

What we have seen:

- MPC can be understood as a model of the optimal action-value function Q^* of real-world MDPs and/or of the optimal policy π^*
- 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 have seen:

- MPC can be understood as a model of the optimal action-value function Q^* of real-world MDPs and/or of the optimal policy π^*
- 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
- Exploration
- Some more explored questions

RL and MPC

Safety, Stability, Belief State

Sebastien Gros

Dept. of Cybernetic, NTNU
Faculty of Information Tech.

Freiburg PhD School

Outline

- 1 SYSID & MPC
- 2 Safe RL via MPC
- 3 Exploration with MPC
- 4 Stability-constrained Learning with MPC
- 5 Explored questions

Outline

- 1 SYSID & MPC
- 2 Safe RL via MPC
- 3 Exploration with MPC
- 4 Stability-constrained Learning with MPC
- 5 Explored questions

Full MPC parametrization:

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

Remarks

- In general, θ such that

$$Q_{\theta}^{\text{MPC}}(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a})$$

is not unique

- Leaves room to align the model

Full MPC parametrization:

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

Remarks

- In general, θ such that

$$Q_{\theta}^{\text{MPC}}(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a})$$

is not unique

- Leaves room to align the model

Multi-objective approach

$$\min_{\theta} \underbrace{J(\pi_{\theta}^{\text{MPC}})}_{\text{RL}} + \omega \cdot \underbrace{\mathbb{E}[\|\mathbf{f}_{\theta}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\|^2]}_{\text{SYSID}}$$

is non-conflicting for a rich parametrization

RL & SYSID for MPC

Full MPC parametrization:

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

Remarks

- In general, θ such that

$$Q_{\theta}^{\text{MPC}}(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a})$$

is not unique

- Leaves room to align the model

Multi-objective approach

$$\min_{\theta} \underbrace{J(\pi_{\theta}^{\text{MPC}})}_{\text{RL}} + \omega \cdot \underbrace{\mathbb{E} [\|\mathbf{f}_{\theta}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\|^2]}_{\text{SYSID}}$$

is non-conflicting for a rich parametrization

RL-prioritized

$$\begin{aligned} \min_{\theta} \quad & \mathbb{E} [\|\mathbf{f}_{\theta}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\|^2] \\ \text{s.t.} \quad & \nabla_{\theta} J(\pi_{\theta}^{\text{MPC}}) = 0 \end{aligned}$$

RL & SYSID for MPC

Full MPC parametrization:

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

Remarks

- In general, θ such that

$$Q_{\theta}^{\text{MPC}}(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a})$$

is not unique

- Leaves room to align the model

Multi-objective approach

$$\min_{\theta} \underbrace{J(\pi_{\theta}^{\text{MPC}})}_{\text{RL}} + \omega \cdot \underbrace{\mathbb{E}[\|\mathbf{f}_{\theta}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\|^2]}_{\text{SYSID}}$$

is non-conflicting for a rich parametrization

RL-prioritized

$$\begin{aligned} \min_{\theta} \quad & \mathbb{E}[\|\mathbf{f}_{\theta}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\|^2] \\ \text{s.t.} \quad & \nabla_{\theta} J(\pi_{\theta}^{\text{MPC}}) = 0 \end{aligned}$$

- Yields different ways of combining SYSID & RL
- SYSID serves as “regularization” for RL
- Especially useful for θ large-dimensional

RL & SYSID for MPC

Full MPC parametrization:

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

Remarks

- In general, θ such that

$$Q_{\theta}^{\text{MPC}}(\mathbf{s}, \mathbf{a}) = Q^*(\mathbf{s}, \mathbf{a})$$

is not unique

- Leaves room to align the model

Multi-objective approach

$$\min_{\theta} \underbrace{J(\pi_{\theta}^{\text{MPC}})}_{\text{RL}} + \omega \cdot \underbrace{\mathbb{E}[\|\mathbf{f}_{\theta}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\|^2]}_{\text{SYSID}}$$

is non-conflicting for a rich parametrization

- Yields different ways of combining SYSID & RL
- SYSID serves as “regularization” for RL
- Especially useful for θ large-dimensional

RL-prioritized

$$\begin{aligned} \min_{\theta} \quad & \mathbb{E}[\|\mathbf{f}_{\theta}(\mathbf{s}, \mathbf{a}) - \mathbf{s}_+\|^2] \\ \text{s.t.} \quad & \nabla_{\theta} J(\pi_{\theta}^{\text{MPC}}) = 0 \end{aligned}$$

Some “draft” theory on how meaningful model $\mathbf{f}_{\theta}$ can be made to be

Outline

- 1 SYSID & MPC
- 2 Safe RL via MPC
- 3 Exploration with MPC
- 4 Stability-constrained Learning with MPC
- 5 Explored questions

Robust MPC - Uncertainty model

True system: $\mathbf{s}_+ \sim \mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$

Deterministic model: $\hat{\mathbf{s}}_+ = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a})$

Robust MPC - Uncertainty model

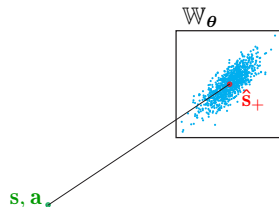
True system: $\mathbf{s}_+ \sim \mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$

Deterministic model: $\hat{\mathbf{s}}_+ = \mathbf{f}_\theta(\mathbf{s}, \mathbf{a})$

Dispersion: $\mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) + \mathbb{W}_\theta$ contains the support of $\mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$, i.e.

$$\mathbf{s}_+ \in \mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) + \mathbb{W}_\theta \quad (1)$$

with probability 1



Robust MPC - Uncertainty model

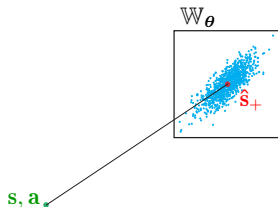
True system: $\mathbf{s}_+ \sim \mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$

Deterministic model: $\hat{\mathbf{s}}_+ = \mathbf{f}_\theta(\mathbf{s}, \mathbf{a})$

Dispersion: $\mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) + \mathbb{W}_\theta$ contains the support of $\mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$, i.e.

$$\mathbf{s}_+ \in \mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) + \mathbb{W}_\theta \quad (1)$$

with probability 1



Remarks:

- Identifying $\mathbb{W}_\theta$ is a set-membership identification problem, well studied
- Obviously $\mathbb{W}_\theta$ is not unique
- Ensuring probability 1 from data is impossible
→ probabilistic guarantees
- Model parameters θ must be such that (1) holds on every known data point

Robust MPC - Uncertainty model

True system: $\mathbf{s}_+ \sim \mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$

Deterministic model: $\hat{\mathbf{s}}_+ = \mathbf{f}_\theta(\mathbf{s}, \mathbf{a})$

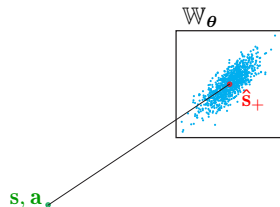
Dispersion: $\mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) + \mathbb{W}_\theta$ contains the support of $\mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$, i.e.

$$\mathbf{s}_+ \in \mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) + \mathbb{W}_\theta \quad (1)$$

with probability 1

Remarks:

- Identifying $\mathbb{W}_\theta$ is a set-membership identification problem, well studied
- Obviously $\mathbb{W}_\theta$ is not unique
- Ensuring probability 1 from data is impossible
→ probabilistic guarantees
- Model parameters θ must be such that (1) holds on every known data point



Condition

$$\mathbf{s}_+ - \mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) \in \mathbb{W}_\theta$$

for all observed triplets $(\mathbf{s}, \mathbf{a}, \mathbf{s}_+)$

→ constraints on θ

Robust MPC - Uncertainty model

True system: $\mathbf{s}_+ \sim \mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$

Deterministic model: $\hat{\mathbf{s}}_+ = \mathbf{f}_\theta(\mathbf{s}, \mathbf{a})$

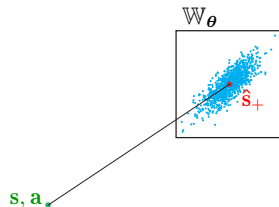
Dispersion: $\mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) + \mathbb{W}_\theta$ contains the support of $\mathbb{P}[\cdot | \mathbf{s}, \mathbf{a}]$, i.e.

$$\mathbf{s}_+ \in \mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) + \mathbb{W}_\theta \quad (1)$$

with probability 1

Remarks:

- Identifying $\mathbb{W}_\theta$ is a set-membership identification problem, well studied
- Obviously $\mathbb{W}_\theta$ is not unique
- Ensuring probability 1 from data is impossible
→ probabilistic guarantees
- Model parameters θ must be such that (1) holds on every known data point



Condition

$$\mathbf{s}_+ - \mathbf{f}_\theta(\mathbf{s}, \mathbf{a}) \in \mathbb{W}_\theta$$

for all observed triplets $(\mathbf{s}, \mathbf{a}, \mathbf{s}_+)$

→ constraints on θ

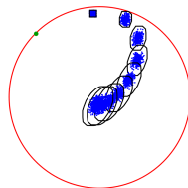
Containing the model-system mismatch becomes constraints in the parameters θ . Constraints can be readily formulated in terms of data.

Safe policies via Robust (N)MPC

Robust (N)MPC delivers policy $\pi_{\theta}(\mathbf{x}_0) = \mathbf{u}_0^*$ from

$$\mathbf{u}^* = \arg \min_{\mathbf{u}} \max_{\mathbf{w} \in \mathbb{W}_{\theta}^N} T_{\theta}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

s.t. $\mathbf{u}_{0,\dots,N} \in \mathbb{U}$



- $\mathbf{x}_{0,\dots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$

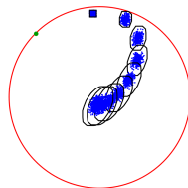
Safe policies via Robust (N)MPC

Robust (N)MPC delivers policy $\pi_{\theta}(\mathbf{x}_0) = \mathbf{u}_0^*$ from

$$\mathbf{u}^* = \arg \min_{\mathbf{u}} \max_{\mathbf{w} \in \mathbb{W}_{\theta}^N} T_{\theta}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\text{s.t. } \mathbf{u}_{0,\dots,N} \in \mathbb{U}$$

$$\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \theta, \mathbf{w}) \in \mathbb{X}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$



- $\mathbf{x}_{0,\dots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \theta, \mathbf{w})$ are the trajectories subject to $\mathbf{w}$ and $\mathbf{f}_{\theta}$
- $\mathbb{X}$ is the “safe” set where the state should be at all time

Safe policies via Robust (N)MPC

Robust (N)MPC delivers policy $\pi_{\theta}(\mathbf{x}_0) = \mathbf{u}_0^*$ from

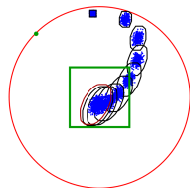
$$\mathbf{u}^* = \arg \min_{\mathbf{u}} \max_{\mathbf{w} \in \mathbb{W}_{\theta}^N} T_{\theta}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\text{s.t. } \mathbf{u}_{0,\dots,N} \in \mathbb{U}$$

$$\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}) \in \mathbb{X}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$

$$\mathbf{x}_N(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}) \in \mathbb{T}_{\theta}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$

- $\mathbf{x}_{0,\dots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w})$ are the trajectories subject to $\mathbf{w}$ and $\mathbf{f}_{\theta}$
- $\mathbb{X}$ is the “safe” set where the state should be at all time
- Terminal set $\mathbb{T}_{\theta}$ (required for recursive feasibility & stability)



Safe policies via Robust (N)MPC

Robust (N)MPC delivers policy $\pi_{\theta}(\mathbf{x}_0) = \mathbf{u}_0^*$ from

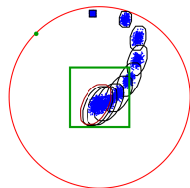
$$\mathbf{u}^* = \arg \min_{\mathbf{u}} \max_{\mathbf{w} \in \mathbb{W}_{\theta}^N} T_{\theta}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\text{s.t. } \mathbf{u}_{0,\dots,N} \in \mathbb{U}$$

$$\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \theta, \mathbf{w}) \in \mathbb{X}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$

$$\mathbf{x}_N(\mathbf{u}, \mathbf{s}, \theta, \mathbf{w}) \in \mathbb{T}_{\theta}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$

- $\mathbf{x}_{0,\dots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \theta, \mathbf{w})$ are the trajectories subject to $\mathbf{w}$ and $\mathbf{f}_{\theta}$
- $\mathbb{X}$ is the “safe” set where the state should be at all time
- Terminal set $\mathbb{T}_{\theta}$ (required for recursive feasibility & stability)
- If θ is such that $\mathbb{W}_{\theta}$ encloses *true* state dispersion, **MPC yields safe policy**



Safe policies via Robust (N)MPC

Robust (N)MPC delivers policy $\pi_{\theta}(\mathbf{x}_0) = \mathbf{u}_0^*$ from

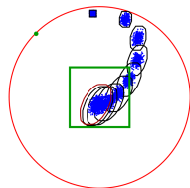
$$\mathbf{u}^* = \arg \min_{\mathbf{u}} \max_{\mathbf{w} \in \mathbb{W}_{\theta}^N} T_{\theta}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\text{s.t. } \mathbf{u}_{0,\dots,N} \in \mathbb{U}$$

$$\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}) \in \mathbb{X}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$

$$\mathbf{x}_N(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}) \in \mathbb{T}_{\theta}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$

- $\mathbf{x}_{0,\dots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w})$ are the trajectories subject to $\mathbf{w}$ and $\mathbf{f}_{\theta}$
- $\mathbb{X}$ is the “safe” set where the state should be at all time
- Terminal set $\mathbb{T}_{\theta}$ (required for recursive feasibility & stability)
- If $\boldsymbol{\theta}$ is such that $\mathbb{W}_{\theta}$ encloses *true* state dispersion, **MPC yields safe policy**



Closed-loop stability under some conditions on $\boldsymbol{\theta}$ (not trivial), need $\gamma = 1$

