2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

A Moreau Envelope Approach for LQR Meta-Policy Estimation

Ashwin Aravind, Mohammad Taha Toghani, and César A. Uribe

Abstract— **We study the problem of policy estimation for the Linear Quadratic Regulator (LQR) in discrete-time linear time-invariant uncertain dynamical systems. We propose a Moreau Envelope-based surrogate LQR cost, built from a finite set of realizations of the uncertain system, to define a meta-policy efficiently adjustable to new realizations. Moreover, we design an algorithm to find an approximate first-order stationary point of the meta-LQR cost function. Numerical results show that the proposed approach outperforms naive averaging of controllers on new realizations of the linear system. We also provide empirical evidence that our method has better sample complexity than Model-Agnostic Meta-Learning (MAML) approaches.**

I. Introduction

Complexity and uncertainty are inherent to the control of modern engineering systems and their applications. Interactions with other socio-technical systems or unpredictability in the operating environment hinder our ability to obtain accurate models as the effects of such complex interactions might become apparent only during task execution [1], [2].

Reinforcement Learning (RL) has emerged as a powerful, data-driven approach in designing controllers, particularly in scenarios where the system model is not fully known or is too complex to be captured by traditional methods [3]–[9]. Such methods involve RL techniques to iteratively learn optimal control policies through interaction with the system, bypassing the need for an explicit model.

The seminal work in [3] introduces a data-driven linear quadratic regulator (LQR) design using RL frameworks and demonstrates global convergence to an optimal control policy without requiring a model for the system. However, the uncertainties in the system while executing the control policy can significantly affect the stability and operating costs. Thus, developing resilient strategies that quickly adapt to new realizations of the uncertain system is crucial when designing control policies.

Meta-learning facilitates quick adaptation to new tasks by applying previously acquired knowledge to new challenges [10], [11]. In RL, this evolves into Meta-RL, which trains agents to efficiently adapt to unknown environments [12]–[16]. The Meta-RL framework aims to develop a meta-policy that allows for rapid learning across various tasks, fostering knowledge adaptation to new scenarios.

AA is with the Centre for Systems and Control, IIT Bombay, Mumbai, MH, India. a.aravind@iitb.ac.in. MTT and CAU are with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. {mttoghani, cauribe}@rice.edu.

AA thanks Mehta Rice Engineering Scholars Program and IoE - IIT Bombay for the support during this work. AA is supported by the Ministry of Human Resource Development, Govt. of India. MTT's research is supported by the Lodieska Stockbridge Vaughn Fellowship. This work is supported by the National Science Foundation under Grants #2211815 and #2213568 and the Google Scholar Research Award.

Model-Agnostic Meta-Learning (MAML), notable for its versatility, focuses on optimizing a model's initial parameters for swift adaptation through minimal adjustments, setting them to be highly responsive to a few policy gradient updates for quick task-specific learning [10], [14]. On the other hand, the Moreau Envelope (ME) approach [17]–[19] introduces a surrogate cost with a regularization term to smooth the optimization process, enhancing the stability of gradient updates and convergence efficiency.

The meta-learning problem for LQR policy design was studied in [20]–[22] and focuses on meta-learning based on the MAML framework. MAML for the LQR problem in a single system and multi-task setting was studied in [22], which was later extended to a multi-system and multi-task setting by [20]. In their setup, a finite set of observed tasks is used to design an LQR meta-policy that can effectively and quickly adapt to unobserved tasks, but only local convergence is shown. Later, the authors in [21] studied system heterogeneity and provided global convergence guarantees.

When applying policy gradient techniques, MAML and ME define different surrogate cost functions and access different computational oracles. MAML uses the estimation of the gradient and Hessian of the cost function, whereas ME uses an approximate solution of an inner optimization problem via a first-order oracle. The inner loop accuracy parameter appears in the convergence guarantee, thereby providing the user with explicit control over the quality of the solution. Complied with this theoretical intuition, recent studies have shown that ME allows for a better empirical performance than MAML in multi-task setups [17]–[19].

This work investigates uncertain linear control systems within the LQR framework, applying meta-learning to handle uncertainties. We introduce a meta-policy estimation method using an ME-based framework [18] tailored for the LQR in both model-free and model-based settings to facilitate flexible policy initialization for rapid adaptation to new realizations of the uncertain system. Our setup is motivated by the challenges a control system may encounter due to uncertainties rather than having to perform multiple LQR tasks. However, both these scenarios result in a similar underlying framework where the cost incurred by any given policy varies due to changing realizations of the uncertainties or a change in the LQR task. Hence, it is possible to use the approaches presented for both settings interchangeably.

The main contribution of this paper is a novel first-order meta-RL algorithm (MEMLQR) that computes LQR meta-policies that can be efficiently adapted for an unseen system realization. By integrating meta-learning principles, our approach improves policy adaptability, ensuring effective

performance in various scenarios and quick convergence.

