A Two-Layer Solution Scheme for Bayesian Reinforcement Learning and a Reduced Case in Dynamic Portfolio Selection

Duan Li
dli226@cityu.edu.hk

City University of Hong Kong

Joint work with Xin Huang

National University of Singapore, Aug 7, 2019

Outline

1. Bayesian Reinforcement Learning (RL) Problems with Unknown System Parameters
   1.1 Introduction
   1.2 Problem Formulation: General Form and LQG with Unknown Gain
   1.3 Progressive Hedging Algorithm (PHA): Background
   1.4 Two-layer Solution Scheme: Naive and Revised PHA Methods
   1.5 Illustrative Example
   1.6 Comparison with Other RL Methods
   1.7 Summary on Part 1

Outline

1. Bayesian Reinforcement Learning (RL) Problems with Unknown System Parameters
   1.1 Introduction
   1.2 Problem Formulation: General Form and LQG with Unknown Gain
   1.3 Progressive Hedging Algorithm (PHA): Background
   1.4 Two-layer Solution Scheme: Naive and Revised PHA Methods
   1.5 Illustrative Example
   1.6 Comparison with Other RL Methods
   1.7 Summary on Part 1
2. Quadratic Dis-utility Portfolio Selection under Lack of Market Information
   1.1 Market Settings: Unknown Mean and Covariance
   1.2 Problem Formulation
   1.3 Solution by Dynamic Programming
   1.4 Special Case: Mean is Unknown While Variance is Known
   1.5 Experimental Results

Part 1: Bayesian Reinforcement Learning (BRL) Problems with Unknown System Parameters

Introduction

The exploitation-exploration tradeoff is crucial and prominent in reinforcement learning (RL), especially under non-episodic setting.

Introduction

The exploitation-exploration tradeoff is crucial and prominent in reinforcement learning (RL), especially under non-episodic setting.

Bayesian RL has been emerged to tackle such cases, through augmenting the state space with probabilistic belief on those unknown parameters.

Introduction

The exploitation-exploration tradeoff is crucial and prominent in reinforcement learning (RL), especially under non-episodic setting.

Bayesian RL has been emerged to tackle such cases, through augmenting the state space with probabilistic belief on those unknown parameters.

Actually, a similar subject termed dual control was invented long ago in Feldbaum (1961). However, to find an optimal policy is always analytically intractable, simply because of the very nonlinearity of inference, even under the Gaussian assumption (Aoki, 1967).

Introduction

The exploitation-exploration tradeoff is crucial and prominent in reinforcement learning (RL), especially under non-episodic setting.

Bayesian RL has been emerged to tackle such cases, through augmenting the state space with probabilistic belief on those unknown parameters.

Actually, a similar subject termed dual control was invented long ago in Feldbaum (1961). However, to find an optimal policy is always analytically intractable, simply because of the very nonlinearity of inference, even under the Gaussian assumption (Aoki, 1967).

As our first attempt, a discrete-time LQG problem with fixed but unknown gain is solved by our novel two-layer solution scheme based on progressive hedging algorithm (PHA) developed by Rockafellar and Wets (1991).

Introduction

The exploitation-exploration tradeoff is crucial and prominent in reinforcement learning (RL), especially under non-episodic setting.

Bayesian RL has been emerged to tackle such cases, through augmenting the state space with probabilistic belief on those unknown parameters.

Actually, a similar subject termed dual control was invented long ago in Feldbaum (1961). However, to find an optimal policy is always analytically intractable, simply because of the very nonlinearity of inference, even under the Gaussian assumption (Aoki, 1967).

As our first attempt, a discrete-time LQG problem with fixed but unknown gain is solved by our novel two-layer solution scheme based on progressive hedging algorithm (PHA) developed by Rockafellar and Wets (1991).

The basic idea of our method is to separate reducible and irreducible uncertainties into two different layers: first decompose and then conquer.

Problem Formulation: General Bayesian RL Problems

State dynamics

$$x_{t+1} = f_t(x_t, u_t, \xi_t|\theta), \;\; t = 0, 1, \cdots, T-1,$$

where $x_t \in \mathbb{R}^n$ is the state (perfectly observed) at time t with x_0 given, T is the finite time horizon, and $u_t \in \mathbb{R}^m$ is the control.

Problem Formulation: General Bayesian RL Problems

State dynamics

$$x_{t+1} = f_t(x_t, u_t, \xi_t|\theta), \;\; t = 0, 1, \cdots, T-1,$$

where $x_t \in \mathbb{R}^n$ is the state (perfectly observed) at time t with x_0 given, T is the finite time horizon, and $u_t \in \mathbb{R}^m$ is the control.

The state transition function f_t is affected by a fixed parameter $\theta \in \mathbb{R}^d$ which is, however, unknown to the agent, who only has some prior belief on its distribution $p_0(\theta)$.

Problem Formulation: General Bayesian RL Problems

State dynamics

$$x_{t+1} = f_t(x_t, u_t, \xi_t|\theta), \;\; t = 0, 1, \cdots, T-1,$$

where $x_t \in \mathbb{R}^n$ is the state (perfectly observed) at time t with x_0 given, T is the finite time horizon, and $u_t \in \mathbb{R}^m$ is the control.

The state transition function f_t is affected by a fixed parameter $\theta \in \mathbb{R}^d$ which is, however, unknown to the agent, who only has some prior belief on its distribution $p_0(\theta)$.

Noise $\xi_t \in \mathbb{R}^n$ is the i.i.d. Gaussian noise with mean 0 and covariance Σ_ξ, and is independent of θ.

Problem Formulation: General Bayesian RL Problems

State dynamics

$$x_{t+1} = f_t(x_t, u_t, \xi_t|\theta), \ \ t = 0, 1, \cdots, T-1,$$

where $x_t \in \mathbb{R}^n$ is the state (perfectly observed) at time t with x_0 given, T is the finite time horizon, and $u_t \in \mathbb{R}^m$ is the control.

The state transition function f_t is affected by a fixed parameter $\theta \in \mathbb{R}^d$ which is, however, unknown to the agent, who only has some prior belief on its distribution $p_0(\theta)$.

Noise $\xi_t \in \mathbb{R}^n$ is the i.i.d. Gaussian noise with mean 0 and covariance Σ_ξ, and is independent of θ.

In our setting, we are able to observe the state x_t when it is realized at time t, while θ is not observable.

Problem Formulation: General Bayesian RL Problems

The goal of the agent is to minimize the expected total cost,

$$\mathcal{L}(\boldsymbol{u}|x_0, p_0(\theta)) = \mathbb{E}_{\theta,\xi}\left[g_T(x_T) + \sum_{t=0}^{T-1} g_t(x_t, u_t, \xi_t) \middle| x_0, p_0(\theta) \right]$$

over all admissible *feedback* policies,

$$\boldsymbol{u} = (u_0, u_1, \cdots, u_{T-1})' \in \mathcal{U}_0 \times \mathcal{U}_1 \times \cdots \times \mathcal{U}_{T-1}.$$

We also assume that when fixing θ at θ_i, $\mathcal{L}(\cdot|x_0, \delta(\theta = \theta_i))$ is convex w.r.t. $\boldsymbol{u}$, as required by PHA. In the following we denote $\delta(\theta = \theta_i)$ by δ_i.

Problem Formulation: General Bayesian RL Problems

The goal of the agent is to minimize the expected total cost,

$$\mathcal{L}(\boldsymbol{u}|x_0, p_0(\theta)) = \mathbb{E}_{\theta,\xi}\left[g_T(x_T) + \sum_{t=0}^{T-1} g_t(x_t, u_t, \xi_t)\middle| x_0, p_0(\theta)\right]$$

over all admissible *feedback* policies,

$$\boldsymbol{u} = (u_0, u_1, \cdots, u_{T-1})' \in \mathcal{U}_0 \times \mathcal{U}_1 \times \cdots \times \mathcal{U}_{T-1}.$$

We also assume that when fixing θ at θ_i, $\mathcal{L}(\cdot|x_0, \delta(\theta = \theta_i))$ is convex w.r.t. $\boldsymbol{u}$, as required by PHA. In the following we denote $\delta(\theta = \theta_i)$ by δ_i.

The uncertainty from θ is due to the lack of knowledge of the agent, and such an uncertainty is reducible by learning, whereas the randomness incurred by system disturbance (ξ_t's) is not.

Problem Formulation: General Bayesian RL Problems

Information set is defined as

$$I^t = \{x_0, x_1, \ldots, x_t, u_0, u_1, \ldots, u_{t-1}\},\ \ t = 1, \cdots, T-1,$$

and $I^0 = \{p_0(\theta), x_0\}$.

Problem Formulation: General Bayesian RL Problems

Information set is defined as

$$I^t = \{x_0, x_1, \ldots, x_t, u_0, u_1, \ldots, u_{t-1}\},\ t = 1, \cdots, T-1,$$

and $I^0 = \{p_0(\theta), x_0\}$.

Applying Bayesian law leads to the recursive relationship between $p_{t+1}(\theta)$ and $p_t(\theta)$,

$$p_t(\theta|I^t) = p_t(\theta|x_t, u_{t-1}, I^{t-1}) \propto \psi(x_t|\theta, x_{t-1}, u_{t-1}) \times p_{t-1}(\theta|I^{t-1}), \qquad (1)$$

for $t = 1, \cdots, T-1$, where $\psi(x_t|\theta, x_{t-1}, u_{t-1})$ is the conditional density of x_t and $p_0(\theta|I^0) = p_0(\theta)$ as prior belief given. We denote the posterior distribution of θ at time t, $p_t(\theta|I^t)$, by $p_t(\theta)$.

Problem Formulation: General Bayesian RL Problems

By augmenting the original state space with belief of model parameters, the Bellman equation for the optimal value function J_t of the augmented system is given by

$$J_t(x_t, p_t(\theta)) = \min_{u_t} \ \mathbb{E}_{\theta,\xi_t}\Big[g_t(x_t, u_t, \xi_t) + J_{t+1}\big(x_{t+1}, p_{t+1}(\theta)\big)\Big| x_t, p_t(\theta)\Big], \quad (2)$$

for $t = 0, \cdots, T-1$ with terminal condition $J_T(x_T, p_T(\theta)) = g_T(x_T)$.

Problem Formulation: General Bayesian RL Problems

By augmenting the original state space with belief of model parameters, the Bellman equation for the optimal value function J_t of the augmented system is given by

$$J_t(x_t, p_t(\theta)) = \min_{u_t} \mathbb{E}_{\theta,\xi_t}\Big[g_t(x_t, u_t, \xi_t) + J_{t+1}\big(x_{t+1}, p_{t+1}(\theta)\big)\Big|x_t, p_t(\theta)\Big], \quad (2)$$

for $t = 0, \cdots, T-1$ with terminal condition $J_T(x_T, p_T(\theta)) = g_T(x_T)$.

