Policy Evaluation in Distributional LQR

Zifan Wang[1] ZIFANW@KTH.SE
Yulong Gao[2] YULONG.GAO@CS.OX.AC.UK
Siyi Wang[3] WANGSIYI199604@GMAIL.COM
Michael M. Zavlanos[4] MICHAEL.ZAVLANOS@DUKE.EDU
Alessandro Abate[2] ALESSANDRO.ABATE@CS.OX.AC.UK
Karl H. Johansson[1] KALLEJ@KTH.SE
[1] *Division of Decision and Control Systems, KTH Royal Institute of Technology, Sweden*
[2] *Department of Computer Science, University of Oxford, UK*
[3] *Chair of Information-oriented Control, Technical University of Munich, Germany*
[4] *Department of Mechanical Engineering and Materials Science, Duke University, USA*

Editors: N. Matni, M. Morari, G. J. Pappas

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 RL. At the same time, a main challenge in DRL is that policy evaluation in DRL typically relies on the representation of the return distribution, which needs to be carefully designed. In this paper, we address this challenge for a special class of DRL problems that rely on discounted linear quadratic regulator (LQR) for control, advocating for a new distributional approach to LQR, which we call *distributional LQR*. Specifically, we provide a closed-form expression of the distribution of the random return which, remarkably, is applicable to all exogenous disturbances on the dynamics, as long as they are independent and identically distributed (i.i.d.). While the proposed exact return distribution consists of infinitely many random variables, we show that this distribution can be approximated by a finite number of random variables, and the associated approximation error can be analytically bounded under mild assumptions. Using the approximate return distribution, we propose a zeroth-order policy gradient algorithm for risk-averse LQR using the Conditional Value at Risk (CVaR) as a measure of risk. Numerical experiments are provided to illustrate our theoretical results.

Keywords: Distributional LQR, distributional RL, policy evaluation, risk-averse control

1. Introduction

In reinforcement learning, 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, Bellemare et al. (2017) 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 (Bellemare et al., 2017), D4PG (Barth-Maron et al., 2018), QR-DQN (Dabney et al., 2018b) and SDPG (Singh et al., 2022), dramatically improve the empirical performance in practical applications over their non-distributional counterpart.

In DRL, the practical effectiveness of algorithms builds on the theory by Bellemare et al. (2017), where the distributional Bellman operator is shown to be a contraction in the (maximum form of) the Wasserstein metric between probability distributions. However, it is usually difficult to characterise the exact return distribution in DRL with finite data. Approximations of the return distribution are thus necessary to make it computable in practice. To address this challenge, Bellemare et al. (2017) propose a categorical method that partitions the return distribution into a finite number of uniformly spaced atoms in a fixed region. One drawback of this method is that it relies on prior knowledge of the range of the returned values. To address this limitation, a quantile function method (Dabney et al., 2018b) and a sample-based method (Singh et al., 2022) have been recently proposed. However, these works cannot provide an analytical expression for the approximation error, and computational cost needs to be decided manually to guarantee approximation accuracy.

In this paper, we characterise the return distribution of the random cost for the classical discounted linear quadratic regulator (LQR) problem, which we term *distributional LQR*. To the best of our knowledge, the return distribution in LQR has not been explored in the literature. Our contributions are summarised as follows:

1. We provide an analytical expression of the random return for distributional LQR problems and prove that this return function is a fixed-point solution of the random variable Bellman equation. 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.

2. We develop an approximation of the distribution of the random return using a finite number of random variables. Under mild assumptions, we theoretically show that the sup of the difference between the exact and approximated return distributions deceases linearly with the numbers of random variables: this is also validated by numerical experiments.

3. The proposed analytical return distribution provides a theoretical foundation for distributional LQR, allowing for general optimality criteria for policy improvement. In this work, we employ the return distribution to analyse risk-averse LQR problems using the Conditional Value at Risk (CVaR) as the risk measure. Since the gradient of CVaR is generally difficult to compute analytically, we propose a risk-averse policy gradient algorithm that relies on the zeroth-order optimisation to seek an optimal risk-averse policy. Numerical experiments are provided to showcase this application.