The contributions of this paper are summarized as follows:

1) We define an augmented cost function where the linear quadratic cost is regularized with Moreau Envelopes to induce personalization amenable to model-based and model-free policy gradient methods for policy design.
2) We propose a first-order iterative algorithm to optimize the defined augmented cost in a client-server setup.
3) We show the convergence of the proposed algorithm to an approximate first-order stationary point. We also show that the policies generated by the algorithm will incur a finite cost for all the available system realizations.
4) We present a set of numerical results that validate the algorithm's efficiency in minimizing adaptation costs at the testing phase, highlighting the practical benefits of the proposed method. Moreover, we show the proposed method outperforms other approaches based on MAML personalization.

The rest of this article is organized as follows. In Section II, we formally introduce the problem we are addressing along with the relevant background and propose a solution. In Section III, we present the assumptions along with results that indicate the convergence of the proposed algorithm. Section IV contains numerical examples to indicate the results provided in Section III. We conclude our article in Section V and provide the scope for future work.

Notation: By $\mathbb{R}$, we denote the set of real numbers, and by $\mathbb{Z}_+$, we represent the set of positive integers. For two real numbers a and b such that $b > a$, by *unif*(a, b), we define a uniform distribution over the interval $[a, b]$. For any square matrix M, $M \succeq 0$ denotes that M is positive semi-definite, and $M \succ 0$ means positive definite. M is Schur if all its eigenvalues lie in the open unit disc.

II. Problem Formulation and Algorithm

In this section, we first introduce the problem setup of finding the optimal policy for a linear system, i.e., single realization. Then, we discuss our approach to finding a meta-policy via personalized costs. Finally, we present our algorithm that minimizes the augmented cost.

A. Preliminaries

Consider the family of discrete-time linear time-invariant uncertain dynamical systems

$$x^{t+1} = \mathcal{A}x^t + \mathcal{B}u^t, \quad t = 0, 1, 2, \cdots, \tag{1}$$

where, $\mathcal{A} := \mathcal{A}_0 + \sum_{i=1}^{p} a_i \mathcal{A}_i, \mathcal{B} := \mathcal{B}_0 + \sum_{j=1}^{q} b_j \mathcal{B}_j$, $x^t \in \mathbb{R}^n$ is the state and $u^t \in \mathbb{R}^m$ is the input to the system at time t, $a := (a_1\ a_2\ \cdots\ a_p) \in \mathbb{R}^p$ and $b := (b_1\ b_2\ \cdots\ b_q) \in \mathbb{R}^q$ are bounded uncertainties that determine the system dynamics along with matrices $\mathcal{A}_0 \in \mathbb{R}^{n\times n}$ and $\mathcal{B}_0 \in \mathbb{R}^{n\times m}$, and sets of matrices, $\bar{\mathcal{A}} := (\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_p) \in (\mathbb{R}^{n\times n})^p$ and $\bar{\mathcal{B}} := (\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_q) \in (\mathbb{R}^{n\times m})^q$. It is assumed that all possible realizations of the $\mathcal{A} - \mathcal{B}$ pair are controllable.

Moreover, consider a set $\mathcal{V}$ of $V = |\mathcal{V}|$ realizations of (1),

$$x_i^{t+1} = A_i x_i^t + B_i u_i^t, \quad i = 1, 2, \cdots, V, \tag{2}$$

where $x_i^t \in \mathbb{R}^n$ is the state and $u_i^t \in \mathbb{R}^m$ is the input for the i^{th} system at time t, and $A_i \in \mathbb{R}^{n\times n}$ and $B_i \in \mathbb{R}^{n\times m}$ are the system matrices, and the initial state x_i^0 is randomly drawn from distribution $\mathcal{D}_i$.

The standard LQR problem for each realization $i \in \mathcal{V}$ is

$$\min_{\{u_i^t\}_{t=0}^{\infty}} \mathbb{E}_{x_i^0 \sim \mathcal{D}_i} \left[\sum_{t=0}^{\infty} {x_i^t}^\top Q x_i^t + {u_i^t}^\top R u_i^t \right], \tag{3}$$

subject to (2), where $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are the cost matrices such that $Q \succeq 0$ and $R \succ 0$. Without loss of generality, it is assumed that the cost matrices Q and R are the same for all the systems. This models a scenario where the controllers have the same goal or are trying to accomplish a similar task. The optimal control policy for the above cost function is a linear feedback policy of form $u_i^t = -K_i x_i^t$ for a policy $K_i \in \mathbb{R}^{m\times n}$ [1, Chapter 3]. Therefore, the cost can be expressed as

$$C_i(K_i) := \mathbb{E}_{x_i^0 \sim \mathcal{D}_i} \left[\sum_{t=0}^{\infty} {x_i^t}^\top (Q + {K_i}^\top R K_i) x_i^t \right]. \tag{4}$$

Note that, for $C_i(K_i) < \infty$ (i.e., K_i is stable), we require that $A_i + B_i K_i$ is Schur.

