2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

2024 IEEE 63rd Conference on Decision and Control (CDC) | 979-8-3503-1633-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/CDC56724.2024.10886131

Dual Ensemble Kalman Filter for Stochastic Optimal Control

Anant A. Joshi, Amirhossein Taghvaei, Prashant G. Mehta, and Sean P. Meyn

***Abstract*— In this paper, stochastic optimal control problems in continuous time and space are considered. In recent years, such problems have received renewed attention from the lens of reinforcement learning (RL) which is also one of our motivation. The main contribution is a simulation-based algorithm – dual ensemble Kalman filter (EnKF) – to numerically approximate the solution of these problems. The paper extends our previous work where the dual EnKF was applied in deterministic settings of the problem. The theoretical results and algorithms are illustrated with numerical experiments.**

I. Introduction

Many types of reinforcement learning (RL) algorithms may be viewed as "simulation-based" where a model of a control system is simulated to evaluate and iteratively improve a policy. In continuous-type continuous-space settings of this paper, the optimal policy may be obtained from solving the Hamilton-Jacobi-Bellman (HJB) equation for the value function (in linear Gaussian settings, the equation reduces to a differential Riccati equation (DRE)). A simulation-based algorithm is useful for approximating the solution of the HJB or the DRE for the cases where the state-space is too large or the model parameters are not explicitly available (even though a simulator for the same is).

In this paper, we consider optimal control problems where the the control system is an Itô stochastic differential equation (SDE) as follows:

$$\mathrm{d}X_t = (a(X_t) + b(X_t)U_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \qquad (1a)$$

$$X_0 = x \qquad (1b)$$

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, $W := \{W_t : t \ge 0 \le t \le T\}$ is a standard Brownian motion (B.M.), and $a(\cdot), b(\cdot), \sigma(\cdot)$ are twice continuously differentiable functions of their arguments. The model is said to be *deterministic* if $\sigma(x) = 0$ for all $x \in \mathbb{R}^d$. The model is said to be linear Gaussian if $a(x) = Ax$, $b(x) = B$, and $\sigma(x) = \sigma$.

Our objective in this paper is to design a simulation-based algorithm to approximate (or learn) the optimal control law. Two types of control objectives are considered for this purpose: (i) stochastic optimal control (SOC); and (ii) risk sensitive control (RSC), both over a finite-time horizon $[0, T]$. The infinite-horizon case is obtained by letting $T \to \infty$.

The help explain the main idea of this paper, consider the SOC problem. Let $v_t(\cdot)$ denote the optimal value function for this problem (for the linear Gaussian model with quadratic cost (LQG), the value function $v_t(x) = x^\intercal P_t x + g_t$ is quadratic). Instead of computing $v_t(x)$ by solving the HJB (or the DRE) equation, our perspective is to view $\exp(-v_t(x))$ as an un-normalized form of a probability density, denoted as $p_t(x)$. That is,

$$p_t(x) = \frac{\exp(-v_t(x))}{\int_{\mathbb{R}^d} \exp(-v_t(z))\mathrm{d}z}, \quad x \in \mathbb{R}^d,\ 0 \le t \le T$$

(For LQG, assuming P_t is invertible, $p_t = \mathcal{N}(0, P_t^{-1})$ is a Gaussian density).

A. Contribution of this paper

Our aim is to approximate the density $p_t(\cdot)$ as an ensemble $\{Y_t^i \in \mathbb{R}^d : 0 \le t \le T, 1 \le i \le N\}$ such that

$$Y_t^i \overset{\text{i.i.d.}}{\sim} p_t(\cdot), \quad i = 1, 2, \ldots, N, \quad 0 \le t \le T \qquad (2)$$

(For LQG, $Y_t^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, P_t^{-1})$ for $0 \le t \le T$).

The proposed simulation-based algorithm is a backward-in-time controlled interacting particle system:

$$\mathrm{d}Y_t^i = \underbrace{a(Y_t^i)\mathrm{d}t + b\mathrm{d}\overleftarrow{\eta}_t^i + \sigma(Y_t^i)\mathrm{d}\overleftarrow{W}_t^i}_{i-\text{th copy of model (1)}} + \underbrace{\mathcal{A}_t(Y_t^i; p_t^{(N)})\mathrm{d}t}_{\text{mean field term}} \qquad (3a)$$

$$Y_T^i \overset{\text{i.i.d.}}{\sim} \frac{\exp(-\mathcal{G}(\cdot))}{\int_{\mathbb{R}^d} \exp(-\mathcal{G}(z))\mathrm{d}z}, \quad i = 1, 2, \ldots, N, \qquad (3b)$$

where $\mathcal{G}$ is the terminal condition for the value function (for LQG it is a quadratic function of the state), and $p_t^{(N)}$ is the empirical distribution of the ensemble. The design problem is to design $\eta := \{(\eta_t^i, : 0 \le t \le T\}$ and $\mathcal{A} := \{\mathcal{A}_t : 0 \le t \le T\}$ such that (2) holds in the mean-field limit ($N = \infty$).

