
Lecture notes on Control Systems and Reinforcement Learning

Written by
Manuel Hinz

mh@mssh.dev or s6mlhinz@uni-bonn.de

Lecturer

Prof. Dr. Jochen Garcke
[garcke\[at\]math.uni-bonn.de](mailto:garcke[at]math.uni-bonn.de)



University of Bonn
Summer semester 2025
Last update: June 17, 2025

Contents

Chapter 0:

Manuel's notes

Warning

These are unofficial lecture notes written by a student. They are messy, will almost surely contain errors, typos and misunderstandings and may not be kept up to date! I do however try my best and use these notes to prepare for my exams. Feel free to email me any corrections to mh@mssh.dev or s6mlhinz@uni-bonn.de.
Happy learning!

Many thanks to Vincent for his feedback and some corrections!

General Information

- Basis: [Basis](#)
- Website: <https://ins.uni-bonn.de/teachings/ss-2025-467-v5e1-advanced-topics/>
- Time slot(s): **Tuesday: 14-16** SR 2.035 and **Thursdays: 16-18** SR 2.035
- Exams: ?
- Deadlines: No exercise sheets / tutorials

0.1 Organization

- Focused on ingredients, won't get to the current state of the art
- Some algorithmic / numerical background (Euler method is fine)
- Control Problems (Steering the bike / car)

Start of lecture 01
(10.4.2025)

The main source for this course is [\[2\]](#). We will follow this somewhat closely, especially in the first part of the course!

Chapter 1:

Introduction to optimal control

1. u is the control (input / action)
2. y observations (outputs)
3. $\phi : Y \rightarrow U$ policy
4. ff feed forward control (plan we had)

Interactions with the outside world might be hidden in the observations. Typically ff is in regard to some reference state. There might be some disturbances (holes in the road, ...). The overall aim is to find a policy ϕ that sticks close to $r(k), k \geq 0$.

$$u(k) = u_{\text{ff}}(k) + U_{\text{fb}}(k)$$

where u_{ff} is the planing to reach the overall goal and u_{fb} actual steering, updated "all the time". Some examples from the book:

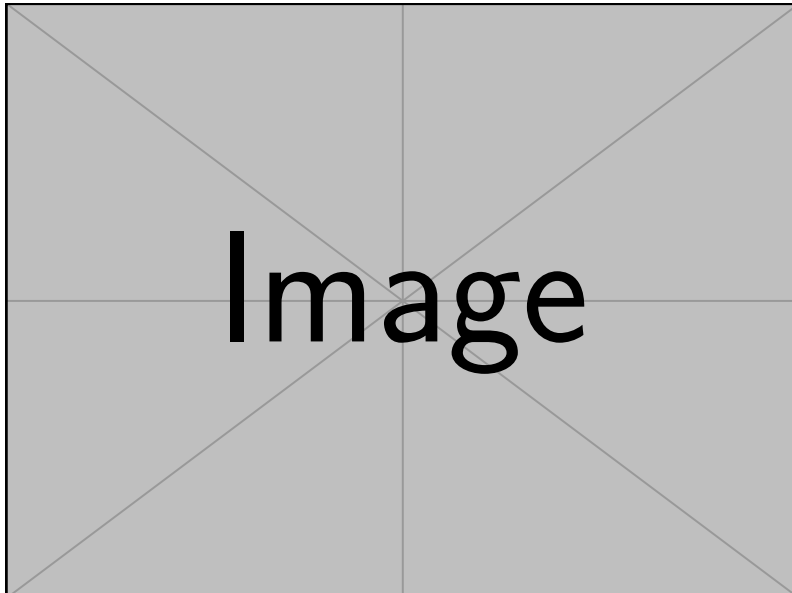


Figure 1.1: Sketch 1.01

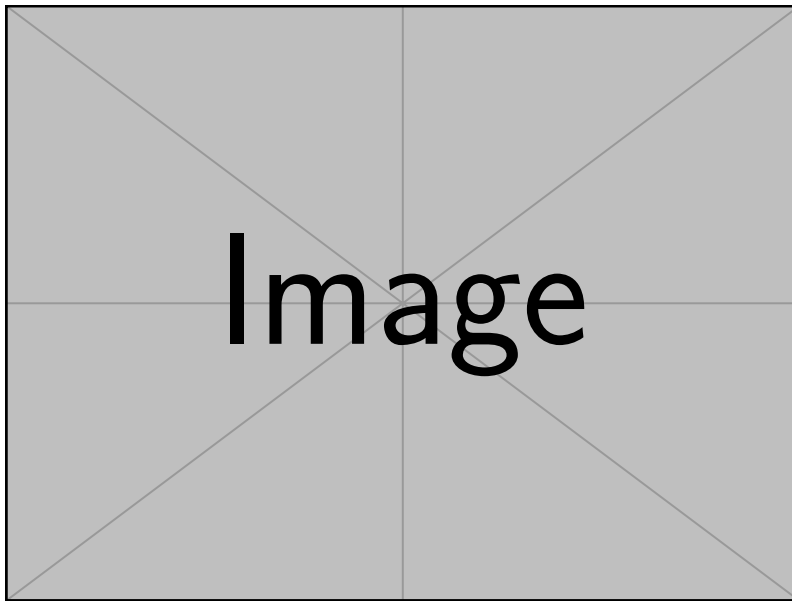


Figure 1.2: Sketch 1.02: Mountain car

Difference: In Reinforcement learning, we don't start with a model / ode.
Some part of reinforcement learning works model-free (i.e. assumes the model only implicitly)

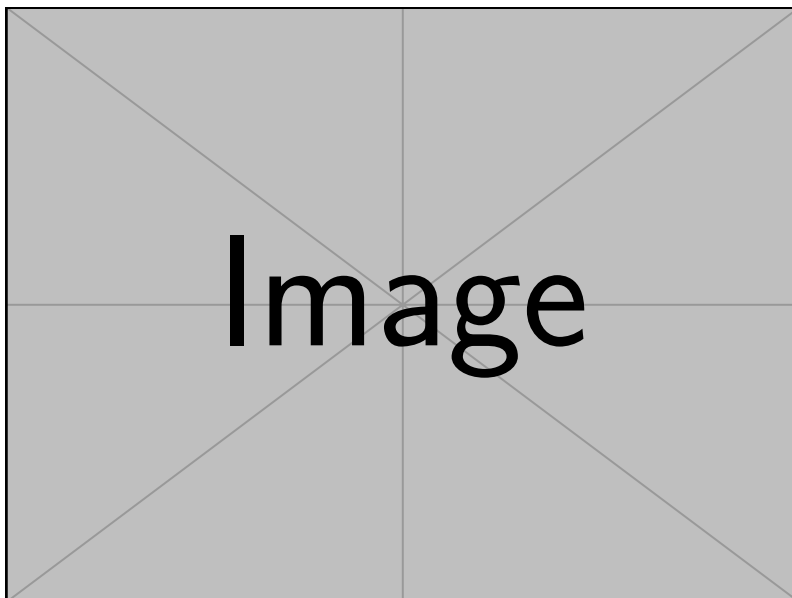


Figure 1.3: Sketch 1.03: cart pole / inverted pendulum

Next example: Acrobot (more than one equilibrium)

1.1 State Space Models

We have some

- state space $X, x \in X$
- action space $U, u \in U$
- action at step $k : u(k) \in U(k)$, i.e. we might have some constraints

- observation space $Y, y \in Y$

Definition 1. Given state, action and observation spaces X, U, Y , a state space model is defined by

$$x(k+1) = \mathcal{F}(x(k), u(k)) \quad (1)$$

$$y(k) = \mathcal{C}(x(k), u(k)) \quad (2)$$

Remark. Overcomplicating problems by loading lots of information into the state space, might make the problem harder!

1.2 Linear State Space Model

$$x(k+1) = Fx(k) + Gu(k) \quad (3)$$

$$y(k) = Cx(k) + Du(k) \quad (4)$$

Remark. The representations (in terms of the matrices) might not be unique!

Common scenario for (??) is to keep $x(k)$ near the origin. You have to think about robustness of the system. Disturbances should be handled by the system.

$$u(k) = -Kx(k).$$

Consider a disturbance under the same control:

$$u(k) = -Kx(k) + v(k)$$

inserting this into (??) yields

$$x(k+1) = (F - GK)x(k) - Gv(k)$$

$$y(k) = (C - DK)x(k) + Dv(k)$$

Closed vs open loop: In closed loops we don't change our course based on observations, while in open loop systems we do.

1.3 State Space Models in continuous Time

$$\frac{d}{dt}x = f(x, u)$$

for $x \in \mathbb{R}^n, u \in \mathbb{R}^m$. We often write u_t, x_t for u, x at time t . If f is linear we get

$$\frac{d}{dt}x = Ax + Bu$$

$$y = Cx + Du$$

$x(k)$ mi
past, m
the stoc



Figure 1.4: Sketch 1.04

To discretize we use the forward Euler method. Given time interval Δ

$$x(k+1) = x(k) + \Delta f(x(k), u(k))$$

so in (??) $\mathcal{F}(x, u) = x + \Delta f(x, u)$. Using Taylor

$$x_{t+\Delta} = x_t \Delta f(x, u) + O(\Delta^2)$$

For the linear model we get $F = I + \Delta A$

$$x(k+1) = x(k) + \Delta A x(k) + \underbrace{\Delta B}_{=:G} u(k)$$

For now fix some policy ϕ , so $u(k) = \phi(x(k))$:

$$x(k+1) = \mathcal{F}(x(k))$$

Assumption 2. The state space X is equal to \mathbb{R}^n or a closed subset of \mathbb{R}^n .

Definition 3. An equilibrium x^e is a state at which is system is frozen:

$$x^e = \mathcal{F}(x^e).$$

Definition 4. Given a cost function $C : X \rightarrow \mathbb{R}_+$ and a policy ϕ we define

$$J_\phi(x) = J(x) = \sum_{k=0}^{\infty} C(x(k)), \quad x(0) = x$$

This is called total cost or value function of the policy ϕ .

Given x^e , we usually assume $C(x^e) = 0$. Generally, we consider a discount factor γ^k in front of $C(x(k))$.

Definition 5. Denote by $\mathcal{X}(k; x_0)$ the state step k with initial condition x_0 and following fixed policy ϕ . The equilibrium x^e is stable in the sense of Lyapunov if for all $\epsilon > 0 \exists \delta > 0$ s.t. $\|x_0 - x^e\| < \delta$, then

$$\|\mathcal{X}(k; x_0) - \mathcal{X}(k; x^e)\| < \epsilon \forall k \geq 0$$

The same
differen
RL und
reward

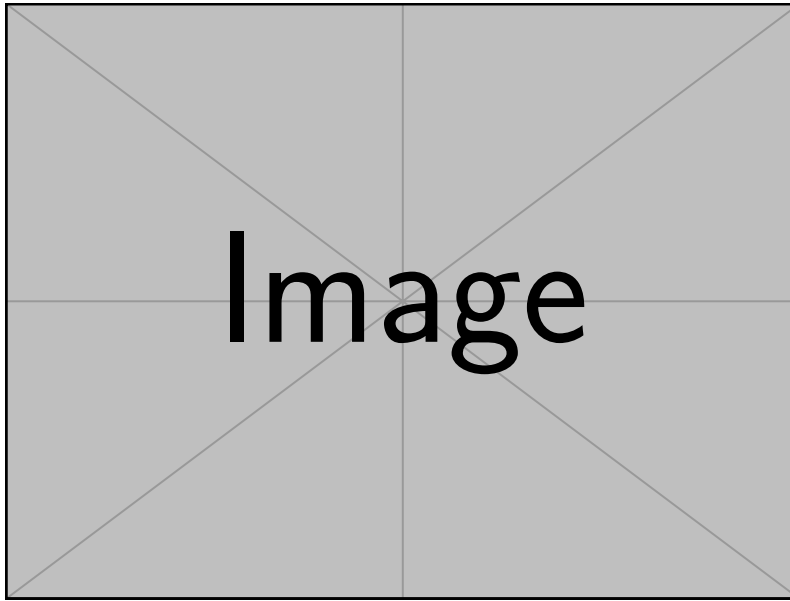


Figure 1.5: Sketch about Lyapunov stability

Definition 6. An equilibrium is said to be asymptotically stable if x^e is stable in the sense of Lyapunov and for some $\delta_0 > 0$, whenever $\|x_0 - x^e\| < \delta_0$, it follows

$$\lim_{k \rightarrow \infty} \mathcal{X}(k, x_0) = x^e.$$

The set of x_0 for which this holds is the region of attraction for x^e . An equilibrium is globally asymptotically stable if the region of attraction is X .

Definition 7 (Lyapunov function). A function $V : X \rightarrow \mathbb{R}_+$ is called Lyapunov function. We frequently assume V is inf-compact, i.e.: it holds

$$\forall x^0 \in X : \{x \in X \mid V(x) \leq V(x^0)\} \text{ is a bounded set.}$$

Remark. There is some variability in the definition of Lyapunov functions! We often assume $V(x)$ is large if x is large.

Sublevel sets:

$$S_V(r) = \{x \in X \mid V(x) \leq r\}.$$

One can see with V being inf-compact $S_V(r)$ is either

- empty
- the whole domain X
- a bounded subset of X .

Usually, $S_V(r) = X$ is impossible, a common assumption is coersiveness:

$$\lim_{\|x\| \rightarrow \infty} V(x) = \infty.$$

Example. • $V(x) = x^2$, coercive

- $V(x) = \frac{x^2}{(1+x)^2}$, not coercive, but inf-compact $r > 1 : S_V(r) = \mathbb{R}$, $r < 1 : S_V(r) = [-a, a]$
with $a = \sqrt{\frac{r}{1+r}}$
- $V(x) = e^x$ is neither

Start of lecture 02
(15.04.2025)

Lemma 8. Suppose that the cost function C and the value function J from definition ?? are non-negative and finite valued.

1. $J(x(k))$ is non-increasing in k and $\lim_{k \rightarrow \infty} J(x(k)) = 0$ for each initial condition.
2. In addition let J be continuous, inf-compact and vanishing only at x^e . Then for each initial condition

$$\lim_{k \rightarrow \infty} x(k) = x^e$$

Proof. Consider $J(x) = \sum_{k=0}^{\infty} c(x(k))$, then

$$\begin{aligned} J(x) &= c(x) + \sum_{k=1}^{\infty} c(x(k)) \\ &= c(x) + \sum_{k=0}^{\infty} c(x^+(k)); \quad x^+(0) = \mathcal{F}(x) \\ &= c(x) + J(\mathcal{F}(x)) \end{aligned}$$

This is the dynamic programming principle for a fixed policy. It is also called Bellmann equation. For 1. from this it follows

$$J(x(k+1)) + c(x) - J(x(k)) = 0$$

summing up from $k = 0$ up to $N - 1$

$$\begin{aligned} J(x) &= J(x(N)) + \sum_{k=0}^{N-1} c(x(k)) \\ &\implies \text{non-increasing} \end{aligned}$$

Taking the limit

$$= \lim_{N \rightarrow \infty} \left[J(x(N)) + \sum_{k=0}^{N-1} c(x(k)) \right] = \left[\lim_{N \rightarrow \infty} J(x(N)) \right] + J(x)$$

using $J(x)$ is finite gives (i).

For 2. with $r = J(x)$, we get $x(k) \in S_J(r) \forall k$. Now suppose $\{x(k_i)\}$ is a convergent subsequence of the trajectory with limit x^∞ . Then $J(x^\infty) = \lim_{i \rightarrow \infty} J(x(k_i)) = 0$ by the continuity of J . We assumed $J(x) = 0 \iff x^e = x \implies x^\infty = x^e$. Finally, the assumption follows, since each convergent subsequence reach the same value x^e .

□

Definition 9 (Poisson's inequality). Let $V, c : X \rightarrow \mathbb{R}_+$ and $\eta \geq 0$. Then Poisson's inequality states that

$$V(\mathcal{F}(x)) \leq V(x) - c(x) + \eta.$$

Proposition 10. Suppose the Poisson inequality holds with $\eta = 0$. Additionally V shall be continuous, inf-compact and it shall have a unique minima at x^e . Then x^e is stable in the sense of Lyapunov (sitsoL).

Proof.

$$\bigcap \{S_V(r) \mid r > V(x^e)\} = \{S_V(r)|_{r=V(x^e)}\}^{\text{unique minimizer}} = \{x^e\}.$$

Using compactness we get: For each $\epsilon > 0$, we can find some $r > V(x^e)$ and some $\delta < \epsilon$ s.t.

$$\{x \in X \mid \|x - x^e\| < \delta\} \subset S_V(r) \subset \{x \in X \mid \|x - x^e\| < \epsilon\}$$

If $\|x_0 - x^e\| < \delta$, then $x_0 \in S_V(r)$ and hence $x(k) \in S_V(r)$ since $V(x(k))$ is non-increasing. With the second inclusion we see

$$\|x(k) - x^e\| < \epsilon \forall k$$

This gives sitsoL.

□

Proposition 11 (Comparison theorem). *Poisson's inequality implies*

1. For each $N \geq 1$ and $x = x(0)$

$$V(x(N)) + \sum_{k=0}^{N-1} c(x(k)) \leq V(x) + N\eta$$

2. If $\eta = 0$, then $J(x) \leq V(x) \forall x$

3. Assume $\eta = 0$ and V, c are continuous. Suppose that c is inf-compact and vanishes only at the equilibrium x^e . Then x^e is globally asymptotically stable.