Theoretically, we could handle Bayesian RL problems by DP, to achieve optimal balance between exploitation and exploration.

Problem Formulation: General Bayesian RL Problems

By augmenting the original state space with belief of model parameters, the Bellman equation for the optimal value function J_t of the augmented system is given by

$$J_t(x_t, p_t(\theta)) = \min_{u_t} \mathbb{E}_{\theta,\xi_t}\Big[g_t(x_t, u_t, \xi_t) + J_{t+1}\big(x_{t+1}, p_{t+1}(\theta)\big)\Big| x_t, p_t(\theta)\Big], \quad (2)$$

for $t = 0, \cdots, T-1$ with terminal condition $J_T(x_T, p_T(\theta)) = g_T(x_T)$.

Theoretically, we could handle Bayesian RL problems by DP, to achieve optimal balance between exploitation and exploration.

The resulting optimal policy, if we are able to derive it, includes an essential feature of active learning, in the sense that taking into account in (2) the effect of future actions and beliefs via conditional planning before we actually observe the future states.

Problem Formulation: General Bayesian RL Problems

By augmenting the original state space with belief of model parameters, the Bellman equation for the optimal value function J_t of the augmented system is given by

$$J_t(x_t, p_t(\theta)) = \min_{u_t} \mathbb{E}_{\theta,\xi_t}\Big[g_t(x_t, u_t, \xi_t) + J_{t+1}\big(x_{t+1}, p_{t+1}(\theta)\big)\Big| x_t, p_t(\theta)\Big], \quad (2)$$

for $t = 0, \cdots, T-1$ with terminal condition $J_T(x_T, p_T(\theta)) = g_T(x_T)$.

Theoretically, we could handle Bayesian RL problems by DP, to achieve optimal balance between exploitation and exploration.

The resulting optimal policy, if we are able to derive it, includes an essential feature of active learning, in the sense that taking into account in (2) the effect of future actions and beliefs via conditional planning before we actually observe the future states.

Unfortunately, due to the high nonlinearity of (1), solving the Bellman equation (2) in general is impossible, at least intractable.

Problem Formulation: LQG with Unknown Gain

We are now focusing on the following LQG problem with unknown gain,

$$
\begin{aligned}
(\mathcal{P}) \qquad & \min_{\boldsymbol{u}} \ \mathbb{E}_{\theta,\xi}\left[\frac{1}{2}x_T' Q x_T + \sum_{t=0}^{T-1}\left(\frac{1}{2}x_t' Q x_t + \frac{1}{2}u_t' R u_t\right)\Big| x_0, p_0(\theta)\right] \\
& \text{s.t.} \ \ x_{t+1} = A x_t + B(\theta) u_t + \xi_t, \ \ t = 0, 1, \cdots, T-1,
\end{aligned}
$$

where A is given, and the gain matrix B is an unknown constant matrix dictated by an unknown parameter θ.

Problem Formulation: LQG with Unknown Gain

We are now focusing on the following LQG problem with unknown gain,

$$(\mathcal{P}) \qquad \min_{\boldsymbol{u}} \ \mathbb{E}_{\theta,\xi}\left[\frac{1}{2}x_T' Q x_T + \sum_{t=0}^{T-1}\left(\frac{1}{2}x_t' Q x_t + \frac{1}{2}u_t' R u_t\right)\Big| x_0, p_0(\theta)\right]$$

$$\text{s.t.} \ \ x_{t+1} = Ax_t + B(\theta)u_t + \xi_t, \ \ t = 0, 1, \cdots, T-1,$$

where A is given, and the gain matrix B is an unknown constant matrix dictated by an unknown parameter θ.

We assume further that θ takes one of N possible values, $\theta_1, \theta_2, \ldots, \theta_N$, with prior belief $p_0(\theta) = (p_{01}, p_{02}, \ldots, p_{0N})'$, where $p_{0i} = \mathbb{P}(\theta = \theta_i \mid I^0)$, $i = 1, \cdots, N$. For simplicity we set $B(\theta_i) = B_i$ for all i.

Problem Formulation: LQG with Unknown Gain

We are now focusing on the following LQG problem with unknown gain,

$$(\mathcal{P}) \qquad \min_{\boldsymbol{u}} \ \mathbb{E}_{\theta,\xi}\left[\frac{1}{2}x_T' Q x_T + \sum_{t=0}^{T-1}\left(\frac{1}{2}x_t' Q x_t + \frac{1}{2}u_t' R u_t\right)\Big| x_0, p_0(\theta)\right]$$

$$\text{s.t.} \ \ x_{t+1} = Ax_t + B(\theta)u_t + \xi_t, \ \ t = 0, 1, \cdots, T-1,$$

where A is given, and the gain matrix B is an unknown constant matrix dictated by an unknown parameter θ.

We assume further that θ takes one of N possible values, $\theta_1, \theta_2, \ldots, \theta_N$, with prior belief $p_0(\theta) = (p_{01}, p_{02}, \ldots, p_{0N})'$, where $p_{0i} = \mathbb{P}(\theta = \theta_i \mid I^0)$, $i = 1, \cdots, N$. For simplicity we set $B(\theta_i) = B_i$ for all i.

Denote $\mathbb{P}(\theta = \theta_i \mid I^{t+1}) = p_{(t+1)i}$. Bayesian law gives for $t = 0, \cdots, T-1$,

$$p_{(t+1)i} = \frac{\phi(x_{t+1}; Ax_t + B_i u_t, \Sigma_\xi)p_{ti}}{\sum_{j=1}^{N}[\phi(x_{t+1}; Ax_t + B_j u_t, \Sigma_\xi)p_{tj}]} \tag{3}$$

which is very nonlinear with respect to the realized x_{t+1}.

PHA: Background

Multistage decision-making problem is given by

$$(\mathcal{P}) \ \min \ F(\mathbf{u}) = \mathbb{E}[f_s(\mathbf{u}(s))] = \sum_{s \in S} p_s f_s(\mathbf{u}(s)) \text{ over all } \mathbf{u} \in \mathcal{C} \cap \mathcal{N},$$

where $\mathcal{C}$ is feasible set and $\mathcal{N}$ is non-anticipative constraint, and f_s is the scenario subproblem (with realizations of randomness) under scenario s.

PHA: Background

Multistage decision-making problem is given by

$$(\mathcal{P}) \;\; \min \;\; F(\mathbf{u}) = \mathbb{E}[f_s(\mathbf{u}(s))] = \sum_{s \in S} p_s f_s(\mathbf{u}(s)) \text{ over all } \mathbf{u} \in \mathcal{C} \cap \mathcal{N},$$

where $\mathcal{C}$ is feasible set and $\mathcal{N}$ is non-anticipative constraint, and f_s is the scenario subproblem (with realizations of randomness) under scenario s. While dynamic programming is a *time-decomposition* solution scheme, PHA is a *scenario-decomposition* solution scheme.

PHA: Background

Multistage decision-making problem is given by

$$(\mathcal{P}) \ \min \ F(\mathbf{u}) = \mathbb{E}[f_s(\mathbf{u}(s))] = \sum_{s \in S} p_s f_s(\mathbf{u}(s)) \text{ over all } \mathbf{u} \in \mathcal{C} \cap \mathcal{N},$$

where $\mathcal{C}$ is feasible set and $\mathcal{N}$ is non-anticipative constraint, and f_s is the scenario subproblem (with realizations of randomness) under scenario s.

While dynamic programming is a *time-decomposition* solution scheme, PHA is a *scenario-decomposition* solution scheme.

Assume that the whole scenario is realized at $s \in S$, we then consider the following scenario sub-problems (for every $s \in S$):

$$(\mathcal{P}_s^{\nu}) \qquad \mathbf{u}^0 = \arg\min \quad f_s(u) \text{ over all } u \in C_s.$$

PHA: Background

Multistage decision-making problem is given by

$$(\mathcal{P}) \ \min \ F(\mathbf{u}) = \mathbb{E}[f_s(\mathbf{u}(s))] = \sum_{s \in S} p_s f_s(\mathbf{u}(s)) \text{ over all } \mathbf{u} \in \mathcal{C} \cap \mathcal{N},$$

where $\mathcal{C}$ is feasible set and $\mathcal{N}$ is non-anticipative constraint, and f_s is the scenario subproblem (with realizations of randomness) under scenario s.

While dynamic programming is a *time-decomposition* solution scheme, PHA is a *scenario-decomposition* solution scheme.

Assume that the whole scenario is realized at $s \in S$, we then consider the following scenario sub-problems (for every $s \in S$):

$$(\mathcal{P}_s^{\nu}) \qquad \mathbf{u}^0 = \arg\min \quad f_s(u) \text{ over all } u \in C_s.$$

Note that the solution from solving f_s is not non-anticipative, as it uses future information. Set $\widehat{\mathbf{u}}^{\nu} = Project_{\mathcal{N}}(\mathbf{u}^{\nu})$, which is non-anticipative.

PHA: Background

Multistage decision-making problem is given by

$$(\mathcal{P}) \ \min \ F(\mathbf{u}) = \mathbb{E}[f_s(\mathbf{u}(s))] = \sum_{s \in S} p_s f_s(\mathbf{u}(s)) \text{ over all } \mathbf{u} \in \mathcal{C} \cap \mathcal{N},$$

where $\mathcal{C}$ is feasible set and $\mathcal{N}$ is non-anticipative constraint, and f_s is the scenario subproblem (with realizations of randomness) under scenario s.

While dynamic programming is a *time-decomposition* solution scheme, PHA is a *scenario-decomposition* solution scheme.

Assume that the whole scenario is realized at $s \in S$, we then consider the following scenario sub-problems (for every $s \in S$):

$$(\mathcal{P}_s^\nu) \qquad \mathbf{u}^0 = \arg\min \quad f_s(u) \text{ over all } u \in C_s.$$

Note that the solution from solving f_s is not non-anticipative, as it uses future information. Set $\widehat{\mathbf{u}}^\nu = Project_{\mathcal{N}}(\mathbf{u}^\nu)$, which is non-anticipative.

PHA: Background (Cont')

For given $\widehat{\mathbf{u}}^{\nu}$ and $\mathbf{w}^{\nu}$ (with initial $\mathbf{w}^{0} = 0$), we then consider the following scenario sub-problems (for every $s \in S$) in an augmented Lagrangian form to get $\widehat{\mathbf{u}}^{\nu+1}$,

$$(\mathcal{P}_s^{\nu}) \qquad \min \quad f_s(u) + u'\mathbf{w}^{\nu}(s) + \frac{1}{2} r |u - \widehat{\mathbf{u}}^{\nu}(s)|^2 \text{ over all } u \in C_s.$$

PHA: Background (Cont')

For given $\widehat{\mathbf{u}}^{\nu}$ and $\mathbf{w}^{\nu}$ (with initial $\mathbf{w}^0 = 0$), we then consider the following scenario sub-problems (for every $s \in S$) in an augmented Lagrangian form to get $\widehat{\mathbf{u}}^{\nu+1}$,

$$(\mathcal{P}_s^{\nu}) \qquad \min \quad f_s(u) + u'\mathbf{w}^{\nu}(s) + \frac{1}{2} r |u - \widehat{\mathbf{u}}^{\nu}(s)|^2 \text{ over all } u \in C_s.$$

Update $\mathbf{w}^{\nu+1} = \mathbf{w}^{\nu} + r(\mathbf{u}^{\nu+1} - \widehat{\mathbf{u}}^{\nu+1})$, where r is penalty parameter.

