Learning + MPC via Reinforcement Learning
Fundamental principles

Sébastien Gros

Dept. of Cybernetic, NTNU
Faculty of Information Tech.

Freiburg PhD School

Forewords

On this topic

- First publication in 2020
- $\sim$40 papers
- Many talks & courses
- Growing portfolio of applications & experiments
- A bit on the "theoretical" side in the field

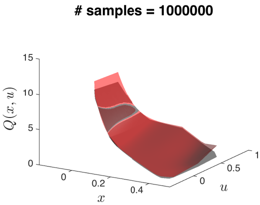
On these lectures

- Give high-level concepts
- Focus on known insights
- What are the current gaps
- New insights (3$^{\text{rd}}$ lecture)

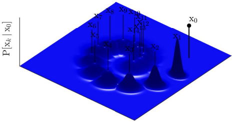
Software for implementation are not mature yet. You will be the first "large" audience playing with them.

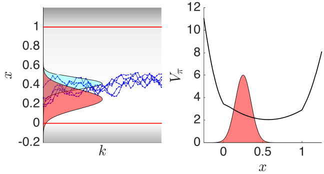
What are we going to discuss?

1. Learning for MPC - A focus on closed-loop performance
2. Safety & stability in Learning for MPC
3. When do "classic" approaches work / When is learning beneficial?



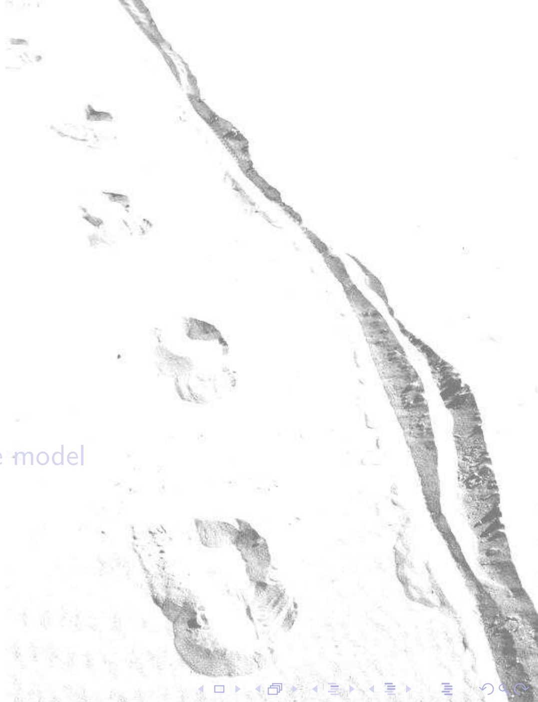
$$Q_{+}(\mathbf{x},\mathbf{u}) \leftarrow L(\mathbf{x},\mathbf{u}) + \gamma \mathbb{E}\left[V\left(\mathbf{x}_{+}\right) \mid \mathbf{x},\mathbf{u}\right]$$





Outline

1. The Basics
2. More background
3. Let's take a deeper dive
4. Parametrization & Role of the model
5. RL over MPC



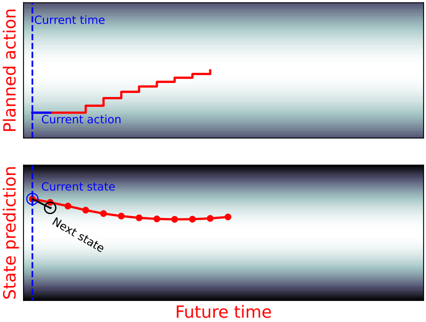
Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



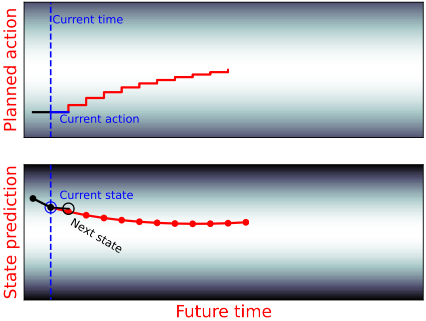
Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

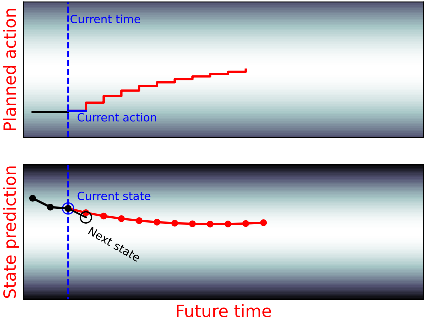
$$\min_{\mathbf{x},\mathbf{u}} \quad T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k)$$

$$\text{s.t.} \quad \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0$$

$$\mathbf{x}_0 = \mathbf{s}$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



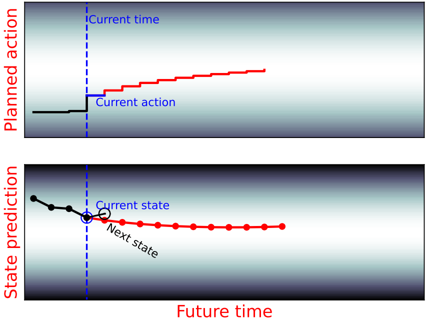
Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



Model Predictive Control (MPC)

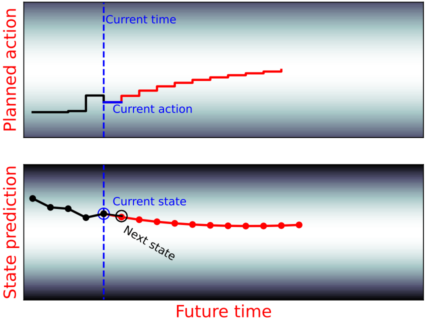
Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

$$\min_{\mathbf{x},\mathbf{u}} \quad T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right)$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

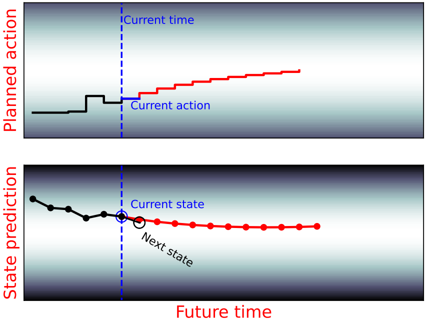
$$\min_{\mathbf{x},\mathbf{u}} \quad T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k)$$

$$\text{s.t.} \quad \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0$$

$$\mathbf{x}_0 = \mathbf{s}$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

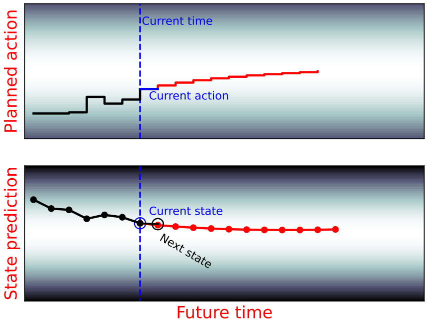
$$\min_{\mathbf{x},\mathbf{u}} \quad T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k)$$

$$\text{s.t.} \quad \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0$$

$$\mathbf{x}_0 = \mathbf{s}$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



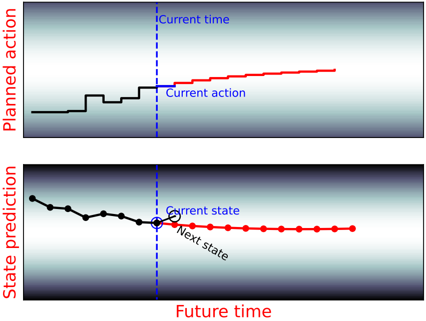
Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



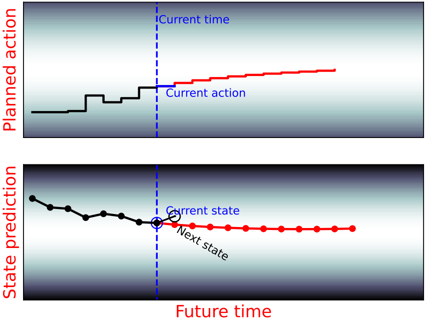
Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\
& \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



MPC

- is **based on planning** the future
- **Policy** from repeated planning

$$\pi^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$$

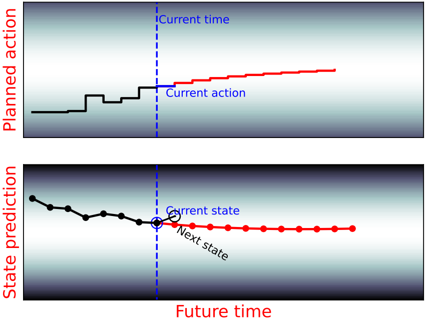
Model Predictive Control (MPC)

Optimize a plan over finite horizon, apply first move, repeat

MPC: at current state $\mathbf{s}$

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

apply action $\mathbf{a} = \mathbf{u}_0^\star$ to the system



MPC

- is **based on planning** the future
- **Policy** from repeated planning

$$\pi^{\mathrm{MPC}}\left(\mathbf{s}\right) = \mathbf{u}_0^\star$$

MPC is a powerful tool to control constrained systems, increasingly used as a practical way of building optimal policies

Theoretical Framework to connect RL and MPC

Model Predictive Control

- Model driven
- Policies from planning
- Constraints oriented

Reinforcement Learning (RL)

- Data driven
- Optimal policies from learning
- Performance oriented

Theoretical Framework to connect RL and MPC

Model Predictive Control

- Model driven
- Policies from planning
- Constraints oriented

Connection?

Reinforcement Learning (RL)

- Data driven
- Optimal policies from learning
- Performance oriented

Theoretical Framework to connect RL and MPC

Model Predictive Control

- Model driven
- Policies from planning
- Constraints oriented

Connection?

Reinforcement Learning (RL)

- Data driven
- Optimal policies from learning
- Performance oriented

Markov Decision Process (MDP)

- Framework to understand optimal policies
- Stochastic, discrete-time problems
- Extremely permissive mathematics
- Powerful abstraction of real-world problems

Theoretical Framework to connect RL and MPC

Model Predictive Control

- Model driven
- Policies from planning
- Constraints oriented

Connection?

Reinforcement Learning (RL)

- Data driven
- Optimal policies from learning
- Performance oriented

Solves MDP from data

Markov Decision Process (MDP)

- Framework to understand optimal policies
- Stochastic, discrete-time problems
- Extremely permissive mathematics
- Powerful abstraction of real-world problems

Theoretical Framework to connect RL and MPC

Model Predictive Control

- Model driven
- Policies from planning
- Constraints oriented

Connection?

Reinforcement Learning (RL)

- Data driven
- Optimal policies from learning
- Performance oriented

Connection?

Solves MDP from data

Markov Decision Process (MDP)

- Framework to understand optimal policies
- Stochastic, discrete-time problems
- Extremely permissive mathematics
- Powerful abstraction of real-world problems

Theoretical Framework to connect RL and MPC

Model Predictive Control

- Model driven
- Policies from planning
- Constraints oriented

Connection?

Reinforcement Learning (RL)

- Data driven
- Optimal policies from learning
- Performance oriented

Connection?

Solves MDP from data

Markov Decision Process (MDP)

- Framework to understand optimal policies
- Stochastic, discrete-time problems
- Extremely permissive mathematics
- Powerful abstraction of real-world problems

Connecting MPC and RL is about connecting MPC to MDPs!!

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)
$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$

A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)
$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)
$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

- **Closed-loop performance**

$$J\left(\boldsymbol{\pi}\right) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k)\right)\right]$$

  with discount $\gamma \in [0, 1]$

A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

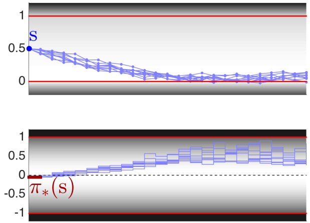
- **Closed-loop performance**

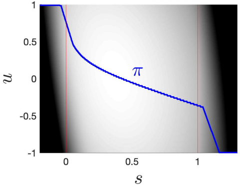
$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k)\right)\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$





A (fairly) general way of describing optimal control

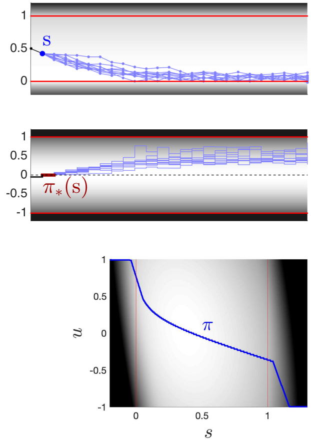
Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)
$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$



- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

is how we act on the system

- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$

A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

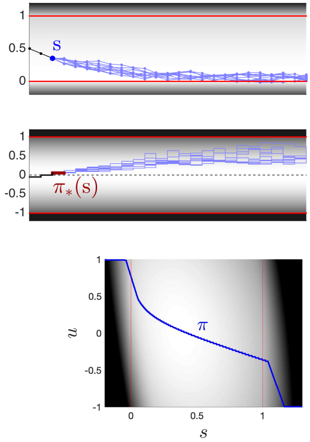
- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$



A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

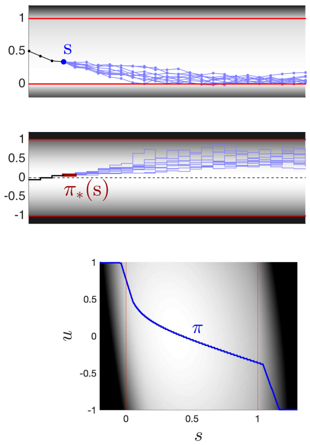
Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$



- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k)\right)\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$

A (fairly) general way of describing optimal control

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)
$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

