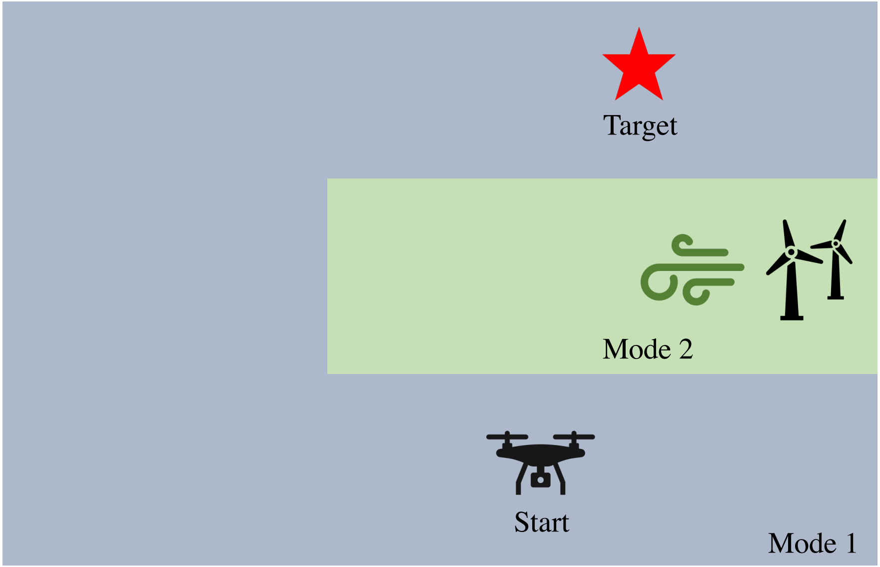
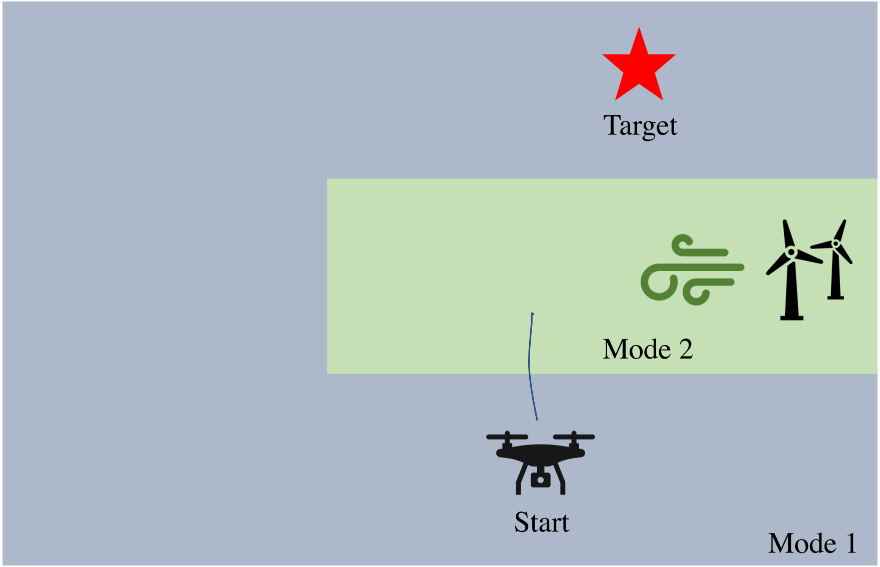
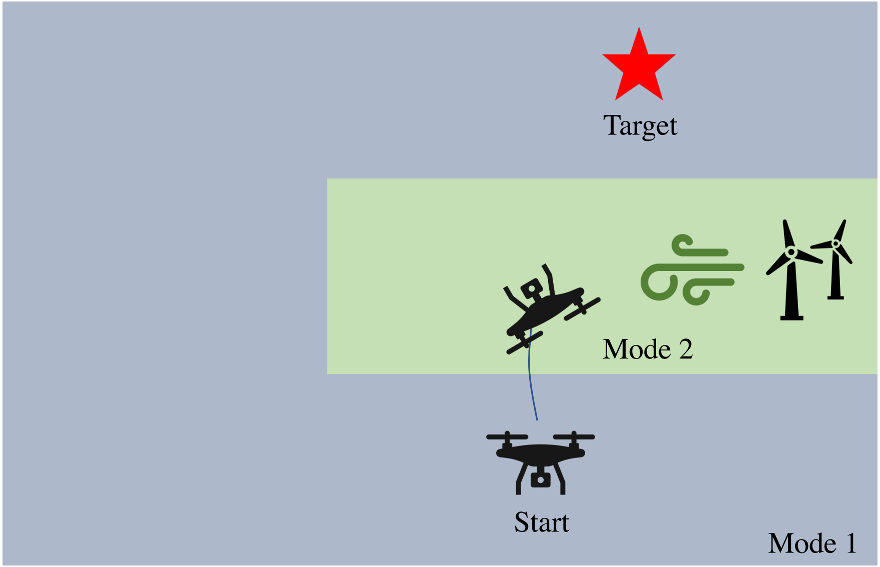
PhD Thesis: Bayesian Learning for Control in Multimodal Dynamical Systems

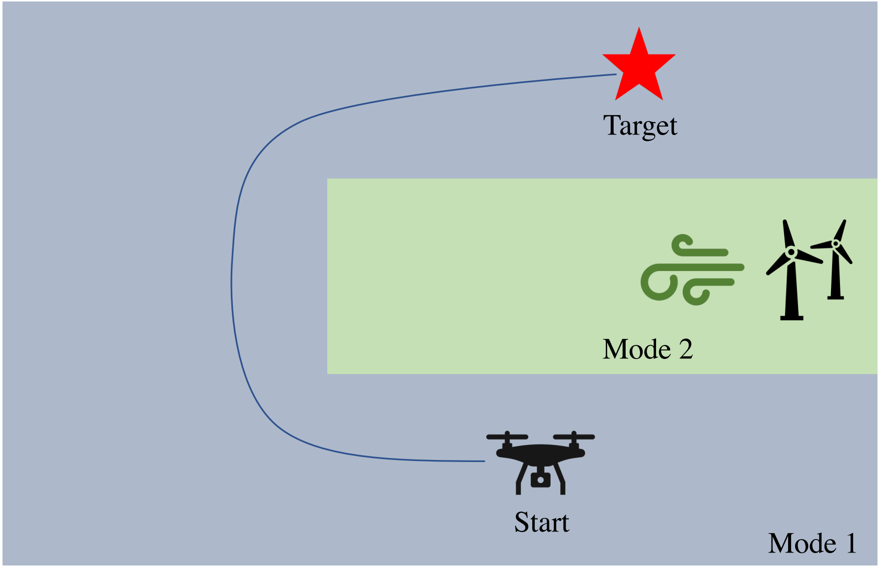
Aidan Scannell | Carl Henrik Ek | Arthur Richards

8th September 2022


Target
Mode 2
Start
Mode 1


Target
Mode 2
Start
Mode 1


Target
Mode 2
Start
Mode 1


Target
Mode 2
Start
Mode 1

Goals

Goal 1 Navigate to the target state x_f

Goal 2 Remain in the operable, desired dynamics mode k^*

Mode remaining navigation problem

$$
\begin{aligned}
\min_{\pi \in \Pi} \quad & \sum_{t=0}^{T} c(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) && \text{(1a)} \\
\text{s.t.} \quad & && \text{(1b)} \\
& && \text{(1c)} \\
& && \text{(1d)} \\
& && \text{(1e)}
\end{aligned}
$$

Mode remaining navigation problem

$$\min_{\pi \in \Pi} \quad \sum_{t=0}^{T} c(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) \tag{1a}$$

$$\text{s.t.} \tag{1b}$$

$$\tag{1c}$$

$$\mathrm{x}_0 = \mathrm{x}_0 \tag{1d}$$

$$\tag{1e}$$

Mode remaining navigation problem

$$\min_{\pi \in \Pi} \quad \sum_{t=0}^{T} c(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) \tag{1a}$$

$$\text{s.t.} \tag{1b}$$

$$\tag{1c}$$

$$\mathrm{x}_0 = \mathrm{x}_0 \tag{1d}$$

$$\mathrm{x}_T = \mathrm{x}_f \tag{1e}$$

Mode remaining navigation problem

$$
\begin{aligned}
\min_{\pi \in \Pi} \quad & \sum_{t=0}^{T} c(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) && \text{(1a)} \\
\text{s.t.} \quad & \mathrm{x}_{t+1} = f_k(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) + \epsilon_k, \quad \text{if } \alpha(\mathrm{x}_t) = k \qquad \forall t \in \{0, \dots, T-1\} && \text{(1b)} \\
& && \text{(1c)} \\
& \mathrm{x}_0 = x_0 && \text{(1d)} \\
& \mathrm{x}_T = x_f && \text{(1e)}
\end{aligned}
$$

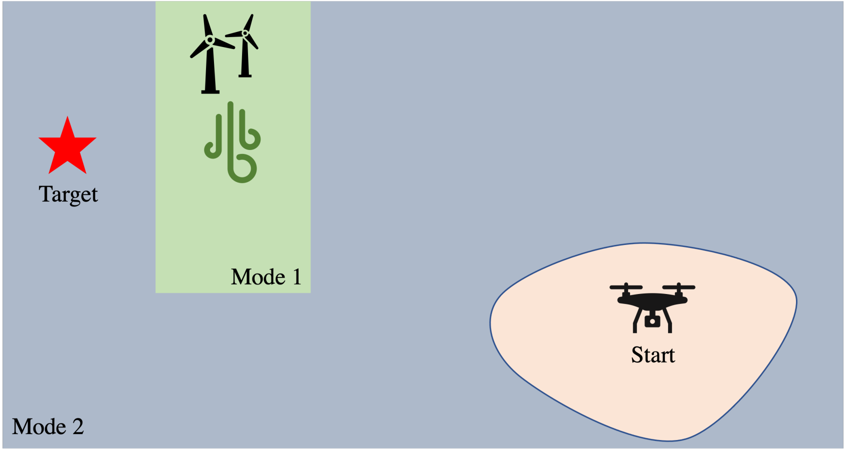
Mode remaining navigation problem

$$
\begin{aligned}
\min_{\pi \in \Pi} \quad & \sum_{t=0}^{T} c(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) && && \text{(1a)} \\
\text{s.t.} \quad & \mathrm{x}_{t+1} = f_k(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) + \epsilon_k, \quad \text{if } \alpha(\mathrm{x}_t) = k && \forall t \in \{0, \ldots, T-1\} && \text{(1b)} \\
& \alpha(\mathrm{x}_t) = k^* && \forall t \in \{0, \ldots, T-1\} && \text{(1c)} \\
& \mathrm{x}_0 = \mathrm{x}_0 && && \text{(1d)} \\
& \mathrm{x}_T = \mathrm{x}_f && && \text{(1e)}
\end{aligned}
$$