The first such algorithm appears in our prior work [2] where explicit forms of η and $\mathcal{A}$ are described for

This work is supported in part by the AFOSR award FA9550-23-1-0060 (Joshi, Mehta), by the National Science Foundation (NSF) award EPCN-2318977 (Taghvaei) and by CCF 2306023 (Meyn).

A. A. Joshi and P. G. Mehta are with the Coordinated Science Laboratory and the Department of Mechanical Science and Engineering at the University of Illinois at Urbana-Champaign (e-mail: anantaj2; mehtapg@illinois.edu). A. Taghvaei is with the Department of Aeronautics and Astronautics at the University of Washington Seattle (email: amirtag@uw.edu). S. P. Meyn is with the Department of Electrical and Computer Engineering at the University of Florida at Gainesville (email: meyn@ece.ufl.edu).

The arXiv version of this paper [1] contains proofs of the technical results and an extended list of references.

Authorized licensed use limited to: University of Illinois. Downloaded on June 26,2025 at 21:25:22 UTC from IEEE Xplore. Restrictions apply.

the deterministic setting of the problem. The resulting algorithm is referred to as the dual ensemble Kalman filter (EnKF). The contribution of this paper is to extend the dual EnKF to stochastic setting for the SOC and RSC problems. The specific types of cost structures are introduced in Sec. II.

B. Relationship to literature

The idea of transforming an optimal control problem into a sampling problem is not new. For the SOC problem, the idea has its roots in the log transform which appears in [3] and is related to the minimum energy duality which is even older [4]. These types of transformations are routinely re-discovered and have been applied for solving sampling/inference problems as optimal control problems and vice-versa (see [1] for more references).

What is perhaps new in [2] are two aspects of (3):

1) Design of η_t^i as a B.M. This may be regarded as the exploration signal in RL.
2) Design of $\mathcal{A}_t(\cdot;\cdot)$. This is the idea of designing interactions between simulations to approximate the solution of the HJB equation.

In numerical performance and theoretical guarantees, these algorithms can be order of magnitude better than the competing approaches (detailed comparisons can be found in [2, Table 1]). There are some recent papers from other groups [5], [6] which also use interacting particle systems to solve stochastic optimal control problems, a detailed comparison to which is presented in the main body of the paper.

C. Organisation of paper

The outline of the remainder of this paper is as follows. The mathematical formulations for the SOC and RSC problems appear in Sec. II. For a detailed historical literature survey of both problems, see [1, Sec. II-E]. The simulation-based algorithm to approximate its solution appears in Sec. III. A detailed comparison to recent literature is included in Sec. III-E. For numerical simulations on an inverted pendulum on cart system, see [1, Sec. IV]

II. Problem Formulation

Notation: Given a symmetric positive semi-definite matrix Q, we let $|z|_Q \coloneqq z^\intercal Q z$. The normal distribution is represented as $\mathcal{N}(\cdot,\cdot)$ where the first argument is the mean and second is the covariance. We let $\mathrm{Cov}(\cdot)$ represent covariance.

A. Deterministic Optimal Control (DOC)

The simplest formulation is the deterministic case obtained when $\sigma = 0$. In this case, the SDE (1) reduces to an ordinary differential equation and X and U are both deterministic processes. The deterministic optimal control objective is as follows:

$$J_T^{\text{DOC}}(U) := \int_0^T \tfrac{1}{2}\left(|c(X_t)|^2 + |U_t|_R^2\right)\mathrm{d}t + \mathcal{G}(X_T)$$

where $c, \mathcal{G}$ are twice continuously differentiably real valued function, with $\mathcal{G}$ taking only non-negative value and R is a symmetric strictly positive definite matrix.

In a prior work [2], a simulation-based algorithm for approximation of the optimal control law is described. The algorithm is referred to as the dual ensemble Kalman filter or dual EnKF for short. The goal of the present paper is to generalize and extend the dual EnKF to the stochastic optimal control problem for the system (1) when $\sigma \neq 0$. For this purpose, the following types of cost structures are considered.

B. Stochastic Optimal Control (SOC)

The stochastic optimal control objective is as follows:

$$J_T^{\text{SOC}}(U) := \mathbb{E}\left[J_T^{\text{DOC}}(U)\right]$$

where randomness enters due to the Brownian motion in (1), and as noted earlier, we require the control to be adapted to the filtration generated by the Brownian motion.

