

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. 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 are 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

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 learned 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

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

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 $i \in [M]$ ¹. The objective of the LQR problem is to design an optimal control sequence $u_t^{(i)} := -K_i^* x_t^{(i)}$ that minimizes a quadratic cost in both state $x_t^{(i)}$ and input $u_t^{(i)}$. The optimal controllers K_i^* solve

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

where $Q^{(i)} \in \mathbb{S}_{>0}^{n_x}$, $R^{(i)} \in \mathbb{S}_{>0}^{n_u}$, 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^* 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_{ML}^* 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_{ML}^* solve

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

where $\bar{\mathcal{K}} := \bigcap_{i \in [M]} \mathcal{K}^{(i)}$ 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:

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

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

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

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

1. 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 $K \in \mathcal{S}_{ML}$, $K - \eta \nabla J_{ML}(K)$ is also MAML stabilizing, i.e., $K - \eta \nabla J_{ML}(K) \in \mathcal{S}_{ML}$. 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 $K \in \mathcal{S}_{ML}$, 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 $P_K^{(i)} \in \mathbb{S}_{>0}^{n_x}$ 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 $K \in \mathcal{S}_{ML}$, 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^* , which in turn is proved to be close to the MAML-LQR optimal controller K_{ML}^* . 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  $\eta_l, \eta$ 
2: for  $n = 0, \dots, N - 1$  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:

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

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:

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

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^* be the optimal controller of task $\mathcal{T}^{(i)}$. Then, it holds that

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

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

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 Appendix 6.1. 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 positive polynomial $h(K)$ over the set of stabilizing controllers \mathcal{S}_{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 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 $K \in \mathcal{S}_{\text{ML}}$. It holds that,*

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

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 $h_{\text{het}}^1(K)$, $h_{\text{het}}^2(K)$, $h_{\text{het}}^3(K)$, and $h_{\text{het}}^4(K)$ are positive polynomials that depend on the problem parameters².

The proof and explicit expressions of $h_{\text{het}}^1(K)$, $h_{\text{het}}^2(K)$, $h_{\text{het}}^3(K)$, and $h_{\text{het}}^4(K)$, are detailed in Appendix 6.2. 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).

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,

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: $\hat{\nabla} J(K) = \frac{n_x n_u}{2r^2 m} \sum_{l=1}^m (J(K_l^1) - J(K_l^2)) U_l$,
 - 6: $\hat{\nabla}^2 J(K) = \frac{n_u^2}{r^2 m} \sum_{l=1}^m (J(K_l^1) - J(K)) (U_l U_l^\top - I_{n_u})$,
 - 7: **Return** $\hat{\nabla} J(K)$, $\hat{\nabla}^2 J(K)$.
-

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 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, $\mathbb{E} [\nabla J^{(i)}(K)] = \mathbb{E} [\frac{n_x n_u}{2r^2} (J^{(i)}(K^1) - J^{(i)}(K^2)) U]$ (Mohammadi et al., 2020; Toso et al., 2023b), and the Hessian with its second-order counterpart, $\mathbb{E} [\nabla^2 J^{(i)}(K)] = \mathbb{E} [\frac{n_u^2}{r^2} (J^{(i)}(K^1) - J^{(i)}(K)) (U U^\top - I_{n_u})]$, (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 carefully 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., $K_n \in \mathcal{S}_{\text{ML}}$, $\forall n$.

2. 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.

