Orientation

What we have seen:

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

Orientation

What we have seen:

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

What we will do next: RL over MPC

- Safe & Stable RL over MPC *(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)*

RL and MPC
Safety, Stability, Belief State

Sebastien Gros

Dept. of Cybernetic, NTNU
Faculty of Information Tech.

Freiburg PhD School

Outline

Outline

Robust MPC - Uncertainty model

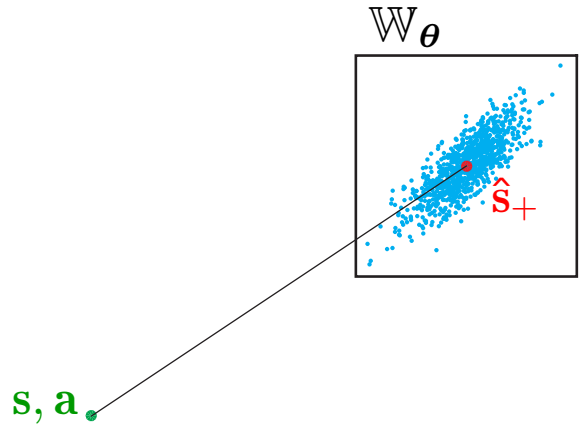
$$\text{True system:} \quad \mathbf{s}_+ \sim \mathbb{P}\left[\cdot \,|\mathbf{s}, \mathbf{a}\right]$$

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

Robust MPC - Uncertainty model

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

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



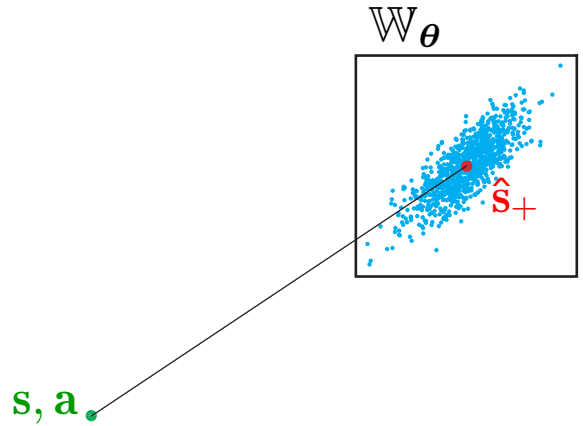
Dispersion: $\mathbf{f}_{\boldsymbol{\theta}}\left(\mathbf{s},\mathbf{a}\right) + \mathbb{W}_{\boldsymbol{\theta}}$ contains the support of $\mathbb{P}\left[\,\cdot\,|\mathbf{s},\mathbf{a}\,\right]$, i.e.

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

with probability 1

Robust MPC - Uncertainty model

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



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

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

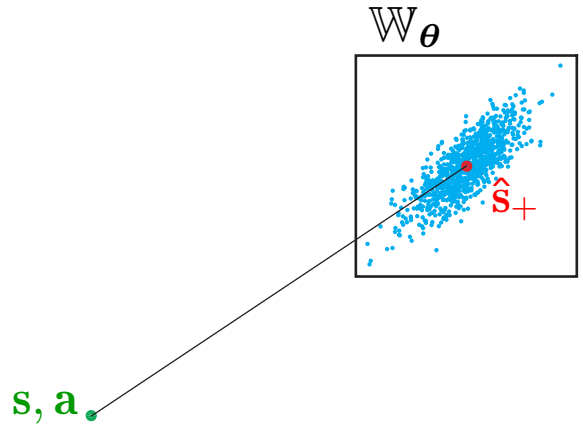
with probability 1

Remarks:

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

Robust MPC - Uncertainty model

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



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

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

with probability 1

Condition

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

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

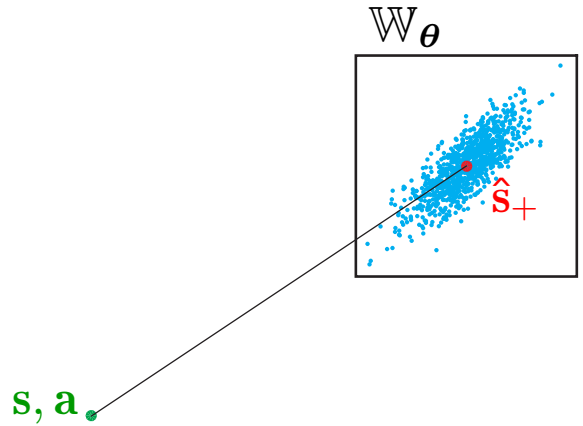
$\rightarrow$ constraints on $\boldsymbol{\theta}$

Remarks:

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

Robust MPC - Uncertainty model

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



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

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

with probability 1

Remarks:

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

Condition

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

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

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

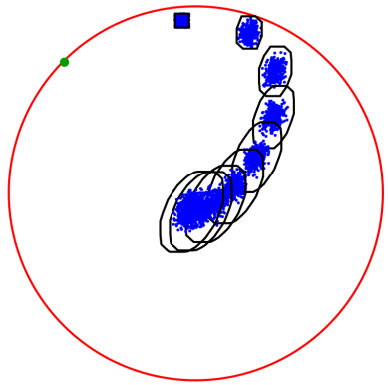
Containing the model-system mismatch becomes constraints in the parameters $\boldsymbol{\theta}$. Constraints can be readily formulated in terms of data.

Safe policies via Robust (N)MPC

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

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

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



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

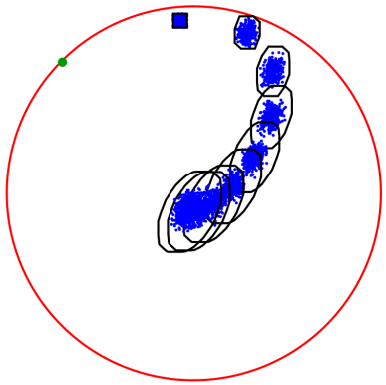
Safe policies via Robust (N)MPC

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

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

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

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

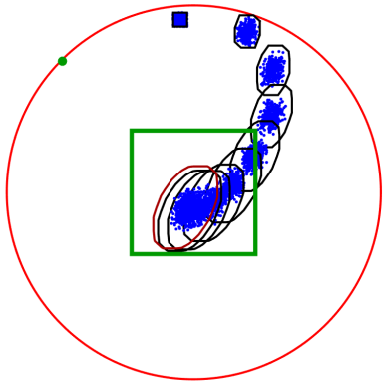


- $\mathbf{x}_{0,\ldots,N}$ is the propagation of the state dispersion
- max cost treats worst-case scenario, required for stability arguments (classical stability)
- $\mathbf{w} = \{\mathbf{w}_0, \ldots, \mathbf{w}_N\}$ is the disturbance with $\mathbf{w}_k \in \mathbb{W}_{\boldsymbol{\theta}}$
- $\mathbf{x}_{1,\ldots,N-1}(\mathbf{u},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

Safe policies via Robust (N)MPC

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

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



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

Safe policies via Robust (N)MPC

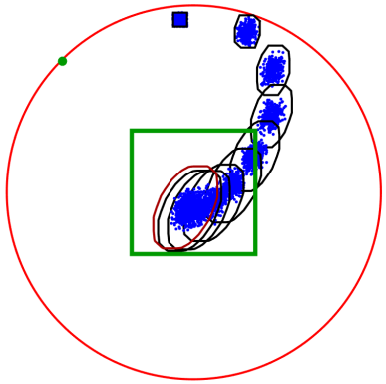
Robust (N)MPC delivers policy $\pi_{\boldsymbol{\theta}}(\mathbf{x}_0) = \mathbf{u}_0^\star$ from

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

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

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

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



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

Safe policies via Robust (N)MPC

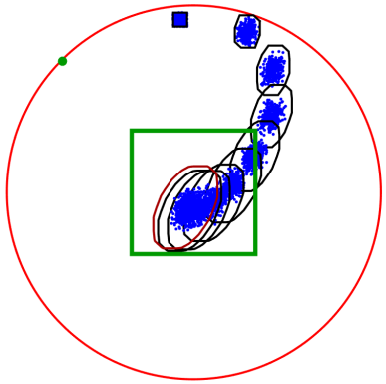
Robust (N)MPC delivers policy $\pi_{\boldsymbol{\theta}}(\mathbf{x}_0) = \mathbf{u}_0^\star$ from

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

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

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

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



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

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

Safe policies via Robust (N)MPC

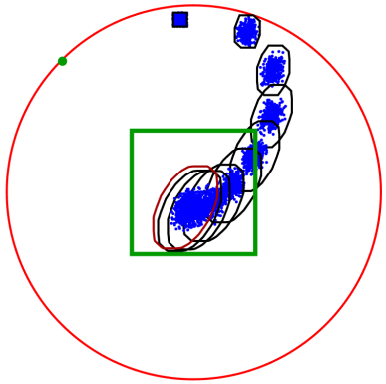
Robust (N)MPC delivers policy $\pi_{\boldsymbol{\theta}}(\mathbf{x}_0) = \mathbf{u}_0^\star$ from

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

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

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

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



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

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

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

Safe Learning via Robust MPC

Robust NMPC parameters $\boldsymbol{\theta}$

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

Condition

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

enforces safety through $\boldsymbol{\theta}$

Safe Learning via Robust MPC

Robust NMPC parameters $\boldsymbol{\theta}$

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

Safe Learning via Robust MPC

Robust NMPC parameters $\boldsymbol{\theta}$

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

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

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts θ for performance

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

Condition

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

enforces safety through θ

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

How to do Safe RL?

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

Safe Learning via Robust MPC

Robust NMPC parameters $\boldsymbol{\theta}$

Policy gradient

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

adjusts $\boldsymbol{\theta}$ for performance

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

Condition

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

enforces safety through $\boldsymbol{\theta}$

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

How to do Safe RL?

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

Also reads as:

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

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

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts θ for performance

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

Condition

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

enforces safety through θ

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

How to do Safe RL?

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

Also reads as:

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

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

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

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

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

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

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts θ for performance

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

Condition

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

enforces safety through θ

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

How to do Safe RL?

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

Also reads as:

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

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

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

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

Safe RL steps seek performance under safety constraints

Safe Learning via Robust MPC

Robust NMPC parameters θ

Policy gradient

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

adjusts θ for performance

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

Condition

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

enforces safety through θ

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

How to do Safe RL?

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

Also reads as:

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

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

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

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

Safe RL steps seek performance under safety constraints

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

Illustrative example

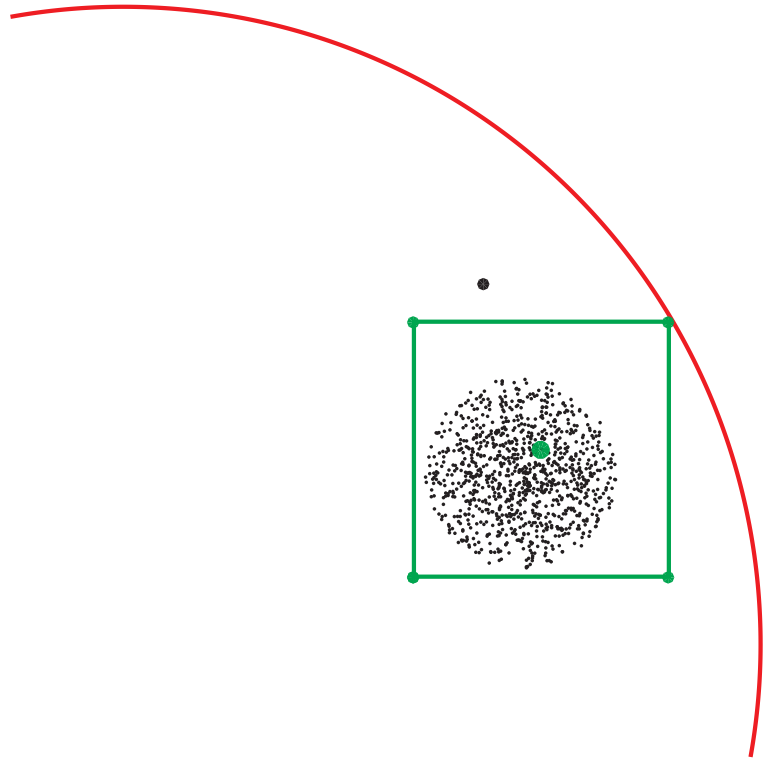
- Real system

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

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

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

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



Illustrative example

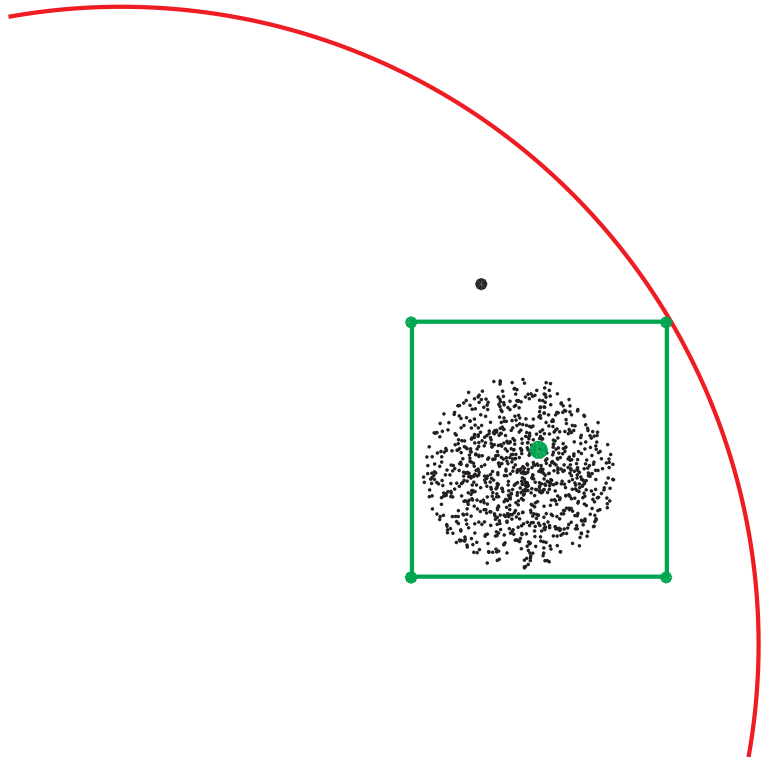
- Real system

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

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

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

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

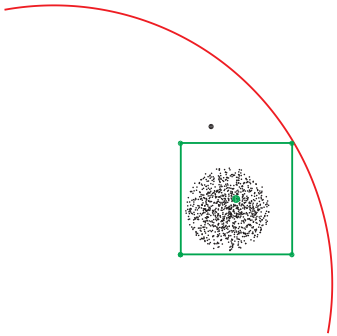


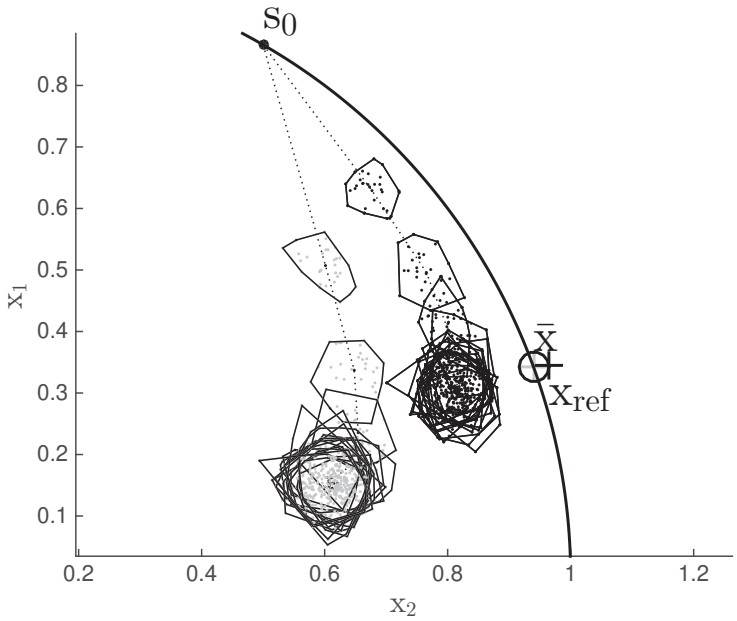
Let's adjusts:

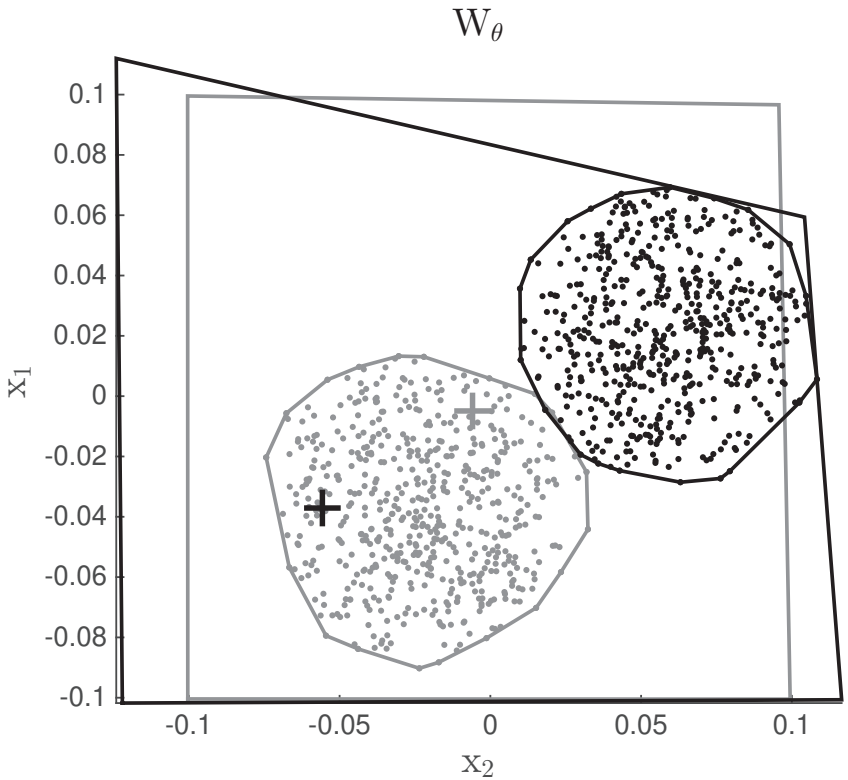
- Set $\mathbb{W}$, while containing process noise
- State and input reference in MPC cost function
- Internal linear feedback (robust MPC internal control)

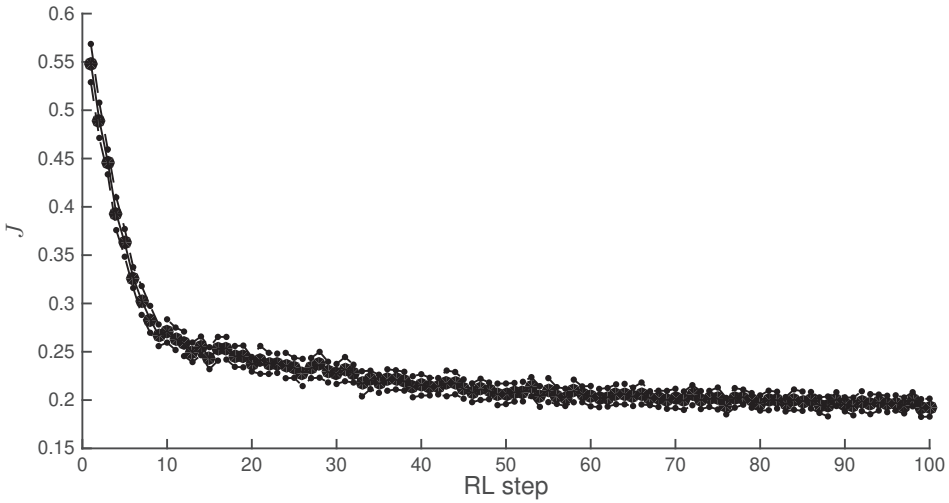
via Safe RL

Illustrative example





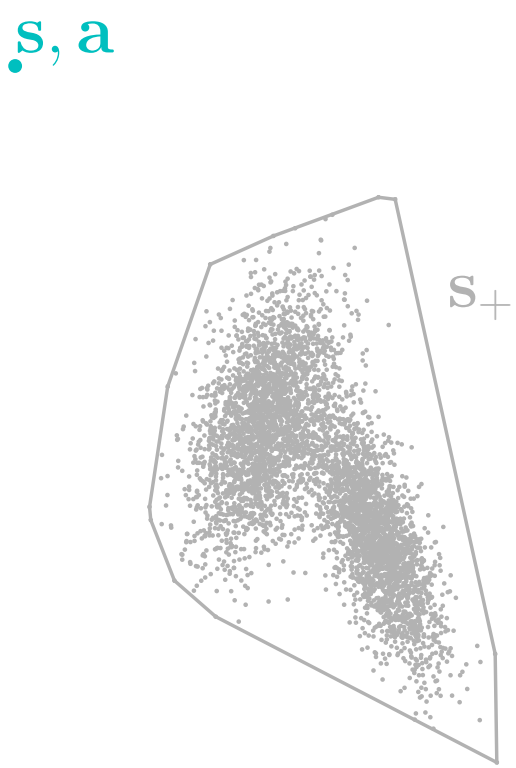




Learning

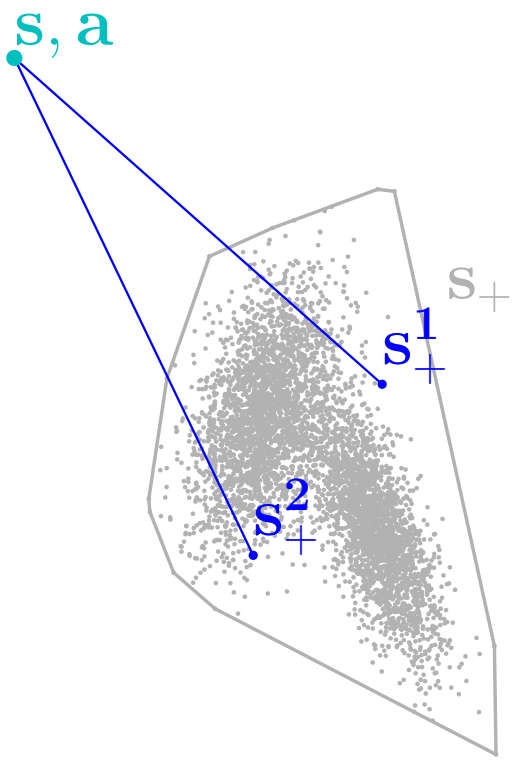
Scenario Tree MPC

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



Scenario Tree MPC

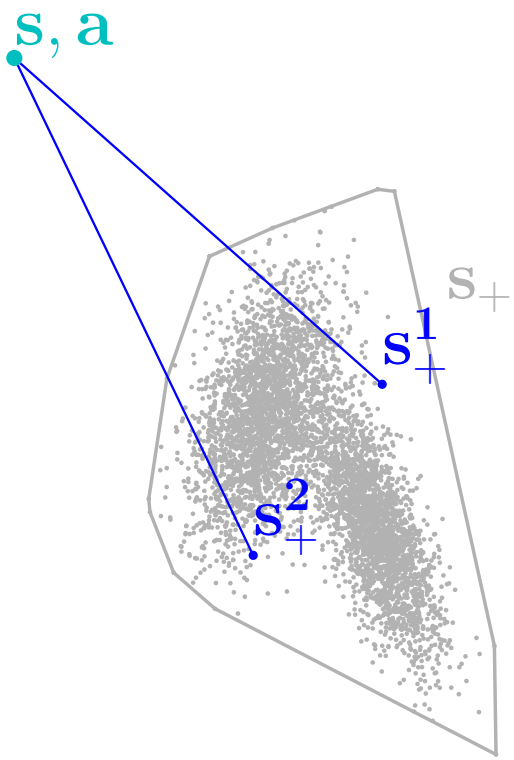
Key idea: approximate distribution $\mathbb{P}[\mathbf{s}_+|\mathbf{s},\mathbf{a}]$ with a finite set of point predictions, develop in a tree of possible future outcomes, with associated decisions. Optimize over the tree.



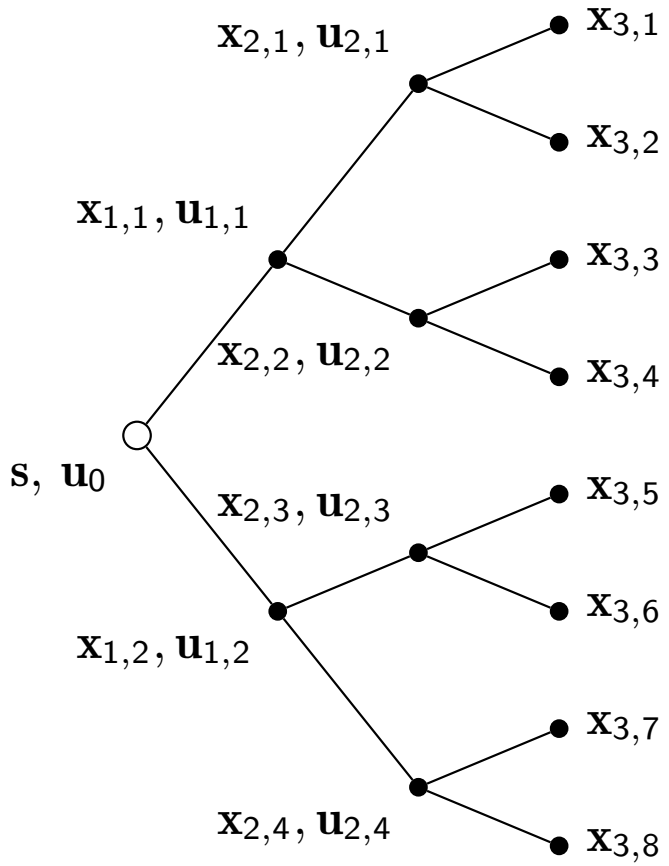
E.g. pick two scenarios

Scenario Tree MPC

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

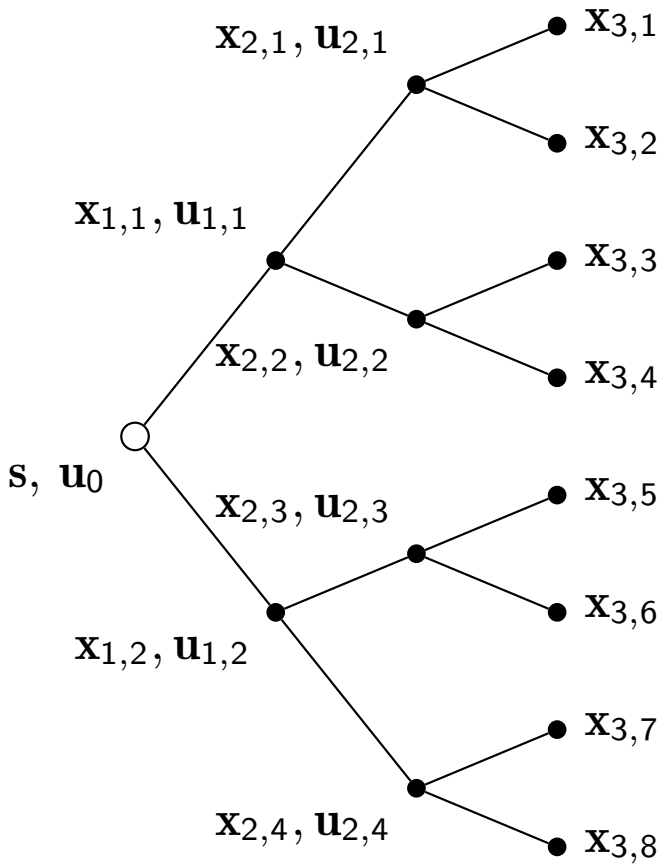


E.g. pick two scenarios



Scenario Tree MPC

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



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

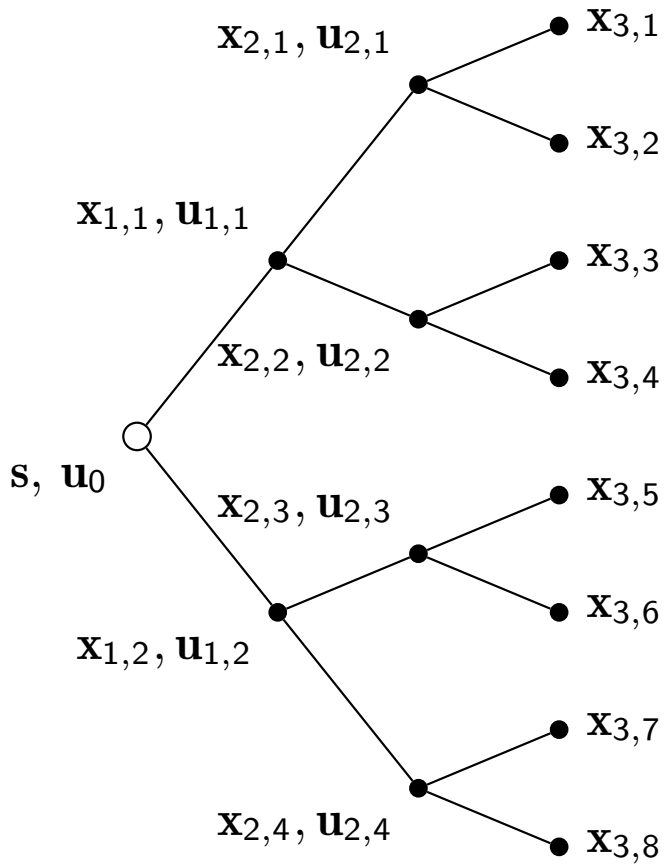
$$
\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & \sum_{i=1}^{n}\sum_{k=0}^{N} \omega_i \, L\left(\mathbf{x}_k^i, \mathbf{u}_k^i\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1}^i = \mathbf{f}_k^i\left(\mathbf{x}_k^i, \mathbf{u}_k^i\right), \quad \mathbf{x}_0^i = \mathbf{s} \\
& \mathbf{h}_i\left(\mathbf{x}_k^i,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

where we need to select

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

Scenario Tree MPC

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



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

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

where we need to select

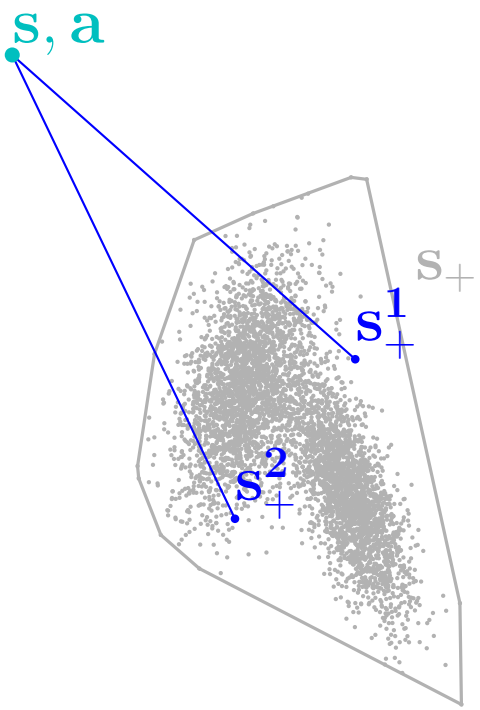
- $\mathbf{f}_k^i$ to form the scenarios
- $\omega_{1,\ldots,n}$ to build $\mathbb{E}[.]$

Difficulties

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

Often shallow trees are used

RL over Scenario Tree MPC



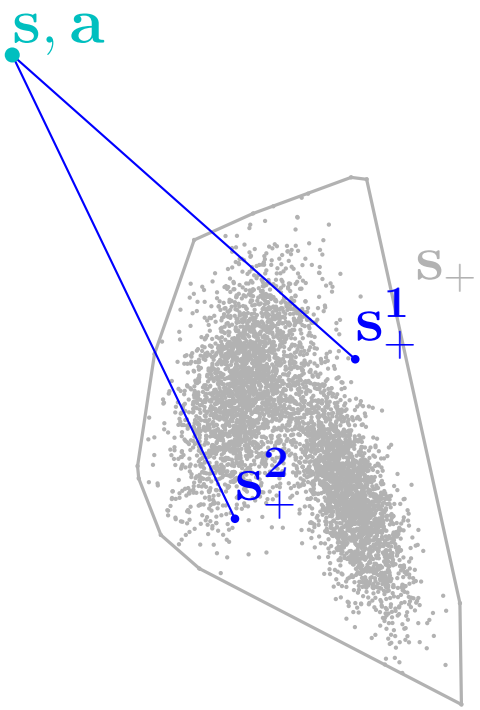
Scenario-Tree MPC

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

Difficulties

- Pick "good" scenarios
- Exploding complexity

RL over Scenario Tree MPC



Difficulties

- Pick "good" scenarios
- Exploding complexity

Scenario-Tree MPC

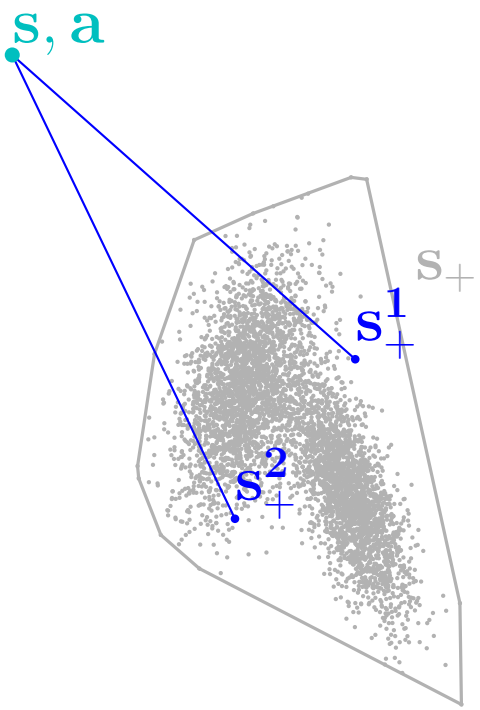
$$\begin{aligned}
\min_{\mathbf{x},\mathbf{u}} \quad & \sum_{i=1}^{n} \sum_{k=0}^{N} \omega_i \, L\left(\mathbf{x}_k^i, \mathbf{u}_k^i\right) \\
\text{s.t.} \quad & \mathbf{x}_{k+1}^i = \mathbf{f}_k^i\left(\mathbf{x}_k^i, \mathbf{u}_k^i\right), \quad \mathbf{x}_0^i = \mathbf{s} \\
& \mathbf{h}_i\left(\mathbf{x}_k^i,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

RL can do:

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

for real-world performance of policy

RL over Scenario Tree MPC



Scenario-Tree MPC

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

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

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

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

RL can do:

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

for real-world performance of policy

Difficulties

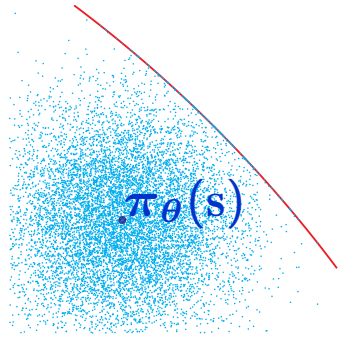
- Pick "good" scenarios
- Exploding complexity

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

Outline

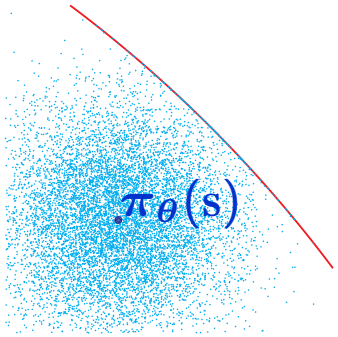
"Safe" (feasible) exploration with MPC

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



"Safe" (feasible) exploration with MPC

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

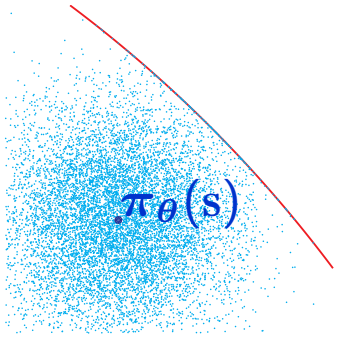


Explore while keeping feasibility?