Related Work: Most closely related to the problem considered in this paper is work on reinforcement learning for LQR, which focuses on learning the expected return through interaction with the environment; see, e.g., Dean et al. (2020); Tu and Recht (2018); Fazel et al. (2018); Malik et al. (2019); Li et al. (2021); Yaghmaie et al. (2022); Zheng et al. (2021). For example, Fazel et al. (2018) propose a model-free policy gradient algorithm for LQR and showed its global convergence with finite polynomial computational and sample complexity. Moreover, Zheng et al. (2021) study model-based reinforcement learning for the Linear Quadratic Gaussian problems, in which a model is first learnt from data and then used to design the policy. However, all these works rely on the expected return instead of the return distribution, hence these methods cannot be applied here.

Since the return distribution captures the intrinsic randomness of the long-term cost, it provides a natural framework to consider more general optimality criteria, e.g., optimal risk-averse policies. There exist recent works on risk-averse policy design for DRL, including Singh et al. (2020);

Dabney et al. (2018a); Tang et al. (2019). For example, the work in Dabney et al. (2018a) use the quantile function to approximate the return distribution, which is then applied to design risk-sensitive policies for Atari games. On the other hand, Singh et al. (2020) show that risk-averse DRL achieves robustness against system disturbances in continuous control tasks. All these works focus on empirical improvements in specific tasks, however, without theoretical analysis. Related to this paper is also work on risk-sensitive LQR, which has been studied in Van Parys et al. (2015); Tsiamis et al. (2021); Kim and Yang (2021); Chapman and Lessard (2021); Kishida and Cetinkaya (2022). Similarly, these methods however do not analyse the return distribution.

2. Problem Statement

Consider a discrete-time linear dynamical system:

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

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^p$, $v_t \in \mathbb{R}^n$ 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.

2.1. Classical LQR

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

$$J(u) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t (x_t^T Q x_t + u_t^T R u_t)\right], \tag{2}$$

where Q, R are positive-definite constant matrices and $\gamma \in (0, 1)$ is a discount parameter. Given a control policy π, let $V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^k (x_t^T Q x_t + u_t^T R u_t)\right]$ 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

$$V^\pi(x) = x^T(Q + K^T R K)x + \gamma \mathop{\mathbb{E}}_{X' = (A+BK)x + v_0}[V^\pi(X')], \tag{3}$$

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

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

2.2. 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 reinforcement learning, which relies on expected returns, DRL (Bellemare et al., 2023) relies on the distribution of random returns. 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

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

It is straightforward to see that the expectation of $G^\pi(x)$ is equivalent to the value function $V^\pi(x)$. The standard Bellman equation in (3) 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

$$G^\pi(x) \overset{D}{=} x^T Q x + \pi(x)^T R \pi(x) + \gamma G^\pi(X'), \quad X' = Ax + B\pi(x) + v_0. \tag{5}$$

Here we use the notation $\overset{D}{=}$ to denote that two random variables Z_1, Z_2 are equal in distribution, i.e., $Z_1 \overset{D}{=} Z_2$. Note that X' denotes a random variable, as in (3). Compared to the expected return in LQR, which is a scalar, here the return distribution is infinite-dimensional and can have a complex form. It is challenging to estimate an infinite-dimensional function exactly with finite data and thus an approximation of the return distribution is necessary in practice.

In this paper, we first analytically characterise the random return for the LQR problem. Then we show how to approximate the distribution of the random return using finite random variables, so that the approximated distribution is computationally tractable and the approximation error is bounded. The proposed distributional LQR framework allows us to consider more general optimality criteria, which we demonstrate by using the proposed return distribution to develop a policy gradient algorithm for risk-averse LQR.

3. Main Results

3.1. Exact Form of the Return Distribution

In this section, we precisely characterise the distribution of the random return that satisfies the distributional Bellman equation (5). 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

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

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

$$G^K(x) \overset{D}{=} x^T Q_K x + \gamma G^K(X'), \quad X' = A_K x + v_0, \tag{6}$$

where $A_K := A + BK$ and $Q_K := Q + K^T R K$. In the following theorem, we provide an explicit expression of the random return $G^K(x)$.

