. *arXiv:2004.13332 [cs, econ, q-fin, stat]*, April 2020. URL http://arxiv.org/abs/2004.13332.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, 8(18):eabk2607, 2022. doi: 10.1126/sciadv.abk2607. URL https://www.science.org/doi/abs/10.1126/sciadv.abk2607.