Policy Evaluation in Distributional LQR

Zifan Wang, Yulong Gao, Siyi Wang, Michael M. Zavlanos, Alessandro Abate, and Karl H. Johansson

***Abstract*—Distributional reinforcement learning (DRL) enhances the understanding of the effects of the randomness in the environment by letting agents learn the distribution of a random return, rather than its expected value as in standard reinforcement learning. Meanwhile, a challenge in DRL is that the policy evaluation typically relies on the representation of the return distribution, which needs to be carefully designed. In this paper, we address this challenge for the special class of DRL problems that rely on a discounted linear quadratic regulator (LQR), which we call *distributional LQR*. Specifically, we provide a closed-form expression for the distribution of the random return, which is applicable for all types of exogenous disturbance as long as it is independent and identically distributed (i.i.d.). We show that the variance of the random return is bounded if the fourth moment of the exogenous disturbance is bounded. Furthermore, we investigate the sensitivity of the return distribution to model perturbations. While the proposed exact return distribution consists of infinitely many random variables, we show that this distribution can be well approximated by a finite number of random variables. The associated approximation error can be analytically bounded under mild assumptions. When the model is unknown, we propose a model-free approach for estimating the return distribution, supported by sample complexity guarantees. Finally, we extend our approach to partially observable linear systems. Numerical experiments are provided to illustrate the theoretical results.**

***Index Terms*—Distributional LQR, distributional RL, distribution sensitivity, policy evaluation, partially observable system**

I. Introduction

In reinforcement learning (RL), the value of implementing a policy at a given state is captured by a value function, which models the expected sum of returns following this prescribed policy. Recently, [1] proposed the notion of distributional reinforcement learning (DRL), which learns the return distribution of a policy from a given state, instead of only its expected return. Compared to the scalar expected value function, the return distribution is infinite-dimensional and contains far more information. It is, therefore, not surprising that a few DRL algorithms, including C51 [1], D4PG [2], QR-DQN [3] and SDPG [4], dramatically improve the empirical performance in practical applications over their non-distributional counterpart. By encompassing the entire distribution, DRL is able to provide a comprehensive framework, for instance, for risk-averse learning, facilitating a deeper understanding and more effective management of uncertainties [5]–[8].

In parallel with the celebrated Bellman equation in the traditional RL, an alternative random variable (or distributional) Bellman equation acts as the theoretical foundation of DRL. It has been shown in [1] that the return distribution satisfies the distributional Bellman equation and the distributional Bellman operator is a contraction in (the maximum form of) the Wasserstein metric between probability distributions. A natural yet fundamental question in DRL is:

Given a policy, how to (exactly) characterise the random return that fulfills the random variable Bellman equation?

The answer to this question provides the structural information of the return distribution, which enables a better understanding of the value of implementing a policy in the DRL setting.

To the best of our knowledge, this problem has received limited attention. One of the challenges is the computational intractability arising from the fact that the return distribution is in an infinite-dimensional space. Approximations thus become necessary for practical implementation - cf. categorical [1], quantile function [3], and sample-based [4] methods. Furthermore, although some recent efforts have been devoted to applying DRL to partially observable systems [9], no theoretical foundations, including the characterisation of the random return, have been built for these partially observable models.

In this paper, we solve the above problem for discrete-time linear systems with stochastic additive disturbances. Specifically, we characterise the random cost for the classical discounted linear quadratic regulator (LQR) problem, which we term *distributional LQR*. We investigate the fundamental properties of the characterised random cost. Furthermore, we explore the extension to partially observable systems and derive fundamental properties of the characterised random cost.

A. Related Work

The problem under investigation falls within the domain of policy evaluation in DRL, specifically focusing on predicting the full probability distribution. This task poses a unique challenge because the full probability distribution is infinitely dimensional, necessitating the use of distribution parametrization techniques to render it computationally feasible. Bellemare

* This work was supported in part by Swedish Research Council Distinguished Professor Grant 2017-01078, Knut and Alice Wallenberg Foundation Wallenberg Scholar Grant, Swedish Strategic Research Foundation SUCCESS Grant FUS21-0026, AFOSR under award #FA9550-19-1-0169, and NSF under award CNS-1932011.

Zifan Wang and Karl H. Johansson are with Division of Decision and Control Systems, School of Electrical Enginnering and Computer Science, KTH Royal Institute of Technology, and also with Digital Futures, SE-10044 Stockholm, Sweden. Email: {zifanw,kallej}@kth.se.

Yulong Gao is with the Department of Electrical and Electronic Engineering, Imperial College London, United Kingdom. Email: yulong.gao@imperial.ac.uk

Siyi Wang is with the School of Computation, Information and Technology, Technical University of Munich, 80333 Munich, Germany. Email: siyi.wang@tum.de

Michael M. Zavlanos is with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA. Email: michael.zavlanos@duke.edu

Alessandro Abate is with Department of Computer Science, University of Oxford, OX1 3QD Oxford, U.K. Email: alessandro.abate@cs.ox.ac.uk

et al. [1] propose a categorical method that discretizes the return distribution by partitioning the return distribution into a finite number of uniformly spaced atoms in a fixed region. Subsequent work [10] delves into the convergence analysis of categorical policy evaluation and shows that the distributional projected Bellman operator with categorical representation is a contraction with respect to the Cramér distance metric. One drawback of the categorical representation is that it relies on prior knowledge of the range of the returned values. To address this limitation, [3] proposes a quantile temporal-difference learning algorithm that learns the quantiles of a probability distribution, and its convergence property is established in [11] using the Wasserstein-∞ metric. However, most of the existing algorithms and analysis of DRL are tailored to address problems with discrete state spaces, which cannot be applied to the linear quadratic control problem with continuous state space. It is worth mentioning that the works [4], [12] investigate DRL with continuous state space and use a reparameterization method to represent the distribution of random variables through a neural network. Despite these significant advancements, there is no theoretical guarantee regarding the quality of the learned distributions in [4], [12]. It still remains an open problem to derive an analytical expression for the return distribution with continuous state space. A challenge that further complicates the problem is represented by the infinitely many decision choices of states.

A related research line is the recent study of RL in the LQR context, which focuses on learning the expected return through interaction with the environment, see [13]–[19]. For example, [15] proposes a model-free policy gradient algorithm for LQR and shows its global convergence with finite polynomial computational and sample complexity. Moreover, [19] studies model-based RL for the linear quadratic Gaussian (LQG) problem, in which a model is first learnt from data and then used to design a policy. In this setup, evaluating the expected return for a policy is easily computed from the Riccati equation, but these methods are not capable of characterising other aspects of distributional information. Works exploring the distributional information include risk-averse control [20]–[27] or distributional robust control [28]–[31]. However, these methods cannot analyse the return distribution.

B. Contributions

This paper aims at studying the return distribution for linear quadratic control problems.

1) We provide an analytical expression of the random return for distributional LQR and prove that this return function is a fixed-point solution to the random variable Bellman equation (Theorem 1). Specifically, we show that the proposed analytical expression consists of infinitely many random variables and holds for arbitrary i.i.d. exogenous disturbances, e.g., non-Gaussian noise or noise with non-zero mean. This characterisation can recover the expected cost, complementing the classical LQR. We remark that the random return naturally contains more information than the expected cost and can thus be particularly useful for policy evaluation in a risk-averse setup [32].
2) We analyse the variance of the random return and show that it is bounded if the fourth moment of the disturbances is bounded (Theorems 2). Furthermore, we investigate the distributional sensitivity with respect to model perturbations. Under mild assumptions, we show that the maximal difference between the exact and perturbed return distributions can be bounded by the extent of model perturbations (Theorem 3).
3) We develop an approximation of the distribution of the random return using a finite number of random variables when the model is known. We show that the maximal difference between the exact and approximated return distributions decreases linearly with the number of random variables (Theorem 4). In the model-free case, we approximate the return distribution using state trajectories. We show that, with high confidence, the distribution approximation error deceases linearly with respect to the trajectory length and sub-linearly with respect to the number of trajectories (Theorem 5).
4) Finally, we derive analytical evidence that most results for distributional LQR have corresponding counterparts for partially observable systems, including exact characterisation of the random return, variance bound, distributional sensitivity under perturbations, and distributional approximation using a finite number of random variables (Corollaries 1–4). These extensions build on the augmented system introduced by a given linear feedback controller and a linear observer, aligning with the well-known separation principle [33]. These results provide insight into extending DRL to partially observable systems.

The work that comes closest to addressing the problems above is our prior work [32]: the current contribution additionally analyses the variance of the random return and the distributional sensitivity with respect to model perturbations. Moreover, this work constructs a confidence bound on the distribution approximation error for the model-free case when the system matrices are unknown. Additionally, we newly derive corresponding counterparts for partially observable models.

C. Organisation and Notations

The paper is organized as follows. In Section II, we provide background on LQR and define our problem. In Section III, we provide the main results for distributional LQR, including the analytical expression of the random return, variance bound, distributional sensitivity under perturbations and model-based and model-free distribution approximations. Section IV provides the main results for partially observable linear systems. In Section V, we experimentally verify our theoretical results. Finally, we conclude the paper in Section VI.

We denote by $\mathbb{R}$ the set of real numbers and $\mathbb{N}$ the set of natural numbers. For a symmetric matrix P, the notation $P > 0$ means that P is positive definite. For a matrix $Q \in \mathbb{R}^{n\times n}$, we denote by $\|Q\|$ and $\|Q\|_F$ its spectral norm and Frobenius norm, respectively. To indicate that two random variables Z_1 and Z_2 are equal in distribution, we use the notation $Z_1 \overset{D}{=} Z_2$. For a random variable Z, $\mathbb{E}[Z]$ denotes its expectation.

II. Problem Statement

Consider a discrete-time linear control system:

$$x_{t+1} = Ax_t + Bu_t + v_t,$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^p$, and $v_t \in \mathbb{R}^n$ are the system state, control input, and the exogenous disturbance, respectively. We assume that the exogenous disturbances v_t with bounded moments, $t \in \mathbb{N}$, are i.i.d. sampled from a distribution $\mathcal{D}$ of arbitrary form.

A. Classical Discounted LQR

The canonical LQR problem aims to find a control policy $\pi : \mathbb{R}^n \to \mathbb{R}^p$ to minimise the objective

$$J(u) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (x_t^{\mathrm{T}} Q x_t + u_t^{\mathrm{T}} R u_t)\right],$$

where Q, R are positive-definite constant matrices and $\gamma \in (0,1)$ is a discount parameter. Given a control policy π, let $V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^k (x_t^{\mathrm{T}} Q x_t + u_t^{\mathrm{T}} R u_t)\right]$ denote the expected return from an initial state $x_0 = x$ with $u_t = \pi(x_t)$. For the static linear policy $\pi(x_t) = Kx_t$, the value function $V^\pi(x)$ satisfies the Bellman equation

$$V^\pi(x) = x^{\mathrm{T}}(Q + K^{\mathrm{T}} R K)x + \gamma \mathop{\mathbb{E}}_{x'=(A+BK)x+v_0}[V^\pi(x')], \tag{1}$$

where the capital letter x' denotes a random variable over which we take the expectation.

When the exogenous disturbance v_t is normally distributed with zero mean, the value function is known to take the quadratic form $V^\pi(x) = x^{\mathrm{T}} P x + q$, where $P > 0$ is the solution of the Lyapunov equation $P = Q + K^{\mathrm{T}} R K + \gamma A_K^{\mathrm{T}} P A_K$ and q is a scalar related to the variance of v_t. In particular, the optimal control feedback gain is $K^* = -\gamma(R + \gamma B^{\mathrm{T}} P B)^{-1} B^{\mathrm{T}} P A$ and P is the solution to the Riccati equation $P = \gamma A^{\mathrm{T}} P A - \gamma^2 A^{\mathrm{T}} P B (R + \gamma B^{\mathrm{T}} P B)^{-1} B^{\mathrm{T}} P A + Q$.

B. Distributional LQR

Motivated by the advantages of DRL in better understanding the effects of the randomness in the environment and in considering more general optimality criteria, in this paper we propose a distributional approach to the LQR problem. Unlike classical RL, which relies on expected returns, DRL [34] relies on the distribution of random returns, which is referred to *return distribution*. The return distribution characterises the probability distribution of different returns generated by a given policy and, as such, it contains much richer information on the performance of a given policy compared to the expected return. In the context of LQR, we denote by $G^\pi(x)$ the random return using the static control strategy $u_t = \pi(x_t)$ from the initial state $x_0 = x$, which is defined as

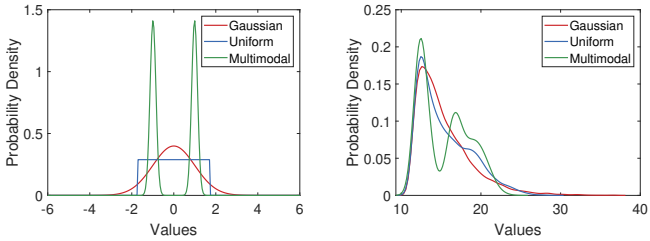
$$G^\pi(x) = \sum_{t=0}^{\infty} \gamma^t (x_t^{\mathrm{T}} Q x_t + u_t^{\mathrm{T}} R u_t), \quad u_t = \pi(x_t), x_0 = x. \tag{2}$$

It is straightforward to see that the expectation of $G^\pi(x)$ is equal to the value function $V^\pi(x)$. The standard Bellman equation in (1) decomposes the long-term expected return into an immediate stage cost plus the expected return of future actions starting at the next step. Similarly, we can define the random variable Bellman equation for the random return as

$$G^\pi(x) \overset{D}{=} x^{\mathrm{T}} Q x + \pi(x)^{\mathrm{T}} R \pi(x) + \gamma G^\pi(x'), \quad x' = Ax + B\pi(x) + v_0. \tag{3}$$

Here we use the notation $Z_1 \overset{D}{=} Z_2$ to denote that two random variables Z_1, Z_2 are equal in distribution. Compared to the expected return in LQR, which is a scalar, here the return distribution is infinite-dimensional.