Safe policies via Robust (N)MPC

Robust (N)MPC delivers policy $\pi_{\theta}(\mathbf{x}_0) = \mathbf{u}_0^*$ from

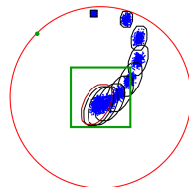
$$\mathbf{u}^* = \arg \min_{\mathbf{u}} \max_{\mathbf{w} \in \mathbb{W}_{\theta}^N} T_{\theta}(\mathbf{x}_N) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

$$\text{s.t. } \mathbf{u}_{0,\dots,N} \in \mathbb{U}$$

$$\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}) \in \mathbb{X}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$

$$\mathbf{x}_N(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}) \in \mathbb{T}_{\theta}, \quad \forall \mathbf{w} \in \mathbb{W}_{\theta}^{N-1}$$

- $\mathbf{x}_{0,\dots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $\mathbf{x}_{1,\dots,N-1}(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w})$ are the trajectories subject to $\mathbf{w}$ and $\mathbf{f}_{\theta}$
- $\mathbb{X}$ is the “safe” set where the state should be at all time
- Terminal set $\mathbb{T}_{\theta}$ (required for recursive feasibility & stability)
- If $\boldsymbol{\theta}$ is such that $\mathbb{W}_{\theta}$ encloses *true* state dispersion, **MPC yields safe policy**



$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_{\mathbf{a}} A_{\pi_{\theta}}]$$

Closed-loop stability under some conditions on $\boldsymbol{\theta}$ (not trivial), need $\gamma = 1$

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_a A_{\pi_{\theta}}]$$

adjusts θ for performance

Condition

$$\mathbf{s}_+ - \mathbf{f}(\mathbf{s}, \mathbf{a}, \theta) \in \mathbb{W}_{\theta}$$

enforces safety through θ

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_a A_{\pi_{\theta}}]$$

adjusts θ for performance

Condition

$$s_+ - \mathbf{f}(s, \mathbf{a}, \theta) \in \mathbb{W}_{\theta}$$

enforces safety through θ

- No clear connection to SYSID
- Sometimes does opposite of SYSID

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_a A_{\pi_{\theta}}]$$

adjusts θ for performance

- No clear connection to SYSID
- Sometimes does opposite of SYSID

Condition

$$\mathbf{s}_+ - \mathbf{f}(\mathbf{s}, \mathbf{a}, \theta) \in \mathbb{W}_{\theta}$$

enforces safety through θ

- Can be interpreted as a form of SYSID (set-membership)

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_a A_{\pi_{\theta}}]$$

adjusts θ for performance

- No clear connection to SYSID
- Sometimes does opposite of SYSID

Condition

$$\mathbf{s}_+ - \mathbf{f}(\mathbf{s}, \mathbf{a}, \theta) \in \mathbb{W}_{\theta}$$

enforces safety through θ

- Can be interpreted as a form of SYSID (set-membership)

How to do Safe RL?

Classic RL steps: $\theta \leftarrow \theta - \alpha \nabla_{\theta} J$

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_{\mathbf{a}} A \pi_{\theta}]$$

adjusts θ for performance

Condition

$$\mathbf{s}_+ - \mathbf{f}(\mathbf{s}, \mathbf{a}, \theta) \in \mathbb{W}_{\theta}$$

enforces safety through θ

- No clear connection to SYSID
- Sometimes does opposite of SYSID
- Can be interpreted as a form of SYSID (set-membership)

How to do Safe RL?

Classic RL steps: $\theta \leftarrow \theta - \alpha \nabla_{\theta} J$

Also reads as:

$$\theta \leftarrow \theta + \Delta \theta$$

$$\Delta \theta = \arg \min_{\Delta \theta} \frac{1}{2\alpha} \|\Delta \theta\|^2 + \nabla_{\theta} J^{\top} \Delta \theta$$

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_{\mathbf{a}} A \pi_{\theta}]$$

adjusts θ for performance

- No clear connection to SYSID
- Sometimes does opposite of SYSID

Condition

$$\mathbf{s}_+ - \mathbf{f}(\mathbf{s}, \mathbf{a}, \theta) \in \mathbb{W}_{\theta}$$

enforces safety through θ

- Can be interpreted as a form of SYSID (set-membership)

How to do Safe RL?

Classic RL steps: $\theta \leftarrow \theta - \alpha \nabla_{\theta} J$

Also reads as:

$$\theta \leftarrow \theta + \Delta\theta$$

$$\Delta\theta = \arg \min_{\Delta\theta} \frac{1}{2\alpha} \|\Delta\theta\|^2 + \nabla_{\theta} J^{\top} \Delta\theta$$

Safe RL steps $\theta \leftarrow \theta + \Delta\theta$:

$$\Delta\theta = \arg \min_{\Delta\theta} \frac{1}{2\alpha} \|\Delta\theta\|^2 + \nabla_{\theta} J^{\top} \Delta\theta$$

$$\begin{aligned} \text{s.t. } & \mathbf{s}_+ - \mathbf{f}(\mathbf{s}, \mathbf{a}, \theta + \Delta\theta) \in \mathbb{W}_{\theta + \Delta\theta} \\ & \forall (\mathbf{s}, \mathbf{a}, \mathbf{s}_+) \text{ in data set} \end{aligned}$$

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_{\mathbf{a}} A \pi_{\theta}]$$

adjusts θ for performance

- No clear connection to SYSID
- Sometimes does opposite of SYSID

Condition

$$\mathbf{s}_+ - \mathbf{f}(\mathbf{s}, \mathbf{a}, \theta) \in \mathbb{W}_{\theta}$$

enforces safety through θ

- Can be interpreted as a form of SYSID (set-membership)

How to do Safe RL?

Classic RL steps: $\theta \leftarrow \theta - \alpha \nabla_{\theta} J$

Also reads as:

$$\theta \leftarrow \theta + \Delta\theta$$

$$\Delta\theta = \arg \min_{\Delta\theta} \frac{1}{2\alpha} \|\Delta\theta\|^2 + \nabla_{\theta} J^{\top} \Delta\theta$$

Safe RL steps $\theta \leftarrow \theta + \Delta\theta$:

$$\Delta\theta = \arg \min_{\Delta\theta} \frac{1}{2\alpha} \|\Delta\theta\|^2 + \nabla_{\theta} J^{\top} \Delta\theta$$

$$\begin{aligned} \text{s.t. } & \mathbf{s}_+ - \mathbf{f}(\mathbf{s}, \mathbf{a}, \theta + \Delta\theta) \in \mathbb{W}_{\theta + \Delta\theta} \\ & \forall (\mathbf{s}, \mathbf{a}, \mathbf{s}_+) \text{ in data set} \end{aligned}$$

Safe RL steps seek performance under safety constraints

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

$$\nabla_{\theta} J = \mathbb{E} [\nabla_{\theta} \pi_{\theta} \nabla_a A \pi_{\theta}]$$

adjusts θ for performance

- No clear connection to SYSID
- Sometimes does opposite of SYSID

Condition

$$s_+ - f(s, a, \theta) \in \mathbb{W}_{\theta}$$

enforces safety through θ

- Can be interpreted as a form of SYSID (set-membership)

How to do Safe RL?

Classic RL steps: $\theta \leftarrow \theta - \alpha \nabla_{\theta} J$

Also reads as:

$$\theta \leftarrow \theta + \Delta\theta$$

$$\Delta\theta = \arg \min_{\Delta\theta} \frac{1}{2\alpha} \|\Delta\theta\|^2 + \nabla_{\theta} J^{\top} \Delta\theta$$

Safe RL steps $\theta \leftarrow \theta + \Delta\theta$:

$$\Delta\theta = \arg \min_{\Delta\theta} \frac{1}{2\alpha} \|\Delta\theta\|^2 + \nabla_{\theta} J^{\top} \Delta\theta$$

$$\begin{aligned} \text{s.t. } & s_+ - f(s, a, \theta + \Delta\theta) \in \mathbb{W}_{\theta + \Delta\theta} \\ & \forall (s, a, s_+) \text{ in data set} \end{aligned}$$

Safe RL steps seek performance under safety constraints

Difficulty: differentiating through the effect of $\mathbb{W}_{\theta}$, big data in Safe RL steps

Illustrative example

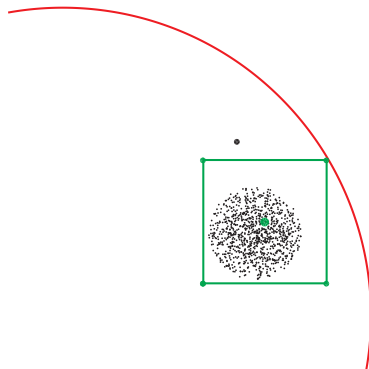
- Real system

$$\mathbf{x}_{k+1} = A_{\text{real}}\mathbf{x}_k + B_{\text{real}}\mathbf{u}_k + \mathbf{n}$$

- Noise $\mathbf{n}$ in a ball
- Robust MPC model

$$\mathbf{x}_{k+1} = A_0\mathbf{x}_k + B_0\mathbf{u}_k \oplus \mathbb{W}$$

- $\mathbb{W}$ is a square
- Quadratic stage cost
- Constraint $\|\mathbf{x}\|^2 \leq 1$



Illustrative example

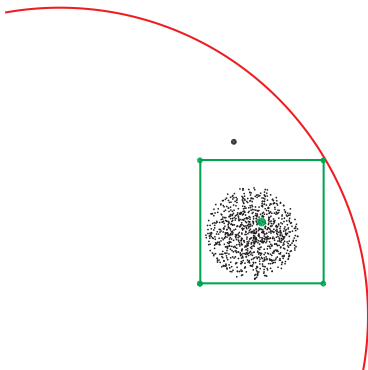
- Real system

$$\mathbf{x}_{k+1} = A_{\text{real}}\mathbf{x}_k + B_{\text{real}}\mathbf{u}_k + \mathbf{n}$$

- Noise $\mathbf{n}$ in a ball
- Robust MPC model

$$\mathbf{x}_{k+1} = A_0\mathbf{x}_k + B_0\mathbf{u}_k \oplus \mathbb{W}$$

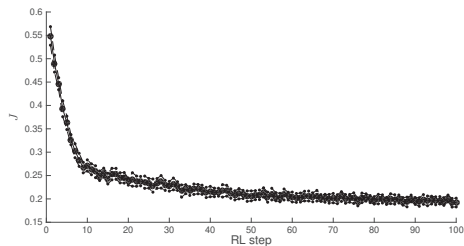
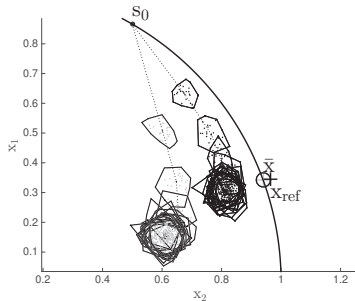
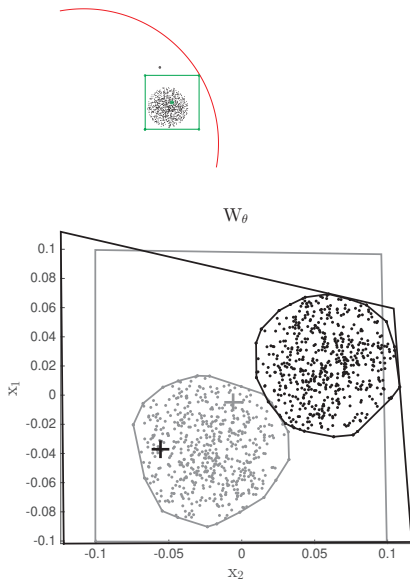
- $\mathbb{W}$ is a square
- Quadratic stage cost
- Constraint $\|\mathbf{x}\|^2 \leq 1$



Let's adjust:

- Set $\mathbb{W}$, while containing process noise
- State and input reference in MPC cost function
- Internal linear feedback (robust MPC internal control)

Illustrative example

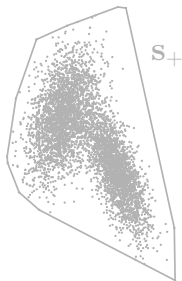


Learning

Scenario Tree MPC

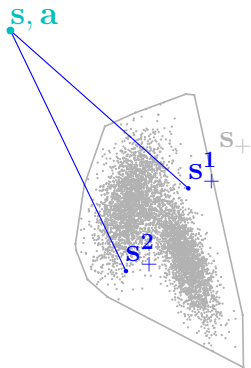
Key idea: approximate distribution $\mathbb{P}[s_+|s, a]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions.
Optimize over the tree.

s, a



Scenario Tree MPC

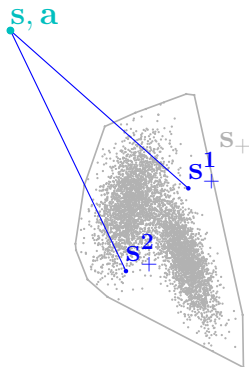
Key idea: approximate distribution $\mathbb{P}[s_+|s, a]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.



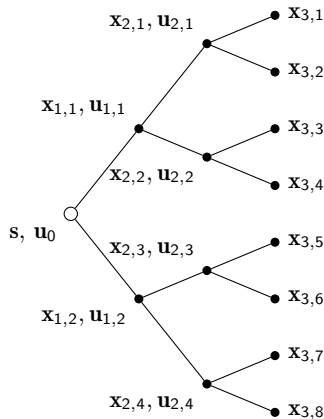
E.g. pick two **scenarios**

Scenario Tree MPC

Key idea: approximate distribution $\mathbb{P}[s_+|s, a]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.



E.g. pick two **scenarios**



Scenario Tree MPC

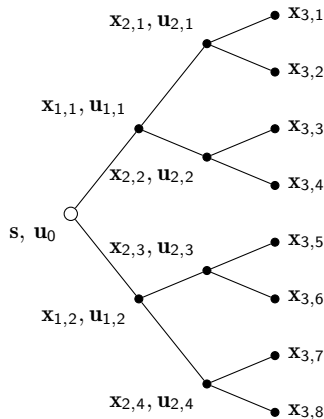
Key idea: approximate distribution $\mathbb{P}[s_+|s, a]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.

