2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Policy Iteration for Multiplicative Noise Output Feedback Control

Benjamin Gravell Matilde Gargiani John Lygeros Tyler H. Summers

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9993098

***Abstract*— We propose a policy iteration algorithm for solving the multiplicative noise linear quadratic output feedback design problem. The algorithm solves a set of coupled Riccati equations for estimation and control arising from a partially observable Markov decision process (POMDP) under a class of linear dynamic control policies. We show in numerical experiments far faster convergence than a value iteration algorithm, formerly the only known algorithm for solving this class of problem. The results suggest promising future research directions for policy optimization algorithms in more general POMDPs, including the potential to develop novel approximate data-driven approaches when model parameters are not available.**

I. Introduction

Multiplicative noise models can be used to represent myriad phenomena where noise or uncertainty depends on the system state, input, or output. These models have a long history in control theory [1] and have been utilized in the context of networked control systems [2], robots with distance-dependent sensors such as lidar and optical cameras [3], turbulent fluid flow [4], climate dynamics [5], biological sensorimotor systems [6], neuronal brain networks [7], portfolio optimization and financial markets [8] power grids with stochastic inertia [9], and aerospace systems [10]. Recently, these models have also been used to represent parametric uncertainty and promote robustness in data-driven control and machine learning via, e.g., bootstrapping [11], domain randomization [12], and dropout [13].

Designing optimal output feedback controllers for multiplicative noise dynamical systems is particularly challenging because, in marked contrast to classical LQG/$\mathcal{H}_2$ and $\mathcal{H}_\infty$ control design, there is no separation between estimation and control. In particular, a canonical linear quadratic problem features a set of Riccati equations involving cost and covariance matrices for the state and state estimate that are coupled by the multiplicative noise. These equations cannot be solved by direct methods for decoupled Riccati equations [14]. The only known solution method is an iterative algorithm akin to value iteration, described in [15].

Policy iteration has been studied extensively in general settings for dynamic programming and reinforcement learning and is closely related to the Newton method [16], [17]. Policy iteration and the Newton method have also been studied extensively for solving the single generalized Riccati equation arising in the linear quadratic state-feedback setting in [18], [19]. However, there is far less known about policy iteration and the Newton method for output feedback control or partially observable Markov decision processes (POMDPs). Various approximate policy iteration schemes have been studied to a limited extent in deterministic and additive noise output feedback linear quadratic problems in [20], [21], [22] and for tabular and nonlinear POMDPs in [23], [24]. None of these address the coupled Riccati equations arising in the output-feedback multiplicative noise setting.

Our main contribution is to propose a policy iteration algorithm for multiplicative noise output feedback control design, which solves a set of coupled Riccati equations for estimation and control. We show in numerical experiments far faster convergence than a value iteration algorithm, the only known algorithm for solving this class of problem. We provide an open source implementation of our algorithm to facilitate further research and reproducibility. The results suggest promising future research directions for policy iteration and data-driven output feedback control algorithms for the multiplicative noise problem and other more general POMDPs.

The paper is organized as follows. §II formulates an optimal output feedback control problem for systems with multiplicative noise. §III describes the coupled Riccati equations that determine an optimal policy in a form that motivates policy iteration. Our policy iteration algorithm is proposed in §IV. §V presents the numerical experiments, and §VI concludes.

Notation: Denote the set of real-valued $n \times m$ matrices as $\mathbb{R}^{n\times m}$ and the set of $n \times n$ symmetric positive semidefinite matrices as $\mathbb{S}^n_+$. Denote the $n \times n$ all-zeros and identity matrices as 0_n and I_n, respectively. For real-valued matrices A and B, denote the transpose as $A^\intercal$, the trace as $\mathrm{Tr}(A)$, and the Frobenius inner product as $\langle A, B\rangle := \mathrm{Tr}(A^\top B)$.

II. Problem Formulation: Output Feedback Control with Multiplicative Noise

We consider output-feedback control of the discrete-time stochastic linear dynamical system

$$x_{t+1} = A_t x_t + B_t u_t + w_t, \tag{1a}$$
$$y_t = C_t x_t + v_t, \tag{1b}$$

where $x_t \in \mathbb{R}^n$ is the system state, $u_t \in \mathbb{R}^m$ is the control input, and $y_t \in \mathbb{R}^p$ is the observed output. The system matrices (A_t, B_t, C_t) are time-varying random matrices

B. Gravell and T. Summers are with the Control, Optimization, and Networks Lab, University of Texas at Dallas (email: {benjamin.gravell}, {tyler.summers}@utdallas.edu). M. Gargiani and J. Lygeros are with the Automatic Control Laboratory, ETH Zürich, Switzerland. This material is based on work supported by the United States Air Force Office of Scientific Research under award number FA2386-19-1-4073, the National Science Foundation under award number ECCS-2047040, and the European Research Council (ERC) under the OCAL project 787845.

