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 $\pi^\star$
- 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 $\pi^\star$
- 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
- Exploration
- Some more explored questions

RL and MPC
Overview of Explored Questions

Sebastien Gros

Dept. of Cybernetic, NTNU
Faculty of Information Tech.

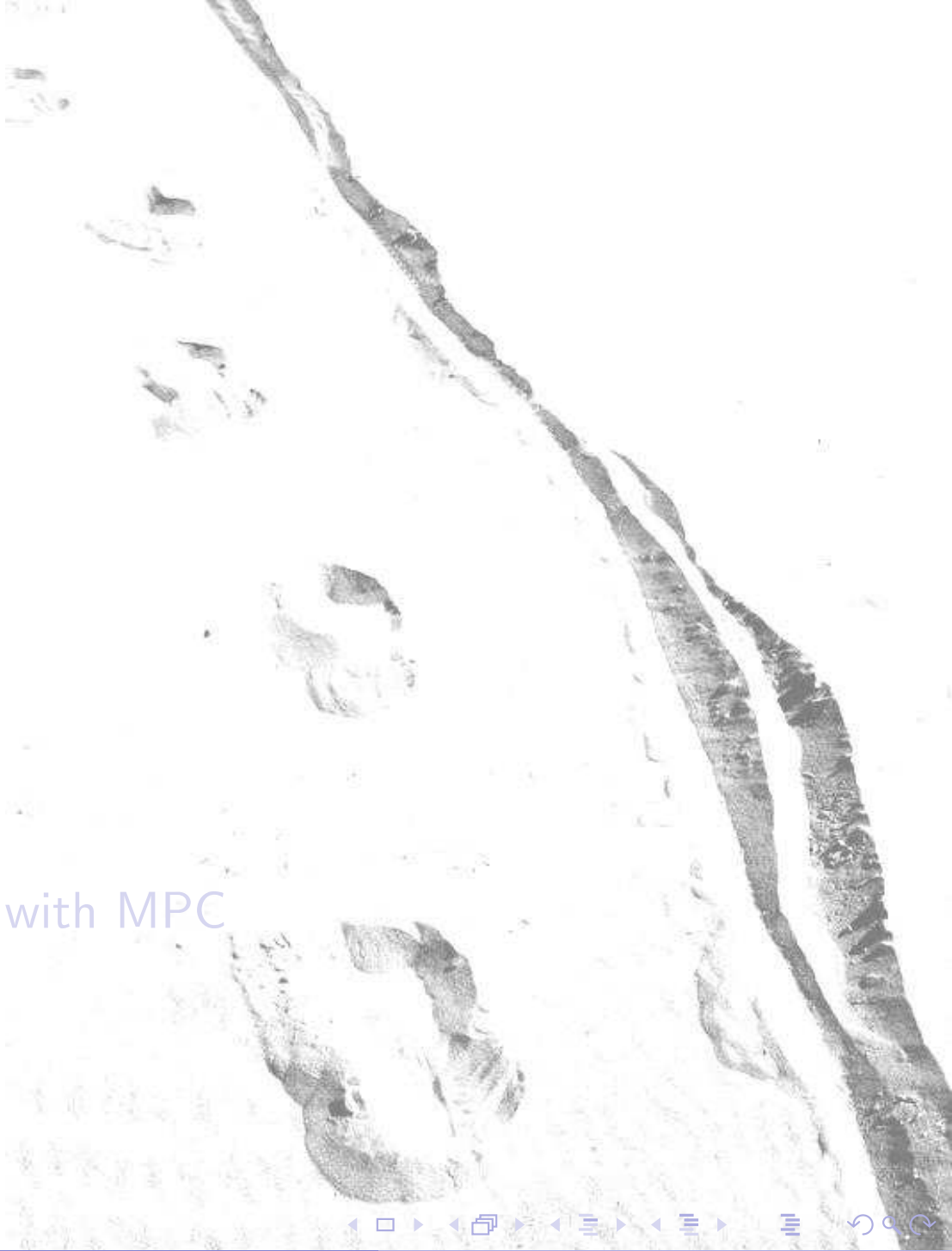
Freiburg PhD School

Outline

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

Outline

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



RL & SYSID for MPC

Full MPC parametrization:

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

Remarks

- In general, $\boldsymbol{\theta}$ such that

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

  is not unique
- Leaves room to align the model

RL & SYSID for MPC

Full MPC parametrization:

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

Remarks

- In general, $\boldsymbol{\theta}$ such that

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

  is not unique
- Leaves room to align the model

Multi-objective approach

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

is non-conflicting for a rich parametrization

RL & SYSID for MPC

Full MPC parametrization:

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

Remarks

- In general, θ such that

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

  is not unique
- Leaves room to align the model

Multi-objective approach

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

is non-conflicting for a rich parametrization

RL-prioritized

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

RL & SYSID for MPC

Full MPC parametrization:

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

Remarks

- In general, $\boldsymbol{\theta}$ such that

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

  is not unique
- Leaves room to align the model

Multi-objective approach

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

is non-conflicting for a rich parametrization

RL-prioritized

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

- Yields different ways of combining SYSID & RL
- SYSID serves as "regularization" for RL
- Especially useful for $\boldsymbol{\theta}$ large-dimensional

RL & SYSID for MPC

Full MPC parametrization:

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

Remarks

- In general, $\boldsymbol{\theta}$ such that

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

  is not unique
- Leaves room to align the model

Multi-objective approach

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

is non-conflicting for a rich parametrization

- Yields different ways of combining SYSID & RL
- SYSID serves as "regularization" for RL
- Especially useful for $\boldsymbol{\theta}$ large-dimensional

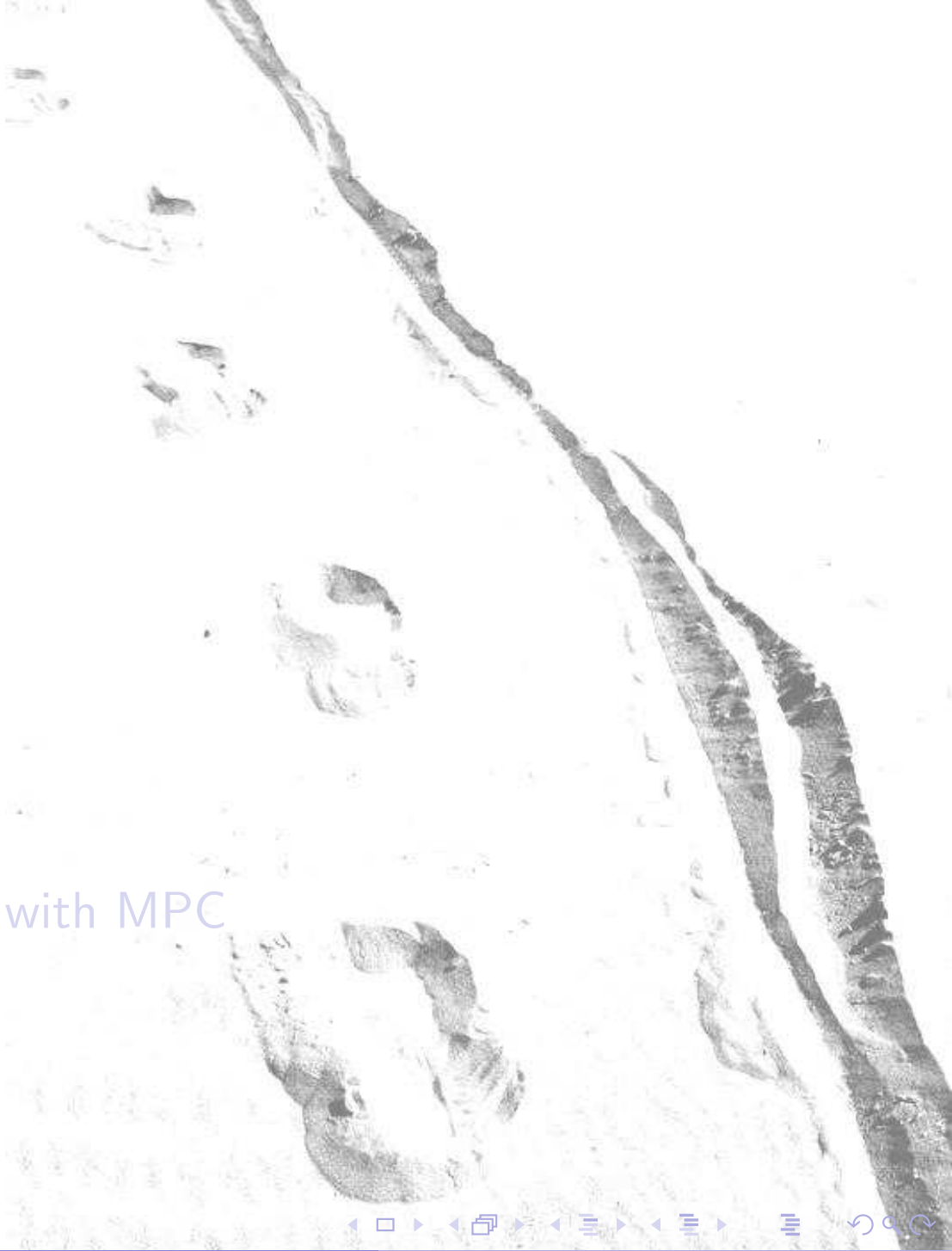
RL-prioritized

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

Some "draft" theory on how meaningful model $\mathbf{f}_{\boldsymbol{\theta}}$ can be made to be. Not yet published.

Outline

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



Robust MPC - Uncertainty model

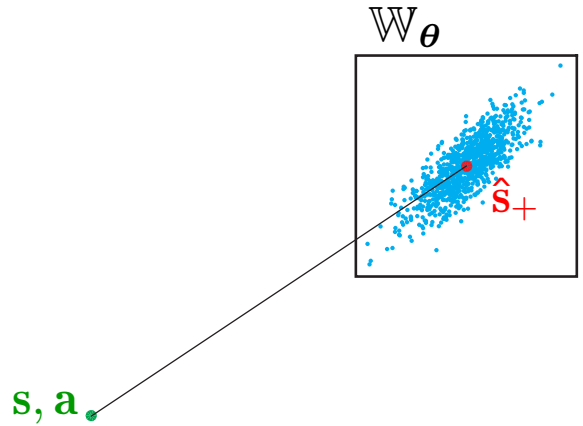
$$\text{True system:} \quad \mathbf{s}_+ \sim \mathbb{P}\left[\cdot \,|\mathbf{s}, \mathbf{a}\right]$$

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

Robust MPC - Uncertainty model

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

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



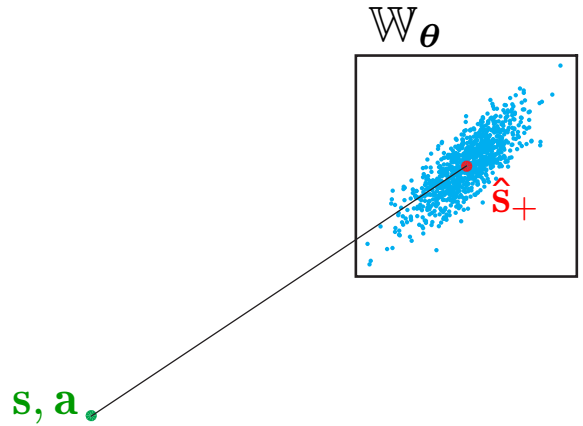
Dispersion: $\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right) + \mathbb{W}_{\boldsymbol{\theta}}$ contains the support of $\mathbb{P}\left[\,\cdot\,|\mathbf{s},\mathbf{a}\,\right]$, i.e.

$$\mathbf{s}_+ \in \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right) + \mathbb{W}_{\boldsymbol{\theta}} \tag{1}$$

with probability 1

Robust MPC - Uncertainty model

$$\text{True system:} \quad \mathbf{s}_+ \sim \mathbb{P}\left[\,\cdot\,|\mathbf{s},\mathbf{a}\right]$$
$$\text{Deterministic model:} \quad \hat{\mathbf{s}}_+ = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right)$$



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

$$\mathbf{s}_+ \in \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right) + \mathbb{W}_{\boldsymbol{\theta}} \tag{1}$$

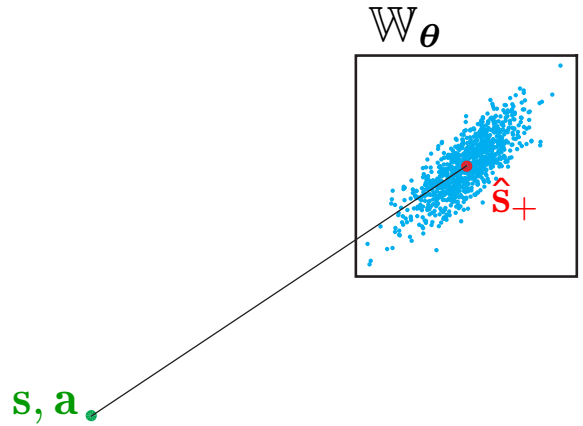
with probability 1

Remarks:

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

Robust MPC - Uncertainty model

$$\text{True system:} \quad \mathbf{s}_+ \sim \mathbb{P}\left[\,\cdot\,|\mathbf{s},\mathbf{a}\right]$$
$$\text{Deterministic model:} \quad \hat{\mathbf{s}}_+ = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right)$$



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

$$\mathbf{s}_+ \in \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right) + \mathbb{W}_{\boldsymbol{\theta}} \qquad (1)$$

with probability 1

Condition

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

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

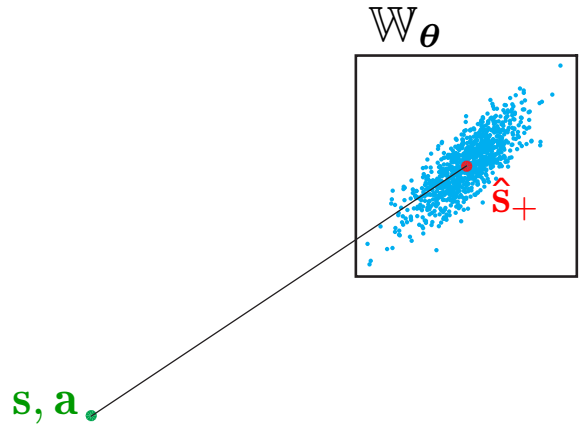
$\rightarrow$ constraints on $\boldsymbol{\theta}$

Remarks:

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

Robust MPC - Uncertainty model

$$\text{True system:} \quad \mathbf{s}_+ \sim \mathbb{P}\left[\,\cdot\,|\mathbf{s},\mathbf{a}\right]$$
$$\text{Deterministic model:} \quad \hat{\mathbf{s}}_+ = \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right)$$



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

$$\mathbf{s}_+ \in \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right) + \mathbb{W}_{\boldsymbol{\theta}} \tag{1}$$

with probability 1

Remarks:

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

Condition

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

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

$\rightarrow$ constraints on $\boldsymbol{\theta}$

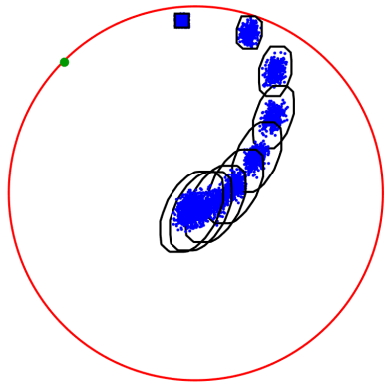
Containing the model-system mismatch becomes constraints in the parameters $\boldsymbol{\theta}$. Constraints can be readily formulated in terms of data.

Safe policies via Robust (N)MPC

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

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

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



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

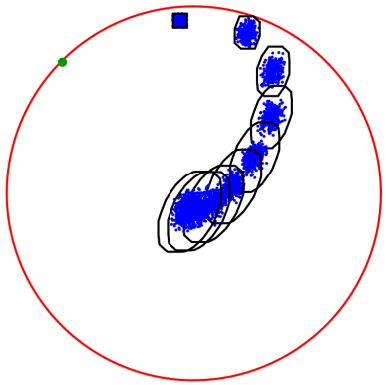
Safe policies via Robust (N)MPC

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

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

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

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

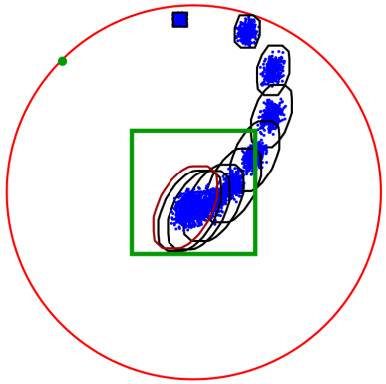


- $\mathbf{x}_{0,\ldots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \ldots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $\mathbf{x}_{1,\ldots,N-1}(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w})$ 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

Safe policies via Robust (N)MPC

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

$$\begin{aligned}\mathbf{u}^\star = \arg\min_{\mathbf{u}} \quad & \max_{\mathbf{w}\in\mathbb{W}_{\boldsymbol{\theta}}{}^N} \; T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{u}_{0,\ldots,N} \in \mathbb{U} \\ & \mathbf{x}_{1,\ldots,N-1}\left(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}\right) \in \mathbb{X}, \quad \forall\, \mathbf{w} \in \mathbb{W}_{\boldsymbol{\theta}}{}^{N-1} \\ & \mathbf{x}_N\left(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}\right) \in \mathbb{T}_{\boldsymbol{\theta}}, \quad \forall\, \mathbf{w} \in \mathbb{W}_{\boldsymbol{\theta}}{}^{N-1}\end{aligned}$$



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

Safe policies via Robust (N)MPC

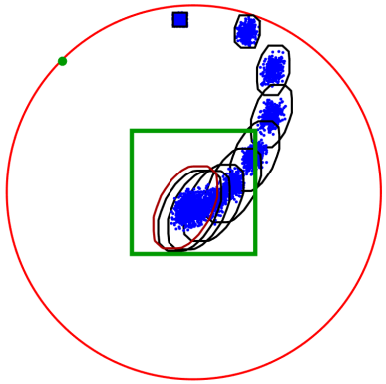
Robust (N)MPC delivers policy $\pi_{\boldsymbol{\theta}}(\mathbf{x}_0) = \mathbf{u}_0^\star$ from

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

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

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

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

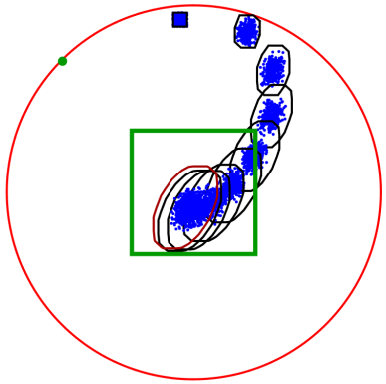


- $\mathbf{x}_{0,\ldots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \ldots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\boldsymbol{\theta}}$
- $\mathbf{x}_{1,\ldots,N-1}(\mathbf{u},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

Safe policies via Robust (N)MPC

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

$$\begin{aligned}\mathbf{u}^\star = \arg\min_{\mathbf{u}} \quad & \max_{\mathbf{w}\in\mathbb{W}_{\boldsymbol{\theta}}{}^N} \; T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{u}_{0,\ldots,N} \in \mathbb{U} \\ & \mathbf{x}_{1,\ldots,N-1}\left(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}\right) \in \mathbb{X}, \quad \forall \mathbf{w} \in \mathbb{W}_{\boldsymbol{\theta}}{}^{N-1} \\ & \mathbf{x}_N\left(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}\right) \in \mathbb{T}_{\boldsymbol{\theta}}, \quad \forall \mathbf{w} \in \mathbb{W}_{\boldsymbol{\theta}}{}^{N-1}\end{aligned}$$



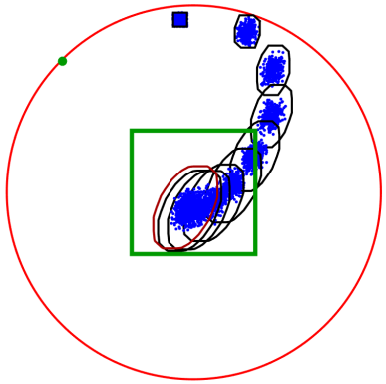
- $\mathbf{x}_{0,\ldots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \ldots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\boldsymbol{\theta}}$
- $\mathbf{x}_{1,\ldots,N-1}\left(\mathbf{u},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

$$\begin{aligned}\mathbf{u}^\star = \arg\min_{\mathbf{u}} \quad & \max_{\mathbf{w}\in\mathbb{W}_{\boldsymbol{\theta}}{}^N} \; T_{\boldsymbol{\theta}}\left(\mathbf{x}_N\right) + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}\left(\mathbf{x}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \mathbf{u}_{0,\ldots,N} \in \mathbb{U} \\ & \mathbf{x}_{1,\ldots,N-1}\left(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}\right) \in \mathbb{X}, \quad \forall \mathbf{w} \in \mathbb{W}_{\boldsymbol{\theta}}{}^{N-1} \\ & \mathbf{x}_N\left(\mathbf{u}, \mathbf{s}, \boldsymbol{\theta}, \mathbf{w}\right) \in \mathbb{T}_{\boldsymbol{\theta}}, \quad \forall \mathbf{w} \in \mathbb{W}_{\boldsymbol{\theta}}{}^{N-1}\end{aligned}$$



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

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

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

Safe Learning via Robust MPC

Robust NMPC parameters $\boldsymbol{\theta}$

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

Condition

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

enforces safety through $\boldsymbol{\theta}$

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

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

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

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

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

How to do Safe RL?

