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

University of Bonn
Summer semester 2025
Last update: July 11, 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

Start of lecture 01 (10.4.2025)

- 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)

The main source for this course is [4]. 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. $\phi : Y \to U$ 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 $r(k), k \geq 0$.

t is continous, k is step by step / iterative

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

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:

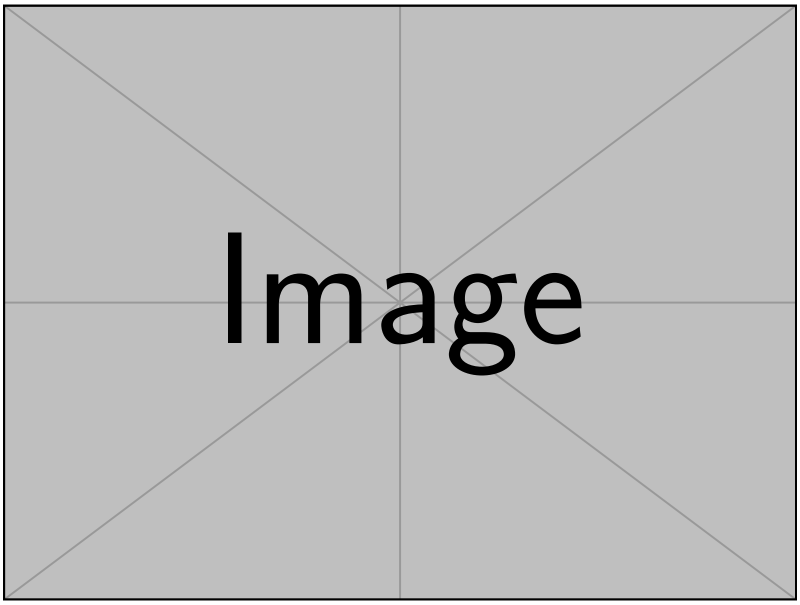


Figure 1.1: Sketch 1.01

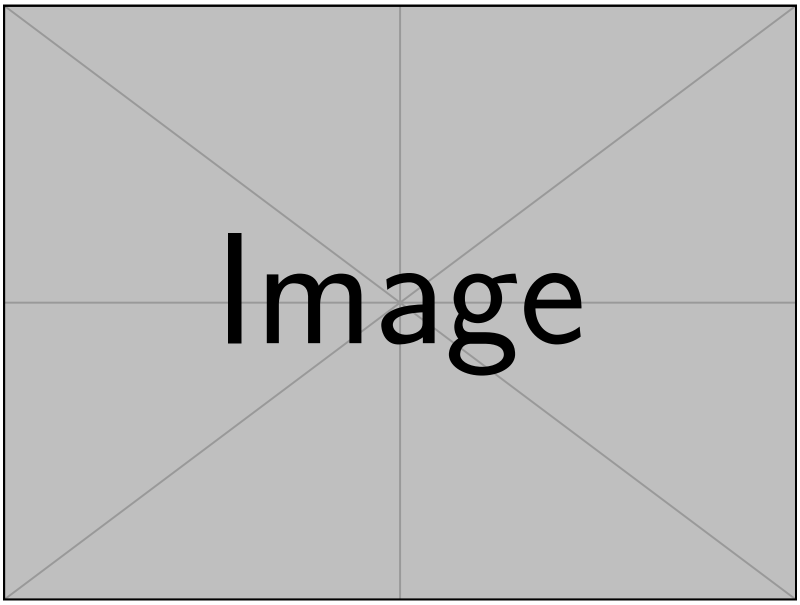


Figure 1.2: Sketch 1.02: Mountain car

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)

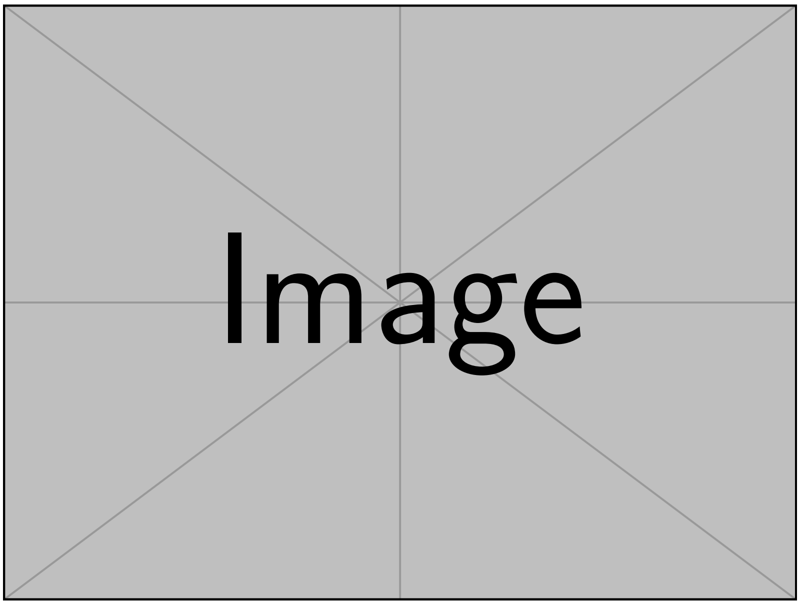


Figure 1.3: Sketch 1.03: cart pole / inverted pendulum

Next example: Acrobot (more then 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*

$$x(k+1) = \mathcal{F}(x(k), u(k)) \tag{1}$$
$$y(k) = \mathcal{C}(x(k), u(k)) \tag{2}$$

$x(k)$ might include the past, might be useful for the stock trading problem

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

1.2 Linear State Space Model

$$x(k+1) = Fx(k) + Gu(k) \tag{3}$$
$$y(k) = Cx(k) + Du(k) \tag{4}$$

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

Common scenario for (3) 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 (3) yields

$$\begin{aligned} x(k+1) &= (F - GK)x(k) - Gv(k) \\ y(k) &= (C - DK)x(k) + Dv(k) \end{aligned}$$

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

$$\begin{aligned} \frac{d}{dt}x &= Ax + Bu \\ y &= Cx + Du \end{aligned}$$

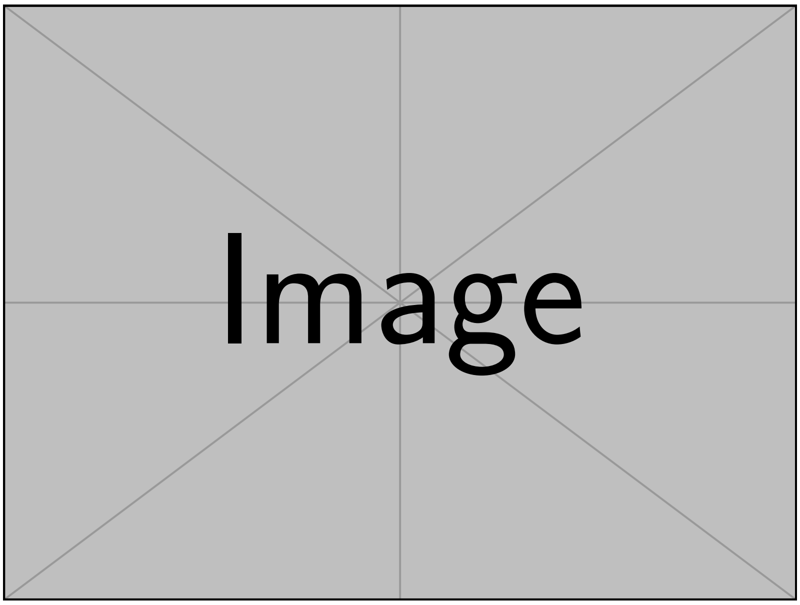


Figure 1.4: Sketch 1.04

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 (1) $\mathcal{F}(x,u) = x + \Delta f(x,u)$. 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*** *$C : X \to \mathbb{R}_+$ and a policy ϕ we define*

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

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

The same concept with a different sign comes up in RL under the term ***reward***

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*

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

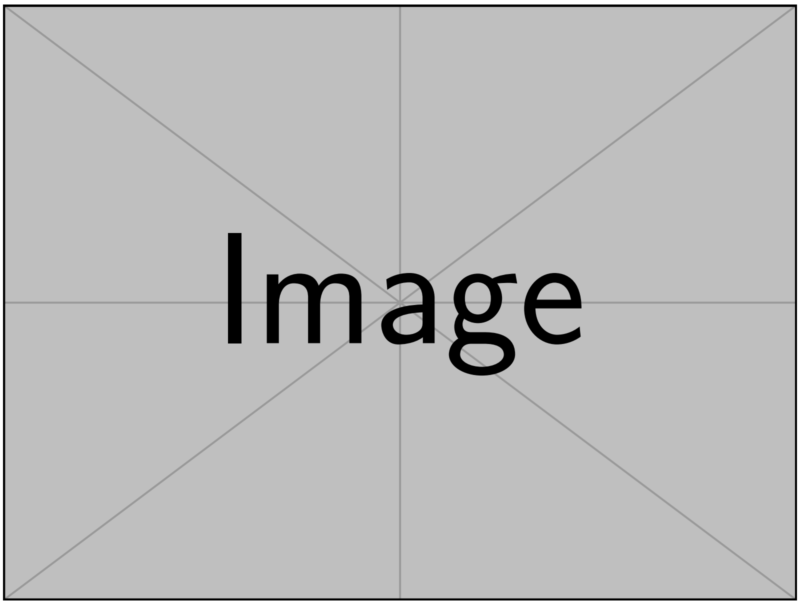


Figure 1.5: Sketch about Lyapunov stability

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*

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

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.*

Start of lecture 02 (15.04.2025)

Definition 7 (Lyapunov function)**.** *A function $V : X \to \mathbb{R}_+$ 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:

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

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

- empty
- the whole domain X — We usually want to avoid this
- a bounded subset of X.

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

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

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*

Lemma 8. *Suppose that the cost function C and the value function J from definition 5 are non-negative and finite valued.*

this is a assumption on the value function

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

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

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

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

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

We are separating one step!
This is the same Bellman from the curse of dimensionality!

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

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

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

$$\implies \text{ non-increasing}$$

Taking the limit

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

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 $J(x^\infty) = \lim_{i\to\infty} J(x(k_i)) = 0$ 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 $V, c : X \to \mathbb{R}_+$ and $\eta \geq 0$. Then* ***Poisson's inequality*** *states that*

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

We often assume $\eta = 0$

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.

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

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*

We don't write that explictlly, but we don't start in x^e!

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 $\sum_{k=0}^{N-1} c(x(k)) \leq V(x(0)) - V(x(k)) \leq V(x(0))$ where the LHS converges to $J(x(0))$ for $N \to \infty$
3. Show sitsoL, with $\eta = 0$ it follows form definition 9 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

This is important!

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

So it is a unique minimum. V has therefore the properties of proposition 10, which gives sitsoL. For global: with 1. we get

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

and assumptions give us by lemma 8 that $x(k) \to x^e$ as $k \to \infty$. 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:

$$V(x(N)) + \overbrace{\sum_{k=0}^{N-1} c(x(k))}^{J(x(N-1)) \overset{N \to \infty}{\to} J(x)} = V(x).$$

Lemma 8 together with the continuity of V implies that

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

This gives

$$V(x^e) + J(x) = V(x)$$

□

Start of lecture 03 (17.04.2025)

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 $c(x) = x^\intercal Sx$, where S is symmetric and positive definite. Observe

$$c(x(k)) = (F^k x)^\intercal SF^k x$$

Summing up yields

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

This satisfies a linear fixed point equation:

$$M = S + F^\intercal MF \tag{5}$$

This is also called ***discrete time Lyapunov equation***

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. *(5) admits a solution M positive semi-definite for any S positive semidefinite.*

Reference: [1]

Consider 1 without y

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

with

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

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

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

This describes the optimal control policy (OCP)

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.

and the correnspending policy

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 $x^\star$ achieving the minimum

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

which is definition 9 with $\eta = 0$ and equality.
Proposition 11 implies, under some conditions, that x^e is globally asymptotically stable.
Under the following assumptions $J^\star$ 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 $\underline{u}$.