decomposed as

$$A_t = A + \sum_{i=1}^{a} \alpha_{t,i} A_{\Delta,i}, \quad B_t = B + \sum_{i=1}^{b} \beta_{t,i} B_{\Delta,i},$$
$$C_t = C + \sum_{i=1}^{c} \gamma_{t,i} C_{\Delta,i},$$

where $A = \mathbb{E}[A_t]$, $B = \mathbb{E}[B_t]$, $C = \mathbb{E}[C_t]$ are mean matrices, $\{\alpha_{t,i}\}_{i=1}^{a}$, $\{\beta_{t,i}\}_{i=1}^{b}$, $\{\gamma_{t,i}\}_{i=1}^{c}$ are random scalars mutually independent and independent across time with zero mean and standard deviations $\{\sigma_{A,i}\}_{i=1}^{a}$, $\{\sigma_{B,i}\}_{i=1}^{b}$, $\{\sigma_{C,i}\}_{i=1}^{c}$ respectively, and $\{A_{\Delta,i}\}_{i=1}^{a}$, $\{B_{\Delta,i}\}_{i=1}^{b}$, $\{C_{\Delta,i}\}_{i=1}^{c}$ are pattern matrices constant across time specifying the directions in which the multiplicative noise acts. This decomposition can be obtained by an eigendecomposition of the covariance structure of the A_t, B_t, C_t, treating the eigenvalues as $\{\sigma_{A,i}\}_{i=1}^{a}$, $\{\sigma_{B,i}\}_{i=1}^{b}$, $\{\sigma_{C,i}\}_{i=1}^{c}$ and eigenvectors as $\{A_{\Delta,i}\}_{i=1}^{a}$, $\{B_{\Delta,i}\}_{i=1}^{b}$, $\{C_{\Delta,i}\}_{i=1}^{c}$ as discussed e.g. in [25]. The process and observation noises w_t and v_t, respectively, are independent across time, have mean zero, and have covariance

$$\mathbb{E}\left[\begin{bmatrix} w_t \\ v_t \end{bmatrix}\begin{bmatrix} w_t \\ v_t \end{bmatrix}^{\intercal}\right] = \begin{bmatrix} W_{xx} & W_{xy} \\ W_{yx} & W_{yy} \end{bmatrix} = W \succ 0.$$

The initial state x_0 is a random vector drawn from a distribution with mean zero and covariance X_0. We consider convex quadratic stage cost in the states and inputs

$$\ell(x_t, u_t) = \begin{bmatrix} x_t \\ u_t \end{bmatrix}^{\intercal} \begin{bmatrix} Q_{xx} & Q_{xu} \\ Q_{ux} & Q_{uu} \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}, \quad Q \succ 0.$$

This yields the infinite-horizon average-cost multiplicative-noise linear quadratic output feedback control problem

$$\min_{\pi} J(\pi) := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \ell(x_t, u_t)\right] \tag{2a}$$
$$\text{subject to (1)}, \tag{2b}$$

where a control policy $u_t = \pi(y_{0:t}, u_{0:t-1})$ dependent on the input-output history $y_{0:t} := [y_0, y_1, ..., y_t], u_{0:t-1} := [u_0, u_1, ..., u_{t-1}]$ is to be designed, and expectation is taken with respect to all random quantities in the problem, namely $\{x_0, \{A_t\}, \{B_t\}, \{C_t\}, \{w_t\}, \{v_t\}\}$.

A. Linear Dynamic Control Policies

A *linear dynamic controller* is a widely used class of policy which combines a linear state estimator with a linear state estimate feedback in the form

$$\hat{x}_{t+1} = F\hat{x}_t + Ly_t, \tag{3a}$$
$$u_t = K\hat{x}_t. \tag{3b}$$

The initial state estimate is chosen as $\hat{x}_0 = 0$, since the initial state has mean zero. Such a controller is fully specified by the triple (F, K, L) where $K \in \mathbb{R}^{m\times n}$ is the control gain, $L \in \mathbb{R}^{n\times p}$ is the state estimator gain, and $F \in \mathbb{R}^{n\times n}$ is the closed-loop model matrix. We will consider the problem (2) with optimization over this special class of linear dynamic controllers. In the classical LQG setting with additive noise-only (where $A_t = A$, $B_t = B$, $C_t = C$), the optimal policy is indeed a linear dynamic controller of the class (3). However, in the multiplicative noise setting this class may not be optimal.[1] Nevertheless, the work of [15] shows that restricting attention to (3) admits useful stability characterizations and optimal control synthesis equations.

B. Closed-Loop Dynamics and Performance Criterion

We now develop some notation and expressions for the closed-loop dynamics and performance criteria, which will be useful later for describing a policy iteration algorithm. Using a controller (F, K, L), the closed-loop system dynamics are

$$\begin{bmatrix} x_{t+1} \\ \hat{x}_{t+1} \end{bmatrix} = \begin{bmatrix} A_t & B_tK \\ LC_t & F \end{bmatrix}\begin{bmatrix} x_t \\ \hat{x}_t \end{bmatrix} + \begin{bmatrix} I_n & 0 \\ 0 & L \end{bmatrix}\begin{bmatrix} w_t \\ v_t \end{bmatrix} \tag{4}$$

