Towards Model-Based Reinforcement Learning on Real Robots

by

Georg Martius

AUTONOMOUS LEARNING
MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

Vision

Dexterous and versatile robots
as assistance to humans

[Kuka]

(frauenhofer ipa)

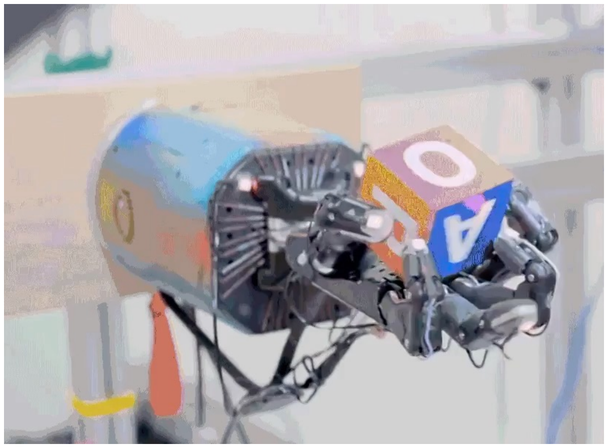
[bergkvistanna karin@tuvie.com]

[NCCR Digital Fabrication]

- learning
- adaptivity
- safety

Reinforcement learning achievements

Robotics, Games: Go, Dota, Starcraft



[OpenAI 2019]

[Deepmind 2019]

Problems

need probitive amount of data
simulations
need very long time

[ionos.com]

[DARPA rescue challenge]

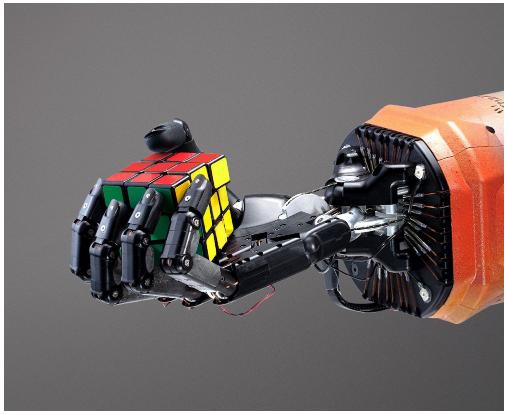
Georg Martius <georg.martius@tue.mpg.de>

Current Situation

Learned robot control in recent **research**

- using simulations of robot and environment
- trained to cope with anticipated variations

✓ works already for difficult tasks
✗ needs a high-fidelity simulation
✗ learning is inefficient (needs domain randomization)
✗ resulting controller is fixed
✗ new task: start from scratch



[OpenAI 2019]

[Hutter lab. ETH, 2020]

How to achieve efficient learning and online adaptation?

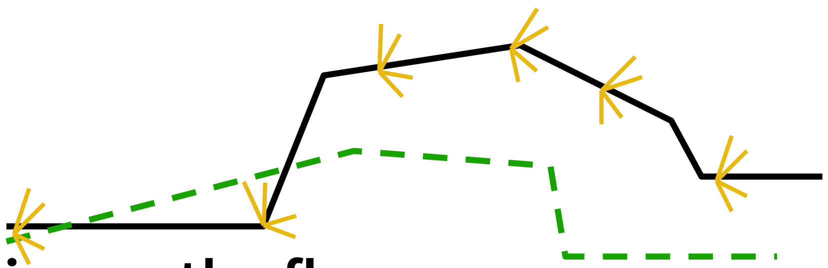
Georg Martius <georg.martius@tue.mpg.de>

Model-based Reinforcement Learning

Two instantiations

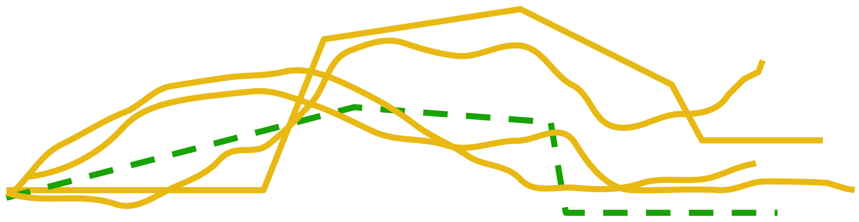
Bellman operator to optimize a value function and the policy

- ➢ use model to collect data nearby real observation
- ➢ learn to solve a specific task
- ➢ global optimimization



Planning to search for a policy on the fly

- ➢ use model for planning
- ➢ perform new task on the fly
- ➢ optimize finite horizon problem

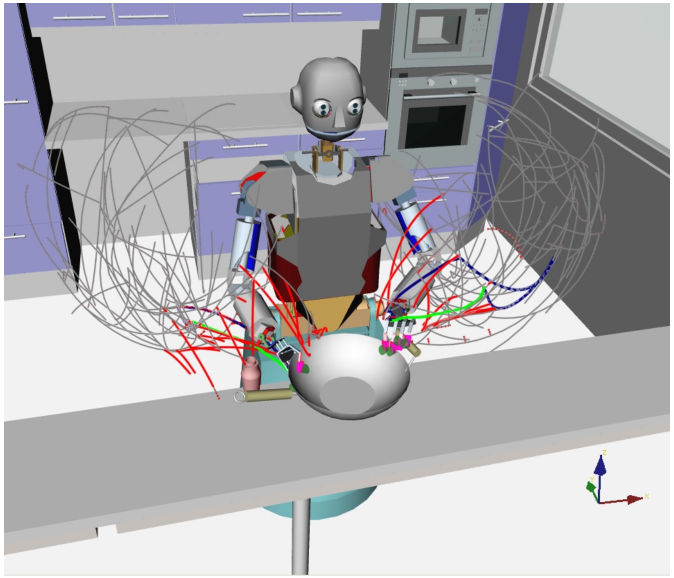


Georg Martius <georg.martius@tue.mpg.de>

Model-based RL with Planning

Challenges:

- Real time planning
  - previous general purpose planner are
  - 1-2 orders of magnitudes too slow
- Good models + uncertainty aware
- Safety



(KIT H^{2}T)

Cristina Pinneri

Sebastian Blaes

Marin Vlastelica

Shambhuraj Sawant

Georg Martius

Model-based Planning

Cross Entropy Method (CEM)

➢ Sampling based optimization

$$a_1, \dots, a_h \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

$$a_1, \dots, a_h = \arg\min_{a_1,\dots,a_h} J(a_1, \dots, a_h)$$

J cost of rollout

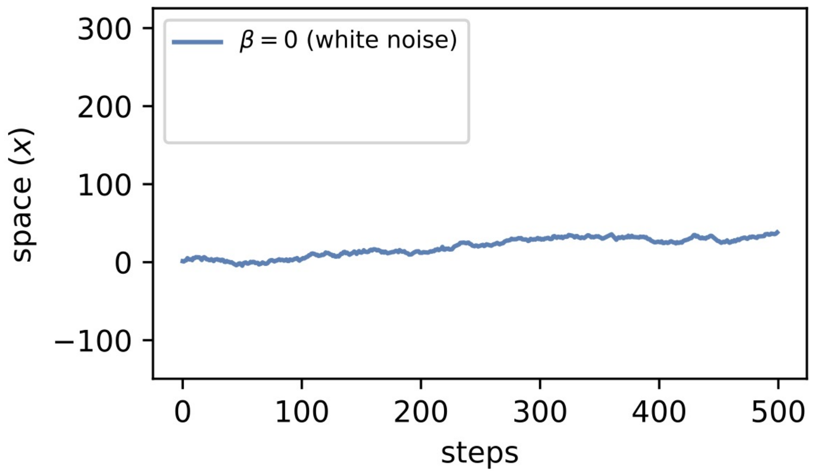
Georg Martius <georg.martius@tue.mpg.de>