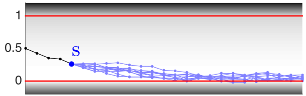
- **Closed-loop performance**

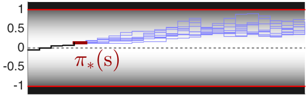
$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k)) \right]$$

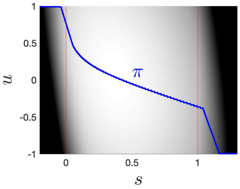
  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$







A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

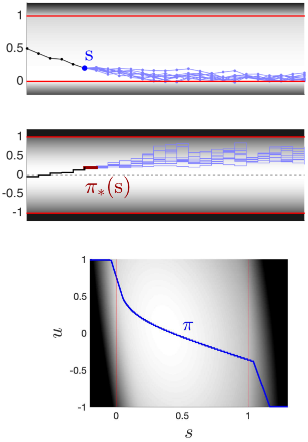
- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$



A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

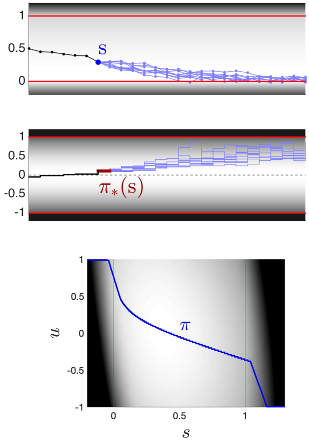
- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$



A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)
$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

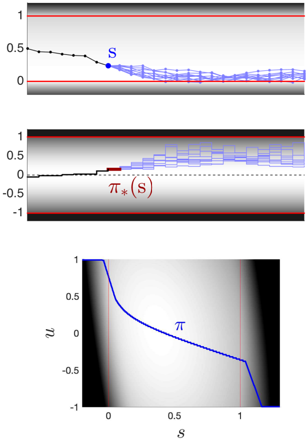
- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$



A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

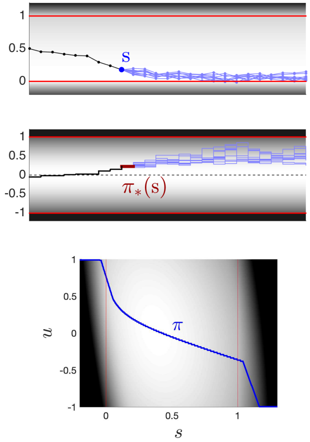
- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k)) \right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$



A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

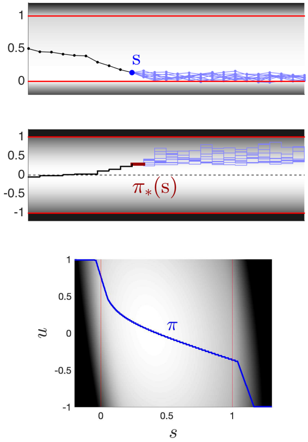
- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$



A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

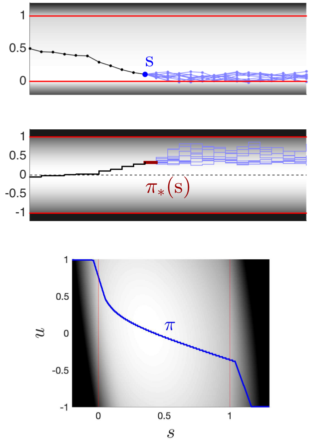
- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$



A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)
$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$

MDP is a go-to framework when considering general optimal control problems, useful for applications with stochastic dynamics.

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$

A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_{+}$$

(state-action $\rightarrow$ next state)

Cost function (instant performance)
$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$

- **Policy**

$$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

- **Closed-loop performance**

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

$$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$

MDP is a go-to framework when considering general optimal control problems, useful for applications with stochastic dynamics.

Solution of an MDP is described by "simple" equations, but solving them is very challenging

A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions

$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(state-action $\rightarrow$ next state)

- **Policy**

  $$\mathbf{a} = \boldsymbol{\pi}(\mathbf{s})$$

  is how we act on the system

- **Closed-loop performance**

  $$J(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \boldsymbol{\pi}(\mathbf{s}_k))\right]$$

  with discount $\gamma \in [0, 1]$

- **Optimal policy**: $\boldsymbol{\pi}^\star$ from

  $$\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi})$$

Cost function (instant performance)

$$L(\mathbf{s}, \mathbf{a}) \in \mathbb{R}$$

MDP is a go-to framework when considering general optimal control problems, useful for applications with stochastic dynamics.

Solution of an MDP is described by "simple" equations, but solving them is very challenging

By doing "re-planning" all the time, MPC generates a policy $\boldsymbol{\pi}^{\mathrm{MPC}}$ that *hopefully* resembles $\boldsymbol{\pi}^\star$

MPC is a heuristic to solve MDPs

why do we use it?

A (fairly) general way of describing optimal control

Purpose of MPC? *According to the MPC community*

Historically MPC focuses on **constraints satisfaction & stability**, track a reference
Tracking MPC

More recent focus is on **closed-loop performance**, e.g. energy, time, money.
Economic MPC

Purpose of MPC? *According to the MPC community*

Historically MPC focuses on **constraints satisfaction & stability**, track a reference
Tracking MPC

More recent focus is on **closed-loop performance**, e.g. energy, time, money.
Economic MPC

E.g. of the form:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix}^\top W \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

- Costs are "designed" to steer the system to reference "$(0, 0)$"
- MPC is not optimizing a specific "physical quantity" (cost unit??)

Purpose of MPC? *According to the MPC community*

Historically MPC focuses on **constraints satisfaction & stability**, track a reference
Tracking MPC

More recent focus is on **closed-loop performance**, e.g. energy, time, money.
Economic MPC

E.g. of the form:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix}^\top W \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

- Costs are "designed" to steer the system to reference "$(0,0)$"
- MPC is not optimizing a specific "physical quantity" (cost unit??)

Generic form:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

- L directly represents something we want to minimize
- MPC is optimizing a specific "physical quantity" (cost has a unit)

Purpose of MPC? *According to the MPC community*

Historically MPC focuses on **constraints satisfaction & stability**, track a reference
Tracking MPC

More recent focus is on **closed-loop performance**, e.g. energy, time, money.
Economic MPC

E.g. of the form:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix}^\top W \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

- Costs are "designed" to steer the system to reference "$(0,0)$"
- MPC is not optimizing a specific "physical quantity" (cost unit??)

Generic form:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

- L directly represents something we want to minimize
- MPC is optimizing a specific "physical quantity" (cost has a unit)

"MPC is for constraints satisfaction..." *(heard in scientific discussions)*

Purpose of MPC? *According to the MPC community*

Historically MPC focuses on **constraints satisfaction & stability**, track a reference
Tracking MPC

E.g. of the form:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix}^\top W \begin{bmatrix} \mathbf{x}_k \\ \mathbf{u}_k \end{bmatrix} \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

- Costs are "designed" to steer the system to reference "$(0,0)$"
- MPC is not optimizing a specific "physical quantity" (cost unit??)

More recent focus is on **closed-loop performance**, e.g. energy, time, money.
Economic MPC

Generic form:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

- L directly represents something we want to minimize
- MPC is optimizing a specific "physical quantity" (cost has a unit)

"MPC is for constraints satisfaction..." *(heard in scientific discussions)*

... it is, but it does not need to be limited to that.

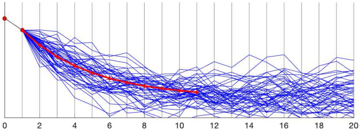
Optimal policies from MPC?

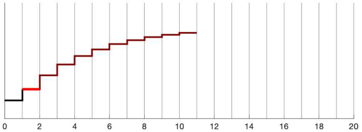
Optimality often cast as minimizing†

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





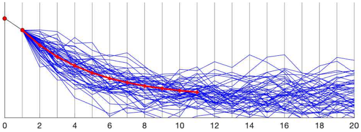
Optimal policies from MPC?

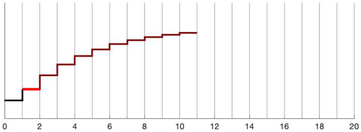
Optimality often cast as minimizing†

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





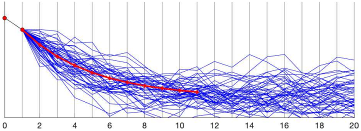
Optimal policies from MPC?

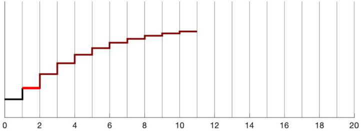
Optimality often cast as minimizing[†]

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

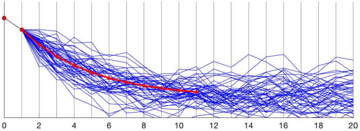
Optimal policies from MPC?

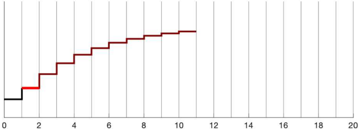
Optimality often cast as minimizing†

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

†Alternative forms of optimality: bias / gain optimal, beyond 1^{st} moment

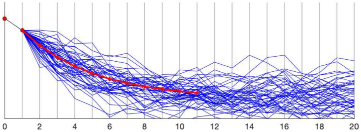
Optimal policies from MPC?

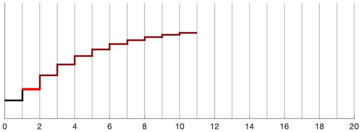
Optimality often cast as minimizing[†]

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

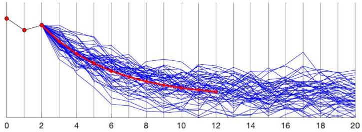
Optimal policies from MPC?

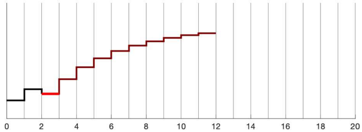
Optimality often cast as minimizing†

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}\end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

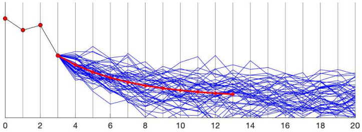
Optimal policies from MPC?

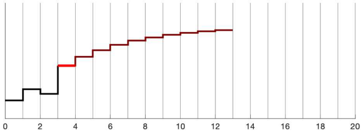
Optimality often cast as minimizing†

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

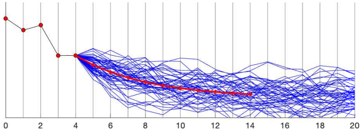
Optimal policies from MPC?

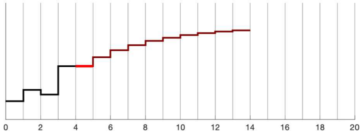
Optimality often cast as minimizing[†]

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

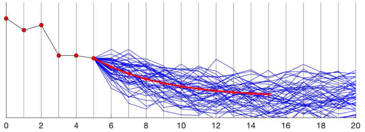
Optimal policies from MPC?

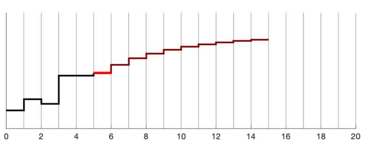
Optimality often cast as minimizing†

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \arg\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

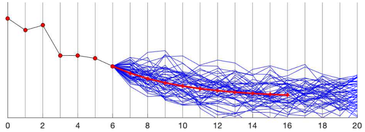
Optimal policies from MPC?

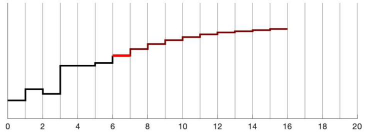
Optimality often cast as minimizing[†]

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \le 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

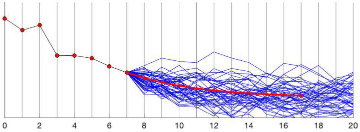
Optimal policies from MPC?

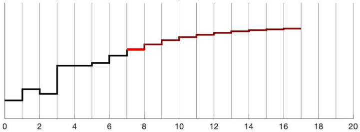
Optimality often cast as minimizing†

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

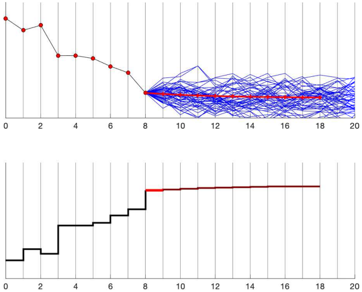
$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

Optimal policies from MPC?



Optimality often cast as minimizing[†]

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

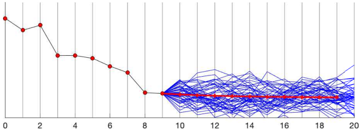
Optimal policies from MPC?

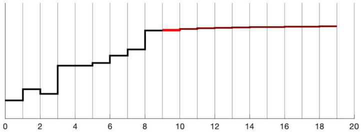
Optimality often cast as minimizing†

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

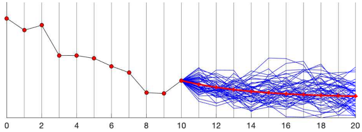
Optimal policies from MPC?

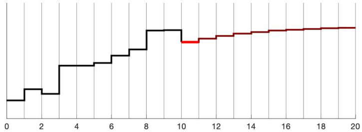
Optimality often cast as minimizing[†]

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

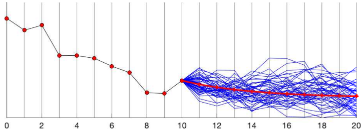
Optimal policies from MPC?

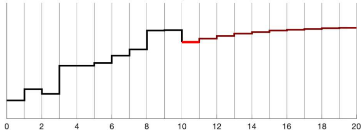
Optimality often cast as minimizing[†]

$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

Optimal policies from MPC?