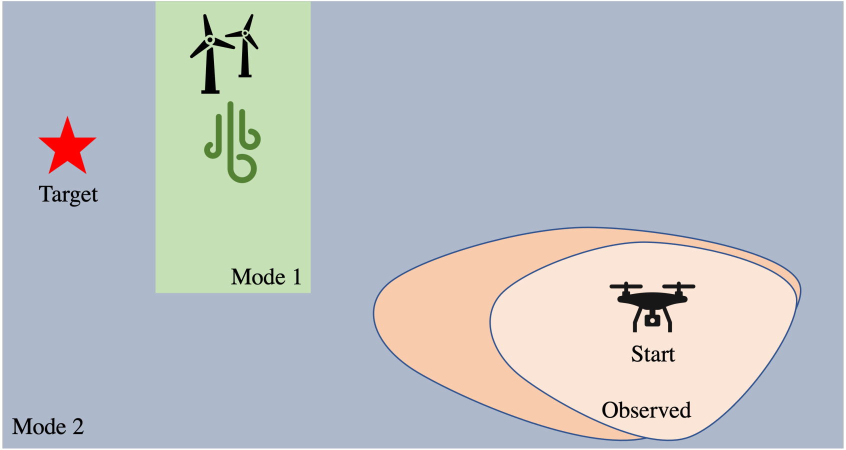
But dynamics are not known a priori...

$$
\begin{aligned}
\min_{\pi \in \Pi} \quad & \sum_{t=0}^{T} c(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) && && \text{(2a)} \\
\text{s.t.} \quad & \mathrm{x}_{t+1} = f_k(\mathrm{x}_t, \pi(\mathrm{x}_t, t)) + \epsilon_k, \quad \text{if } \alpha(\mathrm{x}_t) = k && \forall t \in \{0, \dots, T-1\} && \text{(2b)} \\
& \alpha(\mathrm{x}_t) = k^* && \forall t \in \{0, \dots, T-1\} && \text{(2c)} \\
& \mathrm{x}_0 = \mathrm{x}_0 && && \text{(2d)} \\
& \mathrm{x}_T = \mathrm{x}_f && && \text{(2e)}
\end{aligned}
$$

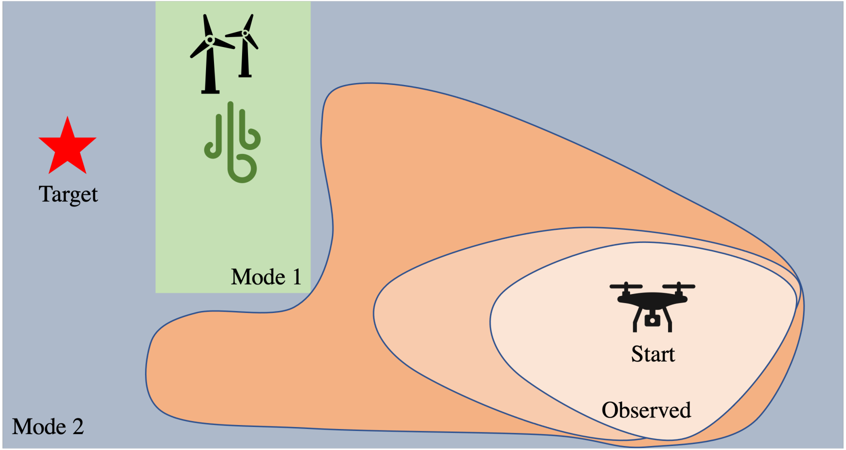
Solve using model-based reinforcement learning?



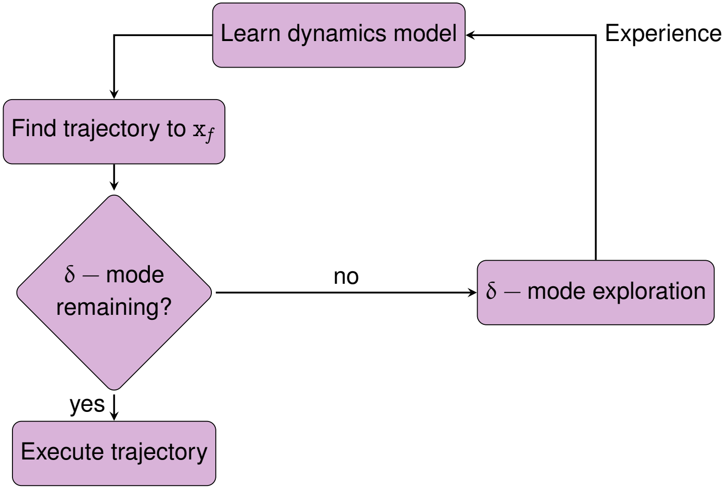
Solve using model-based reinforcement learning?



Solve using model-based reinforcement learning?



Solve using model-based reinforcement learning?



Contributions

1. Model learning

Contributions

1. Model learning
2. Mode remaining trajectory optimisation

Contributions

1. Model learning
2. Mode remaining trajectory optimisation
   - via latent geometry

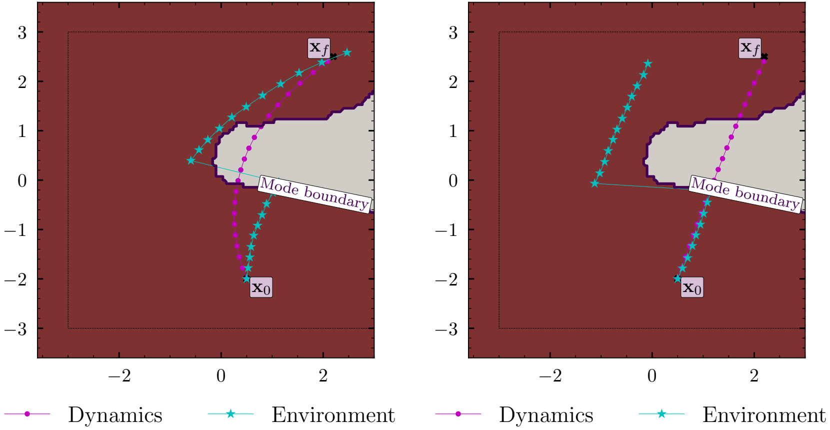
Contributions

1. Model learning
2. Mode remaining trajectory optimisation
   - via latent geometry
   - control as inference

Contributions

1. Model learning
2. Mode remaining trajectory optimisation
   - via latent geometry
   - control as inference
3. Mode remaining exploration for model-based reinforcement learning

Model learning - Gaussian processes don't work...

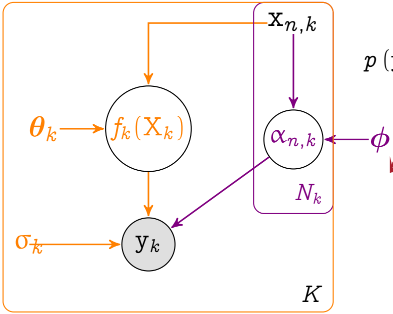


Model learning - mixture models?

MoE marginal likelihood

$$p\left(\mathrm{y} \mid \mathrm{X}, \boldsymbol{\theta}, \phi\right) = \prod_{n=1}^{N} \sum_{k=1}^{K} \underbrace{\mathsf{Pr}\left(\alpha_n = k \mid \mathrm{x}_n, \phi\right)}_{\text{gating network}} \underbrace{p\left(y_n \mid \alpha_n = k, \mathrm{x}_n, \boldsymbol{\theta}_k\right)}_{\text{expert } k}, \tag{3}$$

Model learning - mixtures of nonparametric experts



$$p(\mathrm{y} \mid \mathrm{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{\boldsymbol{\alpha}} p(\boldsymbol{\alpha} \mid \mathrm{X}, \boldsymbol{\phi}) \left[\prod_{k=1}^{K} \underbrace{p(\{y_n : \alpha_n = k\} \mid \{\mathrm{x}_n : \alpha_n = k\}, \boldsymbol{\theta}_k)}_{\text{expert } k} \right]$$

- Sum over exponentially many (K^N) sets of assignments
  - $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_N\}$

Model learning - Parameterise the nonparametric model?

- Like a sparse GP parameterises a GP...

Model learning - Parameterise the nonparametric model?

- Like a sparse GP parameterises a GP...
- GP prior where $\mathrm{X}_k = \{\mathrm{x}_n : \alpha_t = k\}$

$$f_k(\mathrm{X}_k) \sim \mathcal{N}\left(\mu_k(\mathrm{X}_k), k_k(\mathrm{X}_k, \mathrm{X}_k)\right)$$

Model learning - Parameterise the nonparametric model?

- Like a sparse GP parameterises a GP...
- GP prior where $\mathrm{X}_k = \{\mathrm{x}_n : \alpha_t = k\}$

$$f_k(\mathrm{X}_k) \sim \mathcal{N}(\mu_k(\mathrm{X}_k), k_k(\mathrm{X}_k, \mathrm{X}_k))$$

- Augment with inducing points

$$f_k(\zeta_k) \sim \mathcal{N}(\mu_k(\zeta_k), k_k(\zeta_k, \zeta_k))$$

Model learning - Parameterise the nonparametric model?

- Like a sparse GP parameterises a GP...
- GP prior where $\mathrm{X}_k = \{\mathrm{x}_n : \alpha_t = k\}$

$$f_k(\mathrm{X}_k) \sim \mathcal{N}\left(\mu_k(\mathrm{X}_k), k_k(\mathrm{X}_k, \mathrm{X}_k)\right)$$

- Augment with inducing points

$$f_k(\zeta_k) \sim \mathcal{N}\left(\mu_k(\zeta_k), k_k(\zeta_k, \zeta_k)\right)$$

- FITC for MoGPE?

$$p(\mathrm{y} \mid \mathrm{f}(\zeta)) \approx \prod_{n=1}^{N} p(y_n \mid \mathrm{f}(\zeta)) = \prod_{n=1}^{N} \sum_{k=1}^{K} \Pr\left(\alpha_n = k \mid \mathrm{x}_n, \phi\right) \prod_{k=1}^{K} p(y_n \mid f_k(\zeta_k)).$$

Model learning - Parameterise the nonparametric model?

- Like a sparse GP parameterises a GP...
- GP prior where $\mathrm{X}_k = \{\mathrm{x}_n : \alpha_t = k\}$

$$f_k(\mathrm{X}_k) \sim \mathcal{N}\left(\mu_k(\mathrm{X}_k), k_k(\mathrm{X}_k, \mathrm{X}_k)\right)$$

- Augment with inducing points

$$f_k(\zeta_k) \sim \mathcal{N}\left(\mu_k(\zeta_k), k_k(\zeta_k, \zeta_k)\right)$$

- FITC for MoGPE?

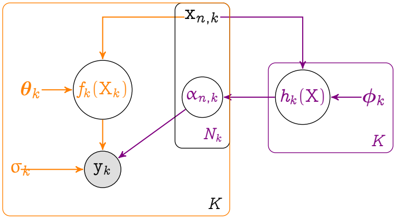
$$p(\mathrm{y} \mid \mathrm{f}(\zeta)) \approx \prod_{n=1}^{N} p(y_n \mid \mathrm{f}(\zeta)) = \prod_{n=1}^{N} \sum_{k=1}^{K} \mathrm{Pr}\left(\alpha_n = k \mid \mathrm{x}_n, \phi\right) \prod_{k=1}^{K} p(y_n \mid f_k(\zeta_k)).$$

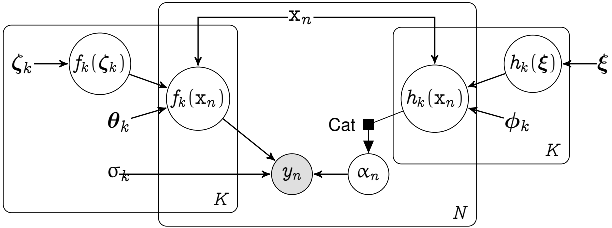
- Assumes inducing variables $\{f_k(\zeta_k)\}_{k=1}^{K}$, are a sufficient statistic for latent function values $\{f_k(\mathrm{X}_k)\}_{k=1}^{K}$ AND the set of assignments $\boldsymbol{\alpha}$.