This is sometimes called **controllability**

Example (Linear Quadratic Regulator)**.** *Consider linear dynamics 3 from the first lecture with quadratic cost $c(x, u) = x^\intercal S x + u^\intercal R u$ with S positive semi-definite and R positive definite.*
Reminder: $u = -Kx$.
If there is a policy for which $J^\star$ is finite, then

$$J^\star(x) = x^\intercal M^\star x$$

with $M^\star$ positive semi-definite and

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

with $K^\star$ depends on $M^\star, R, F, G$.

and implicitly on c

Bellmann equation

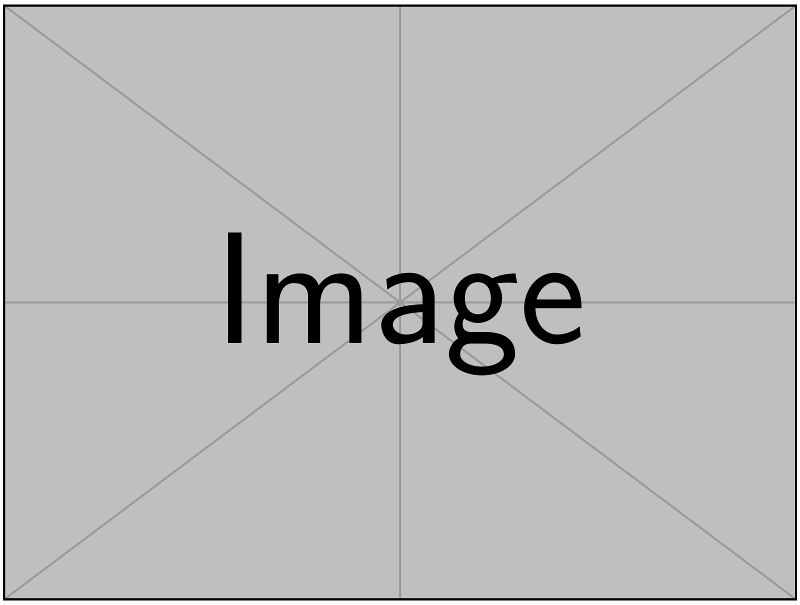


Figure 1.6: Sketch 1.06; Principle of optimality

Observation:

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

This gives

which can be seen as a kind of fix point equation

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

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

Theorem 13 (Bellmann equation, Dynamic Programming equation). *Assume that $J^\star$ is finite and optimal control $u^\star$ solving (6) exists. Then the value function satisfies*

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

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

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

Definition 14 (Q-function). *The function of two variables within the minimum in (7) is called Q-function.*

Definition, which is not so useful for the analysis, but for the pratical application!

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

In the optimal case we write $Q^\star$. Thus

$$J^\star(x) = \min_{\bar{u}} Q^\star(x,\bar{u}).$$

The optimal feedback law is then

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

The Q-function solves the fixed point equation

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

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 $\tilde{J}_\theta$. Aim for the minimum

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

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

scalar product in $\mathbb{R}^n$ (all 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

What changes, or what is the information gain

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

After N samples

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

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 $J^\star$ by a sequence of V^k given an initial value function V^0.

For infinite state spaces we will have to fix this algorithm for memory related reasons

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

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 $J^\star$ is finite valued.*

Then there is $n_0 \geq 1$ such that

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

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

We really exploit the finiteness!

$$(x^\star(k), u^\star(k)) = (x^e, u^e)$$

for $k \geq n_0$. This exists since $J^\star$ is finite.
Using the principle of optimality (6) we can show

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

This gives

$$\begin{aligned} V^n(x) &\leq \sum_{k=0}^{n-1} c(x(k), u(k)) + V^0(x(n)) \;\; \text{for all } u \text{ including } u(k) = \phi^\star(k) \\ &\overset{n \geq n_0}{=} J^\star(x) + V^0(x^e) = J^\star(x) \end{aligned}$$

For such n, the inequality must be an equality, due to (8) 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(9) 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

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

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

Then a corresponding bound holds for each n

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

where η_i is non-increasing:

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

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

This is (connected to?) the Bellman error

$$\eta_n \coloneqq \sup_x B^n(x).$$

Value iteration gives

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

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

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

which gives $B^0(x)\leq\eta\forall x\implies\eta_0\leq\eta$.
For $n\geq1$ 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 11 it follows

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

where $J^\star$ 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

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

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

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

- 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

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

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

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

end while

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

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

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 11. □

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 $x(k) \to x^e$ reasonably rapidly, where we assume $c(x^e, \cdot) = 0$. Typically we have continuity

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

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

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

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

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

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

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

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

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.

Start of lecture 05 (24.04.2025)

1.7 Linear Quadratic Regulator, Revisited

We had $J^\star(x) = x^\intercal M^\star x$ and quadratic costs, $c(x,u) = x^\intercal S x + u^\intercal R u$.
For the Q-function:

$$Q^\star(x,u) = c(x,u) + J^\star(Fx + Gu).$$

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

$$0 = \nabla_u Q^\star(x, u^\star) = 2Ru^\star + 2G^\intercal M^\star(Fx + Gu^\star)$$

Assuming R is positive definite; then $R + G^\intercal MG$ is positive definite and therefore invertible.

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

and

$$\phi^\star(x) = -Kx.$$

To obtain $M^\star$ we can solve a fixed point equation called the **algebraic Riccati equation**

This is a hint, we will prob. revisit this later

$$M^\star = F^\intercal \left(M^\star - M^\star G \left[R + G^\intercal M^\star G \right]^{-1} G^\intercal M^\star F + S \right) \tag{9}$$

1.8 Approximate Q-functions

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

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

where $\psi_i : X \times U \to \mathbb{R},\ 1 \le i \le d$ 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:

Think kernels, finite element basis,...

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 $\phi^{n+1}(x) \coloneqq \phi^{\theta^n}$

Approximation since we do this sample-based in RL

As an alternative, consider dynamic programming equation from definition 14:

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

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

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

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 $Q^\theta = Q^\star$ 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:

$$J^\star(x) = \min_{\underline{\mathrm{u}}} \gamma^k c(x(k), u(k)),\ x(0) \in X$$

where $\gamma \in (0,1)$ is the **discount factor**.
Shortest Path Problem: Given $A \subset X$ define $\tau_A \coloneqq \min\{k \geq 1 \mid x(k) \in A\}$.

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

This is problematic, since we might have longer path with lower cost ...

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

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

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

Proof.

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

$c(x, u(0))$ since we're extracting the first element of the sum

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

To formulate this as a discounted problem

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

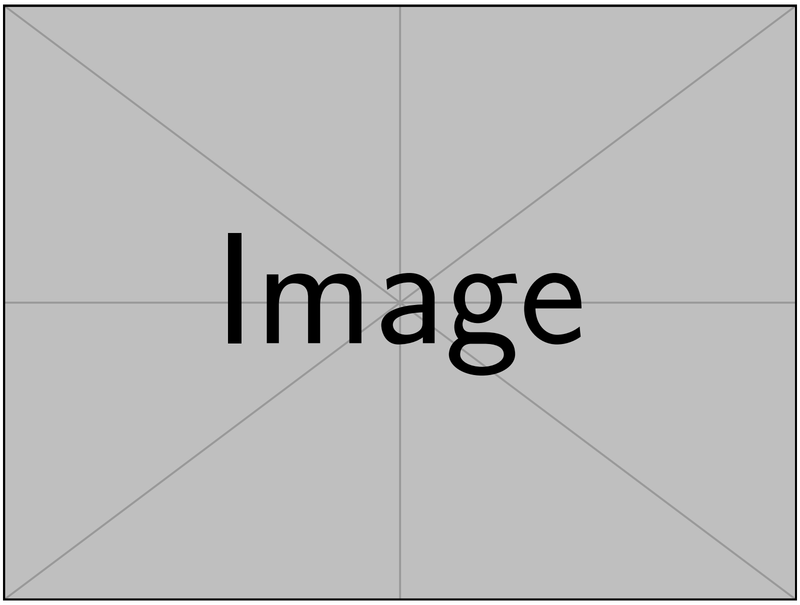


Figure 1.7: Sketch: mountain car value function

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

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

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

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

The Bellmann equation from theorem 13 now becomes

$$J^\star(x,\tau) = \min_u \left\{ c(x,u) 1_{\{\tau \leq N\}} + J^\star(\mathcal{F}(x,u), \tau+1) \right\} \tag{10}$$

For $\tau > N$, it follows that $J^\star(x,\tau) = 0$. This gives kind of a boundary condition

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

So,

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

repeating this backwards in time yields

$$J^\star(x, 0) = J^\star(x^a).$$

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

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

Start of lecture 06 (29.04.2025)

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

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

where $\phi^\star$ from the finite horizon setting (10) for *small* N.
Consider

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

Due to the finite horizon we are not optimal ...

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

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

where $V^0 : X \to \mathbb{R}^+$ 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 15:

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

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

This is also a version of a poisson inequality

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

From the Comparison theorem 11, 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 5, 6 carry over.

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

Definition 20. *A function $V : X \to \mathbb{R}_0^+$ 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 10 and proposition 11 (iii), we can see that these also carry over to the continuous case. So we get

Proposition 21 (Extension of prop 11 (iii))**.** *If there exists a Lyapunov function after definition V 20, then the equilibrium x^e is globally asymptotically stable.*

Since we did not exploit the step-wise nature previously

The continuous version of Poisson's inequality is then

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

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 (11) holds for non-negative c, V, η, then we have*

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

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

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

We extend the Bellmann equation to continuous times

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

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

this is a strong assumption! In principle we would need to talk about viscosity solutions ... Even weak solutions are not enough

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

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

$$0 = \min_u \left[c(x,u) + \nabla J^\star(x)\cdot f(x_0,u_0)\right] \tag{13}$$

The term to minimize has an interpretation as an Hamiltonian

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

One can show

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

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

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

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

1.13 Linear quadratic regulator revisited (once more)

$$
\begin{aligned}
\frac{d}{dt}x &= Fx + Gu,\ x(0) = x_0 \\
c(x,u) &= x^\intercal S x + u^\intercal R u
\end{aligned}
$$

everything we observed so far carries over, assuming $J^\star$ is finite, we have

$$J^\star(x) = x^\intercal M^\star x$$

the HSB (13) gives

$$
\begin{aligned}
\phi^\star(x) &= \operatorname*{argmin}_u \{x^\intercal S x + u^\intercal R u + [2M^\star x]^\intercal [F_x + Gu]\} \\
&= \operatorname*{argmin}_u \{u^\intercal R u + 2x^\intercal M^\star G u\}
\end{aligned}
$$

So,

$$0 = \nabla_u \{u^\intercal R u + 2x^\intercal M^\star G u\}_{|u=\phi^\star(x)}$$

and we get

$$\phi^\star(x) = -R^{-1}G^\intercal M^\star x$$

and

$$\frac{d}{dt}x^\star = \left[F - GR^{-1}G^\intercal M^\star\right]x^\star.$$

HSB (13) further gives

$$
\begin{aligned}
0 &= \{x^\intercal S x + u^\intercal R u + [2M^\star x]^\intercal [Fx + Gu]\}_{|u=\phi^\star(x)} \\
& x^\intercal \left\{S + M^\star G R^{-1} \mathrm{Id} G^\intercal M^\star\right\} x + x^\intercal \left\{2M^\star F + 2M^\star G R^{-1} G^\intercal M^\star\right\} x
\end{aligned}
$$

using $2x^\intercal M^\star F x = x^\intercal \left[M^\star F + F^\intercal M^\star \right]$ we get

$$= x^\intercal \left\{ S + M^\star F + F^\intercal M^\star - M^\star G R^{-1} G^\intercal M^\star \right\} x$$

$$\left\{ S + M^\star F + F^\intercal M^\star - M^\star G R^{-1} G^\intercal M^\star \right\}$$

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

$$0 = S + M^\star F + F^\intercal M^\star - M^\star G R^{-1} G^\intercal M^\star$$

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
$$\bar{f}(\theta^\star) = 0$$
- 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?
$$\theta_{n+1} = \theta_n + \alpha_{n+1}\bar{f}(\theta_n) \tag{1}$$
- Design an algorithm to approximate (1) based on whatever observation is available.

θ for descrete settings, ϑ for continuous settings. Both do the same job

θ_{n+1} is the next iterate, not the next time step!

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** $\theta^\star \in \mathbb{R}^d$, where $\theta^\star$ is an equilibrium
$$f(\theta^\star) = 0.$$

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

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

based on

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

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 (3). in the finite time horizon. Moreover, successive approximation is uniformly convergent:

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

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*

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

Then it holds

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

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

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

Proof. Both proofs can be found in any textbook on ODEs. The second one is also found in [4]. □

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

Not that nice, but at least a bound ...

