

Lecture 2

Markov Decision Processes and Optimal Control

Outline

- State Representation
- Definitions
- Continuous vs Discrete MDP
- Control/Decisions
- Optimal Control
- Optimal Control
- Stability
- Linear Systems

State Space

Dynamic Systems and States

General Dynamic System

1. State: $s_t \in S$
2. Dynamics: $s_{t+1} = f_t(s_t, w_t)$
3. Model of uncertainty: e.g., $\mathbb{P}(\cdot)$ of w_t

Dynamic Systems and States

General Dynamic System

1. State: $s_t \in S$
2. Dynamics: $s_{t+1} = f_t(s_t, w_t)$
3. Model of uncertainty: e.g., $\mathbb{P}(\cdot)$ of w_t

State: A set of variables that summarizes all the necessary information about the system's current and past conditions to predict its future behavior (given past exogenous disturbances), without needing to know the past history of the system.

Dynamic Systems and States

General Dynamic System

1. State: $s_t \in S$
2. Dynamics: $s_{t+1} = f_t(s_t, w_t)$
3. Model of uncertainty: e.g., $\mathbb{P}(\cdot)$ of w_t

State: A set of variables that summarizes all the necessary information about the system's current and past conditions to predict its future behavior (given past exogenous disturbances), without needing to know the past history of the system.

Past $\rightarrow s_t \rightarrow$ Future (Markov property)

Independent Noise

- w_t are independent random variables.
- The system dynamics:

$$s_{t+1} = \frac{1}{2}s_t + w_t$$

Independent Noise

- w_t are independent random variables.
- The system dynamics:

$$s_{t+1} = \frac{1}{2}s_t + w_t$$

Response: Assuming s_0 is fixed,

$$s_t = \left(\frac{1}{2}\right)^t s_0 + \sum_{\tau=0}^{t-1} \left(\frac{1}{2}\right)^{t-\tau-1} w_{\tau}$$

Independent Noise

- w_t are independent random variables.
- The system dynamics:

$$s_{t+1} = \frac{1}{2}s_t + w_t$$

Response: Assuming s_0 is fixed,

$$s_t = \left(\frac{1}{2}\right)^t s_0 + \underbrace{\sum_{\tau=0}^{t-1} \left(\frac{1}{2}\right)^{t-\tau-1} w_\tau}_{\text{Convolution}}$$

Convolution

Discrete State

Let s_t be the number of jobs in queue at time t . w_t models the arrival of jobs, IID process.

Then the queue dynamics follow:

$$s_{t+1} = \max \{0, s_t + w_t - 1\}$$

Assume $w(t)$ are IID

Is this a state-space representation?

Discrete Example: State dependent noise

Suppose:

$$w_t = g(s_t, \bar{w}_t), \quad \bar{w}_t \text{ is i.i.d.}$$

Then the system evolves as:

$$s_{t+1} = f(s_t, w_t) = f(s_t, g(s_t, \bar{w}_t))$$

Converted to Independent Noise Case

We do not need to find g

Example: State dependent noise

Arrival process z_t :

If $s_t < 10$, the person joins. If $s_t \geq 10$, the person flips a coin.

The queue update equation is:

$$s_{t+1} = \max \{0, s_t + z_t - 1\}$$

Example: State dependent noise

Arrival process z_t :

If $s_t < 10$, the person joins. If $s_t \geq 10$, the person flips a coin.

The queue update equation is:

$$s_{t+1} = \max\{0, s_t + z_t - 1\}$$

where z_t is a state-dependent random variable defined by:

$$z_t = g(s_t, \bar{w}_t)$$

with $\bar{w}_t$ being i.i.d. noise.

Kernel Representations

- Let $\mathcal{S}$ be a finite state space, with $|\mathcal{S}| = n$
- Define transition probabilities $P(s'; s)$

Given:

$$s_{t+1} = f(s_t, w_t)$$

The transition probability is:

$$\begin{aligned} P(s'; s) &= \mathbb{P}(s_{t+1} = s' \mid s_t = s) \\ &= \mathbb{P}(f(s, w_t) = s') \end{aligned}$$

Is the converse true?