Proof. 1.

$$V(x(k+1)) - V(x(k)) + c(x(k)) \leq \eta$$

summing up from 0 to $N-1$:

$$V(x(N)) - V(x(0)) + \sum_{k=0}^{N-1} c(x(k)) \leq N\eta$$

2. for $\eta = 0$ the above is ≤ 0 , so $\sum_{k=0}^{N-1} c(x(k)) \leq V(x(0)) - V(x(N)) \leq V(x(0))$ where the LHS converges to $J(x(0))$ for $N \rightarrow \infty$

3. Show sitsoL, with $\eta = 0$ it follows from definition ?? that $V(x) \geq c(x)$, which gives V is also inf-compact. c is vanishing only at x^e , so $V(x(k))$ is strictly decreasing. When $x(k) \neq x^e$, implies $V(x(k)) \downarrow V(x^e)$ for each $x(0)$. Further

$$V(x^e) < V(x(0)) \quad \forall x(0) \in X \setminus \{x^e\}.$$

So it is a unique minimum. V has therefore the properties of proposition ??, which gives sitsoL. For global: with 1. we get

$$\lim_{k \rightarrow \infty} c(x(k)) = 0$$

and assumptions give us by lemma ?? that $x(k) \rightarrow x^e$ as $k \rightarrow \infty$. So, we converge from any initial condition, which gives global asymptotical stability. □

Proposition 12. *Suppose that $V(\mathcal{F}(x)) = V(x) - c(x)$. Further, we assume that*

1. J is continuous, inf-compact, vanishing only at x^e

2. V is continuous

Then $J(x) = V(x) - V(x^e)$.

Proof. As before we sum up:

$$V(x(N)) + \underbrace{\sum_{k=0}^{N-1} c(x(k))}_{J(x(N-1)) \xrightarrow{N \rightarrow \infty} J(x)} = V(x).$$

Lemma ?? together with the continuity of V implies that

$$V(x(N)) \rightarrow V(x^e) \quad \text{as } N \rightarrow \infty.$$

This gives

$$V(x^e) + J(x) = V(x) \quad \square$$

Example (Linear state space model). Setting $x(k+1) = \mathcal{F}(x(k))$, now with linear dynamics:

$$x(k+1) = Fx(k) = F^{k+1}x(0) = F^{k+1}x.$$

Assume quadratic cost $c(x) = x^\top Sx$, where S is symmetric and positive definite. Observe

$$c(x(k)) = (F^k x)^\top S F^k x$$

Summing up yields

$$J(x) = x^\top \underbrace{\left[\sum_{k=0}^{\infty} (F^k)^\top S F^k \right]}_{=:M} x$$

This satisfies a linear fixed point equation:

$$M = S + F^\top M F \quad (5)$$

One can show for the linear state space model, that the following are equivalent:

1. the origin is asymptotically stable
2. the origin is globally asymptotically stable
3. Each eigenvalue λ of F satisfies $|\lambda| < 1$
4. (??) admits a solution M positive semi-definite for any S positive semidefinite.

Reference: [1]

Consider ?? without y

$$y(k+1) = \mathcal{F}(x(k), u(k))$$

with

$$c : X \times U \rightarrow \mathbb{R}_+.$$

The total cost J_ϕ for a given ϕ given $u(k) = \phi(x(k))$ is

$$J_\phi(x) = \sum_{k=0}^{\infty} c(x(k), u(k)).$$

The optimal value function is the minimum over all controls

$$J^*(x) = \min_{\underline{u}=[u(0), u(1), \dots]} \sum_{k=0}^{\infty} c(x(k), u(k)), \quad x(0) = x \in X \quad (6)$$

Remark. The minimizer might not be unique! In harder settings this might need to be an *inf*!

Goal: Find a control sequence that achieves the minimum.

Computationally we can't expect to calculate J_ϕ exactly, but we will approximate it.

Remark. We are in the infinite horizon setting (infinite time steps) to talk about the stability. For this it is important that the equilibrium has cost 0. Without an equilibrium we can also think about discounted value functions

$$J_\phi = \sum_{k=0}^{\infty} \gamma^k c(x(k), u(k))$$

We will see later that it holds for the sequence x^* achieving the minimum

$$J^*(x^*(k)) = c(x^*(k), u^*(k)) + J^*(x^*(k+1))$$

which is definition ?? with $\eta = 0$ and equality.

Proposition ?? implies, under some conditions, that x^e is globally asymptotically stable.

Under the following assumptions J^* is finite:

1. there is a (target) state x^e that is an equilibria for some control $F(x^e, u^e) = x^e$
2. $c \geq 0, c(x^e, u^e) = 0$
3. for any initial condition $x(0) = x$ there is a control sequence \underline{u} and a time T , such that $x(T) = x^e$ for $x(0) = x$ using control \underline{u} .

Example (Linear Quadratic Regulator). Consider linear dynamics ?? from the first lecture with quadratic cost $c(x, u) = x^\top Sx + u^\top Ru$ with S positive semi-definite and R positive definite.

Reminder: $u = -Kx$.

If there is a policy for which J^* is finite, then

$$J^*(x) = x^\top M^* x$$

with M^* positive semi-definite and

$$\phi^*(x) = -K^*(x)$$

with K^* depends on M^*, R, F, G .

Bellmann equation

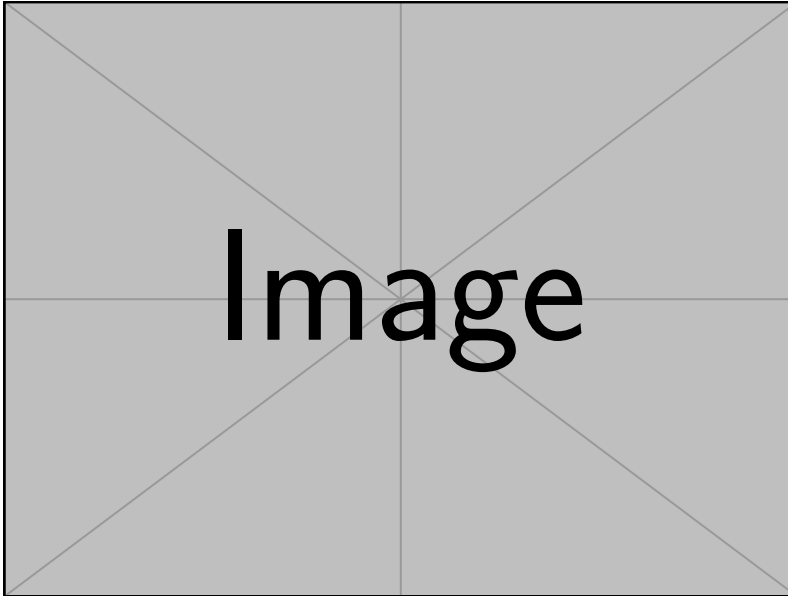


Figure 1.6: Sketch 1.06; Principle of optimality

Observation:

$$\begin{aligned} J^*(x) &= \min_{\underline{u}} \left[\sum_{k=0}^{k_m-1} c(x(k), u(k)) + \sum_{k_m}^{\infty} c(x(k), u(k)) \right] \\ &= \min_{u[0, \dots, k_m-1]} \left[\sum_{k=0}^{k_m-1} c(x(k), u(k)) + \underbrace{\min_{u[k_m, \dots, \infty]} \sum_{k_m}^{\infty} c(x(k), u(k))}_{=J^*(x(k_m))} \right] \end{aligned}$$

This gives

$$J^*(x) = \min_{u[0, \dots, k_m-1]} \left[\sum_{k=0}^{k_m-1} c(x(k), u(k)) \right] + J^*(x(k_m)).$$

With $k_m = 1$ we have shown the following theorem

Theorem 13 (Bellmann equation, Dynamic Programming equation). Assume that J^* is finite and optimal control u^* solving (??) exists. Then the value function satisfies

$$J^*(x) = \min_u \{c(x, u) + J^*(\mathcal{F}(x, u))\} \quad (7)$$

Suppose the minimum is unique for each x and let $\phi^*(x)$ denote the minimum feedback law at x . Then the optimal control is expressed as

$$u^*(k) = \phi^*(x^*(k)).$$

Definition 14 (Q-function). The function of two variables within the minimum in (??) is called Q-function.

$$Q^*(x, u) = c(x, u) + J^*(\mathcal{F}(x, u))$$

In the optimal case we write Q^* . Thus

$$J^*(x) = \min_u Q^*(x, \bar{u}).$$

The optimal feedback law is then

$$\phi^*(x) \in \operatorname{argmin}_u Q^*(x, u).$$

The Q-function solves the fixed point equation

$$Q^*(x, u) = c(x, u) + \min_u Q^*(\mathcal{F}(x, u), u).$$

This already gives a hint for an algorithm coming later next lecture.

Remark. In RL the difference is that we don't know the model, we only observe state action pairs. This motivates the Q-function.

Some concepts from Reinforcement Learning

Actors and critic:

Given is a parameterized family of policies $\{\phi^\theta \mid \theta \in \mathbb{R}^d\}$. the actors. For each θ , observe the trajectories by their states x and actions u determined by their policy.

The critic approximates the associated value function \tilde{J}_θ . Aim for the minimum

$$\theta^* = \operatorname{argmin}_\theta \langle v, \tilde{J}_\theta \rangle,$$

where the weight vector $v \geq 0$ reflects the weighting of the states. $v(x)$ is large for *important* states.

scalar product in \mathbb{R}^n (states?)

Temporal differences:

$$J_\theta(x(k)) = c(x(k), u(k \mid \theta)) + J_\theta(x(k+1))$$

Look for an approximation \hat{J} for which the error is small (w.r.t. the equality above).

Temporal differences are

$$D_{k+1}(\hat{J}) := -\hat{J}(x(k)) + \hat{J}(x(k+1)) + c(x(k), u(k)).$$

After N samples

$$\Gamma(\hat{J}) := \frac{1}{N} \sum_{k=0}^{N-1} D_{k+1}(\hat{J})^2.$$

We can optimize / minimize this.

There is a whole class of TD algorithms and those fit into the actors critic approach!

Start of lecture 04
(22.04.2025)

1.4 Value iteration

We approximate J^* by a sequence of V^k given an initial value function V^0 .

$$V^{k+1}(x) = \min_u \{c(x, u) + V^k(\mathcal{F}(x, u))\}, x \in X, k \geq 0$$

This is called **value iteration** often shortened to VI.

Algorithm 1 Value iteration

Input: Start with an initial value function V^0

Output: Estimates V^{k+1}

$n = 0$

while not good enough **do**

Value function improvement to obtain next value function

$$V^{k+1}(x) = \min_u \{c(x, u) + V^k(\mathcal{F}(x, u))\}, x \in X, k \geq 0$$

end while

Proposition 15. Let V^0 be chosen with non-negative entries and $V^0(x^e) = 0$. Further, we assume

1. X, U are finite sets
2. c is non-negative and vanishes only at (x^e, u^e) , and J^* is finite valued.

Then there is $n_0 \geq 1$ such that

$$V^k(x) = J^*(x), x \in X, k \geq n_0.$$

Proof. Let $\phi^*(x)$ be an optimal policy, and let $n_0 \geq 1$ denote the value such that

$$(x^*(k), u^*(k)) = (x^e, u^e)$$

for $k \geq n_0$. This exists since J^* is finite.

Using the principle of optimality (??) we can show

$$V^n(x) = \min_{u[0, \dots, n-1]} \left\{ \sum_{k=0}^{n-1} c(x(k), u(k)) + V^0(x(n)) \right\}, x(0) \in X \quad (8)$$

This gives

$$V^n(x) \leq \sum_{k=0}^{n-1} c(x(k), u(k)) + V^0(x(n)) \text{ for all } u \text{ including } u(k) = \phi^*(k)$$

$$\stackrel{n \geq n_0}{=} J^*(x) + V^0(x^e) = J^*(x)$$

For such n , the inequality must be an equality, due to (??) and the use of the optimal policy. \square

VI provides a sequence of policies ϕ^n

$$\phi^n(x) \in \operatorname{argmin}_u \{c(x, u) + V^n(\mathcal{F}(x, u))\}.$$

If we assume that V^0 is non-negative and satisfies poisson's inequality(??) for some $\eta \geq 0$

$$V^0(\mathcal{F}(x, u)) \leq V^0(x) - c(x, \phi^0(x)) + \eta, x \in X$$

then we get the following statement

We really exploit finiteness!

Proposition 16. Suppose that V^0 is non-negative and it holds

$$\begin{aligned} \min_u (c(x, u) + V^0(\mathcal{F}(x, u))) &= \{c(x, u) + V^0(\mathcal{F}(x, u))\} |_{u=\phi^0(x)} \\ &\leq V^0(x) + \eta, \quad x \in X \end{aligned}$$

Then a corresponding bound holds for each n

$$\{c(x, u) + V^n(\mathcal{F}(x, u))\} |_{u=\phi^0(x)} \leq V^n(x) + \eta_n, \quad x \in X,$$

where η_i is non-increasing:

$$\eta \geq \eta_0 \geq \eta_1 \dots$$

Proof. Write $B^n(x) = V^{n+1}(x) - V^n(x)$

$$\eta_n := \sup_x B^n(x).$$

Value iteration gives

$$\begin{aligned} \{c(x, u) + V^n(\mathcal{F}(x, u))\} |_{u=\phi^n(x)} &= \min_u \{c(x, u) + V^n(\mathcal{F}(x, u))\} \\ &= V^{n+1}(x) = V^n(x) + B^n(x) \\ &\leq V^n(x) + \eta_n \end{aligned}$$

To show that the η are non-increasing, we consider

$$V^1(x) = \{c(x, u) + V^0(\mathcal{F}(x, u))\} |_{u=\phi^0(x)} \stackrel{\text{Assumption}}{\leq} V^0(x) + \eta$$

which gives $B^0(x) \leq \eta \forall x \implies \eta_0 \leq \eta$.

For $n \geq 1$ The trick is using the old control in the second line:

$$\begin{aligned} V^n(x) &= \{c(x, u) + V^{n-1}(\mathcal{F}(x, u))\} |_{u=\phi^{n-1}(x)} \\ V^{n+1}(x) &\leq \{c(x, u) + V^n(\mathcal{F}(x, u))\} |_{u=\phi^{n-1}(x)} \end{aligned}$$

So,

$$V^{n+1}(x) - V^n(x) \leq \{V^n(\mathcal{F}(x, u)) - V^{n-1}(\mathcal{F}(x, u))\} |_{u=\phi^{n-1}(x)} \leq \eta_{n-1}.$$

Hence, $\eta_n = \sup_x B^n(x) \leq \eta_{n-1}$. □

Now consider $\eta = 0$, so for each n

$$\{c(x, u) + V^n(\mathcal{F}(x, u))\} |_{u=\phi^n(x)} \leq V^n(x)$$

with proposition ?? it follows

$$J^* \leq V^n(x), \quad x \in X,$$

where J^* is the total cost using policy ϕ^n .

One view of policy iteration is the focus on updating the policy function!

1.5 Policy iteration

Start with an initial policy $\phi^0, n = 0$

- Compute the total cost for the policy ϕ^n , this is called policy evaluation

$$J^n(x) = \sum_{k=0}^{\infty} c(x(k), u(k)), \quad u(k) = \phi^n(x(k)) \forall x \in X$$

- perform **policy improvement** to obtain the next policy

$$\phi^{n+1}(x) \in \underset{u}{\operatorname{argmin}} \{c(x, u) + J^n(\mathcal{F}(x, u))\}, \quad x \in X$$

- while *not good enough*

This is sometimes also called Howard's algorithm.

Remark. The first step is some linearization and the second is the update. Like a generalization of Newton's method

Algorithm 2 Policy iteration

Input: Start with an initial policy ϕ^0

Output: Estimates $J^n(x), \phi^{n+1}(x)$

$n = 0$

while not good enough **do**

 Compute the total cost for the policy ϕ^n , this is called policy evaluation

$$J^n(x) = \sum_{k=0}^{\infty} c(x(k), u(k)), \quad u(k) = \phi^n(x(k)) \quad \forall x \in X$$

 perform **policy improvement** to obtain the next policy

$$\phi^{n+1}(x) \in \underset{u}{\operatorname{argmin}} \{c(x, u) + J^n(\mathcal{F}(x, u))\}, \quad x \in X$$

end while

Proposition 17. Suppose that J^0 for ϕ^0 is finite valued. Then for each $n \geq 0$

$$\{c(x, u) + J^n(\mathcal{F}(x, u))\}_{|_{u=\phi^{n+1}(x)}} \leq J^n(x), \quad x \in X$$

and consequently, the value functions are non-increasing

$$J^0(x) \geq J^1(x) \geq \dots$$

Proof. Similar to the proof of proposition ??, where the non-increasing sequence again follows from proposition ??. □

Here we always assumed that we can compute everything, especially \mathcal{F} and the infinite sum.

1.6 Exploration

In RL we learn from observations, each state-action pair, new state and observed cost gives us information. We need *good* and *useful* information.

Consider a policy that is not optimal, but has $x(k) \rightarrow x^e$ reasonably rapidly, where we assume $c(x^e, \cdot) = 0$. Typically we have continuity

$$\begin{aligned} \lim_{k \rightarrow \infty} D_{k+1}(\hat{J}) &= \lim_{k \rightarrow \infty} \left[-\hat{J}(x(k)) + \hat{J}(x(k+1)) + c(x(k), u(k)) \right] \\ &= -\hat{J}(x^e) + \hat{J}(x^e) + 0 = 0. \end{aligned}$$

This is not much information, one cannot further improve the policy!

$$\Gamma^\epsilon(\hat{J}, x^i) = \frac{1}{N_\epsilon} \sum_{k=0}^{N_\epsilon-1} [D_{k+1}(\hat{J})]^2, \quad x(0) = x^i$$

To avoid getting *small* information from long trajectories, one can take a couple of shorter ones.

$$\hat{\Gamma}(\hat{J}) = \frac{1}{M} \sum_{i=1}^M \Gamma^\epsilon(\hat{J}; x^i)$$

How to choose x^i is current research. Much of the theoretical research assume that “every state is assumed regularly”, which is nice for results, but not so nice realistic in most applications.

Another way to get more diverse information is to use [exploration](#). Namely one modifies the trajectories, not strictly follows ϕ^n .