Optimality often cast as minimizing[†]

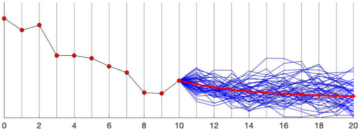
$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

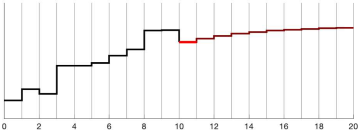
MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \le 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

Can learning help with that?

...different "bets"





Optimal policy $\boldsymbol{\pi}^\star$:

$$\boldsymbol{\pi}^\star = \mathrm{a}\min_{\boldsymbol{\pi}} \; J(\boldsymbol{\pi})$$

Does MPC give $\boldsymbol{\pi}^{\mathrm{MPC}} = \boldsymbol{\pi}^\star$?

1. Infinite vs. finite horizon
2. Discounted ($\gamma < 1$) vs. undiscounted
3. Model $\mathbf{f}$ inaccurate

... in general no

Learning for MPC - Machine Learning in-the-loop

MPC policy

$$\pi_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s})$$

from model*, e.g.

$$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$$

$\mathbf{a}$

$\mathbf{s}$

Learning for MPC - Machine Learning in-the-loop

$\boldsymbol{\theta}$

MPC policy

$\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s})$

from model*, e.g.

$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$

$\mathbf{a}$

$\mathbf{s}_+, \mathbf{s}, \mathbf{a}$

Machine-Learning

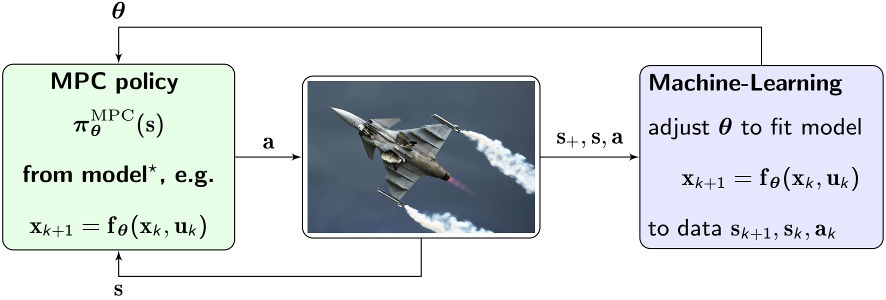
adjust $\boldsymbol{\theta}$ to fit model

$$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$$

to data $\mathbf{s}_{k+1}, \mathbf{s}_k, \mathbf{a}_k$

$\mathbf{s}$

Learning for MPC - Machine Learning in-the-loop



"Machine-Learning" in-the-loop $\mathbf{f}_{\boldsymbol{\theta}}$ from

- **Physics-based**: first principles + SYSID
- **Neural Network**: DNN, LSTM, TFT, ...
- **Statistical**: GP, RKHS, GPC, ARX ...

Learning for MPC - Machine Learning in-the-loop

$\boldsymbol{\theta}$

MPC policy

$\pi_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s})$

from model***, e.g.**

$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$

$\mathbf{a}$

$\mathbf{s}_{+}, \mathbf{s}, \mathbf{a}$

Machine-Learning

adjust $\boldsymbol{\theta}$ to fit model

$$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$$

to data $\mathbf{s}_{k+1}, \mathbf{s}_k, \mathbf{a}_k$

$\mathbf{s}$

"Machine-Learning" in-the-loop $\mathbf{f}_{\boldsymbol{\theta}}$ from

- **Physics-based**: first principles + SYSID
- **Neural Network**: DNN, LSTM, TFT, ...
- **Statistical**: GP, RKHS, GPC, ARX ...

**can replace "model" by any prediction strategies: input-output predictors, multi-step predictors, etc...*

Learning for MPC - Machine Learning in-the-loop

$\boldsymbol{\theta}$

MPC policy

$\pi_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s})$

from model*, **e.g.**

$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$

$\mathbf{a}$

$\mathbf{s}_+, \mathbf{s}, \mathbf{a}$

Machine-Learning

adjust $\boldsymbol{\theta}$ to fit model

$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$

to data $\mathbf{s}_{k+1}, \mathbf{s}_k, \mathbf{a}_k$

$\mathbf{s}$

"Machine-Learning" in-the-loop $\mathbf{f}_{\boldsymbol{\theta}}$ from

- **Physics-based**: first principles + SYSID
- **Neural Network**: DNN, LSTM, TFT, ...
- **Statistical**: GP, RKHS, GPC, ARX ...

Paradigm

- Performance tied to prediction accuracy
- Target accuracy via ML
- Ignore that MPC is a policy

** can replace "model" by any prediction strategies: input-output predictors, multi-step predictors, etc...*

Learning for MPC - Machine Learning in-the-loop

$\boldsymbol{\theta}$

MPC policy

$\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s})$

from model***, e.g.**

$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$

$\mathbf{a}$

$\mathbf{s}_+, \mathbf{s}, \mathbf{a}$

Machine-Learning

adjust $\boldsymbol{\theta}$ to fit model

$\mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k)$

to data $\mathbf{s}_{k+1}, \mathbf{s}_k, \mathbf{a}_k$

$\mathbf{s}$

"Machine-Learning" in-the-loop $\mathbf{f}_{\boldsymbol{\theta}}$ from

- **Physics-based**: first principles + SYSID
- **Neural Network**: DNN, LSTM, TFT, ...
- **Statistical**: GP, RKHS, GPC, ARX ...

**can replace "model" by any prediction strategies: input-output predictors, multi-step predictors, etc...*

Paradigm

- Performance tied to prediction accuracy
- Target accuracy via ML
- Ignore that MPC is a policy

We focus on "breaking" this paradigm
Learning / RL plays a key role

Paradigm shifts...

Paradigm shifts...

Shift 1: focus on performance instead of fitting

- **from**: $\mathbf{f}_{\boldsymbol{\theta}}$ is a model for the system dynamics
- **to**: MPC is a model of optimality (will specify that in a bit...)

Paradigm shifts...

Shift 1: focus on performance instead of fitting

- **from**: $\mathbf{f}_{\boldsymbol{\theta}}$ is a model for the system dynamics
- **to**: MPC is a model of optimality (will specify that in a bit...)

Classic view...

MPC: at current state $\mathbf{s}$ solve

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

gives policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\left(\mathbf{s}\right) = \mathbf{u}_0^\star$

Find θ such that prediction "fits" the data

Paradigm shifts...

Shift 1: focus on performance instead of fitting

- **from**: $\mathbf{f}_{\boldsymbol{\theta}}$ is a model for the system dynamics
- **to**: MPC is a model of optimality (will specify that in a bit...)

Classic view...

MPC: at current state $\mathbf{s}$ solve

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

gives policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\left(\mathbf{s}\right) = \mathbf{u}_0^\star$

Find $\boldsymbol{\theta}$ such that prediction "fits" the data

Shift to...

Find $\boldsymbol{\theta}$ that "fits MPC to optimality" according to the data, e.g. minimizes $J(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}})$

Paradigm shifts...

Shift 1: focus on performance instead of fitting

- **from**: $\mathbf{f}_{\boldsymbol{\theta}}$ is a model for the system dynamics
- **to**: MPC is a model of optimality (will specify that in a bit...)

Classic view...

MPC: at current state $\mathbf{s}$ solve

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

gives policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$

Find $\boldsymbol{\theta}$ such that prediction "fits" the data

Shift to...

Find $\boldsymbol{\theta}$ that "fits MPC to optimality" according to the data, e.g. minimizes $J(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}})$

- $\rightarrow$ Best model for closed-loop performance
- $\neq$ Best model to fit the data!
- *More on this in* 3^{rd} *lecture*

RL is a toolbox to do that...

Paradigm shifts...

Shift 1: focus on performance instead of fitting

- **from**: $\mathbf{f}_{\boldsymbol{\theta}}$ is a model for the system dynamics
- **to**: MPC is a model of optimality (will specify that in a bit...)

Classic view...

MPC: at current state $\mathbf{s}$ solve

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}$$

gives policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\left(\mathbf{s}\right) = \mathbf{u}_0^\star$

Find $\boldsymbol{\theta}$ such that prediction "fits" the data

Shift to...

Find $\boldsymbol{\theta}$ that "fits MPC to optimality" according to the data, e.g. minimizes $J(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}})$

- $\rightarrow$ Best model for closed-loop performance
- $\neq$ Best model to fit the data!
- *More on this in 3$^{\mathrm{rd}}$ lecture*

RL is a toolbox to do that...

But getting $\boldsymbol{\pi}^\star$ places "high demands" on $\mathbf{f}_{\boldsymbol{\theta}}$

Paradigm shifts...

Shift 1: focus on performance instead of fitting

- **from**: $\mathbf{f}_{\boldsymbol{\theta}}$ is a model for the system dynamics
- **to**: MPC is a model of optimality (will specify that in a bit...)

Classic view...

MPC: at current state $\mathbf{s}$ solve

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\
& \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}$$

gives policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$

Find $\boldsymbol{\theta}$ such that prediction "fits" the data

Shift to...

Find $\boldsymbol{\theta}$ that "fits MPC to optimality" according to the data, e.g. minimizes $J(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}})$

- $\rightarrow$ Best model for closed-loop performance
- $\neq$ Best model to fit the data!
- *More on this in 3rd lecture*

RL is a toolbox to do that...

But getting $\boldsymbol{\pi}^\star$ places "high demands" on $\mathbf{f}_{\boldsymbol{\theta}}$

Can we do more? **Yes**...

Paradigm shifts...

Shift 1: focus on performance instead of fitting

- **from**: $\mathbf{f}_{\boldsymbol{\theta}}$ is a model for the system dynamics
- **to**: MPC is a model of optimality (will specify that in a bit...)

Classic view...

MPC: at current state $\mathbf{s}$ solve

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

gives policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$

Find $\boldsymbol{\theta}$ such that prediction "fits" the data

Shift to...

Find $\boldsymbol{\theta}$ that "fits MPC to optimality" according to the data, e.g. minimizes $J(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}})$

Shift 2: "holistic" parametrization

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

gives policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$

How to use this? Reinforcement Learning

Policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\
\mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\
& \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$

- $\min_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$ using data
- $\boldsymbol{\theta} \rightarrow J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$ very implicit
- $J(.)$ is the real-system!

How to use this? Reinforcement Learning

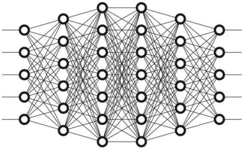
Policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

- $\min_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$ using data
- $\boldsymbol{\theta} \rightarrow J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$ very implicit
- $J(.)$ is the real-system!

Reinforcement Learning

Tools to approximate $\pi^\star$ from data
This is not (necessarily) about DNNs



How to use this? Reinforcement Learning

Policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \le 0, \quad \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

- $\min_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$ using data
- $\boldsymbol{\theta} \to J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$ very implicit
- $J(.)$ is the real-system!

Reinforcement Learning

Tools to approximate $\pi^\star$ from data
This is not (necessarily) about DNNs

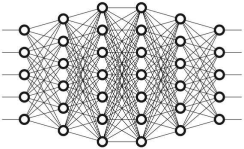
For MPC: tools to find best $\boldsymbol{\theta}$, e.g.

- **Policy Gradient**: estimations of

$$\nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right), \quad \text{possibly} \quad \nabla_{\boldsymbol{\theta}}^2 J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$$

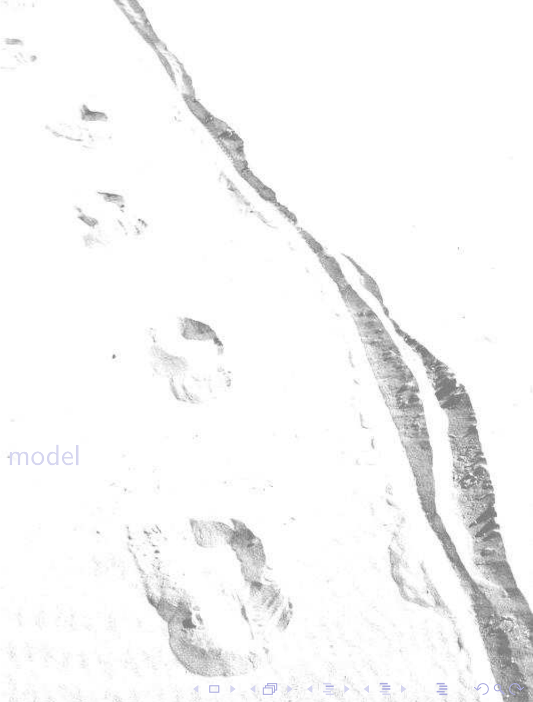
- **Q-learning**: direct "shaping" of MPC

Combination is useful...



