Policy Gradient Bounds in Multitask LQR

Charis Stamouli, *Graduate Student Member, IEEE*, Leonardo F. Toso, *Graduate Student Member, IEEE*, Anastasios Tsiamis, *Member, IEEE*, George J. Pappas, *Fellow, IEEE*, and James Anderson, *Senior Member, IEEE*

***Abstract*—We analyze the performance of policy gradient in multitask linear quadratic regulation (LQR), where the system and cost parameters differ across tasks. The main goal of multitask LQR is to find a controller with satisfactory performance on every task. Prior analyses on relevant contexts fail to capture closed-loop task similarities, resulting in conservative performance guarantees. To account for such similarities, we propose bisimulation-based measures of task heterogeneity. Our measures employ new bisimulation functions to bound the cost gradient distance between a pair of tasks in closed loop with a common stabilizing controller. Employing these measures, we derive suboptimality bounds for both the multitask optimal controller and the asymptotic policy gradient controller with respect to each of the tasks. We further provide conditions under which the policy gradient iterates remain stabilizing for every system. For multiple random sets of certain tasks, we observe that our bisimulation-based measures improve upon baseline measures of task heterogeneity dramatically.**

***Index Terms*—Multitask control, policy gradient, bisimulation.**

I. Introduction

DESIGNING a control policy that performs effectively on tasks with heterogeneous dynamics and objectives is a central problem in multitask reinforcement learning. For a collection of N tasks, a typical approach is to find a controller or policy parameter $K_\star$ that minimizes the average cost:

$$J_{\text{avg}}(K) := \frac{1}{N}\sum_{i=1}^{N} J^{(i)}(K) \tag{1}$$

over a parameter class containing $K_\star$. A popular method for approximating $J_{\text{avg}}(K_\star)$ is provided by policy gradient [1], with applications ranging from autonomous driving [2] to robotic control [3]. Despite the empirical success of policy gradient, its theoretical guarantees remain relatively unexplored, especially for control systems where maintaining the closed-loop stability of all systems is crucial.

In this letter, we analyze the performance of vanilla policy gradient [4] in multitask linear quadratic regulation (LQR) with respect to each task. Our setting involves LQR tasks with heterogeneous system and cost parameters. Our analysis relies on closed-loop measures of task heterogeneity, inspired by classical bisimulation functions [5]. Bisimulation functions, as introduced in [5], are Lyapunov-like functions that provide a principled method for characterizing the output distance of stable systems with vector states. Our main contributions are the following:

- We define a novel notion of bisimulation functions tailored to Lyapunov matrix systems, and provide a systematic design of them via linear matrix inequalities.
- Employing these functions, we introduce bisimulation-based measures that bound the cost gradient discrepancy between two LQR tasks under a common stabilizing controller. Our result establishes the first closed-loop measure of task heterogeneity for our setting.
- Employing these measures, we provide suboptimality bounds for the multitask optimal controller $K_\star$ and the asymptotic policy gradient controller with respect to each task. These bounds depend on the average bisimulation-based measure between the task of interest and the others, evaluated at the respective controller.
- We identify conditions such that the policy gradient iterates remain stabilizing for all systems.

We apply multitask policy gradient LQR across two sets of tasks: one with inverted pendula and another with unicycles. We observe that our bisimulation-based measures can be informative of the multitask policy gradient controller's performance, in cases where previous measures from [6], [7] are vacuous. For multiple random sets of these tasks, our measures dramatically improve upon baselines, effectively mitigating their conservatism. Complete proofs of our results are given in Appendix A of the arXiv version [8]. The related work is summarized below.

Policy Gradient Methods. Recent work has studied variants of our setting with different policy gradient methods. The

Received 12 September 2025; accepted 9 October 2025. Date of publication 27 October 2025; date of current version 6 November 2025. The work of Charis Stamouli and George J. Pappas was supported by NSF under Award SLES-2331880. The work of Leonardo F. Toso was supported by the CAIRFI and Presidential Fellowships. The work of James Anderson was supported in part by NSF under Grant ECCS 2144634 and Grant 2231350, and in part by the Columbia DSI. Recommended by Senior Editor A. P. Aguiar. *(Charis Stamouli and Leonardo F. Toso contributed equally to this work.) (Corresponding author: Charis Stamouli.)*

Charis Stamouli and George J. Pappas are with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: stamouli@seas.upenn.edu; pappasg@seas.upenn.edu).

Leonardo F. Toso and James Anderson are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: lt2879@columbia.edu; james.anderson@columbia.edu).

Anastasios Tsiamis is with the Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland (e-mail: atsiamis@control.ee.ethz.ch).

Digital Object Identifier 10.1109/LCSYS.2025.3625957

2475-1456 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:06:03 UTC from IEEE Xplore. Restrictions apply.

authors of [9], [10] consider the special case where the cost parameters are the same for all tasks. A related line of work focuses on meta-learning LQR [6], which includes additional fine-tuning to each task-specific cost and presents meta-LQR design with heterogeneous systems and objectives. Closest to our setting is the work in [7], which proposes an asynchronous policy gradient approach for multitask LQR under diverse system and cost parameters.

The analyses in [6] and [7] derive suboptimality bounds for policy gradient methods based on a heterogeneity measure that bounds the cost-gradient difference via the maximum norm distance between task parameters. This open-loop formulation overlooks potential closed-loop task similarities, often leading to conservative bounds. In contrast, our bisimulation-based measures provide a principled *closed-loop* notion of task heterogeneity, yielding more informative performance guarantees.

