Interacting Particle Systems for Fast Linear Quadratic RL

Anant A. Joshi ANANTAJ2@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Heng-Sheng Chang HSCHANG2@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Amirhossein Taghvaei AMIRTAG@UW.EDU
University of Washington Seattle

Prashant G. Mehta MEHTAPG@ILLINOIS.EDU
University of Illinois Urbana-Champaign

Sean P. Meyn MEYN@ECE.UFL.EDU
University of Florida at Gainesville

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

This paper is concerned with the design of algorithms based on systems of interacting particles to represent, approximate, and learn the optimal control law for reinforcement learning (RL). The primary contribution is that convergence rates are greatly accelerated by the interactions between particles. The focus is on the linear quadratic stochastic optimal control problem for which a complete and novel theory is presented. Apart from the new algorithm, sample complexity bounds are obtained, and it is shown that the mean square error scales as $1/N$ where N is the number of particles. The theoretical results and algorithms are illustrated with numerical experiments and comparisons with other recent approaches, where the faster convergence of the proposed algorithm is numerically demonstrated.

Keywords: Reinforcement learning, Linear quadratic control, Ensemble kalman filter

1. Introduction

This paper[1] concerns approaches to reinforcement learning (RL) based on the construction of interacting particle systems. The development is in continuous time, and the state is assumed to evolve according to a linear stochastic differential equation (SDE),

$$\mathrm{d}X_t = (AX_t + BU_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \quad X_0 = x \tag{1}$$

where $X := \{X_t : 0 \le t \le T\}$ is the $\mathbb{R}^d$-valued state process, $U := \{U_t : 0 \le t \le T\}$ is the $\mathbb{R}^m$-valued control input, and $W := \{W_t : 0 \le t \le T\}$ is a standard Brownian motion (B.M.), and A, B, σ are matrices of appropriate dimensions.

The proposed approach is related to actor-only methods, also known as the policy optimization (PO) approach, of which Williams' REINFORCE algorithm is most classical Williams (1992). The linear model is the subject of recent work in PO: two types of optimal control objectives have been considered, namely, linear quadratic Gaussian (LQG) (Basei et al., 2022) linear exponential quadratic Gaussian (LEQG) (Zhang et al., 2021b; Roulet et al., 2020), and average cost versions of

1. An extended version of this paper with additional discussion around the technical details appears at Joshi et al. (2024).

these (Krauth et al., 2019; Abbasi-Yadkori et al., 2019; Yang et al., 2019; Cassel and Koren, 2021; Yaghmaie et al., 2023; Hernández-Hernández and Salazar-Sánchez, 2023).

A standard PO approach in the linear quadratic setting is special because the policies $\{\kappa^\theta : \theta \in \mathbb{R}^n\}$ may be chosen deterministic and linear. A basic algorithm is described as the following recursion: Starting from an initial stabilizing gain K^0, a sequence of gains $\{K^j : j = 1, 2, \dots, M\}$ are learnt. During the j-th iteration, the gain K^j is evaluated by simulating N copies of the model over a time-horizon:

$$\mathrm{d}X_t^i = (AX_t^i + BK_t^j X_t^i)\mathrm{d}t + \sigma \mathrm{d}W_t^i, \quad 0 \le t \le T, \quad 1 \le i \le N \tag{2a}$$

$$X_0^i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I), \quad 1 \le i \le N \tag{2b}$$

These evaluations are helpful to compute the gain K^{j+1} through a gradient-descent procedure.

When $N = 1$ the algorithm might be applied using observed samples from a physical system; otherwise, this technique requires a simulator to generate particles. The main message of this paper is that the use of a simulator combined with carefully designed mean-field interactions between simulations (the particles) will ensure far greater efficiency in the learning process.

Algorithm proposed in this paper. Simulate an interacting particle system:

$$\mathrm{d}Y_t^i = \underbrace{AY_t^i\mathrm{d}t + B\mathrm{d}\eta_t^i + \sigma\mathrm{d}W_t^i}_{\text{copy of model}} + \underbrace{\mathcal{A}_t(Y_t^i; p_t^{(N)})\mathrm{d}t}_{\text{mean-field interaction}}, \quad 0 \le t \le T, \quad 1 \le i \le N \tag{3a}$$

$$Y_T^i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathcal{Y}), \quad 1 \le i \le N, \tag{3b}$$

where $p_t^{(N)}$ is the empirical distribution of the ensemble $\{Y_t^i : 1 \le i \le N\}$. The specification of the terminal condition at time $t = T$ means that the system is simulated backward-in-time. The three design variables are as follows:

(i) $\mathcal{Y}$ is the covariance matrix to sample the N particles at the terminal time.

(ii) $\eta := \{\eta^i : 1 \le i \le N\}$ where $\eta^i := \{\eta_t^i, : 0 \le t \le T\}$ is the control input for the i-th particle. These inputs are designed to be independent B.M. with a prescribed covariance.

(iii) $\mathcal{A} := \{\mathcal{A}_t : 0 \le t \le T\}$ is a mean-field process which couples the simulations. The phrase "mean-field" means that the coupling depends *only* upon the (empirical) distribution $p_t^{(N)}$.

The triple $(\mathcal{Y}, \eta, \mathcal{A})$ are designed with the goal that the empirical covariance of the ensemble $\{Y_t^i : 1 \le i \le N\}$ approximates the solution of the differential Riccati equation (DRE) at time t. The resulting system is referred to as the *dual ensemble Kalman filter* (dual EnKF).