Outline

1. The Basics
2. More background
3. Let's take a deeper dive
4. Parametrization & Role of the model
5. RL over MPC



Optimal Value Functions

- **Value function**:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\, \mathbf{s}_0=\mathbf{s},\ \mathbf{a}_k=\boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$

Optimal Value Functions

- **Value function**:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\,\middle|\, \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_k = \boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$

- **Action-Value function**:

$$Q_\star(\mathbf{s}, \mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\,\middle|\, \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_0 = \mathbf{a},\ \mathbf{a}_{k>0} = \boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$, and using action $\mathbf{a}$ as first input (policy $\boldsymbol{\pi}_\star$ after that)

Optimal Value Functions

- **Value function**:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\, \mathbf{s}_0=\mathbf{s},\ \mathbf{a}_k=\boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

  gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$

- **Action-Value function**:

$$Q_\star(\mathbf{s},\mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\, \mathbf{s}_0=\mathbf{s},\ \mathbf{a}_0=\mathbf{a},\ \mathbf{a}_{k>0}=\boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

  gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$, and using action $\mathbf{a}$ as first input (policy $\boldsymbol{\pi}_\star$ after that)

- **Relationship**:

$$V_\star(\mathbf{s}) = \min_{\mathbf{a}}\ Q_\star(\mathbf{s},\mathbf{a})$$

Optimal Value Functions

- **Value function**:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_k = \boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$

- **Action-Value function**:

$$Q_\star(\mathbf{s}, \mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_0 = \mathbf{a},\ \mathbf{a}_{k>0} = \boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$, and using action $\mathbf{a}$ as first input (policy $\boldsymbol{\pi}_\star$ after that)

- **Relationship**:

$$V_\star(\mathbf{s}) = \min_{\mathbf{a}}\ Q_\star(\mathbf{s}, \mathbf{a})$$

- **Optimal Policy**:

$$\boldsymbol{\pi}_\star(\mathbf{s}) = \arg\min_{\mathbf{a}}\ Q_\star(\mathbf{s}, \mathbf{a})$$

Optimal Value Functions

- **Value function**:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_k = \boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$

- **Action-Value function**:

$$Q_\star(\mathbf{s}, \mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_0 = \mathbf{a},\ \mathbf{a}_{k>0} = \boldsymbol{\pi}_\star(\mathbf{s}_k)\right]$$

gives the expected cost for policy $\boldsymbol{\pi}_\star$, starting from given initial conditions $\mathbf{s}$, and using action $\mathbf{a}$ as first input (policy $\boldsymbol{\pi}_\star$ after that)

- **Relationship**:

$$V_\star(\mathbf{s}) = \min_{\mathbf{a}}\ Q_\star(\mathbf{s}, \mathbf{a})$$

- **Optimal Policy**:

$$\boldsymbol{\pi}_\star(\mathbf{s}) = \arg\min_{\mathbf{a}}\ Q_\star(\mathbf{s}, \mathbf{a})$$

Can be computed via the Bellman equations, intractable for “large” state-action spaces

Value Functions

- **Value function**:

$$V_{\pi}(\mathbf{s}) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\,\middle|\, \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

  gives the expected cost for policy $\boldsymbol{\pi}$, starting from given initial conditions $\mathbf{s}$

- **Action-Value function**:

$$Q_{\pi}(\mathbf{s}, \mathbf{a}) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\,\middle|\, \mathbf{s}_0 = \mathbf{s},\ \mathbf{a}_0 = \mathbf{a},\ \mathbf{a}_{k>0} = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

  gives the expected cost for policy $\boldsymbol{\pi}$, starting from given initial conditions $\mathbf{s}$, and using action $\mathbf{a}$ as first input (policy $\boldsymbol{\pi}_\star$ after that)

- **Relationship**:

$$V_{\pi}(\mathbf{s}) = Q_{\pi}(\mathbf{s}, \boldsymbol{\pi}(\mathbf{s}_k))$$

- **Advantage function**:

$$A_{\pi}(\mathbf{s}, \mathbf{a}) = Q_{\pi}(\mathbf{s}, \mathbf{a}) - V_{\pi}(\mathbf{s})$$

  compares $\mathbf{a}$ to policy $\boldsymbol{\pi}$. Instrumental in policy gradient methods.

Note:

$$V_\pi \neq V_\star$$
$$Q_\pi \neq Q_\star$$
$$A_\pi \neq A_\star$$

Can be computed via the Bellman equations, intractable for "large" state-action

MDPs and "forbidden" states

What if the system is not allowed to leave a certain subset of the state space?

MDPs and "forbidden" states

What if the system is not allowed to leave a certain subset of the state space?

- Say there is a "feasible" set:

$$\mathbb{F} = \{\ \mathbf{s} \quad | \quad \mathbf{h}(\mathbf{s}) \leq 0\ \}$$

  where the state of the system should always be.

MDPs and "forbidden" states

What if the system is not allowed to leave a certain subset of the state space?

- Say there is a "feasible" set:

$$\mathbb{F} = \{ \mathbf{s} \quad | \quad \mathbf{h}(\mathbf{s}) \leq 0 \}$$

where the state of the system should always be.

- In the "MDP theory", assign an infinite penalty to leaving $\mathbb{F}$, i.e. add:

$$\mathrm{I}_{\mathbb{F}}(\mathbf{s}, \mathbf{a}) = \begin{cases} 0 & \text{if} \quad \mathbf{s} \in \mathbb{F} \\ +\infty & \text{if} \quad \mathbf{s} \notin \mathbb{F} \end{cases}$$

to stage cost L.

MDPs and "forbidden" states

What if the system is not allowed to leave a certain subset of the state space?

- Say there is a "feasible" set:

$$\mathbb{F} = \{\ \mathbf{s} \quad | \quad \mathbf{h}(\mathbf{s}) \leq 0\ \}$$

  where the state of the system should always be.

- In the "MDP theory", assign an infinite penalty to leaving $\mathbb{F}$, i.e. add:

$$\mathrm{I}_{\mathbb{F}}(\mathbf{s}, \mathbf{a}) = \begin{cases} 0 & \text{if} \quad \mathbf{s} \in \mathbb{F} \\ +\infty & \text{if} \quad \mathbf{s} \notin \mathbb{F} \end{cases}$$

  to stage cost L.

- In RL, ∞ penalties are not meaningful: "There is no backup from death"

MDPs and "forbidden" states

What if the system is not allowed to leave a certain subset of the state space?

- Say there is a "feasible" set:

$$\mathbb{F} = \{ \mathbf{s} \quad | \quad \mathbf{h}(\mathbf{s}) \leq 0 \}$$

  where the state of the system should always be.

- In the "MDP theory", assign an infinite penalty to leaving $\mathbb{F}$, i.e. add:

$$\mathrm{I}_{\mathbb{F}}(\mathbf{s}, \mathbf{a}) = \begin{cases} 0 & \text{if} \quad \mathbf{s} \in \mathbb{F} \\ +\infty & \text{if} \quad \mathbf{s} \notin \mathbb{F} \end{cases}$$

  to stage cost L.

- In RL, ∞ penalties are not meaningful: "There is no backup from death"
- Common approach: assign a "very large" penalty to $\mathrm{s} \notin \mathbb{F}$ instead of $+\infty$.

MDPs and "forbidden" states

What if the system is not allowed to leave a certain subset of the state space?

- Say there is a "feasible" set:

$$\mathbb{F} = \{\ \mathbf{s} \quad | \quad \mathbf{h}(\mathbf{s}) \leq 0\ \}$$

  where the state of the system should always be.

- In the "MDP theory", assign an infinite penalty to leaving $\mathbb{F}$, i.e. add:

$$\mathrm{I}_{\mathbb{F}}(\mathbf{s}, \mathbf{a}) = \begin{cases} 0 & \text{if} \quad \mathbf{s} \in \mathbb{F} \\ +\infty & \text{if} \quad \mathbf{s} \notin \mathbb{F} \end{cases}$$

  to stage cost L.

- In RL, ∞ penalties are not meaningful: "There is no backup from death"
- Common approach: assign a "very large" penalty to $\mathbf{s} \notin \mathbb{F}$ instead of $+\infty$.
- Use of "barrier functions" in RL

Why discounting?

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,\middle|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

Discounting is (in general) needed to make the MDP well defined, is that all?

Can we give an interpretation of discounting?

Why discounting?

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,\middle|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

Discounting is (in general) needed to make the MDP well defined, is that all?

Can we give an interpretation of discounting?

System lifetime: assuming that the system can (irremediably) fail at any time k with probability $1-\gamma$, then discounting accounts for resulting probabilistic lifetime.

Why discounting?

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k,\mathbf{a}_k\,\right]$$

Discounting is (in general) needed to make the MDP well defined, is that all?

Can we give an interpretation of discounting?

System lifetime: assuming that the system can (irremediably) fail at any time k with probability $1-\gamma$, then discounting accounts for resulting probabilistic lifetime.

E.g. a system with a sampling time of 1 second, and a 90% chance of having a lifetime of 20 years, should have $\gamma = 0.999999996349275$

Why discounting?

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

Discounting is (in general) needed to make the MDP well defined, is that all?

Can we give an interpretation of discounting?

Investment model: expected economic growth r (per time unit) implies that earning at time k is worth $(1+r)^{-k}$ the same earning at time 0. Hence $\gamma = (1+r)^{-1}$.

Why discounting?

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

Discounting is (in general) needed to make the MDP well defined, is that all?

Can we give an interpretation of discounting?

Investment model: expected economic growth r (per time unit) implies that earning at time k is worth $(1+r)^{-k}$ the same earning at time 0. Hence $\gamma = (1+r)^{-1}$.

E.g. a system with a sampling time of 1 second and an expected return of 10% per year should have $\gamma = 0.999999999848887$

Why discounting?

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

Discounting is (in general) needed to make the MDP well defined, is that all?

Can we give an interpretation of discounting?

Investment model: expected economic growth r (per time unit) implies that earning at time k is worth $(1+r)^{-k}$ the same earning at time 0. Hence $\gamma = (1+r)^{-1}$.

E.g. a system with a sampling time of 1 second and an expected return of 10% per year should have $\gamma = 0.999999999848887$

<u>Bottom line</u>: on "engineering applications", the discount tends to (should) be extremely close to 1

Why discounting?

Gain optimal MDP:

$$\min_{\boldsymbol{\pi}} \quad \lim_{N \to \infty} \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{k=0}^{N} \frac{1}{N} L\left(\mathbf{s}_k, \mathbf{a}_k\right) \right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,\middle|\, \mathbf{s}_k, \mathbf{a}_k\,\right]$$

What about considering average cost?

Policy $\boldsymbol{\pi}$

- is said to achieve "gain optimality"
- transients are irrelelvant as they have no contribution in the average return
- tends to yield "bang-bang" actions until optimal steady state is reached
- is not unique!

Why discounting?

Gain optimal MDP:

$$\min_{\boldsymbol{\pi}} \quad \lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{N} \frac{1}{N} L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

What about considering average cost?

Policy $\boldsymbol{\pi}$

- is said to achieve "gain optimality"
- transients are irrelelvant as they have no contribution in the average return
- tends to yield "bang-bang" actions until optimal steady state is reached
- is not unique!

... gain optimal policies are of questionable use for control

Why discounting?

Bias optimal MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{N} L\left(\mathbf{s}_k, \mathbf{a}_k\right) - V_{\mathrm{G}}^{\star}\left(\mathbf{s}_0\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,\middle|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

What about "removing" the average cost?

where $V_{\mathrm{G}}^{\star}$ is the value function associated to gain optimal problem.

Policy $\boldsymbol{\pi}$

- is said to achieve "bias optimality"
- "best transient to gain-optimal state"
- there are RL algorithms for bias optimality

Why discounting?

Bias optimal MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{N} L\left(\mathbf{s}_k, \mathbf{a}_k\right) - V_{\mathrm{G}}^{\star}\left(\mathbf{s}_0\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,\middle|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

What about "removing" the average cost?

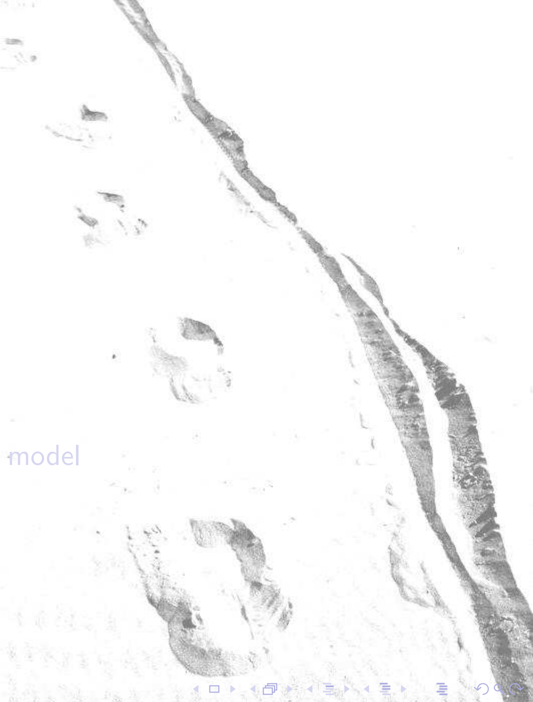
where $V_{\mathrm{G}}^{\star}$ is the value function associated to gain optimal problem.

Policy π

- is said to achieve "bias optimality"
- "best transient to gain-optimal state"
- there are RL algorithms for bias optimality

The ideas we discuss here work for all cases. Discounted problems tend to yield "more meaningful" behavior. Discounting create some challenges for stability theory though. More on this in a bit.