Theorem 1 *Suppose that the feedback gain K is stabilizing, i.e., $A_K = A + BK$ is stable. Let*

$$G^K(x) = x^T P x + \sum_{k=0}^{\infty} \gamma^{k+1} w_k^T P w_k + 2 \sum_{k=0}^{\infty} \gamma^{k+1} w_k^T P A_K^{k+1} x + 2 \sum_{k=1}^{\infty} \gamma^{k+1} w_k^T P \sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau, \tag{7}$$

where P is obtained from the Lyapunov equation $P = Q + K^T RK + \gamma A_K^T P A_K$, 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 (7) is a fixed point solution to the random variable Bellman equation (6).

Proof Recall that $X' = A_K x + v_0$, where v_0 is a random variable sampled from the distribution $\mathcal{D}$ and is independent from w_k, $k \in \mathbb{N}$, in (7). Substituting (7) into the right hand side of the equation (6), we have that

$$
\begin{aligned}
&x^T(Q + K^T RK)x + \gamma G^K(X') \\
=&x^T Q_K x + \gamma X'^T P X' + \sum_{t=0}^{\infty} \gamma^{t+2} w_t^T P w_t + 2\sum_{t=0}^{\infty} \gamma^{t+2} w_t^T P A_K^{t+1} X' \\
&+ 2\sum_{t=1}^{\infty} \gamma^{t+2} w_t^T P A_K \sum_{i=0}^{t-1} A_K^{t-1-i} w_i \\
=&x^T Q_K x + \gamma(A_K x + v_0)^T P(A_K x + v_0) + \gamma^2 \sum_{t=0}^{\infty} \gamma^t w_t^T P w_t + 2\gamma^2 \sum_{t=1}^{\infty} \gamma^t w_t^T P \sum_{i=0}^{t-1} A_K^{t-i} w_i \\
&+ 2\gamma^2 \sum_{t=0}^{\infty} \gamma^t w_t^T P A_K^{t+1}(A_K x + v_0) \\
=&x^T(Q_K + \gamma A_K^T P A_K)x + \underbrace{\gamma v_0^T P v_0 + \gamma^2 \sum_{t=0}^{\infty} \gamma^t w_t^T P w_t}_{:=T_1} + \underbrace{2\gamma v_0^T P A_K x + 2\gamma^2 \sum_{t=0}^{\infty} \gamma^t w_t^T P A_K^{t+2} x}_{:=T_2} \\
&\underbrace{+ 2\gamma^2 \sum_{t=1}^{\infty} \gamma^t w_t^T P \sum_{i=0}^{t-1} A_K^{t-i} w_i + 2\gamma^2 \sum_{t=0}^{\infty} \gamma^t w_t^T P A_K^{t+1} v_0}_{:=T_3}.
\end{aligned}
$$

Define $\xi_0 := v_0$, $\xi_t = w_{t-1}$, $t = 1, 2, \ldots$. From the definition of the term T_1, we have that

$$
T_1 = \gamma v_0^T P v_0 + \gamma^2 \sum_{t=0}^{\infty} \gamma^t w_t^T P w_t \overset{k=t+1}{=} \gamma \xi_0^T P \xi_0 + \gamma \sum_{k=1}^{\infty} \gamma^k \xi_k^T P \xi_k = \gamma \sum_{k=0}^{\infty} \gamma^k \xi_k^T P \xi_k.
$$

For the term T_2, we have that

$$
\begin{aligned}
T_2 &= 2\gamma v_0^T P A_K x + 2\gamma^2 \sum_{t=0}^{\infty} \gamma^t w_t^T P A_K^{t+2} x = 2\gamma \xi_0^T P A_K x + 2\gamma^2 \sum_{t=0}^{\infty} \gamma^t \xi_{t+1}^T P A_K^{t+2} x \\
&\overset{k=t+1}{=} 2\gamma \xi_0^T P A_K x + 2\gamma \sum_{k=1}^{\infty} \gamma^k \xi_k^T P A_K^{k+1} x = 2\gamma \sum_{k=0}^{\infty} \gamma^k \xi_k^T P A_K^{k+1} x.
\end{aligned}
$$