Contributions: The paper builds on Joshi et al. (2022) to include stochastic control systems. The novel aspects of the present paper are three-fold: **(i)** the algorithms and the analysis are extended to stochastic and robust/risk sensitive settings of the problem in a single unified framework; **(ii)** an error analysis is provided for finite but large N interacting particle system; and **(iii)** sample complexity results are obtained and comparison of the same provided against state-of-the-art algorithms for linear quadratic RL (iv) comparisons are presented to the path integral control framework, which is related to our approach and used widely in RL and robotics. (v) This paper includes numerical simulation and comparison for two benchmark example problems from previous works in this area.

Table 1: Expressions for DRE, where $D := BR^{-1}B^{\mathrm{T}}$ and $\Sigma := \sigma\sigma^{\mathrm{T}}$.

Cost	$\mathcal{D}(\Lambda)$	$\mathcal{D}^\dagger(\Lambda)$
LQG	$A^{\mathrm{T}}\Lambda + \Lambda A + C^{\mathrm{T}}C - \Lambda D\Lambda$	$A\Lambda + \Lambda A^{\mathrm{T}} - D + \Lambda C^{\mathrm{T}}C\Lambda$
LEQG	$A^{\mathrm{T}}\Lambda + \Lambda A + C^{\mathrm{T}}C - \Lambda(D - \theta\Sigma)\Lambda$	$A\Lambda + \Lambda A^{\mathrm{T}} - \frac{1}{\lvert\theta\rvert}(D - \theta\Sigma) + \lvert\theta\rvert\Lambda C^{\mathrm{T}}C\Lambda$

The salient features of the proposed algorithm are as follows: **(i)** It is not necessary that the matrix A is Hurwitz or that a stabilizing gain matrix K^0 is known (this is an assumption in many of the prior studies on PO); and **(ii)** convergence theory relies on law of large numbers (LLN) and spectral constant known from the DRE theory. Specifically, as $N \to \infty$, the proposed algorithm yields a learning rate that approximates the exponential rate of convergence of the solution of the DRE; and **(iii)** it alleviates issues like weight collapse and curse of dimensionality that are inherited by algorithms based on the importance sampling paradigm.

2. Problem formulation

Notation: $\|\cdot\|_F$ denotes Frobenius norm for matrices, $|\cdot|$ denotes 2-norm for vectors and $|\cdot|_M$ denotes weighted 2-norm under positive definite matrix M, that is, $|z|_M := z^{\mathrm{T}}Mz$, $\mathcal{N}(\text{mean}, \text{covariance})$ denotes normal distribution, $\mathbb{I}$ is used for identity matrix.

In linear quadratic settings, the cost function is quadratic as follows:

$$c(x,a) = \tfrac{1}{2}|Cx|^2 + \tfrac{1}{2}|a|_R^2, \quad x \in \mathbb{R}^d,\ a \in \mathbb{R}^m$$

Based on this, the following types of stochastic optimal control problems, linear quadratic Gaussian (LQG), linear exponential quadratic Gaussian (LEQG), and their average counterparts are considered (with $\theta \in \mathbb{R} \setminus \{0\}$):

$$J_T^{\text{LQG}}(U) := \mathbb{E}\left[\int_0^T c(X_t, U_t)\mathrm{d}t + \tfrac{1}{2}\,|X_T|_G^2\right], \tag{LQG}$$

$$J_T^{\text{LEQG}}(U) := \theta^{-1}\log\mathbb{E}\left[\exp\theta\left\{\int_0^T c(X_t, U_t)\mathrm{d}t + \tfrac{1}{2}|X_T|_G^2\right\}\right], \tag{LEQG}$$

$$J^{\text{AVG},i}(U) := \limsup_{T\to\infty}\frac{1}{T}J_T^i(U), \qquad i \in \{\text{LQG}, \text{LEQG}\}. \tag{AVG}$$

For the LEQG problem, θ is referred to as the risk parameter: The case $\theta > 0$ is known as risk-averse and $\theta < 0$ as risk-seeking Nagai (2013). The problem is to choose the control U to minimize the respective value $J(U)$ subject to the linear Gaussian dynamics (1). A standard set of assumptions–that are also made here–are now listed.

Assumption 1 *(A, B) is controllable, and $C^{\mathrm{T}}C, R, G \succ 0$. For LEQG, $BR^{-1}B^{\mathrm{T}} - \theta\sigma\sigma^{\mathrm{T}} \succ 0$.*

The main point of difference from the classical treatment is that the linear Gaussian model (1) is available *only* in the form of a simulator.

Definition 1 (Simulator) *A simulator of (1), denoted $\mathcal{S}$, takes the current state $x \in \mathbb{R}^d$, control $a \in \mathbb{R}^m$ and (small) time-step τ as input and gives the following random variable as output*

$$\mathcal{S}(x, a; \tau) = (Ax + Ba)\tau + \sigma \Delta W \quad \textit{where} \;\; \Delta W \overset{i.i.d}{\sim} \mathcal{N}(0, \mathbb{I}\tau).$$

A standard assumption in RL is that the state is available at every time t. Outside of a simulation type setting, it is difficult to describe a system where such an assumption holds: Most real-world systems have partial observation of the states through noisy sensor outputs. Next, many types of RL algorithms implement multiple iterations of the type (2), e.g., (Krauth et al., 2019, Algorithm 1,2), (Yang et al., 2019, Algorithm 2), (Basei et al., 2022, Algorithm 1), (Cassel and Koren, 2021, Algorithm 1) (Yaghmaie et al., 2023, Algorithm 3), (Zhang et al., 2021b, Algorithm 2), (Cui et al., 2023, Algorithm 2), (Lai and Xiong, 2024, Algorithm 1). More discussion in (Joshi et al., 2024, Appendix A).