Denote the following augmented closed-loop matrices

$$\Phi'_t := \begin{bmatrix} A_t & B_tK \\ LC_t & F \end{bmatrix}, \tag{5a}$$
$$Q' := \begin{bmatrix} I_n & 0 \\ 0 & K \end{bmatrix}^{\intercal} \begin{bmatrix} Q_{xx} & Q_{xu} \\ Q_{ux} & Q_{uu} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & K \end{bmatrix}, \tag{5b}$$
$$W' := \begin{bmatrix} I_n & 0 \\ 0 & L \end{bmatrix} \begin{bmatrix} W_{xx} & W_{xy} \\ W_{yx} & W_{yy} \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & L \end{bmatrix}^{\intercal}. \tag{5c}$$

Under the closed-loop dynamics (4), define the value and covariance matrices at time t recursively by

$$P'_{t+1} = \mathbb{E}\left[\Phi_t'^{\intercal} P'_t \Phi'_t\right] + Q', \tag{6a}$$
$$S'_{t+1} = \mathbb{E}\left[\Phi'_t S'_t \Phi_t'^{\intercal}\right] + W'. \tag{6b}$$

with the initial value matrix $P'_0 = Q'$ and the initial second moment matrix $S'_0 = \begin{bmatrix} X_0 & 0 \\ 0 & 0 \end{bmatrix}$. In particular, S'_t is the second moment of the augmented state

$$S'_t = \mathbb{E}\left[\begin{bmatrix} x_t \\ \hat{x}_t \end{bmatrix}\begin{bmatrix} x_t \\ \hat{x}_t \end{bmatrix}^{\intercal}\right],$$

while P'_t is related to the costs $\ell(x_t, u_t)$ through the relation

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \ell(x_t, u_t)\right] = \langle P'_T, S'_0\rangle + \sum_{t=0}^{T} \langle P'_t, W'\rangle.$$

We re-state some definitions from [15] that characterize asymptotic behavior of (4).

Definition 1 (Mean-square stability): System (4) is *mean-square stable (ms-stable)* if there exists finite $S'_\infty \in \mathbb{S}^{2n}_+$ such that $\lim_{k\to\infty} S'_t = S'_\infty$.

Definition 2 (Mean-square compensatability): System (4) is *mean-square compensatable (ms-compensatable)* if there exists a controller (3) which renders system (4) ms-stable.

Assumption. System (4) is *mean-square compensatable*. Define the linear operators $\Psi(\cdot), \Gamma(\cdot) : \mathbb{S}^n \to \mathbb{S}^n$ by

$$\Psi(M) := \mathbb{E}[\Phi_t'^{\intercal} M \Phi'_t]$$
$$\Gamma(N) := \mathbb{E}[\Phi'_t N \Phi_t'^{\intercal}].$$

[1] Due to the multiplicative noise, the state distribution is non-Gaussian even when all primitive distributions are Gaussian, so the Kalman filter is not necessarily the optimal state estimator. To our knowledge, it has not been demonstrated whether there exist conditions under which a nonlinear controller outperforms the optimal linear dynamic controller for the problem (2), but this is beyond our scope here.

These operators govern the second moment dynamics (6) as

$$P'_{t+1} = \Psi(P'_t) + Q',$$
$$S'_{t+1} = \Gamma(S'_t) + W'.$$

Accordingly, the ms-stability of the closed-loop system is characterized by the spectrum of the linear operator $\Psi(\cdot)$, or equivalently of $\Gamma(\cdot)$, since they share the same spectra. Namely, if the spectral radius $\rho(\Psi(\cdot)) < 1$ then the closed-loop system is ms-stable. From a computational viewpoint, the spectral radius of $\Psi(\cdot)$ is equal to the spectral radius of the associated matrix Ψ that satisfies $\Psi M = \mathbb{E}[\mathrm{vec}(\Phi_t'^{\intercal} M \Phi'_t)]$ for any symmetric matrix M, which takes the explicit form

$$\Psi = \Phi'^{\intercal} \otimes \Phi'^{\intercal} + \sum_{i=1}^{a} \sigma_{A,i}^2 A'^{\intercal}_{\Delta,i} \otimes A'^{\intercal}_{\Delta,i} + \sum_{i=1}^{b} \sigma_{B,i}^2 B'^{\intercal}_{\Delta,i} \otimes B'^{\intercal}_{\Delta,i} + \sum_{i=1}^{c} \sigma_{C,i}^2 C'^{\intercal}_{\Delta,i} \otimes C'^{\intercal}_{\Delta,i}, \tag{7}$$

where

$$\Phi' := \mathbb{E}[\Phi'_t] = \begin{bmatrix} A & BK \\ LC & F \end{bmatrix}, \quad A'_{\Delta,i} := \begin{bmatrix} A_{\Delta,i} & 0_n \\ 0_n & 0_n \end{bmatrix},$$
$$B'_{\Delta,i} := \begin{bmatrix} 0_n & B_{\Delta,i}K \\ 0_n & 0_n \end{bmatrix}, \quad C'_{\Delta,i} := \begin{bmatrix} 0_n & 0_n \\ LC_{\Delta,i} & 0_n \end{bmatrix}.$$