Bisimulations for Reinforcement Learning and Control. Behavioral similarity has been central in both robust control (e.g., gap metric [11], [12]) and layered control design (e.g., bisimulation relations [13], [14] and functions [5], [15]). Approximate bisimulation, in particular, employs Lyapunov-like functions to bound the output distance of two stable systems. Unlike our novel notion of bisimulations, classical bisimulations focus only on behavioral similarity with respect to system dynamics and remain agnostic to task objectives. The same limitation applies to the gap metric.

In reinforcement learning, approximate bisimulation has been used to capture state similarity in Markov decision processes with respect to immediate rewards and the distribution of next states under a given policy [16], [17], [18], and has been leveraged for policy transfer across similar states. These measures do not quantify deviations in policy gradient descent directions across tasks, where similarity must jointly capture both dynamics and objectives in the cost gradients. Our work addresses this gap by introducing a bisimulation-based measure that bounds the cost gradient discrepancy across LQR tasks under a common stabilizing controller.

Notation. The norm $\|\cdot\|$ is the Euclidean norm when applied to vectors and the spectral norm when applied to matrices, while $\|\cdot\|_F$ denotes the Frobenius norm. For a square matrix A, let $\mathrm{Tr}(A)$, $\lambda_{\min}(A)$, and $\sigma_{\min}(A)$ denote the trace, minimum eigenvalue, and minimum singular value, respectively. We use the notation $\mathrm{diag}(A, B)$ to denote the block-diagonal matrix with A and B on its diagonal, and $\mathbb{I}_d$ to represent the $d \times d$ identity matrix. The notation $\mathbb{S}_+^d$ refers to the set of positive definite matrices. Expectation with respect to all randomness of the underlying probability space is denoted by $\mathbf{E}$.

II. Problem Formulation

Consider a collection of N discrete-time systems:

$$x_{t+1}^{(i)} = A^{(i)} x_t^{(i)} + B^{(i)} u_t^{(i)},\ i = 1, \ldots, N,$$

where $x_t^{(i)} \in \mathbb{R}^{d_x}$ and $u_t^{(i)} \in \mathbb{R}^{d_u}$ denote the state and control input of the i-th system, respectively. The initial states $x_0^{(i)}$ are zero-mean random variables with covariance $\Sigma_0^{(i)} \succ 0$. The performance of system i under the feedback policy $u_t^{(i)} = -Kx_t^{(i)}$, with $K \in \mathbb{R}^{d_u \times d_x}$, is measured by the quadratic cost:

$$J^{(i)}(K) := \mathbf{E}\left[\sum_{t=0}^{\infty} x_t^{(i)\mathsf{T}} \left(Q^{(i)} + K^{\mathsf{T}} R^{(i)} K\right) x_t^{(i)}\right],$$

where $Q^{(i)} \succeq 0$ and $R^{(i)} \succ 0$. Consider the set of common stabilizing controllers $\mathcal{K}_{\mathrm{stab}} := \cap_{i=1}^N \mathcal{K}_{\mathrm{stab}}^{(i)}$, where $\mathcal{K}_{\mathrm{stab}}^{(i)}$ is the set of stabilizing controllers for system i. For each i, we define the i-th LQR task as $\mathcal{T}^{(i)} = (A^{(i)}, B^{(i)}, Q^{(i)}, R^{(i)})$ and its optimal controller as $K_\star^{(i)} = \arg\min_{K \in \mathcal{K}_{\mathrm{stab}}^{(i)}} J^{(i)}(K)$. The objective of multitask LQR is to find $K_\star \in \mathcal{K}_{\mathrm{stab}}$ that minimizes the average cost over all tasks, given by (1).

In multitask reinforcement learning, a standard method of approximating the minimum value $J_{\mathrm{avg}}(K_\star)$ is that of policy gradient [1]. In our setting, we apply policy gradient with an initial controller $K_0 \in \mathcal{K}_{\mathrm{stab}}$[1] and iterative updates:

$$K_{n+1} = K_n - \alpha \nabla J_{\mathrm{avg}}(K_n), \tag{2}$$

where $\alpha \in \mathbb{R}_+$ is a fixed step size. For each i and any controller $K \in \mathcal{K}_{\mathrm{stab}}$, we define the dynamics $x_{K,t+1}^{(i)} = A_K^{(i)} x_{K,t}^{(i)}$, where $A_K^{(i)} = A^{(i)} - B^{(i)}K$, and $P_K^{(i)}$ as the unique solution of the Lyapunov equation:

$$P_K^{(i)} - A_K^{(i)\mathsf{T}} P_K^{(i)} A_K^{(i)} = Q^{(i)} + K^{\mathsf{T}} R^{(i)} K.$$

From [4, Lemma 1] it follows that for any $K \in \mathcal{K}_{\mathrm{stab}}$, the gradient of the average cost can be expressed as:

$$\nabla J_{\mathrm{avg}}(K) = \frac{1}{N}\sum_{i=1}^N \nabla J^{(i)}(K) = \frac{1}{N}\sum_{i=1}^N E_K^{(i)} \Sigma_K^{(i)}, \tag{3}$$

where for each i:

$$E_K^{(i)} = 2\left((R^{(i)} + B^{(i)\mathsf{T}} P_K^{(i)} B^{(i)})K - B^{(i)\mathsf{T}} P_K^{(i)} A^{(i)}\right), \tag{4}$$

$$\Sigma_K^{(i)} = \mathbf{E}\left[\sum_{t=0}^{\infty} x_{K,t}^{(i)} x_{K,t}^{(i)\mathsf{T}}\right]. \tag{5}$$