2.1. Riccati equation and the Q function

Consider a matrix-valued process $\{P_t : 0 \le t \le T\}$ obtained from solving the DRE as follows:

$$-\frac{\mathrm{d}}{\mathrm{d}t} P_t = \mathcal{D}(P_t), \quad 0 \le t \le T, \quad P_T = G \tag{4}$$

where the expressions for the Riccati operator $\mathcal{D}(\cdot)$ are given in Table 1. While the DRE is the optimality equation for the finite time-horizon, the average cost solution is obtained by letting the time-horizon $T \to \infty$. Because (A, B) is controllable and (A, C) is observable, for any fixed time t, $P_t \to \bar{P}$ which solves the Algebraic Riccati equation (ARE): $\mathcal{D}(\bar{P}) = 0$ ((Kwakernaak and Sivan, 1972, Theorem 3.7)).

Definition 2 (Q-function) *The continuous-time Q-function (or Hamiltonian) is defined as*

$$\mathcal{Q}(x, a; t) := c(x, a) + x^{\mathrm{T}} P_t (Ax + Ba), \quad 0 \le t \le T, \;\; x \in \mathbb{R}^d, \; a \in \mathbb{R}^m \qquad \textit{for LQG, LEQG}$$
$$\bar{\mathcal{Q}}(x, a) := c(x, a) + x^{\mathrm{T}} \bar{P} (Ax + Ba), \quad x \in \mathbb{R}^d, \; a \in \mathbb{R}^m \qquad \textit{for AVG}$$

Then (see Liberzon (2012)),

$$U_t^{\text{opt}} = \begin{cases} \arg\min_{a \in \mathbb{R}^m} \mathcal{Q}(X_t, a; t), & 0 \le t \le T, \quad \text{LQG, LEQG} \\ \arg\min_{a \in \mathbb{R}^m} \bar{\mathcal{Q}}(X_t, a), & \text{AVG} \end{cases}$$

Because the Q function is quadratic, it is easily verified that the optimal control law is linear:

$$U_t^{\text{opt}} = \begin{cases} K_t X_t, \quad K_t := -R^{-1} B^{\mathrm{T}} P_t, & 0 \le t \le T, \quad \text{LQG, LEQG} \\ \bar{K} X_t, \quad \bar{K} := -R^{-1} B^{\mathrm{T}} \bar{P}, & \text{AVG} \end{cases}$$

The analysis of this paper requires consideration of P_t^{-1}. Since $G \succ 0$, it holds that $P_t \succ 0$ for $0 \le t \le T$ (Brockett, 2015, Sec. 24). Therefore, P_t^{-1} is well-defined. For every $t \in [0, T]$,

$$S_t := P_t^{-1} \text{ for LQG}; \qquad \text{and} \qquad S_t := (|\theta| P_t)^{-1} \text{ for LEQG}. \tag{5}$$

Then $\{S_t : 0 \le t \le T\}$ solves the dual DRE (defined in Table 1) as follows:

$$-\frac{\mathrm{d}}{\mathrm{d}t} S_t = \mathcal{D}^\dagger(S_t), \quad 0 \le t \le T, \quad S_T = G^{-1}$$

3. Interacting particle algorithm

In this section, two sets of algorithms are described to approximate the optimal control law based only on the use of the simulator. These are as follows:

• **Offline algorithm for solving DRE.** The goal is to learn an approximation of the Q-function. These approximations for the finite time-horizon and the average cost problems are denoted as $\mathcal{Q}^{(N)}$ and $\bar{\mathcal{Q}}^{(N)}$, respectively. Theoretical justification appears in (Joshi et al., 2024, Appendix A).

• **Online algorithm for computing the optimal control.** For each fixed time t, the optimal control is obtained by taking an arg min of approximate Q-function.

3.1. Dual EnKF for approximating solution of DRE

Consider the interacting particle system (3). The triple $(\mathcal{Y}, \eta, \mathcal{A})$ is designed as follows:

(i) Design of Y_T^i: Sample $Y_T^i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, S_T)$ for $i = 1, 2, \ldots, N$.

(ii) Design of η^i: The input η^i are i.i.d copies of a B.M. η whose covariance is

$$\text{Cov}(\eta) = R^{-1} \text{ for LQG}, \quad \text{and} \quad \text{Cov}(\eta) = (\sqrt{|\theta|}R)^{-1} \text{ for LEQG}. \tag{6}$$

In the context of RL, η has an interpretation as the exploration signal. The form (6) of the covariance means that the cheaper control directions are explored more.