The following example is used to highlight the need of considering the random return.



(a) Probability density values of three types of disturbance v_t.

(b) Probability density values of the random cost $G^\pi(x)$ induced by three disturbances.

Fig. 1: The PDFs of three types of disturbance and of their corresponding random costs in LQR. The PDFs of the random costs are generated by Algorithm 1 in this paper.

Example 1. *Consider the discrete-time scalar linear system $x_{t+1} = x_t + u_t + v_t$ and three different types of disturbance v_t: normal distribution $\mathcal{N}(0,1)$, uniform distribution $U[-\sqrt{3}, \sqrt{3}]$, and multimodal distribution which is characterised by the probability density function (PDF) $(p_1(z) + p_2(z))/2$, where $p_i(z) = \frac{1}{\sqrt{2\pi}\sigma_i} exp\big(-\frac{(z-\mu_i)^2}{2\sigma_i^2}\big)$, $i = 1, 2$, with $\mu_1 = -0.99$, $\mu_2 = 0.99$, $\sigma_1 = \sigma_2 = \sqrt{1 - 0.99^2}$. Their PDFs are shown in Fig. 1(a). It can be verified that the mean of v_t is zero and the variance is 1 for all three types of disturbance. We set the initial state $x = 3$, $Q = R = 1$, and $\gamma = 0.6$. Then, the optimal controller for the three disturbances is the same and given by $u_t = -0.4684 x_t$. The value function $V^\pi(x)$ in (1) of implementing the optimal controller for the three disturbances is the same as well since the variance of v_t is the same.*

However, the distribution of the random return $G^\pi(x)$ varies significantly for the three disturbances, as shown in Fig. 1(b). We observe that the distributions of the random cost for the Gaussian and uniform disturbances are close to chi-square distributions, but the distribution for the multimodal disturbance exhibits multiple peaks. Hence, the distribution of the random return contains more information than in the LQR problem, offering insights into risk analysis, which the mean value alone cannot capture: the random return $G^\pi(x)$ enables us, for instance, to select policies that minimize risks or satisfy probabilistic constraints, which is not possible to do from the value function $V^\pi(x)$.

This paper addresses the following research problems. We first analytically characterise the random return that fulfils the random variable Bellman equation for LQR. Subsequently, we explore the fundamental properties of the random return and its distribution: variance bound, distributional sensitivity under perturbations, and model-based and model-free distribution approximations. Finally, we extend our investigation to encompass partially observable linear systems.

III. Main Results on Distributional LQR

This section focuses on the random return for the LQR problem.

A. Characterisation of the Random Return

In this section, we precisely characterise the distribution of the random return that satisfies the distributional Bellman equation (3). Given a static linear policy $\pi(x_t) = Kx_t$, we denote by $G^K(x)$ the random return $G^\pi(x)$ under the policy $\pi(x_t)$ from the initial state $x_0 = x$, which is defined as

$$G^K(x) = \sum_{t=0}^{\infty} \gamma^t x_t^{\mathrm{T}}(Q + K^{\mathrm{T}}RK)x_t, \quad x_0 = x. \tag{4}$$

The random return $G^K(x)$ satisfies the following random variable Bellman equation

$$G^K(x) \overset{D}{=} x^{\mathrm{T}}Q_K x + \gamma G^K(x'), \quad x' = A_K x + v_0, \tag{5}$$

where $A_K := A + BK$ and $Q_K := Q + K^{\mathrm{T}}RK$. In the following theorem, we provide an explicit expression of the random return $G^K(x)$. The proof can be found in [32].

Theorem 1. *[32] Suppose that the feedback gain K is stabilizing and satisfies $\|A_K\| = \rho_K < 1$. Let*

$$G^K(x) = x^{\mathrm{T}}Px + 2\sum_{k=0}^{\infty} \gamma^{k+1} w_k^{\mathrm{T}} P A_K^{k+1} x + \sum_{k=0}^{\infty} \gamma^{k+1} w_k^{\mathrm{T}} P w_k + 2\sum_{k=1}^{\infty} \gamma^{k+1} w_k^{\mathrm{T}} P \sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau, \tag{6}$$

where P is obtained from the Lyapunov equation $P = Q + K^{\mathrm{T}}RK + \gamma A_K^{\mathrm{T}} P A_K$, and the random variables $w_k \sim \mathcal{D}$ are independent from each other for all $k \in \mathbb{N}$. Then, the random variable $G^K(x)$ defined in (6) *is a fixed point solution to the random variable Bellman equation* (5)*.*

We note that the expression of the random return is meaningful only when the system is stable. When ensuring stability, this analytical expression applies to arbitrary exogenous disturbances including non-Gaussian, uniform noises and noises with non-zero means, as long as the disturbances are i.i.d.. For each realization of the sequence $\{w_k\}_{k=0}^{\infty}$, $G^K(x)$ is represented as an infinite series, which is convergent when these w_k are bounded.

Remark 1. *It is worth mentioning that Theorem 1 holds for a random initial state as long as it is independent of the process noise. That is, when x is random and is independent of the exogenous noise v_t in the system, the random return $G^K(x)$ has the same expression as in Theorem 1. To maintain consistency in our subsequent results, where the upper bounds depend on the initial state x, we present the results for a fixed initial state.*

If we assume $\mathbb{E}[w_k] = 0$, $\mathbb{E}[w_k w_k^{\mathrm{T}}] = \sigma^2 I$, and the disturbances w_k are i.i.d., we have that the expected value of the random return is $x^{\mathrm{T}}Px + \sigma^2 \frac{\gamma}{1-\gamma}\mathrm{Tr}(P)$, which aligns with the classical result in LQR. This observation to some degree validates our characterisation of the random return.

Remark 2. *Recall that the PDF of the sum of two independent random variables is the convolution of their two PDFs. Computing the accurate probability distribution function of $G^K(x)$ in* (6) *is a challenging task due to the potential need for an infinite number of convolution operations. However, we can discuss the approximate shape of this distribution under different conditions. Suppose that the random variable w_k follows a normal distribution. When the initial state is significantly large, the random variable $\sum_{k=0}^{\infty} \gamma^{k+1} w_k^{\mathrm{T}} P A_K^{k+1} x$ dominates the random return. This sum follows a Gaussian distribution, and as a result, the overall distribution of $G^K(x)$ tends to resemble a Gaussian distribution. Conversely, when the value of x is small, the term $\sum_{k=0}^{\infty} \gamma^{k+1} w_k^{\mathrm{T}} P w_k$ becomes dominant. This sum follows a chi-square distribution, and consequently, the entire distribution of $G^K(x)$ takes on a chi-square-like shape. More details can be found in Section V.*

B. Bounded Variance of the Random Return

In this section, we analyse the variance of the random return $G^K(x)$, which is presented in the following theorem. The proof can be found in Appendix A.

Theorem 2. *Assume that $\mathbb{E}[w_k] = 0$ and $\mathbb{E}[\|w_k\|^4] \leq \sigma_4^4$, for all $k \in \mathbb{N}$. Suppose that the feedback gain K satisfies $\|A_K\| = \rho_K < 1$. Then, the variance of the random variable $G^K(x)$ is bounded.*

Although $G^K(x)$ in (6) is composed of infinitely many random variables, Theorem 2 shows that its variance is bounded if the fourth moment of the disturbance is bounded. The fourth moment qualitatively is a measure of the tail of a probability distribution. To ensure a finite variance for the random cost $G^K(x)$, we thus require that the tail of the disturbance distribution is not heavy. This condition seems indispensable, since $G^K(x)$ includes a term $w_k^{\mathrm{T}} P w_k$.

C. Sensitivity Analysis of the Return Distribution

In this section, we investigate how perturbations on matrices influence the distribution of the random return $G^K(x)$. Suppose that we perturb the matrices A, B by an amount ΔA, ΔB, respectively. Let

$$\tilde{A} = A + \Delta A, \quad \tilde{B} = B + \Delta B, \quad A_K = A + BK,$$
$$\tilde{A}_K = \tilde{A} + \tilde{B}K, \quad \Delta A_K = \tilde{A}_K - A_K,$$

and let P and $\tilde{P}$ be the solutions to

$$P - Q - K^{\mathrm{T}}RK = \gamma A_K^{\mathrm{T}} P A_K,$$
$$\tilde{P} - Q - K^{\mathrm{T}}RK = \gamma \tilde{A}_K^{\mathrm{T}} \tilde{P} \tilde{A}_K,$$

respectively. With the introduction of perturbations on matrices, we define the perturbed random variable

$$\tilde{G}^K(x) = x^{\mathrm{T}}\tilde{P}x + 2\sum_{k=0}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}\tilde{P}\tilde{A}_K^{k+1}x + \sum_{k=0}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}\tilde{P}w_k + 2\sum_{k=1}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}\tilde{P}\sum_{\tau=0}^{k-1}\tilde{A}_K^{k-\tau}w_\tau. \quad (7)$$

Let F_x^K and $\tilde{F}_x^K$ denote the cumulative distribution function (CDF) of $G^K(x)$ and $\tilde{G}^K(x)$, respectively. In the following theorem, we show that the sup difference between F_x^K and $\tilde{F}_x^K$ is bounded when the perturbation is reasonably small. The proof can be found in Appendix B.

Theorem 3. *Assume that the PDF of w_k is bounded, and satisfies $\mathbb{E}[w_k^{\mathrm{T}}w_k] \le \sigma^2$, for all $k \in \mathbb{N}$. Suppose that the feedback gain K satisfies $\max\{\|A_K\|, \|\tilde{A}_K\|\} = \rho_K < 1$. Suppose that $l > 2\epsilon$, where $l = \|H^{-1}\|^{-1}$, $H = I \otimes I - \gamma A_K^{\mathrm{T}} \otimes A_K^{\mathrm{T}}$ and $\epsilon = \gamma \|A_K\|_F \|\Delta A_K\|_F + \frac{\gamma}{2}\|\Delta A_K\|_F^2$. Then, we have*

$$\sup_z |F_x^K(z) - \tilde{F}_x^K(z)| \le \tilde{c}_1\|\Delta A_K\| + \tilde{c}_2\|\Delta A_K\|^2, \quad (8)$$

where the constants $\tilde{c}_1, \tilde{c}_2$ (made explicit in the proof) depend on the system matrices, the initial state value x, and the parameters γ, ρ_K, σ.

Traditional sensitivity analysis investigates the impact of perturbations on solutions to the Lyapunov equation, see, e.g., [35]. Building on this result, Theorem 3 shows that we can also bound changes in the perturbed return distribution.

D. Model-Based Approximation of the Return Distribution

The expression of the random return $G^K(x)$ defined in (6) is composed of infinitely many random variables. In this section, we investigate how to approximate the distribution of this random return using a finite number of random variables. A natural idea is to consider only the first N terms in the summations in the expression (6) and disregard the terms for k larger than N, which yields the following:

$$G_N^K(x) = x^{\mathrm{T}}Px + 2\sum_{k=0}^{N-1}\gamma^{k+1}w_k^{\mathrm{T}}PA_K^{k+1}x + \sum_{k=0}^{N-1}\gamma^{k+1}w_k^{\mathrm{T}}Pw_k + 2\sum_{k=1}^{N-1}\gamma^{k+1}w_k^{\mathrm{T}}P\sum_{\tau=0}^{k-1}A_K^{k-\tau}w_\tau. \quad (9)$$

Let $F_{x,N}^K$ denote the CDF of $G_N^K(x)$. The following theorem provides an upper bound on the difference between F_x^K and $F_{x,N}^K$, and shows that the sequence $\{G_N^K(x)\}_{N\in\mathbb{N}}$ converges pointwise in distribution to $G^K(x)$, $\forall x \in \mathbb{R}^n$. The proof can be found in [32].

Theorem 4. *[32] Assume that the PDF of w_k is bounded, and satisfies $\mathbb{E}[w_k^{\mathrm{T}}w_k] \le \sigma^2$, for all $k \in \mathbb{N}$. Suppose that the feedback gain K satisfies $\|A_K\| = \rho_K < 1$. Then, the sup difference between the CDFs F_x^K and $F_{x,N}^K$ is bounded by*

$$\sup_z |F_x^K(z) - F_{x,N}^K(z)| \le c_0\gamma^N, \quad (10)$$

Algorithm 1: Model-free Distributional Policy Evaluation

Require: initial values x, controller K

for iteration $m = 1, \dots, M$ **do**
  Initial state $x_{m,0} = x$;
  for time $t = 0, 1, \dots, T-1$ **do**
    Implement controller $u_{m,t} = Kx_{m,t}$;
    Observe $x_{m,t+1} = Ax_{m,t} + Bu_{m,t} + v_t$;
  end for
  Obtain $G_m^{K,T}(x) = \sum_{t=0}^{T}\gamma^t x_{m,t}^{\mathrm{T}}(Q + K^{\mathrm{T}}RK)x_{m,t}$;
end for
Construct EDF $\hat{F}_{x,M}^{K,T}(z) = \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\{G_m^{K,T}(x) \le z\}$.

where c_0 is a constant (again, made explicit in the proof) that depends on the system matrices, the initial state value x, and the parameters γ, ρ_K, σ.