$u(k) = \hat{\phi}(x(k), \zeta(k))$, where $\zeta(k)$ is some form of noise. Typically

1. $\hat{\phi}(x(k), \zeta(k)) = \phi^\theta(k)$ for *most* k
2. Choose action to explore the state-action space (e.g. randomly) the other times

this is also sometimes called off-policy and on-policy

Generally, the trajectory to gather information stems from a different policy than the current estimate ϕ^θ . This dilemma is called the [exploration-exploitation](#) dilemma.

1.7 Linear Quadratic Regulator, Revisited

We had $J^*(x) = x^\top M^* x$ and quadratic costs, $c(x, u) = x^\top S x + u^\top R u$.

For the Q -function:

$$Q^*(x, u) = c(x, u) + J^*(F x + G u).$$

An optimal policy ϕ is a minimum over Q w.r.t. u :

$$0 = \nabla_u Q^*(x, u^*) = 2R u^* + 2G^\top M^*(F x + G u^*)$$

Assuming R is positive definite; then $R + G^\top M^* G$ is positive definite and therefore invertible.

$$K^* = [R + G^\top M^* G]^{-1} G^\top M^* F$$

and

$$\phi^*(x) = -K x.$$

To obtain M^* we can solve a fixed point equation called the [algebraic Riccati equation](#)

$$M^* = F^\top \left(M^* - M^* G [R + G^\top M^* G]^{-1} G^\top M^* F + S \right) \quad (9)$$

1.8 Approximate Q -functions

Consider a family of Q -functions $\{Q^\theta \mid \theta \in \mathbb{R}^d\}$ to approximate Q^* . Classically used is a linear parametrization

$$Q^\theta(x, u) = \theta^\top \psi(x, u), \quad \theta \in \mathbb{R}^d$$

where $\psi_i : X \times U \rightarrow \mathbb{R}$, $1 \leq i \leq d$ is some set of basis functions. Given Q^θ we have

$\phi^\theta(x) \in \operatorname{argmin}_u Q^\theta(x, u)$, $x \in X$.

Policy iteration for Q -functions:

1. obtain θ^n to get an approximation of Q^{θ^n} where
 $Q^{\theta^n}(x, u) = c(x, u) + Q^{\theta^n}(x^+, u^+)$, $x^+ = \mathcal{F}(x, u)$, $u^+ = \phi^n(x^+)$
2. define new policy $\phi^{n+1}(x) := \phi^{\theta^n}$

As an alternative, consider dynamic programming equation from definition ??:

$$Q^*(x, u) = c(x, u) + \min_{\bar{u}} Q^*(\mathcal{F}(x, u), \bar{u}).$$

We follow a given/ observed state-action trajectory $(x(k), u(k))_{k=0}^N$

$$Q^*(x(k), u(k)) = c(x(k), u(k)) + Q^*(x(k+1), u(k+1))$$

The temporal difference / Bellmann error

$$D_{k+1}(Q^\theta) = -Q^\theta(x(k), u(k)) + c(x(k), u(k)) + Q^\theta(x(k+1), u(k+1))$$

If $Q^\theta = Q^*$ then $D_{k+1}(Q^\theta) = 0 \forall k$. In Q -learning algorithms, one chooses θ^n such that $D_{k+1}(Q^{\theta^n})$ is small in a suitable fashion. So we minimize θ to achieve this, i.e.

$$\Gamma^\epsilon(\theta) = \frac{1}{N} \sum_{i=0}^{N-1} [D_{k+1}(Q^\theta)]^2$$

1.9 Bandits

Theory of multi-armed bandits. One has to accept some loss through exploration in order to achieve(find) the best strategy. One exploits the learned strategy when choosing an action according to it.

In the control of dynamic systems one has for each state x (or $x(k)$) a multi-armed bandit.

1.10 Other control formulations

Discounted cost:

$$J^*(x) = \min_{\underline{u}} \sum_{k=0}^{\infty} \gamma^k c(x(k), u(k)), \quad x(0) \in X$$

where $\gamma \in (0, 1)$ is the discount factor.

Shortest Path Problem: Given $A \subset X$ define $\tau_A := \min\{k \geq 1 \mid x(k) \in A\}$.

$$J^*(x) = \min_u \sum_{k=0}^{\tau_A-1} \gamma^k c(x(k), u(k)), \quad x(0) = x.$$

Proposition 18. *If J^* is finite valued, then it is the solution to the dynamic programming equation in the following sense:*

$$J^*(x) = \min_u \{c(x, u) + \gamma 1_{\{\mathcal{F}(x, u) \in A^c\}} J^*(\mathcal{F}(x, u))\}, \quad x \in X$$

where $1_{\{\dots\}}$ denotes an indicator function.

Proof.

$c(x, u(0))$ since
extracting the f
element of the s

$$\begin{aligned} J^*(x) &= \min_{\underline{u}} \left\{ c(x, \underline{u}) + \sum_{k=1}^{\tau_A-1} \gamma^k c(x(k), u(k)) \right\} \\ \tau_A=1 &\implies \sum=0 \quad \min_{u(0)} \left\{ c(x, u(0)) + \gamma 1_{\{x(1) \in A^c\}} + \min_{u[1, \dots]} \left\{ \sum_{k=1}^{\tau_A-1} \gamma^{k-1} c(x(k), u(k)) \right\} \right\} \\ &= \min_{u(0)} \{c(x, u(0)) + \gamma 1_{\{x(1) \in A^c\}} J^*(x(1))\} \end{aligned}$$

where $x(1) = \mathcal{F}(x, u(0))$. □

To formulate this as a discounted problem

1. modify the cost function $c_A(x, u) = \begin{cases} c(x, u) & x \in A^c \\ 0 & x \in A \end{cases}$
2. modify the state dynamics $\mathcal{F}_A(x, u) = \begin{cases} \mathcal{F}(x, u) & x \in A^c \\ x & x \in A \end{cases}$

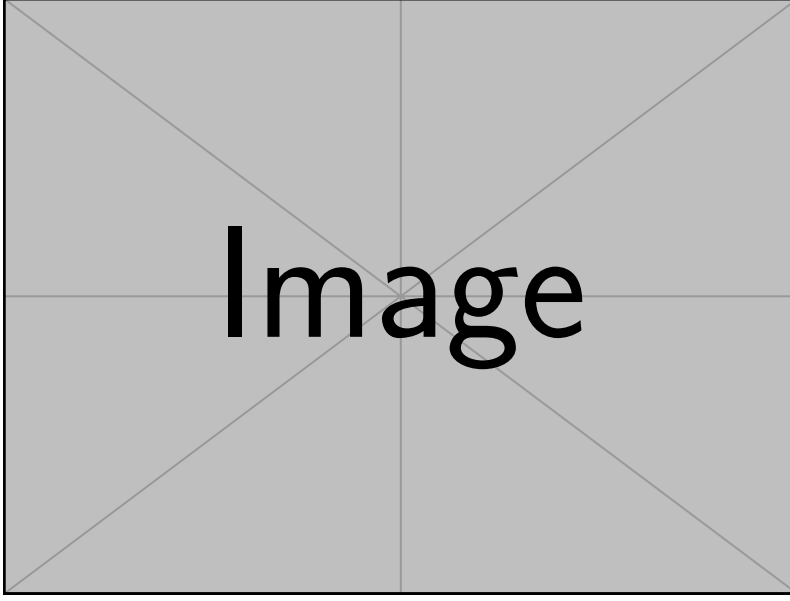


Figure 1.7: Sketch: mountain car value function

Can be numerically very hard, since the value function can be quite discontinuous, but not all value functions are that bad.

Finite Horizon Fix horizon $N \geq 1$ and define

$$J^*(x) = \min_{u[0,N]} \sum_{k=0}^N c(x(k), u(k)), \quad x(0) = x \in X.$$

We can connect to the optimal control problem by

1. enlarging the state space $x^a(k) = (x(k), \tau(k))$, where $\tau(k) = \tau(0) + k$, $k \geq 0$
2. modify the cost function $c^a((x, \tau), u) = \begin{cases} c(x, u) & \tau \leq N \\ 0 & \tau > N \end{cases}$

Then

$$J^*(x^a) = \min_{\underline{u}} \underbrace{\sum_{k=0}^{\infty} c^a(x^a(k), u(k))}_{J^*(x, \tau)}, \quad x^a(0) = (x, 0)$$

The Bellmann equation from theorem ?? now becomes

$$J^*(x, \tau) = \min_u \{c(x, u)1_{\{\tau \leq N\}} + J^*(\mathcal{F}(x, u), \tau + 1)\} \quad (10)$$

For $\tau > N$, it follows that $J^*(x, \tau) = 0$. This gives

$$J^*(x, N) = \min_u c(x, u) = \bar{c}(x).$$

So,

$$J^*(x, N-1) = \min_u \{c(x, u) + \bar{c}(\mathcal{F}(x, u))\}$$

repeating this backwards in time yields

$$J^*(x, 0) = J^*(x^a).$$

For the policy $\phi^*(x, \tau) \in \operatorname{argmin}_u \{c(x, u) + J^*(\mathcal{F}(x, u), \tau + 1)\}, \tau \leq N$ and

$$u^*(k) = \phi^*(x^*(k), k).$$

Model Predictive Control

Here, the policy is computed on-the-fly at each step of the state-action trajectory as a finite horizon problem. The control is

$$u(k) = \phi^{\text{mpc}}(x^*(k)) = \phi^*(x^*(k), 0),$$

where ϕ^* from the finite horizon setting (??) for *small* N .

Consider

$$J^{\text{mpc}}(x) = \sum_{k=0}^{\infty} c(x(k), u(k)), \quad x(0) = x, u(k) = \phi^{\text{mpc}}(x(k)).$$

Due to the fin
we are not op

Proposition 19. Consider $u(k)$ from above with

$$J^*(x; 0) = \min_{u[0, N-1]} \sum_{k=0}^{N-1} c(x(k), u(k)) + V^0(x(N)),$$

where $V^0 : X \rightarrow \mathbb{R}^+$ satisfies the assumption from proposition ?? with $\eta = 0$:

$$\min_u \{c(x, u) + V^0(\mathcal{F}(x, u))\} \leq V^0(x).$$

Then the total cost J^{mpc} is finite everywhere.

Proof. Using an equation from proposition ??:

$$V^N(x) = \min_{u[0, N-1]} \left\{ \sum_{k=0}^{N-1} c(x(k), u(k)) + V^0(x(N)) \right\}$$

and the definition of J^* from above we get $J^*(x, 0) = V^N(x)$ Proposition ?? then gives the bound

$$\{c(x, u) + V(\mathcal{F}(x, u))\}_{|u=\phi^{\text{mpc}}(x)} \leq V(x) = V^N(x)$$

This is also a ve
poisson inequali

From the Comparison theorem ??, it follows that J^{mpc} is finite. □

1.11 Geometry in continuous time

Consider $x(k+1) = \mathcal{F}(x(k))$, now in continuous time:

$$\frac{d}{dt}x_t = f(x_t) \text{ or } \frac{d}{dx}x = f(x)$$

$\mathcal{X}(t, x_0)$ is the solution to the differential equation above. Definition ??, ?? carry over.

$$\lim_{t \rightarrow \infty} \mathcal{X}(t, x_0) = x^e$$

Definition 20. A function $V : X \rightarrow \mathbb{R}_0^+$ is called Lyapunov function for global asymptotic stability if the following conditions hold:

- (i) $V \in C^1$
- (ii) V is *inf-compact*
- (iii) For any solution x , whenever $X_t \neq x^e$

$$\frac{d}{dt}V(x_t) < 0.$$

If $x_t = x^e$, we have $V(x_{t+s}) = V(x^e)$ for all $s \geq 0$, so $\frac{d}{dt}V(x^e) = 0$.

If we look back at the proof of proposition ?? and proposition ?? (iii), we can see that these also carry over to the continuous case. So we get

Proposition 21 (Extension of prop ?? (iii)). *If there exists a Lyapunov function after definition V ??, then the equilibrium x^e is globally asymptotically stable.*

The continuous version of Poisson's inequality is then

$$\langle \nabla V(x), f(x) \rangle \leq -c(x) + \eta \quad (11)$$

using the chain rule we get

$$\frac{d}{dt}V(x) \leq -c(x) + \eta$$

further observing

$$0 \leq V(x_T) = V(x_0) + \int_0^T \frac{d}{dt}V(X_t)dt \leq V(x_0) + T\eta - \int_0^T c(x_t)dt$$

we have shown

Proposition 22 (Continuous Comparison theorem). *If (??) holds for non-negative c, V, η , then we have*

$$V(X_t) + \int_0^T c(x_t)dt \leq V(x) + T\eta, \quad x_0 = x \in X, T > 0 \quad (12)$$

If $\eta = 0$

$$\int_0^\infty c(x_t)dt \leq V(x)$$

the total cost is bounded.

1.12 Optimal control in continuous time

$$\frac{d}{dt}x = f(x, u)$$

with total cost for $\underline{u} = u[0, \infty)$

$$J(\underline{u}) = \int_0^\infty c(x_t, u_t)dt.$$

As before, we minimize over u and want J to be finite. We assume

$$f(x^e, u^e) = 0$$

for some u^e and

$$c(x^e, u^e) = 0$$

which yields that J is finite. As before

$$J^*(x) = \min_u \int_0^\infty c(x_t, u_t)dt, \quad x_0 = x \in X.$$

We extend the Bellmann equation to continuous times

$$\begin{aligned} J^*(x) &= \min_{u[0, \infty]} \left[\int_0^{t_m} c(x_t, u_t)dt + \int_{t_m}^\infty c(x_t, u_t)dt \right] \\ &= \min_{u[0, t_m]} \left[\int_0^{t_m} c(x_t, u_t)dt + \underbrace{\min_{u[t_m, \infty)} \int_{t_m}^\infty c(x_t, u_t)dt}_{J^*(x_{t_m})} \right] \end{aligned}$$

Same principle of optimality: What happens for $t_m \downarrow 0$. We assume $J^* \in C^1$ and write $\Delta x = x_{t_m} - x_0 = x_m - x$. We now use Taylor on the above expression

$$\begin{aligned} J^*(x) &= \min_{u[0, t_m]} \{c(x_t, u_t)t_m + J^*(x) + \nabla J^*(x) \cdot \Delta x + o(t_m)\} \\ \implies 0 &= \min_{u[0, t_m]} \left\{ c(x_t, u_t) \underbrace{\frac{t_m}{t_m}}_{\rightarrow 0} + \nabla J^*(x) \cdot \underbrace{\frac{\Delta x}{t_m}}_{\substack{\frac{d}{dt}|_{t=0} \\ = f(x_0, u_0)}} \right\} + \underbrace{o(1)}_{\rightarrow 0} \\ \implies 0 &= \min_u [c(x, u) + \nabla J^*(x) \cdot f(x_0, u_0)] \end{aligned}$$

Theorem 23. *If the value function J^* has continuous derivatives, then it satisfies the Hamilton-Jacobi-Bellmann equation*

$$0 = \min_u [c(x, u) + \nabla J^*(x) \cdot f(x_0, u_0)] \quad (13)$$

The term to minimize has an interpretation as an Hamiltonian

$$H(x, p, u) = c(x, u) + p^\top f(x, u).$$

One can show

Theorem 24. *Suppose that an optimal state-action pair exists and that $J^* \in C^1$. Then u_t^* must minimize for each t*

$$\min_u H(x_t^*, p_t^*, u) = H(x_t^*, p_t^*, u_t^*)$$

with $p_t^* = \nabla_x J^*(x_t^*)$.

Remark. *Relaxing away from ∇J^* or ∇J can have theoretical and computational advantages.*

1.13 Linear quadratic regulator revisited (once more)

$$\begin{aligned} \frac{d}{dt}x &= Fx + Gu, \quad x(0) = x_0 \\ c(x, u) &= x^\top Sx + u^\top Ru \end{aligned}$$

everything we observed so far carries over, assuming J^* is finite, we have

$$J^*(x) = x^\top M^* x$$

the HSB (??) gives

$$\begin{aligned} \phi^*(x) &= \operatorname{argmin}_u \{x^\top Sx + u^\top Ru + [2M^*x]^\top [Fx + Gu]\} \\ &= \operatorname{argmin}_u \{u^\top Ru + 2x^\top M^*Gu\} \end{aligned}$$

So,

$$0 = \nabla_u \{u^\top Ru + 2x^\top M^*Gu\}|_{u=\phi^*(x)}$$

and we get

$$\phi^*(x) = -R^{-1}G^\top M^*x$$

and

$$\frac{d}{dt}x^* = [F - GR^{-1}G^\top M^*]x^*.$$

HSB (??) further gives

$$\begin{aligned} 0 &= \{x^\top Sx + u^\top Ru + [2M^*x]^\top [Fx + Gu]\}|_{u=\phi^*(x)} \\ &= x^\top \{S + M^*GR^{-1}G^\top M^*\}x + x^\top \{2M^*F + 2M^*GR^{-1}G^\top M^*\}x \end{aligned}$$

using $2x^\top M^* F x = x^\top [M^* F + F^\top M^*]$ we get

$$\begin{aligned} &= x^\top \{S + M^* F + F^\top M^* - M^* G R^{-1} G^\top M^*\} x \\ &\quad \{S + M^* F + F^\top M^* - M^* G R^{-1} G^\top M^*\} \end{aligned}$$

holds for any x and is symmetric, so it follows M^* is a positive definite solution to the algebraic Riccati equation

$$0 = S + M^* F + F^\top M^* - M^* G R^{-1} G^\top M^*$$

Chapter 2:

ODE methods for algorithm design

2.1 ODE methods for algorithm design

Start of lecture 07
(06.05.2025)

Four steps:

- Formulate the algorithmic goal as the root finding problem

$$\bar{f}(\theta^*) = 0$$

- if necessary, refine the design of \bar{f} to ensure that the associated ODE is **globally asymptotically stable**

$$\frac{d}{dt}\vartheta = \bar{f}(\vartheta)$$

- Is an **Euler-approximation** appropriate?

$$\theta_{n+1} = \theta_n + \alpha_{n+1} \bar{f}(\theta_n) \quad (1)$$

- Design an algorithm to approximate (??) based on whatever observation is available.

