Proceedings of Machine Learning Research vol vvv:1–13, 2024

Meta-Learning Linear Quadratic Regulators: A Policy Gradient MAML Approach for the Model-free LQR

Leonardo F. Toso LT2879@COLUMBIA.EDU

Donglin Zhan DZ2478@COLUMBIA.EDU

James Anderson JAMES.ANDERSON@COLUMBIA.EDU

Han Wang HW2786@COLUMBIA.EDU

Columbia University, New York, NY

Abstract

We investigate the problem of learning Linear Quadratic Regulators (LQR) in a multi-task, heterogeneous, and model-free setting. We characterize the stability and personalization guarantees of a Policy Gradient-based (PG) Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) approach for the LQR problem under different task-heterogeneity settings[1]. We show that the MAML-LQR approach produces a stabilizing controller close to each task-specific optimal controller up to a task-heterogeneity bias for both model-based and model-free settings. Moreover, in the *model-based* setting, we show that this controller is achieved with a linear convergence rate, which improves upon sub-linear rates presented in existing MAML-LQR work. In contrast to existing MAML-LQR results, our theoretical guarantees demonstrate that the learned controller can efficiently *adapt* to unseen LQR tasks.

Keywords: Linear Quadratic Regulator; Model-Agnostic Meta-Learning; Model-free Learning

1. Introduction

One of the main successes of Reinforcement Learning (RL) (for example, in the context of robotics) is its ability to learn control policies that rapidly adapt to different agents and environments (Wang et al., 2016; Duan et al., 2016; Rothfuss et al., 2018). This idea of learning a control policy that efficiently adapts to unseen RL tasks is referred to as meta-learning, or learning to learn. The most popular approach is the Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017, 2019). In the context of RL, the role of MAML is to exploit task diversity from RL tasks drawn from a common task distribution to learn a control policy in a multi-task and heterogeneous setting that is only a few policy gradient (PG) steps away from an unseen task-specific optimal policy.

Despite its success in image classification and RL, more needs to be understood about the theoretical convergence guarantees of MAML for both model-based and model-free learning. This is due to the fact that, in general, the MAML objective is non-convex and requires a careful analysis depending on the considered task-setting (e.g., classification and regression (Fallah et al., 2020; Johnson and Mitra, 2011; Abbas et al., 2022; Ji et al., 2022), RL (Fallah et al., 2021; Liu et al., 2022; Beck et al., 2023)). There is a recent body of work on multi-task/agent learning for estimation (Zhang et al., 2023a; Wang et al., 2023a; Zhang et al., 2023b; Toso et al., 2023a; Chen et al., 2023) and control (Wang et al., 2023b; Tang et al., 2023; Wang et al., 2023c), where theoretical

1. This manuscript is a short version of our technical report. For any omitted details, please refer to its appendix.

guarantees are provided for different learning techniques in a variety of control settings. Therefore, with the purpose of characterizing the personalization guarantees of the MAML approach for a baseline and well-established control setting, we consider the model-free MAML-LQR problem.

In the optimal control domain, a highly desired feature is the fast adaptation of a designed controller to unseen situations during deployment, for example, in the setting where a manufacturer (of e.g., robots or drones) is responsible for designing optimal controllers for individual systems' objectives. Designing such controllers from scratch is sample-inefficient since it requires a large amount of trajectory data and several PG steps. Since manufacturers utilize production lines with the purpose of producing near identical systems, a controller should not need to be designed from scratch for every system. As such, the MAML-LQR approach exploits this similarity amongst systems within the same fabrication slot to design an LQR controller that *adapts* to fresh new slots of systems. This significantly reduces the amount of trajectory data required since it relies now on a simple fine-tuning step of the learned MAML-LQR controller to suit each system's objective.

Even for a simple discrete-time control setting, provably guaranteeing that a MAML-LQR approach produces a controller that adapts to unseen LQR tasks is not an easy endeavor and requires careful handling of the task heterogeneity and the stability of the sampled tasks under the learned controller. As well-established in the literature of PG methods for the LQR problem (Fazel et al., 2018; Malik et al., 2019; Gravell et al., 2020; Mohammadi et al., 2019; Hu et al., 2023), some properties of the LQR objective (e.g., gradient dominance and local smoothness) are crucial to derive global convergence guarantees. Although tempting, we cannot simply extend these guarantees to the MAML-LQR approach since those properties are no longer valid for the MAML-LQR objective (Molybog and Lavaei, 2021; Musavi and Dullerud, 2023).

In contrast to (Molybog and Lavaei, 2021; Musavi and Dullerud, 2023), this work establishes personalization guarantees for the MAML-LQR problem in both model-based and model-free settings. In particular, differently from (Molybog and Lavaei, 2021) that only characterizes the convergence of model-based MAML-LQR problem for the single-task setting, we consider the multi-task and heterogeneous setting where multiple task-heterogeneity scenarios are proposed. Moreover, the local convergence guarantees shown in (Musavi and Dullerud, 2023) provide little intuition on how the heterogeneity across the tasks may impact the convergence of the MAML-LQR approach and how efficiently the learned controller adapts to unseen tasks. Therefore, in this work, we address these points and provide meaningful personalization guarantees that support the ability of the learned controller to adapt to unseen LQR tasks under different task-heterogeneity settings.

Contributions: Toward this end, our main contributions are summarized as follows:

- In contrast to existing work on MAML-LQR, this is the first to provide personalization guarantees for model-based and model-free learning. Our convergence bounds characterize the distance between the learned and MAML-LQR optimal controller to each task-specific optimal controller and reveal the ultimate goal of the MAML-LQR, i.e., the adaptation to unseen tasks. In the *model-based* setting, we show that the learned controller is achieved with a linear convergence rate that improves upon existing sub-linear convergence rates in existing work.
- This is the first work to provide stability (Theorem 11) and convergence guarantees (Theorem 13) for the MAML-LQR approach in the model-free setting. In particular, our convergence guarantees demonstrate that the leaned controller stabilizes and is close to each task-specific optimal controller up to a task-heterogeneity bias. Furthermore, our analysis underscores the impact of different heterogeneity settings (i.e., system heterogeneity, cost heterogeneity, system, and cost heterogeneity) on the convergence of the MAML-LQR approach.

