Reinforcement Learning - Lecture 4

Optimal Control and The Rocket Example

ENS M2

October 1, 2025

Contents

1 Recap: Optimal Policy **2**

2 The Rocket Example: LQG Control **2**

2.1 Problem Formulation 2

2.2 Solution by Backward Induction 2

2.3 Generalization by Induction 3

2.4 General Formula 3

3 General LQG Control **4**

3.1 General Formulation 4

4 Evaluation of a Given Policy **4**

4.1 Visit Probabilities 4

4.2 Time Spent in Each State 5

4.3 Avoidance Probability 5

5 Variance Computation by Backward Induction **5**

5.1 Second-Order Moments 5

5.2 Recurrence for Moments 5

6 Distributional Evaluation **6**

6.1 Objective 6

1 Recap: Optimal Policy

Example

Optimal Backward Induction Algorithm

$$\begin{cases} Q(s,a,t) = \sum_{s' \in \mathcal{S}} k(s,a,s')[r(s,a,s') + V^*_{t+1}(s')] \\ V^*_t(s) = \max_{a \in \mathcal{A}} Q(s,a,t) \\ \pi^*_t(s) = \arg\max_{a \in \mathcal{A}} Q(s,a,t) \end{cases}$$

2 The Rocket Example: LQG Control

2.1 Problem Formulation

Example

Rocket Control

- State space: $\mathcal{S} = \mathbb{R}$
- Action space: $\mathcal{A} = \mathbb{R}$
- Reward:
  - $t < T$: $r_t(s,a,s') = -a^2$
  - $t = T$: $r_T(s,a,s') = -\rho(s^* - s)^2$ where s^* is the target
- Transition: $S_{t+1} = S_t + A_t + U_t$ where U_t are i.i.d. with $\mathrm{Var}(U_t) = \sigma^2$
- Cumulated reward: $W_T = -\sum_{t=1}^{T-1} A_t^2 - \rho(s^* - S_T)^2$

2.2 Solution by Backward Induction

We use backward induction:

At $t = T$:

$$V^*(s,T) = -\rho(s^* - s)^2$$

At $t = T-1$:

$$Q(s,a,T-1) = -\rho\mathbb{E}[(s^* - (s + a + U_{T-1}))^2] - a^2$$

$$= -\rho\mathbb{E}[(s^* - s - a - U_{T-1})^2] - a^2$$

$$= -\rho((s^* - s - a)^2 + \mathbb{E}[U_{T-1}^2]) - a^2$$

$$= -\rho((s^* - s - a)^2 + \sigma^2) - a^2$$

To find the optimal action, we differentiate with respect to a:

$$\frac{\partial Q}{\partial a} = -\rho \cdot 2(s^* - s - a)(-1) - 2a = 2\rho(s^* - s - a) - 2a = 0$$

$$\Rightarrow \pi^*(s, T-1) = \frac{\rho(s^* - s)}{\rho + 1}$$

Value function at $t = T - 1$:

$$V^*(s, T-1) = Q(s, \pi^*(s, T-1), T-1)$$

Substituting the optimal action:

$$V^*(s, T-1) = -\rho\left(s^* - s - \frac{\rho(s^* - s)}{\rho + 1}\right)^2 - \left(\frac{\rho(s^* - s)}{\rho + 1}\right)^2 - \rho\sigma^2$$

$$= -\rho(s^* - s)^2\left(1 - \frac{\rho}{\rho + 1}\right)^2 - \frac{\rho^2(s^* - s)^2}{(\rho + 1)^2} - \rho\sigma^2$$

$$= -\rho(s^* - s)^2\left(\frac{1}{\rho + 1}\right)^2 - \frac{\rho^2(s^* - s)^2}{(\rho + 1)^2} - \rho\sigma^2$$

$$= -\frac{\rho(s^* - s)^2}{\rho + 1} - \rho\sigma^2$$

2.3 Generalization by Induction

At $t = T - 2$:

$$Q(s, a, T-2) = \mathbb{E}[V^*(s + a + U_{T-2}, T-1)] - a^2$$

$$= \mathbb{E}\left[-\frac{\rho}{\rho + 1}(s^* - (s + a + U_{T-2}))^2 - \rho\sigma^2\right] - a^2$$

$$= -\frac{\rho}{\rho + 1}((s^* - s - a)^2 + \sigma^2) - \rho\sigma^2 - a^2$$

Differentiating:

$$\frac{\partial Q}{\partial a} = -\frac{\rho}{\rho + 1} \cdot 2(s^* - s - a)(-1) - 2a = \frac{2\rho}{\rho + 1}(s^* - s - a) - 2a = 0$$

$$\Rightarrow a^* = \pi^*(s, T-2) = \frac{\rho(s^* - s)}{2\rho + 1}$$

Value function at $t = T - 2$:

$$V^*(s, T-2) = -\frac{\rho(s^* - s)^2}{1 + 2\rho} - \left(1 + \frac{1}{\rho + 1}\right)\rho\sigma^2$$