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

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

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

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

How to do Safe RL?

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

Also reads as:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$$

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

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

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

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

How to do Safe RL?

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

Also reads as:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$$

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

Safe RL steps $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$:

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

$$\text{s.t. } \mathbf{s}_+ - \mathbf{f}\left(\mathbf{s}, \mathbf{a}, \boldsymbol{\theta} + \Delta\boldsymbol{\theta}\right) \in \mathbb{W}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}$$

$$\forall \left(\mathbf{s}, \mathbf{a}, \mathbf{s}_+\right) \text{ in data set}$$

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

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

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

How to do Safe RL?

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

Also reads as:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$$

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

Safe RL steps $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$:

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

$$\text{s.t. } \mathbf{s}_+ - \mathbf{f}\left(\mathbf{s}, \mathbf{a}, \boldsymbol{\theta} + \Delta\boldsymbol{\theta}\right) \in \mathbb{W}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}$$

$$\forall \left(\mathbf{s}, \mathbf{a}, \mathbf{s}_+\right) \text{ in data set}$$

Safe RL steps seek performance under safety constraints

Safe Learning via Robust MPC

Robust NMPC parameters $\boldsymbol{\theta}$

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

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

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

How to do Safe RL?

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

Also reads as:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$$

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

Safe RL steps $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$:

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

$$\text{s.t. } \mathbf{s}_+ - \mathbf{f}\left(\mathbf{s}, \mathbf{a}, \boldsymbol{\theta} + \Delta\boldsymbol{\theta}\right) \in \mathbb{W}_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}$$

$$\forall \left(\mathbf{s}, \mathbf{a}, \mathbf{s}_+\right) \text{ in data set}$$

Safe RL steps seek performance under safety constraints

Careful: differentiating through Robust MPC requires implicit tube formulation. Gradient is a bit non-smooth.

Illustrative example

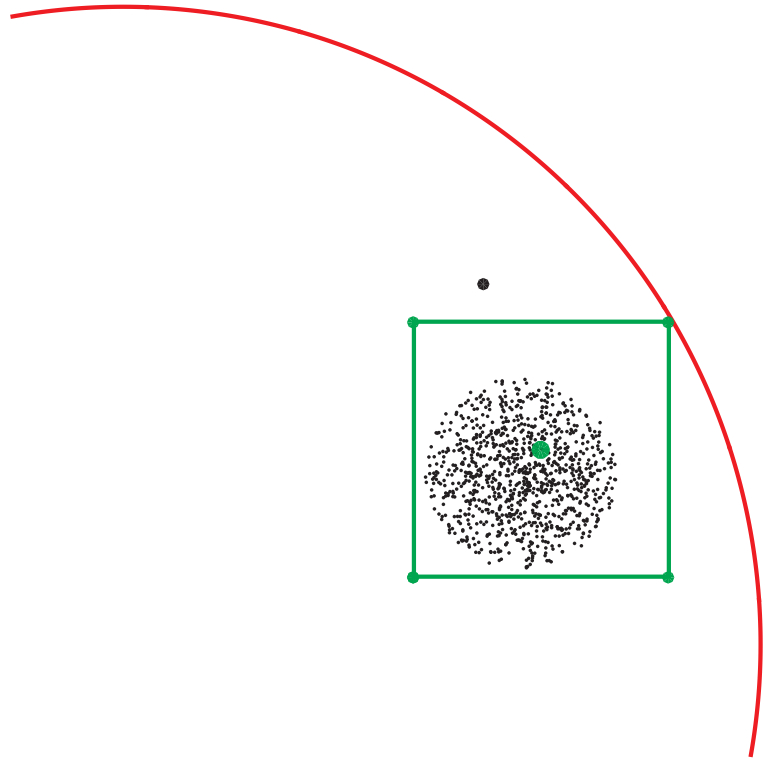
- Real system

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

- 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

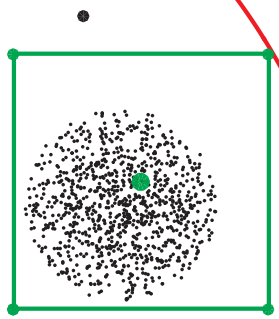
- Real system

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

- 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 $\|\mathbf{x}\|^2 \leq 1$

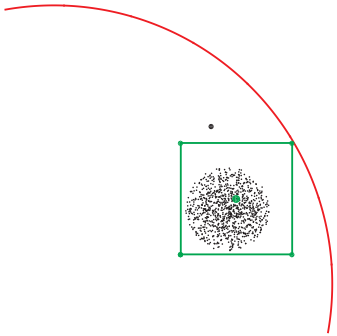


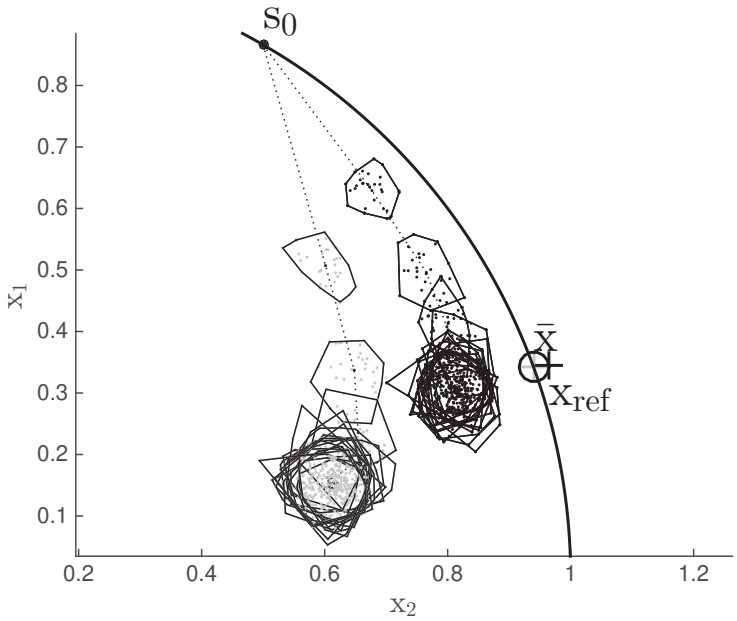
Let's adjusts:

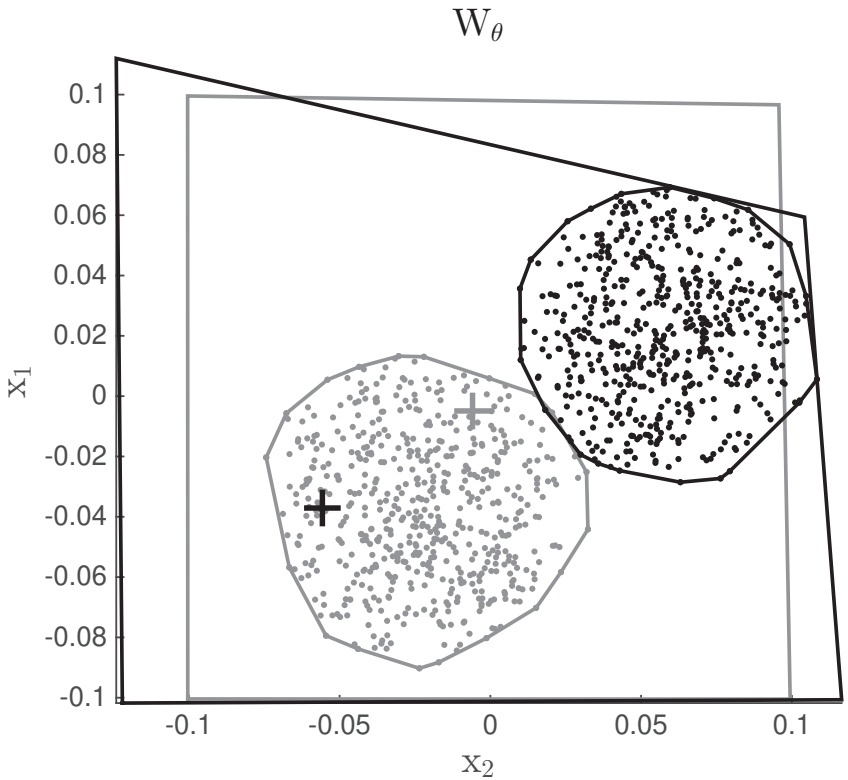
- Set $\mathbb{W}$, while containing process noise
- State and input reference in MPC cost function
- Internal linear feedback (robust MPC internal control)

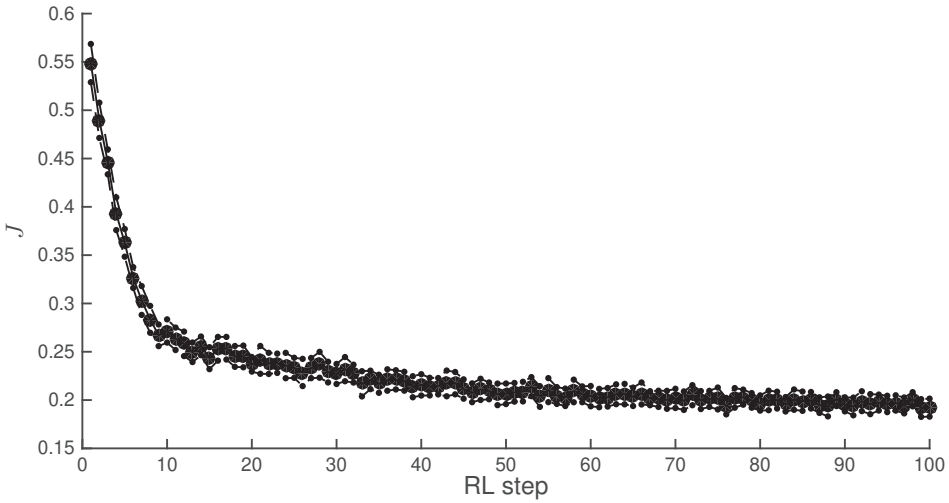
via Safe RL

Illustrative example





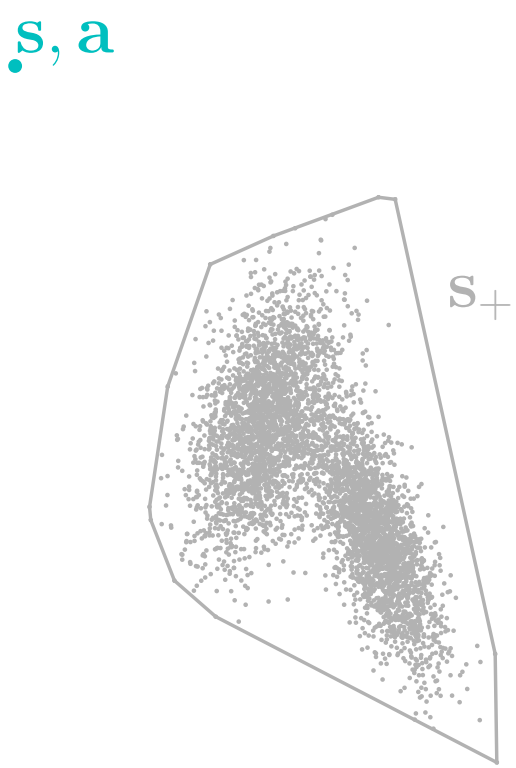




Learning

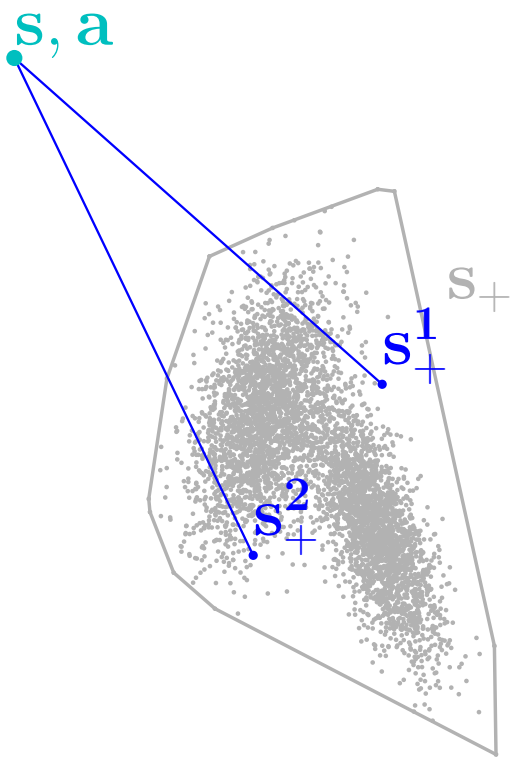
Scenario Tree MPC

Key idea: approximate distribution $\mathbb{P}[\mathbf{s}_+|\mathbf{s},\mathbf{a}]$ 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

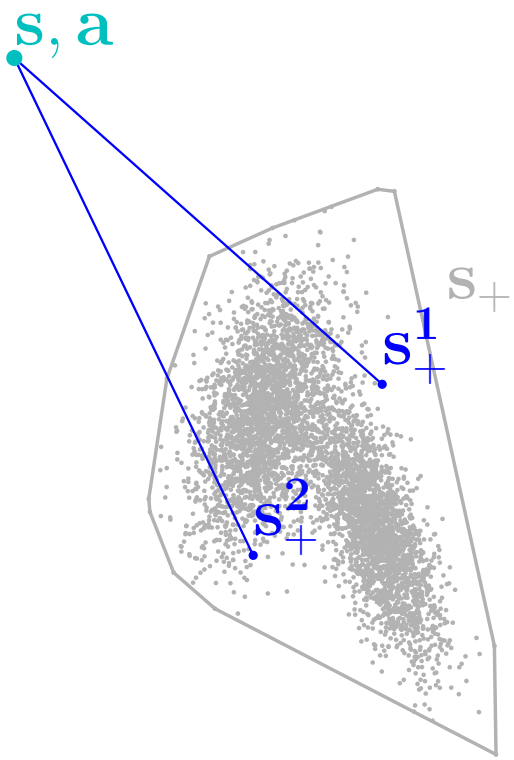
Key idea: approximate distribution $\mathbb{P}[\mathbf{s}_+|\mathbf{s},\mathbf{a}]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.



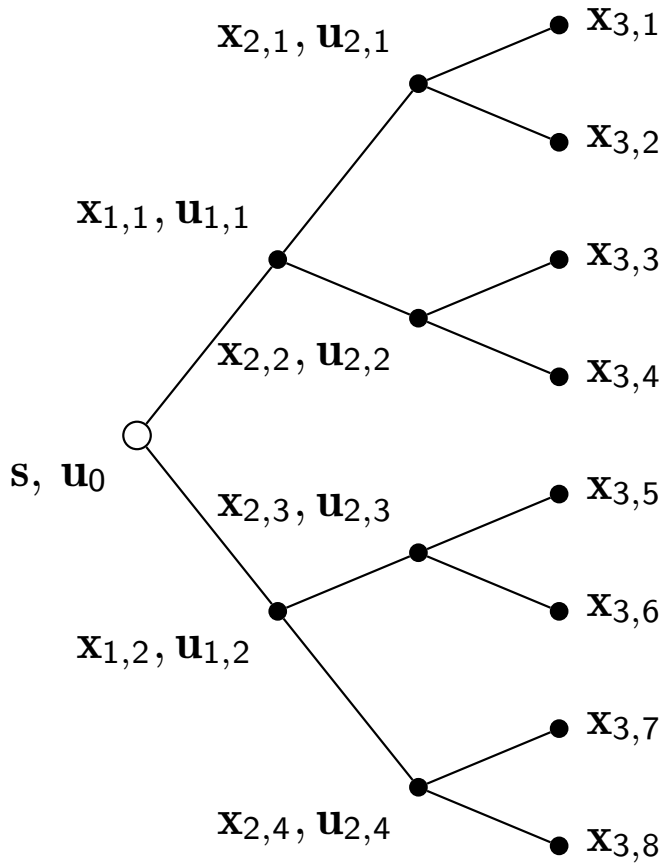
E.g. pick two scenarios

Scenario Tree MPC

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

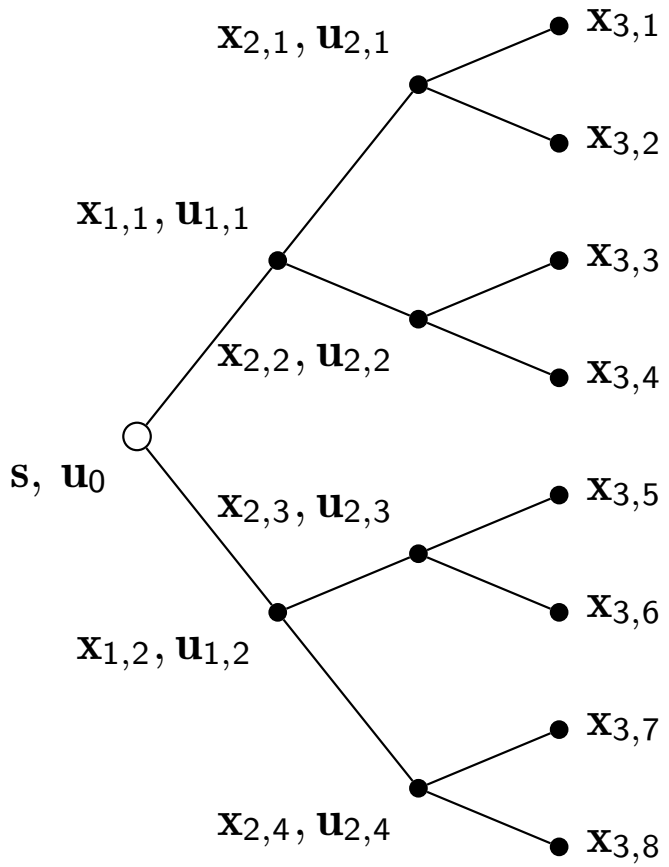


E.g. pick two scenarios



Scenario Tree MPC

Key idea: approximate distribution $\mathbb{P}[\mathbf{s}_+|\mathbf{s},\mathbf{a}]$ 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 $\pi^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