C. Risk Sensitive Control (RSC)

For risk sensitive control, the following objective is of interest [7, Equation (1)]:

$$J_T^{\text{RSC}}(U) = \theta^{-1}\log\mathbb{E}\left[\exp\left(\theta J_T^{\text{DOC}}(U)\right)\right]$$

where $\theta \in \mathbb{R}\setminus\{0\}$ is referred to as the risk sensitive parameter. The case $\theta > 0$ is known as risk averse and $\theta < 0$ as risk seeking. A rigorous treatment of this problem, along with its motivations, appear in [7], [8, Section 8]. We note that the cost is always non-negative for every $\theta \neq 0$ because $\exp\left(\theta J_T^{\text{DOC}}(U)\right) > 1$ if $\theta > 0$ and $\exp\left(\theta J_T^{\text{DOC}}(U)\right) < 1$ when $\theta < 0$ for all U.

D. Linear Quadratic (LQ) Control

Suppose the system (1) has linear time invariant dynamics (LTI), that is, $a(z) = Az$, $b(z)u = Bu$ and $\sigma(z) = \sigma$, and the cost has a quadratic structure, that is, $|c(z)|^2 = |Cz|^2$ and $\mathcal{G}(z) = |z|_G^2$ for matrices A, B, σ, C, G where $G^\intercal = G \succ 0$.

Assumption 1: (A, B) is controllable, (A, C) is observable, and $BR^{-1}B^\intercal - \theta\sigma\sigma^\intercal \succ 0$.

The last assumption is needed to ensure positive definiteness of the solution to the Riccati equation, which is introduced in subsequent sections [9, Equation (90)]. In this case SOC is called linear quadratic Gaussian (LQG) and RSC is called linear quadratic exponential Gaussian (LEQG).

Authorized licensed use limited to: University of Illinois. Downloaded on June 26,2025 at 21:25:22 UTC from IEEE Xplore. Restrictions apply.

	$\mathcal{H}(v)$
SOC	$\frac{1}{2}\|c\|^2 + (\nabla v)^\intercal a - \frac{1}{2}(\nabla v)^\intercal D \nabla v + \frac{1}{2}\mathrm{tr}(\Sigma \nabla^2 v)$
RSC	$\frac{1}{2}\|c\|^2 + (\nabla v)^\intercal a - \frac{1}{2}(\nabla v)^\intercal (D - \theta\Sigma) \nabla v + \frac{1}{2}\mathrm{tr}(\Sigma \nabla^2 v)$

TABLE I: The right-hand-side of the HJB equation (4) for the SOC and RSC problems. The notations $D \coloneqq bR^{-1}b^\intercal$ and $\Sigma \coloneqq \sigma\sigma^\intercal$.

III. Proposed methodology

The proposed methodology is presented in three subsections. In Sec. III-A, we relate the value function for the optimal control problems to a probability density function. In Sec. III-B, we relate the probability density function to a suitably designed stochastic process that can be simulated as an interacting particle system. Finally, in Sec. III-C, we present the specialization to the LQ setting, followed by a Gaussian approximation procedure that can be applied to nonlinear setup.

A. Transforming value function to probability density

Consider the SOC and RSC problems. Value function $\{v_t(x) : 0 \le t \le T, x \in \mathbb{R}^d\}$ is defined as the optimal cost-to-go over the horizon $[t, T]$ when the state $X_t = x$. According to the dynamic programming principle, the value function satisfies the Hamilton Jacobi Bellman (HJB) partial differential equation (PDE)

$$-\frac{\partial v}{\partial t} = \mathcal{H}(v), \quad v_T = \mathcal{G} \tag{4}$$

where $\mathcal{H}$ is given in Table I [7, Equation (3)],[10, Equation (11.2.5)]. The optimal control is obtained as a function of v as

$$U_t^* = -R^{-1} b(X_t)^\intercal v_t(X_t). \tag{5}$$

In this paper, our strategy is to introduce a bijection $\psi : \mathbb{R} \to \mathbb{R}$ so that

$$p_t(x) := \frac{\psi(v_t(x))}{\int \psi(v_t(x))\mathrm{d}x}, \quad 0 \le t \le T,\ x \in \mathbb{R}^d \tag{6}$$

has the meaning of a probability density function. The bijection ψ is selected according to

$$\psi(z) \coloneqq \begin{cases} \exp(-z); & \text{SOC} \\ \exp(-|\theta|z); & \text{RSC} \end{cases} \tag{7}$$

Our choice of ψ is inspired from the log transform which is routinely used in risk sensitive control [8]. Since the value function is always positive, ψ has been appropriately adjusted so that the quantity inside the exponential is negative. We also note that our method is only intended for situations when $\psi(v_t(\cdot)) \in L^1(\mathbb{R})$ for all $0 \le t \le T$.