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

```

1: Input: initial stabilizing controller  $K_0$ , inner and outer step-sizes  $\eta_l, \eta$ , smoothing radius  $r$ ,
   number of samples  $m$ .
2: for  $n = 0, \dots, N - 1$  do
3:   for each task  $i \in [M]$  in  $\mathcal{T}$  compute
4:      $\left[ \widehat{\nabla} J^{(i)}(K_n), \widehat{\nabla}^2 J^{(i)}(K_n) \right] = \text{ZO2P}(K_n, m, r)$ 
5:      $\widehat{K}_n^{(i)} = K_n - \eta_l \widehat{\nabla} J^{(i)}(K_n)$ , and  $\widehat{H}^{(i)}(K_n) = I_{n_u} - \eta_l \widehat{\nabla}^2 J^{(i)}(K_n)$ 
6:   end for
7:    $\widehat{\nabla} J^{(i)}(\widehat{K}_n^{(i)}) = \text{ZO2P}(\widehat{K}_n^{(i)}, m, r)$ 
8:    $K_{n+1} = K_n - \frac{\eta}{M} \sum_{i=1}^M \widehat{H}^{(i)}(K_n) \widehat{\nabla} J^{(i)}(\widehat{K}_n^{(i)})$ 
9: end for
10: 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 $\eta_l \leq \min \left\{ \frac{n_u}{\sqrt{2}h_H}, \frac{1}{\sqrt{2}h_{grad}}, \frac{1}{\sqrt{12(12\bar{h}_{grad}^2 n_u^2 + \bar{h}_H^2)}} \right\}$, $\eta \leq \frac{1}{4h_{grad}}$ and $\bar{f}_z(\bar{\epsilon}) \leq \sqrt{\min_i \frac{\lambda_i \Delta_0^{(i)}}{288n_u^3}}$, 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 $\eta_l \leq \min \left\{ \frac{1}{\bar{h}_G}, \frac{n_u}{\bar{h}_H}, \frac{1}{\bar{h}_{grad}}, \frac{1}{\sqrt{20(12\bar{h}_{grad}^2 n_u^2 + \bar{h}_H^2)}}, \frac{1}{2} \right\}$, $\eta \leq \frac{1}{8h_{grad}}$. In addition, the heterogeneity, smoothing radius and number of samples satisfy $\bar{f}_z(\bar{\epsilon}) \leq \sqrt{\min_i \frac{\lambda_i \Delta_0^{(i)}}{480n_u^3}}$, $r \leq \min \left\{ \underline{h}_r^1 \left(\frac{\sqrt{\psi^{(i)}}}{2} \right), \underline{h}_r^2 \left(\frac{\sqrt{\psi^{(i)}}}{2} \right) \right\}$, and $m \geq \max \left\{ \bar{h}_m^1 \left(\frac{\sqrt{\psi^{(i)}}}{2}, \delta \right), \bar{h}_m^2 \left(\frac{\sqrt{\psi^{(i)}}}{2}, \delta \right) \right\}$ ³, 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 deferred to Sections 6.3 and 6.5 of the appendix. 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 that is 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^* and to the optimal MAML controller K_{ML}^* . For this purpose,

3. 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 Appendix 6.1.

we study the closeness of K_N and K_i^* by bounding $J^{(i)}(K_N) - J^{(i)}(K_i^*)$ and the closeness of K_{ML}^* with $J^{(i)}(K_{ML}^*) - J^{(i)}(K_i^*)$.

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 $\eta_l \leq \min \left\{ \frac{n_u}{\sqrt{2}h_H}, \frac{1}{\sqrt{2}h_{grad}}, \frac{1}{\sqrt{12(12\bar{h}_{grad}^2 n_u^2 + h_H^2)}} \right\}$, and $\eta \leq \frac{1}{4h_{grad}}$, $N \geq \frac{8}{\eta\lambda_i} \log \left(\frac{\Delta_0^{(i)}}{\epsilon'} \right)$, respectively, for some small tolerance $0 < \epsilon' < 1$. Then, it holds that,*

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

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 $\eta_l \leq \min \left\{ \frac{1}{h_G}, \frac{n_u}{h_H}, \frac{1}{h_{grad}}, \frac{1}{\sqrt{20(12\bar{h}_{grad}^2 n_u^2 + h_H^2)}}, \frac{1}{2} \right\}$, $\eta \leq \frac{1}{8h_{grad}}$, $r \leq \min \left\{ \underline{h}_r^1 \left(\frac{\epsilon' \lambda_i}{1296} \right), \underline{h}_r^2 \left(\frac{\epsilon' \lambda_i}{1296} \right) \right\}$, $m \geq \max \left\{ \bar{h}_m^1 \left(\frac{\epsilon' \lambda_i}{1296}, \delta \right), \bar{h}_m^2 \left(\frac{\epsilon' \lambda_i}{1296}, \delta \right) \right\}$, 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,*

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

The proofs of Theorems 12 and 13 are detailed in Sections 6.4 and 6.6 of the appendix. 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).

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^* 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^* , and K_{ML}^* and K_i^* are close, one may conclude that K_N and K_{ML}^* 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).

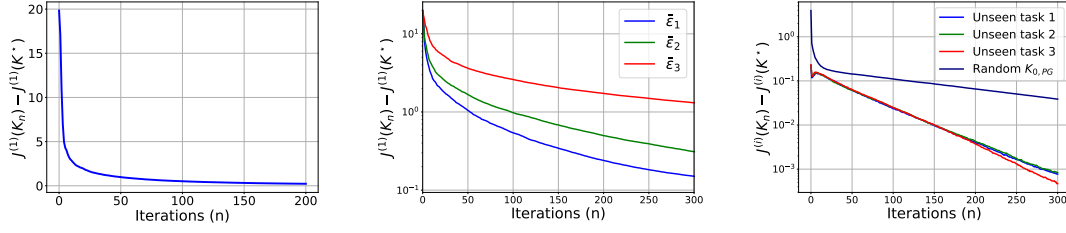


Figure 1: Cost gap between the learned the task-specific optimal controller with respect to iteration. (left) Convergence of the MAML-LQR. (middle) MAML-LQR, $\bar{\epsilon}_1 = (1.2, 1.1, 1.4, 1.2) \times 10^{-3}$, $\bar{\epsilon}_2 = (1.3, 1.1, 1.4, 1.2) \times 10^{-2}$, $\bar{\epsilon}_3 = (1.7, 1.8,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

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 MAML-LQR controller (i.e., $K_{0,PG} = K_N$) 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 Appendix 7.

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 $K_{0,PG}$. 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 abil-

ity 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. Prompt: 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.
- 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, Dhharshan 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.

6. Appendix

We now provide the technical details/proofs of our main results. For this purpose, we first revisit some auxiliary lemmas and norm inequalities that are instrumental in our stability and convergence analysis of the MAML-LQR approach. Besides the lemmas, we also provide the explicit expressions of the positive polynomials $h_G(K)$, $h_c(K)$, $h_H(K)$, $h_\Delta(K)$, $h_{\text{cost}}(K)$, and $h_{\text{grad}}(K)$ that appears in our theoretical guarantees. Lemmas 17 and 18 are used to control the gradient and Hessian estimation errors through a matrix Bernstein-type of inequality (Lemma 16). These lemmas are crucial to provide the stability and convergence guarantees for the model-free setting. Moreover, we provide the experimental setup details utilized to generate the results illustrated in Figure 1.

Notation: Consider a set of M matrices $\{A^{(i)}\}_{i=1}^M$, we denote $\|A\|_{\max} := \max_i \|A^{(i)}\|$, and $\|A\|_{\min} := \min_i \|A^{(i)}\|$. The spectral radius of a square matrix is $\rho(\cdot)$.

6.1. Auxiliary Lemmas and Norm Inequalities

We first introduce some auxiliary lemmas and norm inequality that are essential in proving the main results of this work.

Young's inequality: Given any two matrices $A, B \in \mathbb{R}^{n_x \times n_u}$, for any $\beta > 0$, we have

$$\|A + B\|_2^2 \leq (1 + \beta)\|A\|_2^2 + \left(1 + \frac{1}{\beta}\right)\|B\|_2^2 \leq (1 + \beta)\|A\|_F^2 + \left(1 + \frac{1}{\beta}\right)\|B\|_F^2. \quad (8)$$

Moreover, given any two matrices A, B of the same dimensions, for any $\beta > 0$, we have

$$\langle A, B \rangle \leq \frac{\beta}{2}\|A\|_2^2 + \frac{1}{2\beta}\|B\|_2^2 \leq \frac{\beta}{2}\|A\|_F^2 + \frac{1}{2\beta}\|B\|_F^2. \quad (9)$$

Jensen's inequality: Given M matrices $A^{(1)}, \dots, A^{(M)}$ of identical dimensions, we have that

$$\left\| \sum_{i=1}^M A^{(i)} \right\|_2^2 \leq M \sum_{i=1}^M \|A^{(i)}\|_2^2, \left\| \sum_{i=1}^M A^{(i)} \right\|_F^2 \leq M \sum_{i=1}^M \|A^{(i)}\|_F^2. \quad (10)$$

Lemma 14 (Uniform bounds) *Given a LQR task $i \in [M]$ and an stabilizing controller $K \in \mathcal{S}_{ML}$, the Frobenius norm of gradient $\nabla J^{(i)}(K)$, Hessian $\nabla^2 J^{(i)}(K)$ and control gain K can be bounded as follows:*

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

with

$$h_G(K) = \frac{J_{\max}(K) \sqrt{\frac{\max_i \|R^{(i)} + B^{(i)\top} P_K^{(i)} B^{(i)}\| (J_{\max}(K) - J_{\min}(K))}{\mu}}}{\min_i \sigma_{\min}(Q^{(i)})},$$

$$h_H(K) = \left(2\|R\|_{\max} + \frac{2\|B\|_{\max} J_{\max}(K)}{\mu} + \frac{4\sqrt{2}\tilde{\xi}_{\max}\|B\|_{\max} J_{\max}(K)}{\mu} \right) \frac{J_{\max}(K)n_u}{\|Q\|_{\min}},$$

$$h_c(K) = \frac{\sqrt{\frac{\max_i \|R^{(i)} + B^{(i)\top} P_K^{(i)} B^{(i)}\| (J_{\max}(K) - J_{\min}(K))}{\mu}} + \|B^\top P_K A\|_{\max}}{\sigma_{\min}(R)},$$

$$\text{with } \tilde{\xi}_{\max} := \frac{1}{\|Q\|_{\min}} \left(\frac{(1 + \|B\|_{\max}^2) J_{\max}(K)}{\mu} + \|R\|_{\max} - 1 \right).$$

Proof The proof for the uniform bounds of $\|\nabla J^{(i)}(K)\|_F$, $\|K\|_F$ and the gradient domination are provided in (Fazel et al., 2018; Wang et al., 2023b). In addition, the proof for the uniform bound of $\|\nabla^2 J^{(i)}\|_F$ can be found in (Bu et al., 2019, Lemma 7.9). ■

Lemma 15 (Local smoothness) Given two controller $K, K' \in \mathcal{S}_{ML}$ such that $\|\Delta\| := \|K' - K\| \leq h_\Delta(K) < \infty$. The LQR cost, gradient and Hessian satisfy:

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

for any $i \in [M]$, where

$$\begin{aligned} h_\Delta(K) &= \frac{\max_i \sigma_{\min}(Q^{(i)}) \mu}{4 \|B\|_{\max} J_{\max}(K) (\|A - BK\|_{\max} + 1)}, \\ h_{\text{cost}}(K) &= \frac{4 \text{Tr}(\Sigma_0) J_{\max}(K) \|R\|_{\max}}{\mu \min_i \sigma_{\min}(Q^{(i)})} \left(\|K\| + \frac{h_\Delta(K)}{2} + \|B\|_{\max} \|K\|^2 (\|A - BK\|_{\max} + 1) \nu(K) \right), \\ h_{\text{hess}}(K) &= \sup_{\|X\|_F=1} 2(h_1(K) + 2h_2(K)) \|X\|_F^2, \\ h_{\text{grad}}(K) &= 4 \left(\frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q)} \right) \left[\|R\|_{\max} + \|B\|_{\max} (\|A\|_{\max} + \|B\|_{\max} (\|K\| + h_\Delta(K))) \right. \\ &\quad \times \left(\frac{h_{\text{cost}}(K) J_{\max}(K)}{\text{Tr}(\Sigma_0)} + \|B\|_{\max}^2 \frac{J_{\max}(K)}{\mu} \right] \\ &\quad + 8 \left(\frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q)} \right)^2 \left(\frac{\|B\|_{\max} (\|A - BK\|_{\max} + 1)}{\mu} \right) h_0(K). \end{aligned}$$

$$\text{with } \nu(K) = \frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)}) \mu}, h_0(K) = \sqrt{\frac{\max_i \|R^{(i)} + B^{(i)\top} P_K^{(i)} B^{(i)}\| (J_{\max}(K) - J_{\min}(K))}{\mu}}, \text{ and}$$

$$\begin{aligned} h_1(K) &= h_3(K) \|B\|_{\max}^2 \frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})} + \tilde{\mu} h_4(K) \|B\|_{\max} \frac{J_{\max}(K)}{\mu} + h_4(K) \max_i \text{Tr}(R^{(i)}), \\ h_2(K) &= \|B\|_{\max} J_{\max}(K) \left(\frac{h_6(K) h_4(K) \max_i \text{Tr}(A^{(i)} - B^{(i)} K)}{\mu} + \|B\|_{\max} h_6(K) \tilde{\mu} \nu(K) \right. \\ &\quad \left. + \frac{\tilde{\mu} h_7(K)}{\min_i \sigma_{\min}(Q^{(i)})} \right), \end{aligned}$$

$$h_3(K) = 6 \left(\frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})} \right)^2 \|K\|^2 \|R\|_{\max} \|B\|_{\max} (\|A - BK\|_{\max} + 1)$$

$$\begin{aligned}
& + 6 \left(\frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})} \right) \|K\| \|R\|_{\max}, \\
h_4(K) &= 4 \left(\frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})} \right)^2 \frac{\|B\|_{\max} (\|A - BK\|_{\max} + 1)}{\mu}, \\
h_6(K) &= \sqrt{\frac{1}{\min_i \sigma_{\min}(Q^{(i)})} \left(\|R\|_{\max} + \frac{1 + \|B\|_{\max}^2 J_{\max}(K)}{\mu} \right)} - 1, \\
h_7(K) &= 4 \left(\nu(K) h_8(K) + 8\nu^2(K) \|B\|_{\max} (\|A - BK\|_{\max} + 1) h_9(K) \right), \\
h_8(K) &= \|R\|_{\max} + \|B\|_{\max}^2 \frac{J_{\max}(K)}{\mu} + (\|B\|_{\max} \|A\|_{\max} + \|B\|_{\max}^2 \|K\|_{\max}) h_3(K), \\
h_9(K) &= 2 \left(\|R\|_{\max} \|K\| + \|B\|_{\max} \|A - BK\|_{\max} \frac{J_{\max}(K)}{\mu} \right).
\end{aligned}$$

and $\tilde{\mu} = 1 + \frac{\mu}{h_{\Delta}(K)}$.

Proof The proof of the local smoothness for the cost and gradient are provided in (Wang et al., 2023b, Appendix F), whereas the local smoothness of the Hessian is provided in (Musavi and Dullerud, 2023, Lemma 7) \blacksquare

Lemma 16 (Matrix Bernstein Inequality) (Gravell et al., 2020, Lemma B.5) Let $\{Z_i\}_{i=1}^m$ be a set of m independent random matrices of dimension $d_1 \times d_2$ with $\mathbb{E}[Z_i] = Z$, $\|Z_i - Z\| \leq B_r$ almost surely, and maximum variance

$$\max \left(\left\| \mathbb{E} \left(Z_i Z_i^\top \right) - Z Z^\top \right\|, \left\| \mathbb{E} \left(Z_i^\top Z_i \right) - Z^\top Z \right\| \right) \leq \sigma_r^2,$$

and sample average $\hat{Z} := \frac{1}{m} \sum_{i=1}^m Z_i$. Let a small tolerance $\epsilon \geq 0$ and small probability $0 \leq \delta \leq 1$ be given. If

$$m \geq \frac{2 \min(d_1, d_2)}{\epsilon^2} \left(\sigma_r^2 + \frac{B_r \epsilon}{3 \sqrt{\min(d_1, d_2)}} \right) \log \left[\frac{d_1 + d_2}{\delta} \right]$$

then $\mathbb{P} \left[\|\hat{Z} - Z\|_F \leq \epsilon \right] \geq 1 - \delta$.

Lemma 17 (Gradient Estimation) Given a small tolerance ϵ and probability $0 < \delta < 1$ and suppose the smoothing radius and number of samples is selected according to

$$r \leq \bar{h}_r^1 \left(\frac{\epsilon}{2} \right) := \min \left\{ \bar{h}_{\Delta}, \frac{1}{\bar{h}_{\text{cost}}}, \frac{\epsilon}{2 \bar{h}_{\text{grad}}} \right\}, \quad (11)$$

$$m \geq \bar{h}_m^1 \left(\frac{\epsilon}{2}, \delta \right) := \frac{8 \min\{n_x, n_u\}}{\epsilon^2} \left(\sigma_r^2 + \frac{B_r \epsilon}{6 \sqrt{\min\{n_x, n_u\}}} \right) \log \left[\frac{n_x + n_u}{\delta} \right], \quad (12)$$

where

$$B_r := 2n_x n_u \bar{h}_{\text{cost}} \bar{J}_{\max} + \frac{\epsilon}{2} + \bar{h}_G, \quad \sigma_r^2 := (2n_x n_u \bar{h}_{\text{cost}} \bar{J}_{\max})^2 + \left(\frac{\epsilon}{2} + \bar{h}_G \right)^2,$$

where $\bar{J}_{\max} := \sup_{K \in \mathcal{S}_{ML}} \max_i J^{(i)}(K)$. Then, for any task $\mathcal{T}^{(i)}$, $\|\widehat{\nabla} J^{(i)}(K) - \nabla J^{(i)}(K)\|_F \leq \epsilon$, with probability $1 - \delta$.

Proof To control $\|\widehat{\nabla} J^{(i)}(K) - \nabla J^{(i)}(K)\|$, for any task $\mathcal{T}^{(i)}$, we use triangle inequality to write

$$\|\widehat{\nabla} J^{(i)}(K) - \nabla J^{(i)}(K)\|_F \leq \|\widehat{\nabla} J^{(i)}(K) - \nabla J_r^{(i)}(K)\|_F + \|\nabla J^{(i)}(K) - \nabla J_r^{(i)}(K)\|_F$$

where $\nabla J_r^{(i)}(K) = \mathbb{E}_{U \sim \mathbb{S}_r} [\nabla J^{(i)}(K + U)]$, $\widehat{\nabla} J^{(i)}(K) = \frac{n_x n_u}{2r^2 m} \sum_{l=1}^m (J^{(i)}(K + U_l) - J^{(i)}(K - U_l))U_l$, which implies $\|K + U_l - K\|_F = \|U_l\|_F = r$ and $\|K - U_l - K\|_F = \|U_l\|_F = r$. Therefore, by selecting $r \leq h_{\Delta}(K)$ we have that both $|J(K') - J^{(i)}(K)| \leq h_{\text{cost}}(K)J^{(i)}(K)\|\Delta\|_F$ and $\|\nabla J^{(i)}(K') - \nabla J^{(i)}(K)\|_F \leq h_{\text{grad}}(K)\|\Delta\|_F$ are verified (Lemma 6). Then, by enforcing $r \leq \frac{1}{h_{\text{cost}}(K)}$ we have

$$\begin{aligned} |J^{(i)}(K + U) - J^{(i)}(K)| &\leq J^{(i)}(K) \rightarrow J^{(i)}(K + U) \leq 2J^{(i)}(K), \\ |J^{(i)}(K - U) - J^{(i)}(K)| &\leq J^{(i)}(K) \rightarrow J^{(i)}(K - U) \leq 2J^{(i)}(K) \end{aligned}$$

which ensures the stability of the LQR task under the perturbed controllers. This implies that $J^{(i)}(K + U)$ and $J^{(i)}(K - U)$ are well-defined. Therefore, to control $\|\nabla J^{(i)}(K) - \nabla J_r^{(i)}(K)\|_F$ we can impose $r \leq \frac{\epsilon}{2h_{\text{grad}}(K)}$ to obtain

$$\|\nabla J^{(i)}(K + U) - \nabla J^{(i)}(K)\|_F \leq h_{\text{grad}}(K)r \leq \frac{\epsilon}{2}$$

thus, we use the fact that $\nabla J_r^{(i)}(K) = \mathbb{E}_{U \sim \mathbb{S}_r} [\nabla J^{(i)}(K + U)]$ and Jensen's inequality to write

$$\begin{aligned} \|\nabla J^{(i)}(K) - \nabla J_r^{(i)}(K)\|_F &= \|\mathbb{E}_{U \sim \mathbb{S}_r} (\nabla J^{(i)}(K + U) - \nabla J^{(i)}(K))\|_F \\ &\leq \mathbb{E}_{U \sim \mathbb{S}_r} \|\nabla J^{(i)}(K + U) - \nabla J^{(i)}(K)\|_F \leq \frac{\epsilon}{2}. \end{aligned}$$

On the other hand, to control $\|\widehat{\nabla} J^{(i)}(K) - \nabla J_r^{(i)}(K)\|_F$ we use the Lemma 16. For this purpose, we denote $Z = \nabla J_r^{(i)}(K) - \widehat{\nabla} J^{(i)}(K) = \mathbb{E}_{U \sim \mathbb{S}_r} [\widehat{\nabla} J^{(i)}(K)]$ and each individual sample is denoted by

$$Z_i = \frac{n_x n_u}{2r^2} (J^{(i)}(K + U) - J^{(i)}(K - U))U,$$

then we can write

$$\|Z_i\|_F = \left\| \frac{n_x n_u}{2r^2} (J(K + U) - J(K - U))U \right\|_F \leq \frac{n_x n_u}{2r} |J^{(i)}(K + U) - J^{(i)}(K - U)|$$

and use Lemma 6 to write

$$|J^{(i)}(K + U) - J^{(i)}(K - U)| \leq h_{\text{cost}}(K)J^{(i)}(K - U)\|2U\| = 2h_{\text{cost}}(K)J^{(i)}(K - U)r$$

and obtain

$$\|Z_i\|_F \leq n_x n_u h_{\text{cost}}(K)J^{(i)}(K).$$

Now we can also control $\|Z\|_F$ by using triangle inequality

$$\begin{aligned}\|Z\|_F &= \|\nabla J_r^{(i)}(K)\|_F = \|\nabla J_r^{(i)}(K) - \nabla J^{(i)}(K) + \nabla J^{(i)}(K)\|_F \leq \|\nabla J_r^{(i)}(K) - \nabla J^{(i)}(K)\|_F \\ &\quad + \|\nabla J(K)\|_F \leq \frac{\epsilon}{2} + h_G(K),\end{aligned}$$

and we can use triangle inequality to write

$$\|Z_i - Z\|_F \leq \|Z_i\|_F + \|Z\|_F \leq B_r := n_x n_u h_{\text{cost}}(K) J(K) + \frac{\epsilon}{2} + h_G(K),$$

and for the variance term we have

$$\begin{aligned}\|\mathbb{E}(Z_i Z_i^\top) - Z Z^\top\|_F &\leq \|\mathbb{E}(Z_i Z_i^\top)\|_F + \|Z Z^\top\|_F \\ &\leq \max_{Z_i} (\|Z_i\|_F)^2 + \|Z\|_F^2 \\ &\leq \sigma_r^2 := (n_x n_u h_{\text{cost}}(K) J(K))^2 + \left(\frac{\epsilon}{2} + h_G(K)\right)^2.\end{aligned}$$

■

which completes the proof.

Lemma 18 (Hessian Estimation) *Given a small tolerance ϵ and probability $0 < \delta < 1$ and suppose the smoothing radius r and number of samples m is selected according to*