PHA: Background (Cont')

For given $\widehat{\mathbf{u}}^{\nu}$ and $\mathbf{w}^{\nu}$ (with initial $\mathbf{w}^0 = 0$), we then consider the following scenario sub-problems (for every $s \in S$) in an augmented Lagrangian form to get $\widehat{\mathbf{u}}^{\nu+1}$,

$$(\mathcal{P}_s^{\nu}) \qquad \min \quad f_s(u) + u'\mathbf{w}^{\nu}(s) + \frac{1}{2} r|u - \widehat{\mathbf{u}}^{\nu}(s)|^2 \text{ over all } u \in C_s.$$

Update $\mathbf{w}^{\nu+1} = \mathbf{w}^{\nu} + r(\mathbf{u}^{\nu+1} - \widehat{\mathbf{u}}^{\nu+1})$, where r is penalty parameter. Convergence occurs w.r.t r-norm

$$\| (\mathbf{u}, \mathbf{w}) \|_r = (\| \mathbf{u} \|^2 + r^{-2} \| \mathbf{w} \|^2)^{\frac{1}{2}},$$

if f_s is convex w.r.t u for all s.

Two-layer Solution Scheme: Naive PHA

The ith ($i \in S = \{1, 2, \cdots, N\}$) scenario subproblem (with $\theta = \theta_i$ hence $B = B_i$) is

$$(\mathcal{P}_i) \qquad \min_{\boldsymbol{u}} \ \mathcal{L}(\boldsymbol{u}|x_0, \delta_i) := \mathbb{E}_\xi \Big[\frac{1}{2}x_T' Q x_T + \sum_{t=0}^{T-1} \Big(\frac{1}{2}x_t' Q x_t + \frac{1}{2}u_t' R u_t\Big)\Big| x_0\Big]$$
$$\text{s.t.} \ \ x_{t+1} = A x_t + B_i u_t + \xi_t, \ \ t = 0, 1, \cdots, T-1,$$

which reduces back to a classical LQG problem with known gain and irreducible uncertainty associated with $\{\xi_t\}$.

Two-layer Solution Scheme: Naive PHA

The ith ($i \in S = \{1, 2, \cdots, N\}$) scenario subproblem (with $\theta = \theta_i$ hence $B = B_i$) is

$$(\mathcal{P}_i) \qquad \min_{\boldsymbol{u}} \ \mathcal{L}(\boldsymbol{u}|x_0, \delta_i) := \mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{1}{2}x_T' Q x_T + \sum_{t=0}^{T-1}\Big(\frac{1}{2}x_t' Q x_t + \frac{1}{2}u_t' R u_t\Big)\Big| x_0\Big]$$
$$\text{s.t.} \ \ x_{t+1} = A x_t + B_i u_t + \xi_t, \ \ t = 0, 1, \cdots, T-1,$$

which reduces back to a classical LQG problem with known gain and irreducible uncertainty associated with $\{\xi_t\}$.

The optimal feedback policy at time t (refer to Kirk, 1970, for example)

$$u_{ti}^*(x_t) = -K_{ti} x_t, \tag{4}$$

where

$$\begin{aligned} K_{ti} &= (R + B_i' P_{(t+1)i} B_i)^{-1} B_i' P_{(t+1)i} A \\ P_{ti} &= Q + K_{ti}' R K_{ti} + (A - B_i K_{ti})' P_{(t+1)i} (A - B_i K_{ti}). \end{aligned}$$

Two-layer Solution Scheme: Naive PHA

If we do not update the prior knowledge of parameter θ using future observations, we have a naive version of problem $(\mathcal{P})$,

$$(\mathcal{P}_{Naive}) \qquad \min_{\boldsymbol{u}} \ \sum_{i\in S} p_{0i}\mathcal{L}(\boldsymbol{u}|x_0,\delta_i)$$
$$\text{s.t. } x_{t+1} = Ax_t + B(\theta)u_t + \xi_t, \ \ t = 0,1,\cdots,T-1.$$

Two-layer Solution Scheme: Naive PHA

If we do not update the prior knowledge of parameter θ using future observations, we have a naive version of problem $(\mathcal{P})$,

$$(\mathcal{P}_{Naive}) \qquad \min_{\boldsymbol{u}} \ \sum_{i \in S} p_{0i} \mathcal{L}(\boldsymbol{u} | x_0, \delta_i)$$
$$\text{s.t. } x_{t+1} = Ax_t + B(\theta)u_t + \xi_t, \ \ t = 0, 1, \cdots, T-1.$$

To solve $(\mathcal{P}_{Naive})$, PHA, as a scenario-decomposition method, first decomposes it into N scenario subproblems $(\mathcal{P}_i)$, $i \in S$, and generates the scenario-based feedback policy for each i using (4),

$$\boldsymbol{u}_i^{[0]}(\cdot) = \left(u_{0i}^*(x_0), u_{1i}^*(x_1), \cdots, u_{(T-1)i}^*(x_{N-1})\right)'. \tag{5}$$

Two-layer Solution Scheme: Naive PHA

If we do not update the prior knowledge of parameter θ using future observations, we have a naive version of problem $(\mathcal{P})$,

$$(\mathcal{P}_{Naive}) \qquad \min_{\boldsymbol{u}} \ \sum_{i\in S} p_{0i}\mathcal{L}(\boldsymbol{u}|x_0,\delta_i)$$
$$\text{s.t. } x_{t+1} = Ax_t + B(\theta)u_t + \xi_t, \ \ t = 0,1,\cdots,T-1.$$

To solve $(\mathcal{P}_{Naive})$, PHA, as a scenario-decomposition method, first decomposes it into N scenario subproblems $(\mathcal{P}_i)$, $i \in S$, and generates the scenario-based feedback policy for each i using (4),

$$\boldsymbol{u}_i^{[0]}(\cdot) = \big(u_{0i}^*(x_0), u_{1i}^*(x_1), \cdots, u_{(T-1)i}^*(x_{N-1})\big)'. \tag{5}$$

Then PHA projects all $\boldsymbol{u}_i^{[0]}$'s into a non-anticipative space to get an implementable feedback policy $\hat{\boldsymbol{u}}^{[0]} = \sum_{i\in S} p_{0i}\boldsymbol{u}_i^{[0]}$. Compared with a scenario-specific policy (like $\boldsymbol{u}_i^{[0]}$), an implementable one (like $\hat{\boldsymbol{u}}^{[0]}$) is indifferent to all scenarios.

Two-layer Solution Scheme: Naive PHA

PHA next solves, in parallel for all $i \in S$, the Lagrangian subproblems

$$(\mathcal{P}_i^{[\nu]}) \quad \min_{\boldsymbol{u}} \ \mathcal{L}(\boldsymbol{u}|x_0, \delta_i) + \boldsymbol{u}' \boldsymbol{w}_i^{[\nu]} + \frac{1}{2} r \|\boldsymbol{u} - \hat{\boldsymbol{u}}^{[\nu]}\|^2,$$

in each iteration $\nu = 0, 1, \cdots$, with initials $\hat{\boldsymbol{u}}^{[0]}$ defined before and $\boldsymbol{w}_i^{[0]}$ being zero vector for each i.

Two-layer Solution Scheme: Naive PHA

PHA next solves, in parallel for all $i \in S$, the Lagrangian subproblems

$$(\mathcal{P}_i^{[\nu]}) \quad \min_{\boldsymbol{u}} \ \mathcal{L}(\boldsymbol{u}|x_0, \delta_i) + \boldsymbol{u}' \boldsymbol{w}_i^{[\nu]} + \frac{1}{2} r \|\boldsymbol{u} - \hat{\boldsymbol{u}}^{[\nu]}\|^2,$$

in each iteration $\nu = 0, 1, \cdots$, with initials $\hat{\boldsymbol{u}}^{[0]}$ defined before and $\boldsymbol{w}_i^{[0]}$ being zero vector for each i.

The optimal solution of $(\mathcal{P}_i^{[\nu]})$ is denoted by $\boldsymbol{u}_i^{[\nu+1]}$, and the implementable feedback policy used for next $(\mathcal{P}_i^{[\nu+1]})$ is then given by

$$\hat{\boldsymbol{u}}^{[\nu+1]} = \sum_{i \in S} p_{0i} \boldsymbol{u}_i^{[\nu+1]}.$$

Two-layer Solution Scheme: Naive PHA

The Lagrangian multiplier for scenario i is updated via

$$\boldsymbol{w}_i^{[\nu+1]} = \boldsymbol{w}_i^{[\nu]} + r\big(\boldsymbol{u}_i^{[\nu+1]} - \hat{\boldsymbol{u}}^{[\nu+1]}\big).$$

Two-layer Solution Scheme: Naive PHA

The Lagrangian multiplier for scenario i is updated via

$$\boldsymbol{w}_i^{[\nu+1]} = \boldsymbol{w}_i^{[\nu]} + r\big(\boldsymbol{u}_i^{[\nu+1]} - \hat{\boldsymbol{u}}^{[\nu+1]}\big).$$

The penalty parameter $r > 0$ is predetermined, and $\| \cdot \|$ denotes 2-norm.

Two-layer Solution Scheme: Naive PHA

The Lagrangian multiplier for scenario i is updated via

$$\boldsymbol{w}_i^{[\nu+1]} = \boldsymbol{w}_i^{[\nu]} + r\big(\boldsymbol{u}_i^{[\nu+1]} - \hat{\boldsymbol{u}}^{[\nu+1]}\big).$$

The penalty parameter $r > 0$ is predetermined, and $\|\cdot\|$ denotes 2-norm.

The process repeats until the convergence occurs, which is guaranteed by a convexity of the scenario subproblem w.r.t. the control variable according to the PHA requirement, and $\hat{\boldsymbol{u}}^{[\infty]}$ is actually the optimal policy to the problem $(\mathcal{P}_{Naive})$.

Two-layer Solution Scheme: Revised PHA

In order to incorporate the learning feature into our solution algorithm for $(\mathcal{P})$, we need to update the knowledge about uncertain parameter θ.

Two-layer Solution Scheme: Revised PHA

In order to incorporate the learning feature into our solution algorithm for $(\mathcal{P})$, we need to update the knowledge about uncertain parameter θ.