The seminal work in [3] established that the cost function $C_i(K_i)$ is non-convex with respect to K_i. However, its gradient domination and local smoothness properties make policy gradient iteration methods converge globally in model-based and model-free settings under the assumption that the initial state distribution is such that the covariance matrix given by $\mathbb{E}_{x_i^0 \sim \mathcal{D}_i}\left[x_i^0 {x_i^0}^\top\right]$ is full rank.

The gradient descent iterates for a system $i \in \mathcal{V}$ for finding the optimal policy is stated below:

$$K^{s+1} = K^s - \alpha \nabla C_i(K^s), \tag{5}$$

where, $K^s \in \mathbb{R}^{m\times n}$ is the feedback at s^{th} iteration of the descent, α is the step-size and the gradient is given by

$$\nabla C_i(K^s) = 2\big((R + {B_i}^\top P_{K^s} B_i) K^s - {B_i}^\top P_{K^s} A_i\big) S_{K^s}, \tag{6}$$

with,

$$P_{K^s} = Q + {K^s}^\top R K^s + (A_i - B_i K^s)^\top P_{K^s} (A_i - B_i K^s), \tag{7}$$

$$S_{K^s} = \mathbb{E}_{x_i^0 \sim \mathcal{D}_i} \left[\sum_{t=0}^{\infty} x_i^t {x_i^t}^\top \right]. \tag{8}$$

The iterates in (5) need to be initialized at a stable policy K^0 (i.e., $C_i(K^0) < \infty$). For the case where the model is available, it is easy to calculate the gradient via (6). However, in the case of a model-free setup, the system and task parameters are unknown. Therefore, the gradient may be estimated by a zeroth-order algorithm [3, Algorithm 1].

B. Moreau Envelope-based LQR Meta-Policy Estimation

While the approach presented by the policy gradient iterations in (5) is efficient for the computation of approximate optimal policies, here we focus on the task of learning good initialization that can be efficiently adapted for the LQR problem on unseen realizations of (2). We propose to use

the Moreau Envelope (ME) approach for meta-learning [17]–[19] that defines an optimization problem

$$\min_{K\in\cap\mathcal{K}_i} C_\lambda(K) := \sum_{i\in\mathcal{V}} C_{\lambda,i}(K), \tag{9a}$$

$$\text{with } C_{\lambda,i}(K) := \min_{K_i\in\mathcal{K}_i} C_i(K_i) + \frac{\lambda}{2}\|K_i - K\|^2, \tag{9b}$$

where $\lambda \in (0,\infty)$ is a regularization parameter.

Remark 1 *For any $i \in \mathcal{V}$, the minimizer K_i^* for* (9b) *will be such that $A_i + B_iK_i^*$ is Schur. We define the set of stabilizing policies for any $i \in \mathcal{V}$ as $\mathcal{K}_i := \{K \in \mathbb{R}^{m\times n} | A_i + B_iK \text{ is Schur}\}$.*

Setting $\lambda = 0$ in ME equates the algorithm to local RL, concentrating exclusively on optimizing the current task without leveraging insights from other tasks. Conversely, as λ approaches infinity, the algorithm defaults to a naive averaging of costs, disregarding the unique characteristics of each task and resulting in a broadly applicable but non-personalized policy. This is obtained by augmenting the cost defined in (3) with a regularizer term and summing this augmented cost over all the systems in the set $\mathcal{V}$. We define a cost based on the Moreau Envelopes inspired by the personalized federated learning setup [17], [19].

The meta-policy is given by, $K^* = \arg\min_{K\in\cap\mathcal{K}_i} C_\lambda(K)$. The minimization of this cost $C_\lambda(K)$ is a bilevel optimization problem where the solutions of both outer and inner minimization problems implicitly depend on each other. Here, the inner problem minimizes the cost induced when the LQR cost is regularized with a proximity term.

We propose a gradient-based iterative framework termed *MEMLQR*: Moreau Envelope based Meta Linear Quadratic Regulator, shown in Algorithm 1, to solve the bilevel optimization problem (9a), for which the gradient is

$$\nabla C_\lambda(K) = \sum_{i\in\mathcal{V}} \nabla C_{\lambda,i}(K), \tag{10a}$$

$$\text{such that, } \nabla C_{\lambda,i}(K) = \lambda\big(K - \widehat{K}_i(K)\big), \tag{10b}$$

$$\text{with } \widehat{K}_i(K) = \arg\min_{\breve{K}\in\mathcal{K}_i} C_i(\breve{K}) + \frac{\lambda}{2}\|\breve{K} - K\|^2. \tag{10c}$$