Cost	$\mathcal{D}(\Lambda)$
LQG	$A^\intercal\Lambda + \Lambda A + C^\intercal C - \Lambda B R^{-1} B^\intercal \Lambda$
LEQG	$A^\intercal\Lambda + \Lambda A + C^\intercal C - \Lambda(BR^{-1}B^\intercal - \theta\sigma\sigma^\intercal)\Lambda$

TABLE II: Expression for DRE in (11)

Cost	$\mathcal{D}^\dagger(\Lambda)$
LQG	$A\Lambda + \Lambda A^\intercal - BR^{-1}B^\intercal + \Lambda C^\intercal C \Lambda$
LEQG	$A\Lambda + \Lambda A^\intercal - \frac{1}{\|\theta\|}(BR^{-1}B^\intercal - \theta\sigma\sigma^\intercal) + \|\theta\|\Lambda C^\intercal C\Lambda$

TABLE III: Expression for the Dual DRE in (12)

B. Mean-field process

The goal is to design a (mean-field) stochastic process that has the same density function as $\{p_t : 0 \le t \le T\}$.

Define a stochastic process $Y = \{Y_t \in \mathbb{R}^d : 0 \le t \le T\}$ as a solution of the following backward (in time) SDE:

$$\begin{aligned} \mathrm{d}Y_t = a(Y_t)\mathrm{d}t + b(Y_t)\mathrm{d}\overleftarrow{\eta}_t + \sigma(Y_t)\mathrm{d}\overleftarrow{W}_t \\ + (\mathcal{I}_t(Y_t;\bar{p}_t) + \mathcal{C}_t(Y_t;\bar{p}_t))\mathrm{d}t \end{aligned} \tag{8a}$$

$$Y_T \sim \frac{\psi(\mathcal{G}(\cdot))}{\int \psi(\mathcal{G}(z))\mathrm{d}z} \tag{8b}$$

where $\eta = \{\eta_t \in \mathbb{R}^m : 0 \le t \le T\}$ is a B.M. with a suitably chosen covariance matrix, $\mathcal{I}_t(\cdot;\cdot), \mathcal{C}_t(\cdot;\cdot)$ is a suitably chosen vector field, and $\bar{p}_t$ is the density of Y. Here we have written the mean field term $\mathcal{A}$ introduced earlier as a sum of $\mathcal{I} + \mathcal{C}$, and the reason for which will be made clear later in this exposition. The meaning of the backward arrow on $\mathrm{d}\overleftarrow{\eta}, \mathrm{d}\overleftarrow{W}$ in (8) is that the SDE is simulated backward in time starting from the terminal condition specified at time $t = T$. The reader is referred to [11, Sec. 4.2] for the definition of the backward Itô-integral. The mean-field process is useful because of the following proposition.

Proposition 1: Consider the mean-field process (8). Suppose Cov(η) is Selected according to Table V, $\mathcal{I}$ and $\mathcal{C}$ satisfy the PDEs

$$-\nabla \cdot (\bar{p}_t(\cdot)\mathcal{I}_t(\cdot;\bar{p}_t)) = \bar{p}_t(h_t(\cdot) - \hat{h}_t), \tag{9a}$$

$$-\nabla \cdot (\bar{p}_t(\cdot)\mathcal{C}_t(\cdot;\bar{p}_t)) = \mathcal{V}_t(\cdot), \tag{9b}$$

where $\mathcal{V}_t(\cdot)$, h also appear in Table V, and $\hat{h}_t \coloneqq \int \bar{p}_t(z) h_t(z)\mathrm{d}z$. Then,

$$\bar{p}_t = p_t, \quad \forall t \in [0, T], \tag{10}$$

where $\bar{p}_t$ is the probability density function of Y_t and p_t is defined in (6) in terms of the value function.

Authorized licensed use limited to: University of Illinois. Downloaded on June 26,2025 at 21:25:22 UTC from IEEE Xplore. Restrictions apply.

	$\mathcal{I}_t^{\mathrm{LQ}}(z;n_t,S_t)$	$\mathcal{C}_t^{\mathrm{LQ}}(z;n_t,S_t)$
LQG	$\frac{1}{2}S_tC^\intercal C(z+n_t)$	$\frac{1}{2}\Sigma S_t^{-1}(z-n_t)$
LEQG $\theta>0$	$\frac{\lvert\theta\rvert}{2}S_tC^\intercal C(z+n_t)$	$\Sigma S_t^{-1}(z-n_t)$
LEQG $\theta<0$		0

TABLE IV: Vector fields for LQ case in (13).