Kernel Representations

$$\mathbb{P}(s'; s) \rightsquigarrow s_{t+1} = f(s_t, w_t)$$

Hint: Simulate using a uniform distribution on $[0,1]$.
Divide the access to reflect the probabilities!

Full State	Aggregated State
PORTFOLIO	
p_i = price of asset i σ_i = volatility of asset i ρ = correlation matrix s = sentiment score m = macro indicators f = fund flows Full: $(p_1, \dots, p_n, \sigma_1, \dots, \sigma_n, \rho, s, m, f, \dots)$	V_p = portfolio value σ_p = portfolio volatility w_s = sector weights Aggregated: (V_p, σ_p, w_s)
EPIDEMIC	
S = susceptible E = exposed I = infected R = recovered S_a, E_a, I_a, R_a = age-stratified compartments $v_1, v_2, \dots$ = variant prevalences h_r = hospitalization rate Full: $(S, E, I, R, S_a, E_a, I_a, R_a, v_1, v_2, \dots)$	C = cumulative cases R_c = total recovered I_a = active infections Aggregated: (S, I, R) or (C, R_c, I_a)
SUPPLY CHAIN	
q_i = inventory at node i t_i = in-transit shipment i c_i = supplier capacity at node i d_i = demand forecast at node i δ_i = disruption status at node i ℓ_i = lead time at node i Full: $(q_1, q_2, \dots, t_1, t_2, \dots, c_1, c_2, \dots, d_1, d_2, \dots, \delta_1, \delta_2, \dots, \ell_1, \ell_2, \dots)$	Q_{total} = total network inventory T_{total} = total in-transit $\bar{\ell}$ = average lead time s_{supply} = supply status (sufficient, stressed, critical) Aggregated: $(Q_{total}, T_{total}, \bar{\ell})$
POWER GRID	
v_i = voltage at bus i i_j = current on line j g_i = generation at site i d_i = demand in zone i r_s = spinning reserves r_n = non-spinning reserves e_i = equipment status on line i Full: $(v_1, \dots, v_n, i_1, \dots, i_m, f, g_1, g_2, \dots, d_1, d_2, \dots, r_s, r_n, e_1, \dots, e_m)$	f = grid frequency G_{total} = total generation D_{total} = total demand r = reserve margin Aggregated: $(f, G_{total}, D_{total}, r)$
MEDICAL (HIV/AIDS)	
v_b = viral load in blood v_t = viral load in tissue c_4 = CD4 count c_8 = CD8 count k_f = kidney function l_f = liver function $\alpha_1, \alpha_2, \dots$ = antiretroviral drug levels g = viral genotype μ = resistance mutations γ = treatment adherence Full: $(v_b, v_t, c_4, c_8, k_f, l_f, \alpha_1, \alpha_2, \dots, g, \mu, \gamma, \dots)$	c_4 = CD4 count v_b = viral load stage $\in$ {asymptomatic, symptomatic, AIDS} Aggregated: (c_4, v_b)

Markov Decision Processes (MDP)

Markov Decision Process (MDP) (control systems)

1. State: $s_t \in \mathcal{S}$
2. Dynamics: $s_{t+1} = f_t(s_t, a_t, w_t)$
3. Stage cost: $c_t(s_t, a_t, w_t)$
4. Control constraints: $a_t \in A_t(s_t)$
5. Model of uncertainty: e.g., $\mathbb{P}(\cdot \mid s_t, a_t)$ of w_t

A feedback policy π , is a sequence of functions: $\pi = \{\mu_0, \dots, \mu_{N-1}\}$
where $\mu_t : s_t \mapsto a_t = \mu_t(s_t) \in A_t(s_t)$

Memoryless Feedback

Is it Markov?