This letter aims to analyze the task-specific performance of $K_\star$ and K_n with respect to that of each $K_\star^{(i)}$. More specifically, we are interested in providing upper bounds for:

1) The task-specific optimality gaps of $K_\star$:

$$J^{(i)}(K_\star) - J^{(i)}\left(K_\star^{(i)}\right),\ i = 1, \ldots, N. \tag{6}$$

2) The task-specific asymptotic optimality gaps of K_n:

$$\limsup_{n\to\infty}\left(J^{(i)}(K_n) - J^{(i)}(K_\star^{(i)})\right),\ i = 1, \ldots, N. \tag{7}$$

In similar settings, prior work [6], [7] has established the gradient gaps between pairs of tasks (i, j):

$$g_{ij}(K) := \|\nabla J^{(i)}(K) - \nabla J^{(j)}(K)\|_F \tag{8}$$

as fundamental for bounding the task-specific optimality gaps. Therefore, we say that $g_{ij}(K)$ quantifies the *heterogeneity* between tasks $\mathcal{T}^{(i)}$ and $\mathcal{T}^{(j)}$ under any $K \in \mathcal{K}_{\mathrm{stab}}$.

[1] A controller $K_0 \in \mathcal{K}_{\mathrm{stab}}$ can be efficiently computed [10].

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:06:03 UTC from IEEE Xplore. Restrictions apply.

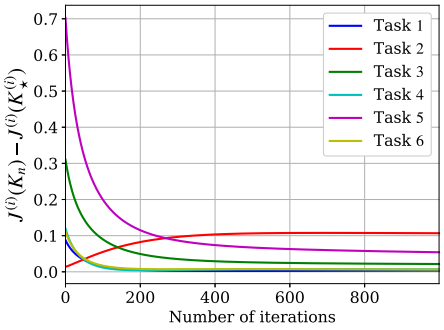


Fig. 1. Evolution of optimality gaps associated with the multitask policy gradient iterates for 6 inverted-pendulum tasks.

A. Limitations of Existing Task Heterogeneity Measures

To measure task heterogeneity, prior works [6], [7] begin by assuming bounded deviations in the system parameters:

$$\max_{i\neq j} \|A^{(i)} - A^{(j)}\| \leq b_A, \quad \max_{i\neq j} \|B^{(i)} - B^{(j)}\| \leq b_B.$$

Bounds b_Q and b_R for the cost parameter deviations are similarly defined. The above deviation bounds are combined into a task heterogeneity measure, $\bar{b}(K) := \bar{b}(K; b_A, b_B, b_Q, b_R)$, which for any $K \in \mathcal{K}_{\text{stab}}$ satisfies:

$$\max_{i\neq j} g_{ij}(K) \leq \bar{b}(K). \tag{9}$$

For the exact expression of $\bar{b}(K)$, we refer the reader to [7, Appendix 6.2]. The measures $\bar{b}(K_\star)$ and $\bar{b}(K_n)$ are used to bound the performance gaps $J^{(i)}(K_\star) - J^{(i)}(K_\star^{(i)})$ and $J^{(i)}(K_n) - J^{(i)}(K_\star^{(i)})$, respectively. However, they are often vacuous in practice, even when the actual performance gaps are small, as illustrated in the following example.

Example 1 (Inverted Pendulum): Consider a collection of 6 LQR tasks corresponding to the linearized inverted pendulum dynamics with system matrices:

$$A^{(i)} = \begin{bmatrix} 1 & dt \\ \frac{g}{\ell_i} dt & 1 \end{bmatrix}, \; B^{(i)} = \begin{bmatrix} 0 \\ \frac{dt}{m_i \ell_i^2} \end{bmatrix},$$

where ℓ_i is the length, m_i is the mass, $g = 10$, and $dt = 0.05$. The LQR cost matrices are $Q^{(i)} = q_i \mathbb{I}_2$ and $R^{(i)} = r_i$, and the initial state covariances are $\Sigma_0^{(i)} = 0.01\mathbb{I}_2$. The parameters ℓ_i, m_i, q_i, and r_i are uniformly sampled from $[0.5, 1]$, $[0.1, 0.5]$, $[0.1, 0.5]$, and $[0.01, 0.05]$, respectively. As shown in Fig. 1, the optimality gaps (7) are below 0.15 at convergence, for a step size of 0.01, indicating satisfactory performance of the multitask policy gradient controller. This is in contrast to the heterogeneity measure $\bar{b}(K_n)$, which evaluates to 2.3×10^6 at convergence. Its conservatism can be explained by the derivation of $\bar{b}(K)$ in terms of parameter deviation bounds. Specifically, despite the dependence of $\bar{b}(K)$ on K, its construction overlooks potential closed-loop task similarities that could lead to smaller task heterogeneity.

B. Toward Closed-Loop Task Heterogeneity Measures

Our goal is to develop closed-loop measures of task heterogeneity, suitable for bounding the performance of both the multitask optimal controller and the policy gradient iterates. Exploiting the closed-loop behavior is intended to mitigate the conservatism incurred by the previous use of parameter deviation bounds to analyze the gaps (6) and (7).

Toward this end, we next present a novel notion of bisimulation-based measures of task heterogeneity, which bound the gradient gaps in (8). In Section IV, we leverage these measures to bound the optimality gaps (6) and (7).