Scenario-Tree MPC gives $\pi^{\text{MPC}}(s) = \mathbf{u}_0^*$ from

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{u}} \quad & \sum_{i=1}^n \sum_{k=0}^N \omega_i L(\mathbf{x}_k^i, \mathbf{u}_k^i) \\ \text{s.t.} \quad & \mathbf{x}_{k+1}^i = \mathbf{f}_k^i(\mathbf{x}_k^i, \mathbf{u}_k^i), \quad \mathbf{x}_0^i = \mathbf{s} \\ & \mathbf{h}_i(\mathbf{x}_k^i, \mathbf{u}_k^i) \leq 0 \\ & \mathbf{c}_k(\mathbf{u}_k^{1, \dots, n}) = 0 \end{aligned}$$

where we need to select

- $\mathbf{f}_k^i$ to form the scenarios
- $\omega_{1, \dots, n}$ to build $\mathbb{E}[\cdot]$



Scenario Tree MPC

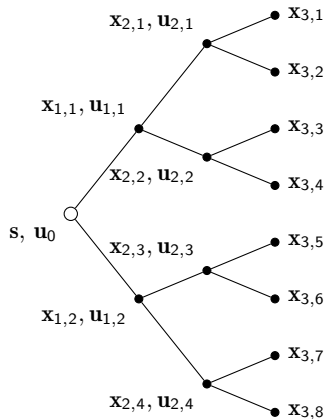
Key idea: approximate distribution $\mathbb{P}[s_+|s, a]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.

Scenario-Tree MPC gives $\pi^{\text{MPC}}(s) = \mathbf{u}_0^*$ from

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{u}} \quad & \sum_{i=1}^n \sum_{k=0}^N \omega_i L(\mathbf{x}_k^i, \mathbf{u}_k^i) \\ \text{s.t.} \quad & \mathbf{x}_{k+1}^i = \mathbf{f}_k^i(\mathbf{x}_k^i, \mathbf{u}_k^i), \quad \mathbf{x}_0^i = \mathbf{s} \\ & \mathbf{h}_i(\mathbf{x}_k^i, \mathbf{u}_k^i) \leq 0 \\ & \mathbf{c}_k(\mathbf{u}_k^{1, \dots, n}) = 0 \end{aligned}$$

where we need to select

- $\mathbf{f}_k^i$ to form the scenarios
- $\omega_{1, \dots, n}$ to build $\mathbb{E}[\cdot]$

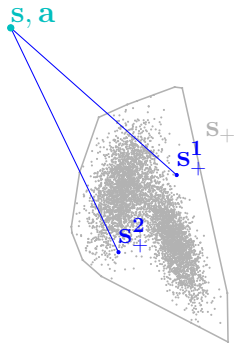


Difficulties

- Pick “good” scenarios
- # scenarios explode with horizon (e.g. $n = 2^N$ here)

Often shallow trees are used

Scenario-Tree MPC



$$\min_{\mathbf{x}, \mathbf{u}} \quad \sum_{i=1}^n \sum_{k=0}^N \omega_i L(\mathbf{x}_k^i, \mathbf{u}_k^i)$$

$$\text{s.t.} \quad \mathbf{x}_{k+1}^i = \mathbf{f}_k^i(\mathbf{x}_k^i, \mathbf{u}_k^i), \quad \mathbf{x}_0^i = \mathbf{s}$$

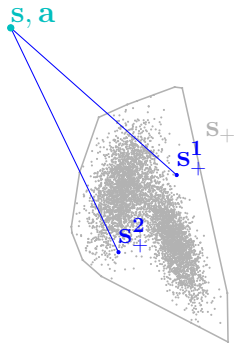
$$\mathbf{h}_i(\mathbf{x}_k^i, \mathbf{u}_k^i) \leq 0$$

$$\mathbf{c}_k(\mathbf{u}_k^{1, \dots, n}) = 0$$

Difficulties

- Pick “good” scenarios
- Exploding complexity

Scenario-Tree MPC



$$\begin{aligned} \min_{\mathbf{x}, \mathbf{u}} \quad & \sum_{i=1}^n \sum_{k=0}^N \omega_i L(\mathbf{x}_k^i, \mathbf{u}_k^i) \\ \text{s.t.} \quad & \mathbf{x}_{k+1}^i = \mathbf{f}_k^i(\mathbf{x}_k^i, \mathbf{u}_k^i), \quad \mathbf{x}_0^i = \mathbf{s} \\ & \mathbf{h}_i(\mathbf{x}_k^i, \mathbf{u}_k^i) \leq 0 \\ & \mathbf{c}_k(\mathbf{u}_k^{1, \dots, n}) = 0 \end{aligned}$$

RL can do:

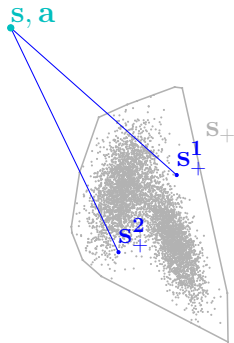
- Optimize scenarios $\mathbf{f}_k^i$
- Optimize weights ω_i
- Can work on shallow trees

Difficulties

- Pick “good” scenarios
- Exploding complexity

for real-world performance of policy

Scenario-Tree MPC



$$\begin{aligned} \min_{\mathbf{x}, \mathbf{u}} \quad & \sum_{i=1}^n \sum_{k=0}^N \omega_i L(\mathbf{x}_k^i, \mathbf{u}_k^i) \\ \text{s.t.} \quad & \mathbf{x}_{k+1}^i = \mathbf{f}_k^i(\mathbf{x}_k^i, \mathbf{u}_k^i), \quad \mathbf{x}_0^i = \mathbf{s} \\ & \mathbf{h}_i(\mathbf{x}_k^i, \mathbf{u}_k^i) \leq 0 \\ & \mathbf{c}_k(\mathbf{u}_k^{1, \dots, n}) = 0 \end{aligned}$$

RL can do:

- Optimize scenarios $\mathbf{f}_k^i$
- Optimize weights ω_i
- Can work on shallow trees

Difficulties

- Pick “good” scenarios
- Exploding complexity

for real-world performance of policy

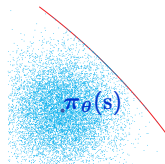
Even a depth-1 tree provides a finer representation than a deterministic model. Theory applies.

Outline

- 1 SYSID & MPC
- 2 Safe RL via MPC
- 3 Exploration with MPC
- 4 Stability-constrained Learning with MPC
- 5 Explored questions

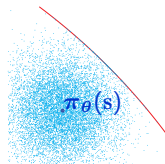
“Safe” (feasible) exploration with MPC

Learning requires exploration. E.g. apply
 $\mathbf{a} = \pi_{\theta}(\mathbf{s}) + \mathbf{d}$ **to the real system where $\mathbf{d}$ is a**
“disturbance”



“Safe” (feasible) exploration with MPC

Learning requires exploration. E.g. apply $\mathbf{a} = \pi_{\theta}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a “disturbance”

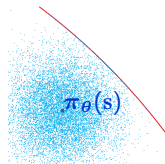


Explore while keeping feasibility?

- Clearly an arbitrary “policy disturbance” $\pi_{\theta}(\mathbf{s}) + \mathbf{d}$ is a poor idea...
- NLP-based policy: “disturb” the cost function instead! (different options)

“Safe” (feasible) exploration with MPC

Learning requires exploration. E.g. apply
 $\mathbf{a} = \pi_{\theta}(\mathbf{s}) + \mathbf{d}$ **to the real system where $\mathbf{d}$ is a**
“disturbance”



Explore while keeping feasibility?

- Clearly an arbitrary “policy disturbance” $\pi_{\theta}(\mathbf{s}) + \mathbf{d}$ is a poor idea...
- NLP-based policy: “disturb” the cost function instead! (different options)

Feasible exploration: $\pi_{\theta}^{\text{e}}(\mathbf{s}) = \mathbf{u}_0^*$:

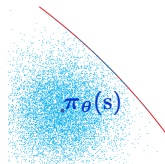
$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta}(\mathbf{x}_N) - \mathbf{d}^{\top} \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

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

satisfies the constraints by construction

“Safe” (feasible) exploration with MPC

Learning requires exploration. E.g. apply
 $\mathbf{a} = \pi_{\theta}(\mathbf{s}) + \mathbf{d}$ **to the real system where $\mathbf{d}$ is a**
“disturbance”



Explore while keeping feasibility?

- Clearly an arbitrary “policy disturbance” $\pi_{\theta}(\mathbf{s}) + \mathbf{d}$ is a poor idea...
- NLP-based policy: **“disturb” the cost function instead!** (different options)

Feasible exploration: $\pi_{\theta}^{\text{e}}(\mathbf{s}) = \mathbf{u}_0^*$:

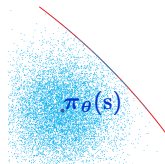
$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta}(\mathbf{x}_N) - \mathbf{d}^{\top} \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

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

satisfies the constraints by construction

“Safe” (feasible) exploration with MPC

Learning requires exploration. E.g. apply $\mathbf{a} = \pi_{\theta}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a “disturbance”



Explore while keeping feasibility?

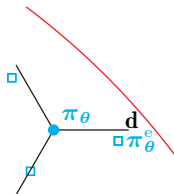
- Clearly an arbitrary “policy disturbance” $\pi_{\theta}(\mathbf{s}) + \mathbf{d}$ is a poor idea...
- NLP-based policy: “disturb” the cost function instead! (different options)

Feasible exploration: $\pi_{\theta}^e(\mathbf{s}) = \mathbf{u}_0^*$:

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta}(\mathbf{x}_N) - \mathbf{d}^{\top} \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

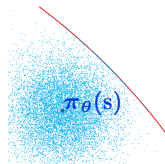
$$\text{s.t.} \quad \mathbf{x}_{k+1} = \mathbf{f}_{\theta}(\mathbf{x}_k, \mathbf{u}_k) \\ \mathbf{h}_{\theta}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}$$

satisfies the constraints by construction



“Safe” (feasible) exploration with MPC

Learning requires exploration. E.g. apply
 $\mathbf{a} = \pi_{\theta}(\mathbf{s}) + \mathbf{d}$ **to the real system where $\mathbf{d}$ is a**
“disturbance”



Explore while keeping feasibility?

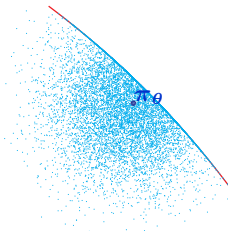
- Clearly an arbitrary “policy disturbance” $\pi_{\theta}(\mathbf{s}) + \mathbf{d}$ is a poor idea...
- NLP-based policy: **“disturb” the cost function instead!** (different options)

Feasible exploration: $\pi_{\theta}^e(\mathbf{s}) = \mathbf{u}_0^*$:

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta}(\mathbf{x}_N) - \mathbf{d}^{\top} \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

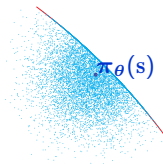
$$\text{s.t.} \quad \mathbf{x}_{k+1} = \mathbf{f}_{\theta}(\mathbf{x}_k, \mathbf{u}_k) \\ \mathbf{h}_{\theta}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}$$

satisfies the constraints by construction



“Safe” (feasible) exploration with MPC

Learning requires exploration. E.g. apply
 $\mathbf{a} = \pi_{\theta}(\mathbf{s}) + \mathbf{d}$ **to the real system where $\mathbf{d}$ is a**
“disturbance”



Explore while keeping feasibility?

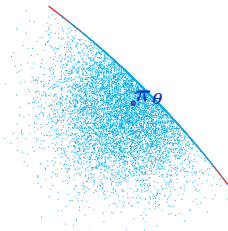
- Clearly an arbitrary “policy disturbance” $\pi_{\theta}(\mathbf{s}) + \mathbf{d}$ is a poor idea...
- NLP-based policy: **“disturb” the cost function instead!** (different options)

Feasible exploration: $\pi_{\theta}^{\circ}(\mathbf{s}) = \mathbf{u}_0^*$:

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta}(\mathbf{x}_N) - \mathbf{d}^{\top} \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

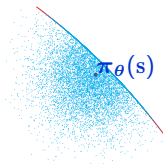
$$\text{s.t.} \quad \mathbf{x}_{k+1} = \mathbf{f}_{\theta}(\mathbf{x}_k, \mathbf{u}_k) \\ \mathbf{h}_{\theta}(\mathbf{x}_k, \mathbf{u}_k) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}$$

satisfies the constraints by construction



“Safe” (feasible) exploration with MPC

Learning requires exploration. E.g. apply
 $\mathbf{a} = \pi_{\theta}(\mathbf{s}) + \mathbf{d}$ **to the real system where $\mathbf{d}$ is a**
“disturbance”



Explore while keeping feasibility?