Remark 3. *The bound on the distribution approximation in* (10) *relies on the conditions of Theorem 4, which ensure that the PDF of G_N^K is continuous and bounded. Note that these conditions are not strict, and indeed hold for many noise distributions commonly used in linear control systems, including the Gaussian and uniform ones.*

E. Model-Free Approximation of the Return Distribution

When the matrices A, B are unknown, one cannot use the exact form of the random return to compute the distribution. In this section, we propose a model-free method to estimate the distribution of the random return.

In the absence of information about the system matrices A and B, one costly yet straightforward approach to estimate the distribution is by directly sampling the random return $G^K(x)$ as defined in (4). This random return represents the sum of discounted rewards over an infinite time horizon. To make the computation practically manageable, we truncate the time horizon and disregard rewards occurring after time step T. Accordingly, we define the random variable

$$G^{K,T}(x) = \sum_{t=0}^{T}\gamma^t x_t^{\mathrm{T}}(Q + K^{\mathrm{T}}RK)x_t, \quad x_0 = x. \quad (11)$$

We denote by $F_x^{K,T}(z)$ the CDF of $G^{K,T}(x)$ and recall that $F_x^K(z)$ is the CDF of $G^K(x)$. Intuitively, $F_x^{K,T}(z)$ closely approximates $F_x^K(z)$ when T is sufficiently large. This is due to the fact that every term beyond time step T becomes negligible after being discounted by γ^t. Therefore, we sample the random return $G^{K,T}(x)$ to estimate the distribution of $G^K(x)$.

The detailed model-free distributional policy evaluation is presented in Algorithm 1. Specifically, at each iteration m, starting from the initial state $x_{m,0} = x$, we repeatedly implement the static controller $u_{m,t} = Kx_{m,t}$ and generate a

total of M trajectories. Given the m-th trajectory $\{x_{m,t}\}_{t=0:T}$ with $x_{m,0} = x$, define the sampling cost

$$G_m^{K,T}(x) = \sum_{t=0}^{T} \gamma^t x_{m,t}^{\mathrm{T}}(Q + K^{\mathrm{T}}RK)x_{m,t}. \tag{12}$$

Using $G_m^{K,T}(x)$, $m = 1, \ldots, M$, the empirical distribution function (EDF) is constructed by

$$\hat{F}_{x,M}^{K,T}(z) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{G_m^{K,T}(x) \le z\}, \tag{13}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Intuitively, the empirical distribution $\hat{F}_{x,M}^{K,T}(z)$ is close to the distribution $F_x^{K,T}(z)$ when M is sufficiently large and to the distribution $F_x^K(z)$ when T is also large. The following theorem provides an upper bound on the difference between the distributions $\hat{F}_{x,M}^{K,T}(z)$ and $F_x^K(z)$. The proof can be found in Appendix C.

Theorem 5. *Assume that the PDF of w_k is bounded, and satisfies $\mathbb{E}[w_k] = 0$ and $\mathbb{E}[w_k^{\mathrm{T}} w_k] \le \sigma^2$, for all $k \in \mathbb{N}$. Suppose that the feedback gain K satisfies $\|A_K\| = \rho_K < 1$. Then, with probability at least $1-\delta$, we have*

$$\sup_z |\hat{F}_{x,M}^{K,T}(z) - F_x^K(z)| \le \sqrt{\frac{\ln(1/\delta)}{2M}} + f_{\max}\|Q_K\|\gamma^{T+1}(c_1\rho_K^{2(T+1)} + c_2\rho_K^{T+1} + c_3), \tag{14}$$

where $f_{\max}$ is the maximum of the PDF of the random variable $G^{K,T}(x)$, $Q_K = Q + K^{\mathrm{T}}RK$, $c_1 = \frac{\|x\|^2}{1-\gamma\rho_K^2}$, $c_2 = \frac{2\|x\|\sigma}{(1-\rho_K)(1-\gamma\rho_K)}$, $c_3 = \frac{\sigma^2}{(1-\gamma)(1-\gamma\rho_K)}$.

Theorem 5 shows that the accuracy of the distribution estimate depends on the choice of two key parameters: the time horizon T and the number of generated trajectories M. When both M and T are sufficiently large, $\hat{F}_{x,M}^{K,T}(z)$ can serve as a reliable approximation for $F_x^K(z)$. Supported by this result, we consider the return distribution learned by Algorithm 1 with sufficiently large values of M, T as the true return distribution in the simulation part. Practically, given a target approximation error ε, we can determine the required values of M and T by ensuring that $\sqrt{\frac{\ln(1/\delta)}{2M}} \le (1-a)\varepsilon$ and $f_{\max}\|Q_K\|\gamma^{T+1}(c_1\rho_K^{2(T+1)} + c_2\rho_K^{T+1} + c_3) \le a\varepsilon$ for any $a \in (0,1)$.

Remark 4. *We note that the random variables $G_N^K(x)$ and $G^{K,T}(x)$ serve as the approximations to the true random return $G^K(x)$ by truncating the number of random variables and the time horizon, respectively. As shown in Theorems 4 and 5, increasing the number of random variables or extending the truncated horizon definitely enhance the approximation accuracy for model-based and model-free methods, respectively. However, the associated costs one needs to pay to obtain their distributions are usually different. As shown in Table I, the model-free method requires a sufficiently large value of M to achieve a reliable distribution estimate with a high probability. In contrast, the model-based method can attain the same level of accuracy with probability 1 using a significantly smaller number of random variables. Hence, when the system matrices and the disturbances are known, the computation of the distribution of $G_N^K(x)$ incurs less costs. It is also worth noting that the model-free method is not sensitive to the discount parameter γ while the model-based method is.*

		MB method	MF method	MF method
γ	UB	N	$T, M(\delta = 5\%)$	$T, M(\delta = 1\%)$
0.6	0.02	11	(100,4000)	(100,6000)
0.6	0.01	12	(100,15000)	(100,23000)
0.8	0.02	24	(100,4000)	(100,6000)
0.8	0.01	27	(100,15000)	(100,23000)
0.95	0.02	112	(200,4000)	(200,6000)
0.99	0.02	698	(1000,4000)	(1000,6000)

TABLE I: Comparison of model-based (MB) and model-free (MF) approximation methods. Here γ is the discount parameter; UB is the bound in the right hand side of (10) and (14) for model-based and model-free methods, respectively; N is the smallest integer such that $c_0\gamma^N \le \mathrm{UB}$; and $T, M(\delta)$ denote the pair comprising the horizon length T and the number of trajectories M required so that the approximation error in (14) is bounded by UB with probability at least $1-\delta$.

Remark 5. *For the discounted infinite-horizon LQR problem, the stability criterion is relaxed to requiring that $\sqrt{\gamma}(A+BK)$ is stable [36], [37]. However, when analyzing the entire return distribution, our results indicate that we need $A + BK$ to be stable, i.e., $\|A + BK\| < 1$, which is a more stringent condition. This is because the random return $G^K(x)$ includes a term $2\sum_{k=1}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}P\sum_{\tau=0}^{k-1}A_K^{k-\tau}w_\tau$, necessitating that $A_K^{k-\tau}$ remains bounded to ensure the convergence of the series. This issue does not arise in the expected return case, since the term involving the process noise disappears when taking the expectation, due to the zero mean of w_k.*

IV. Extension to Partially Observable Systems

In this section, we analyse the case when the state is not fully observable. We show that most of the results for LQR can be extended to this partially observable case.

Consider a partially observable discrete-time linear control system:

$$x_{t+1} = Ax_t + Bu_t + v_t,$$
$$y_t = Cx_t + s_t,$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^p$, $y_t \in \mathbb{R}^l$, $v_t \in \mathbb{R}^n$, and $s_t \in \mathbb{R}^l$ are the system state, control input, system output, process noise, and observation noise, respectively. We assume that the system is observable and controllable. By introducing the feedback gain K and the observer gain L, we define the estimated state and controller

$$\hat{x}_{t+1} = A\hat{x}_t + Bu_t + L(y_t - C\hat{x}_t),$$
$$u_t = K\hat{x}_t.$$

By defining $\tilde{x}_t = x_t - \hat{x}_t$, $\bar{x}_t = [x_t^{\mathrm{T}}, \tilde{x}_t^{\mathrm{T}}]^{\mathrm{T}}$, we get the augmented system

$$\bar{x}_{t+1} = \bar{A}_{KL}\bar{x}_t + \bar{v}_t, \tag{15}$$

where

$$\bar{A}_{KL} = \begin{bmatrix} A+BK & -BK \\ 0 & A-LC \end{bmatrix}, \bar{v}_t = F \begin{bmatrix} v_t \\ s_t \end{bmatrix}, \quad F = \begin{bmatrix} I & 0 \\ I & -L \end{bmatrix}.$$

If $\mathbb{E}[v_t] = \mathbb{E}[s_t] = 0$, and the collection of v_t and s_t is i.i.d., it is easy to verify that $\mathbb{E}[\bar{v}_t] = 0$ and the collection of $\bar{v}_t$ is i.i.d.. We denote the distribution of $\bar{v}_t$ by $\bar{\mathcal{D}}$. Define $\bar{Q}_K := \begin{bmatrix} Q+K^{\mathrm{T}}RK & -K^{\mathrm{T}}RK \\ -K^{\mathrm{T}}RK & K^{\mathrm{T}}RK \end{bmatrix}$ and the random return

$$\begin{aligned} G^{KL}(\bar{x}) &= \sum_{t=0}^{\infty} \gamma^t (x_t^{\mathrm{T}} Q x_t + u_t^{\mathrm{T}} R u_t) \\ &= \sum_{t=0}^{\infty} \gamma^t (x_t^{\mathrm{T}} Q x_t + \hat{x}_t^{\mathrm{T}} K^{\mathrm{T}} R K \hat{x}_t) \\ &= \sum_{t=0}^{\infty} \gamma^t \bar{x}_t^{\mathrm{T}} \bar{Q}_K \bar{x}_t, \quad \bar{x}_0 = \bar{x}, \end{aligned} \tag{16}$$

where $\bar{x} = [x^{\mathrm{T}}, \tilde{x}_0^{\mathrm{T}}]^{\mathrm{T}}$. The random return $G^{KL}(\bar{x})$ satisfies the following random variable Bellman equation

$$G^{KL}(\bar{x}) \overset{D}{=} \bar{x}^{\mathrm{T}} \bar{Q}_K \bar{x} + \gamma G^{KL}(X'), \quad X' = \bar{A}_{KL}\bar{x} + \bar{v}_0. \tag{17}$$

In the following corollary, we provide an explicit expression of the random return $G^{KL}(\bar{x})$. The proof can be obtained by applying the results in Theorem 1 to the augmented system (15) and is omitted.

Corollary 1. *Suppose that the feedback gain K and observer gain L are chosen such that $\left\|\bar{A}_{KL}\right\| = \bar{\rho}_K < 1$. Let*

$$\begin{aligned} G^{KL}(\bar{x}) &= \bar{x}^{\mathrm{T}} \bar{P} \bar{x} + 2\sum_{k=0}^{\infty} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \bar{P} \bar{A}_{KL}^{k+1} \bar{x} \\ &+ \sum_{k=0}^{\infty} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \bar{P} \bar{w}_k + 2\sum_{k=1}^{\infty} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \bar{P} \sum_{\tau=0}^{k-1} \bar{A}_{KL}^{k-\tau} \bar{w}_\tau, \end{aligned} \tag{18}$$

where $\bar{P}$ is obtained from the Lyapunov equation $\bar{P} = \bar{Q}_K + \gamma \bar{A}_{KL}^{\mathrm{T}} \bar{P} \bar{A}_{KL}$, and the random variables $\bar{w}_k \sim \bar{\mathcal{D}}$ are independent from each other for all $k \in \mathbb{N}$. Then, the random variable $G^{KL}(x)$ defined in (16) *is a fixed point solution to the random variable Bellman equation* (17)*.*

The variance bound part is similar to that of the fully observable case, and is presented in the following corollary. The proof can be obtained by following a similar methodology to that employed for Theorem 2 and is omitted.