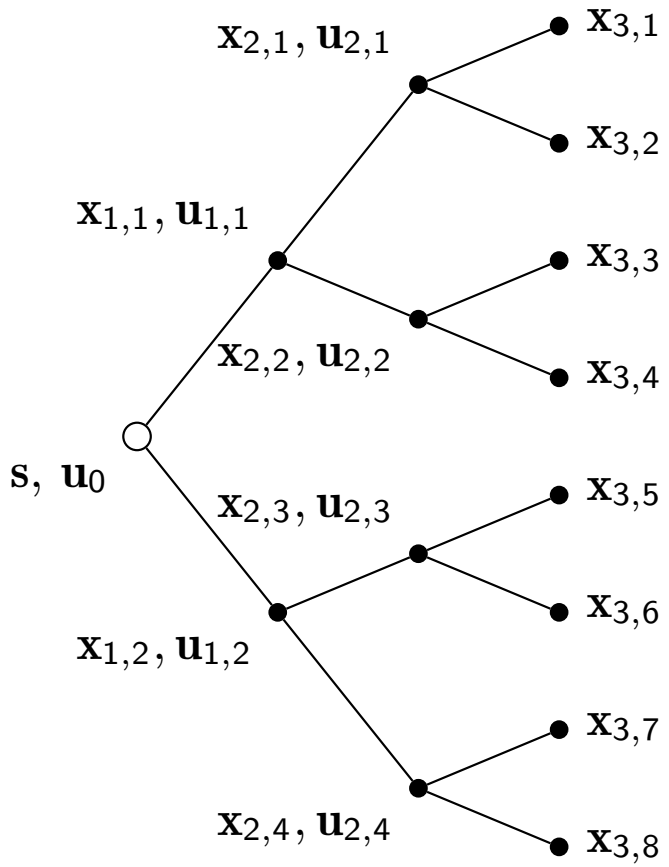
$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & \sum_{i=1}^{n}\sum_{k=0}^{N} \omega_i \, L\left(\mathbf{x}_k^i, \mathbf{u}_k^i\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1}^i = \mathbf{f}_k^i\left(\mathbf{x}_k^i, \mathbf{u}_k^i\right), \quad \mathbf{x}_0^i = \mathbf{s} \\
& \mathbf{h}_i\left(\mathbf{x}_k^i,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

where we need to select

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

Scenario Tree MPC

Key idea: approximate distribution $\mathbb{P}[\mathbf{s}_+|\mathbf{s},\mathbf{a}]$ 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 $\pi^{\mathrm{MPC}}(\mathbf{s}) = \mathbf{u}_0^\star$ from

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

where we need to select

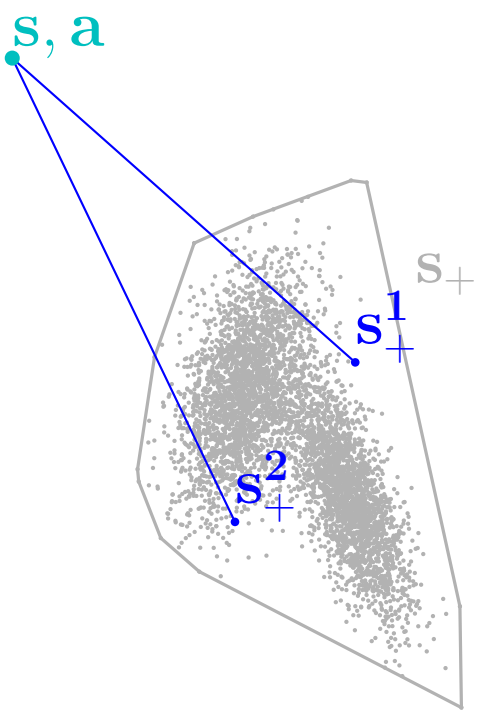
- $\mathbf{f}_k^i$ to form the scenarios
- $\omega_{1,\ldots,n}$ to build $\mathbb{E}[.]$

Difficulties

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

Often shallow trees are used

Scenario-Tree MPC



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

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

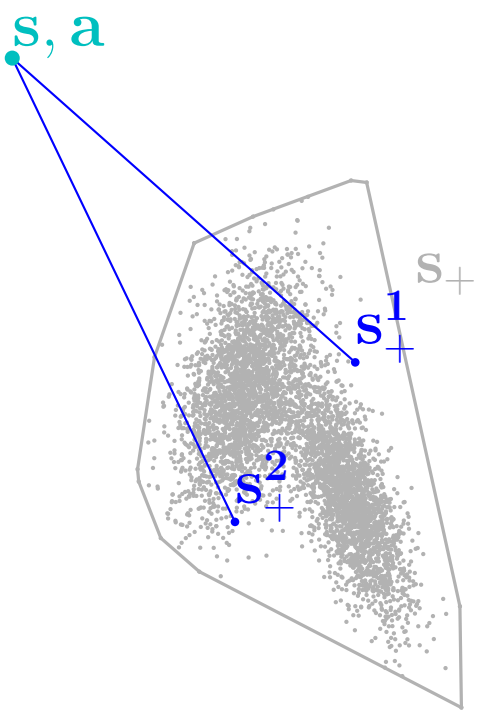
$$\mathbf{h}_i\left(\mathbf{x}_k^i, \mathbf{u}_k^i\right) \leq 0$$

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

Difficulties

- Pick "good" scenarios
- Exploding complexity

RL over Scenario Tree MPC



Scenario-Tree MPC

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

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

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

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

RL can do:

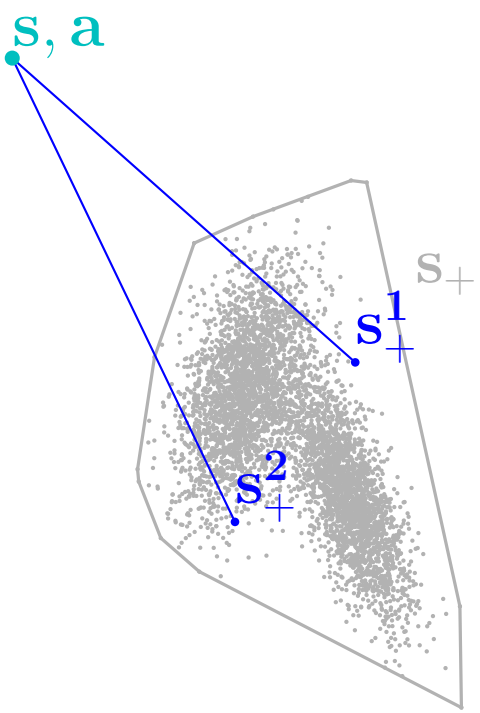
- Optimize scenarios $\mathbf{f}_k^i$
- Optimize weights ω_i
- Can work on shallow trees

for real-world performance of policy

Difficulties

- Pick "good" scenarios
- Exploding complexity

RL over Scenario Tree MPC



Scenario-Tree MPC

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

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

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

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

RL can do:

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

for real-world performance of policy

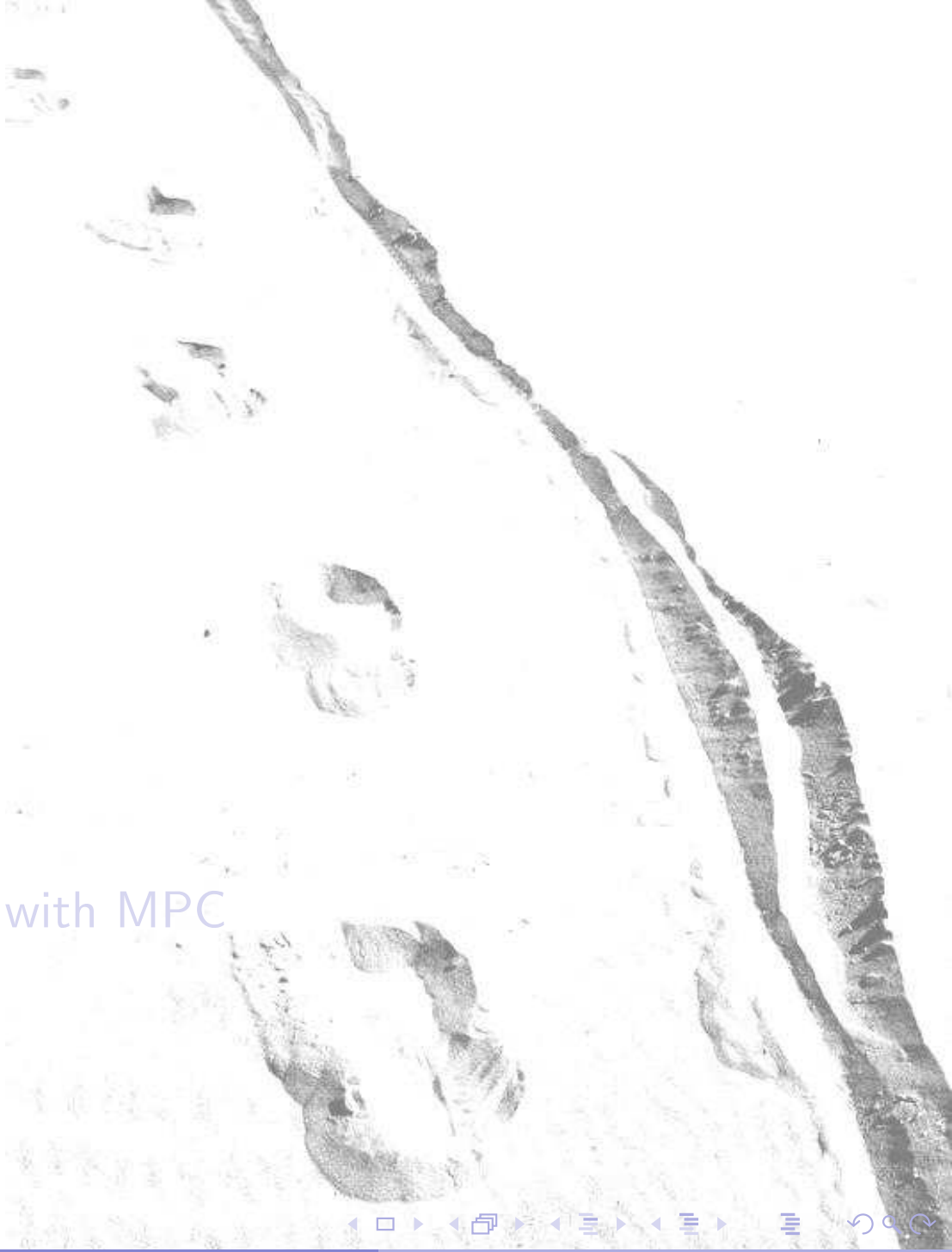
Difficulties

- Pick "good" scenarios
- Exploding complexity

Even a depth-1 tree provides a finer representation than a deterministic model. Theory applies. Conjecture: theory predicts that Monte Carlo MPC can do the job of scenario-tree MPC (to be published)

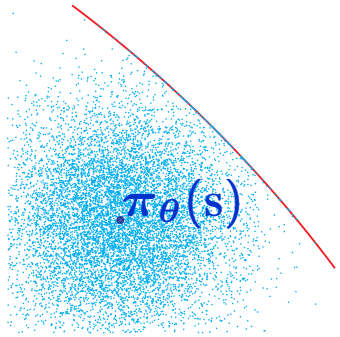
Outline

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



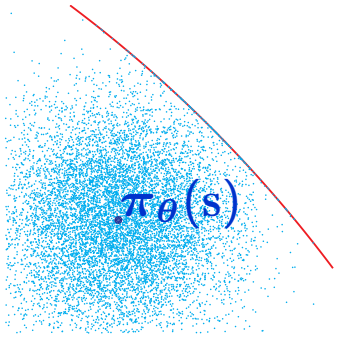
"Safe" (feasible) exploration with MPC

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



"Safe" (feasible) exploration with MPC

Learning requires exploration. E.g. apply $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a "disturbance"

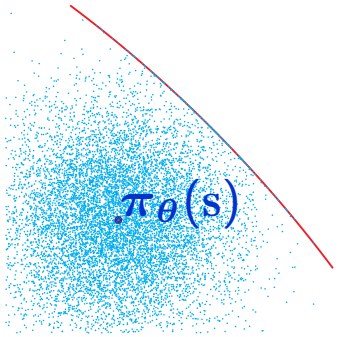


Explore while keeping feasibility?

- Clearly an arbitrary "policy disturbance" $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ 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 $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a "disturbance"



Explore while keeping feasibility?

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

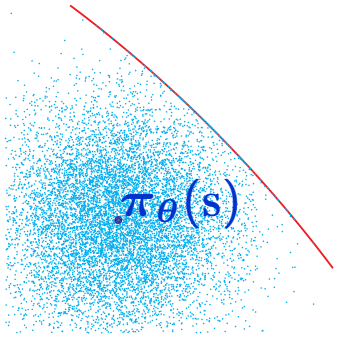
Feasible exploration: $\boldsymbol{\pi}^{\mathrm{e}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{u}_0^\star$:

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

satisfies the constraints by construction

"Safe" (feasible) exploration with MPC

Learning requires exploration. E.g. apply $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a "disturbance"



Explore while keeping feasibility?

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

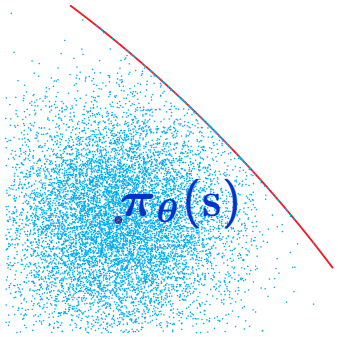
Feasible exploration: $\boldsymbol{\pi}^{\mathrm{e}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{u}_0^\star$:

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

satisfies the constraints by construction

"Safe" (feasible) exploration with MPC

Learning requires exploration. E.g. apply $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a "disturbance"



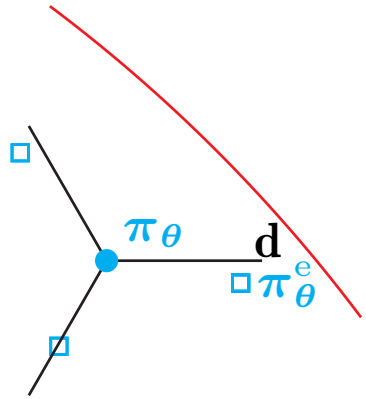
Explore while keeping feasibility?

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

Feasible exploration: $\boldsymbol{\pi}^{\mathrm{e}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{u}_0^\star$:

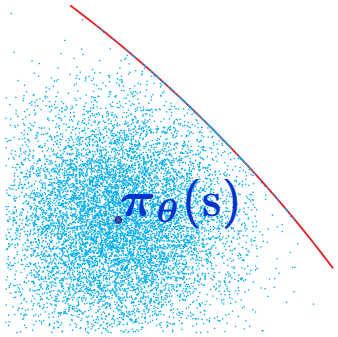
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) - \mathbf{d}^\top \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\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

satisfies the constraints by construction



"Safe" (feasible) exploration with MPC

Learning requires exploration. E.g. apply $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a "disturbance"



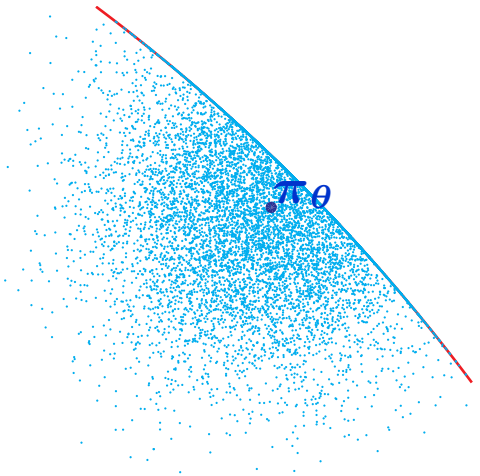
Explore while keeping feasibility?

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

Feasible exploration: $\boldsymbol{\pi}^{\mathrm{e}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{u}_0^\star$:

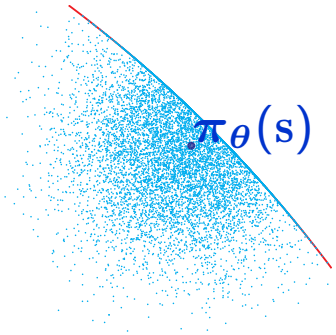
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) - \mathbf{d}^\top \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\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

satisfies the constraints by construction



"Safe" (feasible) exploration with MPC

Learning requires exploration. E.g. apply $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a "disturbance"



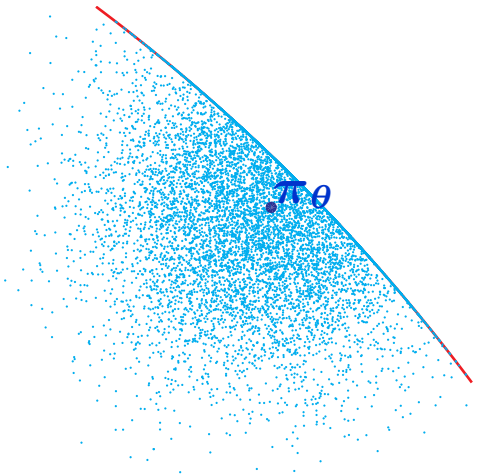
Explore while keeping feasibility?

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

Feasible exploration: $\boldsymbol{\pi}^{\mathrm{e}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{u}_0^\star$:

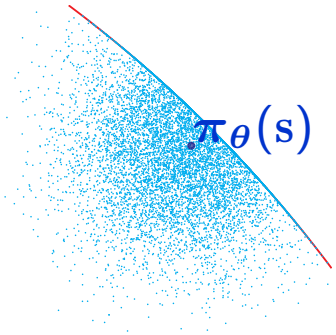
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) - \mathbf{d}^\top \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\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

satisfies the constraints by construction



"Safe" (feasible) exploration with MPC

Learning requires exploration. E.g. apply $\mathbf{a} = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}) + \mathbf{d}$ to the real system where $\mathbf{d}$ is a "disturbance"



Explore while keeping feasibility?

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

Feasible exploration: $\boldsymbol{\pi}^{\mathrm{e}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{u}_0^\star$:

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

satisfies the constraints by construction

Remarks:

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

MPC as a Stochastic Policy

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

MPC as a Stochastic Policy

- **Stochastic policy**: $\mathbf{a} \sim \pi_{\boldsymbol{\theta}}\left[\cdot \,|\, \mathbf{s}\right]$ 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**: $\mathbf{a} \sim \pi_{\boldsymbol{\theta}}\left[\cdot \,|\, \mathbf{s}\right]$ 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):

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

MPC as a Stochastic Policy

- **Stochastic policy**: $\mathbf{a} \sim \pi_{\boldsymbol{\theta}}[\cdot \,|\, \mathbf{s}]$ 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):

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

Requires:

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

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

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

MPC as a Stochastic Policy

- **Stochastic policy**: $\mathbf{a} \sim \pi_{\boldsymbol{\theta}}[\cdot \,|\, \mathbf{s}]$ 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):

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

Requires:

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

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

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

Answer:

- Yes if we, e.g., use $\pi_{\boldsymbol{\theta}}[\cdot \,|\, \mathbf{s}] \equiv \overbrace{\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})}^{\text{MPC}} + \mathbf{w}$ (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 $\pi_{\boldsymbol{\theta}}\left[\cdot|\mathbf{s}\right] \equiv \mathbf{u}_0^\star\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d}\right)$:

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

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

MPC as a Stochastic Policy

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

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

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

Remarks

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

MPC as a Stochastic Policy

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

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

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

Remarks

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

Performance

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

MPC as a Stochastic Policy

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

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

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

Remarks

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

Performance

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

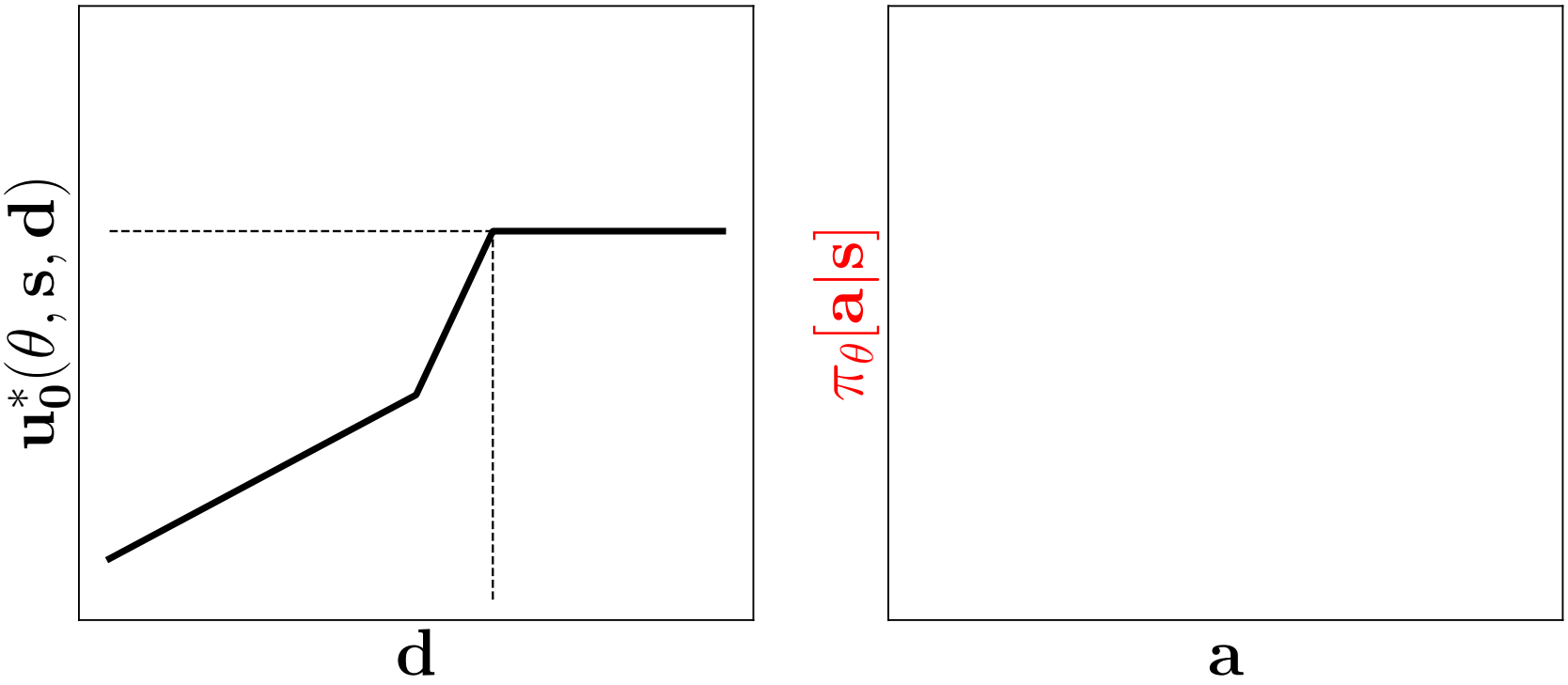
Stochastic Policy Gradient

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

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

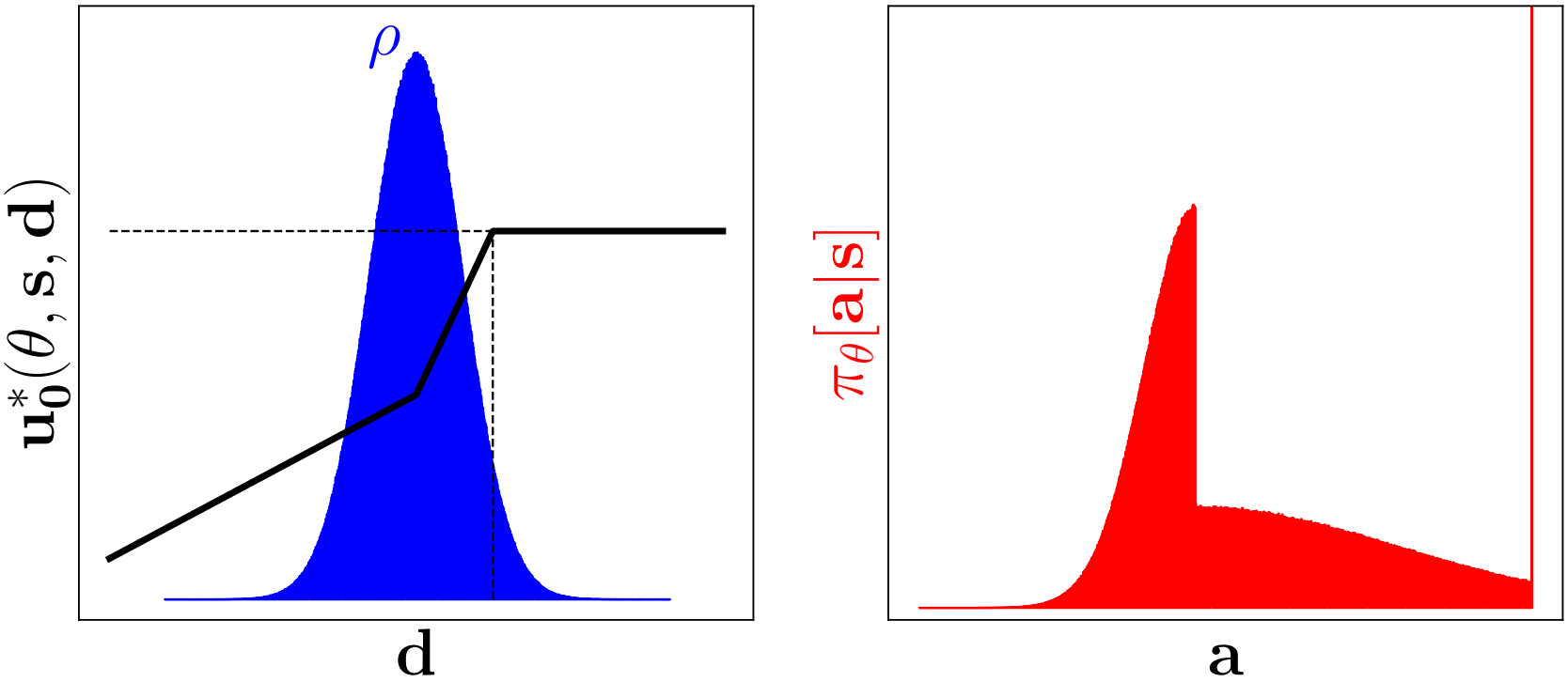
MPC as a Stochastic Policy & Policy Gradient

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



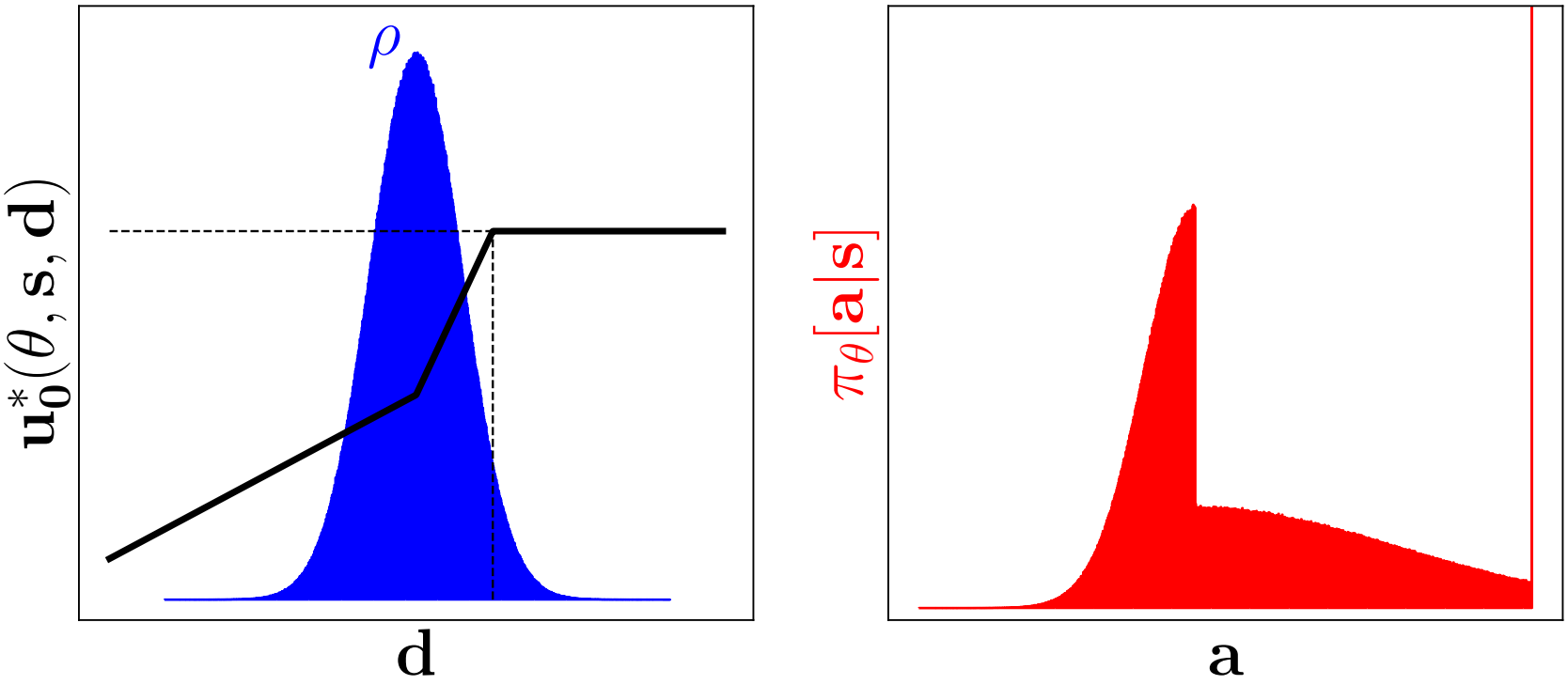
MPC as a Stochastic Policy & Policy Gradient

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



MPC as a Stochastic Policy & Policy Gradient

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

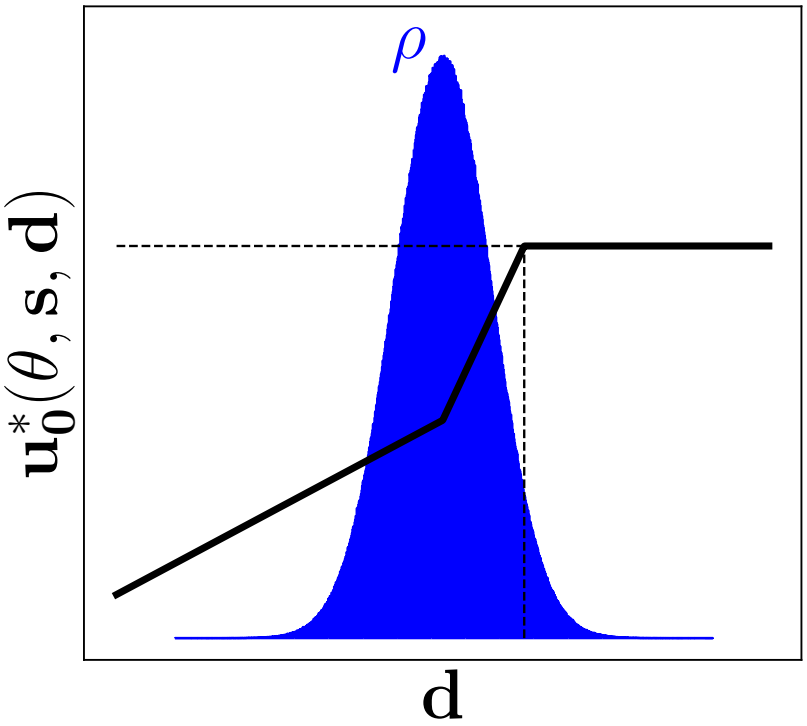


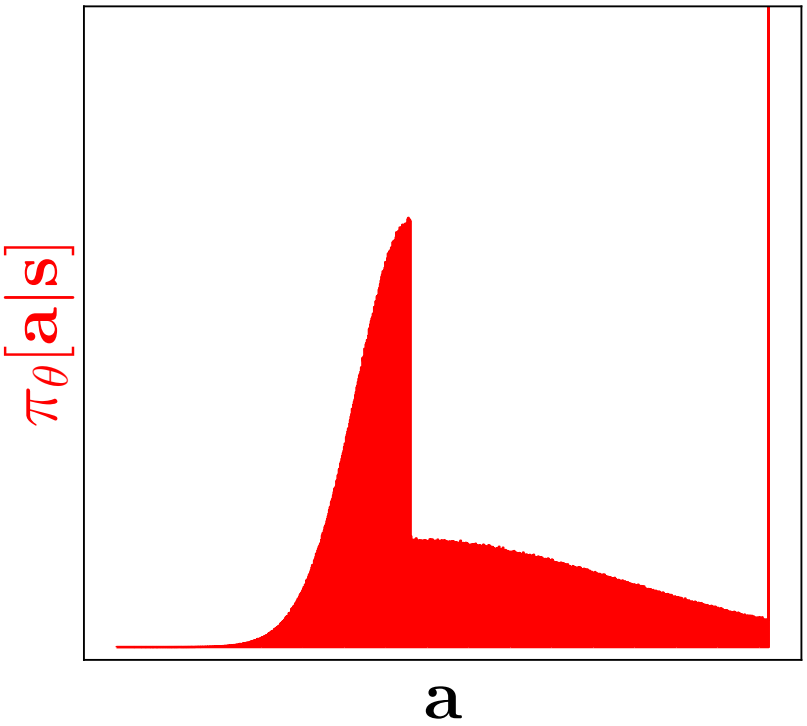
For $\mathbf{d} \to \mathbf{u}_0^\star$ bijective & differentiable

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

MPC as a Stochastic Policy & Policy Gradient

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





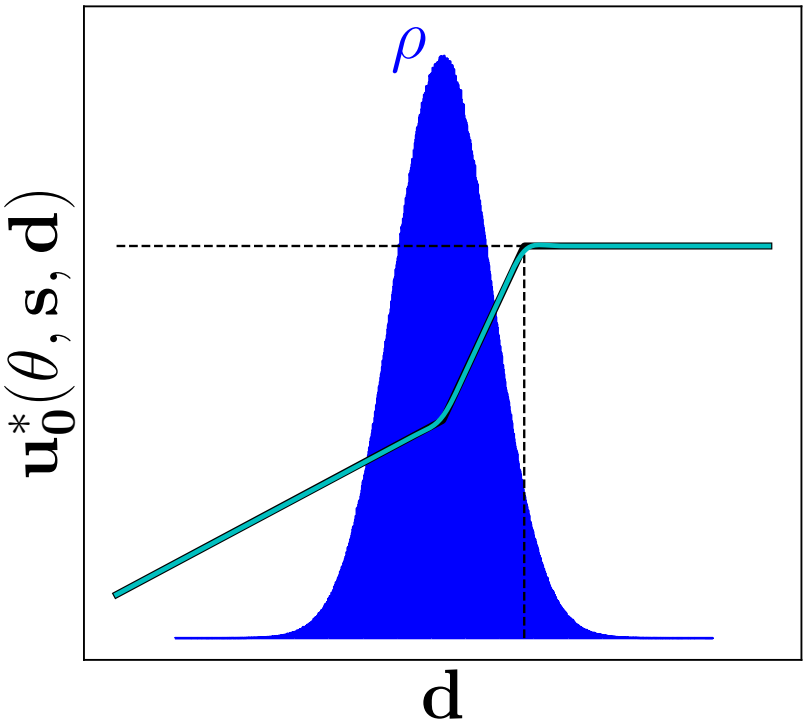
For $\mathbf{d} \to \mathbf{u}_0^\star$ bijective & differentiable

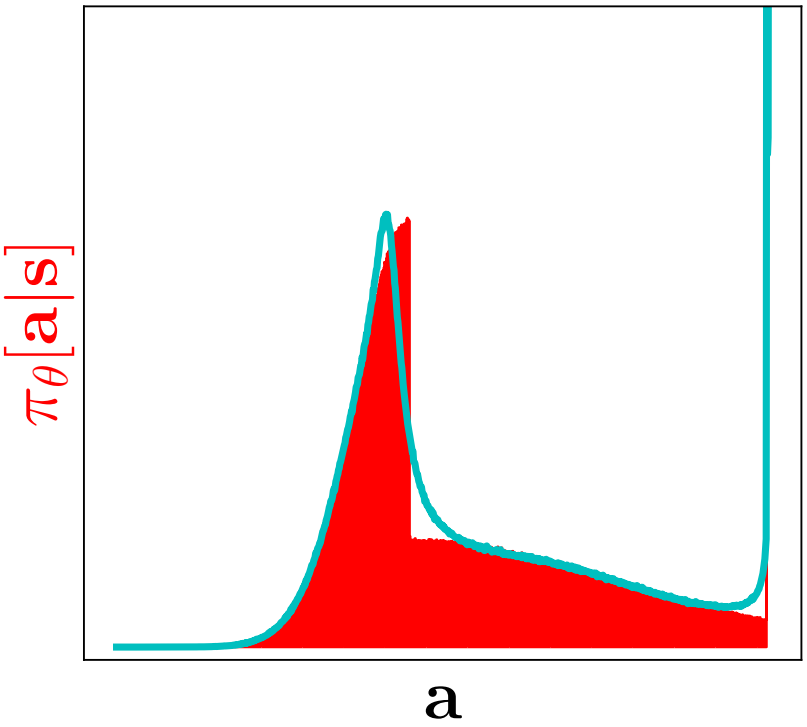
$$\pi_{\boldsymbol{\theta}}\left[\mathbf{a}|\mathbf{s}\right] = \left|\det\left(\nabla_{\mathbf{d}}\mathbf{u}_0^\star\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d}\right)\right)\right|^{-1} \rho\left[\mathbf{d}\right]$$

- 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





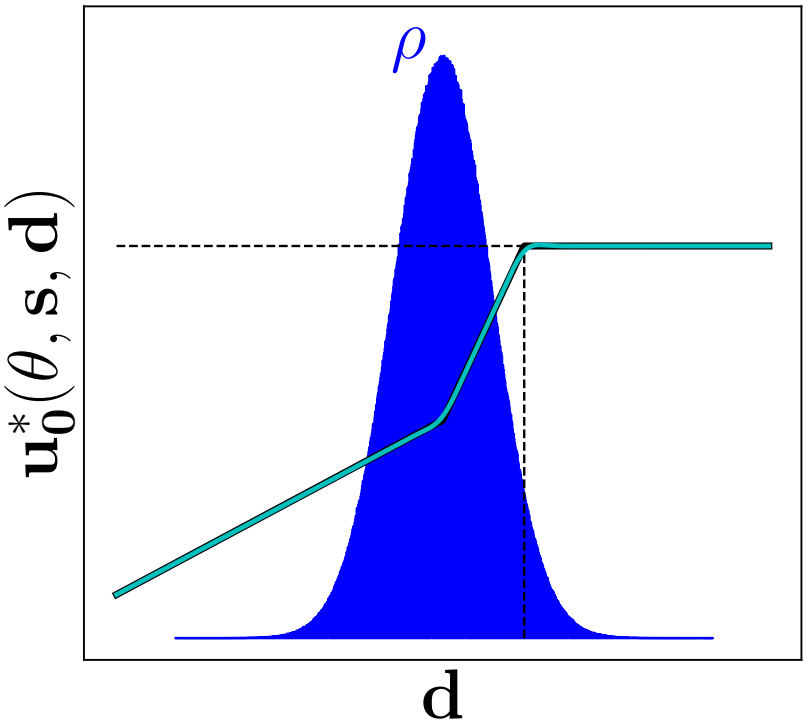
For $\mathbf{d} \to \mathbf{u}_0^\star$ bijective & differentiable

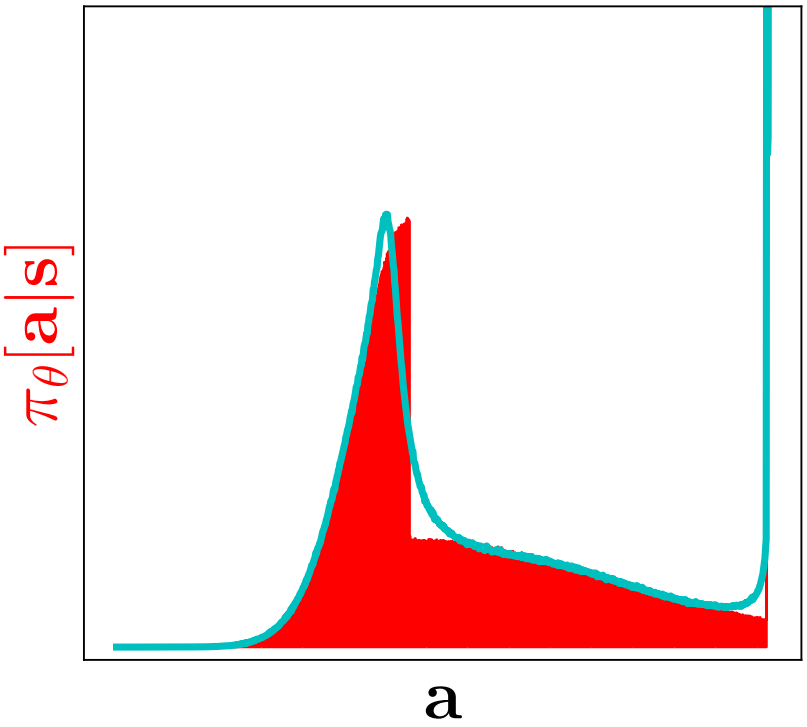
$$\pi_{\boldsymbol{\theta}}\left[\mathbf{a}|\mathbf{s}\right] = \left|\det\left(\nabla_{\mathbf{d}}\mathbf{u}_0^\star\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d}\right)\right)\right|^{-1} \rho\left[\mathbf{d}\right]$$

- 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 $\mathbf{d} \to \mathbf{u}_0^\star$ bijective & differentiable

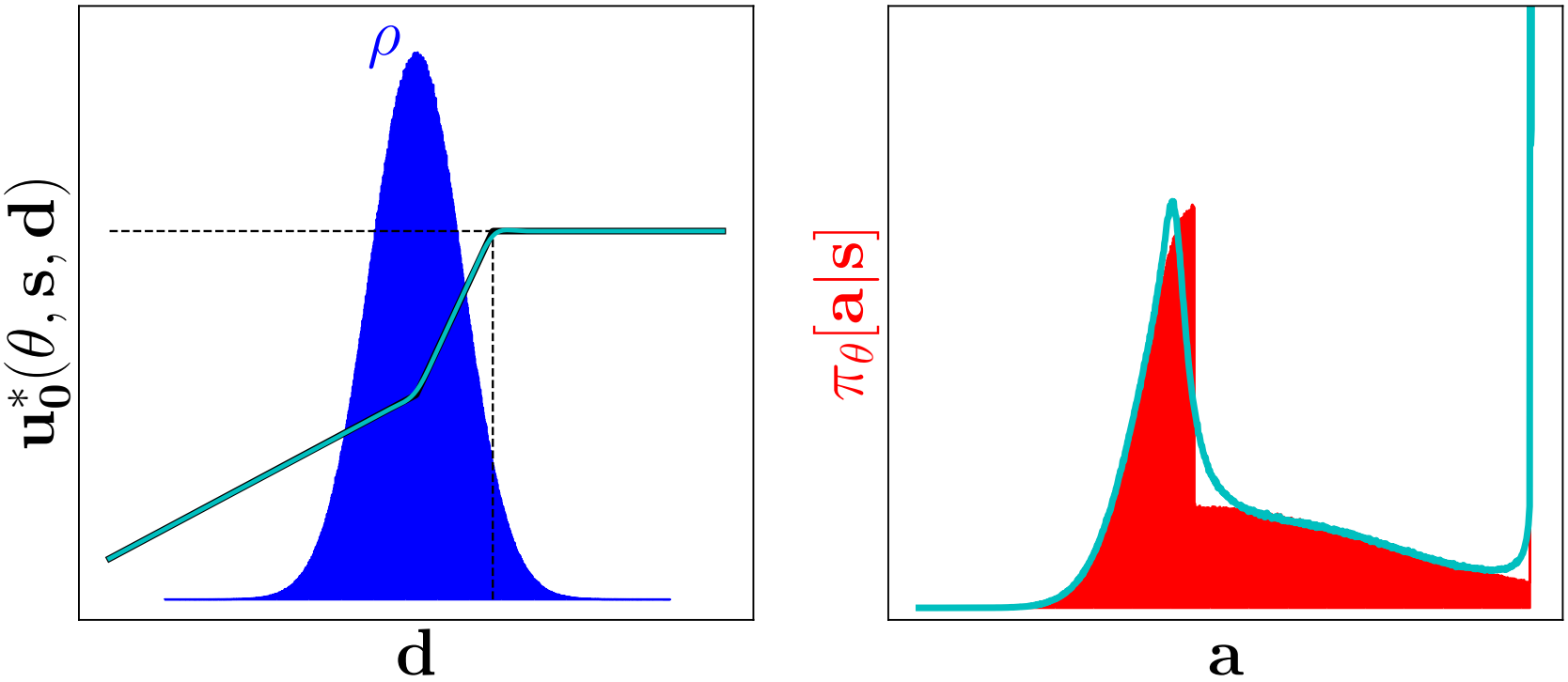
$$\pi_{\boldsymbol{\theta}}[\mathbf{a}|\mathbf{s}] = \left|\det\left(\nabla_{\mathbf{d}} \mathbf{u}_0^\star(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d})\right)\right|^{-1} \rho[\mathbf{d}]$$

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

MPC as a Stochastic Policy & Policy Gradient

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

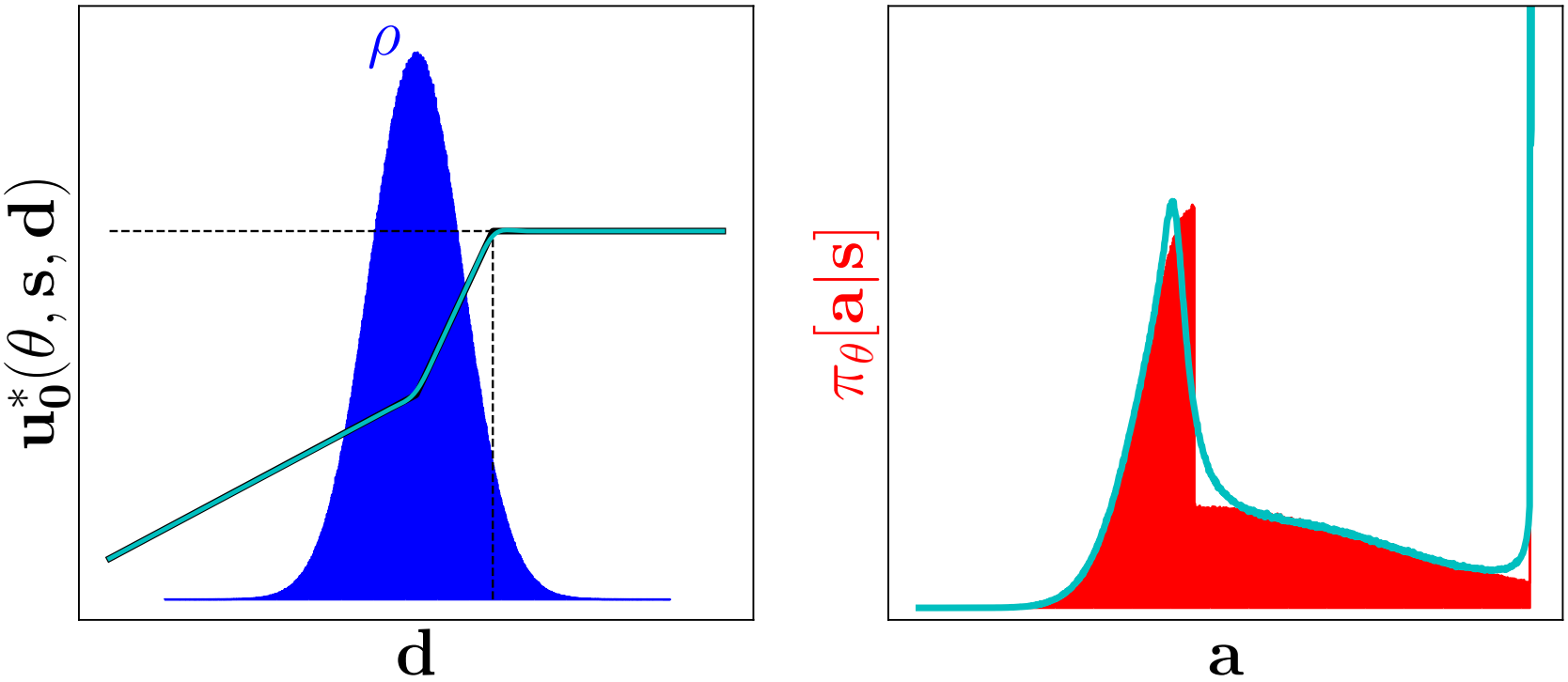


Reparametrization trick gives (smoothing is still *theoretically* good...)

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\pi}_{\boldsymbol{\theta}}) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}} \cdot Q^{\pi_{\boldsymbol{\theta}}}\right] = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \mathbf{u}_0^\star(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d}) \cdot \nabla_{\mathbf{a}} Q^{\pi_{\boldsymbol{\theta}}}(\mathbf{s}, \mathbf{a})\right]$$

MPC as a Stochastic Policy & Policy Gradient

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



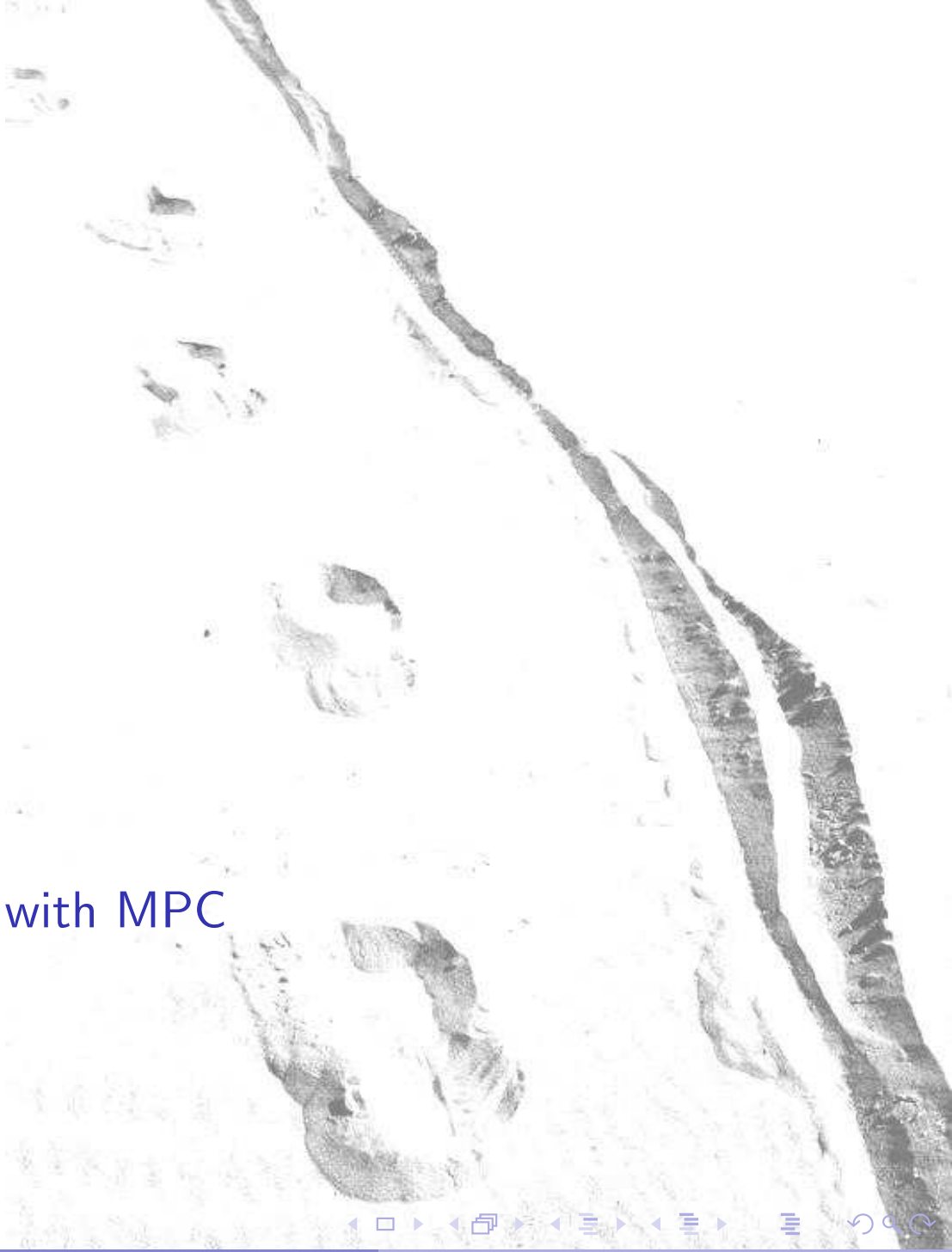
Reparametrization trick gives (smoothing is still *theoretically* good...)

$$\nabla_{\boldsymbol{\theta}} J\left(\boldsymbol{\pi}_{\boldsymbol{\theta}}\right) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}} \cdot Q^{\pi_{\boldsymbol{\theta}}}\right] = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \mathbf{u}_0^\star\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d}\right) \cdot \nabla_{\mathbf{a}} Q^{\pi_{\boldsymbol{\theta}}}\left(\mathbf{s}, \mathbf{a}\right)\right]$$

Open questions on estimation variance w.r.t. sharp variations in Q^{π_θ} and π_θ

Outline

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



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"):

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

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

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

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

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

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

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 RLMPC

Policy π^{MPC} from

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

Stability of MPC - Connection to RLMPC

Policy π^{MPC} from

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

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

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

where κ is K_∞ ($+$conditions on T)

Stability of MPC - Connection to RLMPC

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

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

where κ is K_∞ ($+$conditions on T)

Stability of MPC - Connection to RLMPC

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 $L(\mathbf{s}, \mathbf{a})$, build a **stable policy** $\pi_{\theta}^{\mathrm{MPC}}$ minimizing:

$$J\left(\pi_{\theta}^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L(\mathbf{s}_k, \mathbf{a}_k)$$

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_{\theta}^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L(\mathbf{s}_k, \mathbf{a}_k)$$

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_{\theta}^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L\left(\mathbf{s}_k, \mathbf{a}_k\right)$$

- Learning based on L
- Impose constraint:

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

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

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_{\theta}^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L\left(\mathbf{s}_k, \mathbf{a}_k\right)$$

- Learning based on L
- Impose constraint:

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

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

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

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

Theorem: under some conditions

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

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_{\theta}^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L\left(\mathbf{s}_k, \mathbf{a}_k\right)$$

- Learning based on L
- Impose constraint:

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

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

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

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework:
stability analysis $\to$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_\theta^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L(\mathbf{s}_k, \mathbf{a}_k)$$

Constraint

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

is semi-infinite programming, not trivial

Some solutions:

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

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

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework:
stability analysis $\to$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_\theta^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L\left(\mathbf{s}_k, \mathbf{a}_k\right)$$

Note that λ_θ is redundant for policy gradient, needed for Q-learning...

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

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework:
stability analysis $\to$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_\theta^{\text{MPC}}\right) = \sum_{k=0}^{\infty} L(\mathbf{s}_k, \mathbf{a}_k)$$

Extension to stable policy for MDPs?

- Need stability with discount
- Need "stochastic dissipativity"

Parametrized policy π_θ^{MPC} from:

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework:
stability analysis $\to$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_\theta^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L(\mathbf{s}_k, \mathbf{a}_k)$$

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 $\pi_\theta^{\mathrm{MPC}}$ from:

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework:
stability analysis $\to$ **stable design**

Stability-constrained Learning-based MPC - Deterministic case

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

$$J\left(\pi_\theta^{\mathrm{MPC}}\right) = \sum_{k=0}^{\infty} L\left(\mathbf{s}_k, \mathbf{a}_k\right)$$

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 $\pi_\theta^{\mathrm{MPC}}$ from:

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework:
stability analysis $\to$ **stable design**

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

RL+MPC with Future Information - An important case

Future Information in MPC

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

$$\text{and } \boldsymbol{\pi}\left(\mathbf{s}, \mathbf{I}\right) = \mathbf{u}_0^\star\left(\mathbf{s}, \mathbf{I}\right)$$

RL+MPC with Future Information - An important case

Future Information in MPC

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

$$\text{and } \boldsymbol{\pi}\left(\mathbf{s}, \mathbf{I}\right) = \mathbf{u}_0^\star\left(\mathbf{s}, \mathbf{I}\right)$$

Notes

- Assume $\mathbf{I}_{0,\ldots,\infty}$ 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

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

$$\text{and } \boldsymbol{\pi}\left(\mathbf{s}, \mathbf{I}\right) = \mathbf{u}_0^\star\left(\mathbf{s}, \mathbf{I}\right)$$

Notes

- Assume $\mathbf{I}_{0,\ldots,\infty}$ 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

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

$$\text{and } \boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{s}, \mathbf{I}\right) = \mathbf{u}_0^{\star}\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{I}\right)$$

Notes

- Assume $\mathbf{I}_{0,\ldots,\infty}$ 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

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

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

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

$$\text{and } \boldsymbol{\pi}_{\theta}\left(\mathbf{s}, \mathbf{I}\right) = \mathbf{u}_0^{\star}\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{I}\right)$$

Notes

- Assume $\mathbf{I}_{0,\ldots,\infty}$ 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
- In practice we work on a finite information horizon

RL+MPC with Future Information - An important case

Future Information in MPC

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

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

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

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

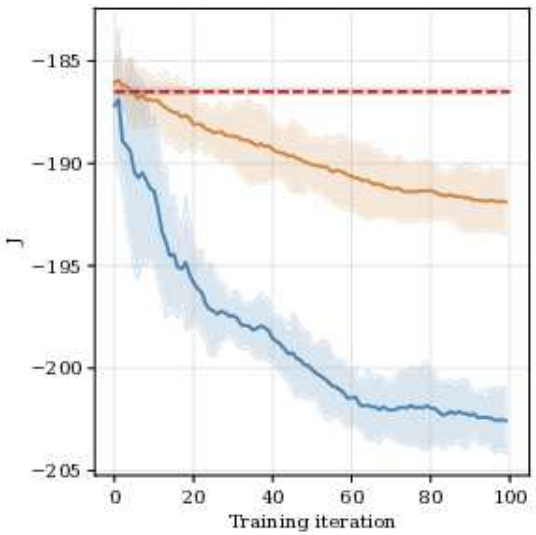
Notes

- Assume $\mathbf{I}_{0,\ldots,\infty}$ 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
- In practice we work on a finite information horizon

Example



RL+MPC with Future Information - An important case

Future Information in MPC

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

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

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

$$\text{and } \boldsymbol{\pi}_{\boldsymbol{\theta}}\left(\mathbf{s}, \mathbf{I}\right) = \mathbf{u}_0^\star\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{I}\right)$$

Notes

- Assume $\mathbf{I}_{0,\ldots,\infty}$ 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"

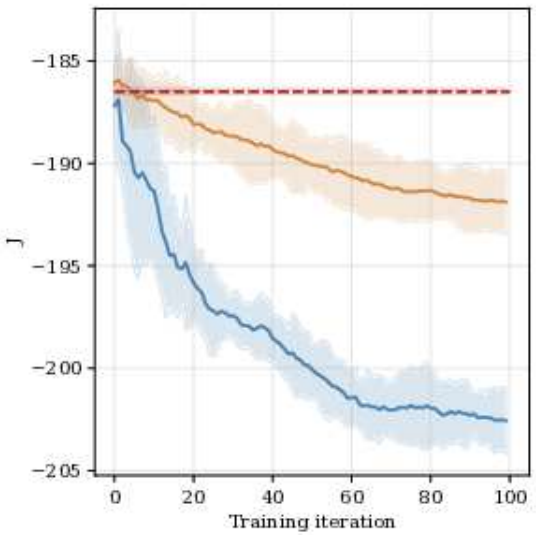
RL+MPC theory applies but:

- $\mathbf{I}$ 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
- In practice we work on a finite information horizon
- Difficulty: high-dimensional $\boldsymbol{\theta}$ and

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

requires a high-dimensional critic $Q^{\boldsymbol{\pi}_{\boldsymbol{\theta}}}$

Example



MPC as Its Own Critic

Full MPC parametrization:

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

Policy gradient

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

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

MPC as Its Own Critic

Full MPC parametrization:

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

Policy gradient

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

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

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:

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

Policy gradient

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

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

Difficulties

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

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

MPC as Its Own Critic

Full MPC parametrization:

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

Policy gradient

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

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

Difficulties

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

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

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 $Q^{\pi_\theta} : S \times A \to \mathbb{R}$

Difficulties

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

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^{π_θ} 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

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

DNN-based MPC parameters

MPC policy

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

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

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

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

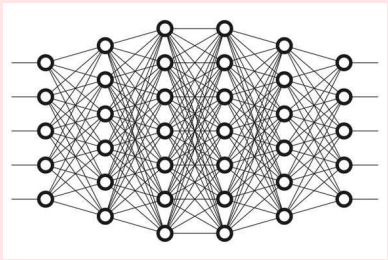


$\mathbf{a}$

$\mathbf{s}$

DNN-based MPC parameters

DNN gives MPC param



$\mathbf{s}$ + context

$\boldsymbol{\theta}$

MPC policy

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

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

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

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

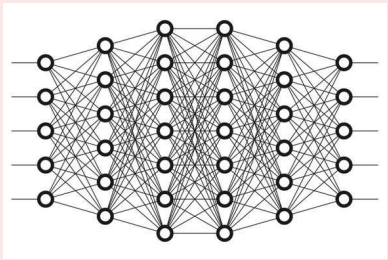
$\mathbf{a}$



$\mathbf{s}$

DNN-based MPC parameters

DNN gives MPC param



ω (DNN weights)

$\mathbf{s}$ + context

$\boldsymbol{\theta}$

MPC policy

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

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

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

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

$\mathbf{a}$



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

RL adjusts ω to achieve

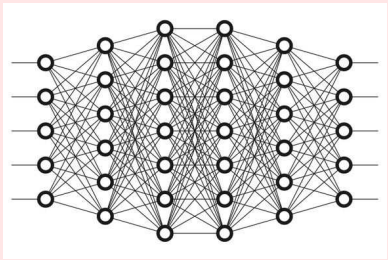
$$\nabla_\omega J(\pi_{\boldsymbol{\theta}}) = 0$$

from data

$\mathbf{s}$

DNN-based MPC parameters

DNN gives MPC param



ω (DNN weights)

$\mathbf{s}$ + context

$\boldsymbol{\theta}$

MPC policy

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

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

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

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

$\mathbf{a}$



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

RL adjusts ω to achieve

$$\nabla_{\omega} J(\pi_{\theta}) = 0$$

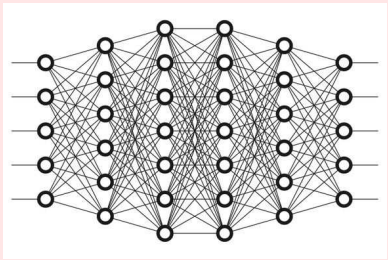
from data

Remarks: $\mathbf{s}$

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

DNN-based MPC parameters

DNN gives MPC param



ω (DNN weights)

$\mathbf{s}$ + context

$\boldsymbol{\theta}$

MPC policy

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

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

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

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

$\mathbf{a}$



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

RL adjusts ω to achieve

$$\nabla_\omega J(\pi_{\boldsymbol{\theta}}) = 0$$

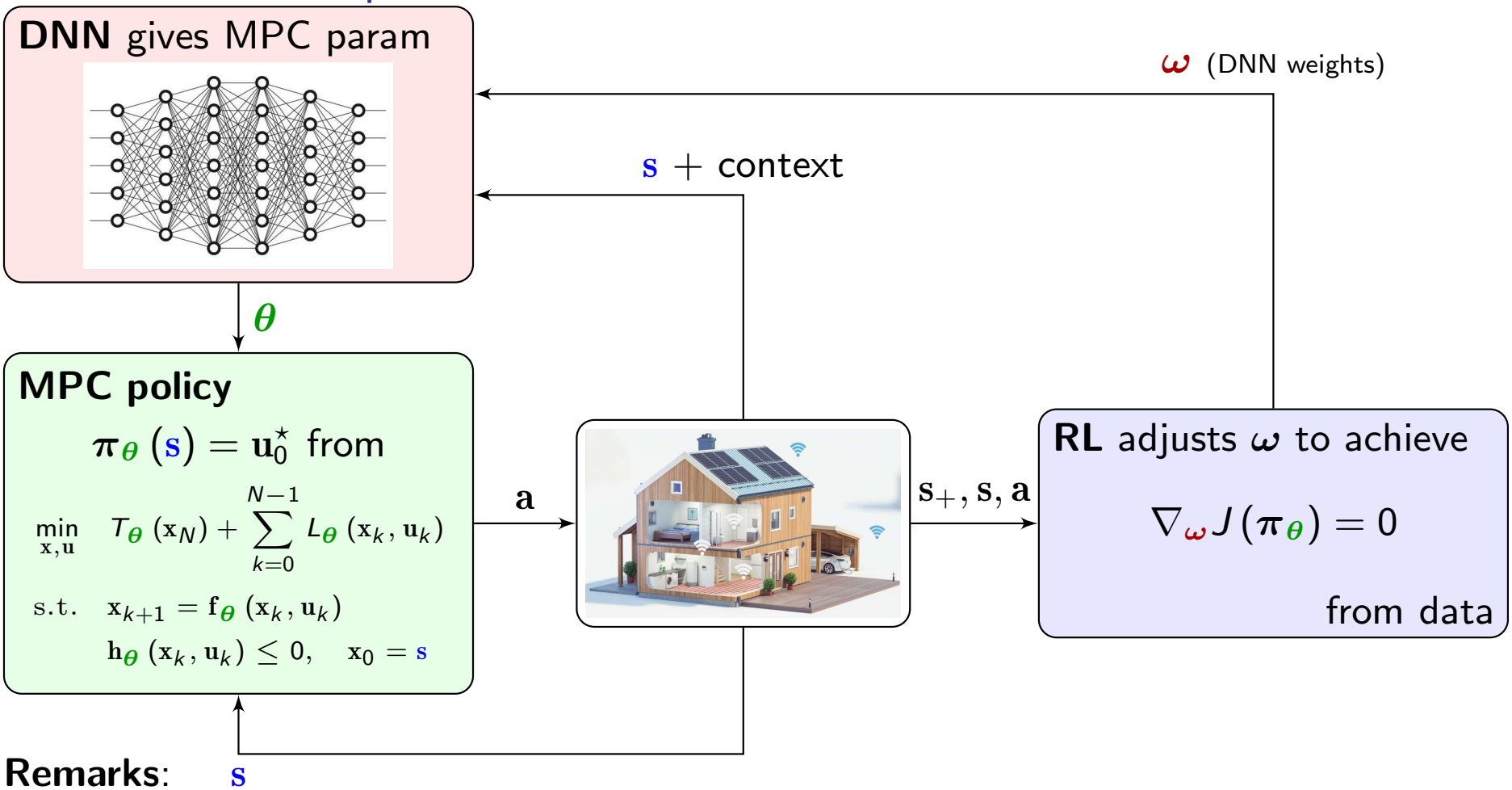
from data

$\mathbf{s}$

Remarks:

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

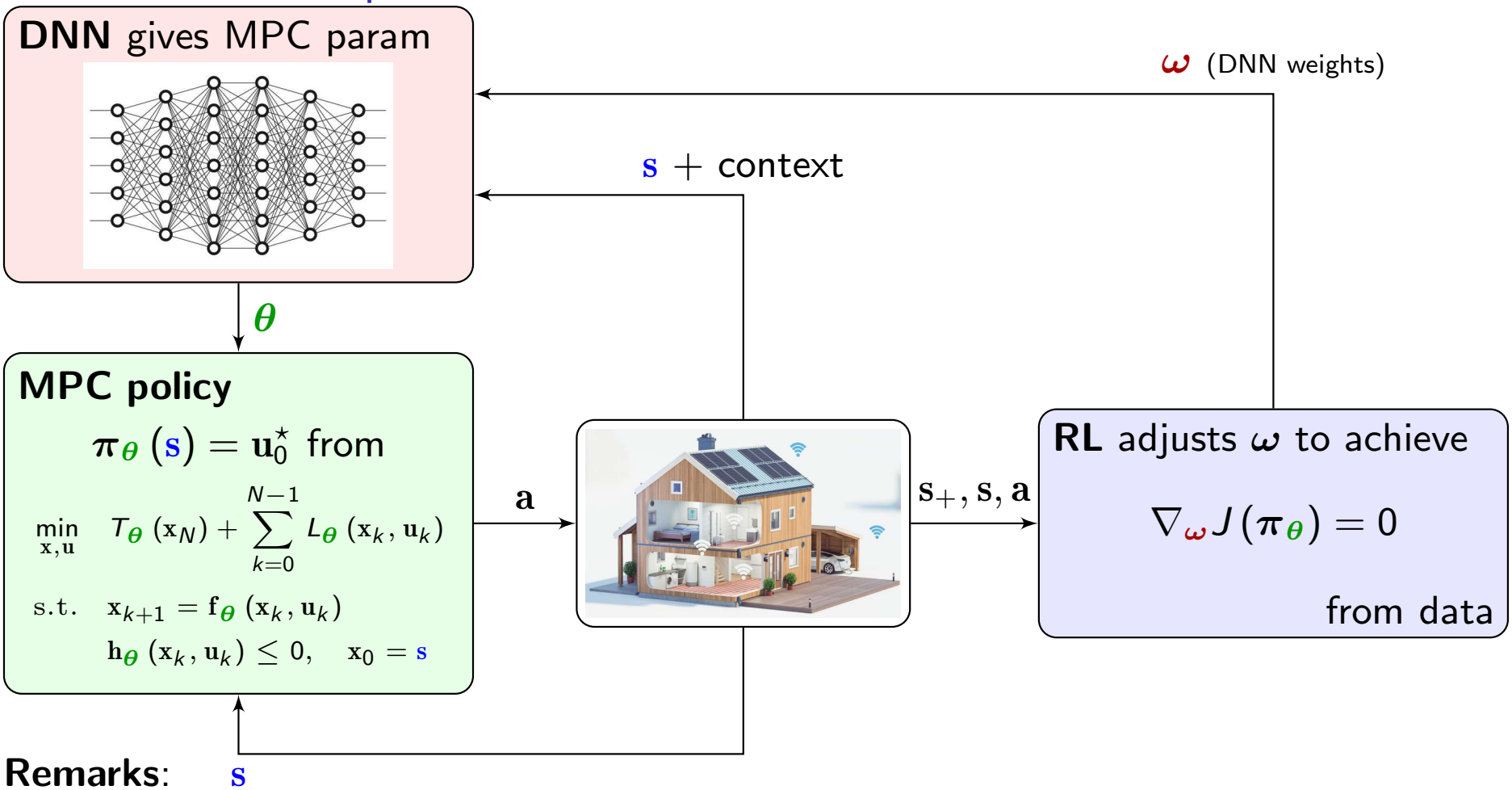
DNN-based MPC parameters



Remarks:

- Chain rule $\nabla_{\omega} \pi_{\boldsymbol{\theta}} = \nabla_{\omega} \boldsymbol{\theta} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}$ 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, $\boldsymbol{\theta}$ remains consistent?

DNN-based MPC parameters



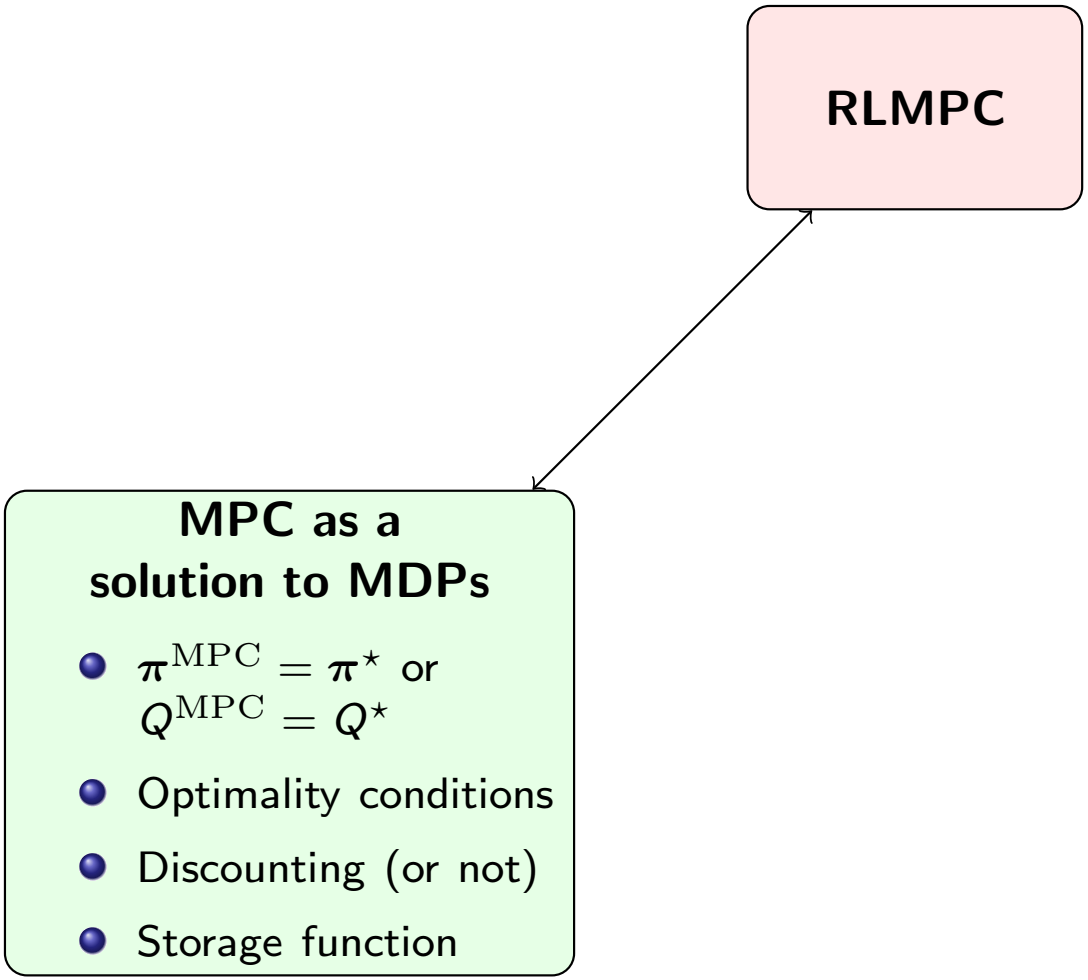
Remarks:

- Chain rule $\nabla_{\omega} \pi_{\boldsymbol{\theta}} = \nabla_{\omega} \boldsymbol{\theta} \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}$ 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, $\boldsymbol{\theta}$ remains consistent?
- Need a theory to explain what that means, how to do it, and why it is useful(?)

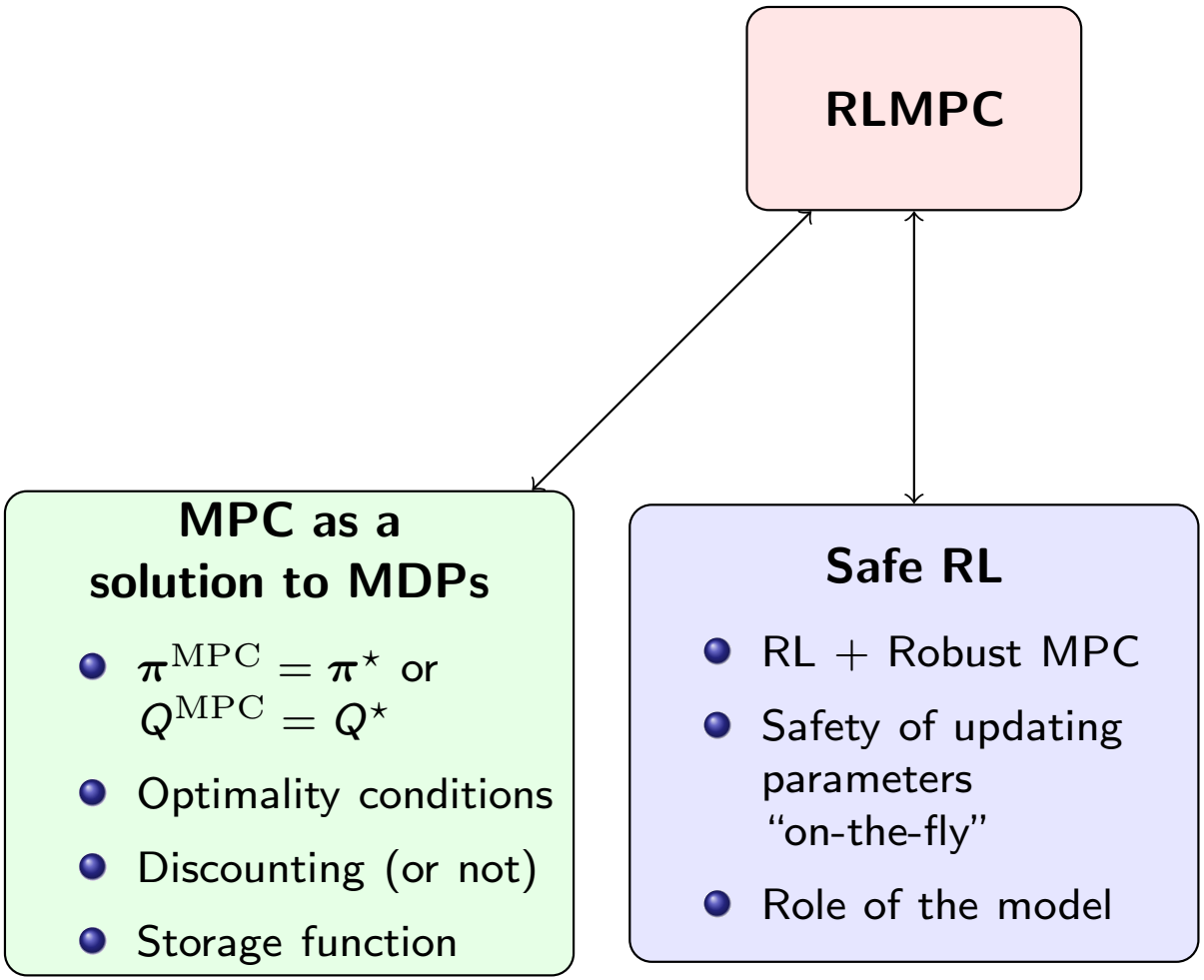
More explored questions in brief - Fundamentals