Model learning - Parameterise the nonparametric model?

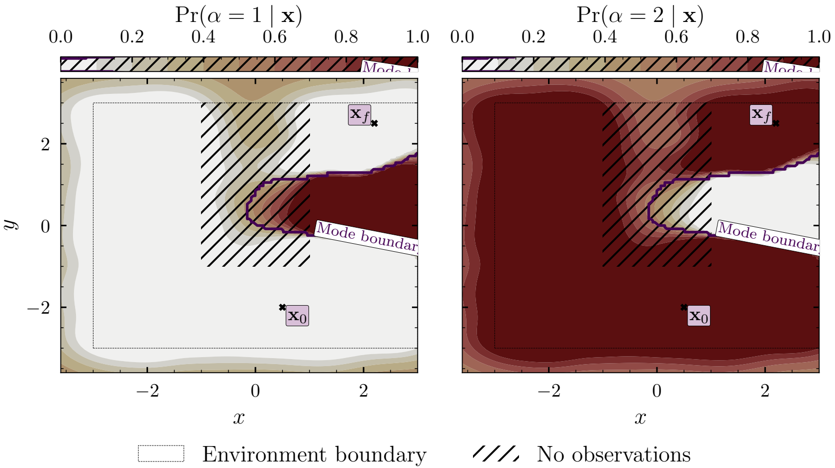
- Approximate marinal likelihood

$$p\,(\mathrm{y} \mid \mathrm{X}) \approx \mathbb{E}_{p(\mathrm{h}(\boldsymbol{\xi}))p(\mathrm{f}(\boldsymbol{\zeta}))}\left[\prod_{n=1}^{N}\sum_{k=1}^{K} \Pr(\alpha_n = k \mid \mathrm{h}(\boldsymbol{\xi}))\, p(y_n \mid f_k(\boldsymbol{\zeta}_k))\right] \tag{4}$$

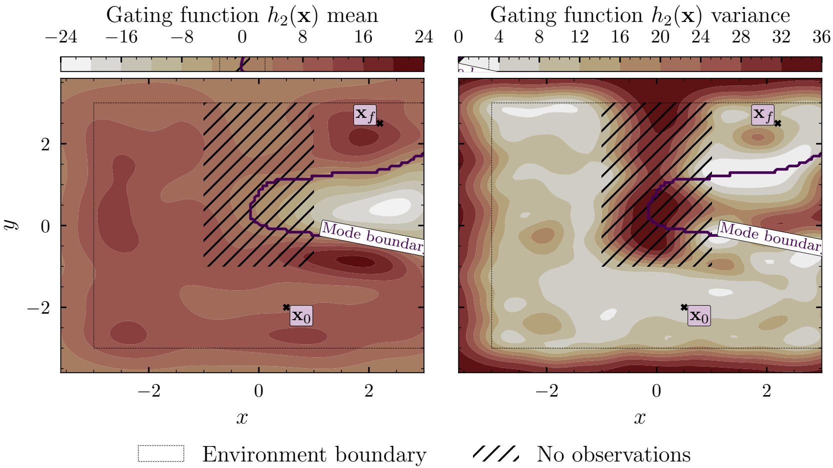




Model learning - latent spaces for planning



Model learning - latent spaces for planning



Mode remaining control

- Goals

Mode remaining control

- Goals
  - Navigate to the target state x_f

Mode remaining control

- Goals
  - Navigate to the target state x_f
  - Remain in desired dynamics mode

Mode remaining control

- Goals
  - Navigate to the target state x_f
  - Remain in desired dynamics mode
- Assumptions

Mode remaining control

- Goals
  - Navigate to the target state $\mathbf{x}_f$
  - Remain in desired dynamics mode
- Assumptions
  - Desired dynamics mode is *known a priori*

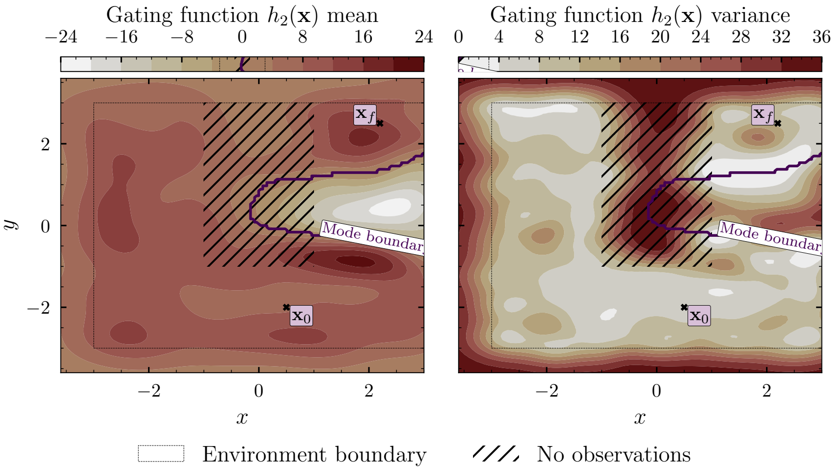
Mode remaining control

- Goals
  - Navigate to the target state $\mathbf{x}_f$
  - Remain in desired dynamics mode
- Assumptions
  - Desired dynamics mode is *known a priori*
  - Prior access to environment

Mode remaining control

- Goals
  - Navigate to the target state x_f
  - Remain in desired dynamics mode
- Assumptions
  - Desired dynamics mode is *known a priori*
  - Prior access to environment
    - Such that a state transition data set $\mathcal{D}$ has been collected

Mode remaining control - via latent geometry



Mode remaining control - via latent geometry

- Desired mode's gating function $h_{k^*} : \mathcal{X} \to \mathbb{R}$

Mode remaining control - via latent geometry

- Desired mode's gating function $h_{k^*} : \mathcal{X} \rightarrow \mathbb{R}$
- State trajectory $\bar{\mathrm{x}} : [t_0, t_f] \rightarrow \mathcal{X}$

Mode remaining control - via latent geometry

- Desired mode's gating function $h_{k^*} : \mathcal{X} \to \mathbb{R}$
- State trajectory $\bar{\mathrm{x}} : [t_0, t_f] \to \mathcal{X}$
- Length minimising trajectories encode mode remaining behaviour

$$\min \text{Length}(h_{k^*}(\bar{\mathrm{x}})) = \min \int_{t_0}^{t_f} \|\dot{\mathrm{x}}(t)\|_{\mathrm{G}(\mathrm{x}(t))} \, \mathrm{d}t \tag{5}$$

where,

$$\|\dot{\mathrm{x}}(t)\|_{\mathrm{G}(\mathrm{x}(t))} = \sqrt{\dot{\mathrm{x}}(t) \mathrm{G}_{\mathrm{x}_t} \dot{\mathrm{x}}(t)} \tag{6}$$

Mode remaining control - via latent geometry

- Desired mode's gating function $h_{k^*} : \mathcal{X} \to \mathbb{R}$
- State trajectory $\bar{\mathbf{x}} : [t_0, t_f] \to \mathcal{X}$
- Length minimising trajectories encode mode remaining behaviour

$$\min \text{Length}(h_{k^*}(\bar{\mathbf{x}})) = \min \int_{t_0}^{t_f} \|\dot{\mathbf{x}}(t)\|_{\mathbf{G}(\mathbf{x}(t))} \, \mathrm{d}t \tag{5}$$

where,

$$\|\dot{\mathbf{x}}(t)\|_{\mathbf{G}(\mathbf{x}(t))} = \sqrt{\dot{\mathbf{x}}(t) \mathbf{G}_{\mathbf{x}_t} \dot{\mathbf{x}}(t)} \tag{6}$$

- But ignores *epistemic uncertainty*...

Mode remaining control - via latent geometry

- Metric depends on Jacobian[1]

$$\mathrm{G}_{\mathbf{x}_t} = \mathrm{J}_{\mathbf{x}_t}^{T}\mathrm{J}_{\mathbf{x}_t} \tag{7}$$

[3] Tosi et al. “Metrics for Probabilistic Geometries”. 2014.

Mode remaining control - via latent geometry

- Metric depends on Jacobian[1]

$$\mathrm{G}_{\mathbf{x}_t} = \mathrm{J}_{\mathbf{x}_t}^{T} \mathrm{J}_{\mathbf{x}_t} \tag{7}$$

- which is Normally distributed

$$\mathrm{J} \sim \mathcal{N}\left(\boldsymbol{\mu}_J, \boldsymbol{\Sigma}_J\right) \tag{8}$$

[3] Tosi et al. “Metrics for Probabilistic Geometries”. 2014.

Mode remaining control - via latent geometry

- Metric depends on Jacobian[1]

$$G_{\mathbf{x}_t} = J_{\mathbf{x}_t}^T J_{\mathbf{x}_t} \tag{7}$$

- which is Normally distributed

$$J \sim \mathcal{N}\left(\boldsymbol{\mu}_J, \boldsymbol{\Sigma}_J\right) \tag{8}$$

- so metric follows non-central Wishart distribution

$$G \sim \mathcal{W}_D\left(p, \boldsymbol{\Sigma}_J, \mathbb{E}\left[J^T\right]\mathbb{E}[J]\right) \tag{9}$$

[3] Tosi et al. “Metrics for Probabilistic Geometries”. 2014.

Mode remaining control - via latent geometry

- Metric depends on Jacobian[1]

$$\mathrm{G}_{\mathbf{x}_t} = \mathrm{J}_{\mathbf{x}_t}^T \mathrm{J}_{\mathbf{x}_t} \tag{7}$$

- which is Normally distributed

$$\mathrm{J} \sim \mathcal{N}\left(\boldsymbol{\mu}_J, \boldsymbol{\Sigma}_J\right) \tag{8}$$

- so metric follows non-central Wishart distribution

$$\mathrm{G} \sim \mathcal{W}_D\left(p, \boldsymbol{\Sigma}_J, \mathbb{E}\left[\mathrm{J}^T\right]\mathbb{E}[\mathrm{J}]\right) \tag{9}$$

- Expected metric increases length of trajectories in regions of high *epistemic uncertainty*

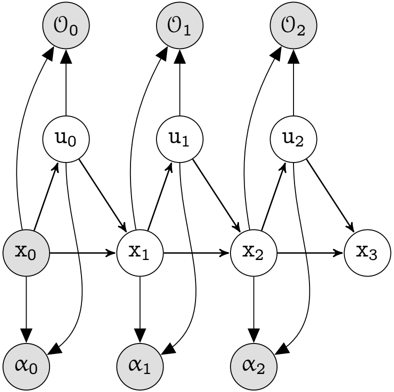
$$\mathbb{E}[\mathrm{G}] = \mathbb{E}[\mathrm{J}^T]\,\mathbb{E}[\mathrm{J}] + \lambda\boldsymbol{\Sigma}_J \tag{10}$$

---

[3] Tosi et al. “Metrics for Probabilistic Geometries”. 2014.