- Clearly an arbitrary “policy disturbance” $\pi_{\theta}(\mathbf{s}) + \mathbf{d}$ is a poor idea...
- NLP-based policy: **“disturb” the cost function instead!** (different options)

Feasible exploration: $\pi_{\theta}^e(\mathbf{s}) = \mathbf{u}_0^*$:

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta}(\mathbf{x}_N) - \mathbf{d}^{\top} \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k)$$

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

satisfies the constraints by construction

Remarks:

- Exploration $\mathbf{e} = \pi_{\theta}^e - \pi_{\theta}$ is not centred-isotropic
- Can create some technical issues with actor-critic methods using linear compatible $A^{\pi_{\theta}}$, yields **biased policy gradient estimation**
- Bias seems small in practice + linear compatible $A^{\pi_{\theta}}$ does not seem to be much used in RL

MPC as a Stochastic Policy

- **Stochastic policy:** $\mathbf{a} \sim \pi_{\theta} [\cdot | \mathbf{s}]$ gives actions as samples of a distribution conditioned on $\mathbf{s}$

MPC as a Stochastic Policy

- **Stochastic policy:** $\mathbf{a} \sim \pi_{\theta} [\cdot | \mathbf{s}]$ gives actions as samples of a distribution conditioned on $\mathbf{s}$
- **Motivation:** modern RL methods for continuous action spaces use stochastic policy gradient approaches (e.g. SAC, PPO)

MPC as a Stochastic Policy

- **Stochastic policy:** $\mathbf{a} \sim \pi_{\theta} [\cdot | \mathbf{s}]$ gives actions as samples of a distribution conditioned on $\mathbf{s}$
- **Motivation:** modern RL methods for continuous action spaces use stochastic policy gradient approaches (e.g. SAC, PPO)
- **Policy gradient** (several variants):

$$\nabla_{\theta} J(\pi_{\theta} [\cdot | \mathbf{s}]) = \mathbb{E} [\nabla_{\theta} \log \pi_{\theta} [\mathbf{a} | \mathbf{s}] Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a})]$$

MPC as a Stochastic Policy

- **Stochastic policy:** $\mathbf{a} \sim \pi_{\theta} [\cdot | \mathbf{s}]$ gives actions as samples of a distribution conditioned on $\mathbf{s}$
- **Motivation:** modern RL methods for continuous action spaces use stochastic policy gradient approaches (e.g. SAC, PPO)
- **Policy gradient** (several variants):

$$\nabla_{\theta} J(\pi_{\theta} [\cdot | \mathbf{s}]) = \mathbb{E} [\nabla_{\theta} \log \pi_{\theta} [\mathbf{a} | \mathbf{s}] Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a})]$$

Requires:

- ▶ Sampling from $\pi_{\theta} [\cdot | \mathbf{s}]$ efficiently $\rightarrow$ as easy as solving MPC
- ▶ Evaluating

$$\nabla_{\theta} \log \pi_{\theta} [\mathbf{a} | \mathbf{s}] = \frac{\nabla_{\theta} \pi_{\theta} [\mathbf{a} | \mathbf{s}]}{\pi_{\theta} [\mathbf{a} | \mathbf{s}]}$$

i.e. $\nabla_{\theta} \pi_{\theta} [\mathbf{a} | \mathbf{s}]$ and $\pi_{\theta} [\mathbf{a} | \mathbf{s}]$ efficiently $\rightarrow$ is that easy?

MPC as a Stochastic Policy

- **Stochastic policy:** $\mathbf{a} \sim \pi_{\theta} [\cdot | \mathbf{s}]$ gives actions as samples of a distribution conditioned on $\mathbf{s}$
- **Motivation:** modern RL methods for continuous action spaces use stochastic policy gradient approaches (e.g. SAC, PPO)
- **Policy gradient** (several variants):

$$\nabla_{\theta} J(\pi_{\theta} [\cdot | \mathbf{s}]) = \mathbb{E} [\nabla_{\theta} \log \pi_{\theta} [\mathbf{a} | \mathbf{s}] Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a})]$$

Requires:

- ▶ Sampling from $\pi_{\theta} [\cdot | \mathbf{s}]$ efficiently $\rightarrow$ as easy as solving MPC
- ▶ Evaluating

$$\nabla_{\theta} \log \pi_{\theta} [\mathbf{a} | \mathbf{s}] = \frac{\nabla_{\theta} \pi_{\theta} [\mathbf{a} | \mathbf{s}]}{\pi_{\theta} [\mathbf{a} | \mathbf{s}]}$$

i.e. $\nabla_{\theta} \pi_{\theta} [\mathbf{a} | \mathbf{s}]$ and $\pi_{\theta} [\mathbf{a} | \mathbf{s}]$ efficiently $\rightarrow$ is that easy?

Answer:

- Yes if we, e.g., use $\pi_{\theta} [\cdot | \mathbf{s}] \equiv \underbrace{\pi_{\theta}(\mathbf{s})}_{\text{MPC}} + \mathbf{w}$ (Random perturbation added to MPC policy)
- But then we can violate the constraints and bounds.
- What about disturbing the MPC cost, ref. exploration slide? **Then no...**

MPC as a Stochastic Policy

Stochastic policy $\pi_{\theta} [\cdot | s] \equiv \mathbf{u}_0^* (\theta, s, \mathbf{d})$:

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta} (\mathbf{x}_N, \mathbf{d}) + \sum_{k=0}^{N-1} L_{\theta} (\mathbf{x}_k, \mathbf{u}_k, \mathbf{d})$$

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

$$\mathbf{h}_{\theta} (\mathbf{x}_k, \mathbf{u}_k, \mathbf{d}) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}$$

where $\mathbf{d} \sim \rho[\cdot]$

MPC as a Stochastic Policy

Stochastic policy $\pi_{\theta} [\cdot | s] \equiv \mathbf{u}_0^* (\theta, s, \mathbf{d})$:

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta} (\mathbf{x}_N, \mathbf{d}) + \sum_{k=0}^{N-1} L_{\theta} (\mathbf{x}_k, \mathbf{u}_k, \mathbf{d})$$

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

$$\mathbf{h}_{\theta} (\mathbf{x}_k, \mathbf{u}_k, \mathbf{d}) \leq 0, \quad \mathbf{x}_0 = s$$

where $\mathbf{d} \sim \rho[\cdot]$

Remarks

- Chose ρ easy to sample from
- Typ. bounded support
- Special case: $\theta = \bar{\theta} + \mathbf{d}$
- To ensure feasibility, leave $\mathbf{f}_{\theta}, \mathbf{h}_{\theta}$ “alone”

MPC as a Stochastic Policy

Stochastic policy $\pi_{\theta} [\cdot | s] \equiv \mathbf{u}_0^* (\theta, s, \mathbf{d})$:

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta} (\mathbf{x}_N, \mathbf{d}) + \sum_{k=0}^{N-1} L_{\theta} (\mathbf{x}_k, \mathbf{u}_k, \mathbf{d})$$

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

$$\mathbf{h}_{\theta} (\mathbf{x}_k, \mathbf{u}_k, \mathbf{d}) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}$$

where $\mathbf{d} \sim \rho[\cdot]$

Remarks

- Chose ρ easy to sample from
- Typ. bounded support
- Special case: $\theta = \bar{\theta} + \mathbf{d}$
- To ensure feasibility, leave $\mathbf{f}_{\theta}, \mathbf{h}_{\theta}$ “alone”

Performance

$$J(\pi_{\theta}) = \mathbb{E}_{\pi_{\theta}} \left[\sum_{k=0}^{\infty} \gamma^k L(s_k, \mathbf{a}_k) \mid \mathbf{a}_k \sim \pi_{\theta} [\cdot | s_k] \right] = \mathbb{E}_{\rho} \left[\sum_{k=0}^{\infty} \gamma^k L(s_k, \mathbf{a}_k) \mid \begin{array}{l} \mathbf{a}_k = \mathbf{u}_0^* (\theta, s_k, \mathbf{d}) \\ \mathbf{d} \sim \rho[\cdot] \end{array} \right]$$

MPC as a Stochastic Policy

Stochastic policy $\pi_{\theta} [\cdot | s] \equiv \mathbf{u}_0^* (\theta, s, \mathbf{d})$:

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta} (\mathbf{x}_N, \mathbf{d}) + \sum_{k=0}^{N-1} L_{\theta} (\mathbf{x}_k, \mathbf{u}_k, \mathbf{d})$$

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

$$\mathbf{h}_{\theta} (\mathbf{x}_k, \mathbf{u}_k, \mathbf{d}) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}$$

where $\mathbf{d} \sim \rho[\cdot]$

Remarks

- Chose ρ easy to sample from
- Typ. bounded support
- Special case: $\theta = \bar{\theta} + \mathbf{d}$
- To ensure feasibility, leave $\mathbf{f}_{\theta}, \mathbf{h}_{\theta}$ “alone”

Performance

$$J(\pi_{\theta}) = \mathbb{E}_{\pi_{\theta}} \left[\sum_{k=0}^{\infty} \gamma^k L(s_k, \mathbf{a}_k) \mid \mathbf{a}_k \sim \pi_{\theta} [\cdot | s_k] \right] = \mathbb{E}_{\rho} \left[\sum_{k=0}^{\infty} \gamma^k L(s_k, \mathbf{a}_k) \mid \begin{array}{l} \mathbf{a}_k = \mathbf{u}_0^* (\theta, s_k, \mathbf{d}) \\ \mathbf{d} \sim \rho[\cdot] \end{array} \right]$$

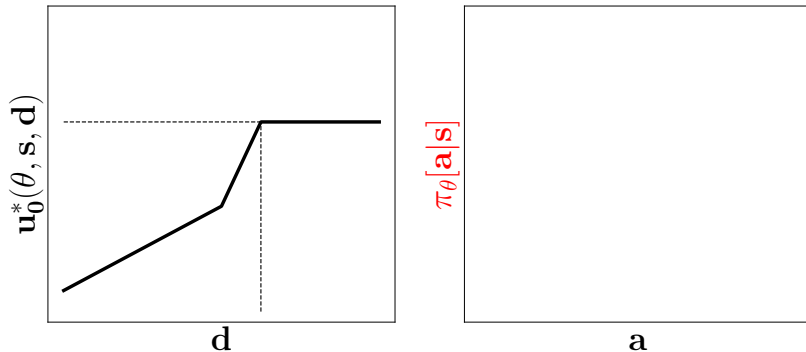
Stochastic Policy Gradient

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} [\nabla_{\theta} \log \pi_{\theta} \cdot Q^{\pi_{\theta}}]$$

... but what about $\nabla_{\theta} \log \pi_{\theta}$?

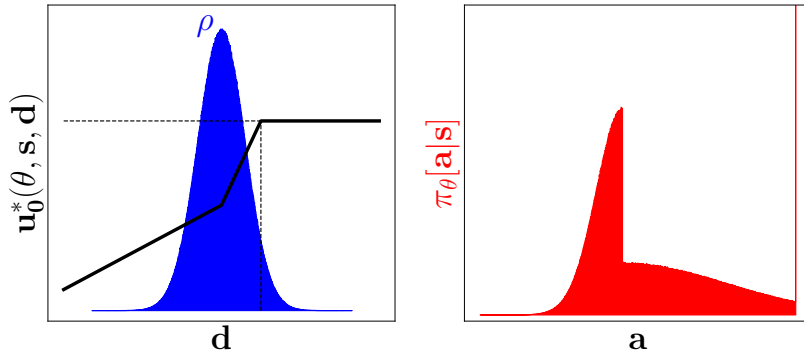
MPC as a Stochastic Policy & Policy Gradient

Illustration – Linear MPC with scalar a , d normal centered



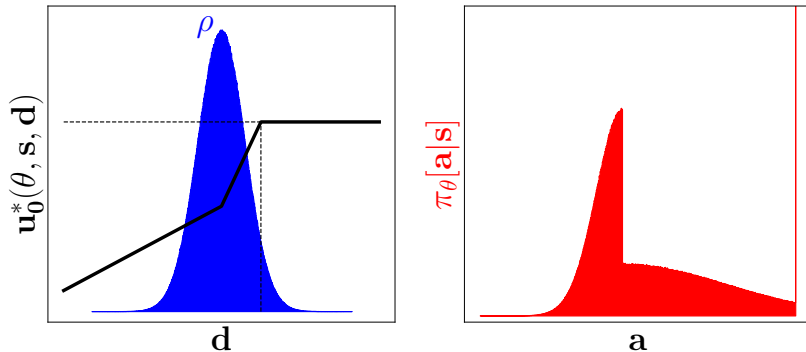
MPC as a Stochastic Policy & Policy Gradient

Illustration – Linear MPC with scalar $\mathbf{a}$, $\mathbf{d}$ normal centered



MPC as a Stochastic Policy & Policy Gradient

Illustration – Linear MPC with scalar $\mathbf{a}$, $\mathbf{d}$ normal centered

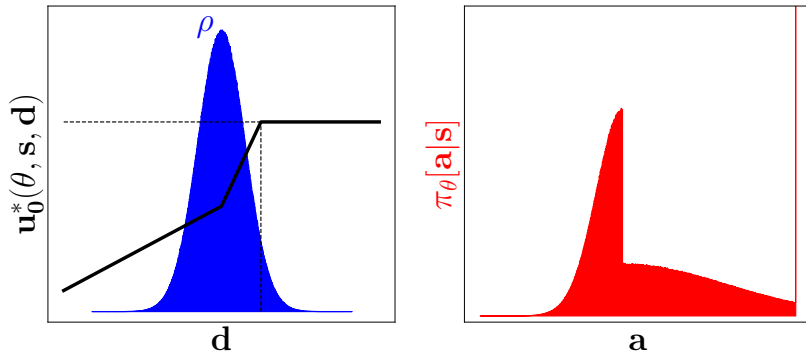


For $\mathbf{d} \rightarrow \mathbf{u}_0^*$ bijective & differentiable

$$\pi_\theta[\mathbf{a}|\mathbf{s}] = |\det(\nabla_{\mathbf{d}} \mathbf{u}_0^*(\theta, \mathbf{s}, \mathbf{d}))|^{-1} \rho[\mathbf{d}]$$

MPC as a Stochastic Policy & Policy Gradient

Illustration – Linear MPC with scalar $\mathbf{a}$, $\mathbf{d}$ normal centered



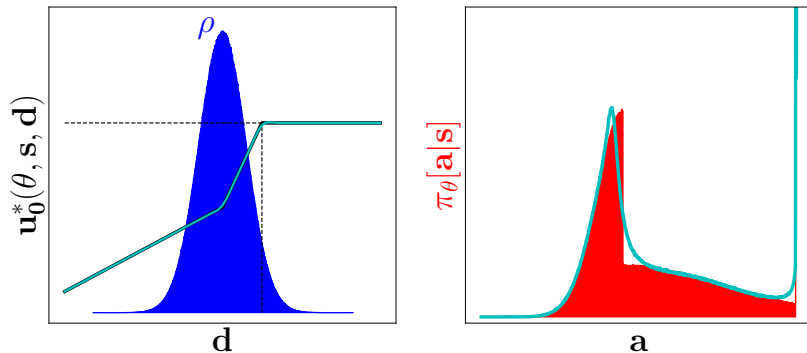
For $\mathbf{d} \rightarrow \mathbf{u}_0^*$ bijective & differentiable

$$\pi_\theta[\mathbf{a}|\mathbf{s}] = |\det(\nabla_{\mathbf{d}} \mathbf{u}_0^*(\theta, \mathbf{s}, \mathbf{d}))|^{-1} \rho[\mathbf{d}]$$