Remark. The idea is to transfer the global stability from the ODE to the algorithm.

Goal: Construct a vector field f such that ϑ_t converges to the **target** $\theta^* \in \mathbb{R}^d$, where θ^* is an equilibrium

$$f(\theta^*) = 0.$$

In ODE theory one uses so called **Picard-Iteration**

$$\vartheta_t^{n+1} = \theta_0 + \int_0^t f(\vartheta_\tau^n) d\tau, \quad 0 \leq t \leq T \quad (2)$$

based on

$$\vartheta_0 + \int_0^t f(\vartheta_\tau) d\tau, \quad 0 \leq t \leq T. \quad (3)$$

Proposition 25. Suppose that the function f is globally Lipschitz continuous:

$$\exists L > 0 : \forall x, y \in \mathbb{R}^d : \|f(x) - f(y)\| \leq L\|x - y\|$$

Then for each θ_0 there exists a unique solution to (??). in the finite time horizon. Moreover, successive approximation is uniformly convergent:

$$\lim_{n \rightarrow \infty} \max_{0 \leq t \leq T} \|\vartheta_t^n - \vartheta_t\| = 0$$

Proposition 26 (Grönwall-Bellman-inequality). Let α, β and z be non-negative functions defined

on $[0, T]$, $T > 0$. Assume that β, z are continuous and that

$$z_t \leq \alpha_t + \int_0^t \beta_s z_s ds, \quad 0 \leq t \leq T$$

Then it holds

$$(i) \quad z_t \leq \alpha_t + \int_0^t \alpha_s \beta_s \exp\left(\int_s^t B_r dr\right) ds$$

(ii) if in addition the function α is non-decreasing, then

$$z_t \leq \alpha_t \exp\left(\int_0^t B_s ds\right), \quad 0 \leq t \leq T$$

Proof. Both proofs can be found in any textbook on ODEs. The second one is also found in [2]. \square

Proposition 27. Consider $\frac{d}{dt}\vartheta = f(\vartheta)$, $\vartheta_0 = \theta_0$ with f globally Lipschitz. Then

(i) There is a constant B_f depending only on f such that, with $t \geq 0$

$$\|\vartheta_t\| \leq (B_f + \|\vartheta_0\|) e^{Lt} - B_f \quad (4)$$

$$\|\vartheta_t - \vartheta_0\| \leq \|B_f + L\|\vartheta_0\| t e^{Lt} \quad (5)$$

(ii) If there is an equilibrium θ^* , then for each initial condition:

$$\|\vartheta_t - \theta^*\| \leq \|\vartheta_0 - \theta^*\| e^{Lt} \quad (6)$$

Proof. (ii): use ?? to get

$$\vartheta_t - \theta^* = \vartheta_0 - \theta^* + \int_0^t f(\vartheta_\tau) d\tau$$

Since $f(\theta^*) = 0$, we see

$$\begin{aligned} \|f(\vartheta_\tau)\| &= \|f(\vartheta_\tau) - f(\theta^*)\| \\ &\leq L \underbrace{\|\vartheta_\tau - \theta^*\|}_{=: z_\tau} \end{aligned}$$

So

$$z_t \leq z_0 + L \int_0^t z_\tau d\tau.$$

Using proposition ?? (ii) with $\beta_t \equiv L$, $\alpha_t \equiv z_0$ we get

$$\|\vartheta_t - \theta^*\| \leq \|\vartheta_0 - \theta_0\| \exp(Lt)$$

(i): take any $\bar{\theta} \in \mathbb{R}^d$ and use the Lipschitz continuity

$$\begin{aligned} \|f(\theta)\| &\leq \|f(\theta) - f(\bar{\theta})\| + \|f(\bar{\theta})\| \\ &\leq L\|\theta - \bar{\theta}\| + \|f(\bar{\theta})\| \\ &\leq L\|\theta\| + L\|\bar{\theta}\| + \|f(\bar{\theta})\|. \end{aligned}$$

For any fixed $\bar{\theta}$, define $B_f = \|\bar{\theta}\| + \|f(\bar{\theta})\|/L$ which gives

$$\|f(\theta)\| \leq L[\|\theta\| + B_f], \quad \theta \in \mathbb{R}^d$$

using (??)

$$\begin{aligned} \|\vartheta_t\| + B_f &\leq \|\vartheta_0\| + B_f + \underbrace{L}_{\beta} \int_0^t \left[\underbrace{\|\vartheta_\tau + B_f\|}_{z_\tau} \right] d\tau \\ &\leq [\|\vartheta_0\| + B_f] \exp(Lt) \end{aligned}$$

where the last step follows by the same trick as in (ii), i.e. by using Grönwall. \square

2.2 Euler's method once more

$$\frac{\hat{\vartheta}_{t_{n+1}} - \hat{\vartheta}_{t_n}}{\alpha_{n+1}} = f(\hat{\vartheta}_{t_n}), \quad \hat{\vartheta}_0 = \vartheta_0 = \theta_0 \quad (7)$$

or

$$\hat{\vartheta}_{t_{n+1}} = \hat{\vartheta}_{t_n} + \alpha_{n+1} f(\hat{\vartheta}_{t_n})$$

It can be shown for f globally Lipschitz

$$\max_{0 \leq t \leq T} \|\hat{\vartheta}_t - \vartheta_t\| \leq \underbrace{K(L, T)}_{\text{exponential in } L, T} \max\{\alpha_k \mid t_k < T\} \quad (8)$$

Explicit Euler, implicit Euler is nicer to analyze

2.3 Optimization

Goal: Find, for some loss function $\Gamma : \mathbb{R}^d \rightarrow \mathbb{R}_+$,

$$\theta^* \in \operatorname{argmin} \Gamma(\theta). \quad (9)$$

Use steepest-descent, formulated as ODE

$$\frac{d}{dt} \vartheta = -\nabla_{\theta} \Gamma(\theta) \quad (10)$$

so called gradient flow.

$$\nabla \Gamma(\theta_0) \perp \{\theta \in \mathbb{R}^d \mid \Gamma(\theta) = \Gamma(\theta_0)\} =: S_{\Gamma}(\theta_0)$$

The gradient flow steers into the interior of $S_{\Gamma}(\theta_0)$.

Definition 28. (i) A set $S \subset \mathbb{R}^d$ is convex if it contains all line segments with endpoints in S

(ii) A function $\Gamma : S \rightarrow \mathbb{R}$ with S convex, is called convex if for any $\theta^0, \theta^1 \in S$ and $\rho \in (0, 1)$

$$\Gamma((1 - \rho)\theta^0 + \rho\theta^1) \leq (1 - \rho)\Gamma(\theta^0) + \rho\Gamma(\theta^1)$$

Γ is strictly convex if this inequality is strict whenever $\theta^0 \neq \theta^1$

(iii) If Γ is differentiable, then it is called strongly convex if for $\delta_0 > 0$

$$\langle \nabla \Gamma(\theta) - \nabla \Gamma(\theta^0), \theta - \theta^0 \rangle \geq \delta_0 \|\theta - \theta_0\|^2, \quad \forall \theta, \theta^0 \in S$$

From numerical optimization we know:

Theorem 29. Suppose that $\Gamma : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex. Then for given $\theta^0 \in \mathbb{R}^d$

(i) if θ^0 is a local minima, then it is also a global minimum

(ii) if Γ is differentiable at θ^0 , with $\nabla \Gamma(\theta) = 0$, then θ^0 is a global minimum

(iii) if either (i) or (ii) hold, and if Γ is strictly convex, then θ^0 is the unique global minimum

Proposition 30. Suppose that Γ is continuously differentiable, convex and coercive, with unique minimizer θ^* . Then the gradient flow

$$\frac{d}{dt} \vartheta = -\nabla \Gamma(\vartheta)$$

is globally asymptotically stable, with unique equilibrium θ^* .

If Γ is strongly convex, then the rate of convergence is exponential

$$\|\vartheta_t - \theta^*\| \leq e^{-\delta_0 t} \|\vartheta_0 - \theta^*\|,$$

where δ_0 comes from theorem ??.

Proof. We use as Lyapunov function $V(\theta) = \frac{1}{2}\|\theta - \theta^*\|^2$. From the chain rule

$$\frac{d}{dt}V(\vartheta_t) = -\nabla_{\theta}\Gamma(\vartheta_t)^{\top} [\vartheta_t - \theta^*]$$

By convexity we get the following bound

$$\Gamma(\theta^*) \geq \Gamma(\vartheta_t) + \nabla_{\theta}\Gamma(\vartheta_t)^{\top} [\theta^* - \vartheta_t]$$

using the support condition this becomes

$$\frac{d}{dt}V(\vartheta_t) \leq \Gamma(\theta^*) - \Gamma(\vartheta_t) \leq 0$$

since θ^* is the minimum. The strict inequality (< 0) holds when $\vartheta_t \neq \theta^*$. V fulfills definition ?? and proposition ?? gives global asymptotic stability.

Under strict convexity

$$\begin{aligned} \frac{d}{dt}V(\vartheta_t) &= - \left[\nabla_{\theta}\Gamma(\vartheta_t) - \underbrace{\nabla_{\theta}\Gamma(\theta^*)}_{=0} \right]^{\top} [\vartheta_t - \theta^*] \\ &\stackrel{\text{strong convexity}}{\leq} -\delta_0 \|\vartheta_t - \theta^*\|^2 = -2\delta_0 V(\vartheta_t) \end{aligned}$$

This implies $V(\vartheta_t) \leq V(\vartheta_0) \exp(-2\delta_0 t) \forall t$ by integrating. \square

Theorem 31. *If the Polyak-Lojasiewicz (PL) inequality*

$$\frac{1}{2}\|\nabla\Gamma(\theta)\|^2 \geq \mu|\Gamma(\theta) - \Gamma(\theta^*)| \quad (11)$$

holds then the gradient flow satisfies for each initial ϑ_0

$$\Gamma(\vartheta_t) - \Gamma^* \leq e^{-\mu t}(\Gamma(\vartheta_0) - \Gamma^*).$$

If in addition Γ is coercive, then the solutions are bounded and any limit point θ_{∞} of $\{\vartheta_t\}$ is an optimizer

$$\Gamma(\theta_{\infty}) = \Gamma^*$$

Proof. We use $V(\theta) = \frac{1}{2}|\Gamma(\theta) - \Gamma^*|$ for the Lyapunov function.

$$\begin{aligned} \implies \frac{d}{dt}V(\vartheta_t) &= \frac{1}{2}\nabla_{\theta}\Gamma(\vartheta_t)^{\top} \frac{d}{dt}\vartheta_t \\ &= -\frac{1}{2}\|\nabla\Gamma(\vartheta_t)\|^2 \leq -\mu V(\vartheta_t) \end{aligned}$$

This implies using the same technique as in the previous proof

$$\begin{aligned} \frac{1}{2}[\Gamma(\vartheta_t) - \Gamma^*] &= V(\vartheta_t) \leq e^{-\mu t}V(\vartheta_0) \\ &= e^{-\mu t} \frac{1}{2}[\Gamma(\vartheta_0) - \Gamma^*] \end{aligned}$$

If Γ is coercive, then trajectories of ϑ evolve in the compact set $S = \{\theta \mid V(\theta) \leq V(\vartheta_0)\}$. If θ_{∞} is a limit point $\theta_{\infty} = \lim_{n \rightarrow \infty} \vartheta_{t_n}$ for $t_n \rightarrow \infty$. Using the continuity of the loss function, this implies optimality:

$$\Gamma(\theta_{\infty}) = \lim_{n \rightarrow \infty} \Gamma(\vartheta_{t_n}) = \Gamma^* \quad \square$$

Consider the Euler method for the gradient flow:

$$\theta_{k+1} = \theta_k - \alpha \nabla\Gamma(\theta_k) \quad (12)$$

Theorem 32. Suppose that Γ satisfies

(i) the L -smooth inequality (LSI)

$$\Gamma(\theta') \leq \Gamma(\theta) + [\theta' - \theta]^\top \nabla \Gamma(\theta) + \frac{1}{2} L \|\theta' - \theta\|^2$$

(ii) the PL inequality ??

Then it holds for $\alpha \leq \frac{1}{2}$

$$\Gamma(\theta_k) - \Gamma^* \leq (1 - \alpha\mu)^k [\Gamma(\theta_0) - \Gamma^*].$$

Proof.

$$\begin{aligned} \Gamma(\theta_{k+1}) - \Gamma(\theta_k) &\stackrel{\text{LSI}}{\leq} [\theta_{k+1} - \theta_k]^\top \nabla \Gamma(\theta_k) + \frac{1}{2} L \|\theta_{k+1} - \theta_k\|^2 \\ &\stackrel{??}{=} -\alpha \|\nabla \Gamma(\theta_k)\|^2 + \frac{1}{2} L \alpha^2 \|\nabla \Gamma(\theta_k)\|^2 \\ &= (-\alpha + \frac{1}{2} L \alpha^2) \|\nabla \Gamma(\theta_k)\|^2 \end{aligned}$$

If $\alpha \leq \frac{1}{L}$ then $(-\alpha + \frac{1}{2} L \alpha^2) \leq \frac{1}{2} \alpha$

$$\begin{aligned} &\leq -\frac{1}{2} \alpha \|\nabla \Gamma(\theta_k)\|^2 \\ &\stackrel{\text{LSI}}{\leq} -\alpha \mu |\Gamma(\theta_k) - \Gamma^*| \end{aligned}$$

and therefore

$$\Gamma(\theta_{k+1}) - \Gamma^* \leq (1 - \alpha\mu)(\Gamma(\theta_k) - \Gamma^*)$$

after iterating $k - 1$ times we obtain the result. \square

Lemma 33. Suppose that $\nabla \Gamma$ is globally Lipschitz

$$\|\nabla \Gamma(\theta') - \nabla \Gamma(\theta)\| \leq L \|\theta' - \theta\|, \quad \forall \theta, \theta' \in S$$

Then

$$(i) \quad |\langle \nabla \Gamma(\theta') - \nabla \Gamma(\theta), \theta' - \theta \rangle| \leq L \|\theta' - \theta\|^2$$

(ii) if S is convex, then Γ is L -smooth

Proof. (i)

$$\begin{aligned} |\langle \nabla \Gamma(\theta') - \nabla \Gamma(\theta), \theta' - \theta \rangle| &\leq \|\nabla \Gamma(\theta') - \nabla \Gamma(\theta)\| \|\theta' - \theta\| \\ &\leq L \|\theta' - \theta\|^2 \end{aligned}$$

(ii) for $\theta', \theta \in S$ denote $S \ni \theta^t := \theta + t(\theta' - \theta)$ and $\xi^t = \Gamma(\theta^t)$.

θ^t in S , since S is convex

$$\begin{aligned} \frac{d}{dt} \xi^t &= \langle \nabla \Gamma(\theta^t), \theta' - \theta \rangle \\ \frac{d}{dt} \xi^t - \frac{d}{dt} \xi^0 &= \langle \nabla \Gamma(\theta^t) - \nabla \Gamma(\theta^0), \theta' - \theta \rangle \\ &\stackrel{(i)}{\leq} t L \|\theta' - \theta\|^2 \end{aligned}$$

Now integrate

$$\begin{aligned} \Gamma(\theta') &= \xi^1 = \xi^0 + \int_0^1 \frac{d}{dt} \xi^t dt \\ &\leq \xi^0 + \frac{d}{dt} \xi^0 + \frac{1}{2} L \|\theta' - \theta\|^2 \\ &= \Gamma(\theta) + \langle \nabla \Gamma(\theta), \theta' - \theta \rangle + \frac{1}{2} L \|\theta' - \theta\|^2 \end{aligned}$$

\square

These are more general version of global Lipschitz and convexity

Remark. *Strong convexity:*

$$\langle \nabla \Gamma(\theta') - \nabla \Gamma(\theta), \theta' - \theta \rangle \geq \delta_0 \|\theta' - \theta\|^2$$

With $D_\Gamma(y \mid x) = \Gamma(y) - \Gamma(x) + \langle \nabla \Gamma(x), y - x \rangle$ is the Bregman divergence.

$$\frac{\mu}{2} \|\theta' - \theta\|^2 \leq D_\Gamma(\theta' \mid \theta) \leq \frac{L}{2} \|\theta' - \theta\|^2$$

2.4 Qausi stochastic approximation

Assume there are observations $\Phi_n \subset \Omega$, which we might consider as realizations of a random variable Φ . We have

$$f : \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}^d$$

$$\bar{f}(\theta) := \mathbb{E}(\underbrace{f(\theta, \Phi)}_{\text{what we observe}}), \theta \in \mathbb{R}^d$$

As before we look for $\bar{f}(\theta^*) = 0$

$$\frac{d}{dt} \vartheta_t = \bar{f}(\vartheta_t)$$

A key assumption is that what happens when following the state dynamics in any step depends only on the current state.

$$\Phi_n = [\cos(\omega n), \sin(\omega n)], \omega > 0$$

Markov chain on the unit circle. We will talk about the probing signal ξ and consider

$$\frac{d}{dt} \hat{\theta}_t = a_t f(\hat{\theta}_t, \xi_t) \quad (13)$$

a quasistochastic approximation(QSA)-ODE, a_t is the step size.