The optimal control is expressed as a function of $\bar{p}_t$ according to

$$U_t^* = \begin{cases} R^{-1}b(X_t)^\intercal \nabla \log \bar{p}_t(X_t); & \text{SOC} \\ (|\theta|R)^{-1}b(X_t)^\intercal \nabla \log \bar{p}_t(X_t); & \text{RSC} \end{cases}$$

Proof: See [1, Appendix I]. ∎

The significance of Prop. 1 is that the optimal control policy $\phi_t(\cdot)$ can now be obtained in terms of the statistics of the random variable Y_t. The PDEs written may not always be analytically tractable, in which case, one has to rely on numerical approximation techniques. We call $\mathcal{I}$ as the interaction term, and $\mathcal{C}$ as the correction term, since the latter accounts for non-constant system model. In other words, if b, σ are constants, then $\mathcal{C}$ becomes 0 for RSC with $\theta < 0$, and otherwise $\mathcal{V}$ simplifies to $\mathrm{tr}(\nabla^2 \bar{p}_t)$ for RSC with $\theta > 0$ and $\frac{1}{2}\mathrm{tr}(\nabla^2 \bar{p}_t)$ for SOC.

C. LQ setting

In this scenario, the value function is obtained as

$$v_t(x) = \tfrac{1}{2}x^\intercal P_t x + g_t$$

where $\{P_t : 0 \le t \le T\}$ is a matrix valued process which is the solution of the backward (in time) differential Riccati equation (DRE)

$$-\frac{\mathrm{d}}{\mathrm{d}t}P_t = \mathcal{D}(P_t), \quad P_T = G \tag{11}$$

where the expressions for the $\mathcal{D}(\cdot)$ are in Table II, and

$$-\dot{g}_t = \mathrm{tr}(\sigma\sigma^\intercal P_t), \quad g_T = 0.$$

Under the Assumption 1, $P_t \succ 0$ for $0 \le t \le T$ whenever $G \succ 0$ [12, Sec. 24], [9, Equation (90)]. Then density p_t obtained from the value function then always takes the form $\mathcal{N}(0, S_t)$, where

$$S_t := \begin{cases} P_t^{-1}; & \text{LQG} \\ (|\theta|P_t)^{-1}; & \text{LEQG} \end{cases}, \quad 0 \le t \le T.$$

It is readily verified that $\{S_t : 0 \le t \le T\}$ also solves a DRE

$$-\frac{\mathrm{d}}{\mathrm{d}t}S_t = \mathcal{D}^\dagger(S_t), \tag{12}$$

which represents the dual of (11), and the expressions for $\mathcal{D}^\dagger$ appear in Table III. The expressions for $\mathcal{I}, \mathcal{C}, \eta$ in this scenario are denoted by a superscript $^{\mathrm{LQ}}$ and appear in Table IV. The derivations of these expressions appear in [1, Appendix I].

The mean-field process is empirically approximated by simulating a system of controlled interacting particles according to

$$\mathrm{d}Y_t^i = \underbrace{AY_t^i\mathrm{d}t + B\mathrm{d}\overleftarrow{\eta}_t^i + \sigma\mathrm{d}\overleftarrow{W}_t^i}_{\text{i-th copy of model (1)}} + \underbrace{\mathcal{I}_t^{\mathrm{LQ}}(Y_t^i; n_t^{(N)}, S_t^{(N)}) + \mathcal{C}_t^{\mathrm{LQ}}(Y_t^i; n_t^{(N)}, S_t^{(N)})\mathrm{d}t}_{\text{data assimilation process}} \tag{13a}$$

$$Y_T^i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, S_T), \quad 1 \le i \le N, \tag{13b}$$

η^i and W^i are an i.i.d copy of η and W respectively, $n_t^{(N)} := N^{-1}\sum_{i=1}^N Y_t^i$, and

$$S_t^{(N)} := \frac{1}{N-1}\sum_{i=1}^N (Y_t^i - n_t^{(N)})(Y_t^i - n_t^{(N)})^\top \tag{14}$$

The data assimilation process serves to couple the particles. Without it, the particles are independent of each other. The finite-N system (13) is referred to as the *dual EnKF*.

Optimal control: Set $X_t^i := (S_t^{(N)})^{-1}(Y_t^i - n_t^{(N)})$. Define

$$\tilde{K}_t^{(N)} := \begin{cases} \frac{1}{N-1}\sum_{i=1}^N X_t^i (X_t^i)^\intercal, & \text{LQG} \\ \frac{1}{(N-1)\lvert\theta\rvert}\sum_{i=1}^N X_t^i (X_t^i)^\intercal, & \text{LEQG} \end{cases} \tag{15}$$

We consider two cases:

(i) If the matrix B is explicitly known then

$$K_t^{(N)} = -R^{-1}B^\intercal \tilde{K}_t^{(N)} \tag{16}$$

from which the optimal control policy is approximated as $\phi_t^{(N)}(x) = K_t^{(N)}x$.