Random Walk Model and Colored Noise

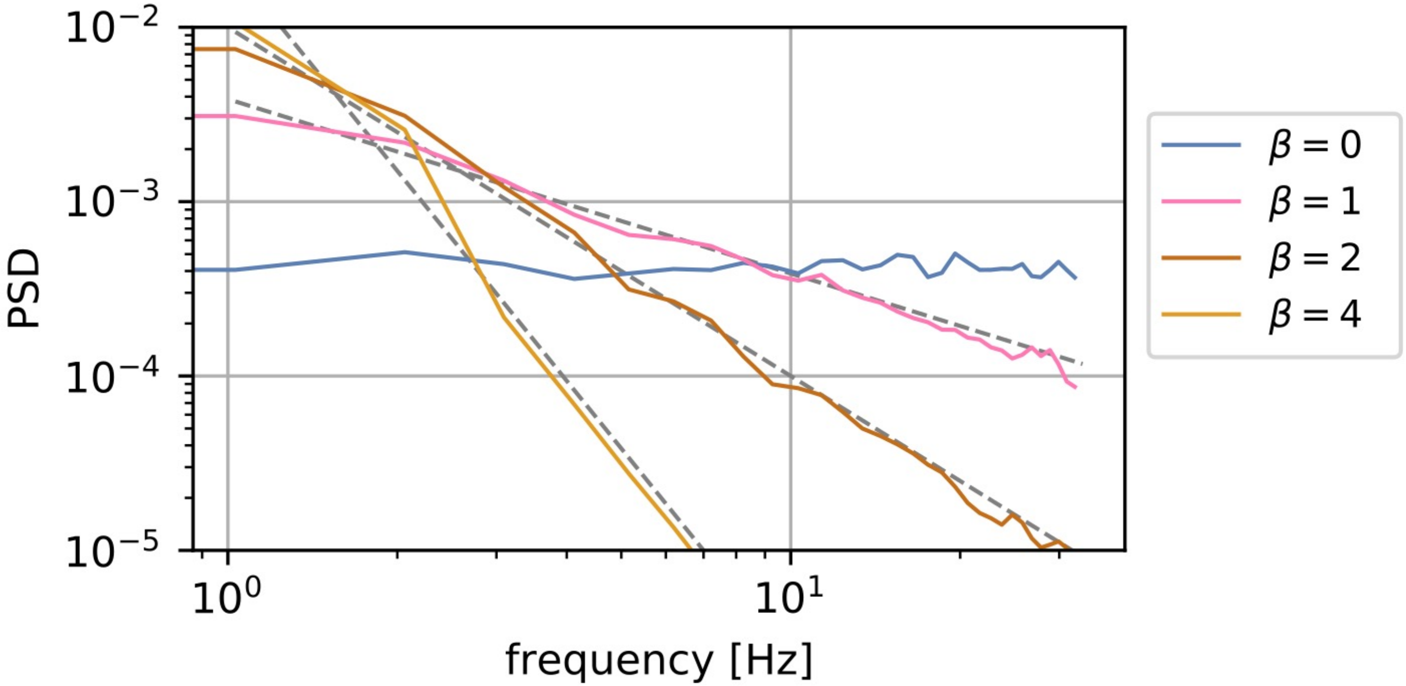
$$x_{t+1} = x_t + a_t$$

Brownian random walk $a_t \sim \mathcal{N}(0,1)$

Lévy walk $a_t \sim \mathcal{CN}^{\beta}(0,1)$



Power spectral density of action sequences



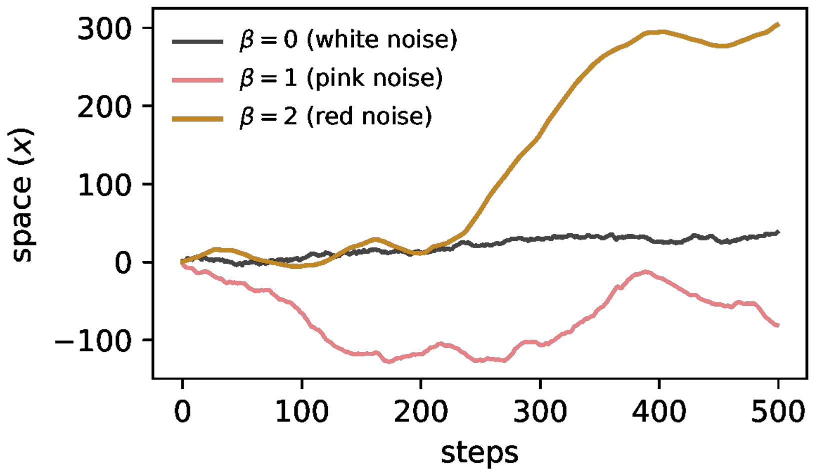
 Georg

Random Walk Model and Colored Noise

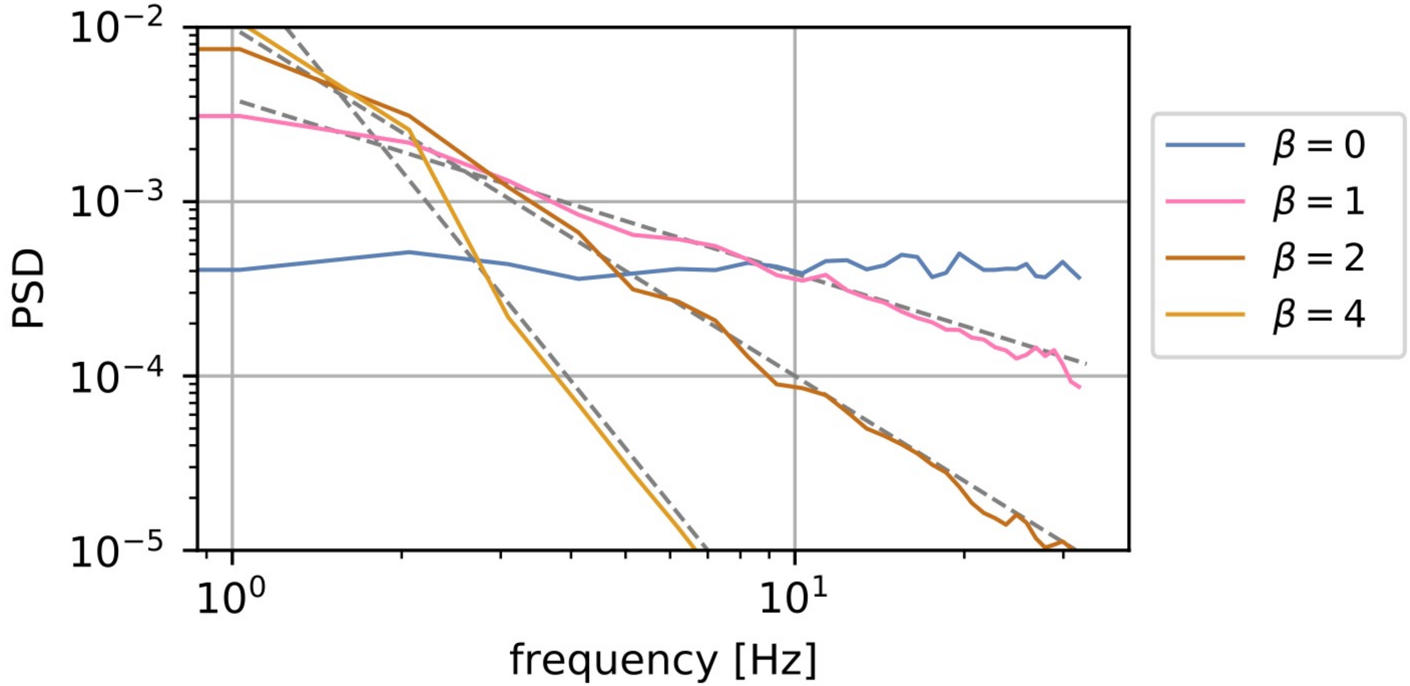
$$x_{t+1} = x_t + a_t$$

Brownian random walk $a_t \sim \mathcal{N}(0,1)$

Lévy walk $a_t \sim \mathcal{CN}^{\beta}(0,1)$



Power spectral density of action sequences



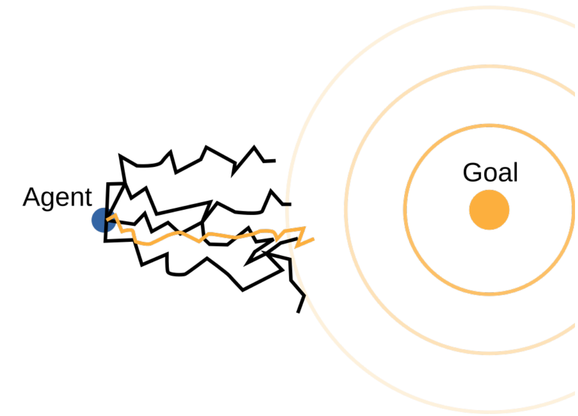
 Georg

Model-based Planning

Cross Entropy Method (CEM)