Note that a dual matrix Γ can be computed by dropping all transpose marks ("$\intercal$") in (7). With such ms-stability, the steady-state value matrix P' and the steady-state second moment S' are found by solving the discrete-time generalized Lyapunov equations

$$P' = \Psi(P') + Q', \tag{8a}$$
$$S' = \Gamma(S') + W'. \tag{8b}$$

With a slight abuse of notation, the performance criterion (2a) can be expressed and computed as

$$J(F, K, L) = \langle P', W' \rangle = \langle S', Q' \rangle. \tag{9}$$

which is finite only when (4) is ms-stable. Solving the generalized Lyapunov equations (8) to evaluate the performance (9) of a given policy will form a basic component of our policy iteration algorithm. Also in preparation for the policy iteration algorithm, given P', S', define

$$P = \begin{bmatrix} I_n & I_n \end{bmatrix} P' \begin{bmatrix} I_n & I_n \end{bmatrix}^{\intercal}, \tag{10a}$$
$$\hat{P} = \begin{bmatrix} 0_n & I_n \end{bmatrix} P' \begin{bmatrix} 0_n & I_n \end{bmatrix}^{\intercal}, \tag{10b}$$
$$S = \begin{bmatrix} I_n & -I_n \end{bmatrix} S' \begin{bmatrix} I_n & -I_n \end{bmatrix}^{\intercal}, \tag{10c}$$
$$\hat{S} = \begin{bmatrix} 0_n & I_n \end{bmatrix} S' \begin{bmatrix} 0_n & I_n \end{bmatrix}^{\intercal}. \tag{10d}$$

In particular, the second moment of state estimation error and state estimate are, respectively,

$$S = \lim_{t\to\infty} \mathbb{E}[(x_t - \hat{x}_t)(x_t - \hat{x}_t)^{\intercal}], \quad \hat{S} = \lim_{t\to\infty} \mathbb{E}[\hat{x}_t \hat{x}_t^{\intercal}],$$

and P, $\hat{P}$ have analogous interpretations in terms of the cost.

III. Optimal Linear Feedback Control Design

The optimal linear dynamic controller for the multiplicative noise problem (2) can be exactly computed by solving a set of coupled Riccati equations for estimation and control [15]. However, in the multiplicative noise setting *there is no separation between estimation and control*, so the optimal controller gains (F, K, L) must be jointly computed. Here we derive the equations in a form that facilitates the development of the policy iteration algorithm developed in the following section. Specifically, the optimal gains can be computed by solving the coupled nonlinear matrix Riccati equation

$$\mathcal{R}(X) = 0 \tag{11}$$

in $X = (P, \hat{P}, S, \hat{S})$ for the Riccati operator $\mathcal{R} : \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n \to \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n \times \mathbb{S}^n$ as

$$\mathcal{R}(X) := \begin{pmatrix} -P + \mathcal{G}_{xx}(X) - \mathcal{G}_{xu}(X)\mathcal{G}_{uu}^{-1}(X)\mathcal{G}_{ux}(X) \\ -\hat{P} + \mathcal{E}(X) + \mathcal{G}_{xu}(X)\mathcal{G}_{uu}^{-1}(X)\mathcal{G}_{ux}(X) \\ -S + \mathcal{H}_{xx}(X) - \mathcal{H}_{xy}(X)\mathcal{H}_{yy}^{-1}(X)\mathcal{H}_{yx}(X) \\ -\hat{S} + \mathcal{F}(X) + \mathcal{H}_{xy}(X)\mathcal{H}_{yy}^{-1}(X)\mathcal{H}_{yx}(X) \end{pmatrix} \tag{12}$$

where we define the operators $\mathcal{G}(X)$ and $\mathcal{H}(X)$ in (13), the closed-loop operators

$$\mathcal{E}(X) := (A - \mathcal{L}(X)C)^{\intercal} \hat{P} (A - \mathcal{L}(X)C)$$
$$\mathcal{F}(X) := (A + B\mathcal{K}(X)) \hat{S} (A + B\mathcal{K}(X))^{\intercal}$$

and the gain matrix operators

$$\mathcal{K}(X) := -\mathcal{G}_{uu}^{-1}(X)\mathcal{G}_{ux}(X), \tag{14a}$$
$$\mathcal{L}(X) := \mathcal{H}_{xy}(X)\mathcal{H}_{yy}^{-1}(X). \tag{14b}$$

$$\mathcal{G}(X) := \begin{bmatrix} Q_{xx} & Q_{xu} \\ Q_{ux} & Q_{uu} \end{bmatrix} + \begin{bmatrix} A^{\intercal}PA & A^{\intercal}PB \\ B^{\intercal}PA & B^{\intercal}PB \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{a} \sigma_{A,i}^2 A_{\Delta,i}^{\intercal} P A_{\Delta,i} & 0 \\ 0 & \sum_{i=1}^{b} \sigma_{B,i}^2 B_{\Delta,i}^{\intercal} P B_{\Delta,i} \end{bmatrix}$$
$$+ \begin{bmatrix} \sum_{i=1}^{a} \sigma_{A,i}^2 A_{\Delta,i}^{\intercal} \hat{P} A_{\Delta,i} + \sum_{i=1}^{c} \sigma_{C,i}^2 C_{\Delta,i}^{\intercal} \mathcal{L}(X)^{\intercal} \hat{P} \mathcal{L}(X) C_{\Delta,i} & 0 \\ 0 & \sum_{i=1}^{b} \sigma_{B,i}^2 B_{\Delta,i}^{\intercal} \hat{P} B_{\Delta,i} \end{bmatrix} \tag{13a}$$