Remark 2 *The computation of* (10b) *requires solving* (10c). *The gradient for the augmented inner cost function is given by $\nabla C_i(\breve{K}) + \lambda(\breve{K} - K)$, where, $\nabla C_i(\breve{K})$ can be obtained using* (6) *or the zeroth-order approach [3, Algorithm 1], and K is the current estimate of the meta feedback policy. Thus, this approach is equally amenable to model-based or model-free scenarios.*

As the cost induced by an estimate of the meta-policy at different systems is independent of each other, a client-server model can be used for computation as individual clients need not share their data. Once the inner optimizer is obtained, a gradient descent step is performed locally for the outer problem, and a local estimate for the meta-policy is obtained. Once all the systems have an estimate of the meta-policy, these estimates are communicated to the server. The meta-policy at the server is updated using estimates from different systems and the previous estimate from the server. This iterative gradient descent is performed until the number of outer iterations guarantees the required estimate accuracy.

Algorithm 1: *MEMLQR*: Moreau Envelope based Meta Linear Quadratic Regulator

Data : Number of outer iterations S, number of inner iterations P, inner step-size α, outer step-size β, accuracy threshold δ, initial value of the policy K^0.

for $s = 0, 1, \cdots, S-1$ **do**
  Communicate the policy at the server, $K_i^{s,0} \leftarrow K^s$, for all $i \in \mathcal{V}$
  for $i \in \mathcal{V}$ **do**
    for $p = 0, 1, 2, \cdots, P-1$ **do**
      Compute $\bar{K}$ s.t.
      $\|\widehat{K}_i(K_i^{s,p}) - \bar{K}\| \leqslant \delta$, c.f. (10c)
      $K_i^{s,p+1} \leftarrow K_i^{s,p} - \alpha\lambda(K_i^{s,p} - \bar{K})$
    end
    $K^{s+1} \leftarrow (1-\beta)K^s + \frac{\beta}{V}\sum_{i\in\mathcal{V}} K_i^{s,P}$
  end
end
return K^S

III. Convergence analysis

This section formally states the associated assumptions and auxiliary results for the convergence analysis for Algorithm 1. Due to space considerations, the proofs have been moved to [23]. The policy gradient method we use as a baseline for designing our algorithm finds the gradient at the current iterate of the feedback policy by simulating the system under that policy from random initial states. The assumption presented below ensures that all the states are visited with a non-zero probability while exploring.

Assumption 3 *(Persistence of excitation-like): Let x_i^0 be the initial state of realization $i \in \mathcal{V}$, then $\mathbb{E}_{x_i^0\sim\mathcal{D}_i}\left[x_i^0 x_i^{0\top}\right]$ is full rank.*

The above assumption is well-known and standard in the control literature [3], [4]. The system should be stable at the initial guess for any policy gradient algorithm to work. The assumption below ensures that none of the system's available realizations become unstable when we initialize the algorithm using a random initial gain. Such instability will cause the gradient to be unavailable at those realizations.

Assumption 4 *(Bound on initial cost): Let the set $\mathcal{K} := \cap\mathcal{K}_i$ be non-empty and $K^0 \in \mathcal{K}$ be the policy used to initialize Algorithm 1, then $C(K^0) < \infty$.*

The following assumption ensures that the systems generated using the available uncertainty realizations are suf-

ficiently close to each other. The bound assumed below is critical for the analysis that follows.

Assumption 5 *(Bounded heterogeneity): For all $K \in \mathcal{K}$ such that $C(K) < \infty$, there exists a constant $\sigma_C > 0$ satisfying,*

$$\frac{1}{V}\sum_{i\in\mathcal{V}} \|\nabla C_i(K) - \nabla C(K)\|^2 \leqslant \sigma_C^2.$$

The motivation for this assumption is similar to that of [21, Lemma 4], where the authors had established a similar bound on the diversity of gradients based on the diversity in the system and cost matrices. In our case, the diversity in system matrices is bounded as the uncertainties a and b are bounded. Therefore, Assumption 5 aligns with the established theory. Algorithm 1 requires the estimation of an optimizer of the inner problem (10c) up to a desired accuracy. The LQR cost as a function of the feedback policy satisfies the PL-inequality [3, Lemma 3], i.e., let K_i^* be the optimal policy for a system realization $i \in \mathcal{V}$, then for any stable policy K and some constant $\mu > 0$, it holds that

$$C_i(K) - C_i(K_i^*) \leqslant \frac{\mu}{2}\|\nabla C_i(K)\|^2, \tag{11}$$

In our approach, we add a quadratic regularizer term to this gradient-dominated function. We have established in Lemma 6 that the regularized cost (9b) also satisfies the PL inequality. This means it is possible to solve the inner minimization problem to any specified accuracy δ using gradient descent (or Gauss-Newton or Natural policy gradient) iterations of order $\mathcal{O}(\log(1/\delta))$ [3, Theorem 7, Theorem 9]. Next, we build a sequence of auxiliary lemmas that will be useful in analyzing Algorithm 1. First, we show the gradient dominance property of the Moreau regularized cost in (9b).