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

"Safe" (feasible) exploration with MPC

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



Explore while keeping feasibility?

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

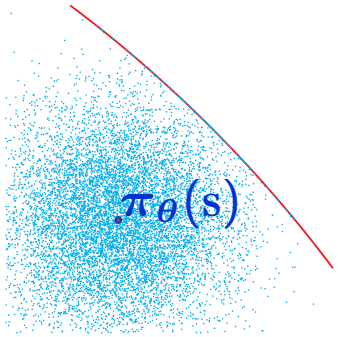
Feasible exploration: $\boldsymbol{\pi}^{\mathrm{e}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{u}_0^\star$:

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

satisfies the constraints by construction

"Safe" (feasible) exploration with MPC

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



Explore while keeping feasibility?

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

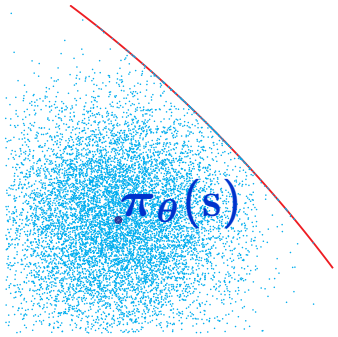
Feasible exploration: $\boldsymbol{\pi}^{\mathrm{e}}_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{u}_0^\star$:

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

satisfies the constraints by construction

"Safe" (feasible) exploration with MPC

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



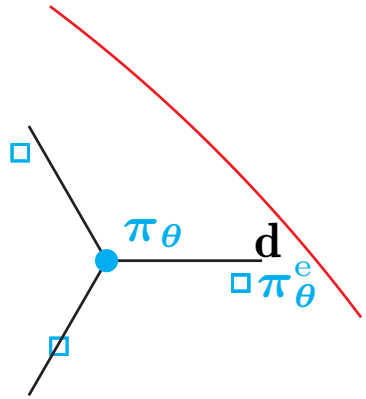
Explore while keeping feasibility?

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

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

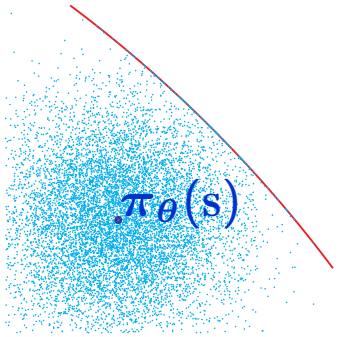
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) - \mathbf{d}^\top \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\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

satisfies the constraints by construction



"Safe" (feasible) exploration with MPC

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



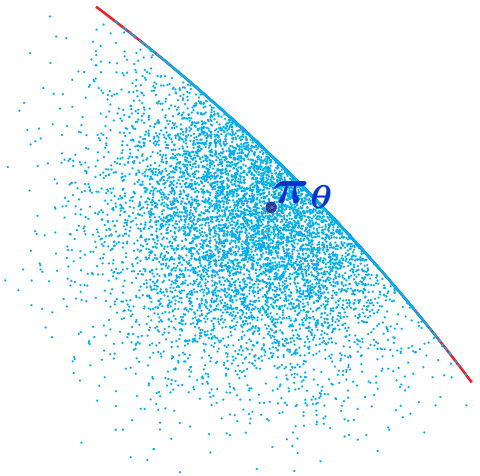
Explore while keeping feasibility?

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

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

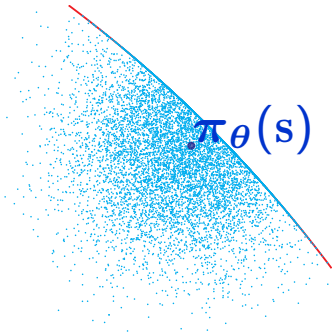
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) - \mathbf{d}^\top \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\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

satisfies the constraints by construction



"Safe" (feasible) exploration with MPC

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



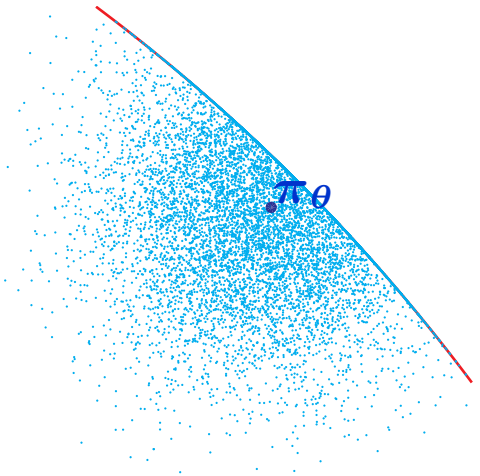
Explore while keeping feasibility?

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

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

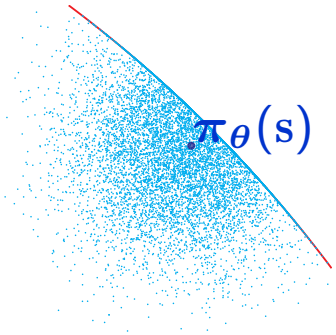
$$\begin{aligned} \min_{\mathbf{x},\mathbf{u}} \quad & T_{\boldsymbol{\theta}}(\mathbf{x}_N) - \mathbf{d}^\top \mathbf{u}_0 + \sum_{k=0}^{N-1} L_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\ \text{s.t.} \quad & \mathbf{x}_{k+1} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{u}_k) \\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

satisfies the constraints by construction



"Safe" (feasible) exploration with MPC

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



Explore while keeping feasibility?

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

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

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

satisfies the constraints by construction

Remarks:

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

MPC as a Stochastic Policy

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

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

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

MPC as a Stochastic Policy

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

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

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

Remarks

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

MPC as a Stochastic Policy

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

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

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

Remarks

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

Performance

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

MPC as a Stochastic Policy

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

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

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

Remarks

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

Performance

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

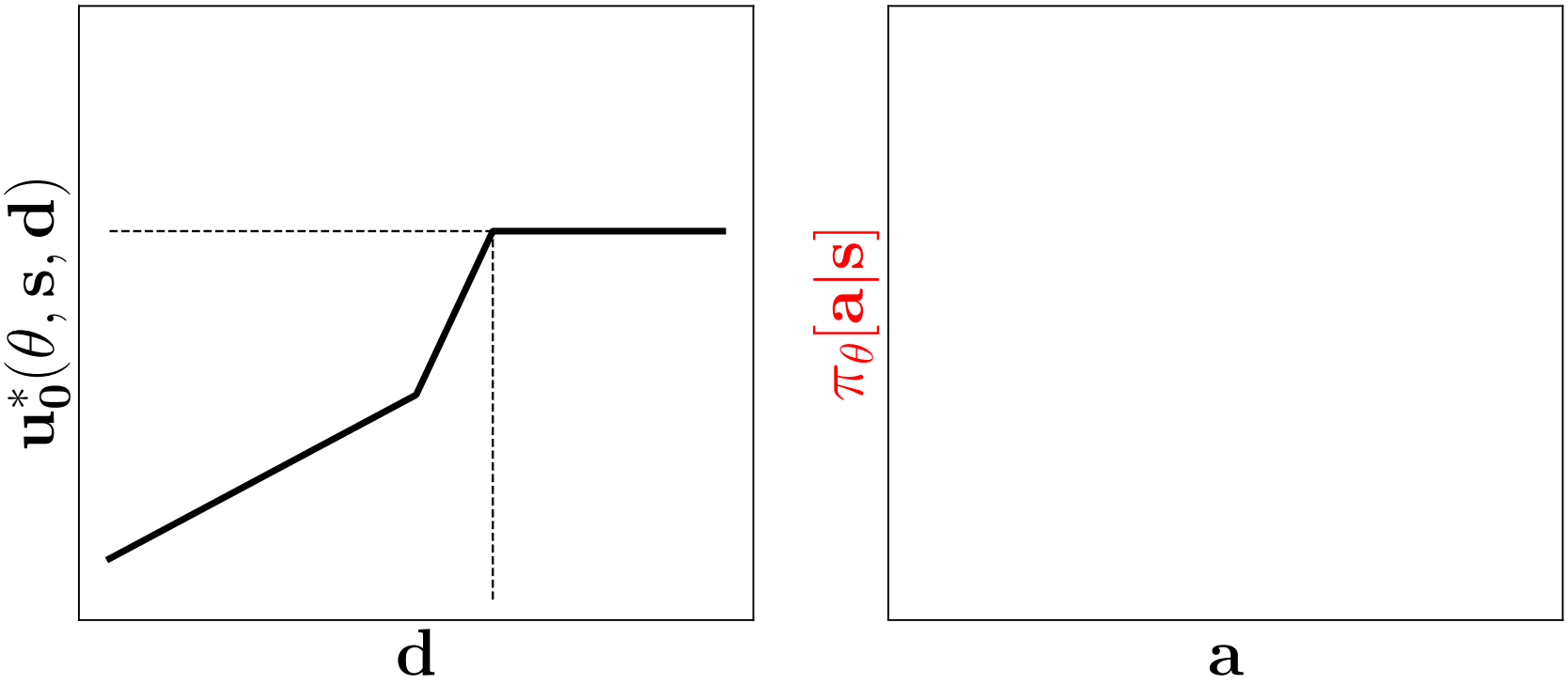
Stochastic Policy Gradient

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

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

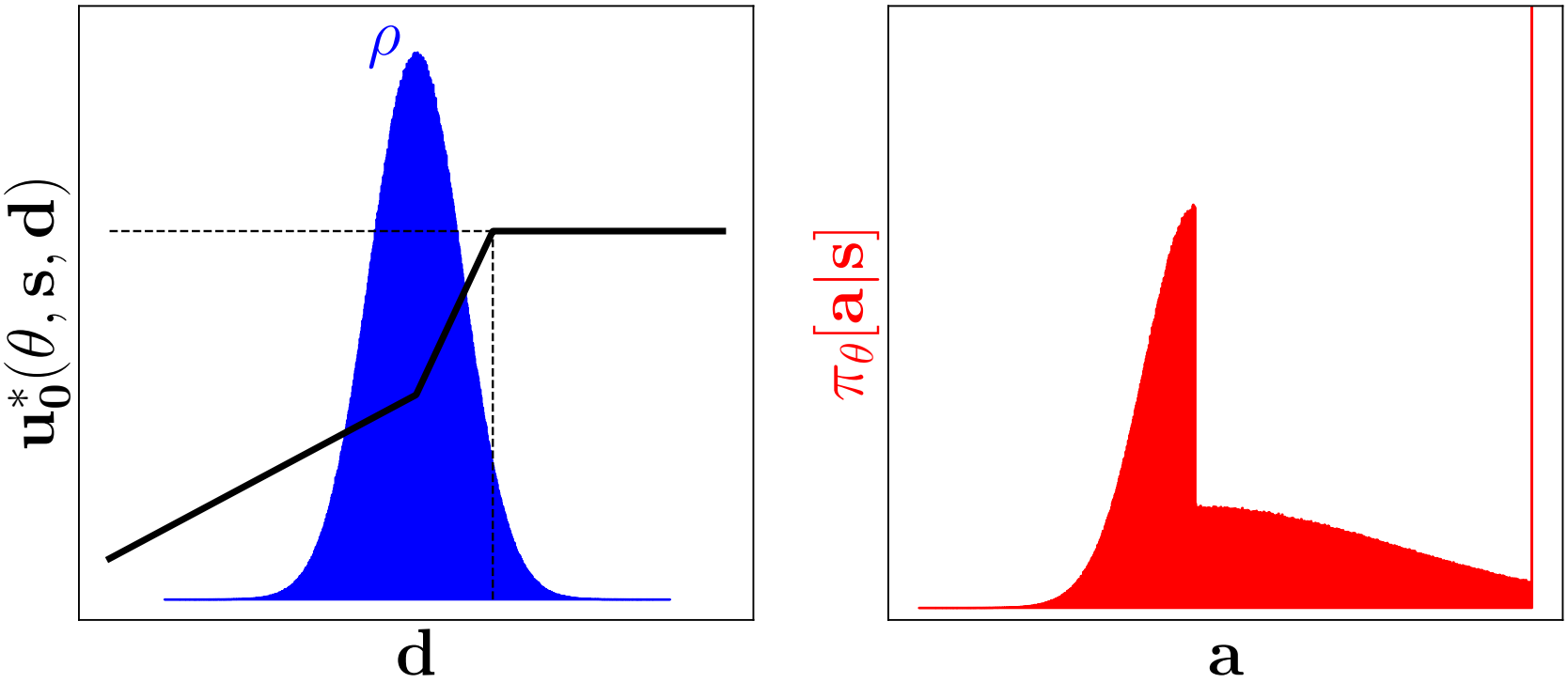
MPC as a Stochastic Policy & Policy Gradient

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



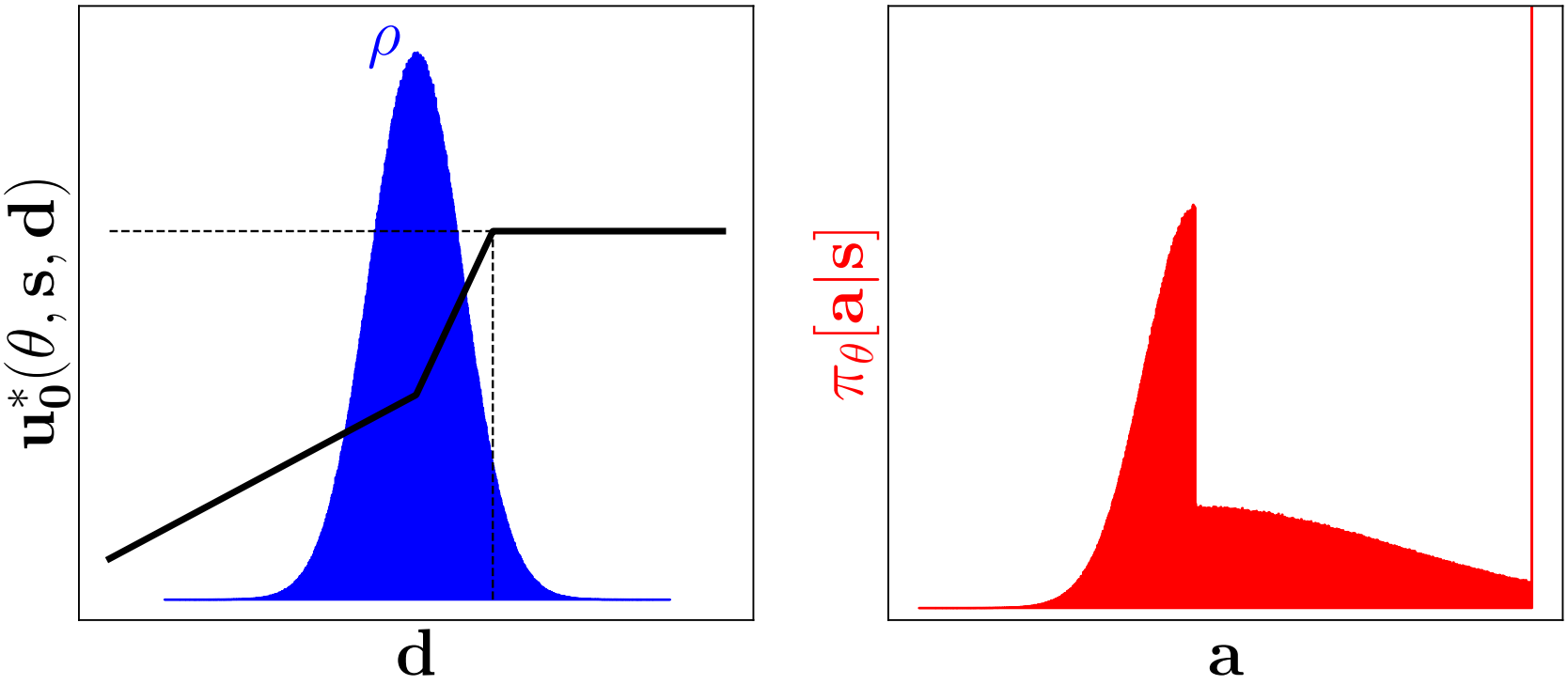
MPC as a Stochastic Policy & Policy Gradient

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



MPC as a Stochastic Policy & Policy Gradient

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

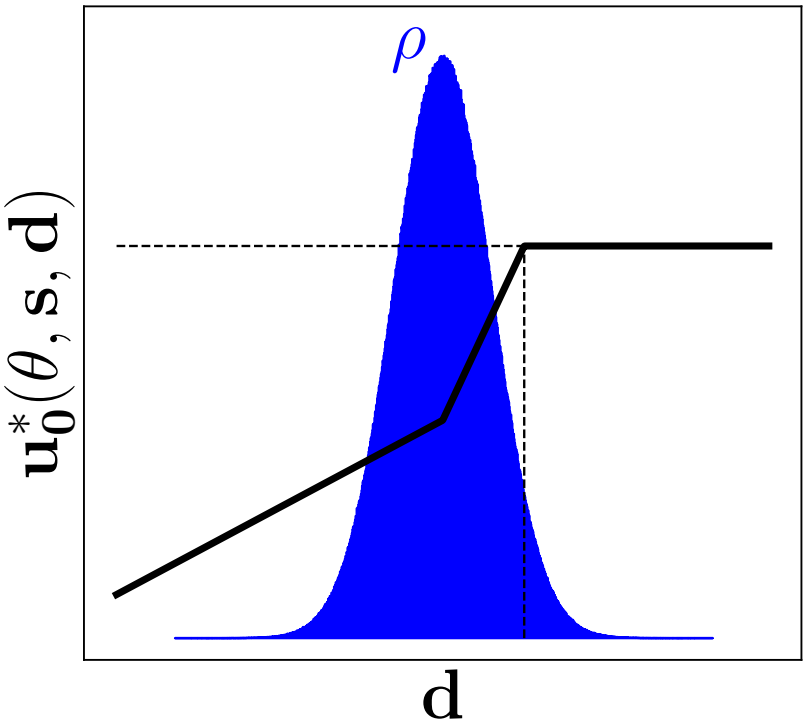


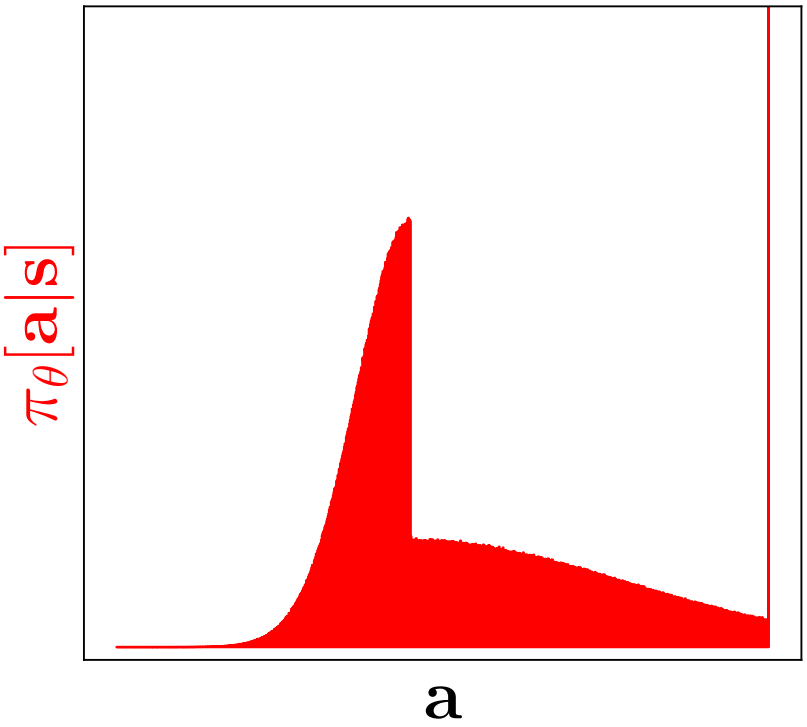
For $\mathbf{d} \to \mathbf{u}_0^\star$ bijective & differentiable

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

MPC as a Stochastic Policy & Policy Gradient

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





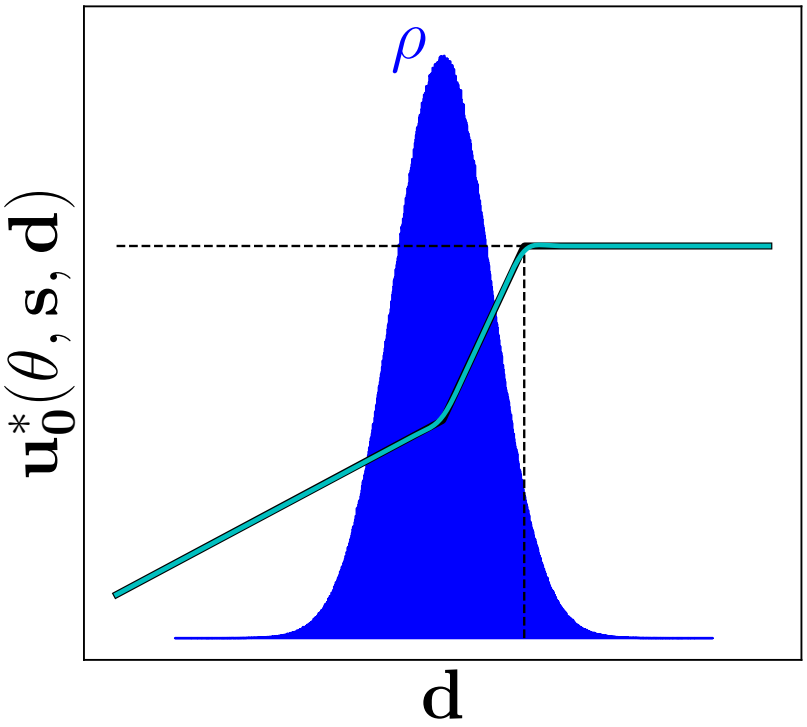
For $\mathbf{d} \to \mathbf{u}_0^\star$ bijective & differentiable

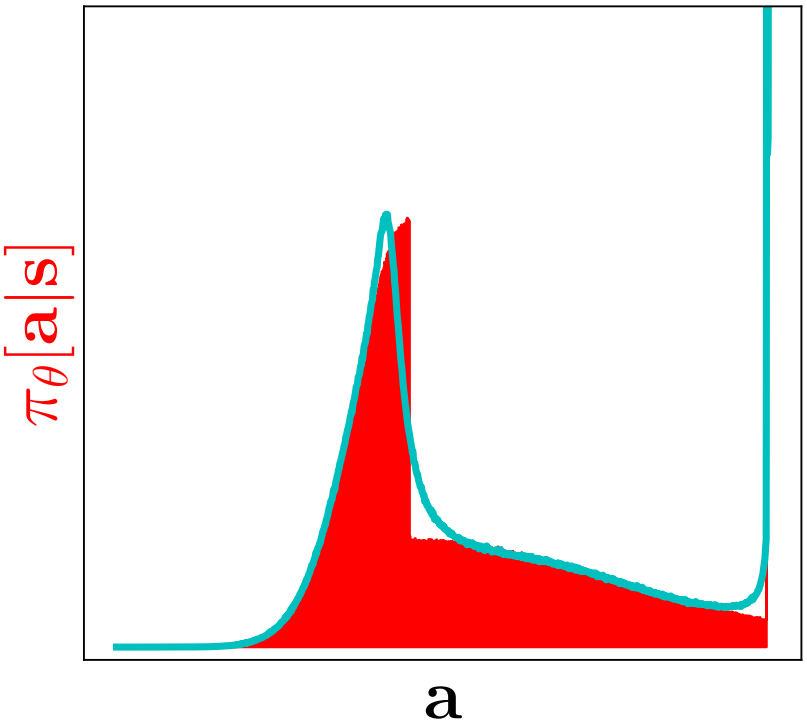
$$\pi_{\boldsymbol{\theta}}\left[\mathbf{a}|\mathbf{s}\right] = \left|\det\left(\nabla_{\mathbf{d}}\mathbf{u}_0^\star\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d}\right)\right)\right|^{-1} \rho\left[\mathbf{d}\right]$$

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

MPC as a Stochastic Policy & Policy Gradient

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





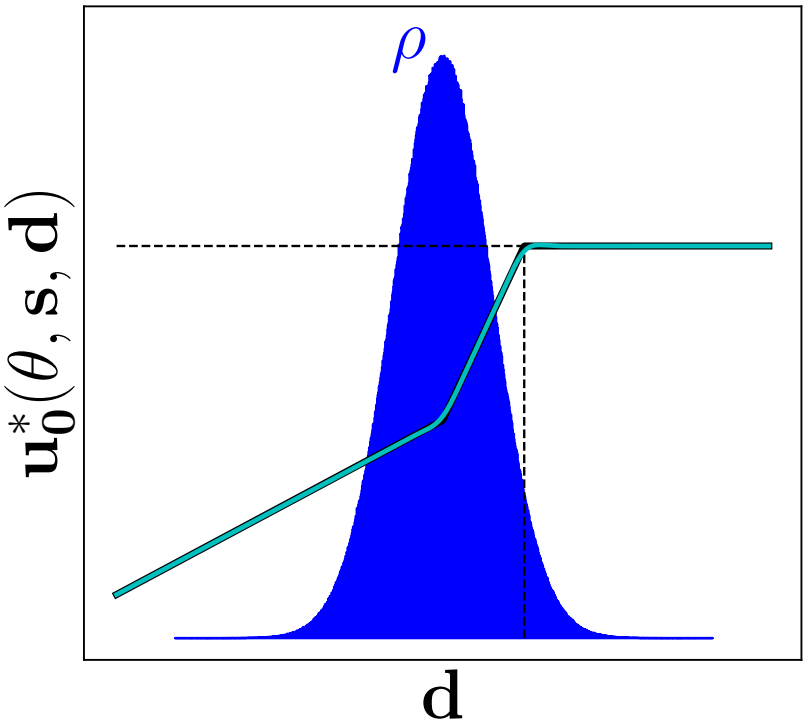
For $\mathbf{d} \to \mathbf{u}_0^\star$ bijective & differentiable

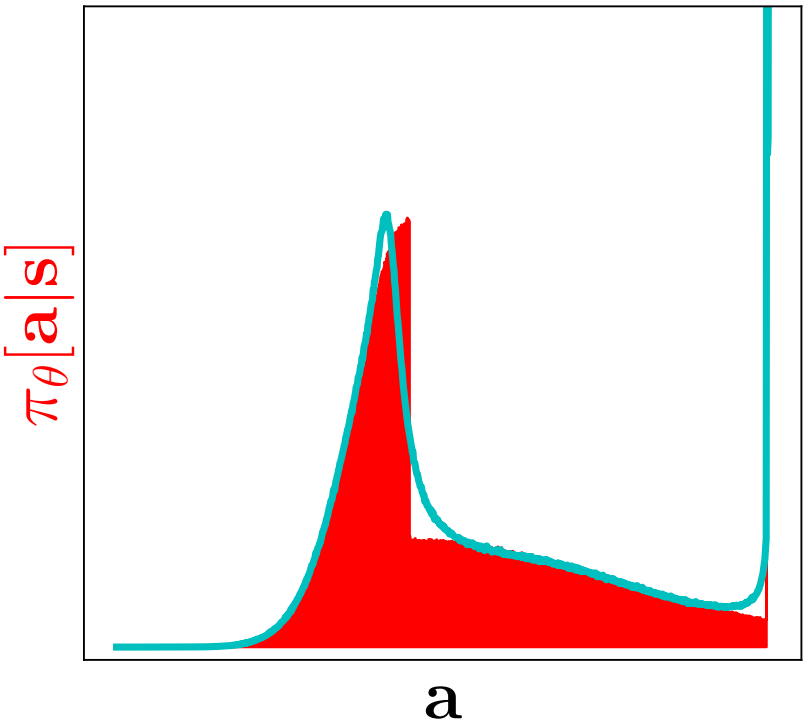
$$\pi_{\boldsymbol{\theta}}\left[\mathbf{a}|\mathbf{s}\right] = \left|\det\left(\nabla_{\mathbf{d}}\mathbf{u}_0^\star\left(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d}\right)\right)\right|^{-1} \rho\left[\mathbf{d}\right]$$

