Orientation

So far we have seen...

- The building blocks
  - MDPs, DP, Bellman, Policy and Value Iteration, Reinforcement Learning, Actor-Critic
  - LQR, Dynamic Optimization, Nonlinear Programming, MPC, Sensitivities
- MPC & RL
  - Synthesis - Different flavors of ML / RL and MPC
  - Software tool leap-c

Orientation

So far we have seen...

- The building blocks
  - MDPs, DP, Bellman, Policy and Value Iteration, Reinforcement Learning, Actor-Critic
  - LQR, Dynamic Optimization, Nonlinear Programming, MPC, Sensitivities
- MPC & RL
  - Synthesis - Different flavors of ML / RL and MPC
  - Software tool leap-c

What we will discuss: parametrized / hierarchical RL over MPC

- The theory that supports it and what does it tell us? *(Now)*

Orientation

So far we have seen...

- The building blocks
  - MDPs, DP, Bellman, Policy and Value Iteration, Reinforcement Learning, Actor-Critic
  - LQR, Dynamic Optimization, Nonlinear Programming, MPC, Sensitivities
- MPC & RL
  - Synthesis - Different flavors of ML / RL and MPC
  - Software tool leap-c

What we will discuss: parametrized / hierarchical RL over MPC

- The theory that supports it and what does it tell us? *(Now)*
- Safe & stable RL over MPC *(In the afternoon)*

Orientation

So far we have seen...

- The building blocks
  - MDPs, DP, Bellman, Policy and Value Iteration, Reinforcement Learning, Actor-Critic
  - LQR, Dynamic Optimization, Nonlinear Programming, MPC, Sensitivities
- MPC & RL
  - Synthesis - Different flavors of ML / RL and MPC
  - Software tool leap-c

What we will discuss: parametrized / hierarchical RL over MPC

- The theory that supports it and what does it tell us? *(Now)*
- Safe & stable RL over MPC *(In the afternoon)*
- RL over MPC with belief states – a future prospect *(In the afternoon)*

Orientation

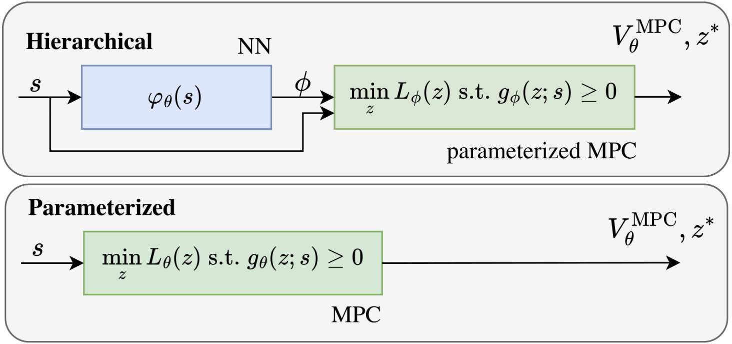
So far we have seen...

- The building blocks
  - MDPs, DP, Bellman, Policy and Value Iteration, Reinforcement Learning, Actor-Critic
  - LQR, Dynamic Optimization, Nonlinear Programming, MPC, Sensitivities
- MPC & RL
  - Synthesis - Different flavors of ML / RL and MPC
  - Software tool leap-c

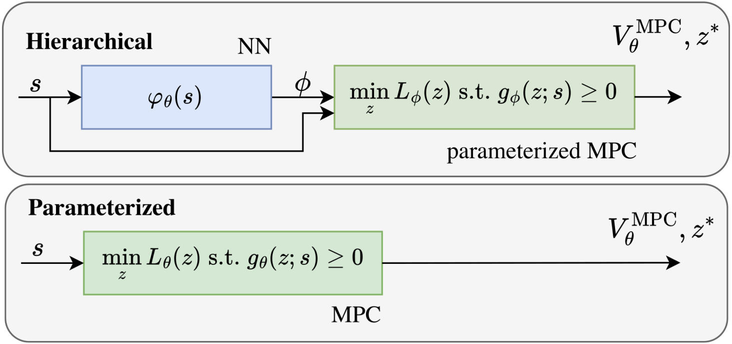
What we will discuss: parametrized / hierarchical RL over MPC

- The theory that supports it and what does it tell us? *(Now)*
- Safe & stable RL over MPC *(In the afternoon)*
- RL over MPC with belief states – a future prospect *(In the afternoon)*
- Beyond MPC – Model-based Decisions and AI for decisions *(Tomorrow)*

Focus of today's lectures (Leap-c structure)



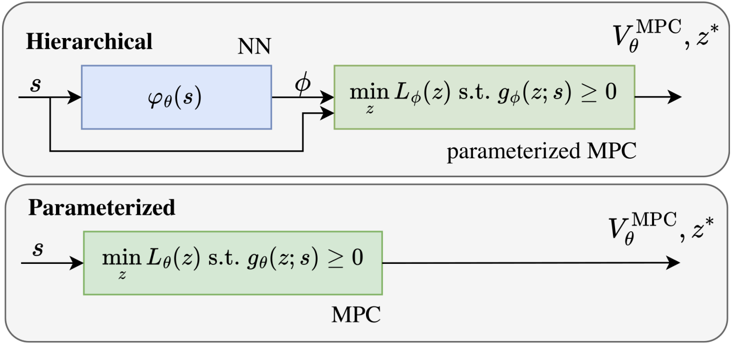
Focus of today's lectures (Leap-c structure)



Normal thinking when using MPC

- Fit MPC model to reality as good as possible (SYSID)
- MPC cost is what we want to minimize (energy, time, money, reference)
- MPC state constraints are what we need to respect (safety)

Focus of today's lectures (Leap-c structure)



Normal thinking when using MPC

- Fit MPC model to reality as good as possible (SYSID)
- MPC cost is what we want to minimize (energy, time, money, reference)
- MPC state constraints are what we need to respect (safety)

We are talking about changing everything!!

Reinforcement Learning Over MPC
Why Does It Work?

Sebastien Gros

Dept. of Cybernetic, NTNU
Faculty of Information Tech.

Freiburg PhD School

Outline

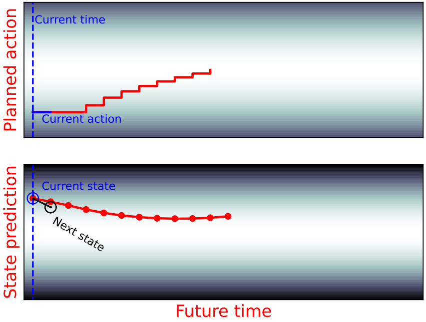
Model Predictive Control (MPC)

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



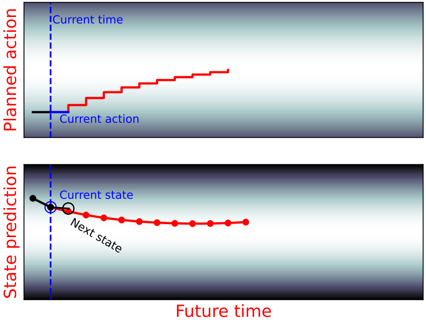
Model Predictive Control (MPC)

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



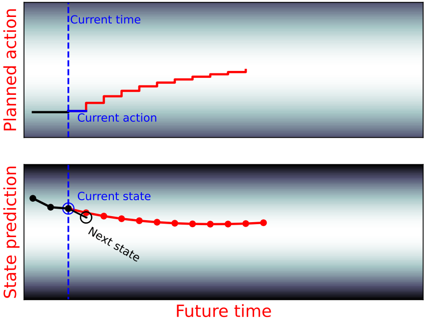
Model Predictive Control (MPC)

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

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

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

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



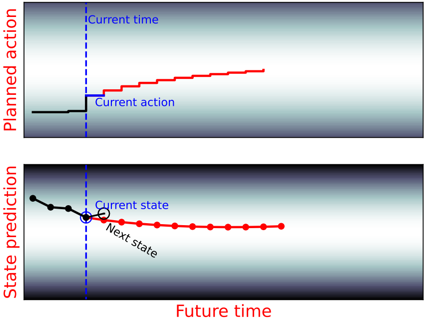
Model Predictive Control (MPC)

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



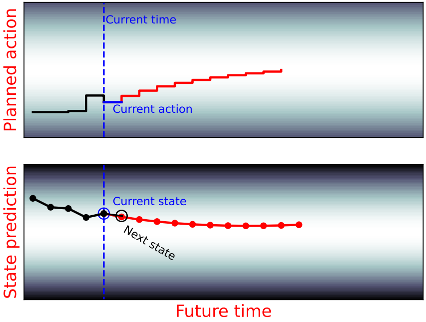
Model Predictive Control (MPC)

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



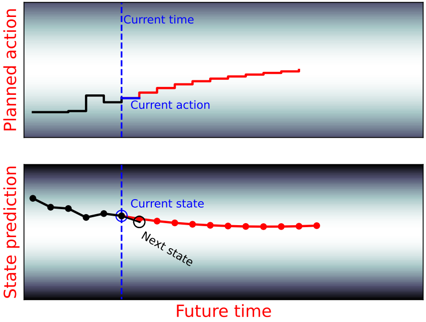
Model Predictive Control (MPC)

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

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

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

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



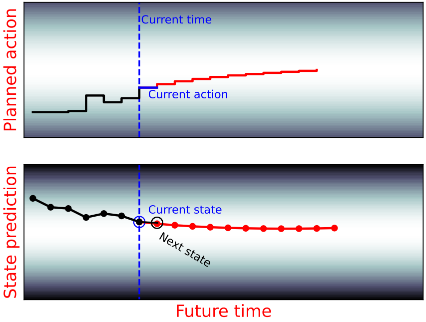
Model Predictive Control (MPC)

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 $\mathbf{s}$

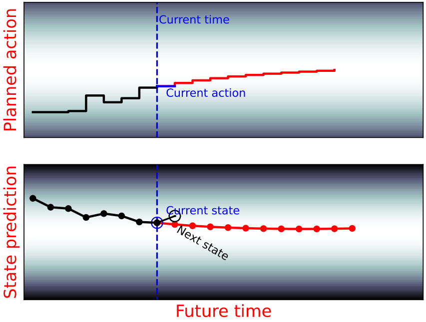
$$\min_{\mathbf{x},\mathbf{u}} \quad T(\mathbf{x}_N) + \sum_{k=0}^{N-1} L(\mathbf{x}_k, \mathbf{u}_k)$$

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

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

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

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



Model Predictive Control (MPC)

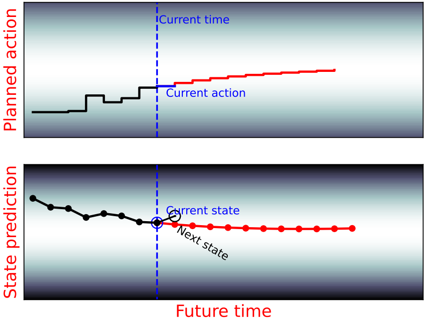
Optimize a plan over finite horizon, apply first move, repeat

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

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

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

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



MPC

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

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

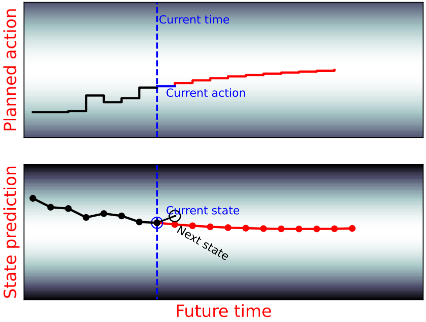
Model Predictive Control (MPC)

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

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

What an odd thing to do: we build and throw away plans all the time knowing that they are wrong all along, but kinda use them anyway...