$$r \leq \underline{h}_r^2 \left(\frac{\epsilon}{2}\right) := \min \left\{ \underline{h}_\Delta, \frac{1}{\bar{h}_{\text{cost}}}, \frac{\epsilon}{2\bar{h}_{\text{hess}}} \right\}, \quad (13)$$

$$m \geq \bar{h}_m^2 \left(\frac{\epsilon}{2}, \delta\right) := \frac{8n_u}{\epsilon^2} \left(\bar{\sigma}_r^2 + \frac{\bar{B}_r \epsilon}{6\sqrt{n_u}} \right) \log \left[\frac{2n_u}{\delta} \right], \quad (14)$$

where

$$\begin{aligned}\bar{B}_r &:= \frac{n_u^2(n_u + r^2)}{r} \bar{h}_{\text{cost}} \bar{J}_{\max} + \frac{\epsilon}{2} + \bar{h}_H, \\ \bar{\sigma}_r^2 &:= \left(\frac{n_u^2(n_u + r^2)}{r} \bar{h}_{\text{cost}} \bar{J}_{\max} \right)^2 + \left(\frac{\epsilon}{2} + \bar{h}_H \right)^2,\end{aligned}$$

then, for any task $\mathcal{T}^{(i)}$, $\|\hat{\nabla}^2 J^{(i)}(K) - \nabla^2 J^{(i)}(K)\|_F \leq \epsilon$ with probability $1 - \delta$.

Proof To control $\|\hat{\nabla}^2 J^{(i)}(K) - \nabla^2 J^{(i)}(K)\|$ we can use triangle inequality to write

$$\|\hat{\nabla}^2 J^{(i)}(K) - \nabla^2 J^{(i)}(K)\|_F \leq \|\hat{\nabla}^2 J^{(i)}(K) - \nabla^2 J_r^{(i)}(K)\|_F + \|\nabla^2 J^{(i)}(K) - \nabla^2 J_r^{(i)}(K)\|_F$$

where $J_r^{(i)}(K) = \mathbb{E}_{U \sim \mathbb{S}_r}[J^{(i)}(K + U)]$, $\nabla^2 J_r^{(i)}(K) = \mathbb{E}_{U \sim \mathbb{S}_r}[\nabla^2 J^{(i)}(K + U)]$, and estimation $\hat{\nabla}^2 J^{(i)}(K) = \frac{n_u^2}{r^2 m} \sum_{l=1}^m (J^{(i)}(K + U_l) - J^{(i)}(K))(U_l U_l^\top - I_{n_u})$. Then, similar to the proof of

Lemma 17, by selecting the smoothing radius according to $r \leq \min \left\{ \frac{1}{h_{\text{cost}}(K)}, h_{\Delta}(K) \right\}$ we have that $J(K+U)$ is well-defined and the local smoothness of the cost and Hessian are satisfied. Then, to control $\|\nabla^2 J^{(i)}(K) - \nabla^2 J_r(K)\|_F$ we select $r \leq \frac{\epsilon}{2h_{\text{hess}}(K)}$ to obtain

$$\|\nabla^2 J^{(i)}(K+U) - \nabla^2 J^{(i)}(K)\|_F \leq h_{\text{hess}}(K)r \leq \frac{\epsilon}{2}$$

and by Jensen's inequality we can write

$$\begin{aligned} \|\nabla^2 J^{(i)}(K) - \nabla^2 J_r^{(i)}(K)\|_F &= \|\mathbb{E}_{U \sim \mathbb{S}_r}(\nabla^2 J^{(i)}(K+U) - \nabla^2 J^{(i)}(K))\|_F \\ &\leq \mathbb{E}_{U \sim \mathbb{S}_r} \|\nabla^2 J^{(i)}(K+U) - \nabla^2 J^{(i)}(K)\|_F \leq \frac{\epsilon}{2}. \end{aligned}$$

Therefore, by using Lemma 16 to control $\|\widehat{\nabla}^2 J^{(i)}(K) - \nabla^2 J_r^{(i)}(K)\|_F$, we have that $Z = \nabla^2 J_r^{(i)}(K) = \mathbb{E}_{U \sim \mathbb{S}_r}[\widehat{\nabla}^2 J^{(i)}(K)]$ where $Z_i = \frac{n_u n_u}{2r^2} (J^{(i)}(K+U) - J^{(i)}(K))(UU^\top - I_{n_u})$, then we have

$$\|Z_i\|_F = \left\| \frac{n_u^2}{2r^2} (J^{(i)}(K+U) - J^{(i)}(K))(UU^\top - I_{n_u}) \right\|_F \leq \frac{n_u^2}{2r^2} |J^{(i)}(K+U) - J^{(i)}(K)| (n_u + r^2)$$

and use Lemma 6 to write

$$|J^{(i)}(K+U) - J^{(i)}(K)| \leq h_{\text{cost}}(K) J^{(i)}(K) \|U\|_F = h_{\text{cost}}(K) J^{(i)}(K) r$$

and obtain

$$\|Z_i\|_F \leq \frac{n_u^2(n_u + r^2)}{r} h_{\text{cost}}(K) J^{(i)}(K).$$

and

$$\begin{aligned} \|Z\|_F &= \|\nabla^2 J_r(K)\|_F = \|\nabla^2 J_r^{(i)}(K) - \nabla^2 J^{(i)}(K) + \nabla^2 J^{(i)}(K)\|_F \leq \|\nabla^2 J_r^{(i)}(K) - \nabla^2 J^{(i)}(K)\|_F \\ &\quad + \|\nabla^2 J^{(i)}(K)\|_F \leq \frac{\epsilon}{2} + h_H(K) \end{aligned}$$

which implies

$$\|Z_i - Z\|_F \leq \|Z_i\|_F + \|Z\|_F \leq \bar{B}_r := \frac{n_u^2(n_u + r^2)}{r} h_{\text{cost}}(K) J^{(i)}(K) + \frac{\epsilon}{2} + h_H(K),$$

and

$$\begin{aligned} \|\mathbb{E}(Z_i Z_i^\top) - Z Z^\top\|_F &\leq \|\mathbb{E}(Z_i Z_i^\top)\|_F + \|Z Z^\top\|_F \\ &\leq \max_{Z_i} (\|Z_i\|_F)^2 + \|Z\|_F^2 \\ &\leq \bar{\sigma}_r^2 := \left(\frac{n_u^2(n_u + r^2)}{r} h_{\text{cost}}(K) J^{(i)}(K) \right)^2 + \left(\frac{\epsilon}{2} + h_H(K) \right)^2. \end{aligned}$$

for the variance term, which completes the proof. ■

6.2. Proof of Lemma 8

For the simplicity of exposure, we derive the gradient heterogeneity bound for the more general task heterogeneity setting, i.e., system and cost heterogeneity which spans the other cases by setting $\epsilon_1 = \epsilon_2 = 0$ (cost heterogeneity) and $\epsilon_3 = \epsilon_4 = 0$ (system heterogeneity).

To derive the expression of the gradient heterogeneity bound, we start by first recalling the expression of the gradient of the LQR cost for $\nabla J^{(i)}(K)$ and $\nabla J^{(j)}(K)$,

$$\nabla J^{(i)}(K) = 2E_K^{(i)}\Sigma_K^{(i)}, \text{ and } \nabla J^{(j)}(K) = 2E_K^{(j)}\Sigma_K^{(j)}$$

where $E_K^{(i)} = R^{(i)}K - B^\top P_K^{(i)}(A^{(i)} - B^{(i)}K)$, $\Sigma_K^{(i)} = \mathbb{E}_{x_0^{(i)} \sim \mathcal{X}_0} \sum_{t=0}^{\infty} x_t x_t^\top$, which can be used to write

$$\begin{aligned} \|\nabla J^{(i)}(K) - \nabla J^{(j)}(K)\| &= 2\|E_K^{(i)}\Sigma_K^{(i)} - E_K^{(j)}\Sigma_K^{(j)}\| \\ &\leq 2\left(\|E_K^{(i)} - E_K^{(j)}\|\|\Sigma_K^{(i)}\| + \|E_K^{(j)}\|\|\Sigma_K^{(i)} - \Sigma_K^{(j)}\|\right). \end{aligned}$$

From (Fazel et al., 2018, Lemma 13) and the definition of $E_K^{(i)}$, we can write

$$\|\Sigma_K^{(i)}\| \leq \frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})}, \quad \|E_K^{(i)}\| \leq \|R\|_{\max}\|K\| + \frac{J_{\max}(K)(\|B\|_{\max}^2\|K\| + \|B\|_{\max}\|A\|_{\max})}{\mu}.$$

where $J_{\max}(K) := \max_i J^{(i)}(K)$. Then, let us first control $\|\Sigma_K^{(i)} - \Sigma_K^{(j)}\|$. For this purpose, similar to (Wang et al., 2023b), which considers only the system heterogeneity setting, we have

$$\begin{aligned} \|\Sigma_K^{(i)} - \Sigma_K^{(j)}\| &= \|\mathcal{T}_K^{(i)}(\Sigma_0) - \mathcal{T}_K^{(j)}(\Sigma_0)\| \leq 2\|\mathcal{T}_K^{(i)}\|^2\|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\|\|\Sigma_0\| \\ &\leq 4\left(\frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})\mu}\right)^2 (\epsilon_1 + \epsilon_2\|K\|)\|A - BK\|_{\max}\|\Sigma_0\| \\ &= 4\epsilon_1\nu^2(K)\|A - BK\|_{\max}\|\Sigma_0\| + 4\epsilon_2\nu^2(K)\|K\|\|A - BK\|_{\max}\|\Sigma_0\|, \end{aligned}$$

and

$$\begin{aligned} \mathcal{F}_K^{(i)}(X) &:= (A^{(i)} - B^{(i)}K)X(A^{(i)} - B^{(i)}K)^\top, \\ \mathcal{T}_K^{(i)}(X) &:= \sum_{t=0}^{\infty} (A^{(i)} - B^{(i)}K)^t X \left[(A^{(i)} - B^{(i)}K)^\top \right]^t, \end{aligned}$$

being linear operators on some symmetric matrix X . We now proceed to control $\|E_K^{(i)} - E_K^{(j)}\|$. To do so, we can first write

$$\|E_K^{(i)} - E_K^{(j)}\| \leq \underbrace{\epsilon_4\|K\|}_{T_1} + \underbrace{\|B^{(i)\top}P_K^{(i)}(A^{(i)} - B^{(i)}K) - B^{(j)\top}P_K^{(j)}(A^{(j)} - B^{(j)}K)\|}_{T_2},$$

where T_1 comes from the cost heterogeneity, and T_2 is similar to the system heterogeneity term in (Wang et al., 2023b). We proceed to control T_2 , by following some similar manipulations as detailed in (Wang et al., 2023b, Appendix E.1). Therefore, we have

$$T_2 \leq \|B\|_{\max}\|P_K\|_{\max}(\epsilon_1 + \epsilon_2\|K\|) + \|B\|_{\max}\|P_K^{(i)} - P_K^{(j)}\|\|A - BK\|_{\max}$$

$$+ \epsilon_2 \|A - BK\|_{\max} \|P_K\|_{\max},$$

with $\|P_K\|_{\max} \leq \frac{J_{\max}(K)}{\mu}$. To control $\|P_K^{(i)} - P_K^{(j)}\|$ we can first write,

$$\begin{aligned} \|P_K^{(i)} - P_K^{(j)}\| &= \|\mathcal{T}_K^{(i)}(Q^{(i)} + K^\top R^{(i)} K) - \mathcal{T}_K^{(j)}(Q^{(j)} + K^\top R^{(j)} K)\| \\ &\leq \|\mathcal{T}_K^{(i)}(Q^{(i)} + K^\top R^{(i)} K) - \mathcal{T}_K^{(j)}(Q^{(i)} + K^\top R^{(i)} K)\| \\ &\quad + \underbrace{\|\mathcal{T}_K^{(j)}(Q^{(i)} + K^\top R^{(i)} K) - \mathcal{T}_K^{(j)}(Q^{(j)} + K^\top R^{(j)} K)\|}_{T_3} \\ &\leq 2\|\mathcal{T}_K^{(i)}\|^2 \|\mathcal{F}_K^{(i)} - \mathcal{F}_K^{(j)}\| \|Q^{(i)} + K^\top R^{(i)} K\| + T_3 \\ &\leq 4 \left(\frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})\mu} \right)^2 (\epsilon_1 + \epsilon_2 \|K\|) \|A - BK\|_{\max} (\|Q\|_{\max} + \|K\|^2 \|R\|_{\max}) + T_3. \end{aligned}$$

To control T_3 , we can write

$$\begin{aligned} T_3 &= \|\mathcal{T}_K^{(j)}(Q^{(i)} + K^\top R^{(i)} K) - \mathcal{T}_K^{(j)}(Q^{(j)} + K^\top R^{(j)} K)\| \\ &\leq \|\mathcal{T}_K^{(j)}(Q^{(i)} + K^\top R^{(i)} K - Q^{(j)} - K^\top R^{(j)} K)\| \\ &\leq \frac{J_{\max}(K)(\epsilon_3 + \epsilon_4 \|K\|^2)}{\min_i \sigma_{\min}(Q^{(i)})\mu}, \end{aligned}$$

then we finally have the following expression for the upper bound of $\|P_K^{(i)} - P_K^{(j)}\|$