- Even for the “simplest” form of MPC, there are challenges

MPC as a Stochastic Policy & Policy Gradient

Illustration – Linear MPC with scalar $\mathbf{a}$, $\mathbf{d}$ normal centered



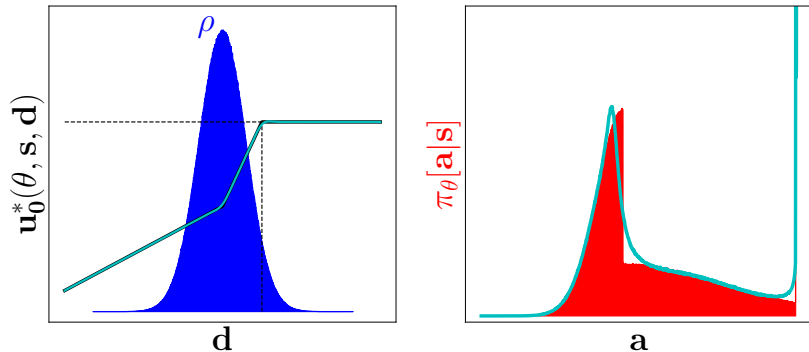
For $\mathbf{d} \rightarrow \mathbf{u}_0^*$ bijective & differentiable

$$\pi_\theta[\mathbf{a}|\mathbf{s}] = |\det(\nabla_{\mathbf{d}} \mathbf{u}_0^*(\theta, \mathbf{s}, \mathbf{d}))|^{-1} \rho[\mathbf{d}]$$

- Even for the “simplest” form of MPC, there are challenges
- Smoothing the MPC, e.g. leveraging on IP methods, solves (many of) them

MPC as a Stochastic Policy & Policy Gradient

Illustration – Linear MPC with scalar $\mathbf{a}$, $\mathbf{d}$ normal centered



For $\mathbf{d} \rightarrow \mathbf{u}_0^*$ bijective & differentiable

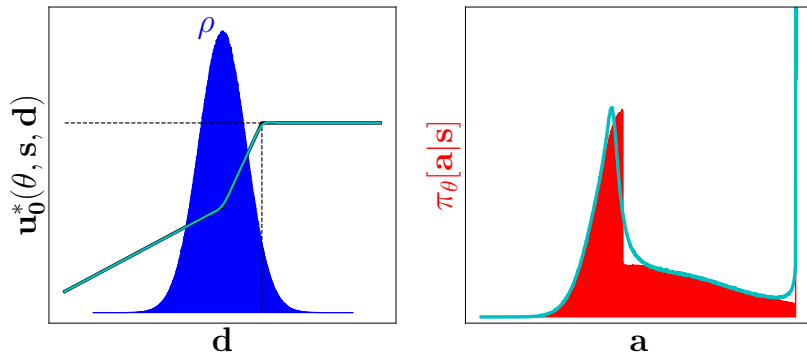
$$\pi_{\theta}[\mathbf{a}|\mathbf{s}] = |\det(\nabla_{\mathbf{d}} \mathbf{u}_0^*(\theta, \mathbf{s}, \mathbf{d}))|^{-1} \rho[\mathbf{d}]$$

- Even for the “simplest” form of MPC, there are challenges
- Smoothing the MPC, e.g. leveraging on IP methods, solves (many of) them

What about $\nabla_{\theta} \log \pi_{\theta}[\mathbf{a}|\mathbf{s}]$? Ok to compute, but needs 2nd-order sensitivities of $\mathbf{u}_0^*$

MPC as a Stochastic Policy & Policy Gradient

Illustration – Linear MPC with scalar $\mathbf{a}$, $\mathbf{d}$ normal centered

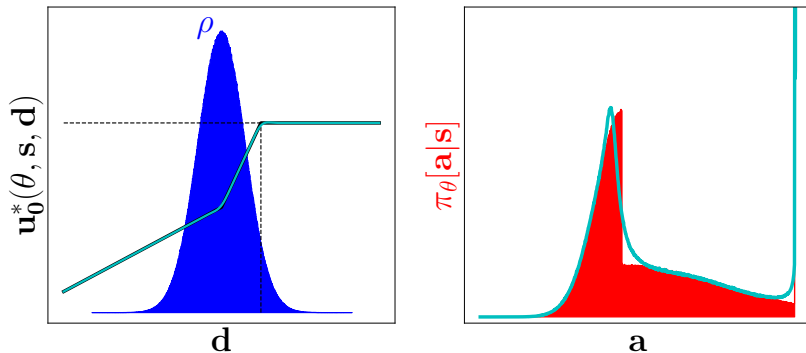


Reparametrization trick gives (smoothing is still *theoretically* good...)

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} [\nabla_{\theta} \log \pi_{\theta} \cdot Q^{\pi_{\theta}}] = \mathbb{E} [\nabla_{\theta} \mathbf{u}_0^*(\theta, \mathbf{s}, \mathbf{d}) \cdot \nabla_{\mathbf{a}} Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a})]$$

MPC as a Stochastic Policy & Policy Gradient

Illustration – Linear MPC with scalar $\mathbf{a}$, $\mathbf{d}$ normal centered



Reparametrization trick gives (smoothing is still *theoretically* good...)

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} [\nabla_{\theta} \log \pi_{\theta} \cdot Q^{\pi_{\theta}}] = \mathbb{E} [\nabla_{\theta} \mathbf{u}_0^*(\theta, \mathbf{s}, \mathbf{d}) \cdot \nabla_{\mathbf{a}} Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a})]$$

Open questions on estimation variance w.r.t. sharp variations in $Q^{\pi_{\theta}}$ and π_{θ}

Outline

- 1 SYSID & MPC
- 2 Safe RL via MPC
- 3 Exploration with MPC
- 4 Stability-constrained Learning with MPC**
- 5 Explored questions

Stability of MPC - Introduction

Policy π_{MPC} from

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

Stability of MPC - Introduction

Policy π_{MPC} from

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

If for some K_∞ function κ (“bowl-shaped”):

$$L(\mathbf{s}, \mathbf{a}) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|), \quad \forall \mathbf{s}, \mathbf{a}$$

holds, then MPC scheme is stabilizing (+conditions on T)

Stability of MPC - Introduction

Policy π_{MPC} from

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

For L not lower bounded by K_∞ , we need λ such that

$$\ell(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \lambda(\mathbf{s}) - \lambda(\mathbf{f}(\mathbf{s}, \mathbf{a})) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|), \quad \forall \mathbf{s}, \mathbf{a}$$

then MPC scheme is stabilizing (+conditions on T)

Stability of MPC - Introduction

Policy π_{MPC} from

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

Equivalent MPC

$$\begin{aligned} \min_{\mathbf{s}, \mathbf{a}} \quad & -\lambda(\mathbf{s}_0) + \tilde{T}(\mathbf{s}_N) + \sum_{k=0}^{N-1} \ell(\mathbf{s}_k, \mathbf{a}_k) \\ \text{s.t.} \quad & \mathbf{s}_{k+1} = \mathbf{f}(\mathbf{s}_k, \mathbf{a}_k) \\ & \mathbf{h}(\mathbf{s}_k, \mathbf{a}_k) \leq 0, \quad \mathbf{s}_N \in \mathbb{T} \end{aligned}$$

where $\ell(\mathbf{s}, \mathbf{a}) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|)$, $\forall \mathbf{s}, \mathbf{a}$

For L not lower bounded by K_∞ , we need λ such that

$$\ell(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \lambda(\mathbf{s}) - \lambda(\mathbf{f}(\mathbf{s}, \mathbf{a})) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|), \quad \forall \mathbf{s}, \mathbf{a}$$

then MPC scheme is stabilizing (+conditions on T)

Stability of MPC - Introduction

Policy π_{MPC} from

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

Equivalent MPC

$$\begin{aligned} \min_{\mathbf{s}, \mathbf{a}} \quad & -\lambda(\mathbf{s}_0) + \tilde{T}(\mathbf{s}_N) + \sum_{k=0}^{N-1} \ell(\mathbf{s}_k, \mathbf{a}_k) \\ \text{s.t.} \quad & \mathbf{s}_{k+1} = \mathbf{f}(\mathbf{s}_k, \mathbf{a}_k) \\ & \mathbf{h}(\mathbf{s}_k, \mathbf{a}_k) \leq 0, \quad \mathbf{s}_N \in \mathbb{T} \end{aligned}$$

where $\ell(\mathbf{s}, \mathbf{a}) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|)$, $\forall \mathbf{s}, \mathbf{a}$

For L not lower bounded by K_∞ , we need λ such that

$$\ell(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \lambda(\mathbf{s}) - \lambda(\mathbf{f}(\mathbf{s}, \mathbf{a})) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|), \quad \forall \mathbf{s}, \mathbf{a}$$

then MPC scheme is stabilizing (+conditions on T)

Remarks:

- No discount $\gamma = 1$
- Exact model, deterministic

Stability of MPC - Introduction

Policy π_{MPC} from

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

Equivalent MPC

$$\begin{aligned} \min_{\mathbf{s}, \mathbf{a}} \quad & -\lambda(\mathbf{s}_0) + \tilde{T}(\mathbf{s}_N) + \sum_{k=0}^{N-1} \ell(\mathbf{s}_k, \mathbf{a}_k) \\ \text{s.t.} \quad & \mathbf{s}_{k+1} = \mathbf{f}(\mathbf{s}_k, \mathbf{a}_k) \\ & \mathbf{h}(\mathbf{s}_k, \mathbf{a}_k) \leq 0, \quad \mathbf{s}_N \in \mathbb{T} \end{aligned}$$

where $\ell(\mathbf{s}, \mathbf{a}) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|)$, $\forall \mathbf{s}, \mathbf{a}$

For L not lower bounded by K_∞ , we need λ such that

$$\ell(\mathbf{s}, \mathbf{a}) = L(\mathbf{s}, \mathbf{a}) + \lambda(\mathbf{s}) - \lambda(\mathbf{f}(\mathbf{s}, \mathbf{a})) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|), \quad \forall \mathbf{s}, \mathbf{a}$$

then MPC scheme is stabilizing (+conditions on T)

Remarks:

- No discount $\gamma = 1$
- Exact model, deterministic

Theory does not apply to MDPs
Can we extend to $\gamma < 1$ and stochastic dynamics?

Stability of MPC - Connection to RL MPC

Policy π^{MPC} from

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

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

Stability of MPC - Connection to RL MPC

Policy π^{MPC} from

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

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

MPC scheme is (nominally) stabilizing if there is λ such that

$$\ell(\mathbf{s}, \mathbf{a}) := L(\mathbf{s}, \mathbf{a}) + \lambda(\mathbf{s}) - \lambda(\mathbf{f}(\mathbf{s}, \mathbf{a})) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|), \quad \forall \mathbf{s}, \mathbf{a}$$

where κ is K_∞ (+conditions on T)

Stability of MPC - Connection to RL MPC

Policy π^{MPC} from

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

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

Equivalent MPC

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

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

is stable

MPC scheme is (nominally) stabilizing if there is λ such that

$$\ell(\mathbf{s}, \mathbf{a}) := L(\mathbf{s}, \mathbf{a}) + \lambda(\mathbf{s}) - \lambda(\mathbf{f}(\mathbf{s}, \mathbf{a})) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|), \quad \forall \mathbf{s}, \mathbf{a}$$

where κ is K_∞ (+conditions on T)

Stability of MPC - Connection to RL MPC

Policy π^{MPC} from

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

Equivalent MPC

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

is stable

MPC scheme is (nominally) stabilizing if there is λ such that

$$\ell(\mathbf{s}, \mathbf{a}) := L(\mathbf{s}, \mathbf{a}) + \lambda(\mathbf{s}) - \lambda(\mathbf{f}(\mathbf{s}, \mathbf{a})) \geq \kappa(\|\mathbf{s} - \mathbf{s}_s\|), \quad \forall \mathbf{s}, \mathbf{a}$$

where κ is K_∞ (+conditions on T)

Remarks

- Modifying the MPC cost is a concept already present in dissipativity theory!
- Aligned with modifying the cost for MPC performance
- $\rightarrow$ Merge the [RL & stability modifications](#) for “stability by design”

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J\left(\pi_{\theta}^{\text{MPC}}\right)=\sum_{k=0}^{\infty} L\left(s_k, a_k\right)$$

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J\left(\pi_{\theta}^{\text{MPC}}\right)=\sum_{k=0}^{\infty} L\left(s_k, a_k\right)$$

Parametrized policy $\pi_{\theta}^{\text{MPC}}$ from:

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

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J\left(\pi_{\theta}^{\text{MPC}}\right)=\sum_{k=0}^{\infty} L\left(s_k, a_k\right)$$

- Learning based on L
- Impose constraint:

$$L_{\theta}(s, a) \geq \kappa\left(\left\|s-s_s\right\|\right), \quad \forall s, a$$

throughout the learning

- L_{θ} different than L due to stability + model error

Parametrized policy $\pi_{\theta}^{\text{MPC}}$ from:

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

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J\left(\pi_{\theta}^{\text{MPC}}\right)=\sum_{k=0}^{\infty} L\left(s_k, a_k\right)$$

- Learning based on L
- Impose constraint:

$$L_{\theta}(s, a) \geq \kappa\left(\left\|s-s_s\right\|\right), \quad \forall s, a$$

throughout the learning

- L_{θ} different than L due to stability + model error

Parametrized policy $\pi_{\theta}^{\text{MPC}}$ from:

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

Theorem: under some conditions

- $\pi_{\theta}^{\text{MPC}} \rightarrow \pi_{\star}$ if $\pi_{\star}$ is stabilizing
- $\pi_{\theta}^{\text{MPC}} \rightarrow$ best stabilizing policy otherwise

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J\left(\pi_{\theta}^{\text{MPC}}\right)=\sum_{k=0}^{\infty} L\left(s_k, a_k\right)$$

- Learning based on L
- Impose constraint:

$$L_{\theta}(s, a) \geq \kappa\left(\|s-s_s\|\right), \quad \forall s, a$$

throughout the learning

- L_{θ} different than L due to stability + model error

Parametrized policy $\pi_{\theta}^{\text{MPC}}$ from:

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

Theorem: under some conditions

- $\pi_{\theta}^{\text{MPC}} \rightarrow \pi_{\star}$ if $\pi_{\star}$ is stabilizing
- $\pi_{\theta}^{\text{MPC}} \rightarrow$ best stabilizing policy otherwise

Change of philosophy from “classic”
dissipativity framework:
stability analysis $\rightarrow$ stable design

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** π_θ^{MPC} minimizing:

$$J(\pi_\theta^{\text{MPC}}) = \sum_{k=0}^{\infty} L(s_k, a_k)$$

Constraint

$$L_\theta(s, a) \geq \kappa(\|s - s_s\|), \quad \forall s, a$$

is semi-infinite programming, not trivial

Some solutions:

- Sum-of-Squares (SOS) prog.
- Convex L_θ (+ radially unbounded)
- Something else?