For deterministic probing signals, we mainly consider two examples

$$\xi_t = \sum_{i=1}^K \overbrace{V^i}^{\in \mathbb{R}^m} \sin(2\pi[\Phi_i + \omega_i t])$$

Mixture of periodic functions, fixed K , phase $\{\Phi_i\}$, frequencies $\{\omega_i\}$.

$$\xi_t = \sum_{i=1}^K V^i [\Phi_i + \omega_i t]_{\text{modulo } 1}$$

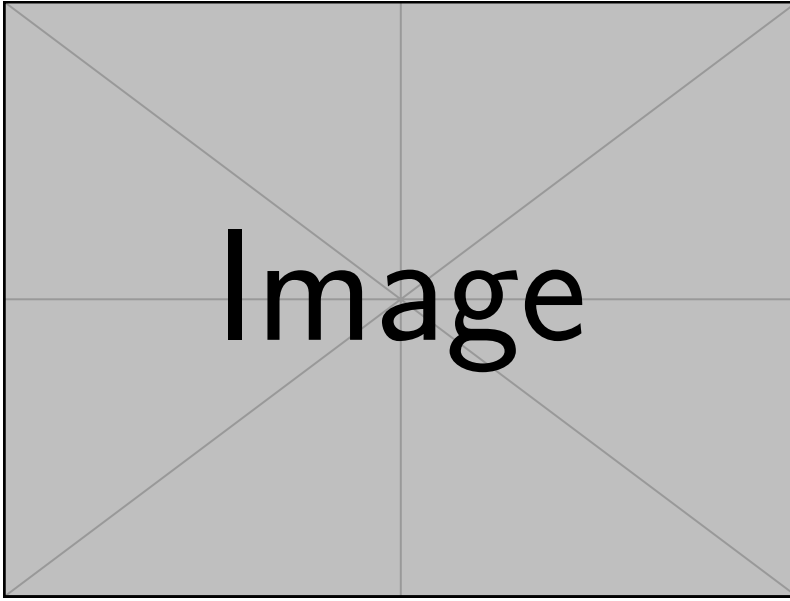


Figure 2.1: Sketch 2.01

These signals have well defined steady-state means and covariance matrices.

Special case: $\xi_t(i) = \sqrt{2} \sin(\omega_i t)$, $1 \leq i \leq m$, $\omega_i \neq \omega_j \forall i \neq j$. Then the steady-state mean

$$\lim_{T \rightarrow \infty} \int_0^T \xi_t dt = 0$$

and covariance

$$\lim_{T \rightarrow \infty} \int_0^T \xi_i \xi_i^\top dt = \text{Id}$$

We now use a slightly different notation $\hat{\theta}$ becomes $\tilde{\theta}$.

Start of lecture 09
(13.05.2025)

$$\frac{d}{dt} \tilde{\theta}_t = a_t f(\tilde{\theta}, \xi_t) \quad (14)$$

a_t non-negative.

Now consider integrating $y : [0, 1] \rightarrow \mathbb{R}$. Basic Monte-Carlo

$$\theta_n = \frac{1}{n} \sum_{i=0}^{n-1} y(\underbrace{\Phi(k)}_{\sim \text{Unif}([0,1])}) \quad (15)$$

A QSA approach is to use the saw tooth function

$$\xi_t = t(\text{modulo } 1).$$

Obtain estimate by

$$\tilde{\theta} = \frac{1}{t} \int_0^t y(\xi_t) dr \quad (16)$$

with a reasonable discretization afterwards.

To use (QSA-ODE (??)) $f(\theta, \xi) = y(\xi) - \theta$ with mean vector field

$$\begin{aligned} \bar{f}(\theta) &= \lim_{T \rightarrow \infty} \int_0^T f(\theta, \xi_t) dt \\ &= \int_0^1 y(\xi_t) dt - \theta \end{aligned}$$

which gives $\theta^* = \int_0^1 y(\xi_t) dt$ as the unique root of \bar{f} . The QSA-ODE ?? is

$$\frac{d}{dt} \tilde{\theta}_t = a_t [y(\xi_t) - \tilde{\theta}_t]$$

(??) can be transformed into

$$\frac{d}{dt} \tilde{\theta}_t = \left[-\frac{1}{t^2} \int_0^t y(\xi_r) dr + \frac{1}{t} y(\xi_t) \right] = \underbrace{\frac{1}{t}}_{\equiv a_t} [y(\xi_t) - \theta_t] \quad (17)$$

Example. $y(\theta) = e^4(\sin(100\theta))$, mean $\theta^* \approx -0.5 \approx -0.48$. Choose $a_t = \frac{g}{1+t}$

2.5 Approximate Policy Improvement

nonlinear state model in continuous time:

$$\frac{d}{dt} x_t = f(x_t, u_t), t \geq 0 \quad (18)$$

$$J^*(x) = \min_{\underline{u}} \int_0^\infty c(x_t, u_t) dt, x = x_0 \quad (19)$$

Given feedback law $u_t = \phi(x_t)$, we have

$$J^\phi(x) = \int_0^\infty c(x_t, \phi(x_t)) dt, x = x_0 \quad (20)$$

Proposition 34. If J is finite, then for each initial condition x_0 and each t

$$\frac{d}{dt} J(x_t) = -c(x_t)$$

If J is continuously differentiable, then the Lyapunov bound $\frac{d}{dt} V(x_t)$ from definition ?? follows with equality

$$\nabla J(x) f(x) = -c(x)$$

Proof. For any $T > 0$, $J(x_0) = \int_0^T c(x_r) dr + J(x_T)$. For $t \geq 0, \delta > 0$ given, use $T = t + \delta$ and $T = t$ and subtract:

$$\begin{aligned} 0 &= J(x_0) - J(x_0) = \int_t^{t+\delta} c(x_r) dr + (J(x_{t+\delta}) - J(x_t)) \\ &= \underbrace{\frac{1}{\delta} \int_t^{t+\delta} c(x_r) dr}_{\xrightarrow{\delta \rightarrow 0} c(x_t)} + \underbrace{\frac{1}{\delta} (J(x_{t+\delta}) - J(x_t))}_{\xrightarrow{\delta \rightarrow 0} \frac{d}{dt} J(x_t)} \\ &\implies \frac{d}{dt} J(x_t) = -c(x_t) \end{aligned}$$

Using the chain rule yields the second equation. □

For J^ϕ we have

$$0 = c(x, \phi(x)) + \nabla J^\phi(x) \cdot f(x, \phi(x))$$

Policy Improvement in continuous time:

$$\phi^+(x) \in \operatorname{argmin}_u \left\{ \underbrace{c(x, u) + \nabla J(x) \cdot f(x, u)}_{\text{need to approximate by } Q^\phi(x, u)} \right\}$$

Now aim for updating of Q -function. Add to the above J^ϕ on both sides

$$J^\phi(x) = J^\phi(x) + c(x, \phi(x)) + \nabla J^\phi(x) \cdot f(x, \phi(x))$$

We solved for the optimal Q -function by using a fixed point equation, with $\underline{Q}^\phi(x) = Q^\phi(x, \phi(x))$ we write

$$Q^\phi(x, u) = \underline{Q}^\phi(x) + c(x, u) + \nabla \underline{Q}^\phi(x) f(x, u).$$

Consider $\{Q^\theta \mid \theta \in \mathbb{R}^d\}$ family of approximations. Bellman errors (Temporal differences expressions?) gives

$$B^\theta(x_t, u_t) = -Q^\theta(x_t, u_t) + \underline{Q}^\theta(x) + c(x_t, u_t) + \underbrace{\nabla \underline{Q}^\theta(x) f(x_t, u_t)}_{= \frac{d}{dt} Q^\theta(x_t)} \quad (21)$$

Everything on the RHS is can be observed for any state-action pair without knowledge of f . Now, find θ^* that minimizes

$$\|B^\theta\|^2 = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [B^\theta(x_t, u_t)]^2 dt$$

Choose feedback law with exploration $u_t = \tilde{\phi}(x_t, \xi_t)$. Assuming bounded state trajectories, such that (??) exists, define $\Gamma(\theta) = \frac{1}{2} \|B^\theta\|^2$, we get

$$0 \stackrel{!}{=} \nabla \Gamma(\theta) = \lim_{t \rightarrow \infty} \int_0^T [B^\theta(x_t, u_t)] \nabla_\theta B^\theta(x_t, u_t) dt$$

Gradient flow

$$\frac{d}{dt} \vartheta_t = -\nabla_\theta \Gamma(\vartheta_t)$$

QSA counterpart is (??) with probing signal

$$\frac{d}{dt} \tilde{\theta}_t = -a_t B^{\tilde{\theta}_t}(x_t, u_t) \kappa_t^{\tilde{\theta}_t}$$

with

$$\begin{aligned} \kappa_t^{\tilde{\theta}_t} &= \nabla_\theta B^{\tilde{\theta}_t}(x_t, u_t) \\ &= -\nabla_\theta Q^\theta(x_t, u_t) + \{\nabla_\theta Q^\theta(x_t, \phi(x_t)) + \frac{d}{dt} \nabla_\theta Q^\theta(x_t, \phi(x_t))\} \end{aligned}$$

assuming we can exchange differentiation w.r.t time and w.r.t θ . (QSA-ODE)

$$\frac{d}{dt} \tilde{\theta}_t = a_t f(\tilde{\theta}_t, \xi_t)$$

aim to relate this to

$$\frac{d}{dt} \vartheta_t = \bar{f}(\vartheta_t).$$

Lemma 35. Define the change of variables

$$\tau = s_t := \int_0^t a_r dr, \quad t \geq t_0.$$

Let $\{\vartheta_\tau \mid \tau \geq \tau_0\}$ the solution to the ODE above initialized to $\tau_0 = s_{t_0}$ with $\vartheta_{\tau_0} = \tilde{\theta}_{t_0}$. The solution to

$$\frac{d}{dt} \bar{\theta}_t = a_t \bar{f}(\bar{\theta}_t), \quad t \geq t_0, \quad \bar{\theta}_{t_0} = \tilde{\theta}_{t_0}$$

is given by $\bar{\theta}_t = \vartheta_\tau$.

Proof. Change of variables and observing that

$$d\tau = a_t dt.$$

□

\underline{Q} for the fixed optimal choice

Recall $\bar{f}(\theta) := \lim_{T \rightarrow \infty} \int_0^T f(\theta, \xi_t) dt$ for all $\theta \in \mathbb{R}^d$. Remember the temporal transformation

Start of lecture 10
(15.05.2025)

$$\tau = s_t = \int_0^t a_r dr$$

and lemma ?? . Define $\hat{\theta}_\tau = \tilde{\theta}(s^{-1}(\tau)) = \tilde{\theta}_t|_{t=s^{-1}(\tau)}$. By the chain rule and observing that $d\tau = a_t dt$ yields

$$\frac{d}{d\tau} \hat{\theta}_\tau = \frac{d}{d\tau} \tilde{\theta}(s^{-1}(\tau)) = f(\tilde{\theta}(s^{-1}(\tau)), \xi(s^{-1}(\tau))).$$

$\hat{\theta}, \tilde{\theta}$ differ only by a time scaling, so convergence of the one yields convergence of the other.

Lemma 36. Consider the original ODE

$$\frac{d}{dt} \vartheta_t = \bar{f}(\vartheta_t) \quad (22)$$

and assume f is locally Lipschitz with constant L_f . Then there exists a constant B_f depending only on f , such that

$$\|\hat{\theta}_t - \hat{\theta}_0\| \leq (B_f + L_f \|\hat{\theta}_0\|) t e^{L_f t}, \quad t \geq 0$$

Proof. Proof of proposition ?? in adapted notation. □

Now, denote by $\vartheta_w^\tau, w \geq \tau$ the unique solution to (??):

$$\frac{\partial}{\partial w} \vartheta_w^\tau = \bar{f}(\vartheta_w^\tau), \quad w \geq \tau, \quad \vartheta_\tau^\tau = \hat{\theta}_\tau$$

with that we get

1. $\vartheta_{\tau+v}^\tau = \hat{\theta}_\tau + \int_0^{\tau+v} \bar{f}(\vartheta_w^\tau) dw, \quad \tau, v \geq 0$
2. $\hat{\theta}_{\tau+v} = \hat{\theta}_\tau + \int_\tau^{\tau+v} f(\hat{\theta}_w, \xi(s^{-1}(w))) dw, \quad \tau, v \geq 0$

The following assumptions will be used in the following:

QSA1 The process a is non-negative, monotonically decreasing and $\lim_{t \rightarrow \infty} a_t = 0, \int_0^\infty a_r dr = \infty$

QSA2 The functions \hat{f}, f are Lipschitz continuous with constant L_f :

$$\begin{aligned} \|\bar{f}(\theta') - \bar{f}(\theta)\| &\leq \|L_f\| \|\theta' - \theta\| \\ \|f(\theta', z) - f(\theta, z)\| &\leq \|L_f\| \|\theta' - \theta\| \end{aligned}$$

for all $\theta, \theta' \in \mathbb{R}^d, z \in \Omega$ and there exists a Lipschitz continuous functions $b_0 : \mathbb{R}^d \rightarrow \mathbb{R}_+$, such that for all $\theta \in \mathbb{R}^d$

$$\left\| \int_{t_0}^{t_1} f(\theta, \xi_t) - \bar{f}(\theta) dt \right\| \leq b_0(\theta), \quad 0 \leq t_1 \leq t_0$$

QSA3 The ODE $\frac{d}{dt} \vartheta_t = \bar{f}(\vartheta_t)$ has a globally asymptotically stable equilibrium θ^*

Consider first, arbitrary θ

Lemma 37. Assume (QSA1), (QSA2) hold for any fixed $T > 0$ and $\theta \in \mathbb{R}^d$.

$$\left\| \int_\tau^{\tau+T} [f(\theta, \xi(s^{-1}(w))) - \bar{f}(\theta)] dw \right\| \leq b_0(\theta) \epsilon_\tau^f,$$

where $\epsilon_\tau^f = 3a_t|_{t=s^{-1}(\tau)}$ and b_0 comes from (QSA2).

quasi-
contin

There is
the law of
...

Proof. Set $\tilde{f}_w(\theta) = f(\theta, \xi_w) - \bar{f}(\theta)$ for each w, θ . Write

$$E_t = \int_0^t \tilde{f}_w(\theta) dw.$$

By assumptions $\|E_t\| \leq b_0(\theta)$, $t \geq 0$.

$$\begin{aligned} \int_{t_0}^{t_1} a_t \tilde{f}_t(\theta) dt &\stackrel{\text{IbP}}{=} a_t E_t \Big|_{t_0}^{t_1} - \int_{t_0}^{t_1} |a'_t| E_t dt \\ \left\| \int_{t_0}^{t_1} a_t \tilde{f}_t(\theta) dt \right\| &\leq a_{t_0} \|E_{t_0}\| + a_{t_1} \|E_{t_0}\| + \int_{t_0}^{t_1} |a'_t| E_t dt \\ &\stackrel{a \text{ decreasing}}{\leq} 2a_{t_0} b_0(\theta) - b_0(\theta) \int_{t_0}^{t_1} a'_t dt \\ &\leq 3a_{t_0} b_0(\theta) \end{aligned}$$

Set $t_0 = s^{-1}(\tau)$, $t_1 = s^{-1}(\tau + T)$, $t = s^{-1}(w)$, giving $dw = a_t dt$

$$\begin{aligned} \left\| \int_{\tau}^{\tau+T} [f(\theta, \xi(s^{-1}(w))) - \bar{f}(\theta)] dw \right\| &= \left\| \int_{t_0}^{t_1} a_t \tilde{f}_t(\theta) dt \right\| \\ &\leq 3a_{t_0} b_0(\theta) = \epsilon_{\tau}^f b_0(\theta) \end{aligned}$$

□

Proposition 38. Assuming that $\hat{\theta}$ is bounded. Then for any $T > 0$

$$\lim_{\tau \rightarrow \infty} \sup_{v \in [0, T]} \left\| \overbrace{\int_{\tau}^{\tau+v} [f(\hat{\theta}_w, \xi(s^{-1}(w))) - \bar{f}(\hat{\theta}_w)] dw}^{E_{\tau+v}^{\tau}} \right\| = 0$$

and

$$\lim_{\tau \rightarrow \infty} \sup_{v \in [0, T]} \left\| \hat{\theta}_{\tau+v} - \vartheta_{\tau+v}^{\tau} \right\| = 0$$

Proof. We use piecewise constant approximation, as in Riemannian integration, and set for $\delta > 0$, $\tau_k = \tau + k\delta$, $k \geq 0$

$$E_{\tau+v}^{\tau} = \sum_{k=0}^{n_v-1} \int_{\tau_k}^{\tau_{k+1}} [f(\hat{\theta}_{\tau_k}, \xi(s^{-1}(w))) - \bar{f}(\hat{\theta}_{\tau_k})] dw + \epsilon_v^{\tau},$$

which holds due to (QSA1), Lipschitz condition, $n_v = \lfloor \frac{v}{\delta} \rfloor$. and

$$\|\epsilon_v^{\tau}\| \leq b_L v \delta$$

for some finite constant b_L . Assuming $\hat{\theta}$ is bounded, this bound is uniform in τ . For fixed $\hat{\theta}_{t_k}$ we can use lemma ??, so

$$\begin{aligned} \|E_{\tau+v}^{\tau}\| &\leq \sum_{k=0}^{n_v-1} \epsilon_{\tau_k}^f b_0(\hat{\theta}_{t_k}) + b_L v \delta \\ &\leq \epsilon_{\tau}^f \sum_{k=0}^{n_v-1} b_0(\hat{\theta}_{\tau_k}) + b_L v \delta \end{aligned}$$

Let $b < \infty$ denote a constant such that $b_0(\hat{\theta}_{\tau_k}) \leq b \forall \tau$, which we can do since $\hat{\theta}$ is bounded, b_0 Lipschitz.

$$\|E_{\tau+v}^{\tau}\| \leq b \frac{v}{\delta} \underbrace{\epsilon_{\tau}^f}_{\xrightarrow{\tau \rightarrow \infty} 0 \text{ by QSA1}} + b_L v \delta$$

For any $T > 0$

$$\lim_{\tau \rightarrow \infty} \sup_{v \in [0, T]} \|E_{\tau+v}^\tau\| \leq 0 + b_L T \delta$$

Since $\delta > 0$ was arbitrary, we have the first statement.