$$\mathcal{H}(X) := \begin{bmatrix} W_{xx} & W_{xy} \\ W_{yx} & W_{yy} \end{bmatrix} + \begin{bmatrix} ASA^{\intercal} & ASC^{\intercal} \\ CSA^{\intercal} & CSC^{\intercal} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{a} \sigma_{A,i}^2 A_{\Delta,i} S A_{\Delta,i}^{\intercal} & 0 \\ 0 & \sum_{i=1}^{c} \sigma_{C,i}^2 C_{\Delta,i} S C_{\Delta,i}^{\intercal} \end{bmatrix}$$
$$+ \begin{bmatrix} \sum_{i=1}^{a} \sigma_{A,i}^2 A_{\Delta,i} \hat{S} A_{\Delta,i}^{\intercal} + \sum_{i=1}^{b} \sigma_{B,i}^2 B_{\Delta,i} \mathcal{K}(X) \hat{S} \mathcal{K}(X)^{\intercal} B_{\Delta,i}^{\intercal} & 0 \\ 0 & \sum_{i=1}^{c} \sigma_{C,i}^2 C_{\Delta,i} \hat{S} C_{\Delta,i}^{\intercal} \end{bmatrix} \tag{13b}$$

Note that $\mathcal{G}$ and $\mathcal{H}$ have a block 2×2 structure whose blocks are referred to with the same subscripts as in Q and W as appropriate. These expressions can be derived from the matrix minimum principle [26] as in [15]. Notice that the gains $\mathcal{K}(X)$ and $\mathcal{L}(X)$ depend only on the $\mathcal{G}_{ux}$, $\mathcal{G}_{uu}$, $\mathcal{H}_{xy}$, $\mathcal{H}_{yy}$ blocks of $\mathcal{G}$ and $\mathcal{H}$. Also notice that, unlike the $\mathcal{G}_{xx}$ and $\mathcal{H}_{xx}$ blocks, $\mathcal{G}_{ux}$, $\mathcal{G}_{uu}$, $\mathcal{H}_{xy}$, $\mathcal{H}_{yy}$ are not specified in terms of $\mathcal{K}(X)$ and $\mathcal{L}(X)$. Hence, $\mathcal{K}(X)$ and $\mathcal{L}(X)$ can be computed explicitly in terms of X, and consequently so can $\mathcal{G}(X)$ and $\mathcal{H}(X)$.

The optimal controller is

$$\big(A+B\mathcal{K}(X^*)-\mathcal{L}(X^*)C,\mathcal{K}(X^*),\mathcal{L}(X^*)\big)$$

where X^* solves (12) [15]. With this controller, the cost and second moment matrices satisfy

$$P'=\begin{bmatrix}P+\hat{P} & -\hat{P}\\ -\hat{P} & \hat{P}\end{bmatrix},\quad S'=\begin{bmatrix}S+\hat{S} & \hat{S}\\ \hat{S} & \hat{S}\end{bmatrix},$$

and thus the optimal controller achieves the optimal cost

$$\begin{aligned}J^*&=\langle Q_{xx},S\rangle+\left\langle\begin{bmatrix}I\\K\end{bmatrix}^{\intercal}\begin{bmatrix}Q_{xx} & Q_{xu}\\ Q_{ux} & Q_{uu}\end{bmatrix}\begin{bmatrix}I\\K\end{bmatrix},\hat{S}\right\rangle\\&=\langle W_{xx},P\rangle+\left\langle\begin{bmatrix}I\\-L\end{bmatrix}\begin{bmatrix}W_{xx} & W_{xy}\\ W_{yx} & W_{yy}\end{bmatrix}\begin{bmatrix}I\\-L\end{bmatrix}^{\intercal},\hat{P}\right\rangle.\end{aligned}$$

When the multiplicative noise terms are zero, the coupled Riccati equations reduce to the familiar two decoupled standard algebraic Riccati equations for optimal linear quadratic control and state estimation, which can be solved via several well-known methods such as the dynamic programming techniques of policy iteration and value iteration [27], convex semidefinite programming [28], and specialized direct linear algebraic methods [14]. In contrast, the only known algorithm for solving the coupled Riccati equations (12) is a value iteration-type algorithm, described in [15]. This turns the Riccati equation (11) into a recursive update according to

$$X^{k+1}=X^k+\mathcal{R}(X^k). \tag{15}$$

Convergence of (15) was proved in [15] using a homotopic continuation argument that continuously deforms a multiplicative noise-free version of the problem (for which convergence of the value iteration algorithm is well-known) to the original multiplicative noise-driven problem. The ms-compensatability test described in [15] is simply that (15) converges.

IV. Policy Iteration for Multiplicative Noise Output Feedback Control

In this section, we describe a novel policy iteration algorithm to solve the coupled Riccati equations (12). Policy iteration is a well-known method for computing optimal policies in the full state-feedback setting, originating with [29], [30] for linear quadratic problems. It consists of two steps: (1) policy evaluation, where the value of the current policy is computed from the dynamics and cost; and (2) policy improvement, where the current policy is improved based on the policy evaluation. However, policy iteration is far less developed in the context of output feedback control problems or POMDPs.