2. Model Agnostic Meta-Learning (MAML) for the LQR problem

Consider M discrete-time and linear time-invariant (LTI) dynamical systems

$$x_{t+1}^{(i)} = A^{(i)}x_t^{(i)} + B^{(i)}u_t^{(i)}, \quad t = 0, 1, 2, \dots, \tag{1}$$

where $A^{(i)} \in \mathbb{R}^{n_x \times n_x}$, $B^{(i)} \in \mathbb{R}^{n_x \times n_u}$, with $n_x \geq n_u$. The initial state of (1) is drawn from an arbitrary distribution $\mathcal{X}_0$ that satisfies $\mathbb{E}_{x_0^{(i)} \sim \mathcal{X}_0}[x_0^{(i)}] = 0$ and $\mathbb{E}_{x_0^{(i)} \sim \mathcal{X}_0}[x_0^{(i)} x_0^{(i)\top}] \succ \mu I_{n_x}$ where $\mu := \sigma_{\min}\left(\mathbb{E}_{x_0^{(i)} \sim \mathcal{X}_0} x_0^{(i)} x_0^{(i)\top}\right)$ for all $i \in [M]$[2]. The objective of the LQR problem is to design an optimal control sequence $u_t^{(i)} := -K_i^\star x_t^{(i)}$ that minimizes a quadratic cost in both state $x_t^{(i)}$ and input $u_t^{(i)}$. The optimal controllers $K_i^\star$ solve

$$K_i^\star := \underset{K \in \mathcal{K}^{(i)}}{\operatorname{argmin}} \left\{ J^{(i)}(K) := \mathbb{E}_{x_0^{(i)} \sim \mathcal{X}_0} \left[\sum_{t=0}^{\infty} x_t^{(i)\top} \left(Q^{(i)} + K^\top R^{(i)} K \right) x_t^{(i)} \right] \right\}, \text{ s.t (1)}, \tag{2}$$

where $Q^{(i)} \in \mathbb{S}_{\succ 0}^{n_x}$, $R^{(i)} \in \mathbb{S}_{\succ 0}^{n_u}$, and $\mathcal{K}^{(i)} := \{K \mid \rho(A^{(i)} - B^{(i)}K) < 1\}$ denotes the set of stabilizing controllers of the i-th system, and $\rho(\cdot)$ denotes the spectral radius of a square matrix.

Definition 1 *The LQR task is a tuple $\mathcal{T}^{(i)} := (A^{(i)}, B^{(i)}, Q^{(i)}, R^{(i)})$ equipped with the objective of designing $K_i^\star$ that minimizes the LQR cost $J^{(i)}(K)$.*

Consider a distribution of LQR tasks denoted by $p(\mathcal{T})$ from which a collection of M LQR tasks $\mathcal{T} := \{\mathcal{T}^{(i)}\}_{i=1}^M$ are sampled. The objective of the MAML approach for the LQR problem is to design a controller $K_{\text{ML}}^\star$ that can *efficiently adapt* to any unseen LQR task originating from $p(\mathcal{T})$, i.e., we aim to find a controller that is only a few PG iterations away from any unseen task-specific optimal controller. Precisely, $K_{\text{ML}}^\star$ solve

$$K_{\text{ML}}^\star := \underset{K \in \bar{\mathcal{K}}}{\operatorname{argmin}} \left\{ J_{\text{ML}}(K) := \frac{1}{M} \sum_{i=1}^{M} J^{(i)}\left(K - \eta_l \nabla J^{(i)}(K)\right) \right\}, \text{ s.t (1) } \forall i \in [M], \tag{3}$$

where $\bar{\mathcal{K}} := \cap_{i \in [M]} \mathcal{K}^{(i)}$ is the MAML stabilizing set and η_l denotes some positive step-size. To solve (3), we exploit a PG-based approach where the update rule is described as follows:

$$K \leftarrow K - \eta \nabla J_{\text{ML}}(K), \text{ where } \nabla J_{\text{ML}}(K) := \frac{1}{M} \sum_{i=1}^{M} H^{(i)}(K) \nabla J^{(i)}(K - \eta_l \nabla J^{(i)}(K)), \tag{4}$$

with $H^{(i)}(K) := I_{n_u} - \eta_l \nabla^2 J^{(i)}(K)$, and η being some positive step-size. Next, we define the task-specific and MAML stabilizing sub-level sets.

Definition 2 *(Stabilizing sub-level set) The task-specific and MAML stabilizing sub-level sets are defined as follows:*

- *Given a task $\mathcal{T}^{(i)}$, the task-specific sub-level set $\mathcal{S}^{(i)} \subseteq \mathcal{K}^{(i)}$ is*

$$\mathcal{S}^{(i)} := \left\{ K \mid J^{(i)}(K) - J^{(i)}(K_i^\star) \leq \gamma_i \Delta_0^{(i)} \right\}.$$

  with $\Delta_0^{(i)} = J^{(i)}(K_0) - J^{(i)}(K_i^\star)$, for any $\gamma^{(i)} \geq 1$.

2. This assumption is standard in PG methods for the LQR problem (Fazel et al., 2018; Malik et al., 2019; Bu et al., 2019; Gravell et al., 2020) and it guarantees that all stationary points are global optima.

- *The MAML-LQR stabilizing sub-level set $\mathcal{S}_{ML} \subseteq \bar{\mathcal{K}}$ is defined as the intersection between each task-specific stabilizing sub-level set, i.e., $\mathcal{S}_{ML} := \cap_{i\in[M]}\mathcal{S}^{(i)}$.*

Remark 3 *Observe that, if $K \in \mathcal{S}_{ML}$, i.e., K stabilizes all LQR tasks in $\mathcal{T}$, one may select a step-size η_l, such that $\bar{K} = K - \eta_l \nabla J^{(i)}(K)$ also stabilizes all the LQR tasks in $\mathcal{T}$, i.e., $\bar{K} \in \mathcal{S}_{ML}$. We prove this fact and provide the condition on η_l to satisfy it in our stability analysis in Theorem 10.*

Assumption 1 *We have access to an initial stabilizing controller $K_0 \in \mathcal{S}_{ML}$.*