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

$$\|\vartheta_t\| \leq \left(B_f + \|\vartheta_{\|}\right) e^{Lt} - B_f \tag{4}$$

$$\|\vartheta_t - \vartheta_0\| |B_f + L\|\vartheta_0\|| t e^{lt} \tag{5}$$

(ii) If there is an equilibrium $\theta^\star$*, then for each initial condition:*

$$\|\vartheta_t - \theta^\star\| \leq \|\vartheta_0 - \theta^\star\| e^{Lt} \tag{6}$$

Proof. **(ii):** use 3 to get

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

Since $f(\theta_\star) = 0$, we see

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

So

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

Using proposition 26 (ii) with $\beta_t \equiv L, \alpha_t \equiv z_0$ we get

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

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

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

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

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

using (3)

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

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

2.2 Euler's method once more

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

Explicit Euler, implicit Euler is nicer to analyze

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

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

2.3 Optimization

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

$$\theta^\star \in \operatorname{argmin} \Gamma(\theta). \tag{9}$$

Use steepest-descent, formulated as ODE

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

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

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

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 $\delta_0 > 0$

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

From numerical optimization we know:

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

(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 $\theta^\star$. Then the gradient flow*

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

is globally asymptotically stable, with unique equilibrium $\theta^\star$.
If Γ is strongly convex, then the rate of convergence is exponential

$$\|\vartheta_t - \theta^\star\| \le e^{-\delta_0 t}\|\vartheta_0 - \theta^\star\|,$$

where δ_0 comes from theorem 29.

Start of lecture 08 (08.05.2025)

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

$$\frac{d}{dt}V(\vartheta_t) = -\nabla_\theta \Gamma(\vartheta_t)^\intercal \left[\vartheta_t - \theta^\star\right]$$

By convexity we get the following bound

$$\Gamma(\theta^\star) \geq \Gamma(\vartheta_t) + \nabla_\theta \Gamma(\vartheta_t)^\intercal \left[\theta^\star - \vartheta_t\right]$$

using the support condition this becomes

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

since $\theta^\star$ is the minimum. The strict inequality (< 0) holds when $\vartheta_t \neq \theta^\star$. V fulfills definition 20 and proposition 21 gives global asymptotic stability.

Coercive, therefore inf-compact

Under strict convexity

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

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

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

Used in stochastic gradient descent

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

holds then the gradient flow satisfies for each initial ϑ_0

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

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

$$\Gamma(\theta_\infty) = \Gamma^\star$$

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

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

This implies using the same technique as in the previous proof

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

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

$$\Gamma(\theta_\infty) = \lim_{n\to\infty}\Gamma(\vartheta_{t_n}) = \Gamma^\star$$ ◻

Consider the Euler method for the gradient flow:

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

Theorem 32. *Suppose that Γ satisfies*

These are more general version of global Lipschitz and convexity

(i) the $L-$smooth inequality (LSI)

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

(ii) the PL inequality 11

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

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

Proof.

$$\begin{aligned}
\Gamma(\theta_{k+1} - \Gamma(\theta_k)) &\overset{\text{LSI}}{\leq} [\theta_{k+1} - \theta_k]^\intercal \nabla\Gamma(\theta_k) + \frac{1}{2}L\|\theta_{k+1} - \theta_k\|^2 \\
&\overset{12}{=} -\alpha\|\nabla\Gamma(\theta_k)\|^2 + \frac{1}{2}L\alpha^2\|\nabla\Gamma(\theta_k)\|^2 \\
&= (-\alpha + \frac{1}{2}L\alpha^2)\|\nabla\Gamma(\theta_k)\|^2
\end{aligned}$$

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

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

and therefore

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

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

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

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

Then

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

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

Proof. **(i)**

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

(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

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

Now integrate

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

□

Remark. *Strong convexity:*

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

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

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

This gives a bound on the loss function from both sides . . .

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

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

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

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

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

I.e. we have the Markov property

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

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

the book uses Θ instead of $\hat{\theta}$

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

a **quasistochastic approximation(QSA)**-ODE, a_t is the **step size**.
For deterministic probing signals, we mainly consider two examples

Mixture of sin functions

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

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

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

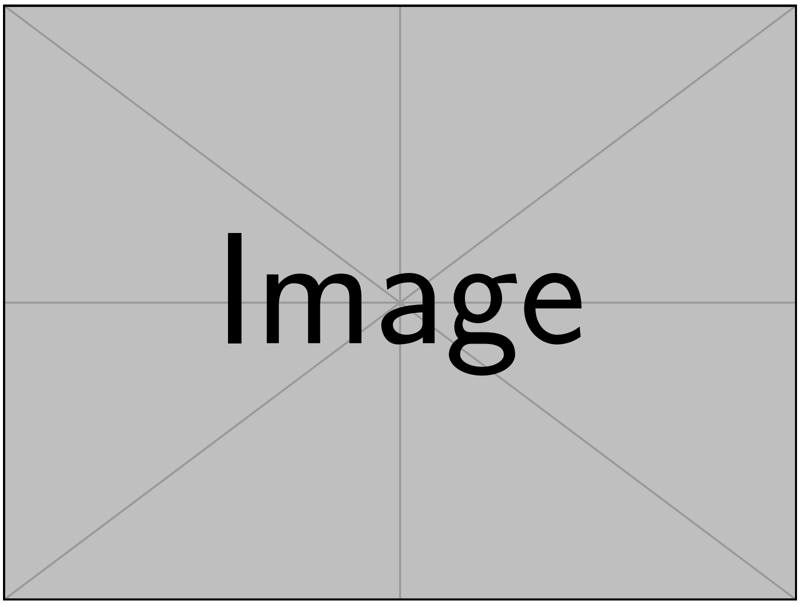


Figure 2.1: Sketch 2.01

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**

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

and **covariance**

$$\lim_{T\to\infty} \int_0^T \xi_i \xi_i^\intercal dt = \mathrm{Id}$$

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

Start of lecture 09 (13.05.2025)

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

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

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

A **QSA** approach is to use the saw tooth function

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

Obtain estimate by

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

with a reasonable discretization afterwards.
To use (QSA-ODE (14)) $f(\theta, \xi) = y(\xi) - \theta$ with mean vector field

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

which gives $\theta^\star = \int_0^1 y(\xi_t)dt$ as the unique root of $\bar{f}$. The QSA-ODE 14 is

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

(16) can be transformed into

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

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

2.5 Approximate Policy Improvement

nonlinear state model in continuous time:

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

$$J^\star(x) = \min_{\underline{\mathrm{u}}} \int_0^\infty c(x_t, u_t)dtx = x_0 \tag{19}$$

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

$$J^\phi(x) = \int_0^\infty c(x_t, \phi(x_t))dt, \; x = x_0 \tag{20}$$

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 20 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:

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

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:

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

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

$\underline{Q}$ for the fixed, but optimal choice of u

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

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

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

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

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

Choose feedback law with exploration $u_t = \tilde{\phi}(x_t, \xi_t)$. Assuming bounded state trajectories, such that (21) exists, define $\Gamma(\theta) = \frac{1}{2}\|B^\theta\|^2$, we get

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

Gradient flow

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

QSA counterpart is (21) 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

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

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*

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

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

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

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

Proof. Change of variables and observing that

$$d\tau = a_t dt.$$

□

Recall $\bar{f}(\theta) \coloneqq \lim_{T\to\infty} \int_0^T f(\theta, \xi_t)dt$ 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 35. Define $\hat{\theta}_\tau = \tilde{\theta}(s^{-1}(\tau)) = \tilde{\theta}_t \mid_{t=s^{-1}(\tau)}$. 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*

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

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

Version of proposition 27

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

Proof. Proof of proposition 27 in adapted notation. □

Now, denote by $\vartheta_w^\tau, w \geq \tau$ the unique solution to (22):

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

with that we get

quasistochastic vs continuous

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

The following assumptions will be used in the following:

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

it does not go to zero to fast

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

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

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

Is my probing covering everything: ergocity, ergodic bound

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

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

Consider first, arbitrary θ

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

There is a connection to the law of large numbers …

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

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

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

large ϵ_t in the book? Prob. $\mathcal{E}$

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

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

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

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 $T > 0$*

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

and

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

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

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

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 37, 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.

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

For any $T > 0$

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

Since $\delta > 0$ was arbitrary, we have the first statement.
For the second limit: $E^\tau_r = \vartheta^\tau_r - \hat{\theta}_r$. The pair of identities after lemma 36 give using Lipschitz condition from (QSA2) we get

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

$$\|E^\tau_{\tau+v}\| \leq \delta^\tau + L_f \int_\tau^{\tau+v} \|E^\tau_w\| dw,$$

where

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

Grönwalls lemma gives

$$\|E^\tau_{\tau+v}\| \leq e^{Lf}\delta^\tau \forall\tau,\ 0 \leq v \leq 1$$

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

Start of lecture 11 (20.05.2025)

Theorem 39 (Boundedness implies convergence). *Suppose that (QSA1-QSA3) hold. 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 (14)*

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

converges to $\theta^\star$ for each initial condition.

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

$$\|\vartheta^\tau_\tau\| = \|\hat{\theta}_t\| \overset{\text{pA}}{\leq} b,\ \tau \geq T_{\theta,z}$$

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

$$\limsup_{\tau\to\infty} \|\hat{\theta}_{\tau+T_\epsilon} - \theta^\star\| \leq \underbrace{\limsup_{\tau\to\infty} \|\hat{\theta}_{\tau+T_\epsilon} - \vartheta^\tau_{\tau+T_\epsilon}\|}_{\to 0 \text{ by proposition } 38} + \underbrace{\limsup_{\tau\to\infty} \|\vartheta^\tau_{\tau+T_\epsilon} - \theta^\star\|}_{\leq\epsilon}$$ □

Lemma 40 (Weaker form of proposition 38 (ii)). *For some $\bar{\delta} < \infty$ and any $0 \leq T \leq 1$*

$$\|\hat{\theta}_{\tau+T} - \vartheta^\tau_{\tau+T}\| \leq e^{L_f} b_0(\hat{\theta}_\tau)\epsilon^f_\tau + \bar{b}(1 + \|\hat{\theta}_\tau\|)T^2$$

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

Proof. Write $E^\tau_r = \vartheta^\tau_r - \hat{\theta}_r,\ r \geq T$. The pair of identities after lemma 36 give, after inserting $\pm\bar{f}(\theta_w)$

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

using (QSA2) we can bound

like last lecture ...

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

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

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

where

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

Using Grönwalls lemma, proposition 26 (ii)

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

Repeating the proof for proposition 27, 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**

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

Definition 41 (ultimately bounded)**.** *The ODE*

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

*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 $\vartheta_t \in S \ \forall t \geq T(\theta_0)$.*

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

Lyapunov function
If we are not in S, we are getting pointed into that direction

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

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

first entrance time T_S

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

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

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

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

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

Assumption (QSV):

QSV1 in the book

There exists a continuous function $V : \mathbb{R}^d \to \mathbb{R}$, 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 is strictly decreasing in that setting

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

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 $\|\vartheta_t\| > c_0$.

Start of lecture 12 (22.05.2025)

Lemma 43. *Assume $V : \mathbb{R}^d \to \mathbb{R}_+$ 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

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

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 $m \in Z_+$, define $\hat\tau_m = \min(\hat\tau, m)$. Then

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

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

$$\hat{\tau} \leq \frac{1}{b\delta_1} \int_{\tau}^{\tau+T} V(\hat{\theta}_w) dw$$

$$\begin{aligned} \int_{\tau}^{\tau+T} V(\hat{\theta}_w) dw &\leq \int_{\tau}^{\tau+T} |V(\hat{\theta}_w) - V(\hat{\theta}_\tau)| + |V(\hat{\theta}_\tau)| dw \\ &\leq \int_{\tau}^{\tau+T} \underbrace{L_V \|\hat{\theta}_w\|}_{\text{prop 27: } \leq (C(V) + \|\hat{\theta}_\tau\|) c(L_V, T)} + L_V \|\hat{\theta}_\tau\| + |V(\hat{\theta}_\tau)| dw \end{aligned}$$

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. □

Proposition 44. *Under (QSV), the solution to (23) is ultimately bounded, i.e. there exists some $b < \infty$ such that for any*

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

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

$$\begin{aligned} V(\hat{\theta}_{\tau+T}) - V(\hat{\theta}_\tau) &= V(\hat{\theta}_{\tau+T}) - V(\vartheta^\tau_{\tau+T}) + V(\vartheta^\tau_{\tau+T}) - \underbrace{V(\vartheta^\tau_\tau)}_{=\hat{\theta}_T} \\ &\leq |V(\hat{\theta}_{\tau+T}) - V(\vartheta^\tau_{\tau+T})| + V(\vartheta^\tau_{\tau+T}) - V(\vartheta^\tau_\tau) \\ &\leq L_V \|\hat{\theta}_{\tau+T} - \vartheta^\tau_{\tau+T}\| - \delta_0 \int_0^T \underbrace{\|\vartheta^\tau_{\tau+T}\|}_{\leq \|\hat{\theta}_\tau\| + \| \int_\tau^{\tau+T} \bar{f}(\vartheta^\tau_w) dw\|} dt \\ &\leq L_V \|\hat{\theta}_{\tau+T} - \vartheta^\tau_{\tau+T}\| - \delta_0 T \|\hat{\theta}_\tau\| \\ &\overset{\text{Lemma 40}}{\leq} L_V (e^{L_f} b_0(\hat{\theta}_\tau) \epsilon^f_\tau + \bar{b}(1 + \|\hat{\theta}_\tau\|) T^2) - \delta_0 T \|\hat{\theta}_\tau\| \end{aligned}$$

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

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

and we can use the lemma 43. □

Now we can ultimate boundedness and therefore convergence!