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

MPC as a Stochastic Policy & Policy Gradient

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





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

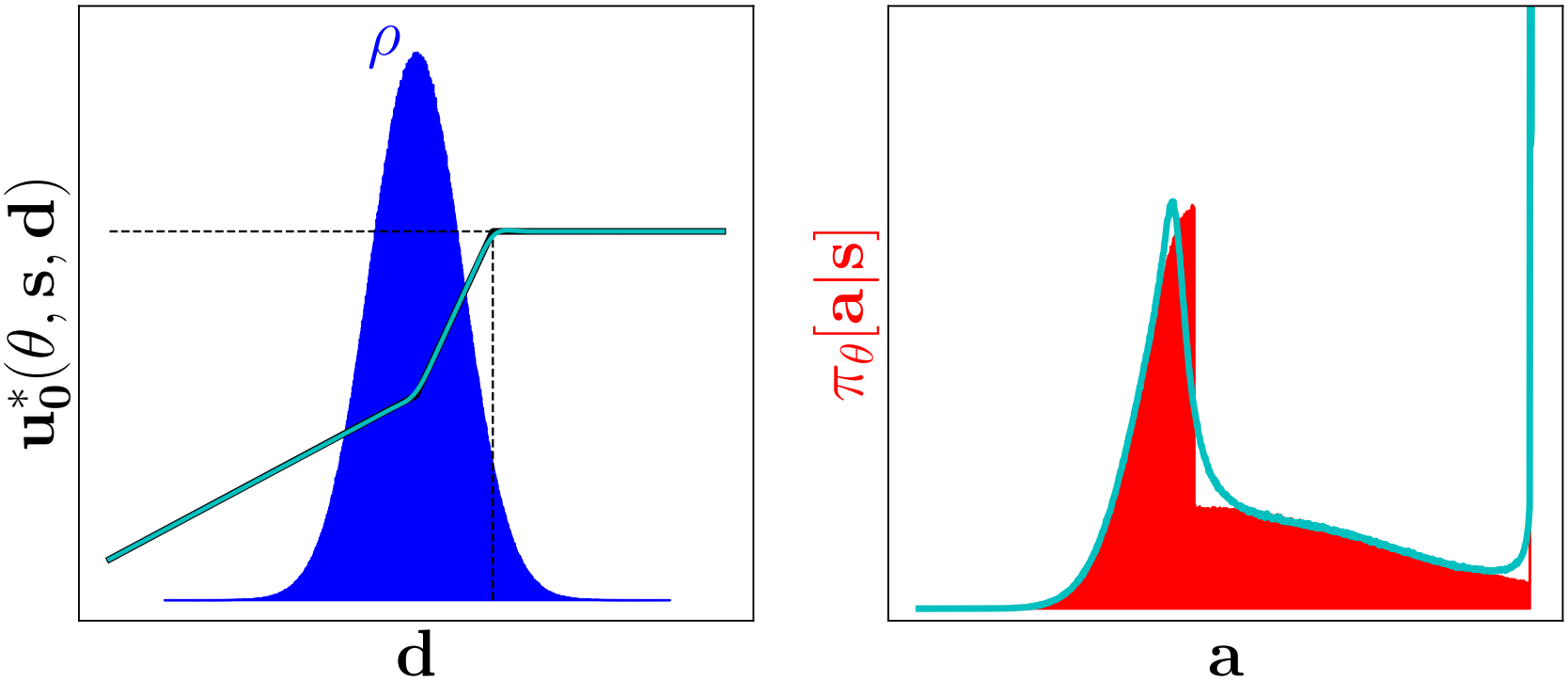
$$\pi_{\boldsymbol{\theta}}[\mathbf{a}|\mathbf{s}] = \left|\det\left(\nabla_{\mathbf{d}} \mathbf{u}_0^\star(\boldsymbol{\theta}, \mathbf{s}, \mathbf{d})\right)\right|^{-1} \rho[\mathbf{d}]$$

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

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

MPC as a Stochastic Policy & Policy Gradient

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

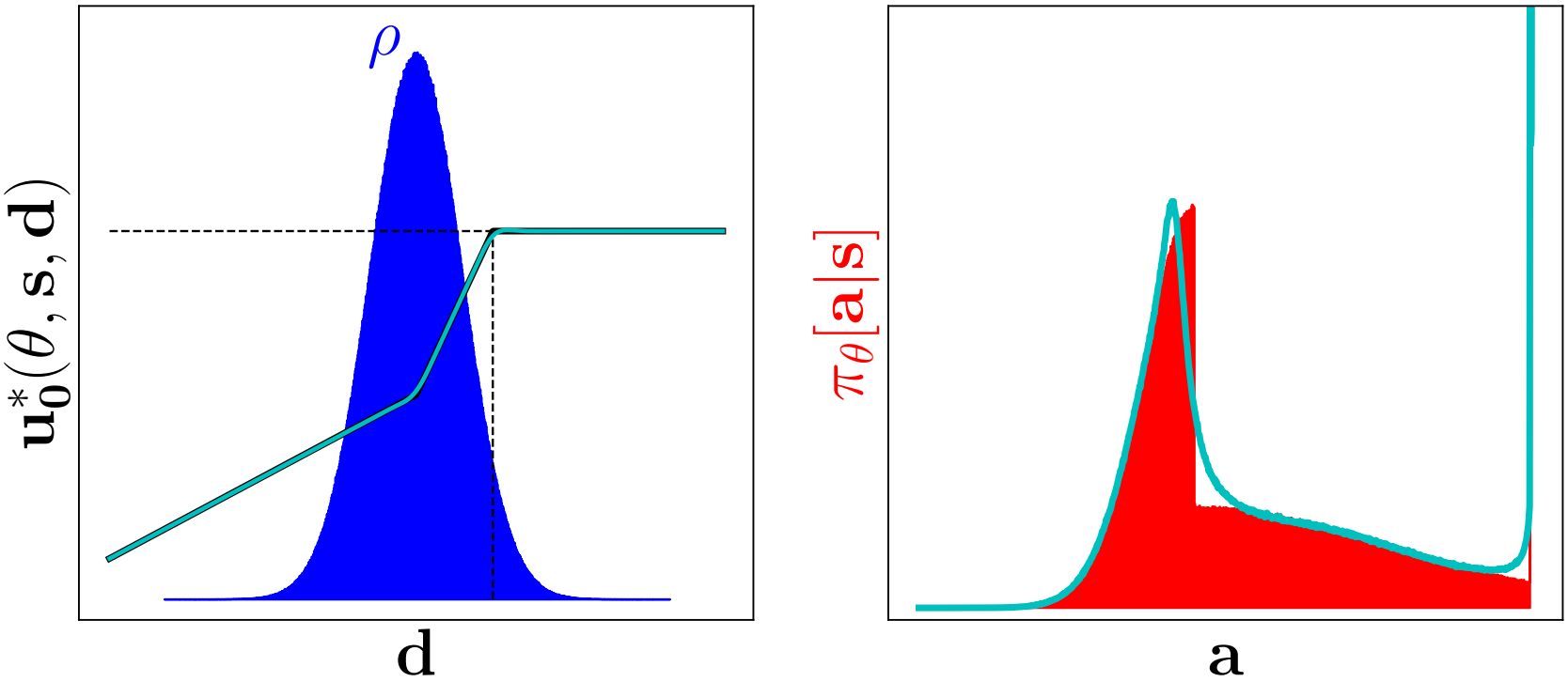


Reparametrization trick gives (smoothing is still *theoretically* good...)

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

MPC as a Stochastic Policy & Policy Gradient

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



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

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

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

Outline

Stability of MPC - Introduction

Policy π_{MPC} from

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

Stability of MPC - Introduction

Policy π_{MPC} from

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

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

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

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

Stability of MPC - Introduction

Policy π_{MPC} from

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

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

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

then MPC scheme is stabilizing ($+$conditions on T)

Stability of MPC - Introduction

Policy π_{MPC} from

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

Equivalent MPC

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

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

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

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

then MPC scheme is stabilizing ($+$conditions on T)

Stability of MPC - Introduction

Policy π_{MPC} from

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

Equivalent MPC

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

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

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

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

then MPC scheme is stabilizing ($+$conditions on T)

Remarks:

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

Stability of MPC - Introduction

Policy π_{MPC} from

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

Equivalent MPC

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

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

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

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

then MPC scheme is stabilizing ($+$conditions on T)

Remarks:

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

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

Stability of MPC - Connection to RLMPC

Policy π^{MPC} from

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

Stability of MPC - Connection to RLMPC

Policy π^{MPC} from

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

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

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

where κ is K_∞ ($+$conditions on T)

Stability of MPC - Connection to RLMPC

Policy π^{MPC} from

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

Equivalent MPC

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

is stable

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

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

where κ is K_∞ (+conditions on T)

Stability of MPC - Connection to RLMPC

Policy π^{MPC} from

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

Equivalent MPC

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

is stable

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

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

where κ is K_∞ (+conditions on T)

Remarks

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

- Learning based on L
- Impose constraint:

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

  throughout the learning
- L_θ different than L due to stability + model error

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

- Learning based on L
- Impose constraint:

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

  throughout the learning
- L_θ different than L due to stability + model error

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

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

Theorem: under some conditions

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

- Learning based on L
- Impose constraint:

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

  throughout the learning
- L_θ different than L due to stability + model error

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

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

Theorem: under some conditions

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

Constraint

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

is semi-infinite programming, not trivial

Some solutions:

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

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

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

Theorem: under some conditions

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

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

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

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

Theorem: under some conditions

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

Extension to stable policy for MDPs?

- Need stability with discount
- Need "stochastic dissipativity"

Parametrized policy π_θ^{MPC} from:

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

Theorem: under some conditions

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

Extension to stable policy for MDPs?

- Need stability with discount
- Need "stochastic dissipativity"

MDP dissipativity: (2x Automatica '22)

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

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

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

Theorem: under some conditions

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

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

Stability-constrained Learning-based MPC - Deterministic case

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

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

Extension to stable policy for MDPs?

- Need stability with discount
- Need "stochastic dissipativity"

MDP dissipativity: (2x Automatica '22)

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

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

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

Theorem: under some conditions

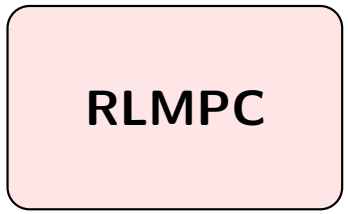
- $\pi_\theta^{\mathrm{MPC}} \to \pi_\star$ if $\pi_\star$ is stabilizing
- $\pi_\theta^{\mathrm{MPC}} \to$ best stabilizing policy otherwise

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

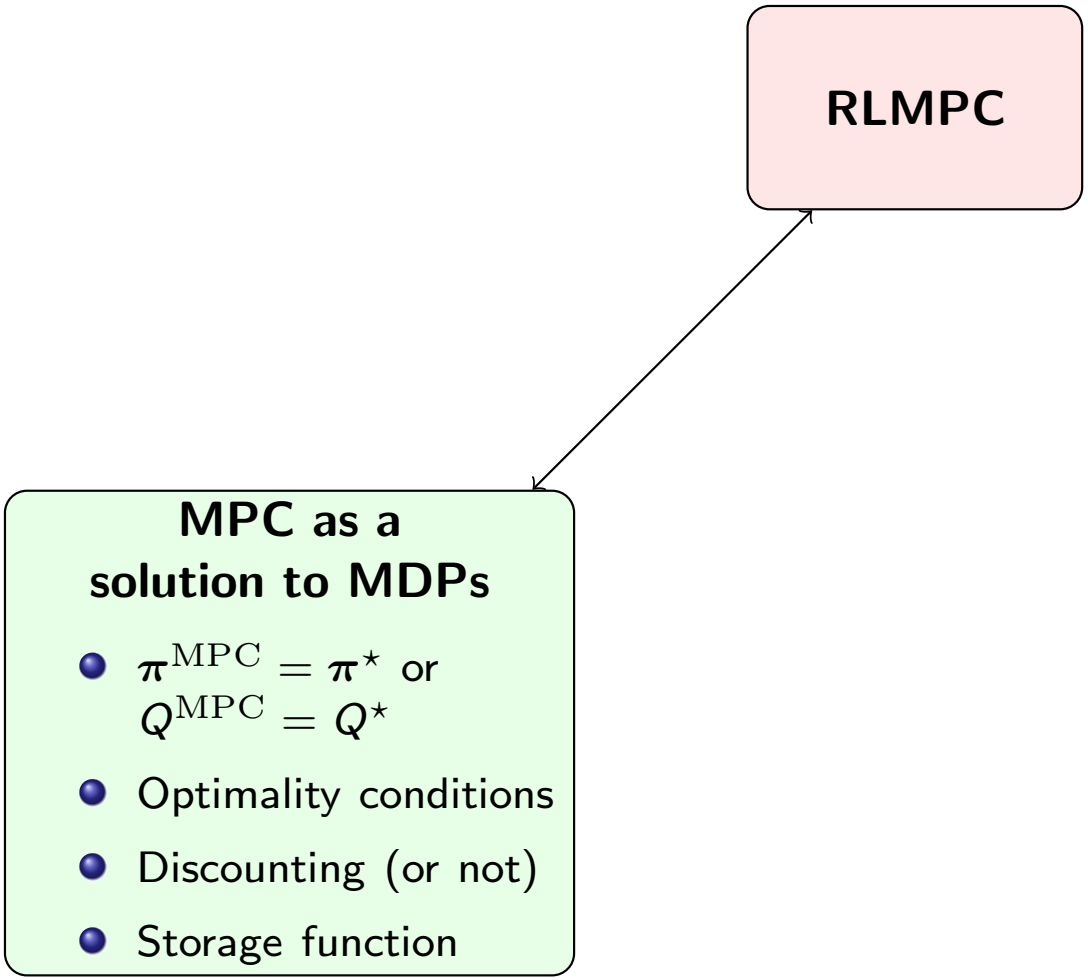
We have the maths to treat this, not yet the algorithms for the stochastic case...

Outline

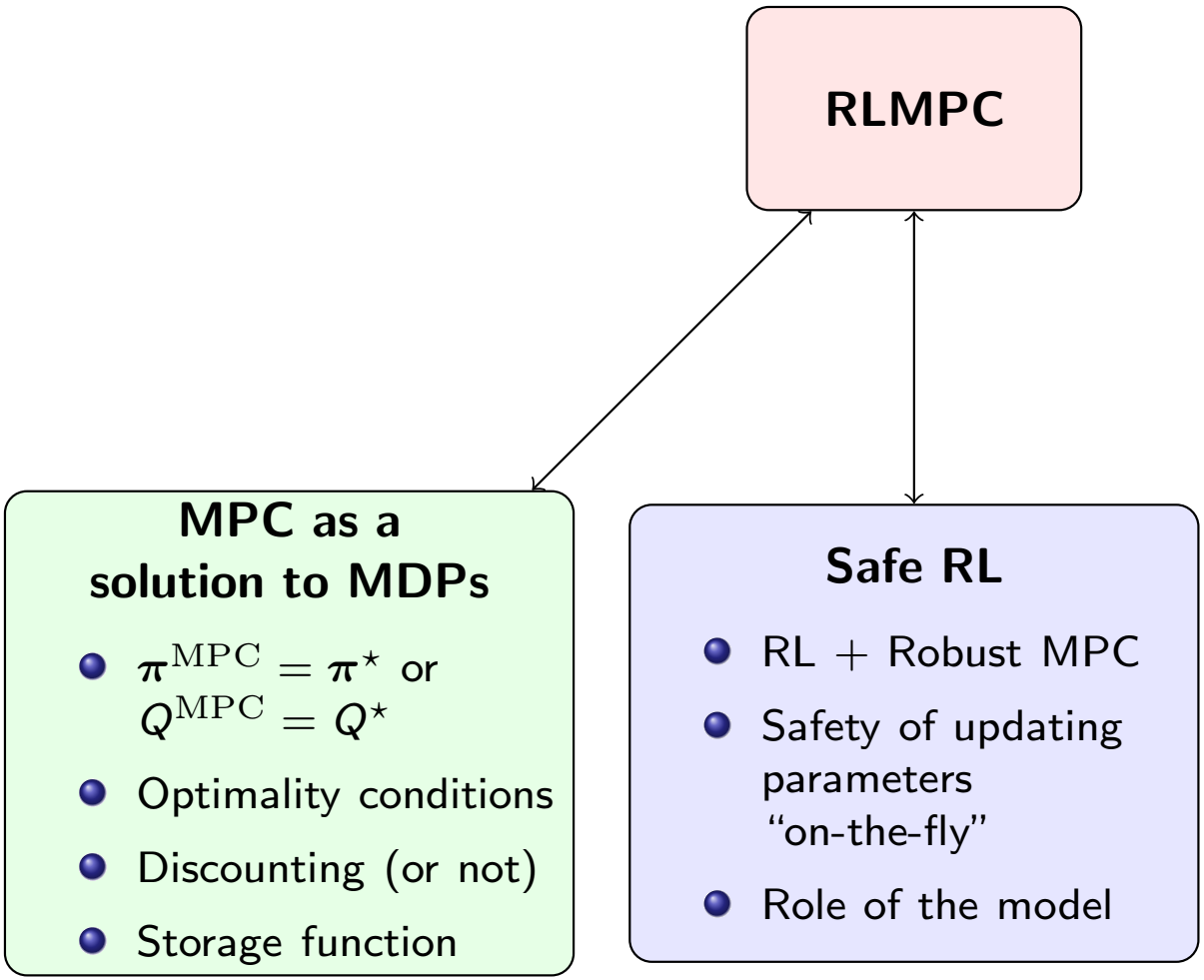
Explored questions - Fundamentals



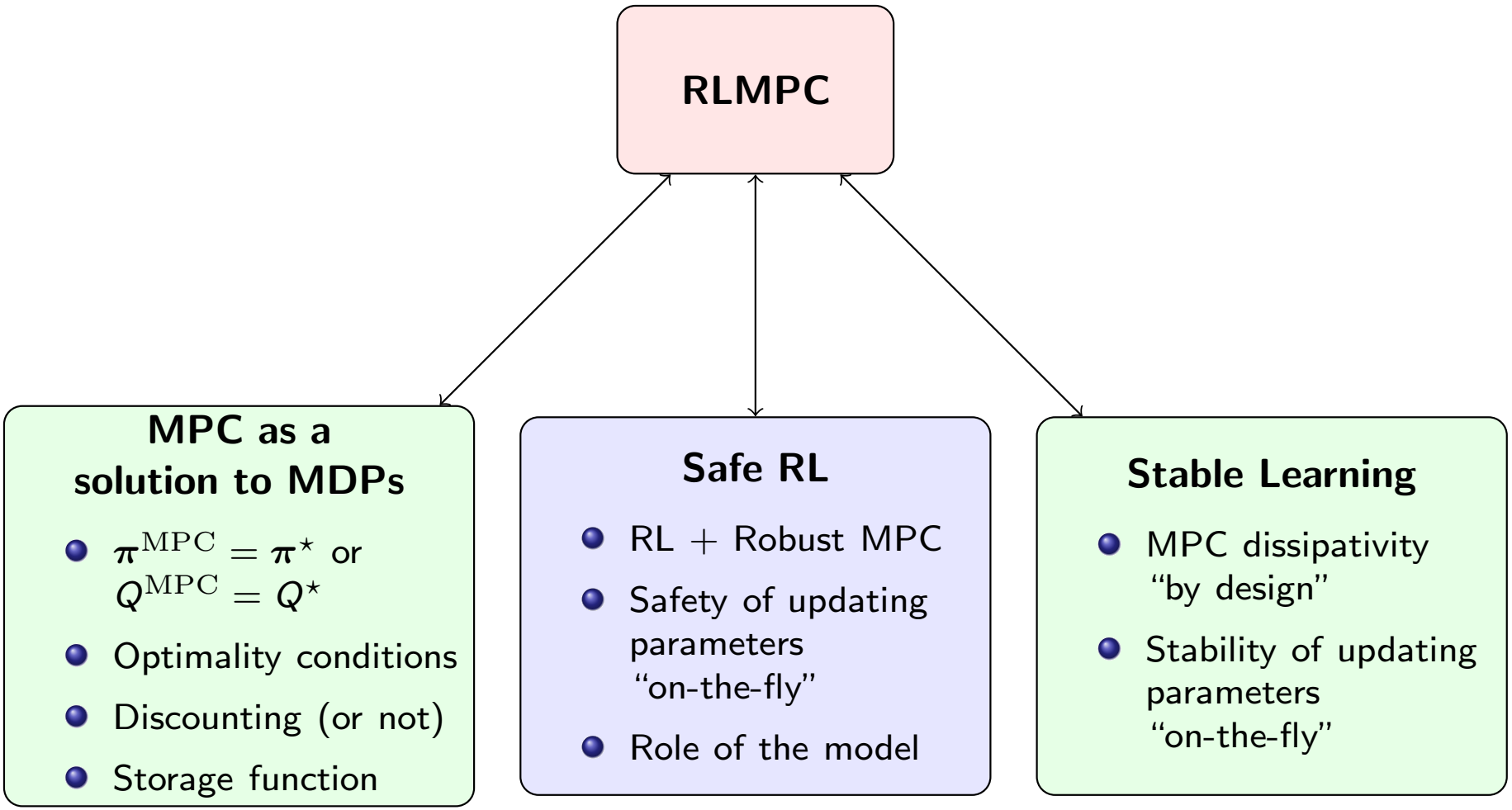
Explored questions - Fundamentals



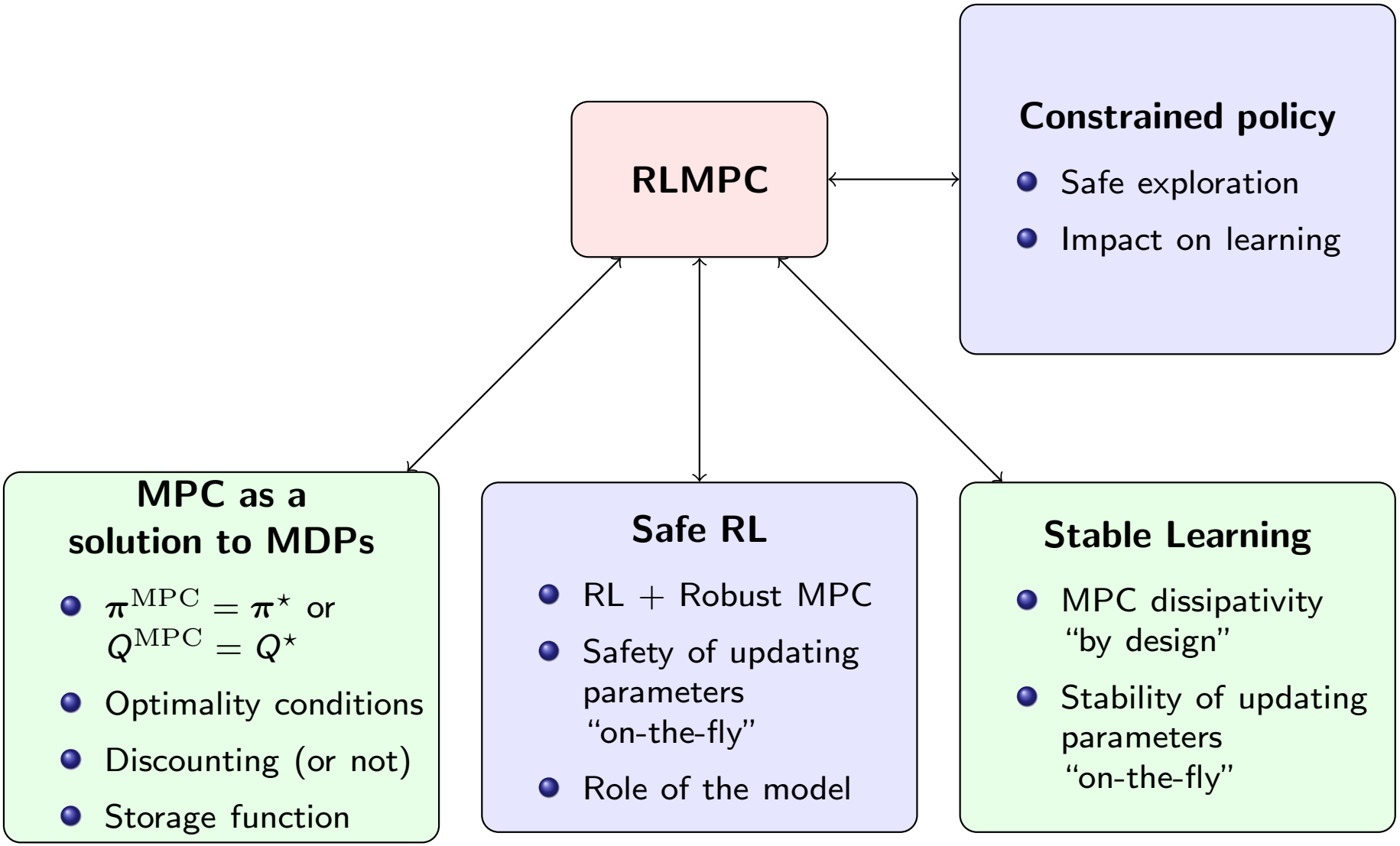
Explored questions - Fundamentals



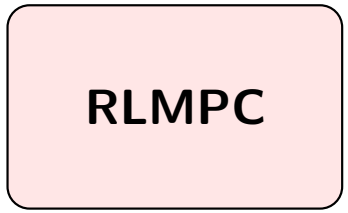
Explored questions - Fundamentals



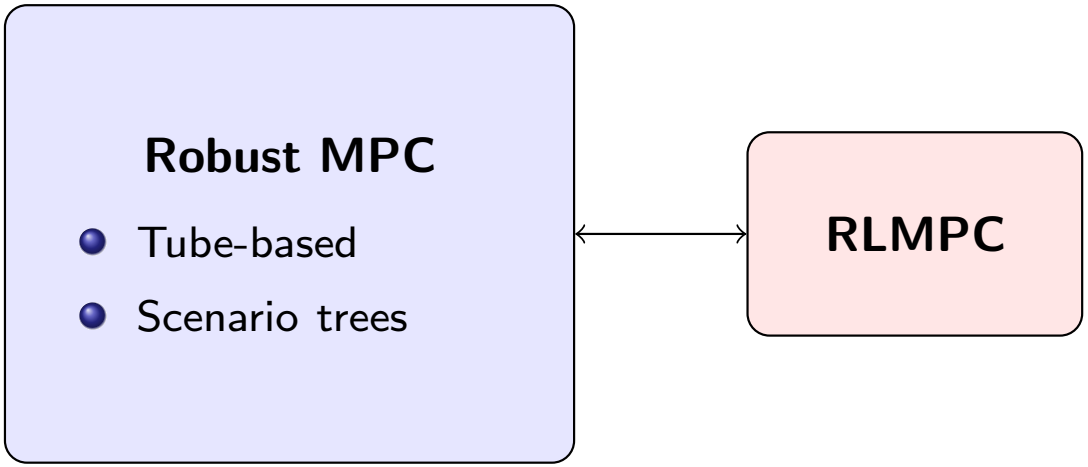
Explored questions - Fundamentals



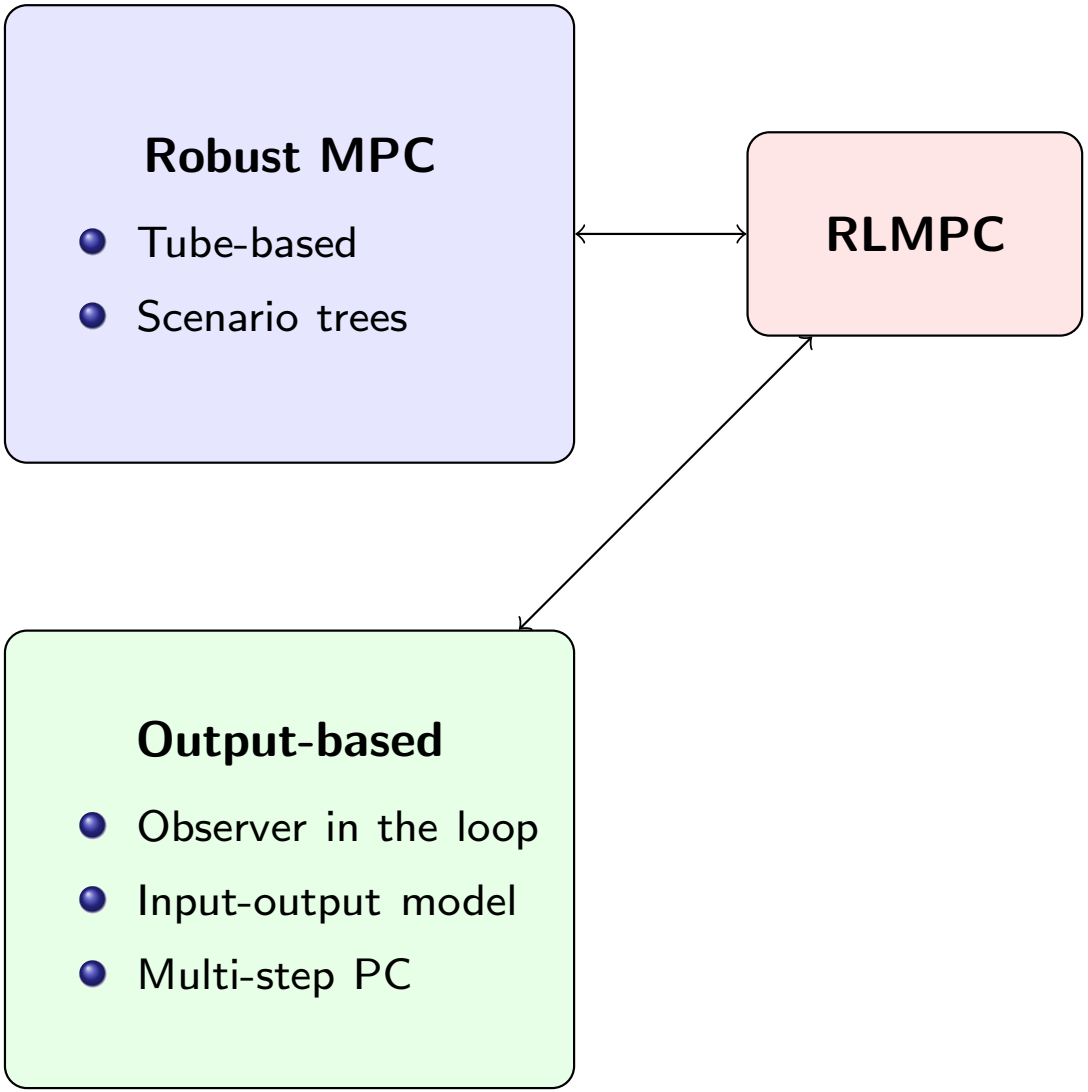
Explored questions - Types of MPC & Learning



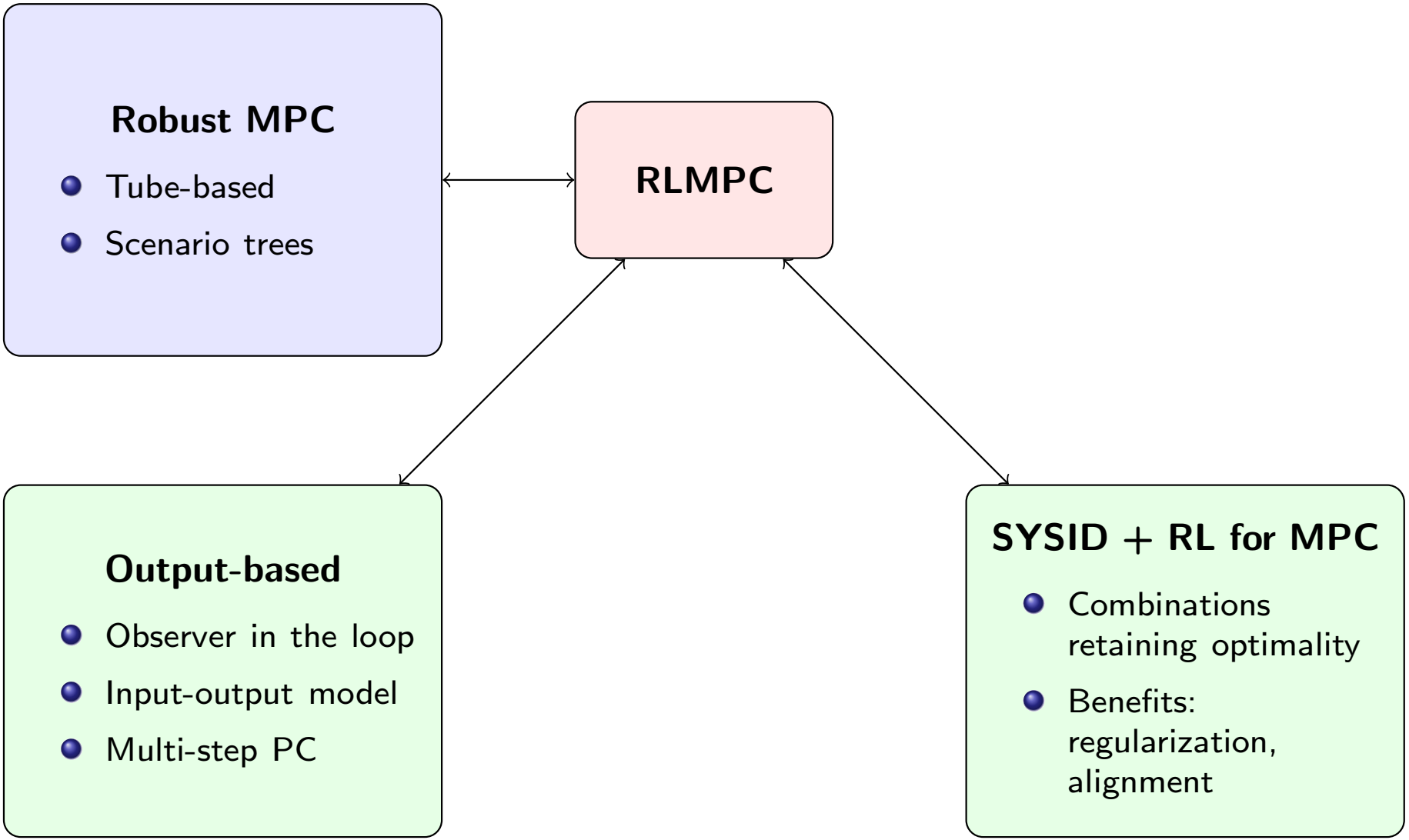
Explored questions - Types of MPC & Learning