Remark 4 *The above assumption is standard in PG methods for the LQR problem (Fazel et al., 2018; Gravell et al., 2020; Wang et al., 2023b; Toso et al., 2023b). If the initial controller K_0 fails to stabilize (1), $\forall i \in [M]$, the MAML-LQR update in (4) cannot produce stabilizing controller, since $\nabla J^{(i)}(K_0)$, $\nabla^2 J^{(i)}(K_0)$ are both undefined for the unstabilized tasks. Moreover, (Perdomo et al., 2021; Ozaslan et al., 2022) shows how to find an initial stabilizing controller.*

As well-established in the literature of PG-LQR $J^{(i)}(K)$ is, in general, non-convex with respect to K. However, by leveraging some properties of the LQR cost (e.g., gradient domination and local smoothness), (Fazel et al., 2018) provides global convergence guarantees of PG methods for both model-based and model-free LQR settings. Although tempting, these properties of the LQR cost cannot simply be extended to the MAML-LQR objective when dealing with task heterogeneity.

In the sequel, we proceed as follows: We provide conditions on the problem parameters to ensure that given any stabilizing controller $K \in \mathcal{S}_{\text{ML}}$, $K - \eta\nabla J_{\text{ML}}(K)$ is also MAML stabilizing, i.e., $K - \eta\nabla J_{\text{ML}}(K) \in \mathcal{S}_{\text{ML}}$. In contrast to (Musavi and Dullerud, 2023) that provides the guarantees for which a model-based MAML-LQR approach finds a stationary solution, we derive meaningful convergence bounds for both model-based and model-free learning. In particular, our convergence guarantees underscore the impact of multiple task-heterogeneity settings on the closeness of learned controller and each task-specific optimal controller, and demonstrate the adaptation of the learned controller to unseen tasks.

2.1. Model-based LQR

In the model-based LQR setting, we assume to have access to the tuple $\mathcal{T}^{(i)} = (A^{(i)}, B^{(i)}, Q^{(i)}, R^{(i)})$. With the ground-truth model in hand, we have closed-form expressions to compute both gradient $\nabla J^{(i)}(K)$ and Hessian $\nabla^2 J^{(i)}(K)$ of the LQR cost.

- **Gradient of the LQR cost** $\nabla J^{(i)}(K)$ (Fazel et al., 2018): Given $\mathcal{T}^{(i)}$ and a stabilizing controller $K \in \mathcal{S}_{\text{ML}}$, the gradient is given by $\nabla J^{(i)}(K) := 2E_K^{(i)}\Sigma_K^{(i)}$, where

$$E_K^{(i)} := R^{(i)}K - B^{(i)\top}P_K^{(i)}(A^{(i)} - B^{(i)}K), \quad \Sigma_K^{(i)} := \mathbb{E}_{x_0^{(i)}\sim\mathcal{X}_0}\sum_{t=0}^{\infty} x_t^{(i)}x_t^{(i)\top},$$

  with $x_t^{(i)}$ subject to the system dynamics in (1), and $P_K^{(i)} \in \mathbb{S}_{\succ 0}^{n_x}$ denoting the solution of the Lyapunov equation $P_K^{(i)} := Q^{(i)} + K^\top R^{(i)}K + (A^{(i)} - B^{(i)}K)^\top P_K^{(i)}(A^{(i)} - B^{(i)}K)$.
- **Hessian of the LQR cost** $\nabla^2 J^{(i)}(K)$ (Bu et al., 2019): Given $\mathcal{T}^{(i)}$ and a stabilizing controller $K \in \mathcal{S}_{\text{ML}}$, the Hessian operator at K acting on some $X \in \mathbb{R}^{n_u\times n_x}$, is given by

$$\nabla^2 J^{(i)}(K)[X] := 2\left(R^{(i)} + B^{(i)\top}P_K^{(i)}B^{(i)}\right)X\Sigma_K^{(i)} - 4B^{(i)\top}\tilde{P}_K^{(i)}[X](A^{(i)} - B^{(i)}K)\Sigma_K^{(i)},$$

  with $\tilde{P}_K^{(i)}[X] := (A^{(i)} - B^{(i)}K)^\top\tilde{P}_K^{(i)}[X](A^{(i)} - B^{(i)}K) + X^\top E_K^{(i)} + E_K^{(i)\top}X$.

Hence, we can exploit these closed-form expressions in order to perform (4) in Algorithm 1 for the model-based MAML-LQR. In step 4, this algorithm computes an one-step inner gradient descent iteration on K_n and $H^{(i)}(K_n) = I_{n_u} - \eta\nabla^2 J^{(i)}(K_n)$, for each task $i \in [M]$ and iteration n. These quantities are then used to update the controller K_{n+1} in step 6. By repeating this process for N iterations, Algorithm 1 returns K_N. We further prove that K_N is close to each task-specific optimal controller $K_i^\star$, which in turn is proved to be close to the MAML-LQR optimal controller $K_{\text{ML}}^\star$. To rigorously prove those claims, we first revisit some properties of the LQR cost function.

Algorithm 1 `MAML-LQR`: Model-Agnostic Meta-Learning for LQR tasks **(Model-based)**

1: **Input:** initial stabilizing controller K_0, inner and outer step-sizes η_l, η
2: **for** $n = 0, \ldots, N-1$ **do**
3: **for** each task $i \in [M]$ in $\mathcal{T}$ **compute**
4: $\bar{K}_n^{(i)} = K_n - \eta_l \nabla J^{(i)}(K_n)$, and $H^{(i)}(K_n) = I_{n_u} - \eta_l \nabla^2 J^{(i)}(K_n)$
5: **end for**
6: $K_{n+1} = K_n - \frac{\eta}{M}\sum_{i=1}^{M} H^{(i)}(K_n)\nabla J^{(i)}(\bar{K}_n^{(i)})$
7: **end for**
8: **Output:** K_N

Lemma 5 *(Uniform bounds) Given $\mathcal{T}^{(i)}$ and a stabilizing controller $K \in \mathcal{S}_{ML}$, the Frobenius norm of the gradient $\nabla J^{(i)}(K)$, Hessian $\nabla^2 J^{(i)}(K)$ and controller K are bounded as follows:*

$$\|\nabla J^{(i)}(K)\|_F \leq h_G(K),\ \|\nabla^2 J^{(i)}(K)\|_F \leq h_H(K),\ \text{and}\ \|K\|_F \leq h_c(K),$$

where $h_G(K)$, $h_H(K)$, and $h_c(K)$ are polynomials on the problem parameters.