Mode remaining control as probabilistic inference

$$\Pr(\mathcal{O}_t = 1 \mid \mathbf{x}_t, \mathbf{u}_t) \propto \exp(-\gamma c(\mathbf{x}_t, \mathbf{u}_t))$$



Mode remaining control as probabilistic inference

- Goal $p(\bar{\mathrm{x}}, \bar{\mathrm{u}} | \, \mathrm{x}_0, \mathcal{O}_{0:T} = 1, \alpha_{0:T} = k^*)$

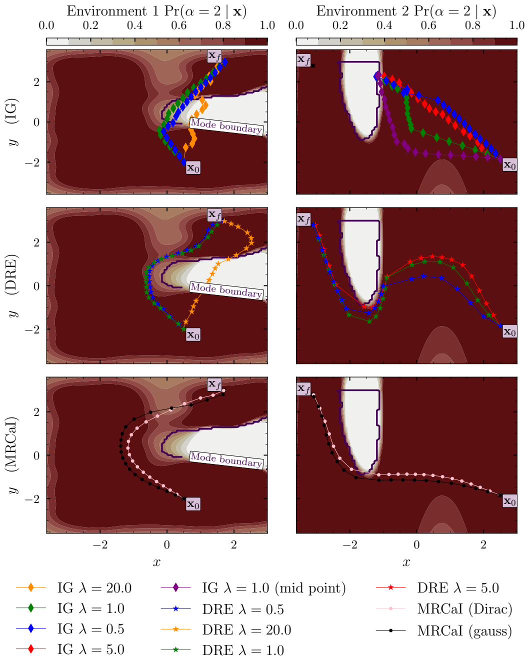
Mode remaining control as probabilistic inference

- Goal $p(\bar{\mathrm{x}}, \bar{\mathrm{u}} | \, \mathrm{x}_0, \mathcal{O}_{0:T} = 1, \alpha_{0:T} = k^*)$
- Variational inference (lower bound $p(\mathcal{O}_{0:T} = 1, \alpha_{0:T} = k^* \mid \mathrm{x}_0))$

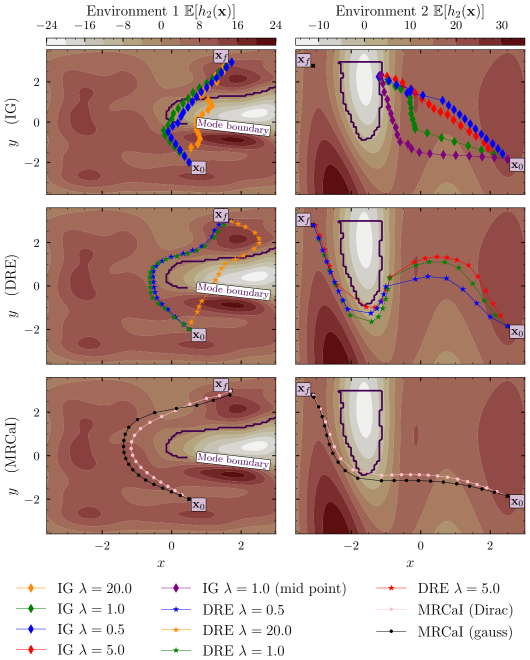
$$\mathcal{L}_{\text{mode}} = - \sum_{t=0}^{T} \underbrace{\mathbb{E}_{q(\mathrm{x}_t | \mathrm{x}_0, \alpha_{0:T} = k^*_{0:t-1}) q(\mathrm{u}_t)} \left[c(\mathrm{x}_t, \mathrm{u}_t) \right]}_{\text{expected cost}} \tag{11}$$

$$+ \sum_{t=0}^{T} \underbrace{\mathbb{E}_{q(\mathrm{x}_t | \mathrm{x}_0, \alpha_{0:T} = k^*_{0:t-1})} \left[\log \Pr(\alpha_t = k^* \mid \mathrm{x}_t) \right]}_{\text{mode remaining term}}$$

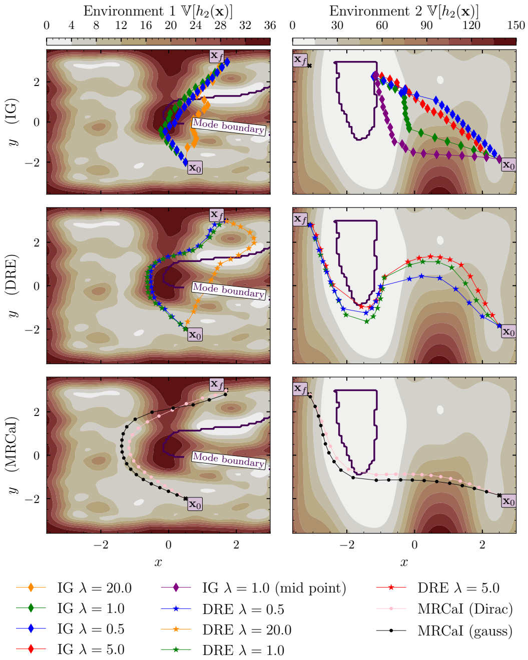
$$+ \sum_{t=0}^{T-1} \underbrace{H[\mathrm{u}_t]}_{\text{entropy}} \tag{12}$$



Environment 1 Pr(α = 2 | x)
Environment 2 Pr(α = 2 | x)
0.0
0.2
0.4
0.6
0.8
1.0
x_f
x_0
Mode boundary
y (IG)
y (DRE)
y (MRCaI)
x
IG λ = 20.0
IG λ = 1.0
IG λ = 0.5
IG λ = 5.0
IG λ = 1.0 (mid point)
DRE λ = 0.5
DRE λ = 20.0
DRE λ = 1.0
DRE λ = 5.0
MRCaI (Dirac)
MRCaI (gauss)



Environment 1 $\mathbb{E}[h_2(\mathbf{x})]$
Environment 2 $\mathbb{E}[h_2(\mathbf{x})]$
y (IG)
y (DRE)
y (MRCaI)
x
$\mathbf{x}_f$
$\mathbf{x}_0$
Mode boundary
IG $\lambda = 20.0$
IG $\lambda = 1.0$
IG $\lambda = 0.5$
IG $\lambda = 5.0$
IG $\lambda = 1.0$ (mid point)
DRE $\lambda = 0.5$
DRE $\lambda = 20.0$
DRE $\lambda = 1.0$
DRE $\lambda = 5.0$
MRCaI (Dirac)
MRCaI (gauss)



Environment 1 $\mathbb{V}[h_2(\mathbf{x})]$
Environment 2 $\mathbb{V}[h_2(\mathbf{x})]$
y (IG)
y (DRE)
y (MRCaI)
x
Mode boundary
$\mathbf{x}_f$
$\mathbf{x}_0$
IG $\lambda = 20.0$
IG $\lambda = 1.0$
IG $\lambda = 0.5$
IG $\lambda = 5.0$
IG $\lambda = 1.0$ (mid point)
DRE $\lambda = 0.5$
DRE $\lambda = 20.0$
DRE $\lambda = 1.0$
DRE $\lambda = 5.0$
MRCaI (Dirac)
MRCaI (gauss)

Mode remaining exploration for MBRL

- Goals

Mode remaining exploration for MBRL

- Goals
  - Navigate to the target state $\mathbf{x}_f$

Mode remaining exploration for MBRL

- Goals
  - Navigate to the target state $\mathbf{x}_f$
  - Remain in desired dynamics mode

Mode remaining exploration for MBRL

- Goals
  - Navigate to the target state $\mathbf{x}_f$
  - Remain in desired dynamics mode
- Assumptions

Mode remaining exploration for MBRL

- Goals
  - Navigate to the target state x_f
  - Remain in desired dynamics mode
- Assumptions
  - Desired dynamics mode is *known a priori*

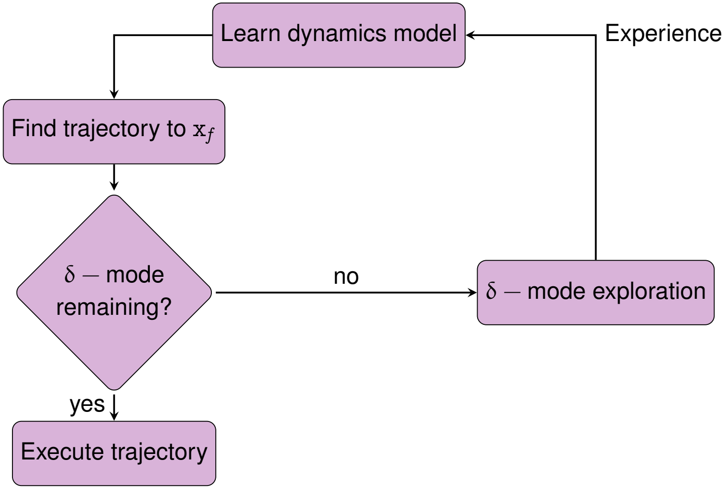
Mode remaining exploration for MBRL

- Goals
  - Navigate to the target state $\mathbf{x}_f$
  - Remain in desired dynamics mode
- Assumptions
  - Desired dynamics mode is *known a priori*
  - No access to environment a priori

Mode remaining exploration for MBRL

- Goals
  - Navigate to the target state x_f
  - Remain in desired dynamics mode
- Assumptions
  - Desired dynamics mode is *known a priori*
  - No access to environment a priori
    - Only a local state transition data set around start state

Mode remaining exploration for MBRL - the loop



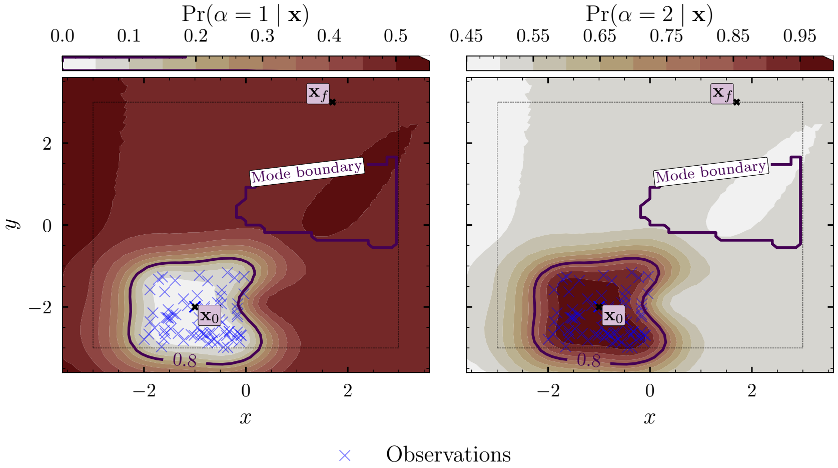
Mode remaining exploration for MBRL - information-based objective

$$\max_{\pi\in\Pi} \underbrace{\mathcal{H}\left[h_{k^*}(\bar{\mathbf{x}}) \mid \bar{\mathbf{x}}, \mathcal{D}_{0:i-1}\right]}_{\text{joint gating entropy}} + \sum_{t=1}^{T-1} \mathbb{E}\left[\underbrace{-(\mathbf{x}_t - \mathbf{x}_f)^T \mathbf{Q}(\mathbf{x}_t - \mathbf{x}_f)}_{\text{state difference}} - \underbrace{\mathbf{u}_t^T \mathbf{R} \mathbf{u}_t}_{\text{control cost}}\right] \tag{13a}$$