Every step is the same as in PHA for the naive version $(\mathcal{P}_{Naive})$, except that, when forming the implementable feedback policy at iteration ν, we need to take conditional expectations using the posterior probabilities at each time $t = 0, 1, \cdots, T-2$,

$$\hat{u}^{[\nu]}_{t+1}(\cdot) = \sum_{i \in S} p_{(t+1)i} u^{[\nu]}_{(t+1)i}(\cdot), \tag{6}$$

where $u^{[\nu]}_{(t+1)i}(\cdot)$ as element of $\boldsymbol{u}^{[\nu]}_i(\cdot)$ comes from solving the ith Lagrangian subproblem $(\mathcal{P}^{[\nu-1]}_i)$ $(\nu \geq 1)$, with initial $\boldsymbol{u}^{[0]}_i$ being the solution of scenario subproblem (5).

Two-layer Solution Scheme: Revised PHA

Note that $p_{(t+1)i}$ depends on I^{t+1}. If we directly substitute it into (6), $\hat{u}^{[\nu]}_{t+1}(\cdot)$ becomes nonlinear in state, which in turn leads to the intractability when dealing with $(\mathcal{P}^{[\nu]}_i)$.

Two-layer Solution Scheme: Revised PHA

Note that $p_{(t+1)i}$ depends on I^{t+1}. If we directly substitute it into (6), $\hat{u}_{t+1}^{[\nu]}(\cdot)$ becomes nonlinear in state, which in turn leads to the intractability when dealing with $(\mathcal{P}_i^{[\nu]})$.

We bypass this difficulty by setting $p_{(t+1)i}$ at its nominal value,

$$\bar{p}_{(t+1)i}^{[\nu]} = \frac{\phi(\bar{x}_{t+1}^{[\nu]}; \bar{\mu}_{ti}^{[\nu]}, \Sigma_\xi)\bar{p}_{ti}^{[\nu]}}{\sum_{j\in S}[\phi(\bar{x}_{t+1}^{[\nu]}; \bar{\mu}_{tj}^{[\nu]}, \Sigma_\xi)\bar{p}_{tj}^{[\nu]}]} \tag{7}$$

where $\bar{\mu}_{ti}^{[\nu]} = A\bar{x}_t^{[\nu]} + B_i\hat{u}_t^{[\nu]}(\bar{x}_t^{[\nu]})$ and the nominal state is determined sequentially by

$$\begin{aligned} \bar{x}_{t+1}^{[\nu]} &= \mathbb{E}_{\theta,\xi_t}\left[A\bar{x}_t^{[\nu]} + B(\theta)\hat{u}_t^{[\nu]}(\bar{x}_t^{[\nu]}) + \xi_t \mid \bar{x}_t^{[\nu]}, \bar{p}_t^{[\nu]}(\theta)\right] \\ &= A\bar{x}_t^{[\nu]} + \left(\sum_{i\in S}\bar{p}_{ti}^{[\nu]}B_i\right)\hat{u}_t^{[\nu]}(\bar{x}_t^{[\nu]}) \end{aligned} \tag{8}$$

for $t = 0, 1, \cdots, T-1$ with nominal initial state $\bar{x}_0^{[\nu]} = x_0$ and nominal prior distribution $\bar{p}_0^{[\nu]}(\theta) = p_0(\theta)$ held for every ν.

Two-layer Solution Scheme: Revised PHA

Then the implementable feedback policy obtained at $t+1$ becomes linear w.r.t the state compared with (6),

$$\hat{u}_{t+1}^{[\nu]}(\cdot) = \sum_{i \in S} \bar{p}_{(t+1)i}^{[\nu]} u_{(t+1)i}^{[\nu]}(\cdot). \tag{9}$$

Two-layer Solution Scheme: Revised PHA

Then the implementable feedback policy obtained at $t+1$ becomes linear w.r.t the state compared with (6),

$$\hat{u}_{t+1}^{[\nu]}(\cdot) = \sum_{i \in S} \bar{p}_{(t+1)i}^{[\nu]} u_{(t+1)i}^{[\nu]}(\cdot)\,. \tag{9}$$

This relaxation to a linear policy enables us to proceed the iteration until converging to a final approximate feedback policy of our two-layer (TL) method,

$$u_t^{TL}(x_t) = \hat{u}_t^{[\nu]}(x_t) = -\hat{K}_t^{[\nu]} x_t, \text{ as } \nu \to \infty, \text{ for all } t. \tag{10}$$

Two-layer Solution Scheme: Revised PHA

Then the implementable feedback policy obtained at $t+1$ becomes linear w.r.t the state compared with (6),

$$\hat{u}_{t+1}^{[\nu]}(\cdot) = \sum_{i \in S} \bar{p}_{(t+1)i}^{[\nu]} u_{(t+1)i}^{[\nu]}(\cdot). \tag{9}$$

This relaxation to a linear policy enables us to proceed the iteration until converging to a final approximate feedback policy of our two-layer (TL) method,

$$u_t^{TL}(x_t) = \hat{u}_t^{[\nu]}(x_t) = -\hat{K}_t^{[\nu]} x_t, \text{ as } \nu \to \infty, \text{ for all } t. \tag{10}$$

In practice, the algorithm will stop when the predetermined convergence tolerance level (tol) is satisfied, namely, $err < tol$, where the error is

$$err := \sqrt{\|\hat{\boldsymbol{u}}^{[\nu+1]} - \hat{\boldsymbol{u}}^{[\nu]}\|^2 + \frac{1}{r^2} \sum_{i \in S} \|\boldsymbol{w}_i^{[\nu+1]} - \boldsymbol{w}_i^{[\nu]}\|^2}. \tag{11}$$

Two-layer Solution Scheme: Revised PHA

Confining ourselves on the nominal trajectory, on the one hand, we are able to forwardly calculate an implementable policy and the nominal posterior distribution along the time horizon in each iteration, on the other hand, however, since Bellman equation in our case considers the entire (continuous) state space, the converged nominal-based policy is only suboptimal.

Two-layer Solution Scheme: Revised PHA

Confining ourselves on the nominal trajectory, on the one hand, we are able to forwardly calculate an implementable policy and the nominal posterior distribution along the time horizon in each iteration, on the other hand, however, since Bellman equation in our case considers the entire (continuous) state space, the converged nominal-based policy is only suboptimal.

Nevertheless, as we will demonstrate later, our newly-derived approximation performs better in an average sense than the prevalent passive learning method and others borrowed from traditional RL algorithms.

Illustrative Example of a Scalar System

Consider a simple scalar system with system dynamics

$$x_{t+1} = ax_t + b(\theta)u_t + \xi_t, \ \ t = 0, 1, \cdots, T-1,$$

where we denote $b(\theta_i) = b_i$, $\forall i \in S = \{1, \cdots, N\}$ and the i.i.d. system random disturbance ξ_t follows $\mathcal{N}(0, \sigma^2)$, together with other usual assumptions for $(\mathcal{P})$.

Illustrative Example of a Scalar System

Consider a simple scalar system with system dynamics

$$x_{t+1} = ax_t + b(\theta)u_t + \xi_t, \;\; t = 0, 1, \cdots, T-1,$$

where we denote $b(\theta_i) = b_i$, $\forall i \in S = \{1, \cdots, N\}$ and the i.i.d. system random disturbance ξ_t follows $\mathcal{N}(0, \sigma^2)$, together with other usual assumptions for $(\mathcal{P})$.

Solving the scenario subproblem $\min_{\boldsymbol{u}} \; \mathcal{L}(\boldsymbol{u}|x_0, \delta_i)$ by DP for each i at the lower layer, we obtain the scenario-specific feedback policy

$$u_{ti}^{[0]}(x_t) = -K_{ti}^{[0]} x_t$$

for all t, with backward recursions

$$\begin{aligned} K_{ti}^{[0]} &= (ab_i\Gamma_{(t+1)i}^{[0]})/(R + b_i^2\Gamma_{(t+1)i}^{[0]}), \\ \Gamma_{ti}^{[0]} &= Q + R(K_{ti}^{[0]})^2 + (a - b_iK_{ti}^{[0]})^2\Gamma_{(t+1)i}^{[0]}, \end{aligned}$$

and the boundary condition $\Gamma_{Ti}^{[0]} \equiv Q$ for all i.

Illustrative Example of a Scalar System

The optimal cost-to-go function is then given by

$$J_{ti}^{[0]}(x_t) = \frac{1}{2}\Gamma_{ti}^{[0]}x_t^2 + \Lambda_{ti}^{[0]},$$

with

$$\Lambda_{ti}^{[0]} = \Lambda_{(t+1)i}^{[0]} + \frac{1}{2}\Gamma_{(t+1)i}^{[0]}\sigma^2$$

and the boundary condition $\Lambda_{Ti}^{[0]} \equiv 0$.

Illustrative Example of a Scalar System

The optimal cost-to-go function is then given by

$$J_{ti}^{[0]}(x_t) = \frac{1}{2}\Gamma_{ti}^{[0]} x_t^2 + \Lambda_{ti}^{[0]},$$

with

$$\Lambda_{ti}^{[0]} = \Lambda_{(t+1)i}^{[0]} + \frac{1}{2}\Gamma_{(t+1)i}^{[0]}\sigma^2$$

and the boundary condition $\Lambda_{Ti}^{[0]} \equiv 0$.

At the upper layer, we first initialize the nominal initial state $\bar{x}_0^{[0]} = x_0$ and the nominal prior distribution $\bar{p}_0^{[0]}(\theta) = p_0(\theta)$, and then calculate the implementable control at $t = 0$:

$$\hat{u}_0^{[0]}(\bar{x}_0^{[0]}) = \sum_{i \in S} \bar{p}_{0i}^{[0]} u_{0i}^{[0]}(\bar{x}_0^{[0]}) = -\hat{K}_0^{[0]} \bar{x}_0^{[0]}$$

where

$$\hat{K}_0^{[0]} = \textstyle\sum_{i \in S} \bar{p}_{0i}^{[0]} K_{0i}^{[0]}.$$

Illustrative Example of a Scalar System

The next nominal state is obtained via (8):

$$\bar{x}_1^{[0]} = \left[a - (\sum_{i \in S} \bar{p}_{0i}^{[0]} b_i) \hat{K}_0^{[0]} \right] \bar{x}_0^{[0]},$$

with the nominal posterior probabilities $\bar{p}_{1i}^{[0]}$'s updated numerically through (7) for various i, which in turn yields

$$\hat{u}_1^{[0]}(x_1) = -\hat{K}_1^{[0]} x_1$$

with $\hat{K}_1^{[0]} = \sum_{i \in S} \bar{p}_{1i}^{[0]} K_{1i}^{[0]}$, which is still linear in state. We then apply the control $\hat{u}_1^{[0]}(\bar{x}_1^{[0]})$ to get the nominal state and posterior distribution at $t = 2$. We conduct the above procedure till the end of time horizon and finally obtain a feedback policy $\hat{\boldsymbol{u}}^{[0]}$ for $\nu = 0$.