The form of the coupled Riccati equations in (12) suggests a policy iteration algorithm analogous to the full state-feedback setting. The operators $\mathcal{G}(X)$ and $\mathcal{H}(X)$ play a role analogous to the state-action value function (also called the Q-function) in reinforcement learning. In particular, for a given X, the gain operators (14) can be viewed as an update to the control and estimator gains that improves the corresponding value. Combining this with a policy evaluation step from solving the generalized Lyapunov equations (8) leads to a policy iteration algorithm for the multiplicative noise output feedback control problem detailed in Algorithm 1. Note that the initial policy must be ms-stabilizing so that solutions to (8) exist; this is a standard assumption in policy iteration [29], [30] and policy optimization algorithms [25]. Such a policy may be known from special structural knowledge of the system e.g. open-loop ms-stability or notions of ms-passivity, or can be found by iterating (15) a sufficient number of times starting from $X=(0,0,0,0)$. Also note that due to the assumptions $Q\succ 0$ and $W\succ 0$, the inverses $\mathcal{G}_{uu}^{-1}(X)$ and $\mathcal{H}_{yy}^{-1}(X)$ remain well-defined throughout the iterations of Algorithm 1.

Algorithm 1 Policy iteration for optimal dynamic output feedback of linear systems with multiplicative noise

Input: ms-stabilizing policy $(A+BK^0-L^0C,K^0,L^0)$, convergence threshold $\epsilon>0$.

1: Initialize $X^0=(0,0,0,0)$, $X^1=(\infty,\infty,\infty,\infty)$, $k=0$.

2: **while** $\|X^{k+1}-X^k\|>\epsilon$ **do**

3: **Policy Evaluation:** Compute the value $X^k=(P^k,\hat{P}^k,S^k,\hat{S}^k)$ of the current policy $(A+BK^k-L^kC,K^k,L^k)$ by finding the solutions P'^k, S'^k to the Lyapunov equations (8) and using the relations (10).

4: **Policy Improvement:** Update the policy according to

$$K^{k+1}=\mathcal{K}(X^k),\quad L^{k+1}=\mathcal{L}(X^k)$$

5: $k\leftarrow k+1$

Output: Nearly optimal policy $(A+BK^k-L^kC,K^k,L^k)$

V. Numerical Experiments

The algorithms were implemented in Python and executed on a desktop PC with a quad-core Intel i7 6700K 4.0GHz CPU and 16GB RAM; no GPU computing was utilized. Code supporting §V is available in the GitHub repository

https://github.com/TSummersLab/policy-iteration-mlqc

A. Pendulum system

As a first example, we examined a particular system representing a forward Euler discretization of the continuous-time dynamics of a pendulum with torque actuation and control-dependent noise. The first and second states represent the angular position and velocity, respectively. The system dimensions were $n=2,m=1,p=1,b=1$, and $a=c=0$, indicating A_t and C_t were unaffected by multiplicative noise.

The problem data were

$$A = \begin{bmatrix} 1.0 & 0.1 \\ 1.0 & 0.95 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix}, \quad B_{\Delta,1} = \begin{bmatrix} 0 \\ 1.0 \end{bmatrix},$$
$$C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \sigma_{B,1} = 1.$$

The penalty and additive noise covariance matrices were chosen as $Q = I$ and $W = \mathrm{diag}(0, 0.01, 0.001)$ respectively. The multiplicative noise value $\sigma_{B,1}$ was varied by scaling by a noise level factor $\eta \in [0, 1]$. Note that the system was open-loop ms-stable; therefore we selected as our initial policy the open-loop policy $(A, 0, 0)$. The algorithms were terminated once the convergence criterion $\|X^k - X^{k-1}\| \leq 10^{-12}$ was achieved.

Figure 1 compares the policy iteration Algorithm 1 with the value iteration method of [15]. The metric used for comparison was

$$e^k = \max\{\delta(P^k), \delta(\hat{P}^k), \delta(S^k), \delta(\hat{S}^k)\},$$
$$\text{where} \quad \delta(M^k) = \|M^k - M^*\| / \|M^0 - M^*\|,$$

which measures the error of the current solution relative to that of the initial solution. Iteration count refers to the index k of each algorithm, while wall clock time refers to the total elapsed time from the beginning of the algorithm to the end of the current iteration; the wall clock time accounts for per-iteration computation time while iteration count does not.

The results demonstrate two important issues which arise and are addressed by the proposed policy iteration algorithm:

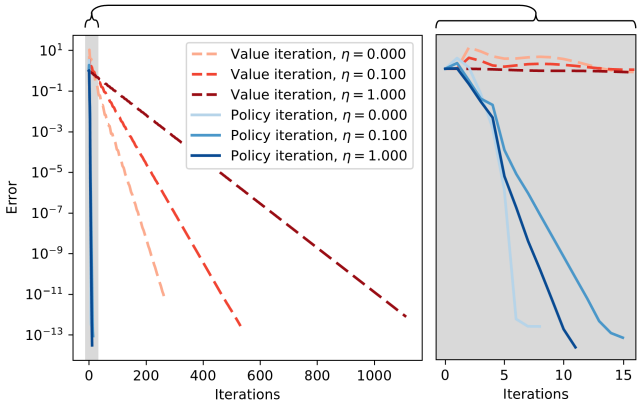
1) The value iteration algorithm converges much more slowly compared to policy iteration, both in the multiplicative noise-free setting ($\eta = 0$) as well as in the multiplicative noise setting ($\eta > 0$).
2) The presence of multiplicative noise causes value iteration to converge significantly more slowly, while the proposed policy iteration suffers a much less dramatic slowdown as the noise level η increases.