Using similar techniques for the next term, we obtain that $T_3 = 2\gamma \sum_{k=1}^{\infty} \gamma^k \xi_k^T P A_K \sum_{i=0}^{k-1} A_K^{k-1-i} \xi_i$. Due to the fact that $P = Q + K^T RK + \gamma A_K^T P A_K$, we have

$$
\begin{aligned}
&x^T Q_K x + \gamma G^K(X') = x^T P x + T_1 + T_2 + T_3 \\
=&x^T P x + \gamma \sum_{k=0}^{\infty} \gamma^k \xi_k^T P \xi_k + 2\gamma \sum_{k=0}^{\infty} \gamma^k x^T P A_K^{k+1} \xi_k + 2\gamma \sum_{k=1}^{\infty} \gamma^k \xi_k^T P A_K \sum_{i=0}^{k-1} A_K^{k-1-i} \xi_i, \qquad (8)
\end{aligned}
$$

which is in the same form as in (7). Since $\{\xi_k\}_{k=0}^{\infty}$ and $\{w_k\}_{k=0}^{\infty}$ are i.i.d., we have that the two random variables (7) and (8) have the same distribution, i.e., $G^K(x) \overset{D}{=} x^T Q_K x + \gamma G^K(X')$. ■

3.2. Approximation of the Return Distribution with Finite Parameters

In this section, we show how to approximate the random return defined in (7) using a finite number of random variables. Considering only the first N terms in the summations in the expression (7) and disregarding the terms for k larger than N yields the following:

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

Let F_x^K and $F_{x,N}^K$ denote the cumulative distribution function (CDF) of $G^K(x)$ and $G_N^K(x)$, respectively. 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 $G^K(x)$, $\forall x \in \mathbb{R}^n$.

Theorem 2 *Assume that the probability density functions of w_k exist and are bounded, and satisfy $\mathbb{E}[w_k^T w_k] \le \sigma_0^2$, for $\forall k \in \mathbb{N}$. Suppose that the feedback gain K is stabilizing such that $\|A_K\|_2 = \rho_K < 1$. Then, the sup difference between the CDFs F_x^K and $F_{x,N}^K$ is bounded by*

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

where C is a constant that depends on the matrices A, B, Q, R, K, the initial state value x, and the parameters γ, ρ_K, σ_0.

Proof Define $Y_N := G^K(x) - G_N^K(x)$, we have

$$\begin{aligned}
&\sup_z |F_x^K(z) - F_{x,N}^K(z)| = \sup_z |\mathbb{P}(G_N^K(x) \le z) - \mathbb{P}(G^K(x) \le z)| \\
=&\sup_z |\mathbb{P}(G_N^K(x) \le z) - \mathbb{P}(G_N^K(x) + Y_N \le z)| \\
=&\sup_z \Big|\mathbb{P}(G_N^K(x) \le z)\int_{-\infty}^{\infty} \mathbb{P}(Y_N = t)dt - \int_{-\infty}^{\infty} \mathbb{P}(G_N^K(x) \le z - t)\mathbb{P}(Y_N = t)dt\Big| \\
=&\sup_z \Big|\int_{-\infty}^{\infty} \mathbb{P}(Y_N = t)\big(F_{x,N}^K(z) - F_{x,N}^K(z-t)\big)dt\Big|.
\end{aligned} \tag{11}$$

Since the random variables w_t are i.i.d for all $t > 0$ and the probability density function of w_t exists, the function $F_{x,N}^K$ is continuous and differentiable. Applying the mean value theorem, when $t > 0$ there exists a point $z' \in [z-t, z]$ such that $F_{x,N}^K(z) - F_{x,N}^K(z-t) = f_{x,N}^K(z')t$, where $f_{x,N}^K$ is the probability density function of $G_N^K(x)$. Since the probability density function of w_t is bounded, it further follows that $f_{x,N}^K$ is bounded. Then, we have that $|F_{x,N}^K(z) - F_{x,N}^K(z-t)| = |f_{x,N}^K(z')t| \le L_0|t|$, where L_0 is an upper bound of the probability function $f_{x,N}^K$. Following a similar argument, we can show that this inequality holds when $t \le 0$. Substituting this inequality into (11), we obtain

$$\sup_z |F_x^K(z) - F_{x,N}^K(z)| \le \sup_z \Big|\int_{-\infty}^{\infty} \mathbb{P}(Y_N = t)L_0|t|dt\Big| = L_0\mathbb{E}|Y_N|. \tag{12}$$