2.6 Gradient free Optimization

Reminder:

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

we assume it has a unique minimizer $\theta^\star$.

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

we look for $\theta^\star$ with $\bar{f}(\theta^\star) = 0$. 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

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

and construct and ODE

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

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

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \xi_t dt = 0, \; \lim_{T\to\infty} \frac{1}{T}\int_0^T \xi \cdot \xi^\intercal dt = Id$$

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
$f(\theta_t, \xi_t) = -\frac{1}{\epsilon} G\xi_t \Gamma(\theta_t + \epsilon\xi_t)$ If$\Gamma \in C^2$:

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

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

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

For $G = \mathrm{Id}$ 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 $\tilde{\theta}_0 \in \mathbb{R}^d$

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

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

$$f(\theta, \xi_t) = -G\xi_t\xi_t^\intercal \nabla\Gamma(\theta) + o(\epsilon), \; \lim_{T\to\infty} \int_0^T f(\theta, \xi_t) = -G\nabla\Gamma(\theta) + o(\epsilon)$$

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*

Control on both sides of the loss function ...

1. *(QSA1) holds*
2. *The probing signal satisfies*
$$\int_0^T \xi_t\xi_t^\intercal dt = Id$$
3. *$\nabla\Gamma$ is globally Lipschitz continuous, and Γ is strongly convex with unique minimizer $\theta^\star \in \mathbb{R}^d$*

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 $\theta^\star_\epsilon$ of $\bar{f}_\epsilon$, satisfying

$$\|\theta^\star_\epsilon - \theta^\star\| = O(\epsilon)$$

Moreover, convergence holds from each initial condition:

$$\lim_{t\to\infty} \theta_t = \theta^\star_\epsilon$$

Proof. The assumptions imply that (QSA2) holds for Exploid $\nabla\Gamma$ is convex

$$f(\theta, \xi) = -G\xi\xi^\intercal \nabla\Gamma(\theta) + O(\epsilon)$$
$$\bar{f}_\epsilon(\theta) = \lim_{T\to\infty} \frac{1}{T}\int_0^T f(\theta, \xi) dt$$

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 $\|z\| \le \epsilon_0$. From this (QSA3, the asymptotic stability condition), can be established for $\epsilon > 0$ small enough.
Theorem 39 yields that for each $\epsilon > 0, \theta_t$ converges to the unique root $\theta^\star_\epsilon$ of $\bar{f}_\epsilon$ satisfying

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

Ffrom there, strong convexity gives

$$\Gamma(\theta^\star) \ge \Gamma(\theta^\star_\epsilon) + \nabla\Gamma(\theta^\star_\epsilon)^\intercal(\theta^\star - \theta^\star_\epsilon) + \frac{\eta}{2}\|\theta^\star_\epsilon - \theta^\star\|^2$$

for some $\eta > 0$.

$$\begin{aligned}\frac{\eta}{2}\|\theta^\star_\epsilon - \theta^\star\|^2 &\le \underbrace{\Gamma(\theta^\star) - \Gamma(\theta^\star_\epsilon)}_{\le 0} + \nabla\Gamma(\theta^\star_\epsilon)^\intercal(\theta^\star - \theta^\star_\epsilon) \\ &\le \|\nabla\Gamma(\theta^\star_\epsilon)\|\|\theta^\star - \theta^\star_\epsilon\|\end{aligned}$$

which gives

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

□

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:

For more information about kernels, you can look at my lecture notes for scientific computing 2 (also held by Garcke)

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

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

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

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

We usually use **regularization** to avoid **overfitting**.
Always reserve samples for evaluating the quality of the prediction.

3.2 Reinforcement Learning

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

We have a sequence of state-action pairs

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

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

where

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

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

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

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

Since $D_{k+1}(Q^\theta)$ will be small ...

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

This looks like a regression problem:

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

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

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

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

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

is the approximation of the Q-function.
We have a positive definite quadratic objective, so the solution to (1) can be obtained by solving for $\nabla\Gamma(\theta) \stackrel{!}{=} 0$.

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

The regularization W is introduced to ensure a unique solution.

Start of lecture 14 (03.06.2025)

Proposition 47 (Redundant Parametrization)**.** *Suppose that* $R_N = \frac{1}{N}\sum_{i=1}^{N} \tilde{\gamma}_k \tilde{\gamma}_k^\intercal$ *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 $u(k) = \psi(x(k))$
$$v^\intercal \Psi(x(0), u(0)) = v^\intercal \Psi(x(k), u(k)).$$

(i) really is about the interplay of recorded respones and our representation and not about an identification problem in the statistical sense.

Proof. R_N does not have full rank, therefore there exists $v \neq 0$ s.t.
$$0 = v^\intercal R_N v = \frac{1}{N}\sum_{i=1}^{N} (v^\intercal \tilde{\gamma}_k)^2.$$
Therefore, $v^\intercal \tilde{\gamma}_k = 0$ for every observed sample.
$$0 = v^\intercal \Psi(x(k), u(k)) - v^\intercal \Psi(x(k+1), \phi(x(x+1))),\ 0 \leq k \leq N-1 \tag{2}$$
So,
$$\begin{aligned} D_{k+1}(Q^{\theta'}) &= -Q^{\theta'}(x(k), u(k)) + c(x(k), u(k)) + Q^{\theta'}(x(k+1), \phi(x(k+1))) \\ &= c(x(k), u(k)) + [\theta + rv]\left[-\Psi(x(k), u(k)) + \Psi(x(k+1), \phi(x(k)))\right] \\ &\overset{2}{=} c(x(k), u(k)) + \theta\left[-\Psi(x(k), u(k)) + \Psi(x(k+1), \phi(x(k)))\right], \end{aligned}$$
which yields (i).
If $u(k) = \phi(x(k))$, use (2)
$$v^\intercal \Psi(x(k), u(k)) = v^\intercal \Psi(x(k+1), u(k+1))$$
repeated use for every k gives (ii).

□

To avoid the convergence of the $\Gamma(\theta) \to 0$ 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
$$\left\{x^i(k), u^i(k) \mid 0 \leq k \leq N, 1 \leq i \leq M\right\}$$
with user defined initial conditions
$$\{x^i(0) \mid 1 \leq i \leq M\}$$
and with action
$$u^i(k) = \tilde{\phi}(x^i(k), \xi^i(k))$$

It is fine not to probe at all

the approximation $Q^{\theta_N^{\text{LSTD}}} = \Psi^\intercal \theta_N^{\text{LSTD}}$ is obtained. Here
$$\theta_N^{\text{LSTD}} = \underset{\theta}{\operatorname{argmin}}\, \Gamma_N^i(\theta),\ \Gamma_N(\theta) = \frac{1}{M}\sum_{i=1}^{M} \Gamma_N^i$$
and
$$\Gamma_N^i(\theta) = \theta^\intercal W \theta + \sum_{i=1}^{N-1} \left[\gamma_k^i - \xi \tilde{\gamma}_{k+1}^{i\intercal} \theta\right]$$

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

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

3.2.3 Galerkin relaxation

Basis $\{\psi_i\}, h^\theta(z) = \sum_{i=1}^d \theta_i \psi_i(z)$, we want $0 \overset{!}{=} \nabla_\theta \Gamma(h^\theta)$.
For Bellman 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 $\{\zeta(k)\},\ \zeta(k) \in \mathbb{R}^{d_\zeta}$ constraints

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

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

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

$\zeta(k)$ does not depend on θ, $\zeta(k) \neq \zeta^\theta(k)$, *maybe* $\zeta(k) \approx \zeta^\theta(k), \theta \in$ *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

Motivated by the solution of the Bellman equation

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

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

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

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

Proposition 48. *Suppose that the following hold*

(i) $\mathcal{H} = G$

(ii) $\mathcal{H}$ *is a linear function class, i.e.* $a_1h_1 + a_2h_2 \in \mathcal{H}$ *for* $h_1, h_2 \in \mathcal{H},\ a_1, a_2 \in \mathbb{R}$

(iii) The mapping $P_{\mathcal{H}}$ *is linear. For* $h_1, h_2 \in \mathcal{H},\ a_{1,2} \in \mathbb{R}$*:*

$$P_{\mathcal{H}}(a_1h_1 + a_2h_2) = a_1P_{\mathcal{H}}(h_1) + a_2P_{\mathcal{H}}(h_2)$$

Then the solution to (3) and (4) coincide.

Start of lecture 15 (05.06.2025)

Proof. Trivial □

We assume for $g \in G$: $g : Z \to \mathbb{R}$, 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

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

For $\hat{h} \in G$

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

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

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

stack them to get $\zeta : Z \to \mathbb{R}^d$ and define $G\{g = \theta^\intercal\zeta \mid \theta \in \mathbb{R}^d\}$. $\zeta(k) := \zeta(\Phi(k))$ 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 $\|\zeta^\intercal\zeta\|_\omega = 0$. For each $h \in L_2(\omega)$, the projection exists, is unique, and given by*

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

with $\theta^\star = [R^\zeta]^{-1}\bar{\psi}^h$, $\bar{\psi}^h \in \mathbb{R}^d, \bar{\psi}^h_i = \langle\zeta_i, h\rangle_\omega$ $\mathbb{R}^z eta \in \mathbb{R}^{d\times d}, R^\zeta_{ij} = \langle\zeta_i, \zeta_j\rangle_\omega$.

Sketch. The orthogonality principle gives

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

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

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

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

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

We saw this last time as well

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

ψ like a Q-function

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

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

More concrete Galerkin estimation

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

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 $\{x(k), u(k) \mid 0 \le k \le N\}$, the sequence of estimates is defined by three coupled equations

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

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

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

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

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

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

Observe that $\zeta^0_n = \Psi_{(n)}$ for a linear function class, so this is a consistent generalization.
To extend, we use instead of the so far fixed policy ϕ

$\lambda = 0$ means we don't have a history at all! TDλ is for a fixed policy

$$\phi^\theta_{(n)} = \operatorname*{argmin}_u Q^\theta(x, u)$$

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 $\theta^\star$ will save $\bar{f}(\theta^\star) = 0$ with

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

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 48 to conclude that $Q^{\theta^\star}$ solves

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

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.

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

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

$$0 \leq n \leq B-1 \tag{7}$$

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 6 solves thef 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

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

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

In this can we can rearrange and invert

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

For α small enough, we can observe that

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

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:

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

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:*
$$\bar{f} = \mathbb{E}_\omega\left[\zeta(k) D_{k+1}(\theta)\right]$$
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 G$Q(\lambda)$-Learning

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

The G prob. stands for generalized

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

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

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

choosing $M = \mathbb{E}\left[\zeta_n \zeta_n^\intercal\right]^{-1}$, one can derive the **D$Q(\lambda)$-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

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

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.

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

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

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

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

The approximation is successful if

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

Problems

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

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

  if $A^\star$ 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

Start of lecture 17 (24.06.2025)

3.5 Summary

To summarize, inside TD taxonomy, we have seen

(i) approximate PIA using LSTD or TD(λ). We can be sure it converges under two conditions:

   (a) linearity: the function class is linear

   (b) the function class is complete, in the sense that we have $Q^{\theta_n} = Q^{\psi^n}$ for each n

(ii) Galerkin relaxations of the dynamic programming (DP) equation are obtained using $Q(\lambda)$-learning, DQN or batch $Q(\lambda)$-learning. There theory is almost nonexistent

(iii) Generalized Q-learning, to obtain the minimal mean square Bellman error. We are assured success, if $Q^\star$ lies inside our function class and the objective satisfies conditions aligned with gradient descent, e.g. the PL condition from the earlier chapter

Next couple of lecture also have parts from other books

3.6 Exploration

We assume $u(k) = \breve{\psi}(x(k), \xi(k))$, where $\underline{\xi}$ is a bounded sequence on a set $\Omega \subset \mathbb{R}^p$ for some $p > 1$
We assume an autonomous state space model for ξ

$$\xi(k+1) = H(\xi(k)),\ H \text{ continuous.}$$

. $\Phi(k) := (x(k), u(k), \xi(k))$ has an analogous form in the state space Z.
Remember (QSA2), ergodic limit, Z, average of observations. Denote for $g : Z \to \mathbb{R}$, g continuous, $N \geq 1$

$$\bar{g}_N = \frac{1}{N}\sum_{k=1}^{N} g(\Phi(k)).$$

We will assume the existence of

$$E_\omega\left[g(\Phi)\right] := \lim_{N\to\infty} \bar{g}_N. \tag{8}$$

Often, we have ω as a probability measure with density ρ, s.t.

$$E_\omega(g(\Phi)) = \int_Z g(z)\rho(z)dz$$

Lemma 53. *Consider the probing signal $\xi(k) = \sin(2\pi k/T)$, $k \geq 0$, provided that T is an irrational* *number, for any continuous function $g : \mathbb{R} \to \mathbb{R}$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} g(\xi(k)) = \int_0^1 g(\sin(2\pi k)dt) = \int_{-1}^1 g(t)\rho(t)dt$$

where $\rho(t) = \left[\pi\sqrt{1-t^2}\right]^{-1}$ is known as the arcsine density.