These observations apply both in terms of iteration count and wall clock time.

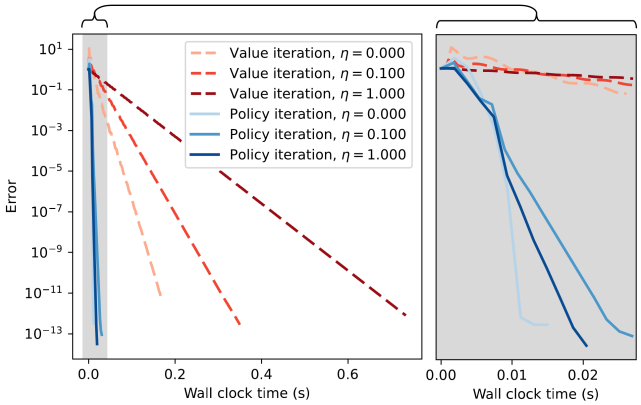
B. Random systems

Next, we report results for systems with randomly generated parameters. The dimensions of each problem were chosen as $n = 2, m = 1, p = 1$ and $a = b = c = 1$. The entries of A, B C, $A_{\Delta,1}$, $B_{\Delta,1}$, $C_{\Delta,1}$ were randomly drawn from a standard normal distribution. A was scaled to have $\rho(A)$ drawn uniformly random between 0 and 1. The variances $\sigma^2_{A,1}$, $\sigma^2_{B,1}$, $\sigma^2_{C,1}$ were drawn uniform randomly between 0 and 1, then scaled such that the open-loop system with $(A, 0, 0)$ had $\sqrt{\rho(\Psi)} = 1$. Finally, the variances $\sigma^2_{A,1}$, $\sigma^2_{B,1}$, $\sigma^2_{C,1}$ were scaled by a factor η drawn uniform randomly between 0 and 1. Hence, the systems were always open-loop ms-stable by construction; therefore we selected as our initial policy the open-loop policy $(A, 0, 0)$. The penalty and additive noise covariance matrices were chosen as $Q = I$ and $W = 0.01I$ respectively.

The results are plotted in Figure 2, where each solved problem instance is represented by a scatter point. It is evident



(a) Relative error vs iteration count.



(b) Relative error vs elapsed wall clock time.

Fig. 1: Performance of value iteration and policy iteration for the pendulum system with various noise levels η.

that the proposed policy iteration algorithm converges in far fewer total iterations on almost every problem instance. Due to greater per-iteration computation costs, the benefit in time elapsed is not as large for the policy iteration algorithm overall. However, per-iteration costs can be significantly reduced by using specialized solvers for the generalized Lyapunov equations based on e.g. alternating-direction implicit preconditioning and Krylov subspaces [31] and other schemes that take advantage of sparsity in the linear matrix equation. Nevertheless, the greatest benefit was achieved when the multiplicative noise was high relative to the maximum uncertainty that allows mean-square compensatability, which can be seen in the trend of Figure 2, mirroring the results observed for the pendulum system in §V-A.

The observations of §V-A and §V-B together suggest that the least computationally costly method of solving (2) may be problem-instance dependent; a reliable indicator of when our policy iteration Algorithm 1 will outperform the value iteration (15) remains to be discovered.

VI. Conclusions

We proposed a policy iteration algorithm for multiplicative noise optimal output feedback control design that, empirically, converges far faster than value iteration.

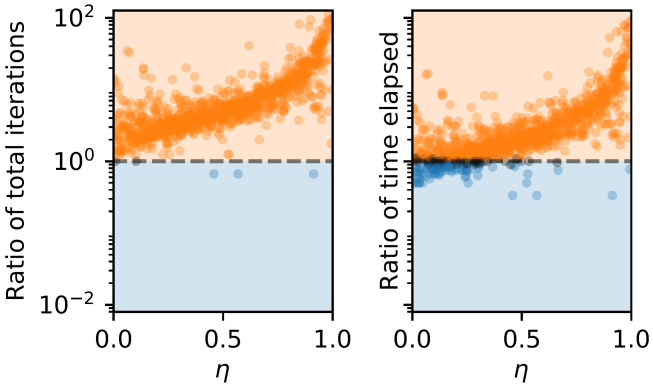


Fig. 2: Ratio of total iterations and time elapsed using value iteration to that using policy iteration.

Given the intimate connections between policy iteration and the Newton method, we believe there is a connection between our proposed algorithm and the Newton method (or some variation thereof) applied to the coupled Riccati equations (12). Interpreting our algorithm through a Newton lens may facilitate theoretical convergence results, in particular a quadratic rate of convergence as in [29], [30], [32]. Furthermore, establishing a rate of convergence for the value iteration algorithm (15), which we conjecture is a linear rate, is, to our knowledge, also an open problem whose solution is required to establish theoretical relative performance claims. We leave this for future work.