Corollary 2. *Assume that $\mathbb{E}[\bar{w}_k] = 0$ and $\mathbb{E}[\|\bar{w}_k\|^4] \le \bar{\sigma}_4$, for all $k \in \mathbb{N}$. Suppose that the feedback gain K and observer gain L are chosen such that $\left\|\bar{A}_{KL}\right\| = \bar{\rho}_K < 1$. Then, the variance of the random variable $G^{KL}(x)$ is bounded.*

The sensitivity analysis for the partially observable case is similar to that of the fully observable case. Suppose that we perturb the matrix $\bar{A}_{KL}$ by an amount $\Delta\bar{A}_{KL}$. Define the matrix $\check{A}_{KL} = \bar{A}_{KL} + \Delta\bar{A}_{KL}$ and the perturbed random variable

$$\begin{aligned} \tilde{G}^{KL}(\bar{x}) &= \bar{x}^{\mathrm{T}} \check{P} \bar{x} + 2\sum_{k=0}^{\infty} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \check{P} \check{A}_K^{k+1} x \\ &+ \sum_{k=0}^{\infty} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \check{P} \bar{w}_k + 2\sum_{k=1}^{\infty} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \check{P} \sum_{\tau=0}^{k-1} \check{A}_K^{k-\tau} \bar{w}_\tau. \end{aligned} \tag{19}$$

where $\check{P} = \bar{Q}_K + \gamma \check{A}_{KL}^{\mathrm{T}} \check{P} \check{A}_{KL}$. Let F_x^{KL} and $\tilde{F}_x^{KL}$ denote the CDF of $G^{KL}(x)$ and $\tilde{G}^{KL}(x)$, respectively. We obtain the perturbation for partially observable case in the following corollary. The proof can be adapted from that of Theorem 3 and is omitted.

Corollary 3. *Assume that the PDF of $\bar{w}_k$ is bounded, and satisfy $\mathbb{E}[\bar{w}_k^{\mathrm{T}} \bar{w}_k] \le \bar{\sigma}^2$, for all $k \in \mathbb{N}$. Suppose that the feedback gain K and the observer L are chosen such that $\max\{\left\|\bar{A}_{KL}\right\|, \left\|\check{A}_{KL}\right\|\} = \bar{\rho}_K < 1$. Suppose that $\bar{l} > 2\bar{\epsilon}$, where $\bar{l} = \left\|\bar{H}^{-1}\right\|^{-1}$, $\bar{H} = I \otimes I - \gamma \bar{A}_{KL}^{\mathrm{T}} \otimes \bar{A}_{KL}^{\mathrm{T}}$ and $\bar{\epsilon} = \gamma \left\|\bar{A}_{KL}\right\|_F \left\|\Delta\bar{A}_{KL}\right\|_F + \frac{\gamma}{2} \left\|\Delta\bar{A}_{KL}\right\|_F^2$. Then, we have*

$$\sup_z |F_x^{KL}(z) - \tilde{F}_x^{KL}(z)| \le \bar{c}_1 \left\|\Delta\bar{A}_{KL}\right\| + \bar{c}_2 \left\|\Delta\bar{A}_{KL}\right\|^2,$$

where the constants $\bar{c}_1, \bar{c}_2$ depend on the system matrices, the initial state value $\bar{x}$, and the parameters $\gamma, \bar{\rho}_K, \bar{\sigma}$.

The approximation part is similar to that of the fully observable case. Let

$$\begin{aligned} G_N^{KL}(\bar{x}) &= \bar{x}^{\mathrm{T}} \bar{P} \bar{x} + 2\sum_{k=0}^{N} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \bar{P} \bar{A}_{KL}^{k+1} \bar{x} \\ &+ \sum_{k=0}^{N} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \bar{P} \bar{w}_k + 2\sum_{k=1}^{N} \gamma^{k+1} \bar{w}_k^{\mathrm{T}} \bar{P} \sum_{\tau=0}^{k-1} \bar{A}_{KL}^{k-\tau} \bar{w}_\tau. \end{aligned} \tag{20}$$

Let F_x^{KL} and $F_{x,N}^{KL}$ denote the CDF of $G^{KL}(x)$ and $G_N^{KL}(x)$, respectively. In the following theorem, we show that the approximation error with a finite number of random variables can be bounded in the partially observable case. The proof can be adapted from that of Theorem 4 and is omitted.

Corollary 4. *Assume that the PDF of $\bar{w}_k$ is bounded, and satisfy $\mathbb{E}[\bar{w}_k^{\mathrm{T}} \bar{w}_k] \le \bar{\sigma}^2$, for all $k \in \mathbb{N}$. Suppose that the feedback gain K and observer gain L are chosen such that $\left\|\bar{A}_{KL}\right\| = \bar{\rho}_K < 1$. Then, the sup difference between the CDFs F_x^{KL} and $F_{x,N}^{KL}$ is bounded by*

$$\sup_z |F_x^{KL}(z) - F_{x,N}^{KL}(z)| \le \bar{c}_0 \gamma^N, \tag{21}$$

where $\bar{c}_0$ is a constant that depends on the system matrices, the initial state value $\bar{x}$, and the parameters $\gamma, \bar{\rho}_K, \bar{\sigma}$.

We remark that the model-free approximation (Theorem 5) is not applicable for partially observable systems. Unlike the distributional LQR where the state is directly measurable, it is nontrivial to achieve an accurate estimation of the states such that the cumulative estimation error can be controlled arbitrarily small using only the observation sequence $\{y_t\}$ and control sequence $\{u_t\}$ (when the system model is unknown). Thus, we leave this problem for future work.

V. Experiments

In this section, we consider an idealised example of data center cooling with three sources coupled to their own cooling devices [13], [18], [38] with the dynamics $x_{t+1} = Ax_t + Bu_t + v_t$, where

$$A = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We select $Q = I$ and $R = I$. The exogenous disturbances have standard normal distributions with zero mean.

Even for this linear system, it is impossible to simplify the expression of the exact return distribution, which still depends on an infinite number of random variables. Thus, as a baseline for the return distribution, we generate an empirical distribution by Algorithm 1 with a sufficiently large amount of samples that approximates the true distribution of the random return. Specifically, we run Algorithm 1 with the parameters $T = 3000$ and $M = 30000$. By Theorem 5, the maximal difference between the generated empirical distribution and the true one is bounded by 0.0088 with probability at least 99%, which means that the generated empirical distribution is reliably close to the true one. We use the sample frequency over evenly-divided regions as an approximation of the PDF.

A. LQR

We first consider the fully observable case. We select different values of γ and x_0, and fix the optimal controller gain $K = -\gamma(R + \gamma B^T P B)^{-1} P A$, where P is the solution to the classic Riccati equation $P = \gamma A^T P A - \gamma^2 A^T P B (R + \gamma B^T P B)^{-1} B^T P A + Q$. The controller is given by

$$K = -0.01 \begin{bmatrix} 56.19 & 0.7692 & 0.0027 \\ 0.7692 & 56.20 & 0.7692 \\ 0.0027 & 0.7692 & 56.19 \end{bmatrix}.$$

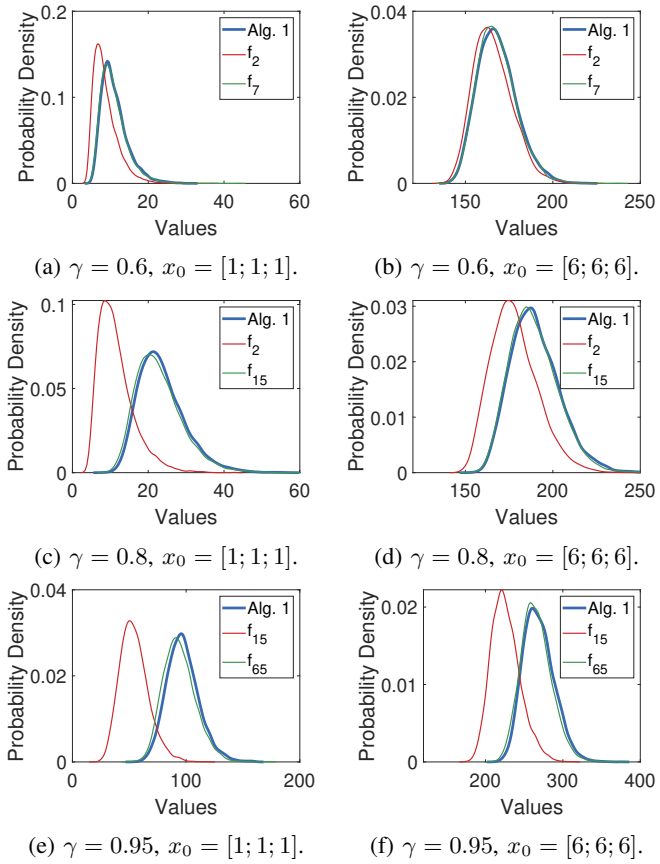
In what follows, we verify the results in Theorem 4 by evaluating the quality of the approximation of the return distribution using different numbers of random variables. We denote here by f_N the distribution of the approximated random return $G_N^K(x_0)$ in (9) obtained considering N random variables. We compute the constant c_0 in equation (10) and the required number of random variables that guarantees $\sup_z |F_x^K(z) - F_{x,N}^K(z)| \leq 0.01$, meaning that the estimate distribution is sufficiently close to the true distribution. As shown in Table II, an increasing number of random variables is needed when dealing with larger values of γ and/or x_0. The simulation results are shown in Fig. 2. Specifically, Fig. 2 (a) and (c) show that when γ is small, the return distribution can be well approximated using only a few random variables ($N = 7$ works well). However, when γ approaches 1, more random variables are needed for an accurate approximation: as shown in Fig. 2 (b) and (d), we need $N = 15$ to have a good approximation of the return distribution in the case of $\gamma = 0.8$.

Moreover, the value of the initial state x_0 has an influence on the shape of the return distribution. When x_0 is large, the random variable $w_k^T P A_K^{k+1} x_0$ dominates and, therefore, its distribution is close to a Gaussian distribution, as shown in Fig. 2 (c) and (d). If instead x_0 is small, then the random variable $w_k^T P w_k$ plays a leading role, so the overall distribution is close to the chi-square one, as shown in Fig. 2 (a) and (b).

Next we perturb the matrices A, B by an amount $\epsilon_A A$ and $\epsilon_B B$, respectively. We select $x_0 = [1; 1; 1]$. We compute the constants of $\tilde{c}_1$, $\tilde{c}_2$, the true sup difference between original and perturbed distributions, and the upper bounds in (8). The results are shown in Table III. We observe that the perturbed distribution becomes significantly distinct from the original distribution when γ, ϵ_A, and ϵ_B take on larger values. We also note that our computational upper bound becomes conservative

γ	x_0	c_0	N_0
0.6	[1;1;1]	0.5447	8
0.8	[1;1;1]	0.5917	19
0.6	[6;6;6]	1.7550	11
0.8	[6;6;6]	2.6134	25

TABLE II: Constant c_0 in (10) and required number N_0 to obtain a good estimate for different values of γ and x_0 in LQR, where N_0 is the smallest integer such that $\sup_z |F_x^K(z) - F_{x,N_0}^K(z)| \leq c_0 \gamma^{N_0} \leq 0.01$.

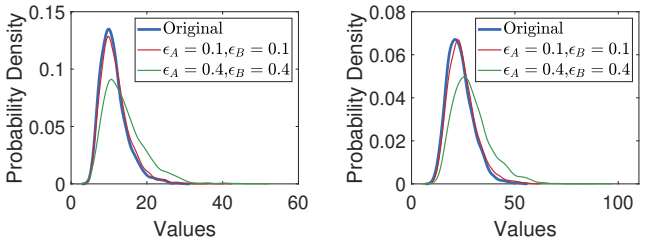


(a) $\gamma = 0.6$, $x_0 = [1; 1; 1]$. (b) $\gamma = 0.6$, $x_0 = [6; 6; 6]$.

(c) $\gamma = 0.8$, $x_0 = [1; 1; 1]$. (d) $\gamma = 0.8$, $x_0 = [6; 6; 6]$.

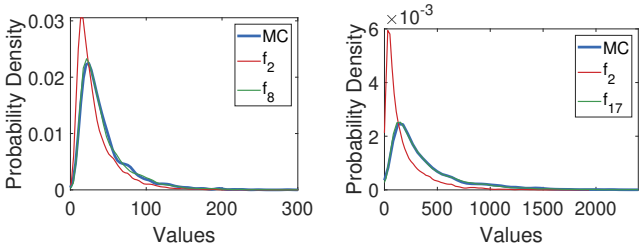
(e) $\gamma = 0.95$, $x_0 = [1; 1; 1]$. (f) $\gamma = 0.95$, $x_0 = [6; 6; 6]$.

Fig. 2: Return distribution and its approximation with finite number of random variables for different values of γ and x_0 in LQR. Alg. 1 denotes the distribution returned by Algorithm 1 and f_N denotes the distribution of the approximated random return $G_N^K(x_0)$.



(a) $\gamma = 0.6$. (b) $\gamma = 0.8$.

Fig. 3: Original and perturbed return distributions for different values of γ, ϵ_A and ϵ_B in LQR.



(a) $\gamma = 0.6$. (b) $\gamma = 0.8$.

Fig. 4: Return distribution and its approximation with finite number of random variables for different values of γ in LQG. MC denotes the distribution estimated using the Monte Carlo method and f_N denotes the distribution of the approximated random return $G_N^{KL}(\bar{x})$.