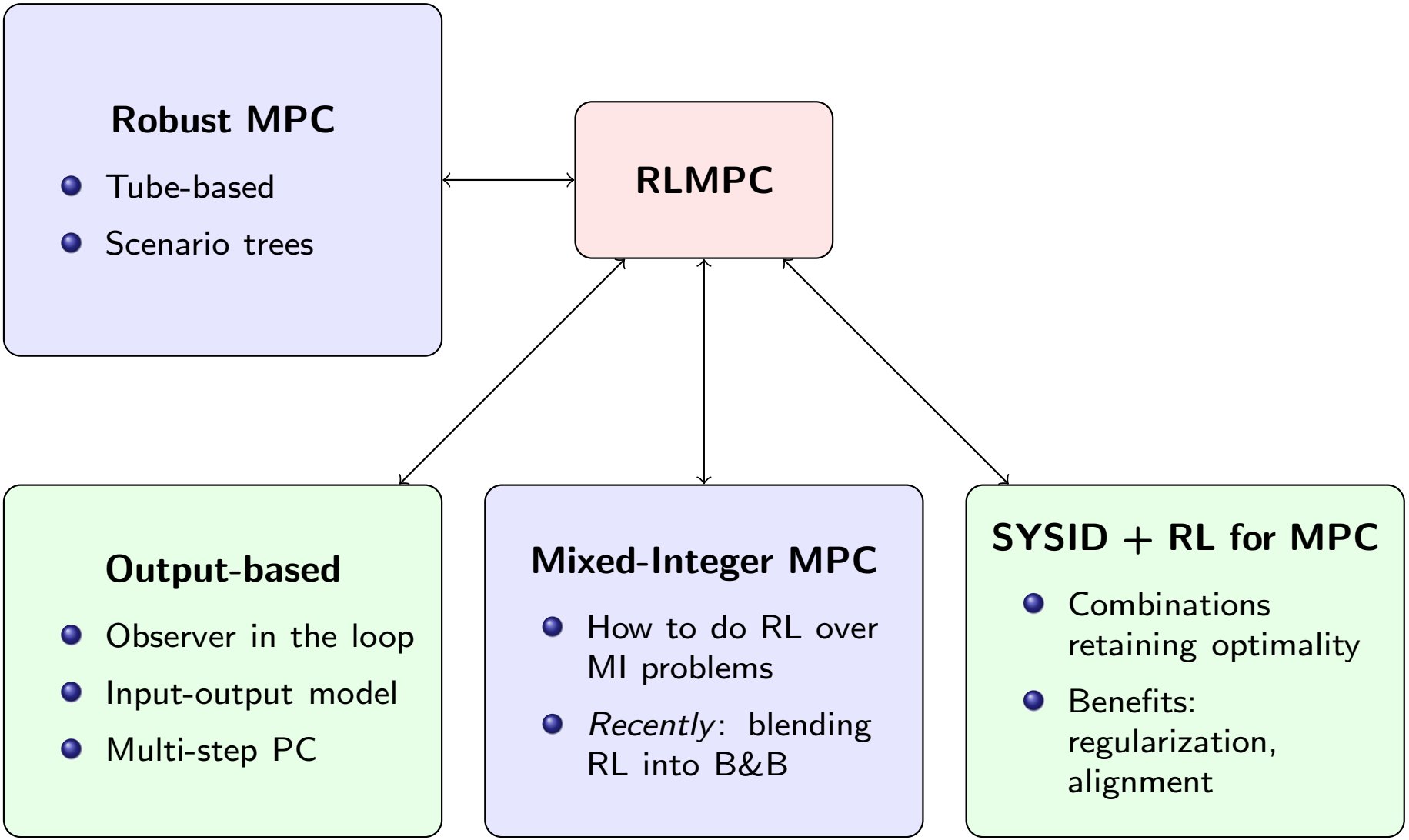
Explored questions - Types of MPC & Learning



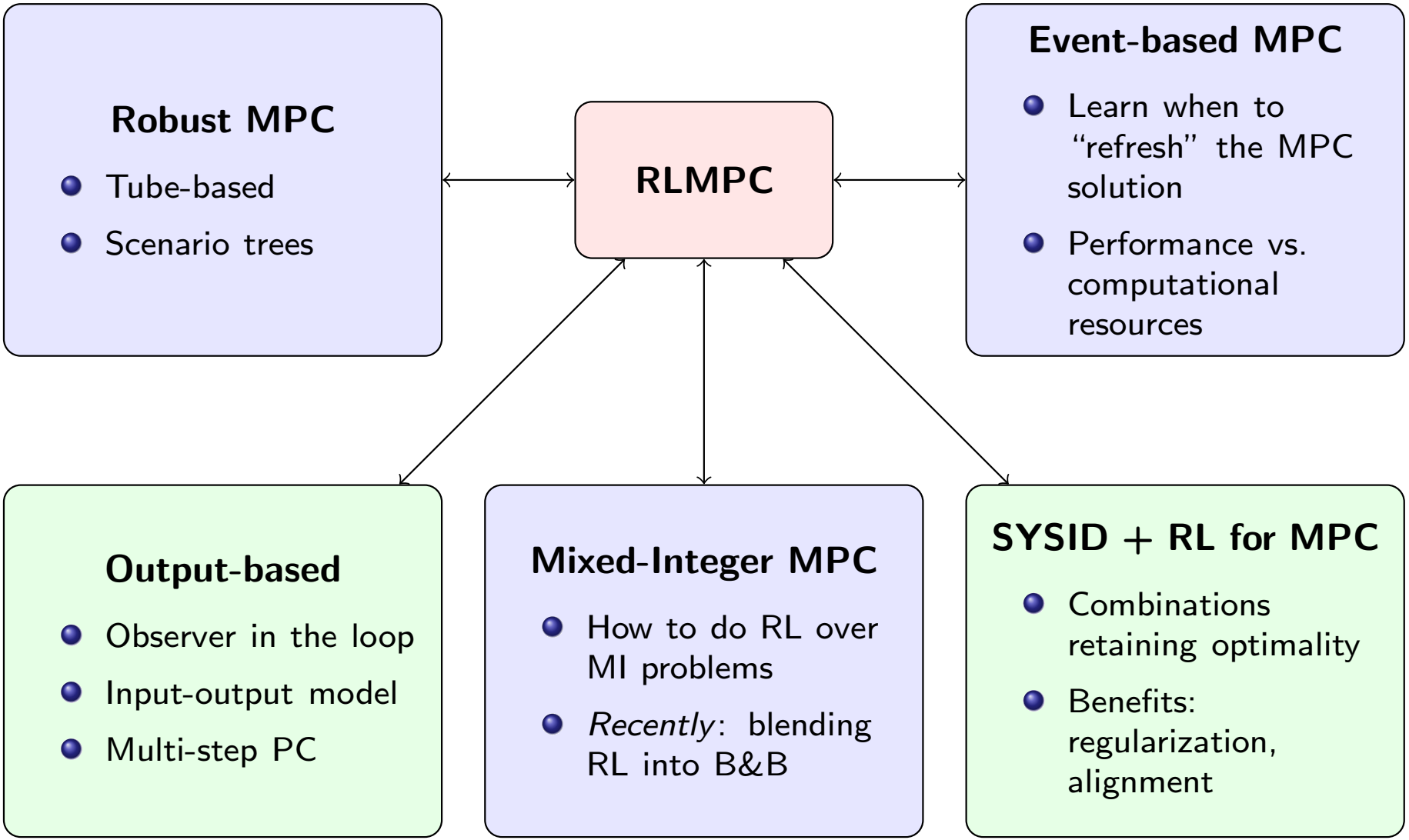
Explored questions - Types of MPC & Learning



Explored questions - Types of MPC & Learning



Explored questions - Types of MPC & Learning



Explored questions - Algorithms

RLMPC

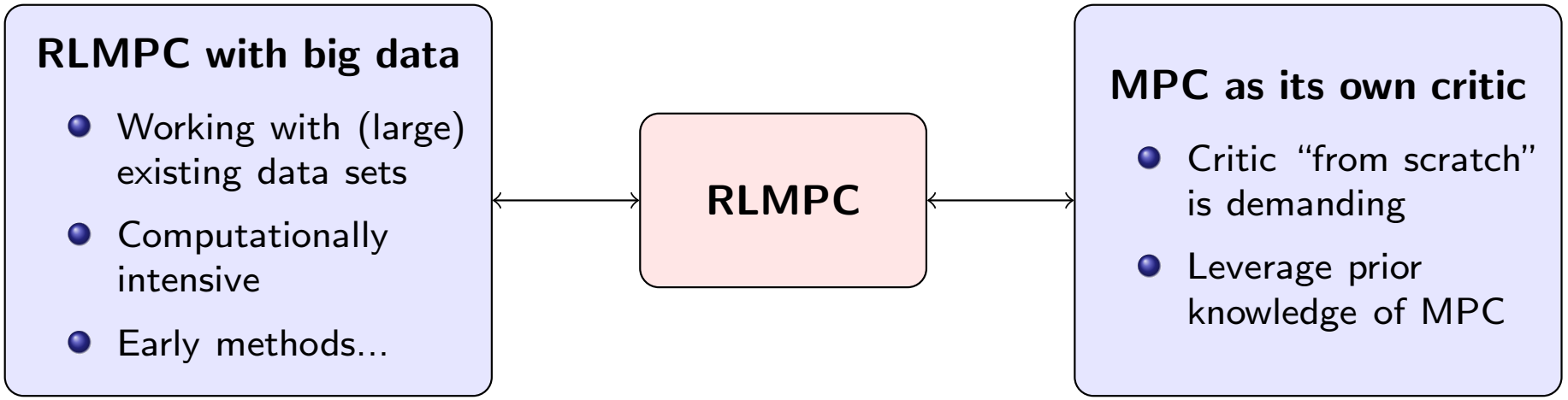
Explored questions - Algorithms

RLMPC with big data

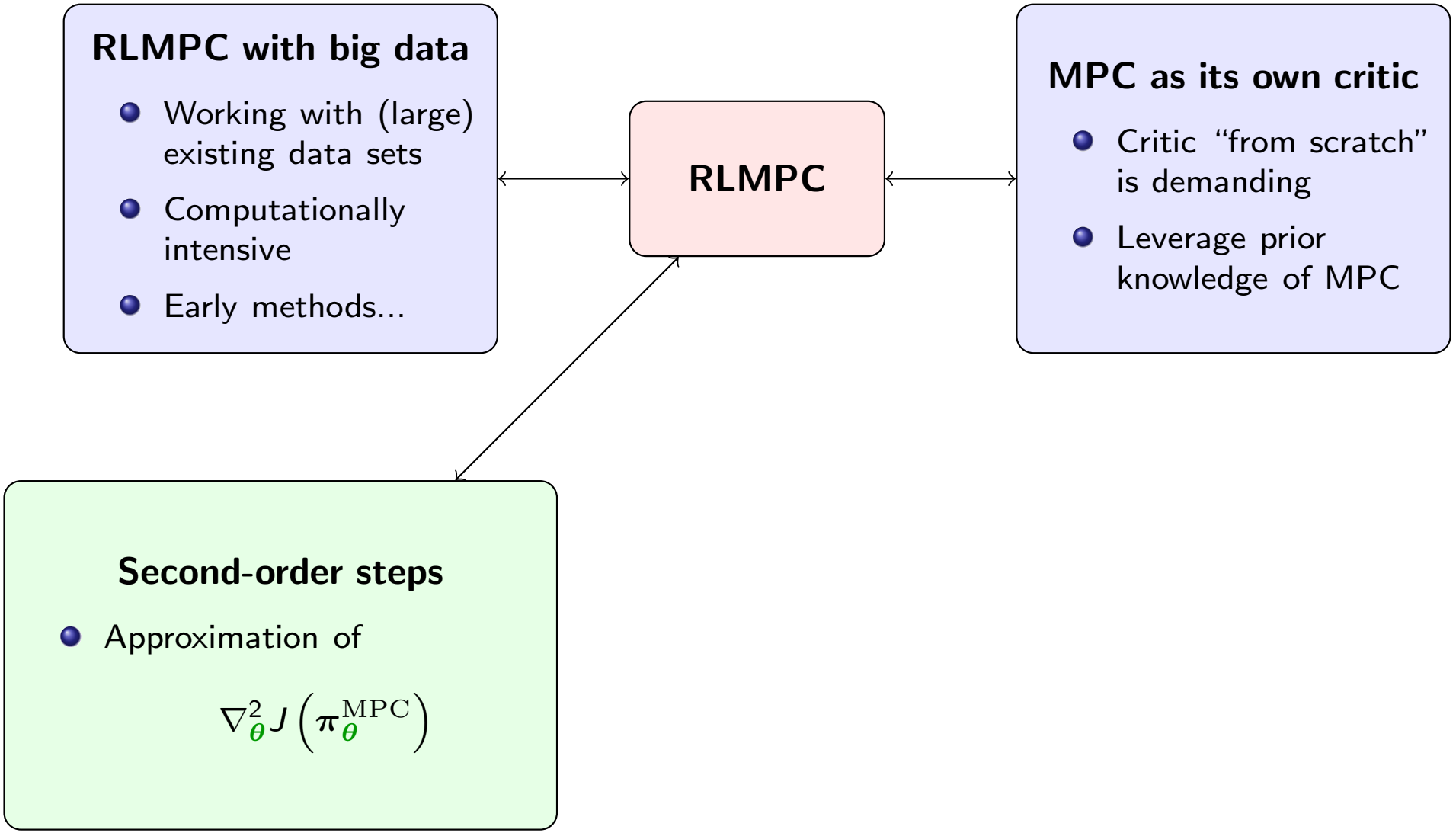
- Working with (large) existing data sets
- Computationally intensive
- Early methods...

RLMPC

Explored questions - Algorithms



Explored questions - Algorithms



Explored questions - Algorithms

RLMPC with big data

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

RLMPC

MPC as its own critic

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

Second-order steps

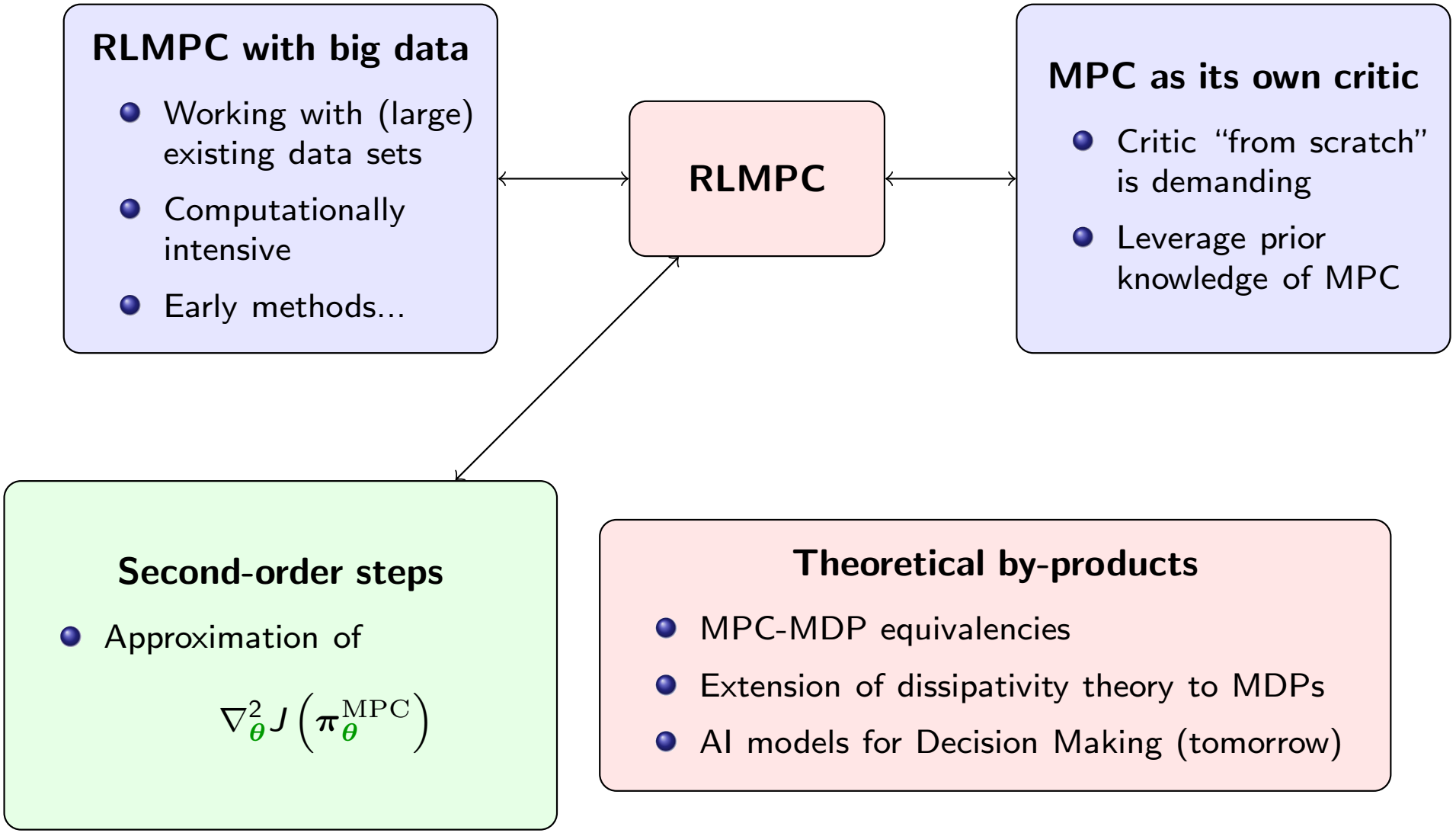
- Approximation of

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

Theoretical by-products

- MPC-MDP equivalencies
- Extension of dissipativity theory to MDPs
- AI models for Decision Making (tomorrow)

Explored questions - Algorithms



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

Outline

Belief States – What is it about?

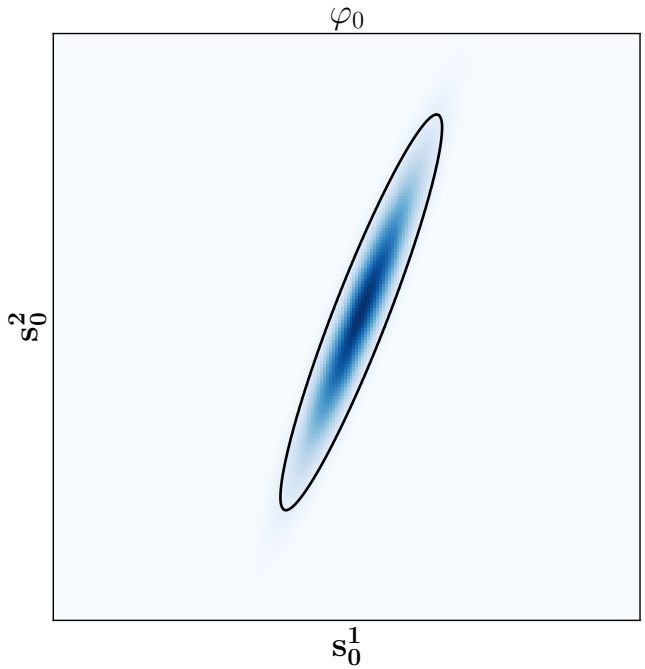
- In **Regular MDP**s, state $\mathbf{s}$ is known
- Policy $\mathbf{a} = \boldsymbol{\pi}^\star(\mathbf{s})$
- Can work with $\mathbf{s}$ =raw recent data
- Often there is a latent space construction
- Markovian property more or less explicitly promoted
- Estimate $\mathbf{s}_k$ from observations $\mathbf{o}_{0,\ldots,k}$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

imperfect knowledge of $\mathbf{s}_k$

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



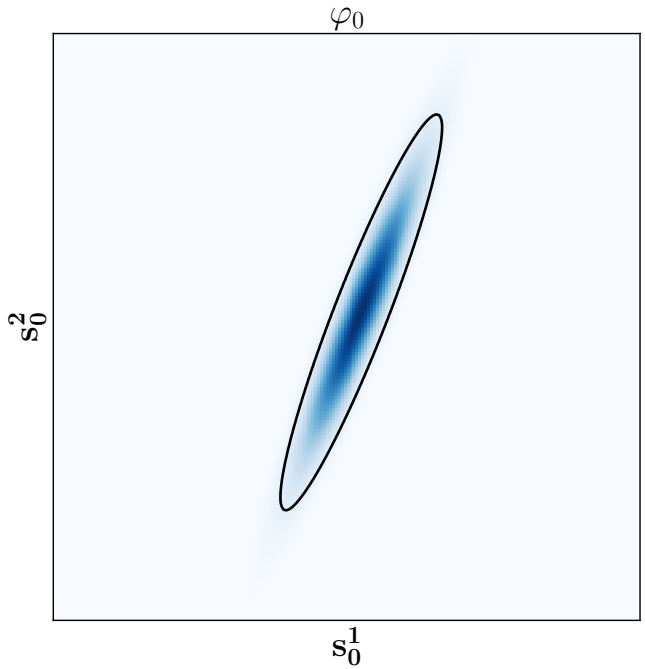
How to "encode" that imperfect knowledge in decision making?

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



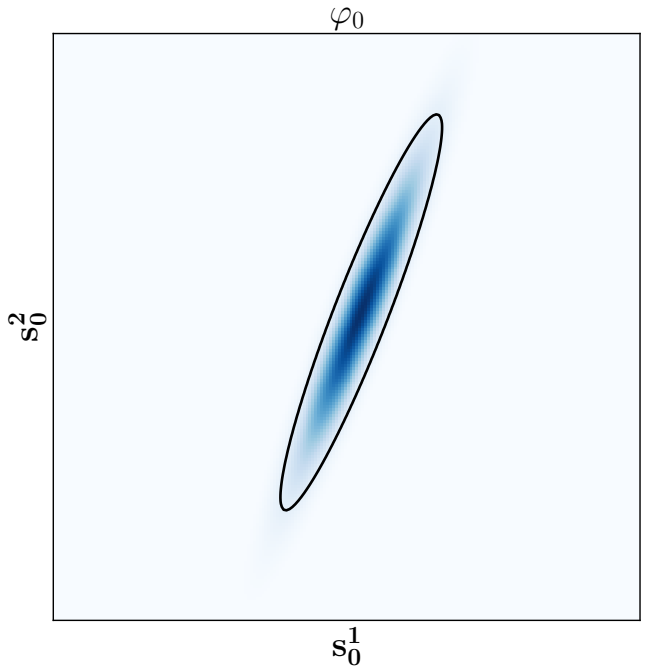
How to "encode" that imperfect knowledge in decision making?

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



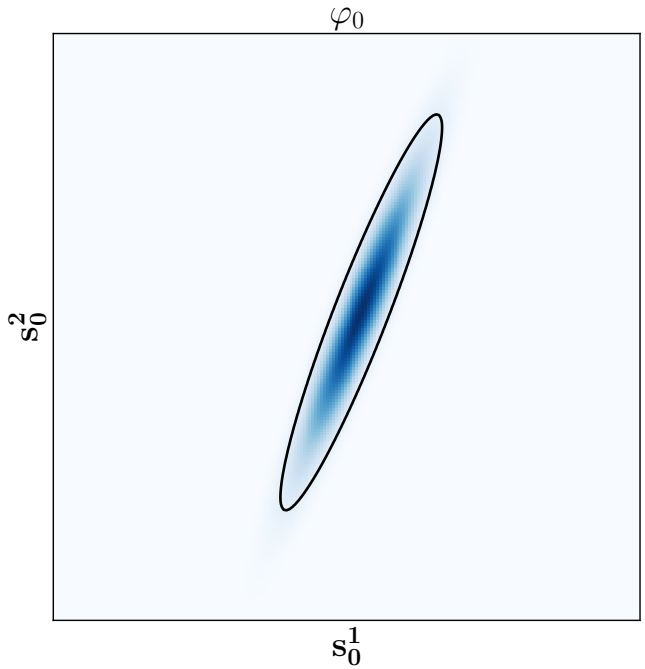
How to "encode" that imperfect knowledge in decision making?

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

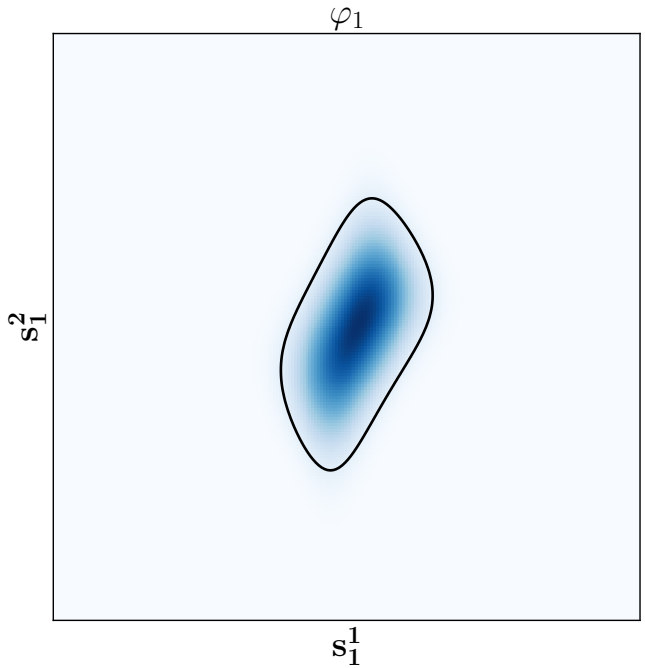
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

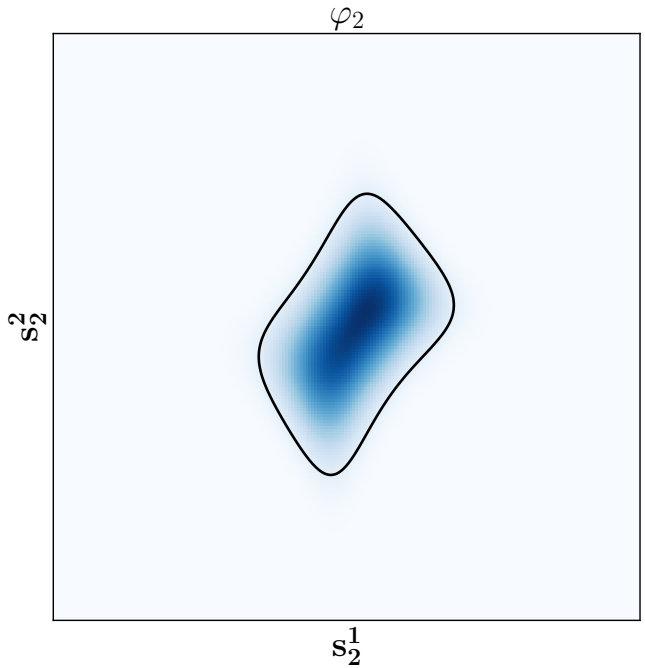
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

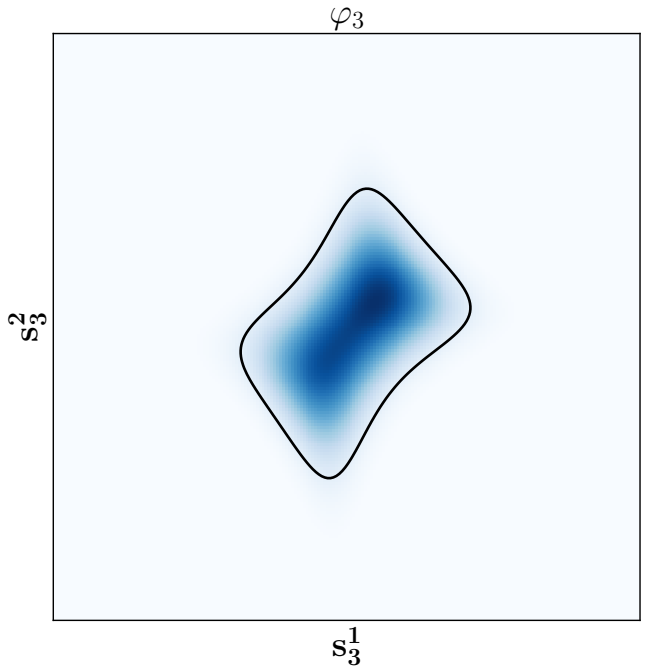
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

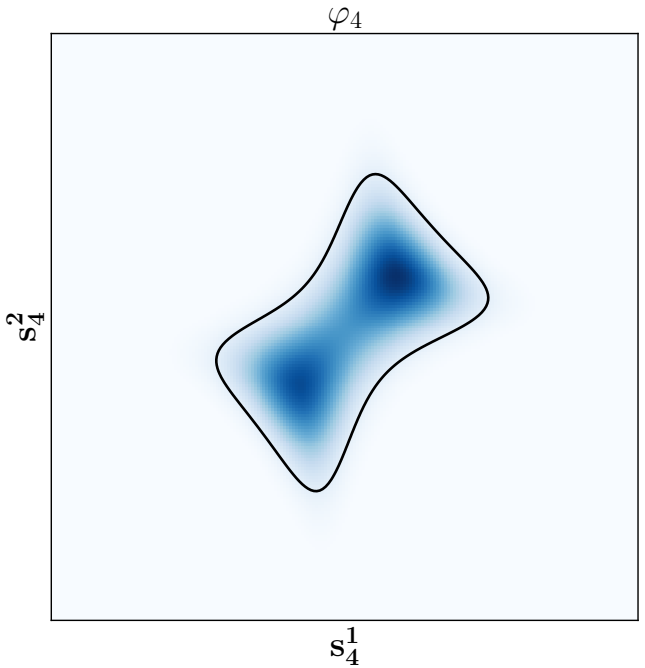
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

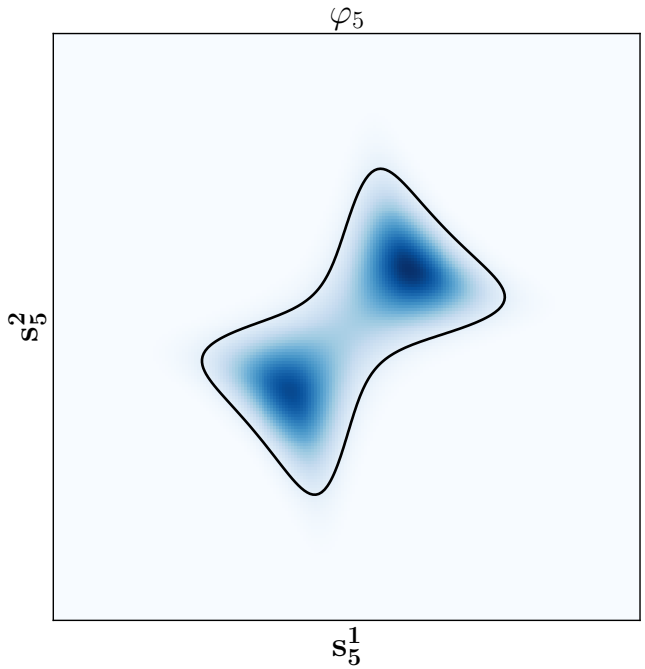
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

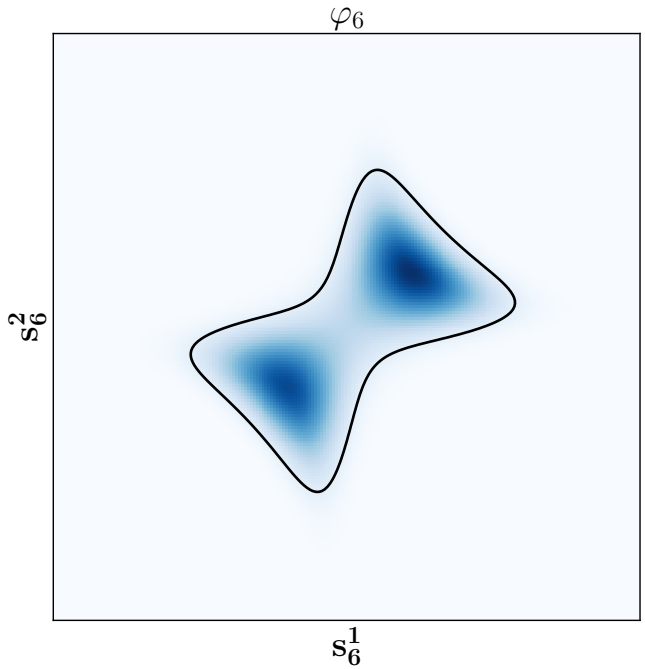
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

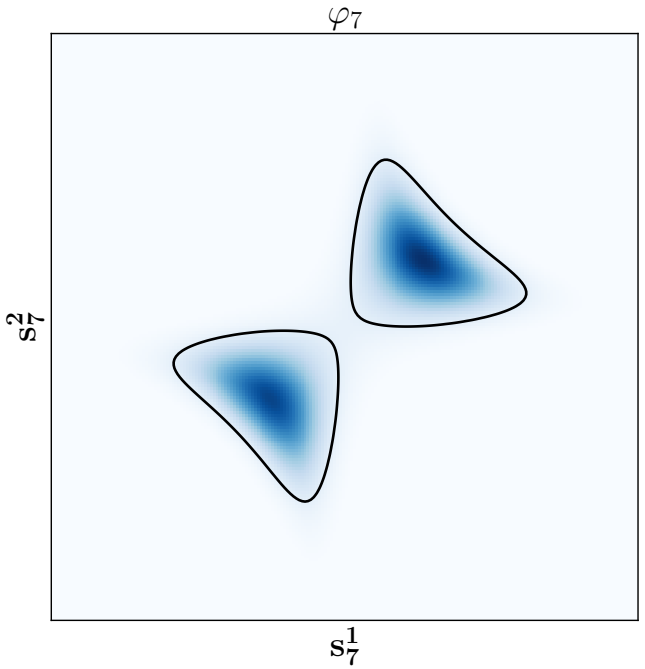
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

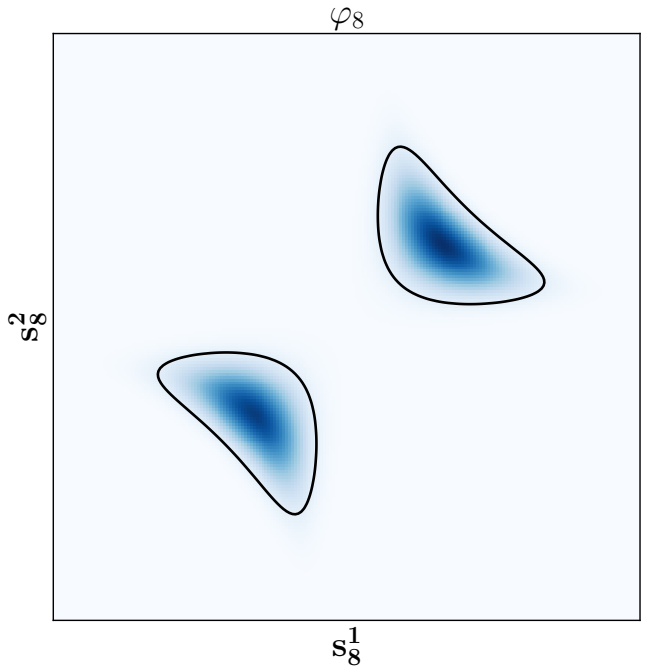
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

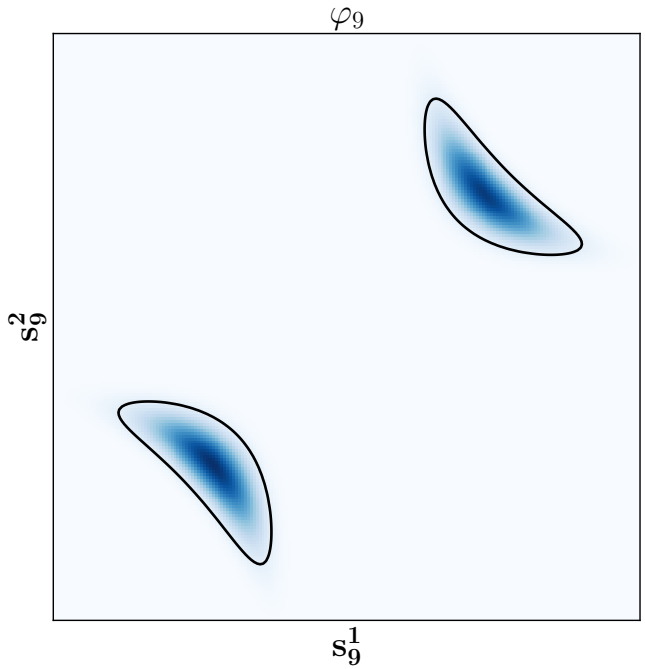
- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- Bayesian perspective: φ_k represents imperfect knowledge, it is not frequentist
- Kalman filter: φ_k is Gaussian, described via mean and covariance

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

Belief States – What is it about?

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

imperfect knowledge of $\mathbf{s}_k$



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

- Belief state: $\mathbf{s}_k \sim \varphi_k(\cdot)$
- 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 $\varphi_k \rightarrow \varphi_{k+1}$, tied to $\mathbf{a}_k$ and $\mathbf{o}_{0,\ldots,k+1}$ (and φ_0)

Belief State Dynamics – How does it work?

Dynamics model

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

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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

Belief State Dynamics – How does it work?

Dynamics model

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

Observation model

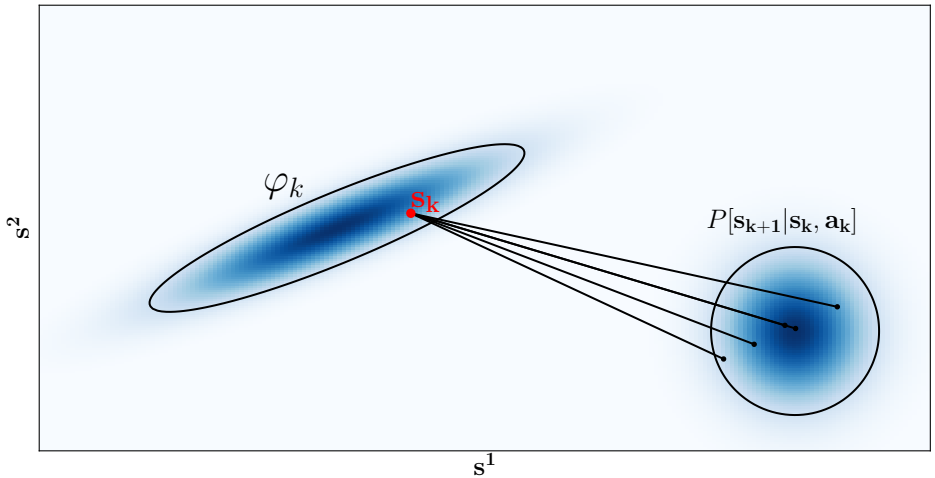
$$\mathbf{o}_k \sim \mathcal{O}\left[\,\cdot \,|\, \mathbf{s}_k\,\right]$$

State $\mathbf{s}_k$ is "partially known", i.e.

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

Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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



Belief State Dynamics – How does it work?

Dynamics model

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

Observation model

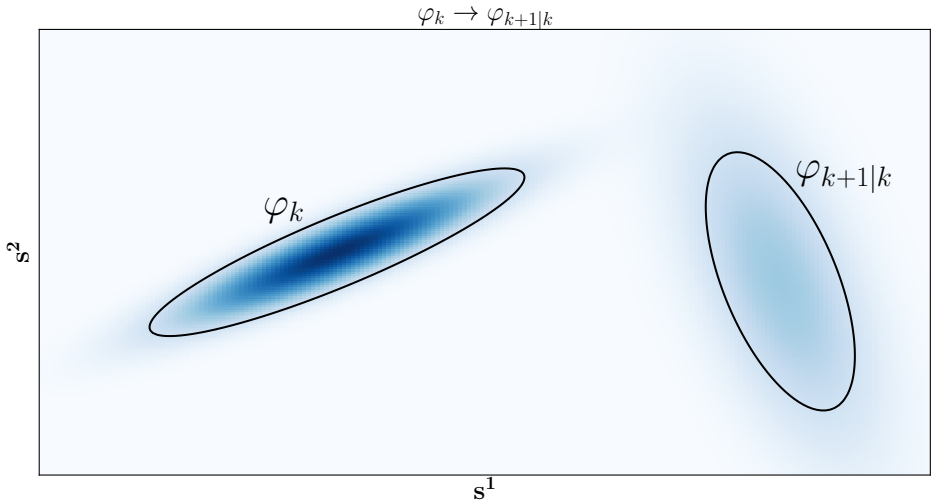
$$\mathbf{o}_k \sim \mathcal{O}\left[\,\cdot \,|\, \mathbf{s}_k\,\right]$$

State $\mathbf{s}_k$ is "partially known", i.e.

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

Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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



Belief State Dynamics – How does it work?

Dynamics model

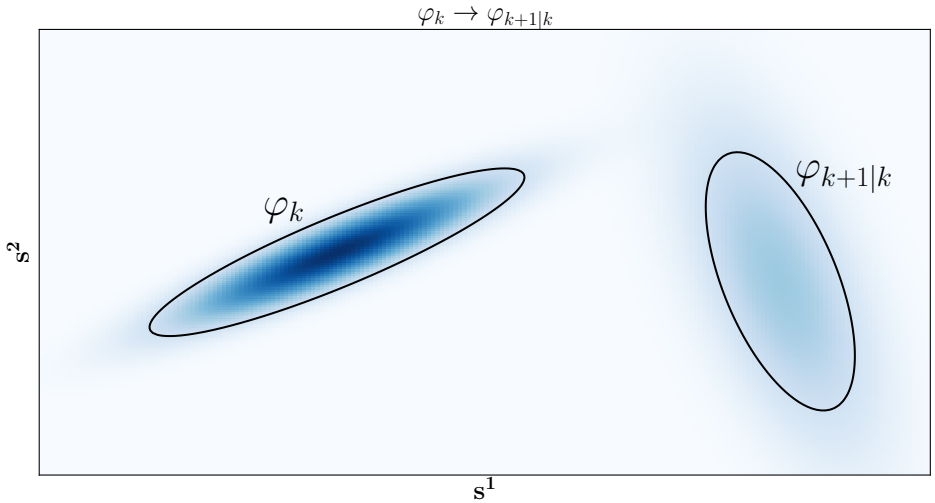
$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k\right]$$

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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



Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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

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

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

Belief State Dynamics – How does it work?

Dynamics model

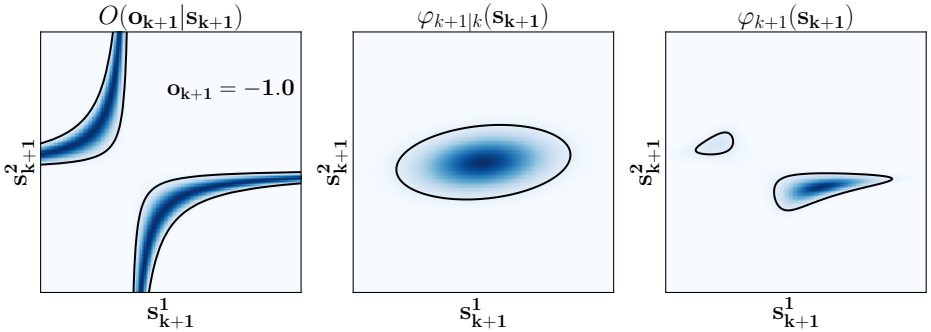
$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k\right]$$

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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



Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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

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

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

Belief State Dynamics – How does it work?

Dynamics model

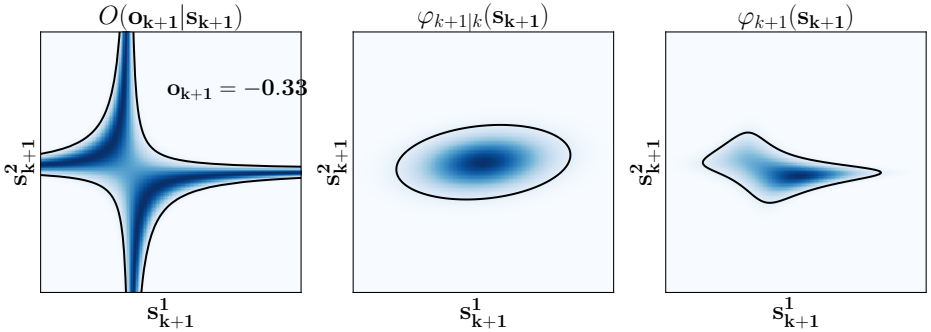
$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k\right]$$

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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



Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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

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

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

Belief State Dynamics – How does it work?

Dynamics model

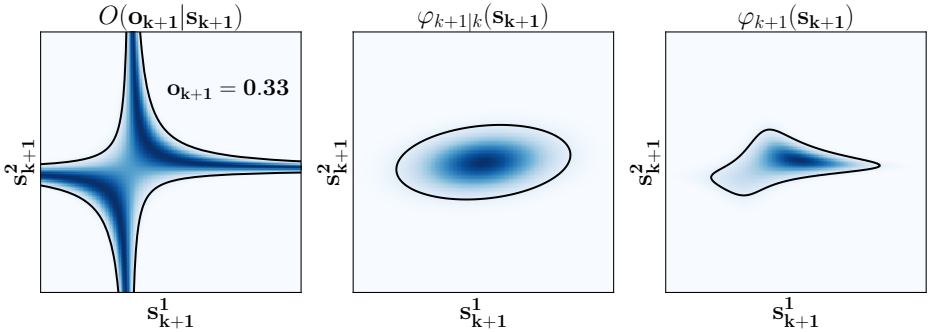
$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k\right]$$

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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



Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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

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

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

Belief State Dynamics – How does it work?

Dynamics model

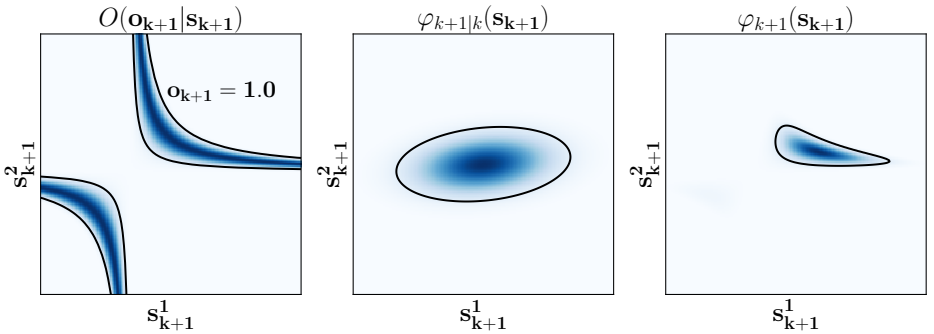
$$\mathbf{s}_{k+1} \sim \mathbb{P}\left[\,\cdot \,|\, \mathbf{s}_k, \mathbf{a}_k\right]$$

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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



Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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

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

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

Belief State Dynamics – How does it work?

Dynamics model

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

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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

Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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

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 $\mathbf{o}_{k+1}$, new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

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

Belief State Dynamics – How does it work?

Dynamics model

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

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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

Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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

Remarks state transition

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

- Is in general nonlinear
- Is deterministic for $\mathbf{o}_{k+1}$ known

Posterior to observation $\mathbf{o}_{k+1}$, new info *corrects* $\varphi_{k+1|k} \rightarrow \varphi_{k+1}$:

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

Belief State Dynamics – How does it work?

Dynamics model

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

Observation model

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

State $\mathbf{s}_k$ is "partially known", i.e.

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

Prior to observation $\mathbf{o}_{k+1}$, φ_{k+1} is

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

Remarks state transition

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

- Is in general nonlinear
- Is deterministic for $\mathbf{o}_{k+1}$ known
- Is stochastic prior to $\mathbf{o}_{k+1}$ 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 $\mathbf{o}_{k+1}$, new info *corrects* $\varphi_{k+1|k} \to \varphi_{k+1}$:

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

Illustrative example - Belief state dynamics

Dynamics model

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

Observation model (nonlinear)

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

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

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

Illustrative example - Belief state dynamics

Dynamics model

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

Observation model (nonlinear)

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

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

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

Posterior to observation $\mathbf{o}_{k+1}$, belief state $\varphi_{k+1} = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right)$ is not Gaussian

Illustrative example - Belief state dynamics

Dynamics model

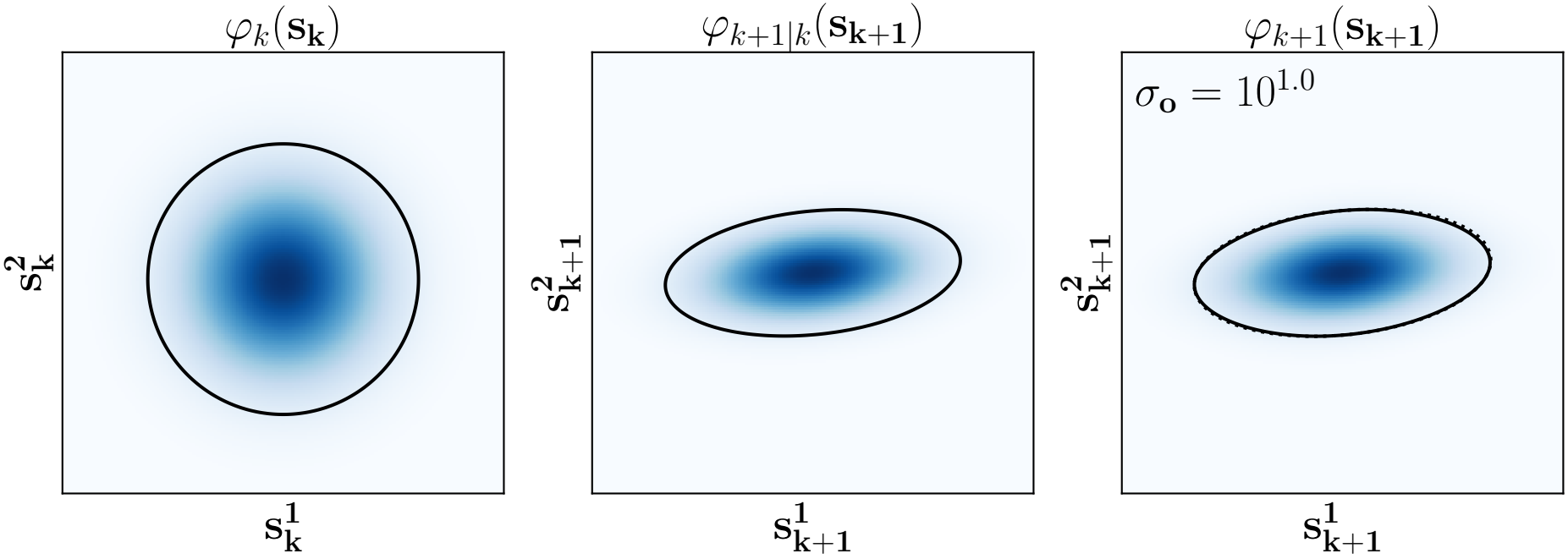
$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k,\, \Sigma_{\mathbf{s}}\right)$$

Observation model (nonlinear)

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

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

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



Posterior to observation $\mathbf{o}_{k+1}$, belief state $\varphi_{k+1} = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right)$ is not Gaussian

Illustrative example - Belief state dynamics

Dynamics model

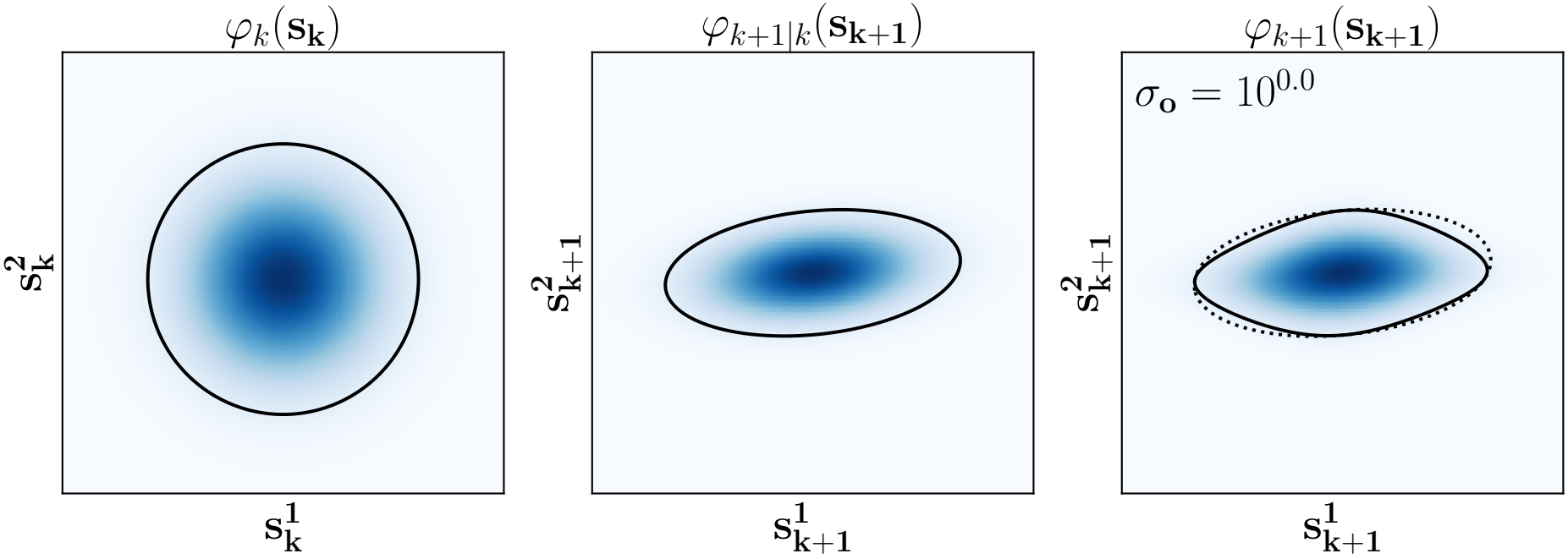
$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k,\, \Sigma_{\mathbf{s}}\right)$$

Observation model (nonlinear)

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

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

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



Posterior to observation $\mathbf{o}_{k+1}$, belief state $\varphi_{k+1} = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right)$ is not Gaussian

Illustrative example - Belief state dynamics

Dynamics model

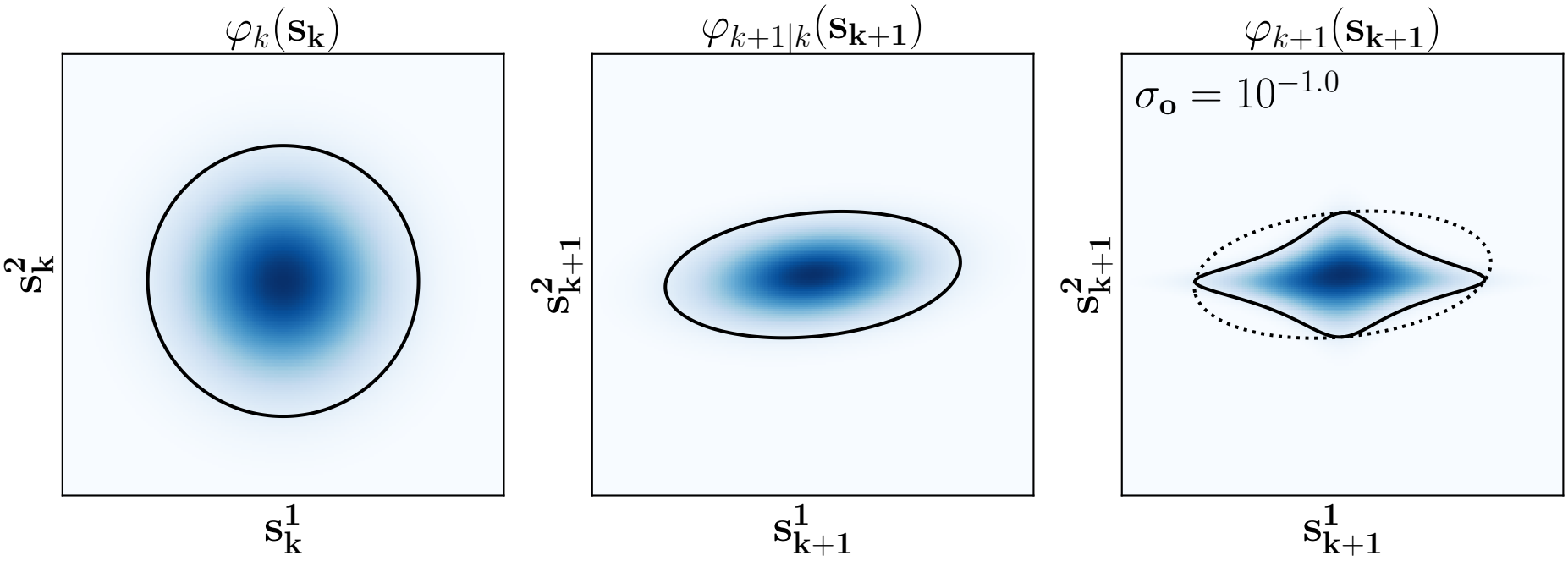
$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k,\, \Sigma_{\mathbf{s}}\right)$$

Observation model (nonlinear)

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

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

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



Posterior to observation $\mathbf{o}_{k+1}$, belief state $\varphi_{k+1} = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right)$ is not Gaussian

Illustrative example - Belief state dynamics

Dynamics model

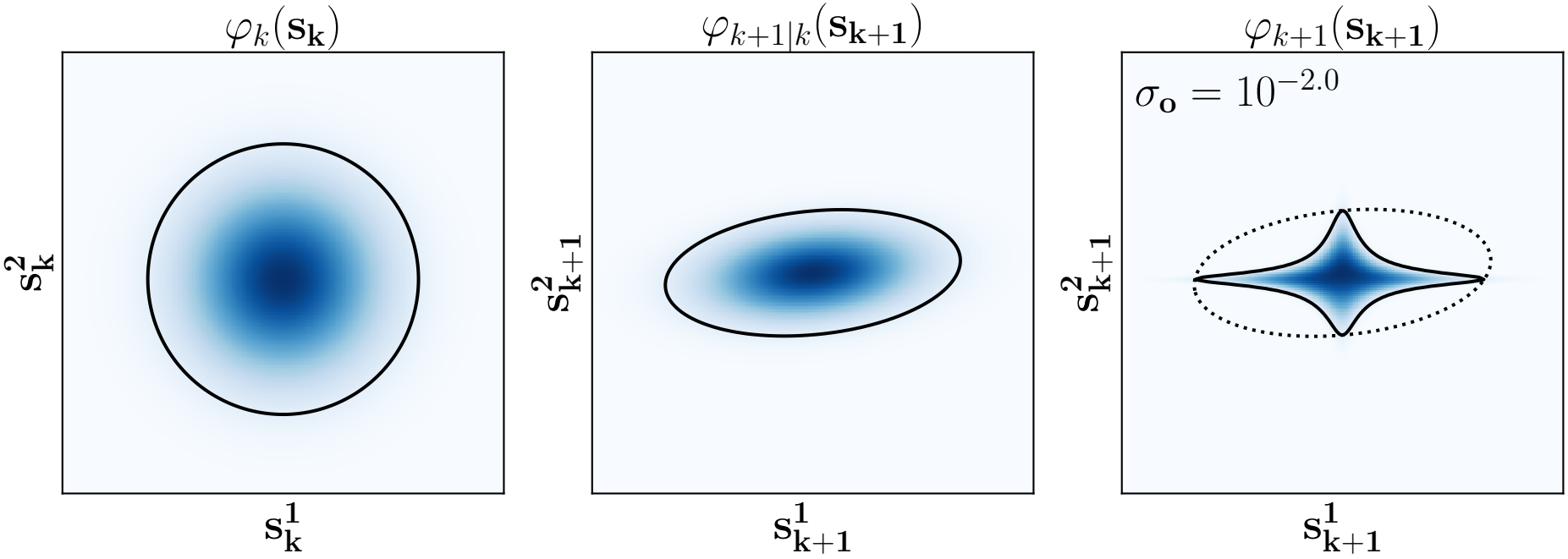
$$\mathbf{s}_{k+1} \sim \mathcal{N}\left(A\mathbf{s}_k + B\mathbf{a}_k,\, \Sigma_{\mathbf{s}}\right)$$

Observation model (nonlinear)

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

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

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



Posterior to observation $\mathbf{o}_{k+1}$, belief state $\varphi_{k+1} = \mathcal{B}\left(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k\right)$ is not Gaussian

Illustrative example - Policy & belief state

Dynamics model

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

Observation model

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

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

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

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)

Illustrative example - Policy & belief state

Dynamics model

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

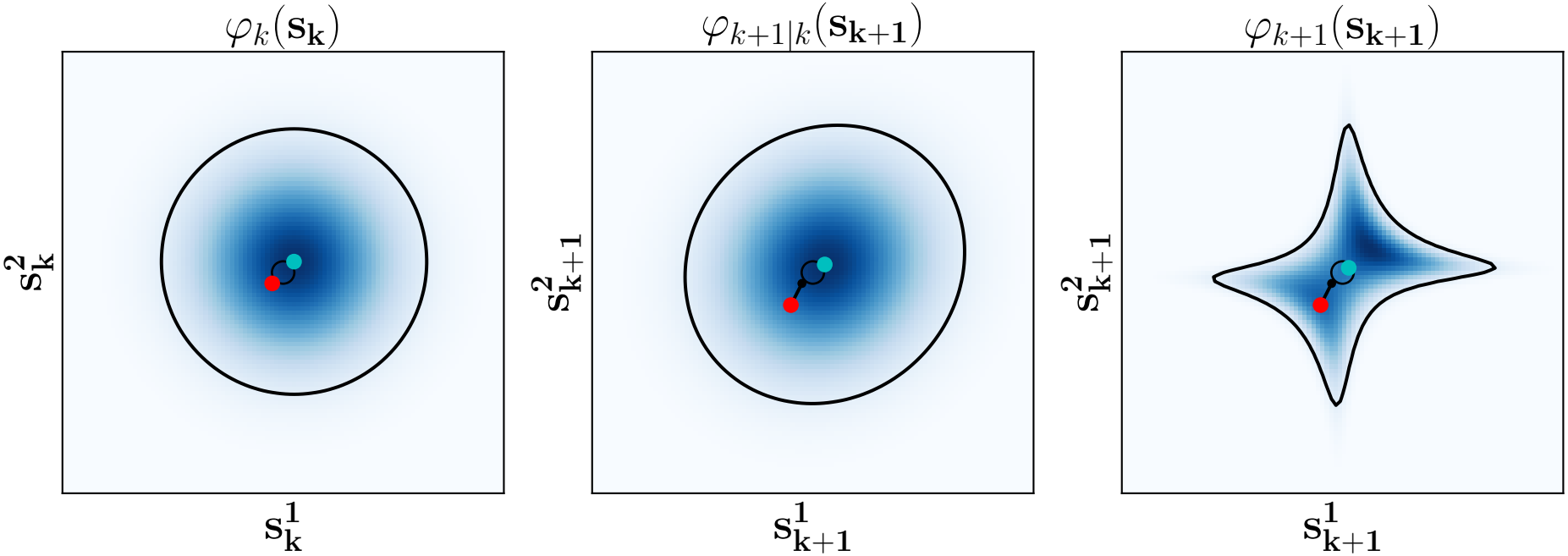
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

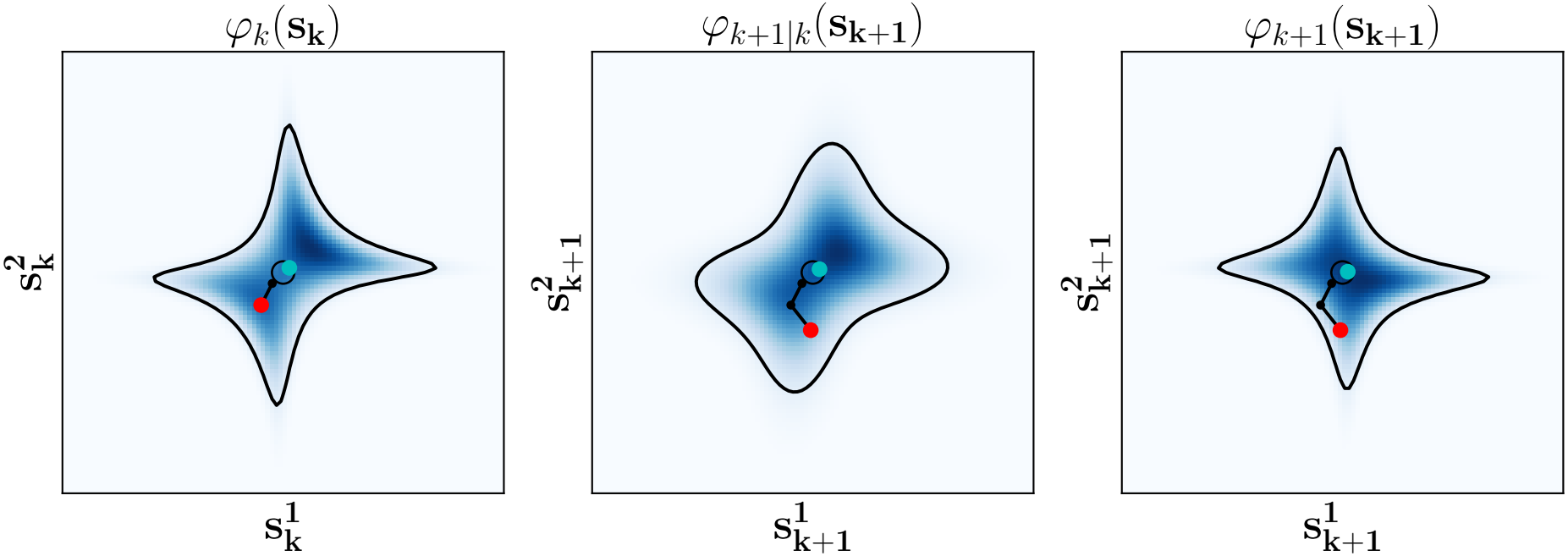
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

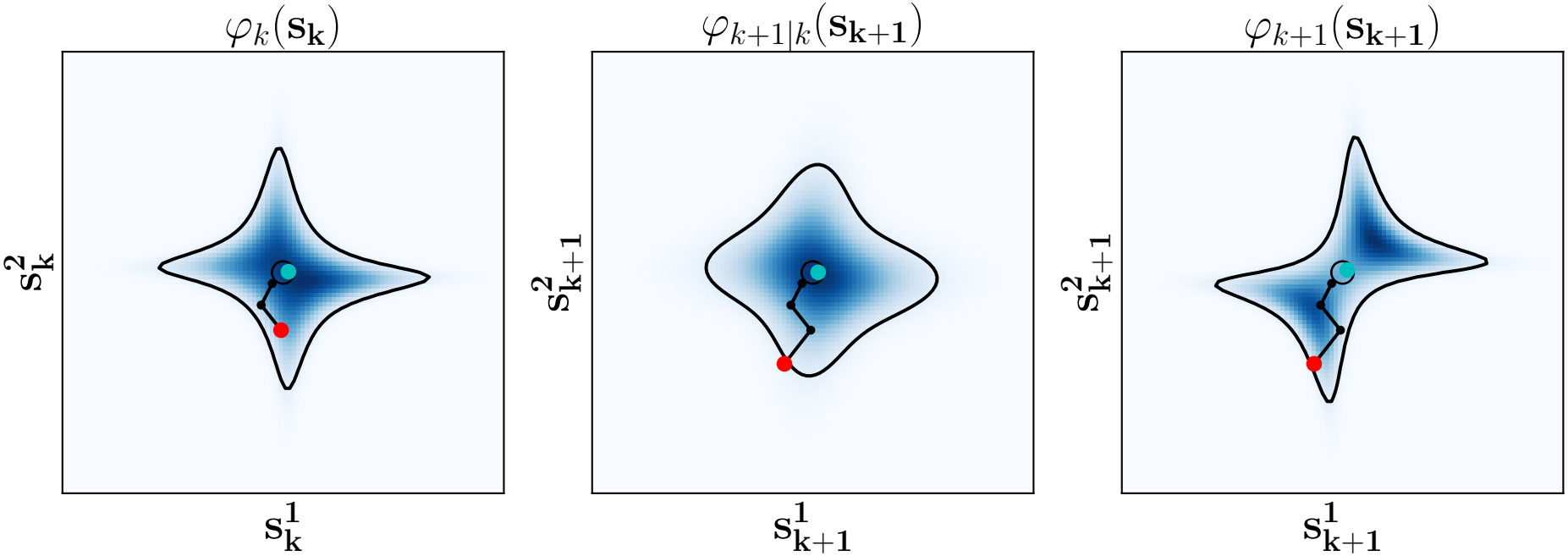
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

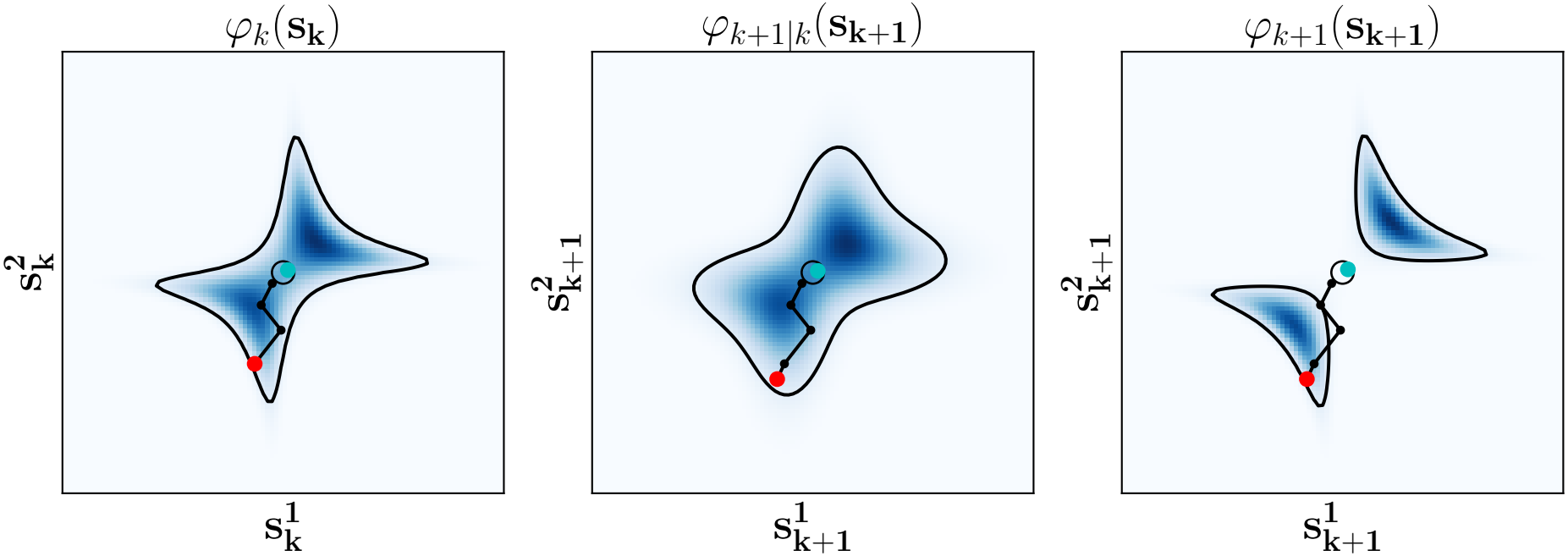
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

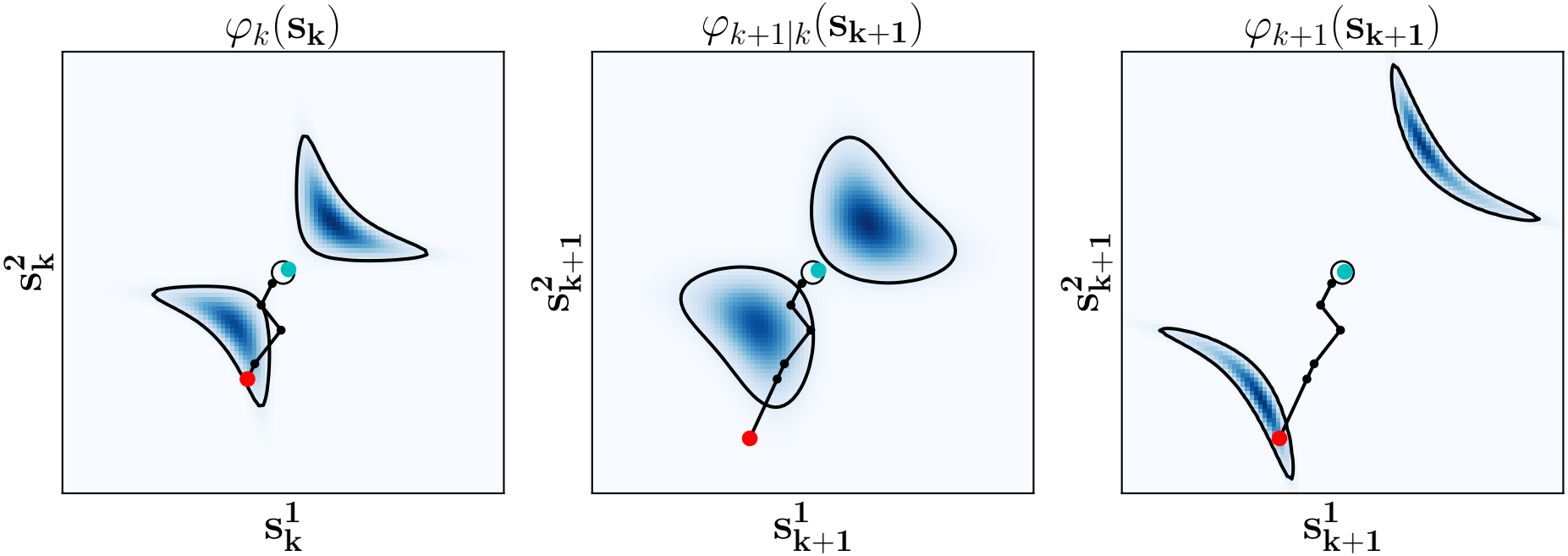
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

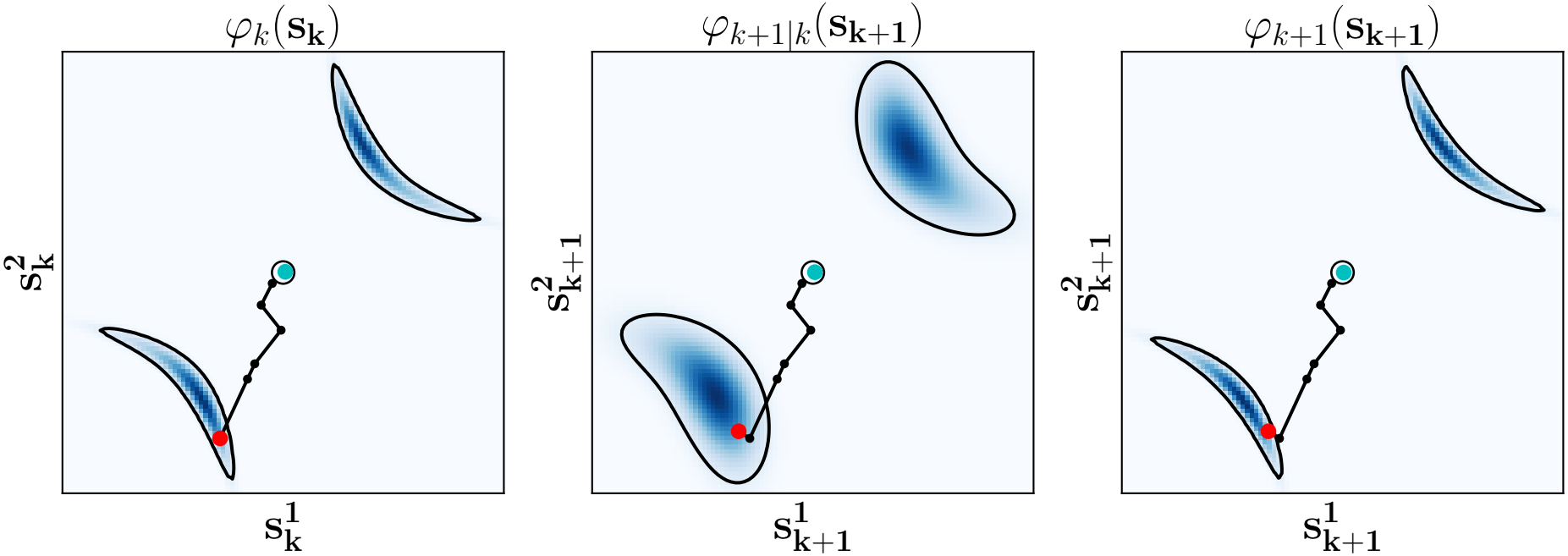
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

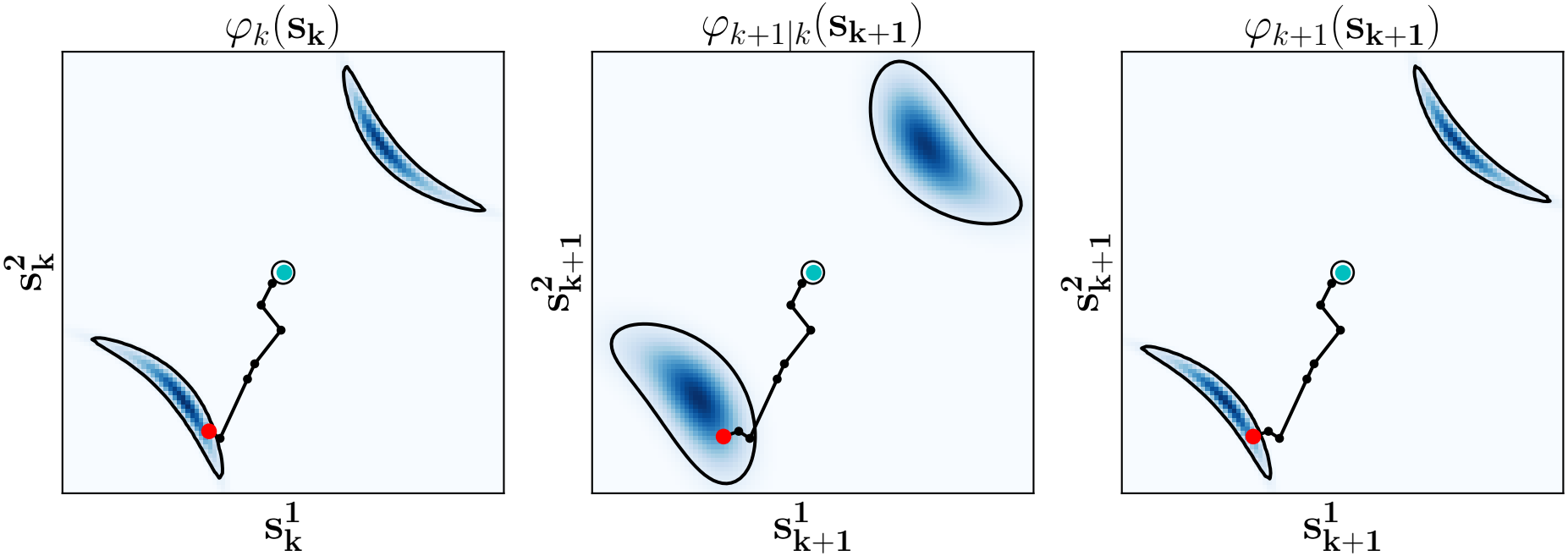
$$A = \begin{bmatrix} 1 & .1 \\ 0 & .1 \end{bmatrix},\; B = \begin{bmatrix} 0 \\ .1 \end{bmatrix},\; M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Observation model

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

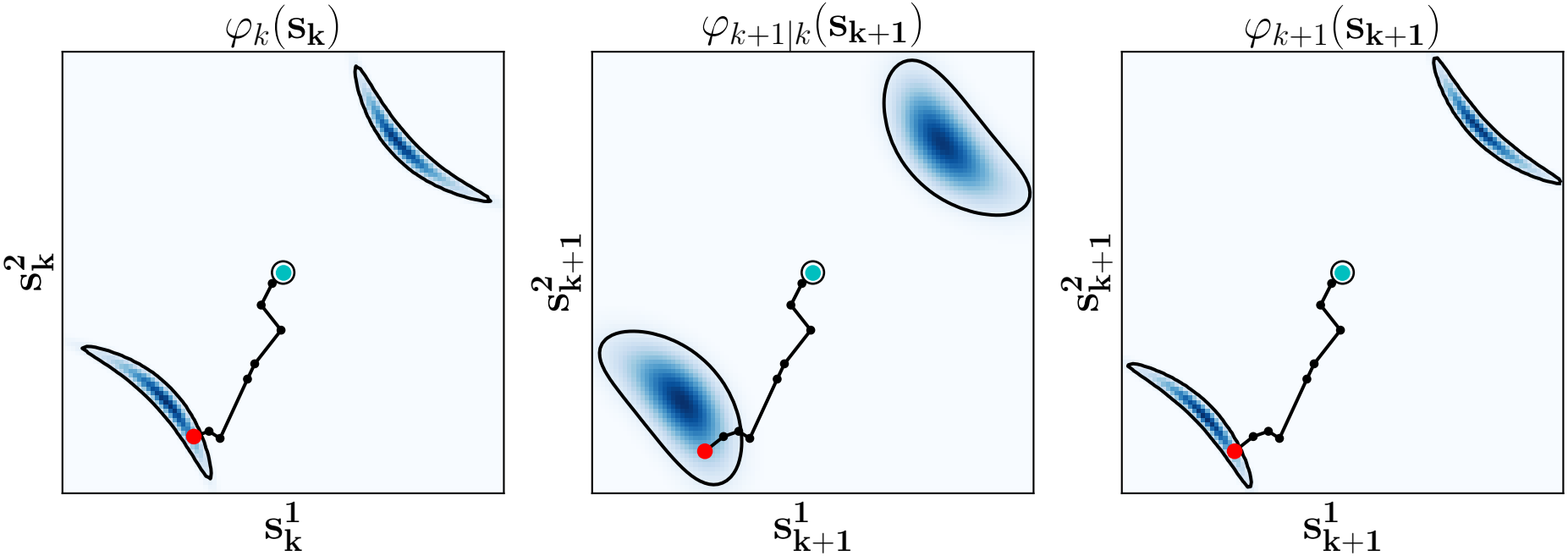
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

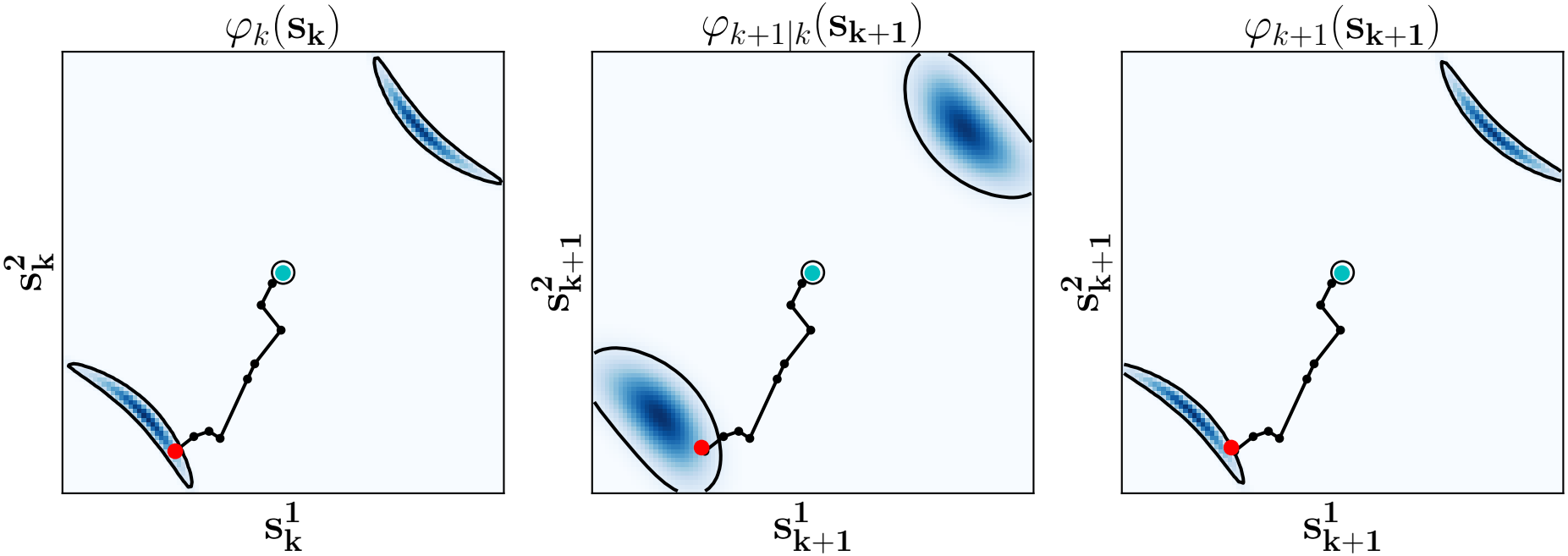
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

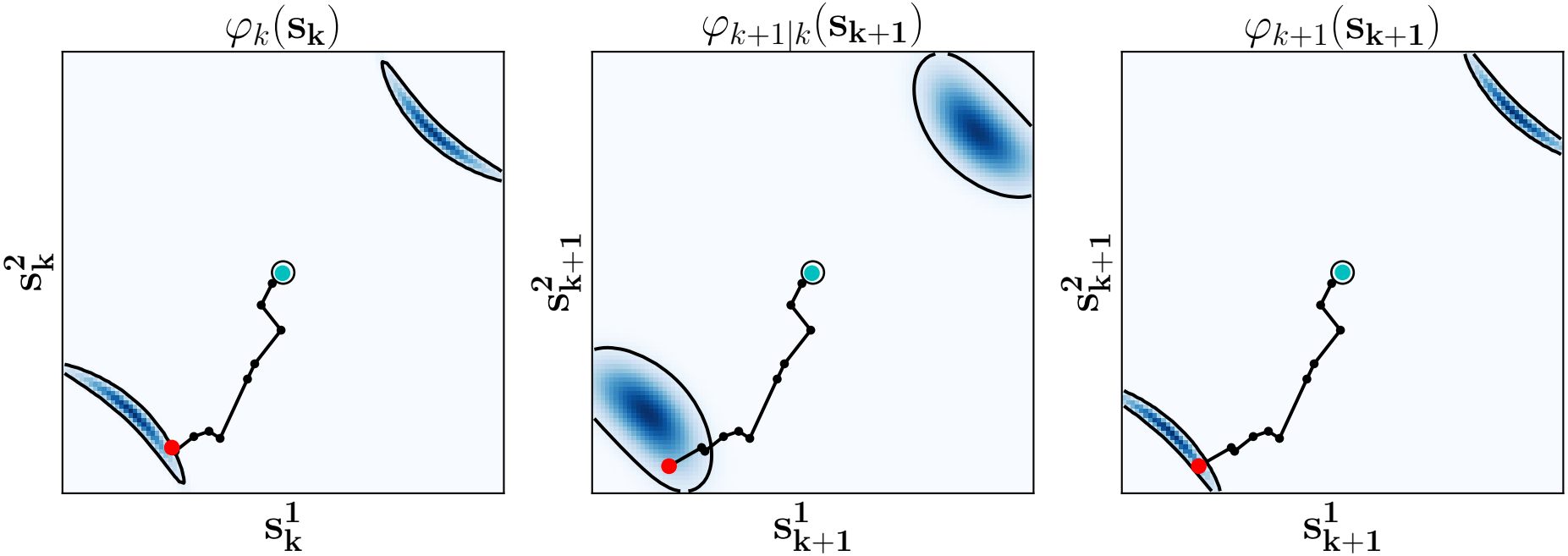
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

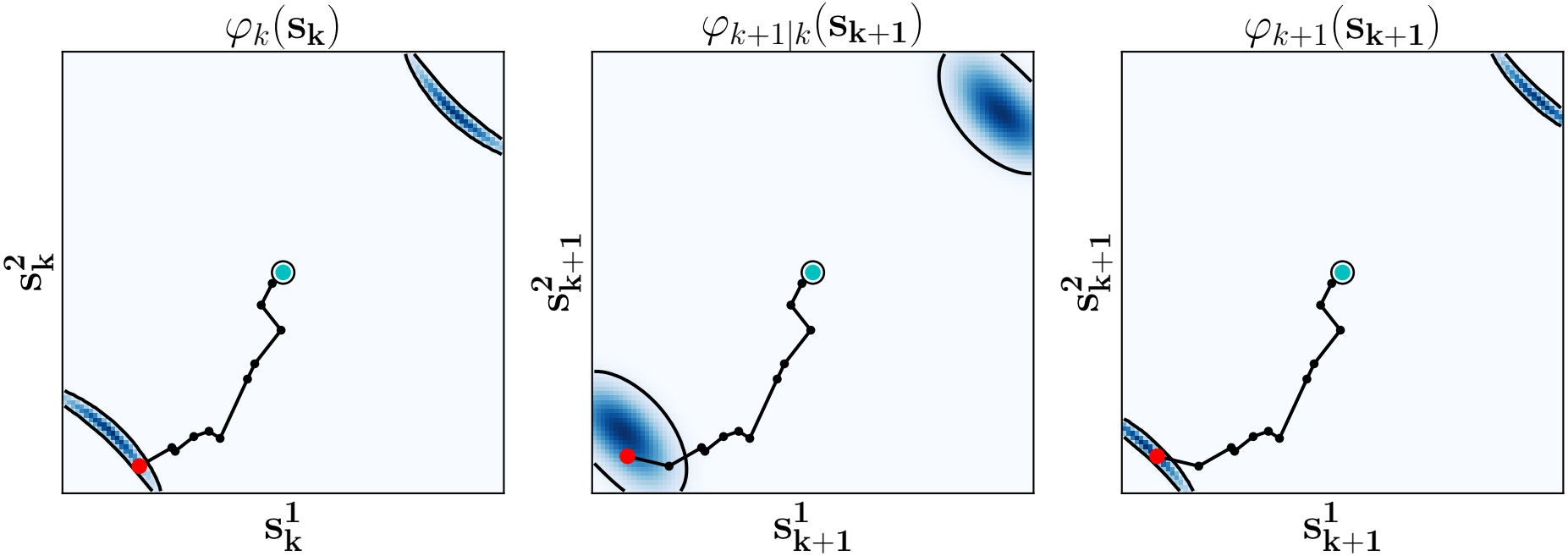
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

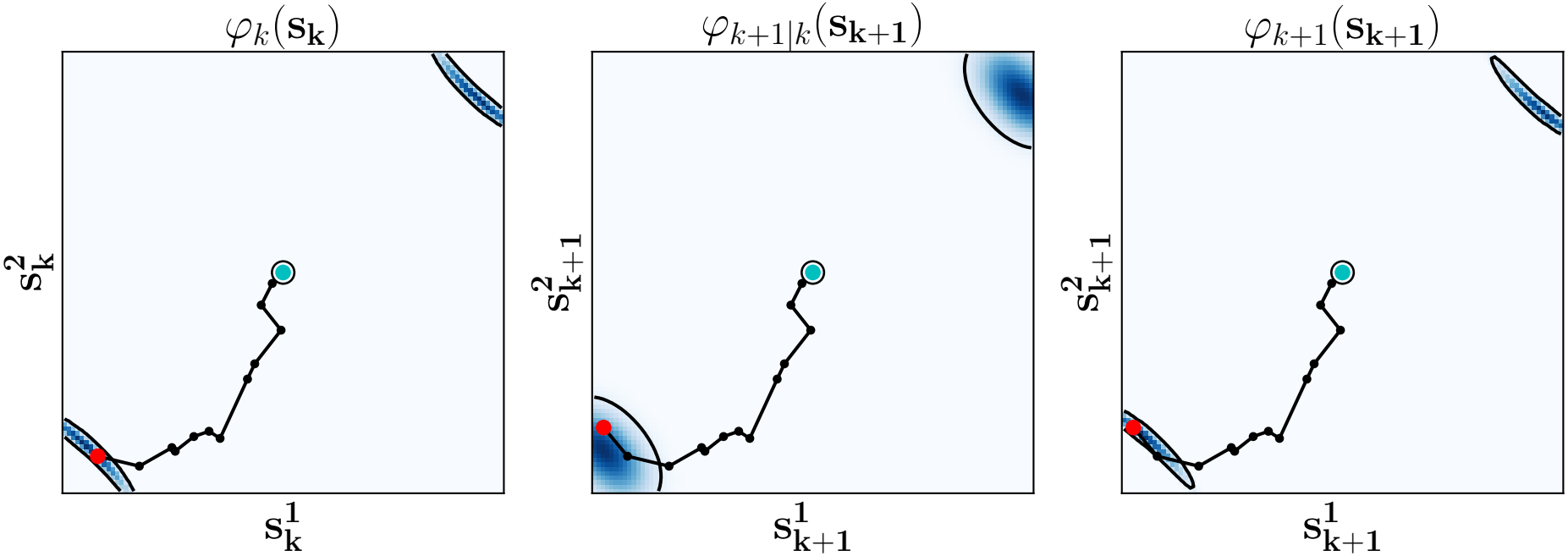
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

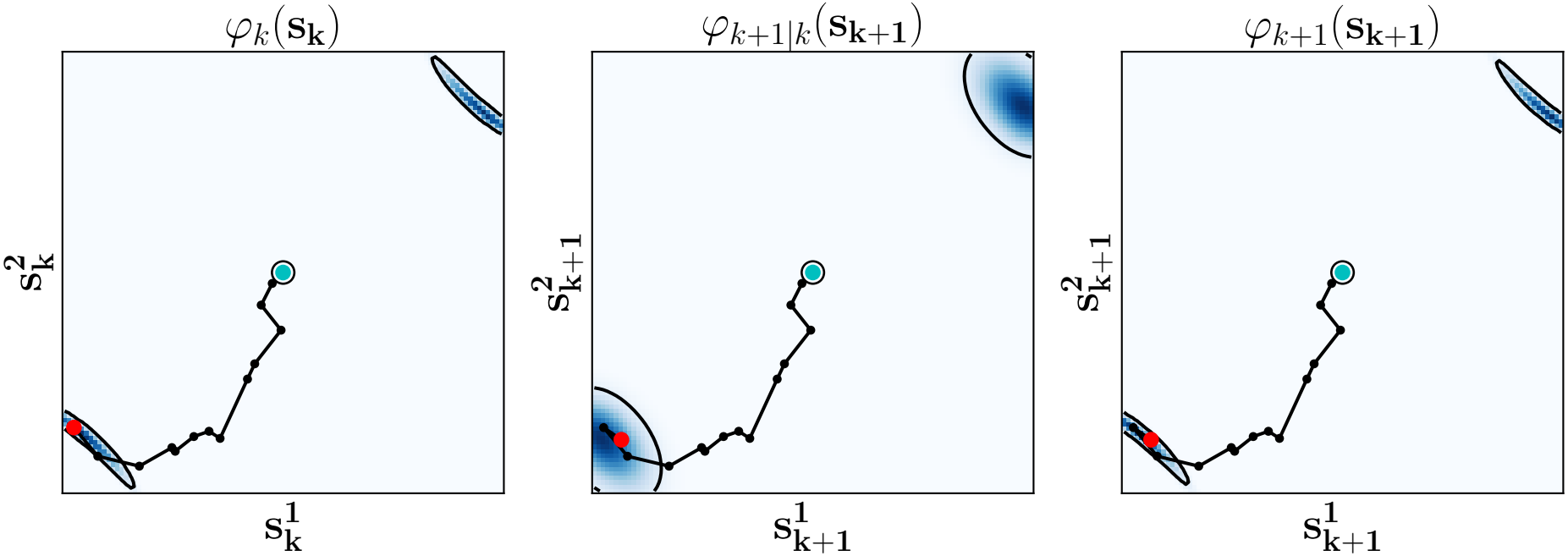
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

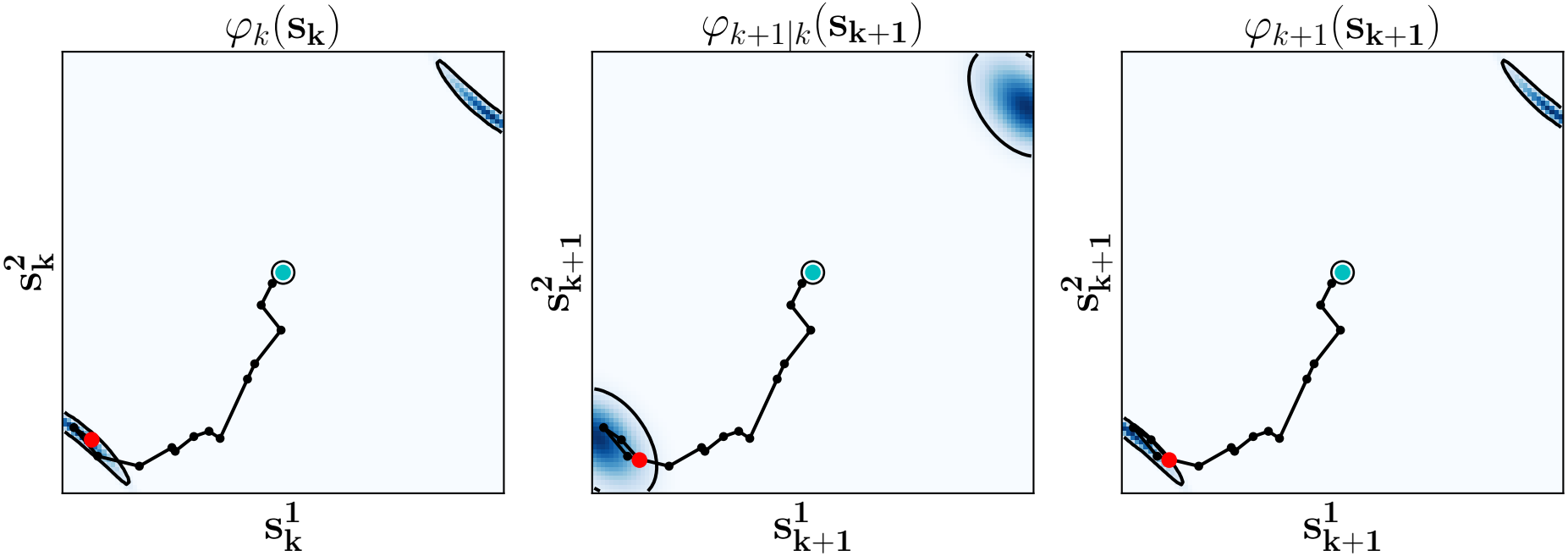
Observation model

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

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

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

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 $\circ = (0, 0)$

Illustrative example - Policy & belief state

Dynamics model

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

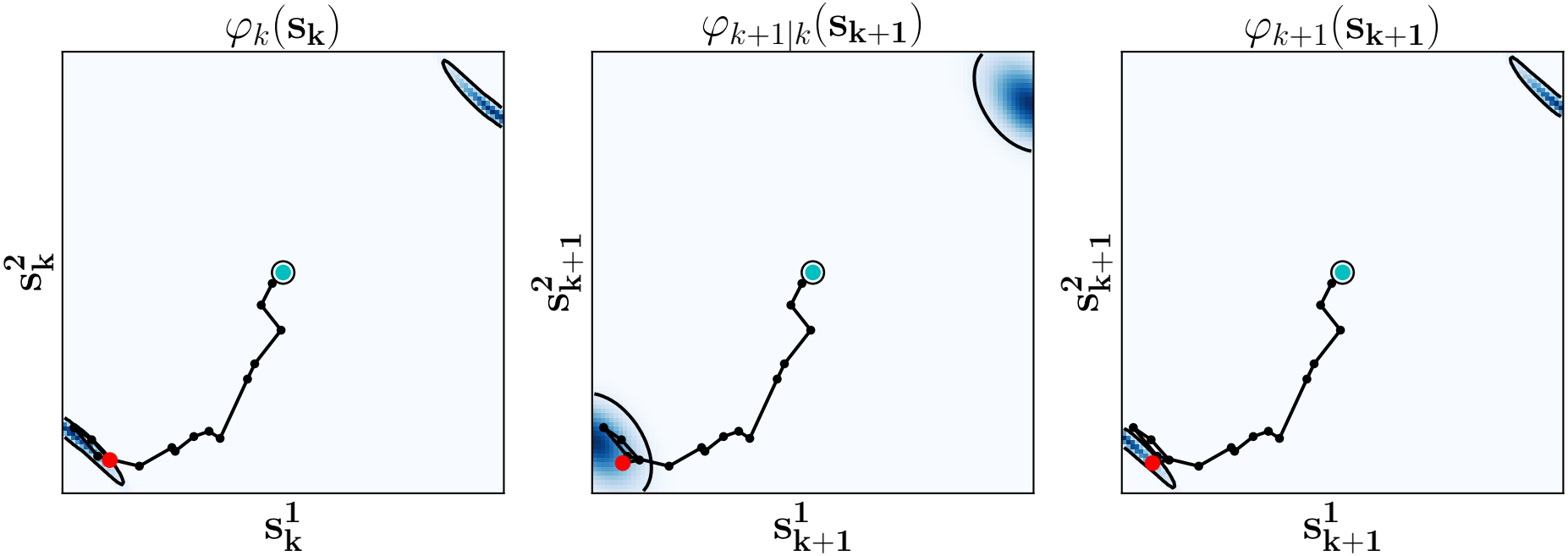
Observation model

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

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

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

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 $\circ = (0, 0)$

MDPs with Belief States

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

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

MDPs with Belief States

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

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

Underlying value functions

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

MDPs with Belief States

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

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

Underlying value functions

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

where

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

MDPs with Belief States

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

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

Underlying value functions

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

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

where

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

MDPs with Belief States

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

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

Underlying value functions

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

A belief state MDP is not only accounting for the uncertainty on $\mathbf{s}$. 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

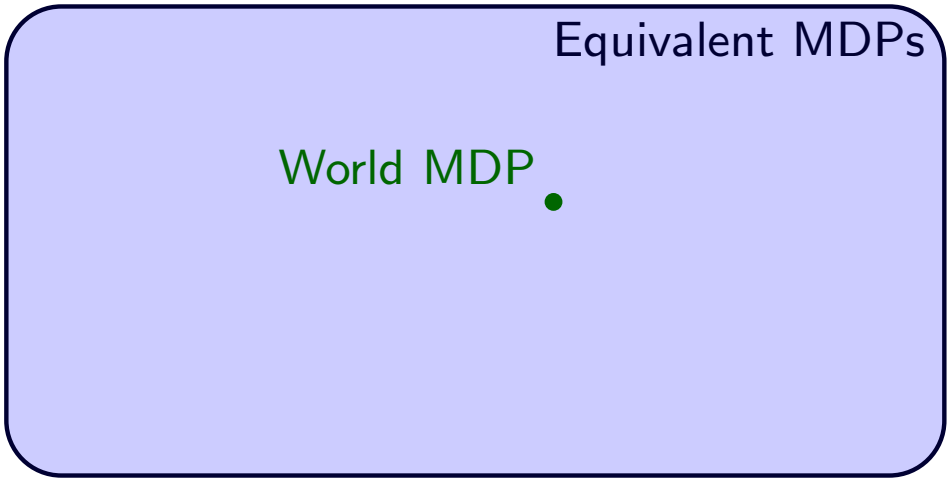
$$\mathbb{E}[V^\star(\varphi_{k+1}) \mid \varphi_k, \mathbf{a}_k] = \int V^\star(\mathcal{B}(\varphi_k, \mathbf{o}_{k+1}, \mathbf{a}_k)) \, \rho[\mathbf{o}_{k+1} \mid \varphi_k, \mathbf{a}_k] \, \mathrm{d}\mathbf{o}_{k+1}$$

Belief States & MPC

World MDP

RLMPC theoretical pathway

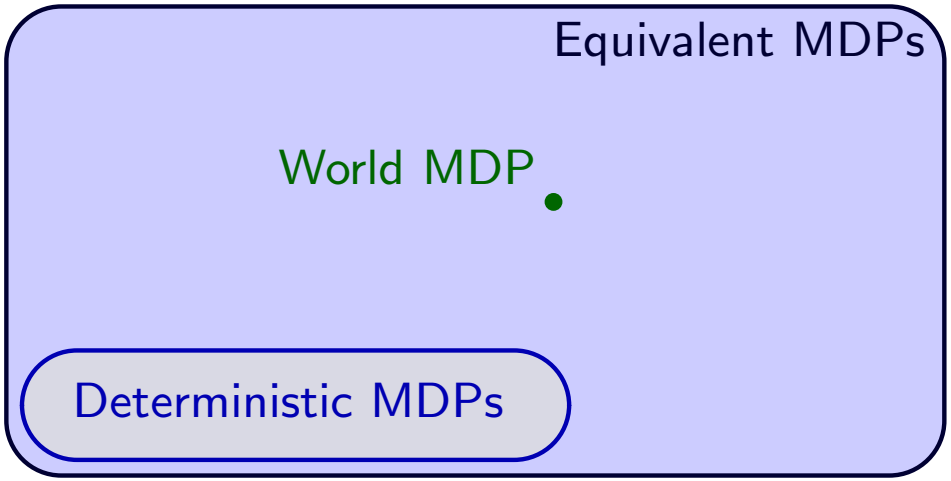
Belief States & MPC



RLMPC theoretical pathway

- Establish the set of equivalent MDPs

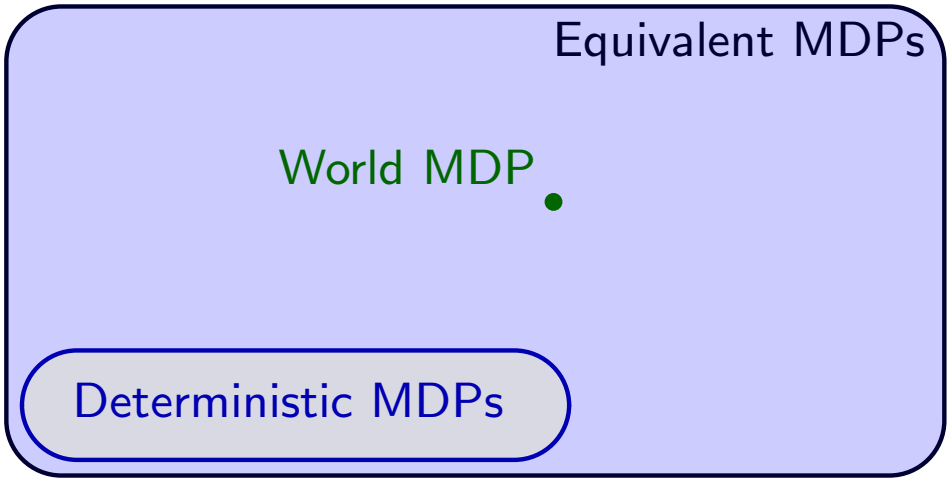
Belief States & MPC



RLMPC theoretical pathway

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

Belief States & MPC

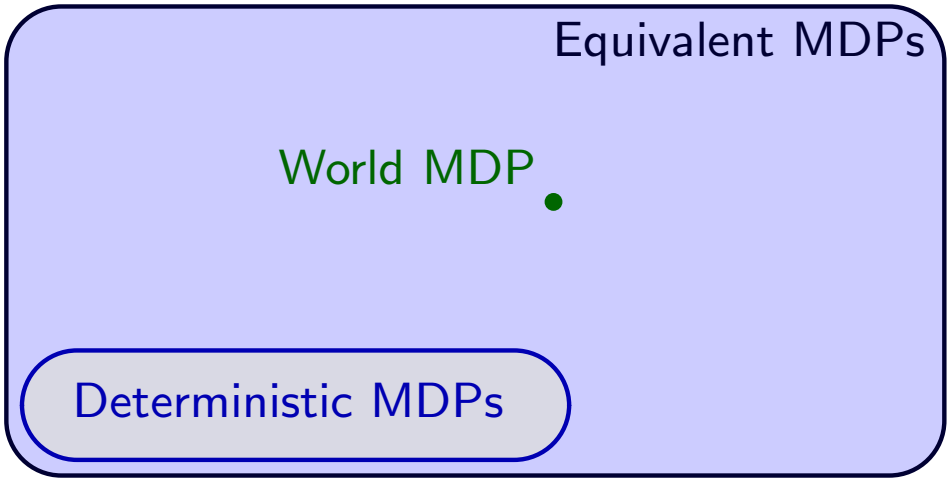


RLMPC theoretical pathway

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

That subset gives your MPC structure

Belief States & MPC



RLMPC theoretical pathway

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

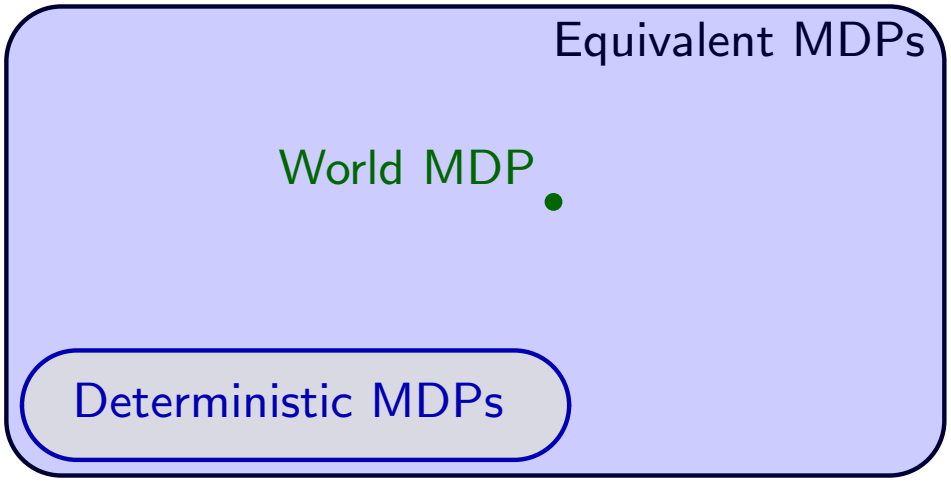
That subset gives your MPC structure

Remarks:

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

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

Belief States & MPC



RLMPC theoretical pathway

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

That subset gives your MPC structure

Remarks:

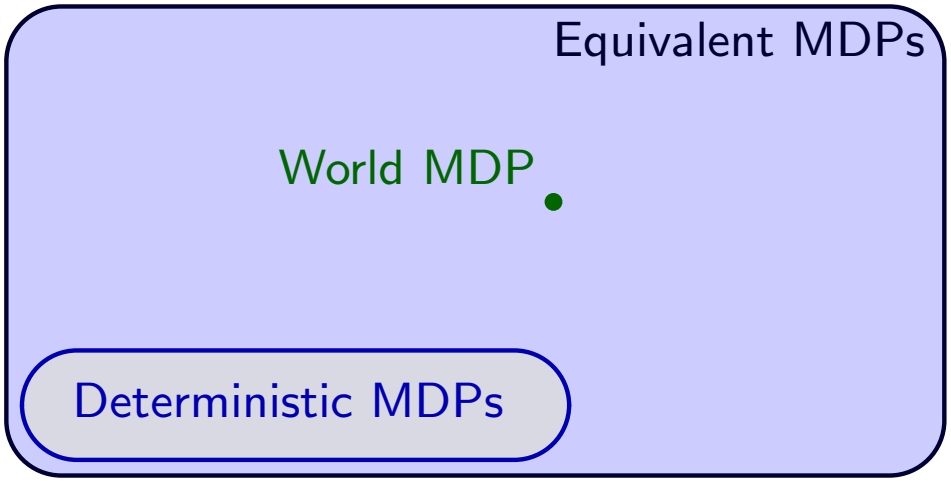
- Belief state MDPs have a set of equivalent MDPs (same structure as regular MDPs)
- "Degrade" the stochasticity of

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

Deterministic model e.g.

- Work in expected values:
  $\hat{\varphi}_{k+1} = \mathbb{E}[\mathcal{B}(\hat{\varphi}_k, \mathbf{o}_{k+1}, \mathbf{a}_k)] = \mathcal{T}_{\mathbf{a}_k}\hat{\varphi}_k$?
- Expected observation:
  $\hat{\varphi}_{k+1} = \mathcal{B}(\hat{\varphi}_k, \mathbb{E}[\mathbf{o}_{k+1}], \mathbf{a}_k)$?
- ML of observation:
  $\hat{\varphi}_{k+1} = \mathcal{B}(\hat{\varphi}_k, \mathrm{ML}[\mathbf{o}_{k+1}], \mathbf{a}_k)$?
- Generic model $\hat{\mathbf{o}}_{k+1} = \mathcal{M}(\hat{\varphi}_k)$?

Belief States & MPC



RLMPC theoretical pathway

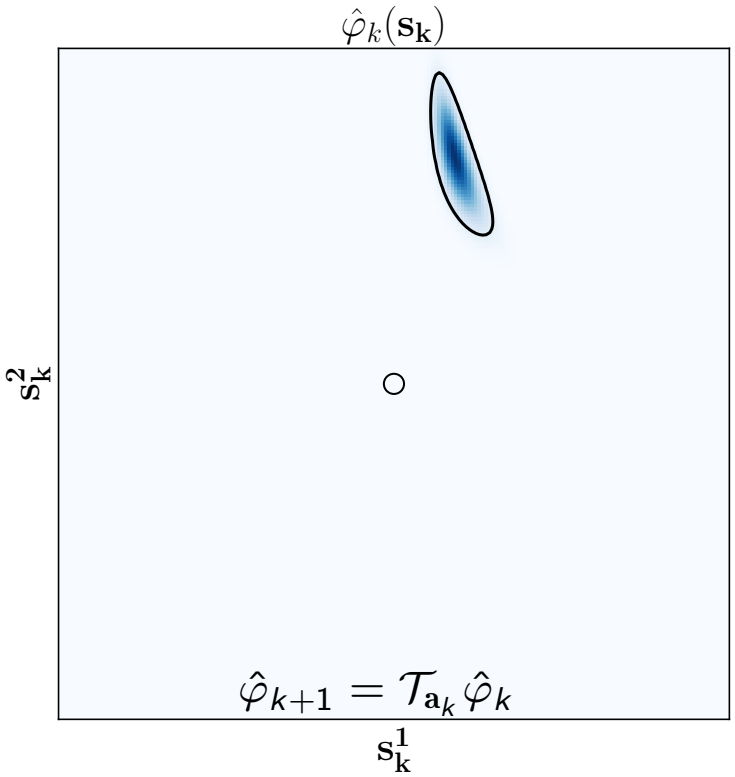
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

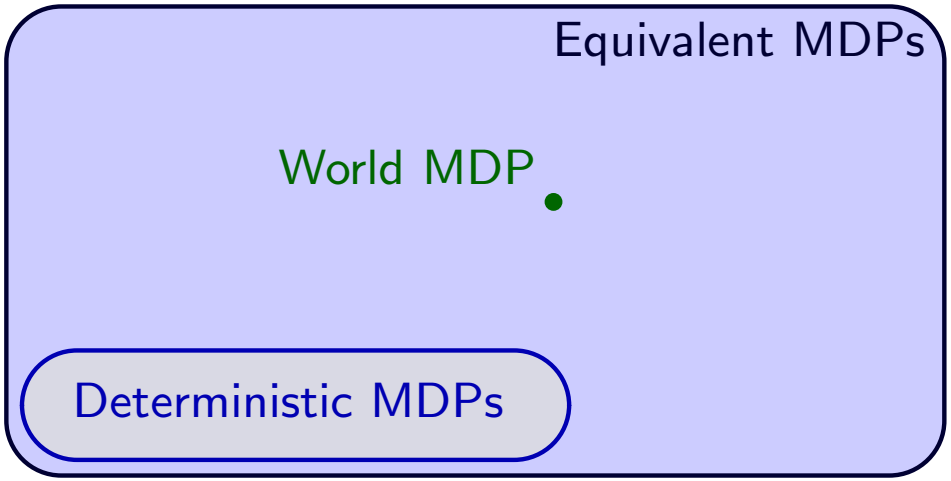
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

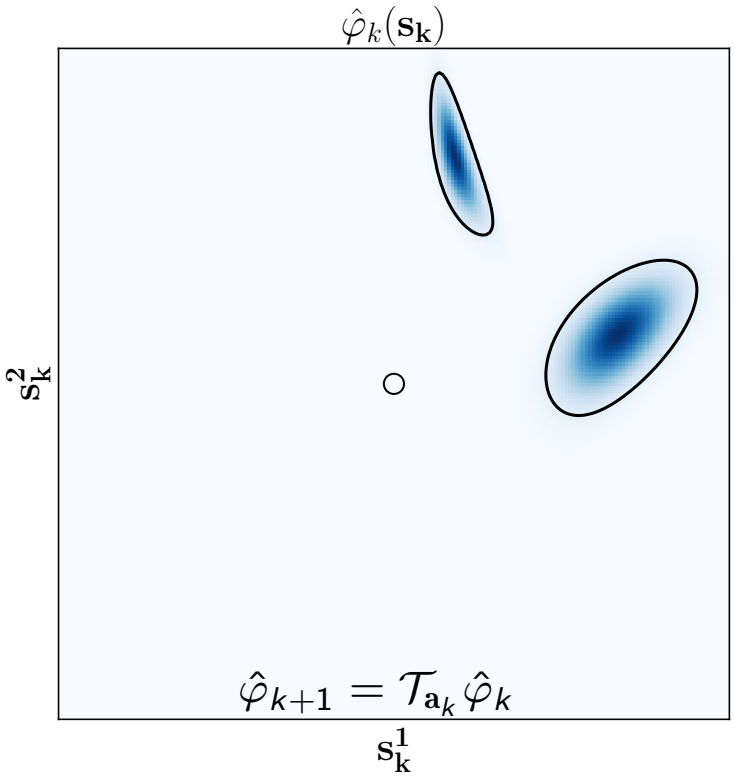
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

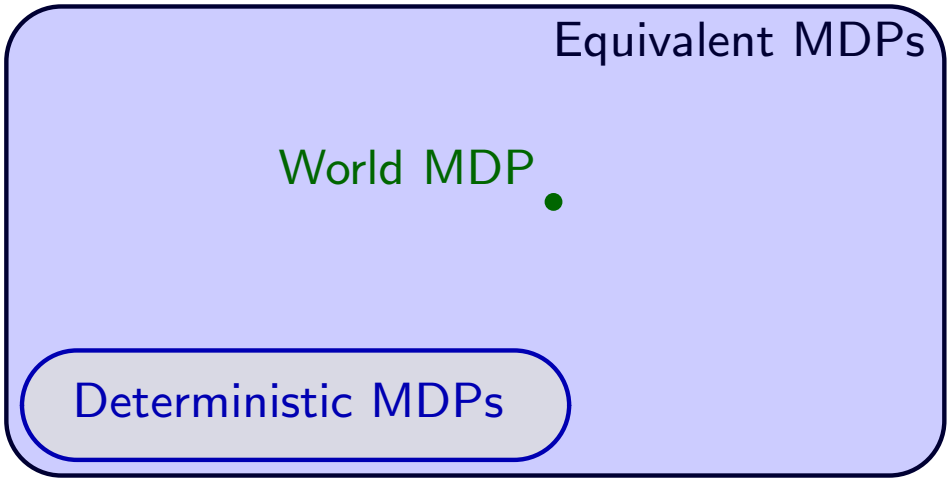
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

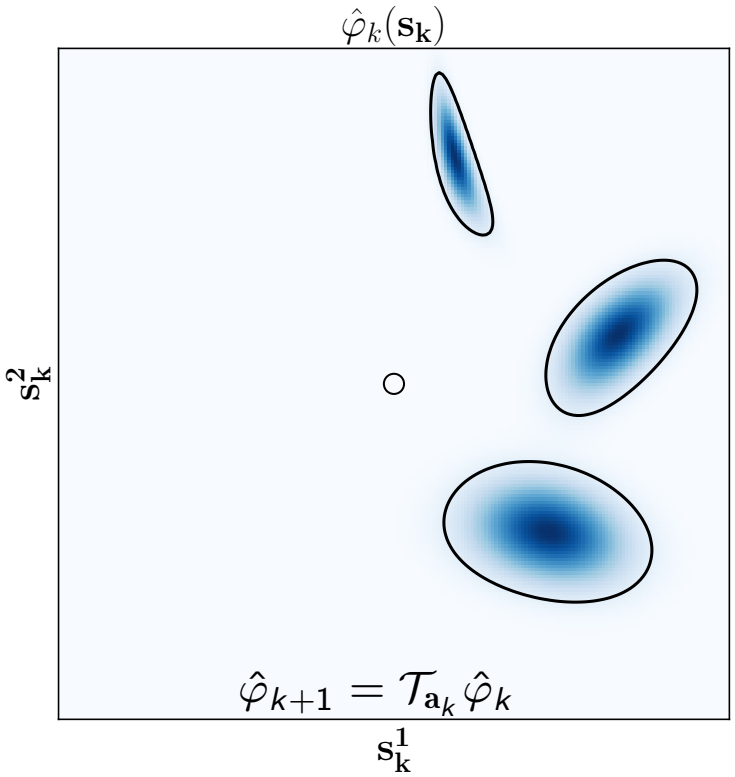
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

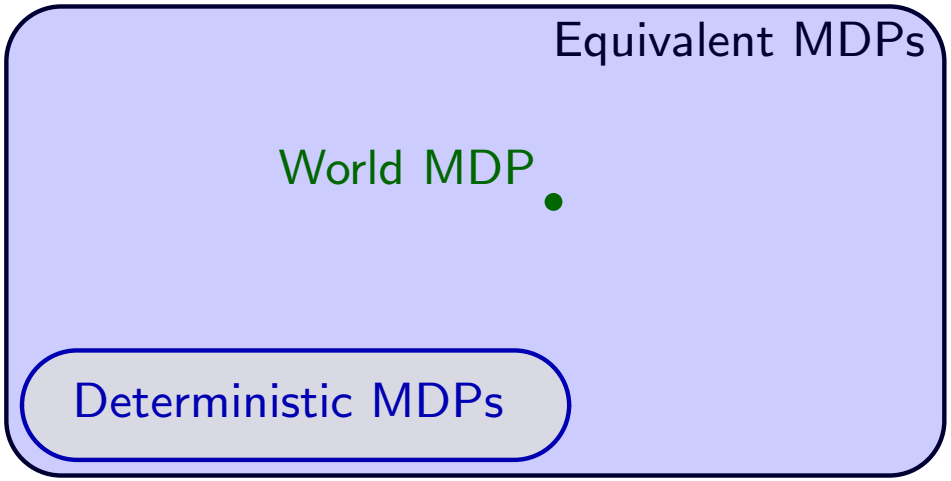
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

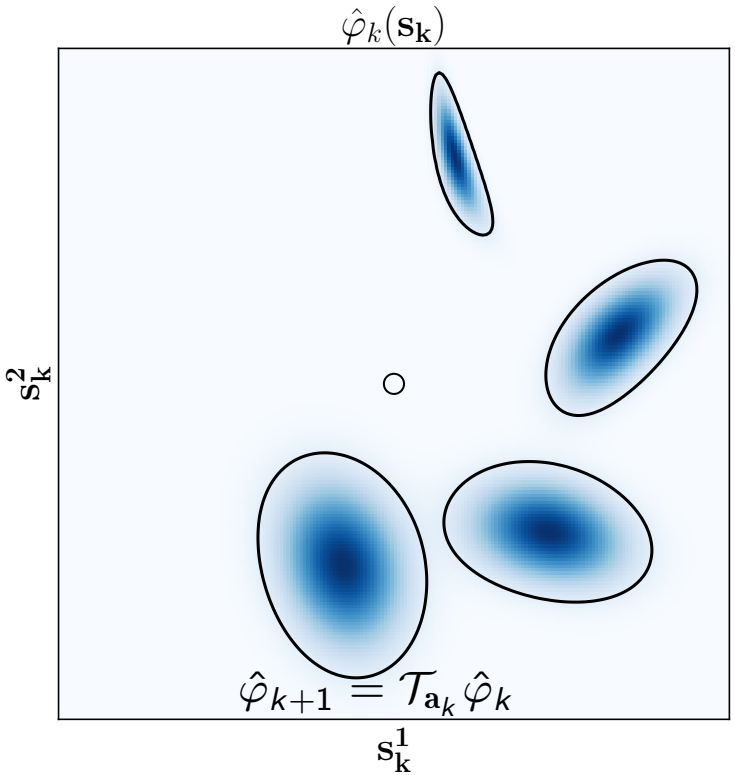
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

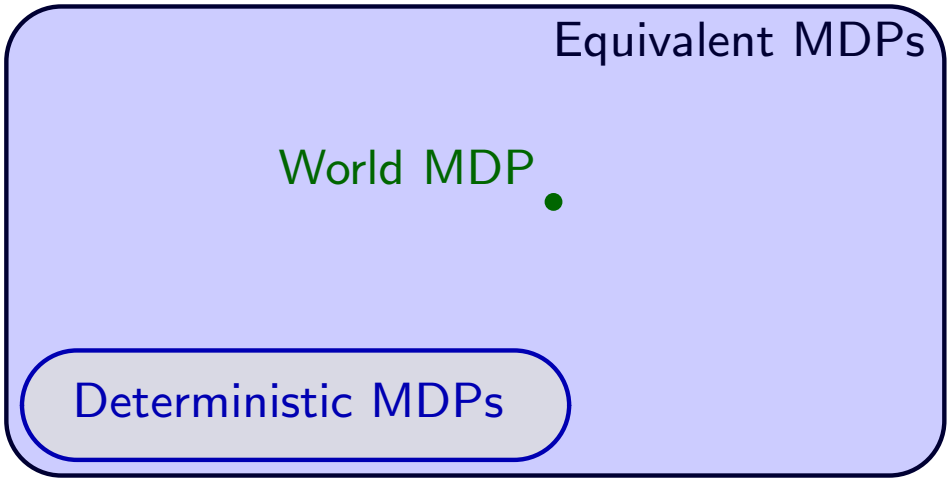
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

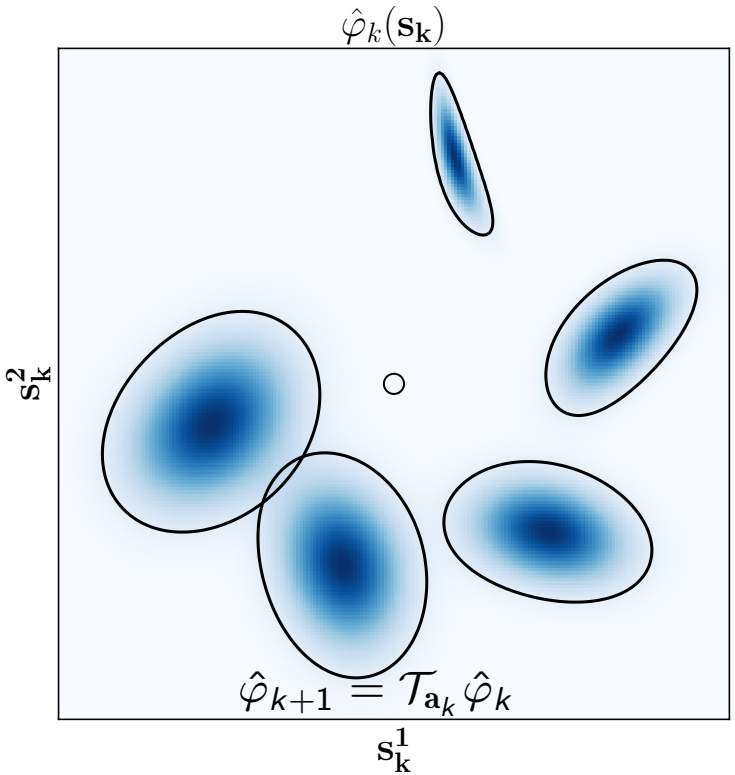
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

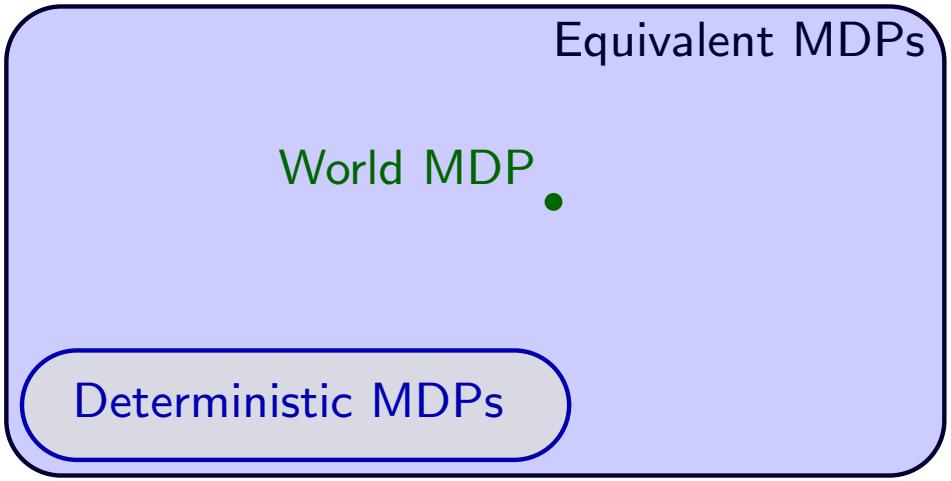
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

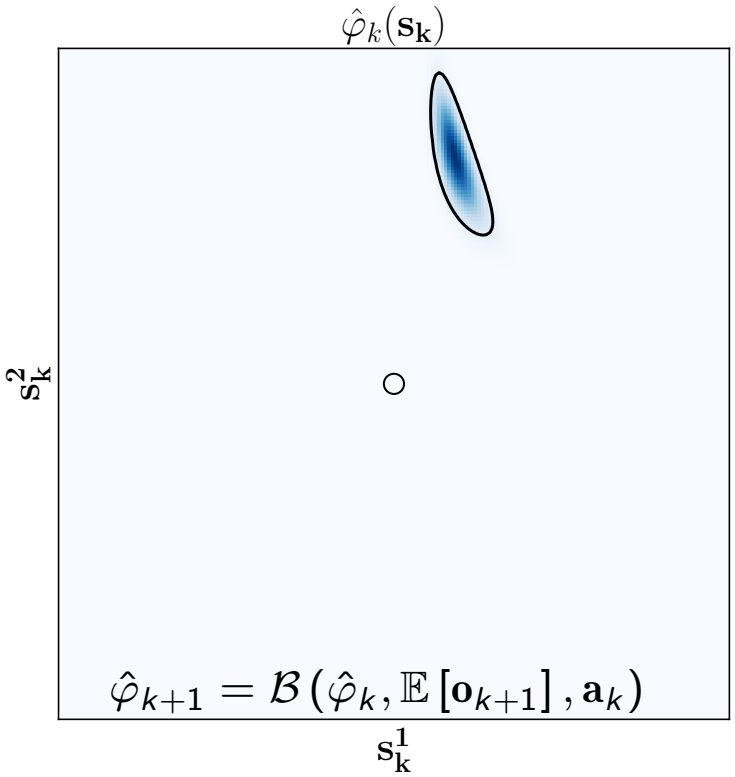
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

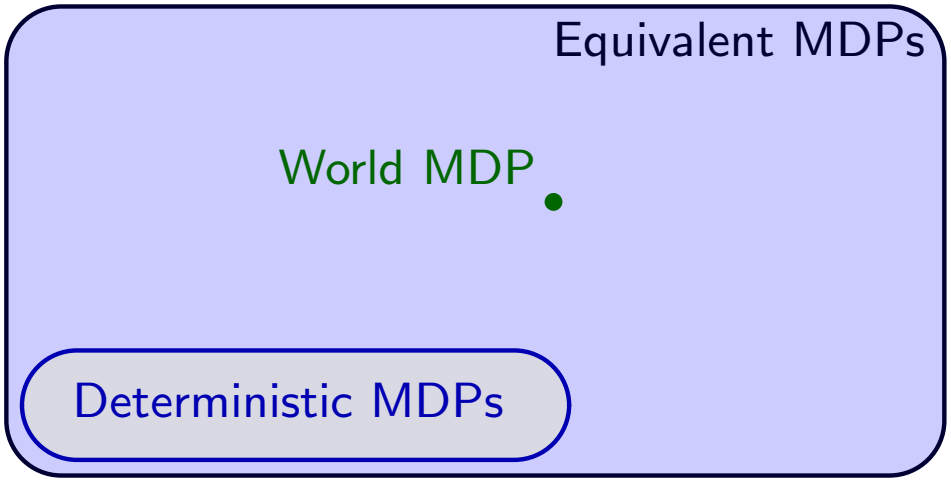
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

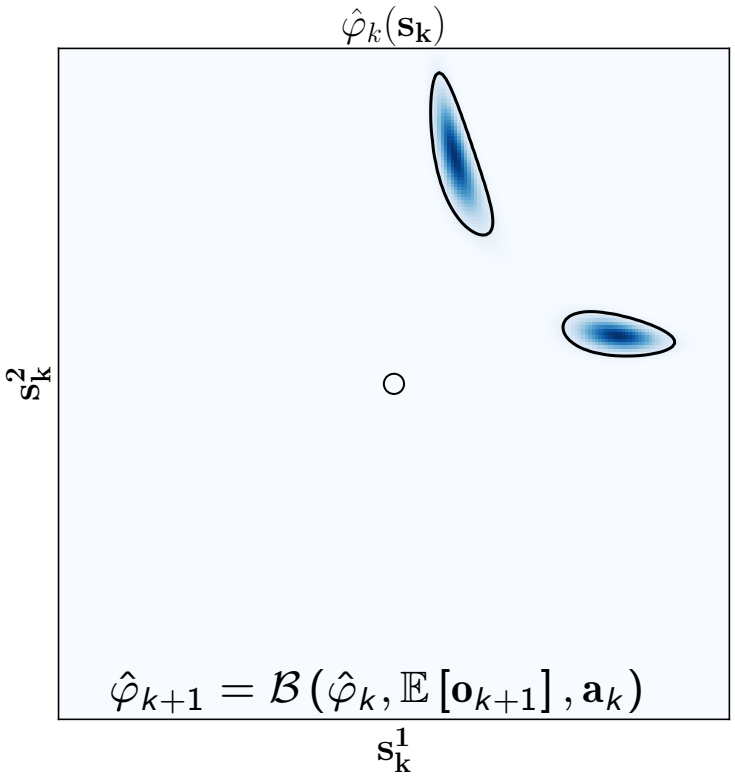
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

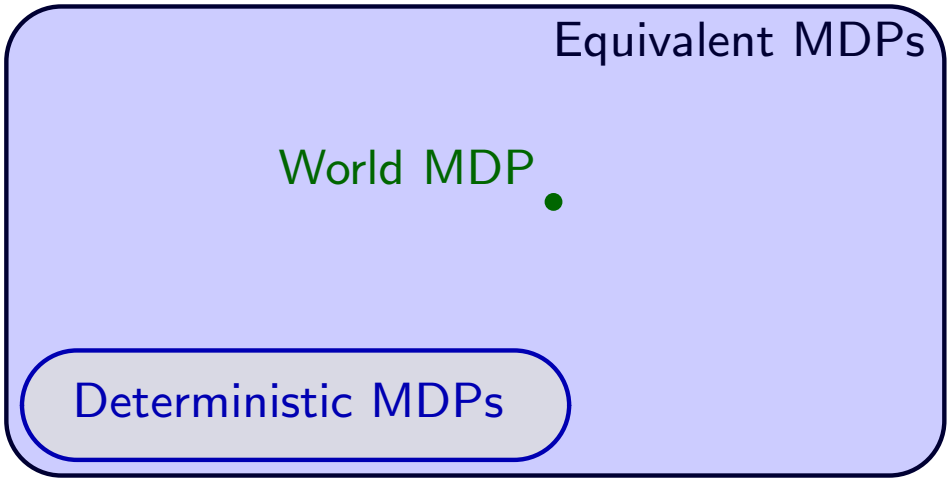
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

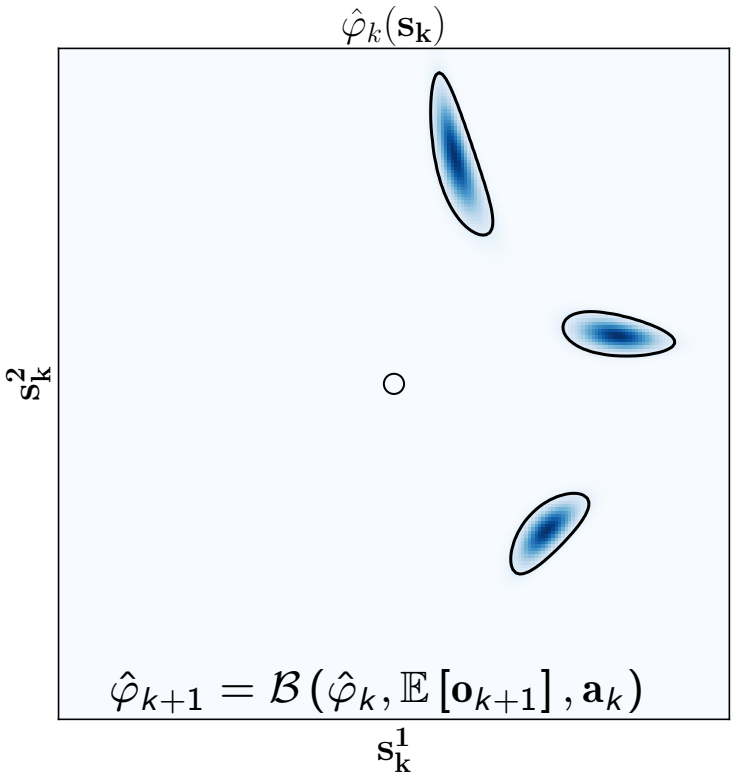
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

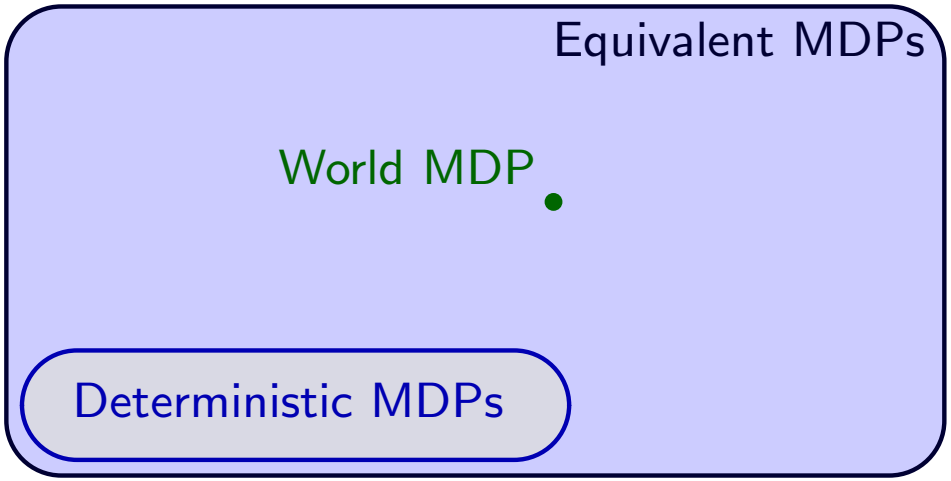
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

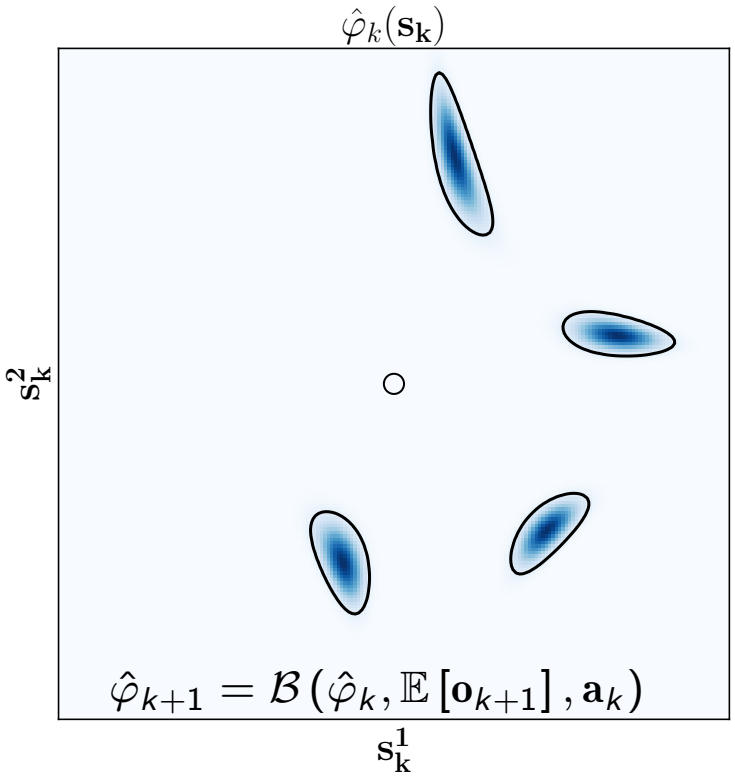
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

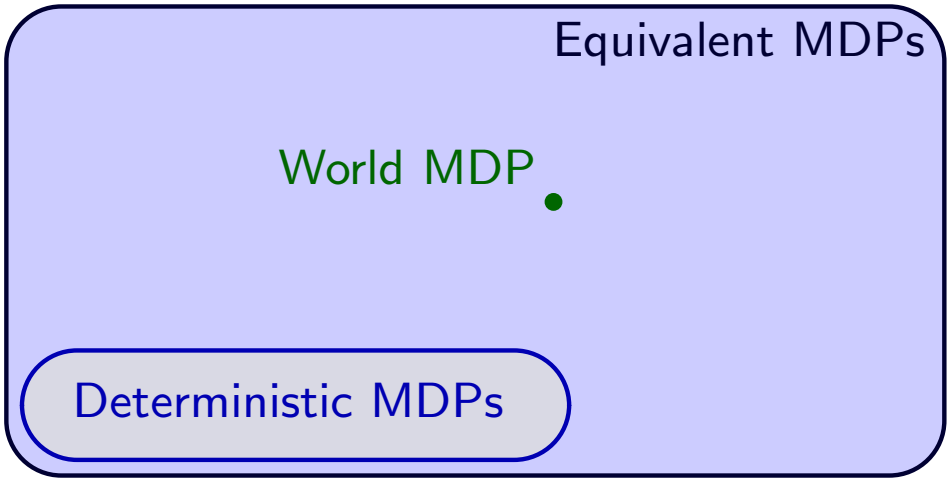
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

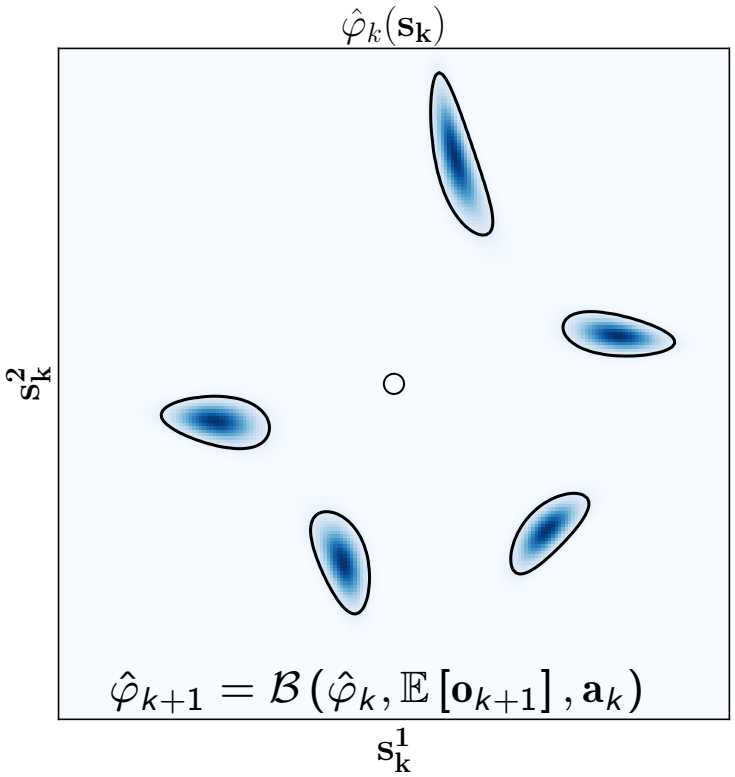
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

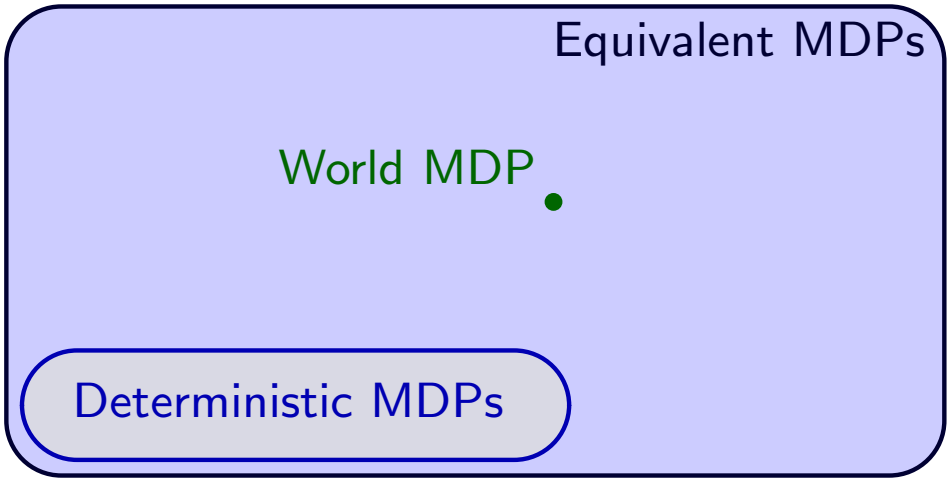
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

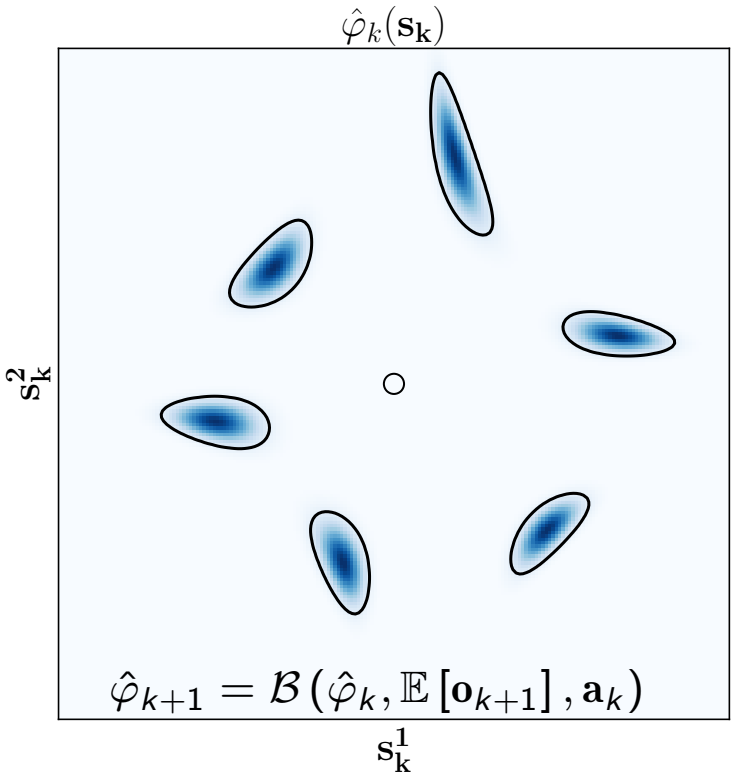
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

That subset gives your MPC structure

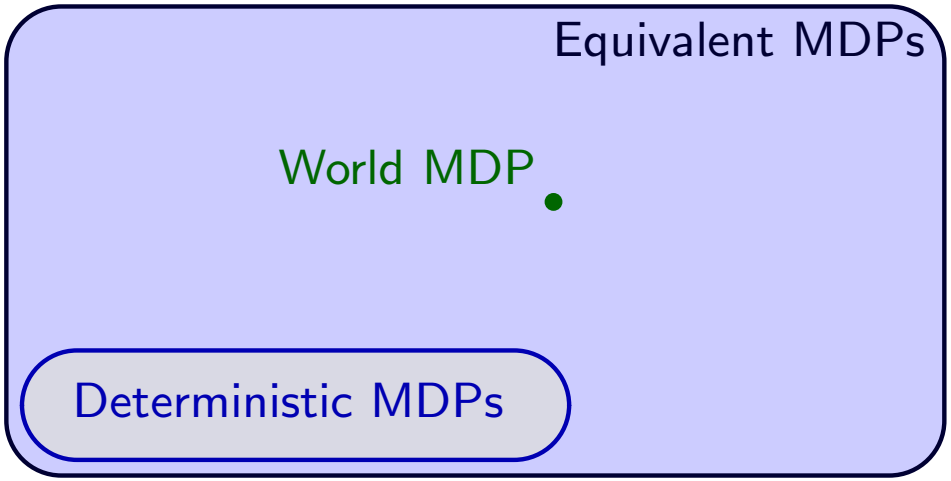
Remarks:

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

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



Belief States & MPC



RLMPC theoretical pathway

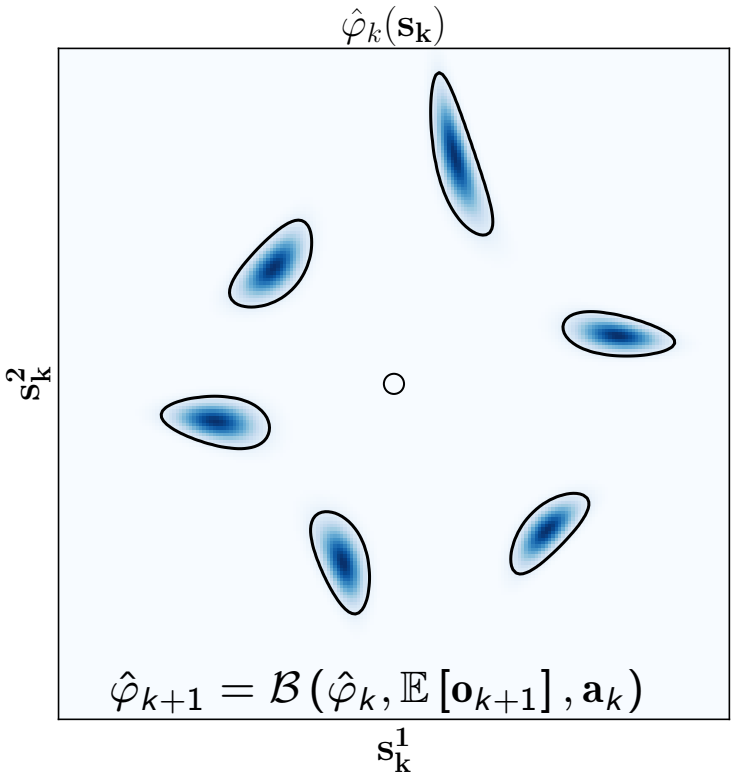
- Establish the set of equivalent MDPs
- Find a deterministic subset in that set

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

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



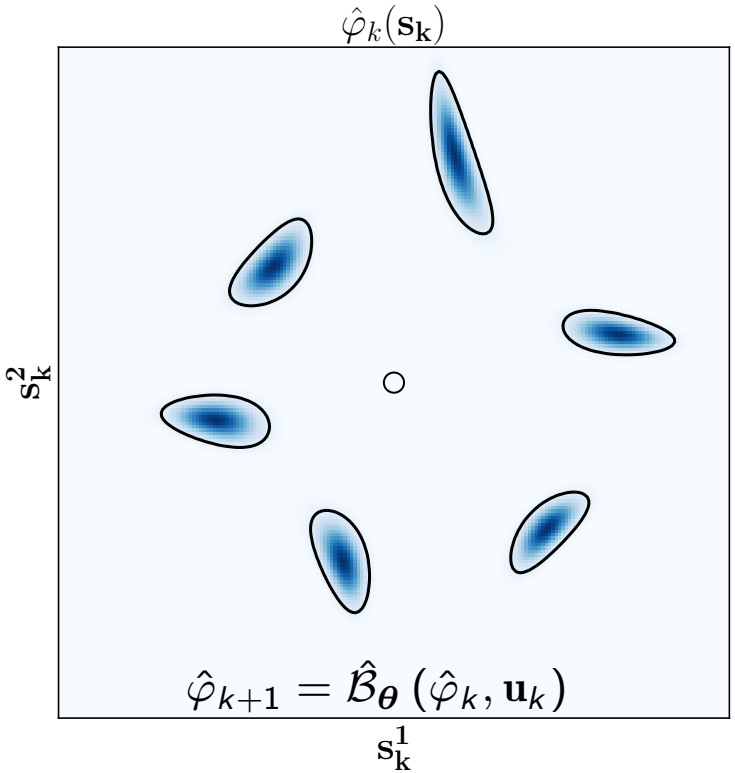
What is the best choice? Is there a good choice? *Open question*

MPC over Belief State MDPs

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 over Belief State MDPs

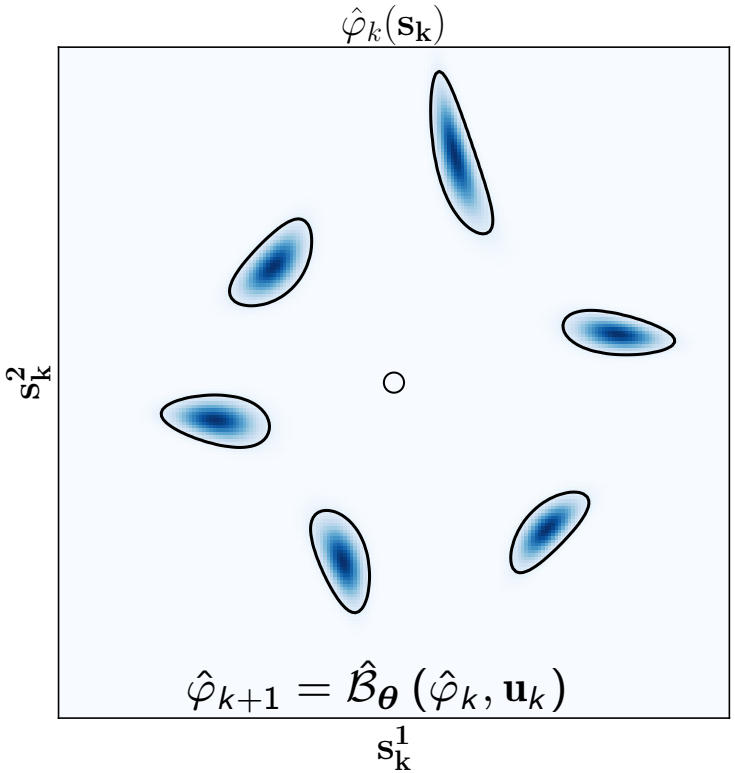
MPC model

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

"resolving" stochasticity over $\mathbf{o}_k$

MPC scheme gives $\boldsymbol{\pi}^{\mathrm{MPC}}(\varphi) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\varphi, \mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\hat{\varphi}_N\right) + \sum_{k=0}^{N-1} \ell_{\boldsymbol{\theta}}\left(\hat{\varphi}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\boldsymbol{\theta}}\left(\hat{\varphi}_k, \mathbf{u}_k\right), \quad \hat{\varphi}_0 = \varphi \end{aligned}$$



MPC over Belief State MDPs

MPC model

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

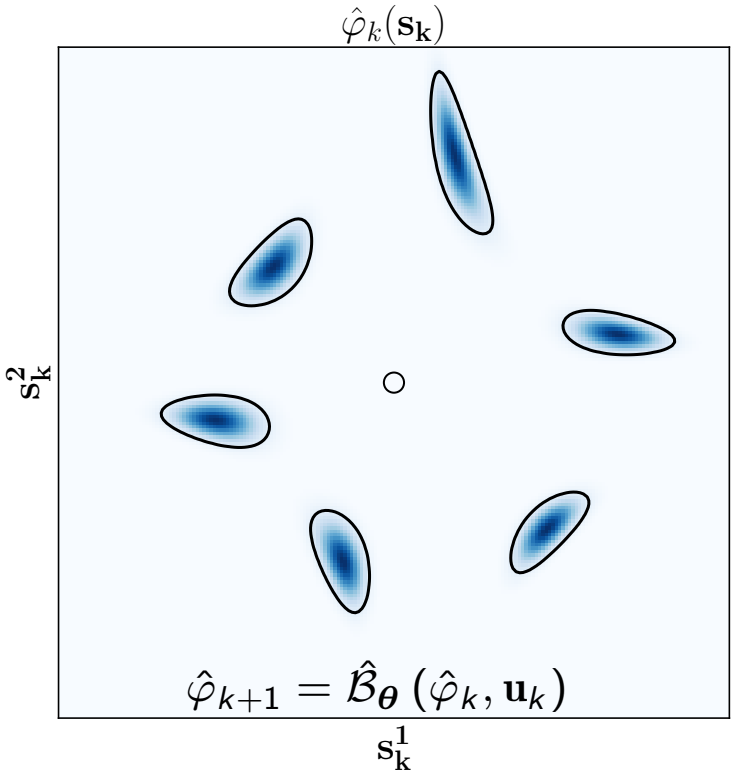
"resolving" stochasticity over $\mathbf{o}_k$

MPC scheme gives $\boldsymbol{\pi}^{\mathrm{MPC}}(\varphi) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\varphi, \mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\hat{\varphi}_N\right) + \sum_{k=0}^{N-1} \ell_{\boldsymbol{\theta}}\left(\hat{\varphi}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\boldsymbol{\theta}}\left(\hat{\varphi}_k, \mathbf{u}_k\right), \quad \hat{\varphi}_0 = \varphi \end{aligned}$$

Remark: optimal $\ell_{\boldsymbol{\theta}}$ is nonlinear in $\hat{\varphi}_k$ i.e.

$$\ell_{\boldsymbol{\theta}}\left(\hat{\varphi}_k, \mathbf{u}_k\right) \neq \mathbb{E}_{\mathbf{s}\sim\hat{\varphi}_k}\left[L_{\boldsymbol{\theta}}\left(\mathbf{s}, \mathbf{u}_k\right)\right]$$



MPC over Belief State MDPs

Remark: optimal $\ell_{\boldsymbol{\theta}}$ is nonlinear in $\hat{\varphi}_k$ i.e.

$$\ell_{\boldsymbol{\theta}}\left(\hat{\varphi}_k, \mathbf{u}_k\right) \neq \mathbb{E}_{\mathbf{s}\sim\hat{\varphi}_k}\left[L_{\boldsymbol{\theta}}\left(\mathbf{s}, \mathbf{u}_k\right)\right]$$

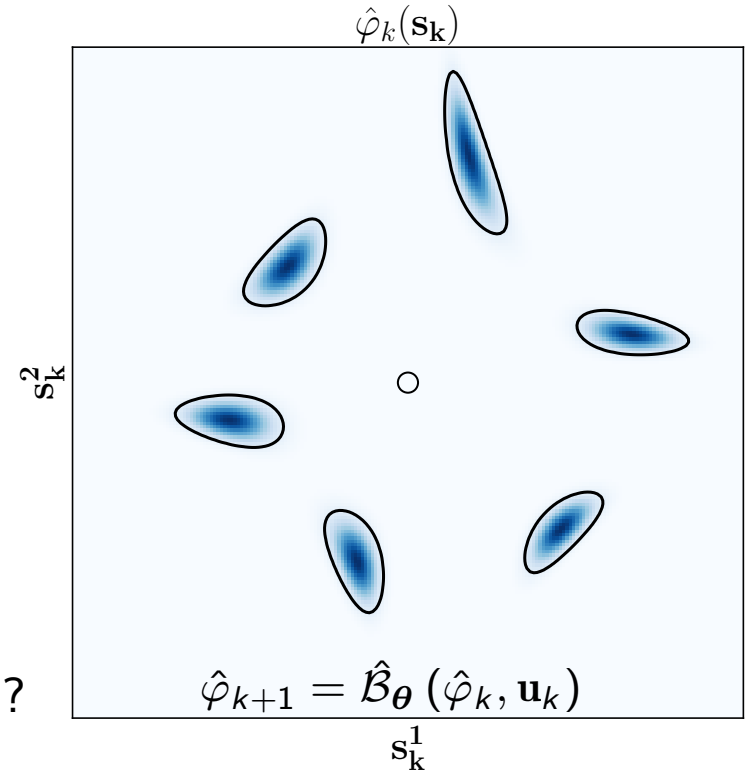
MPC model

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

"resolving" stochasticity over $\mathbf{o}_k$

MPC scheme gives $\boldsymbol{\pi}^{\mathrm{MPC}}\left(\varphi\right) = \mathbf{u}_0^\star$ from

$$\begin{aligned} \min_{\varphi,\mathbf{u}} \quad & T_{\boldsymbol{\theta}}\left(\hat{\varphi}_N\right) + \sum_{k=0}^{N-1} \ell_{\boldsymbol{\theta}}\left(\hat{\varphi}_k, \mathbf{u}_k\right) \\ \text{s.t.} \quad & \hat{\varphi}_{k+1} = \hat{\mathcal{B}}_{\boldsymbol{\theta}}\left(\hat{\varphi}_k, \mathbf{u}_k\right), \quad \hat{\varphi}_0 = \varphi \end{aligned}$$



Questions

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

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: RL over MPC

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

Thanks for your attention!

ResearchGate

Google Scholar