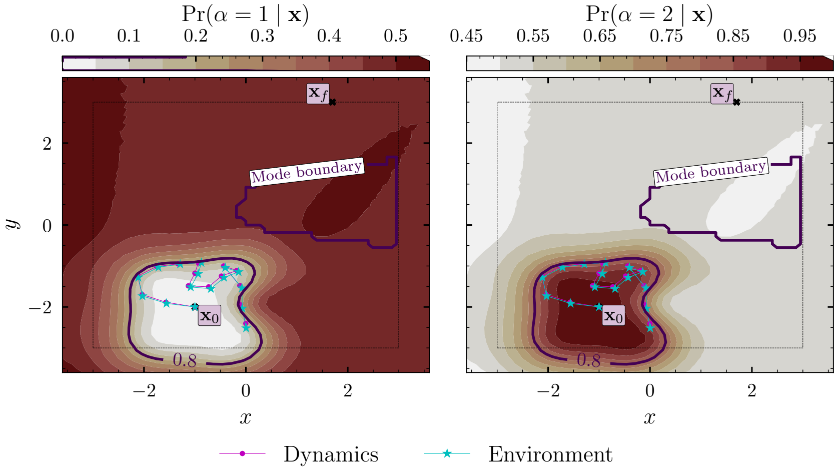
Mode remaining exploration for MBRL - chance constraints

$$\Pr(\alpha_t = k^* \mid \mathbf{x}_0, \mathbf{u}_{0:t}, \boldsymbol{\alpha}_{0:t-1} = k^*) \geqslant 1 - \delta \quad \forall t \in \{0, \dots, T\}$$

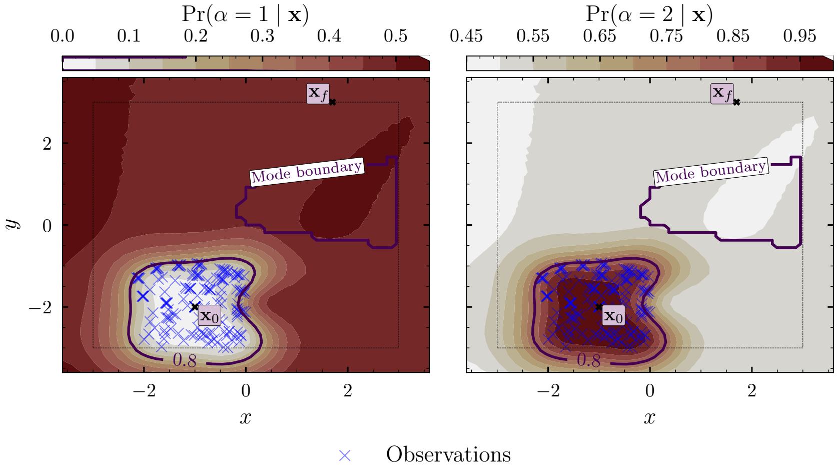
Mode remaining exploration for MBRL - iteration 0



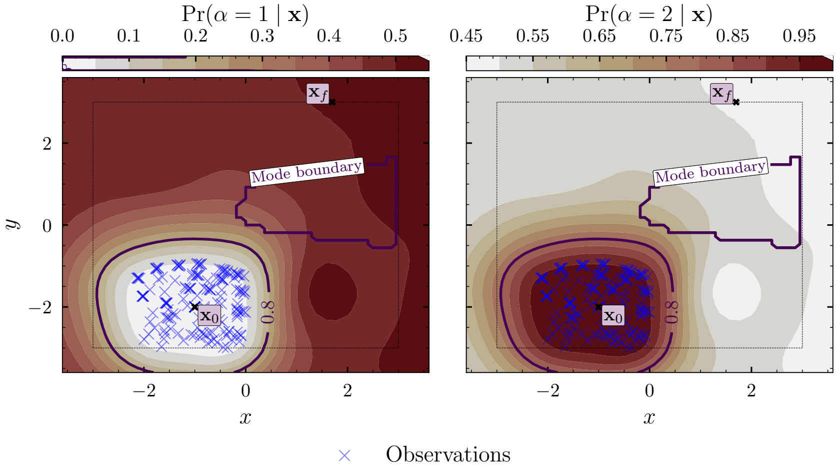
Mode remaining exploration for MBRL - iteration 0



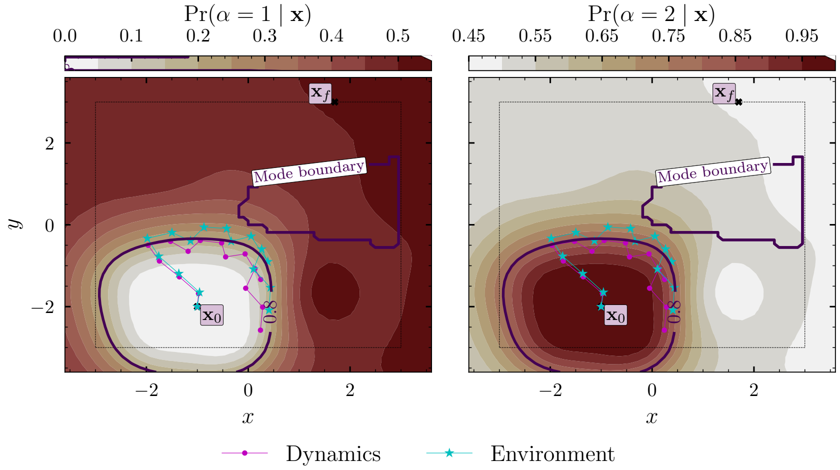
Mode remaining exploration for MBRL - iteration 0



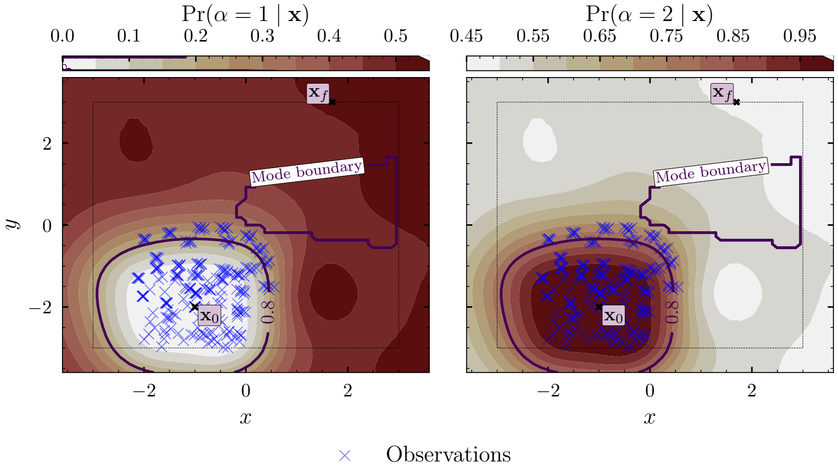
Mode remaining exploration for MBRL - iteration 1



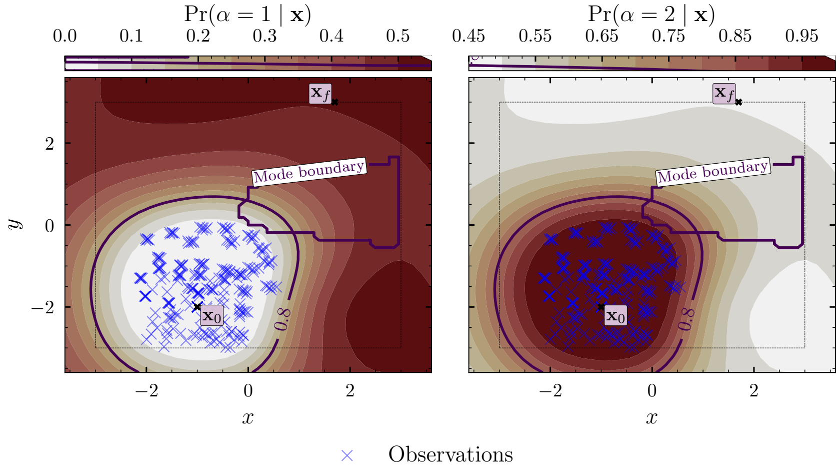
Mode remaining exploration for MBRL - iteration 1



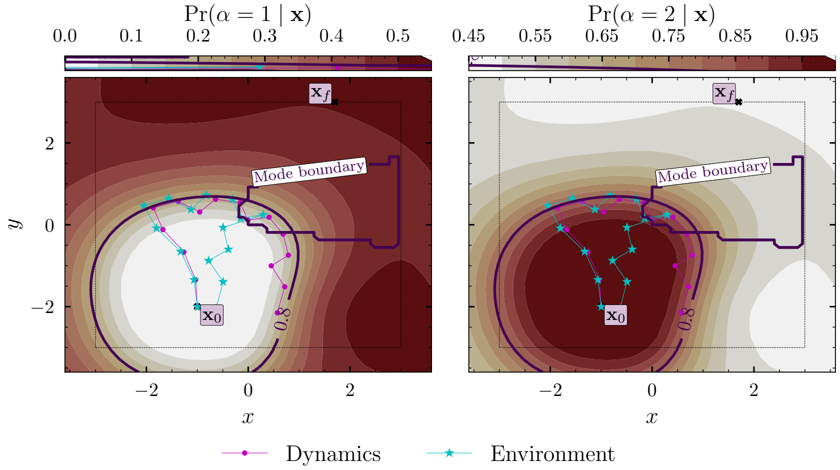
Mode remaining exploration for MBRL - iteration 1



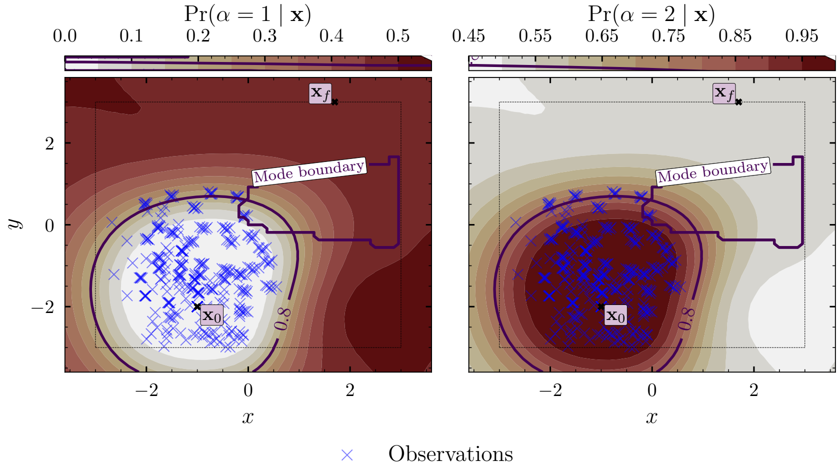
Mode remaining exploration for MBRL - iteration 2