γ	ϵ_A	ϵ_B	$\tilde{c}_1$	$\tilde{c}_2$	Sup difference	UB
0.6	0.1	0.1	6.5	4.6	0.051	0.33
0.6	0.4	0.4	20.3	11.9	0.24	0.52
0.8	0.1	0.1	12.4	9.9	0.056	0.53
0.8	0.4	0.4	30.5	20.9	0.26	0.80

TABLE III: Computation of the actual maximal difference between the perturbed and the original distributions, and computational upper bound (UB) for different values of γ, ϵ_A and ϵ_B in LQR. The constants $\tilde{c}_1$ and $\tilde{c}_2$ are those in (8). Sup difference is the value of $\sup_z |F_x^K(z) - \tilde{F}_x^K(z)|$ while UB is the value of $\tilde{c}_1 \|\Delta A_K\| + \tilde{c}_2 \|\Delta A_K\|^2$.

when γ is close to 1. The perturbed return distributions for different values of ϵ_A and ϵ_B are shown in Fig. 3. We observe that large perturbations change the distributions dramatically.

B. LQG

In this section, we assume that the system is partially observable and we have the observation $y_t = Cx_t + s_t$, where $C = [1, 0, 0; 0, 1, 0]$. We assume that the disturbance s_t is normally distributed with zero mean. We design the state estimator and controller

$$\hat{x}_{t+1} = A\hat{x}_t + Bu_t + L(y_t - C\hat{x}_t),$$
$$u_t = K\hat{x}_t.$$

where the controller is selected the same as that in LQR and the observer is selected as $L = [0.21, 0.01; 0.01, 0.32; 0, 2.32]$. We set $x_0 = [1; 1; 1]$, $\hat{x}_0 = [0; 0; 0]$. The simulation results for LQG are presented in Fig. 4. Similarly, we denote by f_N the distribution of the approximated random return $G_N^{KL}(\bar{x})$ in (20) obtained based on N random variables. We use the Monte Carlo (MC) method with sufficiently many data to construct an empirical distribution that serves as the baseline distribution for comparison. As shown in Fig. 4, when $\gamma = 0.6$, we need $N = 8$ number of random variables to obtain a good approximation of the return distribution. When $\gamma = 0.8$, a greater number $N = 17$ is needed to achieve reliable approximation of the return distribution.

VI. Conclusions

We have proposed a new distributional approach to the classic discounted LQR problem. Specifically, we have first provided an analytic expression for the exact random return that depends on infinitely many random variables. In this context, we have shown that the variance remains bounded if the fourth moment of the disturbance is bounded. Furthermore, we have conducted an analysis of distribution sensitivity. Besides, we have proposed a model-free method for evaluating the return distribution, with theoretical analysis of its sample complexity. Since the computation of this expression is difficult in practice, we have also proposed an approximate expression for the distribution of the random return that only depends on a finite number of random variables, and have further characterised the approximation error. Moreover, we have extended most of the above results for LQR to the partially observable case.

This work provides a framework for distributional LQR: it inherits the advantages of DRL methods compared to standard RL ones that rely on the expected return to evaluate a given policy, but it also provides an analytic expression for the return distribution, an aspect where current DRL methods significantly lack. Our framework provides richer information for linear control systems, i.e., the whole distribution of the random return, and enables us to consider more general objectives, e.g., risk-averse control. Future research includes exploring policy improvement for risk-averse control using the learned return distribution.

Appendix

A. Proof of Theorem 2

By virtue of Jensen's Inequality, we have $\mathbb{E}^2[\|w_k\|^2] \leq \mathbb{E}[\|w_k\|^4]$ and $\mathbb{E}^2[\|w_k\|] \leq \mathbb{E}[\|w_k\|^2]$. Therefore, we have $\mathbb{E}[\|w_k\|^2] \leq \sigma_4^2$ and $\mathbb{E}[\|w_k\|] \leq \sigma_4$. Since $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$, we have

$$\begin{aligned}
&\mathbb{E}\Big[G^K(x)G^K(x)\Big] \\
&\leq 4\mathbb{E}\Big[(x^{\mathrm{T}}Px)^2 + \big(\sum_{k=0}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}Pw_k\big)^2 \\
&\quad + \big(2\sum_{k=0}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}PA_K^{k+1}x\big)^2 \\
&\quad + \big(2\sum_{k=1}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}P\sum_{\tau=0}^{k-1}A_K^{k-\tau}w_\tau\big)^2\Big]. \qquad (22)
\end{aligned}$$

We handle the terms one by one. The first term can be easily bounded by

$$(x^{\mathrm{T}}Px)^2 \le \|x\|^4 \|P\|^2 . \tag{23}$$

By virtue of the Cauchy Product that $\left(\sum_{k=0}^{\infty} a_k\right)^2 = \sum_{k=0}^{\infty}\sum_{l=0}^{k} a_l a_{k-l}$, the second term can be bounded by

$$\begin{aligned}
&\mathbb{E}\Big[\big(\sum_{k=0}^{\infty}\gamma^{k+1} w_k^{\mathrm{T}} P w_k\big)^2\Big] \\
&\le \mathbb{E}\Big[\big(\sum_{k=0}^{\infty}\gamma^{k+1}\|w_k\|^2\|P\|\big)^2\Big] \\
&= \|P\|^2 \mathbb{E}\Big[\sum_{k=0}^{\infty}\gamma^{k+2}\sum_{l=0}^{k}\|w_l\|^2\|w_{k-l}\|^2\Big] \\
&= \|P\|^2 \sum_{k=0}^{\infty}\gamma^{k+2}\sum_{l=0}^{k}\mathbb{E}\Big[\|w_l\|^2\|w_{k-l}\|^2\Big] \\
&\le \|P\|^2 \sum_{k=0}^{\infty}\gamma^{k+2}(k+1)\sigma_4^4 \\
&= \sigma_4^4\|P\|^2\frac{\gamma^2}{(1-\gamma)^2}.
\end{aligned} \tag{24}$$

The first inequality holds since $w_k P w_k \ge 0$. The second inequality holds since $\mathbb{E}[\|w_l\|^2\|w_{k-l}\|^2] = \mathbb{E}[\|w_l\|^2]\mathbb{E}[\|w_{k-l}\|^2] \le \sigma_4^4$ when $l \ne k-l$ and $\mathbb{E}[\|w_l\|^2\|w_{k-l}\|^2] = \mathbb{E}[\|w_l\|^4] \le \sigma_4^4$ when $l = k-l$. The last equality holds since $\sum_{k=0}^{\infty}\gamma^{k+2}(k+1) = \frac{\gamma^2}{(1-\gamma)^2}$.

For the third term, we have

$$\begin{aligned}
&\mathbb{E}\Big[\big(2\sum_{k=0}^{\infty}\gamma^{k+1} w_k^{\mathrm{T}} P A_K^{k+1} x\big)^2\Big] \\
&= 4\mathbb{E}\Big[\sum_{k=0}^{\infty}\sum_{l=0}^{k}\gamma^{l+1} w_l^{\mathrm{T}} P A_K^{l+1} x \gamma^{k-l+1} w_{k-l}^{\mathrm{T}} P A_K^{k-l+1} x\Big] \\
&\le 4\|P\|^2\|x\|^2\mathbb{E}\Big[\sum_{k=0}^{\infty}\gamma^{k+2}\sum_{l=0}^{k}\|w_l\|\big\|A_K^{l+1}\big\|\|w_{k-l}\| \\
&\quad\times\big\|A_K^{k-l+1}\big\|\Big] \\
&\le 4\|P\|^2\|x\|^2\sum_{k=0}^{\infty}(\gamma\rho)^{k+2}\sum_{l=0}^{k}\mathbb{E}\Big[\|w_l\|\|w_{k-l}\|\Big] \\
&\le 4\|P\|^2\|x\|^2\sigma_4^2\sum_{k=0}^{\infty}(k+1)(\gamma\rho_K)^{k+2} \\
&= 4\|P\|^2\|x\|^2\sigma_4^2\frac{\gamma^2\rho_K^2}{(1-\gamma\rho_K)^2}.
\end{aligned} \tag{25}$$

The first equality follows from the Cauchy Product. The first inequality follows from $w_l^{\mathrm{T}} P A_K^l x \le \|P\|\big\|A_K^l\big\|\|w_l\|\|x\|$ and $w_{k-l}^{\mathrm{T}} P A_K^{k-l+1} x \le \|P\|\big\|A_K^{k-l+1}\big\|\|w_{k-l}\|\|x\|$. The second inequality holds since $\big\|A_K^l\big\| \le \|A_K\|^l \le \rho_K^l$. The third inequality holds since $\mathbb{E}[\|w_l\|\|w_{k-l}\|] \le \sigma_4^2$ when $l = k-l$ and $l \ne k-l$. The last equality holds since $\sum_{k=0}^{\infty}(\gamma\rho)^{k+2}(k+1) = \frac{\gamma^2\rho^2}{(1-\gamma\rho)^2}$.

For the fourth term, by virtue of the Cauchy Product, we have

$$\begin{aligned}
&\mathbb{E}\Big[\big(2\sum_{k=1}^{\infty}\gamma^{k+1} w_k^{\mathrm{T}} P\sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau\big)^2\Big] \\
&= \mathbb{E}\Big[\big(2\sum_{k=0}^{\infty}\gamma^{k+2} w_{k+1}^{\mathrm{T}} P\sum_{\tau=0}^{k} A_K^{k+1-\tau} w_\tau\big)^2\Big] \\
&= 4\mathbb{E}\Big[\sum_{k=0}^{\infty}\gamma^{k+4}\sum_{l=0}^{k} w_{l+1}^{\mathrm{T}} P\big(\sum_{\tau=0}^{l} A_K^{l+1-\tau} w_\tau\big) \\
&\quad\times w_{k-l+1}^{\mathrm{T}} P\big(\sum_{\tau=0}^{k-l} A_K^{k-l+1-\tau} w_\tau\big)\Big].
\end{aligned} \tag{26}$$

Let $\xi := w_{l+1}^{\mathrm{T}} P(\sum_{\tau=0}^{l} A_K^{l+1-\tau} w_\tau) \times w_{k-l+1}^{\mathrm{T}} P(\sum_{\tau=0}^{k-l} A_K^{k-l+1-\tau} w_\tau)$. Recall that the random variables w_k are independent from each other and $\mathbb{E}[w_k] = 0$ for all $k \in \mathbb{N}$. It yields that when $l > k-l$, $\mathbb{E}[\xi] = 0$, and when $l < k-l$, $\mathbb{E}[\xi] = 0$. Thus, (26) can be simplified to be with the items when $k = 2l$, i.e.,

$$\begin{aligned}
&\mathbb{E}\Big[\big(2\sum_{k=1}^{\infty}\gamma^{k+1} w_k^{\mathrm{T}} P\sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau\big)^2\Big] \\
&= 4\mathbb{E}\Big[\sum_{k=0}^{\infty}\gamma^{k+4}\sum_{l=0}^{k} w_{l+1}^{\mathrm{T}} P\big(\sum_{\tau=0}^{l} A_K^{l+1-\tau} w_\tau\big) \\
&\quad\times w_{k-l+1}^{\mathrm{T}} P\big(\sum_{\tau=0}^{k-l} A_K^{k-l+1-\tau} w_\tau\big)\Big] \\
&= 4\mathbb{E}\Big[\sum_{l=0}^{\infty}\gamma^{2l+4}\big(w_{l+1}^{\mathrm{T}} P\big(\sum_{\tau=0}^{l} A_K^{l+1-\tau} w_\tau\big)\big)^2\Big] \\
&\le 4\|P\|^2\sum_{l=0}^{\infty}\gamma^{2l+4}\mathbb{E}\Big[\|w_{l+1}\|^2\Big]\mathbb{E}\Big[\Big\|\sum_{\tau=0}^{l} A_K^{l+1-\tau} w_\tau\Big\|^2\Big] \\
&\le 4\|P\|^2\sigma_4^2\sum_{l=0}^{\infty}\gamma^{2l+4}\mathbb{E}\Big[\Big\|\sum_{\tau=0}^{l} A_K^{l+1-\tau} w_\tau\Big\|^2\Big] \\
&\le 4\|P\|^2\sigma_4^2\sum_{l=0}^{\infty}\gamma^{2l+4}\mathbb{E}\Big[\big(\sum_{\tau=0}^{l} w_\tau^{\mathrm{T}}(A_K^{l+1-\tau})^{\mathrm{T}}\big) \\
&\quad\times\big(\sum_{\tau=0}^{l} A_K^{l+1-\tau} w_\tau\big)\Big] \\
&= 4\|P\|^2\sigma_4^2\mathbb{E}\Big[\sum_{l=0}^{\infty}\gamma^{2l+4}\sum_{\tau=0}^{l}\sum_{\kappa=0}^{l}\big(w_\tau^{\mathrm{T}}(A_K^{l+1-\tau})^{\mathrm{T}} A_K^{l+1-\kappa} w_\kappa\big)\Big] \\
&= 4\|P\|^2\sigma_4^2\sum_{l=0}^{\infty}\gamma^{2l+4}\sum_{\tau=0}^{l}\sum_{\kappa=0}^{l}\mathbb{E}\Big[\big(w_\tau^{\mathrm{T}}(A_K^{l+1-\tau})^{\mathrm{T}} A_K^{l+1-\kappa} w_\kappa\big)\Big] \\
&= 4\|P\|^2\sigma_4^2\sum_{l=0}^{\infty}\gamma^{2l+4}\sum_{\tau=0}^{l}\mathbb{E}\Big[\big(w_\tau^{\mathrm{T}}(A_K^{l+1-\tau})^{\mathrm{T}}\times A_K^{l+1-\tau} w_\tau\big)\Big] \\
&\le 4\|P\|^2\sigma_4^2\sum_{l=0}^{\infty}\gamma^{2l+4}\sum_{\tau=0}^{l}\rho_K^{2(l+1-\tau)}\mathbb{E}\Big[\|w_\tau\|^2\Big] \\
&\le 4\|P\|^2\sigma_4^4\sum_{l=0}^{\infty}\gamma^{2l+4}\sum_{\tau=0}^{l}\rho_K^{2(l+1-\tau)}
\end{aligned}$$

$$\leq 4\left\|P\right\|^2\sigma_4^4\sum_{l=0}^{\infty}\gamma^{2l+4}\frac{\rho_K^2}{1-\rho_K^2}$$
$$\leq \frac{4\left\|P\right\|^2\sigma_4^4\rho_K^2\gamma^4}{(1-\rho_K^2)(1-\gamma^2)}. \tag{27}$$

The first inequality holds since w_{l+1} and w_τ are independent for all $\tau=0,\ldots,l$. The last equality holds since w_τ and w_κ are independent for $\tau\neq\kappa$ and $\mathbb{E}[w_\tau]=0$ for all $\tau=0,\ldots,l$. The second to last inequality holds since $\sum_{\tau=0}^{l}\rho_K^{2(l+1-\tau)}\leq\frac{\rho_K^2}{1-\rho_K^2}$.