III. Bisimulation-Based Task Heterogeneity

We now propose bisimulation-based measures that allow us to bound the discrepancy in cost gradients between tasks, as quantified by (8). Unlike existing measures of task heterogeneity in [6], [7], our measures provide a *closed-loop* notion of heterogeneity between different tasks. In Section III-A, we define bisimulation functions that can be used to bound the gradient discrepancies (8). In Section III-B, we develop an effective derivation of these functions and formalize our bisimulation-based measures.

A. Bisimulation Functions for Cost Gradient Discrepancies

For every task $\mathcal{T}^{(i)}$ and any $K \in \mathcal{K}_{\text{stab}}$, we consider the associated matrix system:

$$S_K^{(i)} : \begin{cases} \Sigma_{K,t+1}^{(i)} = A_K^{(i)} \Sigma_{K,t}^{(i)} A_K^{(i)\intercal} + \Sigma_0^{(i)} \\ Y_{K,t}^{(i)} = E_K^{(i)} \Sigma_{K,t}^{(i)} \end{cases}, \tag{10}$$

where $\Sigma_{K,t}^{(i)} \in \mathbb{S}_+^{d_x}$ and $E_K^{(i)}$ is defined by (4). The dynamics in (10) represent those of system i's covariance under controller K. By using [4, Lemma 1] and the fact that $K \in \mathcal{K}_{\text{stab}}$, we can deduce that the output $Y_{K,t}^{(i)}$ converges to the cost gradient $\nabla J^{(i)}(K)$. Therefore, we have:

$$\|\nabla J^{(i)}(K) - \nabla J^{(j)}(K)\|_F = \lim_{t\to\infty} \|Y_{K,t}^{(i)} - Y_{K,t}^{(j)}\|_F \tag{11}$$

and can focus on bounding the asymptotic output distance of systems $S_K^{(i)}$ and $S_K^{(j)}$.

The concept of bisimulation functions, as introduced in [5], provides a principled method for characterizing the output distance between two stable continuous-time systems with vector states. Next, we extend the definition of bisimulation functions to affine matrix systems of the form (10).

Definition 1 (Bisimulation Functions for Lyapunov Matrix Systems): A continuous function $V : \mathbb{S}_+^{d_x} \times \mathbb{S}_+^{d_x} \to \mathbb{R}_+$ is a bisimulation function between $S_K^{(i)}$ and $S_K^{(j)}$ if there exists $\lambda \in (0, 1)$ such that for all $(\Sigma_{K,t}^{(i)}, \Sigma_{K,t}^{(j)}) \in \mathbb{S}_+^{d_x} \times \mathbb{S}_+^{d_x}$:

$$V\left(\Sigma_{K,t}^{(i)}, \Sigma_{K,t}^{(j)}\right) \geq \|E_K^{(i)} \Sigma_{K,t}^{(i)} - E_K^{(j)} \Sigma_{K,t}^{(j)}\|_F, \quad \text{and} \tag{12a}$$

$$V\left(\Sigma_{K,t+1}^{(i)}, \Sigma_{K,t+1}^{(j)}\right) - V\left(\Sigma_{K,t}^{(i)}, \Sigma_{K,t}^{(j)}\right) \leq -\lambda V\left(\Sigma_{K,t}^{(i)}, \Sigma_{K,t}^{(j)}\right) + V\left(\Sigma_0^{(i)}, \Sigma_0^{(j)}\right). \tag{12b}$$

Condition (12a) implies that a bisimulation function $V(\cdot, \cdot)$ provides an upper bound for the output distance of systems $S_K^{(i)}$ and $S_K^{(j)}$, given any pair of states. Condition (12b) can be easily shown to guarantee that $V(\Sigma_{K,t}^{(i)}, \Sigma_{K,t}^{(j)})$ is ultimately bounded within a ball of radius $V(\Sigma_0^{(i)}, \Sigma_0^{(j)})/\lambda$.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:06:03 UTC from IEEE Xplore. Restrictions apply.

Lemma 1: Let $V : \mathbb{S}_+^{d_x} \times \mathbb{S}_+^{d_x} \to \mathbb{R}_+$ denote a bisimulation function between $S_K^{(i)}$ and $S_K^{(j)}$. It holds that:

$$\lim_{t\to\infty} \|Y_{K,t}^{(i)} - Y_{K,t}^{(j)}\|_F \leq V\Big(\Sigma_0^{(i)}, \Sigma_0^{(j)}\Big)/\lambda. \tag{13}$$

The proof follows from condition (12a) and a recursive expansion of (12b). Combining (11) and (13), we obtain:

$$\|\nabla J^{(i)}(K) - \nabla J^{(j)}(K)\|_F \leq V\Big(\Sigma_0^{(i)}, \Sigma_0^{(j)}\Big)/\lambda. \tag{14}$$

Next, we show that for any $K \in \mathcal{K}_{\text{stab}}$, a bisimulation function $V(\cdot,\cdot)$ can be effectively designed for $S_K^{(i)}$ and $S_K^{(j)}$.

B. Task Heterogeneity via Bisimulation Function Design

In this subsection, we develop a systematic procedure for designing bisimulation functions between systems $S_K^{(i)}$ and $S_K^{(j)}$. Then, we leverage these functions to introduce bisimulation-based measures of task heterogeneity.