(iii) Design of $\mathcal{A}_t$: The interaction term is a mean-field type linear control law as follows:

$$\mathcal{A}_t(z; p_t^{(N)}) := \begin{cases} \frac{1}{2} L_t^{(N)} C(z + n_T^{(N)}) + \frac{1}{2}\Sigma(S_t^{(N)})^{-1}(z - n_t^{(N)}); & \text{LQG} \\ \frac{|\theta|}{2} L_t^{(N)} C(z + n_T^{(N)}) + \text{sgn}(\theta)\Sigma(S_t^{(N)})^{-1}(z - n_t^{(N)}); & \text{LEQG} \end{cases} \tag{7}$$

where $\Sigma := \sigma\sigma^{\mathrm{T}}$, $n_t^{(N)} := N^{-1}\sum_i Y_t{}^i$, and

$$L_t^{(N)} := \tfrac{1}{N-1}\sum_{i=1}^N (Y_t{}^i - n_t^{(N)})(CY_t{}^i - Cn_t^{(N)})^{\mathrm{T}},\ S_t^{(N)} := \tfrac{1}{N-1}\sum_{i=1}^N (Y_t^i - n_t^{(N)})(Y_t^i - n_t^{(N)})^{\mathrm{T}}.$$

From (5), provided the right-hand side is well-defined,

$$P_t^{(N)} := (S_t^{(N)})^{-1} \text{ for LQG}, \quad \text{and} \quad P_t^{(N)} := (|\theta| S_t^{(N)})^{-1}; \text{ for LEQG}, \tag{8}$$

and for the average cost problem, $\bar{P}^{(N)} := P_0^{(N)}$. The error analysis (based on Bishop and Moral (2019)) is the subject of the following main result of this paper.

Theorem 3 *Consider the dual EnKF (3) under Assumption 1. Then for $N \geq d+1$, for each fixed t,*

$$\textit{(Finite-horizon)} \quad \mathbb{E}[\|S_t^{(N)} - S_t\|_F^2] \leq \frac{C_1}{N}, \quad \mathbb{E}[\|P_t^{(N)} - P_t\|_F^2] \leq \frac{C_4}{N}, \qquad 0 \leq t \leq T, \tag{9a}$$

$$\textit{(Average cost)} \quad \mathbb{E}[\|S_t^{(N)} - \bar{S}\|_F^2] \leq \frac{C_2}{N}, \quad \mathbb{E}[\|P_t^{(N)} - \bar{P}\|_F^2] \leq \frac{C_5}{N}, \qquad \textit{as } T \to \infty \tag{9b}$$

(where C_1, C_2, C_3, C_4 are model dependent but time-independent constants). For the average cost problem, there exists a constant $\lambda > 0$ such that exponential convergence to the stationary solution is obtained as follows:

$$\mathbb{E}[\|S_t^{(N)} - \bar{S}\|_F^2] \leq \frac{C_2}{N} + C_3 e^{-2\lambda(T-t)} \mathbb{E}[\|S_T^{(N)} - \bar{S}\|_F^2], \qquad 0 \leq t \leq T \tag{10}$$

Proof The proof appears in (Joshi et al., 2024, Appendix C). ■

Formula (10) is important because λ is the rate for learning the optimal solution. The constant λ is the spectral constant related to the exponential convergence of the solution of the DRE to the solution of the ARE (Kwakernaak and Sivan, 1972). As seen in the proof, the interaction term $\mathcal{A}$ is responsible for this property. The formula is useful to see the relation between the simulation horizon T and the error. For $\varepsilon > 0$, let $t = 0$ in (10), $N > O(\frac{1}{\varepsilon^2})$ and $T > O(\log(\frac{1}{\varepsilon}))$, then error is smaller than ε. The offline dual EnKF algorithm (Algorithm 1 in (Joshi et al., 2024, Appendix B)) presents a method to simulate the interacting particle system (3) using only access to a simulator. For the numerical approximation of the SDE, a first order Euler-Maruyama method is used and may be replaced with a higher order method.

3.2. Algorithm for approximating optimal control

If the matrix B is available, then the optimal control input at time t is approximated as follows:

$$U_t^{(N)} = \begin{cases} K_t^{(N)} X_t, & K_t^{(N)} := -R^{-1} B^{\mathrm{T}} P_t^{(N)}, \quad 0 \le t \le T, & \text{LQG, LEQG} \\ \bar{K}^{(N)} X_t, & \bar{K}^{(N)} := -R^{-1} B^{\mathrm{T}} \bar{P}^{(N)}, & \text{AVG} \end{cases}$$

For the case where an explicit form of B is not known, then the simulator is used to obtain an empirical approximation of the Q-function as follows:

Definition 4 (Empirical Q-function) *The empirical approximations are defined as*

$$\begin{aligned} \mathcal{Q}^{(N)}(x, a; t, \tau) &:= c(x, a)\tau + x^{\mathrm{T}} P_t^{(N)} \mathcal{S}(x, a; \tau), \quad 0 \le t \le T, \ \ x \in \mathbb{R}^d, \ a \in \mathbb{R}^m \\ \bar{\mathcal{Q}}^{(N)}(x, a; \tau) &:= c(x, a)\tau + x^{\mathrm{T}} \bar{P}^{(N)} \mathcal{S}(x, a; \tau), \quad x \in \mathbb{R}^d, \ a \in \mathbb{R}^m \end{aligned} \tag{11}$$

Based on the empirical Q-function, the optimal control is given by

$$U_t^{(N)} \tau = \begin{cases} \arg\min_{a \in \mathbb{R}^m} \mathbb{E}\left[\mathcal{Q}^{(N)}(X_t, a; t, \tau) | X_t \right], \quad 0 \le t \le T, & \text{LQG, LEQG} \\ \arg\min_{a \in \mathbb{R}^m} \mathbb{E}\left[\bar{\mathcal{Q}}(X_t, a; \tau) | X_t \right], & \text{AVG} \end{cases}$$

The expectation on the right-hand side is necessary because the simulator is noisy. A most straightforward implementation is to simply replace the expectation with a single sample—as one does in a stochastic gradient descent procedure. With additional computational budget, the expectation is approximated through N_e evaluations in a batch.