From the definition of Y_N, we obtain that

$$\begin{aligned}
Y_N &= \sum_{k=N}^{\infty} \gamma^{k+1} w_k^T P w_k + 2\sum_{k=N}^{\infty} \gamma^{k+1} w_k^T P A_K^{k+1} x + 2\sum_{k=N}^{\infty} \gamma^{k+1} w_k^T P \sum_{\tau=0}^{k-1} A_K^{k-\tau} w_\tau \\
&\overset{t=k-N}{=} \gamma^N \Big(\sum_{t=0}^{\infty} \gamma^{t+1} w_{t+N}^T P w_{t+N} + 2\sum_{t=0}^{\infty} \gamma^{t+1} w_{t+N}^T P A_K^{t+N+1} x \\
&\quad + 2\sum_{t=0}^{\infty} \gamma^{t+1} w_{t+N}^T P \sum_{\tau=0}^{t+N-1} A_K^{t+N-\tau} w_\tau \Big).
\end{aligned}$$

Taking the expectation of the absolute value of Y_N, we have

$$\begin{aligned}
\mathbb{E}|Y_N| \leq& \gamma^N \Big(\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}|w_{t+N}^T P w_{t+N}| + 2\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}|w_{t+N}^T P A_K^{t+N+1} x| \\
&+ 2\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}|w_{t+N}^T P \sum_{\tau=0}^{t+N-1} A_K^{t+N-\tau} w_\tau| \Big).
\end{aligned}$$

We handle the terms in the above inequality one by one. For the first term, we have that

$$\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}|w_{t+N}^T P w_{t+N}| \leq \sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}|\lambda_{\max}(P) w_{t+N}^T w_{t+N}| \leq \lambda_{\max}(P)\sigma_0^2 \frac{\gamma}{1-\gamma}. \tag{13}$$

By virtue of Jensen's Inequality, it gives $\mathbb{E}^2[\|w_k\|_2] \leq \mathbb{E}[\|w_k\|_2^2] \leq \sigma_0^2$. Then, for the second term, we have

$$\begin{aligned}
&2\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}|w_{t+N}^T P A_K^{t+N+1} x| \leq 2\sigma_0 \sum_{t=0}^{\infty} \gamma^{t+1} \|P\|_2 \left\|A_K^{t+N+1}\right\|_2 \|x\|_2 \\
\leq& 2\sigma_0 \sum_{t=0}^{\infty} \gamma^{t+1} \|P\|_2 \rho_K^{t+N-1} \|x\|_2 \leq 2\sigma_0 \|P\|_2 |x| \frac{\gamma \rho_K^{N-1}}{1-\gamma\rho_K} \leq 2\sigma_0 \|P\|_2 \|x\|_2 \frac{\gamma}{1-\gamma\rho_K},
\end{aligned} \tag{14}$$

where the second inequality is due to the fact that $\left\|A_K^{t+N+1}\right\|_2 \leq (\|A_K\|_2)^{t+N+1} \leq \rho_K^{t+N+1}$ and the last inequality follows from the fact that $N \geq 1$. For the third term, we have that

$$\begin{aligned}
&2\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}|w_{t+N}^T P \sum_{\tau=0}^{t+N-1} A_K^{t+N-\tau} w_\tau| \leq 2\sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}\left[\left\|w_{t+N}^T\right\|_2 \|P\|_2 \left\| \sum_{\tau=0}^{t+N-1} A_K^{t+N-\tau} w_\tau \right\|_2 \right] \\
\leq& 2\sigma_0 \|P\|_2 \sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}\left[\left\| \sum_{\tau=0}^{t+N-1} A_K^{t+N-\tau} w_\tau \right\|_2 \right] \leq 2\sigma_0 \|P\|_2 \sum_{t=0}^{\infty} \gamma^{t+1} \mathbb{E}\left[\sum_{\tau=0}^{t+N-1} \left\|A_K^{t+N-\tau}\right\|_2 \|w_\tau\|_2 \right] \\
\leq& 2\sigma_0^2 \|P\|_2 \sum_{t=0}^{\infty} \gamma^{t+1} \sum_{\tau=0}^{t+N-1} \rho_K^{t+N-\tau} \leq 2\sigma_0^2 \|P\|_2 \sum_{t=0}^{\infty} \gamma^{t+1} \frac{\rho_K}{1-\rho_K} \leq 2\sigma_0^2 \|P\|_2 \frac{\gamma\rho_K}{(1-\gamma)(1-\rho_K)},
\end{aligned} \tag{15}$$