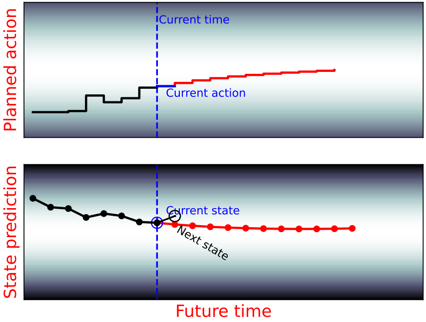
Model Predictive Control (MPC)

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

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

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

We have connected MPC and RL from an "implementation" point of view

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

We have connected MPC and RL from an “implementation” point of view
But understanding what we are doing is about connecting MPC to MDPs!!

Markov Decision Processes (MDP)

Stochastic state transitions (real world)

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

(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 (real world)

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

(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 (real world)

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

(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**

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

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

A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

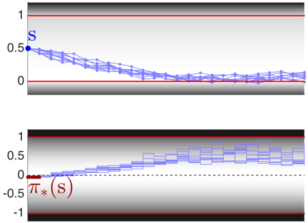
Stochastic state transitions (real world)

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

(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

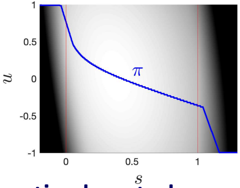
- **Closed-loop performance**

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

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)

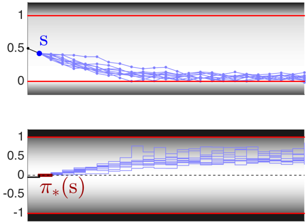
Stochastic state transitions (real world)

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

(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

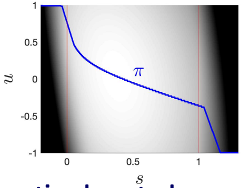
- **Closed-loop performance**

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

  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)

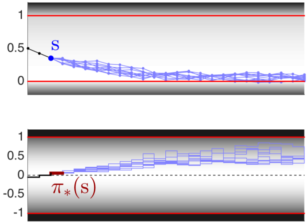
Stochastic state transitions (real world)

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

(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

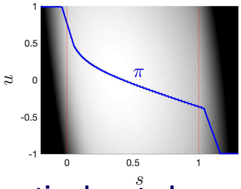
- **Closed-loop performance**

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

  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

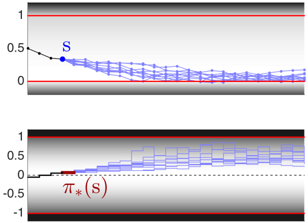
Markov Decision Processes (MDP)

Stochastic state transitions (real world)

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

(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

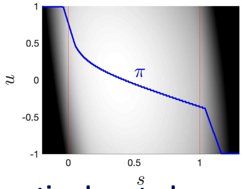
- **Closed-loop performance**

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

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 (real world)

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

(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

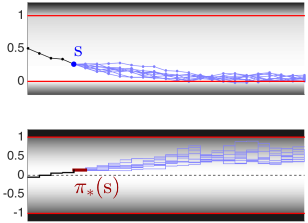
- **Closed-loop performance**

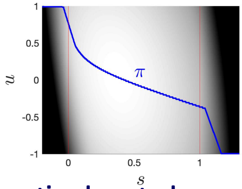
$$J\left(\boldsymbol{\pi}\right) = \mathbb{E}\left[\left.\sum_{k=0}^{\infty}\gamma^k L\left(\mathbf{s}_k, \mathbf{a}_k\right)\right| \boldsymbol{\pi}\right]$$

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

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

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





A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

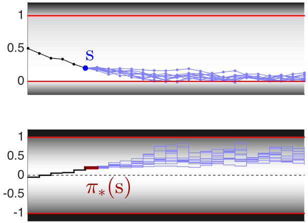
Stochastic state transitions (real world)

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

(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

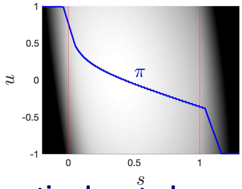
- **Closed-loop performance**

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

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)

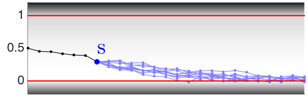
Stochastic state transitions (real world)

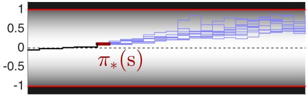
$$\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$$

(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

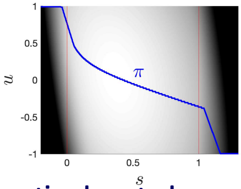
- **Closed-loop performance**

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

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)

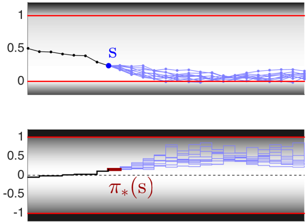
Stochastic state transitions (real world)

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

(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

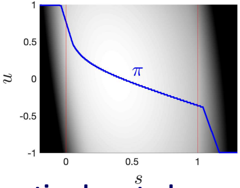
- **Closed-loop performance**

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

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)

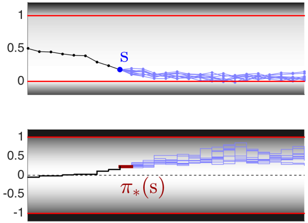
Stochastic state transitions (real world)

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

(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

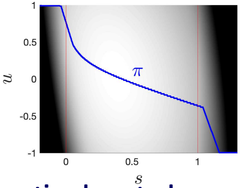
- **Closed-loop performance**

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

  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 (real world)

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

(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

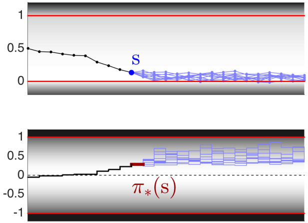
- **Closed-loop performance**

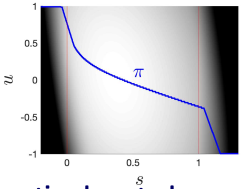
$$J(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \boldsymbol{\pi}\right]$$

  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

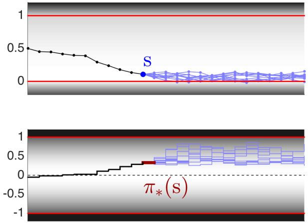
Markov Decision Processes (MDP)

Stochastic state transitions (real world)

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

(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

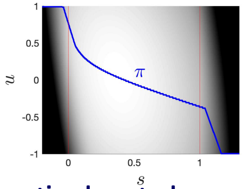
- **Closed-loop performance**

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

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

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

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



A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions (real world)

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

(state-action $\rightarrow$ next state)

Cost function (instant performance)

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

Impose hard constraints $\mathbf{h}(\mathbf{s}, \mathbf{a}) \leq 0$?

$$L(\mathbf{s}, \mathbf{a}) = \begin{cases} \ell(\mathbf{s}, \mathbf{a}) & \text{if} \quad \mathbf{h}(\mathbf{s}, \mathbf{a}) \leq 0 \\ \infty & \text{if} \quad \mathbf{h}(\mathbf{s}, \mathbf{a}) > 0 \end{cases}$$

I will use the same view in MPC for a bit

- **Policy**

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

  is how we act on the system

- **Closed-loop performance**

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

  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 (real world)

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

(state-action $\rightarrow$ next state)

- **Policy**

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

is how we act on the system

- **Closed-loop performance**

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

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

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

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

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

Impose hard constraints $\mathbf{h}(\mathbf{s}, \mathbf{a}) \leq 0$?

$$L(\mathbf{s}, \mathbf{a}) = \begin{cases} \ell(\mathbf{s}, \mathbf{a}) & \text{if} \quad \mathbf{h}(\mathbf{s}, \mathbf{a}) \leq 0 \\ \infty & \text{if} \quad \mathbf{h}(\mathbf{s}, \mathbf{a}) > 0 \end{cases}$$

I will use the same view in MPC for a bit

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

A (fairly) general way of describing optimal control

Markov Decision Processes (MDP)

Stochastic state transitions (real world)

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

(state-action $\rightarrow$ next state)

- **Policy**

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

  is how we act on the system

- **Closed-loop performance**

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

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

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

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

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

Impose hard constraints $\mathbf{h}(\mathbf{s}, \mathbf{a}) \leq 0$?

$$L(\mathbf{s}, \mathbf{a}) = \begin{cases} \ell(\mathbf{s}, \mathbf{a}) & \text{if} \quad \mathbf{h}(\mathbf{s}, \mathbf{a}) \leq 0 \\ \infty & \text{if} \quad \mathbf{h}(\mathbf{s}, \mathbf{a}) > 0 \end{cases}$$

I will use the same view in MPC for a bit

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

From Policy to Repeated Planning (MPC)

Infinite horizon & discounted

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

Policy $\boldsymbol{\pi}$: state $\rightarrow$ action
belongs to a function space

Let's transform an MDP into an MPC and understand the approximations we make

From Policy to Repeated Planning (MPC)

Infinite horizon & discounted

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

Policy $\boldsymbol{\pi}$: state $\rightarrow$ action
belongs to a function space

Let's transform an MDP into an MPC and understand the approximations we make

Finite-horizon equivalent:

$$\boldsymbol{\pi}^\star_{0,\dots,N-1} = \arg\min_{\boldsymbol{\pi}_{0,\dots,N-1}} \mathbb{E}\left[T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\right]$$

If $T = V^\star$, then $\boldsymbol{\pi}^\star_{0,\dots,N-1} = \boldsymbol{\pi}^\star_\infty$

From Policy to Repeated Planning (MPC)

Infinite horizon & discounted

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

Policy $\boldsymbol{\pi}$: state $\rightarrow$ action
belongs to a function space

Let's transform an MDP into an MPC and understand the approximations we make

Finite-horizon equivalent:

$$\boldsymbol{\pi}^\star_{0,\ldots,N-1} = \arg\min_{\boldsymbol{\pi}_{0,\ldots,N-1}} \mathbb{E}\left[T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\right]$$

If $T = V^\star$, then $\boldsymbol{\pi}^\star_{0,\ldots,N-1} = \boldsymbol{\pi}^\star_\infty$

Planning instead of a policy:

$$\min_{\mathbf{a}_{0,\ldots,N-1}} \mathbb{E}\left[T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{s}_0 = \mathbf{s}\right]$$

i.e. restrict policies to fixed $\mathbf{a}_{0,\ldots,N-1}$

From Policy to Repeated Planning (MPC)

Infinite horizon & discounted

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

Policy $\boldsymbol{\pi}$: state $\rightarrow$ action
belongs to a function space

Let's transform an MDP into an MPC and understand the approximations we make

Finite-horizon equivalent:

$$\boldsymbol{\pi}^\star_{0,\ldots,N-1} = \arg\min_{\boldsymbol{\pi}_{0,\ldots,N-1}} \mathbb{E}\left[T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\right]$$

If $T = V^\star$, then $\boldsymbol{\pi}^\star_{0,\ldots,N-1} = \boldsymbol{\pi}^\star_\infty$

Planning instead of a policy:

$$\min_{\mathbf{a}_{0,\ldots,N-1}} \mathbb{E}\left[T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{s}_0 = \mathbf{s}\right]$$

i.e. restrict policies to fixed $\mathbf{a}_{0,\ldots,N-1}$

Deterministic model, policy

$$\begin{aligned} \min_{\boldsymbol{\pi}_{0,\ldots,N-1}} \quad & T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \\ \text{s.t} \quad & \mathbf{s}_{k+1} = \mathbf{f}(\mathbf{s}_k, \mathbf{a}_k) \\ & \mathbf{s}_0 = \mathbf{s} \end{aligned}$$

i.e. adopt deterministic model

From Policy to Repeated Planning (MPC)

Infinite horizon & discounted

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

Policy $\boldsymbol{\pi}$: state $\rightarrow$ action
belongs to a function space

Let's transform an MDP into an MPC and understand the approximations we make

Finite-horizon equivalent:

$$\boldsymbol{\pi}^\star_{0,\ldots,N-1} = \arg\min_{\boldsymbol{\pi}_{0,\ldots,N-1}} \mathbb{E}\left[T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k)\right]$$

If $T = V^\star$, then $\boldsymbol{\pi}^\star_{0,\ldots,N-1} = \boldsymbol{\pi}^\star_\infty$

Why attacking the problem in these ways?