Illustrative Example of a Scalar System

The next nominal state is obtained via (8):

$$\bar{x}_1^{[0]} = \left[a - (\sum_{i \in S} \bar{p}_{0i}^{[0]} b_i) \hat{K}_0^{[0]} \right] \bar{x}_0^{[0]},$$

with the nominal posterior probabilities $\bar{p}_{1i}^{[0]}$'s updated numerically through (7) for various i, which in turn yields

$$\hat{u}_1^{[0]}(x_1) = -\hat{K}_1^{[0]} x_1$$

with $\hat{K}_1^{[0]} = \sum_{i \in S} \bar{p}_{1i}^{[0]} K_{1i}^{[0]}$, which is still linear in state. We then apply the control $\hat{u}_1^{[0]}(\bar{x}_1^{[0]})$ to get the nominal state and posterior distribution at $t = 2$. We conduct the above procedure till the end of time horizon and finally obtain a feedback policy $\hat{\boldsymbol{u}}^{[0]}$ for $\nu = 0$.

The initial Lagrangian multiplier is set to be zero, which is equivalent as a linear function of state, namely, $w_{ti}^{[0]}(x_t) = W_{ti}^{[0]} x_t$ with $W_{ti}^{[0]} \equiv 0$ for all t and i.

Illustrative Example of a Scalar System

We are now ready to define Lagrangian subproblems for $\nu = 0$:

$$(\mathcal{Q}_i^{[0]}): \ \min_{\boldsymbol{u}} \ \mathcal{L}(\boldsymbol{u}|x_0, \delta_i) + \sum_{t=0}^{T-1} w_{ti}^{[0]} u_t + \frac{1}{2} r \sum_{t=0}^{T-1} (u_t - \hat{u}_t^{[0]})^2.$$

The above multistage optimization problem (subject to the linear state dynamics) can be analytically solved by DP and the optimal solution is linear in state.

Illustrative Example of a Scalar System

We are now ready to define Lagrangian subproblems for $\nu = 0$:

$$(\mathcal{Q}_i^{[0]}): \ \min_{\boldsymbol{u}} \ \mathcal{L}(\boldsymbol{u}|x_0, \delta_i) + \sum_{t=0}^{T-1} w_{ti}^{[0]} u_t + \frac{1}{2} r \sum_{t=0}^{T-1} (u_t - \hat{u}_t^{[0]})^2.$$

The above multistage optimization problem (subject to the linear state dynamics) can be analytically solved by DP and the optimal solution is linear in state.

It is easy to prove, by mathematical induction, that starting with linear forms of $\hat{u}_t^{[0]}(\cdot)$ and $w_{ti}^{[0]}(\cdot)$, all $(\mathcal{Q}_i^{[\nu]})$'s, $\nu \geq 0$, keep the same quadratic forms, with optimal cost-to-go function at t satisfying the new Bellman

$$\begin{aligned} J_{ti}^{[\nu+1]}(x_t) &= \min_{u_t} \ \mathbb{E}_\xi \left[\frac{1}{2} Q x_t^2 + \frac{1}{2} R u_t^2 + w_{ti}^{[\nu]} u_t + \frac{1}{2} r (u_t - \hat{u}_t^{[\nu]})^2 \right. \\ & \left. + J_{(t+1)i}^{[\nu+1]}\Big(a x_t + b_i u_t + \xi_t \Big) \ \Big| \ x_t \right], \end{aligned}$$

with boundary condition $J_{Ti}^{[\nu+1]}(x_T) = \frac{1}{2} x_T^2$.

Illustrative Example of a Scalar System

Then optimal solution is

$$u_{ti}^{[\nu+1]}(x_t) = -K_{ti}^{[\nu+1]} x_t,$$

where

$$K_{ti}^{[\nu+1]} = (ab_i \Gamma_{(t+1)i}^{[\nu+1]} + r\hat{K}_t^{[\nu]} + W_{ti}^{[\nu]})/(R + r + b_i^2 \Gamma_{(t+1)i}^{[\nu+1]}).$$

Illustrative Example of a Scalar System

Then optimal solution is

$$u_{ti}^{[\nu+1]}(x_t) = -K_{ti}^{[\nu+1]} x_t,$$

where

$$K_{ti}^{[\nu+1]} = (ab_i \Gamma_{(t+1)i}^{[\nu+1]} + r\hat{K}_t^{[\nu]} + W_{ti}^{[\nu]})/(R + r + b_i^2 \Gamma_{(t+1)i}^{[\nu+1]}).$$

Accordingly,

$$J_{ti}^{[\nu+1]}(x_t) = \frac{1}{2}\Gamma_{ti}^{[\nu+1]} x_t^2 + \Lambda_{ti}^{[\nu+1]},$$

where

$$\begin{aligned} \Gamma_{ti}^{[\nu+1]} &= a^2 \Gamma_{(t+1)i}^{[\nu+1]} + Q + r\big(\hat{K}_t^{[\nu]}\big)^2 - \big(R + r + b_i^2 \Gamma_{(t+1)i}^{[\nu+1]}\big)\big(K_{ti}^{[\nu+1]}\big)^2 \\ \Lambda_{ti}^{[\nu+1]} &= \Lambda_{(t+1)i}^{[\nu+1]} + \frac{1}{2}\Gamma_{(t+1)i}^{[\nu+1]} \sigma^2, \end{aligned} \tag{12}$$

with boundary conditions $\Gamma_{Ti}^{[\nu+1]} = Q$ and $\Lambda_{Ti}^{[\nu+1]} = 0$.

Illustrative Example of a Scalar System

Finally, the Lagrangian multiplier is updated via

$$w_{ti}^{[\nu+1]}(x_t) = w_{ti}^{[\nu]}(x_t) + r\big[u_{ti}^{[\nu+1]}(x_t) - \hat{u}_t^{[\nu+1]}(x_t)\big] = W_{ti}^{[\nu+1]} x_t,$$

where

$$W_{ti}^{[\nu+1]} = W_{ti}^{[\nu]} + r(\hat{K}_t^{[\nu+1]} - K_{ti}^{[\nu+1]}).$$

Illustrative Example of a Scalar System

Finally, the Lagrangian multiplier is updated via

$$w_{ti}^{[\nu+1]}(x_t) = w_{ti}^{[\nu]}(x_t) + r\big[u_{ti}^{[\nu+1]}(x_t) - \hat{u}_t^{[\nu+1]}(x_t)\big] = W_{ti}^{[\nu+1]}x_t,$$

where

$$W_{ti}^{[\nu+1]} = W_{ti}^{[\nu]} + r(\hat{K}_t^{[\nu+1]} - K_{ti}^{[\nu+1]}).$$

The iteration terminates when the stopping criterion is satisfied, and results in a linear feedback policy as in (10):

$$u_t^{TL}(x_t) = \hat{u}_t^{[\nu]}(x_t) = -\hat{K}_t^{[\nu]}x_t, \text{ as } \nu \to \infty, \text{ for all } t.$$

Experimental Comparison and Results

We now verify the efficiency of our proposed two-layer (TL) scheme, compared with other algorithms including DP, DUL (the prevalent passive learning approach of Deshpande et al. (1973)), and three other methods leveraging ideas from traditional RL algorithms: the greedy method, ϵ-greedy, and Thompson sampling.

Experimental Comparison and Results

We now verify the efficiency of our proposed two-layer (TL) scheme, compared with other algorithms including DP, DUL (the prevalent passive learning approach of Deshpande et al. (1973)), and three other methods leveraging ideas from traditional RL algorithms: the greedy method, ϵ-greedy, and Thompson sampling.

While DP, as the theoretical best, provides a benchmark for comparison, it is *only* applicable when $T = 2$, where analytical optimal policy can be obtained at $t = 1$, and *numerical method* has to be invoked at $t = 0$, for example by MATLAB.

Experimental Comparison and Results

We now verify the efficiency of our proposed two-layer (TL) scheme, compared with other algorithms including DP, DUL (the prevalent passive learning approach of Deshpande et al. (1973)), and three other methods leveraging ideas from traditional RL algorithms: the greedy method, ϵ-greedy, and Thompson sampling.

While DP, as the theoretical best, provides a benchmark for comparison, it is *only* applicable when $T = 2$, where analytical optimal policy can be obtained at $t = 1$, and *numerical method* has to be invoked at $t = 0$, for example by MATLAB.

As for DUL, it assumes that the expectation and the minimization operators in the original problem $(\mathcal{P})$ can be exchanged, i.e.,

$$\min_{\boldsymbol{u}} \ \mathbb{E}_\theta \left\{ \mathbb{E}_{\boldsymbol{\xi}} \left[\cdots \ \middle| p_0(\theta) \right] \right\} \approx \mathbb{E}_\theta \Big\{ \min_{\boldsymbol{u}} \ \mathbb{E}_{\boldsymbol{\xi}} \left[\cdots \ \middle| p_0(\theta) \right] \Big\}.$$

The DUL algorithm is basically a rolling horizon approach.

Experimental Comparison and Results

Adopting similar idea of rolling horizon, we may also think out other three algorithms (rooted originally in classical RL problems) that are applicable to non-episodic cases.

Experimental Comparison and Results

Adopting similar idea of rolling horizon, we may also think out other three algorithms (rooted originally in classical RL problems) that are applicable to non-episodic cases.

The first one is similar to the *greedy* method (named GRE) by selecting the scenario-specific policy with largest posterior probability at time t.

Experimental Comparison and Results

Adopting similar idea of rolling horizon, we may also think out other three algorithms (rooted originally in classical RL problems) that are applicable to non-episodic cases.

The first one is similar to the *greedy* method (named GRE) by selecting the scenario-specific policy with largest posterior probability at time t.

As a variation of GRE, the ϵ-*greedy* type strategy (termed ϵ-GRE) perturbs the greedy policy a bit by a randomized policy of selecting the greedy policy with probability $(1-\epsilon)$ or a randomly chosen policy with probability ϵ.

Experimental Comparison and Results

Adopting similar idea of rolling horizon, we may also think out other three algorithms (rooted originally in classical RL problems) that are applicable to non-episodic cases.

The first one is similar to the *greedy* method (named GRE) by selecting the scenario-specific policy with largest posterior probability at time t.

As a variation of GRE, the ϵ-*greedy* type strategy (termed ϵ-GRE) perturbs the greedy policy a bit by a randomized policy of selecting the greedy policy with probability $(1-\epsilon)$ or a randomly chosen policy with probability ϵ.

The last algorithm follows the idea of *Thompson sampling* (labelled TS here) that a policy at time t is selected by randomly sampling a scenario-specific policy based on the posterior distribution.