Lemma 6 *(Local smoothness) Given $\mathcal{T}^{(i)}$ and two stabilizing controllers $K, K' \in \mathcal{S}_{ML}$ such that $\|\Delta\| := \|K' - K\| \leq h_\Delta(K) < \infty$. The LQR cost, gradient and Hessian satisfy:*

$$\begin{aligned}
&\left|J^{(i)}\left(K'\right) - J^{(i)}(K)\right| \leq h_{cost}(K) J^{(i)}(K)\|\Delta\|_F,\\
&\left\|\nabla J^{(i)}\left(K'\right) - \nabla J^{(i)}(K)\right\|_F \leq h_{grad}(K)\|\Delta\|_F,\\
&\left\|\nabla^2 J^{(i)}\left(K'\right) - \nabla^2 J^{(i)}(K)\right\|_F \leq h_{hess}(K)\|\Delta\|_F
\end{aligned}$$

where $h_\Delta(K)$, $h_{cost}(K)$, $h_{hess}(K)$ and $h_{grad}(K)$ are polynomials on the problem parameters.

Lemma 7 *(Gradient Domination) Given $\mathcal{T}^{(i)}$ and a stabilizing controller $K \in \mathcal{S}_{ML}$. Let $K_i^\star$ be the optimal controller of task $\mathcal{T}^{(i)}$. Then, it holds that*

$$J^{(i)}(K) - J^{(i)}(K_i^\star) \leq \frac{1}{\lambda_i}\|\nabla J^{(i)}(K)\|_F^2$$

where $\lambda_i := 4\mu^2\sigma_{\min}(R^{(i)})/\left\|\Sigma_{K_i^\star}\right\|$.

The uniform bounds of $\|\nabla J^{(i)}(K)\|_F$ and $\|K\|_F$, and the gradient domination property are proved in (Fazel et al., 2018; Wang et al., 2023a). Moreover, the uniform bound of $\|\nabla^2 J^{(i)}(K)\|_F$ can be found in (Bu et al., 2019, Lemma 7.9). In addition, the proofs for the local smoothness of the

cost and gradient are detailed in (Wang et al., 2023b, Appendix F), whereas the local smoothness of the Hessian is proved in (Musavi and Dullerud, 2023, Appendix B). The explicit expressions of $h_G(K)$, $h_c(K)$, $h_H(K)$, $h_\Delta(K)$, $h_{\text{cost}}(K)$, and $h_{\text{grad}}(K)$ are revisited in our technical report (Toso et al., 2023c). Throughout the paper, we use $\bar{h} := \sup_{K\in\mathcal{S}_{\text{ML}}} h(K)$ and $\underline{h} := \inf_{K\in\mathcal{S}_{\text{ML}}} h(K)$ to denote the supremum and infimum of a polynomial $h(K)$ over the set of stabilizing controllers $\mathcal{S}_{\text{ML}}$.

2.2. Task Heterogeneity

In contrast to (Musavi and Dullerud, 2023), we consider multiple task-heterogeneity settings. We do so to understand the impact of different types of heterogeneity in the convergence of the MAML-LQR approach for both model-based and model-free settings. In particular, we consider a general task-heterogeneity setting characterized by the combination of system and cost heterogeneity. That is, we assume that there exist positive scalars ϵ_1, ϵ_2, ϵ_3 and ϵ_4, such that

$$\max_{i\neq j}\|A^{(i)}-A^{(j)}\|\leq\epsilon_1,\max_{i\neq j}\|B^{(i)}-B^{(j)}\|\leq\epsilon_2,\max_{i\neq j}\|Q^{(i)}-Q^{(j)}\|\leq\epsilon_3,\max_{i\neq j}\|R^{(i)}-R^{(j)}\|\leq\epsilon_4.$$

Observe that this setting spans three different types of task heterogeneity: 1) system heterogeneity, with $\epsilon_3=\epsilon_4=0$, i.e., $Q^{(i)}=Q$, $R^{(i)}=R$, $\forall i\in[M]$. 2) cost heterogeneity, with $\epsilon_1=\epsilon_2=0$, i.e., $A^{(i)}=A$, $B^{(i)}=B$, $\forall i\in[M]$. 3) system and cost heterogeneity, where ϵ_1, ϵ_2, ϵ_3 and ϵ_4 are non-zero. Next, we bound the norm of the gradient difference between two distinct tasks.

Lemma 8 *(Gradient heterogeneity) For any two distinct LQR tasks $\mathcal{T}^{(i)}=(A^{(i)},B^{(i)},Q^{(i)},R^{(i)})$ and $\mathcal{T}^{(j)}=(A^{(j)},B^{(j)},Q^{(j)},R^{(j)})$ and stabilizing controller $K\in\mathcal{S}_{ML}$. It holds that,*

$$\|\nabla J^{(i)}(K)-\nabla J^{(i)}(K)\|\leq f_z(\bar{\epsilon}):=\epsilon_1 h^1_{het}(K)+\epsilon_2 h^2_{het}(K)+\epsilon_3 h^3_{het}(K)+\epsilon_4 h^4_{het}(K),\tag{5}$$

for any $i\neq j\in[M]$, where $z\in\{1,2,3\}$, and $\bar{\epsilon}=\{\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4\}$, where $h^1_{het}(K),h^2_{het}(K),h^3_{het}(K)$, and $h^4_{het}(K)$ are positive polynomials that depend on the problem parameters[3].

The proof and explicit expressions of $h^1_{\text{het}}(K),h^2_{\text{het}}(K),h^3_{\text{het}}(K)$, and $h^4_{\text{het}}(K)$, are detailed in our technical report (Toso et al., 2023c). We observe that as long as the heterogeneity level $\epsilon_1,\epsilon_2,\epsilon_3,\epsilon_4$ is small, the gradient descent direction of task $\mathcal{T}^{(i)}$ (4) is close to the one of task $\mathcal{T}^{(j)}$. Moreover, by combining (5) along with the uniform bound of the Hessian (Lemma 5), we observe that $\nabla J_{\text{ML}}(K)$ is also close to $\nabla J^{(i)}(K)$. This is a crucial step we use in our stability and convergence analysis for both model-based and model-free settings.