➢ Sampling based optimization

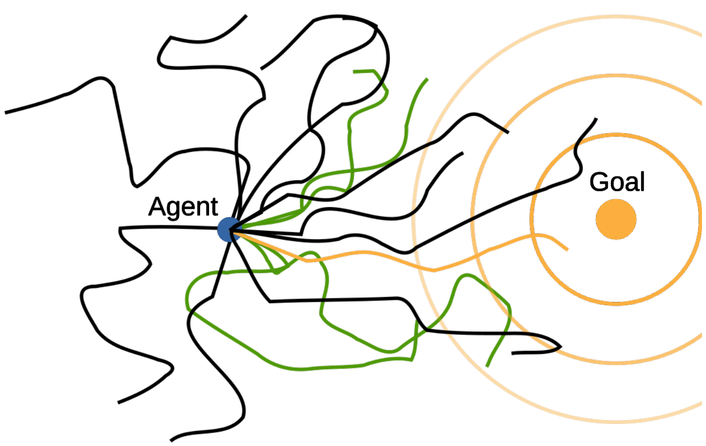
$$a_{t,\dots,t+H} \sim \mathcal{N}(\mu_i, \sigma_i^2)$$



improved Cross Entropy Method

- Memory
- Colored noise: temporal correlation

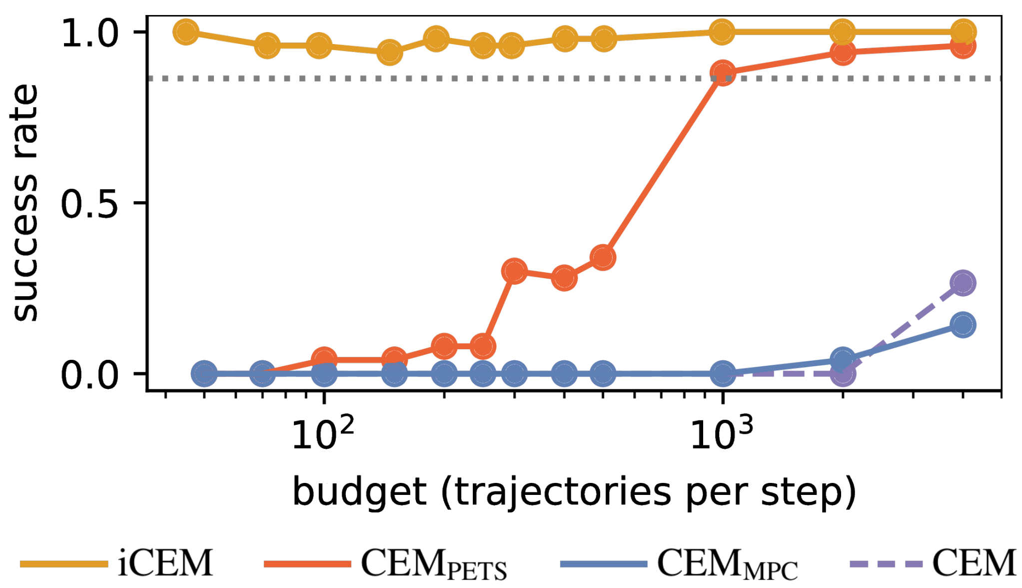
  power-law spectrum: $\mathrm{PSD} \propto \frac{1}{f^\beta}$
- small improvements



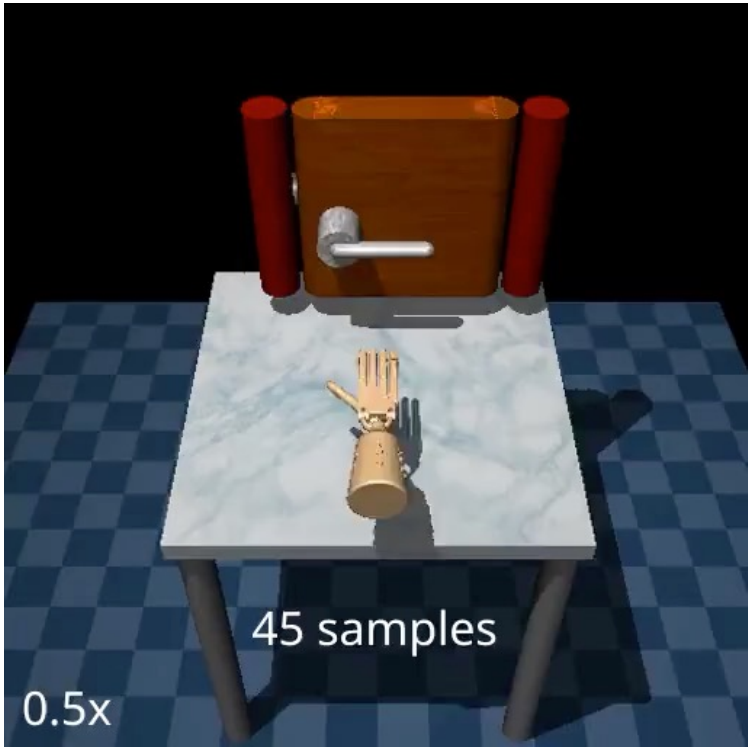
Georg Martius <georg.martius@tue.mpg.de>
Pinneri, Sawant, Blaes, Achterhold, Stückler, GM. *CORL* 2020