Combining (23), (24), (25), (27) and (22), we have

$$\mathbb{E}\Big[G^K(x)G^K(x)\Big]$$
$$\leq 4\left\|x\right\|^4\left\|P\right\|^2+\frac{4\sigma_4^4\left\|P\right\|^2\gamma^2}{(1-\gamma)^2}+\frac{16\left\|P\right\|^2\left\|x\right\|^2\sigma_4^2\gamma^2\rho_K^2}{(1-\gamma\rho_K)^2}$$
$$+\frac{16\left\|P\right\|^2\sigma_4^4\rho_K^2\gamma^4}{(1-\rho_K^2)(1-\gamma^2)}.$$

Since $\mathbb{E}\big[G^K(x)\big]$ is bounded, the variance of $G^K(x)$ is bounded. The proof is complete.

B. Proof of Theorem 3

Before analyzing the effect of perturbations on the return distribution, it is necessary to investigate how perturbations affect the solution to the Lyapunov equation. The following lemma presents the well-known sensitivity result for LQR.

Lemma 1. *[35] Let X be the unique solution of the Lyapunov equation $X=Q+A^{\mathrm{T}}XA$ for a stable matrix A. Let $\tilde{X}$ be the unique solution of the perturbed Lyapunov equation $\tilde{X}=Q+\tilde{A}^{\mathrm{T}}\tilde{X}\tilde{A}$ for a stable matrix $\tilde{A}=A+\Delta A$. Then, when $l_0>2\epsilon_0$, where $l_0=\left\|H_0^{-1}\right\|^{-1}$, $H_0=I\otimes I-A^{\mathrm{T}}\otimes A^{\mathrm{T}}$, and $\epsilon_0=\left\|A\right\|_F\left\|\Delta A\right\|_F+\frac{1}{2}\left\|\Delta A\right\|_F^2$, we have*

$$\left\|X-\tilde{X}\right\|_F\leq\frac{2\left\|X\right\|_F\epsilon_0}{l_0-2\epsilon_0}.$$

Lemma 1 analyzes the sensitivity of the Lyapunov equation for the canonical form of LQR. For discounted LQR, the sensitivity analysis of the Lyapunov equation is presented in the following lemma.

Lemma 2. *Let $l=\left\|H^{-1}\right\|^{-1}$, $H=I\otimes I-\gamma A_K^{\mathrm{T}}\otimes A_K^{\mathrm{T}}$, $\epsilon=\gamma\left\|A_K\right\|_F\left\|\Delta A_K\right\|_F+\frac{\gamma}{2}\left\|\Delta A_K\right\|_F^2$. If $l>2\epsilon$, we have*

$$\left\|P-\tilde{P}\right\|\leq\left\|P-\tilde{P}\right\|_F\leq\frac{2\left\|P\right\|_F\epsilon}{l-2\epsilon}. \tag{28}$$

Proof. It directly follows from Lemma 1 and $\left\|M\right\|_2\leq\left\|M\right\|_F\leq\sqrt{n}\left\|M\right\|_2$, for any matrix $M\in\mathbb{R}^{n\times n}$. □

Back to the sensitivity of perturbations on the return distribution, we define $\tilde{Y}:=G^K(x)-\tilde{G}^K(x)$, we have

$$\sup_z\left|F_x^K(z)-\tilde{F}_x^K(z)\right|$$
$$=\sup_z\left|\mathbb{P}(\tilde{G}^K(x)\leq z)-\mathbb{P}(G^K(x)\leq z)\right|$$
$$=\sup_z\left|\mathbb{P}(\tilde{G}^K(x)\leq z)-\mathbb{P}(\tilde{G}^K(x)+\tilde{Y}\leq z)\right|$$
$$=\sup_z\left|\mathbb{P}(\tilde{G}^K(x)\leq z)\int_{-\infty}^{\infty}\mathbb{P}(\tilde{Y}=t)dt\right.$$
$$\left.-\int_{-\infty}^{\infty}\mathbb{P}(\tilde{G}^K(x)\leq z-t)\mathbb{P}(\tilde{Y}=t)dt\right|$$
$$=\sup_z\left|\int_{-\infty}^{\infty}\mathbb{P}(\tilde{Y}=t)\big(\tilde{F}_x^K(z)-\tilde{F}_x^K(z-t)\big)dt\right|$$
$$\leq\sup_z\left|\int_{-\infty}^{\infty}\mathbb{P}(\tilde{Y}=t)\tilde{f}_{\max}|t|dt\right|$$
$$=\tilde{f}_{\max}\mathbb{E}\big[|\tilde{Y}|\big], \tag{29}$$

where $\tilde{f}_{\max}$ is an upper bound of the PDF of $\tilde{G}^K(x)$ and the last inequality follows from the mean value theorem.

From the definition of $\tilde{Y}$, we have

$$\mathbb{E}|\tilde{Y}|=\mathbb{E}\Big[\Big|x^{\mathrm{T}}(P-\tilde{P})x+\sum_{k=0}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}(P-\tilde{P})w_k$$
$$+2\sum_{k=0}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}(PA_K^{k+1}-\tilde{P}\tilde{A}_K^{k+1})x$$
$$+2\sum_{k=1}^{\infty}\gamma^{k+1}w_k^{\mathrm{T}}\big(P\sum_{\tau=0}^{k-1}A_K^{k-\tau}w_\tau-\tilde{P}\sum_{\tau=0}^{k-1}\tilde{A}_K^{k-\tau}w_\tau\big)\Big|\Big]$$
$$\leq|x^{\mathrm{T}}(\tilde{P}-P)x|+\sum_{k=0}^{\infty}\gamma^{k+1}\mathbb{E}\Big[\Big|w_k^{\mathrm{T}}(\tilde{P}-P)w_k\Big|\Big]$$
$$+2\sum_{k=0}^{\infty}\gamma^{k+1}\mathbb{E}\Big[\Big|w_k^{\mathrm{T}}(\tilde{P}\tilde{A}_K^{k+1}-PA_K^{k+1})x\Big|\Big]$$
$$+2\sum_{k=1}^{\infty}\gamma^{k+1}\mathbb{E}\Big[\Big|w_k^{\mathrm{T}}\big(\tilde{P}\sum_{\tau=0}^{k-1}\tilde{A}_K^{k-\tau}-P\sum_{\tau=0}^{k-1}A_K^{k-\tau}\big)w_\tau\Big|\Big]. \tag{30}$$

We handle the terms in the above inequality one by one. By virtue of the Holder's inequality, the first term can be bounded by

$$|x^{\mathrm{T}}(\tilde{P}-P)x|\leq\left\|x\right\|^2\left\|\tilde{P}-P\right\|. \tag{31}$$

Similarly, the second term can be bounded by

$$\sum_{k=0}^{\infty}\gamma^{k+1}\mathbb{E}\Big[\Big|w_k^{\mathrm{T}}(\tilde{P}-P)w_k\Big|\Big]$$
$$\leq\sum_{k=0}^{\infty}\gamma^{k+1}\sigma^2\left\|\tilde{P}-P\right\|\leq\frac{\sigma^2\gamma}{1-\gamma}\left\|\tilde{P}-P\right\|. \tag{32}$$

For the third term, we have

$$2\sum_{k=0}^{\infty}\gamma^{k+1}\mathbb{E}\Big[\Big|w_k^{\mathrm{T}}(\tilde{P}\tilde{A}_K^{k+1}-PA_K^{k+1})x\Big|\Big]$$
$$=2\sum_{k=0}^{\infty}\gamma^{k+1}\mathbb{E}\Big[\Big|w_k^{\mathrm{T}}\big(\tilde{P}-P\big)\tilde{A}_K^{k+1}x$$
$$+w_k^{\mathrm{T}}P\big(\tilde{A}_K^{k+1}-A_K^{k+1}\big)x\Big|\Big]$$
$$\leq2\sum_{k=0}^{\infty}\gamma^{k+1}\mathbb{E}\Big[\left\|w_k\right\|\left\|\tilde{P}-P\right\|\left\|\tilde{A}_K^{k+1}\right\|\left\|x\right\|\Big]$$
$$+2\sum_{k=0}^{\infty}\gamma^{k+1}\mathbb{E}\Big[\left\|w_k\right\|\left\|P\right\|\left\|\tilde{A}_K^{k+1}-A_K^{k+1}\right\|\left\|x\right\|\Big]$$
$$\leq2\sigma\left\|\tilde{P}-P\right\|\left\|x\right\|\sum_{k=0}^{\infty}\gamma^{k+1}\left\|\tilde{A}_K^{k+1}\right\|$$

$$+ 2\sigma \|P\| \|x\| \sum_{k=0}^{\infty} \gamma^{k+1} \left\| \tilde{A}_K^{k+1} - A_K^{k+1} \right\|. \tag{33}$$

By decomposing the term $\tilde{A}_K^{k+1} - A_K^{k+1} = (\tilde{A}_K - A_K)(\sum_{i=0}^{k} \tilde{A}_K^{k-i} A_K^i)$, we have

$$\begin{aligned}
&\left\| \tilde{A}_K^{k+1} - A_K^{k+1} \right\| \\
&\le \left\| \tilde{A}_K - A_K \right\| \left\| \sum_{i=0}^{k} \tilde{A}_K^{k-i} A_K^i \right\| \\
&\le \left\| \tilde{A}_K - A_K \right\| (\sum_{i=0}^{k} \left\| \tilde{A}_K^{k-i} A_K^i \right\|) \\
&\le \|\Delta A_K\| (\sum_{i=0}^{k} \left\| \tilde{A}_K^{k-i} \right\| \|A_K^i\|) \\
&\le \|\Delta A_K\| (k+1) \rho_K^k \le \|\Delta A_K\| U.
\end{aligned} \tag{34}$$

Define the function $f(k) = (k+1)\rho_K^k$, and the constant $U = \max\{1, \frac{1}{\rho_0} \rho_K^{(1-\rho_0)/\rho_0}\}$, $\rho_0 = \ln(1/\rho_K)$. It is easy to verify that the function $f(k)$ obtains the maximum at $k = 0$ if $\rho_0 \ge 1$ and at $k = \frac{1-\rho_0}{\rho_0}$ if $\rho_0 < 1$. Therefore, the last inequality follows since $f(k) \le \max\{1, \frac{1}{\rho_0} \rho_K^{(1-\rho_0)/\rho_0}\} = U$ for all $k \ge 0$. Substituting (34) into (33), we have

$$\begin{aligned}
&2 \sum_{k=0}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\Big| w_k^{\mathrm{T}} (\tilde{P} \tilde{A}_K^{k+1} - P A_K^{k+1}) x \Big| \Big] \\
&\le \frac{2\sigma \|P\| \|x\| U \gamma}{1-\gamma} \|\Delta A_K\| \\
&\quad + 2\sigma \left\| \tilde{P} - P \right\| \|x\| \sum_{k=0}^{\infty} \gamma^{k+1} \left\| \tilde{A}_K^{k+1} \right\| \\
&\le \frac{2\sigma \|P\| \|x\| U \gamma}{1-\gamma} \|\Delta A_K\| + 2\sigma \left\| \tilde{P} - P \right\| \|x\| \sum_{k=0}^{\infty} (\gamma \rho_K)^{k+1} \\
&\le \frac{2\sigma \|P\| \|x\| U \gamma}{1-\gamma} \|\Delta A_K\| + \frac{2\sigma\gamma\rho_K}{1-\gamma\rho_K} \|x\| \left\| \tilde{P} - P \right\|,
\end{aligned} \tag{35}$$

where the second inequality follows since $\left\| \tilde{A}_K^{k+1} \right\| \le \left\| \tilde{A}_K \right\|^{k+1} \le \rho_K^{k+1}$. For the fourth term, we have