2.3. Model-free LQR

We now consider the setting where the tuple $\mathcal{T}^{(i)}=(A^{(i)},B^{(i)},Q^{(i)},R^{(i)})$ is unknown. Therefore, computing gradient and Hessian through closed-form expressions is no longer possible. This leads to resorting to methods that approximate such quantities. Following numerous work in the literature of model-free PG-LQR (Fazel et al., 2018; Malik et al., 2019; Gravell et al., 2020; Mohammadi et al., 2019; Wang et al., 2023b; Toso et al., 2023b), we focus on zeroth-order methods to estimate the gradient and Hessian of the LQR cost. In particular, we consider a two-point estimation scheme since it has a lower estimation variance compared to its one-point counterpart (Malik et al., 2019).

3. We use z to denote the type of heterogeneity, namely, $z=1$ refers to system heterogeneity, $z=2$ to cost heterogeneity and $z=3$ system and cost heterogeneity.

Zeroth-order methods with two-point estimation solely rely on querying cost values at symmetric perturbed controllers to construct a biased estimation of both gradient and Hessian. In particular, zeroth-order estimation is a Gaussian smoothing approach (Nesterov and Spokoiny, 2017) based on Stein's identity (Stein, 1972) that relates gradient and Hessian to cost queries.

Algorithm 2 `ZO2P`: Zeroth-order with two-point estimation

1: **Input:** controller K, number of samples m and smoothing radius r,
2: **for** $l = 1, \dots, m$ **do**
3: Sample controllers $K_l^1 = K + U_l$ and $K_l^2 = K - U_l$, where U_l is drawn uniformly at random over matrices with Frobenius norm r.
4: **end for**
5: $\widehat{\nabla} J(K) = \frac{n_x n_u}{2r^2 m} \sum_{l=1}^{m} (J(K_l^1) - J(K_l^2)) U_l$,
6: $\widehat{\nabla}^2 J(K) = \frac{n_u^2}{r^2 m} \sum_{l=1}^{m} (J(K_l^1) - J(K)) (U_l U_l^\top - I_{n_u})$,
7: **Return** $\widehat{\nabla} J(K)$, $\widehat{\nabla}^2 J(K)$.

Algorithm 2 describes the zeroth-order estimation of the gradient and Hessian of the LQR cost. Firstly, a pair of symmetric perturbed controllers are sampled $K^1 = K + U$, $K^2 = K - U$ by perturbing $K \in \mathcal{S}_{\text{ML}}$ with $U \sim \mathbb{S}_r$, where $\mathbb{S}_r$ denotes a distribution of $n_u \times n_x$ real matrices such that $\|U\|_F = r$, where r is referred to as smoothing radius. A cost oracle provides the true cost assessed at the original and perturbed controllers. The gradient is estimated via the empirical value of the first-order Gaussian Stein's identity, $\mathbb{E}\left[\nabla J^{(i)}(K)\right] = \mathbb{E}\left[\frac{n_x n_u}{2r^2}(J^{(i)}(K^1) - J^{(i)}(K^2))U\right]$ (Mohammadi et al., 2020; Toso et al., 2023b), and the Hessian with its second-order counterpart, $\mathbb{E}\left[\nabla^2 J^{(i)}(K)\right] = \mathbb{E}\left[\frac{n_u^2}{r^2}(J^{(i)}(K^1) - J^{(i)}(K))(UU^\top - I_{n_u})\right]$, (Balasubramanian and Ghadimi, 2022), both with m samples.

Remark 9 *(Cost oracle) For simplicity, we assume to have access to the true cost, provided by a cost oracle, as in (Malik et al., 2019; Toso et al., 2023b). We emphasize that our work can be readily extended to the setting where only finite-horizon approximation of the cost is available. This is the case since the finite-horizon approximation of the true cost is upper-bounded by its true value, with the approximation error controlled by the horizon length, as in (Gravell et al., 2020, Appendix B).*

With the gradient and Hessian zeroth-order estimators in hand, Algorithm 3 follows the same structure as Algorithm 1, except for steps 4 and 7 where the gradient and Hessian computations are replaced by a zeroth-order estimation. Although the presence of an error in these estimations may impact the convergence of Algorithm 3, we careful control this error in our theoretical analysis.

3. Theoretical Guarantees

We now provide the theoretical guarantees for the stability and convergence of the MAML-LQR for both model-based and model-free settings, i.e., Algorithms 1 and 3.

3.1. Stability Analysis

The objective of the stability analysis is to provide the conditions on the step-sizes η_l, η, heterogeneity $\bar{f}_z(\bar{\epsilon})$, and zeroth-order estimation parameters m and r, such that for every iteration of Algorithms 1 and 3, the current obtained controller is MAML-LQR stabilizing, i.e., $K_n \in \mathcal{S}_{\text{ML}}$, $\forall n$.

Algorithm 3 `MAML-LQR`: Model-Agnostic Meta-Learning for LQR tasks **(Model-free)**

Input: initial stabilizing controller K_0, inner and outer step-sizes η_l, η, smoothing radius r, number of samples m.
for $n = 0, \dots, N-1$ **do**
  for each task $i \in [M]$ in $\mathcal{T}$ **compute**
    $\left[\widehat{\nabla} J^{(i)}(K_n), \widehat{\nabla}^2 J^{(i)}(K_n)\right] = \texttt{ZO2P}(K_n, m, r)$
    $\widehat{K}_n^{(i)} = K_n - \eta_l \widehat{\nabla} J^{(i)}(K_n)$, and $\widehat{H}^{(i)}(K_n) = I_{n_u} - \eta_l \widehat{\nabla}^2 J^{(i)}(K_n)$
  end for
  $\widehat{\nabla} J^{(i)}(\widehat{K}_n^{(i)}) = \texttt{ZO2P}(\widehat{K}_n^{(i)}, m, r)$
  $K_{n+1} = K_n - \frac{\eta}{M} \sum_{i=1}^{M} \widehat{H}^{(i)}(K_n) \widehat{\nabla} J^{(i)}(\widehat{K}_n^{(i)})$
end for
Output: K_N