Planning instead of a policy:

$$\min_{\mathbf{a}_{0,\ldots,N-1}} \mathbb{E}\left[T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k) \,\middle|\, \mathbf{s}_0 = \mathbf{s}\right]$$

i.e. restrict policies to fixed $\mathbf{a}_{0,\ldots,N-1}$

Deterministic model, planning

$$\min_{\mathbf{a}_{0,\ldots,N-1}} T(\mathbf{s}_N) + \sum_{k=0}^{N-1} \gamma^k L(\mathbf{s}_k, \mathbf{a}_k)$$

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

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

i.e. adopt deterministic model

MPC as a policy

Deterministic MPC:

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

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

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

Defines policy:

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

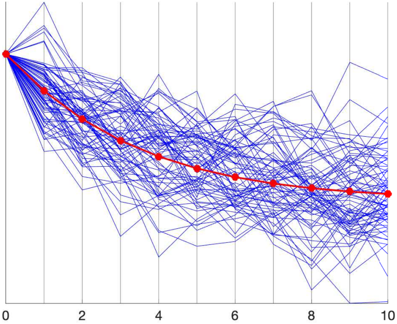
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

Defines policy:

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

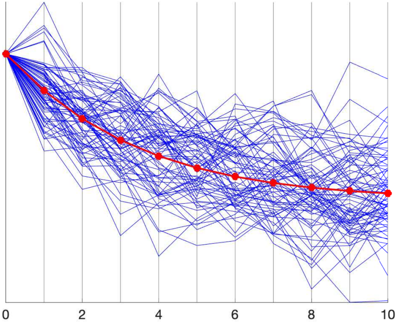
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

Defines policy:

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

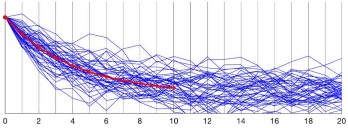
How does $\boldsymbol{\pi}^{\text{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

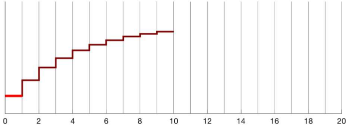
No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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





MPC as a policy

Deterministic MPC:

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

Defines policy:

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

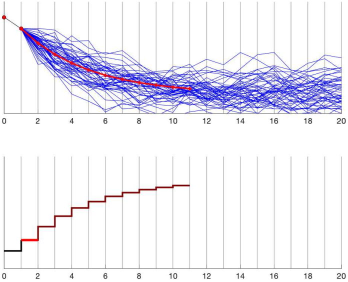
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

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

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

Defines policy:

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

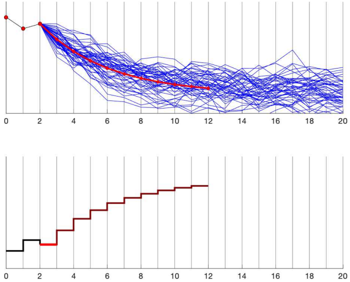
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

Defines policy:

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

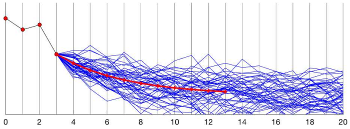
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

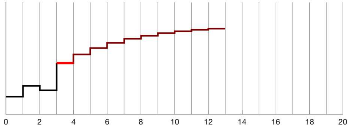
No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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





MPC as a policy

Deterministic MPC:

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

Defines policy:

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

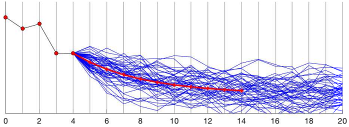
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

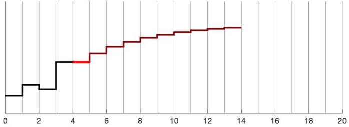
No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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





MPC as a policy

Deterministic MPC:

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

Defines policy:

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

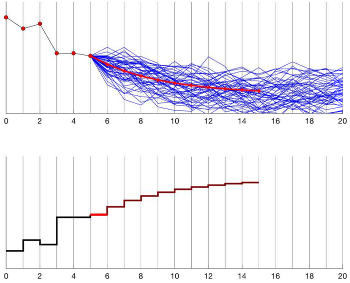
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

Defines policy:

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

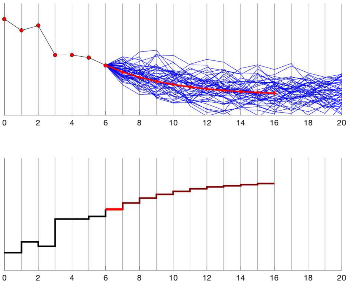
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

Defines policy:

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

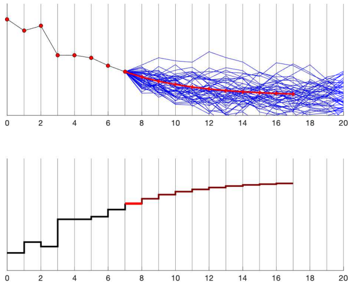
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

Defines policy:

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

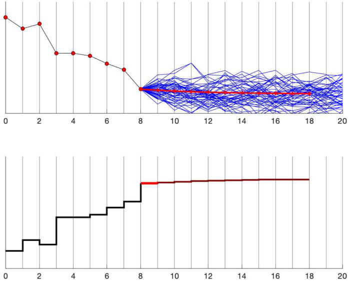
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

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

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

Defines policy:

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

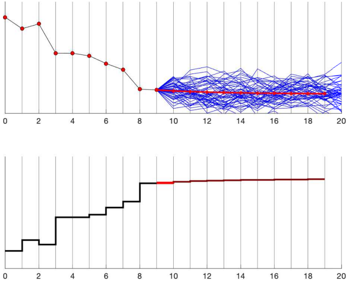
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

Defines policy:

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

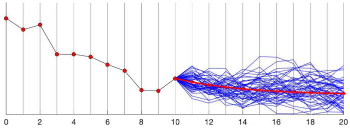
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

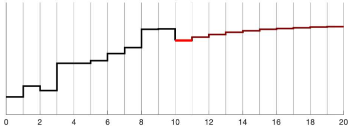
No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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





MPC as a policy

Deterministic MPC:

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

Defines policy:

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

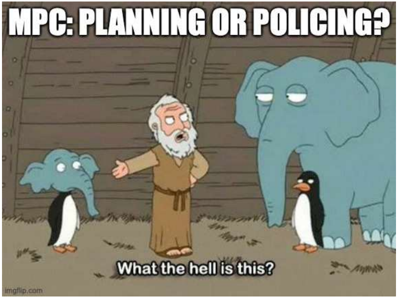
How does $\boldsymbol{\pi}^{\text{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



MPC as a policy

Deterministic MPC:

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

Defines policy:

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

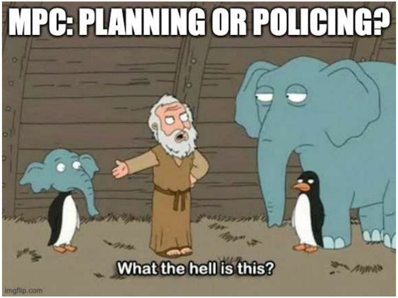
How does $\boldsymbol{\pi}^{\mathrm{MPC}}$ relate to $\boldsymbol{\pi}^\star$?

No reason to match:

- Planning rather than policing
- Model approximates stochasticity, often deterministic

Infinite horizon & discounted

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



Can we clarify the relationship?

Some more context on MPC for performance

Historically MPC focuses on **constraints satisfaction & stability**. Cost is for reference tracking, not representative of the system performance.

Some more context on MPC for performance

Historically MPC focuses on **constraints satisfaction & stability**. Cost is for reference tracking, not representative of the system performance.

- "Tracking MPC"
- Classic stability theory
- Uncertainty via
  - Robust MPC
  - Stochastic MPC
- *"MPC is for constraints satisfaction"* (undisclosed speaker)

Some more context on MPC for performance

Historically MPC focuses on **constraints satisfaction & stability**. Cost is for reference tracking, not representative of the system performance.

More recently, focus shifted to **closed-loop performance**, e.g. energy, time, money. Cost is generic, representative of the system performance.

- "Tracking MPC"
- Classic stability theory
- Uncertainty via
  - Robust MPC
  - Stochastic MPC
- *"MPC is for constraints satisfaction"* (undisclosed speaker)

Some more context on MPC for performance

Historically MPC focuses on **constraints satisfaction & stability**. Cost is for reference tracking, not representative of the system performance.

- "Tracking MPC"
- Classic stability theory
- Uncertainty via
  - Robust MPC
  - Stochastic MPC
- *"MPC is for constraints satisfaction"* (undisclosed speaker)

More recently, focus shifted to **closed-loop performance**, e.g. energy, time, money. Cost is generic, representative of the system performance.

- "Economic MPC"
- Dissipativity theory
- Uncertainty via
  - Robust MPC
  - Stochastic MPC
- MPC can optimize the system performance...

Some more context on MPC for performance

Historically MPC focuses on **constraints satisfaction & stability**. Cost is for reference tracking, not representative of the system performance.

- "Tracking MPC"
- Classic stability theory
- Uncertainty via
  - Robust MPC
  - Stochastic MPC
- *"MPC is for constraints satisfaction"* (undisclosed speaker)

More recently, focus shifted to **closed-loop performance**, e.g. energy, time, money. Cost is generic, representative of the system performance.

- "Economic MPC"
- Dissipativity theory
- Uncertainty via
  - Robust MPC
  - Stochastic MPC
- MPC can optimize the system performance...

MPC for closed-loop performance

- is not a very old topic
- "historical robustness" of MPC does not hold in the presence of stochasticity

Some more context on MPC for performance

Historically MPC focuses on **constraints satisfaction & stability**. Cost is for reference tracking, not representative of the system performance.

- "Tracking MPC"
- Classic stability theory
- Uncertainty via
  - Robust MPC
  - Stochastic MPC
- *"MPC is for constraints satisfaction"* (undisclosed speaker)

More recently, focus shifted to **closed-loop performance**, e.g. energy, time, money. Cost is generic, representative of the system performance.

- "Economic MPC"
- Dissipativity theory
- Uncertainty via
  - Robust MPC
  - Stochastic MPC
- MPC can optimize the system performance...