Proof. Consider $\xi^0(k) = [k/T]_1 = k/T - \lfloor k/T \rfloor$, which is the fractional part of k/T. $\xi^0(k)$ samples uniformly in $[0,1]$, for continuous functions $h : \mathbb{R} \to \mathbb{R}$ it then holds

$$\frac{1}{N} \sum_{k=1}^{N} h(\xi^0(k)) = \int_0^1 h(r)dr$$

with $h(\xi^0(k)) = g(\sin(2\pi\xi^0(k))) = g(\xi(k))$ the first equality follows, the second equality is standard calculus. □

Assumption Aξ:
The state and action spaces are each subsets of Euclidean space $F : X \times U \to X$, describing $x(k+1) = F(x(k), u(k))$ $\breve{\phi}, H$ from above are continuous. The state process Φ has the following properties

1. Φ evolves on a closed subset of Euclidean space, denoted Z, and $(x(k), u(k), \xi(k)) = w(\Phi(k))$ for each k, where $w : Z \to X \times U \times \Omega$ is Lipschitz
2. there is a probability measure ω, s.t. for any continuous function $g : Z \to \mathbb{R}$ the ergodic mean (8) exists for each initial condition
3. the limit (8) is uniform on

   $$G_L := \{g \mid \|g(z') - g(z)\| \le L\|z - z'\|, \ \forall z, z' \in Z\}$$

   for each $L < \infty$.

   $$\lim_{N\to\infty} \sup_{g\in G_L} |\bar{g}_N - E_\omega[g(\Phi)]| = 0$$

3.7 ODE approximation

Consider a recursion

$$\theta_{n+1} = \theta_n + \alpha_{n+1} f_{n+1}(\theta_n) \tag{9}$$

, here $\{f_n\}$ is a sequence of functions that admit an ergodic limit

$$\bar{f}(\theta) := \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f_k(\theta), \ \theta \in \mathbb{R}^d$$

We associate an ODE

$$\frac{d}{dt}\vartheta_t = \bar{f}(\vartheta_t) \tag{10}$$

We can use the Euler scheme, $\tau_0, \tau_n = \sum_{k=1}^n \alpha_k$, for $n \ge 1$. To get a continuous time process, we set $\hat{\theta}_{\tau_n} = \theta_n$ and extend by piecewise linear interpolation. Let $\{\vartheta_t^n \mid t \ge \tau_n\}$ denote the solution to (10) with starting condition $\vartheta^n_{\tau_n} = \theta_n$.
The recursion (9) is said to admit an ODE approximation, if the error

$$\lim_{n\to\infty} \sup_{\tau_n \le \tau \le \tau_n + N} \|\hat{\theta}_\tau - \vartheta^n_\tau\| = 0$$

If $\{\theta_n\}$ is bounded, convergence can be shown similar to proposition 38, which allows to use the ideas behind proposition 39to establish convergence if (10) is globally asymptotically stable.

3.8 Convergence rates

The rate of convergence is $1/t^{\rho_0}$ if

$$\limsup_{t\to\infty} t^{\rho}\|\tilde{\theta}_t\| = \begin{cases}\infty & \rho > \rho_0\\ 0 & \rho < \rho_0\end{cases}$$

where $\tilde{\theta}_t = \theta_t - \theta^\star$. In our context, one can achieve $\rho_0 = 1$, which is optimal in most cases.
Generally, there is an influence of the gain α on the convergence. Consider, a standard choice $a_t = g/(1+t)^{\rho}$, where $g > 0,\ 0 < \rho \leq 1$ are fixed.
The time scaling $\tau = s_t := \int_0^t a_r dr$ results in

$$\tau = \begin{cases} g\log(1+t) & \rho = 1\\ g\frac{1}{1-\rho}(1+t)^{1-\rho} & 0<\rho<1\end{cases} \tag{11}$$

$\frac{d}{dt}\vartheta_t = \bar{f}(\vartheta_t)$ and assume exponential asymptotically:
there exists $\rho_0 > 0$, $B_0 < \infty$ s.t. for any solution to the ODE and any $t \geq 0$

$$\|\vartheta_t - \theta^\star\| \leq B_0\|\vartheta_0 - \theta^\star\|\exp(-\rho_0 t).$$

Remember from lemma 35 $\frac{d}{dt}\bar{\theta}_t = a_t\bar{f}(\bar{\theta}_t)$ that $\theta_t = \vartheta_\tau, t \geq t_0$.
So that $\|\vartheta_\tau - \theta^\star\| = \|\bar{\theta}_t - \theta^\star\|$. One can see two different aspects.
$\rho < 1$: $\{\bar{\theta}_t\}$ converges to $\theta^\star$ very quickly. But, the boundedness of $\frac{1}{a_t}(\theta_t - \bar{\theta}_t)$ implies a suboptimal rate

$$\|\theta_t - \bar{\theta}_t\| \leq B\frac{1}{(1+t)^{\rho}},$$

where B is a function of the initial condition θ_0.
$\rho = 1$ the above bound is ideal, but with (11) we can observe

$$\|\bar{\theta}_t - \theta^\star\| \leq B_0\|\bar{\theta}_0 - \theta^\star\|\frac{1}{(1+t)^{g\rho_0}}.$$

So the rate of convergence of $\{\bar{\theta}_t\}$ depends on g. For the optimal one $1/t$, one needs $g \geq \frac{1}{\rho_0}$.
So, g can be large, which can lead to large transients/ vector fields. By averaging techniques: $\theta_T^{PR} := \frac{1}{T-T_0}\int_{T_0}^{T}\theta_t dt$. One can achieve the optimal rate of 1 overall.
F.s. $T_0 = T - T/5$, averages over last 20%.
This lecture is mostly based on [5]

Start of lecture 18 (26.06.2025)

3.9 Examples of Off-policy divergence

$$\begin{tikzcd} & \theta \arrow[r] & 2\theta \\ \text{values} \arrow[ur] & \text{feature 1} \arrow[u] & \text{feature 2} \arrow[u] \end{tikzcd}$$

two states, whose estimated values are $\theta, 2\theta, \theta \in \mathbb{R}^n$. Feature vectors are $1, 2$. In state 1, only action is going to state 2 with cost 0. We consider discounted

$$\sum_{k=1}^{\infty}\gamma^k c(x(k),(k))$$

and TD(0). Assume that $\theta : 0 = 10,\ \gamma \approx 1$.

$$\begin{aligned}\theta_{n+1} &= \theta_n + \alpha D_{n+1}\Psi_{(n+1)}\\ D_{n+1} &= -\underbrace{Q^{\theta_n}(x(k),u(k))}_{10} + 0 + \underbrace{\gamma Q^{\theta_n}(x(k),\phi)}_{20}\end{aligned}$$

If $\alpha = 0.1,\ \theta_1 \approx 11.$, do it once more to get $\theta_2 \approx 12.1$

$$\begin{aligned} D_{n+1}(Q) &= -\theta_n + 0 + \gamma 2\theta_n = (2\gamma - 1)\theta_n \\ \theta_{n+1} &= \theta_n + \alpha(2\gamma - 1)\theta_n \cdot 1 = \underbrace{(1 + \alpha(2\gamma - 1))}_{>1 \text{ if } \gamma > 0.5} \theta_n \end{aligned}$$

In the off-policy training, we do not follow the currently best action, whatever it may be. In off-policy training, one usually uses **importance sampling** or **reweighting** between target and behavior policy.
the update becomes

$$\theta_{n+1} = \theta_n + \alpha \delta_n D_{n+1} \Psi_{(n+1)}$$

with

$$\delta_n = \mathbb{P}\left[\frac{\text{target policy takes u at x(n)}}{\text{behavior policy takes u at x(n)}}\right]$$

So $\delta_n = 0$ if ϕ^b takes something, which ϕ^t never would.

3.9.1 Baird's counter examples

Cost is 0 on all transitions, so the true value function is constant 0, which can be achieved by $\theta = 0$, but it is not unique. due to our parametrization

$$\theta \in \mathbb{R}^n \Psi(1) = \begin{pmatrix} 2 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \qquad \qquad \ldots$$

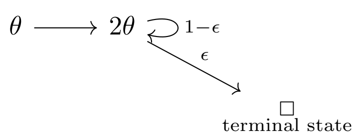
vectors linearly independent.
Consider exemplarily the solid transitions

$$D_{n+1} = \begin{cases} -\left[\theta_n^8 + 2\theta_n^k\right] + \gamma\left[2\theta_n^8 + \theta_n^7\right] & x(n) = k \leq 6 \\ -\left[2\theta_n^8 + 2\theta_n^7\right] + \gamma\left[2\theta_n^8 + \theta_n^7\right] & x(n) = 7 \end{cases}$$

$$\theta_{n+1}^8 = \theta_n^8 + \begin{cases} \alpha[(2\gamma - 1)\theta_n^8 + \gamma\theta_n^7 - 2\theta_n^k] & x(n) = k \leq 6 \\ \alpha[-(1-\gamma)[2\theta_n^8 + \theta_n^7 -]] & x(n) = 7 \end{cases}$$

Insert $\frac{1}{7}$ for importance sampling.
6 of 7 times, $(2\gamma - 1)$ for $\gamma > 0.5$ amplifies θ_n^8. TD(λ) does not change the behavior.
A DP-like algorithm with gradient updates and averaging, or expectation, over all states does not change the behavior.
Similar counterexamples exist for Q-learning. Generally, a behavior policy *close enough* to the target does not result in divergence, but so far there is no theory.

3.9.2 Tsitsiklis and Van Roy's counter example



$$\theta_{n+1} = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \sum_{x \in X} \left(V^\theta(x) - \mathbb{E}(0 + \gamma V^\theta(x') \mid x' = F(x))\right)^2$$

So θ_{n+1} minimizes the MSE at each step between the approximation and the expected return.

$$
\begin{aligned}
\theta_{n+1} &= \operatorname*{argmin}_{\theta\in\mathbb{R}} (\theta - \gamma 2\theta_n)^2 + (2\theta - (1-\epsilon)\gamma 2\theta_n)^2 \\
&= \frac{6-4\epsilon}{5}\gamma\theta_n
\end{aligned}
$$

When $\gamma > \frac{5}{6-4-\epsilon}$ and $\theta_0 \neq 0$ the sequence diverges.

3.9.3 The deadly triad

Non-Bootstrap methods would not have these problems?

1. Function approximation: scalable way to work with a state space much larger then the memory and compute resources
2. Bootstrapping: Updates that are based on current estimates, as in TD or DP methods. One could alternatively use so called Monte Carlo methods, which use only actual rewards and compute returns
3. Off-policy learning: Training on a distribution of transitions different to that of the target policy.

Divergence arises if all three are present. If there are only two present, instability can be avoided.

The note this is not necessarily specific to the RL-setting and hint to operations research

1. For large problems function approximation cannot be avoided
2. Bootstrapping can be avoided at the cost of computational and data efficiency. One advantage is the direct updating after each transition, or couple of transitions. It (bootstrapping) typically results in faster learning w.r.t data efficiency
3. Often on-policy is adequate, as long as the state-action space is reasonably covered
   (a) data re-use, in particular if data is costly. One would like to do **experience replay**, i.e. re-use data from *earlier* policies
   (b) learning multiple RL agents, i.e. several value functions and policies

   The behavior policy likely reflects only one task of many, i.e. there is only one target policy, but it may overlap partly with other tasks.

This lecture is based on [3] and [2] (fourth edition) chapter 6.1 and 6.2.

Start of lecture 19 (01.07.2025)

3.10 Monte Carlo Sampling / Simulation

Aim: Generate trajectories, use observed states, actions and costs. Use that to directly estimate V^ϕ, J^ϕ.