$$\begin{aligned} \|P_K^{(i)} - P_K^{(j)}\| &\leq 4 \left(\frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})\mu} \right)^2 (\epsilon_1 + \epsilon_2 \|K\|) \|A - BK\|_{\max} (\|Q\|_{\max} + \|K\|^2 \|R\|_{\max}) \\ &\quad + \frac{J_{\max}(K)(\epsilon_3 + \epsilon_4 \|K\|^2)}{\min_i \sigma_{\min}(Q^{(i)})\mu}. \end{aligned}$$

Therefore, by using the above expression in T_2 we can derive the upper bound of $\|E_K^{(i)} - E_K^{(j)}\|$ as follows

$$\begin{aligned} \|E_K^{(i)} - E_K^{(j)}\| &\leq \epsilon_1 \|B\|_{\max} (\|P_K\|_{\max} + 4\|A - BK\|_{\max}^2 (\|Q\|_{\max} + \|K\|^2 \|R\|_{\max}) \nu^2(K)) \\ &\quad + \epsilon_2 (\|A - BK\|_{\max} \|P_K\|_{\max} + \|B\|_{\max} \|P_K\|_{\max} \|K\| \\ &\quad \quad \quad + 4\|K\| \|A - BK\|_{\max}^2 (\|Q\|_{\max} + \|K\|^2 \|R\|_{\max}) \nu^2(K)) \\ &\quad + \epsilon_3 \|B\|_{\max} \|A - BK\|_{\max} \nu(K) + \epsilon_4 (\|K\| + \|K\|^2 \|B\|_{\max} \|A - BK\|_{\max} \nu(K)), \end{aligned}$$

which can then be combined with the upper bound of $\|\Sigma_K^{(i)} - \Sigma_K^{(j)}\|$, $\|\Sigma_K^{(i)}\|$, and $\|E_K^{(j)}\|$ to obtain

$$\|\nabla J^{(i)}(K) - \nabla J^{(j)}(K)\| \leq \bar{f}_z(\bar{\epsilon}) := \epsilon_1 h_{\text{het}}^1(K) + \epsilon_2 h_{\text{het}}^2(K) + \epsilon_3 h_{\text{het}}^3(K) + \epsilon_4 h_{\text{het}}^4(K),$$

with,

$$h_{\text{het}}^1(K) = \frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})} \|B\|_{\max} (\|P_K\|_{\max} + 4\|A - BK\|_{\max}^2 (\|Q\|_{\max} + \|K\|^2 \|R\|_{\max}) \nu^2(K))$$

$$\begin{aligned}
& + 4 \left(\|R\|_{\max} \|K\| + \frac{J_{\max}(K)(\|B\|_{\max}^2 \|K\| + \|B\|_{\max} \|A\|_{\max})}{\mu} \right) \nu^2(K) \|A - BK\|_{\max} \|\Sigma_0\|, \\
h_{\text{het}}^2(K) &= \frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})} (\|A - BK\|_{\max} \|P_K\|_{\max} + \|B\|_{\max} \|P_K\|_{\max} \|K\| \\
& + 4 \|K\| \|A - BK\|_{\max}^2 (\|Q\|_{\max} + \|K\|^2 \|R\|_{\max}) \nu^2(K)) \\
& + 4 \left(\|R\|_{\max} \|K\| + \frac{J_{\max}(K)(\|B\|_{\max}^2 \|K\| + \|B\|_{\max} \|A\|_{\max})}{\mu} \right) \nu^2(K) \|K\| \|A - BK\|_{\max} \|\Sigma_0\|, \\
h_{\text{het}}^3(K) &= \frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})} \|B\|_{\max} \|A - BK\|_{\max} \nu(K), \\
h_{\text{het}}^4(K) &= \frac{J_{\max}(K)}{\min_i \sigma_{\min}(Q^{(i)})} (\|K\| + \|K\|^2 \|B\|_{\max} \|A - BK\|_{\max} \nu(K)).
\end{aligned}$$

6.3. Proof of Theorem 10

Objective: Given an initial stabilizing controller $K_0 \in \mathcal{S}_{\text{ML}}$, our aim is to provide the conditions on the step-sizes η_l , η and heterogeneity $\bar{f}_z(\bar{\epsilon})$, such that $K_{n+1} \in \mathcal{S}_{\text{ML}}$ for all $n \in \{0, 1, \dots, N-1\}$.

We first observe that, if $K \in \mathcal{S}_{\text{ML}}$, one may select η_l such that $\bar{K} = K - \eta_l \nabla J^{(i)}(K)$ is also in the MAML stabilizing sub-level set \mathcal{S}_{ML} . To verify it, we can use the local smoothness property of the LQR cost (Lemma 6) to obtain

$$\begin{aligned}
J^{(i)}(\bar{K}) - J^{(i)}(K) &\leq \langle \nabla J^{(i)}(K), \bar{K} - K \rangle + \frac{h_{\text{grad}}(K)}{2} \|\bar{K} - K\|_F^2 \\
&\leq -\eta_l \langle \nabla J^{(i)}(K), \nabla J^{(i)}(K) \rangle + \frac{h_{\text{grad}}(K) \eta_l^2}{2} \|\nabla J^{(i)}(K)\|_F^2 \\
&\leq \left(\frac{h_{\text{grad}}(K) \eta_l^2}{2} - \eta_l \right) \|\nabla J^{(i)}(K)\|_F^2 \\
&\stackrel{(i)}{\leq} -\frac{\eta_l}{2} \|\nabla J^{(i)}(K)\|_F^2,
\end{aligned}$$

where (i) follows from $\eta_l \leq \frac{1}{h_{\text{grad}}(K)}$. We now use the gradient domination property of the LQR cost (Lemma 7) to write

$$J^{(i)}(\bar{K}) - J^{(i)}(K_i^*) \leq \left(1 - \frac{\eta_l \lambda_i}{2} \right) (J^{(i)}(K) - J^{(i)}(K_i^*)),$$

which implies that $J^{(i)}(\bar{K}) - J^{(i)}(K_i^*) \leq J^{(i)}(K) - J^{(i)}(K_i^*)$. Then, since $K \in \mathcal{S}_{\text{ML}}$ we conclude that $J^{(i)}(\bar{K}) - J^{(i)}(K_i^*) \leq J^{(i)}(K_0) - J^{(i)}(K_i^*)$, which leads to $\bar{K} \in \mathcal{S}_{\text{ML}}$. Therefore, given $K \in \mathcal{K}$ and by selecting the step-size according to $\eta_l \leq \frac{1}{h_{\text{grad}}(K)}$ we have $\bar{K} = K - \eta_l \nabla J^{(i)}(K) \in \mathcal{S}_{\text{ML}}$. We now proceed to show that $K_1 \in \mathcal{S}_{\text{ML}}$. For this purpose, we consider the first iteration of the MAML-LQR algorithm (Algorithm 1) as the base case. Then, at the first iteration, we have

$$K_1 = K_0 - \frac{\eta}{M} \sum_{j=1}^M H^{(j)}(K_0) \nabla J^{(j)}(K_0 - \eta_l \nabla J^{(j)}(K_0)),$$

and by using the local smoothness of the LQR cost, we can write

$$\begin{aligned}
 J^{(i)}(K_1) - J^{(i)}(K_0) &\leq \langle \nabla J^{(i)}(K_0), K_1 - K_0 \rangle + \frac{\bar{h}_{\text{grad}}}{2} \|K_1 - K_0\|_F^2 \\
 &= \langle \nabla J^{(i)}(K_0), -\eta \nabla J_{\text{ML}}(K_0) - \eta \nabla J^{(i)}(K_0) + \eta \nabla J^{(i)}(K_0) \rangle + \frac{\bar{h}_{\text{grad}} \eta^2}{2} \|\nabla J_{\text{ML}}(K_0)\|_F^2 \\
 &\leq -\frac{\eta}{2} \|\nabla J^{(i)}(K_0)\|_F^2 + \frac{\eta}{2} \|\nabla J_{\text{ML}}(K_0) - \nabla J^{(i)}(K_0)\|_F^2 + \frac{\bar{h}_{\text{grad}} \eta^2}{2} \|\nabla J_{\text{ML}}(K_0)\|_F^2
 \end{aligned}$$

and by using Young's inequality (8) along with the selection of the step-size $\eta \leq \frac{1}{4\bar{h}_{\text{grad}}}$ we can write

$$J^{(i)}(K_1) - J^{(i)}(K_0) \leq -\frac{\eta}{4} \|\nabla J^{(i)}(K_0)\|_F^2 + \frac{3\eta}{4} \|\nabla J_{\text{ML}}(K_0) - \nabla J^{(i)}(K_0)\|_F^2.$$

As we show in details in the proof of Theorem 12, we control the term $\|\nabla J_{\text{ML}}(K_0) - \nabla J^{(i)}(K_0)\|_F^2$ as follows

$$\|\nabla J_{\text{ML}}(K_0) - \nabla J^{(i)}(K_0)\|_F^2 \leq 24n_u^3 f_z^2(\bar{\epsilon}) + 2\eta_l^2 (12\bar{h}_{\text{grad}}^2 n_u^2 + h_H^2) \|\nabla J^{(i)}(K_0)\|_F^2$$

where we impose $\eta_l \leq \min \left\{ \frac{n_u}{\sqrt{2}h_H}, \frac{1}{\sqrt{2}h_{\text{grad}}} \right\}$. Then, we obtain

$$J^{(i)}(K_1) - J^{(i)}(K_0) \leq -\frac{\eta}{8} \|\nabla J^{(i)}(K_0)\|_F^2 + 18\eta n_u^3 f_z^2(\bar{\epsilon})$$

by selecting $\eta_l \leq \frac{1}{\sqrt{12(12\bar{h}_{\text{grad}}^2 n_u^2 + \bar{h}_H^2)}}$. Then, we use the gradient domination property of the LQR cost to write

$$J^{(i)}(K_1) - J^{(i)}(K_i^*) \leq \left(1 - \frac{\eta \lambda_i}{8}\right) (J^{(i)}(K_0) - J^{(i)}(K_i^*)) + 18\eta n_u f_z^2(\bar{\epsilon})$$

and by assuming that

$$18n_u^3 f_z^2(\bar{\epsilon}) \leq \min_{i \in [M]} \frac{\lambda_i}{16} (J^{(i)}(K_0) - J^{(i)}(K_i^*))$$

we obtain

$$J^{(i)}(K_1) - J^{(i)}(K_i^*) \leq \left(1 - \frac{\eta \lambda_i}{16}\right) (J^{(i)}(K_0) - J^{(i)}(K_i^*))$$

which implies

$$J^{(i)}(K_1) - J^{(i)}(K_i^*) \leq J^{(i)}(K_0) - J^{(i)}(K_i^*).$$

Therefore, we know that under the above selection of step-sizes and heterogeneity we have that $K_1, K_1 - \eta_l \nabla J^{(i)}(K_1) \in \mathcal{S}_{\text{ML}}$, for all $i \in [M]$. The stability analysis is completed by applying an induction step on the base case for all iterations $n \in \{0, 1, \dots, N-1\}$.

6.4. Proof of Theorem 12

Objective: Given a set of LQR tasks \mathcal{T} , where each task-specific optimal controller is K_i^* . We aim to derive the gap between K_N (i.e., the controller obtained Algorithm 1) and K_i^* . We do so, by deriving the bound for $J^{(i)}(K_N) - J^{(i)}(K_i^*)$, for all $i \in [M]$. Moreover, we also characterize the gap between K_{ML}^* and K_i^* through $J^{(i)}(K_{ML}^*) - J^{(i)}(K_i^*)$. With both bounds in hands, we are able to conclude about the local convergence of Algorithm 1.

Remark 19 *Observe that, in general, we cannot provide any guarantee about the global convergence for the MAML-LQR approach. This is due to the summation in (3) over heterogeneous LQR tasks, and it prevent us of having a nice MAML-LQR optimization landscape, i.e., a landscape where gradient domination type of properties are satisfied.*

Controlling $J^{(i)}(K_N) - J^{(i)}(K_i^*)$: We recall that the MAML-LQR update (4) is given by $K_{n+1} = K_n - \eta \nabla J_{ML}(K_n)$ with $\nabla J_{ML}(K_n) = \frac{1}{M} \sum_{j=1}^M H^{(j)}(K_n) \nabla J^{(j)}(K_n - \eta_l \nabla J^{(j)}(K_n))$. Therefore, to proceed we can use the fact that $K_n, K_{n+1} \in \mathcal{S}_{ML}$ along with the local smoothness property of each LQR task's objective to write