MPC for closed-loop performance

- is not a very old topic
- "historical robustness" of MPC does not hold in the presence of stochasticity

Soft claim: MPC can be used as a practical toolbox to model the solution of MDPs. This view is the "best way" for understanding what we are doing with economic MPC

Outline

1. Let's rebuild some background – MPC and MDPs
2. MPC as a solution to MDPs
3. When is RL (most) beneficial for MPC?
4. A Deeper Look at the Theory

Learning for MPC - Machine Learning in-the-loop

MPC policy

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

from model*, e.g.

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

$\mathbf{a}$

$\mathbf{s}$

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

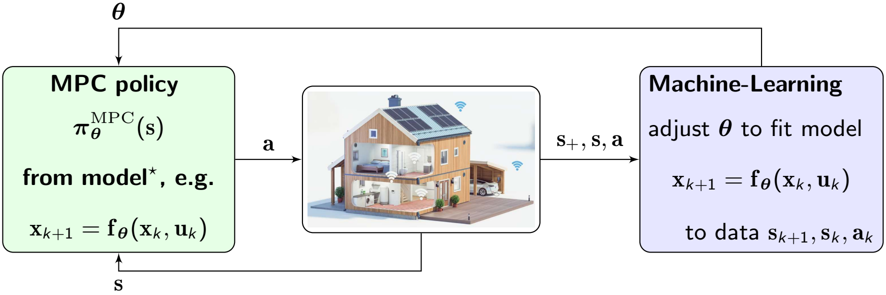
adjust $\boldsymbol{\theta}$ to fit model

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

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

$\mathbf{s}$

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, GPC, Probabilistic AI ...

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, GPC, Probabilistic AI ...

**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, GPC, Probabilistic AI ...

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

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

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, GPC, Probabilistic AI ...

**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

How can MPC model an MDP?

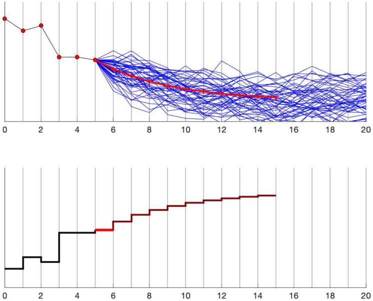
MDP optimal policy

$$\boldsymbol{\pi}^\star = \arg\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k)\right]$$

is entirely described by $Q^\star(\mathrm{s}, \mathrm{a})$

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

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



How can MPC model an MDP?

MDP optimal policy

$$\pi^\star = \arg\min_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k)\right]$$

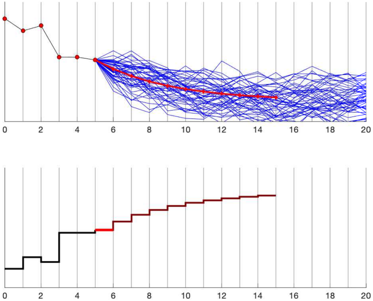
is entirely described by $Q^\star(\mathbf{s}, \mathbf{a})$

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

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

MPC as a model of the MDP

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



How can MPC model an MDP?

MDP optimal policy

$$\pi^\star = \arg\min_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k)\right]$$

is entirely described by $Q^\star(\mathbf{s}, \mathbf{a})$

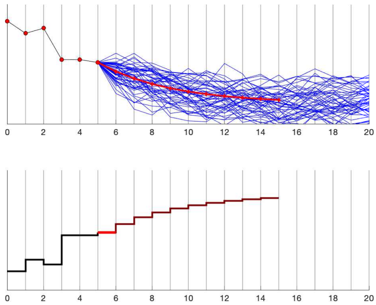
MPC policy $\pi^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

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

MPC as a model of the MDP

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

- Solving MDP $\leftarrow$ building model of $Q^\star$
- E.g. Q-learning does that from data
- MPC can model a value function...
- Can it model an action-value function?



How can MPC model an MDP?

MDP optimal policy

$$\boldsymbol{\pi}^\star = \arg\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k)\right]$$

is entirely described by $Q^\star(\mathbf{s}, \mathbf{a})$

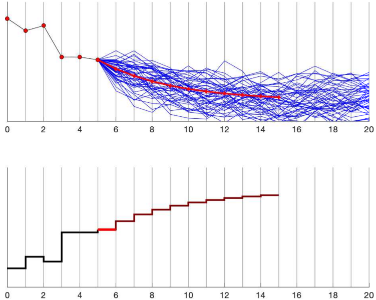
MPC policy $\boldsymbol{\pi}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

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

MPC as a model of the MDP

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

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



How can MPC model an MDP?

MDP optimal policy

$$\pi^\star = \arg\min_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k)\right]$$

is entirely described by $Q^\star(\mathbf{s}, \mathbf{a})$

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

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

MPC as a model of the MDP

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

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

MPC is consistent (for correct T):

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

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

How can MPC model an MDP?

MDP optimal policy

$$\boldsymbol{\pi}^\star = \arg\min_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k L(\mathbf{x}_k, \mathbf{u}_k)\right]$$

is entirely described by $Q^\star(\mathbf{s}, \mathbf{a})$

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

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

MPC as a model of the MDP

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

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

MPC is consistent (for correct T):

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

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

MPC is a complete model of MDP if:

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

for all $\mathbf{s}, \mathbf{a}$. Then **optimality** holds:

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

Modelling the MDP solution with MPC? Paradigm shifts...

Modelling the MDP solution with 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 the MDP solution

Modelling the MDP solution with 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 the MDP solution

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

Modelling the MDP solution with 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 the MDP solution

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

Find θ such that prediction "fits" the data

Shift to...

Find θ that "fits MPC to optimality" according to the data

($Q^{\mathrm{MPC}} \rightarrow Q^\star$ or at least $\boldsymbol{\pi}^{\mathrm{MPC}} \rightarrow \boldsymbol{\pi}^\star$)

Modelling the MDP solution with 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 the MDP solution

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
($Q^{\mathrm{MPC}} \rightarrow Q^\star$ or at least $\boldsymbol{\pi}^{\mathrm{MPC}} \rightarrow \boldsymbol{\pi}^\star$)

- $\rightarrow$ Best model for closed-loop performance
- $\neq$ Best model to fit the data!

Modelling the MDP solution with 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 the MDP solution

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 θ such that prediction "fits" the data

Shift to...

Find θ that "fits MPC to optimality" according to the data
($Q^{\mathrm{MPC}} \rightarrow Q^\star$ or at least $\boldsymbol{\pi}^{\mathrm{MPC}} \rightarrow \boldsymbol{\pi}^\star$)

- $\rightarrow$ Best model for closed-loop performance
- $\neq$ Best model to fit the data!

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

Can we do better?

Modelling the MDP solution with 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 the MDP solution

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
($Q^{\mathrm{MPC}} \to Q^\star$ or at least $\boldsymbol{\pi}^{\mathrm{MPC}} \to \boldsymbol{\pi}^\star$)

- $\to$ Best model for closed-loop performance
- $\neq$ Best model to fit the data!

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

Can we do better? **Yes**!

Modelling the MDP solution with 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 the MDP solution

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
($Q^{\mathrm{MPC}} \rightarrow Q^\star$ or at least $\boldsymbol{\pi}^{\mathrm{MPC}} \rightarrow \boldsymbol{\pi}^\star$)

Shift 2: "holistic" parametrization

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

i.e. cost and constraints are part of the model

Modelling the MDP solution with 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 the MDP solution

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

Find θ such that prediction "fits" the data

Shift to...

Find θ that "fits MPC to optimality" according to the data

($Q^{\mathrm{MPC}} \to Q^\star$ or at least $\boldsymbol{\pi}^{\mathrm{MPC}} \to \boldsymbol{\pi}^\star$)

"Holistic" parametrization - Is that formally justified? Yes...

Full MPC parametrization:

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

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

“Holistic” parametrization - Is that formally justified? Yes...

Full MPC parametrization:

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

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

Theorem: under some technical conditions and for a “rich” parametrization of the MPC, there is a $\boldsymbol{\theta}$ such that

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

even if the model $\mathbf{f}_{\boldsymbol{\theta}}$ *cannot* describe the real system accurately

"Holistic" parametrization - Is that formally justified? Yes...

Full MPC parametrization:

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

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

Theorem: under some technical conditions and for a "rich" parametrization of the MPC, there is a $\boldsymbol{\theta}$ such that

$$\begin{aligned} Q_{\boldsymbol{\theta}}^{\mathrm{MPC}}\left(\mathbf{s}, \mathbf{a}\right) &= Q^\star\left(\mathbf{s}, \mathbf{a}\right) \\ \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\left(\mathbf{s}\right) &= \boldsymbol{\pi}^\star\left(\mathbf{s}\right) \end{aligned}$$

even if the model $\mathbf{f}_{\boldsymbol{\theta}}$ *cannot* describe the real system accurately

Remarks

- Compensate MPC model deficiencies in the cost + constraints

“Holistic” parametrization - Is that formally justified? Yes...

Full MPC parametrization:

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

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

Theorem: under some technical conditions and for a “rich” parametrization of the MPC, there is a $\boldsymbol{\theta}$ such that

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

even if the model $\mathbf{f}_{\boldsymbol{\theta}}$ *cannot* describe the real system accurately

Remarks

- Compensate MPC model deficiencies in the cost + constraints
- Generic: works for robust MPC, stochastic MPC, economic MPC, Multi-Step PC, etc.

"Holistic" parametrization - Is that formally justified? Yes...

Full MPC parametrization:

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

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

Theorem: under some technical conditions and for a "rich" parametrization of the MPC, there is a $\boldsymbol{\theta}$ such that

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

even if the model $\mathbf{f}_{\boldsymbol{\theta}}$ *cannot* describe the real system accurately

Remarks

- Compensate MPC model deficiencies in the cost + constraints
- Generic: works for robust MPC, stochastic MPC, economic MPC, Multi-Step PC, etc.
- Sanity check: technical conditions are mild but forbid MPC model to be "very absurd"

“Holistic” parametrization - Is that formally justified? Yes...

Full MPC parametrization:

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

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

MPC is a model of the optimal policy
not a policy approximation using open-loop (model-based) predictions

Theorem: under some technical conditions and for a “rich” parametrization of the MPC, there is a $\boldsymbol{\theta}$ such that

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

even if the model $\mathbf{f}_{\boldsymbol{\theta}}$ *cannot* describe the real system accurately

Remarks

- Compensate MPC model deficiencies in the cost + constraints
- Generic: works for robust MPC, stochastic MPC, economic MPC, Multi-Step PC, etc.
- Sanity check: technical conditions are mild but forbid MPC model to be “very absurd”

How to use this? Reinforcement Learning

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

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

- $\min_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$ 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

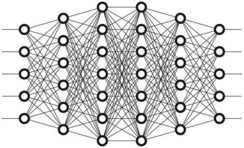
Policy $\pi^{\mathrm{MPC}}_{\theta}(\mathbf{s}) = \mathbf{u}_0^\star$ from

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

- $\min_\theta J\left(\pi^{\mathrm{MPC}}_{\theta}\right)$ using data
- $\theta \rightarrow J\left(\pi^{\mathrm{MPC}}_{\theta}\right)$ 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

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

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

Reinforcement Learning

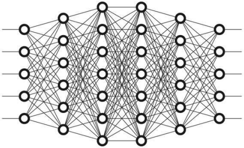
Tools to approximate $\pi^\star$ from data
This is not (necessarily) about DNNs

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

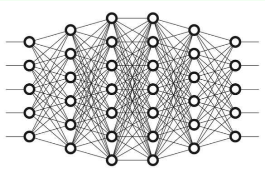


Reinforcement Learning & MPC

Classical Reinforcement Learning

$\boldsymbol{\theta}$

$\boldsymbol{\pi}_{\boldsymbol{\theta}}$ given by



$\mathbf{a}$

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

RL adjusts $\boldsymbol{\theta}$ to achieve

$$\nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right) = 0$$

from data

$\mathbf{s}$

Reinforcement Learning over MPC

$\boldsymbol{\theta}$

MPC policy

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

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

$\mathbf{a}$

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

RL adjusts $\boldsymbol{\theta}$ to achieve

$$\nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right) = 0$$

from data

$\mathbf{s}$

Reinforcement Learning & MPC

Reinforcement Learning over MPC

$\boldsymbol{\theta}$

MPC policy

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

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

$\mathbf{a}$

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

RL adjusts $\boldsymbol{\theta}$ to achieve

$$\nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right) = 0$$

from data

$\mathbf{s}$

Why is it useful?