Transition Kernel Representation

$$s_{t+1} = f_t(s_t, a_t, w_t) \iff \mathbb{P}(s' \mid s, a)$$

$$\mathbb{P}(f_t(s, a, w_t) = s' \mid s, a) = p_t(s'; s, a)$$

If $|\mathcal{S}| = n$ (finite state space), we write the kernel as:

$$p_t(j; i, u)$$

To compute expectations:

$$\mathbb{E}[g(s_{t+1}) \mid s_t = s, a_t = a] = \sum_{s'} g(s') p_t(s'; s, a)$$

Example: A queuing control problem

- ▶ A service facility receives customers at **random times**
- ▶ The manager **controls the service rate** by allocating resources:
 - ▶ Add or remove servers
 - ▶ Allocate compute or labor
 - ▶ Prioritize customer classes
- ▶ **Objective:** Balance two costs
 - ▶ Cost of waiting: Long queues $\Rightarrow$ dissatisfied customers, lost revenue
 - ▶ Cost of service: Higher service rate $\Rightarrow$ labor, infrastructure
- ▶ This is a **control problem**

Example: Model and State

Queue evolution:

$$n_{t+1} = \max(0, n_t + A_t - S_t)$$

where:

- ▶ $A_t \sim \text{Poisson}(\lambda)$ = arrivals during interval $[t, t + 1]$
- ▶ $S_t \sim \text{Poisson}(\mu_t)$ = services completed during $[t, t + 1]$
- ▶ λ = fixed arrival rate

Transition probability:

$$P(n'|n, \mu) = P(A_t - S_t = n' - n)$$

Is this a state representation?

Example: Memoryless Distribution

- ▶ Arrivals and service completions are **memoryless** (independent)
- ▶ The past does not affect the future conditional on n_t

For $n' \geq 0$:

$$\begin{aligned} P(n'|n, \mu) &= \sum_{k=0}^{\infty} P(A_t = k) \cdot P(S_t = k - n + n') \\ &= \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \cdot \frac{e^{-\mu} \mu^{k-n+n'}}{(k - n + n')!} \end{aligned}$$

Interpretation:

- ▶ Net flow: $n' - n = A_t - S_t$
- ▶ We need exactly k arrivals and $k - n + n'$ services
- ▶ Probability decreases in μ (faster service $\Rightarrow$ more likely to drain queue)
- ▶ Probability increases in λ (more arrivals $\Rightarrow$ larger queue)

MDP (general strategies)

Control Strategy

$$\mu_t(s_\tau, a_\tau, s_{\tau+1}, a_{\tau+1} \dots s_t)$$

Is there a problem with this?

$$\text{Dynamics: } s_{t+1} = f_t(s_t, \mu_t(s_\tau, a_\tau \dots, s_t), w_t)$$

Is it Markov? What is the state?

Example: Autoregressive Models with Inputs

$$y_t = \sum_{i=1}^p d_i y_{t-i} + b a_{t-1} + \varepsilon_{t-1},$$

state: $S_t = \begin{bmatrix} y_{t-p} \\ \vdots \\ y_t \end{bmatrix},$

$$S_{t+1} = \begin{bmatrix} y_{t-p+1} \\ \vdots \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ d_p & d_{p-1} & \cdots & d_2 & d_1 \end{bmatrix} S_t + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ b \end{bmatrix} a_t + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \varepsilon_t.$$

Example: Autoregressive with moving average models

$$y_t = d_1 y_{t-1} + \cdots + d_p y_{t-p} + b_1 a_{t-1} + \cdots + b_{m_1} a_{t-m_1} \\ + \epsilon_t + c_1 \epsilon_{t-1} + \cdots + c_{m_2} \epsilon_{t-m_2}$$

Use superposition: Linearity allows the realization of each separate system

Dynamic Programming and Optimal Control

Introduction to Dynamic Programming

A mathematical framework of **sequential** decision making under uncertainty

1. State: $s_t \in S$
 2. Dynamics: $s_{t+1} = f_t(s_t, a_t, w_t)$
 3. Stage cost: $c_t(x_t, a_t, w_t)$
 4. Control constraints: $a_t \in A_t(x_t)$
 5. Model of uncertainty: e.g., $\mathbb{P}(\cdot \mid s_t, a_t)$ of w_t
-