Outline



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\,\mathbf{a}_k=\boldsymbol{\pi}(\mathbf{s}_k)\right]$$

Value functions:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\right]$$

$$Q_\star(\mathbf{s},\mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\,\mathbf{a}_0=\mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} Q_\star(\mathbf{s},\mathbf{a})$$

MPC

How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

Value functions:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\right]$$

$$Q_\star(\mathbf{s}, \mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_0 = \mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} Q_\star(\mathbf{s}, \mathbf{a})$$

MPC

$$\begin{aligned} V^{\mathrm{MPC}}(\mathbf{s}) = \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

yields $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ as by-product

How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \middle| \mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

Value functions:

$$V_\star\left(\mathbf{s}\right) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

$$Q_\star\left(\mathbf{s}, \mathbf{a}\right) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \middle| \mathbf{a}_0 = \mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star\left(\mathbf{s}\right) = \mathrm{a}\min_{\mathbf{a}} Q_\star\left(\mathbf{s}, \mathbf{a}\right)$$

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) = \min_{\mathbf{x},\mathbf{u}} \; & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\, \mathbf{a}_k=\boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

Value functions:

$$V_\star\left(\mathbf{s}\right)=\mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\right]$$

$$Q_\star\left(\mathbf{s},\mathbf{a}\right)=\mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\,\mathbf{a}_0=\mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star\left(\mathbf{s}\right)=\mathrm{a}\min_{\mathbf{a}} Q_\star\left(\mathbf{s},\mathbf{a}\right)$$

MPC

$$\begin{aligned}Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right)=\min_{\mathbf{x},\mathbf{u}}\;& T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1}L\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ \mathrm{s.t.}\quad & \mathbf{x}_{k+1}=\mathbf{f}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ & \mathbf{h}\left(\mathbf{x}_k,\mathbf{u}_k\right)\leq 0\\ & \mathbf{x}_0=\mathbf{s},\quad \mathbf{u}_0=\mathbf{a}\end{aligned}$$

MPC is **consistent**, i.e.

$$V^{\mathrm{MPC}}\left(\mathbf{s}\right)=\min_{\mathbf{a}}\; Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right)$$

$$\boldsymbol{\pi}^{\mathrm{MPC}}\left(\mathbf{s}\right)=\arg\min_{\mathbf{a}}\; Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right)$$

$\rightarrow$ **“sound representation” of MDP**

How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

MPC is optimal if:

$$\boldsymbol{\pi}^{\mathrm{MPC}}\left(\mathbf{s}\right) = \boldsymbol{\pi}_\star\left(\mathbf{s}\right)$$

for all $\mathbf{s}$

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) = \min_{\mathbf{x},\mathbf{u}} \; & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

MPC is **consistent**, i.e.

$$\begin{aligned} V^{\mathrm{MPC}}\left(\mathbf{s}\right) &= \min_{\mathbf{a}} \; Q^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) \\ \boldsymbol{\pi}^{\mathrm{MPC}}\left(\mathbf{s}\right) &= \arg\min_{\mathbf{a}} \; Q^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) \end{aligned}$$

$\rightarrow$ **"sound representation" of MDP**

How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC is optimal if:

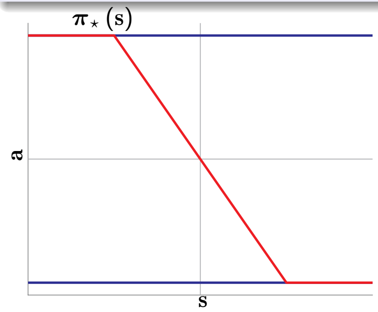
$$\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \boldsymbol{\pi}_\star(\mathbf{s})$$

for all $\mathbf{s}$

MPC

$$Q^{\mathrm{MPC}}(\mathbf{s}, \mathbf{a}) = \min_{\mathbf{x},\mathbf{u}} T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k)$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

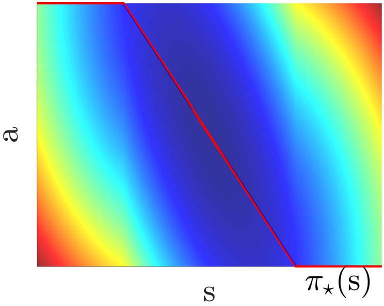
MPC is optimal if:

$$\boldsymbol{\pi}^{\mathrm{MPC}}\left(\mathbf{s}\right) = \boldsymbol{\pi}_\star\left(\mathbf{s}\right)$$

for all s

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) = \min_{\mathbf{x},\mathbf{u}} \; & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

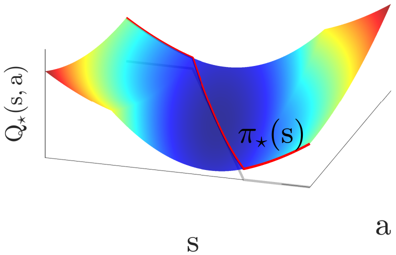
MPC is optimal if:

$$\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \boldsymbol{\pi}_\star(\mathbf{s})$$

for all $\mathbf{s}$

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}(\mathbf{s}, \mathbf{a}) = \min_{\mathbf{x},\mathbf{u}} \; & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

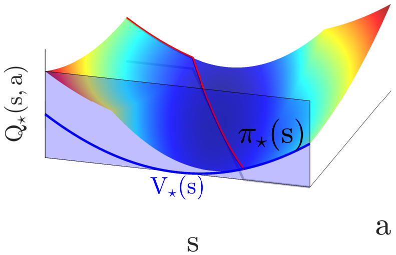
MPC is optimal if:

$$\boldsymbol{\pi}^{\mathrm{MPC}}\left(\mathbf{s}\right) = \boldsymbol{\pi}_\star\left(\mathbf{s}\right)$$

for all $\mathbf{s}$

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) = \min_{\mathbf{x},\mathbf{u}} \quad & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\,\mathbf{a}_k=\boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

MPC is optimal if:

$$\boldsymbol{\pi}^{\mathrm{MPC}}\left(\mathbf{s}\right)=\boldsymbol{\pi}_\star\left(\mathbf{s}\right)$$

for all $\mathbf{s}$

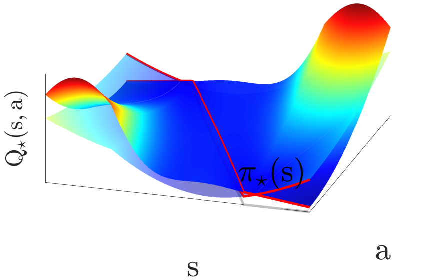
MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right)=\min_{\mathbf{x},\mathbf{u}}\ & T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1}L\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ \mathrm{s.t.}\quad & \mathbf{x}_{k+1}=\mathbf{f}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ & \mathbf{h}\left(\mathbf{x}_k,\mathbf{u}_k\right)\leq 0\\ & \mathbf{x}_0=\mathbf{s},\quad \mathbf{u}_0=\mathbf{a}\end{aligned}$$

But optimality implies only

$$\arg\max_{\mathbf{a}} Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right)=\arg\max_{\mathbf{a}} Q_\star\left(\mathbf{s},\mathbf{a}\right)$$

Optimal MPC can still be an "incomplete" model of the MDP, i.e. not a model of the value of states and actions.



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\,\mathbf{a}_k=\boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

MPC is optimal if:

$$\boldsymbol{\pi}^{\mathrm{MPC}}\left(\mathbf{s}\right)=\boldsymbol{\pi}_\star\left(\mathbf{s}\right)$$

for all $\mathbf{s}$

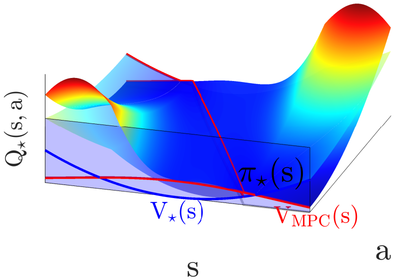
MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right)=\min_{\mathbf{x},\mathbf{u}}\; & T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1}L\left(\mathbf{x}_k,\mathbf{u}_k\right) \\ \mathrm{s.t.}\quad & \mathbf{x}_{k+1}=\mathbf{f}\left(\mathbf{x}_k,\mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k,\mathbf{u}_k\right)\leq 0 \\ & \mathbf{x}_0=\mathbf{s},\quad \mathbf{u}_0=\mathbf{a} \end{aligned}$$

But optimality implies only

$$\arg\max_{\mathbf{a}} Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right)=\arg\max_{\mathbf{a}} Q_\star\left(\mathbf{s},\mathbf{a}\right)$$

Optimal MPC can still be an "incomplete" model of the MDP, i.e. not a model of the value of states and actions.



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \middle| \mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

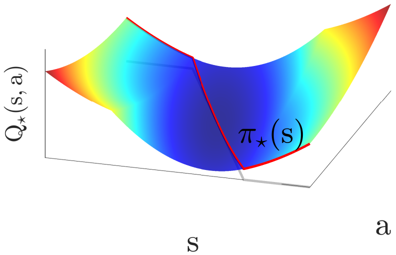
MPC is a complete MDP model if:

$$Q^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) = Q_\star\left(\mathbf{s}, \mathbf{a}\right)$$

for all $\mathbf{s}, \mathbf{a}$

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) = \min_{\mathbf{x},\mathbf{u}} \; & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\,\mathbf{a}_k=\boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}(\mathbf{s},\mathbf{a}) = \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N)+\sum_{k=0}^{N-1}L(\mathbf{x}_k,\mathbf{u}_k) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1}=\mathbf{f}(\mathbf{x}_k,\mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k,\mathbf{u}_k)\leq 0 \\ & \mathbf{x}_0=\mathbf{s},\quad \mathbf{u}_0=\mathbf{a} \end{aligned}$$

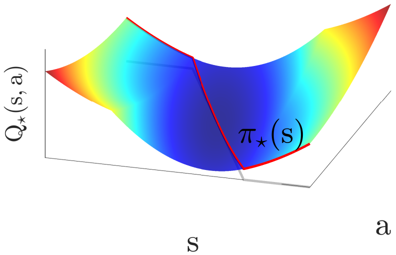
MPC is a complete MDP model if:

$$Q^{\mathrm{MPC}}(\mathbf{s},\mathbf{a}) = Q_\star(\mathbf{s},\mathbf{a})$$

for all $\mathbf{s},\mathbf{a}$

Completeness implies optimality i.e.

$$\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \boldsymbol{\pi}_\star(\mathbf{s})$$



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

MPC

$$Q^{\mathrm{MPC}}(\mathbf{s}, \mathbf{a}) = \min_{\mathbf{x},\mathbf{u}} T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k)$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

MPC is a complete MDP model if:

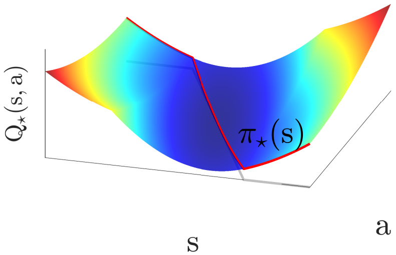
$$Q^{\mathrm{MPC}}(\mathbf{s}, \mathbf{a}) = Q_\star(\mathbf{s}, \mathbf{a})$$

for all $\mathbf{s}, \mathbf{a}$

Completeness implies optimality i.e.

$$\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \boldsymbol{\pi}_\star(\mathbf{s})$$

Matching the MPC action-value function to the optimal one is desirable



How does MPC model an MDP?

Markov Decision Process:

$$\min_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{a}_k = \pi(\mathbf{s}_k)\right]$$

MPC

$$\begin{aligned} Q_{\theta}^{\mathrm{MPC}}(\mathbf{s}, \mathbf{a}) = \min_{\mathbf{x},\mathbf{u}} \quad & T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\theta}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

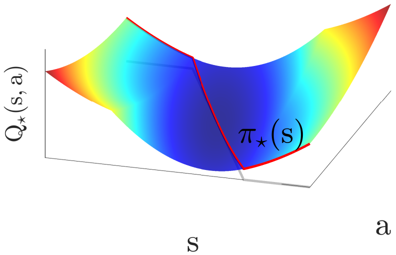
Find θ such that

$$Q_{\theta}^{\mathrm{MPC}}(\mathbf{s}, \mathbf{a}) = Q_{\star}(\mathbf{s}, \mathbf{a})$$

for all $\mathbf{s}, \mathbf{a}$?

High demand in the MPC model !

But more on that in Lecture 3...



More on the full MPC parametrization

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

MPC

Value functions:

$$V_\star\left(\mathbf{s}\right) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

$$Q_\star\left(\mathbf{s}, \mathbf{a}\right) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_0 = \mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star\left(\mathbf{s}\right) = \mathrm{a}\min_{\mathbf{a}} Q_\star\left(\mathbf{s}, \mathbf{a}\right)$$

More on the full MPC parametrization

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\, \mathbf{a}_k=\boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

Value functions:

$$V_\star\left(\mathbf{s}\right)=\mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\right]$$

$$Q_\star\left(\mathbf{s},\mathbf{a}\right)=\mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\,\mathbf{a}_0=\mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star\left(\mathbf{s}\right)=\mathrm{a}\min_{\mathbf{a}} Q_\star\left(\mathbf{s},\mathbf{a}\right)$$

MPC

$$\begin{aligned} V^{\mathrm{MPC}}\left(\mathrm{s}\right)=\min_{\mathbf{x},\mathbf{u}}\quad & T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1}L\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ \mathrm{s.t.}\quad & \mathbf{x}_{k+1}=\mathbf{f}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ & \mathbf{h}\left(\mathbf{x}_k,\mathbf{u}_k\right)\leq 0\\ & \mathbf{x}_0=\mathbf{s}\end{aligned}$$

yields $\boldsymbol{\pi}^{\mathrm{MPC}}\left(\mathrm{s}\right)=\mathbf{u}_0^\star$ as by-product

More on the full MPC parametrization

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

Value functions:

$$V_\star\left(\mathbf{s}\right) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\right]$$

$$Q_\star\left(\mathbf{s},\mathbf{a}\right) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\, \mathbf{a}_0 = \mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star\left(\mathbf{s}\right) = \mathrm{a}\min_{\mathbf{a}} Q_\star\left(\mathbf{s},\mathbf{a}\right)$$

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right) = \min_{\mathbf{x},\mathbf{u}} \quad & T\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L\left(\mathbf{x}_k,\mathbf{u}_k\right) \\ \mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k,\mathbf{u}_k\right) \\ & \mathbf{h}\left(\mathbf{x}_k,\mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

More on the full MPC parametrization

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\,\mathbf{a}_k=\boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

Value functions:

$$V_\star\left(\mathbf{s}\right)=\mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\right]$$

$$Q_\star\left(\mathbf{s},\mathbf{a}\right)=\mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\,\mathbf{a}_0=\mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star\left(\mathbf{s}\right)=\mathrm{a}\min_{\mathbf{a}} Q_\star\left(\mathbf{s},\mathbf{a}\right)$$

MPC

$$\begin{aligned} Q^{\mathrm{MPC}}\left(\mathbf{s},\mathbf{a}\right)=\min_{\mathbf{x},\mathbf{u}}\quad & T\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1}L\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ \mathrm{s.t.}\quad & \mathbf{x}_{k+1}=\mathbf{f}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ & \mathbf{h}\left(\mathbf{x}_k,\mathbf{u}_k\right)\leq 0\\ & \mathbf{x}_0=\mathbf{s},\quad \mathbf{u}_0=\mathbf{a}\end{aligned}$$

In general

$$\boldsymbol{\pi}^{\mathrm{MPC}}\neq\boldsymbol{\pi}_\star,\ V^{\mathrm{MPC}}\neq V_\star,\ Q^{\mathrm{MPC}}\neq Q_\star$$

but...

More on the full MPC parametrization

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\,\mathbf{a}_k=\boldsymbol{\pi}(\mathbf{s}_k)\right]$$

Value functions:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\right]$$

$$Q_\star(\mathbf{s},\mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\,\mathbf{a}_0=\mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} Q_\star(\mathbf{s},\mathbf{a})$$

MPC

$$\begin{aligned} Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) = \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k,\mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k,\mathbf{u}_k) \\ & \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k,\mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

In general

$$\boldsymbol{\pi}^{\mathrm{MPC}} \neq \boldsymbol{\pi}_\star,\ V^{\mathrm{MPC}} \neq V_\star,\ Q^{\mathrm{MPC}} \neq Q_\star$$

but...

More on the full MPC parametrization

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\,\mathbf{a}_k=\boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

Value functions:

$$V_\star\left(\mathbf{s}\right)=\mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\right]$$

$$Q_\star\left(\mathbf{s},\mathbf{a}\right)=\mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\,\middle|\,\mathbf{a}_0=\mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star\left(\mathbf{s}\right)=\mathrm{a}\min_{\mathbf{a}} Q_\star\left(\mathbf{s},\mathbf{a}\right)$$

MPC

$$\begin{aligned} Q_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right)=\min_{\mathbf{x},\mathbf{u}}\quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right)+\sum_{k=0}^{N-1}L_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ \text{s.t.}\quad & \mathbf{x}_{k+1}=\mathbf{f}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ & \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right)\leq 0\\ & \mathbf{x}_0=\mathbf{s},\quad \mathbf{u}_0=\mathbf{a}\end{aligned}$$

Theorem: under some assumptions

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}=\boldsymbol{\pi}_\star,\quad V_{\boldsymbol{\theta}}=V_\star,\quad Q_{\boldsymbol{\theta}}=Q_\star$$

hold for some $T_{\boldsymbol{\theta}}$, $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$

More on the full MPC parametrization

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\,\mathbf{a}_k=\boldsymbol{\pi}(\mathbf{s}_k)\right]$$

Value functions:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\right]$$

$$Q_\star(\mathbf{s},\mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\,\mathbf{a}_0=\mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} Q_\star(\mathbf{s},\mathbf{a})$$

MPC

$$\begin{aligned} Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) = \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k,\mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k,\mathbf{u}_k) \\ & \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k,\mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

Theorem: under some assumptions

$$\boldsymbol{\pi}_{\boldsymbol{\theta}} = \boldsymbol{\pi}_\star, \quad V_{\boldsymbol{\theta}} = V_\star, \quad Q_{\boldsymbol{\theta}} = Q_\star$$

hold for some $T_{\boldsymbol{\theta}}$, $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$

- MPC can "capture" $\boldsymbol{\pi}_\star$, $Q_\star$, $V_\star$, **even if MPC model is inaccurate**
- Requires modifications of the stage cost & constraints
- Valid for all MPC schemes (classic, robust, stochastic, economic, etc)

More on the full MPC parametrization

Markov Decision Process:

$$\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)\right]$$

Value functions:

$$V_\star\left(\mathbf{s}\right) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

$$Q_\star\left(\mathbf{s}, \mathbf{a}\right) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right) \,\middle|\, \mathbf{a}_0 = \mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star\left(\mathbf{s}\right) = \mathrm{a}\min_{\mathbf{a}} Q_\star\left(\mathbf{s}, \mathbf{a}\right)$$

MPC

$$\begin{aligned} Q_{\boldsymbol{\theta}}\left(\mathbf{s}, \mathbf{a}\right) = \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

Theorem: under some assumptions

$$\boldsymbol{\pi}_{\boldsymbol{\theta}} = \boldsymbol{\pi}_\star, \quad V_{\boldsymbol{\theta}} = V_\star, \quad Q_{\boldsymbol{\theta}} = Q_\star$$

hold for some $T_{\boldsymbol{\theta}}$, $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$

Learning+MPC where cost & constraints are adjusted is formally justified

“Holistic” view of MPC: model for $Q_\star$, cost & constraints are part of that

More on the full MPC parametrization

Markov Decision Process:

$$\min_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\, \mathbf{a}_k = \boldsymbol{\pi}(\mathbf{s}_k)\right]$$

Value functions:

$$V_\star(\mathbf{s}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\right]$$

$$Q_\star(\mathbf{s},\mathbf{a}) = \mathbb{E}_{\boldsymbol{\pi}_\star}\left[\sum_{k=0}^{\infty}\gamma^k L(\mathbf{s}_k,\mathbf{a}_k)\,\middle|\, \mathbf{a}_0 = \mathbf{a}\right]$$

$$\boldsymbol{\pi}_\star(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} Q_\star(\mathbf{s},\mathbf{a})$$

MPC

$$\begin{aligned} Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) = \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k,\mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k,\mathbf{u}_k) \\ & \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k,\mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}, \quad \mathbf{u}_0 = \mathbf{a} \end{aligned}$$

Theorem: under some assumptions

$$\boldsymbol{\pi}_{\boldsymbol{\theta}} = \boldsymbol{\pi}_\star, \quad V_{\boldsymbol{\theta}} = V_\star, \quad Q_{\boldsymbol{\theta}} = Q_\star$$

hold for some $T_{\boldsymbol{\theta}}$, $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$

Learning+MPC where cost & constraints are adjusted is formally justified

“Holistic” view of MPC: model for $Q_\star$, cost & constraints are part of that

Outline

What MPC parametrization?

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$

What MPC parametrization?

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$

Theory says:

$$
L_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right) = L\left(\mathbf{x}, \mathbf{u}\right) + \Delta\left(\mathbf{x}, \mathbf{u}\right)
$$

$$
\Delta\left(\mathbf{x}, \mathbf{u}\right) = \underbrace{\mathbb{E}\left[V_\star\left(\mathbf{x}_+\right) \mid \mathbf{x}, \mathbf{u}\right]}_{\text{Real system}} - \underbrace{V_\star\left(\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right)\right)}_{\text{Model}}
$$

$\mathbf{h}_{\boldsymbol{\theta}} > 0 \;\leftrightarrow\; \infty$ values in modification

What MPC parametrization?

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

Theory says:

$$L_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right) = L\left(\mathbf{x}, \mathbf{u}\right) + \Delta\left(\mathbf{x}, \mathbf{u}\right)$$

$$\Delta\left(\mathbf{x}, \mathbf{u}\right) = \underbrace{\mathbb{E}\left[V_\star\left(\mathbf{x}_+\right) \,|\, \mathbf{x}, \mathbf{u}\right]}_{\text{Real system}} - \underbrace{V_\star\left(\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right)\right)}_{\text{Model}}$$

$\mathbf{h}_{\boldsymbol{\theta}} > 0 \ \leftrightarrow \ \infty$ values in modification

Remarks:

- In practice, Δ (or $L_{\boldsymbol{\theta}}$) parametrized in a chosen class of functions, and "learned"
- $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$ convex is very beneficial, maybe restrictive
- When is the model optimal as is? We will come back to that later...

What MPC parametrization?

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\
& \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$

Theory is not very restrictive on admissible models $\mathbf{f}_{\boldsymbol{\theta}}$. Should we be worried? Not necessarily... this theory is not the end of the story

Theory says:

$$L_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \Delta(\mathbf{x}, \mathbf{u})$$

$$\Delta(\mathbf{x}, \mathbf{u}) = \underbrace{\mathbb{E}\left[V_\star(\mathbf{x}_+) \,|\, \mathbf{x}, \mathbf{u}\right]}_{\text{Real system}} - \underbrace{V_\star(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u}))}_{\text{Model}}$$

$\mathbf{h}_{\boldsymbol{\theta}} > 0 \;\leftrightarrow\; \infty$ values in modification

Remarks:

- In practice, Δ (or $L_{\boldsymbol{\theta}}$) parametrized in a chosen class of functions, and "learned"
- $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$ convex is very beneficial, maybe restrictive
- When is the model optimal as is? We will come back to that later...

Role of the MPC model?

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\theta}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\theta}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\theta}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}_{\theta}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$

Role of the MPC model?

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

Remarks:

- Theory not very restrictive on admissible models $\mathbf{f}_{\boldsymbol{\theta}}$
- Examples where $\mathbf{f}_{\boldsymbol{\theta}}$ is "very wrong" but MPC gives $Q^\star, V^\star, \boldsymbol{\pi}^\star$
- $\boldsymbol{\theta}$ such that MPC gives $Q^\star, V^\star, \boldsymbol{\pi}^\star$ may be non-unique
- Best "SYSID model" is not necessarily the best MPC model

What is the role of the model?

Role of the MPC model?

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}$$

Remarks:

- Theory not very restrictive on admissible models $\mathbf{f}_{\boldsymbol{\theta}}$
- Examples where $\mathbf{f}_{\boldsymbol{\theta}}$ is "very wrong" but MPC gives $Q^\star, V^\star, \boldsymbol{\pi}^\star$
- $\boldsymbol{\theta}$ such that MPC gives $Q^\star, V^\star, \boldsymbol{\pi}^\star$ may be non-unique
- Best "SYSID model" is not necessarily the best MPC model

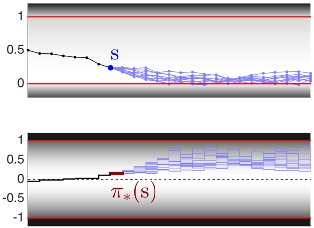
What is the role of the model?

Reflections: *depending on how we view this, it can become "philosophical"*

- MPC plan provides **explainability**. Wrong model $\Rightarrow$ wrong plan $\Rightarrow$ no explainability.
- MPC model associated to **safety** *(more on that soon)*

Role of the model - Explainable RL

Benefit of MPC over "black-box" RL



Role of the model - Explainable RL

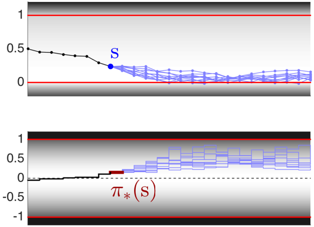
Benefit of MPC over "black-box" RL

MPC provides explainability...

... if model "makes sense"

- Not required by theory
- Not necessarily done by RL for MPC

How to keep the model sensible?



Role of the model - Explainable RL

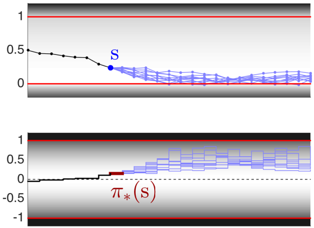
Benefit of MPC over "black-box" RL

MPC provides explainability...

... if model "makes sense"

- Not required by theory
- Not necessarily done by RL for MPC

How to keep the model sensible?



Theory says:

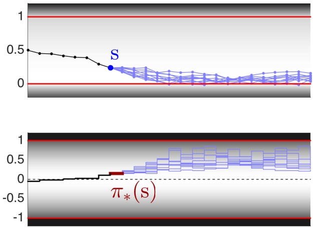
$$L_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{u}) = L(\mathbf{x},\mathbf{u}) + \Delta(\mathbf{x},\mathbf{u})$$

$$\Delta(\mathbf{x},\mathbf{u}) = \underbrace{\mathbb{E}\left[V_\star(\mathbf{x}_+)\mid \mathbf{x},\mathbf{u}\right]}_{\text{Real system}} - \underbrace{V_\star(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{u}))}_{\text{Model}}$$

Role of the model - Explainable RL

Benefit of MPC over "black-box" RL

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$



Theory says:

$$
L_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right) = L\left(\mathbf{x}, \mathbf{u}\right) + \Delta\left(\mathbf{x}, \mathbf{u}\right)
$$

$$
\Delta\left(\mathbf{x}, \mathbf{u}\right) = \underbrace{\mathbb{E}\left[V_\star\left(\mathbf{x}_+\right) \,|\, \mathbf{x}, \mathbf{u}\right]}_{\text{Real system}} - \underbrace{V_\star\left(\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right)\right)}_{\text{Model}}
$$

Adjust $\boldsymbol{\theta}$ such that

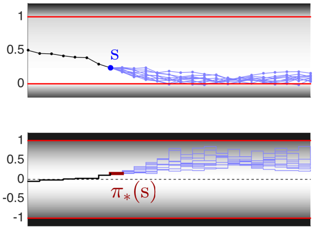
- $\mathbb{P}\left[\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s}, \mathbf{a}\right) \,|\, \mathbf{s}, \mathbf{a}\right]$ (likelihood) is "high"
- $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$ gives optimal MPC

for all $\mathbf{s}, \mathbf{a}$ in data

Role of the model - Explainable RL

Benefit of MPC over "black-box" RL

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$



Theory says:

$$
\begin{aligned}
L_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right) &= L\left(\mathbf{x}, \mathbf{u}\right) + \Delta\left(\mathbf{x}, \mathbf{u}\right) \\
\Delta\left(\mathbf{x}, \mathbf{u}\right) &= \underbrace{\mathbb{E}\left[V_\star\left(\mathbf{x}_+\right) \mid \mathbf{x}, \mathbf{u}\right]}_{\text{Real system}} - \underbrace{V_\star\left(\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right)\right)}_{\text{Model}}
\end{aligned}
$$

Adjust $\boldsymbol{\theta}$ such that

- $\mathbb{P}\left[\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s}, \mathbf{a}\right) \mid \mathbf{s}, \mathbf{a}\right]$ (likelihood) is "high"
- $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$ gives optimal MPC

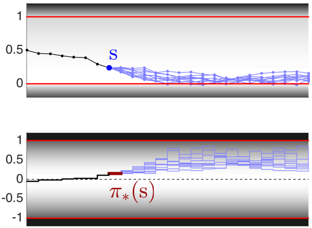
for all $\mathbf{s}, \mathbf{a}$ in data

RL & SYSID ought to combine without contradiction. If performance is key, RL should superseded SYSID.

Role of the model - Explainable RL

Benefit of MPC over "black-box" RL

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ & \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s}\end{aligned}$$



Theory says:

$$L_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}, \mathbf{u}) + \Delta(\mathbf{x}, \mathbf{u})$$

$$\Delta(\mathbf{x}, \mathbf{u}) = \underbrace{\mathbb{E}\left[V_\star(\mathbf{x}_+) \mid \mathbf{x}, \mathbf{u}\right]}_{\text{Real system}} - \underbrace{V_\star(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u}))}_{\text{Model}}$$

Adjust $\boldsymbol{\theta}$ such that

- $\mathbb{P}\left[\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \mid \mathbf{s}, \mathbf{a}\right]$ (likelihood) is "high"
- $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$ gives optimal MPC

for all $\mathbf{s}, \mathbf{a}$ in data

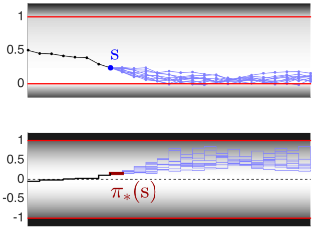
SYSID & RL can do

- Need to "harmonize" the two methods
- Take "SYSID steps" in the null space of $\nabla^2_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}^{\mathrm{MPC}}_{\boldsymbol{\theta}}\right)$

Role of the model - Explainable RL

Benefit of MPC over "black-box" RL

$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}
$$



Theory says:

$$
\begin{aligned}
L_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right) &= L\left(\mathbf{x}, \mathbf{u}\right) + \Delta\left(\mathbf{x}, \mathbf{u}\right) \\
\Delta\left(\mathbf{x}, \mathbf{u}\right) &= \underbrace{\mathbb{E}\left[V_\star\left(\mathbf{x}_+\right) \mid \mathbf{x}, \mathbf{u}\right]}_{\text{Real system}} - \underbrace{V_\star\left(\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}, \mathbf{u}\right)\right)}_{\text{Model}}
\end{aligned}
$$

Adjust $\boldsymbol{\theta}$ such that

- $\mathbb{P}\left[\mathbf{f}_{\boldsymbol{\theta}}\left(\mathrm{s}, \mathrm{a}\right) \mid \mathrm{s}, \mathrm{a}\right]$ (likelihood) is "high"
- $L_{\boldsymbol{\theta}}$, $\mathbf{h}_{\boldsymbol{\theta}}$ gives optimal MPC

for all s, a in data

Reflection:
Do we need a concept of "explainability" for MPC? What fundamental properties should the MPC model have to be deemed "explaining"?

Outline

Classic RL vs. RL-MPC

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k,\mathbf{a}_k\,\right]$$

MPC:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ & \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right) \leq 0\\ & \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

yields $\boldsymbol{\pi}^{\mathrm{MPC}}$, V^{MPC}, and Q^{MPC}

Classic RL vs. RL-MPC

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

RL with DNN

- correct structure is unknown
- good initialization is difficult
- respecting constraints is difficult & implicit

MPC:

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
\mathrm{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\
& \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\
& \mathbf{x}_0 = \mathbf{s}
\end{aligned}$$

yields $\boldsymbol{\pi}^{\mathrm{MPC}}$, V^{MPC}, and Q^{MPC}

Classic RL vs. RL-MPC

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k,\mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k,\mathbf{a}_k\,\right]$$

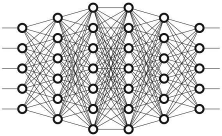
MPC:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right)\\ & \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k,\mathbf{u}_k\right) \leq 0\\ & \mathbf{x}_0 = \mathbf{s}\end{aligned}$$

yields $\boldsymbol{\pi}^{\mathrm{MPC}}$, V^{MPC}, and Q^{MPC}

RL with DNN

- correct structure is unknown
- good initialization is difficult
- respecting constraints is difficult & implicit



Classic RL vs. RL-MPC

MDP:

$$\min_{\boldsymbol{\pi}} \quad \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right]$$

where $\mathbf{a}_k = \boldsymbol{\pi}\left(\mathbf{s}_k\right)$ and system dynamics

$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot\,|\,\mathbf{s}_k, \mathbf{a}_k\,\right]$$

MPC:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ & \mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \leq 0 \\ & \mathbf{x}_0 = \mathbf{s} \end{aligned}$$

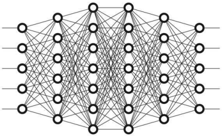
yields $\boldsymbol{\pi}^{\mathrm{MPC}}$, V^{MPC}, and Q^{MPC}

RL with DNN

- correct structure is unknown
- good initialization is difficult
- respecting constraints is difficult & implicit

MPC

- Structure and initialization given
- Constraints enforced explicitly
- Theory says that we can get $V_\star$, $Q_\star$, $\boldsymbol{\pi}_\star$ from MPC



RL methods & MPC

Form function approximators:

$$Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}), \quad V_{\boldsymbol{\theta}}(\mathbf{s}), \quad \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$$

via ad-hoc parametrization

RL methods & MPC

Form function approximators:

$$Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}), \quad V_{\boldsymbol{\theta}}(\mathbf{s}), \quad \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$$

via ad-hoc parametrization

- *Q*-**learning methods** adjust $\boldsymbol{\theta}$ to get

$$Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) \approx Q_\star(\mathbf{s},\mathbf{a})$$

Yields policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} \; Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) \approx \mathrm{a}\min_{\mathbf{a}} \; Q_\star(\mathbf{s},\mathbf{a}) = \boldsymbol{\pi}_\star(\mathbf{s})$$

E.g. basic Q-learning uses:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\delta\nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}(\mathbf{s}_k,\mathbf{a}_k)$$

$$\delta = L(\mathbf{s}_k,\mathbf{a}_k) + \gamma V_{\boldsymbol{\theta}}(\mathbf{s}_{k+1}) - Q_{\boldsymbol{\theta}}(\mathbf{s}_k,\mathbf{a}_k)$$

RL methods & MPC

Form function approximators:

$$Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}),\quad V_{\boldsymbol{\theta}}(\mathbf{s}),\quad \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$$

via ad-hoc parametrization

- ***Q*-learning methods** adjust $\boldsymbol{\theta}$ to get

$$Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) \approx Q_\star(\mathbf{s},\mathbf{a})$$

Yields policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}}\ Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) \approx \mathrm{a}\min_{\mathbf{a}}\ Q_\star(\mathbf{s},\mathbf{a}) = \boldsymbol{\pi}_\star(\mathbf{s})$$

E.g. basic Q-learning uses:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\delta\nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}(\mathbf{s}_k,\mathbf{a}_k)$$
$$\delta = L(\mathbf{s}_k,\mathbf{a}_k) + \gamma V_{\boldsymbol{\theta}}(\mathbf{s}_{k+1}) - Q_{\boldsymbol{\theta}}(\mathbf{s}_k,\mathbf{a}_k)$$

- **Policy gradient methods** adjust $\boldsymbol{\theta}$ to get

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = 0$$

yields policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \approx \boldsymbol{\pi}_\star(\mathbf{x})$ directly.

RL methods & MPC

Form function approximators:

$$Q_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}), \quad V_{\boldsymbol{\theta}}(\mathbf{s}), \quad \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$$

via ad-hoc parametrization

- ***Q*-learning methods** adjust $\boldsymbol{\theta}$ to get

  $$Q_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \approx Q_\star(\mathbf{s}, \mathbf{a})$$

  Yields policy:

  $$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} \ Q_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \approx \mathrm{a}\min_{\mathbf{a}} \ Q_\star(\mathbf{s}, \mathbf{a}) = \boldsymbol{\pi}_\star(\mathbf{s})$$

  E.g. basic Q-learning uses:

  $$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\delta\nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}(\mathbf{s}_k, \mathbf{a}_k)$$
  $$\delta = L(\mathbf{s}_k, \mathbf{a}_k) + \gamma V_{\boldsymbol{\theta}}(\mathbf{s}_{k+1}) - Q_{\boldsymbol{\theta}}(\mathbf{s}_k, \mathbf{a}_k)$$

- **Policy gradient methods** adjust $\boldsymbol{\theta}$ to get

  $$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = 0$$

  yields policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \approx \boldsymbol{\pi}_\star(\mathbf{x})$ directly. E.g.

  $$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{a}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right]$$

- **Derivative-free methods**
  - Build a surrogate of $J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$
  - Optimize over that model
  - Difficult over large parameter spaces

RL methods & MPC

Form function approximators:

$$Q_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}), \quad V_{\boldsymbol{\theta}}(\mathbf{s}), \quad \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$$

via ad-hoc parametrization

Derivative-based methods require $Q_{\boldsymbol{\theta}}$, $V_{\boldsymbol{\theta}}$, $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ and computing their sensitivities

- **Q-learning methods** adjust $\boldsymbol{\theta}$ to get

$$Q_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \approx Q_{\star}(\mathbf{s}, \mathbf{a})$$

Yields policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} Q_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \approx \mathrm{a}\min_{\mathbf{a}} Q_{\star}(\mathbf{s}, \mathbf{a}) = \boldsymbol{\pi}_{\star}(\mathbf{s})$$

E.g. basic Q-learning uses:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\delta\nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}(\mathbf{s}_k, \mathbf{a}_k)$$

$$\delta = L(\mathbf{s}_k, \mathbf{a}_k) + \gamma V_{\boldsymbol{\theta}}(\mathbf{s}_{k+1}) - Q_{\boldsymbol{\theta}}(\mathbf{s}_k, \mathbf{a}_k)$$

- **Policy gradient methods** adjust $\boldsymbol{\theta}$ to get

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = 0$$

yields policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \approx \boldsymbol{\pi}_{\star}(\mathbf{x})$ directly. E.g.

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{a}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right]$$

- **Derivative-free methods**
  - Build a surrogate of $J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$
  - Optimize over that model
  - Difficult over large parameter spaces

RL methods & MPC

Form function approximators:

$$Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}), \quad V_{\boldsymbol{\theta}}(\mathbf{s}), \quad \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$$

via ad-hoc parametrization

Derivative-based methods require $Q_{\boldsymbol{\theta}}$, $V_{\boldsymbol{\theta}}$, $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ and computing their sensitivities

In the RL-MPC context, $Q_{\boldsymbol{\theta}}$, $V_{\boldsymbol{\theta}}$, $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ are coming from an **MPC scheme, typically cast as Nonlinear Program. What about the sensitivities?**

- ***Q*-learning methods** adjust $\boldsymbol{\theta}$ to get

$$Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) \approx Q_\star(\mathbf{s},\mathbf{a})$$

Yields policy:

$$\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathrm{a}\min_{\mathbf{a}} \; Q_{\boldsymbol{\theta}}(\mathbf{s},\mathbf{a}) \approx \mathrm{a}\min_{\mathbf{a}} \; Q_\star(\mathbf{s},\mathbf{a}) = \boldsymbol{\pi}_\star(\mathbf{s})$$

E.g. basic Q-learning uses:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\delta\nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}(\mathbf{s}_k,\mathbf{a}_k)$$
$$\delta = L(\mathbf{s}_k,\mathbf{a}_k) + \gamma V_{\boldsymbol{\theta}}(\mathbf{s}_{k+1}) - Q_{\boldsymbol{\theta}}(\mathbf{s}_k,\mathbf{a}_k)$$

- **Policy gradient methods** adjust $\boldsymbol{\theta}$ to get

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = 0$$

yields policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{x}) \approx \boldsymbol{\pi}_\star(\mathbf{x})$ directly. E.g.

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\pi}_{\boldsymbol{\theta}} \nabla_{\mathbf{a}} Q_{\boldsymbol{\pi}_{\boldsymbol{\theta}}}\right]$$

- **Derivative-free methods**
  - Build a surrogate of $J(\boldsymbol{\pi}_{\boldsymbol{\theta}})$
  - Optimize over that model
  - Difficult over large parameter spaces

MPC Sensitivities?