(ii) If B is unknown, define the Hamiltonian

$$H^{(N)}(x,\alpha,t) := \underbrace{\tfrac{1}{2}|Cx|^2 + \tfrac{1}{2}\alpha^\intercal R\alpha}_{\text{cost function}} + x^\intercal \tilde{K}_t^{(N)} \underbrace{(Ax + B\alpha)}_{\text{model (1)}}$$

from which the optimal control policy is approximated as

$$\phi_t^{(N)}(x) = \operatorname*{arg\,min}_{a\in\mathbb{R}^m} H^{(N)}(x,a,t)$$

We obtain $(Ax + B\alpha)$ by averaging the model (1) N_s many times. The error in estimation would scale as $1/N_s$ from the strong law of large numbers. There are several zeroth-order approaches to

Authorized licensed use limited to: University of Illinois. Downloaded on June 26,2025 at 21:25:22 UTC from IEEE Xplore. Restrictions apply.

Cost	$\mathcal{V}(\cdot)$	$h_t(\cdot)$	$\mathrm{Cov}(\eta)$
LQG	$\frac{1}{2}\nabla^2\cdot(\Sigma\bar{p}_t)$	$\frac{1}{2}\lvert c\rvert^2+\nabla\cdot a+\frac{1}{2}\mathrm{tr}((D-\Sigma)\nabla^2\log(\bar{p}_t))$	R^{-1}
LEQG $\theta<0$	$\nabla\cdot(\bar{p}_t\nabla\cdot(\Sigma-\frac{1}{\theta}D))$	$-\frac{\theta}{2}\lvert c\rvert^2+\nabla\cdot a+\frac{1}{2}\nabla^2\cdot(\frac{1}{\theta}D-\Sigma)-\frac{1}{2\theta}\mathrm{tr}(D\nabla^2\log\bar{p}_t)$	$(\sqrt{\lvert\theta\rvert}R)^{-1}$
LEQG $\theta>0$	$\nabla^2\cdot(\bar{p}_t\Sigma)+\nabla\cdot(\bar{p}_t\nabla\cdot(\frac{1}{\theta}D-\Sigma))$	$\frac{\theta}{2}\lvert c\rvert^2+\nabla\cdot a-\nabla^2\cdot(\frac{1}{\theta}D-\Sigma)+\frac{1}{2}\mathrm{tr}((\frac{1}{\theta}D-2\Sigma)\nabla^2\log\bar{p}_t)$	$(\sqrt{\lvert\theta\rvert}R)^{-1}$

TABLE V: Details for the Poisson equation (9) in Proposition 1.

solve the minimization problem, e.g., by constructing 2-point estimators for the gradient. Since the objective function is quadratic and the matrix R is known, m queries of $H^{(N)}(x,\cdot,t)$ are sufficient to compute $\phi_t^{(N)}(x)$.

The interacting particle system (13) is simulated using Euler-Maruyama discretization scheme, where the direction of time is reversed. The discretization scheme is similar to the dual EnKF algorithm that appears in [2].

The correction and interaction terms are simplified under certain assumption about the model, as described in the following result.

Proposition 2: Consider the mean-field process in the LQ setting.

1) For LQG, if $BR^{-1}B^\intercal=\sigma\sigma^\intercal+B\tilde{R}B^\intercal$ for some $\tilde{R}\succeq 0$, then set $\mathrm{Cov}(\eta)=\tilde{R}$ and $\mathcal{C}_t\equiv 0$. In particular, if $\sigma=BR^{-\frac{1}{2}}$, then $\eta\equiv 0$.
2) For LEQG with $\theta>0$, if $BR^{-1}B^\intercal=2\theta\sigma\sigma^\intercal+\theta B\tilde{R}B^\intercal$ for some $\tilde{R}\succeq 0$, then set $\mathrm{Cov}(\eta)=\tilde{R}$ and $\mathcal{C}_t\equiv 0$. In particular, if $\sigma=B(2\theta R)^{-\frac{1}{2}}$, then $\eta\equiv 0$.

Proof: See [1, Appendix I]. ∎

Remark 1: We make some observations on making our algorithm model-free. In the earlier work [2], emphasis was on designing algorithms which can be implemented without having access to the model parameters in (1), but with only access to model evaluations. The LEQG for $\theta<0$ is model free, since the vector field $\mathcal{I}^{\mathrm{LQ}}$ does not involve any of the model parameters A,B,σ. Similarly, the situations considered in Proposition 2 can be implemented in a model free manner.

D. Gaussian Approximation

For a numerical approximation of the solution of the Poisson equations, we notice that the terms simplify in the following case:

1) $a(x)$ is conservative, i.e. $\nabla\cdot a(x)=0$.
2) If $b(x)=B$ then $\nabla\cdot D=0$.
3) If $\sigma(x)=\sigma$ then $\nabla\cdot\Sigma=0$.