Lemma 6 *(Gradient dominance of Moreau Envelope cost): Consider a policy $\breve{K} \in \mathbb{R}^{m\times n}$, then*

$$C_{\lambda,i}(\breve{K}) - \min_{K\in\mathcal{K}_i} C_{\lambda,i}(K) \leqslant \left(\frac{\mu}{2}+\frac{1}{2\lambda}\right)\|\nabla C_{\lambda,i}(\breve{K})\|^2. \tag{12}$$

The following lemma shows how the cost function's smoothness property is inherited from the Moreau regularized cost.

Lemma 7 *(Local smoothness of cost functions): For any given system realization $i \in \mathcal{V}$ and (stable) policies $K_1, K_2 \in \mathcal{K}_i$ such that $\|K_1 - K_2\| \leqslant \Delta_K$, there exists $L > 0$ such that,*

$$\|\nabla C_i(K_1) - \nabla C_i(K_2)\| \leqslant L\|K_1 - K_2\| \tag{13}$$

and, for some constant $\kappa > 1$ if $\lambda > \kappa L$ it follows that,

$$\|\nabla C_{\lambda,i}(K_1) - \nabla C_{\lambda,i}(K_2)\| \leqslant L_C\|K_1 - K_2\|, \tag{14}$$

where, $L_C := L/(\kappa - 1)$.

We have assumed a diversity (c.f. heterogeneity) bound for the gradients of LQR costs in Assumption 5. Still, our approach minimizes a cost function, the sum of Moreau regularized LQR costs. The estimate of the policy at $s+1$ outer iteration can be written as

$$K^{s+1} = K^s - \theta\nabla\widehat{C}_\lambda^s, \tag{15}$$

where, $\theta := \alpha\beta P$ and $\nabla\widehat{C}_\lambda^s := \frac{1}{VP}\sum_{i\in\mathcal{V}}\sum_{p=0}^{P}\nabla\widehat{C}_{\lambda,i}^{s,p}$, such that $\nabla\widehat{C}_{\lambda,i}^{s,p} := \lambda\big(K_i^{s,p} - \widehat{K}_i^{s,p}\big)$. The following proposition shows that Moreau regularized cost for a realization $i \in \mathcal{V}$ decreases during the iterations of the inner loop.

Proposition 8 *(Decrease in the Moreau Envelope cost): Let Assumptions 3, 4 and 5 hold, and $\alpha \leqslant 1/L_C$ with $P \geqslant 1$. Consider an iterate K^s inside Algorithm 1, such that $C_i(K^s) < \infty$, then $C_{\lambda,i}(K_i^{s,P}) \leqslant C_{\lambda,i}(K^s)$ and $C_i(K_i^{s,P}) < \infty$.*

Next, we present a corollary that establishes that the policy generated by local iterations for a system realization, when used for another system realization, incurs a finite cost.

Corollary 9 *(Finite cost by local policies): Let the conditions in Proposition 8 be satisfied. Then, consider the local estimate $K_i^{s,P}$ at a system realization $i \in \mathcal{V}$ for an outer iteration s of Algorithm 1, it follows that $C_j(K_i^{s,P}) < \infty$ for all $j \in \mathcal{V}$.*

Algorithm 1 depends on the gradient calculation at the current iterate, which requires the cost incurred at the iterate to be finite for all the available system realizations. The following result establishes this for the policies Algorithm 1 generates at every outer iteration.

Corollary 10 *(Finite cost at all iterations): Let the conditions in Proposition 8 be satisfied. Then, for any iterate K^s generated by Algorithm 1, $C_i(K^s) < \infty$ for all $i \in \mathcal{V}$.*

Remark 11 *Proposition 8, Corollary 9, and Corollary 10 are analogous to the stability results in [21, Theorem 1, Theorem 2]. Through these, we show that the cost incurred by policies generated inside Algorithm 1 remains finite for all available system realizations.*

The next theorem presents our main result; it shows the convergence of the iterates produced by Algorithm 1 to an approximate first-order stationary point of (9a).