Lemma 2: Consider a controller $K \in \mathcal{K}_{\text{stab}}$ and a pair of task indices (i, j). There exist $M \in \mathbb{S}_+^{2d_x}$ and $\lambda \in (0, 1)$ such that the following linear matrix inequalities (LMIs) hold:

$$M \succeq E_K^{(ij)\intercal} E_K^{(ij)}, \tag{15a}$$

$$A_K^{(ij)\intercal} M A_K^{(ij)} - M \preceq -\lambda M, \tag{15b}$$

where $E_K^{(ij)} = [E_K^{(i)} \ -E_K^{(j)}]$ and $A_K^{(ij)} = \text{diag}(A_K^{(i)}, A_K^{(j)})$.

The proof is similar to that of [19, Lemma 2]. We next use the linear matrix inequalities above to systematically characterize bisimulation functions.

Theorem 1 (Bisimulation Functions for Lyapunov Matrix Systems): Consider a controller $K \in \mathcal{K}_{\text{stab}}$ and a pair of task indices (i, j). Let $M \in \mathbb{S}_+^{2d_x}$ and $\lambda \in (0, 1)$ satisfy (15a) and (15b). The function $V{:}\mathbb{S}_+^{d_x} \times \mathbb{S}_+^{d_x} \to \mathbb{R}_+$ defined by:

$$V\Big(\Sigma_{K,t}^{(i)}, \Sigma_{K,t}^{(j)}\Big) = \frac{\sqrt{2}\text{Tr}\Big(M\,\text{diag}\Big(\Sigma_{K,t}^{(i)}, \Sigma_{K,t}^{(j)}\Big)\Big)}{\sqrt{\lambda_{\min}(M)}}, \tag{16}$$

is a bisimulation function between systems $S_K^{(i)}$ and $S_K^{(j)}$.

Combining (14) and Theorem 1 we can deduce that:

$$\|\nabla J^{(i)}(K) - \nabla J^{(j)}(K)\|_F \leq \frac{\sqrt{2}\text{Tr}\Big(M\,\text{diag}\Big(\Sigma_0^{(i)}, \Sigma_0^{(j)}\Big)\Big)}{\lambda\sqrt{\lambda_{\min}(M)}}, \tag{17}$$

for any $M \in \mathbb{S}_+^{2d_x}$ and $\lambda \in (0, 1)$ satisfying (15a) and (15b).

Definition 2 (Bisimulation-Based Task Heterogeneity): Consider a controller $K \in \mathcal{K}_{\text{stab}}$ and a pair of distinct task indices (i, j). Let $M_K^{(ij)}$ and $\lambda_K^{(ij)}$ denote the minimizers of the bound in (17) that satisfy the conditions of Lemma 2. The bisimulation-based heterogeneity between tasks $\mathcal{T}^{(i)}$ and $\mathcal{T}^{(j)}$, under the policy $u_t^{(i)} = -Kx_t^{(i)}$, is given by:

$$b_{ij}(K) := \frac{\sqrt{2}\text{Tr}\Big(M_K^{(ij)}\,\text{diag}\Big(\Sigma_0^{(i)}, \Sigma_0^{(j)}\Big)\Big)}{\lambda_K^{(ij)}\sqrt{\lambda_{\min}\Big(M_K^{(ij)}\Big)}}. \tag{18}$$

The measure (18) provides a closed-loop notion of task heterogeneity, as the LMI (15b) considers systems i and j in closed loop with the controller K. The quantities $M_K^{(ij)}$ and $\lambda_K^{(ij)}$ in (18) are obtained by solving a conic program (see Appendix B of our arXiv version [8]), whose computational cost scales polynomially with the state dimension d_x. Standard conic program solvers are provided in CVXPY [20].

Remark 1: The conditions in Lemma 2 are the discrete-time analogues of those derived for continuous-time linear systems with vector states in [5]. The bisimulation-based task heterogeneity (18) is the minimum bisimulation bound for the difference in the cost gradient responses of the i-th and j-th task, under a common $K \in \mathcal{K}_{\text{stab}}$, as in (17). The value of $\lambda_K^{(ij)}$ is determined by the stability margin of $A_K^{(ij)}$; specifically, the larger the stability margin, the larger $\lambda_K^{(ij)}$ becomes (see (15b)). Moreover, $M_K^{(ij)}$ corresponds to the solution of a Lyapunov equation for $A_K^{(ij)}$, which dominates a matrix related to $E_K^{(i)}$ and $E_K^{(j)}$ (see (15a) and (15b)). These matrices are zero when K is equal to $K_\star^{(i)}$ and $K_\star^{(j)}$, respectively. Note also the dependence of $b_{ij}(K)$ on the initial state covariances $\Sigma_0^{(i)}$ and $\Sigma_0^{(j)}$, which directly affect the cost of the respective LQR tasks [21]. Subsequently, we empirically observe how the measures $b_{ij}(K)$ influence the optimality gaps (6) and (7), while reducing the conservatism of the previous task heterogeneity bounds in (9). A thorough theoretical comparison of our measures with the previous measures is reserved for future work.

IV. Policy Gradient Bounds in Multitask LQR via Bisimulations

We now analyze the task-specific optimality gaps in (6) and (7) based on the bisimulation-based task heterogeneity measures (18). Before introducing our main results, we present a few necessary definitions. For every task $\mathcal{T}^{(i)}$ and any $K \in \mathcal{K}_{\text{stab}}$, we define $b_i(K)$ as:

$$b_i(K) = \frac{1}{N}\sum_{j\neq i} b_{ij}(K), \tag{19}$$

where $b_{ij}(K)$ are given by (18). We note that $b_i(K)$ describes the average bisimulation-based heterogeneity between the task $\mathcal{T}^{(i)}$ and each of the other tasks. We proceed with defining the stabilizing subset $\mathcal{K} \subset \mathbb{R}^{m\times n}$.