Making these simplifications, and considering a Gaussian approximation for the density p_t we get

$$h_t(x)=\frac{1}{2}|c(x)|^2+\text{(constant)}$$

which makes the solution of (9) simpler. This is useful to obtain a dual EnKF algorithm:

$$\begin{aligned}\mathrm{d}Y_t^i&=\underbrace{a(Y_t^i)\mathrm{d}t+b(Y_t^i)\mathrm{d}\overleftarrow{\eta}_t^i+\sigma\mathrm{d}\overleftarrow{W}_t^i}_{\text{i-th copy of model (1)}}\\&+\underbrace{\mathcal{I}_t^{\mathrm{GA}}(Y_t^i;n_t^{(N)},S_t^{(N)})+\mathcal{C}_t^{\mathrm{LQ}}(Y_t^i;n_t^{(N)},S_t^{(N)})\mathrm{d}t}_{\text{data assimilation process}}\\Y_T^i&\overset{\text{i.i.d}}{\sim}\mathcal{N}(0,S_T),\quad 1\le i\le N,\end{aligned}$$

where the vector field $\mathcal{C}^{\mathrm{LQ}}$ is the same as the linear quadratic case (Table IV), $\eta^i:=\{\eta_t^i\in\mathbb{R}^m:i:0\le t\le T\}$ is an independent copy of η, and

$$\mathcal{I}_t^{\mathrm{GA}}(Y_t^i;n_t^{(N)},S_t^{(N)}):=\begin{cases}\frac{1}{2(N-1)}\tilde{\mathcal{I}}; & \text{SOC}\\ \frac{1}{2|\theta|(N-1)}\tilde{\mathcal{I}}; & \text{RSC}\end{cases}$$

where

$$\tilde{\mathcal{I}}:=\sum_{j=1}^N(Y_t^j-n_t^{(N)})(c(Y_t^j)-\hat{c}_t^{(N)})^\top(c(Y_t^i)+\hat{c}^{(N)})$$

and $\hat{c}_t^{(N)}:=N^{-1}\sum_{i=1}^N c(Y_t^i)$. One may interpret the above as the dual counterpart of the FPF algorithm with a constant gain approximation [13, Example 2].

The optimal control may be approximated as earlier via the Hamiltonian (with $\tilde{K}_t^{(N)}$ as in (15)),

$$\begin{aligned}H^{(N)}(x,\alpha,t):=\tfrac{1}{2}|c(x)|^2+\tfrac{1}{2}\alpha^\intercal R\alpha\\+x^\intercal\tilde{K}_t^{(N)}\underbrace{(a(x)+b(x)\alpha)}_{i-\text{th copy of model (1)}}\end{aligned}$$

where as before $X_t^i:=(S_t^{(N)})^{-1}(Y_t^i-n_t^{(N)})$.

E. Comparison to recent literature

In recent years, there has been work on using particle based methods to approximate the solution of Fokker Planck equation [14], and the solution of stochastic non-linear affine and quadratic in control problems [5],

Authorized licensed use limited to: University of Illinois. Downloaded on June 26,2025 at 21:25:22 UTC from IEEE Xplore. Restrictions apply.

[6]. In [14], authors express the Fokker Planck equation as a Liouville equation by incorporating the score function (defined as the gradient of log of density) in the dynamics of the original system. Then they adopt a variational representation for the score function and propose a particle based approach to estimate it.

Closely related to our approach are [5], [6] which are also based on the same fundamental idea of turning value functions to probability density functions using the exponential transform. However, in [5] and [6], the density obtained from the value function p is expressed as ratio of two densities q and ρ, which is very much like smoothing [15]. To be precise, they write $q_t(x) = \rho_t(x)p_t(x)$, where ρ and q propagates forward in time. The PDEs governing the two densities are as follows

$$\frac{\partial \rho_t(x)}{\partial t} = -\nabla \cdot (a\rho_t) + \mathrm{tr}(\Sigma \nabla^2 \rho_t) - \rho_t c$$
$$\frac{\partial q_t(x)}{\partial t} = -\nabla \cdot ((a + bU_t^*)\rho_t) + \mathrm{tr}(\Sigma \nabla^2 q_t)$$

and then they replace $U_t^* = R^{-1}b^\intercal \nabla \log p_t = R^{-1}b^\intercal(\nabla \log q_t - \nabla \log \rho_t)$. The two densities ρ and q are each simulated as a coupled interacting particle system that involves approximation of the so-called score function, i.e. gradient of the logarithm of density. Although the PDEs for the densities they obtain are very similar to the PDEs that we get (see [1, Table VI]), a significant difference is that they need two separate interacting particle systems and their simulation for q utilises the result of the simulation for ρ. In contrast, our proposed approach only involves one interacting particle system and is also applicable to risk sensitive case. Although we divide our solution into two parts $\mathcal{I}$ and $\mathcal{C}$, we do it for convenience of notation and comprehension, and they may very well be combined into a single vector field by summing them. Moreover, we lay emphasis on the model-free setting under certain scenarios as stated in Remark 1 while their approach needs access to the model.