Experimental Comparison and Results

For simplicity, the model is set with $a = Q = R = \sigma = x_0 = 1$, and $N = 2$ meaning that b takes two possible values. The penalty parameter $r = 1$ and the tolerance level $tol = 10^{-5}$. $\epsilon = 10\%$.

Experimental Comparison and Results

For simplicity, the model is set with $a = Q = R = \sigma = x_0 = 1$, and $N = 2$ meaning that b takes two possible values. The penalty parameter $r = 1$ and the tolerance level $tol = 10^{-5}$. $\epsilon = 10\%$.

We do 12 experiments in total for different time horizons ($T = 2$ when DP works, and $T = 3, 5$ when DP fails and we also adopt rolling version TL_R) and different assignments on $b = \{b_1, b_2\}$ and $p_0(\theta) = \{p_{01}, p_{02}\}$.

Experimental Comparison and Results

For simplicity, the model is set with $a = Q = R = \sigma = x_0 = 1$, and $N = 2$ meaning that b takes two possible values. The penalty parameter $r = 1$ and the tolerance level $tol = 10^{-5}$. $\epsilon = 10\%$.

We do 12 experiments in total for different time horizons ($T = 2$ when DP works, and $T = 3, 5$ when DP fails and we also adopt rolling version TL_R) and different assignments on $b = \{b_1, b_2\}$ and $p_0(\theta) = \{p_{01}, p_{02}\}$.

For each experiment, we compute the TL feedback gain $\hat{K}_t$ in (10), and generate ten thousand simulations that are shared by all the seven considered algorithms.

Experimental Comparison and Results

For simplicity, the model is set with $a = Q = R = \sigma = x_0 = 1$, and $N = 2$ meaning that b takes two possible values. The penalty parameter $r = 1$ and the tolerance level $tol = 10^{-5}$. $\epsilon = 10\%$.

We do 12 experiments in total for different time horizons ($T = 2$ when DP works, and $T = 3, 5$ when DP fails and we also adopt rolling version TL_R) and different assignments on $b = \{b_1, b_2\}$ and $p_0(\theta) = \{p_{01}, p_{02}\}$.

For each experiment, we compute the TL feedback gain $\hat{K}_t$ in (10), and generate ten thousand simulations that are shared by all the seven considered algorithms.

Every simulation is characterized by two parts $(\theta, \{\xi_t\}_t)$, where θ is sampled by $p_0(\theta)$ and each ξ_t is sampled from the assumed i.i.d. Gaussian noise, in order to calculate and compare the total costs induced by different policies in the average sense.

Experimental Comparison and Results

No.	p_{01}	b		DP	TL	TL$_R$	DUL	GRE	ϵ-GRE	TS
For $T = 2$										
(i)	$\frac{1}{3}$	$b_1 = 1$	$b_2 = 2$	**1.8170**	1.8172	1.8171	1.8204	1.8213	1.8236	1.8409
(ii)		$b_1 = 1$	$b_2 = 5$	**1.8199**	1.8261	1.8203	1.9974	2.0491	2.1060	2.4745
(iii)		$b_1 = 1$	$b_2 = 10$	1.8793	1.8833	**1.8785**	2.7875	3.7598	4.0188	5.7482
(iv)	1/2	$b_1 = 1$	$b_2 = 2$	**1.9052**	1.9060	1.9055	1.9095	1.9296	1.9314	1.9310
(v)	2/3			**1.9383**	1.9395	1.9389	1.9427	1.9614	1.9611	1.9589
(vi)	1			**2.0276**	**2.0276**	**2.0276**	**2.0276**	**2.0276**	**2.0276**	**2.0276**
For $T = 3$										
(vii)	1/3	$b_1 = 1$	$b_2 = 2$	N/A	2.5349	**2.5333**	2.5371	2.5517	2.5545	2.5837
$(viii)$	1/2			N/A	2.6541	**2.6511**	2.6542	2.6949	2.6916	2.6932
(ix)	2/3			N/A	2.7140	**2.7106**	2.7139	2.7384	2.7419	2.7506
For $T = 5$										
(x)	1/3	$b_1 = 1$	$b_2 = 2$	N/A	3.8848	**3.8779**	3.8804	3.9070	3.9134	3.9671
(xi)	1/2			N/A	4.0923	**4.0762**	4.0789	4.1359	4.1364	4.1559
(xii)	2/3			N/A	4.2734	**4.2546**	4.2558	4.3022	4.3039	4.3159

Experimental Comparison and Results

We can see from the table

- When $T = 2$, DP always ranks the top (with one exception, which could be due to that MATLAB can only identify a local minimum for a possible non-convex value function at $t = 0$) and TL approximates the true optimal policy pretty well as evidenced by its lower average total cost compared to others (except for its rolling variant and DP).

Experimental Comparison and Results

We can see from the table

- When $T = 2$, DP always ranks the top (with one exception, which could be due to that MATLAB can only identify a local minimum for a possible non-convex value function at $t = 0$) and TL approximates the true optimal policy pretty well as evidenced by its lower average total cost compared to others (except for its rolling variant and DP).
- For $T = 3$ where DP no longer works, TL almost maintains superior over the rest (except for TL_R) even without utilizing any *online* posterior information which other approaches rely on.

Experimental Comparison and Results

We can see from the table

- When $T = 2$, DP always ranks the top (with one exception, which could be due to that MATLAB can only identify a local minimum for a possible non-convex value function at $t = 0$) and TL approximates the true optimal policy pretty well as evidenced by its lower average total cost compared to others (except for its rolling variant and DP).
- For $T = 3$ where DP no longer works, TL almost maintains superior over the rest (except for TL_R) even without utilizing any *online* posterior information which other approaches rely on.
- As time goes by, reference to newly-updated belief becomes more and more necessary. Based on this recognition, our TL_R essentially beats all the rest when T goes beyond 2.

Experimental Comparison and Results

We also observe some interesting findings that should be naturally expected

- The longer horizon, the larger the total cost.

Experimental Comparison and Results

We also observe some interesting findings that should be naturally expected

- The longer horizon, the larger the total cost.
- In experiment (*vi*) where we are *certain* about the system parameter with a one-point distribution for the prior belief, all the algorithms lead to the same (actually optimal) policy and yield the same cost, since the problem, in such a case, reduces to a pure stochastic decision problem with *full* knowledge on parameters.

Experimental Comparison and Results

We also observe some interesting findings that should be naturally expected

- The longer horizon, the larger the total cost.
- In experiment (*vi*) where we are *certain* about the system parameter with a one-point distribution for the prior belief, all the algorithms lead to the same (actually optimal) policy and yield the same cost, since the problem, in such a case, reduces to a pure stochastic decision problem with *full* knowledge on parameters.
- We can see from experiment (*i*) to (*iii*) that the larger the variance of b, the worse the passive learning DUL and others perform. In other words, the inherent active learning feature in TL (TL_R) and DP becomes much more demanding when the uncertainty in b is large.

Summary

- We develop a novel solution approach to a type of Bayesian reinforcement learning (RL) problem under the non-episodic setting, especially the discrete-time linear-quadratic-Gaussian (LQG) problem with fixed but unknown gain as one concrete example, to which the classical dynamic programming (DP) fails.

Summary

- We develop a novel solution approach to a type of Bayesian reinforcement learning (RL) problem under the non-episodic setting, especially the discrete-time linear-quadratic-Gaussian (LQG) problem with fixed but unknown gain as one concrete example, to which the classical dynamic programming (DP) fails.
- Our new solution approximates the optimal policy directly, thus bypassing the stage of approximating the value function.

Summary

- We develop a novel solution approach to a type of Bayesian reinforcement learning (RL) problem under the non-episodic setting, especially the discrete-time linear-quadratic-Gaussian (LQG) problem with fixed but unknown gain as one concrete example, to which the classical dynamic programming (DP) fails.
- Our new solution approximates the optimal policy directly, thus bypassing the stage of approximating the value function.
- Most importantly, our scheme separates the non-episodic problem into two different layers according to different types of uncertainties, and combines the time-decomposition based method DP at the lower layer and the revised scenario-decomposition based approach progressive hedging algorithm (PHA) at the upper layer, to strike a balance between exploitation and exploration.

Summary

- By separating the reducible uncertainty from the irreducible one, we may take advantage of DP to generate an analytical solution for scenario-specific subproblems with reducible uncertainty fixed at a certain scenario. The revised PHA at the upper level, on the other hand, aggregates the solutions from all scenario subproblems to generate an implementable one, which finally converges to a suboptimal policy to approximate the optimal one of the primal Bayesian RL problem, as shown in our experiments.

Summary

- By separating the reducible uncertainty from the irreducible one, we may take advantage of DP to generate an analytical solution for scenario-specific subproblems with reducible uncertainty fixed at a certain scenario. The revised PHA at the upper level, on the other hand, aggregates the solutions from all scenario subproblems to generate an implementable one, which finally converges to a suboptimal policy to approximate the optimal one of the primal Bayesian RL problem, as shown in our experiments.
- One future research topics are to investigate deeper the convergence property of our revised-PHA based two-layer solution algorithm, and to study how to generate nominal trajectory or even multiple ones in order to simulate more learning environment in advance.

Part 2: Quadratic Dis-utility Portfolio Selection under Lack of Market Information

Market Settings: Unknown Mean and Covariance

There are n risky assets and one risk-free asset in the market. Suppose that random total return $\boldsymbol{e}_t$ is i.i.d and follows $\mathcal{N}(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$, where $\mu^{\boldsymbol{e}} \in \mathbb{R}^n$ and $\Sigma^{\boldsymbol{e}} \in \mathbb{R}^{n \times n}$ are unknown at the beginning.

Market Settings: Unknown Mean and Covariance

There are n risky assets and one risk-free asset in the market. Suppose that random total return $\boldsymbol{e}_t$ is i.i.d and follows $\mathcal{N}(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$, where $\mu^{\boldsymbol{e}} \in \mathbb{R}^n$ and $\Sigma^{\boldsymbol{e}} \in \mathbb{R}^{n \times n}$ are unknown at the beginning.

We further assume that $(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$ follows normal-inverse-Wishart (NIW) distribution with four initial values as prior belief, denoted by

$$(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}) \sim \mathcal{NIW}(\mu_0, \kappa_0, \Psi_0, \nu_0),$$

with $\kappa_0 > 0$ and $\nu_0 > n + 1$, and density being given as

$$f_{\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}}(x, y | \mu_0, \kappa_0, \Psi_0, \nu_0) = \psi(x | \mu_0, \frac{1}{\kappa_0} y) \mathcal{W}^{-1}(y | \Psi_0, \nu_0),$$

where $\phi(\cdot | z_1, z_2)$ is multivariate normal density with mean z_1 and covariance z_2, and $\mathcal{W}^{-1}(\cdot | z_3, z_4)$ is inverse Wishart density with hyperparameters z_3 and z_4.