To evaluate the arg min, one may use a zero order optimization framework (Bach and Perchet, 2016). A simpler algorithm is obtained by noting that, like the Q function, the empirical Q function is also a quadratic function of the state, of the form $\frac{1}{2} a^{\mathrm{T}} R a + B^{\mathrm{T}} a + \varphi(x)$ where $\varphi(\cdot)$ is now a random function. For the case when the number of control inputs m is small, optimal control is approximated by evaluating the Q function for $a = R^{-1} e_i$ where $\{e_1, e_2, \ldots, e_m\}$ are basis vectors in $\mathbb{R}^m$. The online approximation of optimal control input is tabulated as Algorithm 2 in Appendix B and an error analysis of the gain appears in Appendix C.3 of Joshi et al. (2024).

Table 2: Complexity bounds in terms of error ε. These estimates are reported for the error in approximating gain in (Zhang et al., 2021b, Theorem 4.3) and (Krauth et al., 2019, Theorem 2.2); and the error in approximating the optimal cost in (Cassel and Koren, 2021, Lemma 6) and (Yang et al., 2019, Theorem 4.3).

Algorithm	particles/samples	simulation time	iterations
dual EnKF	$O(1/\varepsilon^2)$	$O(1/\log(\varepsilon))$	1
Zhang et al. (2021b)	$\tilde{O}(1/\varepsilon^4)$	$O(1)$	$O(1/\varepsilon)$
Cassel and Koren (2021)	$\tilde{O}(1/\varepsilon^4)$	$O(1)$	$O(1/\varepsilon)$
Krauth et al. (2019)	1	$O(1/\varepsilon^2)$	$O(1/\log(\varepsilon))$
Yang et al. (2019)	1	$O(1/\varepsilon^5)$	$O(\log(1/\varepsilon))$

3.3. Comparison of sample complexity to related works

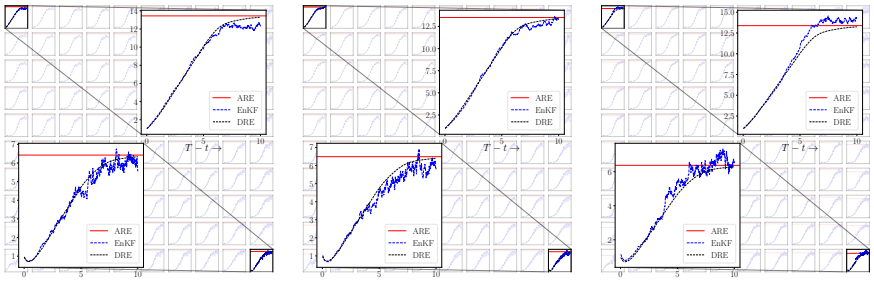
There are two types of errors for which analysis has been reported in recent literature: (i) the error in approximating the optimal value function; and (ii) the error in approximating the optimal gain matrix. Most of these results are for the stationary average cost case in the stochastic setting of the problem or for the infinite-horizon linear quadratic regulator (LQR) in the deterministic ($\sigma = 0$) setting. The quantitative comparisons with prior work are tabulated in Table 2.

In Krauth et al. (2019), an off policy method is used to estimate the Q function for discrete time average cost LQG. A linear function approximation is used with quadratic basis functions. The system is run for some fixed time using an exploration policy. At the end of each episode, the Q function is estimated using least squares. The error bounds in approximating the optimal gain are reported in (Krauth et al., 2019, Theorem 2.2). These results are closest to our work in terms of sample complexity requiring $O(\log(1/\varepsilon))$ training episodes and $O(1/\varepsilon^2)$ simulation time for error of ε (see (Krauth et al., 2019, Theorem 2.2)).

In Yang et al. (2019), a policy gradient algorithm is described. The actor is a gradient descent over the space of gains, where the policy gradient theorem is used to obtain the gradient. Error bounds are obtained for the error in value function (Yang et al., 2019, Theorem 4.3) which is related to error in solution of Riccati equation (Yang et al., 2019, Theorem 4.3). The algorithm needs $O(\log(1/\varepsilon))$ iterations, and a simulation horizon of the order $O(1/\varepsilon)$ for an ε error from the optimal value (Yang et al., 2019, Theorem 4.3).

In Cassel and Koren (2021), a zero order policy gradient algorithm is given for regret minimization in discrete time LQG. The idea is to perturb the gain in random directions to estimate the gradient of the value function with respect to the gain. Based on (Cassel and Koren, 2021, Lemma 6), $\tilde{O}(1/\varepsilon^4)$ samples are needed for gradient estimation and $O(1/\varepsilon)$ gradient descent iterations are needed for ε error in approximating the optimal value.