where the second inequality is due to the fact that w_τ and w_{t+N} are independent and the second to last inequality follows from the fact that $\sum_{\tau=0}^{t+N-1} \rho_K^{t+N-\tau} = \sum_{\tau=1}^{t+N} \rho_K^{\tau} \leq \frac{\rho_K}{1-\rho_K}$. Combining (13), (14) and (15), we have that

$$\begin{aligned}
&\sup_z |F_x^K(z) - F_{x,N}^K(z)| \leq L_0 \mathbb{E}|Y_N| \\
\leq & L_0 \gamma^N \left(\lambda_{\max}(P) \sigma_0^2 \frac{\gamma}{1-\gamma} + 2\sigma_0 \|P\|_2 \|x\|_2 \frac{\gamma}{1-\gamma\rho_K} + 2\sigma_0^2 \|P\|_2 \frac{\gamma\rho_K}{(1-\gamma)(1-\rho_K)} \right) := C\gamma^N.
\end{aligned}$$

The proof is complete and also yields the expression of the constant C. ■

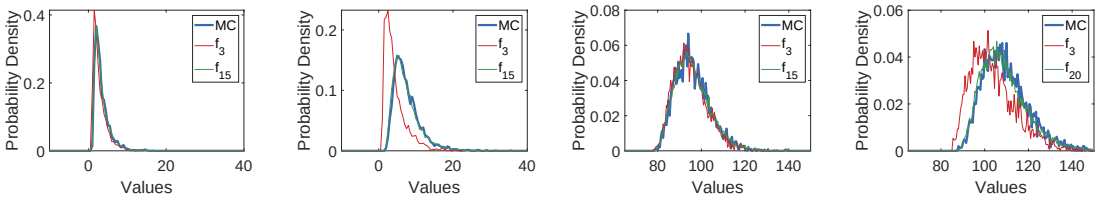
Remark 3 *The bound on the distribution approximation in (10) relies on the conditions of Theorem 2, which ensure that the PDF of G_N^K is continuous and bounded. Note that these conditions are not strict, and indeed hold for many noise distributions commonly used in linear dynamical systems, including Gaussian and uniform. Future work will investigate relaxations of these conditions.*

3.3. Numerical Experiments on Quality of the Approximation of the Return Distribution

In the following experiment, we consider a scalar model with matrices $A = B = 1$. Similarly, the weighting matrices in the LQR cost are chosen as $Q = R = 1$. The exogenous disturbances are standard normal distributions with zero mean.

Even for this scalar 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 that approximates the true distribution of the random return. More specifically, we use the Monte Carlo (MC) method to obtain 10000 samples of the random return and use the sample frequency over evenly-divided regions as an approximation of the probability density function. According to the law of large numbers, the empirical distribution approaches the real one as the number of trials increases. Note that, although the MC method provides an alternative way to approximate the return distribution, it relies on using sufficiently many samples that can be time-consuming, and its (statistical) approximation error is generally difficult to analyse. Thus, the MC method is not applicable for practical policy evaluation of distributional LQR, and in this experiment, it is used only to verify our approximate return distribution. In comparison, the approximate return distribution using finite number of random variables in this paper is analytical for policy evaluation and the corresponding approximation error can be bounded: as such, it is further usable for policy optimisation, as shown in Section 4. We denote here by f_N the distribution of the approximated random return $G_N^K(x_0)$ obtained considering N random variables.

We fix the feedback gain as $K = -0.4684$ and select different values of γ and x_0. The results are shown in Fig. 1. Specifically, Fig. 1 (a) and (c) show that when γ is small, the return distribution can be well approximated using only few random variables ($N = 3$ works well). However, when γ approaches 1, more random variables are needed for an accurate approximation: we employ $N = 15$ and $N = 20$ random variables in the case of $\gamma = 0.8$ and $\gamma = 0.85$, respectively, as shown in Fig. 1 (b) and (d). Moreover, the value of the initial state x_0 has an influence on the shape of the return distribution, which can be clearly observed from the scalar case. 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. 1 (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. 1 (a) and (b).