Theorem 12 *(Convergence to a first-order stationary point): Let Assumptions 3, 4 and 5 hold, L and L_C from Lemma 7 with $\lambda > 2\sqrt{2}L$, $S \geqslant 16L_C^2(1+72\lambda^2)^2$, $P \geqslant 1$, $\alpha \leqslant 1/(2L_C P)$, $\beta \geqslant \max\{0.5, 2L_C/\sqrt{S}\}$, and $\theta = 1/\sqrt{S}$. Then, the sequence $\{K^s\}_{s=0}^S$ generated by Algorithm 1 has the following property:*

$$\frac{1}{4S}\sum_{s=0}^{S}\|\nabla C_\lambda(K^s)\|^2 \leqslant \frac{1}{\sqrt{S}}\big(C_\lambda(K^0) - C_\lambda(K^*)\big) + \frac{\Omega_2}{S} + \Omega_1 \tag{16}$$

where, $\Omega_1 = 2\lambda^2\delta^2$ and $\Omega_2 = 4\Big(\frac{2\lambda^2\delta^2}{P} + \frac{6\lambda^2}{\lambda^2 - 8L^2}\sigma_C^2\Big)$.

The following corollary shows the iteration complexity of Algorithm 1 by quantifying the number of outer iterations S and the required inner accuracy δ to reach a $\epsilon > 0$ accuracy in the approximate first-order stationary point.

Corollary 13 *Let Assumptions 3, 4, and 5 hold, and the parameters of Algorithm 1 are set as stated in Theorem 12. Moreover, let $0 < \epsilon \leqslant 1/(256L_C^4(1+72\lambda^2)^4)$, total number*

of iterations $S = \mathcal{O}(1/\epsilon^2)$ and the required inner accuracy $\delta = \mathcal{O}(\sqrt{\epsilon})$. Then, the output by Algorithm 1 is an $\epsilon-$optimal first-order stationary point of (9a).

Corollary 13 shows that as the required accuracy increases (i.e., ϵ decreases), there is a quadratic increase in the number of outer iterations required, and the required inner accuracy also increases as $\sqrt{\epsilon}$. Note that the convergence result in [21] is up to a non-vanishing constant ball around the optimal solution ([21, Theorem 3, Theorem 4]) while the precision of our method is arbitrary by the precision of the inner optimization problem.

IV. Numerical experiments

In this section, we present numerical experiments to illustrate the performance of the proposed MEMLQR algorithm. To this end, we divide our experiments into two parts; the first introduces a model for uncertain linear systems. We use this model to study the convergence properties of the MEMLQR algorithm and demonstrate adaptation to a system realization. In the second part, we compare our setup to that of [20] and [21]. We also present the results of a numerical experiment conducted based on the example provided in [21].

A. Uncertain linear system

We used the following model parameters for the experiments in (1). The uncertainties in the system are $a \sim \text{Unif}(-1,1)^2$ and $b \sim \text{Unif}(-1,1)^2$, where $\text{Unif}(-1,1)$ represents a uniform distribution between -1 and 1. The matrices generating the system are $\mathcal{A}_0 = \begin{bmatrix} 0.7 & -0.3 & 0.0 & 0.1 \\ 0.5 & -0.4 & 0.3 & 0.0 \\ 0.0 & 0.4 & 0.2 & -0.1 \\ 0.2 & 0.0 & 0.4 & 0.6 \end{bmatrix}$, $\mathcal{A}_1$ and $\mathcal{A}_2$ are lower and upper triangular matrices, respectively, with all entries as 0.1 and

$$\mathcal{B}_0 = \begin{bmatrix} 0.3 & 0.2 \\ 0.1 & 0.5 \\ 0.4 & 0.1 \\ 0.0 & 0.1 \end{bmatrix}, \mathcal{B}_1 = \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.1 \\ 0.1 & 0.0 \\ 0.0 & 0.1 \end{bmatrix}, \mathcal{B}_2 = \begin{bmatrix} 0.0 & 0.1 \\ 0.1 & 0.0 \\ 0.0 & 0.1 \\ 0.1 & 0.0 \end{bmatrix}.$$

The initial state is given by $x^0 \sim (\textit{unif}(-10,10))^4$ and the task is specified by a cost of form (3) where the cost matrices are given by $Q = \textit{diag}(1,2,3,4)$ and $R = \textit{diag}(1,2)$.

Four system realizations were generated using the parameters provided above. The MEMLQR algorithm was implemented with $S = 300$, $P = 2$, $\alpha = 0.1$, $\beta = 1$ and $\lambda = 0.2$. To solve the optimization problem in Step 5, we use the model-based gradient descent algorithm from [3]. We show the convergence characteristics of the algorithm by plotting the Moreau regularized cost $(C_\lambda(K^s))$ against the number of outer iterations (s) in Fig. 1a. Note that we slightly abuse our notations here by representing the cost incurred while adapting to a new system realization z as C_z. In Fig. 1b we plot the accuracy $\left(1 - |C_z(K^N) - C_z(K^n)|/C_z(K^N)\right)$ for the cost incurred while adapting to a new system, by initializing using the policy that minimizes the cumulative cost and the MEMLQR policy generated previously. It can be observed that the MEMLQR policy adapts faster to the optimal policy of the unseen realization z. We also plot in Fig. 1c the state and input trajectories of the unseen system realization z for the policy generated when a policy gradient algorithm is initialized at the MEMLQR policy.