Todays focus is on value functions

Policy evaluation, both the same

The uniform choice implies that X is bounded in some sense I think

Algorithm 3 Episodes first-visit MC policy evaluation

initialize return(x_i) as an empty list for all states x_i
while stopping criteria not fullfilled **do**
  choose $x(0) \in X$ uniformly random among possible start positions
  Sample ϕ to generate a trajectory $x(0), u(0), c(0), \dots, x(l-1), u(l-1), c(l-1)$
  target $\leftarrow 0$
  for $i \in l-1, \dots, 0$ **do**
    target $\leftarrow c(i+1) + \gamma\cdot$target
    if $x(i) \notin \{x(0), \dots, x(i-1)\}$ **then**
      append target to return($x(i)$)
      $V(x(i)) \leftarrow$ average(return($x(i)$))
    end if
  end for
end while

Remark. *ϕ is fixed, so sampling it means following the path (which might be random).*

So, we compute in each update

$$\tilde{c}(x(i)) = \sum_{k=i}^{l-1} \gamma^{k-i} c(k-1)$$

which is an estimate of $V^{\phi}(x_i)$. Using only the first visit, one can see that we have independent identically distributed estimates, convergence of the average to V^{ϕ} follows by the law of large numbers.
Consider a $q \times d$ matrix Ψ, which we can view as some basis representation, and some subspace S spanned by

$$S = \{\Psi \cdot \theta \mid \theta \in \mathbb{R}^d\}$$

. The projected Bellman equation

$$\Psi\theta = \pi T^{\phi}(\Psi\theta)$$

where T^{ϕ} is the Bellman operator and π is the projection onto S w.r.t. $\|\cdot\|$.
This solves approximately $J^{\phi} = T^{\phi} J^{\phi}$. In the terminology of [2] this us called the **indirect approach**.
The **direct approach** is finding $\tilde{J} \in S$ via

$$\min_{J \in S} \|J^{\phi} - \tilde{J}\|$$

or

$$\min_{\theta \in \mathbb{R}^d} \|J^{\phi} - \Psi\theta.$$

If Ψ has independent columns, then the solution $\theta^{\star}$ is unique.
Now consider $\|\cdot\|_{\xi}$,

not totally different to the previous ξ

$$\xi_i \geq 0,\ i = 1, \ldots, q,\ \|J\|_{\xi}^2 = \sum_{i=1}^{q} \xi_i (J_i)^2.$$

$$\theta^{\star} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{q} \xi(\Psi(i)^{\intercal}) \cdot \theta - J_i,$$

where $\Psi(i)$ is the ith row of Ψ.
Setting the gradient to 0 gives

$$\theta^{\star} = \underbrace{\left(\sum_{i=1}^{q} \xi_i \Psi(i)\Psi(i)^{\intercal}\right)^{-1}}_{\hat{A}} \underbrace{\sum_{i=1}^{q} \xi_i \Psi(i) J_i}_{\hat{B}}.$$

Now assume ξ is a probability distribution, so we can consider both terms as expected values and can approximate them by Monte Carlo estimates.
So, we generate a sequence of samples of indices $i_t,\ t = 1, \ldots, K$ according to ξ and obtain

$$A = \frac{1}{K} \sum_{t=1}^{K} \Psi(i_t)\Psi(i_t) \approx \hat{A}$$
$$B = \frac{1}{K} \sum_{t=1}^{K} \Psi(i_t) J_{i_t} \approx \hat{B}$$

Generally, $\hat{\phi}_k \to \theta^{\star}$ as k is increasing. For that

$$\xi_i = \lim_{k \to \infty} \frac{1}{K} \sum_{t=1}^{K} \delta(i_t = i),\ i = 1, \ldots, q$$

the long term empirical frequencies should be consistent with the probabilities ξ_i.

ξ_i does not need to tbe pre-determined, i.e. it can be implicitly defined as above, given a reasonable sampling scheme.
We can also solve

$$\hat{\theta}_K = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{t=1}^{K} (\Psi(i_t)^\intercal \theta - J(i_t))^2.$$

So, we reduce q dimensional linear algebra operations to d-dimensional ones, using Monte Carlo sampling, and perform lower dimensional linear algebra.

3.10.1 MC estimation and the solution of linear equations systems

Generally

$$Cr = d,$$

where $C,\ d$ might be difficult to compute directly. We aim for $\hat{C},\ \hat{d}$ as simulation generated estimates.
In our context, we aim to find $\tilde{J}(\cdot, \theta)$ as an approximation of J^ϕ.

$$\min_\theta \sum_{k=1}^{q} \left(J_i^\phi - \tilde{J}(i, \theta)\right)^2$$

We use a subset $\tilde{I}$ of *representative* states
For each $i \in \tilde{I}$ we obtain $M(i)$ samples of J_i^ϕ, with $c(i, m)$ denoting the m-th sample.

$$c(i, m) \approx J_i^\phi + \text{ noise } + \text{ simulation error}$$

$$\min_\theta \sum_{i \in \tilde{I}} \sum_{m=1}^{M(i)} \left(c(i, m) - \tilde{J}(i, \theta)\right)^2$$

this is now actually computable and the key idea of this lecture

If $\tilde{J}(i, \theta) = \Psi(i)^\intercal \theta$, solve

$$\sum_{i \in \tilde{I}} \sum_{m=1}^{m(i)} (c(i, m) - \tilde{J}(i, \theta))^2$$

If $\tilde{I} = \{1, \ldots, q\}$, then for $M(i) \to \infty$, $\Psi\theta$ converges to the projection of J^Φ into $\{\Psi\theta \mid \theta \in \mathbb{R}^d\}$ w.r.t some weighted Euclidean norm. The weights of the norm are specified by the relative frequencies of the different states:

$$\lim_{K \to \infty} \frac{M(i)}{\sum_{i=1}^{q}} M(i)$$

here $K = \sum_{i=1}^{q} M(i)$

3.10.2 Importance Sampling

$$\|\Psi\theta - J\|_\xi^2$$

Projection is an expected value according to ξ, where there are multiple alternative distribution according to which we may represent the error above as an expected value.
It should be more effective to sample *important* terms/ states more often, i.e. large vs small size of J_i^ϕ. This is known as **important samplings**.
Generally, consider

$$z = \sum_{w \in W} v(w),$$

where W is a finite set and $v : W \to \mathbb{R}$. Consider a sampling distribution ξ over W, and sample according to it. Write first

$$z = \sum_{w \in W} \xi(w) \frac{v(w)}{\xi(w)}$$

and estimate it by

$$\hat{z}_K = \frac{1}{K}\sum_{i=1}^{K}\frac{v(w(i))}{\xi(w(i))}. \tag{12}$$

For this to valid, we want $\xi(w) = \lim_{K\to\infty}\frac{1}{K}\sum_{i=1}^{K}\delta(w_i = w)\ \forall w \in W$.
The expression (12) suggests that ξ should be chosen s.t. the variance of the random variable $\frac{v(w)}{\xi(w)}$ is small. In the extreme the variance is 0 and $\xi(w) = v(w)/z\ \forall w \in W$ and $v(w) > 0$, then a single sample is enough.

Start of lecture 20 (03.07.2025)

3.11 Gradient Methods for direct Policy Evaluation

$$\min_\theta \sum_{i\in I}\sum_{m=1}^{M(i)}\left(\tilde{J}(i,\theta) - c(i,m)\right)^2$$

Now, use gradient descent to solve this, and use the data in batches.
We have a N-transition portion $(i_0,\dots,i_N)$ of a sampled trajectory, a **batch**.

$$\sum_{t=k}^{N-1}\gamma^{t-k}c(i_t,u(i_t),i_{t+1}),\ \ k = 0,\dots,N-1$$

are **cost samples** as we had in Monte Carlo Policy evaluation (MC PE).
To approximate in L^2-sense

$$\min_\theta \sum_{k=0}^{N-1}\left(\tilde{J}(i_k,\theta) - \sum_{t=k}^{N-1}\gamma^{t-k}c(i_t,u(i_t),i_{t+1})\right)^2$$

$\to$ use gradient descent to update

$$\theta_{n+1} = \theta_n - \alpha\sum_{n=0}^{N-1}\nabla\tilde{J}(i_k,\theta)\left(\tilde{J}(i_k,\theta) - \sum_{t=k}^{N-1}\gamma^{t-k}c(i_t,u(i_t),i_{t+1})\right)$$

In traditional gradient descent, this iteration is repeated until convergence. In part, this one N-transition is used. Balancing

- Large $N \to$ reduce sample error, and to obtain multiple estimates per state and cover all states

Or a representative subset

- Small $N \to$ to keep effort per GD step small

In RL, batches may be changed after (some) iterations. Batches might come from different sampling strategies, might be part of a long trajectory, might overlap,...
Clearly this connects to aspect of exploration (which we have previously seen).

Remark. *Convergence analysis taking the stochastic nature into account is possible, but can be mathematically involved (due to the several aspects: sampling, stochastic aspects,... and the interactions)*

3.11.1 Incremental Gradient Method for direct Policy Evaluation

Instead of updating θ after N transitions, processing all N at once, we can incrementally update θ N times. After each transition (i_k, i_{k+1}):

1. Evaluate $\nabla\tilde{J}(i_k,\theta)$
2. Sum all terms that involve (i_k, i_{k+1}) and update

$$\theta' = \theta - \alpha\left[\nabla\tilde{J}(i_k,\theta)\tilde{J}(i_k,\theta) - \left(\sum_{t=0}^{k}\gamma^{k-t}\nabla\tilde{J}(i_t,\theta)\right)c(i_k,u(i_k),i_{k+1})\right]$$

After N transitions all the terms of the batch iteration have been accumulated. Here, θ is updated during the batch processing, and $\nabla\tilde{J}$ is evaluated at a different (updated) θ after each transition.
Since θ is updated all the time, the location of the end of the batch becomes less relevant. As before, the $\|\cdot\|_\xi$, will be implicitly weighted in proportion to the frequency of occurrence of each state.
Connection to TD Error: $-D_{k+1} = td_k = \tilde{J}(i_k,\theta) - \gamma\tilde{J}(i_{k+1},\theta) - c(i_k, u(i_k), i_{k+1})$
with $td_{N-1} = \tilde{J}(i_{N-1},\theta) - c(i_{N-1}, u(i_{N-1}), i_N)$. We can write this as:

$$td_k + \gamma td_{k+1} + \cdots + \gamma^{N-1-k} td_{N-1}$$

With that we can implement the batch iteration as

- after (i_0, i_1) set
$$\theta' = \theta - \alpha td_0 \nabla\tilde{J}(i_0,\tilde{\theta})$$
- after i_1, i_2 set $\theta = \theta'$ and
$$\theta' = \theta - \alpha td_1 \left(\gamma\nabla\tilde{J}(i_0,\tilde{\theta}) + \nabla\tilde{J}(i_1,\tilde{\theta})\right)$$

Repeating gives after (i_{N-1}, i_N) set

$$\theta' = \theta - \alpha td_{N-1} \left(\gamma^{N-1}\nabla\tilde{J}(i_0,\tilde{\theta}) + \gamma^{N-2}\nabla\tilde{J}(i_1,\tilde{\theta}) + \cdots + \nabla\tilde{J}(i_{N-1}\tilde{\theta})\right)$$

Here $\tilde{\theta} = \theta$ a the beginning of the batch. In the incremental version $\tilde{\theta} = \theta$ at transition (i_k, i_{k+1}) for each $\nabla\tilde{J}(i_k,\tilde{\theta})$.
In particular, start with θ_0 and for $k = 0, \ldots, N_1$ set

$$\theta_{k+1} = \theta_k - \alpha td_k \sum_{t=0}^{k} \gamma^{k-t}\nabla\tilde{J}(i_t,\theta_t).$$

For linear approximation $\tilde{J}(i,\theta) = \Psi(i)^\intercal\theta,\ i = 1,\ldots,q,\ \Psi(i) \in \mathbb{R}^5$.

$$\theta_{k+1} = \theta_k - \alpha td_k \sum_{t=0}^{k} \Psi(i_t).$$

This is TD(1)!

in slightly different notation . . .