(*a*) $\gamma = 0.6$, $x_0 = 1$. (*b*) $\gamma = 0.8$, $x_0 = 1$. (*c*) $\gamma = 0.6$, $x_0 = 8$. (*d*) $\gamma = 0.85$, $x_0 = 8$.

Figure 1: Return distribution and its approximation with finite number of random variables for different γ and x_0. MC denotes the distribution returned by the Monte Carlo method and f_N denotes the distribution of the approximated random return $G_N^K(x_0)$.

Algorithm 1 Risk-Averse Policy Gradient

Require: initial values K_0, x, step size η, smoothing parameter δ, and dimension n

1: **for** *episode* $t = 1, \dots, T$ **do**
2: Sample $\hat{K}_t = K_t + U_t$, where U_t is drawn at random over matrices whose norm is δ;
3: Compute the distribution of the random variable $G_N^{\hat{K}_t}$;
4: Compute $\hat{\mathcal{C}}_N(\hat{K}_t)$;
5: $K_{t+1} = K_t - \eta g_t$, where $g_t = \frac{n}{\delta^2}\Big(\hat{\mathcal{C}}_N(\hat{K}_t) - \hat{\mathcal{C}}_N(\hat{K}_{t-1})\Big)U_t$.
6: **end for**

In conclusion, when N is large, the approximate distribution is closer to the distribution obtained from the MC method, and thus to the true distribution.

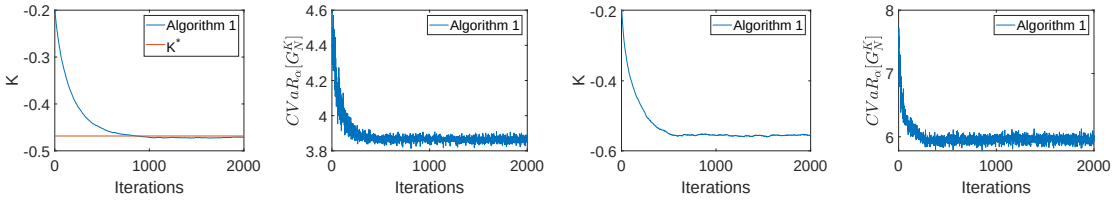
4. Application to Risk-Averse LQR

In this section, we consider a risk-averse LQR problem and leverage the closed-form expression of the random return $G^K(x)$ to obtain an optimal policy. Since the distribution of the random return $G^K(x)$ consists of an infinite number of random variables, it is computationally unwieldy. Instead, we employ the approximate random return $G_N^K(x)$ proposed in Section 3.2. As a risk measure for the problem at hand, we select the well-known Conditional Value at Risk (CVaR) (Rockafellar et al., 2000). We then construct an approximate risk-averse objective function, as $\hat{\mathcal{C}}_N(K) := \mathrm{CVaR}_\alpha\left[G_N^K(x)\right]$. For a random variable Z with the CDF F and a risk level $\alpha \in (0, 1]$, the CVaR value is defined as $\mathrm{CVaR}_\alpha[Z] = \mathbb{E}_F[Z|Z > Z^\alpha]$, where Z^α is the $1 - \alpha$ quantile of the distribution of the random variable Z. Given this objective function, the goal is to find the optimal risk-averse controller, that is, to select the feedback gain K that minimises $\hat{\mathcal{C}}_N(K)$.

4.1. Risk-Averse Policy Gradient Algorithm

In what follows, we propose a policy gradient method to solve this problem. We assume that the matrices A, B, Q, R are known. The first-order gradient descent step is hard to compute as it hinges on the gradient of the CVaR function. Therefore, we rely on zeroth-order optimisation to derive the policy gradient, as detailed in Algorithm 1.