Definition 3 (Stabilizing Subset [6]): Given a controller $K_0 \in \mathcal{K}_{\text{stab}}$ and constants $\beta_1, \ldots, \beta_N \geq 1$, the stabilizing subset is defined as $\mathcal{K} := \cap_{i=1}^N \mathcal{K}^{(i)}$, where:

$$\mathcal{K}^{(i)} = \Big\{K \in \mathbb{R}^{d_u\times d_x} \ : \ J^{(i)}(K) - J^{(i)}\Big(K_\star^{(i)}\Big) \leq \beta_i\Big(J^{(i)}(K_0) - J^{(i)}(K_\star^{(i)})\Big)\Big\}.$$

A few observations about the set $\mathcal{K}$ are in order. First, recall that $\mathcal{K}_{\text{stab}}$ is the set of common stabilizing controllers for all systems and note that $\mathcal{K} \subseteq \mathcal{K}_{\text{stab}}$, since the cost of each task is finite for any controller in $\mathcal{K}$. Second, observe the implicit dependence of $\mathcal{K}$ on the given initial policy gradient controller K_0. Third, note that we can always pick constants β_i large enough such that $K_\star \in \mathcal{K}$. Therefore, the use of $\mathcal{K}$ instead of $\mathcal{K}_{\text{stab}}$ does not limit our results, but rather simplifies our analysis, due to the compactness of $\mathcal{K}$ shown in [22]. Specifically, from [22, Th. 1] it follows that the functions $J^{(i)}(\cdot)$ are smooth on the compact set $\mathcal{K}$. The use of $\mathcal{K}$ is also standard in prior work [6], [7].

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:06:03 UTC from IEEE Xplore. Restrictions apply.

Theorem 2 (Bisimulation-Based Suboptimality Bounds for Multitask Optimal LQR): Let $K_\star \in \mathcal{K}_{\text{stab}}$ be a minimizer of (1). For every task $\mathcal{T}^{(i)}$ and any $K \in \mathcal{K}_{\text{stab}}$, let $b_i(K)$ be given by (19). Then, for each task index i, the task-specific optimality gap associated with $K_\star$ is bounded as:

$$J^{(i)}(K_\star) - J^{(i)}\left(K_\star^{(i)}\right) \leq \frac{2\left\|\Sigma^{(i)}_{K_\star^{(i)}}\right\| b_i(K_\star)^2}{\lambda_{\min}\left(\Sigma_0^{(i)}\right)^2 \sigma_{\min}\left(R^{(i)}\right)}. \quad (20)$$

The proof results from adapting that of [6, Th. 3] to the setting of multitask LQR. We point out the dependence of the bound in (20) on the squared bisimulation-based heterogeneity measure (19), evaluated at $K_\star$.

Theorem 3 (Bisimulation-Based Suboptimality Asymptotics for Multitask Policy Gradient LQR): Let $K_0 \in \mathcal{K}_{\text{stab}}$ and consider the sequence $\{K_n\}_{n=0}^{\infty}$ of policy gradient iterates given by (2). Consider $\beta_1, \ldots, \beta_N \geq 1$, and let $\mathcal{K}$ denote the stabilizing subset from Definition 3. For each i, let L_i denote the smoothness constant of the i-th LQR cost function $J^{(i)}(\cdot)$ over the set $\mathcal{K}$. For every task $\mathcal{T}^{(i)}$ and any $K \in \mathcal{K}_{\text{stab}}$, let $b_i(K)$ be given by (19). Assume that:

$$\alpha < \min\left\{1/\left(4\max_i L_i\right), 4/\max_i \gamma_i\right\},$$

where $L_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} L_i$, $\gamma_i = \frac{4\lambda_{\min}(\Sigma_0^{(i)})^2 \sigma_{\min}(R^{(i)})}{\|\Sigma^{(i)}_{K_\star^{(i)}}\|}$, for all $i = 1, \ldots, N$. Moreover, assume that $\sup_{K\in\mathcal{K}} b_i(K)^2 \leq \gamma_i(J^{(i)}(K_0) - J^{(i)}(K_\star^{(i)}))/6$. Then, we have $K_n \in \mathcal{K}$, for all $n \in \mathbb{N}$, and for each task index i, the task-specific asymptotic optimality gap associated with $\{K_n\}_{n=0}^{\infty}$ is bounded as:

$$\limsup_{n\to\infty} \left(J^{(i)}(K_n) - J^{(i)}(K_\star^{(i)})\right) \leq \limsup_{n\to\infty} \frac{3\left\|\Sigma^{(i)}_{K_\star^{(i)}}\right\| b_i(K_n)^2}{4\lambda_{\min}\left(\Sigma_0^{(i)}\right)^2 \sigma_{\min}\left(R^{(i)}\right)}. \quad (21)$$

The proof results from combining a modified version of the proof of [6, Th. 3] with a novel extension of [22, Th. 1] to the setting of multitask LQR. The conditions of Theorem 3 on a small enough step size and a low bisimulation-based heterogeneity compared to the initial optimality gap $J^{(i)}(K_0) - J^{(i)}(K_\star^{(i)})$ are reasonable and standard in prior work [6], [7]. We note the dependence of the bound in (21) on the squared bisimulation-based heterogeneity measure (19), evaluated at the limit superior of policy gradient iterations. If the policy gradient iterates converge to some controller K_∞, then for all tasks $i = 1, \ldots, N$, inequality (21) implies that:

$$J^{(i)}(K_\infty) - J^{(i)}\left(K_\star^{(i)}\right) \leq \frac{3\left\|\Sigma^{(i)}_{K_\star^{(i)}}\right\| b_i(K_\infty)^2}{4\lambda_{\min}\left(\Sigma_0^{(i)}\right)^2 \sigma_{\min}\left(R^{(i)}\right)}, \quad (22)$$