Model-based Planning

Door (sparse reward)



Ground Truth Models

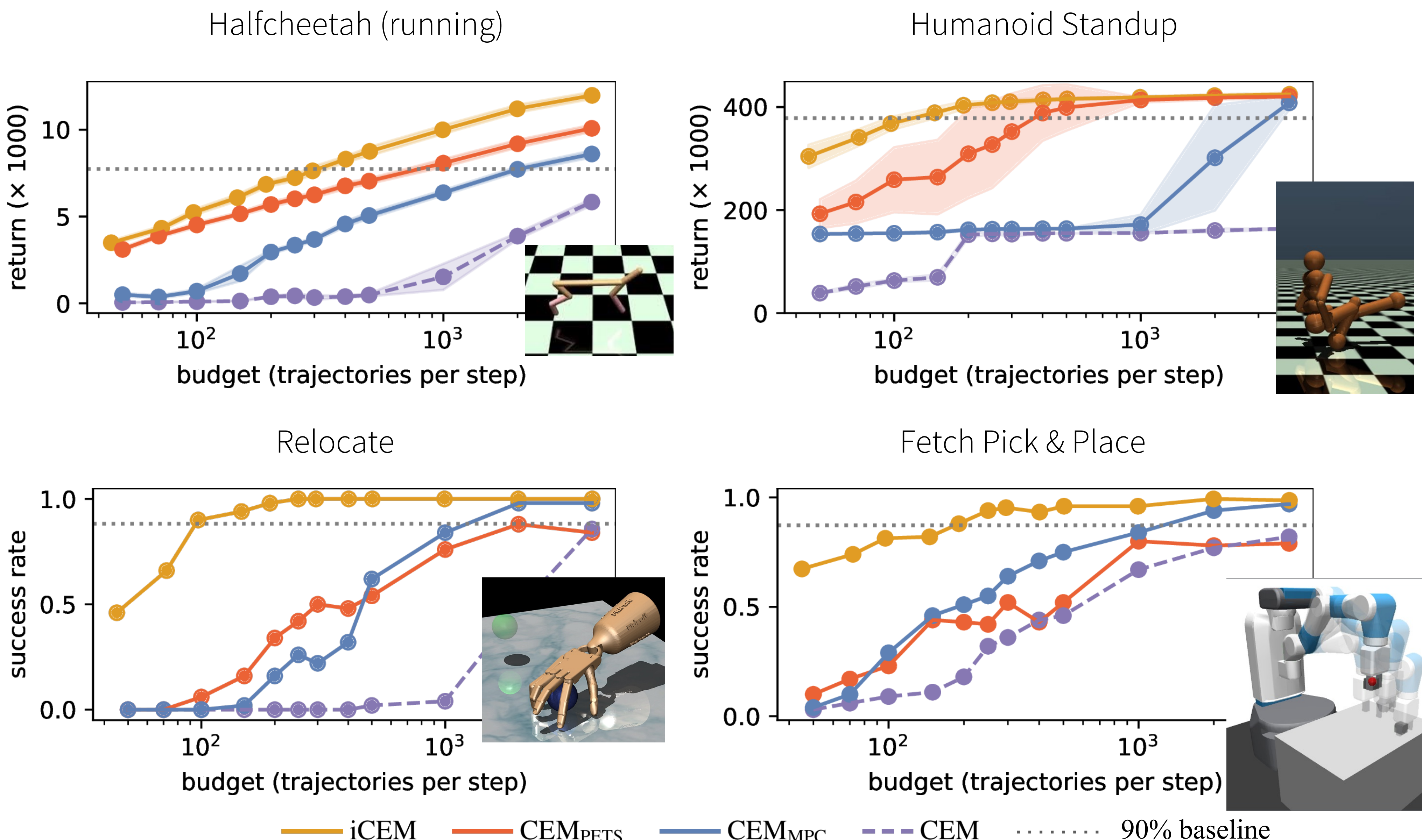


(environment from DAPG project)