Theorem 10 *(Model-based) Given an initial stabilizing controller $K_0 \in \mathcal{S}_{ML}$. We suppose that the step-sizes and heterogeneity satisfy $\eta_l \leq \min\left\{\frac{n_u}{\sqrt{2}h_H}, \frac{1}{\sqrt{2}\bar{h}_{grad}}, \frac{1}{\sqrt{12(12\bar{h}_{grad}^2 n_u^2 + \bar{h}_H^2)}}\right\}$, $\eta \leq \frac{1}{4\bar{h}_{grad}}$ and $\bar{f}_z(\bar{\epsilon}) \leq \sqrt{\min_i \frac{\lambda_i \Delta_0^{(i)}}{288 n_u^3}}$, respectively. Then, $\bar{K}_n, K_n \in \mathcal{S}_{ML}$, for every iteration of Algorithm 1.*

Theorem 11 *(Model-free) Given an initial stabilizing controller $K_0 \in \mathcal{S}_{ML}$ and probability δ. Suppose that the step-sizes satisfy $\eta_l \leq \min\left\{\frac{1}{h_G}, \frac{n_u}{h_H}, \frac{1}{\bar{h}_{grad}}, \frac{1}{\sqrt{20(12\bar{h}_{grad}^2 n_u^2 + \bar{h}_H^2)}}, \frac{1}{2}\right\}$, $\eta \leq \frac{1}{8\bar{h}_{grad}}$. In addition, the heterogeneity, smoothing radius and number of samples satisfy $\bar{f}_z(\bar{\epsilon}) \leq \sqrt{\min_i \frac{\lambda_i \Delta_0^{(i)}}{480 n_u^3}}$, $r \leq \min\left\{\underline{h}_r^1\left(\frac{\sqrt{\psi^{(i)}}}{2}\right), \underline{h}_r^2\left(\frac{\sqrt{\psi^{(i)}}}{2}\right)\right\}$, and $m \geq \max\left\{\bar{h}_m^1\left(\frac{\sqrt{\psi^{(i)}}}{2}, \delta\right), \bar{h}_m^2\left(\frac{\sqrt{\psi^{(i)}}}{2}, \delta\right)\right\}$[4], with $\psi^{(i)} := \frac{\lambda_i \Delta_0^{(i)}}{1296}$. Then, with probability, $1-\delta$, $\bar{K}_n, K_n \in \mathcal{S}_{ML}$, for every iteration of Algorithm 3.*

The proof of Theorems 10 and 11 are detailed in (Toso et al., 2023c). The proof strategy follows from an induction argument where the base case is the first iteration. We combine the local smoothness of each task-specific LQR cost (Lemma 6) along with the gradient heterogeneity bound (Lemma 8) and the definition of the MAML-LQR stabilizing sub-level set to show that $J^{(i)}(K_1) \leq J^{(i)}(K_0)$ for any $i \in [M]$. These results provide the conditions for which the learned controller K_N is MAML stabilizing. This is essential to guarantee that the learned controller K_N in Algorithms 1 and 3 can be promptly utilized to stabilize an unseen LQR task drawn from $p(\mathcal{T})$.

3.2. Convergence Analysis

We now provide the conditions on the step-sizes η_l, η and zeroth-order estimation parameters m, and r, such that we can ensure that the learned MAML-LQR controller K_N is sufficiently close to each task-specific optimal controller $K_i^\star$ and to the optimal MAML controller $K_{\text{ML}}^\star$. For this purpose, we study the closeness of K_N and $K_i^\star$ by bounding $J^{(i)}(K_N) - J^{(i)}(K_i^\star)$ and the closeness of $K_{\text{ML}}^\star$ with $J^{(i)}(K_{\text{ML}}^\star) - J^{(i)}(K_i^\star)$.

4. The expressions of the positive polynomials $\underline{h}_r^1(\cdot), \underline{h}_r^2(\cdot), \bar{h}_m^1(\cdot)$ and $\bar{h}_m^2(\cdot)$ are deferred to our technical report.

Theorem 12 *(Model-based) Given an initial stabilizing controller $K_0 \in \mathcal{S}_{ML}$. Suppose that the step-sizes and number of iterations satisfy $\eta_l \leq \min\left\{\frac{n_u}{\sqrt{2}\bar{h}_H}, \frac{1}{\sqrt{2}\bar{h}_{grad}}, \frac{1}{\sqrt{12(12\bar{h}_{grad}^2 n_u^2 + h_H^2)}}\right\}$, and $\eta \leq \frac{1}{4\bar{h}_{grad}}$, $N \geq \frac{8}{\eta\lambda_i}\log\left(\frac{\Delta_0^{(i)}}{\epsilon'}\right)$, respectively, for some small tolerance $0 < \epsilon' < 1$. Then, it holds that,*

$$J^{(i)}(K_N) - J^{(i)}(K_i^\star) \leq \epsilon' + \frac{144 n_u^3 \bar{f}_z^2(\bar{\epsilon})}{\lambda_i},$$
$$J^{(i)}(K_{ML}^\star) - J^{(i)}(K_i^\star) \leq n_u^3 \bar{f}_z^2(\bar{\epsilon})\left(\frac{144}{\lambda_i} + \frac{96\bar{J}_{\max}}{\mu^2 \min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)})}\right), \tag{6}$$

with $\bar{J}_{\max} := \max_i J^{(i)}(K_0)$.