Finally, our policy iteration algorithm opens the door to investigate novel approximate policy iteration algorithms that use input-output data to approximately execute policy evaluation and policy improvement steps to solve POMDPs. For example, it may be possible to use input-output data generated from a given policy to form a least-squares estimate of the cost and covariance operators, and then perform an approximate policy improvement step based on this estimate. We will also explore this and other variations in future work.

REFERENCES

[1] W. M. Wonham, "Optimal stationary control of a linear system with state-dependent noise," *SIAM Journal on Control*, vol. 5, no. 3, pp. 486–500, 1967.

[2] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1453–1464, 2004.

[3] N. E. Du Toit and J. W. Burdick, "Robot motion planning in dynamic, uncertain environments," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 101–115, 2011.

[4] J. L. Lumley, *Stochastic Tools in Turbulence*. Courier Corporation, 2007.

[5] A. J. Majda, I. Timofeyev, and E. V. Eijnden, "Models for stochastic climate prediction," *Proceedings of the National Academy of Sciences*, vol. 96, no. 26, pp. 14 687–14 691, 1999.

[6] E. Todorov, "Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system," *Neural computation*, vol. 17, no. 5, pp. 1084–1108, 2005.

[7] M. Breakspear, "Dynamic models of large-scale brain activity," *Nature neuroscience*, vol. 20, no. 3, p. 340, 2017.

[8] J. A. Primbs, "Portfolio optimization applications of stochastic receding horizon control," in *American Control Conference*. IEEE, 2007, pp. 1811–1816.

[9] Y. Guo and T. H. Summers, "A performance and stability analysis of low-inertia power grids with stochastic system inertia," in *2019 American Control Conference, ACC 2019, Philadelphia, PA, USA, July 10-12, 2019*. IEEE, 2019, pp. 1965–1970.

[10] D. E. Gustafson and J. L. Speyer, "Design of linear regulators for nonlinear stochastic systems," *Journal of Spacecraft and Rockets*, vol. 12, no. 6, pp. 351–358, 1975.

[11] B. Gravell and T. Summers, "Robust learning-based control via bootstrapped multiplicative noise," in *Proceedings of Machine Learning Research*, vol. 120. The Cloud: PMLR, 2020, pp. 599–607.

[12] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[14] S. Bittanti, A. J. Laub, and J. C. Willems, *The Riccati Equation*. Springer Science & Business Media, 2012.

[15] W. L. De Koning, "Compensatability and optimal compensation of systems with white parameters," *IEEE Transactions on Automatic Control*, vol. 37, no. 5, pp. 579–588, May 1992.

[16] M. L. Puterman and S. L. Brumelle, "On the convergence of policy iteration in stationary dynamic programming," *Mathematics of Operations Research*, vol. 4, no. 1, pp. 60–69, 1979.

[17] D. Bertsekas, "Lessons from alphazero for optimal, model predictive, and adaptive control," *arXiv preprint arXiv:2108.10315*, 2021.

[18] D. Kleinman, "On the stability of linear stochastic systems," *IEEE Transactions on Automatic Control*, vol. 14, no. 4, pp. 429–430, August 1969.

[19] T. Damm, *Rational matrix equations in stochastic control*. Springer Science & Business Media, 2004, vol. 297.

[20] F. L. Lewis and K. G. Vamvoudakis, "Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 14–25, 2011.

[21] S. A. Asad Rizvi and Z. Lin, "Output feedback reinforcement learning control for the continuous-time linear quadratic regulator problem," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 3417–3422.

[22] F. Adib Yaghmaie, F. Gustafsson, and L. Ljung, "Linear quadratic control using model-free reinforcement learning," *IEEE Transactions on Automatic Control*, pp. 1–1, 2022.

[23] E. A. Hansen, "Solving POMDPs by searching in policy space," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, p. 211–219.

[24] W. Gao, Y. Jiang, Z.-P. Jiang, and T. Chai, "Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming," *Automatica*, vol. 72, pp. 37–45, 2016.

[25] B. Gravell, P. M. Esfahani, and T. Summers, "Learning optimal controllers for linear systems with multiplicative noise via policy gradient," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5283–5298, 2021.

[26] M. Athans, "The matrix minimum principle," *Information and Control*, vol. 11, no. 5, pp. 592–606, 1967.

[27] D. Bertsekas, *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012, vol. 1.

[28] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*. SIAM, 1994, vol. 15.

[29] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114–115, 1968.

[30] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," *IEEE Transactions on Automatic Control*, vol. 16, no. 4, pp. 382–384, 1971.

[31] T. Damm, "Direct methods and ADI-preconditioned krylov subspace methods for generalized lyapunov equations," *Numerical Linear Algebra with Applications*, vol. 15, no. 9, pp. 853–871, 2008.

[32] T. Damm and D. Hinrichsen, "Newton's method for a rational matrix equation occurring in stochastic control," *Linear Algebra and its Applications*, vol. 332-334, pp. 81–109, 2001.