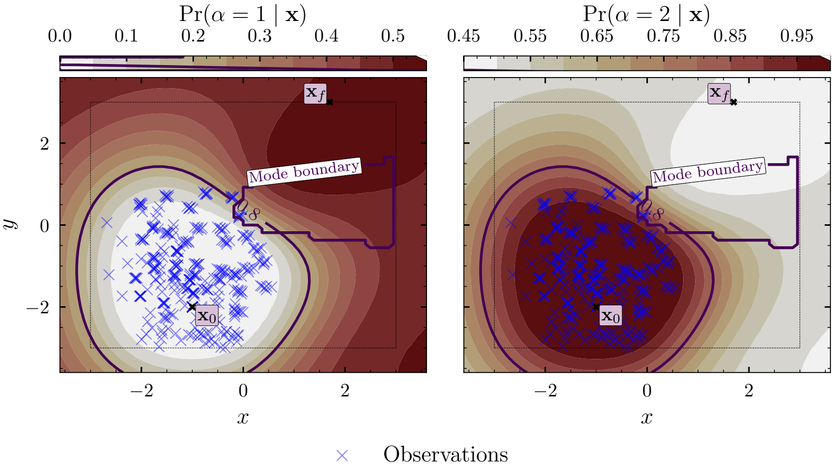
Mode remaining exploration for MBRL - iteration 2



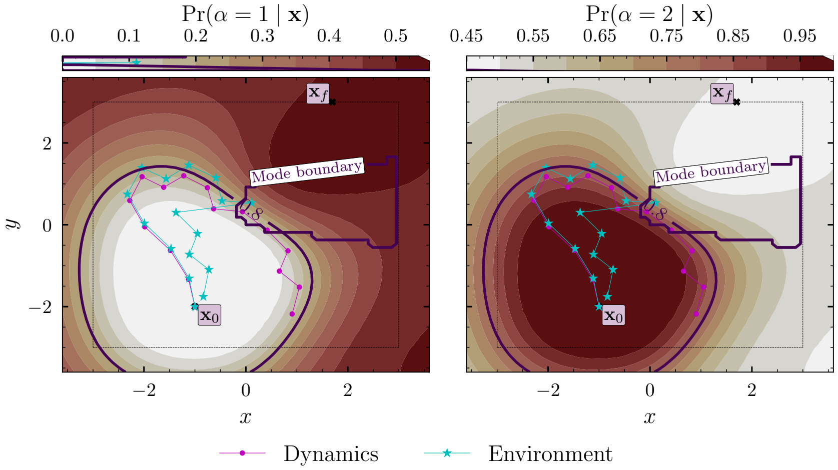
Mode remaining exploration for MBRL - iteration 2



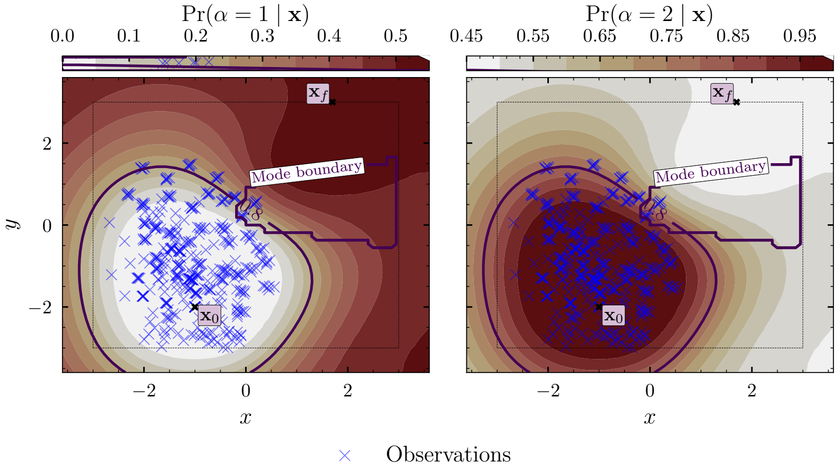
Mode remaining exploration for MBRL - iteration 3



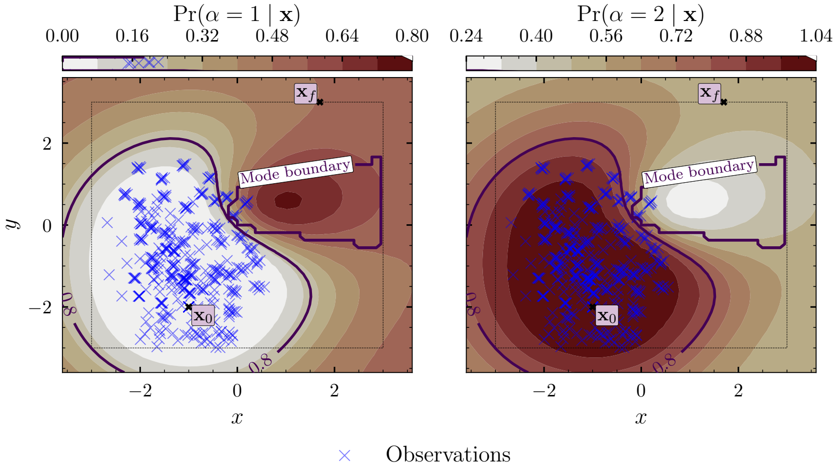
Mode remaining exploration for MBRL - iteration 3



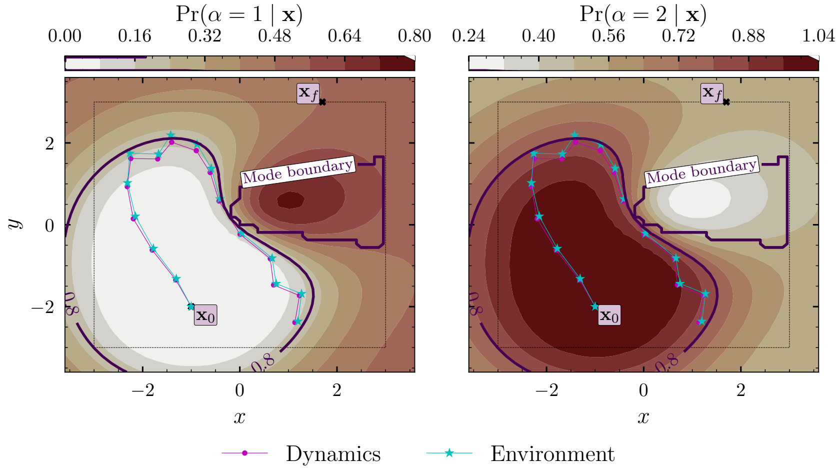
Mode remaining exploration for MBRL - iteration 3



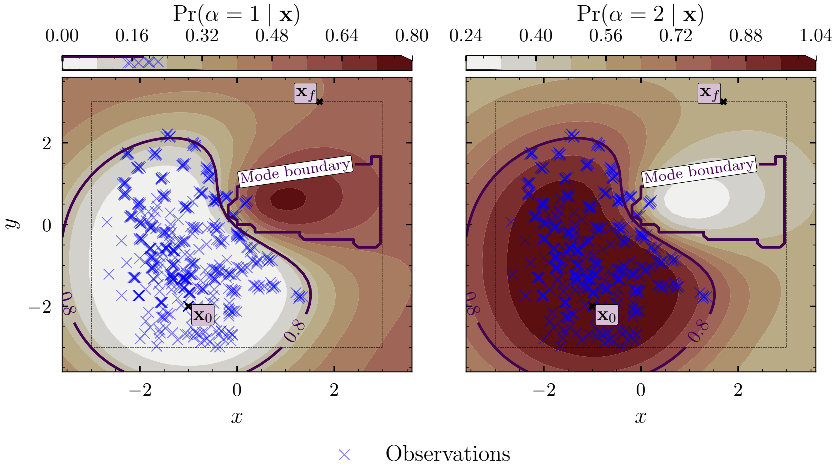
Mode remaining exploration for MBRL - iteration 4



Mode remaining exploration for MBRL - iteration 4

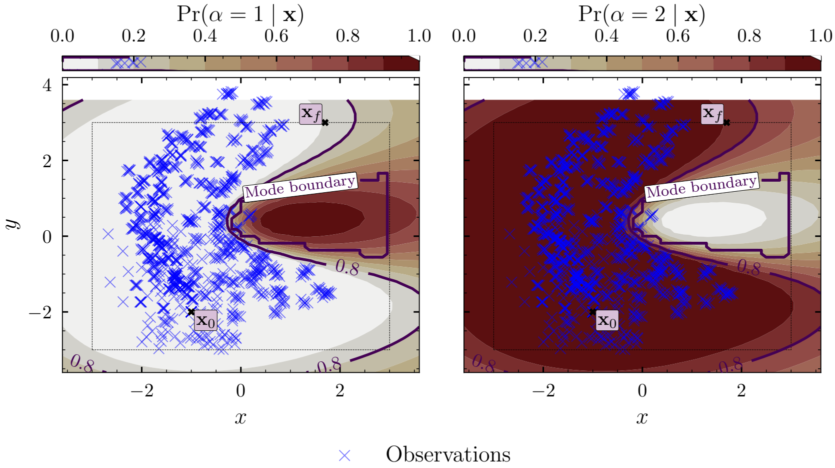


Mode remaining exploration for MBRL - iteration 4



And so on, until

And so on, until



Future work

- Bayesian treatment of inducing inputs

Future work

- Bayesian treatment of inducing inputs
- Dynamically add inducing points during exploration

Future work

- Bayesian treatment of inducing inputs
- Dynamically add inducing points during exploration
- Exploration guarantees

Future work

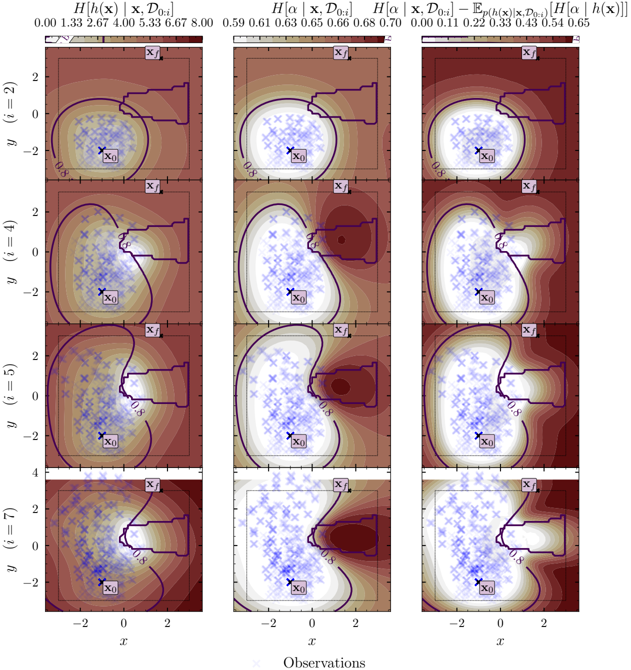
- Bayesian treatment of inducing inputs
- Dynamically add inducing points during exploration
- Exploration guarantees
- External sensing / higher dimensional inputs

Future work

- Bayesian treatment of inducing inputs
- Dynamically add inducing points during exploration
- Exploration guarantees
- External sensing / higher dimensional inputs
- Real-time feedback control, e.g. learn a policy

Future work

- Bayesian treatment of inducing inputs
- Dynamically add inducing points during exploration
- Exploration guarantees
- External sensing / higher dimensional inputs
- Real-time feedback control, e.g. learn a policy
- Better information criterion?