3.11.2 Multistep methods with sampling

Fixed point view: $J = TJ,\ J = \Pi TJ$
We can replace T by either $T^l,\ l > 1$ or consider

$$T^{(\lambda)} = (1-\lambda)\sum_{l=0}^{\infty} \lambda^l T^{l+1}$$

One can show under natural assumptions that$T^{(\lambda)}$ and $\Pi T^{(\lambda)}$ are contractions of modulus $\gamma_\lambda = \frac{\gamma(1-\lambda)}{1-\gamma\lambda}$:

On can show $\lambda \to 1$ connects to TD(1) as previously considered

$$\|T^{(\lambda)}y - T^{(\lambda)}z\| \leq \gamma_\lambda\|y - z\|.$$

Furthermore

$$\|J^\phi - Psi\theta^\star_\lambda\|_\xi \leq \frac{1}{\sqrt{1-\gamma_\lambda^2}}\|J^\phi - \Pi J^phi\|_\xi$$

We can see from this, that we want λ close to 1, as $\gamma_\lambda \overset{\lambda\to 1}{\to} 1$ and $T^{(\lambda)}$ is a contraction for any given norm for λ close enough to 1. Same goes for $\Pi T^{(\lambda)}$. Further with $\lambda \to 1$ the error bound becomes better.

3.11.3 Bias-Variance Tradeoff

We can consider $\Psi\theta^\star - \Pi J^\phi$ as a form of bias. But, one can observe that the sampling error becomes larger as λ increases.

$$T^{(\lambda)} = \sum_{l=0}^{\infty} \underbrace{(1-\lambda)\lambda^l}_{\text{increases with } \lambda} \underbrace{T^{l+1}}_{\text{noise variance due to the approximation of the lth power. Increases with } l}$$

Therefore one needs to balance bias-variance, and experiment with λ.

It is not clear how good one needs to approximate the value function to update the policy? Adding a constant worsens the error, but does not hange the policies!

Start of lecture 21 (08.07.2025)

3.12 Policy Gradient Methods

Fix a class of policies, parametrized by $\theta \in \mathbb{R}^d$, i.e. we have ϕ^θ.
Goal: minimize

$$\theta \mapsto V^{\phi^\theta}(x) = J_x(\theta).$$

We use gradient procedures, these are called policy gradient methods. We assume $\theta \mapsto \phi^\theta$ is differentiable in θ for all states of X.If we use Q-view, also for all actions.

$$\theta_{n+1} = \theta_n + \alpha \nabla J(\theta_n) \tag{13}$$

PG methods are typically stochastic. Consider a finite state action space, $d = |X| \cdot |U|$. Denote $\theta = (\theta_{x,u})_{x\in X,\ u\in U}$ and define softmax policy

$$\begin{aligned}\phi^\theta(u;x) &= \mathbb{P}(U_t = u \mid X_t = x, \theta_t = \theta)\\ &= \mathbb{P}(U = u \mid X = x, \theta = \theta)\\ &= \frac{e^{\theta_{x,u}}}{\sum_{u'\in U} e^{\theta_{x,u'}}}\end{aligned}$$

Let $C_t^T = \sum_{k=t}^{T-1} c(x(k), u(k))$ be the cost after time t.

Theorem 54 (Policy Gradient Theorem)**.** *Assume that we have T-step MDP with finite state action spaces and consider (stationary, in the sense of (13)) differentiable family of policies $\phi^\theta,\ \theta \in \mathbb{R}^d$. Then the gradient of the value function is*

$$\nabla_\theta J_x(\theta) = \mathbb{E}_x^{\phi^\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \phi^\theta(u(t), x(t)) C_t^T\right]$$

Proof idea. Basically calculus / chain rule, re-arranging terms and using the log trick

$$\nabla \log = \frac{\nabla f}{f},\ \ \nabla f \frac{f}{f} = (\nabla \log f) f$$

□

Using $\mathbb{E}_x^{\phi^\theta}\left[C_t^T \mid x_t = x, u_t = u\right] = Q_t^{\phi^\theta}(x,u)$, we can write

$$\nabla_\theta J_x(\theta) = \mathbb{E}_x^{\phi^\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \left(\log \phi^\theta(u(t)); x(t) Q_t^{\phi^\theta}(x(t), u(t))\right)\right]$$

Definition. *If ϕ^θ is a policy, then the vector $\nabla_\theta \log \phi^\theta(u;x)$ is called the score function of the policy.*

We sample to estimate the expectation and perform stochastic gradient descent using K trajectories $(x^i(0), u^i(0), x^i(1), c^i(1), u^i(1), \dots, u^i(T-1), x^i(T), c^i(T))$ sampled according to the policy ϕ^θ

$$\tilde{\nabla}_\theta J_x(\theta) = \frac{1}{K}\sum_{i=1}^{K}\left[\sum_{t=0}^{T-1} \nabla_\theta \left(\log \phi^\theta(u^i(t), x^i(t)) \sum_{t'=t}^{T-1} c(t'+1)\right)\right]$$

Algorithm 4 REINFORCE-(Batch) Stochastic Gradient Algorithm

Given: $\theta_0, K \geq 1$ initial state distribution μ

$l = 0$

while stopping criteria not fullfilled **do**

  for $i = 1, \dots K$ **do**

    sample trajectory i: $(x^i(0), u^i(0), x^i(1), c^i(1), u^i(1), \dots, u^i(T-1), x^i(T), c^i(T))$

  end for

  choose α

$$\tilde{\nabla}_\theta J_x(\theta) = \frac{1}{K} \sum_{i=1}^{K} \left[\sum_{t=0}^{T-1} \nabla_\theta \left(\log \phi^\theta(u^i(t), x^i(t)) \sum_{t'=t}^{T-1} c(t'+1) \right) \right]$$

with $\theta = \theta_l$

  $\theta_{l+1} = \theta_k - \alpha \tilde{\nabla} J(\theta_l)$

  $l = l + 1$

end while

3.12.1 Infinite Horizon

We measure the **discounted state visitations** and denote $\mathbb{P}^\phi_\mu(X_t = x') = \mathbb{P}(\mu \to x' \mid t, \phi)$

$$\rho_\mu(x') = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\phi_\mu(X_t = x') = \mathbb{E}^\phi_\mu \left[\sum_{t=0}^{\infty} \gamma^t 1_{X_t = x'} \right]$$

We define a measure from it

$$d^\phi_\mu(x) = \frac{\rho^\phi_\mu(x)}{\sum_{x'} \rho^\phi_\mu(x')} = (1-\gamma)\rho^\phi_\mu(x)$$

$$\sum_{x'} \mathbb{E}(1_{X_t = x'}) = \frac{1}{1-\gamma}$$

Using dynamic programming equation, repeating chain rule, one can show

Theorem 55. *Under the assumption that $J_x(\theta)$ is differentiable for every state $x \in X$ it holds that*

$$\nabla_\theta J_x(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{\substack{X \sim dx \\ u \sim \phi^\theta(x)}} \left[\nabla \log \left(\phi^\theta(U; X) \right) Q^{\phi^\theta}(X, U) \right]$$

Proof sketch.

$$\begin{aligned}
\nabla J_x(\theta) &= \sum_{x' \in X} \sum_{u \in U_{x'}} \overbrace{\rho^{\phi^\theta}_x}^{\frac{1}{1-\gamma} d^{\phi^\theta}_x(x')} \nabla \phi^\theta(u; x') Q^{\phi^\theta} \\
&= \frac{1}{1-\gamma} \sum_{x' \in X} \sum_{u \in U} \nabla \log \left(\phi^\theta(u; x') \right) Q^{\phi^\theta}(x', u) \phi^\theta(u; x') d^{\phi^\theta}_x(x') \\
&= \frac{1}{1-\gamma} \mathbb{E}_{x \sim d^{\phi^\theta}_x, U \sim \phi^\theta(\cdot; x)} \left[\nabla \log(\phi^\theta(u, x')) Q^{\phi^\theta}(x', u) \right]
\end{aligned}$$

□

Sampling from d^ϕ can be achieved as follows:

This is esentially bootstrap sampling, but worse?

1. follow rollout until an independent time according to Geom$(1-\gamma)$
2. estimate empirical distribution by counting the number of visits
3. sample from the thereby estimated occupancy measure

Theorem 56. *Suppose that* $(x,u) \mapsto \nabla_\theta \left(\log \phi^\theta(u;x)\right) Q^{\phi^\theta}$ *is bounded. Then*

$$\nabla_\theta J_x(\theta) = \mathbb{E}_x^{\phi^\theta}\left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta (\log \phi^\theta(U_t, X_t)) Q^{\phi^\theta}(X_t, U_t)\right]$$

The boundedness does not hold in general and might be hard to check. For softmax policies the score function can be computed and is bounded for bounded feature vectors.. If the rewards are bounded, so is Q.

Proof.

$$\begin{aligned}
\nabla J_x(\theta) &= \sum_{t=0}^{\infty} \gamma^t \sum_{x' \in X} \mathbb{P}_x^{\phi^\theta}(X_t = x') \sum_{u \in U_{x'}} \nabla \phi^\theta(u;x') Q(x',u) \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_x^{\phi^\theta}\left[\sum_{u \in U} \underbrace{\nabla \phi^\theta(u;X_t)}_{\phi^\theta(\dots)\log\phi^\theta} Q^{\phi^\theta}(X_t,u)\right] \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_x^{\phi^\theta}\left[\nabla \log \phi^\theta(U_t;X_t) Q^{\phi^\theta}(X_t,U_t)\right] \\
&= \mathbb{E}_x^{\phi^\theta}\left[\sum_{t=0}^{\infty} \gamma^t \nabla \log \phi^\theta(U_t;X_t) Q^{\phi^\theta}(X_t,U_t)\right]
\end{aligned}$$

□

With that, one can use the REINFORCE algorithm 4 up to some large T, use the truncated series as an estimate for the gradient. This estimator is biased!

Start of lecture 22 (10.07.2025)

Assumptions for this lecture (APP):

Assumption on Parametrized Policy

The policy ϕ^θ is differentiable w.r.t. θ an $\nabla(\phi^\theta(u;x))$ exists is globally Lipschitz w.r.t. θ, i.e. it is L_θ-smooth and has bounded norm for any $x, u \in X \times U$.
This is fulfilled for linear softmax parametrization.

Proposition 57. *Suppose that* ϕ^θ *fulfills APP and that* $T \sim Geo(1-\gamma)$, $T' \sim Geo(1-\gamma^{\frac{1}{2}})$ *are independent of each other and the MDP. Then*

$$\nabla J_x(\theta) = \frac{1}{1-\gamma} \mathbb{E}_x^{\phi^\theta}\left[\nabla \log \phi^\theta(U_T;X_T)\right] \sum_{t=T}^{T+T'} \gamma^{(t-T)/2} c(X_t,U_t)$$

This way, an unbiased estimator can be obtained by a finite but random number of steps.

Algorithm 5 Minibatch-REINFORCE for finite Time Horizon

Given: $\theta_0, K \geq 1$ initial state distribution μ

$l = 0$
while stopping criteria not fullfilled **do**
  for $i = 1, \dots K$ **do**
    sample $T_i \sim \text{Geo}(1-\gamma)$
    sample trajectory i: $(x^i(0), u^i(0), x^i(1), c^i(1), u^i(1), \dots, u^i(T_i - 1), x^i(T_i), c^i(T_i))$
    sample $T_i \sim \text{Geo}(1-\gamma^{\frac{1}{2}})$
    sample trajectory i: $(\tilde{x}^i(0) = x^i(T_i), u^i(0) = u^i(T_i), \tilde{x}^i(1), \tilde{c}^i(1), \tilde{u}^i(1), \dots, \tilde{u}^i(\tilde{T}_i - 1), \tilde{x}^i(\tilde{T}_i), \tilde{c}^i(\tilde{T}_i))$
  end for
  choose α

$$\tilde{\nabla}_\theta J_x(\theta) = \frac{1}{K}\sum_{i=1}^{K}\left[\nabla \log \phi^\theta(U_T; X_T)\right]\sum_{t=T}^{T+T'} \gamma^{(t-T)/2} c(X_t, U_t)$$

with $\theta = \theta_l$
  $\theta_{l+1} = \theta_k - \alpha \tilde{\nabla} J(\theta_l)$
  $l = l + 1$
end while

Like burn-in

Theorem 58 (Stochastic Gradient Descent (SGD) almost sure conv. for L-smooth functions)**.** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a L-smooth with $F_\star = \inf_{\theta \in \mathbb{R}^d} F(\theta) > -\infty$. It satisfies*

$$\mathbb{E}_{z\sim\mu_\theta}\left[\|\nabla f(\theta, z) - \mathbb{E}\left[\nabla_\theta f(\theta, z)\right]\|^2\right] \leq c(1 + (F(\theta) - F_\star))$$

where $\mathbb{E}_{z\sim\mu_\theta}\left[\nabla_\theta f(\theta, z)\right] = \nabla F(\theta)$, this is called ergodic. Consider SGD with step