$$\begin{aligned} J^{(i)}(K_{n+1}) - J^{(i)}(K_n) &\leq \langle \nabla J^{(i)}(K_n), K_{n+1} - K_n \rangle + \frac{\bar{h}_{\text{grad}}}{2} \|K_{n+1} - K_n\|_F^2 \\ &= \langle \nabla J^{(i)}(K_n), -\eta \nabla J_{ML}(K_n) - \eta \nabla J^{(i)}(K_n) + \eta \nabla J^{(i)}(K_n) \rangle + \frac{\bar{h}_{\text{grad}} \eta^2}{2} \|\nabla J_{ML}(K_n)\|_F^2 \\ &\leq -\frac{\eta}{2} \|\nabla J^{(i)}(K_n)\|_F^2 + \frac{\eta}{2} \|\nabla J_{ML}(K_n) - \nabla J^{(i)}(K_n)\|_F^2 + \frac{\bar{h}_{\text{grad}} \eta^2}{2} \|\nabla J_{ML}(K_n)\|_F^2, \end{aligned}$$

where we use Young's inequality (8) in the last inequality. We re-write $\|\nabla J_{ML}(K_n)\|_F^2$ as follows:

$$\begin{aligned} \|\nabla J_{ML}(K_n)\|_F^2 &= \|\nabla J_{ML}(K_n) - \nabla J^{(i)}(K_n) + \nabla J^{(i)}(K_n)\|_F^2 \\ &\leq 2\|\nabla J_{ML}(K_n) - \nabla J^{(i)}(K_n)\|_F^2 + 2\|\nabla J^{(i)}(K_n)\|_F^2, \end{aligned}$$

by using Young's inequality (8). We use the expression of $\|\nabla J_{ML}(K_n)\|_F^2$ to obtain

$$\begin{aligned} J^{(i)}(K_{n+1}) - J^{(i)}(K_n) &\leq \left(-\frac{\eta}{2} + \bar{h}_{\text{grad}} \eta^2\right) \|\nabla J^{(i)}(K_n)\|_F^2 \\ &\quad + \left(\frac{\eta}{2} + \bar{h}_{\text{grad}} \eta^2\right) \|\nabla J_{ML}(K_n) - \nabla J^{(i)}(K_n)\|_F^2 \\ &\leq -\frac{\eta}{4} \|\nabla J^{(i)}(K_n)\|_F^2 + \frac{3\eta}{4} \|\nabla J_{ML}(K_n) - \nabla J^{(i)}(K_n)\|_F^2, \end{aligned}$$

where the last inequality follows from the selection $\eta \leq \frac{1}{4\bar{h}_{\text{grad}}}$. We now proceed to control the gradient term $\|\nabla J_{ML}(K_n) - \nabla J^{(i)}(K_n)\|_F^2$. We do it for a general stabilizing policy $K \in \mathcal{S}_{ML}$ below.

$$\|\nabla J_{ML}(K) - \nabla J^{(i)}(K)\|_F^2 = \left\| \frac{1}{M} \sum_{j=1}^M H^{(j)}(K) \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2$$

$$\leq \frac{1}{M} \sum_{j=1}^M \left\| H^{(j)}(K) \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2,$$

with $\bar{K} := K - \eta_l \nabla J^{(j)}(K)$. Then, since $\eta_l \leq \frac{1}{h_{\text{grad}}}$ we also have $\bar{K} \in \mathcal{S}_{\text{ML}}$. We emphasize that the last inequality is due to Jensen's inequality. We then proceed to bound the term inside the sum.

$$\begin{aligned} \left\| H^{(j)}(K) \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2 &= \left\| H^{(j)}(K) \nabla J^{(j)}(\bar{K}) - H^{(j)}(K) \nabla J^{(i)}(K) \right. \\ &\quad \left. + H^{(j)}(K) \nabla J^{(i)}(K) - \nabla J^{(i)}(K) \right\|_F^2 \\ &\leq 2 \left\| H^{(j)}(K) \right\|_F^2 \left\| \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2 \\ &\quad + 2 \left\| (H^{(j)}(K) - I) \nabla J^{(i)}(K) \right\|_F^2 \\ &= 2 \left\| H^{(j)}(K) \right\|_F^2 \left\| \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2 \\ &\quad + 2\eta_l^2 \left\| \nabla^2 J^{(j)}(K) \right\|_F^2 \left\| \nabla J^{(i)}(K) \right\|_F^2 \\ &\stackrel{(i)}{\leq} 4(n_u^2 + \eta_l^2 \bar{h}_H^2) \left\| \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2 \\ &\quad + 2\eta_l^2 \bar{h}_H^2 \left\| \nabla J^{(i)}(K) \right\|_F^2, \end{aligned}$$

where (i) is due to the fact $\left\| H^{(j)}(K) \right\|_F^2 = \left\| I_{n_u} - \eta_l \nabla^2 J^{(j)}(K) \right\|_F^2 \leq 2(n_u^2 + \eta_l^2 \left\| \nabla^2 J^{(j)}(K) \right\|_F^2)$ along with the uniform bound for the Hessian in Lemma 5. Then, we now proceed to control $\left\| \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2$.

$$\begin{aligned} \left\| \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2 &= \left\| \nabla J^{(j)}(\bar{K}) - \nabla J^{(j)}(K) + \nabla J^{(j)}(K) - \nabla J^{(i)}(K) \right\|_F^2 \\ &\leq 2 \underbrace{\left\| \nabla J^{(j)}(\bar{K}) - \nabla J^{(j)}(K) \right\|_F^2}_{\text{grad. Lipschitz}} + 2 \underbrace{\left\| \nabla J^{(j)}(K) - \nabla J^{(i)}(K) \right\|_F^2}_{\text{grad. heterogeneity}} \\ &\leq 2\bar{h}_{\text{grad}}^2 \eta_l^2 \left\| \nabla J^{(j)}(K) \right\|_F^2 + 2n_u f_z^2(\bar{\epsilon}) \end{aligned}$$

where in the last term we use the relationship between Frobenius and spectral norm, since the gradient heterogeneity bound was previously derived in the spectral norm. We can also use the fact that $\left\| \nabla J^{(j)}(K) \right\|_F^2 \leq 2 \left\| \nabla J^{(j)}(K) - \nabla J^{(i)}(K) \right\|_F^2 + 2 \left\| \nabla J^{(i)}(K) \right\|_F^2$ to obtain

$$\left\| \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2 \leq 2n_u f_z^2(\bar{\epsilon}) (2\bar{h}_{\text{grad}}^2 \eta_l^2 + 1) + 4\bar{h}_{\text{grad}}^2 \eta_l^2 \left\| \nabla J^{(i)}(K) \right\|_F^2,$$

which implies

$$\begin{aligned} \left\| H^{(j)}(K) \nabla J^{(j)}(\bar{K}) - \nabla J^{(i)}(K) \right\|_F^2 &\leq 8(n_u^2 + \eta_l^2 \bar{h}_H^2) (2\bar{h}_{\text{grad}}^2 \eta_l^2 + 1) n_u f_z^2(\bar{\epsilon}) \\ &\quad + [16(n_u^2 + \eta_l^2 \bar{h}_H^2) \bar{h}_{\text{grad}}^2 \eta_l^2 + 2\eta_l^2 \bar{h}_H^2] \left\| \nabla J^{(i)}(K) \right\|_F^2, \\ &\stackrel{(i)}{\leq} 24n_u^3 f_z^2(\bar{\epsilon}) + 2\eta_l^2 (12\bar{h}_{\text{grad}}^2 n_u^2 + \bar{h}_H^2) \left\| \nabla J^{(i)}(K) \right\|_F^2, \end{aligned}$$

where (i) follows from the selection $\eta_l \leq \min \left\{ \frac{n_u}{\sqrt{2}h_H}, \frac{1}{\sqrt{2}h_{\text{grad}}} \right\}$. Then, we obtain

$$\left\| \nabla J_{\text{ML}}(K) - \nabla J^{(i)}(K) \right\|_F^2 \leq 24n_u^3 f_z^2(\bar{\epsilon}) + 2\eta_l^2 (12\bar{h}_{\text{grad}}^2 n_u^2 + \bar{h}_H^2) \left\| \nabla J^{(i)}(K) \right\|_F^2.$$

which yields to

$$\begin{aligned} J^{(i)}(K_{n+1}) - J^{(i)}(K_n) &\leq \left(-\frac{\eta}{4} + \frac{3\eta\eta_l^2}{2}(12\bar{h}_{\text{grad}}^2 n_u^2 + h_H^2) \right) \|\nabla J^{(i)}(K_n)\|_F^2 + 18\eta n_u^3 f_z^2(\bar{\epsilon}) \\ &\stackrel{(i)}{\leq} -\frac{\eta}{8} \|\nabla J^{(i)}(K_n)\|_F^2 + 18\eta n_u^3 f_z^2(\bar{\epsilon}) \end{aligned}$$

where (i) follows from the selection $\eta_l \leq \frac{1}{\sqrt{12(12\bar{h}_{\text{grad}}^2 n_u^2 + \bar{h}_H^2)}}$. Therefore, we use the gradient domination property of the LQR task cost (Lemma 7) to write

$$J^{(i)}(K_{n+1}) - J^{(i)}(K_i^*) \leq \left(1 - \frac{\eta\lambda_i}{8} \right) (J^{(i)}(K_n) - J^{(i)}(K_i^*)) + 18\eta n_u^3 f_z^2(\bar{\epsilon})$$

and by telescoping the above expression over $n = \{0, 1, \dots, N-1\}$ we finally obtain

$$\begin{aligned} J^{(i)}(K_N) - J^{(i)}(K_i^*) &\leq \left(1 - \frac{\eta\lambda_i}{8} \right)^N (J^{(i)}(K_0) - J^{(i)}(K_i^*)) + 18\eta n_u^3 f_z^2(\bar{\epsilon}) \\ &= \left(1 - \frac{\eta\lambda_i}{8} \right)^N \Delta_0^{(i)} + \frac{144}{\lambda_i} n_u^3 f_z^2(\bar{\epsilon}) \end{aligned}$$

and by selecting the number of iterations $N \geq \frac{8}{\eta\lambda_i} \log \left(\frac{\Delta_0^{(i)}}{\epsilon'} \right)$ for a small tolerance ϵ' , we have

$$J^{(i)}(K_N) - J^{(i)}(K_i^*) \leq \epsilon' + \frac{144}{\lambda_i} n_u^3 f_z^2(\bar{\epsilon}).$$

Controlling $J^{(i)}(K_{\text{ML}}^*) - J^{(i)}(K_i^*)$: With the bound for $J^{(i)}(K_N) - J^{(i)}(K_i^*)$ in hands, we now proceed to derive the one for K_{ML}^* and K_i^* . To do so we can follow a similar approach as presented in (Wang et al., 2023b). First we recall that

$$\limsup_{n \rightarrow \infty} J^{(i)}(K_n) - J^{(i)}(K_i^*) \leq \frac{144}{\lambda_i} n_u^3 f_z^2(\bar{\epsilon})$$

where we can use the gradient domination property as proved in (Fazel et al., 2018) to write

$$\begin{aligned} J^{(i)}(K_{\text{ML}}^*) - \limsup_{n \rightarrow \infty} J^{(i)}(K_n) &\leq \liminf_{n \rightarrow \infty} \text{Tr} \left(\Sigma_{K_n}^{(i)} E_{K_{\text{ML}}^*}^{(i)\top} \left(R^{(i)} + B^{(i)\top} P_{K_{\text{ML}}^*}^{(i)} B^{(i)} \right)^{-1} E_{K_{\text{ML}}^*}^{(i)} \right) \\ &\leq \liminf_{n \rightarrow \infty} \left\| \Sigma_{K_n}^{(i)} \right\| \text{Tr} \left(E_{K_{\text{ML}}^*}^{(i)\top} \left(R^{(i)} + B^{(i)\top} P_{K_{\text{ML}}^*}^{(i)} B^{(i)} \right)^{-1} E_{K_{\text{ML}}^*}^{(i)} \right) \\ &\leq \liminf_{n \rightarrow \infty} \frac{\bar{J}_{\text{max}}}{\min_i \sigma_{\min}(Q^{(i)})} \left\| \left(R^{(i)} + B^{(i)\top} P_{K_{\text{ML}}^*}^{(i)} B^{(i)} \right)^{-1} \right\| \text{Tr} \left(E_{K_{\text{ML}}^*}^{(i)\top} E_{K_{\text{ML}}^*}^{(i)} \right) \\ &\leq \frac{\bar{J}_{\text{max}}}{\min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)})} \text{Tr} \left(E_{K_{\text{ML}}^*}^{(i)\top} E_{K_{\text{ML}}^*}^{(i)} \right) \\ &= \frac{\bar{J}_{\text{max}}}{\min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)})} \text{Tr} \left(\left(\Sigma_{K_{\text{ML}}^*}^{(i)} \right)^{-1} \nabla J^{(i)\top}(K_{\text{ML}}^*) \nabla J^{(i)}(K_{\text{ML}}^*) \left(\Sigma_{K_{\text{ML}}^*}^{(i)} \right)^{-1} \right) \end{aligned}$$