Market Settings: Unknown Mean and Covariance

According to Bayes' law, the posterior belief on $(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$ is still normal-inverse-Wishart, as it belongs to conjugate family. See, for example, Murphy (2007) for more details.

Market Settings: Unknown Mean and Covariance

According to Bayes' law, the posterior belief on $(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$ is still normal-inverse-Wishart, as it belongs to conjugate family. See, for example, Murphy (2007) for more details.

That is, $(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})|I^t \sim \mathcal{NIW}(\mu_t, \kappa_t, \Psi_t, \nu_t)$ where I^t is the information set at time t, and hyperparameters $(\mu_t, \kappa_t, \Psi_t, \nu_t)_t$ are updated recursively by

$$\begin{aligned}
\mu_{t+1} &= \frac{\kappa_t \mu_t + \tilde{R}_{t+1}}{\kappa_t + 1}, \\
\kappa_{t+1} &= \kappa_t + 1, \\
\Psi_{t+1} &= \Psi_t + \frac{\kappa_t}{\kappa_t + 1}(\tilde{R}_{t+1} - \mu_t)(\tilde{R}_{t+1} - \mu_t)', \\
\nu_{t+1} &= \nu_t + 1,
\end{aligned} \tag{13}$$

where $\tilde{R}_{t+1}$ is the real total return vector at $t+1$ that can be computed from observed assets prices in the market.

Problem Formulation

We are interested in solving a discrete-time portfolio selection problem with quadratic dis-utility objective which is closely related to the mean-variance objective,

$$(\mathcal{A}(\lambda,\omega)) \quad \min_{\boldsymbol{u}_t,\ \forall t} \ \mathbb{E}_{\mathcal{M}}\left[\omega x_T^2 - \lambda x_T \middle| I_0\right]$$

$$\text{s.t. } x_{t+1} = s_t x_t + \boldsymbol{P}_t' \boldsymbol{u}_t, \ \ t = 0, 1, \cdots, T-1,$$

Problem Formulation

We are interested in solving a discrete-time portfolio selection problem with quadratic dis-utility objective which is closely related to the mean-variance objective,

$$(\mathcal{A}(\lambda, \omega)) \quad \min_{\boldsymbol{u}_t, \ \forall t} \ \mathbb{E}_{\mathcal{M}} \left[\omega x_T^2 - \lambda x_T \middle| I_0 \right]$$

$$\text{s.t. } x_{t+1} = s_t x_t + \boldsymbol{P}_t' \boldsymbol{u}_t, \ \ t = 0, 1, \cdots, T-1,$$

where $x_t \in \mathbb{R}$ is wealth level, $\boldsymbol{u}_t \in \mathbb{R}^n$ represents amounts of dollar invested on n risky assets, $s_t \in \mathbb{R}$ is risk-free rate, and

$$\boldsymbol{P}_t' = (\boldsymbol{e}_t - s_t \mathbf{1})' = (e_t^1 - s_t, \cdots, e_t^n - s_t)' \in \mathbb{R}^n$$

is random excess return at time t with $\mathbf{1}$ being the all-one vector, and $I_0 = \{x_0, \mu_0, \kappa_0, \Psi_0, \nu_0\}$.

Problem Formulation

We are interested in solving a discrete-time portfolio selection problem with quadratic dis-utility objective which is closely related to the mean-variance objective,

$$(\mathcal{A}(\lambda,\omega)) \quad \min_{\boldsymbol{u}_t,\ \forall t} \ \mathbb{E}_{\mathcal{M}}\left[\omega x_T^2 - \lambda x_T \middle| I_0\right]$$

$$\text{s.t. } x_{t+1} = s_t x_t + \boldsymbol{P}_t' \boldsymbol{u}_t,\ \ t = 0, 1, \cdots, T-1,$$

where $x_t \in \mathbb{R}$ is wealth level, $\boldsymbol{u}_t \in \mathbb{R}^n$ represents amounts of dollar invested on n risky assets, $s_t \in \mathbb{R}$ is risk-free rate, and

$$\boldsymbol{P}_t' = (\boldsymbol{e}_t - s_t\mathbf{1})' = (e_t^1 - s_t, \cdots, e_t^n - s_t)' \in \mathbb{R}^n$$

is random excess return at time t with $\mathbf{1}$ being the all-one vector, and $I_0 = \{x_0, \mu_0, \kappa_0, \Psi_0, \nu_0\}$.

$\mathcal{M} = \{\boldsymbol{e}, \mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}\}$ is used to emphasize the expectation not only on irreducible uncertainty of $\boldsymbol{e}$ but also on that of $(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$ which is reducible through online Bayesian learning.

Problem Formulation

Note that the reducible uncertainty in $x_{t+1} = s_t x_t + \boldsymbol{P}_t' \boldsymbol{u}_t$ is (indirectly) observable, compared with the Bayesian RL example of Part 1.

Problem Formulation

Note that the reducible uncertainty in $x_{t+1} = s_t x_t + \boldsymbol{P}'_t \boldsymbol{u}_t$ is (indirectly) observable, compared with the Bayesian RL example of Part 1.

Different from settings of Li and Ng (2000), the first two central moments of $\boldsymbol{P}_t$ here are unknown but conditionally given by

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{P}_t|\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}] &= \mathbb{E}[\boldsymbol{e}_t - s_t \mathbf{1}|\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}] = \mu^{\boldsymbol{e}} - s_t \mathbf{1} := \mu_t^{\boldsymbol{P}}(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}), \\
\mathbb{E}[\boldsymbol{P}_t \boldsymbol{P}'_t|\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}] &= \mathbb{E}[(\boldsymbol{e}_t - s_t \mathbf{1})(\boldsymbol{e}_t - s_t \mathbf{1})'|\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}] \\
&= \mathbb{E}[\boldsymbol{e}_t \boldsymbol{e}_t{}'|\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}] - s_t \mathbb{E}[\boldsymbol{e}_t|\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}]\mathbf{1}' \\
&\qquad - s_t \mathbf{1} \mathbb{E}'[\boldsymbol{e}_t|\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}] + s_t^2 I \\
&= \mu^{\boldsymbol{e}}(\mu^{\boldsymbol{e}})' + \Sigma^{\boldsymbol{e}} - s_t \mu^{\boldsymbol{e}} \mathbf{1}' - s_t \mathbf{1}(\mu^{\boldsymbol{e}})' + s_t^2 I \\
&:= \Delta_t^{\boldsymbol{P}}(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}}),
\end{aligned}
$$

as functions of unknown $(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$ at each time, and I is the identity matrix.

Problem Formulation

Note that the reducible uncertainty in $x_{t+1} = s_t x_t + \boldsymbol{P}'_t \boldsymbol{u}_t$ is (indirectly) observable, compared with the Bayesian RL example of Part 1.

Different from settings of Li and Ng (2000), the first two central moments of $\boldsymbol{P}_t$ here are unknown but conditionally given by

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{P}_t|\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}] &= \mathbb{E}[\boldsymbol{e}_t - s_t\mathbf{1}|\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}] = \mu^{\boldsymbol{e}} - s_t\mathbf{1} := \mu_t^{\boldsymbol{P}}(\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}),\\
\mathbb{E}[\boldsymbol{P}_t\boldsymbol{P}'_t|\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}] &= \mathbb{E}[(\boldsymbol{e}_t - s_t\mathbf{1})(\boldsymbol{e}_t - s_t\mathbf{1})'|\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}]\\
&= \mathbb{E}[\boldsymbol{e}_t\boldsymbol{e}_t{}'|\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}] - s_t\mathbb{E}[\boldsymbol{e}_t|\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}]\mathbf{1}'\\
&\qquad -s_t\mathbf{1}\mathbb{E}'[\boldsymbol{e}_t|\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}] + s_t^2 I\\
&= \mu^{\boldsymbol{e}}(\mu^{\boldsymbol{e}})' + \Sigma^{\boldsymbol{e}} - s_t\mu^{\boldsymbol{e}}\mathbf{1}' - s_t\mathbf{1}(\mu^{\boldsymbol{e}})' + s_t^2 I\\
&:= \Delta_t^{\boldsymbol{P}}(\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}}),
\end{aligned}
$$

as functions of unknown $(\mu^{\boldsymbol{e}},\Sigma^{\boldsymbol{e}})$ at each time, and I is the identity matrix.

It turns out that solving $(\mathcal{A}(\lambda,\omega))$ with Bayesian learning yields the same form as in the full-knowledge case (Li et al., 1998).

Solution by Dynamic Programming

Theorem

The optimal policy of $(\mathcal{A}(\lambda, \omega))$ *at time* $t = 0, 1, \cdots, T-1$ *is given by*

$$\boldsymbol{u}_t(x_t) = -(\Delta_t^{\boldsymbol{P}})^{-1} \mu_t^{\boldsymbol{P}} (s_t x_t - \frac{\lambda_{t+1}}{2\omega_{t+1}}), \tag{14}$$

where $\mu_t^{\boldsymbol{P}} = \mu_t - s_t \mathbf{1}$ *and*

$$\Delta_t^{\boldsymbol{P}} = \left(\mu_t(\mu_t)' + \frac{\Psi_t}{\kappa_t(\nu_t - n - 1)}\right) + \frac{\Psi_t}{\nu_t - n - 1} - s_t \mu_t \mathbf{1}' - s_t \mathbf{1}(\mu_t)' + s_t^2 I, \tag{15}$$

$$\begin{aligned} \lambda_t &= \lambda_{t+1} s_t^2 (1 - \mu_t^{\boldsymbol{P}} (\Delta_t^{\boldsymbol{P}})^{-1} \mu_t^{\boldsymbol{P}}), \ \lambda_T = \lambda, \\ \omega_t &= \omega_{t+1} s_t (1 - \mu_t^{\boldsymbol{P}} (\Delta_t^{\boldsymbol{P}})^{-1} \mu_t^{\boldsymbol{P}}), \ \omega_T = \omega, \end{aligned}$$

with hyperparameters $(\mu_t, \kappa_t, \Psi_t, \nu_t)$ *updated forwardly through (13) based on online observations, given initial* $(\mu_0, \kappa_0, \Psi_0, \nu_0)$.