For the second limit: $E_r^\tau = \vartheta_r^\tau - \hat{\theta}_r$. The pair of identities after lemma ?? give using Lipschitz condition from (QSA2) we get

$$E_{\tau+v}^\tau = 0 + \int_\tau^{\tau+v} \bar{f}(\hat{\theta}_w) - f(\hat{\theta}_w, \xi(s^{-1}(w))) dw + \int_\tau^{\tau+v} \underbrace{\left[\bar{f}(\vartheta_w^\tau) - \bar{f}(\hat{\theta}_w) \right]}_{\|\dots\| \leq L_f \|E_w^\tau\|} dw$$

$$\|E_{\tau+v}^\tau\| \leq \delta^\tau + L_f \int_\tau^{\tau+v} \|E_w^\tau\| dw,$$

where

$$\delta^\tau := \sup_{\tau' \geq \tau} \max_{0 \leq v \leq T} \left\| \int_{\tau'}^{\tau'+v} \left[\bar{f}(\hat{\theta}_w) - f(\hat{\theta}_w, \xi(s^{-1}(w))) \right] dw \right\|$$

Grönwall's lemma gives

$$\|E_{\tau+v}^\tau\| \leq e^{L_f v} \delta^\tau, \quad 0 \leq v \leq 1$$

$\delta^\tau \rightarrow 0$ for $\tau \rightarrow \infty$ due to the first statement. \square

Start of lecture 11
(20.05.2025)

Theorem 39 (Boundedness implies convergence). *Suppose that (QSA1-QSA3) hold. Further assume **ultimate boundedness**, i.e. that a $b < \infty$ exists, such that for each $\theta \in \mathbb{R}^d$ and $z \in \Omega$ there is a $T_{\theta,z}$, such that $\|\hat{\theta}_\tau\| \leq b$ for all $\tau \geq T_{\theta,z}$, where $\hat{\theta}_0 = \theta, \xi_0 = z$. Then the solution to (??)*

$$\frac{d}{dt} \tilde{\theta}_t = a_t f(\tilde{\theta}_t, \xi_t)$$

converges to θ^ for each initial condition.*

Proof. Consider the time scaled $\hat{\theta}_t$

$$\|\vartheta_\tau^\tau\| = \|\hat{\theta}_t\| \stackrel{\text{pA}}{\leq} b, \quad \tau \geq T_{\theta,z}$$

Using (QSA3), i.e. $\frac{d}{dt} \vartheta_t = \bar{f}(\vartheta_t)$ has a globally asymptotically stable equilibrium θ^* , we have that for every $\epsilon > 0$, there exists $T_\epsilon > 0$ s.t. $\|\vartheta_{\tau+v}^\tau - \theta^*\| < \epsilon \quad \forall v \geq T_\epsilon$, whenever $\|\vartheta_\tau^\tau\| \leq b$. Then

$$\limsup_{\tau \rightarrow \infty} \|\hat{\theta}_{\tau+T_\epsilon} - \theta^*\| \leq \underbrace{\limsup_{\tau \rightarrow \infty} \|\hat{\theta}_{\tau+T_\epsilon} - \vartheta_{\tau+T_\epsilon}^\tau\|}_{\rightarrow 0 \text{ by proposition ??}} + \underbrace{\limsup_{\tau \rightarrow \infty} \|\vartheta_{\tau+T_\epsilon}^\tau - \theta^*\|}_{\leq \epsilon} \quad \square$$

Lemma 40 (Weaker form of proposition ?? (ii)). *For some $\bar{\delta} < \infty$ and any $0 \leq T \leq 1$*

$$\|\hat{\theta}_{\tau+T} - \vartheta_{\tau+T}^\tau\| \leq e^{L_f T} b_0(\hat{\theta}_\tau) \epsilon_\tau^f + \bar{b}(1 + \|\hat{\theta}_\tau\|) T^2$$

where $b_0(\theta)$ and L_f are from (QSA2).

Proof. Write $E_r^\tau = \vartheta_r^\tau - \hat{\theta}_r$, $r \geq T$. The pair of identities after lemma ?? give, after inserting $\pm \bar{f}(\theta_w)$

$$E_{\tau+T}^\tau = 0 + \int_\tau^{\tau+T} \left[\bar{f}(\hat{\theta}_w) - f(\hat{\theta}_w, \xi(s^{-1}(w))) \right] dw + \int_\tau^{\tau+T} \left[\bar{f}(\vartheta_w^\tau) - \bar{f}(\hat{\theta}_w) \right] dw$$

using (QSA2) we can bound

$$\begin{aligned}\|\bar{f}(\hat{\theta}_w) - \bar{f}(\hat{\theta}_\tau)\| &\leq L_f \|\hat{\theta}_w - \hat{\theta}_\tau\| \\ \|f(\hat{\theta}_w, \xi(s^{-1}(w))) - f(\hat{\theta}_\tau, \xi(s^{-1}(w)))\| &\leq L_f \|\hat{\theta}_w - \hat{\theta}_\tau\| \\ \|\bar{f}(\vartheta_w^\tau) - \bar{f}(\vartheta_w)\| &\leq L_f \|E_w^\tau\|\end{aligned}$$

With that, for any $T > 0$ by inserting terms with $\hat{\theta}_\tau$

$$\begin{aligned}\|E_{\tau+T}^\tau\| &\leq \left\| \int_\tau^{\tau+T} [\bar{f}(\hat{\theta}_\tau) - f(\hat{\theta}_\tau, \xi(s^{-1}(w)))] dw \right\| + 2L_f \int_\tau^{\tau+T} \|\hat{\theta}_w - \hat{\theta}_\tau\| + L_f \int_\tau^{\tau+T} \|E_w^\tau\| dw \\ &\leq \alpha_T^\tau + L_f \int_\tau^{\tau+T} \|E_w^\tau\| dw,\end{aligned}$$

where

$$\alpha_T^\tau := \underbrace{b_0(\hat{\theta}_\tau)}_{\text{from (QSA2)}} \epsilon_\tau^f + 2L_f \int_0^T \|\hat{\theta}_{\tau+w} - \hat{\theta}_\tau\| dw$$

Using Grönwall's lemma, proposition ?? (ii)

$$\|E_{\tau+T}^\tau\| \leq \alpha_T^\tau e^{L_f T}$$

Repeating the proof for proposition ??, we get

$$\|\hat{\theta}_{\tau+w} - \hat{\theta}_\tau\| \leq (B_f + L_f \|\hat{\theta}_\tau\|) w e^{L_f w}.$$

Increase $e^{L_f w}$ to $e^{L_f T}$ to get

$$\begin{aligned}2 \int_0^T \|\hat{\theta}_{\tau+w} - \hat{\theta}_\tau\| dw &\leq 2(B_f + L_f \|\hat{\theta}_\tau\|) e^{L_f T} \int_0^T w dw \\ &= (B_f + L_f \|\hat{\theta}_\tau\|) T^2 e^{L_f T}\end{aligned}$$

Hence

$$\alpha_T^\tau \leq b_0(\hat{\theta}_\tau) \epsilon_\tau^f + L_f (B_f + L_f \|\hat{\theta}_\tau\|) T^2 e^{L_f T}$$

Since $0 \leq T \leq 1$, we can find $\bar{b} < \infty$ to bound $L_f (B_f + L_f \|\hat{\theta}_\tau\|) T^2 e^{L_f T}$ by $\bar{b}(1 + \|\hat{\theta}_\tau\|) T^2$, where \bar{b} depends on fixed B_f, L_f . \square

Reminder, **drift condition**

$$\langle \nabla f(\theta), f(\theta) \rangle < 0, \quad \theta \neq \theta^*$$

Definition 41 (ultimately bounded). *The ODE*

$$\frac{d}{d\vartheta_t} = f(\vartheta_t), \quad \vartheta_0 = \theta_0$$

is called **ultimately bounded** if there exists a bounded set $S \subset \mathbb{R}^d$, such that for each initial condition θ_0 , there is a time $T(\theta_0)$ such that $\vartheta_t \in S \quad \forall t \geq T(\theta_0)$.

Proposition 42. *Assume that there is a continuously differentiable function $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ satisfying the Lyapunov condition*

$$\langle \nabla V(\theta), f(\theta) \rangle \leq -\delta_0, \quad \theta \in S^c$$

for some $\delta_0 > 0$ and some set $S \subset \mathbb{R}^d$. Then $T_S(\theta) \leq \delta_0^{-1} V(\theta)$ for $\theta \in \mathbb{R}^d$, where

$$T_S(\theta) = \min\{t \mid \vartheta_t \in S\}, \quad \vartheta_0 = \theta \in \mathbb{R}^d.$$

If in addition S is compact and V inf-compact, then the corresponding ODE is ultimately bounded.

Proof. Assume $\delta_0 = 1$ w.l.o.g., we interpret the condition as *along a path*

$$\frac{d}{dt}V(\vartheta_t) \leq 1,$$

for $0 \leq t \leq T_S(\theta)$, $\vartheta_0 = \theta \in \mathbb{R}^d$. $T_N = \min(N, T_S(\theta))$, integrate both sides from $t = 0$ to $t = T_N$.

$$-V(\vartheta_0) \leq V(\vartheta_{T_N}) - V(\vartheta_0) \leq \int_0^{T_N} \frac{d}{dt}V(\vartheta_t)dt \leq -T_N$$

or $\min(N, T_S(\theta)) \leq V(\vartheta_0)$. Choosing $N \geq V(\vartheta_0)$ gives the bound on the first entrance time:

$$T_S(\theta) \leq \delta_0^{-1}V(\theta).$$

Now we need to show that it stays in some S . Now, S is compact, V is inf-compact, so there exists $N < \infty$ such that $S \subset S_V(N) = \{\theta \mid V(\theta) \leq N\}$, with $S_V(N)$ compact as well. Hence

$$\langle \nabla V(\theta), f(\theta) \rangle \leq -1, \quad \theta \in \mathbb{R}^d, \quad V(\theta) \geq N$$

writing $V(\theta) > N$ corresponds to $\theta \in S_V(N)^c$.

Now, $V(\vartheta_t)$ is therefore decreasing, whenever $\vartheta_t \in S_V(N)^c$, this shows that the set $S_V(N)$ is **absorbing**, which gives that

$$\vartheta_t \in S_V(N) \quad \forall t \geq T_S(\theta).$$

□

Assumption (QSV):

There exists a continuous function $V : \mathbb{R}^d \rightarrow \mathbb{R}$, and constants $c_0 > 0, \delta_0$ s.t. for any initial condition ϑ_0 of the ODE and $0 \leq T \leq 1$ it holds for $\|\vartheta_s\| > c_0$, that

$$V(\vartheta_{s+T}) - V(\vartheta_s) \leq -\delta_0 \int_0^T \|\vartheta_{s+t}\| dt.$$

The Lyapunov function V is Lipschitz continuous with constant L_V .

If V is differentiable, then QSV implies

$$\frac{d}{dt}V(\vartheta_t) \leq -\delta_0 \|\vartheta_t\|,$$

whenever $\|\vartheta_t\| > c_0$.

Lemma 43. Assume $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ is Lipschitz continuous and that for some constant $T > 0$, $0 < \delta_1 < 1$ and $\tau_0, b < \infty$ it holds

$$V(\hat{\theta}_{\tau+T}) - V(\hat{\theta}_\tau) \leq -\delta_1 \|\hat{\theta}_\tau\|$$

for all $\tau \geq \tau_0$, $\|\hat{\theta}_\tau\| > b$. Then the solution to the time-scaled ODE

$$\frac{d}{d\tau} \hat{\theta}_\tau = f(\hat{\theta}(s^{-1}(\tau)), \xi(s^{-1}(\tau))) \quad (23)$$

is ultimately bounded.

Proof. For each initial condition $\hat{\theta}_0 = \theta$ and $\tau \geq \tau_0$, denote by

$\hat{\tau} = \hat{\tau}(\theta, \tau) := \min(v \geq 0 \mid \|\hat{\theta}_{\tau+v}\| \leq b)$, where τ_0, b as before. For clarity, if $\|\hat{\theta}_{\tau+v}\| > b$ for all $v \geq 0$, set $\hat{\tau} = \infty$.

For $m \in \mathbb{Z}_+$, define $\hat{\tau}_m = \min(\hat{\tau}, m)$. Then

$$\begin{aligned} -\hat{\tau}_m b \delta_1 &\geq -\delta_1 \int_\tau^{\tau+\hat{\tau}_m} \underbrace{\|\hat{\theta}_w\|}_{\leq b} dw \\ &\geq \int_\tau^{\tau+\hat{\tau}_m} (V(\hat{\theta}_{w+T}) - V(\hat{\theta}_w)) dw \\ &= \int_{\tau+\hat{\tau}_m}^{\tau+\hat{\tau}_m+T} V(\hat{\theta}_w) dw - \int_\tau^{\tau+T} V(\hat{\theta}_w) dw \\ &\geq - \int_\tau^{\tau+T} V(\hat{\theta}_w) dw \end{aligned}$$

This bound is independent of m , holds for all $\hat{\tau}_m$. Therefore

$$\begin{aligned}\hat{\tau} &\leq \frac{1}{b\delta_1} \int_{\tau}^{\tau+T} V(\hat{\theta}_w) dw \\ \int_{\tau}^{\tau+T} V(\hat{\theta}_w) dw &\leq \int_{\tau}^{\tau+T} |V(\hat{\theta}_w) - V(\hat{\theta}_{\tau})| + |V(\hat{\theta}_{\tau})| dw \\ &\leq \int_{\tau}^{\tau+T} \underbrace{L_V \|\hat{\theta}_w - \hat{\theta}_{\tau}\|}_{\text{prop ??: } \leq (C(V) + \|\hat{\theta}_{\tau}\|)c(L_V, T)} + L_V \|\hat{\theta}_{\tau}\| + |V(\hat{\theta}_{\tau})| dw\end{aligned}$$

So the integral can be bounded by constants depending on fixed values. So we can obtain a bound

$$\hat{\tau} \leq b_V(1 + \|\hat{\theta}_{\tau}\|).$$

Hence $\hat{\tau}(\theta, \tau)$ is everywhere finite.

Denote by $b_1 \sup\{\|\hat{\theta}_{\tau+v}\| \mid \tau \geq \tau_0, v \leq \hat{\tau}(\theta, \tau), \|\hat{\theta}_{\tau}\| \leq b+1\}$. In words, b_1 bounds the maximum norm of any breakout at time τ if $\hat{\theta}_{\tau} \in S = \{\theta \mid \|\theta\| \leq b+1\}$ and ends at the arrival time to the set

$$S_0 := \{\theta \mid \|\theta\| \leq b\}$$

denoted $\tau + \hat{\tau}(\theta, \tau)$.

Now every trajectory enters $S_0 \subset S$ for some $\tau \geq \tau_0$, so it fulfills that $\|\hat{\theta}_{\tau}\| \leq b_1$ for all τ sufficiently large, which gives ultimate boundedness. \square

Proposition 44. *Under (QSV), the solution to (??) is ultimately bounded, i.e. there exists some $b < \infty$ such that for any*

$$\hat{\theta}_0 = \theta, \limsup_{\tau \rightarrow \infty} \|\hat{\theta}_{\tau}\| \leq b$$

Proof. V is from (QSV) and c_0 the constant. For $0 \leq T \leq 1$, $\|\hat{\theta}_{\tau}\| \geq c_0 + 1$

$$\begin{aligned}V(\hat{\theta}_{\tau+T}) - V(\hat{\theta}_{\tau}) &= V(\hat{\theta}_{\tau+T}) - V(\vartheta_{\tau+T}^{\tau}) + V(\vartheta_{\tau+T}^{\tau}) - \underbrace{V(\vartheta_{\tau}^{\tau})}_{= \hat{\theta}_{\tau}} \\ &\leq |V(\hat{\theta}_{\tau+T}) - V(\vartheta_{\tau+T}^{\tau})| + V(\vartheta_{\tau+T}^{\tau}) - V(\vartheta_{\tau}^{\tau}) \\ &\leq L_V \|\hat{\theta}_{\tau+T} - \vartheta_{\tau+T}^{\tau}\| - \delta_0 \int_0^T \underbrace{\|\vartheta_{\tau+T}^{\tau}\|}_{\leq \|\hat{\theta}_{\tau}\| + \|\int_{\tau}^{\tau+T} \bar{f}(\vartheta_w^{\tau}) dw\|} dt \\ &\leq L_V \|\hat{\theta}_{\tau+T} - \vartheta_{\tau+T}^{\tau}\| - \delta_0 T \|\hat{\theta}_{\tau}\| \\ &\stackrel{\text{Lemma ??}}{\leq} L_V(e^{L_f} b_0(\hat{\theta}_{\tau}) \epsilon_{\tau}^f + \bar{b}(1 + \|\hat{\theta}_{\tau}\|)T^2) - \delta_0 T \|\hat{\theta}_{\tau}\|\end{aligned}$$

So, we can choose $T > 0$ small enough and τ_0 large enough, so that

$$V(\hat{\theta}_{\tau+T}) - V(\hat{\theta}_{\tau}) \leq -\frac{1}{2}\delta_0 T \|\hat{\theta}_{\tau}\|, \quad \tau \geq \tau_0, \quad \|\hat{\theta}_{\tau}\| \geq c_0 + 1$$

and we can use the lemma ??.