- *sample z_{+1} from μ_θ*
- *update $\theta_{l+1} \leftarrow theta_l - \alpha \nabla_\theta f(\theta_l, z_{l+1})$*

suppose for $(\alpha_l)_{l\in\mathbb{N}}$ it holds $\alpha_l \geq 0$ and that it holds almost surely

$$\sum_{l=0}^{\infty} \alpha_l = \infty, \ \sum_{l=0}^{\infty} \alpha_l^2 < \infty.$$

Let θ_0 be a random variable s.t. $\mathbb{E}(F(\theta)) < \infty$ and $(\theta_l)_{l\in\mathbb{N}}$ is the sequence of random variable generated by SGD. Then $(F(\theta_l))_l$ converges almost surely to some finite r. v. F_∞ and

$$\lim_{n\to\infty} \|\nabla_\theta F(\theta_l)\|^2 = 0 \ a.s.$$

Like the probing signal previously

Lemma 59. *Under the assumption APP, the objective $J_x(\theta)$ is L-smooth, more precisely global Lipschitz, with $L = \frac{c_\star L_\theta}{(1-\gamma)^2} + (1+\gamma)\frac{c_\star B_\theta^2}{(1-\gamma)^3}$, where $c_\star$ is the maximal cost from the bounded cost assumption.*

Proof. This proof was a nightmare to write down, as details are not asked in the exam I will leave it like this ...
From theorem 55

$$\nabla J_{x_0}(\theta) = \frac{1}{1-\gamma}\sum_{\substack{x\in X\\u\in U}} d_{x_0}^{\phi^\theta}\phi^\theta(u;x)\nabla\log(\phi^\theta(u;x))Q^{\phi^\theta}(x,u)$$

$$Q^{\phi^\theta} = \sum_{t'=0}^{\infty}\gamma^{t'}\sum_{\substack{x'\in X\\u\in U}} p((x,u) \to x_i' t_i' \phi^\theta)\phi^\theta(u';x')c(x',u')$$

Together

$$\nabla J_{x_0}(\theta) = \sum_{t=0}^{\infty}\sum_{t'=0}^{\infty}\gamma^{t+t'} \sum_{\substack{x\in X\\ u\in U\\ x'\in X\\ u'\in U}} \underbrace{p((x,u)\to x'_i t'_i \phi^\theta)\phi^\theta(u';x')p(x_0\to x_i t_i \phi^\theta)\phi^\theta(u;x)}_{f_{t,t'}^{x_0,\theta}(x,u,x',u')} \nabla\log(\phi^\theta(u;x))c(x',u')$$

$$\begin{aligned}
\|\nabla J_x(\theta_1) - \nabla J_x(\theta_2)\| &= \|\sum_{t=0}^{\infty}\sum_{t'=0}^{\infty}\gamma^{t+t'}\left\{\sum fortheta1 - \sum fortheta2\right\} \\
&= \|\sum_{t=0}^{\infty}\sum_{t'=0}^{\infty}\gamma^{t+t'}\Big\{\sum f_{t,t'}^{x_0,\theta_1}\left(\nabla\log(\phi^{\theta_1}(u;x) - \nabla\log(\phi^{\theta_2}(u;x)))c(x'-u')\right) + \sum\left(f_{t,t'}^{x_0,\theta_1} - f_{t,t'}^{x_0,\theta_2}\right)\nabla\log\left(\phi^{\theta_2}\right. \\
&\le \sum_{t=0}^{\infty}\sum_{t'=0}^{\infty}\gamma^{t+t'}\Big\{\sum f_{t,t'}^{x_0,\theta_1}\left\|\nabla\log(\phi^{\theta_1}(u;x) - \nabla\log(\phi^{\theta_2}(u;x)))c(x'-u')\right\| + \sum\left(f_{t,t'}^{x_0,\theta_1} - f_{t,t'}^{x_0,\theta_2}\right)\underbrace{\|\nabla\log\phi^{\theta_2}(}_{B_\theta}
\end{aligned}$$

$\mathcal{T}_t = \{\tau = \{x_0, u_0, x_1, u_1, \ldots, x_t, u_t\} \mid u_i \in U_{x_i}\}$

$$f_{t,t'}^{x_0,\theta} = \sum_{\tau\in\mathcal{T}_{t+t'}} 1_{X_t=x,U_t=u,X_{t'}=x',U_{t'}=u'} \prod_{n=0}^{t+t'}\phi^\theta(u_n;x_n)\prod_{n=0}^{t+t'-1}p(X_{n+1},x_n,u_n)$$

$$f_{t,t'}^{x_0,\theta_1} - f_{t,t'}^{x_0,\theta_2} = \sum_{\tau\in\mathcal{T}_{t+t'}} 1_{X_t=x,U_t=u,X_{t'}=x',U_{t'}=u'}\left(\prod_{n=0}^{t+t'}\phi^{\theta_1}(u_n;x_n) - \prod_{n=0}^{t+t'}\phi^{\theta_2}(u_n;x_n)\right)\prod_{n=0}^{t+t'-1}p(X_{n+1},x_n,u_n)$$

Using Taylor expansion (or the higher dimensional mean value theorem) of $\theta\mapsto\prod\phi^\theta$

$$\begin{aligned}
|\prod\phi^{\theta_1} - \prod\phi^{\theta_2}| &\le |(\theta_1-\theta_2)^\intercal\nabla_\theta\left(\prod\phi^\theta(u_n;x_n)\right)_{|_{\theta=\tilde\theta}}| \\
&\le \|\theta_1-\theta_2\|\left\|\sum_{n=0}^{t+t'}\nabla\phi^{\tilde\theta}(u_n;x_n)\prod_{m=0,m\neq n}^{t+t'}\phi^{\tilde\theta}(u_n;x_n)\right\| \\
&\overset{\le}{\text{log trick}} \|\theta_1-\theta_2\|\underbrace{\sum_{n=0}^{t+t'}\|\nabla\log\left(\phi^{\tilde\theta}(u_n;x_n)\right)\|}_{(t+t'+1)B_\theta}\prod_{m=0}^{t+t'}\phi^{\tilde\phi}(u_n;x_n)
\end{aligned}$$

Then the second term from the whole expression can be bounded

$$\sum_{\tau\in\mathcal{T}_{t+t'}}|\prod\phi^{\theta_1} - \prod\phi^{\theta_2}\|B\cdot B_\theta\cdot C_\star$$

$$\le\|\theta_1-\theta_2\|B_\theta^2 C_\star\underbrace{\sum_{\tau\in\mathcal{T}_{t+t'}}\prod_{m=0}^{t+t'}\phi^\theta(u_m;x_m)\prod_{n=0}^{t+t'}p(x_{n+1},x_n,u_n)}_{=1}$$

since the sum over all trajectories of the probabilities of the paths is 1. B is something he circled on the board and then erased. Together

$$\|\nabla J_x(\theta_1) - \nabla J_x(\theta_2)\| \le \sum_{t=0}^{\infty}\sum_{t'=0}^{\infty}\gamma^{t+t'}\left(c_\star L_\theta\|\theta_1-\theta_2\| + \|\theta_1-\theta_2\|\left((t+t'+1)B_\theta^2 c_\star\right)\right)$$

using $\sum\sum\gamma^{t+t'}(t+t'+1) = \frac{1+\gamma}{(1-\gamma)^2}$ and $L = \frac{C_\star L_\theta}{(1-\gamma^2)+\frac{(1+\gamma)c_\star B_\theta^2}{(1-\gamma^3)}}$. $\square$

Journal

- Lecture 01: Covering: Introduction, (linear, continuous) State space models, equilibrium, (Lyapunov,asymptotically) stable, region of attraction, globally asymptotically stable .
Starting in 'Organization' on page 3 and ending in 'State Space Models in continuous Time' on page 8. Spanning 5 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 'State Space Models in continuous Time' on page 8 and ending in 'State Space Models in continuous Time' on page 10. Spanning 2 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 'State Space Models in continuous Time' on page 10 and ending in 'Some concepts from Reinforcement Learning' on page 13. Spanning 3 pages
- Lecture 04: Covering: Value iteration, policy iteration, exploration-exploitation .
Starting in 'Some concepts from Reinforcement Learning' on page 13 and ending in 'Exploration' on page 17. Spanning 4 pages
- Lecture 05: Covering: Approximate Q-functions, Bandits, discounted cost, shortest path, finite horizon and translations between them .
Starting in 'Exploration' on page 17 and ending in 'Other control formulations' on page 20. Spanning 3 pages
- Lecture 06: Covering: Model predictive control, continuous time formulations of previous results .
Starting in 'Other control formulations' on page 20 and ending in 'Linear quadratic regulator revisited (once more)' on page 23. Spanning 3 pages
- Lecture 07: Covering: Picard-Iteration, Grönwall-Bellma inequality, Euler's method, gradient flows .
Starting in 'ODE methods for algorithm design' on page 24 and ending in 'Optimization' on page 26. Spanning 2 pages
- Lecture 08: Covering: Polyak-Lojasiewicz inequality, L-smooth inequality, Bregman divergence, quasi stochastic approximation .
Starting in 'Optimization' on page 26 and ending in 'Qausi stochastic approximation' on page 30. Spanning 4 pages
- Lecture 09: Covering: QSA continued, approximate policy improvement .
Starting in 'Qausi stochastic approximation' on page 30 and ending in 'Approximate Policy Improvement' on page 33. Spanning 3 pages
- Lecture 10: Covering: QSA1-QSA3, some convergence results .
Starting in 'Approximate Policy Improvement' on page 33 and ending in 'Approximate Policy Improvement' on page 35. Spanning 2 pages

- Lecture 11: Covering: Boundedness implies convergence, ultimate boundedness, first entrance times, QSV assumption .
Starting in 'Approximate Policy Improvement' on page 35 and ending in 'Approximate Policy Improvement' on page 37. Spanning 2 pages
- Lecture 12: Covering: Using QSV to show ODE solutions are ultimately bounded, Gradient free optimization: QSGD1, QSDG3 .
Starting in 'Approximate Policy Improvement' on page 37 and ending in 'Algorithm: qSDG #3' on page 39. Spanning 2 pages
- Lecture 13: Covering: Global consistency, very short crash course in ML, reinforcement learning, least squares temporal difference learning .
Starting in 'Algorithm: qSDG #3' on page 39 and ending in 'Algorithm: Least Squares Temporal Difference Learning (LSTD)' on page 43. Spanning 4 pages
- Lecture 14: Covering: Redundant Parametrization, Galerkin relaxation, projected bellman equation .
Starting in 'Algorithm: Least Squares Temporal Difference Learning (LSTD)' on page 43 and ending in 'Projected Bellman equation' on page 44. Spanning 1 pages
- Lecture 15: Covering: Eligibility vectors, Galerkin relaxation in the L_2 setting, TD(λ),TD(λ) with non-linear function approximation, Q-learning .
Starting in 'Projected Bellman equation' on page 44 and ending in 'Algorithm: Q-learning' on page 46. Spanning 2 pages
- Lecture 16: Covering:
.
Starting in 'Deep Q-Networks and Batch methods' on page 47 and ending in 'Algorithm: G$Q(\lambda)$ Learning for linear function approximation' on page 49. Spanning 2 pages
- Lecture 17: Covering:
.
Starting in 'Algorithm: G$Q(\lambda)$ Learning for linear function approximation' on page 49 and ending in 'Convergence rates' on page 51. Spanning 2 pages
- Lecture 18: Covering:
.
Starting in 'Convergence rates' on page 51 and ending in 'The deadly triad' on page 53. Spanning 2 pages
- Lecture 19: Covering:
.
Starting in 'The deadly triad' on page 53 and ending in 'Importance Sampling' on page 56. Spanning 3 pages
- Lecture 20: Covering:
.
Starting in 'Importance Sampling' on page 56 and ending in 'Bias-Variance Tradeoff' on page 58. Spanning 2 pages
- Lecture 21: Covering:
.
Starting in 'Bias-Variance Tradeoff' on page 58 and ending in 'Infinite Horizon' on page 60. Spanning 2 pages
- Lecture 22: Covering:
.
Starting in 'Infinite Horizon' on page 60 and ending in 'Infinite Horizon' on page 62. Spanning 2 pages

Bibliography

[1] Tamer Basar, Sean Meyn, and William R. Perkins. *Lecture Notes on Control System Theory and Design*. 2024. arXiv: 2007.01367 [math.OC]. URL: https://arxiv.org/abs/2007.01367.

[2] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. 2nd. Athena Scientific, 2000. ISBN: 1886529094.

[3] Bastian Bohn, Jochen Garcke, and Michael Griebel. *Algorithmic Mathematics in Machine Learning*. Data Science. Philadelphia, PA, USA: SIAM, 2024. DOI: 10.1137/1.9781611977882. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611977882.

[4] Sean Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.

[5] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second. The MIT Press, 2018. URL: https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf.