Orientation

What we have seen:

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

Orientation

What we have seen:

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

What we will do next: RL over MPC

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

4□ > 4□ > 4□ > 4□ > 4□ > 4□

1/33

RL and MPC Safety, Stability, Belief State

Sebastien Gros

Dept. of Cybernetic, NTNU Faculty of Information Tech.

Freiburg PhD School

Outline Safe RL via MPC Exploration with MPC 3 Stability-constrained Learning with MPC **Explored questions** Future Prospect – Belief State in RLMPC?

Outline Safe RL via MPC Stability-constrained Learning with MPC Future Prospect – Belief State in RLMPCK

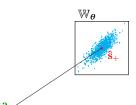
True system: $\mathbf{s}_+ \sim \mathbb{P}\left[\,\cdot\,|\mathbf{s},\mathbf{a}\,
ight]$

Deterministic model: $\hat{s}_{+} = f_{\theta}\left(s,a\right)$

S. Gros (NTNU) MPC & RL Fall 2025 5/33

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

Deterministic model: $\hat{s}_{+} = f_{\theta}\left(s,a\right)$



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

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

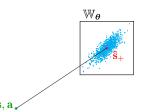
with probability 1

<ロ > ← □ > ← □ > ← □ > ← □ = − の へ ○

5/33

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

Deterministic model: $\hat{s}_{+} = f_{\theta}\left(s,a\right)$



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

$$\mathbf{s}_{+}\in\mathbf{f}_{m{ heta}}\left(\mathbf{s},\mathbf{a}
ight)+\mathbb{W}_{m{ heta}}$$
 (1) with probability 1

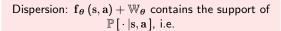
Remarks:

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

◆ロト ◆御 ト ◆ 恵 ト ◆ 恵 ・ 夕 Q ○

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

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

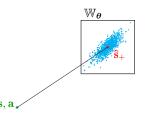


$$\mathbf{s}_{+} \in \mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s}, \mathbf{a}\right) + \mathbb{W}_{\boldsymbol{\theta}}$$
 (1)

with probability 1

Remarks:

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



Condition

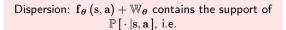
$$\mathbf{s}_{+}-\mathbf{f}_{oldsymbol{ heta}}\left(\mathbf{s},\mathbf{a}
ight)\in\mathbb{W}_{oldsymbol{ heta}}$$

for all observed triplets (s,a,s_+)

ightarrow constraints on $oldsymbol{ heta}$

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

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

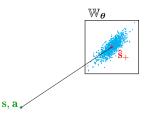


$$\mathbf{s}_{+}\in\mathbf{f}_{oldsymbol{ heta}}\left(\mathbf{s},\mathbf{a}
ight)+\mathbb{W}_{oldsymbol{ heta}}$$
 (1)

with probability 1

Remarks:

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



Condition

$$\mathbf{s}_{+}-\mathbf{f}_{oldsymbol{ heta}}\left(\mathbf{s},\mathbf{a}
ight)\in\mathbb{W}_{oldsymbol{ heta}}$$

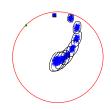
for all observed triplets (s,a,s_+)

ightarrow constraints on $oldsymbol{ heta}$

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

Robust (N)MPC delivers policy $oldsymbol{\pi}_{oldsymbol{ heta}}(x_0) = \mathbf{u}_0^{\star}$ from

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



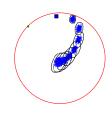
- ullet $x_{0,...,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$

S. Gros (NTNU) MPC & RL Fall 2025 6 / 33

Robust (N)MPC delivers policy
$$\pi_{ heta}(x_0) = \mathbf{u}_0^{\star}$$
 from

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

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



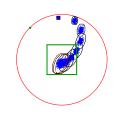
- ullet $x_{0,...,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $\mathbf{x}_{1,...,N-1}(\mathbf{u},\mathbf{s},\boldsymbol{\theta},\mathbf{w})$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

6/33

S. Gros (NTNU) MPC & RL Fall 2025

Robust (N)MPC delivers policy $\pi_{ heta}(x_0) = \mathbf{u}_0^{\star}$ from

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

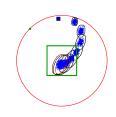


- $\mathbf{x}_{0,...,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $\mathbf{x}_{1,...,N-1}(\mathbf{u},\mathbf{s},\boldsymbol{\theta},\mathbf{w})$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

6/33

Robust (N)MPC delivers policy $\pi_{ heta}(x_0) = \mathbf{u}_0^{\star}$ from

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

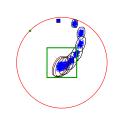


- ullet $x_{0,...,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \dots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\theta}$
- $x_{1,...,N-1}(\mathbf{u},\mathbf{s},\boldsymbol{\theta},\mathbf{w})$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

6/33

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

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



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

Closed-loop stability under some conditions on θ (not trivial), need $\gamma=1$

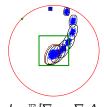
S. Gros (NTNU) MPC & RL Fall 2025 6/33

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

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

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

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



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

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

Closed-loop stability under some conditions on heta (not trivial), need $\gamma=1$

S. Gros (NTNU) MPC & RL Fall 2025 6/33

Robust NMPC parameters θ

Policy gradient

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

adjusts $oldsymbol{ heta}$ for performance

Condition

$$\mathbf{s}_{+}-\mathbf{f}\left(\mathbf{s},\mathbf{a},oldsymbol{ heta}
ight)\in\mathbb{W}_{oldsymbol{ heta}}$$

enforces safety through heta

Robust NMPC parameters θ

Policy gradient

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

adjusts $oldsymbol{ heta}$ for performance

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

Condition

$$\mathbf{s}_{+}-\mathbf{f}\left(\mathbf{s},\mathbf{a},oldsymbol{ heta}
ight)\in\mathbb{W}_{oldsymbol{ heta}}$$

enforces safety through heta

7/33

Robust NMPC parameters θ

Policy gradient

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

adjusts heta for performance

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

Condition

$$\mathbf{s}_{+}-\mathbf{f}\left(\mathbf{s},\mathbf{a},oldsymbol{ heta}
ight)\in\mathbb{W}_{oldsymbol{ heta}}$$

enforces safety through heta

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

Robust NMPC parameters θ

Policy gradient

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

adjusts heta for performance

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

Condition

$$\mathbf{s}_{+} - \mathbf{f}\left(\mathbf{s}, \mathbf{a}, oldsymbol{ heta}
ight) \in \mathbb{W}_{oldsymbol{ heta}}$$

enforces safety through heta

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

How to do Safe RL?

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

Robust NMPC parameters θ

Policy gradient

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

adjusts heta for performance

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

Condition

$$\mathbf{s}_{+} - \mathbf{f}\left(\mathbf{s}, \mathbf{a}, oldsymbol{ heta}
ight) \in \mathbb{W}_{oldsymbol{ heta}}$$

enforces safety through heta

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

How to do Safe RL?

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

Also reads as:

$$oldsymbol{ heta} \leftarrow oldsymbol{ heta} + \Delta oldsymbol{ heta}$$

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

S. Gros (NTNU)

MPC & RL

Fall 2025

Robust NMPC parameters θ

Policy gradient

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

adjusts heta for performance

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

Condition

$$\mathbf{s}_{+} - \mathbf{f}\left(\mathbf{s}, \mathbf{a}, oldsymbol{ heta}
ight) \in \mathbb{W}_{oldsymbol{ heta}}$$

enforces safety through heta

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

How to do Safe RL?

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

Also reads as:

$$oldsymbol{ heta} \leftarrow oldsymbol{ heta} + \Delta oldsymbol{ heta}$$

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

Safe RL steps $oldsymbol{ heta} \leftarrow oldsymbol{ heta} + \Delta oldsymbol{ heta}$:

$$\begin{split} \Delta \boldsymbol{\theta} &= \arg \min_{\Delta \boldsymbol{\theta}} \frac{1}{2\alpha} \left\| \Delta \boldsymbol{\theta} \right\|^2 + \nabla_{\boldsymbol{\theta}} \boldsymbol{J}^\top \Delta \boldsymbol{\theta} \\ &\mathrm{s.t.} \ \, \mathbf{s}_+ - \mathbf{f} \left(\mathbf{s}, \mathbf{a}, \boldsymbol{\theta} + \Delta \boldsymbol{\theta} \right) \in \mathbb{W}_{\boldsymbol{\theta} + \Delta \boldsymbol{\theta}} \\ & \quad \forall \left(\mathbf{s}, \mathbf{a}, \mathbf{s}_+ \right) \ \, \text{in data set} \end{split}$$

S. Gros (NTNU) MPC & RL

Robust NMPC parameters θ

Policy gradient

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

adjusts heta for performance

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

Condition

$$\mathbf{s}_{+} - \mathbf{f}\left(\mathbf{s}, \mathbf{a}, oldsymbol{ heta}
ight) \in \mathbb{W}_{oldsymbol{ heta}}$$

enforces safety through heta

 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:

$$oldsymbol{ heta} \leftarrow oldsymbol{ heta} + \Delta oldsymbol{ heta}$$

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

Safe RL steps $oldsymbol{ heta} \leftarrow oldsymbol{ heta} + \Delta oldsymbol{ heta}$:

$$\begin{split} \Delta \boldsymbol{\theta} &= \arg \min_{\Delta \boldsymbol{\theta}} \frac{1}{2\alpha} \left\| \Delta \boldsymbol{\theta} \right\|^2 + \nabla_{\boldsymbol{\theta}} \boldsymbol{J}^\top \Delta \boldsymbol{\theta} \\ \text{s.t. } \mathbf{s}_+ - \mathbf{f} \left(\mathbf{s}, \mathbf{a}, \boldsymbol{\theta} + \Delta \boldsymbol{\theta} \right) \in \mathbb{W}_{\boldsymbol{\theta} + \Delta \boldsymbol{\theta}} \\ \forall \left(\mathbf{s}, \mathbf{a}, \mathbf{s}_+ \right) \text{ in data set} \end{split}$$

Safe RL steps seek performance under safety constraints

Robust NMPC parameters θ

Policy gradient

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

adjusts heta for performance

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

Condition

$$\mathbf{s}_{+} - \mathbf{f}\left(\mathbf{s}, \mathbf{a}, oldsymbol{ heta}
ight) \in \mathbb{W}_{oldsymbol{ heta}}$$

enforces safety through heta

7/33

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

How to do Safe RL?

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

Also reads as:

$$oldsymbol{ heta} \leftarrow oldsymbol{ heta} + \Delta oldsymbol{ heta}$$

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

Safe RL steps $oldsymbol{ heta} \leftarrow oldsymbol{ heta} + \Delta oldsymbol{ heta}$:

$$\Delta \boldsymbol{\theta} = \arg \min_{\Delta \boldsymbol{\theta}} \frac{1}{2\alpha} \|\Delta \boldsymbol{\theta}\|^2 + \nabla_{\boldsymbol{\theta}} \boldsymbol{J}^{\top} \Delta \boldsymbol{\theta}$$
 s.t. $\mathbf{s}_{+} - \mathbf{f} (\mathbf{s}, \mathbf{a}, \boldsymbol{\theta} + \Delta \boldsymbol{\theta}) \in \mathbb{W}_{\boldsymbol{\theta} + \Delta \boldsymbol{\theta}}$
$$\forall (\mathbf{s}, \mathbf{a}, \mathbf{s}_{+}) \text{ in data set}$$

Safe RL steps seek performance under safety constraints Difficulty: differentiating through the effect of \mathbb{W}_{θ} , big data in Safe RL steps

Illustrative example

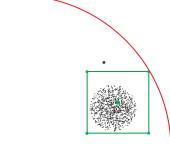
Real system

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

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

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

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



Illustrative example

Real system

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

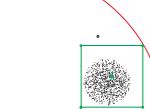
- Noise n in a ball
- Robust MPC model

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

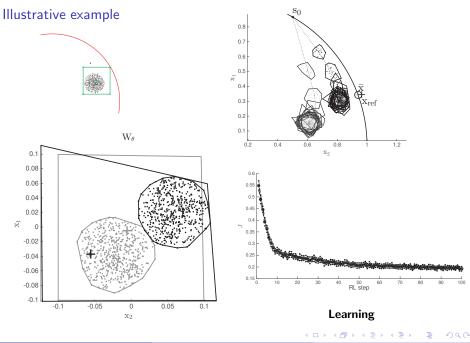
- W is a square
- Quadratic stage cost
- Constraint $||\mathbf{x}||^2 \le 1$

Let's adjusts:

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

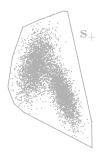


S. Gros (NTNU) MPC & RL Fall 2025 8/33



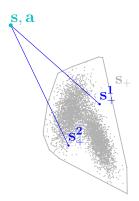
Key idea: approximate distribution $\mathbb{P}\left[\mathbf{s}_{+}|\mathbf{s},\mathbf{a}\right]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.

 \mathbf{s}, \mathbf{a}



S. Gros (NTNU) MPC & RL Fall 2025 10 / 33

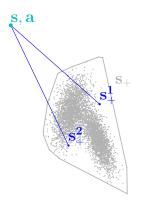
Key idea: approximate distribution $\mathbb{P}\left[s_{+}|s,a\right]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.



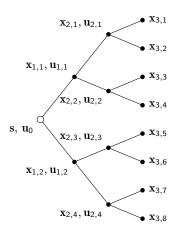
E.g. pick two scenarios

S. Gros (NTNU) MPC & RL Fall 2025 10 / 33

Key idea: approximate distribution $\mathbb{P}\left[s_{+}|s,a\right]$ 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



S. Gros (NTNU) MPC & RL Fall 2025 10 / 33

Key idea: approximate distribution $\mathbb{P}\left[s_{+}|s,a\right]$ 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 $\pi^{\mathrm{MPC}}\left(s\right)=u_{0}^{\star}$ from

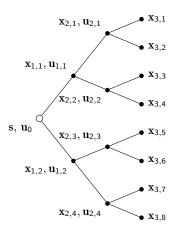
$$\min_{\mathbf{x}, \mathbf{u}} \quad \sum_{i=1}^{n} \sum_{k=0}^{N} \omega_{i} L\left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right)$$
s.t.
$$\mathbf{x}_{k+1}^{i} = \mathbf{f}_{k}^{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right), \quad \mathbf{x}_{0}^{i} = \mathbf{s}$$

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

$$\mathbf{c}_{k} \left(\mathbf{u}_{k}^{1, \dots, n}\right) = 0$$

where we need to select

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



10 / 33

S. Gros (NTNU) MPC & RL Fall 2025

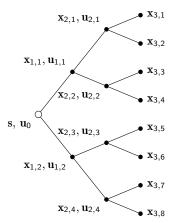
Key idea: approximate distribution $\mathbb{P}\left[s_{+}|s,a\right]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.

Scenario-Tree MPC gives $oldsymbol{\pi}^{\mathrm{MPC}}\left(s\right)=u_{0}^{\star}$ from

$$\begin{aligned} & \underset{\mathbf{x}, \mathbf{u}}{\min} & & \sum_{i=1}^{n} \sum_{k=0}^{N} \omega_{i} L\left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right) \\ & \text{s.t.} & & \mathbf{x}_{k+1}^{i} = \mathbf{f}_{k}^{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right), & & \mathbf{x}_{0}^{i} = \mathbf{s} \\ & & \mathbf{h}_{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right) \leq 0 \\ & & & \mathbf{c}_{k} \left(\mathbf{u}_{k}^{1, \dots, n}\right) = 0 \end{aligned}$$

where we need to select

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



Difficulties

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

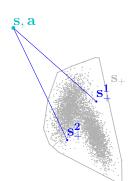
Often shallow trees are used

4□ > 4問 > 4 = > 4 = > ■ 90

10 / 33

S. Gros (NTNU) MPC & RL Fall 2025

RL over Scenario Tree MPC



Scenario-Tree MPC

$$\begin{aligned} & \underset{\mathbf{x}, \mathbf{u}}{\text{min}} & & \sum_{i=1}^{n} \sum_{k=0}^{N} \omega_{i} L\left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right) \\ & \text{s.t.} & & \mathbf{x}_{k+1}^{i} = \mathbf{f}_{k}^{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right), \quad \mathbf{x}_{0}^{i} = \mathbf{s} \\ & & & \mathbf{h}_{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right) \leq 0 \\ & & & & \mathbf{c}_{k} \left(\mathbf{u}_{k}^{1, \dots, n}\right) = \mathbf{0} \end{aligned}$$

Difficulties

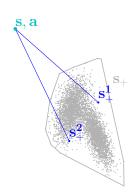
- Pick "good" scenarios
- Exploding complexity

S. Gros (NTNU)

MPC & RL

Fall 2025

RL over Scenario Tree MPC



Difficulties

- Pick "good" scenarios
- Exploding complexity

Scenario-Tree MPC

$$\begin{aligned} & \underset{\mathbf{x}, \mathbf{u}}{\text{min}} & & \sum_{i=1}^{n} \sum_{k=0}^{N} \omega_{i} L\left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right) \\ & \text{s.t.} & & \mathbf{x}_{k+1}^{i} = \mathbf{f}_{k}^{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right), & & \mathbf{x}_{0}^{i} = \mathbf{s} \\ & & & \mathbf{h}_{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right) \leq 0 \\ & & & & \mathbf{c}_{k} \left(\mathbf{u}_{k}^{1, \dots, n}\right) = 0 \end{aligned}$$

RL can do:

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

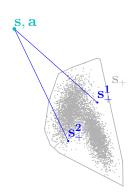
for real-world performance of policy

S. Gros (NTNU)

MPC & RL

Fall 2025

RL over Scenario Tree MPC



Difficulties

- Pick "good" scenarios
- Exploding complexity

Scenario-Tree MPC

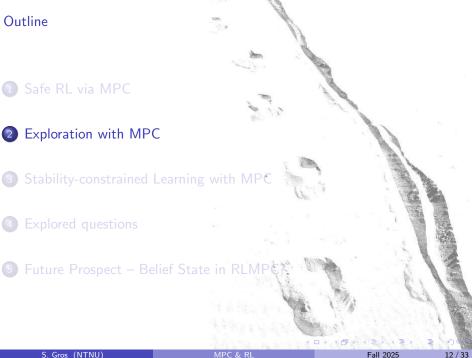
$$\begin{aligned} & \underset{\mathbf{x}, \mathbf{u}}{\text{min}} & & \sum_{i=1}^{n} \sum_{k=0}^{N} \omega_{i} L\left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right) \\ & \text{s.t.} & & \mathbf{x}_{k+1}^{i} = \mathbf{f}_{k}^{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right), \quad \mathbf{x}_{0}^{i} = \mathbf{s} \\ & & & \mathbf{h}_{i} \left(\mathbf{x}_{k}^{i}, \mathbf{u}_{k}^{i}\right) \leq 0 \\ & & & & \mathbf{c}_{k} \left(\mathbf{u}_{k}^{1, \dots, n}\right) = 0 \end{aligned}$$

RL can do:

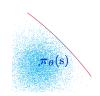
- Optimize scenarios f_k^i
- Optimize weights ω_i
- Can work on shallow trees

for real-world performance of policy

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



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



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



Explore while keeping feasibility?

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

Learning requires exploration. E.g. apply $\mathbf{a}=\pi_{m{ heta}}\left(\mathbf{s}\right)+\mathbf{d}$ to the real system where \mathbf{d} is a "disturbance"



Explore while keeping feasibility?

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

Feasible exploration:
$$m{\pi}^{\mathrm{e}}_{m{ heta}}(\mathrm{s}) = \mathbf{u}^{\star}_{0}$$
:

$$\min_{\mathbf{x},\mathbf{u}} \quad T_{oldsymbol{ heta}}\left(\mathbf{x}_{oldsymbol{N}}
ight) - \mathbf{d}^{ op}\mathbf{u}_{0} + \sum_{k=0}^{oldsymbol{N}-1} L_{oldsymbol{ heta}}\left(\mathbf{x}_{k},\mathbf{u}_{k}
ight)$$

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

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

satisfies the constraints by construction

◆ロト ◆問 ▶ ◆ 重 ト ◆ 重 ・ 夕 Q ©

S. Gros (NTNU) MPC & RL

Learning requires exploration. E.g. apply $\mathbf{a}=\pi_{m{ heta}}\left(\mathbf{s}\right)+\mathbf{d}$ to the real system where \mathbf{d} is a "disturbance"



Explore while keeping feasibility?

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

Feasible exploration:
$$m{\pi}^{\mathrm{e}}_{m{ heta}}(\mathrm{s}) = \mathbf{u}^{\star}_{0}$$
:

$$\min_{\mathbf{x},\mathbf{u}} \quad T_{oldsymbol{ heta}}\left(\mathbf{x}_{oldsymbol{N}}
ight) - \mathbf{d}^{ op}\mathbf{u}_{0} + \sum_{k=0}^{oldsymbol{N}-1} L_{oldsymbol{ heta}}\left(\mathbf{x}_{k},\mathbf{u}_{k}
ight)$$

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

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

satisfies the constraints by construction

◆ロト ◆問 ▶ ◆ 重 ト ◆ 重 ・ 夕 Q ©

S. Gros (NTNU) MPC & RL

Learning requires exploration. E.g. apply $\mathbf{a}=\pi_{m{ heta}}\left(\mathbf{s}\right)+\mathbf{d}$ to the real system where \mathbf{d} is a "disturbance"



Explore while keeping feasibility?

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

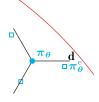
Feasible exploration:
$$oldsymbol{\pi}_{oldsymbol{ heta}}^{\mathrm{e}}(\mathrm{s}) = \mathrm{u}_{\mathrm{0}}^{\star}$$
:

$$\min_{\mathbf{x},\mathbf{u}} \quad T_{\boldsymbol{ heta}}\left(\mathbf{x}_{N}
ight) - \mathbf{d}^{ op}\mathbf{u}_{0} + \sum_{k=0}^{N-1} L_{oldsymbol{ heta}}\left(\mathbf{x}_{k},\mathbf{u}_{k}
ight)$$

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

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

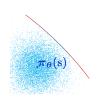
satisfies the constraints by construction



13 / 33

S. Gros (NTNU) MPC & RL Fall 2025

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



Explore while keeping feasibility?

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

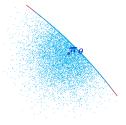
Feasible exploration:
$$oldsymbol{\pi}_{oldsymbol{ heta}}^{\mathrm{e}}(\mathrm{s}) = \mathbf{u}_{\mathrm{0}}^{\star}$$
:

$$\min_{\mathbf{x},\mathbf{u}} \quad \mathcal{T}_{oldsymbol{ heta}}\left(\mathbf{x}_{oldsymbol{N}}
ight) - \mathbf{d}^{ op}\mathbf{u}_{0} + \sum_{k=0}^{N-1} L_{oldsymbol{ heta}}\left(\mathbf{x}_{k},\mathbf{u}_{k}
ight)$$

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

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

satisfies the constraints by construction



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



Explore while keeping feasibility?

- ullet Clearly an arbitrary "policy disturbance" $oldsymbol{\pi_{ heta}}\left(\mathbf{s}
 ight)+\mathbf{d}$ is a poor idea...
- NLP-based policy: "disturb" the cost function instead! (different options)

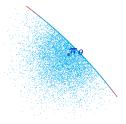
Feasible exploration:
$$m{\pi}^{\mathrm{e}}_{m{ heta}}(\mathrm{s}) = \mathbf{u}^{\star}_{0}$$
:

$$\min_{\mathbf{x},\mathbf{u}} \quad \mathcal{T}_{oldsymbol{ heta}}\left(\mathbf{x}_{oldsymbol{N}}
ight) - \mathbf{d}^{ op}\mathbf{u}_{0} + \sum_{k=0}^{N-1} L_{oldsymbol{ heta}}\left(\mathbf{x}_{k},\mathbf{u}_{k}
ight)$$

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

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

satisfies the constraints by construction



13 / 33

S. Gros (NTNU) MPC & RL Fall 2025

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



Explore while keeping feasibility?

- ullet Clearly an arbitrary "policy disturbance" $oldsymbol{\pi_{ heta}}\left(\mathbf{s}
 ight)+\mathbf{d}$ is a poor idea...
- NLP-based policy: "disturb" the cost function instead! (different options)

Feasible exploration: $oldsymbol{\pi}_{oldsymbol{ heta}}^{\mathrm{e}}(\mathrm{s}) = \mathrm{u}_{\mathrm{0}}^{\star}$:

$$\min_{\mathbf{x},\mathbf{u}} \quad T_{oldsymbol{ heta}}\left(\mathbf{x}_{oldsymbol{N}}
ight) - \mathbf{d}^{ op}\mathbf{u}_{0} + \sum_{k=0}^{oldsymbol{N}-1} L_{oldsymbol{ heta}}\left(\mathbf{x}_{k},\mathbf{u}_{k}
ight)$$

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

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

satisfies the constraints by construction

Remarks:

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

Stochastic policy $\pi_{\boldsymbol{\theta}}\left[\,\cdot\,|\mathbf{s}
ight] \equiv \mathbf{u}_{0}^{\star}\left(\boldsymbol{\theta},\mathbf{s},\mathbf{d}\right)$:

$$\min_{\mathbf{x},\mathbf{u}} \quad T_{\theta}\left(\mathbf{x}_{\textit{N}}, \textcolor{red}{\textbf{d}}\right) + \sum_{k=0}^{\textit{N}-1} \textit{L}_{\theta}\left(\mathbf{x}_{k}, \mathbf{u}_{k}, \textcolor{red}{\textbf{d}}\right)$$

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

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

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

∢ロ ▶ ∢御 ▶ ∢ 種 ▶ ∢ 種 ▶ ○ 種 ○ 夕 ◎

Stochastic policy
$$\pi_{\boldsymbol{\theta}}\left[\,\cdot\,|\mathbf{s}
ight] \equiv \mathbf{u}_{0}^{\star}\left(\boldsymbol{\theta},\mathbf{s},\mathbf{d}\right)$$
:

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

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

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

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

Remarks

- ullet Chose ho easy to sample from
- Typ. bounded support
- Special case: $\theta = \bar{\theta} + \mathbf{d}$
- ullet To ensure feasibility, leave $f_{m{ heta}}, h_{m{ heta}}$ "alone"

Stochastic policy $\pi_{\boldsymbol{\theta}}\left[\,\cdot\,|\mathbf{s} ight] \equiv \mathbf{u}_{0}^{\star}\left(\boldsymbol{\theta},\mathbf{s},\mathbf{d}\right)$:

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

$$\mathbf{h}_{\theta}\left(\mathbf{x}_{k},\mathbf{u}_{k},\mathbf{d}\right)<0,\quad\mathbf{x}_{0}=\mathbf{s}$$

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

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

Remarks

- ullet Chose ho easy to sample from
- Typ. bounded support
- Special case: $\theta = \bar{\theta} + \mathbf{d}$
- \bullet To ensure feasibility, leave f_{θ}, h_{θ} "alone"

Performance

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

◆ロト ◆部 ▶ ◆ 恵 ト ・ 恵 ・ 夕 Q (^)

S. Gros (NTNU) MPC & RL

Stochastic policy $\pi_{\boldsymbol{\theta}}\left[\,\cdot\,|\mathbf{s} ight] \equiv \mathbf{u}_{0}^{\star}\left(\boldsymbol{\theta},\mathbf{s},\mathbf{d}\right)$:

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

$$\mathbf{h}_{\boldsymbol{\theta}}\left(\mathbf{x}_{k},\mathbf{u}_{k},\mathbf{d}\right)<0,\quad\mathbf{x}_{0}=\mathbf{s}$$

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

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

Remarks

- ullet Chose ho easy to sample from
- Typ. bounded support
- Special case: $\theta = \bar{\theta} + \mathbf{d}$
- ullet To ensure feasibility, leave $f_{m{ heta}}, h_{m{ heta}}$ "alone"

Performance

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

Stochastic Policy Gradient

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

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

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

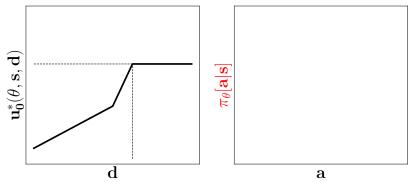


Illustration – Linear MPC with scalar $\mathbf{a},\ \mathbf{d}$ normal centered

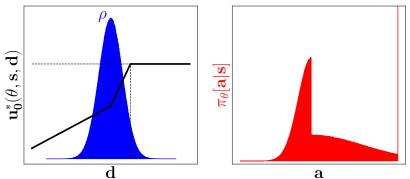
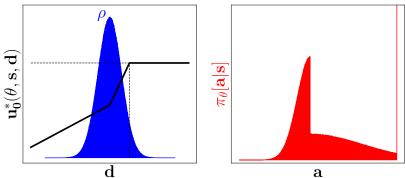


Illustration – Linear MPC with scalar $\mathbf{a},\,\mathbf{d}$ normal centered



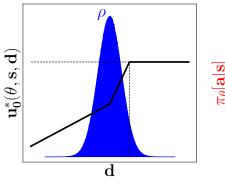
For $\mathbf{d} \to \mathbf{u}_0^{\star}$ bijective & differentiable

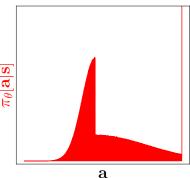
$$\pi_{\boldsymbol{\theta}}\left[\mathbf{a}|\mathbf{s}\right] = \left|\det\left(\nabla_{\mathbf{d}}\mathbf{u}_{0}^{\star}\left(\boldsymbol{\theta},\mathbf{s},\mathbf{d}\right)\right)\right|^{-1}\rho\left[\mathbf{d}\right]$$

15/33

S. Gros (NTNU) MPC & RL Fall 2025

Illustration - Linear MPC with scalar a, d normal centered





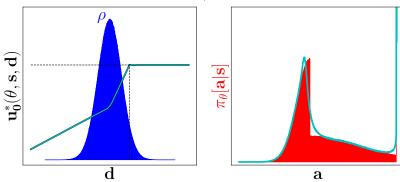
For $\mathbf{d} \to \mathbf{u}_0^{\star}$ bijective & differentiable

$$\pi_{\boldsymbol{\theta}}\left[\mathbf{a}|\mathbf{s}\right] = \left|\det\left(\nabla_{\mathbf{d}}\mathbf{u}_{0}^{\star}\left(\boldsymbol{\theta},\mathbf{s},\mathbf{d}\right)\right)\right|^{-1}\rho\left[\mathbf{d}\right]$$

 Even for the "simplest" form of MPC, there are challenges

15/33

Illustration - Linear MPC with scalar a, d normal centered



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

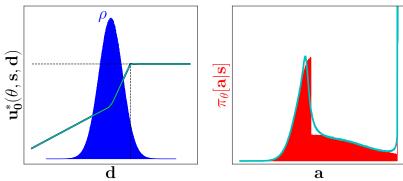
$$\pi_{\boldsymbol{\theta}}\left[\mathbf{a}|\mathbf{s}\right] = \left|\det\left(\nabla_{\mathbf{d}}\mathbf{u}_{0}^{\star}\left(\boldsymbol{\theta},\mathbf{s},\mathbf{d}\right)\right)\right|^{-1}\rho\left[\mathbf{d}\right]$$

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

15/33

S. Gros (NTNU) MPC & RL Fall 2025

Illustration - Linear MPC with scalar a, d normal centered



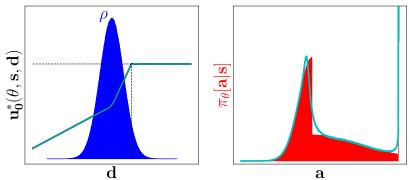
For $\mathbf{d} \to \mathbf{u}_0^\star$ bijective & differentiable

$$\pi_{\boldsymbol{\theta}}\left[\mathbf{a}|\mathbf{s}\right] = \left|\det\left(\nabla_{\mathbf{d}}\mathbf{u}_{0}^{\star}\left(\boldsymbol{\theta},\mathbf{s},\mathbf{d}\right)\right)\right|^{-1}\rho\left[\mathbf{d}\right]$$

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

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

Illustration - Linear MPC with scalar a, d normal centered

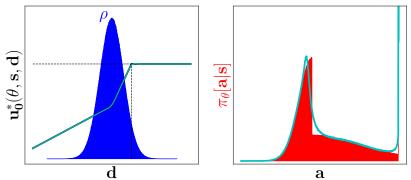


Reparametrization trick gives (smoothing is still theoretically good...)

$$\nabla_{\theta} J\left(\pi_{\theta}\right) = \mathbb{E}\left[\nabla_{\theta} \log \pi_{\theta} \cdot Q^{\pi_{\theta}}\right] = \mathbb{E}\left[\nabla_{\theta} \mathbf{u}_{0}^{\star}\left(\theta, \mathbf{s}, \mathbf{d}\right) \cdot \nabla_{\mathbf{a}} Q^{\pi_{\theta}}\left(\mathbf{s}, \mathbf{a}\right)\right]$$

◆ロト ◆御 ト ◆ 恵 ト ◆ 恵 ・ 夕久で

Illustration - Linear MPC with scalar a, d normal centered



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

$$\nabla_{\theta} J(\pi_{\theta}) = \mathbb{E}\left[\nabla_{\theta} \log \pi_{\theta} \cdot Q^{\pi_{\theta}}\right] = \mathbb{E}\left[\nabla_{\theta} \mathbf{u}_{0}^{\star}\left(\theta, \mathbf{s}, \mathbf{d}\right) \cdot \nabla_{\mathbf{a}} Q^{\pi_{\theta}}\left(\mathbf{s}, \mathbf{a}\right)\right]$$

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

15 / 33

Outline 3 Stability-constrained Learning with MPC Future Prospect - Belief State in RLMPC.

$$\begin{split} & \textbf{Policy } \boldsymbol{\pi}_{\mathrm{MPC}} \ \textbf{from} \\ & \underset{s,\mathbf{a}}{\text{min}} \quad \boldsymbol{T}\left(\mathbf{s}_{\textit{N}}\right) + \sum_{k=0}^{N-1} L\left(\mathbf{s}_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) \\ & \text{s.t.} \quad \mathbf{s}_{k+1} = \mathbf{f}\left(\mathbf{s}_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) \\ & \quad \mathbf{h}\left(\mathbf{s}_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) \leq 0, \quad \mathbf{s}_{\textit{N}} \in \mathbb{T} \end{split}$$

$$\begin{split} & \text{Policy } \boldsymbol{\pi}_{\mathrm{MPC}} \text{ from} \\ & \underset{s,a}{\text{min}} \quad \mathcal{T}\left(s_{\mathcal{N}}\right) + \sum_{k=0}^{\mathcal{N}-1} L\left(s_{k}, a_{k}\right) \\ & \text{s.t.} \quad s_{k+1} = \mathbf{f}\left(s_{k}, a_{k}\right) \\ & \quad \mathbf{h}\left(s_{k}, a_{k}\right) \leq 0, \quad s_{\mathcal{N}} \in \mathbb{T} \end{split}$$

If for some K_{∞} function κ ("bowl-shaped"):

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

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

《□▶ 《圖▶ 《圖▶ 《圖▶ ■ 釣@@

$$\begin{split} & \textbf{Policy } \boldsymbol{\pi}_{\mathrm{MPC}} \ \textbf{from} \\ & \underset{s,\mathbf{a}}{\text{min}} \quad \mathcal{T}\left(s_{\mathcal{N}}\right) + \sum_{k=0}^{\mathcal{N}-1} L\left(s_{k}, \mathbf{a}_{k}\right) \\ & \text{s.t.} \quad s_{k+1} = \mathbf{f}\left(s_{k}, \mathbf{a}_{k}\right) \\ & \quad \mathbf{h}\left(s_{k}, \mathbf{a}_{k}\right) \leq \mathbf{0}, \quad s_{\mathcal{N}} \in \mathbb{T} \end{split}$$

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

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

then MPC scheme is stabilizing (+conditions on T)

Policy π_{MPC} from

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

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

 $\mathbf{h}(\mathbf{s}_k, \mathbf{a}_k) < 0, \quad \mathbf{s}_N \in \mathbb{T}$

$$=(S_K, u_K) \subseteq S_K \subseteq S_K$$

Equivalent MPC

$$\min_{\mathbf{s},\mathbf{a}} \quad -\lambda\left(\mathbf{s}_{0}\right) + \tilde{\mathcal{T}}\left(\mathbf{s}_{N}\right) + \sum_{k=0}^{N-1} \ell\left(\mathbf{s}_{k},\mathbf{a}_{k}\right)$$

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

 $\mathbf{h}(\mathbf{s}_k, \mathbf{a}_k) \le 0, \quad \mathbf{s}_N \in \mathbb{T}$

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

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

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

then MPC scheme is stabilizing (+conditions on T)

◆ロト ◆部 ト ◆ 差 ト ◆ 差 ・ 夕 Q (*)

$\begin{aligned} & \textbf{Policy } \; \boldsymbol{\pi}_{\mathrm{MPC}} \; \textbf{from} \\ & \underset{\mathbf{s}, \mathbf{a}}{\text{min}} \quad \mathcal{T}\left(\mathbf{s}_{\textit{N}}\right) + \sum_{k=0}^{N-1} L\left(\mathbf{s}_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) \\ & \text{s.t.} \quad \mathbf{s}_{k+1} = \mathbf{f}\left(\mathbf{s}_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) \\ & \quad \quad \mathbf{h}\left(\mathbf{s}_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) < 0, \quad \mathbf{s}_{\textit{N}} \in \mathbb{T} \end{aligned}$

Equivalent MPC

$$egin{aligned} \min_{\mathbf{s},\mathbf{a}} & -\lambda\left(\mathbf{s}_{0}
ight) + ilde{\mathcal{T}}\left(\mathbf{s}_{N}
ight) + \sum_{k=0}^{N-1}\ell\left(\mathbf{s}_{k},\mathbf{a}_{k}
ight) \\ \mathrm{s.t.} & \mathbf{s}_{k+1} = \mathbf{f}\left(\mathbf{s}_{k},\mathbf{a}_{k}
ight) \\ & \mathbf{h}\left(\mathbf{s}_{k},\mathbf{a}_{k}
ight) \leq 0, \quad \mathbf{s}_{N} \in \mathbb{T} \end{aligned}$$

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

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

then MPC scheme is stabilizing (+conditions on T)

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

Remarks:

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

17/33

S. Gros (NTNU) MPC & RL Fall 2025

$$\begin{split} & \text{Policy } \boldsymbol{\pi}_{\mathrm{MPC}} \text{ from} \\ & \underset{s, \mathbf{a}}{\text{min}} \quad \mathcal{T}\left(s_{\textit{N}}\right) + \sum_{k=0}^{N-1} L\left(s_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) \\ & \text{s.t.} \quad \mathbf{s}_{k+1} = \mathbf{f}\left(s_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) \\ & \quad \mathbf{h}\left(s_{\textit{k}}, \mathbf{a}_{\textit{k}}\right) \leq 0, \quad s_{\textit{N}} \in \mathbb{T} \end{split}$$

Equivalent MPC

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

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

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

then MPC scheme is stabilizing (+conditions on T)

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

Remarks:

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

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

$$\begin{split} & \textbf{Policy } \boldsymbol{\pi}^{\mathrm{MPC}} \ \textbf{from} \\ & \underset{\mathbf{x},\mathbf{u}}{\text{min}} \quad \mathcal{T}\left(\mathbf{x}_{\textit{N}}\right) + \sum_{\textit{k}=0}^{\textit{N}-1} \textit{L}\left(\mathbf{x}_{\textit{k}},\mathbf{u}_{\textit{k}}\right) \\ & \text{s.t.} \quad \mathbf{x}_{\textit{k}+1} = \mathbf{f}\left(\mathbf{x}_{\textit{k}},\mathbf{u}_{\textit{k}}\right), \ \mathbf{x}_{0} = \mathbf{s} \\ & \quad \mathbf{h}\left(\mathbf{x}_{\textit{k}},\mathbf{u}_{\textit{k}}\right) \leq 0 \end{split}$$

$$\begin{split} & \text{Policy } \boldsymbol{\pi}^{\mathrm{MPC}} \text{ from} \\ & \underset{x,u}{\text{min}} \quad \mathcal{T}\left(\mathbf{x}_{\textit{N}}\right) + \sum_{k=0}^{\textit{N}-1} \textit{L}\left(\mathbf{x}_{\textit{k}}, \mathbf{u}_{\textit{k}}\right) \\ & \mathrm{s.t.} \quad \mathbf{x}_{\textit{k}+1} = \mathbf{f}\left(\mathbf{x}_{\textit{k}}, \mathbf{u}_{\textit{k}}\right), \ \mathbf{x}_{0} = \mathbf{s} \\ & \quad \mathbf{h}\left(\mathbf{x}_{\textit{k}}, \mathbf{u}_{\textit{k}}\right) \leq 0 \end{split}$$

MPC scheme is (nominally) stabilizing if there is λ such that $\ell(\mathbf{s}, \mathbf{a}) := L(\mathbf{s}, \mathbf{a}) + \lambda(\mathbf{s}) - \lambda(\mathbf{f}(\mathbf{s}, \mathbf{a})) \ge \kappa(\|\mathbf{s} - \mathbf{s}_{\mathbf{s}}\|), \quad \forall \, \mathbf{s}, \mathbf{a}$ where κ is K_{∞} (+conditions on T)

$$\begin{split} & \textbf{Policy } \boldsymbol{\pi}^{\mathrm{MPC}} \ \textbf{from} \\ & \underset{x,u}{\text{min}} \quad \mathcal{T}\left(\mathbf{x}_{\textit{N}}\right) + \sum_{\textit{k}=0}^{\textit{N}-1} \textit{L}\left(\mathbf{x}_{\textit{k}},\mathbf{u}_{\textit{k}}\right) \\ & \text{s.t.} \quad \mathbf{x}_{\textit{k}+1} = \mathbf{f}\left(\mathbf{x}_{\textit{k}},\mathbf{u}_{\textit{k}}\right), \ \mathbf{x}_{0} = \mathbf{s} \\ & \quad \mathbf{h}\left(\mathbf{x}_{\textit{k}},\mathbf{u}_{\textit{k}}\right) \leq 0 \end{split}$$

Equivalent MPC

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

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

is stable

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

$$\ell\left(\mathbf{s},\mathbf{a}
ight) := L\left(\mathbf{s},\mathbf{a}
ight) + \lambda\left(\mathbf{s}
ight) - \lambda\left(\mathbf{f}\left(\mathbf{s},\mathbf{a}
ight)
ight) \geq \kappa\left(\left\|\mathbf{s}-\mathbf{s}_{\mathbf{s}}
ight\|
ight), \quad \forall \, \mathbf{s},\mathbf{a}$$
where κ is K_{∞} (+conditions on T)

$$\begin{split} & \textbf{Policy } \boldsymbol{\pi}^{\mathrm{MPC}} \ \textbf{from} \\ & \underset{x,u}{\text{min}} \quad \mathcal{T}\left(x_{\textit{N}}\right) + \sum_{\textit{k}=0}^{\textit{N}-1} \textit{L}\left(x_{\textit{k}}, \mathbf{u}_{\textit{k}}\right) \\ & \mathrm{s.t.} \quad x_{\textit{k}+1} = \mathbf{f}\left(x_{\textit{k}}, \mathbf{u}_{\textit{k}}\right), \ x_{0} = \mathbf{s} \\ & \quad \mathbf{h}\left(x_{\textit{k}}, \mathbf{u}_{\textit{k}}\right) < 0 \end{split}$$

Equivalent MPC

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

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

is stable

18 / 33

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

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

$$\text{where } \kappa \text{ is } \mathcal{K}_{\infty} \text{ (+conditions on } T)$$

Remarks

- Modifying the MPC cost is a concept already present in dissipativity theory!
- Aligned with modifying the cost for MPC performance
- → Merge the RL & stability modifications for "stability by design"

S. Gros (NTNU) Fall 2025

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

$$J\left(\pi_{m{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} L\left(\mathbf{s}_k, \mathbf{a}_k
ight)$$

◆ロト ◆部ト ◆差ト ◆差ト を めなび

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

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

$$\begin{array}{ll} \text{Parametrized policy } \pi_{\theta}^{\mathrm{MPC}} \text{ from:} \\ \underset{x,u}{\text{min}} & -\lambda_{\theta}\left(s\right) + \mathcal{T}_{\theta}\left(x_{\textit{N}}\right) + \sum_{k=0}^{N-1} \textit{L}_{\theta}\left(x_{k}, u_{k}\right) \end{array}$$

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

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

19 / 33

S. Gros (NTNU) Fall 2025

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

$$J\left(oldsymbol{\pi}_{oldsymbol{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} oldsymbol{L}\left(\mathbf{s}_{k}, \mathbf{a}_{k}
ight)$$

- Learning based on L
- Impose constraint:

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

throughout the learning

L_θ different than L due to stability
 + model error

$$\begin{split} & \text{Parametrized policy } \boldsymbol{\pi}_{\theta}^{\mathrm{MPC}} \text{ from:} \\ & \underset{x,u}{\text{min}} \quad -\lambda_{\theta}\left(s\right) + T_{\theta}\left(x_{\textit{N}}\right) + \sum_{\textit{k}=0}^{\textit{N}-1} \textit{L}_{\theta}\left(x_{\textit{k}}, u_{\textit{k}}\right) \\ & \text{s.t.} \quad x_{\textit{k}+1} = f_{\theta}\left(x_{\textit{k}}, u_{\textit{k}}\right), \ x_{0} = s \end{split}$$

 $\mathbf{h}_{\theta}\left(\mathbf{x}_{k},\mathbf{u}_{k}\right)<0$

19/33

S. Gros (NTNU) MPC & RL Fall 2025

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

$$J\left(oldsymbol{\pi}_{oldsymbol{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} oldsymbol{L}\left(\mathbf{s}_k, \mathbf{a}_k
ight)$$

- Learning based on L
- Impose constraint:

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

throughout the learning

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

$$\begin{array}{ll} \textbf{Parametrized policy} \ \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\mathrm{MPC}} \ \text{from:} \\ \underset{\boldsymbol{x},\boldsymbol{u}}{\text{min}} \quad - \lambda_{\boldsymbol{\theta}} \left(\boldsymbol{s} \right) + \mathcal{T}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_{\textit{N}} \right) + \sum_{k=0}^{N-1} \mathcal{L}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_{k}, \boldsymbol{u}_{k} \right) \end{array}$$

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

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

Theorem: under some conditions

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

19 / 33

S. Gros (NTNU) MPC & RL Fall 2025

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

$$J\left(oldsymbol{\pi}_{oldsymbol{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} oldsymbol{L}\left(\mathbf{s}_{k}, \mathbf{a}_{k}
ight)$$

- Learning based on L
- Impose constraint:

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

throughout the learning

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

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

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

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

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

Theorem: under some conditions

- ullet $oldsymbol{\pi}_{oldsymbol{ heta}}^{\mathrm{MPC}}
 ightarrow oldsymbol{\pi}_{\star}$ if $oldsymbol{\pi}_{\star}$ is stabilizing
- $m{\phi} m{\pi}^{\mathrm{MPC}}_{m{ heta}}
 ightarrow \mathsf{best}$ stabilizing policy otherwise

Change of philosophy from "classic" dissipativity framework: stability analysis → stable design

◆ロト ◆昼 ト ◆ 重 ト ・ 重 ・ 夕 Q ○

19 / 33

S. Gros (NTNU) MPC & RL Fall 2025

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

$$J\left(oldsymbol{\pi}_{oldsymbol{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} oldsymbol{L}\left(\mathbf{s}_{k}, \mathbf{a}_{k}
ight)$$

Constraint

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

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:

$$\min_{\mathbf{x},\mathbf{u}} -\lambda_{\theta}(\mathbf{s}) + T_{\theta}(\mathbf{x}_{N}) + \sum_{k=0}^{N-1} L_{\theta}(\mathbf{x}_{k}, \mathbf{u}_{k})$$
s.t. $\mathbf{x}_{k+1} = \mathbf{f}_{\theta}(\mathbf{x}_{k}, \mathbf{u}_{k}), \ \mathbf{x}_{0} = \mathbf{s}$

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

Theorem: under some conditions

- ullet $\pi_{m{ heta}}^{\mathrm{MPC}}
 ightarrow m{\pi}_{\star}$ if $m{\pi}_{\star}$ is stabilizing
- $m{\phi} m{\pi}^{\mathrm{MPC}}_{m{ heta}} o \mathsf{best} \; \mathsf{stabilizing} \; \mathsf{policy} \; \mathsf{otherwise}$

Change of philosophy from "classic" dissipativity framework: stability analysis → stable design

19 / 33

S. Gros (NTNU) MPC & RL Fall 2025

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

$$J\left(oldsymbol{\pi}_{oldsymbol{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} oldsymbol{L}\left(\mathbf{s}_{k}, \mathbf{a}_{k}
ight)$$

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

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

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

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

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

Theorem: under some conditions

- ullet $\pi_{ heta}^{\mathrm{MPC}}
 ightarrow \pi_{\star}$ if π_{\star} is stabilizing
- $m{\phi} m{\pi}^{\mathrm{MPC}}_{m{ heta}}
 ightarrow \mathsf{best}$ stabilizing policy otherwise

Change of philosophy from "classic" dissipativity framework: stability analysis → stable design

◆ロト ◆部ト ◆恵ト ◆恵ト ・恵 ・ め Q (*)

19 / 33

S. Gros (NTNU) MPC & RL Fall 2025

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

$$J\left(oldsymbol{\pi}_{oldsymbol{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} oldsymbol{L}\left(\mathbf{s}_{k}, \mathbf{a}_{k}
ight)$$

Extension to stable policy for MDPs?

- Need stability with discount
- Need "stochastic dissipativity"

$$\begin{array}{ll} \text{Parametrized policy } \boldsymbol{\pi}_{\theta}^{\mathrm{MPC}} \text{ from:} \\ \underset{\boldsymbol{x},\boldsymbol{u}}{\text{min}} & -\lambda_{\theta}\left(\boldsymbol{s}\right) + \mathcal{T}_{\theta}\left(\boldsymbol{x}_{\textit{N}}\right) + \sum_{k=0}^{N-1} L_{\theta}\left(\boldsymbol{x}_{k},\boldsymbol{u}_{k}\right) \end{array}$$

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

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework: stability analysis → stable design

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

$$J\left(oldsymbol{\pi}_{oldsymbol{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} oldsymbol{L}\left(\mathbf{s}_{k}, \mathbf{a}_{k}
ight)$$

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:
$$\min_{\mathbf{x},\mathbf{u}} \quad -\lambda_{\theta}\left(\mathbf{s}\right) + \mathcal{T}_{\theta}\left(\mathbf{x}_{\mathit{N}}\right) + \sum_{k=0}^{N-1} \mathcal{L}_{\theta}\left(\mathbf{x}_{\mathit{k}},\mathbf{u}_{\mathit{k}}\right)$$

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

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework: stability analysis → stable design

4□▶ 4□▶ 4□▶ 4□▶ ■ 900

19 / 33

S. Gros (NTNU) MPC & RL Fall 2025

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

$$J\left(oldsymbol{\pi}_{oldsymbol{ heta}}^{ ext{MPC}}
ight) = \sum_{k=0}^{\infty} oldsymbol{L}\left(\mathbf{s}_{k}, \mathbf{a}_{k}
ight)$$

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:
$$\min_{\mathbf{x},\mathbf{u}} \quad -\lambda_{\theta}\left(\mathbf{s}\right) + \mathcal{T}_{\theta}\left(\mathbf{x}_{\textit{N}}\right) + \sum_{k=0}^{N-1} \mathcal{L}_{\theta}\left(\mathbf{x}_{k},\mathbf{u}_{k}\right)$$

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

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

Theorem: under some conditions

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

Change of philosophy from "classic" dissipativity framework: stability analysis → stable design

4 0 1 4 6 1 4 5 1 4 5 1 5

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

S. Gros (NTNU) Fall 2025 19 / 33

Outline Stability-constrained Learning with MPC **Explored questions** Future Prospect – Belief State in RLMPCK MPC & RL

RLMPC

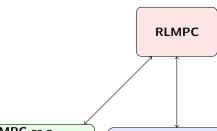
RLMPC

MPC as a solution to MDPs

- $oldsymbol{\sigma}^{\mathrm{MPC}} = oldsymbol{\pi}^{\star} ext{ or } O^{\mathrm{MPC}} = O^{\star}$
- Optimality conditions
- Discounting (or not)
- Storage function



S. Gros (NTNU) MPC & RL



MPC as a solution to MDPs

- $oldsymbol{\pi}^{\mathrm{MPC}} = oldsymbol{\pi}^{\star} ext{ or } Q^{\mathrm{MPC}} = Q^{\star}$
- Optimality conditions
- Discounting (or not)
- Storage function

Safe RL

- \bullet RL + Robust MPC
- Safety of updating parameters "on-the-fly"
- Role of the model

S. Gros (NTNU) MPC & RL



MPC as a solution to MDPs

- $oldsymbol{\pi}^{\mathrm{MPC}} = oldsymbol{\pi}^{\star} ext{ or } \ Q^{\mathrm{MPC}} = Q^{\star}$
- Optimality conditions
- Discounting (or not)
- Storage function

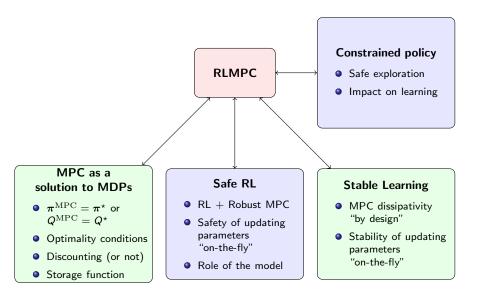
Safe RL

- RL + Robust MPC
- Safety of updating parameters "on-the-fly"
- Role of the model

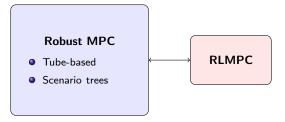
Stable Learning

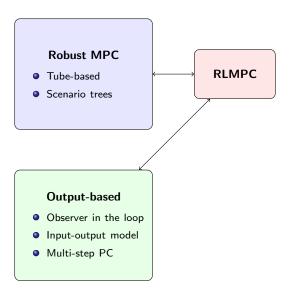
- MPC dissipativity "by design"
- Stability of updating parameters "on-the-fly"

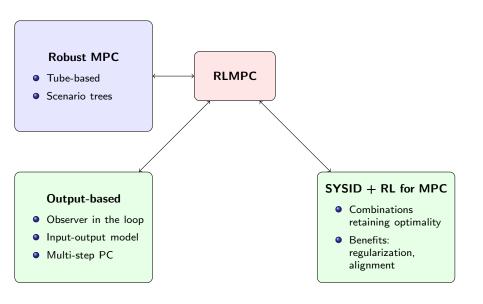
21/33

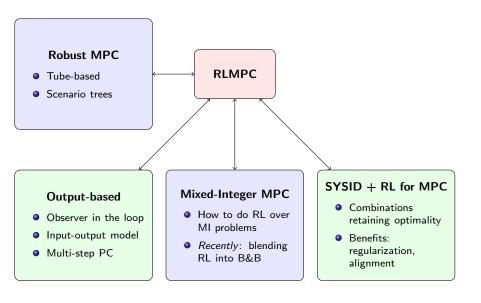


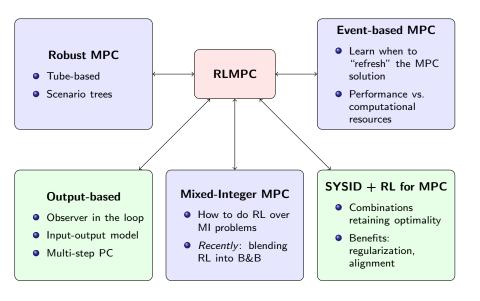
RLMPC





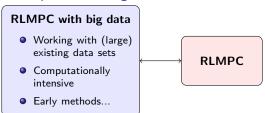


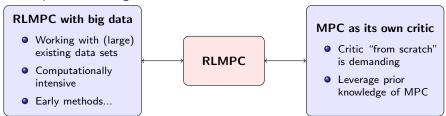




RLMPC







RLMPC with big data Working with (large) existing data sets Computationally intensive Early methods... RLMPC RLMPC MPC as its own critic Critic "from scratch" is demanding Leverage prior knowledge of MPC

Second-order steps

Approximation of

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

|ロ > 4回 > 4回 > 4回 > 回 かへぐ

RLMPC with big data

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

MPC as its own critic

- Critic "from scratch" is demanding
- Leverage prior knowledge of MPC

Second-order steps

Approximation of

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

Theoretical by-products

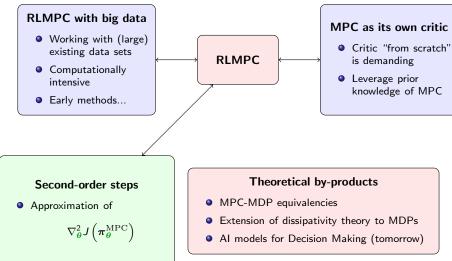
MPC-MDP equivalencies

RLMPC

- Extension of dissipativity theory to MDPs
- Al models for Decision Making (tomorrow)

23 / 33

S. Gros (NTNU) MPC & RL



>50 papers, but most aspects could use more developments, algorithms, software, applications. Welcome onboard!

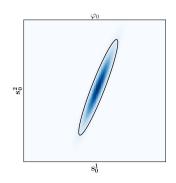
Outline Stability-constrained Learning with MPC Future Prospect – Belief State in RLMPC?

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate s_k from observations $o_{0,...,k}$
- Estimation $\mathbf{o}_{0,...,k} \to \hat{\mathbf{s}}_k$ yields

imperfect knowledge of \mathbf{s}_k

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- ullet Estimate \mathbf{s}_k from observations $\mathbf{o}_{0,...,k}$
- Estimation $\mathbf{o}_{0,...,k} \to \hat{\mathbf{s}}_k$ yields

imperfect knowledge of \mathbf{s}_k



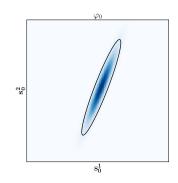
How to "encode" that imperfect knowledge in decision making?

• Belief state: $\mathbf{s}_k \sim \varphi_k \left(\cdot \right)$

<ロ > < 部 > < き > < き > の < で

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate s_k from observations $o_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k

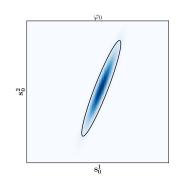


How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k (\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate s_k from observations $o_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k

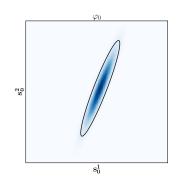


How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k (\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate \mathbf{s}_k from observations $\mathbf{o}_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



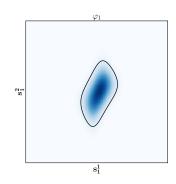
How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k \left(\cdot \right)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to \mathbf{a}_k and $\mathbf{o}_{0,\dots,k+1}$ (and φ_0)

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate s_k from observations $o_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



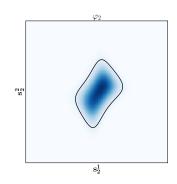
How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to \mathbf{a}_k and $\mathbf{o}_{0,\dots,k+1}$ (and φ_0)

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate s_k from observations $o_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



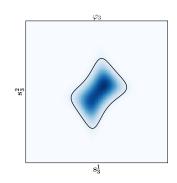
How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to a_k and $o_{0,...,k+1}$ (and φ_0)

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate \mathbf{s}_k from observations $\mathbf{o}_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



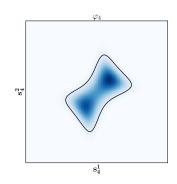
How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to a_k and $o_{0,...,k+1}$ (and φ_0)

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate \mathbf{s}_k from observations $\mathbf{o}_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



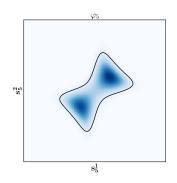
How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to \mathbf{a}_k and $\mathbf{o}_{0,\dots,k+1}$ (and φ_0)

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate \mathbf{s}_k from observations $\mathbf{o}_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



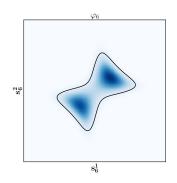
How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to a_k and $o_{0,...,k+1}$ (and φ_0)

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate \mathbf{s}_k from observations $\mathbf{o}_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



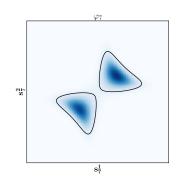
How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to \mathbf{a}_k and $\mathbf{o}_{0,\dots,k+1}$ (and φ_0)

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^* (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate \mathbf{s}_k from observations $\mathbf{o}_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



How to "encode" that imperfect knowledge in decision making?

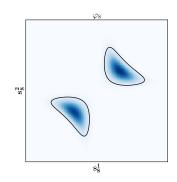
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to a_k and $o_{0,\dots,k+1}$ (and φ_0)

Belief States - What is it about?

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^{\star} (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate s_k from observations $o_{0,...,k}$
- ullet Estimation $\mathbf{o}_{0,...,k}$ $ightarrow \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



How to "encode" that imperfect knowledge in decision making?

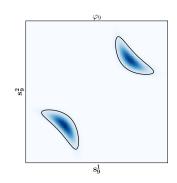
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to \mathbf{a}_k and $\mathbf{o}_{0,\dots,k+1}$ (and φ_0)

Belief States - What is it about?

- In Regular MDPs, state s is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^* (\mathbf{s})$
- Can work with s = raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate \mathbf{s}_k from observations $\mathbf{o}_{0,...,k}$
- Estimation $\mathbf{o}_{0,...,k} \to \hat{\mathbf{s}}_k$ yields

imperfect knowledge of s_k



How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

Belief state has its own dynamics $\varphi_k \to \varphi_{k+1}$, tied to a_k and $o_{0,...,k+1}$ (and φ_0)

Dynamics model

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

Observation model

$$\mathbf{o}_k \sim \mathcal{O}\left[\cdot \,|\, \mathbf{s}_k\, \right]$$

State s_k is "partially known", i.e.

$$\mathbf{s}_{\mathit{k}} \sim \varphi_{\mathit{k}}(\,\cdot\,)$$
 is a belief state

Dynamics model

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

Observation model

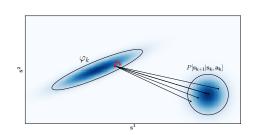
$$\mathbf{o}_k \sim \mathcal{O}\left[\cdot \mid \mathbf{s}_k\right]$$

State s_k is "partially known", i.e.

$$\mathbf{s}_{\it k} \sim arphi_{\it k}(\,\cdot\,)$$
 is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$\varphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot | \mathbf{s}_k, \mathbf{a}_k\right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$
$$:= \mathcal{T}_{\mathbf{a}_k} \varphi_k$$



Dynamics model

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

Observation model

$$\mathbf{o}_k \sim \mathcal{O}\left[\cdot \mid \mathbf{s}_k\right]$$

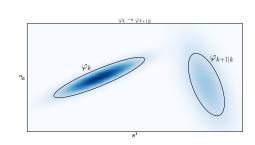
State s_k is "partially known", i.e.

$$\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$$
 is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$arphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot | \mathbf{s}_k, \mathbf{a}_k \right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$

:= $\mathcal{T}_{\mathbf{a}_k} \varphi_k$



Dynamics model

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

Observation model

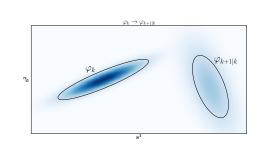
$$\mathbf{o}_k \sim \mathcal{O}\left[\cdot \mid \mathbf{s}_k\right]$$

State s_k is "partially known", i.e.

$$\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$$
 is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$\varphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot | \mathbf{s}_k, \mathbf{a}_k\right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$
$$:= \mathcal{T}_{\mathbf{a}_k} \varphi_k$$



Posterior to observation o_{k+1} , new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

$$\varphi_{k+1}(\,\cdot\,) = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right) = \alpha \underbrace{\mathcal{O}\left[\mathbf{o}_{k+1} | \cdot\,\right] \varphi_{k+1|k}(\,\cdot\,)}_{\text{Bayesian inference}} \qquad \alpha \text{ is a normalization}$$

Dynamics model

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

Observation model

$$\mathbf{o}_k \sim \mathcal{O}\left[\,\cdot\,|\,\mathbf{s}_k\,
ight]$$

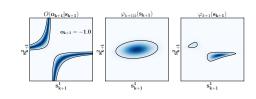
State s_k is "partially known", i.e.

$$\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$$
 is a belief state

 $s_k \sim \varphi_k(\cdot)$ is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$\varphi_{k+1|k}(\cdot) = \int \mathbb{P}[\cdot | \mathbf{s}_k, \mathbf{a}_k] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$
$$:= \mathcal{T}_{\mathbf{a}_k} \varphi_k$$



Posterior to observation o_{k+1} , new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

$$\varphi_{k+1}(\,\cdot\,) = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right) = \alpha \underbrace{\mathcal{O}\left[\mathbf{o}_{k+1}|\,\cdot\,\right] \varphi_{k+1|k}(\,\cdot\,)}_{\text{Bayesian inference}} \qquad \alpha \text{ is a normalization}$$

Dynamics model

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

Observation model

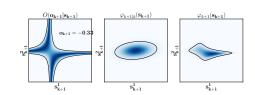
$$\mathbf{o}_k \sim \mathcal{O}\left[\,\cdot\,|\,\mathbf{s}_k\,
ight]$$

State s_k is "partially known", i.e.

$$\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$$
 is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$\varphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot \mid \mathbf{s}_k, \mathbf{a}_k\right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$
$$:= \mathcal{T}_{\mathbf{a}_k} \varphi_k$$



Posterior to observation o_{k+1} , new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

$$\varphi_{k+1}(\,\cdot\,) = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right) = \alpha \underbrace{\mathcal{O}\left[\mathbf{o}_{k+1} | \cdot\,\right] \varphi_{k+1|k}(\,\cdot\,)}_{\text{Bayesian inference}} \qquad \alpha \text{ is a normalization}$$

Dynamics model

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

Observation model

$$\mathbf{o}_k \sim \mathcal{O}\left[\,\cdot\,|\,\mathbf{s}_k\,
ight]$$

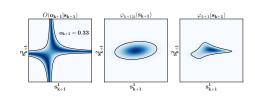
State s_k is "partially known", i.e.

$$\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$$
 is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$arphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot | \mathbf{s}_k, \mathbf{a}_k\right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$

:= $\mathcal{T}_{\mathbf{a}_k} \varphi_k$



Posterior to observation o_{k+1} , new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

$$\varphi_{k+1}(\,\cdot\,) = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right) = \alpha \underbrace{\mathcal{O}\left[\mathbf{o}_{k+1} | \cdot\,\right] \varphi_{k+1|k}(\,\cdot\,)}_{\text{Bayesian inference}} \qquad \alpha \text{ is a normalization}$$

Dynamics model

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

Observation model

$$\mathbf{o}_k \sim \mathcal{O}\left[\,\cdot\,|\,\mathbf{s}_k\,
ight]$$

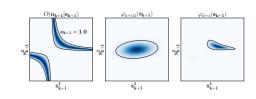
State s_k is "partially known", i.e.

$$\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$$
 is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$arphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot | \mathbf{s}_k, \mathbf{a}_k\right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$

:= $\mathcal{T}_{\mathbf{a}_k} \varphi_k$



Posterior to observation o_{k+1} , new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

$$\varphi_{k+1}(\,\cdot\,) = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right) = \alpha \underbrace{\mathcal{O}\left[\mathbf{o}_{k+1} | \cdot\,\right] \varphi_{k+1|k}(\,\cdot\,)}_{\text{Bayesian inference}} \qquad \alpha \text{ is a normalization}$$

Dynamics model

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

Observation model

$$\mathbf{o}_k \sim \mathcal{O}\left[\cdot \mid \mathbf{s}_k\right]$$

State s_k is "partially known", i.e.

 $\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$ is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$\varphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot | \mathbf{s}_k, \mathbf{a}_k\right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$

$$:= \mathcal{T}_{\mathbf{a}_k} \varphi_k$$

Remarks state transition

$$\varphi_{k+1} = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right)$$

Is in general nonlinear

Posterior to observation o_{k+1} , new info corrects $\varphi_{k+1|k} \to \varphi_{k+1}$:

$$\varphi_{k+1}(\,\cdot\,) = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right) = \alpha \underbrace{\mathcal{O}\left[\mathbf{o}_{k+1} | \cdot\,\right] \, \varphi_{k+1|k}(\,\cdot\,)}_{\text{Bayesian inference}} \qquad \alpha \text{ is a normalization}$$

Dynamics model

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

Observation model

$$\mathbf{o}_k \sim \mathcal{O}\left[\,\cdot\,|\,\mathbf{s}_k\,
ight]$$

State s_k is "partially known", i.e.

 $\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$ is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$\varphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot | \mathbf{s}_k, \mathbf{a}_k\right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$

$$:= \mathcal{T}_{\mathbf{a}_k} \varphi_k$$

Remarks state transition

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$

- Is in general nonlinear
- Is deterministic for o_{k+1} known

Posterior to observation o_{k+1} , new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

$$\varphi_{k+1}(\,\cdot\,) = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right) = \alpha \underbrace{\mathcal{O}\left[\mathbf{o}_{k+1}|\,\cdot\,\right] \, \varphi_{k+1|k}(\,\cdot\,)}_{\text{Bayesian inference}} \qquad \alpha \text{ is a normalization}$$

Dynamics model

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

Observation model

$$\mathbf{o}_k \sim \mathcal{O}\left[\cdot \mid \mathbf{s}_k\right]$$

State s_k is "partially known", i.e.

 $\mathbf{s}_k \sim \varphi_k(\,\cdot\,)$ is a belief state

Prior to observation o_{k+1} , φ_{k+1} is

$$\varphi_{k+1|k}(\cdot) = \int \mathbb{P}\left[\cdot | \mathbf{s}_k, \mathbf{a}_k\right] \varphi_k(\mathbf{s}_k) d\mathbf{s}_k$$
$$:= \mathcal{T}_{\mathbf{a}_k} \varphi_k$$

Remarks state transition

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$

- Is in general nonlinear
- Is deterministic for o_{k+1} known
- Is stochastic prior to o_{k+1} known, with

$$\mathbb{E}_{\mathbf{o}_{k+1}|\varphi_k,\mathbf{a}_k}\left[\varphi_{k+1}\right] = \varphi_{k+1|k}$$

Posterior to observation o_{k+1} , new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

$$\varphi_{k+1}(\,\cdot\,) = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right) = \alpha \underbrace{\mathcal{O}\left[\mathbf{o}_{k+1}|\,\cdot\,\right] \varphi_{k+1|k}(\,\cdot\,)}_{\text{Bayesian inference}} \qquad \alpha \text{ is a normalization}$$

Dynamics model

$$s_{k+1} \sim \mathcal{N}\left(As_k + Ba_k, \, \Sigma_s\right)$$

Observation model (nonlinear)

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\Sigma_s = \frac{1}{5} \emph{I}, \quad \sigma_o = 1$$

27 / 33

S. Gros (NTNU) MPC & RL Fall 2025

Dynamics model

$$s_{\mathit{k}+1} \sim \mathcal{N}\left(As_{\mathit{k}} + Ba_{\mathit{k}}, \, \Sigma_{\mathit{s}}\right)$$

Observation model (nonlinear)

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} \mathbf{M} \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\Sigma_{
m s}=rac{1}{5}I,\quad \sigma_{
m o}=1$$

Posterior to observation o_{k+1} , belief state $\varphi_{k+1} = \mathcal{B}(\varphi_k, o_{k+1}, a_k)$ is not Gaussian

4□ N 4 ₫ N 4 ₹ N ₹ N ₹ N 0 (N)

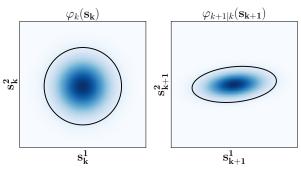
Dynamics model

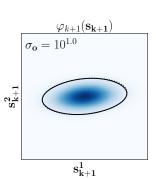
$$s_{k+1} \sim \mathcal{N}\left(\textit{A}s_k + \textit{B}a_k,\, \Sigma_s\right)$$

Observation model (nonlinear)

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} \mathsf{M} \mathbf{s}_k, \, \sigma_{\mathbf{o}} \right)$$

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$\Sigma_{s} = \frac{1}{5}I, \quad \sigma_{o} = 1$$





Posterior to observation o_{k+1} , belief state $\varphi_{k+1} = \mathcal{B}(\varphi_k, o_{k+1}, a_k)$ is not Gaussian

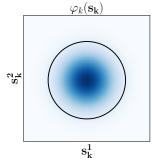
Dynamics model

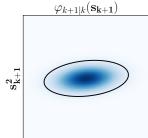
$$s_{k+1} \sim \mathcal{N}\left(As_k + Ba_k, \, \Sigma_s\right)$$

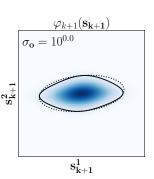
Observation model (nonlinear)

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} \mathsf{M} \mathbf{s}_k, \, \sigma_{\mathbf{o}} \right)$$

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$\Sigma_{s} = \frac{1}{5}I, \quad \sigma_{o} = 1$$







Posterior to observation o_{k+1} , belief state $\varphi_{k+1} = \mathcal{B}(\varphi_k, o_{k+1}, a_k)$ is not Gaussian

S. Gros (NTNU) Fall 2025 27/33

 $\mathbf{s}^1_{\mathbf{k}_{\perp 1}}$

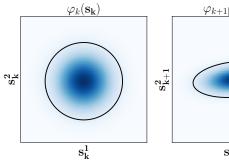
Dynamics model

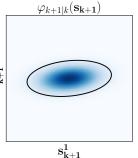
$$s_{k+1} \sim \mathcal{N}\left(\textit{A}s_k + \textit{B}a_k,\, \Sigma_s\right)$$

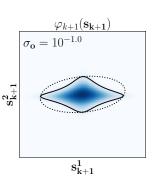
Observation model (nonlinear)

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} \mathsf{M} \mathbf{s}_k, \, \sigma_{\mathbf{o}} \right)$$

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$\Sigma_{s} = \frac{1}{5}I, \quad \sigma_{o} = 1$$







Posterior to observation o_{k+1} , belief state $\varphi_{k+1} = \mathcal{B}(\varphi_k, o_{k+1}, a_k)$ is not Gaussian

S. Gros (NTNU) 27/33

Dynamics model

$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k,\, \Sigma_s\right)$$

Observation model (nonlinear)

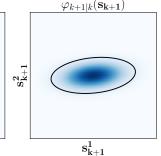
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

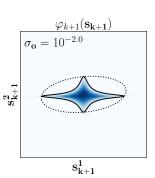
$$\varphi_k(\mathbf{s}_k)$$

 \mathbf{s}^1_{k}

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$\Sigma_{s} = \frac{1}{5}I, \quad \sigma_{o} = 1$$

$$\varphi_{k+1|k}(\mathbf{s_{k+1}})$$





Posterior to observation o_{k+1} , belief state $\varphi_{k+1} = \mathcal{B}(\varphi_k, o_{k+1}, a_k)$ is not Gaussian

Dynamics model

$$s_{\mathit{k}+1} \sim \mathcal{N}\left(\mathit{A}s_{\mathit{k}} + \mathit{B}a_{\mathit{k}},\, \Sigma_{s}\right)$$

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Observation model

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} \mathbf{M} \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_s = \frac{1}{5} I, \quad \sigma_o = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}\left[\mathbf{s}\right] \quad (o \mathsf{Kalman filter})$

28 / 33

S. Gros (NTNU) MPC & RL Fall 2025

Dynamics model

$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k, \, \Sigma_s\right)$$

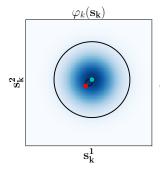
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

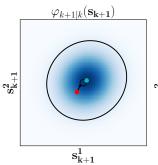
Observation model

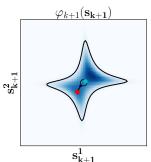
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_{\rm s} = rac{1}{5}I, \quad \sigma_{
m o} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$ (→ Kalman filter)







Dynamics model

$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k,\, \Sigma_{\mathrm{s}}\right)$$

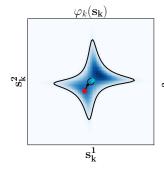
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $\Sigma_{s} = \frac{1}{5}I, \quad \sigma_{o} = 1$

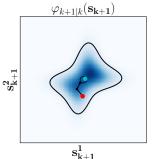
Observation model

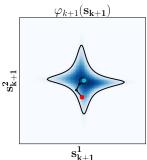
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_{\rm s} = \frac{1}{5}I, \quad \sigma_{\rm o}$$

Policy: e.g. DLQR over expected state, i.e. $a_k = -K \cdot \mathbb{E}_{s \sim \varphi_k}[s]$ (\to Kalman filter)







Dynamics model

$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k,\, \Sigma_s\right)$$

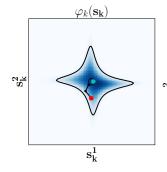
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

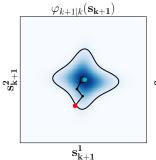
Observation model

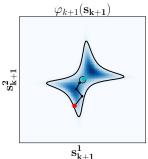
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_{\mathrm{s}} = \frac{1}{5}I, \quad \sigma_{\mathrm{o}} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$ (→ Kalman filter)







Dynamics model

$$s_{\mathit{k}+1} \sim \mathcal{N}\left(\mathit{A}s_{\mathit{k}} + \mathit{B}a_{\mathit{k}},\, \Sigma_{s}\right)$$

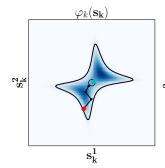
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

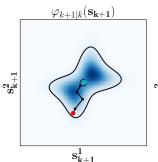
Observation model

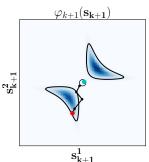
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} M \mathbf{s}_k, \, \sigma_\mathbf{o}\right)$$

$$\Sigma_{\rm s} = rac{1}{5}I, \quad \sigma_{
m o} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$ (→ Kalman filter)







28 / 33

True state and $\circ = (0,0)$

S. Gros (NTNU)

Fall 2025

Dynamics model

$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k, \, \Sigma_{\mathrm{s}}\right)$$

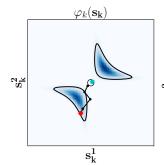
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

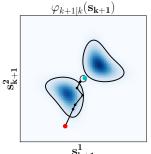
Observation model

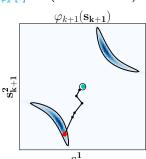
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_{\rm s} = rac{1}{5}I, \quad \sigma_{
m o} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$ (→ Kalman filter)







 $\mathbf{s}_{\mathbf{k}+1}^{1}$

 $\mathbf{s}_{\mathbf{k}+1}^{1}$

Dynamics model

$$s_{\mathit{k}+1} \sim \mathcal{N}\left(\mathit{A}s_{\mathit{k}} + \mathit{B}a_{\mathit{k}},\, \Sigma_{\mathrm{s}}\right)$$

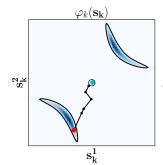
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

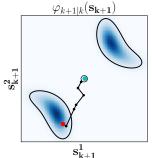
Observation model

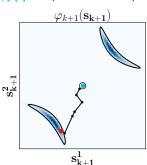
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_{\rm s} = \frac{1}{5}I, \quad \sigma_{\rm o} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$ (→ Kalman filter)







Dynamics model

$$s_{\mathit{k}+1} \sim \mathcal{N}\left(\mathit{A}s_{\mathit{k}} + \mathit{B}a_{\mathit{k}},\, \Sigma_{\mathrm{s}}\right)$$

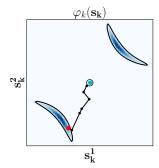
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $\Sigma_{\rm s} = \frac{1}{5}I, \quad \sigma_{\rm o} = 1$

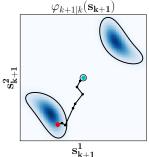
Observation model

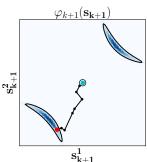
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} M \mathbf{s}_k, \ \sigma_{\mathbf{o}}\right)$$

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$ (→ Kalman filter)







Dynamics model

$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k, \, \Sigma_s\right)$$

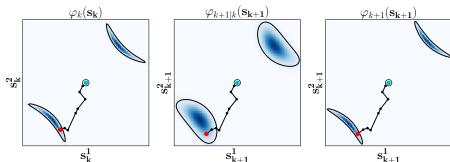
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Observation model

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_{\mathrm{s}} = \frac{1}{5}I, \quad \sigma_{\mathrm{o}} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}\left[\mathbf{s}\right] \quad \ (\rightarrow \mathsf{Kalman} \; \mathsf{filter})$



True state and $\circ = (0,0)$

28 / 33

S. Gros (NTNU) MPC & RL Fall 2025

Dynamics model

$$s_{\mathit{k}+1} \sim \mathcal{N}\left(\mathit{A}s_{\mathit{k}} + \mathit{B}a_{\mathit{k}},\, \Sigma_{\mathit{s}}\right)$$

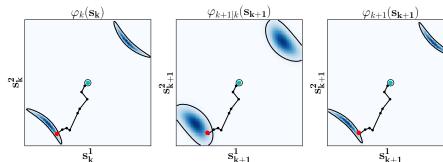
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Observation model

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} \mathbf{M} \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_{\rm s} = rac{1}{5}I, \quad \sigma_{
m o} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}\left[\mathbf{s}\right] \quad (o \mathsf{Kalman} \; \mathsf{filter})$



True state and
$$\circ = (0,0)$$

Dynamics model

$$s_{k+1} \sim \mathcal{N}\left(As_k + Ba_k, \, \Sigma_s\right)$$

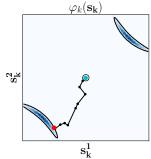
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

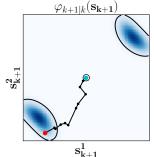
Observation model

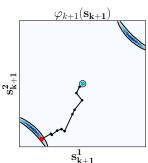
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

$$\Sigma_{\mathrm{s}} = \frac{1}{5}I, \quad \sigma_{\mathrm{o}} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$ (→ Kalman filter)







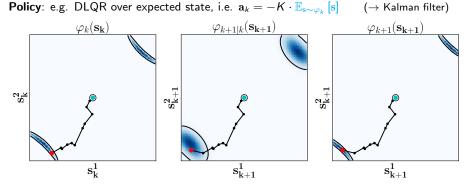
Dynamics model

$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k, \, \Sigma_{\mathrm{s}}\right)$$

 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $\Sigma_{s} = \frac{1}{5}I, \quad \sigma_{o} = 1$

Observation model

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$



True state and $\circ = (0,0)$

28 / 33

S. Gros (NTNU) MPC & RL Fall 2025

Dynamics model

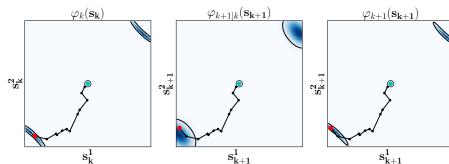
$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k, \, \Sigma_{\mathrm{s}}\right)$$

Observation model $\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{\top} \mathcal{M} \mathbf{s}_k, \ \sigma_{\mathbf{o}}\right)$

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\Sigma_s = \frac{1}{5} \emph{I}, \quad \sigma_o = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k} [\mathbf{s}]$ (\rightarrow Kalman filter)



True state and $\circ = (0,0)$

28 / 33

S. Gros (NTNU) MPC & RL Fall 2025

Dynamics model

$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k, \, \Sigma_{\mathrm{s}}\right)$$

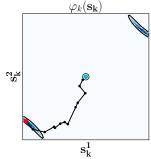
 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $\Sigma_{\rm s} = \frac{1}{5}I, \quad \sigma_{\rm o} = 1$

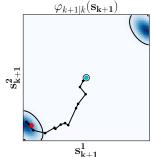
Observation model

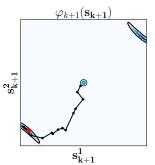
$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

Policy: e.g. DLQR over expected state, i.e.
$$\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$$
 (\rightarrow Kalm

(→ Kalman filter)







True state and $\circ = (0,0)$

S. Gros (NTNU)

Fall 2025

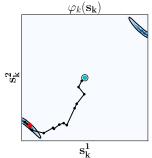
Dynamics model

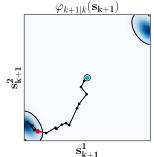
$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k, \, \Sigma_{\mathrm{s}}\right)$$

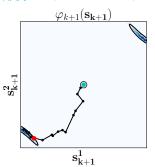
Observation model $\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} M \mathbf{s}_k, \, \sigma_\mathbf{o}\right)$

$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$\Sigma_{s} = \frac{1}{5}I, \quad \sigma_{o} = 1$$

Policy: e.g. DLQR over expected state, i.e. $\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k}[\mathbf{s}]$ (→ Kalman filter)







Dynamics model

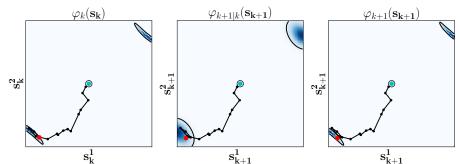
$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k,\, \Sigma_s\right)$$

 $A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ .1 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $\Sigma_{s} = \frac{1}{5}I, \quad \sigma_{o} = 1$

Observation model

$$\mathbf{o}_k \sim \mathcal{N}\left(\mathbf{s}_k^{ op} M \mathbf{s}_k, \, \sigma_{\mathbf{o}}\right)$$

Policy: e.g. DLQR over expected state, i.e.
$$\mathbf{a}_k = -K \cdot \mathbb{E}_{\mathbf{s} \sim \varphi_k} [\mathbf{s}]$$
 (\rightarrow Kalman filter)



True state and $\circ = (0,0)$

MDPs with Belief States

Policy π should select actions \mathbf{a}_k for the belief state φ_k

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

◆ロト ◆部ト ◆差ト ◆差ト を めらる

Policy π should select actions \mathbf{a}_k for the belief state φ_k

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

Underlying value functions

$$\begin{aligned} Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) &= \mathbb{E}_{\mathbf{s}_{k}\sim\varphi_{k}}\left[L\left(\mathbf{s}_{k},\mathbf{a}_{k}\right)\right] \\ &+ \gamma \mathbb{E}\left[\left.V^{\star}\left(\varphi_{k+1}\right)\right|\,\varphi_{k},\mathbf{a}_{k}\right] \\ V^{\star}\left(\varphi_{k}\right) &= \min_{\mathbf{a}_{k}}\,Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) \\ \pi^{\star}\left(\varphi_{k}\right) &= \arg\min_{\mathbf{a}_{k}}\,Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) \end{aligned}$$

◆ロト ◆個ト ◆差ト ◆差ト を めるの

Policy π should select actions \mathbf{a}_k for the belief state φ_k

$$J(oldsymbol{\pi}) = \mathbb{E}\left[\left.\sum_{k=0}^{\infty}\,\gamma^{k}L\left(\mathbf{s}_{k},\mathbf{a}_{k}
ight)\,
ight|\,\mathbf{a}_{k} = oldsymbol{\pi}\left(arphi_{k}
ight)
ight]$$

Underlying value functions

$$egin{aligned} Q^{\star}\left(arphi_{k},\mathbf{a}_{k}
ight) &= \mathbb{E}_{\mathbf{s}_{k}\simarphi_{k}}\left[L\left(\mathbf{s}_{k},\mathbf{a}_{k}
ight)
ight] \ &+ \gamma\mathbb{E}\left[\left.V^{\star}\left(arphi_{k+1}
ight)
ight|\left.arphi_{k},\mathbf{a}_{k}
ight] \end{aligned} \ V^{\star}\left(arphi_{k}
ight) &= \min_{\mathbf{a}_{k}} \,Q^{\star}\left(arphi_{k},\mathbf{a}_{k}
ight) \ egin{aligned} oldsymbol{\pi}^{\star}\left(arphi_{k}
ight) &= rg\min_{\mathbf{a}_{k}} \,Q^{\star}\left(arphi_{k},\mathbf{a}_{k}
ight) \end{aligned}$$

where

$$\mathbb{E}\left[V^{\star}\left(\varphi_{k+1}\right) \mid \varphi_{k}, \mathbf{a}_{k}\right] = \int V^{\star}\left(\mathcal{B}\left(\varphi_{k}, \mathbf{o}_{k+1}, \mathbf{a}_{k}\right)\right) \rho\left[\mathbf{o}_{k+1} \mid \varphi_{k}, \mathbf{a}_{k}\right] \mathrm{d}\mathbf{o}_{k+1}$$

Policy π should select actions $\mathbf{a}_{\mathbf{k}}$ for the belief state $\varphi_{\mathbf{k}}$

$$J(oldsymbol{\pi}) = \mathbb{E}\left[\left.\sum_{k=0}^{\infty}\,\gamma^{k}L\left(\mathbf{s}_{k},\mathbf{a}_{k}
ight)\,
ight|\,\mathbf{a}_{k} = oldsymbol{\pi}\left(arphi_{k}
ight)
ight]$$

A belief state MDP is not only accounting for the uncertainty on s. It also selects actions that optimize "information gathering", conducive to taking better decisions later.

Underlying value functions

$$\begin{aligned} Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) &= \mathbb{E}_{\mathbf{s}_{k}\sim\varphi_{k}}\left[L\left(\mathbf{s}_{k},\mathbf{a}_{k}\right)\right] \\ &+ \gamma \mathbb{E}\left[V^{\star}\left(\varphi_{k+1}\right) \mid \varphi_{k},\mathbf{a}_{k}\right] \\ V^{\star}\left(\varphi_{k}\right) &= \min_{\mathbf{a}_{k}} \ Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) \\ \pi^{\star}\left(\varphi_{k}\right) &= \arg\min_{\mathbf{a}_{k}} \ Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) \end{aligned}$$

where

$$\mathbb{E}\left[V^{\star}\left(\varphi_{k+1}\right) \mid \varphi_{k}, \mathbf{a}_{k}\right] = \int V^{\star}\left(\mathcal{B}\left(\varphi_{k}, \mathbf{o}_{k+1}, \mathbf{a}_{k}\right)\right) \rho\left[\mathbf{o}_{k+1} \mid \varphi_{k}, \mathbf{a}_{k}\right] d\mathbf{o}_{k+1}$$

Policy π should select actions $\mathbf{a}_{\mathbf{k}}$ for the belief state $\varphi_{\mathbf{k}}$

$$J(oldsymbol{\pi}) = \mathbb{E}\left[\left.\sum_{k=0}^{\infty}\,\gamma^{k}L\left(\mathbf{s}_{k},\mathbf{a}_{k}
ight)\,
ight|\,\mathbf{a}_{k} = oldsymbol{\pi}\left(arphi_{k}
ight)
ight]$$

Underlying value functions

$$\begin{split} Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) &= \mathbb{E}_{\mathbf{s}_{k}\sim\varphi_{k}}\left[L\left(\mathbf{s}_{k},\mathbf{a}_{k}\right)\right] \\ &+ \gamma \mathbb{E}\left[\left.V^{\star}\left(\varphi_{k+1}\right)\right|\,\varphi_{k},\mathbf{a}_{k}\right] \\ V^{\star}\left(\varphi_{k}\right) &= \min_{\mathbf{a}_{k}}\,Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) \\ \pi^{\star}\left(\varphi_{k}\right) &= \operatorname{\mathsf{arg\,min}}\,Q^{\star}\left(\varphi_{k},\mathbf{a}_{k}\right) \end{split}$$

A belief state MDP is not only accounting for the uncertainty on s. It also selects actions that optimize "information gathering", conducive to taking better decisions later.

Can we carry the MDP solution in a repeated planning framework (MPC-like)?

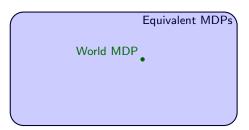
where

$$\mathbb{E}\left[V^{\star}\left(\varphi_{k+1}\right) \mid \varphi_{k}, \mathbf{a}_{k}\right] = \int V^{\star}\left(\mathcal{B}\left(\varphi_{k}, \mathbf{o}_{k+1}, \mathbf{a}_{k}\right)\right) \rho\left[\mathbf{o}_{k+1} \mid \varphi_{k}, \mathbf{a}_{k}\right] d\mathbf{o}_{k+1}$$

World MDP

 $\ensuremath{\mathsf{RLMPC}}$ theoretical pathway

《□▶ 《圖▶ 《圖▶ 《圖▶ ■ 釣@@



RLMPC theoretical pathway

Establish the set of equivalent MDPs



RLMPC theoretical pathway

- Establish the set of equivalent MDPs
- Find a deterministic subset in that set



RLMPC theoretical pathway

- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

S. Gros (NTNU)

Equivalent MDPs World MDP Deterministic MDPs

RLMPC theoretical pathway

- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?

Equivalent MDPs

World MDP

Deterministic MDPs

RLMPC theoretical pathway

- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?

Deterministic model e.g.

- Work in expected values:
 - $\hat{\varphi}_{k+1} = \mathbb{E}\left[\mathcal{B}\left(\hat{\varphi}_{k}, \mathbf{o}_{k+1}, \mathbf{a}_{k}\right)\right] = \mathcal{T}_{\mathbf{a}_{k}}\hat{\varphi}_{k}$?
- Expected observation:

$$\hat{\varphi}_{k+1} = \mathcal{B}\left(\hat{\varphi}_k, \mathbb{E}\left[\mathbf{o}_{k+1}\right], \mathbf{a}_k\right)$$
 ?

ML of observation:

$$\hat{\varphi}_{k+1} = \mathcal{B}\left(\hat{\varphi}_{k}, \operatorname{ML}\left[\mathbf{o}_{k+1}\right], \mathbf{a}_{k}\right) \quad ?$$

• Generic model $\hat{\mathbf{o}}_{k+1} = \mathcal{M}(\hat{\varphi}_k)$?



RLMPC theoretical pathway

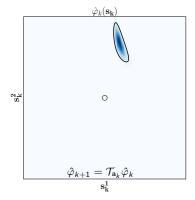
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

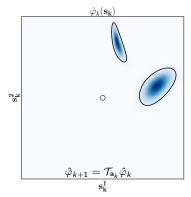
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

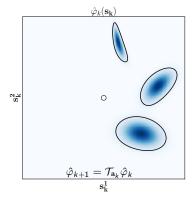
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

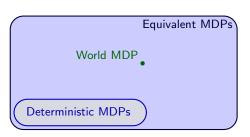
That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

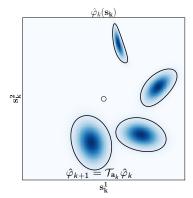
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

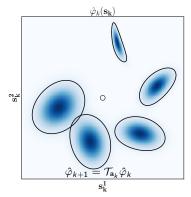
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

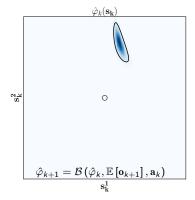
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

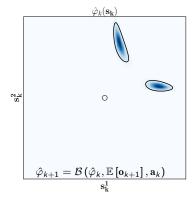
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

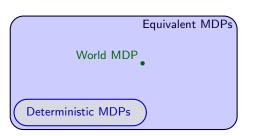
That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

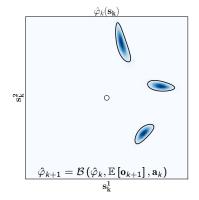
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

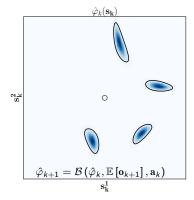
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

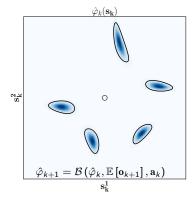
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

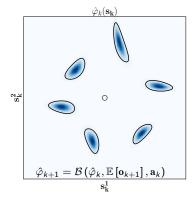
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?





RLMPC theoretical pathway

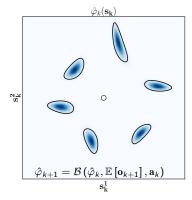
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

Remarks:

- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

$$\varphi_{k+1} = \mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)$$
 ?

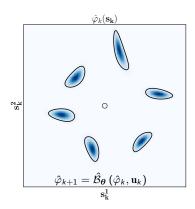


What is the best choice? Is there a good choice? Open question

MPC model

$$\hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\boldsymbol{\theta}} \left(\hat{\varphi}_k, \mathbf{a}_k \right)$$

"resolving" stochasticity over \mathbf{o}_k



MPC model

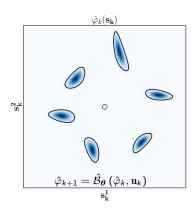
$$\hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\boldsymbol{\theta}} \left(\hat{\varphi}_k, \mathbf{a}_k \right)$$

"resolving" stochasticity over o_k

MPC scheme gives $oldsymbol{\pi}^{\mathrm{MPC}}\left(arphi
ight)=\mathbf{u}_{0}^{\star}$ from

$$\min_{\varphi,\mathbf{u}} \quad T_{\boldsymbol{\theta}}\left(\hat{\varphi}_{N}\right) + \sum_{k=0}^{N-1} \ell_{\boldsymbol{\theta}}\left(\hat{\varphi}_{k}, \mathbf{u}_{k}\right)$$

s.t.
$$\hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\theta} (\hat{\varphi}_k, \mathbf{u}_k), \quad \hat{\varphi}_0 = \varphi$$



MPC model

$$\hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\boldsymbol{\theta}} \left(\hat{\varphi}_k, \mathbf{a}_k \right)$$

"resolving" stochasticity over o_k

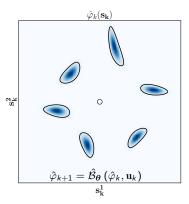
MPC scheme gives $oldsymbol{\pi}^{\mathrm{MPC}}\left(arphi
ight)=\mathbf{u}_{0}^{\star}$ from

$$\min_{\varphi,\mathbf{u}} \quad T_{\theta}\left(\hat{\varphi}_{N}\right) + \sum_{k=0}^{N-1} \ell_{\theta}\left(\hat{\varphi}_{k}, \mathbf{u}_{k}\right)$$

s.t.
$$\hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\theta} (\hat{\varphi}_k, \mathbf{u}_k), \quad \hat{\varphi}_0 = \varphi$$

Remark: optimal ℓ_{θ} is nonlinear in $\hat{\varphi}_k$ i.e.

$$\ell_{oldsymbol{ heta}}\left(\hat{arphi}_{k},\mathbf{u}_{k}
ight)
eq\mathbb{E}_{\mathbf{s}\sim\hat{arphi}_{k}}[L_{oldsymbol{ heta}}\left(\mathbf{s},\mathbf{u}_{k}
ight)]$$



MPC model

$$\hat{arphi}_{k+1} = \hat{\mathcal{B}}_{oldsymbol{ heta}}\left(\hat{arphi}_{k}, \mathbf{a}_{k}
ight)$$

"resolving" stochasticity over o_k

MPC scheme gives $oldsymbol{\pi}^{\mathrm{MPC}}\left(arphi
ight)=\mathbf{u}_{0}^{\star}$ from

$$\min_{\varphi,\mathbf{u}} \quad T_{\theta}\left(\hat{\varphi}_{N}\right) + \sum_{k=0}^{N-1} \ell_{\theta}\left(\hat{\varphi}_{k}, \mathbf{u}_{k}\right)$$

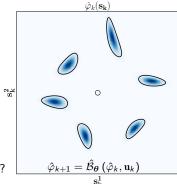
s.t.
$$\hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\theta} (\hat{\varphi}_k, \mathbf{u}_k), \quad \hat{\varphi}_0 = \varphi$$

Questions

- Algorithmic to carry $\hat{\varphi}_{0,...,N}$ in optimization?
- And to evaluate ℓ_{θ} ($\hat{\varphi}_k, \mathbf{u}_k$)?
- Good choice of $\hat{\mathcal{B}}_{\theta}$?
- Integration with state observer?

Remark: optimal ℓ_{θ} is nonlinear in $\hat{\varphi}_{k}$ i.e.

$$\ell_{oldsymbol{ heta}}\left(\hat{arphi}_{k},\mathbf{u}_{k}
ight)
eq\mathbb{E}_{\mathbf{s}\sim\hat{arphi}_{k}}[L_{oldsymbol{ heta}}\left(\mathbf{s},\mathbf{u}_{k}
ight)]$$



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

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: RI over MPC

Beyond MPC – Model-based Decisions and AI for decisions (Tomorrow)

Thanks for your attention!



ResearchGate



Google Scholar