On the LEQG problem, Zhang et al. (2021b) extends the previous work of Zhang et al. (2020), Zhang et al. (2021a), and studies model free policy gradient methods for finite-horizon discrete-time LEQG. The work utilizes the equivalence between LEQG and linear quadratic min-max game to describe a "double-loop scheme". The approach is to write the optimization on the space of gains, and then apply a zeroth order policy optimization method to approximate the gradient flow. A sample



(*a*) (*b*) (*c*)

Figure 1: Comparison of the numerical solutions obtained from the EnKF, the DRE, and the ARE. The plots are in order: (a) LQG, (b) LEQG ($\theta > 0$) (c) LEQG ($\theta < 0$).

complexity analysis is given that quantifies the error bounds based on number of iterations and number of samples needed. The algorithm requires $\tilde{O}(1/\varepsilon^4)$ samples to estimate the gradient, and $O(1/\varepsilon)$ of gradient descent iterations for ε error in gain (Zhang et al., 2021b, Theorem 4.3).

The trade-off between EnKF and policy gradient type or least-squares type algorithms is as follows. The latter class of methods typically require multiple iterations (episodes) for simulating a system over a finite time-horizon, albeit with a relatively smaller number of particles, while EnKF needs only a single iteration but with a larger number of particles. Notably, the work in Yang et al. (2019); Krauth et al. (2019); Lai and Xiong (2024); Yaghmaie et al. (2023) needs only a single copy of the system. The EnKF particles are simulated in parallel, giving rise to much more efficient and faster implementation. Moreover, EnKF does not require an initial feasible (stabilizing) gain, while Krauth et al. (2019), Yang et al. (2019), Yaghmaie et al. (2023), Cassel and Koren (2021), Zhang et al. (2021b), Cui et al. (2023), Lai and Xiong (2024) need one.

3.4. Conceptual comparison to path integral control

While the focus of this paper is on designing the interaction term $\mathcal{A}$ in (3a) for the purpose of learning the value function, a related idea is the path integral approach (Kappen (2005), Theodorou and Todorov (2012), Thijssen and Kappen (2015)) and model predictive path integral control (MPPI) (Williams et al. (2016), Williams et al. (2018)) In this class of algorithms, one works with the LQG problem under the assumption that $\Sigma = \lambda D$ for some $\lambda > 0$ (for simplicity we present formulas with $b = \sigma$ and $R = \mathbb{I}$). One simulates multiple trajectories $\mathrm{d}X_t^i = (AX_t^i + BU_t^i)\mathrm{d}t + \mathrm{d}W_t^i$, and approximates the value function $v(x,t)$ as

$$\exp(-v(x,t)) = \mathbb{E}\left[\exp\left(\int_s^T |CX_t|^2 \mathrm{d}t - g^U(s,x)\right)\right] \approx \frac{1}{N}\sum_{i=1}^N w_i$$

$$U_s^*(x) - U_s \approx \frac{\sum_{i=1}^N w_i W_\tau^{(i)}}{\tau \sum_{i=1}^N w_i}, \quad w_i := \exp\left(\int_s^T |CX_t^i|^2 \mathrm{d}t - g^{U^i}(s,x)\right)$$

with $g^U(s,T) := \int_s^T \frac{1}{2} U_t^{\mathrm{T}} U_t dt + U_t^{\mathrm{T}} dW_t$. The following are key differences between the two approaches: **(i)** In our approach, all particles have equal weight $1/N$. However, importance sampling is well known to suffer from particle degeneracy. The issue becomes severe in higher dimensions and is known as curse of dimensionality, which our algorithm avoids (see Taghvaei and Mehta (2023)

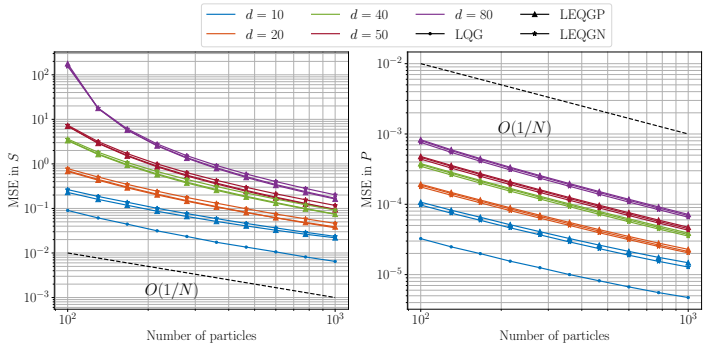


Figure 2: Relative error in approximating the solution of the ARE by the dual EnKF.

for a theoretical comparison). **(ii)** Another major point of distinction is that the path integral control is a fully model-based algorithm (see for example, formula (25) (where implementation of those terms assumes knowldge of B and σ) and Section IV-B in Williams et al. (2016) and Section VI-B in Williams et al. (2018))) while we focus on a simulator-based setting through the design of interactions. Moreover, MPPI assumes a special relation between control cost and noise which we do not need (see for example, the equation above (24) in Williams et al. (2018) or (Williams et al., 2016, equation (6))). **(iii)** A third major point is distinction between how trajctories are utilised. Since MPPI needs to iteratively evaluate expectation under the controlled measure, they need to run multiple copies of the system in for each iteration, while we need to run them only once. Our algorithm has interaction between particles, while MPPI uses importance sampling based approaches.