Parametrized policy π_θ^{MPC} from:

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

Theorem: under some conditions

- $\pi_\theta^{\text{MPC}} \rightarrow \pi_\star$ if $\pi_\star$ is stabilizing
- $\pi_\theta^{\text{MPC}} \rightarrow$ best stabilizing policy otherwise

Change of philosophy from “classic”
dissipativity framework:
stability analysis $\rightarrow$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J\left(\pi_{\theta}^{\text{MPC}}\right)=\sum_{k=0}^{\infty} L\left(s_k, a_k\right)$$

Note that λ_{θ} is redundant for policy gradient, needed for Q-learning...

Parametrized policy $\pi_{\theta}^{\text{MPC}}$ from:

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

Theorem: under some conditions

- $\pi_{\theta}^{\text{MPC}} \rightarrow \pi_{\star}$ if $\pi_{\star}$ is stabilizing
- $\pi_{\theta}^{\text{MPC}} \rightarrow$ best stabilizing policy otherwise

Change of philosophy from “classic”
dissipativity framework:
stability analysis $\rightarrow$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J\left(\pi_{\theta}^{\text{MPC}}\right)=\sum_{k=0}^{\infty} L\left(s_k, a_k\right)$$

Extension to stable policy for MDPs?

- Need stability with discount
- Need “stochastic dissipativity”

Parametrized policy $\pi_{\theta}^{\text{MPC}}$ from:

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

Theorem: under some conditions

- $\pi_{\theta}^{\text{MPC}} \rightarrow \pi_{\star}$ if $\pi_{\star}$ is stabilizing
- $\pi_{\theta}^{\text{MPC}} \rightarrow$ best stabilizing policy otherwise

Change of philosophy from “classic”
dissipativity framework:
stability analysis $\rightarrow$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J(\pi_{\theta}^{\text{MPC}}) = \sum_{k=0}^{\infty} L(s_k, a_k)$$

Extension to stable policy for MDPs?

- Need stability with discount
- Need “stochastic dissipativity”

MDP dissipativity: (2x Automatica '22)

- Use Strong Discounted Strict Dissipativity conditions
- Form the dissipativity equations in the measure space of the MDP

Parametrized policy $\pi_{\theta}^{\text{MPC}}$ from:

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

Theorem: under some conditions

- $\pi_{\theta}^{\text{MPC}} \rightarrow \pi_{\star}$ if $\pi_{\star}$ is stabilizing
- $\pi_{\theta}^{\text{MPC}} \rightarrow$ best stabilizing policy otherwise

Change of philosophy from “classic”
dissipativity framework:
stability analysis $\rightarrow$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

Given arbitrary stage cost $L(s, a)$, build a **stable policy** $\pi_{\theta}^{\text{MPC}}$ minimizing:

$$J(\pi_{\theta}^{\text{MPC}}) = \sum_{k=0}^{\infty} L(s_k, a_k)$$

Extension to stable policy for MDPs?

- Need stability with discount
- Need “stochastic dissipativity”

MDP dissipativity: (2x Automatica '22)

- Use Strong Discounted Strict Dissipativity conditions
- Form the dissipativity equations in the measure space of the MDP

Parametrized policy $\pi_{\theta}^{\text{MPC}}$ from:

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

Theorem: under some conditions

- $\pi_{\theta}^{\text{MPC}} \rightarrow \pi_{\star}$ if $\pi_{\star}$ is stabilizing
- $\pi_{\theta}^{\text{MPC}} \rightarrow$ best stabilizing policy otherwise

Change of philosophy from “classic”
dissipativity framework:
stability analysis $\rightarrow$ **stable design**

We have the maths to treat this, not yet the algorithms for the stochastic case...

Outline

- 1 SYSID & MPC
- 2 Safe RL via MPC
- 3 Exploration with MPC
- 4 Stability-constrained Learning with MPC
- 5 Explored questions

RL+MPC with Future Information - An important case

Future Information in MPC

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

RL+MPC with Future Information - An important case

Future Information in MPC

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

Notes

- Assume $\mathbf{I}_{0, \dots, \infty}$ exogenous, independent of $\mathbf{s}$ and $\mathbf{a}$
- Examples: reference trajectory, prices, weather forecasts, etc.
- In practice it is known up to a certain time in the future
- $\mathbf{I}_{k > N}$ typ. regresses to its “ground statistics”

RL+MPC with Future Information - An important case

Future Information in MPC

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

Notes

- Assume $\mathbf{I}_{0, \dots, \infty}$ exogenous, independent of $\mathbf{s}$ and $\mathbf{a}$
- Examples: reference trajectory, prices, weather forecasts, etc.
- In practice it is known up to a certain time in the future
- $\mathbf{I}_{k > N}$ typ. regresses to its “ground statistics”

RL+MPC theory applies but:

- $\mathbf{I}$ is part of the state, i.e. $\{\mathbf{s}, \mathbf{I}\}$ is the state we need in RL

RL+MPC with Future Information - An important case

Future Information in MPC

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

Notes

- Assume $\mathbf{I}_{0, \dots, \infty}$ exogenous, independent of $\mathbf{s}$ and $\mathbf{a}$
- Examples: reference trajectory, prices, weather forecasts, etc.
- In practice it is known up to a certain time in the future
- $\mathbf{I}_{k > N}$ typ. regresses to its “ground statistics”

RL+MPC theory applies but:

- $\mathbf{I}$ is part of the state, i.e. $\{\mathbf{s}, \mathbf{I}\}$ is the state we need in RL
- Theory requires the MPC modifications to be non-localized

RL+MPC with Future Information - An important case

Future Information in MPC

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

Notes

- Assume $\mathbf{I}_{0, \dots, \infty}$ exogenous, independent of $\mathbf{s}$ and $\mathbf{a}$
- Examples: reference trajectory, prices, weather forecasts, etc.
- In practice it is known up to a certain time in the future
- $\mathbf{I}_{k \gg N}$ typ. regresses to its “ground statistics”

RL+MPC theory applies but:

- $\mathbf{I}$ is part of the state, i.e. $\{\mathbf{s}, \mathbf{I}\}$ is the state we need in RL
- Theory requires the MPC modifications to be non-localized
- In practice we work on a finite information horizon

RL+MPC with Future Information - An important case

Future Information in MPC

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta}(\mathbf{x}_N, \mathbf{I}_{N, \dots, N+M}) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{I}_{k, \dots, k+M})$$

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

$$\mathbf{h}_{\theta}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{I}_{k, \dots, k+M}) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}$$

$$\text{and } \pi_{\theta}(\mathbf{s}, \mathbf{I}) = \mathbf{u}_0^*(\theta, \mathbf{s}, \mathbf{I})$$

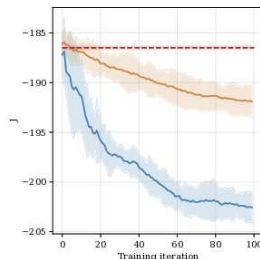
Notes

- Assume $\mathbf{I}_{0, \dots, \infty}$ exogenous, independent of $\mathbf{s}$ and $\mathbf{a}$
- Examples: reference trajectory, prices, weather forecasts, etc.
- In practice it is known up to a certain time in the future
- $\mathbf{I}_{k \gg N}$ typ. regresses to its “ground statistics”

RL+MPC theory applies but:

- $\mathbf{I}$ is part of the state, i.e. $\{\mathbf{s}, \mathbf{I}\}$ is the state we need in RL
- Theory requires the MPC modifications to be non-localized
- In practice we work on a finite information horizon

Example



RL+MPC with Future Information - An important case

Future Information in MPC

$$\min_{\mathbf{x}, \mathbf{u}} \quad T_{\theta}(\mathbf{x}_N, \mathbf{I}_{N, \dots, N+M}) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{I}_{k, \dots, k+M})$$

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

$$\mathbf{h}_{\theta}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{I}_{k, \dots, k+M}) \leq 0, \quad \mathbf{x}_0 = \mathbf{s}$$

$$\text{and } \pi_{\theta}(\mathbf{s}, \mathbf{I}) = \mathbf{u}_0^*(\theta, \mathbf{s}, \mathbf{I})$$

Notes

- Assume $\mathbf{I}_{0, \dots, \infty}$ exogenous, independent of $\mathbf{s}$ and $\mathbf{a}$
- Examples: reference trajectory, prices, weather forecasts, etc.
- In practice it is known up to a certain time in the future
- $\mathbf{I}_{k \gg N}$ typ. regresses to its “ground statistics”

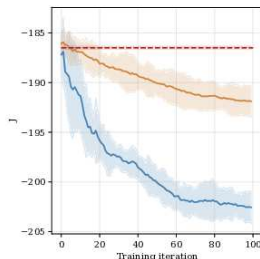
RL+MPC theory applies but:

- $\mathbf{I}$ is part of the state, i.e. $\{\mathbf{s}, \mathbf{I}\}$ is the state we need in RL
- Theory requires the MPC modifications to be non-localized
- In practice we work on a finite information horizon
- Difficulty: high-dimensional θ and

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} \left[\nabla_{\theta} \pi_{\theta}(\mathbf{s}, \mathbf{I}) \nabla_{\mathbf{a}} Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{I}, \mathbf{a}) \big|_{\mathbf{a}=\pi_{\theta}(\mathbf{s}, \mathbf{I})} \right]$$

requires a high-dimensional critic $Q^{\pi_{\theta}}$

Example



MPC as Its Own Critic

Full MPC parametrization:

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

Policy gradient

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} \left[\nabla_{\theta} \pi_{\theta}(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a})|_{\mathbf{a}=\pi_{\theta}(\mathbf{s})} \right]$$

requires the critic $Q^{\pi_{\theta}} : S \times A \rightarrow \mathbb{R}$

MPC as Its Own Critic

Full MPC parametrization:

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

Policy gradient

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} \left[\nabla_{\theta} \pi_{\theta}(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a})|_{\mathbf{a}=\pi_{\theta}(\mathbf{s})} \right]$$

requires the critic $Q^{\pi_{\theta}} : S \times A \rightarrow \mathbb{R}$

Difficulties

- For a large state space, $Q^{\pi_{\theta}}$ is high-dimensional
- For some applications, “a lot is happening” in some small parts of $S \times A$

MPC as Its Own Critic

Full MPC parametrization:

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

Methods

- In “classical” RL, critic is a DNN, needs to be learned “from scratch”, end-to-end
- Q_{θ} from MPC has often already a shape close to the critic

Policy gradient

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} \left[\nabla_{\theta} \pi_{\theta}(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a})|_{\mathbf{a}=\pi_{\theta}(\mathbf{s})} \right]$$

requires the critic $Q^{\pi_{\theta}} : S \times A \rightarrow \mathbb{R}$

Difficulties

- For a large state space, $Q^{\pi_{\theta}}$ is high-dimensional
- For some applications, “a lot is happening” in some small parts of $S \times A$

MPC as Its Own Critic

Full MPC parametrization:

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

Methods

- In “classical” RL, critic is a DNN, needs to be learned “from scratch”, end-to-end
- Q_{θ} from MPC has often already a shape close to the critic

Key idea (simplified, for deterministic policy gradient)

- Approximate $Q^{\pi_{\theta}}$ with a Taylor expansion of Q_{θ}

$$Q^{\pi_{\theta}} \approx Q_{\theta} + \nabla_{\theta} Q_{\theta} \Delta \theta + \Delta \mathbf{a}^{\top} \nabla_{\mathbf{a}_{\theta}} Q_{\theta} \Delta \theta$$

where $\Delta \theta$ is the **critic parameters**, estimated via classical RL methods

Policy gradient

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} \left[\nabla_{\theta} \pi_{\theta}(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a}) |_{\mathbf{a}=\pi_{\theta}(\mathbf{s})} \right]$$

requires the critic $Q^{\pi_{\theta}} : S \times A \rightarrow \mathbb{R}$

Difficulties

- For a large state space, $Q^{\pi_{\theta}}$ is high-dimensional
- For some applications, “a lot is happening” in some small parts of $S \times A$

MPC as Its Own Critic

Full MPC parametrization:

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

Methods

- In “classical” RL, critic is a DNN, needs to be learned “from scratch”, end-to-end
- Q_{θ} from MPC has often already a shape close to the critic

Key idea (simplified, for deterministic policy gradient)

- Approximate $Q^{\pi_{\theta}}$ with a Taylor expansion of Q_{θ}

$$Q^{\pi_{\theta}} \approx Q_{\theta} + \nabla_{\theta} Q_{\theta} \Delta \theta + \Delta \mathbf{a}^{\top} \nabla_{\mathbf{a}_{\theta}} Q_{\theta} \Delta \theta$$

where $\Delta \theta$ is the **critic parameters**, estimated via classical RL methods

- Policy gradient becomes

$$\nabla_{\theta} J(\pi_{\theta}) \approx \mathbb{E} \left[\nabla_{\theta} \pi_{\theta}(\mathbf{s}) \nabla_{\mathbf{a}_{\theta}} Q_{\theta} \Delta \theta \Big|_{\mathbf{a}=\pi_{\theta}(\mathbf{s})} \right]$$