MPC is a Nonlinear Program

Optimal value

$$\begin{aligned} V_{\boldsymbol{\theta}}(\mathrm{s}) = \min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \mathbf{g}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) = 0 \\ & \mathbf{h}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \leq 0 \end{aligned}$$

Optimal solution

$$\begin{aligned} \mathbf{w}^{\star}_{\boldsymbol{\theta}}(\mathrm{s}) = \mathrm{a}\min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \dots \end{aligned}$$

MPC Sensitivities?

MPC is a Nonlinear Program

Optimal value

$$
\begin{aligned}
V_{\boldsymbol{\theta}}(\mathrm{s}) = \min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\
\text{s.t.} \quad & \mathbf{g}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) = 0 \\
& \mathbf{h}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \leq 0
\end{aligned}
$$

Optimal solution

$$
\begin{aligned}
\mathbf{w}^\star_{\boldsymbol{\theta}}(\mathrm{s}) = \mathrm{a}\min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\
\text{s.t.} \quad & \ldots
\end{aligned}
$$

How to obtain:

$$\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}}, \ \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}, \ \nabla_{\boldsymbol{\theta}} \mathbf{w}^\star_{\boldsymbol{\theta}}$$

MPC Sensitivities?

MPC is a Nonlinear Program

Optimal value

$$\begin{aligned} V_{\boldsymbol{\theta}}(\mathrm{s}) = \min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \mathbf{g}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) = 0 \\ & \mathbf{h}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \leq 0 \end{aligned}$$

Optimal solution

$$\begin{aligned} \mathbf{w}^\star_{\boldsymbol{\theta}}(\mathrm{s}) = \mathrm{a}\min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \dots \end{aligned}$$

How to obtain:

$$\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}}, \;\; \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}, \;\; \nabla_{\boldsymbol{\theta}} \mathbf{w}^\star_{\boldsymbol{\theta}}$$

NLP solution satisfies (KKT conditions)

$$\mathbf{r} = \begin{bmatrix} \nabla_{\mathbf{w}} \mathcal{L} \\ \mathbf{g} \\ \mathbf{h}_i \boldsymbol{\mu}_i \end{bmatrix} = 0$$

$$\mathbf{h} \leq 0, \; \boldsymbol{\mu} \geq 0$$

where Lagrange function is

$$\mathcal{L} = \Phi + \boldsymbol{\lambda}^\top \mathbf{g} + \boldsymbol{\mu}^\top \mathbf{h}$$

and $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are the dual variables

MPC Sensitivities?

MPC is a Nonlinear Program

Optimal value

$$\begin{aligned} V_{\boldsymbol{\theta}}(\mathrm{s}) = \min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \mathbf{g}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) = 0 \\ & \mathbf{h}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \leq 0 \end{aligned}$$

Optimal solution

$$\begin{aligned} \mathbf{w}^{\star}_{\boldsymbol{\theta}}(\mathrm{s}) = \mathrm{a}\min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \ldots \end{aligned}$$

How to obtain:

$$\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}}, \ \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}, \ \nabla_{\boldsymbol{\theta}} \mathbf{w}^{\star}_{\boldsymbol{\theta}}$$

NLP solution satisfies (KKT conditions)

$$\mathbf{r} = \begin{bmatrix} \nabla_{\mathbf{w}} \mathcal{L} \\ \mathbf{g} \\ \mathbf{h}_i \boldsymbol{\mu}_i \end{bmatrix} = 0$$

$$\mathbf{h} \leq 0, \ \boldsymbol{\mu} \geq 0$$

where Lagrange function is

$$\mathcal{L} = \Phi + \boldsymbol{\lambda}^{\top} \mathbf{g} + \boldsymbol{\mu}^{\top} \mathbf{h}$$

and $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are the dual variables

MPC Sensitivities?

MPC is a Nonlinear Program

Optimal value

$$
\begin{aligned}
V_{\boldsymbol{\theta}}(\mathrm{s}) = \min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\
\text{s.t.} \quad & \mathbf{g}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) = 0 \\
& \mathbf{h}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \leq 0
\end{aligned}
$$

Optimal solution

$$
\begin{aligned}
\mathbf{w}^{\star}_{\boldsymbol{\theta}}(\mathrm{s}) = \mathrm{a}\min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\
\text{s.t.} \quad & \ldots
\end{aligned}
$$

How to obtain:

$$\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}}, \ \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}, \ \nabla_{\boldsymbol{\theta}} \mathbf{w}^{\star}_{\boldsymbol{\theta}}$$

NLP solution satisfies (KKT conditions)

$$\mathbf{r} = \begin{bmatrix} \nabla_{\mathbf{w}} \mathcal{L} \\ \mathbf{g} \\ \mathbf{h}_i \boldsymbol{\mu}_i \end{bmatrix} = 0$$

$$\mathbf{h} \leq 0, \ \boldsymbol{\mu} \geq 0$$

where Lagrange function is

$$\mathcal{L} = \Phi + \boldsymbol{\lambda}^{\top} \mathbf{g} + \boldsymbol{\mu}^{\top} \mathbf{h}$$

and $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are the dual variables

Solve NLP for $\mathrm{s}, \boldsymbol{\theta}$, provides $\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}$, then:

$$\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}}(\mathrm{s}) = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

is a simple function evaluation

MPC Sensitivities?

MPC is a Nonlinear Program

Optimal value

$$\begin{aligned} V_{\boldsymbol{\theta}}(\mathrm{s}) = \min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \mathbf{g}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) = 0 \\ & \mathbf{h}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \leq 0 \end{aligned}$$

Optimal solution

$$\begin{aligned} \mathbf{w}^\star_{\boldsymbol{\theta}}(\mathrm{s}) = \mathrm{a}\min_{\mathbf{w}} \quad & \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \dots \end{aligned}$$

How to obtain:

$$\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}}, \ \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}, \ \nabla_{\boldsymbol{\theta}} \mathbf{w}^\star_{\boldsymbol{\theta}}$$

NLP solution satisfies (KKT conditions)

$$\mathbf{r} = \begin{bmatrix} \nabla_{\mathbf{w}} \mathcal{L} \\ \mathbf{g} \\ \mathbf{h}_i \boldsymbol{\mu}_i \end{bmatrix} = 0$$

$$\mathbf{h} \leq 0, \ \boldsymbol{\mu} \geq 0$$

where Lagrange function is

$$\mathcal{L} = \Phi + \boldsymbol{\lambda}^\top \mathbf{g} + \boldsymbol{\mu}^\top \mathbf{h}$$

and $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are the dual variables

Solve NLP for $\mathrm{s}, \boldsymbol{\theta}$, provides $\mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}$, then:

$$\frac{\partial \mathbf{w}^\star_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} = -\frac{\partial \mathbf{r}}{\partial \mathbf{w}}^{-1} \frac{\partial \mathbf{r}}{\partial \boldsymbol{\theta}}$$

where $\frac{\partial \mathbf{r}}{\partial \mathbf{w}}^{-1}$ is already built in the solver, works if LICQ / SOSC

MPC Sensitivities?

MPC is a Nonlinear Program

Optimal value

$$V_{\boldsymbol{\theta}}(\mathbf{s}) = \min_{\mathbf{w}} \quad \Phi(\mathbf{w}, \mathbf{s}, \boldsymbol{\theta})$$
$$\text{s.t.} \quad \mathbf{g}(\mathbf{w}, \mathbf{s}, \boldsymbol{\theta}) = 0$$
$$\mathbf{h}(\mathbf{w}, \mathbf{s}, \boldsymbol{\theta}) \leq 0$$

Optimal solution

$$\mathbf{w}^\star_{\boldsymbol{\theta}}(\mathbf{s}) = \mathrm{a}\min_{\mathbf{w}} \quad \Phi(\mathbf{w}, \mathbf{s}, \boldsymbol{\theta})$$
$$\text{s.t.} \quad \dots$$

How to obtain:

$$\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}}, \;\; \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}, \;\; \nabla_{\boldsymbol{\theta}} \mathbf{w}^\star_{\boldsymbol{\theta}}$$

NLP solution satisfies (KKT conditions)

$$\mathbf{r} = \begin{bmatrix} \nabla_{\mathbf{w}} \mathcal{L} \\ \mathbf{g} \\ \mathbf{h}_i \boldsymbol{\mu}_i \end{bmatrix} = 0$$
$$\mathbf{h} \leq 0, \; \boldsymbol{\mu} \geq 0$$

where Lagrange function is

$$\mathcal{L} = \Phi + \boldsymbol{\lambda}^\top \mathbf{g} + \boldsymbol{\mu}^\top \mathbf{h}$$

and $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are the dual variables

Sensitivities do not exist for all $\mathbf{s}, \mathbf{a}$.
Does that matter?

MPC Sensitivities?

MPC is a Nonlinear Program

Optimal value

$$V_{\boldsymbol{\theta}}(\mathrm{s}) = \min_{\mathbf{w}} \quad \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \quad \text{s.t.} \quad \mathbf{g}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) = 0, \quad \mathbf{h}(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \leq 0$$

Optimal solution

$$\mathbf{w}^\star_{\boldsymbol{\theta}}(\mathrm{s}) = \mathrm{a}\min_{\mathbf{w}} \quad \Phi(\mathbf{w}, \mathrm{s}, \boldsymbol{\theta}) \quad \text{s.t.} \quad \ldots$$

How to obtain:

$$\nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}}, \quad \nabla_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}, \quad \nabla_{\boldsymbol{\theta}} \mathbf{w}^\star_{\boldsymbol{\theta}}$$

NLP solution satisfies (KKT conditions)

$$\mathbf{r} = \begin{bmatrix} \nabla_{\mathbf{w}} \mathcal{L} \\ \mathbf{g} \\ \mathbf{h}_i \boldsymbol{\mu}_i \end{bmatrix} = 0$$

$$\mathbf{h} \leq 0, \; \boldsymbol{\mu} \geq 0$$

where Lagrange function is

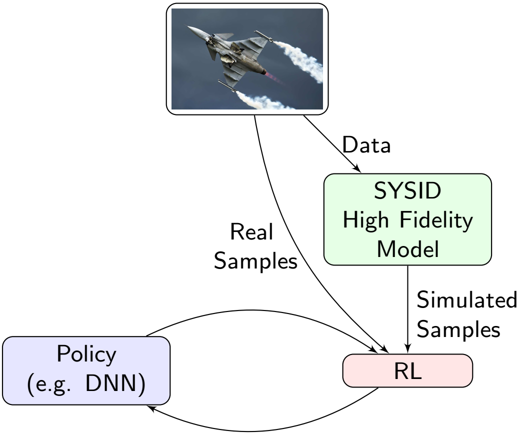
$$\mathcal{L} = \Phi + \boldsymbol{\lambda}^\top \mathbf{g} + \boldsymbol{\mu}^\top \mathbf{h}$$

and $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are the dual variables

Sensitivities do not exist for all s, a.
Does that matter?

In general no: they exist *almost everywhere*, and always appear inside $\mathbb{E}[\cdot]$. If the MDP has underlying densities, then we are good.

Model-based RL methods vs. RL-MPC: Data flow



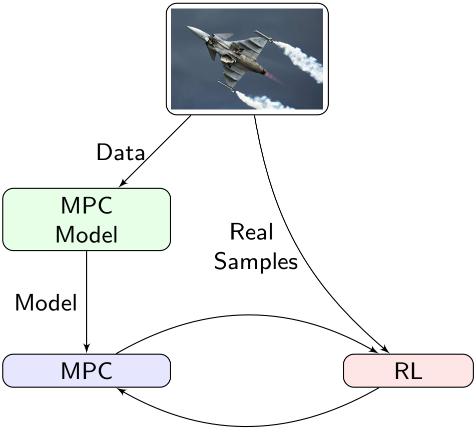
Common setup for "classic RL:

- Build statistical model of the real system
- Generate simulated samples
- Feed RL with real and simulated samples

Remarks:

- Simulated data much cheaper than real ones, most data will be simulated ones
- With mostly simulated data:
  - ≈equivalent to approximate DP
  - policy optimality relies on model quality

Model-based RL methods vs. RL-MPC: Data flow



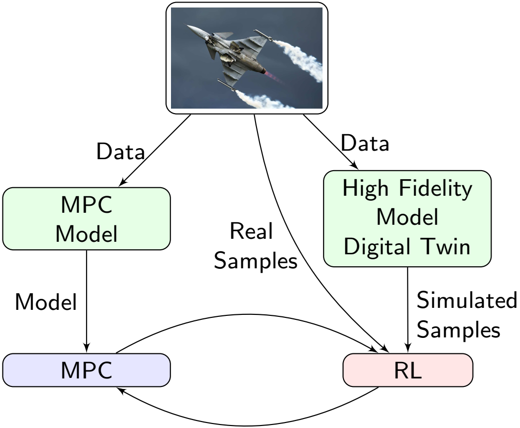
Basic setup for "RL-MPC":

- Build MPC model of the real system
- Pass it to MPC scheme
- Feed RL with real samples

Remarks:

- RL tunes MPC for real system
- MPC model may be "detuned" from SYSID version
- Real data are expensive...

Model-based RL methods vs. RL-MPC: Data flow



"Mixed" setup for "RL-MPC":

- Build MPC model of the real system
- MPC model is typically "simple"
- Build statistical model of the real system
- Generate simulated samples
- Feed RL with real and simulated samples

Remarks:

- Simple MPC model
- Complex simulation model
- MPC model may be "detuned" from SYSID version

Next lecture

- MPC as a path for safety and stability in RL
- More results & ideas

Thanks for your attention!