Introduction to Dynamic Programming

A mathematical framework of **sequential** decision making under uncertainty

1. State: $s_t \in S$
2. Dynamics: $s_{t+1} = f_t(s_t, a_t, w_t)$
3. Stage cost: $c_t(s_t, a_t, w_t)$
4. Control constraints: $a_t \in A_t(s_t)$
5. Model of uncertainty: e.g., $\mathbb{P}(\cdot \mid s_t, a_t)$ of w_t

$$\mathbb{E} \left[c_N(s_N) + \sum_{t=0}^{N-1} c_t(s_t, a_t, w_t) \right]$$

Introduction to Dynamic Programming

A mathematical framework of **sequential** decision making under uncertainty

1. State: $s_t \in S$
2. Dynamics: $s_{t+1} = f_t(s_t, a_t, w_t)$
3. Stage cost: $c_t(s_t, a_t, w_t)$
4. Control constraints: $a_t \in A_t(s_t)$
5. Model of uncertainty: e.g., $\mathbb{P}(\cdot \mid s_t, a_t)$ of w_t

$$\mathbb{E} \left[c_N(s_N) + \sum_{t=0}^{N-1} c_t(s_t, a_t, w_t) \right]$$

■ A feedback policy π , is a sequence of functions: $\pi = \{\mu_0, \dots, \mu_{N-1}\}$
where $\mu_t : s_t \mapsto a_t = \mu_t(s_t) \in A_t(s_t)$

$$V_N^\pi(s_0) = \mathbb{E} \left[c_N(s_N) + \sum_{t=0}^{N-1} c_t(s_t, \mu_t(s_t), w_t) \right]$$

Introduction to Dynamic Programming

A mathematical framework of **sequential** decision making under uncertainty

1. State: $s_t \in S$
2. Dynamics: $s_{t+1} = f_t(s_t, a_t, w_t)$
3. Stage cost: $c_t(s_t, a_t, w_t)$
4. Control constraints: $a_t \in A_t(s_t)$
5. Model of uncertainty: e.g., $\mathbb{P}(\cdot \mid s_t, a_t)$ of w_t

$$\mathbb{E} \left[c_N(s_N) + \sum_{t=0}^{N-1} c_t(s_t, a_t, w_t) \right]$$

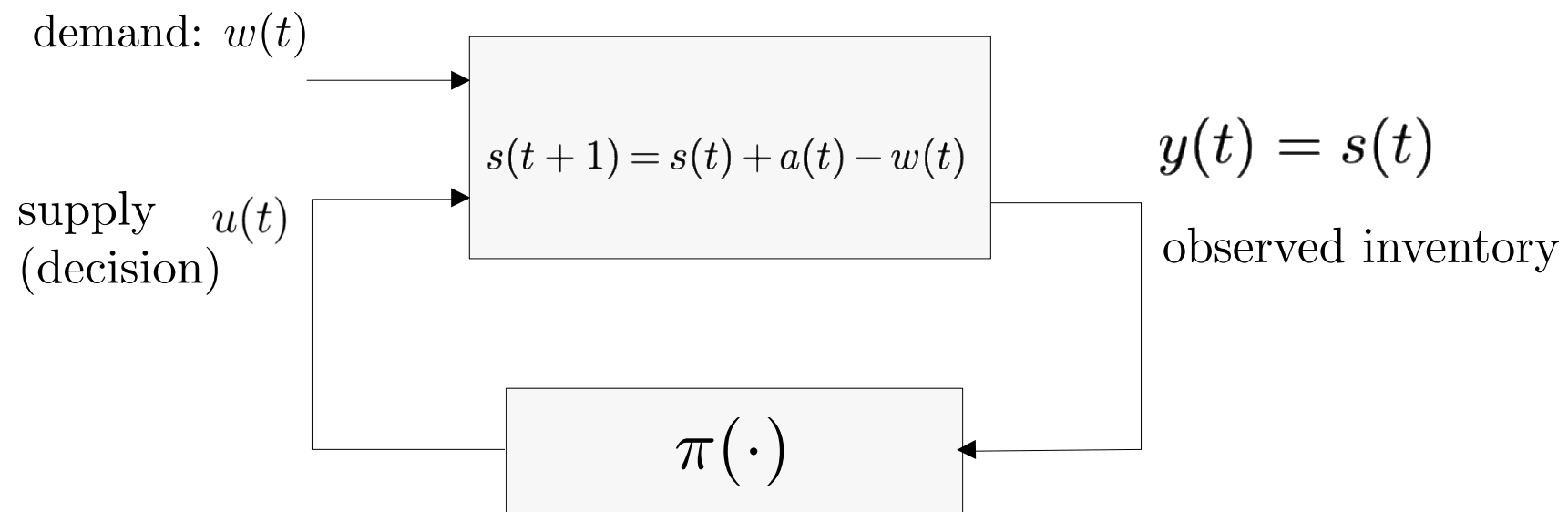
■ A feedback policy π , is a sequence of functions: $\pi = \{\mu_0, \dots, \mu_{N-1}\}$
where $\mu_t : s_t \mapsto a_t = \mu_t(s_t) \in A_t(s_t)$