- ✓ Tunes your optimization model for real-world performance
- ✓ MPC: 25 years of results on formal guarantees & methods
- ✓ Easy to inject knowledge
- ✓ Learning does not start from scratch
- ✓ Planning provides explainability
- ✓ Software now available

Reinforcement Learning & MPC

Reinforcement Learning over MPC

$\boldsymbol{\theta}$

MPC policy

$$\pi_{\boldsymbol{\theta}}^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star \text{ from}$$

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

$\mathbf{a}$

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

RL adjusts $\boldsymbol{\theta}$ to achieve

$$\nabla_{\boldsymbol{\theta}} J\left(\pi_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right) = 0$$

from data

$\mathbf{s}$

Why is it useful?

- ✓ Tunes your optimization model for real-world performance
- ✓ MPC: 25 years of results on formal guarantees & methods
- ✓ Easy to inject knowledge
- ✓ Learning does not start from scratch
- ✓ Planning provides explainability
- ✓ Software now available

Why can it be difficult?

- ✗ Optimization is computationally more expensive than a DNN
- ✗ Deployment on GPUs is lagging
- ✗ Non-convexity can get in the way...

Reinforcement Learning & MPC

Reinforcement Learning over MPC

$\boldsymbol{\theta}$

MPC policy

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

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

$\mathbf{a}$

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

RL adjusts $\boldsymbol{\theta}$ to achieve

$$\nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right) = 0$$

from data

$\mathbf{s}$

Why is it useful?

- ✓ Tunes your optimization model for real-world performance
- ✓ MPC: 25 years of results on formal guarantees & methods
- ✓ Easy to inject knowledge
- ✓ Learning does not start from scratch
- ✓ Planning provides explainability
- ✓ Software now available

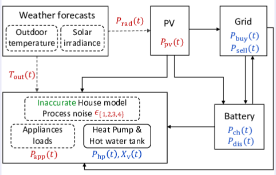
Why can it be difficult?

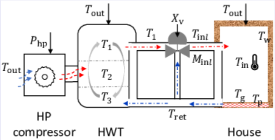
- ✘ Optimization is computationally more expensive than a DNN
- ✘ Deployment on GPUs is lagging
- ✘ Non-convexity can get in the way...

RL over MPC can tune your MPC for performance beyond what classical methods can do!

Example - Home Energy Management (simulated)

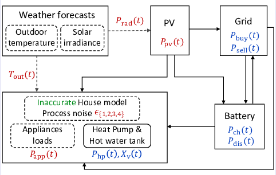
Setup

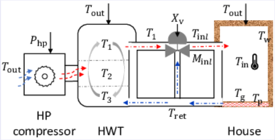




Example - Home Energy Management (simulated)

Setup



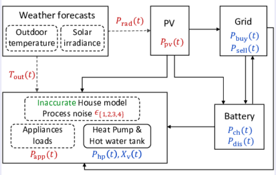


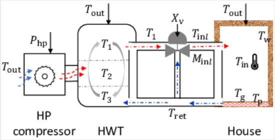
Learning process: aligned followed with closed-loop

Real world data

Example - Home Energy Management (simulated)

Setup



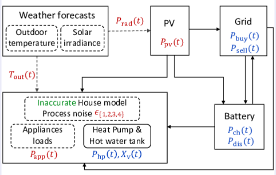


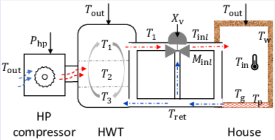
Learning process: aligned followed with closed-loop

Real world data —Data→ Classical model fitting to build MPC model

Example - Home Energy Management (simulated)

Setup



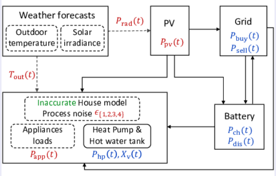


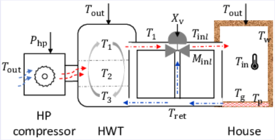
Learning process: aligned followed with closed-loop

Real world data $\xrightarrow{\text{Data}}$ Classical model fitting to build MPC model $\xrightarrow{\boldsymbol{\theta}}$ Reinforcement Learning over MPC

Example - Home Energy Management (simulated)

Setup

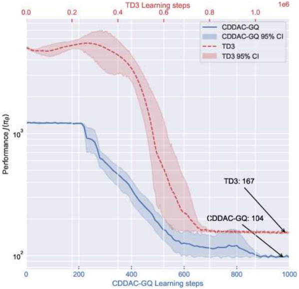




Learning process: aligned followed with closed-loop

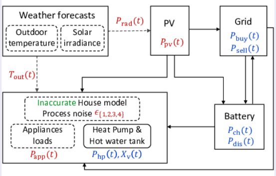
Real world data $\xrightarrow{\text{Data}}$ Classical model fitting to build MPC model $\xrightarrow{\boldsymbol{\theta}}$ Reinforcement Learning over MPC

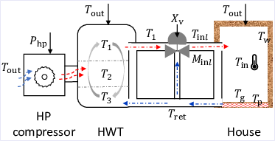
RL phase (2 RL algo.)



Example - Home Energy Management (simulated)

Setup

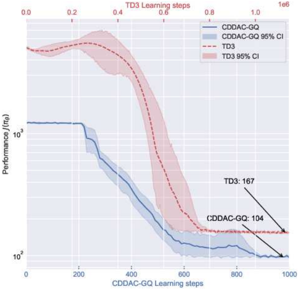




Learning process: aligned followed with closed-loop

Real world data $\xrightarrow{\text{Data}}$ Classical model fitting to build MPC model $\xrightarrow{\boldsymbol{\theta}}$ Reinforcement Learning over MPC

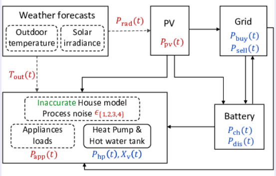
RL phase (2 RL algo.)

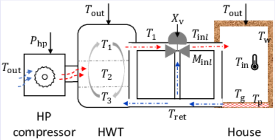


The RL step improves the MPC performance significantly from only model fitting

Example - Home Energy Management (simulated)

Setup

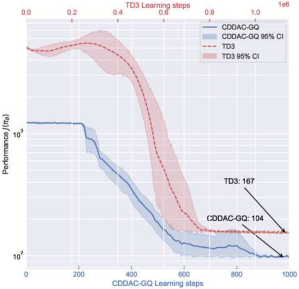




Learning process: aligned followed with closed-loop

Real world data $\xrightarrow{\text{Data}}$ Classical model fitting to build MPC model $\xrightarrow{\boldsymbol{\theta}}$ Reinforcement Learning over MPC

RL phase (2 RL algo.)



The RL step improves the MPC performance significantly from only model fitting

RL over MPC can tune your MPC for performance beyond what classical methods can do!

Outline

Model fitting?

Proposed paradigm

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

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

Model fitting?

A step back: **model adjustment?**

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

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

Adjust $\boldsymbol{\theta}$ according to (e.g.)

1. $\min_{\boldsymbol{\theta}} \mathbb{E}\left[\left\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\right\|^2\right]$ vs.
2. $\min_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$

... **is that different?**

Model fitting?

A step back: **model adjustment?**

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

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

Adjust $\boldsymbol{\theta}$ according to (e.g.)

1. $\min_{\boldsymbol{\theta}} \mathbb{E}\left[\left\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\right\|^2\right]$ vs.
2. $\min_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}}\right)$

... **is that different?**

Empirical answers:

- In general it is different...
- Good to do 1. first, then 2.
- Is step 2 necessary?
  - Improves closed-loop performance
  - But gain is *very* case dependent (also with full parametrization)

What causes that?

Sufficient Condition for MPC Optimality (for MDPs)

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

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

Sufficient Condition for MPC Optimality (for MDPs)

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

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

Sufficient condition of optimality of MPC
relates model $\mathbf{f}$ to optimal value function and conditional distribution $\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$ of the MDP (+ condition on T)

Sufficient Condition for MPC Optimality (for MDPs)

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

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

Sufficient condition of optimality of MPC
relates model $\mathbf{f}$ to optimal value function and conditional distribution $\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$ of the MDP ($+$ condition on T)

In general

- ridge regression

$$\mathbf{f}(\mathbf{s}, \mathbf{a}) = \mathbb{E}\left[\mathbf{s}_+ \,|\, \mathbf{s}, \mathbf{a}\right]$$

- max likelihood

$$\mathbf{f}(\mathbf{s}, \mathbf{a}) = \max \rho\left[\mathbf{s}_+ \,|\, \mathbf{s}, \mathbf{a}\right]$$

do not satisfy the optimality conditions

Sufficient Condition for MPC Optimality (for MDPs)

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

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

Sufficient condition of optimality of MPC relates model $\mathbf{f}$ to optimal value function and conditional distribution $\mathbf{s}, \mathbf{a} \to \mathbf{s}_+$ of the MDP ($+$ condition on T)

In general

- ridge regression

$$\mathbf{f}(\mathbf{s}, \mathbf{a}) = \mathbb{E}[\mathbf{s}_+ \,|\, \mathbf{s}, \mathbf{a}]$$

- max likelihood

$$\mathbf{f}(\mathbf{s}, \mathbf{a}) = \max \rho[\mathbf{s}_+ \,|\, \mathbf{s}, \mathbf{a}]$$

do not satisfy the optimality conditions

Model from "classic SYSID" does not necessarily yield the best MPC policy

Sufficient Condition for MPC Optimality (for MDPs)

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

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

Sufficient condition of optimality of MPC relates model $\mathbf{f}$ to optimal value function and conditional distribution $\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$ of the MDP (+ condition on T)

In general

- ridge regression

$$\mathbf{f}\,(\mathbf{s}, \mathbf{a}) = \mathbb{E}\left[\,\mathbf{s}_+ \,|\, \mathbf{s}, \mathbf{a}\,\right]$$

- max likelihood

$$\mathbf{f}\,(\mathbf{s}, \mathbf{a}) = \max \rho\left[\,\mathbf{s}_+ \,|\, \mathbf{s}, \mathbf{a}\,\right]$$

do not satisfy the optimality conditions

Model from "classic SYSID" does not necessarily yield the best MPC policy

Notable exceptions: model gives $\mathbb{E}\left[\,\mathbf{s}_+ \,|\, \mathbf{s}, \mathbf{a}\,\right]$ and

- LQR with process noise
- By extension, locally optimal policies for
  - dissipative MDP and
  - $V^\star \approx$ quadratic in positive invariant set

Smooth tracking MPC, fixed reference away from constraints, "reasonable" stochasticity

Sufficient Condition for MPC Optimality (for MDPs)

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

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

Sufficient condition of optimality of MPC
relates model $\mathbf{f}$ to optimal value function and conditional distribution $\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$ of the MDP
($+$ condition on T)

When is RL (most) useful for MPC? *"practical" view*

- Not much if smooth tracking problem, spend most of the time close to fixed steady state, rare transients away from the constraints $\rightarrow$ small performance gain

Sufficient Condition for MPC Optimality (for MDPs)

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

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

Sufficient condition of optimality of MPC relates model $\mathbf{f}$ to optimal value function and conditional distribution $\mathbf{s}, \mathbf{a} \rightarrow \mathbf{s}_+$ of the MDP (+ condition on T)

When is RL (most) useful for MPC? *"practical" view*

- Not much if smooth tracking problem, spend most of the time close to fixed steady state, rare transients away from the constraints $\rightarrow$ small performance gain
- Model is not rich enough to predict the (relevant) expected state transition
- Economic problem with "low dissipativity" (trajectories spread over the state space)
- Value function curvature changes a lot, problem is non-smooth
- Optimal steady-state near / at constraints
- Varying exogeneous inputs (e.g. varying prices in energy-related problems, changing reference)
- Task-based problems: racing, minimum time, termination conditions, etc.