2.4 General Formula

Theorem

LQG Solution
By induction, we can deduce the general formula:

$$V^*(s, T-k) = -\frac{\rho(s^* - s)^2}{1 + k\rho} - \left(1 + \frac{1}{\rho+1} + \cdots + \frac{1}{(\rho+1)^{k-1}}\right)\rho\sigma^2$$

$$\pi^*(s, T-k) = \frac{\rho(s^* - s)}{1 + k\rho}$$

3 General LQG Control

3.1 General Formulation

Definition

Linear Quadratic Gaussian (LQG) Control
Class of problems with:

- Linear dynamics: $S_{t+1} = AS_t + BA_t + U_t$
- Quadratic cost: $r(s, a, s') = -a^T Q a$ with Q positive definite
- Gaussian noise: $U_t \sim \mathcal{N}(0, \Sigma)$

Example

Example: Heating

$$S_{t+1} = \lambda S_t + (1 - \lambda)A_t + U_t$$

where the cost is:

$$-\rho(S_T - s^*)^2 - \sum_{t=1}^{T-1} A_t^2$$

Temperature evolves according to linear dynamics with inertia λ and control $(1 - \lambda)$.

4 Evaluation of a Given Policy

4.1 Visit Probabilities

For a fixed policy π and initial state $s \in \mathcal{S}$:

$$R_s^\pi = \mathbb{P}\left(\bigcup_{t=1}^{T} \{S_t = \sigma\} \middle| S_1 = s\right)$$

for $\sigma \in \mathcal{S}$ is the probability of visiting state σ.

4.2 Time Spent in Each State

$$T^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{T} \mathbf{1}_{S_t=\sigma}\middle|S_1 = s\right] = T^\pi(s,1)$$

With the recurrence:

$$\begin{cases} T^\pi(s,t) = \sum_{s'\in\mathcal{S}} k(s,\pi(s),s')(\mathbf{1}_{s'=\sigma} + T^\pi(s',t+1)) \\ T^\pi(s,T) = \mathbf{1}_{s=\sigma} \end{cases}$$

4.3 Avoidance Probability

$$A^\pi(s) = 1 - R^\pi(s) = \mathbb{P}\left(\bigcap_{t=1}^{T}\{S_t \neq \sigma\}\middle|S_1 = s\right)$$

is the probability of not visiting state σ.

With the recurrence:

$$\begin{cases} A(s,t) = \sum_{s'\in\mathcal{S}} k(s,\pi(s),s')\mathbf{1}_{s'\neq\sigma} \cdot A(s',t+1) \\ A(s,T) = \mathbf{1}_{s\neq\sigma} \end{cases}$$

$$R(s,t) = 1 - A(s,t)$$

5 Variance Computation by Backward Induction

5.1 Second-Order Moments

$$\mathrm{Var}^\pi(W_t) = \mathbb{E}^\pi[W_t^2] - \mathbb{E}^\pi[W_t]^2$$

$$W_t^2 = \left(\sum_{k=1}^{T} R_k\right)^2$$

At $t = T$:

$$\mathbb{E}[R_T^2|S_T = s] = \mathbb{E}(r(s,\pi(s),S_{T+1})^2|S_T = s) = \sum_{s'\in\mathcal{S}} k(s,\pi(s),s')r(s,\pi(s),s')^2$$

5.2 Recurrence for Moments

$$m_2(s,T-1) = \mathbb{E}[(R_{T-1} + R_T)^2|S_{T-1} = s]$$

$$= \mathbb{E}[R_{T-1}^2 + 2R_{T-1}R_T + R_T^2|S_{T-1} = s]$$

$$= \sum_{s'\in\mathcal{S}} k(s,\pi(s),s')[r(s,\pi(s),s')^2 + 2r(s,\pi(s),s')\mathbb{E}[R_T|S_T = s'] + \mathbb{E}[R_T^2|S_T = s']]$$

Definition

General Recurrence

$$m_2(s,t) = \sum_{s' \in \mathcal{S}} k(s,\pi(s),s')[r(s,\pi(s),s')^2 + 2r(s,\pi(s),s')m_1(s',t+1) + m_2(s',t+1)]$$

$$m_1(s,t) = \sum_{s' \in \mathcal{S}} k(s,\pi(s),s')[r(s,\pi(s),s') + m_1(s',t+1)]$$

with $m(s,T+1) = (0,0)$ and:

$$m(s,t) = \sum_{s' \in \mathcal{S}} k(s,\pi(s),s') \begin{pmatrix} r(s,\pi(s),s') + m_1(s',t+1) \\ r(s,\pi(s),s')^2 + 2r(s,\pi(s),s')m_1(s',t+1) + m_2(s',t+1) \end{pmatrix}$$

$$\mathrm{Var}^\pi(W_t) = m_2(s,1) - m_1(s,1)^2$$

6 Distributional Evaluation

6.1 Objective

Note

Goal
Evaluate the law of $\nu(s,t) =$ law of $R_t + R_{t+1} + \cdots + R_T$ given $S_t = s \in \mathcal{M}_1(\mathbb{R})$.

$\nu(s,T) =$ law of $R_T = r(S_T, \pi(S_T), S_{T+1})$ given $S_T = s$.

In a windy environment, this equals a mixture of Dirac distributions.

$\nu(s,t) =$ law of $R_t + R_{t+1} + \cdots + R_T$ given $S_t = s$

$=$ mixture law of $r(s,\pi(s),s') + R_{t+1} + \cdots + R_T$ given $S_{t+1} = s' \sim \nu(s',t+1)$