$H[h(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}_{0:i}]$
0.00 1.33 2.67 4.00 5.33 6.67 8.00
$H[\alpha \mid \mathbf{x}, \mathcal{D}_{0:i}]$
0.59 0.61 0.63 0.65 0.66 0.68 0.70
$H[\alpha \mid \mathbf{x}, \mathcal{D}_{0:i}] - \mathbb{E}_{p(h(\mathbf{x})\mid\mathbf{x},\mathcal{D}_{0:i})}[H[\alpha \mid h(\mathbf{x})]]$
0.00 0.11 0.22 0.33 0.43 0.54 0.65
$y \quad (i=2)$
$y \quad (i=4)$
$y \quad (i=5)$
$y \quad (i=7)$
$\mathbf{x}_f$
$\mathbf{x}_0$
0.8
x
Observations

What's next

- MBRL with BNN dynamics

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole
  - When do ensembles fail?

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole
  - When do ensembles fail?
  - When does Laplace approx fail?

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole
  - When do ensembles fail?
  - When does Laplace approx fail?
    - Due to unimodal posterior?

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole
  - When do ensembles fail?
  - When does Laplace approx fail?
    - Due to unimodal posterior?
    - Idea Approximate posterior as Gaussian mixture?

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole
  - When do ensembles fail?
  - When does Laplace approx fail?
    - Due to unimodal posterior?
    - Idea Approximate posterior as Gaussian mixture?
- Multi-step dynamics models

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole
  - When do ensembles fail?
  - When does Laplace approx fail?
    - Due to unimodal posterior?
    - Idea Approximate posterior as Gaussian mixture?
- Multi-step dynamics models
  - If multi-step models can outperform single-step models

What's next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole
  - When do ensembles fail?
  - When does Laplace approx fail?
    - Due to unimodal posterior?
    - Idea Approximate posterior as Gaussian mixture?
- Multi-step dynamics models
  - If multi-step models can outperform single-step models
  - And, BNNs (e.g. Laplace) can work for MBRL

What’s next

- MBRL with BNN dynamics
  - Compare Laplace / MC dropout / ensembles / BNN posterior via sampling
    - Visualise *epistemic uncertainty* e.g. position vs angle for cartpole
  - When do ensembles fail?
  - When does Laplace approx fail?
    - Due to unimodal posterior?
    - Idea Approximate posterior as Gaussian mixture?
- Multi-step dynamics models
  - If multi-step models can outperform single-step models
  - And, BNNs (e.g. Laplace) can work for MBRL
  - Idea use marginal likelihood to set multi-step model’s horizon?

What's next

- Adaptivity in MBRL, i.e. switching reward functions

[2] Khan et al. "Knowledge-Adaptation Priors". 2021.

What's next

- Adaptivity in MBRL, i.e. switching reward functions
  - Latent dynamics learn reward functions from latent, e.g. $r_\theta : \mathcal{Z} \times \mathcal{A} \rightarrow \mathbb{R}$

[2] Khan et al. "Knowledge-Adaptation Priors". 2021.

What's next

- Adaptivity in MBRL, i.e. switching reward functions
  - Latent dynamics learn reward functions from latent, e.g. $r_\theta : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$
  - PlaNet/DreamerV2/MuZero fail when changing reward function

[2] Khan et al. "Knowledge-Adaptation Priors". 2021.

What's next

- Adaptivity in MBRL, i.e. switching reward functions
  - Latent dynamics learn reward functions from latent, e.g. $r_\theta : \mathbb{Z} \times \mathcal{A} \to \mathbb{R}$
  - PlaNet/DreamerV2/MuZero fail when changing reward function
    - Replay buffers lead to interference from old task

[2] Khan et al. "Knowledge-Adaptation Priors". 2021.

What's next

- Adaptivity in MBRL, i.e. switching reward functions
  - Latent dynamics learn reward functions from latent, e.g. $r_\theta : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$
  - PlaNet/DreamerV2/MuZero fail when changing reward function
    - Replay buffers lead to interference from old task
    - Clearing replay buffer leads to catastrophic forgetting

[2] Khan et al. "Knowledge-Adaptation Priors". 2021.

What's next

- Adaptivity in MBRL, i.e. switching reward functions
  - Latent dynamics learn reward functions from latent, e.g. $r_\theta : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$
  - PlaNet/DreamerV2/MuZero fail when changing reward function
    - Replay buffers lead to interference from old task
    - Clearing replay buffer leads to catastrophic forgetting
    - Idea Place K-priors[2] on dynamics?

[2] Khan et al. "Knowledge-Adaptation Priors". 2021.

What's next

- Safety function $c : \mathcal{X} \times \mathcal{A} \to \{0 = \text{safe}, 1 = \text{unsafe}\}$

What's next

- Safety function $c : \mathcal{X} \times \mathcal{A} \to \{0 = \text{safe}, 1 = \text{unsafe}\}$
- Idea 1 Learn safety function using BNN, e.g. $c_\theta : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$

What's next

- Safety function $c : \mathcal{X} \times \mathcal{A} \to \{0 = \text{safe}, 1 = \text{unsafe}\}$
- Idea 1 Learn safety function using BNN, e.g. $c_\theta : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$
  - Probabilistic constraints

$$\prod_{t=1}^{T} \Pr(c_t = 0 \mid \mathbf{x}_t, \mathbf{u}_t) \geqslant 1 - \delta \tag{14}$$

What's next

- Safety function $c : \mathcal{X} \times \mathcal{A} \to \{0 = \text{safe}, 1 = \text{unsafe}\}$
- Idea 1 Learn safety function using BNN, e.g. $c_\theta : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$
  - Probabilistic constraints

$$\prod_{t=1}^{T} \Pr(c_t = 0 \mid \mathbf{x}_t, \mathbf{u}_t) \geqslant 1 - \delta \tag{14}$$

- which consider *epistemic uncertainty* of learned constraints function: $p(\theta \mid \mathcal{D})$

$$\Pr(c_t = 0 \mid \mathbf{x}_t, \mathbf{u}_t) = \int \Pr(c_t = 0 \mid \mathbf{x}_t, \mathbf{u}_t, \theta) p(\theta \mid \mathcal{D}) d\theta \tag{15}$$

What's next

- Safety function $c : \mathcal{X} \times \mathcal{A} \to \{0 = \text{safe}, 1 = \text{unsafe}\}$
- Idea 1 Learn safety function using BNN, e.g. $c_\theta : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$
  - Probabilistic constraints

$$\prod_{t=1}^{T} \Pr(c_t = 0 \mid \mathbf{x}_t, \mathbf{u}_t) \geqslant 1 - \delta \tag{14}$$

    - which consider *epistemic uncertainty* of learned constraints function: $p(\theta \mid \mathcal{D})$

$$\Pr(c_t = 0 \mid \mathbf{x}_t, \mathbf{u}_t) = \int \Pr(c_t = 0 \mid \mathbf{x}_t, \mathbf{u}_t, \theta) p(\theta \mid \mathcal{D}) \mathrm{d}\theta \tag{15}$$

  - Can be extended to latent space dynamics models, e.g. $c_\theta : \mathcal{Z} \times \mathcal{A} \to \{0, 1\}$

What's Next

- Extending Bayesian Active Learning Disagreement[3] to RL

[1] Houlsby et al. "Bayesian Active Learning for Classification and Preference Learning". 2011.

What's Next

- Extending Bayesian Active Learning Disagreement[3] to RL
- In action value function space?

[1] Houlsby et al. "Bayesian Active Learning for Classification and Preference Learning". 2011.

What's Next

- Extending Bayesian Active Learning Disagreement[3] to RL
- In action value function space?
  - Learn Q function $q_\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

$$\pi = \arg\max_{\mathrm{u}_t} \underbrace{q_\theta(\mathrm{x}_t, \mathrm{u}_t)}_{\text{greedy}} + \underbrace{H\left[q_\theta(\mathrm{x}_t, \mathrm{u}_t) \mid \mathrm{x}_t, \mathrm{u}_t, \mathcal{D}_{0:i}\right] - \mathbb{E}_{\theta \sim p(\theta \mid \mathcal{D}_{0:i})}\left[H\left[q_\theta(\mathrm{x}_t, \mathrm{u}_t) \mid \mathrm{x}_t, \mathrm{u}_t, \theta\right]\right]}_{\text{exploration}} \tag{16}$$

[1] Houlsby et al. “Bayesian Active Learning for Classification and Preference Learning”. 2011.

What's Next

- Extending Bayesian Active Learning Disagreement[3] to RL
- In action value function space?
  - Learn Q function $q_\theta : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$

$$\pi = \arg\max_{\mathrm{u}_t} \underbrace{q_\theta(\mathrm{x}_t, \mathrm{u}_t)}_{\text{greedy}} + \underbrace{H\left[q_\theta(\mathrm{x}_t, \mathrm{u}_t) \mid \mathrm{x}_t, \mathrm{u}_t, \mathcal{D}_{0:i}\right] - \mathbb{E}_{\theta \sim p(\theta \mid \mathcal{D}_{0:i})}\left[H\left[q_\theta(\mathrm{x}_t, \mathrm{u}_t) \mid \mathrm{x}_t, \mathrm{u}_t, \theta\right]\right]}_{\text{exploration}} \tag{16}$$

- Or in reward space?

[1] Houlsby et al. "Bayesian Active Learning for Classification and Preference Learning". 2011.

What's Next

- Extending Bayesian Active Learning Disagreement[3] to RL
- In action value function space?
  - Learn Q function $q_\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

$$\pi = \arg\max_{\mathrm{u}_t} \underbrace{q_\theta(\mathrm{x}_t, \mathrm{u}_t)}_{\text{greedy}} + \underbrace{H\left[q_\theta(\mathrm{x}_t, \mathrm{u}_t) \mid \mathrm{x}_t, \mathrm{u}_t, \mathcal{D}_{0:i}\right] - \mathbb{E}_{\theta \sim p(\theta \mid \mathcal{D}_{0:i})}\left[H\left[q_\theta(\mathrm{x}_t, \mathrm{u}_t) \mid \mathrm{x}_t, \mathrm{u}_t, \theta\right]\right]}_{\text{exploration}} \tag{16}$$

- Or in reward space?
  - Learn reward $r_\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

$$r'(\mathrm{x}_t, \mathrm{u}_t) = \underbrace{r_\theta(\mathrm{x}_t, \mathrm{u}_t)}_{\text{greedy}} + \underbrace{H\left[r_\theta(\mathrm{x}_t, \mathrm{u}_t) \mid \mathrm{x}_t, \mathrm{u}_t \mathcal{D}_{0:i}\right] - \mathbb{E}_{\theta \sim p(\theta \mid \mathcal{D}_{0:i})}\left[H\left[r_\theta(\mathrm{x}_t, \mathrm{u}_t) \mid \mathrm{x}_t, \mathrm{u}_t, \theta\right]\right]}_{\text{exploration}}. \tag{17}$$

[1] Houlsby et al. "Bayesian Active Learning for Classification and Preference Learning". 2011.

Thanks for listening

Questions?