Outline

Fully Parametrized MPC Can Describe the MDP Solution

Fully Parametrized MPC Can Describe the MDP Solution

The theory is about equivalences between MDPs

Fully Parametrized MPC Can Describe the MDP Solution

The theory is about equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\,\cdot\,|\mathbf{s}, \mathbf{a}]$

Fully Parametrized MPC Can Describe the MDP Solution

The theory is about equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\,\cdot\,|\mathbf{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathbf{s}) = \min_{\mathbf{a}} Q^\star(\mathbf{s}, \mathbf{a})$$

$$Q^\star(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathbf{s}_+ \sim \rho}\left[V^\star(\mathbf{s}_+)\,|\mathbf{s}, \mathbf{a}\right]$$

Fully Parametrized MPC Can Describe the MDP Solution

The theory is about equivalences between MDPs

World MDP

World MDP definition

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathrm{s}) = \min_{\mathbf{a}} Q^\star(\mathrm{s}, \mathbf{a})$$

$$Q^\star(\mathrm{s}, \mathbf{a}) = L(\mathrm{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathrm{s}_+ \sim \rho}[V^\star(\mathrm{s}_+)|\mathrm{s}, \mathbf{a}]$$

Optimal policy

$$\boldsymbol{\pi}^\star(\mathrm{s}) = \arg\min_{\mathbf{a}} Q^\star(\mathrm{s}, \mathbf{a})$$

Fully Parametrized MPC Can Describe the MDP Solution

The theory is about equivalences between MDPs

World MDP

World MDP definition

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathrm{s}) = \min_{\mathbf{a}} Q^\star(\mathrm{s}, \mathbf{a})$$

$$Q^\star(\mathrm{s}, \mathbf{a}) = L(\mathrm{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathrm{s}_+ \sim \rho}[V^\star(\mathrm{s}_+)|\mathrm{s}, \mathbf{a}]$$

Optimal policy

$$\boldsymbol{\pi}^\star(\mathrm{s}) = \arg\min_{\mathbf{a}} Q^\star(\mathrm{s}, \mathbf{a})$$

Model MDP

Model MDP definition

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Optimal value functions

$$\hat{V}^\star(\mathrm{s}) = \min_{\mathbf{a}} \hat{Q}^\star(\mathrm{s}, \mathbf{a})$$

$$\hat{Q}^\star(\mathrm{s}, \mathbf{a}) = \hat{L}(\mathrm{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathrm{s}_+ \sim \hat{\rho}}[\, V^\star(\hat{\mathrm{s}}_+)|\, \mathrm{s}, \mathbf{a}]$$

Optimal policy

$$\hat{\boldsymbol{\pi}}^\star(\mathrm{s}) = \arg\min_{\mathbf{a}} \hat{Q}^\star(\mathrm{s}, \mathbf{a})$$

Fully Parametrized MPC Can Describe the MDP Solution

The theory is about equivalences between MDPs

World MDP

World MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s},\mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s},\mathbf{a}]$

Optimal value functions

$$V^\star(\mathbf{s}) = \min_{\mathbf{a}}\, Q^\star(\mathbf{s},\mathbf{a})$$

$$Q^\star(\mathbf{s},\mathbf{a}) = L(\mathbf{s},\mathbf{a}) + \gamma\mathbb{E}_{\mathbf{s}_+\sim\rho}\left[V^\star(\mathbf{s}_+)\,|\mathbf{s},\mathbf{a}\right]$$

Optimal policy

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

Model MDP

Model MDP definition

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s},\mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s},\mathbf{a}]$

Optimal value functions

$$\hat{V}^\star(\mathbf{s}) = \min_{\mathbf{a}}\, \hat{Q}^\star(\mathbf{s},\mathbf{a})$$

$$\hat{Q}^\star(\mathbf{s},\mathbf{a}) = \hat{L}(\mathbf{s},\mathbf{a}) + \gamma\mathbb{E}_{\mathbf{s}_+\sim\hat{\rho}}\left[\,V^\star(\hat{\mathbf{s}}_+)\right|\,\mathbf{s},\mathbf{a}]$$

Optimal policy

$$\hat{\boldsymbol{\pi}}^\star(\mathbf{s}) = \arg\min_{\mathbf{a}}\, \hat{Q}^\star(\mathbf{s},\mathbf{a})$$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

Fully Parametrized MPC Can Describe the MDP Solution

The theory is about equivalences between MDPs

World MDP

World MDP definition

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Optimal value functions

$$V^\star(\mathrm{s}) = \min_{\mathbf{a}} Q^\star(\mathrm{s}, \mathbf{a})$$

$$Q^\star(\mathrm{s}, \mathbf{a}) = L(\mathrm{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathrm{s}_+ \sim \rho}\left[V^\star(\mathrm{s}_+)|\mathrm{s}, \mathbf{a}\right]$$

Optimal policy

$$\boldsymbol{\pi}^\star(\mathrm{s}) = \arg\min_{\mathbf{a}} Q^\star(\mathrm{s}, \mathbf{a})$$

Model MDP

Model MDP definition

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Optimal value functions

$$\hat{V}^\star(\mathrm{s}) = \min_{\mathbf{a}} \hat{Q}^\star(\mathrm{s}, \mathbf{a})$$

$$\hat{Q}^\star(\mathrm{s}, \mathbf{a}) = \hat{L}(\mathrm{s}, \mathbf{a}) + \gamma \mathbb{E}_{\mathrm{s}_+ \sim \hat{\rho}}\left[\left. V^\star(\hat{\mathrm{s}}_+)\right| \mathrm{s}, \mathbf{a}\right]$$

Optimal policy

$$\hat{\boldsymbol{\pi}}^\star(\mathrm{s}) = \arg\min_{\mathbf{a}} \hat{Q}^\star(\mathrm{s}, \mathbf{a})$$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

Proof: telescopic sum, some non-trivial assumptions to prevent $\infty - \infty$ cancellations

More on MDP Equivalences

World MDP

- States s and actions a
- Cost $L(\mathrm{s},\mathrm{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s},\mathrm{a}]$

Model MDP

- States s and actions a
- Cost $\hat{L}(\mathrm{s},\mathrm{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s},\mathrm{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

More on MDP Equivalences

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$

More on MDP Equivalences

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)

More on MDP Equivalences

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L\left(\mathbf{s},\mathbf{a}\right)$
- Transition $\mathbf{s}_+ \sim \rho\left[\,\cdot\,|\mathbf{s},\mathbf{a}\right]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}\left(\mathbf{s},\mathbf{a}\right)$
- Transition $\mathbf{s}_+ \sim \hat{\rho}\left[\,\cdot\,|\mathbf{s},\mathbf{a}\right]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

More on MDP Equivalences

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s},\mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s},\mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s},\mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s},\mathbf{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ “optimality” $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

MPC is a “weird” undiscounted Model MDP

- Deterministic model is a trivial $\hat{\rho}$, then optimal policy for Model MDP $\equiv$ planning
- Finite horizon: ok if correct terminal cost (strong assumption, but learning can tune it)
- Scenario tree: rough $\hat{\rho}$ and policy.
- Robust MPC: carries support $\hat{\rho}$, worst-case cost is a bit off...

More on MDP Equivalences

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it

More on MDP Equivalences

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then $\mathbf{f}(\mathbf{s}, \mathbf{a}) \subset$ support of ρ

More on MDP Equivalences

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho\left[\,\cdot\,|\mathbf{s}, \mathbf{a}\right]$

Model MDP

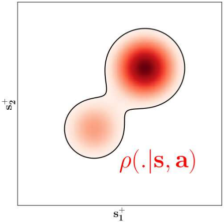
- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}\left[\,\cdot\,|\mathbf{s}, \mathbf{a}\right]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then $\mathbf{f}(\mathbf{s}, \mathbf{a}) \subset$ support of ρ



More on MDP Equivalences

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot\,|\mathrm{s}, \mathbf{a}]$

Model MDP

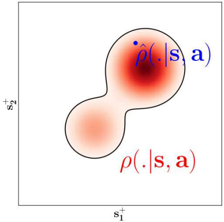
- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot\,|\mathrm{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then $\mathbf{f}(\mathrm{s}, \mathbf{a}) \subset$ support of ρ



More on MDP Equivalences

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s},\mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s},\mathbf{a}]$

Model MDP

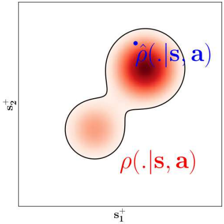
- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s},\mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s},\mathbf{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then $\mathbf{f}(\mathbf{s},\mathbf{a}) \subset$ support of ρ
- *More simply said*: we can build a "plausible" optimal deterministic model $\mathbf{f}$ (predictions have prob. > 0).



More on MDP Equivalences

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

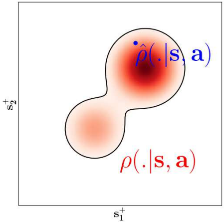
- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that

- Under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$
- For $\hat{L} = L$ there is a set of models $\hat{\rho}$ such that $\hat{Q}^\star = Q^\star$
- Conditions for model $\hat{\rho}$ "optimality" $\neq$ min of classical loss functions (except. LQR)
- World MDP and Model MDP do not need to use the same discount γ

Remarks

- Non-unique optimal model $\hat{\rho}$ leaves room for aligning it
- If $V^\star$ is continuous and support of ρ is bounded(?) and path-connected, then $\mathbf{f}(\mathbf{s}, \mathbf{a}) \subset$ support of ρ
- *More simply said*: we can build a "plausible" optimal deterministic model $\mathbf{f}$ (predictions have prob. > 0).
- Cost modification promotes "simplicity".



Is there a catch?

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

Technical assumptions? They are very technical and not intuitive at all. But not very demanding, and there is a "self-fulfilling" effect in learning, i.e. RL "learns" to satisfy them.

Is there a catch?

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s},\mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\,\cdot\,|\mathrm{s},\mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s},\mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\,\cdot\,|\mathrm{s},\mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

But we are making a strong assumption (in plain sight)

Is there a catch?

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

But we are making a strong assumption (in plain sight)

- Theory assumes that World MDP and Model MDP have the same state $\mathbf{s}$
- I.e. we need to know the state of the World MDP to build a correct Model MDP
- That is often unrealistic...

Is there a catch?