4. Numerical experiments and comparisons

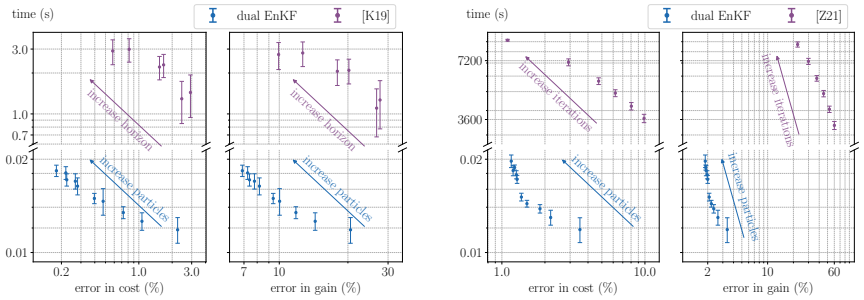
4.1. Numerical illustration of error formulas (9),(10)

An attractive feature of dual EnKF is that with large N, learning rate is inherited from the DRE convergence theory (see formula (10)). A numerical illustration of this formula, showing convergence of the d^2 entries of the P matrix, is depicted in Figure 1. The model is $d = 10$ dimensional where the entries of the A matrix are randomly sampled. Five of the total ten eigenvalues of A have positive real parts for the particular realization used in generating Figure 1.

In order to investigate scaling with increasing state dimension d, a spring mass damper model was introduced in Mohammadi et al. (2019). For this model, all three controllers are evaluated (LQG and LEQG for θ positive and negative). Figure 2 depicts the scaling as a function of N for the following metrics: $\frac{\mathbb{E}[\|\bar{S}^{(N)}-\bar{S}\|_F^2]}{\|\bar{S}\|_F^2}$ and $\frac{\mathbb{E}[\|\bar{P}^{(N)}-\bar{P}\|_F^2]}{\|\bar{P}\|_F^2}$ Consistent with (9), both the errors go down as $\frac{1}{N}$. See (Joshi et al., 2024, Appendix E) for model parameters and simulation details.

4.2. Numerical comparisons with prior work

Policy optimization: For this study, a three dimensional discrete-time system from Zhang et al. (2021b) is considered. For this model, comparisons are made with the following: (i) Finite-horizon LEQG in Zhang et al. (2021b), denoted [Z21]; and (ii) Average cost LQG in Krauth et al. (2019), denoted [K19]. Comparison is done for relative error in approximation of the optimal gain (with respect to the optimal gain) and relative error in the cost incurred by the control (with respect to



(*a*) (*b*)

Figure 3: Comparison of dual EnKF with : (a) [K19] for infinite horizon LQG; and (b) [Z21] for finite horizon LEQG. See Section 4.2 for details.

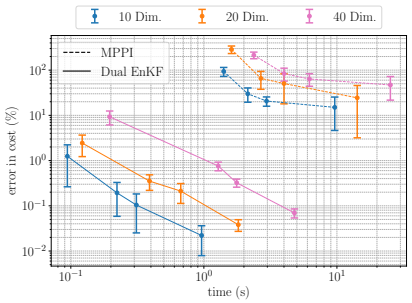


Figure 4: Comparison of dual EnKF with path integral control for spring mass damper system.

the optimal cost) given by the algorithm. Figure 3 depicts the numerically computed relationship between the relative error and the computational time.

Path integral control: We compare relative error in the cost incurred by the control (with respect to the optimal cost) given by the path integral approach (Williams et al., 2018), on the spring mass damper system for various dimensions of the state in Figure 4.

For each algorithm, the error becomes smaller with increasing computational time. For the dual EnKF, this tradeoff is obtained by increasing the number of particles. For [Z21] and [K19], the tradeoff is obtained by increasing the number of iterations and the time horizon. We observe that EnKF needs simulation times which are at least an order of magnitude lower than the other algorithms. See (Joshi et al., 2024, Appendix D) for additional information on the optimal control problem and the simulation parameters, and additional discussion on these studies.

5. Conclusion and future work

In this paper, an interacting particle system is described applicable to LQ stochastic optimal control problems in RL. The main contribution is that convergence rates are accelerated through appropriate design of interactions between particles (simulations) as shown in error bound (10) and illustrated using numerical comparisons with related algorithms. Two prominent areas for future work are extension to nonlinear non-Gaussian models, and to the partial state observation setting.

References

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvari. Model-free linear quadratic control via reduction to expert prediction. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3108–3117. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/abbasi-yadkori19a.html.

Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 257–283, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/bach16.html.

Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 23(178):1–34, 2022. URL http://jmlr.org/papers/v23/20-664.html.

Adrian N. Bishop and Pierre Del Moral. On the stability of matrix-valued Riccati diffusions. *Electronic Journal of Probability*, 24(none):1 – 40, 2019. doi: 10.1214/19-EJP342. URL https://doi.org/10.1214/19-EJP342.

R. W Brockett. *Finite dimensional linear systems*. SIAM, 2015.

Asaf B Cassel and Tomer Koren. Online policy gradient for model free learning of linear quadratic regulators with $\sqrt{t}$ regret. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1304–1313. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/cassel21a.html.

Leilei Cui, Tamer Basar, and Zhong-Ping Jiang. A reinforcement learning look at risk-sensitive linear quadratic gaussian control. In Nikolai Matni, Manfred Morari, and George J. Pappas, editors, *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pages 534–546. PMLR, 15–16 Jun 2023. URL https://proceedings.mlr.press/v211/cui23c.html.