\square

2.6 Gradient free Optimization

Reminder:

$$\min_{\theta \in \mathbb{R}^d} \Gamma(\theta)$$

we assume it has a unique minimizer θ^* .

$$\bar{f}(\theta) = \nabla \Gamma(\theta)$$

we look for θ^* with $\bar{f}(\theta^*) = 0$. But, we are using $f(\theta, \xi_t)$ due to lack of information. Generally, we design some $\tilde{\nabla}_\Gamma(t)$ to approximate the above in an average sense

$$\int_{T_0}^{T_1} a_t \tilde{\nabla}_\Gamma(t) dt \approx \int_{T_0}^{T_1} a_t \nabla \Gamma(\tilde{\theta}_t) dt, \quad T_1 > T_0 > 0$$

and construct an ODE

$$\frac{d}{dt} \tilde{\theta}_t = -a_t \tilde{\nabla}_\Gamma(t) \quad (24)$$

We now assume $\psi_t = \tilde{\theta}_t + \epsilon \xi_t$, $t \geq 0, \epsilon \geq 0$ and we observe $\Gamma(\psi_t)$ for each t . Here ψ_t is a d -dimensional probing signal.

We had

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \xi_t dt = 0, \quad \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \xi \cdot \xi^\top dt = Id$$

2.6.1 Algorithm: quasi Stochastic Gradient Descent #1: qSGD #1

Input: $d \times d$ pos. def. matrix G , $\tilde{\theta}_0 \in \mathbb{R}^d$

$\frac{d}{dt} \tilde{\theta}_t = -a_t \frac{1}{\epsilon} G \xi_t \Gamma(\psi_t)$, where $\psi_t = \tilde{\theta}_t + \epsilon \xi_t$. In QSA-ODE we have therefore $f(\theta_t, \xi_t) = -\frac{1}{\epsilon} G \xi_t \Gamma(\theta_t + \epsilon \xi_t)$ iff $\Gamma \in C^2$:

$$\Gamma(\theta + \epsilon \xi_t) = \Gamma(\theta) + \epsilon \xi_t^\top \nabla \Gamma(\theta) + \frac{1}{2} \epsilon^2 \xi_t^\top \nabla^2 \Gamma(\theta) \xi_t + o(\epsilon^2).$$

$$f(\theta, \xi_t) = -\frac{1}{\epsilon} G \xi_t \Gamma(\theta) - G \xi_t \xi_t^\top \nabla \Gamma(\theta) + O(\epsilon)$$

$$\underbrace{\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T f(\theta, \xi_t) dt}_{\bar{f}_\epsilon(\theta)} = 0 - G \nabla \Gamma(\theta) + O(\epsilon)$$

For $G = Id$ qSGD#1 will approximate the steepest descent algorithm. In (QSA2) we assumed that f, \bar{f} are Lipschitz, but while $\nabla \Gamma$ usually is Lipschitz, Γ often is not!

2.6.2 Algorithm: qSDG #3

For a given $d \times d$ pos. def. matrix G and $\tilde{\theta}_0 \in \mathbb{R}^d$

$$\frac{d}{dt} \tilde{\theta}_t = -a_t \frac{1}{2\epsilon} G \xi_t \left[\Gamma(\tilde{\theta}_t + \epsilon \xi_t) - \Gamma(\tilde{\theta}_t - \epsilon \xi_t) \right] \quad \text{--: } a_t f(\tilde{\theta}_t, \xi_t)$$

f can be shown to be Lipschitz in θ , whenever $\nabla \Gamma$ is Lipschitz. In this case

$$f(\theta, \xi_t) = -G \xi_t \xi_t^\top \nabla \Gamma(\theta) + o(\epsilon), \quad \lim_{T \rightarrow \infty} \int_0^T f(\theta, \xi_t) dt = -G \nabla \Gamma(\theta) + o(\epsilon)$$

Start of lecture 13
(27.05.2025)

Proposition 45 (Global consistency). *Suppose that the following hold for Γ and the algorithm parameters in QSGD#3*

1. (QSA1) holds
2. The probing signal satisfies

$$\int_0^T \xi_t \xi_t^\top dt = Id$$

3. $\nabla \Gamma$ is globally Lipschitz continuous, and Γ is strongly convex with unique minimizer $\theta^* \in \mathbb{R}^d$

Control on both sides of
the loss function ...

4. the corresponding QSA-ODE is ultimately bounded

Then there exists $\bar{\epsilon} > 0$ s.t. for all $\epsilon \in (0, \bar{\epsilon})$ there is a unique root θ_ϵ^* of \bar{f}_ϵ , satisfying

$$\|\theta_\epsilon^* - \theta^*\| = O(\epsilon)$$

Moreover, convergence holds from each initial condition:

$$\lim_{t \rightarrow \infty} \theta_t = \theta_\epsilon^*$$

Proof. The assumptions imply that (QSA2) holds for

$$f(\theta, \xi) = -G\xi\xi^\top \nabla \Gamma(\theta) + O(\epsilon)$$

$$\bar{f}_\epsilon(\theta) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T f(\theta, \xi) dt$$

Γ is strongly convex, therefore there is an $\epsilon_0 > 0$ s.t. there is a unique solution to $G\nabla \Gamma(\theta) = z$, whenever $\|z\| \leq \epsilon_0$. From this (QSA3, the asymptotic stability condition), can be established for $\epsilon > 0$ small enough.

Theorem ?? yields that for each $\epsilon > 0$, θ_t converges to the unique root θ_ϵ^* of \bar{f}_ϵ satisfying

$$\|\nabla \Gamma(\theta_\epsilon)\| = O(\epsilon)$$

From there, strong convexity gives

$$\Gamma(\theta^*) \geq \Gamma(\theta_\epsilon^*) + \nabla \Gamma(\theta_\epsilon^*)^\top (\theta^* - \theta_\epsilon^*) + \frac{\eta}{2} \|\theta_\epsilon^* - \theta^*\|^2$$

for some $\eta > 0$.

$$\begin{aligned} \frac{\eta}{2} \|\theta_\epsilon^* - \theta^*\|^2 &\leq \underbrace{\Gamma(\theta^*) - \Gamma(\theta_\epsilon^*)}_{\leq 0} + \nabla \Gamma(\theta_\epsilon^*)^\top (\theta^* - \theta_\epsilon^*) \\ &\leq \|\nabla \Gamma(\theta_\epsilon^*)\| \|\theta^* - \theta_\epsilon^*\| \end{aligned}$$

which gives

$$\|\theta_\epsilon^* - \theta^*\|^2 = O(\epsilon).$$

□

Remark. For the exam: About the structure of the proof / is it long / technical / which results does it use?

Chapter 3:

Value and Q -Function approximation

3.1 A very short crash course in machine learning

How can we represent functions?

Goal:

$$h(x) = \sum_{i=1}^d \theta_i \psi_i(x).$$

We could also use neural networks, or kernels:

$$h(x) = \sum_{i=1}^d \theta_i k(x, x_i)$$

$$K_{ij} = k(x_i, x_j)$$

is a positive (semi)-definite matrix for each dataset.

1. We need a way to represent a function $h \in \mathcal{H}$

- linear
- neural networks
- piecewise polynomials
- kernels

2. loss $\Gamma(h)$, $\Gamma(h) = \Gamma(h(z_1), h(z_2), \dots, h(z_N))$ evaluated at some samples z_i , $1 \leq i \leq N$

3. algorithm to obtain $\operatorname{argmin}_{h \in \mathcal{H}} \Gamma(h)$

Training data $\{(z_i, y_i)\}_{i=1}^N$, $y_i = h^*(z_i) + \epsilon_i$,

$$\Gamma(h) = \frac{1}{N} \sum_{i=1}^N (y_i - h(z_i))^2$$

We usually use **regularization** to avoid **overfitting**.

Always reserve samples for evaluating the quality of the prediction.

3.2 Reinforcement Learning

$$\mathcal{D}_{k+1}(Q^\theta) = -Q^\theta(x(k), u(k)) + c(x(k), u(k)) + \underbrace{Q^\theta(x(k+1))}_{=\min_u Q^\theta(x, u) \text{ or } Q^\theta(x(k+1), \phi(x(k+1)))}$$

We have a sequence of state-action pairs

$$\underbrace{\{x(k), u(k)\}}_{z_k} \mid 0 \leq k \leq N$$

$$\Gamma(h) = \frac{1}{N} \sum_{k=1}^N D_k(h(z_k), h(z_{k+1}))^2$$

where

$$D(h(z_k), h(z_{k+1})) := -h(x(k-1), u(k-1)) + c(x(k-1), u(k-1)) + \underline{h}(x(k))$$

with $\underline{h}(x) = \min_u h(x, u)$.

$$Q^\theta(x, u) = \theta^\top \Psi(x, u), \quad \theta \in \mathbb{R}^d$$

and Ψ a collection of basis functions ψ_i . Write

$$\begin{aligned} \gamma_k &= c(x(k), u(k)) \\ \tilde{\gamma}_{k+1} &= \Psi(x(k), u(k)) - \Psi(x(k+1), \phi(x(k+1))). \end{aligned}$$

Rewrite $D_{k+1}(Q^\theta)$ as

$$\gamma_k = \tilde{\gamma}_{k+1}^\top \theta + \underbrace{D_{k+1}(Q^\theta)}_{:=\epsilon_k}$$

This looks like a regression problem:

$$\Gamma(\theta) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} \left[\underbrace{\gamma_k - \tilde{\gamma}_{k+1}^\top \theta}_{=D_{k+1}(Q^\theta)} \right]^2$$

Look for $\theta^* = \operatorname{argmin}_\theta \Gamma(\theta)$.

3.2.1 Algorithm: Least Squares Temporal Difference Learning (LSTD)

For a given $d \times d$ regularization matrix W , W psd, integer N , and obtained samples $\{(x(k), u(k)) \mid 0 \leq k \leq N\}$, the minimizer is obtained.

One of three streams in RL

$$\theta_N^{\text{LSTD}} = \operatorname{argmin}_\theta \Gamma_N(\theta), \quad \Gamma_N(\theta) = \theta^\top W \theta + \frac{1}{N} \sum_{k=0}^{N-1} [\gamma_k - \tilde{\gamma}_{k+1}^\top(\theta)]^2 \quad (1)$$

$$Q_N^{\theta^{\text{LSTD}}} = \sum_{i=1}^d \theta_N^{\text{LSTD}}(\psi(i)) \psi(i)$$

is the approximation of the Q -function.

We have a positive definite quadratic objective, so the solution to (??) can be obtained by solving for $\nabla \Gamma(\theta) \stackrel{!}{=} 0$.

Proposition 46. Define $R_N = \frac{1}{N} \sum_{i=1}^N \tilde{\gamma}_i \tilde{\gamma}_i^\top$, $\bar{\psi}_N^\gamma = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{\gamma}_{k+1} \gamma_k$. Then $\theta_N^{LSTD} = [\frac{1}{N} W + R_N]^{-1} \bar{\Psi}_N^\gamma$

The regularization W is introduced to ensure a unique solution.

Proposition 47 (Redundant Parametrization). Suppose that $R_N = \frac{1}{N} \sum_{i=1}^N \tilde{\gamma}_i \tilde{\gamma}_i^\top$ has rank less than d . Then there is a non zero vector $v \in \mathbb{R}^d$ for which the following two statements hold for each $0 \leq k \leq N-1$:

(i) For any $\theta \in \mathbb{R}^d$ and $r \in \mathbb{R}$:

$$D_{k+1}(Q^\theta) = D_{k+1}(Q^{\theta'}),$$

where $\theta' = \theta + rv$.

(ii) From the on-policy implementation $u(k) = \psi(x(k))$

$$v^\top \Psi(x(0), u(0)) = v^\top \Psi(x(k), u(k)).$$

Proof. R_N does not have full rank, therefore there exists $v \neq 0$ s.t.

$$0 = v^\top R_N v = \frac{1}{N} \sum_{i=1}^N (v^\top \tilde{\gamma}_i)^2.$$

Therefore, $v^\top \tilde{\gamma}_k = 0$ for every observed sample.

$$0 = v^\top \Psi(x(k), u(k)) - v^\top \Psi(x(k+1), \phi(x(k+1))), \quad 0 \leq k \leq N-1 \quad (2)$$

So,

$$\begin{aligned} D_{k+1}(Q^{\theta'}) &= -Q^{\theta'}(x(k), u(k)) + c(x(k), u(k)) + Q^{\theta'}(x(k+1), \phi(x(k+1))) \\ &= c(x(k), u(k)) + [\theta + rv] [-\Psi(x(k), u(k)) + \Psi(x(k+1), \phi(x(k+1)))] \\ &\stackrel{??}{=} c(x(k), u(k)) + \theta [-\Psi(x(k), u(k)) + \Psi(x(k+1), \phi(x(k+1)))] , \end{aligned}$$

which yields (i).

If $u(k) = \phi(x(k))$, use (??)

$$v^\top \Psi(x(k), u(k)) = v^\top \Psi(x(k+1), u(k+1))$$

repeated use for every k gives (ii). □

To avoid the convergence of the $\Gamma(\theta) \rightarrow 0$ for long trajectories, one can do restarts.

3.2.2 Algorithms: LSTD-Learning with restarts

For a given $d \times d$ matrix $W > 0$, integers N, M , and observed samples

$$\{x^i(k), u^i(k) \mid 0 \leq k \leq N, 1 \leq i \leq M\}$$

with user defined initial conditions

$$\{x^i(0) \mid 1 \leq i \leq M\}$$

and with action

$$u^i(k) = \tilde{\phi}(x^i(k), \xi^i(k))$$

the approximation $Q_N^{\theta^{LSTD}} = \Psi^\top \theta_N^{LSTD}$ is obtained. Here

$$\theta_N^{LSTD} = \underset{\theta}{\operatorname{argmin}} \Gamma_N^i(\theta), \quad \Gamma_N(\theta) = \frac{1}{M} \sum_{i=1}^M \Gamma_N^i$$

and

$$\Gamma_N^i(\theta) = \theta^\top W \theta + \sum_{i=1}^{N-1} [\gamma_k^i - \xi \tilde{\gamma}_{k+1}^\top \theta]$$

Remark. The LSTD algorithm can be formulated as a recursive algorithm

$$\theta_{N+1} = \theta_N + G_N \tilde{\gamma}_{N+1} (\gamma_N - \tilde{\gamma}_{N+1} \theta_N)$$

where

$$G_{N+1} = G_N - \frac{1}{K_{N+1}} G_N \tilde{\gamma}_{N+1} \tilde{\gamma}_{N+1}^T G_N$$

$$K_{N+1} = 1 + \tilde{\gamma}_{N+1}^T G_N \tilde{\gamma}_{N+1}$$

3.2.3 Galerkin relaxation

Basis $\{\psi_i\}$, $h^\theta(z) = \sum_{i=1}^d \theta_i \psi_i(z)$, we want $0 \stackrel{!}{=} \nabla_\theta \Gamma(h^\theta)$.
For Bellmann error

$$0 = \frac{1}{N} \sum_{k=1}^N D_k(h^\theta(z_k), h^\theta(z_{k+1})) \zeta^\theta(k)$$

$$\zeta^\theta(k) = \nabla_\theta D_k(h^\theta(z_k), h^\theta(z_{k+1}))$$

Alternative is so-called Galerkin-relaxation, We construct a sequence $\{\zeta(k)\}$, $\zeta(k) \in \mathbb{R}^{d_\zeta}$

$$0 = \frac{1}{N} \sum_{k=1}^N D_k(h^\theta(z_k), h^\theta(z_{k+1})) \zeta_i(k) \quad 1 \leq i \leq d_\zeta$$

We relax $D_k(h^\theta(z_k), h^\theta(z_{k+1})) = 0 \quad \forall k$

$\{\zeta(k)\}$ are called eligibility vectors in RL.

$\zeta(k)$ does not depend on θ , $\zeta(k) \neq \zeta^\theta(k)$, maybe $\zeta(k) \approx \zeta^\theta(k)$, $\theta \in \text{region of interest}$. It can make sense to have $d_\zeta = d$, if $\theta \in \mathbb{R}^d$.

One can introduce them in at least one other way...

3.3 Projected Bellman equation

Consider $h^* = T(h^*)$.

Reminder $Q^n(x, u) = c(x, u) + Q^n(x^+, u^+)$, where $x^+ = F(x, u)$, $u^+ = \phi(x^+)$. In our notation $Q^\theta(x, u)$:

$$T(h)|_{(x,u)} = c(x, u) + h(x^+, u^+),$$

so $Q^\theta = T(Q^\theta)$. Consider an approximation in a function class \mathcal{H} .

$$\hat{h} = \hat{T}(\hat{h}) = P_{\mathcal{H}}(T(\hat{h})) \quad (3)$$

with $P_{\mathcal{H}}(h) \in \mathcal{H}$ for $h \in \mathcal{H}$.

Or, consider a second function class \mathcal{G} and solve for $\hat{h} \in \mathcal{H}$:

$$0 = P_{\mathcal{G}}(\hat{h} - T(\hat{h})) \quad (4)$$

Proposition 48. Suppose that the following hold

- (i) $\mathcal{H} = G$
- (ii) \mathcal{H} is a linear function class, i.e. $a_1 h_1 + a_2 h_2 \in \mathcal{H}$ for $h_1, h_2 \in \mathcal{H}$, $a_1, a_2 \in \mathbb{R}$
- (iii) The mapping $P_{\mathcal{H}}$ is linear. For $h_1, h_2 \in \mathcal{H}$, $a_1, a_2 \in \mathbb{R}$:

$$P_{\mathcal{H}}(a_1 h_1 + a_2 h_2) = a_1 P_{\mathcal{H}}(h_1) + a_2 P_{\mathcal{H}}(h_2)$$

Then the solution to (??) and (??) coincide.