Theorem 13 *(Model-free) Given an initial stabilizing controller $K_0 \in \mathcal{S}_{ML}$ and probability δ. Suppose that the step-sizes, smoothing radius, number of samples and number of iterations satisfy $\eta_l \leq \min\left\{\frac{1}{\bar{h}_G}, \frac{n_u}{\bar{h}_H}, \frac{1}{\bar{h}_{grad}}, \frac{1}{\sqrt{20(12\bar{h}_{grad}^2 n_u^2 + h_H^2)}}, \frac{1}{2}\right\}$, $\eta \leq \frac{1}{8\bar{h}_{grad}}$, $r \leq \min\left\{\underline{h}_r^1\left(\frac{\epsilon'\lambda_i}{1296}\right), \underline{h}_r^2\left(\frac{\epsilon'\lambda_i}{1296}\right)\right\}$, $m \geq \max\left\{\bar{h}_m^1\left(\frac{\epsilon'\lambda_i}{1296}, \delta\right), \bar{h}_m^2\left(\frac{\epsilon'\lambda_i}{1296}, \delta\right)\right\}$, and $N \geq \frac{8}{\eta\lambda_i}\log\left(\frac{2\Delta_0^{(i)}}{\epsilon'}\right)$, respectively, for some small tolerance $0 < \epsilon' < 1$. Then, with probability $1 - \delta$, it holds that,*

$$J^{(i)}(K_N) - J^{(i)}(K_i^\star) \leq \epsilon' + \frac{240 n_u^3 \bar{f}_z^2(\bar{\epsilon})}{\lambda_i},$$
$$J^{(i)}(K_{ML}^\star) - J^{(i)}(K_i^\star) \leq n_u^3 \bar{f}_z^2(\bar{\epsilon})\left(\frac{240}{\lambda_i} + \frac{96\bar{J}_{\max}}{\mu^2 \min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)})}\right). \tag{7}$$

The proofs of Theorems 12 and 13 are deferred to (Toso et al., 2023c). The proof strategy follows from the local smoothness of the LQR cost (Lemma 6), gradient domination (Lemma 7), and the gradient heterogeneity bound (Lemma 8). The model-free setting also involves controlling the estimation error through a matrix Bernstein-type of inequality (Tropp, 2012; Gravell et al., 2020).

These results characterize the convergence of the MAML-LQR for both model-based and model-free settings. We emphasize that both Algorithms 1 and 3 produce a controller K_N that is provably close to each task-specific optimal controller $K_i^\star$ up to a heterogeneity bias. This indicates that under a low heterogeneity regime K_N will serve as a good initialization for any unseen task that is also drawn from $p(\mathcal{T})$ (i.e., an unseen task that satisfy the same task-heterogeneity level as the ones used in the MAML-LQR learning process). Moreover, in contrast to (Musavi and Dullerud, 2023), our convergence bounds (6) and (7) emphasize the impact of task heterogeneity on the convergence of Algorithms 1 and 3. In a low heterogeneity regime, where K_N and $K_i^\star$, and $K_{\text{ML}}^\star$ and $K_i^\star$ are close, one may conclude that K_N and $K_{\text{ML}}^\star$ are also sufficiently close. We also emphasize that, in the model-based setting, the learned controller K_N is achieved with linear convergence rate on the iteration count which improves upon the sub-linear rate in (Musavi and Dullerud, 2023).

4. Experimental Results

Numerical results are now provided to illustrate and assess the convergence and personalization of the model-free MAML-LQR approach. In particular, we show that initializing from the learned

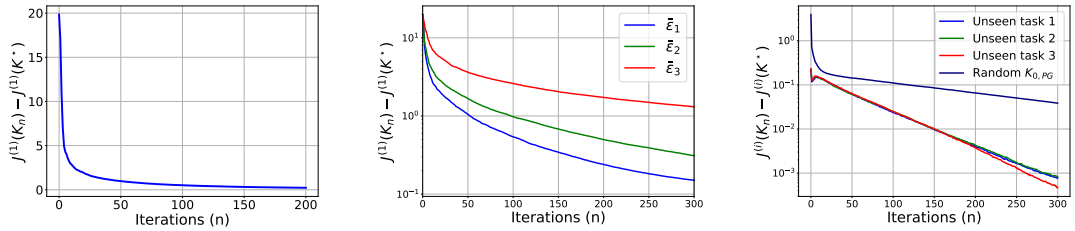

Figure 1: Cost gap between the learned the task-specific optimal controller with respect to iteration. (left) Convergence of the `MAML-LQR`. (middle) `MAML-LQR`, $\bar{\epsilon}_1 = (1.2, 1.1, 1.4, 1.2) \times 10^{-3}$, $\bar{\epsilon}_2 = (1.3, 1.1, 1.4, 1.2) \times 10^{-2}$, $\bar{\epsilon}_3 = (1.7, 1.8, 1.9, 1.7) \times 10^{-2}$. (right) `PG-LQR` (Fazel et al., 2018).

MAML-LQR controller (i.e., $K_{0,\text{PG}} = K_N$) enables a model-free PG-LQR approach (Fazel et al., 2018, Section 4.2) to be close to the task-specific optimal controller within just a few PG iterations, for an unseen task. To illustrate so, we consider an unstable modification of the Boeing system from (Hong et al., 2021) to define a nominal LQR task, which is used to generate multiple tasks. The technical details regarding the experimental setup are deferred to (Toso et al., 2023c).

Figure 1 depicts the cost gap between the current learned controller and the nominal task (i.e., $\mathcal{T}^{(1)}$) optimal controller with respect to iterations of Algorithm 3. In alignment with Theorem 13, Figure 1-(left) shows that the learned controller closely converges to the nominal task's optimal controller up to a small bias characterized by $\bar{\epsilon} = (1.2, 1.1, 1.4, 1.2) \times 10^{-3}$. Moreover, Figure 1-(middle) shows that the learned MAML-LQR controller drastically deviates from the nominal task's optimal controller when it faces a large heterogeneity level. This aligns with Theorem 13, where we demonstrate that the learned MAML-LQR controller is close to each task-specific optimal controller up to a task heterogeneity bias, where the bias increases when ϵ_1, ϵ_2, ϵ_3, and ϵ_4 increase.

Figure 1-(right) illustrates the adaptation of the learned MAML-LQR controller to an unseen task drawn from $p(\mathcal{T})$ (i.e., the same task distribution used in the MAML-LQR learning process). With unseen tasks 1, 2 and 3, this figure shows that, by initializing the PG-LQR approach (Fazel et al., 2018) from the learned controller K_{200}, it takes only a few PG iterations to achieve a controller that is sufficiently close to the unseen tasks' optimal, which is significantly fewer than initializing from a randomly sampled initial stabilizing controller $K_{0,\text{PG}}$. This aligns with Theorem 13, showing that the learned controller is close to the task-specific optimal controller for unseen tasks.