$$V_N^\pi(s_0) = \mathbb{E} \left[c_N(s_N) + \sum_{t=0}^{N-1} c_t(s_t, \mu_t(s_t), w_t) \right]$$

■ $V^*(s_0) = \min_{\pi} V^\pi(s_0)$

■ $\pi^* \in \arg \min_{\pi} V^\pi(s_0)$ i.e., $V^{\pi^*}(s_0) = V^*(s_0)$

Motivation: Inventory control



Cumulative cost over N periods:

$$\mathbb{E} \left[c(s(N)) + \sum_{t=0}^{N-1} \{c(s_t, a_t)\} \right]$$

How should we make supply order decisions to minimize this cost / maximize profit?

Motivation: Inventory control

$$\min_{\pi} \quad \mathbb{E} \left[r(s_N) + \sum_{t=0}^{N-1} da_t + g(s_t) \right]$$

$$\text{s.t.} \quad s_{t+1} = s_t + a_t - w_t, \quad a_t \geq 0, \quad w_t \sim \text{i.i.d.}$$

$$g(s) = p \max(hs, -qs)$$

Motivation: Stopping-time Problem

State Evolution:

$$S_{t+1} = S_t + W_t, \quad S_0 = s_0 \text{ (given)}$$

where: W_t = random price change (e.g., $W_t \in \{-1, 0, +1\}$ or $N(0, \sigma^2)$)

Stopping-time Problem: A call Option

State Evolution:

$$S_{t+1} = S_t + W_t, \quad S_0 = s_0 \text{ (given)}$$

where: W_t = random price change (e.g., $W_t \in \{-1, 0, +1\}$ or $N(0, \sigma^2)$)

Parameters: Strike price p , time horizon T , discount factor β

Action: $a_t \in \{0, 1\}$ (0 = hold, 1 = exercise)

Cost Function (Reward Max):

$$J = \mathbb{E} \left[\sum_{t=0}^T a_t (S_t - p) \right]$$

Constraint: $\sum_t a_t \leq 1$ (exercise at most once)

Queuing Example

Stage Cost:

$$g_t(n_t, \mu_t) = h \cdot n_t + c(\mu_t)$$

where:

- ▶ $h > 0$ = holding cost per customer per period (congestion penalty)
- ▶ $c(\mu_t)$ = operational cost of service rate μ_t

Service Cost Specifications:

- ▶ **Linear:** $c(\mu) = c \cdot \mu$ (cost proportional to rate)
- ▶ **Affine:** $c(\mu) = c_0 + c_1\mu$ (fixed + variable)
- ▶ **Quadratic:** $c(\mu) = c_0 + c_1\mu + \frac{c_2}{2}\mu^2$ (capacity constraints, overtime)

Total Cost:

$$J = \mathbb{E} \left[\sum_{t=0}^{T-1} (h \cdot n_t + c(\mu_t)) + h_T \cdot n_T \right]$$

where $h_T \cdot n_T$ = terminal cost (penalize remaining queue)

Questions

More on Dynamical Systems

Linear Systems

Linear System

$$x_{t+1} = A_t x_t + B_t u_t + w_t$$

$$y = C_t x_t$$

$$x_t \in \mathbf{R}^n$$

$$u_t \in \mathbf{R}^p$$

$$y_t \in \mathbf{R}^q$$

$$w_t \in \mathbf{R}^n$$

Solution

$$x_1 = A_0 x_0 + B_0 u_0 + w_0$$

$$x_2 = A_1 A_0 x_0 + A_1 B_0 u_0 + A_1 w_0 + B_1 u_1 + w_1$$

$$x_3 = A_2 A_1 A_0 x_0 + A_2 A_1 B_0 u_0 + A_2 B_1 u_1 + A_2 A_1 w_0 + A_2 w_1 + B_2 u_2 + w_2$$