Special Case: Mean is Unknown While Variance is Known

Propostion

Normal-inverse-Wishart belief with known $\Sigma^{\boldsymbol{e}} = \Sigma$ reduces to the normal belief $\mathcal{N}(\mu_t, \frac{1}{\kappa_t}\Sigma)$ on $\mu^{\boldsymbol{e}}$ as a special case, with the same updating rules

$$\begin{aligned} \mu_{t+1} &= \frac{\kappa_t \mu_t + \tilde{R}_{t+1}}{\kappa_t + 1}, \\ \kappa_{t+1} &= \kappa_t + 1. \end{aligned}$$

The optimal policy takes the same structure

$$\boldsymbol{u}_t(x_t) = -(\Delta_t^{\boldsymbol{P}})^{-1} \mu_t^{\boldsymbol{P}} (s_t x_t - \frac{\lambda_{t+1}}{2\omega_{t+1}}), \tag{16}$$

where $\mu_t^{\boldsymbol{P}} = \mu_t - s_t \mathbf{1}$, *but* $\Delta_t^{\boldsymbol{P}} = \left(\mu_t(\mu_t)' + \frac{1}{\kappa_t}\Sigma\right) + \Sigma - s_t \mu_t \mathbf{1}' - s_t \mathbf{1}(\mu_t)' + s_t^2 I.$

Special Case: Mean is Unknown While Variance is Known

Propostion

Normal-inverse-Wishart belief with known $\Sigma^{\boldsymbol{e}} = \Sigma$ reduces to the normal belief $\mathcal{N}(\mu_t, \frac{1}{\kappa_t}\Sigma)$ on $\mu^{\boldsymbol{e}}$ as a special case, with the same updating rules

$$\mu_{t+1} = \frac{\kappa_t \mu_t + \tilde{R}_{t+1}}{\kappa_t + 1},$$
$$\kappa_{t+1} = \kappa_t + 1.$$

The optimal policy takes the same structure

$$\boldsymbol{u}_t(x_t) = -(\Delta_t^{\boldsymbol{P}})^{-1}\mu_t^{\boldsymbol{P}}(s_t x_t - \frac{\lambda_{t+1}}{2\omega_{t+1}}), \tag{16}$$

where $\mu_t^{\boldsymbol{P}} = \mu_t - s_t\mathbf{1}$, *but* $\Delta_t^{\boldsymbol{P}} = \left(\mu_t(\mu_t)' + \frac{1}{\kappa_t}\Sigma\right) + \Sigma - s_t\mu_t\mathbf{1}' - s_t\mathbf{1}(\mu_t)' + s_t^2 I.$

Note that the belief uncertainty on $\mu^{\boldsymbol{e}}$: $1/\kappa_t\Sigma \to 0$ as $t \to \infty$, independent of observations.

Experimental Results

There are three risky assets ($n = 3$) in the market and the real random total return $\boldsymbol{e}$ is normally distributed with mean $\mu = (1.162\ 1.246\ 1.228)'$ and covariance

$$\Sigma = \begin{pmatrix} 0.0146 & 0.0187 & 0.0145 \\ 0.0187 & 0.0854 & 0.0104 \\ 0.0145 & 0.0104 & 0.0289 \end{pmatrix}.$$

Experimental Results

There are three risky assets ($n = 3$) in the market and the real random total return $\boldsymbol{e}$ is normally distributed with mean $\mu = (1.162\ 1.246\ 1.228)'$ and covariance

$$\Sigma = \begin{pmatrix} 0.0146 & 0.0187 & 0.0145 \\ 0.0187 & 0.0854 & 0.0104 \\ 0.0145 & 0.0104 & 0.0289 \end{pmatrix}.$$

An investor does not know neither the true μ nor the true Σ, who also adopts NIW prior on $(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$ with $\mu_0 = (1.1\ 1.2\ 1.3)', \kappa_0 = 2, \nu_0 = 5$, and

$$\Psi_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Experimental Results

There are three risky assets ($n = 3$) in the market and the real random total return $\boldsymbol{e}$ is normally distributed with mean $\mu = (1.162\ 1.246\ 1.228)'$ and covariance

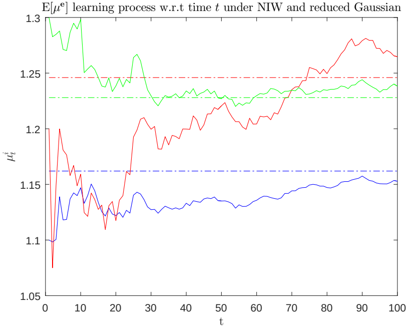
$$\Sigma = \begin{pmatrix} 0.0146 & 0.0187 & 0.0145 \\ 0.0187 & 0.0854 & 0.0104 \\ 0.0145 & 0.0104 & 0.0289 \end{pmatrix}.$$

An investor does not know neither the true μ nor the true Σ, who also adopts NIW prior on $(\mu^{\boldsymbol{e}}, \Sigma^{\boldsymbol{e}})$ with $\mu_0 = (1.1\ 1.2\ 1.3)', \kappa_0 = 2, \nu_0 = 5$, and

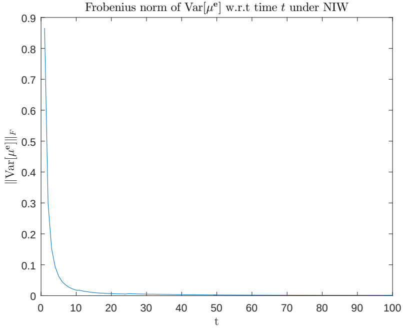
$$\Psi_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The risk-free rate s_t is set to be 1.04 all the time. Besides, $\omega = 2, \lambda = 1$, and $x_0 = 1$ as initial wealth. In order to clearly see the learning process on the unknown parameters, we set $T = 100$, and empirical total return $\tilde{R}_t$ is sampled from $\mathcal{N}(\mu, \Sigma)$ at each time by MATLAB.

Experimental Results



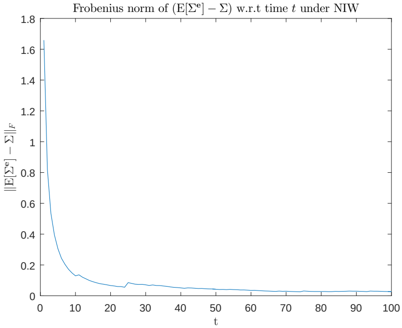
(a)



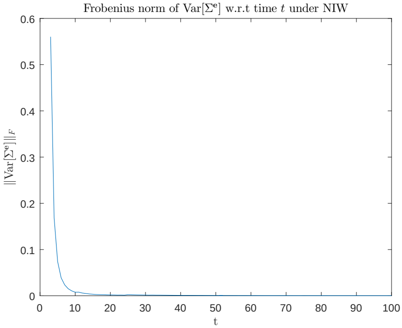
(b)

Figure: 1. μ^e learning under NIW and reduced Gaussian belief.

Experimental Results



(a)



(b)

Figure: 2. Σ^e learning under NIW.

Experimental Results

Figure 1(a) illustrates how μ_t^i (solid lines) approaches to real μ^i (dotted lines) for each asset i as time goes by, and Figure 1(b) exhibits its variance of t-distribution as marginal distribution of NIW, measured by Frobenius norm, reduces to zero at the same time.

Experimental Results

Figure 1(a) illustrates how μ_t^i (solid lines) approaches to real μ^i (dotted lines) for each asset i as time goes by, and Figure 1(b) exhibits its variance of t-distribution as marginal distribution of NIW, measured by Frobenius norm, reduces to zero at the same time.

Note that the unknown $\mu^{\boldsymbol{e}}$ updatings under both NIW and the corresponding reduced Gaussian are the same.

Experimental Results

Figure 1(a) illustrates how μ_t^i (solid lines) approaches to real μ^i (dotted lines) for each asset *i* as time goes by, and Figure 1(b) exhibits its variance of *t*-distribution as marginal distribution of NIW, measured by Frobenius norm, reduces to zero at the same time.

Note that the unknown $\mu^{\boldsymbol{e}}$ updatings under both NIW and the corresponding reduced Gaussian are the same.

Likewise, Figures 2(a) and 2(b) show distance between expectation of $\Sigma^{\boldsymbol{e}}$ and true Σ, and variance of $\Sigma^{\boldsymbol{e}}$ from inverse Wishart as marginal distribution of NIW, respectively, which both converge to zero.

Experimental Results

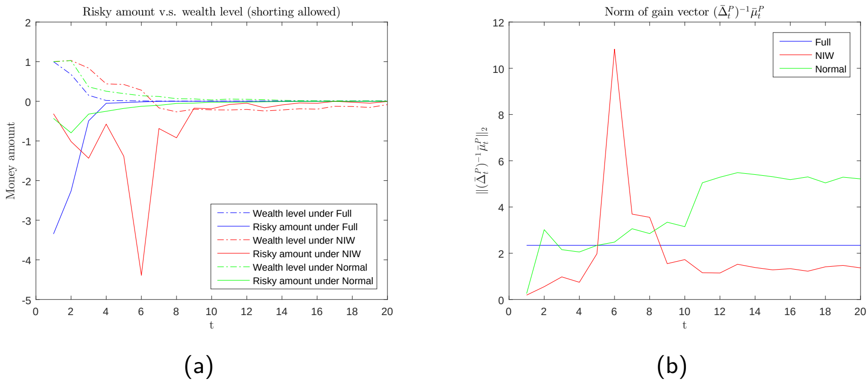


Figure: 3. Policy behavior among Full information, NIW belief and reduced Gaussian belief.

Experimental Results

Figure 3 shows policy behavior under different cases for first twenty time points.

Experimental Results

Figure 3 shows policy behavior under different cases for first twenty time points.

From Figure 3(a) we can see at the beginning the cautious property of policies from lack of knowledge situation (NIW in red and reduced Gaussian in green), compared with full information case (blue line), in terms of *lower* proportional amount in shorting risky assets (solid lines) from total wealth (dotted lines).

Experimental Results

Figure 3 shows policy behavior under different cases for first twenty time points.

From Figure 3(a) we can see at the beginning the cautious property of policies from lack of knowledge situation (NIW in red and reduced Gaussian in green), compared with full information case (blue line), in terms of *lower* proportional amount in shorting risky assets (solid lines) from total wealth (dotted lines).

This phenomenon can be also seen from Figure 3(b), where we measure Frobenius norms of different gain vectors.

Selected References

Aoki, M. (1967). *Optimization of Stochastic Systems: Topics in Discrete-Time Systems*, volume 32. Academic Press.

Deshpande, J., Upadhyay, T., and Lainiotis, D. (1973). Adaptive control of linear stochastic systems. *Automatica*, 9(1):107–115.

Feldbaum, A. (1960–1961). Dual control theory I–IV. *Avtomatika i Telemekhanika*, 21(9), 21(11), 22(1), 22(2).

Kirk, D. E. (1970). *Optimal Control Theory: An Introduction*. Springer.

Li, D., Chan, T.-F., and Ng, W.-L. (1998). Safety-first dynamic portfolio selection. *Dynamics of Continuous, Discrete and Impulsive Systems Series B: Applications and Algorithms*, 4(4):585–600.

Li, D. and Ng, W.-L. (2000). Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Mathematical Finance*, 10(3):387–406.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution. Technical report, University of British Columbia.

Rockafellar, R. T. and Wets, R. J.-B. (1991). Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16(1):119–147.

Thank you for your attention!