Specifically, at each episode t, we sample an approximate feedback gain $\hat{K}_t = K_t + U_t$, where U_t is drawn uniformly at random from the set of matrices with norm δ. Given $\hat{K}_t$, we compute the



(*a*) The K values when $\alpha = 1$. (*b*) The CVaR values when $\alpha = 1$. (*c*) The K values when $\alpha = 0.4$. (*d*) The CVaR values when $\alpha = 0.4$.

Figure 2: Risk-averse control using Algorithm 1. The solid lines are averages over 20 runs.

approximate distribution of the random return $G_N^{\hat{K}_t}(x)$ in (9) and the value of $\hat{\mathcal{C}}_N(\hat{K}_t)$. Then, we can perform the feedback gain update as $K_{t+1} = K_t - \eta g_t$, where $g_t = \frac{n}{\delta^2}\Big(\hat{\mathcal{C}}_N(\hat{K}_t) - \hat{\mathcal{C}}_N(\hat{K}_{t-1})\Big)U_i$. Here, the zeroth-order residual feedback technique proposed in Zhang et al. (2022) is used to reduce the variance. The theoretical analysis of this algorithm is left as our future work.

4.2. Numerical Experiments

Next, we consider a risk-averse LQR problem and experimentally illustrate the performance of Algorithm 1. We illustrate our approach for the same scalar system with the same cost function as in Section 3.3. The other parameters are selected as $\gamma = 0.6$, $\delta = 0.1$, $\eta = 0.0004$, $N = 10$, respectively. The initial controller is set as $K_0 = -0.2$, which is a stable one.

We first set $\alpha = 1$: in this case, the risk-averse control problem is reduced to a risk-neutral control problem. Therefore, we can use traditional LQR techniques to compute the optimal feedback gain $K^* = -0.468$. We run the proposed risk-averse policy gradient Algorithm 1 and the simulation results are presented in Fig. 2 (a) and (b). Specifically, in Fig. 2 (a), the controller K returned by Algorithm 1 converges to K^*, which verifies our proposed method for the risk-neutral case. Fig. 2 (b) illustrates the values of CVaR achieved by Algorithm 1. Additionally, we select $\alpha = 0.4$ to find the optimal risk-averse controller. The simulation results are presented in Fig. 2 (c) and (d). We see that K converges to -0.55, which leads to a smaller $A + BK$ compared to $K^* = -0.468$.

5. Conclusions

We have proposed a new distributional approach to the classic discounted LQR problem. Specifically, we first provided an analytic expression for the exact random return that depends on infinitely many random variables. Since the computation of this expression is difficult in practice, we 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 error between these two distributions. Finally, we utilised the proposed random return to obtain an optimal controller for a risk-averse LQR problem using the CVaR as a measure of risk. To the best of our knowledge, this is a first framework for distributional LQR: it inherits the advantages of DRL methods compared to standard RL methods that rely on the expected return to evaluate the effect of a given policy, but it also provides an analytic expression for the return distribution, an area where current DRL methods significantly lack. Future research includes analyzing the theoretical convergence of risk-averse policy-gradient algorithms and exploring a model-free setup where the system matrices are unknown.

Acknowledgments

This work is supported in part by the Knut and Alice Wallenberg Foundation, the Swedish Strategic Research Foundation, the Swedish Research Council, AFOSR under award #FA9550-19-1-0169, and NSF under award CNS-1932011.

References

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.

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.

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

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

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of International Conference on Machine Learning*, pages 1096–1105. PMLR, 2018a.

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, 2018b.

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.

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.

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

Masako Kishida and Ahmet Cetinkaya. Risk-aware linear quadratic control using conditional value-at-risk. *IEEE Transactions on Automatic Control*, 2022.

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*, 2021.

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.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

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.

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.

Yichuan Charlie Tang, Jian Zhang, and Ruslan Salakhutdinov. Worst case policy gradients. *arXiv preprint arXiv:1911.03618*, 2019.

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

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.

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.

Farnaz Adib Yaghmaie, Fredrik Gustafsson, and Lennart Ljung. Linear quadratic control using model-free reinforcement learning. *IEEE Transactions on Automatic Control*, 2022.

Yan Zhang, Yi Zhou, Kaiyi Ji, and Michael M Zavlanos. A new one-point residual-feedback oracle for black-box learning and control. *Automatica*, 136:110006, 2022.

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.