Solution of Linear Systems

Solving a linear system

$$\Phi(t, \tau) = A_{t-1}A_{t-2} \cdots A_{\tau}$$

$$x_t = \Phi(t, 0)x_0 + \sum_{\tau=1}^{t-1} \Phi(t, \tau)B_{\tau}u_{\tau} + \sum_{\tau=1}^{t-1} \Phi(t, \tau)w_{\tau}$$

$$\Phi(t, t) = I$$

$$x_t = \Phi(t, 0)x_0 + \sum_{\tau=1}^{t-1} \Phi(t, \tau) (B_{\tau}u_{\tau} + w_{\tau})$$

LTI system case:

$$\Phi(t, \tau) = A^{t-\tau}$$

$$x_t = A^t x_0 + \sum_{\tau=1}^{t-1} A^{t-\tau} B_{\tau} u_{\tau} + \sum_{\tau=1}^{t-1} A^{t-\tau} w_{\tau}$$

Stability of Linear Systems

A linear system is said to be 'Asymptotically Stable' if:

$$\forall x_0, \quad x_t \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty$$

For time-varying systems, using the state transition matrix $\Phi(t, \tau)$, stability implies:

$$\Phi(t, \tau) \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty \quad \forall \tau$$

Stability of Linear Systems

A linear system is said to be 'Asymptotically Stable' if:

$$\forall x_0, \quad x_t \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty$$

For time-varying systems, using the state transition matrix $\Phi(t, \tau)$, stability implies:

$$\Phi(t, \tau) \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty \quad \forall \tau$$

LTI Case:

In a time-invariant system, $\Phi(t, \tau) = A^{t-\tau}$. Stability is guaranteed if:

$$A^t \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty$$

which occurs if and only if all eigenvalues of A lie strictly inside the unit circle (i.e., have magnitude less than 1).

Stability dynamical systems

Consider the system

$$s_{k+1} = f(s_k), \quad f(0) = 0.$$

- **Stable (Lyapunov stable):** For every $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\|s_0\| < \delta \implies \|s_k\| < \varepsilon \quad \forall k \geq 0.$$