Remark 2 (Result Interpretation): The suboptimality bounds in (20) and (22) are equal up to universal constants and the bisimulation-based measures $b_i(K)$, which are evaluated at $K_\star$ and K_∞, respectively. Their similarity is notable as the analysis of Theorem 2 is algorithm-independent and corresponds to the multitask optimal controller, while the analysis of Theorem 3 relies on the policy gradient algorithm defined by (2). Beyond the measures $b_i(K)$, the optimality gaps are bounded in terms of: i) the steady-state covariance of system i under the task-specific optimal controller $K_\star^{(i)}$, ii) the initial state covariance $\Sigma_0^{(i)}$, and iii) the input cost matrix $R^{(i)}$.

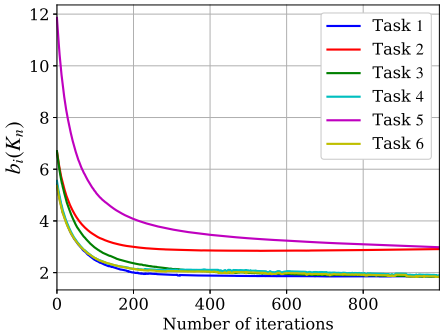


Fig. 2. Evolution of bisimulation-based measures from (19) in multitask policy gradient LQR for 6 inverted-pendulum tasks.

In the next section, we observe that our closed-loop measures (19) significantly reduce the conservatism of previous heterogeneity measures and better capture the performance of multitask policy gradient LQR.

V. Numerical Validation

We next present experiments[2] on the inverted pendulum and unicycle dynamics to demonstrate the effectiveness of our proposed closed-loop task heterogeneity measures. In particular, we observe that our bisimulation-based measures are less conservative than those in [6], [7], which bound the cost gradient discrepancies (8) in terms of the task parameter deviation bounds b_A, b_B, b_Q, and b_R.

Inverted Pendulum. We revisit Example 1 to demonstrate the effectiveness of our bisimulation-based measures (19) compared to the previous task heterogeneity measures $\bar{b}(K) := \bar{b}(K; b_A, b_B, b_Q, b_R)$ employed in (9).

Suboptimality bounds similar to those in (20) and (22) can be derived in terms of $\bar{b}(K_\star)$ and $\bar{b}(K_\infty)$, respectively (see Section II-A). Recall that $\bar{b}(K_\infty)$ evaluates to 2.3×10^6 in this example, despite the small optimality gaps observed in Fig. 2. In contrast, our bisimulation-based measures depicted in Fig. 2 are below 3 at convergence, reflecting the favorable task-specific performance of the multitask controller.

We repeat the computation of both measures over 100 collections of inverted-pendulum tasks, which are generated as in Example 1. In this case, we observe an average reduction of 99.9998%, which suggests a substantial decrease of conservatism with respect to the previous bounds.

Unicycle. We consider a collection of 6 LQR tasks corresponding to the linearized unicycle dynamics. The continuous-time kinematics are given by:

$$\dot{p}_x = v\cos\theta, \dot{p}_y = v\sin\theta, \text{ and } \dot{\theta} = \omega,$$

[2]Code for reproduction can be found at our GitHub repository.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:06:03 UTC from IEEE Xplore. Restrictions apply.

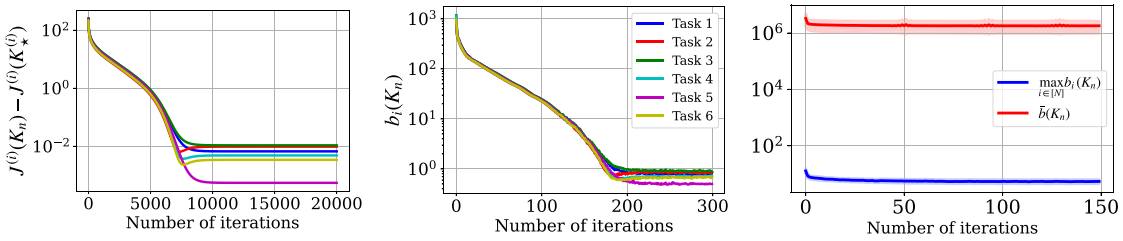


Fig. 3. Evolution of optimality gaps (left), bisimulation-based measures from (19) (middle), and cost gradient discrepancy bounds (right) with respect to policy gradient iterations for multitask LQR with unicycle dynamics. On the right, we compare our bounds with those proposed in [6], [7].

with state $x = (p_x, p_y, \theta)$ and input $u = (v, \omega)$, where (p_x, p_y) is the robot's position on the x, y-plane, θ is the orientation angle, v is the forward velocity, and ω is the yaw rate. We linearize the dynamics at operating points $(v_{0,i}, \theta_{0,i})$ and then discretize with step size $dt = 0.05$ to obtain the matrices $A^{(i)}$ and $B^{(i)}$ (see our arXiv version for details [8]). We uniformly draw $v_{0,i}$, $\theta_{0,i}$, q_i, and r_i from the intervals $[0.1, 1.75]$, $[0, \pi/2]$, $[0.1, 0.5]$, $[0.01, 0.05]$, respectively. We set $Q^{(i)} = q_i \mathbb{I}_3$, $R^{(i)} = r_i \mathbb{I}_2$, and initial state covariances $\Sigma_0^{(i)} = \mathbb{I}_3$, for all tasks.