$$\begin{aligned}
&2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\Big| w_k^{\mathrm{T}} \big(\tilde{P} \sum_{\tau=0}^{k-1} \tilde{A}_K^{k-\tau} w_\tau - P \sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau \big) \Big| \Big] \\
&= 2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\Big| w_k^{\mathrm{T}} \Big(\tilde{P} \sum_{\tau=0}^{k-1} \tilde{A}_K^{k-\tau} w_\tau - P \sum_{\tau=0}^{k-1} \tilde{A}_K^{k-\tau} w_\tau \\
&\quad + P \sum_{\tau=0}^{k-1} \tilde{A}_K^{k-\tau} w_\tau - P \sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau \Big) \Big| \Big] \\
&\le 2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\Big| w_k^{\mathrm{T}} (\tilde{P} - P) \sum_{\tau=0}^{k-1} \tilde{A}_K^{k-\tau} w_\tau \Big| \Big] \\
&\quad + 2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\Big| w_k^{\mathrm{T}} P \sum_{\tau=0}^{k-1} (\tilde{A}_K^{k-\tau} - A_K^{k-\tau}) w_\tau \Big| \Big] \\
&\le 2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\|w_k\| \left\| \tilde{P} - P \right\| \left\| \sum_{\tau=0}^{k-1} \tilde{A}_K^{k-\tau} w_\tau \right\| \Big] \\
&\quad + 2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\|w_k\| \|P\| \left\| \sum_{\tau=0}^{k-1} (\tilde{A}_K^{k-\tau} - A_K^{k-\tau}) w_\tau \right\| \Big] \\
&\le 2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\|w_k\| \|P\| \sum_{\tau=0}^{k-1} \left\| \tilde{A}_K^{k-\tau} - A_K^{k-\tau} \right\| \|w_\tau\| \Big] \\
&\quad + 2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\|w_k\| \left\| \tilde{P} - P \right\| \sum_{\tau=0}^{k-1} \left\| \tilde{A}_K^{k-\tau} \right\| \|w_\tau\| \Big] \\
&\le 2\sigma^2 \|P\| \sum_{k=1}^{\infty} \gamma^{k+1} \sum_{\tau=0}^{k-1} \left\| \tilde{A}_K^{k-\tau} - A_K^{k-\tau} \right\| \\
&\quad + 2\sigma^2 \left\| \tilde{P} - P \right\| \sum_{k=1}^{\infty} \gamma^{k+1} \sum_{\tau=0}^{k-1} \left\| \tilde{A}_K^{k-\tau} \right\|,
\end{aligned} \tag{36}$$

where the last inequality follows since w_k and w_τ, $\tau = 0, 1, \ldots, k-1$ are independent. By decomposing the term $\tilde{A}_K^{k-\tau} - A_K^{k-\tau}$, we have

$$\begin{aligned}
&\sum_{\tau=0}^{k-1} \left\| \tilde{A}_K^{k-\tau} - A_K^{k-\tau} \right\| \\
&= \sum_{\tau=0}^{k-1} \left\| (\tilde{A}_K - A_K) \sum_{i=0}^{k-\tau-1} \tilde{A}_K^{k-\tau-1-i} A_K^i \right\| \\
&\le \sum_{\tau=0}^{k-1} \left\| \tilde{A}_K - A_K \right\| \left\| \sum_{i=0}^{k-\tau-1} \tilde{A}_K^{k-\tau-1-i} A_K^i \right\| \\
&\le \left\| \tilde{A}_K - A_K \right\| \sum_{\tau=0}^{k-1} \sum_{i=0}^{k-\tau-1} \rho_K^{k-\tau-1} \\
&= \|\Delta A_K\| \sum_{\tau=0}^{k-1} (k-\tau) \rho_K^{k-\tau-1} \\
&= \|\Delta A_K\| \sum_{i=1}^{k} i \rho_K^{i-1} \\
&\le \frac{\|\Delta A_K\|}{(1-\rho_K)^2},
\end{aligned} \tag{37}$$

where the last inequality follows since $S_k := \sum_{i=1}^{k} i\rho_K^{i-1} = \frac{1-\rho_K^k}{(1-\rho_K)^2} - \frac{k\rho_K^k}{1-\rho_K} \le \frac{1}{(1-\rho_K)^2}$. Substituting (37) into (36), we have

$$\begin{aligned}
&2 \sum_{k=1}^{\infty} \gamma^{k+1} \mathbb{E}\Big[\Big| w_k^{\mathrm{T}} \big(\tilde{P} \sum_{\tau=0}^{k-1} \tilde{A}_K^{k-\tau} w_\tau - P \sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau \big) \Big| \Big] \\
&\le 2\sigma^2 \|P\| \sum_{k=1}^{\infty} \gamma^{k+1} \frac{\|\Delta A_K\|}{(1-\rho_K)^2} \\
&\quad + 2\sigma^2 \left\| \tilde{P} - P \right\| \sum_{k=1}^{\infty} \gamma^{k+1} \sum_{\tau=0}^{k-1} \left\| \tilde{A}_K^{k-\tau} \right\| \\
&\le \frac{2\sigma^2 \|P\| \gamma^2}{(1-\gamma)(1-\rho_K)^2} \|\Delta A_K\| \\
&\quad + 2\sigma^2 \left\| \tilde{P} - P \right\| \sum_{k=1}^{\infty} \gamma^{k+1} \sum_{\tau=0}^{k-1} \rho^{k-\tau} \\
&\le \frac{2\sigma^2 \|P\| \gamma^2 \|\Delta A_K\|}{(1-\gamma)(1-\rho_K)^2} + \frac{2\sigma^2 \left\| \tilde{P} - P \right\|}{(1-\gamma)(1-\rho_K)}.
\end{aligned} \tag{38}$$

Combining (31), (32), (35), (38) and (30), we have

$$\begin{aligned}\mathbb{E}\big[|\tilde{Y}|\big] &\le \big(\|x\|^2+\frac{\sigma^2\gamma}{1-\gamma}\big)\left\|\tilde{P}-P\right\| \\ &\quad+\frac{2\sigma\|P\|\|x\|U\gamma}{1-\gamma}\|\Delta A_K\|+\frac{2\sigma\gamma\rho_K\|x\|}{1-\gamma\rho_K}\left\|\tilde{P}-P\right\| \\ &\quad+\frac{2\sigma^2\|P\|\gamma^2\|\Delta A_K\|}{(1-\gamma)(1-\rho_K)^2}+\frac{2\sigma^2\left\|\tilde{P}-P\right\|}{(1-\gamma)(1-\rho_K)} \\ &:= \tilde{c}_3\left\|\tilde{P}-P\right\|+\tilde{c}_4\|\Delta A_K\|, \end{aligned} \tag{39}$$

where

$$\begin{aligned}\tilde{c}_3 &= \|x\|^2+\frac{\sigma^2\gamma}{1-\gamma}+\frac{2\sigma\gamma\rho_K\|x\|}{1-\gamma\rho_K}+\frac{2\sigma^2}{(1-\gamma)(1-\rho_K)}, \\ \tilde{c}_4 &= \frac{2\sigma\|P\|\|x\|U\gamma}{1-\gamma}+\frac{2\sigma^2\|P\|\gamma^2}{(1-\gamma)(1-\rho_K)^2}.\end{aligned}$$

Substituting (39) into (29) and using (28), we have

$$\begin{aligned}&\sup_z|F_x^K(z)-\tilde{F}_x^K(z)| \\ &\le \tilde{f}_{\max}(\tilde{c}_3\left\|\tilde{P}-P\right\|+\tilde{c}_4\|\Delta A_K\|) \\ &\le \tilde{f}_{\max}\tilde{c}_3\frac{2\|P\|_F}{l-2\epsilon}\big(\gamma\|A_K\|_F\|\Delta A_K\|_F+\frac{\gamma}{2}\|\Delta A_K\|_F^2\big) \\ &\quad+\tilde{f}_{\max}\tilde{c}_4\|\Delta A_K\| \\ &\le \Big(\frac{2\tilde{f}_{\max}\tilde{c}_3\sqrt{n}\gamma\|A_K\|_F\|P\|_F}{l-2\epsilon}+\tilde{f}_{\max}\tilde{c}_4\Big)\|\Delta A_K\| \\ &\quad+\frac{\tilde{f}_{\max}\tilde{c}_3 n\gamma\|P\|_F}{l-2\epsilon}\|\Delta A_K\|^2 \\ &:= \tilde{c}_1\|\Delta A_K\|+\tilde{c}_2\|\Delta A_K\|^2,\end{aligned}$$

where the last inequality follows from $\|M\|_F \le \sqrt{n}\|M\|_2$ for any matrix $M\in\mathbb{R}^{n\times n}$. The proof is complete and also yields the expression of the constants $\tilde{c}_1,\tilde{c}_2$.

C. Proof of Theorem 5

It follows that

$$\begin{aligned}&\sup_z|\hat{F}_{x,M}^{K,T}(z)-F_x^K(z)| \\ &\le \sup_z|\hat{F}_{x,M}^{K,T}(z)-F_x^{K,T}(z)|+\sup_z|F_x^{K,T}(z)-F_x^K(z)|.\end{aligned} \tag{40}$$

Note that $\hat{F}_{x,M}^{K,T}$ and $F_x^{K,T}$ are the EDF and CDF of the random variable $G^{K,T}(x)$, respectively. By virtue of the Dvoretzky–Kiefer–Wolfowitz inequality, we have

$$\sup_z|\hat{F}_M^{K,T}(z)-F^{K,T}(z)| \le \sqrt{\frac{\ln(1/\delta)}{2M}}, \tag{41}$$

with probability at least $1-\delta$. Define the random variable $Z_T = G^K(x)-G^{K,T}(x) = \sum_{T+1}^{\infty}\gamma^t x_t^{\mathrm{T}}(Q+K^{\mathrm{T}}RK)x_t$ Further, we have

$$\begin{aligned}&\sup_z|F_x^{K,T}(z)-F_x^K(z)| \\ &= \sup_z\Big|\mathbb{P}\{G^{K,T}(x)\le z\}-\mathbb{P}\{G^K(x)\le z\}\Big| \\ &= \sup_z\Big|\mathbb{P}\{G^{K,T}(x)\le z\}-\mathbb{P}\{G^{K,T}(x)+Z_T\le z\}\Big| \\ &= \sup_z\Big|\mathbb{P}(G^{K,T}(x)\le z)\int_{-\infty}^{\infty}\mathbb{P}(Z_T=t)dt \\ &\quad-\int_{-\infty}^{\infty}\mathbb{P}(G^{K,T}(x)\le z-t)\mathbb{P}(Z_T=t)dt\Big| \\ &= \sup_z\Big|\int_{-\infty}^{\infty}\mathbb{P}(Z_T=t)\big(F_x^{K,T}(z)-F_x^{K,T}(z-t)\big)dt\Big| \\ &\le \sup_z\Big|\int_{-\infty}^{\infty}\mathbb{P}(Z_T=t)f_{\max}|t|dt\Big| \\ &= f_{\max}\mathbb{E}\big[|Z_T|\big].\end{aligned} \tag{42}$$

Now we focus on $\mathbb{E}|Z_T|$. From its definition, it gives

$$\begin{aligned}\mathbb{E}\big[|Z_T|\big] &= \mathbb{E}\Big[\sum_{t=T+1}^{\infty}\gamma^t x_t^{\mathrm{T}}(Q+K^{\mathrm{T}}RK)x_t\Big] \\ &\le \mathbb{E}\Big[\sum_{t=T+1}^{\infty}\gamma^t\big\|Q+K^{\mathrm{T}}RK\big\|\|x_t\|^2\Big].\end{aligned} \tag{43}$$