B. Comparison to the MAML approach

To empirically compare the performance of the MAML-based approach to that of ours, we borrowed the example presented in [21]. We generated ten different system realizations and obtained meta policies using the MAML-LQR algorithm from [21] and our MEMLQR algorithm with three different values, i.e., $\lambda \in \{0.02, 0.2, 2\}$. These policies are evaluated by comparing the convergence characteristics while performing model-free policy gradient for three previously unseen system realizations. It can be observed from Fig. 2 that at initialization, the MEMLQR policy has a lower cost than the MAML-LQR policy for all three realizations, thus leading to a faster improvement in the cost incurred.

V. Conclusion

In this article, we explored personalization using Moreau envelopes to obtain an initialization for policy gradient algorithms (model-based/free) when dealing with uncertain linear systems, given that there is access to a finite number of uncertainty realizations. An iterative first-order framework, Algorithm 1, was presented to optimize this personalized cost function in a client-server setup, and its analysis showed the algorithm's convergence to a first-order stationary point. Numerical experiments were provided to empirically show the advantage of designing and using the proposed framework's meta-policy rather than the total cost. We compared the results of our approach to that of an approach based on the MAML framework and empirically showed that our approach incurred a lower cost for unseen scenarios.

References

[1] D. Bertsekas, *Dynamic programming and optimal control: Volume I*, vol. 4. Athena scientific, 2012.

[2] I. S. Khalil, J. C. Doyle, and K. Glover, *Robust and optimal control*. Prentice hall, 1996.

[3] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 1467–1476, PMLR, 2018.

[4] B. Gravell, P. M. Esfahani, and T. Summers, "Learning optimal controllers for linear systems with multiplicative noise via policy gradient," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5283–5298, 2020.

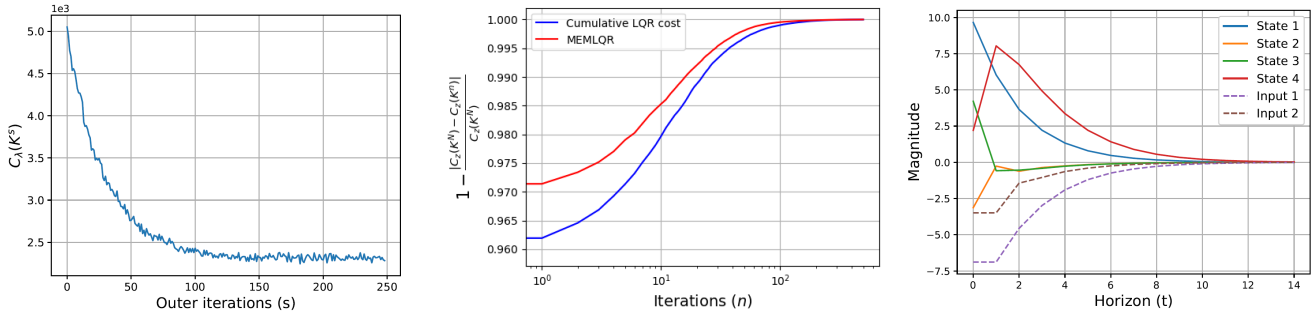
[5] B. Pang and Z. P. Jiang, "Robust reinforcement learning: A case study in linear quadratic regulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 9303–9311, 2021.

[6] G. Jing, H. Bai, J. George, A. Chakrabortty, and P. K. Sharma, "Learning distributed stabilizing controllers for multi-agent systems," *IEEE Control Systems Letters*, vol. 6, pp. 301–306, 2021.

[7] H. Wang, L. F. Toso, A. Mitra, and J. Anderson, "Model-free learning with heterogeneous dynamical systems: A federated LQR approach," *arXiv preprint arXiv:2308.11743*, 2023.

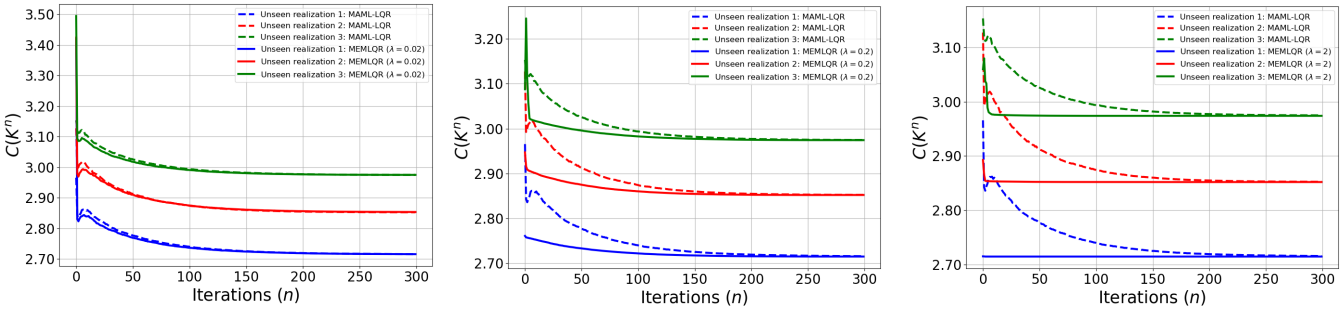
[8] M. Giegrich, C. Reisinger, and Y. Zhang, "Convergence of policy gradient methods for finite-horizon exploratory linear-quadratic control problems," *SIAM Journal on Control and Optimization*, vol. 62, no. 2, pp. 1060–1092, 2024.