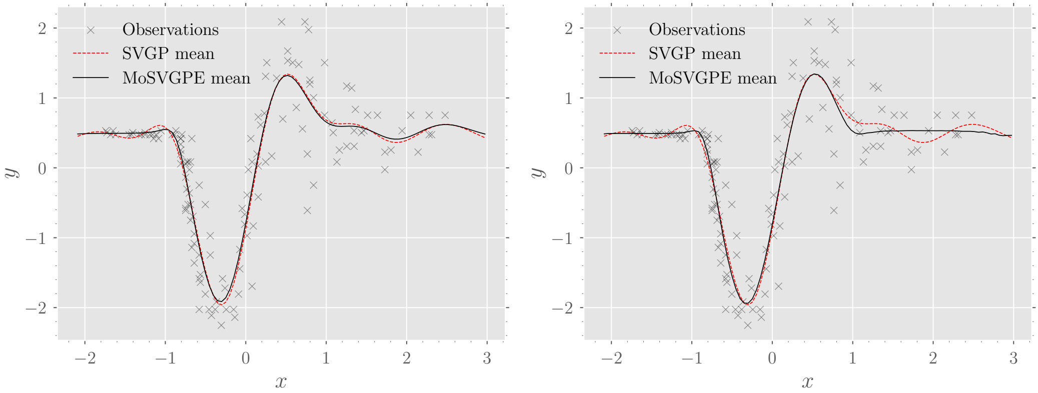
$<_k$all> $<_k$all>

[1] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. "Bayesian Active Learning for Classification and Preference Learning". Dec. 24, 2011. arXiv: 1112.5745 [cs, stat].

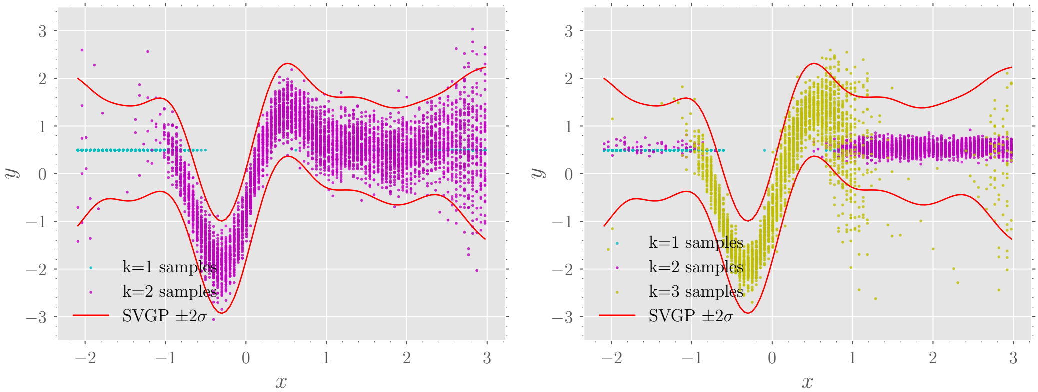
[2] Mohammad Emtiyaz E Khan and Siddharth Swaroop. "Knowledge-Adaptation Priors". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 19757–19770.

[3] Alessandra Tosi, Søren Hauberg, Alfredo Vellido, and Neil D Lawrence. "Metrics for Probabilistic Geometries". In: *Proceedings of the 30th Conference*. Uncertainty in Artificial Intelligence. 2014, pp. 800–808.

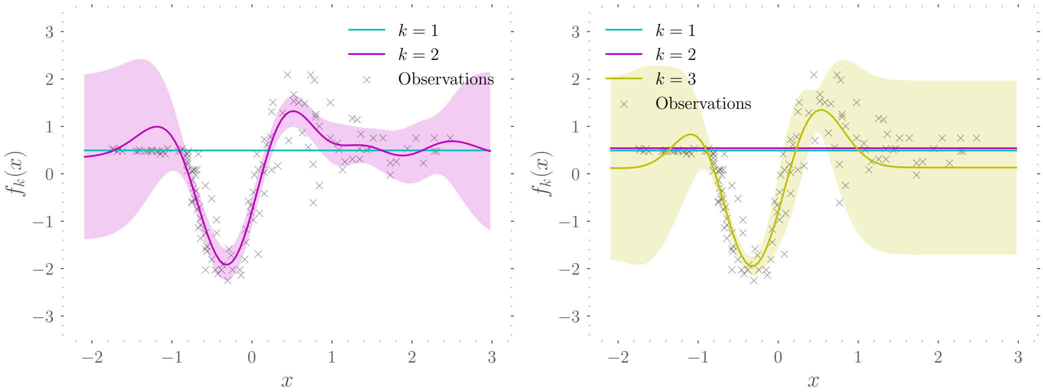
Motorcycle dataset | 2 experts vs 3 experts | Mean



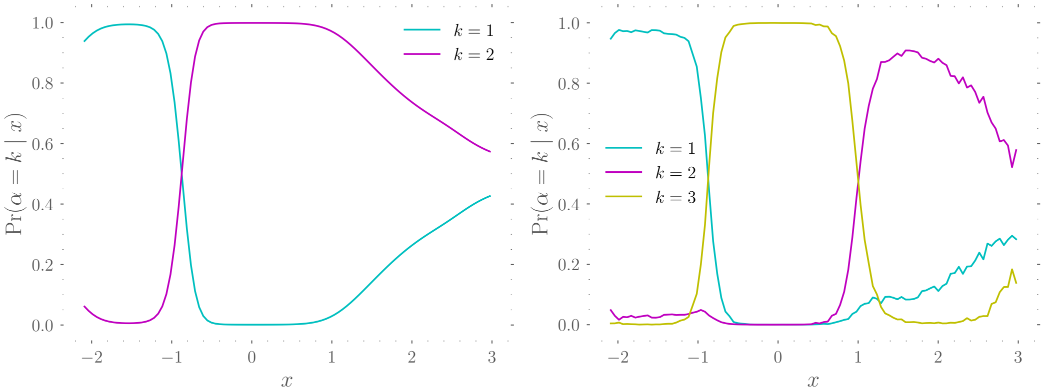
Motorcycle dataset | 2 vs 3 experts | Posterior samples



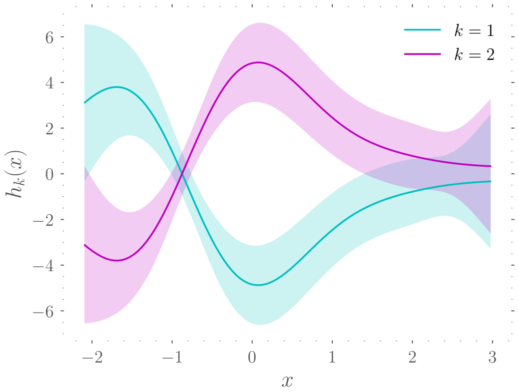
Motorcycle dataset | 2 vs 3 experts | Experts' GP posteriors

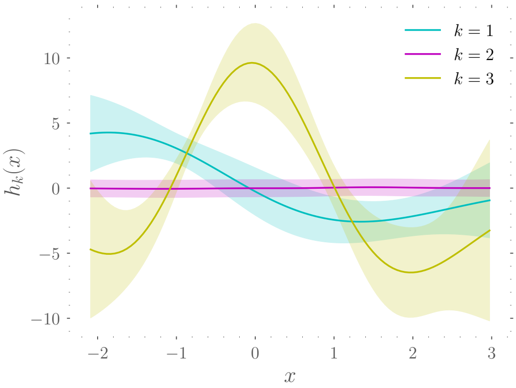


Motorcycle dataset | 2 vs 3 experts | Mixing probabilities



Motorcycle dataset | 2 vs 3 experts | Gating GP posteriors





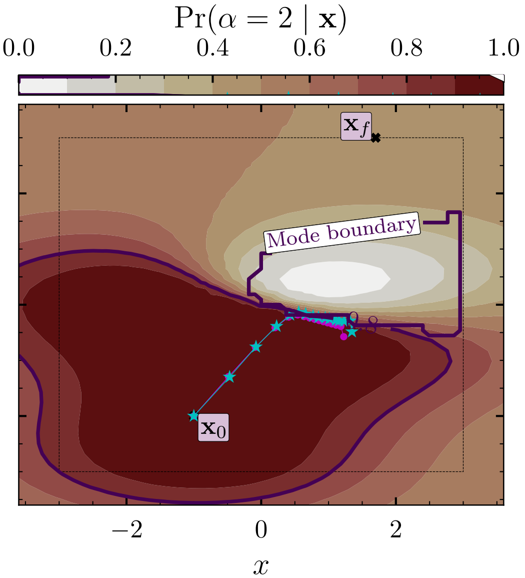
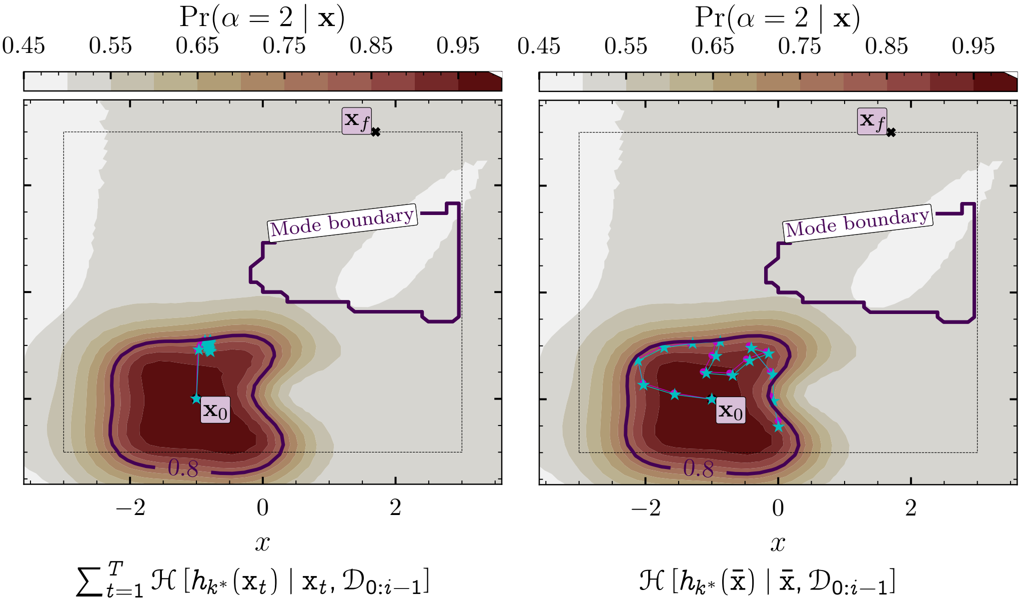


Figure: State difference only $\sum_{t=1}^{T-1} \mathbb{E}\left[-(\mathrm{x}_t - \mathrm{x}_f)^T \mathrm{Q}(\mathrm{x}_t - \mathrm{x}_f)\right]$



$\Pr(\alpha = 2 \mid \mathbf{x})$
0.45
0.55
0.65
0.75
0.85
0.95
$\mathbf{x}_f$
Mode boundary
$\mathbf{x}_0$
0.8
−2
0
2
x
$\sum_{t=1}^{T} \mathcal{H}\left[h_{k^*}(\mathbf{x}_t) \mid \mathbf{x}_t, \mathcal{D}_{0:i-1}\right]$
$\Pr(\alpha = 2 \mid \mathbf{x})$
0.45
0.55
0.65
0.75
0.85
0.95
$\mathbf{x}_f$
Mode boundary
$\mathbf{x}_0$
0.8
−2
0
2
x
$\mathcal{H}\left[h_{k^*}(\bar{\mathbf{x}}) \mid \bar{\mathbf{x}}, \mathcal{D}_{0:i-1}\right]$