- **Asymptotically stable:** The equilibrium is stable, and in addition

$$\|s_0\| < \delta \implies \lim_{k \rightarrow \infty} s_k = 0.$$

- **Exponentially stable:** There exist $c > 0$, $\lambda \in (0, 1)$ such that

$$\|s_k\| \leq c\lambda^k \|s_0\|, \quad \forall k \geq 0.$$

Lyapunov Functions

Consider the system

$$s_{k+1} = f(s_k), \quad f(0) = 0.$$

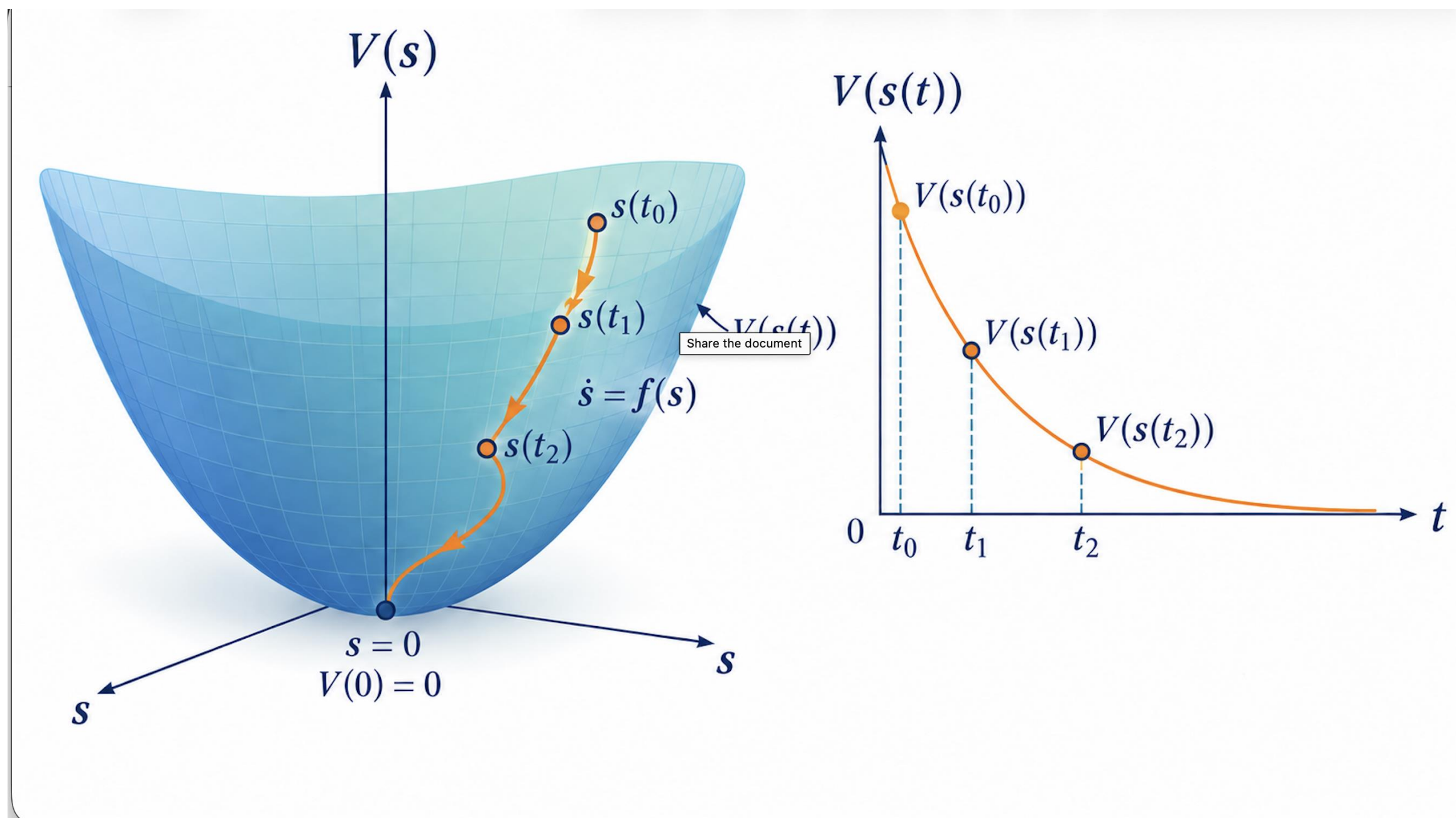
A function $V : \mathbb{R}^n \rightarrow \mathbb{R}$ is called a **Lyapunov function** if:

- $V(s) > 0$ for all $s \neq 0$, and $V(0) = 0$.
- With

$$\Delta V(s) = V(f(s)) - V(s),$$

we have:

- $\Delta V(s) \leq 0 \implies$ the equilibrium is Lyapunov stable.
- $\Delta V(s) < 0 \implies$ the equilibrium is asymptotically stable.



Systems with Noise

Consider the stochastic system

$$s_{k+1} = f(s_k, w_k), \quad f(0, 0) = 0,$$

with $\{w_k\}$ a sequence of random disturbances, independent conditioned on the past.

A function $V : \mathbb{R}^n \rightarrow \mathbb{R}_+$ is a **stochastic Lyapunov function** if:

- $V(s) > 0$ for all $s \neq 0$, and $V(0) = 0$.
- The conditional drift

$$\Delta V(s) = \mathbb{E}[V(f(s, w)) \mid s_k = s] - V(s)$$

satisfies:

- $\Delta V(s) \leq 0 \implies$ stochastic stability in probability,
- $\Delta V(s) < 0 \implies$ asymptotic stability in expectation.

Linear Systems with Noise

For the linear system $s_{k+1} = As_k + w_k$, with zero-mean noise, a quadratic candidate $V(s) = s^\top Ps$, $P \succ 0$, yields

$$\mathbb{E}[V(s_{k+1}) \mid s_k] - V(s_k) = s_k^\top (A^\top PA - P)s_k + \text{tr}(PW).$$

If there exists $P \succ 0$ such that $A^\top PA - P \prec 0$, the system is mean-square stable.

Questions