The idea of the log transform employed in this paper builds on our previous work [2], in which we make detailed comparisons to the body of literature in nonlinear filtering theory. In a companion paper [16] we show error analysis with respect to the number of particles N for the linear quadratic case wherein we quantify the error in estimating $S_t^{(N)}$, which changes as $1/N$. We also compare our approach to policy gradient type reinforcement learning approaches which solve the same problem. We would like to emphasize here that we employ a simulator based approach, as opposed to a model identification approach based on estimating the system dynamics and solving the HJB equation. For the general nonlinear case, our work involves solving the Poisson equation (9), see e.g. [17] and references therein.

ACKNOWLEDGMENTS

We would like to thank Prof. Vivek Borkar for suggesting this direction of work and for many helpful discussions.

REFERENCES

[1] A. A. Joshi, A. Taghvaei, P. G. Mehta, and S. P. Meyn, "Dual ensemble kalman filter for stochastic optimal control," 2024. [Online]. Available: https://arxiv.org/abs/2404.06696

[2] ——, "Controlled interacting particle algorithms for simulation-based reinforcement learning," *Systems & Control Letters*, vol. 170, p. 105392, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167691122001694

[3] W. H. Fleming and S. K. Mitter, "Optimal Control and Nonlinear Filtering for Nondegenerate Diffusion Processes," *Stochastics*, vol. 8, no. 1, pp. 63–77, January 1982. [Online]. Available: https://doi.org/10.1080/17442508208833228

[4] R. E. Mortensen, "Maximum-likelihood recursive nonlinear filtering," *Journal of Optimization Theory and Applications*, vol. 2, no. 6, pp. 386–394, Nov 1968. [Online]. Available: https://doi.org/10.1007/BF00925744

[5] D. Maoutsa and M. Opper, "Deterministic particle flows for constraining stochastic nonlinear systems," *Phys. Rev. Res.*, vol. 4, p. 043035, Oct 2022. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevResearch.4.043035

[6] S. Reich, "Particle-based algorithm for stochastic optimal control," 2024. [Online]. Available: https://arxiv.org/abs/2311.06906

[7] H. Nagai, *Risk-Sensitive Stochastic Control*. London: Springer London, 2013, pp. 1–9. [Online]. Available: https://doi.org/10.1007/978-1-4471-5102-9_233-1

[8] W. H. Fleming and H. Soner, *Logarithmic Transformations and Risk Sensitivity*. New York, NY: Springer New York, 2006, pp. 227–259. [Online]. Available: https://doi.org/10.1007/0-387-31071-1_6

[9] D. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Transactions on Automatic Control*, vol. 18, no. 2, pp. 124–131, 1973.

[10] A. Bensoussan, *Estimation and Control of Dynamical Systems*, ser. Interdisciplinary Applied Mathematics. Springer International Publishing, 2018.

[11] D. Nualart and É. Pardoux, "Stochastic calculus with anticipating integrands," *Probability Theory and Related Fields*, vol. 78, no. 4, pp. 535–581, 1988.

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

[13] T. Yang, R. Laugesen, P. Mehta, and S. Meyn, "Multivariable feedback particle filter," *Automatica*, vol. 71, pp. 10–23, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000510981630142X

[14] D. Maoutsa, S. Reich, and M. Opper, "Interacting particle solutions of fokker–planck equations through gradient–log–density estimation," *Entropy*, vol. 22, no. 8, 2020. [Online]. Available: https://www.mdpi.com/1099-4300/22/8/802

[15] J. W. Kim and P. G. Mehta, "An optimal control derivation of nonlinear smoothing equations," in *Advances in Dynamics, Optimization and Computation*, O. Junge, O. Schütze, G. Froyland, S. Ober-Blöbaum, and K. Padberg-Gehle, Eds. Cham: Springer International Publishing, 2020, pp. 295–311.

[16] A. A. Joshi, H.-S. Chang, A. Taghvaei, P. G. Mehta, and S. P. Meyn, "Design of interacting particle systems for fast and efficient reinforcement learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.11057

[17] A. Taghvaei, P. G. Mehta, and S. P. Meyn, "Diffusion map-based algorithm for gain function approximation in the feedback particle filter," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 8, no. 3, pp. 1090–1117, 2020. [Online]. Available: https://doi.org/10.1137/19M124513X

Authorized licensed use limited to: University of Illinois. Downloaded on June 26,2025 at 21:25:22 UTC from IEEE Xplore. Restrictions apply.