Policy gradient

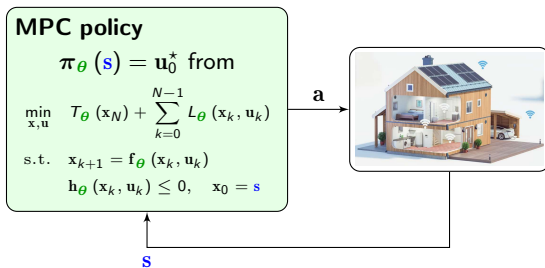
$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E} \left[\nabla_{\theta} \pi_{\theta}(\mathbf{s}) \nabla_{\mathbf{a}} Q^{\pi_{\theta}}(\mathbf{s}, \mathbf{a}) \Big|_{\mathbf{a}=\pi_{\theta}(\mathbf{s})} \right]$$

requires the critic $Q^{\pi_{\theta}} : S \times A \rightarrow \mathbb{R}$

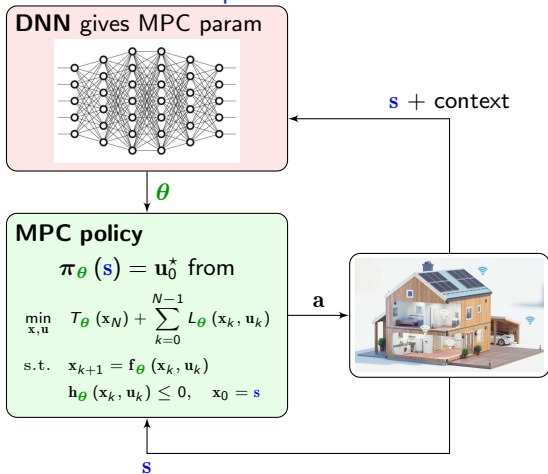
Difficulties

- For a large state space, $Q^{\pi_{\theta}}$ is high-dimensional
- For some applications, “a lot is happening” in some small parts of $S \times A$

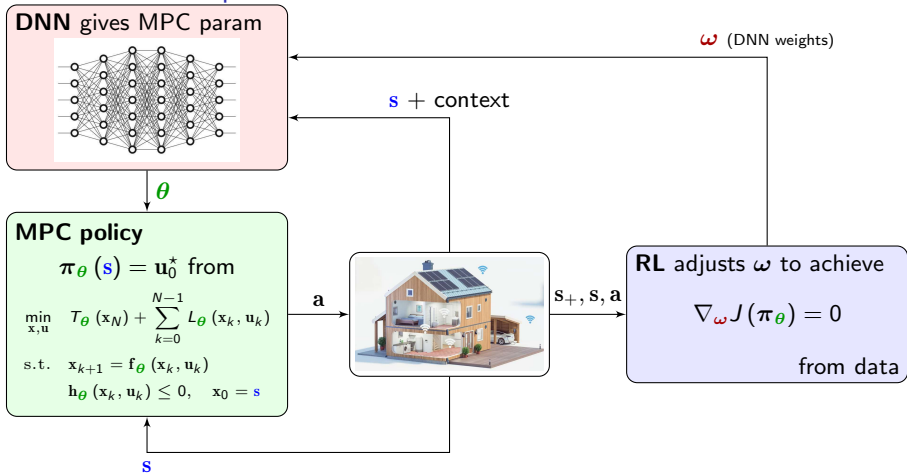
DNN-based MPC parameters



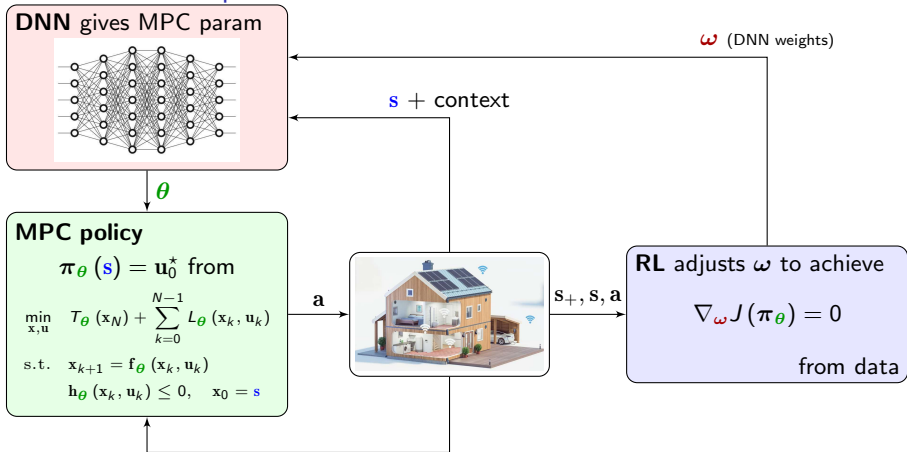
DNN-based MPC parameters



DNN-based MPC parameters



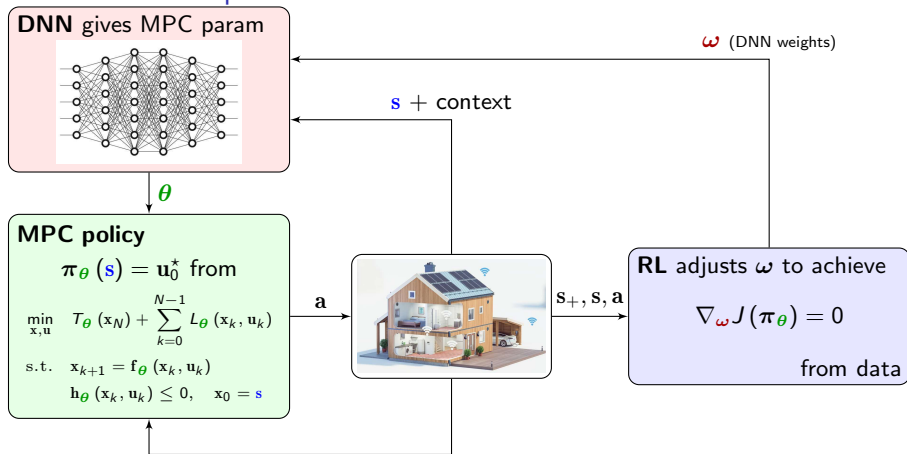
DNN-based MPC parameters



Remarks: s

- Chain rule $\nabla_{\omega} \pi_{\theta} = \nabla_{\omega} \theta \nabla_{\theta} \pi_{\theta}$ is simple and efficient

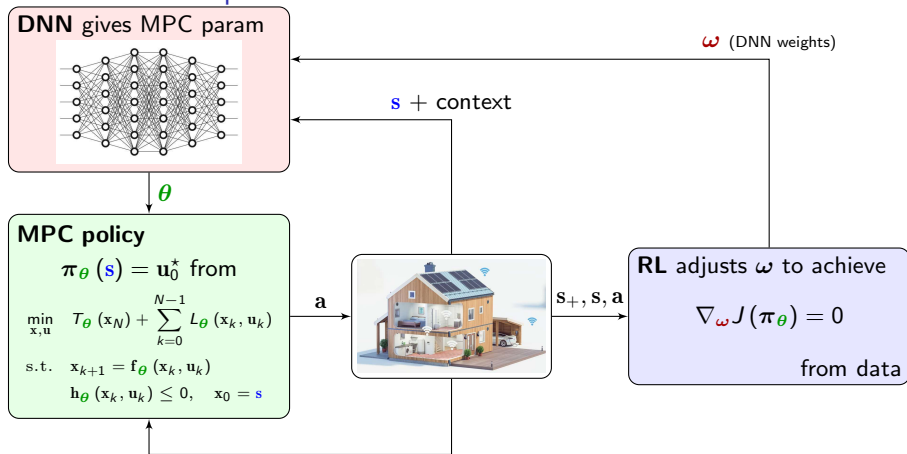
DNN-based MPC parameters



Remarks: s

- Chain rule $\nabla_{\omega} \pi_{\theta} = \nabla_{\omega} \theta \nabla_{\theta} \pi_{\theta}$ is simple and efficient
- Achieves great performance on investigated examples (e.g. car racing)

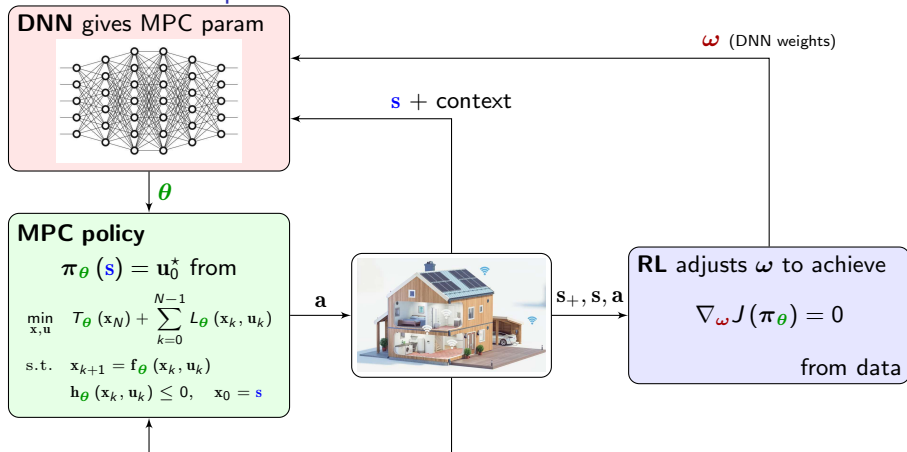
DNN-based MPC parameters



Remarks: s

- Chain rule $\nabla_{\omega} \pi_{\theta} = \nabla_{\omega} \theta \nabla_{\theta} \pi_{\theta}$ is simple and efficient
- Achieves great performance on investigated examples (e.g. car racing)
- Regularize DNN such that if s is close to expected dynamics, θ remains consistent?

DNN-based MPC parameters

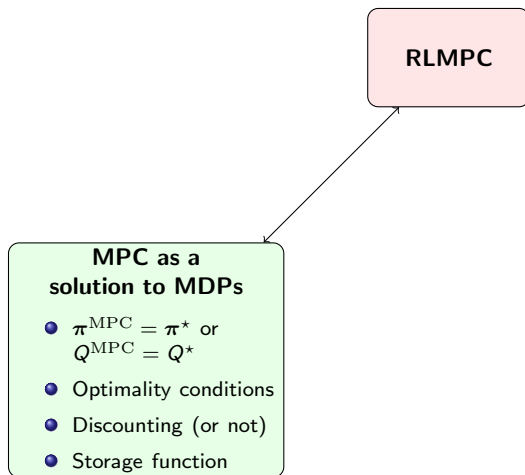


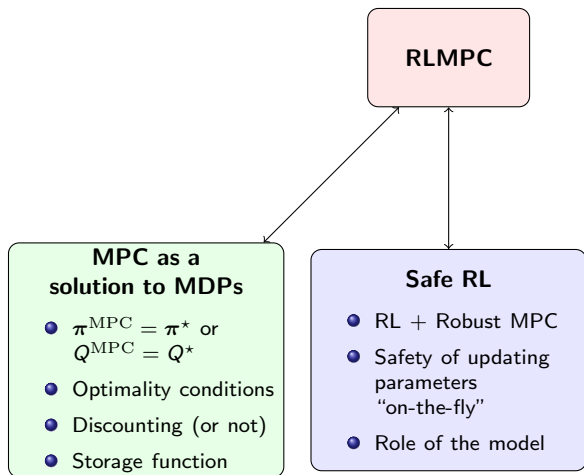
Remarks: $\mathbf{s}$

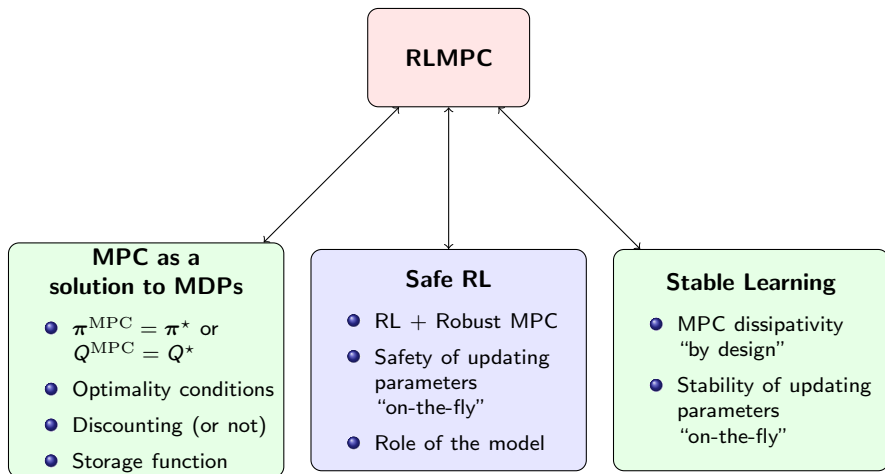
- Chain rule $\nabla_{\omega} \pi_{\theta} = \nabla_{\omega} \theta \nabla_{\theta} \pi_{\theta}$ is simple and efficient
- Achieves great performance on investigated examples (e.g. car racing)
- Regularize DNN such that if $\mathbf{s}$ is close to expected dynamics, θ remains consistent?
- Need a theory to explain what that means, how to do it, and why it is useful(?)