Fig. 3 depicts the results for the multitask unicycle setting. On the left, we observe that the policy gradient controller achieves performance close to the task-specific optima, illustrating a setting favorable for collaborative control design. In the middle, we note that our bisimulation-based measures are below 1 at convergence, suggesting that the tasks are similar in closed loop with the designed common controller. This indicates the effectiveness of our task heterogeneity measure on capturing collaborative settings where the designed controller can be readily applied across distinct tasks. Further evaluation of our measures is reserved for future work.

On the right plot, we compare $\max_i b_i(K_n)$ with the measure $\bar{b}(K_n)$ from prior works [6], [7], averaged over 75 random collections of two unicycle tasks. We note that our measure is significantly less conservative, with a reduction of 99.9996% at convergence, leading to non-vacuous task-specific optimality bounds. In contrast, the previous heterogeneity measure is overly pessimistic and fails to capture the observed similarities.

VI. Future Work

Moving forward, our bisimulation-based measures could be applied to relax the task heterogeneity assumptions and enhance the performance bounds in related settings (e.g., LQR with domain randomization [10], meta-learning LQR [6]). Moreover, we aim to leverage our new task heterogeneity measures to guide the learning of latent shared dynamics among multiple systems, providing a more favorable setting for applying multitask LQR. Another interesting direction is to employ our bisimulation-based measures in the design of policy gradient updates and provide robustness guarantees for multitask LQR design. Our results can also be extended for the analysis of the model-free setting.

Acknowledgment

The authors thank Bruce D. Lee for insightful discussions on this letter.

References

[1] Y. Teh et al., "Distral: Robust multitask reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–13.
[2] Q. Qi et al., "Scalable parallel task scheduling for autonomous driving using multi-task deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13861–13874, Nov. 2020.
[3] D. Kalashnikov et al., "MT-Opt: Continuous multi-task robotic reinforcement learning at scale," 2021, *arXiv:2104.08212*.
[4] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1467–1476.
[5] A. Girard and G. Pappas, "Approximate bisimulations for constrained linear systems," in *Proc. 44th IEEE Conf. Decis. Control*, 2005, pp. 4700–4705.
[6] L. F. Toso, D. Zhan, J. Anderson, and H. Wang, "Meta-learning linear quadratic regulators: A policy gradient MAML approach for model-free LQR," in *Proc. 6th Annu. Learn. Dyn. Control Conf.*, 2024, pp. 902–915.
[7] L. F. Toso, H. Wang, and J. Anderson, "Asynchronous heterogeneous linear quadratic regulator design," in *Proc. IEEE 63rd Conf. Decis. Control (CDC)*, 2024, pp. 801–808.
[8] C. Stamouli, L. F. Toso, A. Tsiamis, G. J. Pappas, and J. Anderson, "Policy gradient bounds in multitask LQR," 2025, *arXiv:2509.19266*.
[9] H. Wang, L. F. Toso, A. Mitra, and J. Anderson, "Model-free learning with heterogeneous dynamical systems: A federated LQR approach," 2023, *arXiv:2308.11743*.
[10] T. Fujinami, B. D. Lee, N. Matni, and G. J. Pappas, "Policy gradient for LQR with domain randomization," 2025, *arXiv:2503.24371*.
[11] T. T. Georgiou and M. C. Smith, "Optimal robustness in the gap metric," in *Proc. 28th IEEE Conf. Decis. Control*, 1989, pp. 2331–2336.
[12] K. Zhou and J. Doyle, "Essentials of robust control," 1998. [Online]. Available: https://sharif.edu/ namvar/index_files/robust.pdf
[13] E. Haghverdi, P. Tabuada, and G. Pappas, "Bisimulation relations for dynamical and control systems," *Electron. Notes Theor. Comput. Sci.*, vol. 69, pp. 120–136, Feb. 2003.
[14] A. Van der Schaft, "Equivalence of dynamical systems by bisimulation," *IEEE Trans. Autom. Control*, vol. 49, no. 12, pp. 2160–2172, Dec. 2004.
[15] A. Girard and G. J. Pappas, "Approximate bisimulation: A bridge between computer science and control theory," *Eur. J. Control*, vol. 17, nos. 5–6, pp. 568–578, 2011.
[16] P. Castro and D. Precup, "Using bisimulation for policy transfer in MDPs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 24, no. 1, 2010, pp. 1065–1070.
[17] N. Ferns and D. Precup, "Bisimulation metrics are optimal value functions," in *Proc. 30th Conf. Uncertainty Artif. Intell.*, 2014, p. 10.
[18] M. Kemertas and A. Jepson, "Approximate policy iteration with bisimulation metrics," 2022, *arXiv:2202.02881*.
[19] C. Stamouli, A. Tsiamis, M. Morari, and G. J. Pappas, "Layered multi-rate control of constrained linear systems," 2025, *arXiv:2504.10461*.
[20] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, 2016.
[21] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Chelmsford, MA, USA: Courier Corp., 2007.
[22] B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel, and T. Başar, "Toward a theoretical foundation of policy optimization for learning control policies," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 6, no. 1, pp. 123–158, 2023.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:06:03 UTC from IEEE Xplore. Restrictions apply.