Daniel Hernández-Hernández and Pedro Salazar-Sánchez. Risk-sensitive lqg discounted control problems and their asymptotic behavior. *SIAM Journal on Control and Optimization*, 61(3):1136–1161, 2023. doi: 10.1137/21M1459253. URL https://doi.org/10.1137/21M1459253.

A. A. Joshi, A. Taghvaei, P. G. Mehta, and S. P. Meyn. Controlled interacting particle algorithms for simulation-based reinforcement learning. *Systems & Control Letters*, 170:105392, 2022. ISSN 0167-6911. doi: https://doi.org/10.1016/j.sysconle.2022.105392. URL https://www.sciencedirect.com/science/article/pii/S0167691122001694.

Anant A Joshi, Heng-Sheng Chang, Amirhossein Taghvaei, Prashant G Mehta, and Sean P. Meyn. Design of interacting particle systems for fast linear quadratic rl, 2024. URL https://arxiv.org/abs/2406.11057.

Hilbert J. Kappen. Linear theory for control of nonlinear stochastic systems. *Phys. Rev. Lett.*, 95: 200201, Nov 2005. doi: 10.1103/PhysRevLett.95.200201. URL https://link.aps.org/doi/10.1103/PhysRevLett.95.200201.

Karl Krauth, Stephen Tu, and Benjamin Recht. Finite-time analysis of approximate policy iteration for the linear quadratic regulator. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/aaebdb8bb6b0e73f6c3c54a0ab0c6415-Paper.pdf.

Huibert Kwakernaak and Raphael Sivan. *Linear optimal control systems*. Wiley Interscience, New York, 1972. ISBN 0471511102.

Jing Lai and Junlin Xiong. Reinforcement learning for linear exponential quadratic gaussian problem. *Systems & Control Letters*, 185:105749, 2024. ISSN 0167-6911. doi: https://doi.org/10.1016/j.sysconle.2024.105749. URL https://www.sciencedirect.com/science/article/pii/S0167691124000379.

Daniel Liberzon. *Calculus of Variations and Optimal Control Theory*. Princeton University Press, Princeton, NJ, 2012. ISBN 978-0-691-15187-8. A concise introduction.

H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanovic. Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7474–7479, December 2019. doi: 10.1109/CDC40024.2019.9029985. ISSN: 2576-2370.

Hideo Nagai. Risk-sensitive stochastic control. In John Baillieul and Tariq Samad, editors, *Encyclopedia of Systems and Control*, pages 1–9. Springer London, London, 2013. ISBN 978-1-4471-5102-9. doi: 10.1007/978-1-4471-5102-9_233-1. URL https://doi.org/10.1007/978-1-4471-5102-9_233-1.

Vincent Roulet, Maryam Fazel, Siddhartha Srinivasa, and Zaid Harchaoui. On the convergence of the iterative linear exponential quadratic gaussian algorithm to stationary points. In *2020 American Control Conference (ACC)*, pages 132–137, 2020. doi: 10.23919/ACC45564.2020.9147694.

Amirhossein Taghvaei and Prashant G. Mehta. A survey of feedback particle filter and related controlled interacting particle systems (cips). *Annual Reviews in Control*, 55:356–378, 2023. ISSN 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2023.03.006. URL https://www.sciencedirect.com/science/article/pii/S136757882300010X.

Evangelos A. Theodorou and Emanuel Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 1466–1473, 2012. doi: 10.1109/CDC.2012.6426381.

Sep Thijssen and H. J. Kappen. Path integral control and state-dependent feedback. *Phys. Rev. E*, 91:032104, Mar 2015. doi: 10.1103/PhysRevE.91.032104. URL https://link.aps.org/doi/10.1103/PhysRevE.91.032104.

Grady Williams, Paul Drews, Brian Goldfain, James M. Rehg, and Evangelos A. Theodorou. Aggressive driving with model predictive path integral control. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1433–1440, 2016. doi: 10.1109/ICRA.2016.7487277.

Grady Williams, Paul Drews, Brian Goldfain, James M. Rehg, and Evangelos A. Theodorou. Information-theoretic model predictive control: Theory and applications to autonomous driving. *IEEE Transactions on Robotics*, 34(6):1603–1622, 2018. doi: 10.1109/TRO.2018.2865891.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

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, 2023. doi: 10.1109/TAC.2022.3145632.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/9713faa264b94e2bf346a1bb52587fd8-Paper.pdf.

Kaiqing Zhang, Bin Hu, and Tamer Basar. Policy Optimization for $\mathcal{H}_2$ Linear Control with $\mathcal{H}_\infty$ Robustness Guarantee: Implicit Regularization and Global Convergence. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, pages 179–190. PMLR, July 2020. URL https://proceedings.mlr.press/v120/zhang20a.html. ISSN: 2640-3498.

Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy Optimization for $\mathcal{H}_2$ Linear Control with $\mathcal{H}_\infty$ Robustness Guarantee: Implicit Regularization and Global Convergence, February 2021a. URL http://arxiv.org/abs/1910.09496. arXiv:1910.09496 [cs, eess, math].

Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer Basar. Derivative-Free Policy Optimization for Linear Risk-Sensitive and Robust Control Design: Implicit Regularization and Sample Complexity. In *Advances in Neural Information Processing Systems*, volume 34, pages 2949–2964. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper/2021/hash/1714726c817af50457d810aae9d27a2e-Abstract.html.