Explored questions - Fundamentals

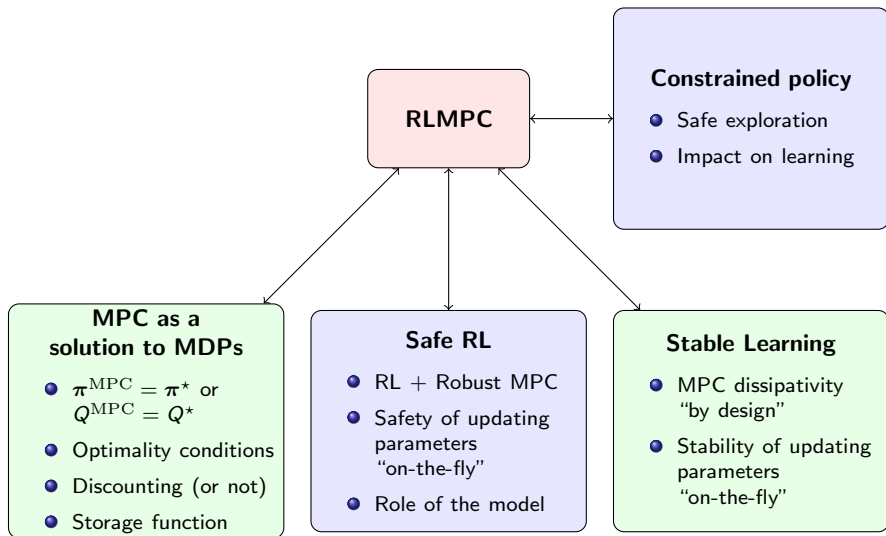
RLMPC







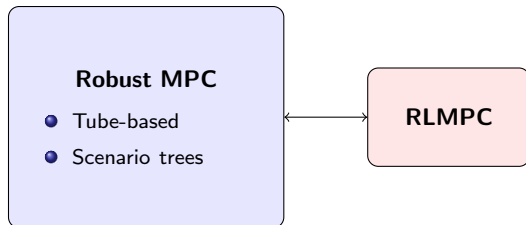
Explored questions - Fundamentals



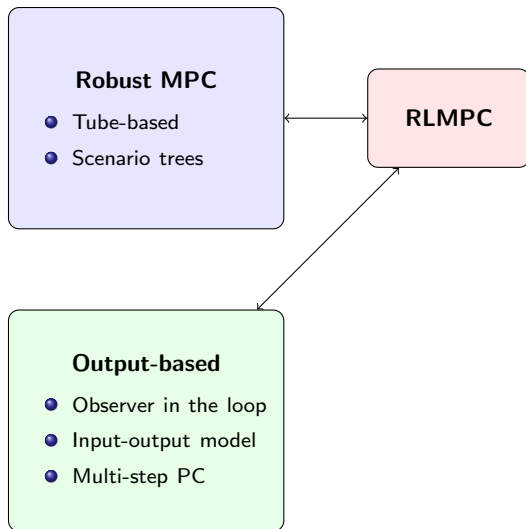
Explored questions - Types of MPC & Learning

RLMPC

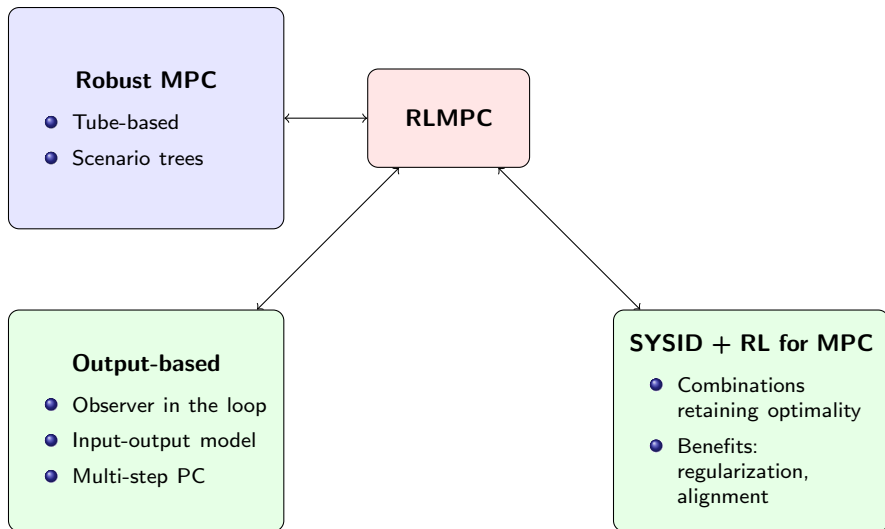
Explored questions - Types of MPC & Learning



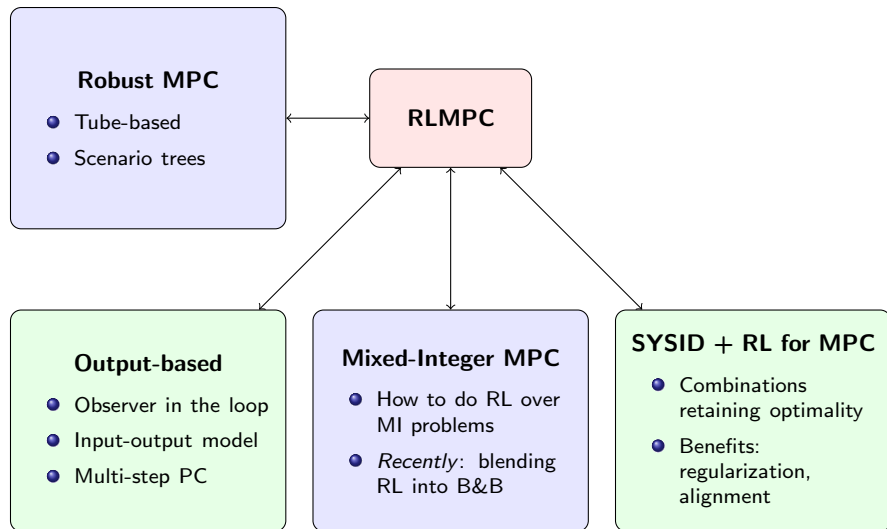
Explored questions - Types of MPC & Learning



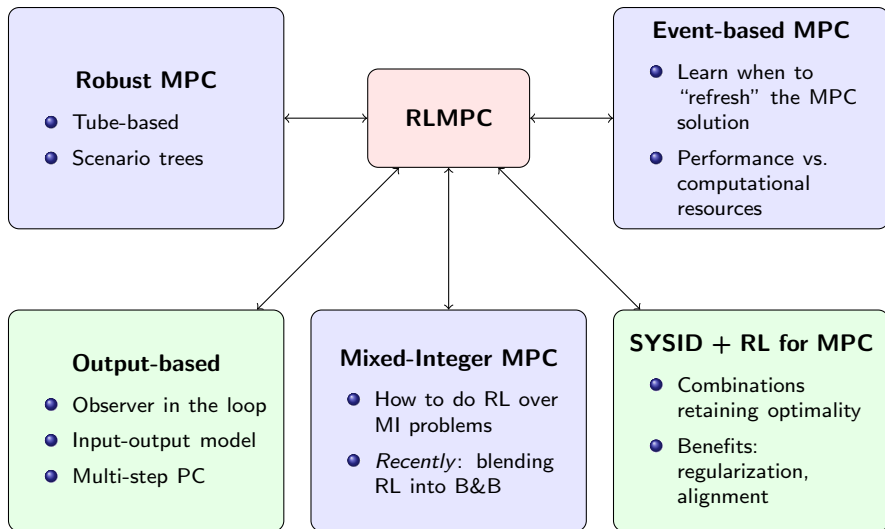
Explored questions - Types of MPC & Learning



Explored questions - Types of MPC & Learning



Explored questions - Types of MPC & Learning



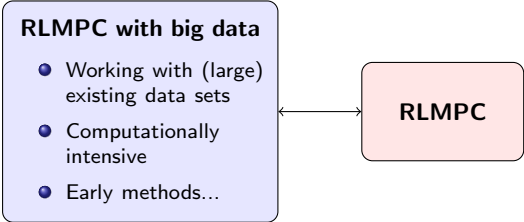
RLMPC

Explored questions - Algorithms

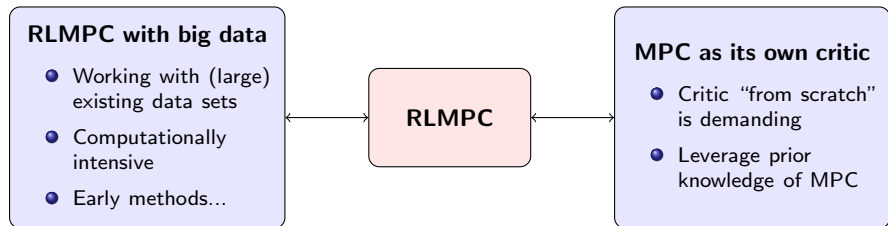
RLMPC with big data

- Working with (large) existing data sets
- Computationally intensive
- Early methods...

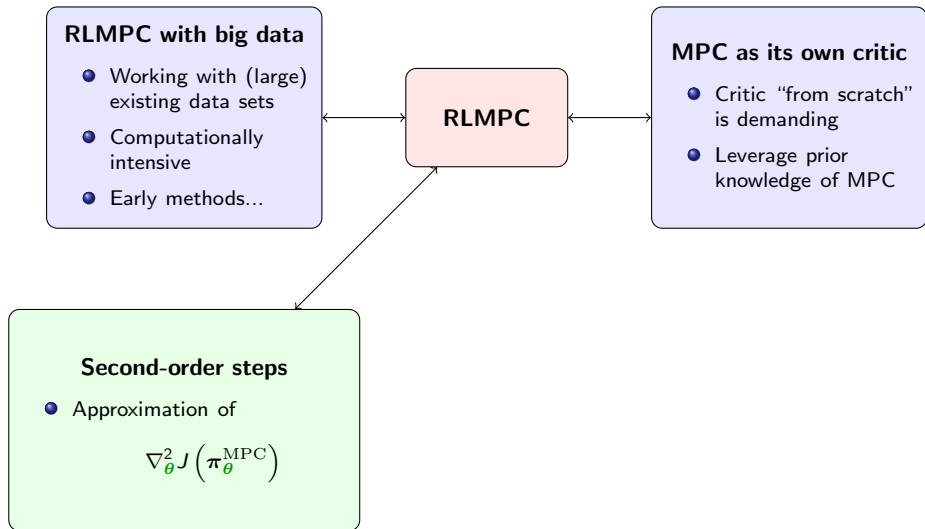
RLMPC



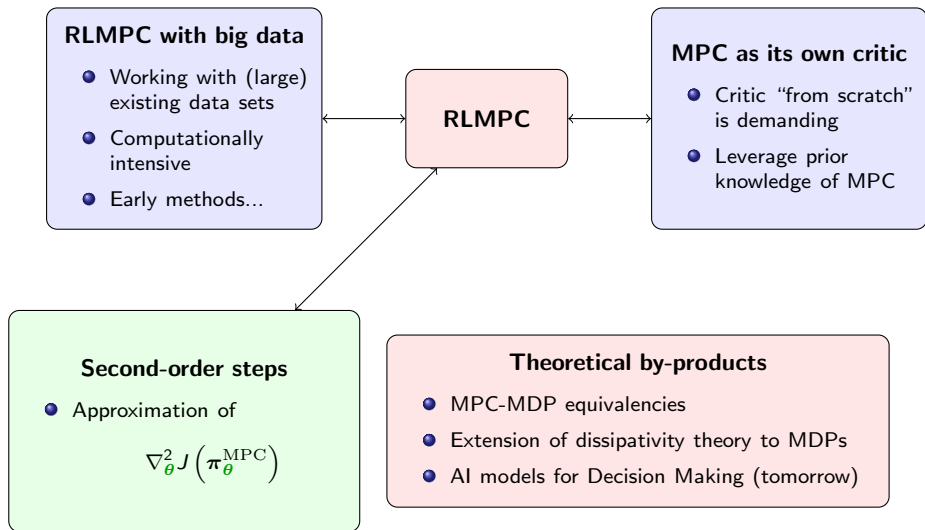
Explored questions - Algorithms



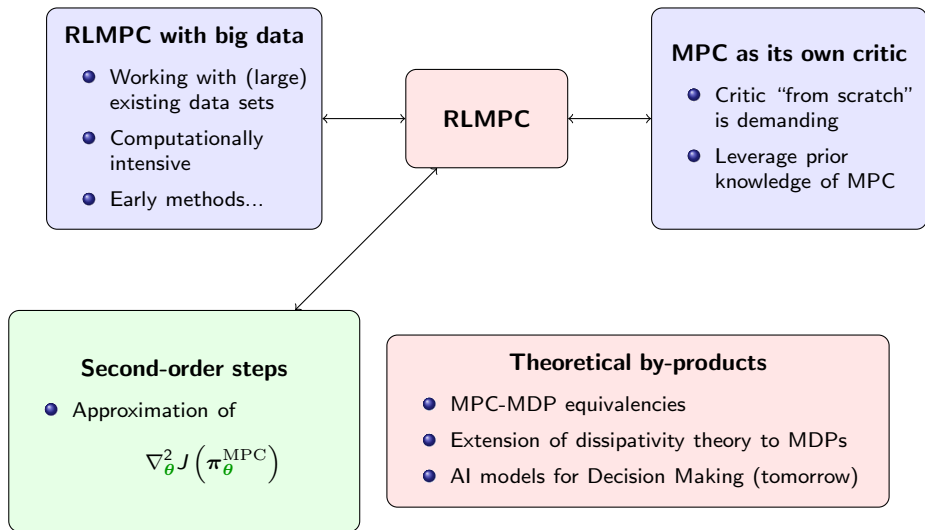
Explored questions - Algorithms



Explored questions - Algorithms



Explored questions - Algorithms



>60 papers, but most aspects could use more developments, algorithms, software, applications. Welcome onboard!

What we have seen:

- Robust MPC can provide a safe policy, RL can tune that policy for performance
- Some (standard) limitations apply
- MPC enables safe (as in feasible) exploration, some challenges though
- RL over MPC provides a pathway to enforcing stability in the learning process
- Some open research questions for MDPs
- RL over MPC with belief states – Theory seems to add up, implementation is an open question

What we have seen:

- Robust MPC can provide a safe policy, RL can tune that policy for performance
- Some (standard) limitations apply
- MPC enables safe (as in feasible) exploration, some challenges though
- RL over MPC provides a pathway to enforcing stability in the learning process
- Some open research questions for MDPs
- RL over MPC with belief states – Theory seems to add up, implementation is an open question

What we will do next: RL over MPC

- Open research questions and hard problems (*Afternoon, untested lecture*)
- Beyond MPC – AI for decisions, Co-Design (*Tomorrow, untested lecture*)