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

But we are making a strong assumption (in plain sight)

- Theory assumes that World MDP and Model MDP have the same state s
- I.e. we need to know the state of the World MDP to build a correct Model MDP
- That is often unrealistic...

The end of a nice theory? No, but it is a word of caution...

Is there a catch?

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

Eventually the Markov state s "boils down to" the history of past observations

How to embed that in a Model MDP?

Is there a catch?

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

Eventually the Markov state s "boils down to" the history of past observations

How to embed that in a Model MDP?

1. "Input-output models": ARX, multi-step predictors (linear or not), etc.

Is there a catch?

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \rho[\,\cdot\,|\mathrm{s}, \mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_+ \sim \hat{\rho}[\,\cdot\,|\mathrm{s}, \mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

Eventually the Markov state s "boils down to" the history of past observations

How to embed that in a Model MDP?

1. "Input-output models": ARX, multi-step predictors (linear or not), etc.
2. Latent states (a.k.a. embeddings or compressed representations) giving an AI-driven lower-dimensional version of that history

Is there a catch?

World MDP

- States s and actions $\mathbf{a}$
- Cost $L(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_{+} \sim \rho[\cdot|\mathrm{s}, \mathbf{a}]$

Model MDP

- States s and actions $\mathbf{a}$
- Cost $\hat{L}(\mathrm{s}, \mathbf{a})$
- Transition $\mathrm{s}_{+} \sim \hat{\rho}[\cdot|\mathrm{s}, \mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^{\star} = Q^{\star}$

Eventually the Markov state s "boils down to" the history of past observations

How to embed that in a Model MDP?

1. "Input-output models": ARX, multi-step predictors (linear or not), etc.
2. Latent states (a.k.a. embeddings or compressed representations) giving an AI-driven lower-dimensional version of that history
3. State observers giving a model-based meaningful low-dimensional version of that history

Is there a catch?

World MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $L(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \rho[\cdot|\mathbf{s}, \mathbf{a}]$

Model MDP

- States $\mathbf{s}$ and actions $\mathbf{a}$
- Cost $\hat{L}(\mathbf{s}, \mathbf{a})$
- Transition $\mathbf{s}_+ \sim \hat{\rho}[\cdot|\mathbf{s}, \mathbf{a}]$

Theory says that - under some technical conditions - there is a $\hat{L}$ such that $\hat{Q}^\star = Q^\star$

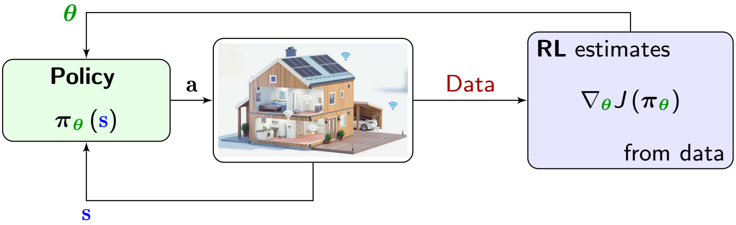
Eventually the Markov state $\mathbf{s}$ "boils down to" the history of past observations

How to embed that in a Model MDP?

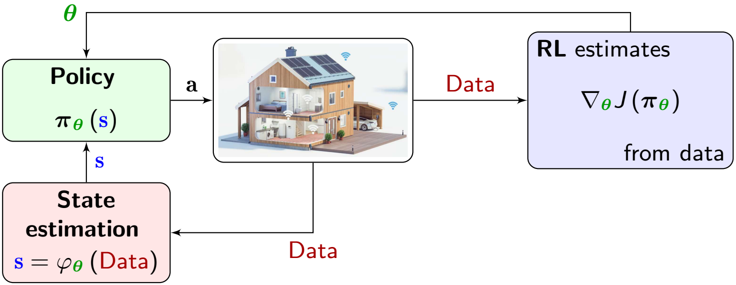
1. "Input-output models": ARX, multi-step predictors (linear or not), etc.
2. Latent states (a.k.a. embeddings or compressed representations) giving an AI-driven lower-dimensional version of that history
3. State observers giving a model-based meaningful low-dimensional version of that history

Options 2 and 3 are part of the decision-making process, back-propagate through policy + state "constructor".

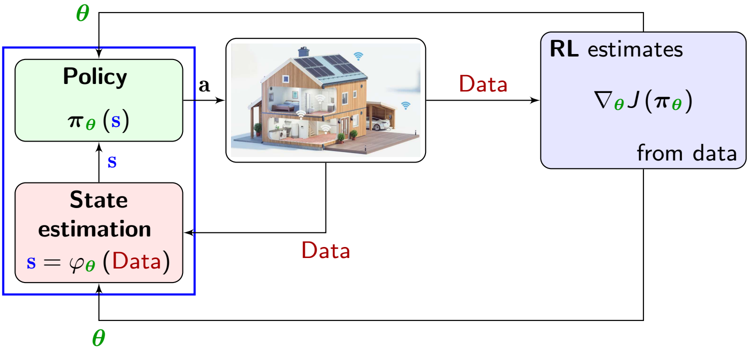
State Estimation in the Learning Loop



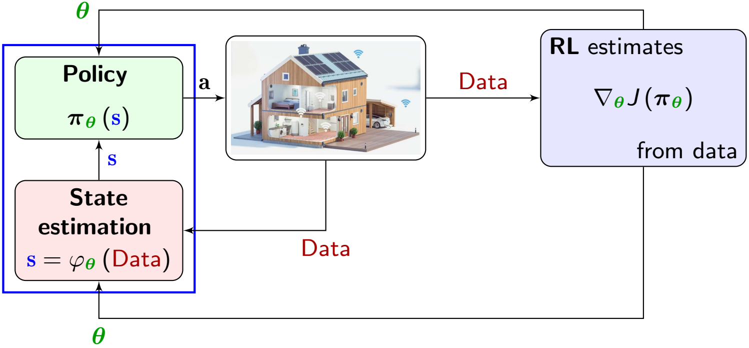
State Estimation in the Learning Loop



State Estimation in the Learning Loop



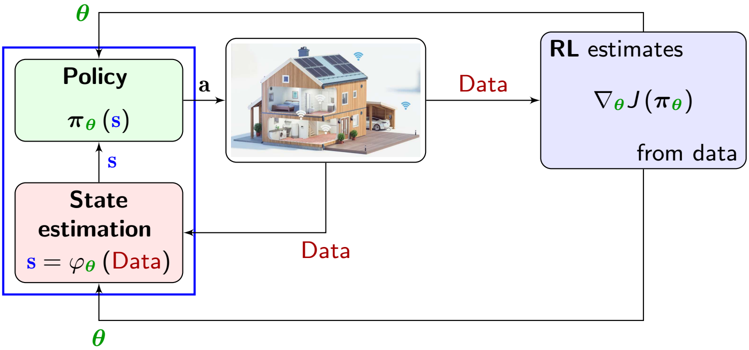
State Estimation in the Learning Loop



Remarks

- Policy is $\boldsymbol{\pi}^{\text{D}}_{\boldsymbol{\theta}}(\text{Data}) = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\varphi_{\boldsymbol{\theta}}(\text{Data}))$

State Estimation in the Learning Loop

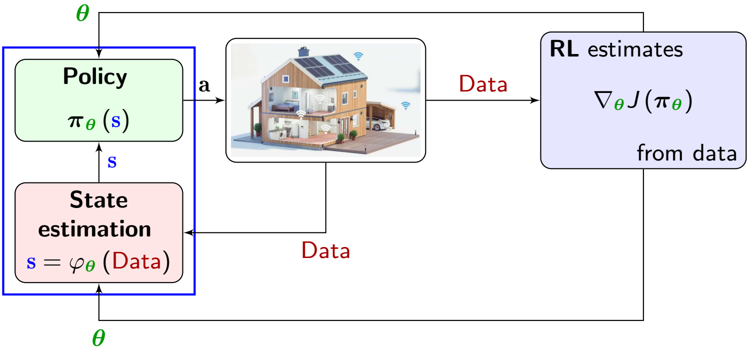


Remarks

- Policy is $\boldsymbol{\pi}^{\mathsf{D}}_{\boldsymbol{\theta}}(\mathsf{Data}) = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\varphi_{\boldsymbol{\theta}}(\mathsf{Data}))$
- Gradient of the policy now is:

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}^{\mathsf{D}}_{\boldsymbol{\theta}}(\mathsf{Data}) = \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \nabla_{\boldsymbol{\theta}} \varphi_{\boldsymbol{\theta}}(\mathsf{Data}) \nabla_{\mathbf{s}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$$

State Estimation in the Learning Loop



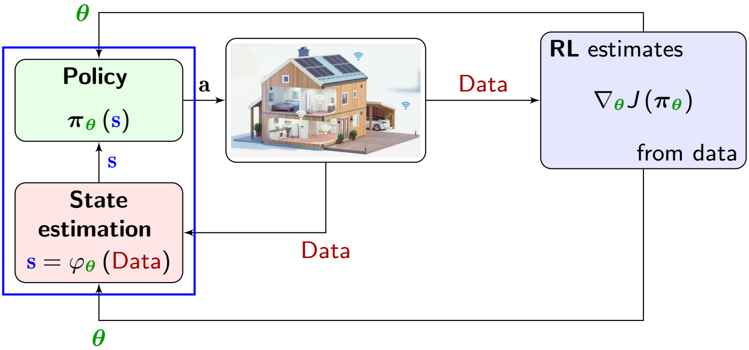
Remarks

- Policy is $\boldsymbol{\pi}^{\mathsf{D}}_{\boldsymbol{\theta}}(\text{Data}) = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\varphi_{\boldsymbol{\theta}}(\text{Data}))$
- Gradient of the policy now is:

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}^{\mathsf{D}}_{\boldsymbol{\theta}}(\text{Data}) = \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \nabla_{\boldsymbol{\theta}} \varphi_{\boldsymbol{\theta}}(\text{Data}) \nabla_{\mathbf{s}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})$$

- Critic is looking at Data $\xrightarrow{\boldsymbol{\theta}} \mathbf{a}$, i.e. $Q^{\boldsymbol{\pi}^{\mathsf{D}}_{\boldsymbol{\theta}}}(\text{Data}, \mathbf{a})$

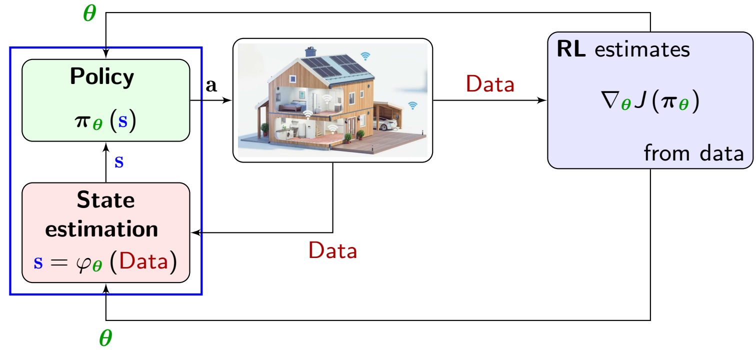
State Estimation in the Learning Loop



If φ_{θ} (Data) is

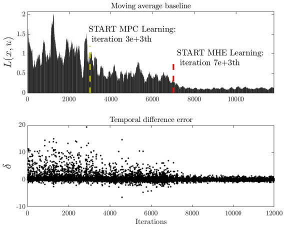
- MHE, then its gradient is computed using NLP sensitivities (same as MPC)
- DNN, then efficient sensitivity computations exist
- Bayesian inference? Belief state? Open research topic...