$$\begin{aligned}
 &\leq \frac{\bar{J}_{\max}}{\min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)}) \sigma_{\min}(\Sigma_{K^*}^{(i)})^2} \text{Tr} \left(\nabla J^{(i)\top}(K_{\text{ML}}^*) \nabla J^{(i)}(K_{\text{ML}}^*) \right) \\
 &\leq \frac{\bar{J}_{\max}}{\mu^2 \min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)})} \|\nabla J^{(i)}(K_{\text{ML}}^*)\|_F^2,
 \end{aligned}$$

now we proceed to control $\|\nabla J^{(i)}(K_{\text{ML}}^*)\|_F^2$. To do so, we first write

$$\begin{aligned}
 \|\nabla J^{(i)}(K_{\text{ML}}^*)\|_F^2 &\leq 2 \underbrace{\|\nabla J_{\text{ML}}(K_{\text{ML}}^*)\|_F^2}_{=0} + 2 \|\nabla J^{(i)}(K_{\text{ML}}^*) - \nabla J_{\text{ML}}(K_{\text{ML}}^*)\|_F^2 \\
 &\leq 48n_u^3 f_z^2(\bar{\epsilon}) + 4\eta_l^2 (12\bar{h}_{\text{grad}}^2 n_u^2 + h_H^2) \|\nabla J^{(i)}(K_{\text{ML}}^*)\|_F^2
 \end{aligned}$$

where the last inequality follows from the upper bound of $\|\nabla J_{\text{ML}}(K) - \nabla J^{(i)}(K)\|_F^2$. Therefore, we can select $\eta_l \leq \frac{1}{\sqrt{8(12\bar{h}_{\text{grad}}^2 n_u^2 + h_H^2)}}$ to write

$$\|\nabla J^{(i)}(K_{\text{ML}}^*)\|_F^2 \leq 96n_u^3 f_z^2(\bar{\epsilon})$$

which can be used to write

$$J^{(i)}(K_{\text{ML}}^*) - \limsup_{n \rightarrow \infty} J^{(i)}(K_n) \leq \frac{96\bar{J}_{\max} n_u^3 f_z^2(\bar{\epsilon})}{\mu^2 \min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)})}$$

which implies

$$J^{(i)}(K_{\text{ML}}^*) - J^{(i)}(K_i^*) \leq n_u^3 f_z^2(\bar{\epsilon}) \left(\frac{144}{\lambda_i} + \frac{96\bar{J}_{\max}}{\mu^2 \min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)})} \right),$$

and completes the proof.

6.5. Proof of Theorem 11

We now move our attention to the model-free setting, where both gradient and Hessian are estimated through a zeroth-order estimation with two-point estimation. Similar to the model-based setting, we begin the stability analysis by setting the base case as the first iteration Algorithm 3. Before proceeding to show that $K_1 \in \mathcal{S}_{\text{ML}}$, we can first prove that given an initial stabilizing controller $K_0 \in \mathcal{S}_{\text{ML}}$, one may select η_l , r and m to ensure that $\hat{K}_0 := K_0 - \eta_l \hat{\nabla} J^{(i)}(K_0) \in \mathcal{S}_{\text{ML}}$. To do so, we use the local smoothness of the LQR cost (Lemma 6) to write

$$\begin{aligned}
 J^{(i)}(\hat{K}_0) - J^{(i)}(K_0) &\leq \langle \nabla J^{(i)}(K_0), \hat{K}_0 - K_0 \rangle + \frac{\bar{h}_{\text{grad}}}{2} \|\hat{K}_0 - K_0\|_F^2 \\
 &= \langle \nabla J^{(i)}(K_0), -\eta_l \hat{\nabla} J^{(i)}(K_0) \rangle + \frac{\bar{h}_{\text{grad}} \eta_l^2}{2} \|\hat{\nabla} J^{(i)}(K_0)\|_F^2 \\
 &\leq -\frac{\eta_l}{2} \|\nabla J^{(i)}(K_0)\|_F^2 + \frac{\eta_l}{2} \|\hat{\nabla} J^{(i)}(K_0) - \nabla J^{(i)}(K_0)\|_F^2 + \frac{\bar{h}_{\text{grad}} \eta_l^2}{2} \|\hat{\nabla} J^{(i)}(K_0)\|_F^2 \\
 &\leq \left(\bar{h}_{\text{grad}} \eta_l^2 - \frac{\eta_l}{2} \right) \|\nabla J^{(i)}(K_0)\|_F^2 + \left(\bar{h}_{\text{grad}} \eta_l^2 + \frac{\eta_l}{2} \right) \|\hat{\nabla} J^{(i)}(K_0) - \nabla J^{(i)}(K_0)\|_F^2
 \end{aligned}$$

$$\stackrel{(i)}{\leq} -\frac{\eta_l}{4} \|\nabla J^{(i)}(K_0)\|_F^2 + \frac{3\eta_l}{4} \|\widehat{\nabla} J^{(i)}(K_0) - \nabla J^{(i)}(K_0)\|_F^2,$$

where (i) follows from the selection $\eta_l \leq \frac{1}{4h_{\text{grad}}}$. Therefore, by using the gradient domination property of the LQR cost (Lemma 7), we obtain

$$\begin{aligned} J^{(i)}(\widehat{K}_0) - J^{(i)}(K_i^*) &\leq \left(1 - \frac{\eta_l \lambda_i}{4}\right) \left(J^{(i)}(K_0) - J^{(i)}(K_i^*)\right) + \frac{3\eta_l}{4} \|\widehat{\nabla} J^{(i)}(K_0) - \nabla J^{(i)}(K_0)\|_F^2 \\ &= \left(1 - \frac{\eta_l \lambda_i}{4}\right) \Delta_0^{(i)} + \frac{3\eta_l}{4} \|\widehat{\nabla} J^{(i)}(K_0) - \nabla J^{(i)}(K_0)\|_F^2, \end{aligned}$$

and since $K_0 \in \mathcal{S}_{\text{ML}}$, we can select the smoothing radius r and the number of samples m according to Lemma 17 such that $\|\widehat{\nabla} J^{(i)}(K_0) - \nabla J^{(i)}(K_0)\|_F^2 \leq \psi := \frac{\lambda_i \Delta_0^{(i)}}{6}$. Then, we obtain

$$r \leq \min \left\{ h_r^1 \left(\frac{\sqrt{\psi}}{2} \right) \right\}, \quad m \geq \max \left\{ \bar{h}_m^1 \left(\frac{\sqrt{\psi}}{2}, \delta \right) \right\}.$$

with probability $1 - \delta$. This implies that

$$\begin{aligned} J^{(i)}(\widehat{K}_0) - J^{(i)}(K_i^*) &\leq \left(1 - \frac{\eta_l \lambda_i}{8}\right) \left(J^{(i)}(K_0) - J^{(i)}(K_i^*)\right) \\ &\leq J^{(i)}(K_0) - J^{(i)}(K_i^*). \end{aligned}$$

which yields to $K_0 - \eta_l \widehat{\nabla} J^{(i)}(K_0) \in \mathcal{S}_{\text{ML}}$. We then proceed to show that $K_1 \in \mathcal{S}_{\text{ML}}$. By following similar derivations as in the convergence analysis, we can write

$$J^{(i)}(K_1) - J^{(i)}(K_0) \leq -\frac{\eta}{8} \|\nabla J^{(i)}(K_0)\|_F^2 + 30\eta n_u^3 f_z^2(\bar{\epsilon}) + \frac{9\eta}{8} \|\widehat{\nabla} J_{\text{ML}}(K_0) - \nabla J_{\text{ML}}(K_0)\|_F^2,$$

by selecting $\eta_l \leq \frac{1}{\sqrt{20(12\bar{h}_{\text{grad}}^2 n_u^2 + \bar{h}_H^2)}}$ and $\eta \leq \frac{1}{8h_{\text{grad}}}$. Therefore, by using the gradient domination property of the LQR cost (Lemma 7) we can write

$$\begin{aligned} J^{(i)}(K_1) - J^{(i)}(K_i^*) &\leq \left(1 - \frac{\eta \lambda_i}{8}\right) \left(J^{(i)}(K_0) - J^{(i)}(K_i^*)\right) + 30\eta n_u^3 f_z^2(\bar{\epsilon}) \\ &\quad + \frac{9\eta}{8} \|\widehat{\nabla} J_{\text{ML}}(K_0) - \nabla J_{\text{ML}}(K_0)\|_F^2, \end{aligned}$$

where we suppose that the heterogeneity satisfy $f_z^2(\bar{\epsilon}) \leq \frac{\lambda_i \Delta_0^{(i)}}{480n_u^3}$ to obtain

$$J^{(i)}(K_1) - J^{(i)}(K_i^*) \leq \left(1 - \frac{\eta \lambda_i}{16}\right) \Delta_0^{(i)} + \frac{9\eta}{8} \|\widehat{\nabla} J_{\text{ML}}(K_0) - \nabla J_{\text{ML}}(K_0)\|_F^2,$$

which implies

$$J^{(i)}(K_1) - J^{(i)}(K_i^*) \leq \left(1 - \frac{\eta \lambda_i}{32}\right) \Delta_0^{(i)}$$

by selecting the smoothing radius r and number of samples m according to

$$r \leq \min \left\{ \bar{h}_r^1 \left(\frac{\sqrt{\bar{\psi}}}{2} \right), \bar{h}_r^2 \left(\frac{\sqrt{\bar{\psi}}}{2} \right) \right\}, \quad m \geq \max \left\{ \bar{h}_m^1 \left(\frac{\sqrt{\bar{\psi}}}{2}, \delta \right), \bar{h}_m^2 \left(\frac{\sqrt{\bar{\psi}}}{2}, \delta \right) \right\}.$$

which follows from Lemmas 17 and 18, with $\bar{\psi} := \frac{\lambda_i \Delta_0^{(i)}}{1296}$. This implies that $K_1 \in \mathcal{S}_{\text{ML}}$. The stability analysis is completed by applying a induction step for all iterations $n \in \{0, 1, \dots, N-1\}$.

6.6. Proof of Theorem 13

Controlling $J^{(i)}(K_N) - J^{(i)}(K_i^*)$: We recall that model-free MAML-LQR update is given by $K_{n+1} = K_n - \eta \widehat{\nabla} J_{\text{ML}}(K_n)$ with $\widehat{\nabla} J_{\text{ML}}(K_n) = \frac{1}{M} \sum_{j=1}^M \widehat{H}^{(j)}(K_n) \widehat{\nabla} J^{(j)}(K_n - \eta_l \widehat{\nabla} J^{(j)}(K_n))$.

$$\begin{aligned} J^{(i)}(K_{n+1}) - J^{(i)}(K_n) &\leq \langle \nabla J^{(i)}(K_n), K_{n+1} - K_n \rangle + \frac{\bar{h}_{\text{grad}}}{2} \|K_{n+1} - K_n\|_F^2 \\ &= \langle \nabla J^{(i)}(K_n), -\eta \widehat{\nabla} J_{\text{ML}}(K_n) - \eta \nabla J^{(i)}(K_n) + \eta \nabla J^{(i)}(K_n) \rangle + \frac{\bar{h}_{\text{grad}} \eta^2}{2} \|\widehat{\nabla} J_{\text{ML}}(K_n)\|_F^2 \\ &\leq -\frac{\eta}{2} \|\nabla J^{(i)}(K_n)\|_F^2 + \frac{\eta}{2} \|\widehat{\nabla} J_{\text{ML}}(K_n) - \nabla J^{(i)}(K_n)\|_F^2 + \frac{\bar{h}_{\text{grad}} \eta^2}{2} \|\widehat{\nabla} J_{\text{ML}}(K_n)\|_F^2 \\ &\leq -\frac{\eta}{4} \|\nabla J^{(i)}(K_n)\|_F^2 + \frac{5\eta}{4} \|\nabla J_{\text{ML}}(K_n) - \nabla J^{(i)}(K_n)\|_F^2 + \frac{9\eta}{8} \|\widehat{\nabla} J_{\text{ML}}(K_n) - \nabla J_{\text{ML}}(K_n)\|_F^2 \end{aligned}$$

where the last inequality follows from the selection of the step-size $\eta \leq \frac{1}{8\bar{h}_{\text{grad}}}$ along with Young's inequality (8). Therefore, in contrast to the model-free setting, we need to control both $\|\nabla J_{\text{ML}}(K_n) - \nabla J^{(i)}(K_n)\|_F^2$ and $\|\widehat{\nabla} J_{\text{ML}}(K_n) - \nabla J_{\text{ML}}(K_n)\|_F^2$. From our previous derivations for the model-based setting we know that $\|\nabla J_{\text{ML}}(K_n) - \nabla J^{(i)}(K_n)\|_F^2$ can be controlled as follows:

$$\|\nabla J_{\text{ML}}(K) - \nabla J^{(i)}(K)\|_F^2 \leq 24n_u^3 f_z^2(\bar{\epsilon}) + 2\eta_l^2 (12\bar{h}_{\text{grad}}^2 n_u^2 + \bar{h}_H^2) \|\nabla J^{(i)}(K)\|_F^2$$

which can be used to obtain

$$J^{(i)}(K_{n+1}) - J^{(i)}(K_n) \leq -\frac{\eta}{8} \|\nabla J^{(i)}(K_n)\|_F^2 + 30\eta n_u^3 f_z^2(\bar{\epsilon}) + \frac{9\eta}{8} \|\widehat{\nabla} J_{\text{ML}}(K_n) - \nabla J_{\text{ML}}(K_n)\|_F^2$$

by selecting $\eta_l \leq \frac{1}{\sqrt{20(12\bar{h}_{\text{grad}}^2 n_u^2 + \bar{h}_H^2)}}$. On the other hand, to control $\|\widehat{\nabla} J_{\text{ML}}(K_n) - \nabla J_{\text{ML}}(K_n)\|_F^2$ we can first write

$$\begin{aligned} \|\widehat{\nabla} J_{\text{ML}}(K_n) - \nabla J_{\text{ML}}(K_n)\|_F &= \left\| \frac{1}{M} \sum_{i=1}^M \widehat{H}^{(i)}(K_n) \widehat{\nabla} J^{(i)}(K_n - \eta_l \widehat{\nabla} J^{(i)}(K_n)) \right. \\ &\quad \left. - H^{(i)}(K_n) \nabla J^{(i)}(K_n - \eta_l \nabla J^{(i)}(K_n)) \right\|_F \\ &\stackrel{(i)}{\leq} \frac{1}{M} \sum_{i=1}^M \|\widehat{H}^{(i)}(K_n) \widehat{\nabla} J^{(i)}(\widehat{K}_n) - H^{(i)}(K_n) \nabla J^{(i)}(\bar{K}_n)\|_F, \end{aligned}$$

where we control $(I) := \|\widehat{H}^{(i)}(K_n)\widehat{\nabla}J^{(i)}(\widehat{K}_n) - H^{(i)}(K_n)\nabla J^{(i)}(\bar{K}_n)\|_F$ as follows:

$$\begin{aligned}
(I) &= \|\widehat{H}^{(i)}(K_n)\widehat{\nabla}J^{(i)}(\widehat{K}_n) - H^{(i)}(K_n)\widehat{\nabla}J^{(i)}(\widehat{K}_n) + H^{(i)}(K_n)\widehat{\nabla}J^{(i)}(\widehat{K}_n) - H^{(i)}(K_n)\nabla J^{(i)}(\bar{K}_n)\|_F \\
&= \|\left(\widehat{H}^{(i)}(K_n) - H^{(i)}(K_n)\right)\widehat{\nabla}J^{(i)}(\widehat{K}_n) + H^{(i)}(K_n)\left(\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\right)\|_F \\
&\leq \|\widehat{H}^{(i)}(K_n) - H^{(i)}(K_n)\|_F \|\widehat{\nabla}J^{(i)}(\widehat{K}_n)\|_F + \|H^{(i)}(K_n)\|_F \|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F \\
&\stackrel{(i)}{\leq} \|\widehat{\nabla}^2 J^{(i)}(K_n) - \nabla^2 J^{(i)}(K_n)\|_F \|\widehat{\nabla}J^{(i)}(\widehat{K}_n)\|_F + (n_u + \eta_l \bar{h}_H) \|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F \\
&\stackrel{(ii)}{\leq} \eta_l \bar{h}_G \|\widehat{\nabla}^2 J^{(i)}(K_n) - \nabla^2 J^{(i)}(K_n)\|_F \\
&\quad + \left[(n_u + \eta_l \bar{h}_H) + \eta_l \|\widehat{\nabla}^2 J^{(i)}(K_n) - \nabla^2 J^{(i)}(K_n)\|_F\right] \|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F,
\end{aligned}$$

where (i) follows from the definitions of $H^{(i)}(K_n) = I_{n_u} - \eta \nabla^2 J^{(i)}(K_n)$ and $\widehat{H}^{(i)}(K_n) = I_{n_u} - \eta \widehat{\nabla}^2 J^{(i)}(K_n)$, along with $\|H^{(i)}(K_n)\|_F \leq n_u + \eta_l \bar{h}_H$. (ii) is due to $\|\widehat{\nabla}J^{(i)}(\widehat{K}_n)\|_F \leq \|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F + \|\nabla J^{(i)}(\bar{K}_n)\|_F$ and $\|\nabla J^{(i)}(\bar{K}_n)\|_F \leq \bar{h}_G$. Therefore, we proceed to bound $\|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F$. For this purpose, we can write

$$\begin{aligned}
\|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F &= \|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\widehat{K}_n) + \nabla J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F \\
&\leq \|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\widehat{K}_n)\|_F + \|\nabla J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F \\
&\stackrel{(i)}{\leq} \|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\widehat{K}_n)\|_F + \bar{h}_{\text{grad}}\eta_l \|\widehat{\nabla}J^{(i)}(K_n) - \nabla J^{(i)}(K_n)\|_F,
\end{aligned}$$

where (i) follow from the local smoothness property of the LQR cost (Lemma 6). Hence, since $K_n, \widehat{K}_n \in \mathcal{S}_{\text{ML}}$, we have with $1 - \delta$,

$$\|\widehat{\nabla}J^{(i)}(\widehat{K}_n) - \nabla J^{(i)}(\bar{K}_n)\|_F \leq \epsilon(1 + \bar{h}_{\text{grad}}\eta_l),$$

with r and m satisfying the conditions in Lemma 17.

$$\begin{aligned}
(I) &\leq \eta_l \bar{h}_G \|\widehat{\nabla}^2 J^{(i)}(K_n) - \nabla^2 J^{(i)}(K_n)\|_F \\
&\quad + \epsilon \left[(n_u + \eta_l \bar{h}_H) + \eta_l \|\widehat{\nabla}^2 J^{(i)}(K_n) - \nabla^2 J^{(i)}(K_n)\|_F\right] (1 + \bar{h}_{\text{grad}}\eta_l),
\end{aligned}$$

and by setting r and m according to the conditions in Lemma 18, we obtain

$$(I) \leq \eta_l \bar{h}_G \epsilon + \epsilon \left[(n_u + \eta_l \bar{h}_H) + \eta_l \epsilon\right] (1 + \bar{h}_{\text{grad}}\eta_l),$$

which implies that

$$\|\widehat{\nabla}J_{\text{ML}}(K_n) - \nabla J_{\text{ML}}(K_n)\|_F \leq \eta_l \bar{h}_G \epsilon + \epsilon \left[(n_u + \eta_l \bar{h}_H) + \eta_l \epsilon\right] (1 + \bar{h}_{\text{grad}}\eta_l) \stackrel{(i)}{\leq} 6n_u \epsilon,$$

where (i) follows from the selection $\eta_l \leq \min \left\{ \frac{1}{\bar{h}_G}, \frac{n_u}{\bar{h}_H}, \frac{1}{\bar{h}_{\text{grad}}}, \frac{1}{2} \right\}$ and from the fact that $0 < \epsilon < 1$. Therefore, we obtain

$$J^{(i)}(K_{n+1}) - J^{(i)}(K_n) \leq -\frac{\eta}{8} \|\nabla J^{(i)}(K_n)\|_F^2 + 30\eta n_u^3 f_z^2(\bar{\epsilon}) + \frac{324n_u \eta \epsilon}{8},$$

where we can use the gradient domination property of the LQR cost to write

$$J^{(i)}(K_{n+1}) - J^{(i)}(K_i^*) \leq \left(1 - \frac{\eta\lambda_i}{8}\right) \left(J^{(i)}(K_n) - J^{(i)}(K_i^*)\right) + 30\eta n_u^3 f_z^2(\bar{\epsilon}) + \frac{324n_u\eta\epsilon}{8},$$

and we can unroll the above expression over $n = \{0, 1, \dots, N-1\}$ to obtain

$$J^{(i)}(K_N) - J^{(i)}(K_i^*) \leq \left(1 - \frac{\eta\lambda_i}{8}\right)^N \left(J^{(i)}(K_0) - J^{(i)}(K_i^*)\right) + \frac{240}{\lambda_i} n_u^3 f_z^2(\bar{\epsilon}) + \frac{324n_u\epsilon}{\lambda_i},$$

where by selecting $N \geq \frac{8}{\eta\lambda_i} \log\left(\frac{2\Delta_0^{(i)}}{\epsilon'}\right)$ and $\epsilon := \frac{\epsilon'\lambda_i}{648n_u}$, we obtain

$$J^{(i)}(K_N) - J^{(i)}(K_i^*) \leq \epsilon' + \frac{240n_u^3 f_z^2(\bar{\epsilon})}{\lambda_i}.$$

Controlling $J^{(i)}(K_{\text{ML}}^*) - J^{(i)}(K_i^*)$: By following similar derivation as in the model-based setting, we obtain

$$J^{(i)}(K_{\text{ML}}^*) - J^{(i)}(K_i^*) \leq n_u^3 f_z^2(\bar{\epsilon}) \left(\frac{240}{\lambda_i} + \frac{96\bar{J}_{\max}}{\mu^2 \min_i \sigma_{\min}(R^{(i)}) \min_i \sigma_{\min}(Q^{(i)})} \right),$$

which completes the proof.

7. Experimental Results - Setup Details

We now provide some details on the experimental setup considered in this work.

Task Generative Process: Given a nominal LQR task (A, B, Q, R) we generate M similar but not identical LQR tasks as follows

1. We first generate random scalar factors $a^{(i)} \sim \mathcal{U}(0, \epsilon_1)$, $b^{(i)} \sim \mathcal{U}(0, \epsilon_2)$, $q^{(i)} \sim \mathcal{U}(0, \epsilon_3)$, and $r^{(i)} \sim \mathcal{U}(0, \epsilon_4)$, $\forall i \in [M]$, with $\epsilon_1, \epsilon_2, \epsilon_3$ and ϵ_4 being the heterogeneity levels.
2. These random vectors are combined with modification masks $Z_a \in \mathbb{R}^{n_x \times n_x}$, $Z_b \in \mathbb{R}^{n_x \times n_u}$, $Z_q \in \mathbb{R}^{n_u \times n_x}$ and $Z_r \in \mathbb{R}^{n_u \times n_u}$ to generate multiple LQR tasks $(A^{(i)}, B^{(i)}, Q^{(i)}, R^{(i)})$ for all $i \in [M]$.
3. The LQR tasks $(A^{(i)}, B^{(i)}, Q^{(i)}, R^{(i)})$ are then constructed according to: $A^{(i)} = A + a^{(i)}Z_a$, $B^{(i)} = B + b^{(i)}Z_b$, $Q^{(i)} = Q + q^{(i)}Z_q$ and $R^{(i)} = R + r^{(i)}Z_r$.

Modified Boeing system (Hong et al., 2021): We consider a modification of the Boeing system described in (Hong et al., 2021) to obtain an unstable nominal LQR task described by

$$A = \begin{bmatrix} 1.22 & 0.03 & -0.02 & -0.32 \\ 0.01 & 0.47 & 4.70 & 0.00 \\ 0.02 & -0.06 & 0.40 & 0.00 \\ 0.01 & -0.04 & 0.72 & 1.55 \end{bmatrix}, B = \begin{bmatrix} 0.01 & 0.99 \\ -3.44 & 1.66 \\ -0.83 & 0.44 \\ -0.47 & 0.25 \end{bmatrix}, Q = I_4, R = I_2.$$

We implement Algorithm 3 with $\mathcal{X}_0 \stackrel{d}{=} \mathcal{N}(0, \frac{1}{4}I_{n_x})$, $\eta_l = \eta = 8 \times 10^{-6}$, $r = 1 \times 10^{-2}$, $m = 20$, and initial stabilizing controller given by

$$K_0 = \begin{bmatrix} 0.613 & -1.535 & 0.303 & 0.396 \\ 0.888 & 0.604 & -0.147 & -0.582 \end{bmatrix}.$$

Personalization: We assess the personalization of Algorithm 3 in Figure 1-(right) by generating \bar{M} tasks for the modified Boeing system and use 80% of the tasks to learn K_N and with the 20% remaining we construct a held-out set of tasks where we randomly sample unseen tasks to assess the generalization of the learned controller.