From $x_{t+1}=A_Kx_t+w_t$, where $A_K=A+BK$, we have

$$x_t = A_K^t x+\sum_{\tau=0}^{t-1}A_K^{t-1-\tau}w_\tau. \tag{44}$$

Hence,

$$\begin{aligned}\mathbb{E}[\|x_t\|^2] &= \mathbb{E}\Big[\Big\|A_K^tx+\sum_{\tau=0}^{t-1}A_K^{t-1-\tau}w_\tau\Big\|^2\Big] \\ &\le \mathbb{E}\Big[\big\|A_K^tx\big\|^2+\Big\|\sum_{\tau=0}^{t-1}A_K^{t-1-\tau}w_\tau\Big\|^2 \\ &\quad+2\big\|A_K^tx\big\|\Big\|\sum_{\tau=0}^{t-1}A_K^{t-1-\tau}w_\tau\Big\|\Big] \\ &\le \rho_K^{2t}\|x\|^2+\mathbb{E}\Big[\Big\|\sum_{\tau=0}^{t-1}A_K^{t-1-\tau}w_\tau\Big\|^2 \\ &\quad+2\rho_K^t\|x\|\Big\|\sum_{\tau=0}^{t-1}A_K^{t-1-\tau}w_\tau\Big\|\Big],\end{aligned} \tag{45}$$

where the last inequality follows from $\|A_K^tx\|\le\|A_K^t\|\|x\|\le\rho_K^t\|x\|$. Further, we have

$$\begin{aligned}\mathbb{E}\left[\Big\|\sum_{\tau=0}^{t-1}A_K^{t-1-\tau}w_\tau\Big\|\right] &\le \mathbb{E}\left[\sum_{\tau=0}^{t-1}\big\|A_K^{t-1-\tau}\big\|\|w_\tau\|\right] \\ \le \sigma\sum_{\tau=0}^{t-1}\big\|A_K^{t-1-\tau}\big\| &\le \sigma\sum_{\tau=0}^{t-1}\rho_K^{t-1-\tau} \le \frac{\sigma}{1-\rho_K},\end{aligned} \tag{46}$$

where the first inequality follows from the Cauchy–Schwarz inequality, the second inequality follows from $\mathbb{E}^2[\|w_k\|]\le\mathbb{E}[\|w_k\|^2]\le\sigma^2$, and the third inequality follows from $\big\|A_K^{t+N+1}\big\|\le(\|A_K\|)^{t+N+1}\le\rho_K^{t+N+1}$. Further,

$$\begin{aligned}&\mathbb{E}\left[\Big\|\sum_{\tau=0}^{t-1}A_K^{t-1-\tau}w_\tau\Big\|^2\right] \\ &= \mathbb{E}\left[\sum_{\tau=0}^{t-1}w_\tau^{\mathrm{T}}(A_K^{t-1-\tau})^{\mathrm{T}}A_K^{t-1-\tau}w_\tau\right]\end{aligned}$$

$$\leq \mathbb{E}\left[\sum_{\tau=0}^{t-1} \left\|(A_K^{t-1-\tau})^{\mathrm{T}} A_K^{t-1-\tau}\right\| \|w_\tau\|^2\right]$$
$$\leq \sigma^2 \sum_{\tau=0}^{t-1} \left\|(A_K^{t-1-\tau})^{\mathrm{T}} A_K^{t-1-\tau}\right\|$$
$$\leq \sigma^2 \sum_{\tau=0}^{t-1} \rho_K^{2(t-1-\tau)} \leq \frac{\sigma^2}{1-\rho_K^2}, \tag{47}$$

where the first equality follows from the fact that the random variables w_τ, $\tau \in \mathbb{N}$, are i.i.d. and with zero mean. The first inequality follows from the Cauchy-Schwarz inequality and the third inequality follows from $\left\|(A_K^{t-1-\tau})^{\mathrm{T}} A_K^{t-1-\tau}\right\| \leq \left\|(A_K^{t-1-\tau})^{\mathrm{T}}\right\| \left\|A_K^{t-1-\tau}\right\| \leq \rho_K^{2(t-1-\tau)}$. Substituting (46) and (47) into (45), we have

$$\mathbb{E}[\|x_t\|^2] \leq \rho_K^{2t} \|x\|^2 + \frac{\sigma^2}{1-\rho_K^2} + \frac{2\rho_K^t \|x\| \sigma}{1-\rho_K}. \tag{48}$$

Substituting (48) into (43), we have

$$\begin{aligned}
&\mathbb{E}\big[|Z_T|\big] \\
&\leq \left\|Q + K^{\mathrm{T}} R K\right\| \sum_{t=T+1}^{\infty} \gamma^t \Big(\rho_K^{2t} \|x\|^2 + \frac{\sigma^2}{1-\rho_K^2} \\
&\quad + \frac{2\rho_K^t \|x\| \sigma}{1-\rho_K}\Big) \\
&\leq \left\|Q + K^{\mathrm{T}} R K\right\| \Big(\frac{\gamma^{T+1} \rho_K^{2(T+1)} \|x\|^2}{1-\gamma\rho_K^2} \\
&\quad + \frac{2\gamma^{T+1}\rho_K^{T+1} \|x\| \sigma}{(1-\rho_K)(1-\gamma\rho_K)} + \frac{\gamma^{T+1}\sigma^2}{(1-\gamma)(1-\rho_K^2)}\Big) \\
&= \|Q_K\| \gamma^{T+1} (c_1 \rho_K^{2(T+1)} + c_2 \rho_K^{T+1} + c_3).
\end{aligned} \tag{49}$$

Combining (40), (41), (42) and (49), it gives

$$\begin{aligned}
&\sup_z |\hat{F}_M^{K,T}(z) - F^K(z)| \\
\leq &\sup_z |\hat{F}_M^{K,T}(z) - F^{K,T}(z)| + \sup_z |F^{K,T}(z) - F^K(z)| \\
\leq &f_{\max} \|Q_K\| \gamma^{T+1} (c_1 \rho_K^{2(T+1)} + c_2 \rho_K^{T+1} + c_3) + \sqrt{\frac{\ln(1/\delta)}{2M}},
\end{aligned}$$

which completes the proof.

REFERENCES

[1] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[2] Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva Tb, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018.

[3] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[4] Rahul Singh, Keuntaek Lee, and Yongxin Chen. Sample-based distributional policy gradient. In *Proceedings of Learning for Dynamics and Control Conference*, pages 676–688. PMLR, 2022.

[5] Danial Kamran, Tizian Engelgeh, Marvin Busch, Johannes Fischer, and Christoph Stiller. Minimizing safety interference for safe and comfortable automated driving with distributional reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1236–1243. IEEE, 2021.

[6] Jianyi Zhang and Paul Weng. Safe distributional reinforcement learning. In *Distributed Artificial Intelligence: Third International Conference, DAI 2021, Shanghai, China, December 17–18, 2021, Proceedings 3*, pages 107–128. Springer, 2022.

[7] Hao Liang and Zhi-Quan Luo. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *arXiv preprint arXiv:2210.14051*, 2022.

[8] Kyushik Min, Hayoung Kim, and Kunsoo Huh. Deep distributional reinforcement learning based high-level driving policy determination. *IEEE Transactions on Intelligent Vehicles*, 4(3):416–424, 2019.

[9] Shuyuan Xu, Qiao Liu, Yuhui Hu, Mengtian Xu, and Jiachen Hao. Decision-making models on perceptual uncertainty with distributional reinforcement learning. *Green Energy and Intelligent Transportation*, 2(2):100062, 2023.

[10] Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

[11] Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *arXiv preprint arXiv:2301.04462*, 2023.

[12] Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. In *Proceedings of Learning for Dynamics and Control Conference*, pages 958–968. PMLR, 2020.

[13] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

[14] Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *Proceedings of International Conference on Machine Learning*, pages 5005–5014. PMLR, 2018.

[15] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

[16] 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 *Proceedings of 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925. PMLR, 2019.

[17] Yingying Li, Yujie Tang, Runyu Zhang, and Na Li. Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach. *IEEE Transactions on Automatic Control*, 67(12):6429–6444, 2021.

[18] Farnaz Adib Yaghmaie, Fredrik Gustafsson, and Lennart Ljung. Linear quadratic control using model-free reinforcement learning. *IEEE Transactions on Automatic Control*, 68(2):737–752, 2022.

[19] Yang Zheng, Luca Furieri, Maryam Kamgarpour, and Na Li. Sample complexity of linear quadratic Gaussian (LQG) control for output feedback systems. In *Proceedings of Learning for Dynamics and Control Conference*, pages 559–570. PMLR, 2021.

[20] Bart PG Van Parys, Daniel Kuhn, Paul J Goulart, and Manfred Morari. Distributionally robust control of constrained stochastic systems. *IEEE Transactions on Automatic Control*, 61(2):430–442, 2015.

[21] Anastasios Tsiamis, Dionysios S Kalogerias, Alejandro Ribeiro, and George J Pappas. Linear quadratic control with risk constraints. *arXiv preprint arXiv:2112.07564*, 2021.

[22] Kihyun Kim and Insoon Yang. Distributional robustness in minimax linear quadratic control with Wasserstein distance. *arXiv preprint arXiv:2102.12715*, 2021.

[23] Margaret P Chapman, Riccardo Bonalli, Kevin M Smith, Insoon Yang, Marco Pavone, and Claire J Tomlin. Risk-sensitive safety analysis using conditional value-at-risk. *IEEE Transactions on Automatic Control*, 67(12):6521–6536, 2021.

[24] Margaret P Chapman, Michael Fauß, and Kevin M Smith. On optimizing the conditional value-at-risk of a maximum cost for risk-averse safety analysis. *IEEE Transactions on Automatic Control*, 2022.

[25] Margaret P Chapman and Laurent Lessard. Toward a scalable upper bound for a CVaR-LQ problem. *IEEE Control Systems Letters*, 6:920–925, 2021.

[26] Margaret P Chapman and Dionysios S Kalogerias. Risk-aware stability of discrete-time systems. *arXiv preprint arXiv:2211.12416*, 2022.

[27] Masako Kishida and Ahmet Cetinkaya. Risk-aware linear quadratic control using conditional value-at-risk. *IEEE Transactions on Automatic Control*, 68(1):416–423, 2022.

[28] Kihyun Kim and Insoon Yang. Distributional robustness in minimax linear quadratic control with wasserstein distance. *SIAM Journal on Control and Optimization*, 61(2):458–483, 2023.

[29] Astghik Hakobyan and Insoon Yang. Wasserstein distributionally robust control of partially observable linear stochastic systems. *arXiv preprint arXiv:2212.04644*, 2022.

[30] Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 66(8):3863–3870, 2020.

[31] Bahar Taşkesen, Dan A Iancu, Çağıl Koçyiğit, and Daniel Kuhn. Distributionally robust linear quadratic control. *arXiv preprint arXiv:2305.17037*, 2023.

[32] Zifan Wang, Yulong Gao, Siyi Wang, Michael M Zavlanos, Alessandro Abate, and Karl Henrik Johansson. Policy evaluation in distributional LQR. In *Learning for Dynamics and Control Conference*, pages 1245–1256. PMLR, 2023.

[33] Karl J Åström. *Introduction to Stochastic Control Theory*. Courier Corporation, 2012.

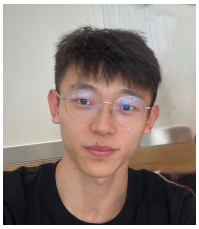
[34] Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

[35] Pascal M. Gahinet, Charles S. Kenney, and Gary A. Hewer. Sensitivity of the stable discrete-time lyapunov equation. *IEEE Transactions on Automatic Control*, 35:1209–1217, 1990.

[36] Dimitri Bertsekas. *Dynamic Programming and Optimal Control: Volume II*, volume 4. Athena scientific, 2012.

[37] Andrew Lamperski. Computing stabilizing linear controllers via policy iteration. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1902–1907. IEEE, 2020.

[38] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.



Zifan Wang received the B.S. and M.S. degrees in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2019 and 2021, respectively. He is currently working toward the Ph.D. degree in electrical engineering with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden.

His current research interests lie in decision making under uncertainty and generative models.

Yulong Gao received the B.E. degree in Automation in 2013, the M.E. degree in Control Science and Engineering in 2016, both from Beijing Institute of Technology, and the joint Ph.D. degree in Electrical Engineering in 2021 from KTH Royal Institute of Technology and Nanyang Technological University. He was a Researcher at KTH from 2021 to 2022 and a postdoctoral researcher at Oxford from 2022 to 2023. He is an Assistant Professor at the Department of Electrical and Electronic Engineering, Imperial College London, from 2024. His research interests include formal verification and control, machine learning, and applications to safety-critical systems.

Siyi Wang received the B.S. degree in electrical engineering and the M.S. degree in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2017 and 2019, respectively. She is currently working toward the Ph.D. degree within the School of Computation, Information and Technology, Technical University of Munich, Munich, Germany.

Her current research interests lie in networked control, value of information optimal control and risk-averse learning.

Michael M. Zavlanos received the Diploma in mechanical engineering from the National Technical University of Athens, Greece, in 2002, and the M.S.E. and Ph.D. degrees in electrical and systems engineering from the University of Pennsylvania, Philadelphia, PA, in 2005 and 2008, respectively.

He is currently the Yoh Family Professor of the Department of Mechanical Engineering and Materials Science at Duke University, Durham, NC. He also holds a secondary appointment in the Department of Electrical and Computer Engineering and the Department of Computer Science. His research focuses on control theory, optimization, and learning with applications in robotics and autonomous systems, cyber-physical systems, and healthcare/medicine. Dr. Zavlanos is a recipient of various awards including the 2014 ONR YIP Award and the 2011 NSF CAREER Award.

Alessandro Abate (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, Berkeley, CA, USA. He is Professor of Verification and Control with the Department of Computer Science, University of Oxford, Oxford, UK.

Karl Henrik Johansson (Fellow, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering from Lund University, Lund, Sweden, in 1992 and 1997, respectively. He is a Professor with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. He has held visiting positions at UC Berkeley, Caltech, NTU, HKUST Institute of Advanced Studies, and NTNU. His research interests include networked control systems, cyber-physical systems, and applications in transportation, energy, and automation. He has served on the IEEE Control Systems Society Board of Governors, the IFAC Executive Board, and the European Control Association Council.

Dr. Johansson was a recipient of several best paper awards and other distinctions from IEEE and ACM. He has been awarded Distinguished Professor with the Swedish Research Council and Wallenberg Scholar with the Knut and Alice Wallenberg Foundation. He was also a recipient of the Future Research Leader Award from the Swedish Foundation for Strategic Research and the Triennial Young Author Prize from IFAC. He is Fellow of the Royal Swedish Academy of Engineering Sciences, and he is IEEE Control Systems Society Distinguished Lecturer.