State Estimation in the Learning Loop



If φ_θ (Data) is

- MHE, then its gradient is computed using NLP sensitivities (same as MPC)
- DNN, then efficient sensitivity computations exist
- Bayesian inference? Belief state? Open research topic...



Multi-Step Predictive Control

Systems with

- $\sim$Linear dynamics
- Input-output data
- Significant stochasticity
- Modelling is difficult

Multi-Step Predictive Control

Systems with

- $\sim$Linear dynamics
- Input-output data
- Significant stochasticity
- Modelling is difficult

Multi-step linear predictors

$$\hat{\mathbf{y}} = \Phi \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

- Recent history of input-output sequence $\mathbf{u}, \mathbf{y}$
- Planned input sequence $\mathbf{u}$
- Predicted output sequence $\hat{\mathbf{y}}$

Multi-Step Predictive Control

SPC for $\mathbf{u}$, $\mathbf{y}$ given

$$\min_{\mathbf{u},\,\hat{\mathbf{y}}} \quad \sum_{k=0}^{N} L\left(\hat{\mathbf{y}}_k, \mathbf{u}_k\right)$$

$$\text{s.t.} \quad \hat{\mathbf{y}} = \Phi \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

$$\mathbf{h}\left(\hat{\mathbf{y}}_k, \mathbf{u}_k\right) \leq 0$$

yields policy $\boldsymbol{\pi}\left(\mathbf{u}, \mathbf{y}\right) = \mathbf{u}_0^\star$

Multi-step linear predictors

$$\hat{\mathbf{y}} = \Phi \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

- Recent history of input-output sequence $\mathbf{u}, \mathbf{y}$
- Planned input sequence $\mathbf{u}$
- Predicted output sequence $\hat{\mathbf{y}}$

Multi-Step Predictive Control

SPC for $\mathbf{u}$, $\mathbf{y}$ given

$$\min_{\mathbf{u},\,\hat{\mathbf{y}}} \quad \sum_{k=0}^{N} L\left(\hat{\mathbf{y}}_k, \mathbf{u}_k\right)$$

$$\text{s.t.} \quad \hat{\mathbf{y}} = \Phi \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

$$\mathbf{h}\left(\hat{\mathbf{y}}_k, \mathbf{u}_k\right) \leq 0$$

yields policy $\boldsymbol{\pi}\left(\mathbf{u}, \mathbf{y}\right) = \mathbf{u}_0^\star$

Multi-step linear predictors

$$\hat{\mathbf{y}} = \Phi \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

- Recent history of input-output sequence $\mathbf{u}, \mathbf{y}$
- Planned input sequence $\mathbf{u}$
- Predicted output sequence $\hat{\mathbf{y}}$
- Measured output sequences $\mathbf{y}$

matrix Φ can be **built from past data** $\mathcal{D}$, e.g.

$$\min_{\Phi} \quad \sum_{i\in\mathcal{D}} \frac{1}{2} \left\| \mathbf{y}_i - \Phi \begin{bmatrix} \mathbf{u}_i \\ \mathbf{y}_i \\ \mathbf{u}_i \end{bmatrix} \right\|^2 + R(\Phi)$$

$$\text{s.t.} \quad \Phi \text{ causal}$$

or alternative loss functions (e.g. quantile)

Multi-Step Predictive Control

SPC for $\mathbf{u}$, $\mathbf{y}$ given

$$\min_{\mathbf{u},\,\hat{\mathbf{y}}} \quad \sum_{k=0}^{N} L\left(\hat{\mathbf{y}}_k, \mathbf{u}_k\right)$$

$$\text{s.t.} \quad \hat{\mathbf{y}} = \Phi \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

$$\mathbf{h}\left(\hat{\mathbf{y}}_k, \mathbf{u}_k\right) \leq 0$$

yields policy $\boldsymbol{\pi}\left(\mathbf{u}, \mathbf{y}\right) = \mathbf{u}_0^\star$

If R promotes a dynamic system structure, regression well-posed even with limited data (and Φ high-dimensional)

But best model fit $\not\Rightarrow$ best policy

Multi-step linear predictors

$$\hat{\mathbf{y}} = \Phi \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

- Recent history of input-output sequence $\mathbf{u}, \mathbf{y}$
- Planned input sequence $\mathbf{u}$
- Predicted output sequence $\hat{\mathbf{y}}$
- Measured output sequences $\mathbf{y}$

matrix Φ can be **built from past data** $\mathcal{D}$, e.g.

$$\min_{\Phi} \quad \sum_{i\in\mathcal{D}} \frac{1}{2} \left\| \mathbf{y}_i - \Phi \begin{bmatrix} \mathbf{u}_i \\ \mathbf{y}_i \\ \mathbf{u}_i \end{bmatrix} \right\|^2 + R(\Phi)$$

$$\text{s.t.} \quad \Phi \text{ causal}$$

or alternative loss functions (e.g. quantile)

Multi-Step Predictive Control

SPC for $\mathbf{u}$, $\mathbf{y}$ given

$$\min_{\mathbf{u},\,\hat{\mathbf{y}}} \quad \sum_{k=0}^{N} L\left(\hat{\mathbf{y}}_k, \mathbf{u}_k\right)$$

$$\text{s.t.} \quad \hat{\mathbf{y}} = \Phi \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

$$\mathbf{h}\left(\hat{\mathbf{y}}_k, \mathbf{u}_k\right) \leq 0$$

yields policy $\boldsymbol{\pi}\left(\mathbf{u}, \mathbf{y}\right) = \mathbf{u}_0^\star$

But best model fit $\not\Rightarrow$ best policy

Can we close the gap via RL? Yes!

- RL-MPC theory applies with some twists
- State becomes $\mathbf{u}$, $\mathbf{y}$ (window of input-output)

Multi-Step Predictive Control

MSPC for $\mathbf{u}$, $\mathbf{y}$ given

$$\min_{\mathbf{u},\,\hat{\mathbf{y}}} \quad \Psi_{\boldsymbol{\theta}}\left(\mathbf{u},\, \hat{\mathbf{y}}, \mathbf{u}, \mathbf{y}\right)$$

$$\text{s.t.} \quad \hat{\mathbf{y}} = \Phi_{\boldsymbol{\theta}} \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

$$\mathbf{H}_{\boldsymbol{\theta}}\left(\mathbf{u},\, \hat{\mathbf{y}}, \mathbf{u}, \mathbf{y}\right) \leq 0$$

yields policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{u}, \mathbf{y}\right) = \mathbf{u}_0^\star$

Can we close the gap via RL? Yes!

- RL-MPC theory applies with some twists
- State becomes $\mathbf{u}$, $\mathbf{y}$ (window of input-output)
- Modifications in principle not localized in time

But best model fit $\not\Rightarrow$ best policy

Multi-Step Predictive Control

MSPC for $\mathbf{u}$, $\mathbf{y}$ given

$$\min_{\mathbf{u},\,\hat{\mathbf{y}}} \quad \Psi_{\boldsymbol{\theta}}\left(\mathbf{u},\, \hat{\mathbf{y}}, \mathbf{u}, \mathbf{y}\right)$$

$$\text{s.t.} \quad \hat{\mathbf{y}} = \Phi_{\boldsymbol{\theta}} \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

$$\mathbf{H}_{\boldsymbol{\theta}}\left(\mathbf{u},\, \hat{\mathbf{y}}, \mathbf{u}, \mathbf{y}\right) \leq 0$$

yields policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{u}, \mathbf{y}\right) = \mathbf{u}_0^\star$

But best model fit $\not\Rightarrow$ best policy

Can we close the gap via RL? Yes!

- RL-MPC theory applies with some twists
- State becomes $\mathbf{u}$, $\mathbf{y}$ (window of input-output)
- Modifications in principle not localized in time
- High-dimensional parameter space for RL

Multi-Step Predictive Control

MSPC for $\mathbf{u}$, $\mathbf{y}$ given

$$\begin{aligned} \min_{\mathbf{u},\,\hat{\mathbf{y}}} \quad & \Psi_{\boldsymbol{\theta}}\left(\mathbf{u},\, \hat{\mathbf{y}}, \mathbf{u}, \mathbf{y}\right) \\ \text{s.t.} \quad & \hat{\mathbf{y}} = \Phi_{\boldsymbol{\theta}} \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix} \\ & \mathbf{H}_{\boldsymbol{\theta}}\left(\mathbf{u},\, \hat{\mathbf{y}}, \mathbf{u}, \mathbf{y}\right) \leq 0 \end{aligned}$$

yields policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{u}, \mathbf{y}\right) = \mathbf{u}_0^\star$

But best model fit $\not\Rightarrow$ best policy

Can we close the gap via RL? Yes!

- RL-MPC theory applies with some twists
- State becomes $\mathbf{u}$, $\mathbf{y}$ (window of input-output)
- Modifications in principle not localized in time
- High-dimensional parameter space for RL

We need an add-on in Leap-C for high-speed MPC + sensitivity on this type of model structure! *For now we are down to using CVXPY.*

Multi-Step Predictive Control

MSPC for $\mathbf{u}$, $\mathbf{y}$ given

$$\min_{\mathbf{u},\,\hat{\mathbf{y}}} \quad \Psi_{\boldsymbol{\theta}}\left(\mathbf{u},\, \hat{\mathbf{y}}, \mathbf{u}, \mathbf{y}\right)$$

$$\text{s.t.} \quad \hat{\mathbf{y}} = \Phi_{\boldsymbol{\theta}} \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \\ \mathbf{u} \end{bmatrix}$$

$$\mathbf{H}_{\boldsymbol{\theta}}\left(\mathbf{u},\, \hat{\mathbf{y}}, \mathbf{u}, \mathbf{y}\right) \leq 0$$

yields policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{u}, \mathbf{y}\right) = \mathbf{u}_0^{\star}$

But best model fit $\not\Rightarrow$ best policy

Can we close the gap via RL? Yes!

- RL-MPC theory applies with some twists
- State becomes $\mathbf{u}$, $\mathbf{y}$ (window of input-output)
- Modifications in principle not localized in time
- High-dimensional parameter space for RL

We need an add-on in Leap-C for high-speed MPC + sensitivity on this type of model structure! *For now we are down to using CVXPY.*

Orientation

What we have seen:

- MPC can be understood as a model of the optimal action-value function $Q^\star$ of real-world MDPs and/or of the optimal policy $\boldsymbol{\pi}^\star$
- MPC cost (and constraints) become part of that model
- Model that best fits the real-world does not (necessarily) yield the best policy
- RL is a toolbox to tune the MPC as a model of the MDP solution
- MPC state space should match the real world, strong assumption that can be alleviated

Orientation

What we have seen:

- MPC can be understood as a model of the optimal action-value function $Q^\star$ of real-world MDPs and/or of the optimal policy $\boldsymbol{\pi}^\star$
- MPC cost (and constraints) become part of that model
- Model that best fits the real-world does not (necessarily) yield the best policy
- RL is a toolbox to tune the MPC as a model of the MDP solution
- MPC state space should match the real world, strong assumption that can be alleviated

What we will do next: RL over MPC

- Safe & Stable RL over MPC *(In the afternoon)*
- RL over MPC with belief states – a future prospect *(In the afternoon)*
- Beyond MPC – Model-based Decisions and AI for decisions *(Tomorrow)*

Thanks for your attention!

ResearchGate

Google Scholar