RLMPC

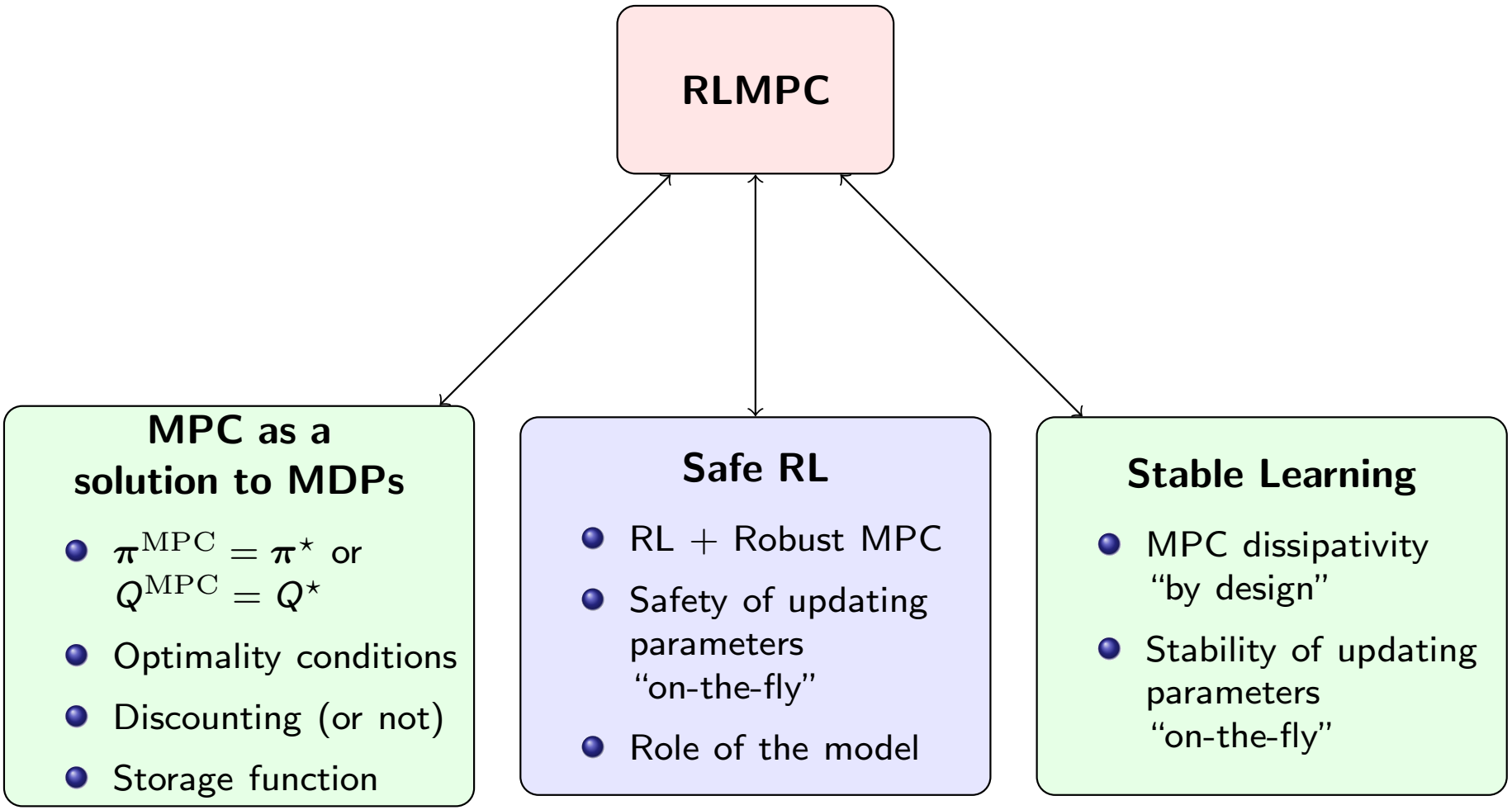
More explored questions in brief - Fundamentals



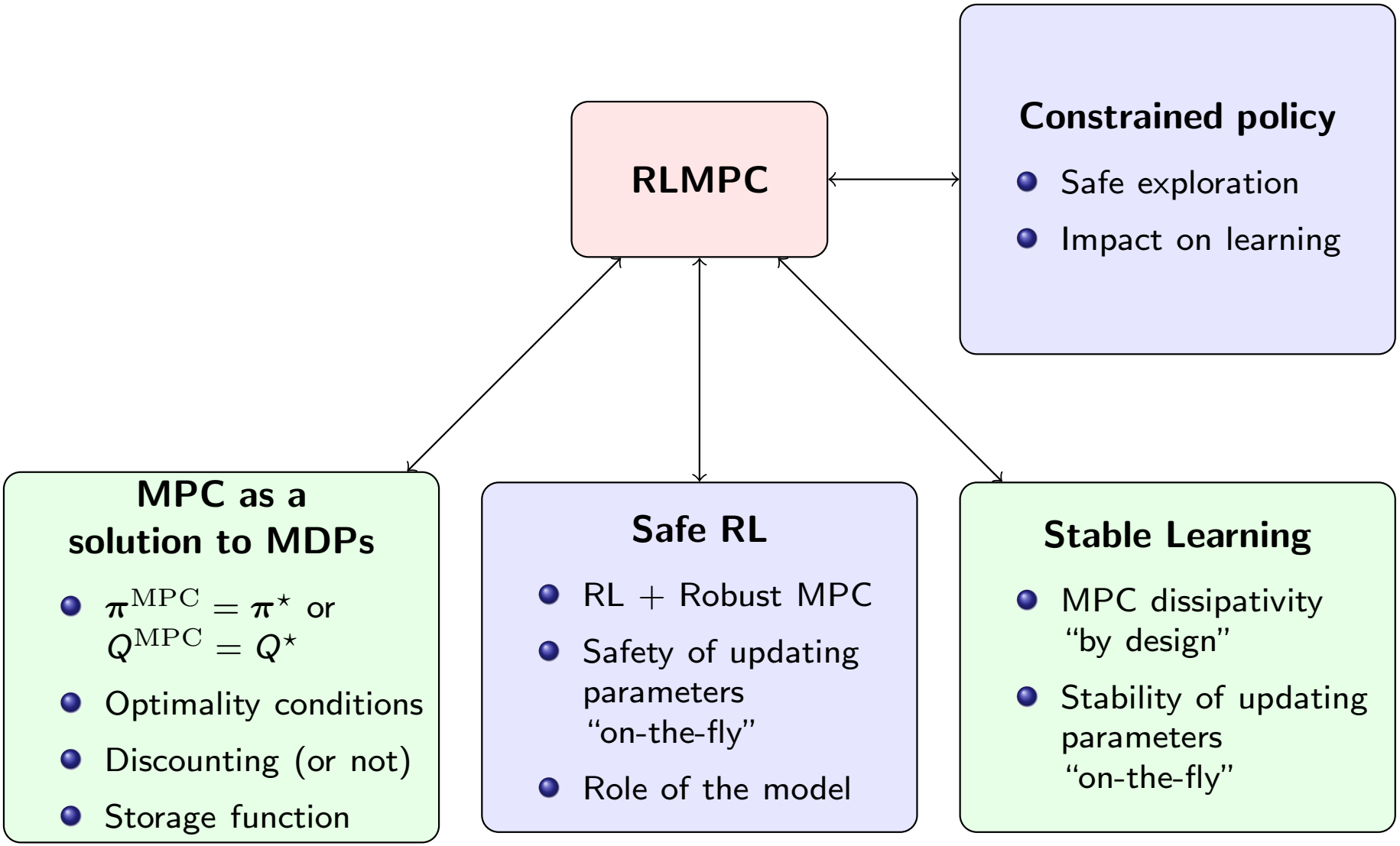
More explored questions in brief - Fundamentals



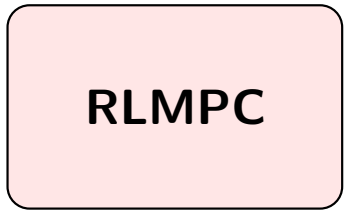
More explored questions in brief - Fundamentals



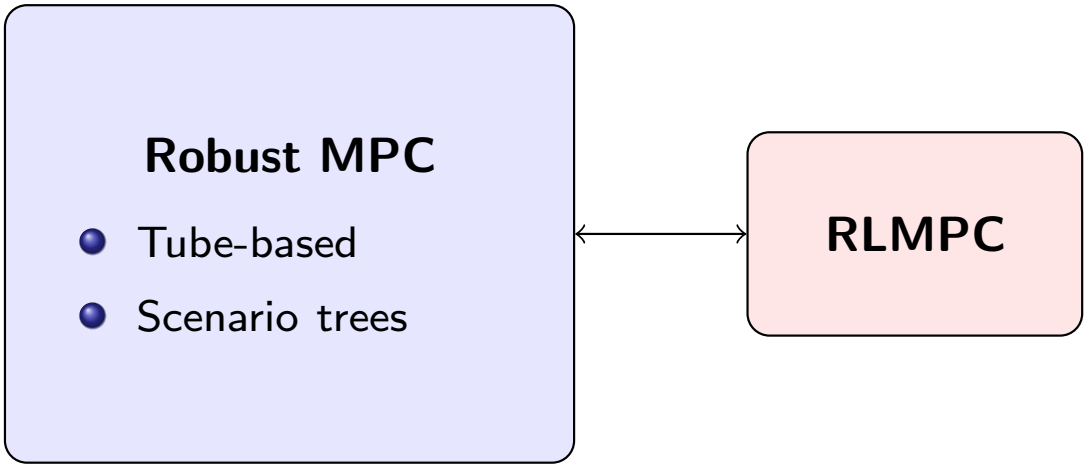
More explored questions in brief - Fundamentals



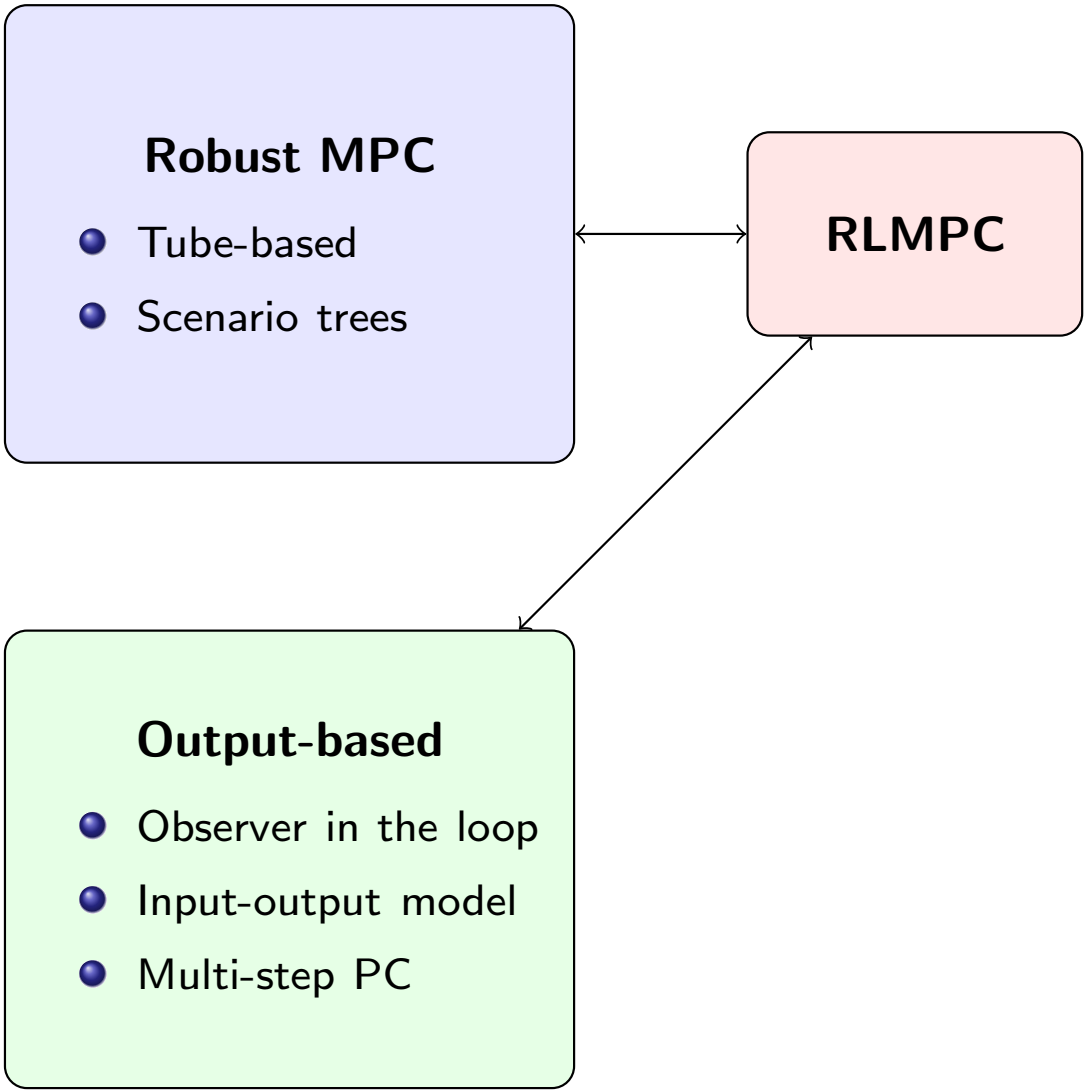
More explored questions in brief - Types of MPC & Learning



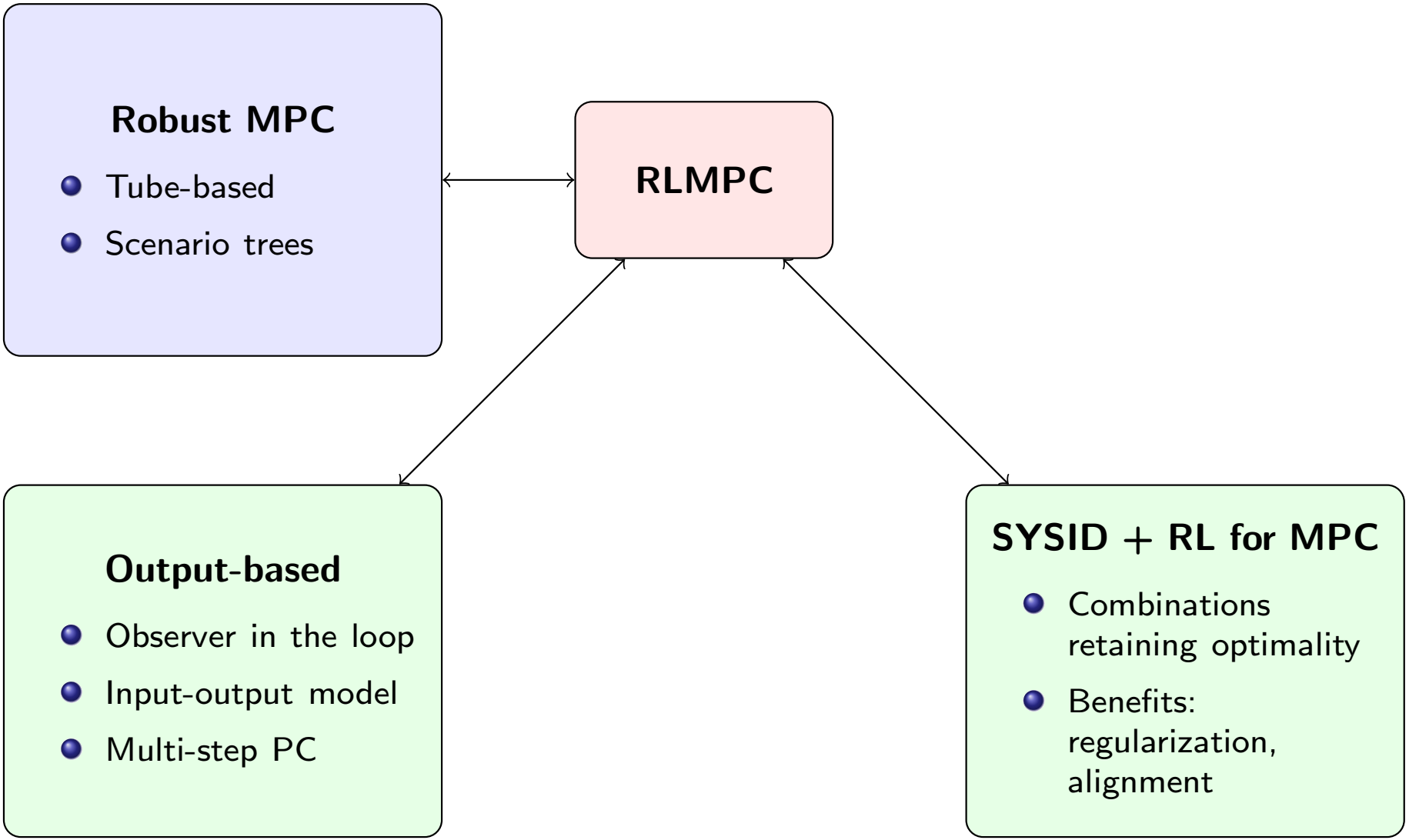
More explored questions in brief - Types of MPC & Learning