5. Conclusions and Future Work

We investigated the problem of meta-learning linear quadratic regulators in a heterogeneous and model-free setting, characterizing the stability and convergence of a MAML-LQR approach. We provided theoretical guarantees to ensure task-specific stability under the learned controller for both model-based and model-free settings. We established gradient heterogeneity bounds for three different task heterogeneity cases and offered convergence guarantees showing that the learned controller is close to each task-specific optimal controller up to a task-heterogeneity bias, emphasizing its ability to adapt to unseen tasks. Numerical experiments demonstrated the effect of task heterogeneity on the convergence of the MAML-LQR approach and assessed the adaptation of the learned controller to unseen tasks. Future work may explore variance-reduced approaches to reduce the variance of the zeroth-order gradient and Hessian estimation to improve the model-free sample complexity further.

Acknowledgments

Leonardo F. Toso is funded by the Columbia Presidential Fellowship. James Anderson is partially funded by NSF grants ECCS 2144634 and 2231350 and the Columbia Data Science Institute. Han Wang is funded by the Wei Family Fellowship.

REFERENCES

Momin Abbas, Quan Xiao, Lisha Chen, Pin-Yu Chen, and Tianyi Chen. Sharp-maml: Sharpness-aware model-agnostic meta learning. In *International conference on machine learning*, pages 10–32. PMLR, 2022.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42, 2022.

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, and Shimon Whiteson. A survey of meta-reinforcement learning. *arXiv preprint arXiv:2301.08028*, 2023.

Jingjing Bu, Afshin Mesbahi, Maryam Fazel, and Mehran Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

Yiting Chen, Ana M Ospina, Fabio Pasqualetti, and Emiliano Dall'Anese. Multi-Task System Identification of Similar Linear Time-Invariant Dynamical Systems. *arXiv preprint arXiv:2301.01430*, 2023.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092. PMLR, 2020.

Alireza Fallah, Kristian Georgiev, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of debiased model-agnostic meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 34:3096–3107, 2021.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

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

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019.

Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2020.

J Hong, N Moehle, and S Boyd. Introduction to Matrix Methods. In *Lecture notes*, 2021.

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a Theoretical Foundation of Policy Optimization for Learning Control Policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:123–158, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Theoretical convergence of multi-step model-agnostic meta-learning. *The Journal of Machine Learning Research*, 23(1):1317–1357, 2022.

Taylor T Johnson and Sayan Mitra. Safe flocking in spite of actuator faults using directional failure detectors. *Journal of Nonlinear Systems and Applications*, 2(1-2):73–95, 2011.

Bo Liu, Xidong Feng, Jie Ren, Luo Mai, Rui Zhu, Haifeng Zhang, Jun Wang, and Yaodong Yang. A theoretical understanding of gradient bias in meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 35:31059–31072, 2022.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd international conference on artificial intelligence and statistics*, pages 2916–2925. PMLR, 2019.

Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7474–7479. IEEE, 2019.

Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R Jovanović. On the linear convergence of random search for discrete-time LQR. *IEEE Control Systems Letters*, 5(3):989–994, 2020.

Igor Molybog and Javad Lavaei. When does maml objective have benign landscape? In *2021 IEEE Conference on Control Technology and Applications (CCTA)*, pages 220–227. IEEE, 2021.

Negin Musavi and Geir E Dullerud. Convergence of Gradient-based MAML in LQR. *arXiv preprint arXiv:2309.06588*, 2023.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Ibrahim K Ozaslan, Hesameddin Mohammadi, and Mihailo R Jovanović. Computing stabilizing feedback gains via a model-free policy gradient method. *IEEE Control Systems Letters*, 7:407–412, 2022.

Juan Perdomo, Jack Umenberger, and Max Simchowitz. Stabilizing dynamical systems via policy gradient methods. *Advances in neural information processing systems*, 34:29274–29286, 2021.

Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. Promp: Proximal meta-policy search. *arXiv preprint arXiv:1810.06784*, 2018.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6, pages 583–603. University of California Press, 1972.

Yujie Tang, Zhaolin Ren, and Na Li. Zeroth-order feedback optimization for cooperative multi-agent systems. *Automatica*, 148:110741, 2023.

Leonardo F. Toso, Han Wang, and James Anderson. Learning Personalized Models with Clustered System Identification. *arXiv preprint arXiv:2304.01395*, 2023a.

Leonardo F. Toso, Han Wang, and James Anderson. Oracle Complexity Reduction for Model-free LQR: A Stochastic Variance-Reduced Policy Gradient Approach. *arXiv preprint arXiv:2309.10679*, 2023b.

Leonardo F. Toso, Donglin Zang, James Anderson, and Han Wang. Meta-Learning Linear Quadratic Regulators: A Policy Gradient MAML Approach for the Model-free LQR. 2023c. URL https://github.com/jd-anderson/MAML-LQR/blob/main/MAML_LQR_technical_report.pdf.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Han Wang, Leonardo F. Toso, and James Anderson. Fedsysid: A federated approach to sample-efficient system identification. In *Learning for Dynamics and Control Conference*, pages 1308–1320. PMLR, 2023a.

Han Wang, Leonardo F. Toso, Aritra Mitra, and James Anderson. Model-free Learning with Heterogeneous Dynamical Systems: A Federated LQR Approach. *arXiv preprint arXiv:2308.11743*, 2023b.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Lirui Wang, Kaiqing Zhang, Allan Zhou, Max Simchowitz, and Russ Tedrake. Fleet Policy Learning via Weight Merging and An Application to Robotic Tool-Use. *arXiv preprint arXiv:2310.01362*, 2023c.

Thomas T Zhang, Katie Kang, Bruce D Lee, Claire Tomlin, Sergey Levine, Stephen Tu, and Nikolai Matni. Multi-task imitation learning for linear dynamical systems. In *Learning for Dynamics and Control Conference*, pages 586–599. PMLR, 2023a.

Thomas TCK Zhang, Leonardo F Toso, James Anderson, and Nikolai Matni. Meta-Learning Operators to Optimality from Multi-Task Non-IID Data. *arXiv preprint arXiv:2308.04428*, 2023b.