[9] L. Sforni, G. Carnevale, I. Notarnicola, and G. Notarstefano, "Stability-certified on-policy data-driven lqr via recursive learning and policy gradient," *arXiv preprint arXiv:2403.05367*, 2024.



(a) (b) (c)

Fig. 1: Depiction of properties of Algorithm 1: (a) Convergence of the MEMLQR algorithm: evolution of the Moreau envelope regularized cost $C_\lambda(K^s)$ as the number of outer iterations (s) increase. (b) The evolution of accuracy $\left(1 - |C_z(K^N) - C_z(K^n)|/C_z(K^N)\right)$ of the policy generated for an unseen realization z by a model-based policy gradient framework when initialized using a policy obtained by minimizing the total LQR cost $C(\cdot)$ and by using the K^S generated by the MEMLQR algorithm. Note that the cost calculation here is over 50 randomly initial states. (c) State and input trajectories of the unseen realization z for the estimate of the optimal policy (K^N) generated after $N = 250$ iterations of the model-based policy gradient algorithm, which was initialized at K^S generated by the MEMLQR algorithm.



(a) (b) (c)

Fig. 2: Convergence to the optimal policy after initialization using the policy generated by MAML-LQR and MEMLQR ($\lambda = 0.02, 0.2, 2$) for three random system realizations in a model-free setting. It can be observed that the cost incurred by policy generated by the MEMLQR approach is closer to the optimal value initially, thus aiding in faster convergence. Also, it may be noted that for higher values of λ, the cost incurred is lower.

[10] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, pp. 1126–1135, PMLR, 2017.

[11] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Provably convergent policy gradient methods for model-agnostic meta-reinforcement learning," *arXiv preprint arXiv:2002.05135*, 2020.

[12] I. Clavera, J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel, "Model-based reinforcement learning via meta-policy optimization," in *Conference on Robot Learning*, pp. 617–629, PMLR, 2018.

[13] A. Nagabandi, I. Clavera, S. Liu, R. S. Fearing, P. Abbeel, S. Levine, and C. Finn, "Learning to adapt in dynamic, real-world environments through meta-reinforcement learning," *arXiv preprint arxiv:1803.11347*, 2019.

[14] A. Fallah, K. Georgiev, A. Mokhtari, and A. Ozdaglar, "On the convergence theory of debiased model-agnostic meta-reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3096–3107, 2021.

[15] M. T. Toghani, S. Lee, and C. A. Uribe, "PARS-Push: Personalized, asynchronous and robust decentralized optimization," *IEEE Control Systems Letters*, vol. 7, pp. 361–366, 2022.

[16] J. Beck, R. Vuorio, E. Z. Liu, Z. Xiong, L. Zintgraf, C. Finn, and S. Whiteson, "A survey of meta-reinforcement learning," *arXiv preprint arxiv:2301.08028*, 2024.

[17] M. T. Toghani, S. Lee, and C. A. Uribe, "PersA-FL: personalized asynchronous federated learning," *Optimization Methods and Software*, pp. 1–38, 2023.

[18] M. T. Toghani, S. P. Salazar, and C. A. Uribe, "On First-Order Meta-Reinforcement Learning with Moreau Envelopes," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 4176–4181, 2023.

[19] C. T. Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with Moreau envelopes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21394–21405, 2020.

[20] N. Musavi and G. E. Dullerud, "Convergence of gradient-based MAML in LQR," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 7362–7366, IEEE, 2023.

[21] L. F. Toso, D. Zhan, J. Anderson, and H. Wang, "Meta-learning linear quadratic regulators: A policy gradient MAML approach for the model-free LQR," *arXiv preprint arXiv:2401.14534*, 2024.

[22] I. Molybog and J. Lavaei, "When does MAML objective have benign landscape?," in *2021 IEEE Conference on Control Technology and Applications (CCTA)*, pp. 220–227, IEEE, 2021.

[23] A. Aravind, M. T. Toghani, and C. A. Uribe, "A Moreau envelope approach for LQR meta-policy estimation," *arXiv preprint arXiv:2403.17364*, 2024.