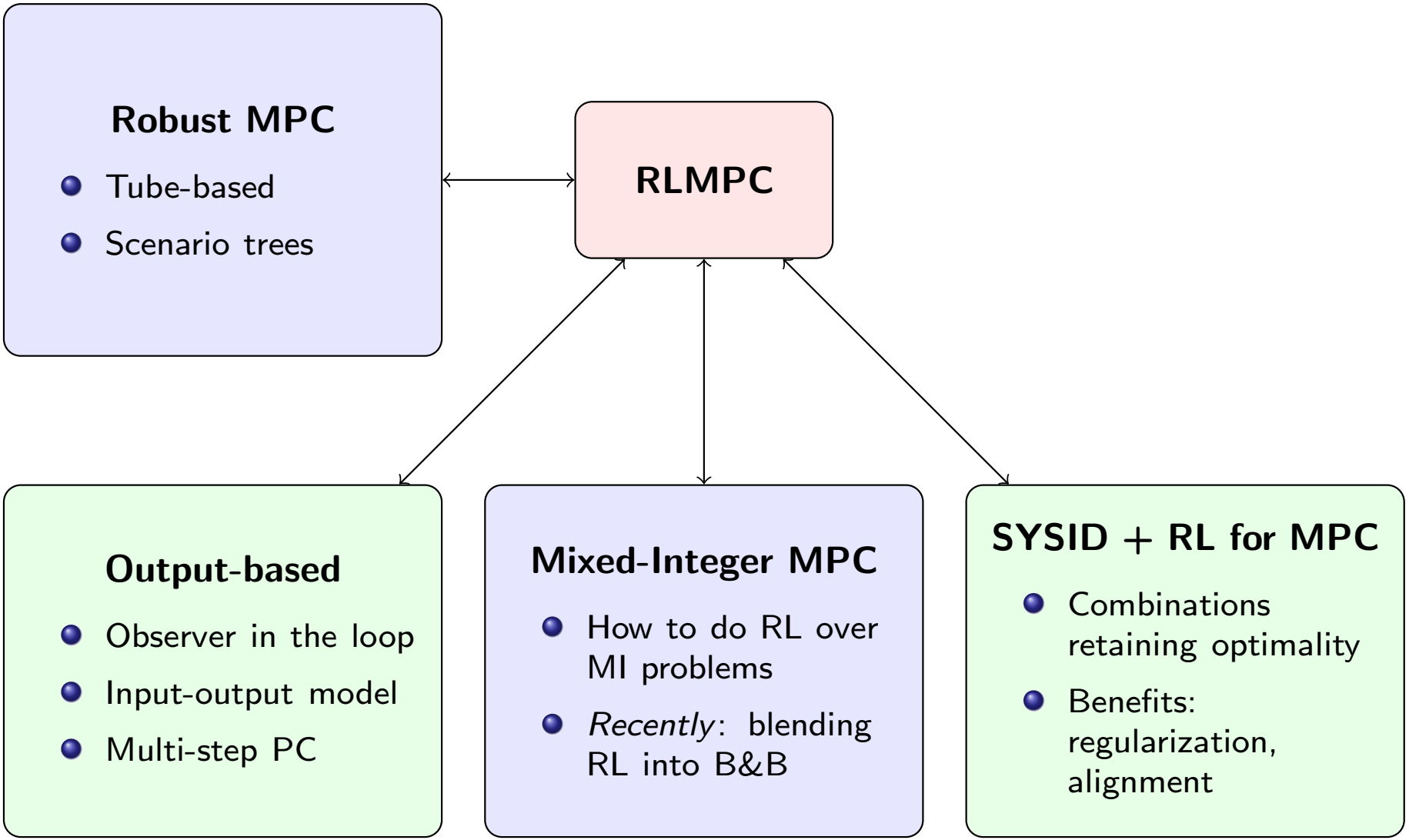
More explored questions in brief - Types of MPC & Learning



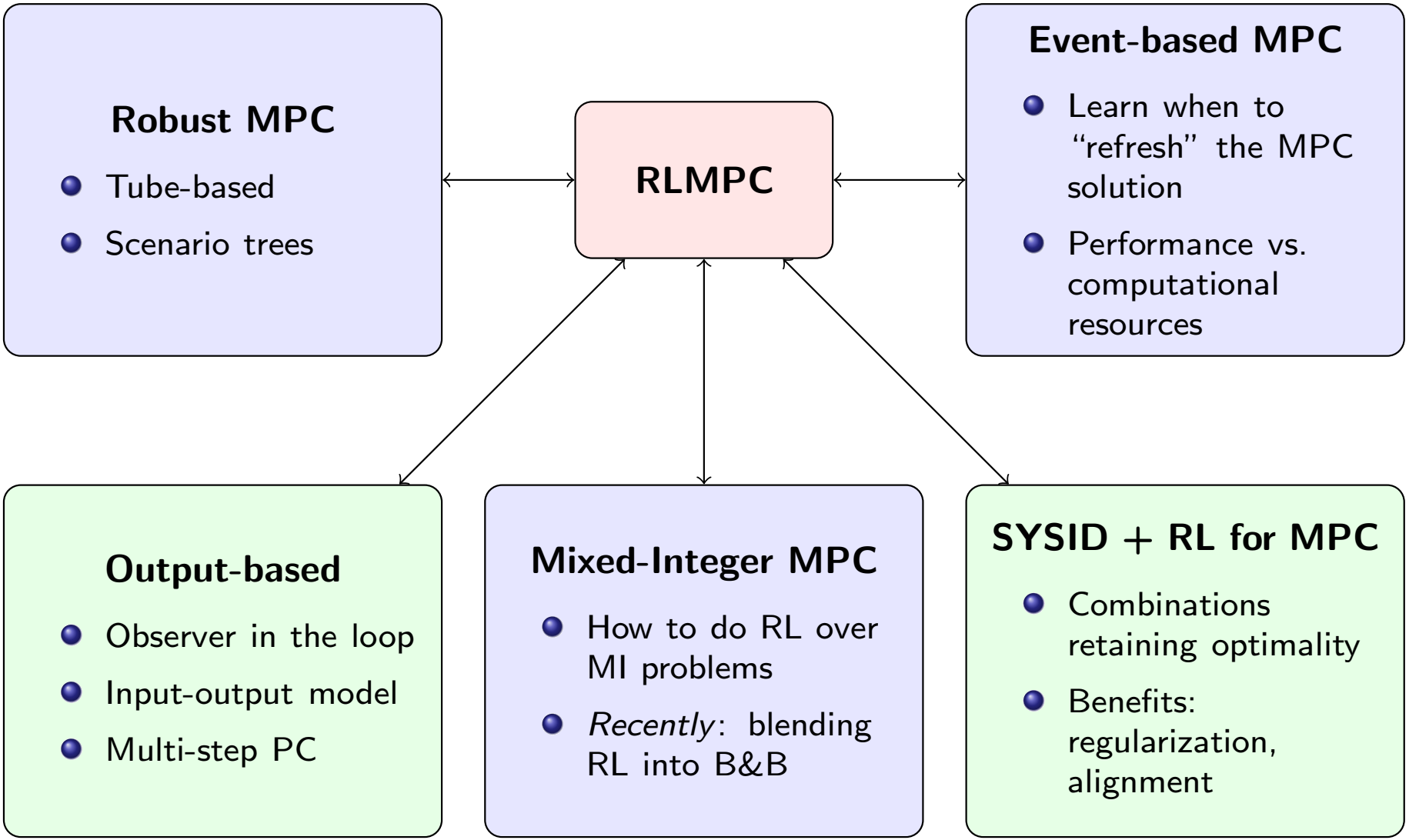
More explored questions in brief - Types of MPC & Learning



More explored questions in brief - Types of MPC & Learning



More explored questions in brief - Types of MPC & Learning



More explored questions in brief - Algorithms

RLMPC

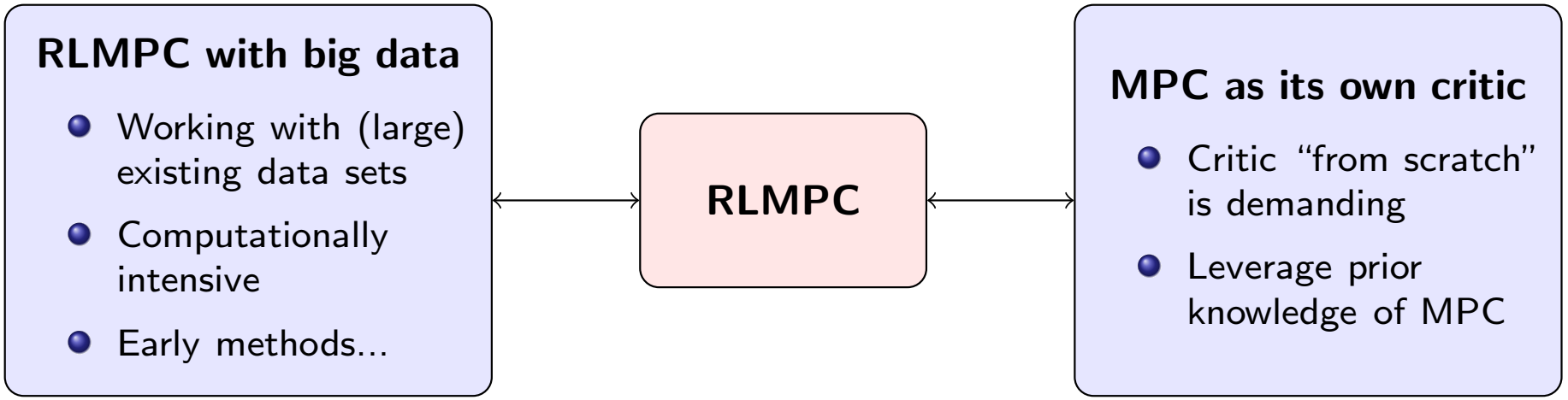
More explored questions in brief - Algorithms

RLMPC with big data

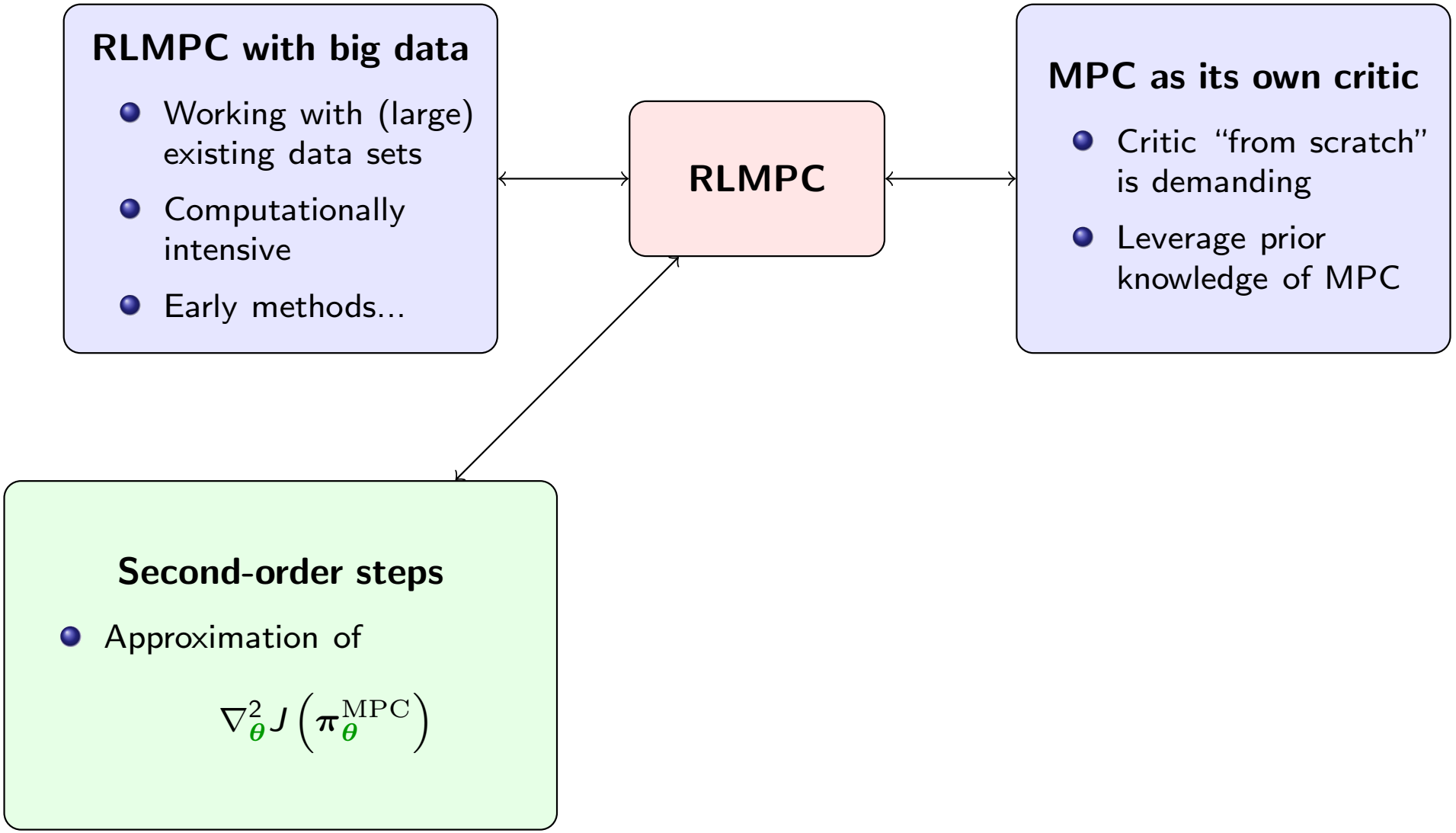
- Working with (large) existing data sets
- Computationally intensive
- Early methods...

RLMPC

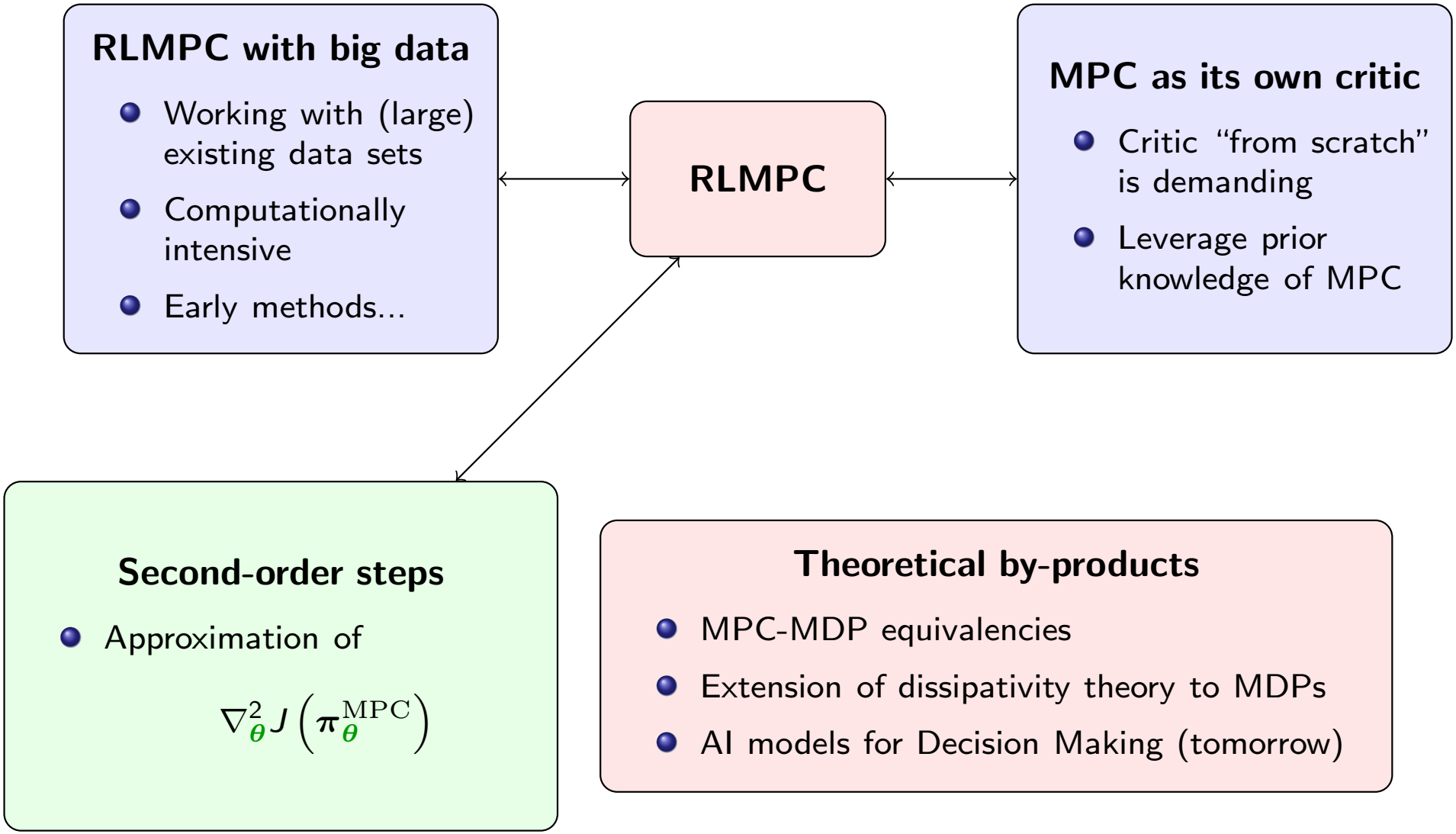
More explored questions in brief - Algorithms



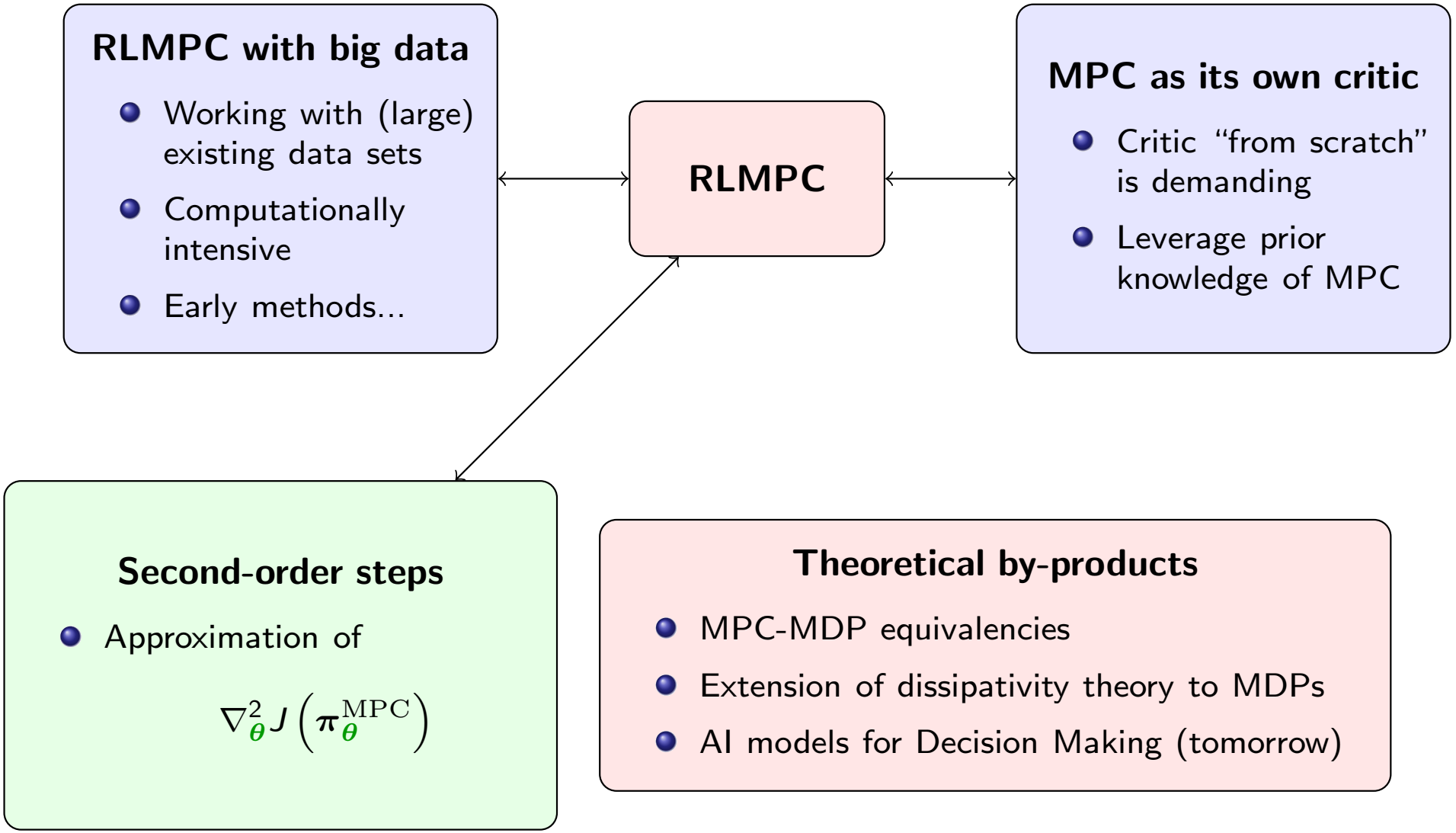
More explored questions in brief - Algorithms



More explored questions in brief - Algorithms



More explored questions in brief - Algorithms



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

Orientation

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)*