Proof. Trivial □

We assume for $g \in G$: $g : Z \rightarrow \mathbb{R}$, and G is a linear function class. We further assume there is a state-process Φ on Z , where $(x(k), u(k), \xi(k)) = w(\Phi(k))$, where w is Lipschitz. We define for a probability measure ω with density ρ

$$\langle h_1, h_2 \rangle_\omega = \mathbb{E}_\omega(h_1(\Phi), h_2(\Phi)) = \int_Z h_1(z) h_2(z) \rho(z) dz$$

$$\|h\|_\omega = \sqrt{\langle h, h \rangle_\omega}.$$

$$L_2(\omega) = \{h \mid \|h\|_\omega < \infty\}.$$

For any $h \in L_2(\omega)$, we define projection onto G as

$$\hat{h} = P_G(h) = \underset{g \in G}{\operatorname{argmin}} \{\|g - h\|_\omega\}.$$

For $\hat{h} \in G$

$$\langle h - \hat{h}, g \rangle_\omega = 0, \quad g \in G$$

In particular, we assume that G has finite dimension. We choose d functions

$$\{\zeta_i \mid 1 \leq i \leq d\}$$

stack them to get $\zeta : Z \rightarrow \mathbb{R}^d$ and define $G\{g = \theta^\top \zeta \mid \theta \in \mathbb{R}^d\}$. $\zeta(k) := \zeta(\Phi(k))$ is the sequence of **eligibility vectors**.

We do not assume that d is the dimension of G in general

Proposition 49. Suppose that $\zeta_i \in L_2(\omega)$ for each i and that the functions are linear independent in L_2^ω . That is $\|\zeta^\top \zeta\|_\omega = 0$. For each $h \in L_2(\omega)$, the projection exists, is unique, and given by

$$\hat{h} = (\omega^\star)^\top \zeta$$

with $\theta^\star = [R^\zeta]^{-1} \bar{\psi}^h$, $\bar{\psi}^h \in \mathbb{R}^d$, $\bar{\psi}_i^h = \langle \zeta_i, h \rangle_\omega$ *Lemma* $\in \mathbb{R}^{d \times d}$, $R_{ij}^\zeta = \langle \zeta_i, \zeta_j \rangle_\omega$.

Sketch. The orthogonality principle gives

$$\langle h - \hat{h}, \zeta_i \rangle_\omega = 0$$

we use this identity with $\hat{h} = (\theta^\star)^\top \zeta$ □

Proposition 50. $0 = P_G(\hat{h} - T(\hat{h}))$ holds if and only if

$$0 = \langle \zeta_i, \hat{h} - T(\hat{h}) \rangle_\omega \quad 1 \leq i \leq d.$$

This is the **Galerkin relaxation** of $h^\star = T(h^\star)$ in the $L_2(\omega)$ setting.

Consider $\mathcal{H} = \{h = \theta^\top \psi \mid \theta \in \mathbb{R}^d\}$, where $\psi : X \times U \rightarrow \mathbb{R}$. Now, we use the above on the Bellman operator.

$$0 = \mathbb{E}(\zeta_i(k)(\hat{h}(x(k), u(k)) - [c(x(k), u(k)) + \hat{h}(x(k+1), \phi(x(k+1)))]))$$

Solutions of this root finding problem define $Q^{\theta^\star} \in \mathcal{H}$.

Recall D_{k+1} , we can write equivalently

$$0 = \mathbb{E}(\zeta(k) D_{k+1}(Q^\theta))|_{\theta=\theta^\star}.$$

Given N observations, we approximate this by

$$0 = \frac{1}{N} \sum_{k=0}^{N-1} \zeta(k) D_{k+1}(Q^\theta)|_{\theta=\theta^\star}.$$

3.3.1 Algorithm: TD(λ)

Notation: $\psi_{(k)} = \psi(x(k), u(k))$, $c(k) = c(x(k), u(k))$, $\zeta_k = \zeta(k)$

For a given $\lambda \in [0, 1]$, nonnegative step size sequence $\{\alpha_n\}$, initial conditions θ_0, ζ_0 and observed samples $\{x(k), u(k) \mid 0 \leq k \leq N\}$, the sequence of estimates is defined by three coupled equations

$$\begin{aligned}\theta_{n+1} &= \theta_n + \alpha_{n+1} D_{n+1} \zeta_n \\ D_{n+1} &= -Q^{\theta_n}(x(n), u(n)) + c_n - Q^{\theta_n}(x(n+1), \phi(x(n+1))) \\ \zeta_{n+1} &= \lambda \zeta_n + \psi_{(n+1)}\end{aligned}$$

This defines the approximation of the Q-function $Q^{\theta_n} = \sum_{i=1}^d (\theta_n)_i \phi_i$.

We extend the state process

$$\Phi(k) = (x(k), u(k), \xi(k), \zeta(k)).$$

This means that $\zeta(k)$ is a linear function of the state process $\Phi(k)$.

Denote $\bar{f}_\lambda(\theta) = \mathbb{E}_\omega [\zeta(k) D_{k+1}(Q^\theta)]$. TD(λ) is an approximation of the ODE

$$\frac{d}{dt} \vartheta = \bar{f}_\lambda(\vartheta) \quad (5)$$

$\bar{f}_\lambda(\theta) = A(\theta - \theta^*)$, where $A = \mathbb{E}_\omega [\zeta(k) [-\psi_{(k)} + \psi(x(k+1), \phi(k+1))]^\top]$.

For linear systems QSV-assumptions can be shown if all eigenvalues of the systemmatrix have strictly negative real parts, i.e. A is Hurwitz.

This can be shown for the on-policy approach, so the algorithm converges. There is a counter example in the book if we are off-policy., so convergence of TD(λ) is not guaranteed in the off-policy setting.

3.3.2 Algorithm TD(λ)-learning with nonlinear function approximation

In the setup as before in TD(λ), θ_{n+1}, D_{n+1} are as for the linear case.

$$\begin{aligned}\zeta_{n+1} &= \lambda \zeta_n + \zeta_{n+1}^0 \\ \zeta_{n+1}^0 &= \nabla_\theta Q^\theta(x(n), u(n))|_{\theta=\theta_n}\end{aligned}$$

Observe that $\zeta_n^0 = \Psi_{(n)}$ for a linear function class, so this is a consistent generalization.

To extend, we use instead of the so far fixed policy ϕ

$$\phi_{(n)}^\theta = \underset{u}{\operatorname{argmin}} Q^\theta(x, u)$$

3.3.3 Algorithm: Q-learning

The change in comparison to TD(λ) is

$$D_{n+1}(Q^{\theta_n}) = -Q^{\theta_n}(x(n), u(n)) + c(k) - Q^{\theta_n}(x(n+1), \phi^{\theta_{n+1}})$$

A limit θ^* will save $\bar{f}(\theta^*) = 0$ with

$$\bar{f}(\theta) = \mathbb{E}_\omega [\zeta(k), D_{k+1}(Q^\theta)].$$

At first glance this looks as for TD(λ), but the last term of the update to D_{n+1} is different! For $\lambda = 0$, we can apply proposition ?? to conclude that Q^{θ^*} solves

$$Q^{\theta^*} = P_{\mathcal{H}}(T(Q^{\theta^*})),$$

where $T(Q)|_{(x,u)} = c(x, u) + \min_{u^+} Q(x^+, u^+)$ and $x^+ = F(x, u)$.

Theory for existence of a solution or stability (in the sense of global asymptotic stability) is so far lacking in the context of ODE analysis.

This was introduced differently, maybe we v also see this later. In a there are three views

3.4 Deep Q-Networks and Batch methods

Instead of the purely recursive form going over all N , we break this into batches
 $T_0 = 0 < T_1 < T_2 < T_B = N$

Start of lecture 16
 (17.06.2025)

3.4.1 Algorithm: DQN

With $\theta_0 \in \mathbb{R}^d$ given, and a sequence of positive scalars $\{\alpha_n\}$ we Define

$$\theta_{n+1} = \underset{\theta}{\operatorname{argmin}} \Gamma_n^\epsilon(\theta) + \frac{1}{\alpha_{n+1}} \|\theta - \theta_n\|^2, \quad (6)$$

$$0 \leq n \leq B-1 \quad (7)$$

where for each n , $r_n = T_{n+1} - T_n$

$$\Gamma_n^\epsilon(\theta) = \frac{1}{2} \frac{1}{r_n} \sum_{k=T_n}^{T_{n+1}} \left[-Q^\theta(x(k), u(k)) + c_k + \underline{Q}^{\theta_n}(x(k+1)) \right]^2$$

where $\underline{Q}^{\theta_n}(x) := Q^{\theta_n}(x, \phi^{\theta_n}(x))$.

We collect some natural properties which hold for linear and nonlinear scenarios.

Proposition 51. Suppose that $\{Q^\theta(x, u) \mid \theta \in \mathbb{R}^d\}$ is continuously differentiable in θ for each x, u . Then

1. The solution to ?? solves the fixed point equation

$$\theta_{n+1} = \theta_n + \alpha_{n+1} \frac{1}{r_n} \sum_{k=T_n}^{T_{n+1}} \left[-Q^\theta(x(k), u(k)) + \gamma_n(k) \right] + \nabla_\theta Q^\theta(x(k), u(k))|_{\theta=\theta_{n+1}}$$

with $\gamma_n(k) = c_k + \underline{Q}^{\theta_n}(x(k+1))$

2. if the parametrization is linear, so that

$$\nabla_\theta Q^\theta(x(k), u(k)) = \Psi_{(k)},$$

then

$$\theta_{n+1} = \theta_n + \alpha_{n+1} [A_n \theta_{n+1} - b_n]$$

with $A_n = -\frac{1}{r_n} \sum_{k=T_n}^{T_{n+1}-1} \Psi_{(k)} \Psi_{(k)}^\top$, $b_n = -\frac{1}{r_n} \sum_{k=T_n}^{T_{n+1}-1} \gamma_n(k) \Psi_{(k)}$.

In this case we can rearrange and invert

$$\theta_{n+1} = [I - \alpha_{n+1} A_n]^{-1} (\theta_n - \alpha_{n+1} b_n).$$

For α small enough, we can observe that

$$[I - \alpha_{n+1} A_n]^{-1} \approx I + \alpha_{n+1} A_n$$

which gives

$$\begin{aligned} \theta_{n+1} &\approx [I + \alpha_{n+1} A_n] (\theta_n - \alpha_{n+1} b_n) \\ &\approx \theta_n + \alpha_{n+1} (A_n \theta_n - b_n) \end{aligned}$$

Similarly, we aim for an approximation in the nonlinear case. For $Q^\theta \in C^1$, we have $\|\theta_{n+1} - \theta_n\| \leq K \alpha_{n+1}$ for some fixed $K < \infty$, whenever $\{\theta_n\}$ is bounded. Consequently,

$$\theta_{n+1} = \theta_n + \alpha_{n+1} \frac{1}{r_n} \sum_{k=T_n}^{T_{n+1}-1} \left[-Q^\theta(x(k), u(k)) + \gamma_n(k) + \epsilon_{n+1} \right] + \nabla_\theta Q^{\theta_n}(x(k), u(k)),$$

where $\|\epsilon_{n+1}\| = O(\alpha_{n+1})$.

3.4.2 Algorithm: Batch $Q(0)$ learning

With $\theta_0 \in \mathbb{R}^d$ given, along with $\{\alpha_n\}$, $\alpha_n > 0$ define recursively:

$$\begin{aligned}\theta_{n+1} &= \theta_n + \alpha_{n+1} \frac{1}{r_n} \sum_{k=T_n}^{T_{n+1}-1} D_{k+1}(\theta_n) \nabla_{\theta} Q^{\theta_n}(x(k), u(k)) \\ D_{n+1}(\theta_n) &= -Q^{\theta_n}(x(k), u(k)) + c_k - \underline{Q}^{\theta_n}(x(k+1))\end{aligned}$$

Proposition 52. Consider the DQN algorithm with a possibly nonlinear function approximation. Assuming $Q^{\theta} \in C^1$ and that its gradient is Lipschitz globally with constant independent of (x, u) . Suppose that $B = \infty$, that the nonnegative $\{\alpha_n\}$ satisfy $\sum \alpha_n = \infty$, $\sum \alpha_n^2 < \infty$ and suppose that the $\{\theta_n\}$ obtained by our algorithm converge to a $\theta_{\infty} \in \mathbb{R}^d$.

1. $\bar{f}(\theta_{\infty}) = 0$ with \bar{f} as before:

$$\bar{f} = \mathbb{E}_{\omega} [\zeta(k) D_{k+1}(\theta)]$$

$$\text{and } \zeta(n) = \nabla_{\theta} Q^{\theta}(x(k), u(k))|_{\theta=\theta_n}$$

2. The algorithm admits the ODE approximation

$$\frac{d}{dt} \vartheta_t = \bar{f}(\vartheta_t).$$

Note, the states if we have convergence, then the behavior is consistent with the ODE view. Generally, we do not know if \bar{f} as defined above has a root. Even if we would know, the existence of $\bar{f}(\theta_{\infty}) = 0$ does not tell us if the ODE is stable, nor if θ_{∞} has desirable properties.

3.4.3 $GQ(\lambda)$ -Learning

Instead of aiming for $\bar{f}(\theta^*) = 0$, aim to minimize

$$\min_{\theta} \Gamma(\theta) = \min_{\theta} \frac{1}{2} \bar{f}^{\top} M \bar{f}(\theta)$$

for some $d \times d$ matrix M spd.

$$\frac{d}{dt} \vartheta_t = - [\partial_{\theta} \bar{f}(\vartheta_t)^{\top} M \bar{f}(\vartheta_t)]$$

choosing $M = \mathbb{E} [\zeta_n \zeta_n^{\top}]^{-1}$, one can derive the **DQ(λ)-Learning** algorithm.

To avoid matrix inversion, one can use a two-time scale approach:

Obtain first an ODE approximation of $M \bar{f} \vartheta_t$ using

$$\frac{d}{dt} w_t = b_t [\bar{f}(\vartheta) - R w_t]$$

where $R = M^{-1}$.

Provided $\{b_t\}$ chosen very large, and ϑ_t is bounded, one can derive that $w_t \approx M \bar{f}(\vartheta_t)$ after some t .

The G prob. stands for generalized

This b is not the same as earlier, here it is a scalar

3.4.4 Algorithm: $GQ(\lambda)$ Learning for linear function approximation

With the same starting point of $Q(\lambda)$ and an additional initialization w_0 , we iterate:

$$\begin{aligned}\theta_{n+1} &= \theta_n - \alpha_{n+1} A_{n+1}^{\top} w_n \\ w_{n+1} &= w_n + b_{n+1} (f_{n+1}(\theta_n) - \zeta_{n+1} \zeta_{n+1}^{\top} w_n) \\ \zeta_{n+1} &= \lambda \zeta_n + \Psi_{(n+1)} \\ D_{n+1} &= -Q^{\theta_n}(x(n), u(n)) + c_n - \underline{Q}^{\theta_n}(x(k+1)) \\ f_{n+1}(\theta_n) &= D_{n+1} \zeta_{n+1}, \quad A_{n+1} = \partial_{\theta} f_{n+1}(\theta_n) = \zeta_n (-\Psi_n \bar{\Psi}_{n+1})^{\top} \\ \Psi_{(n+1)} &= \Psi(x(n+1), u(n+1)), \quad \bar{\Psi}_{(n+1)} = \Psi(x(n+1), \phi^{\theta_n}(x(n+1)))\end{aligned}$$

The approximation is successful if

$$\lim_{n \rightarrow \infty} \frac{b_n}{\alpha_n} = \infty$$

Problems

- Γ is not convex, so difficult to get global minima
- Even if $\bar{f}(\theta^*) = 0$ does have a solution, there are numerical challenges. Consider

$$\Gamma(\theta) = \Gamma(\theta^*) + \underbrace{0}_{\bar{f}(\theta^*)=0} + (\theta - \theta^*) \left[A^{*\top} \right] (\theta - \theta^*)$$

if A^* has a large condition number the observed condition number is squared, so even worse. Maybe M can be chosen to avoid this.

- It is not obvious why minimizing $\Gamma(\theta)$ is a reasonable goal

Journal

- **Lecture 01:** Covering: Introduction, (linear, continuous) State space models, equilibrium, (Lyapunov, asymptotically) stable, region of attraction, globally asymptotically stable .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 02:** Covering: Lyapunov function, inf-compactness and coerciveness, sublevel sets, Poisson’s inequality, comparison theorem, a few propositions connecting the value function, equilibria and Lyapunov functions .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 03:** Covering: discrete time Lyapunov equation, optimal control policy, controllability, linear quadratic regulator, Bellmann equation, principle of optimality, Q-function and some concepts from Reinforcement Learning .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 04:** Covering: Value iteration, policy iteration, exploration-exploitation .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 05:** Covering: Approximate Q-functions, Bandits, discounted cost, shortest path, finite horizon and translations between them .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 06:** Covering: Model predictive control, continuous time formulations of previous results .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 07:** Covering: Picard-Iteration, Grönwall-Bellman inequality, Euler’s method, gradient flows .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 08:** Covering: Polyak-Lojasiewicz inequality, L-smooth inequality, Bregman divergence, quasi stochastic approximation .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 09:** Covering: QSA continued, approximate policy improvement .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 10:** Covering: QSA1-QSA3, some convergence results .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 11:** Covering: Boundedness implies convergence, ultimate boundedness, first entrance times, QSV assumption .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 12:** Covering: Using QSV to show ODE solutions are ultimately bounded, Gradient free optimization: QSGD1, QSDG3 .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 13:** Covering: Global consistency, very short crash course in ML, reinforcement learning, least squares temporal difference learning .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages

- **Lecture 14:** Covering: Redundant Parametrization, Galerkin relaxation, projected bellman equation .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 15:** Covering: Eligibility vectors, Galerkin relaxation in the L_2 setting, $TD(\lambda)$, $TD(\lambda)$ with non-linear function approximation, Q-learning .
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages
- **Lecture 16:** Covering:
.
Starting in ‘??’ on page ?? and ending in ‘??’ on page ?? . Spanning 0 pages

Bibliography

- [1] Tamer Basar, Sean Meyn, and William R. Perkins. *Lecture Notes on Control System Theory and Design*. 2024. arXiv: [2007.01367](https://arxiv.org/abs/2007.01367) [math.OG]. URL: <https://arxiv.org/abs/2007.01367>.
- [2] Sean Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.