Georg Martius <georg.martius@tue.mpg.de>

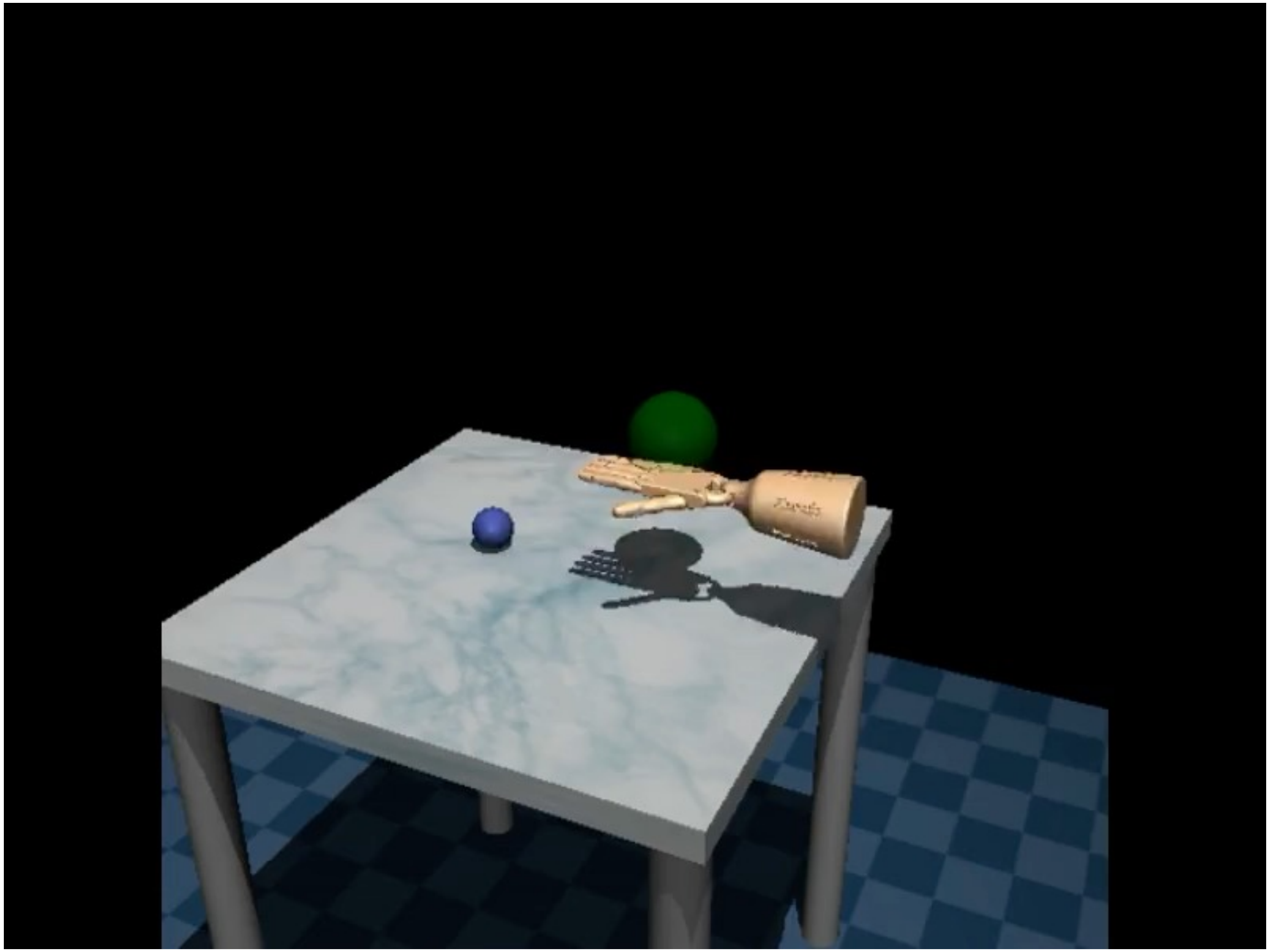
Pinneri, Sawant, Blaes, Achterhold, Stückler, GM. *CORL* 2020

Model-based Planning

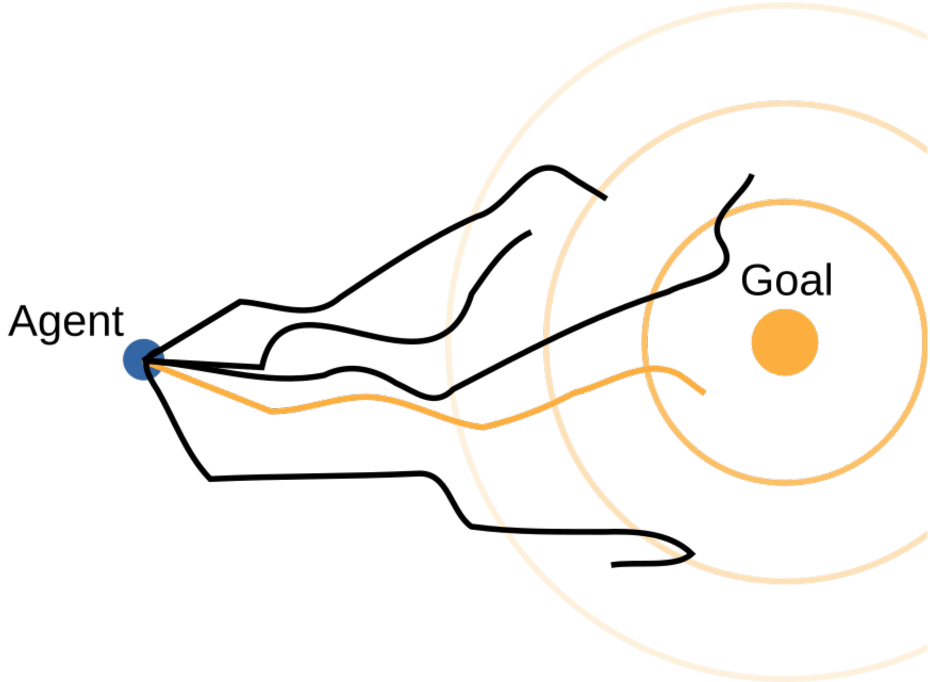


Pinneri, Sawant, Blaes, Achterhold, Stückler, GM. *CORL* 2020

Action



Georg Martius <georg.martius@tue.mpg.de>



Fast planner – iCEM

extract policies
for fast execution

learned models +
safety aware planning

Georg Martius <georg.martius@tue.mpg.de>

Learn Policy from Plans

We can create solutions for complicated control problems within seconds, but:

- ➢ need a lot of run-time compute
- ➢ mostly with Ground Truth models (simulations)

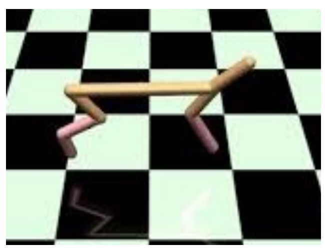
Goal of our method:

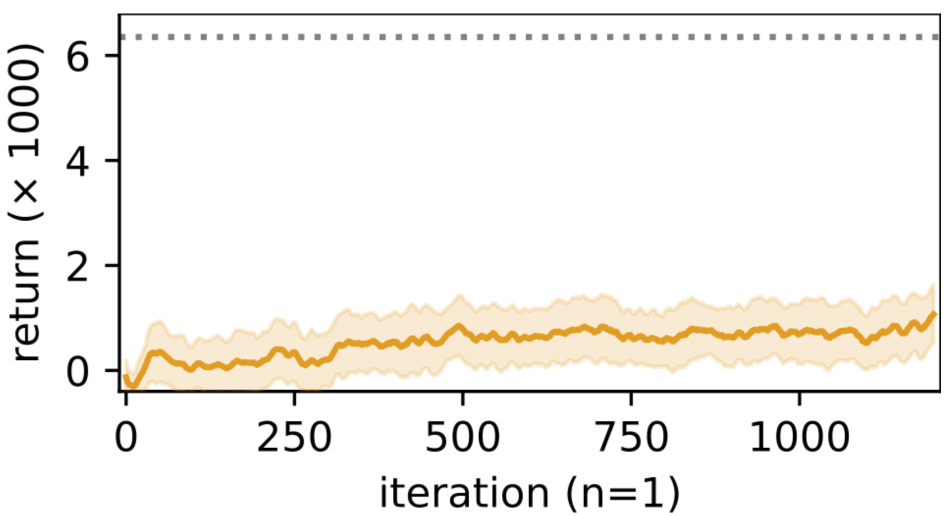
- ➢ train policy from planner data
- ➢ make policy and planner mutually improve themselves
- ➢ solve tasks that standard RL struggles with

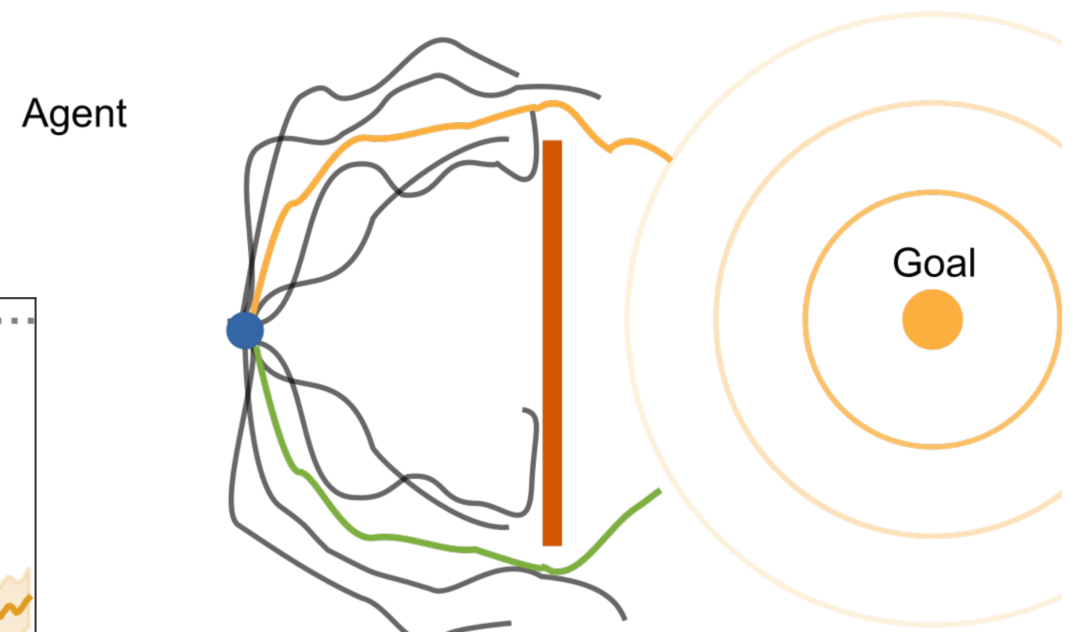
Challenges of learning policy from planner's data

Let's do simple behavioral cloning:

- ➢ does not work!
- ➢ multimodality + combounding errors







Georg Martius <georg.martius@tue.mpg.de>

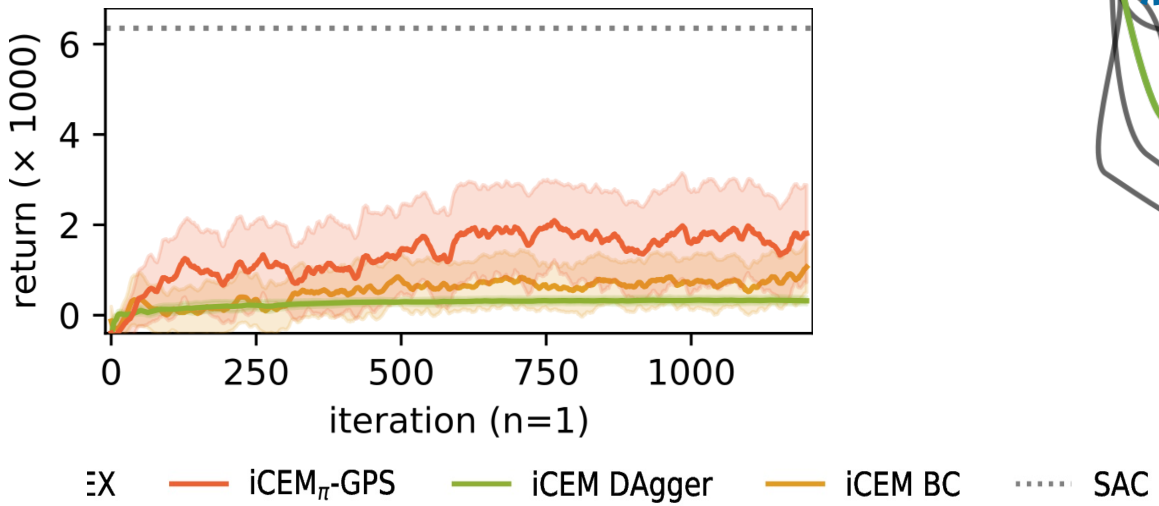
Pinneri*, Sawant*, Blaes, GM. ICLR 2021

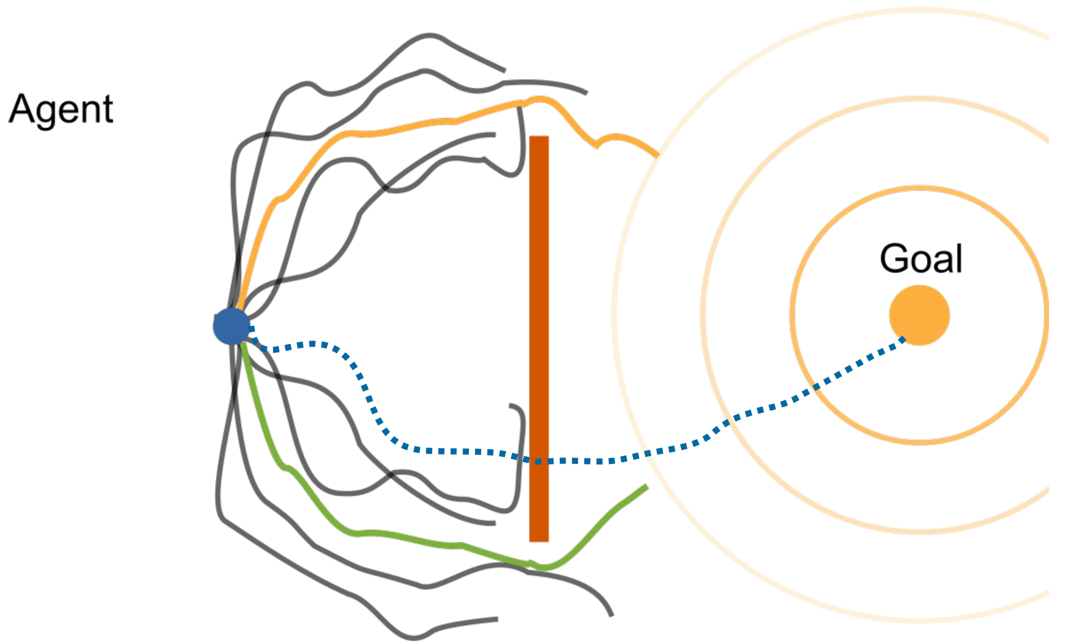
Challenges of learning policy from planner's data

Okay, use guidance (*guided policy search)*

$$a_1, \ldots, a_h = \underset{a_1,\ldots,a_h}{\arg\min}\, f(a_1, \ldots, a_h) + \lambda \sum_i \|a_i - \pi(s_i)\|$$

better, but not ideal!





Georg Martius <georg.martius@tue.mpg.de>
Pinneri*, Sawant*, Blaes, GM. ICLR 2021

Challenges of learning policy from planner's data

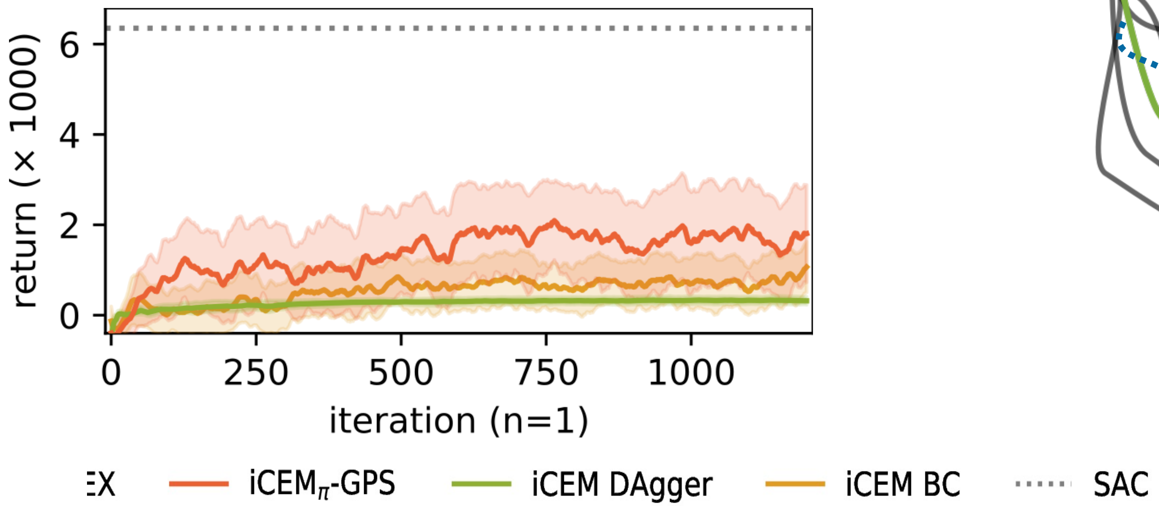
Okay, use guidance (*guided policy search)*

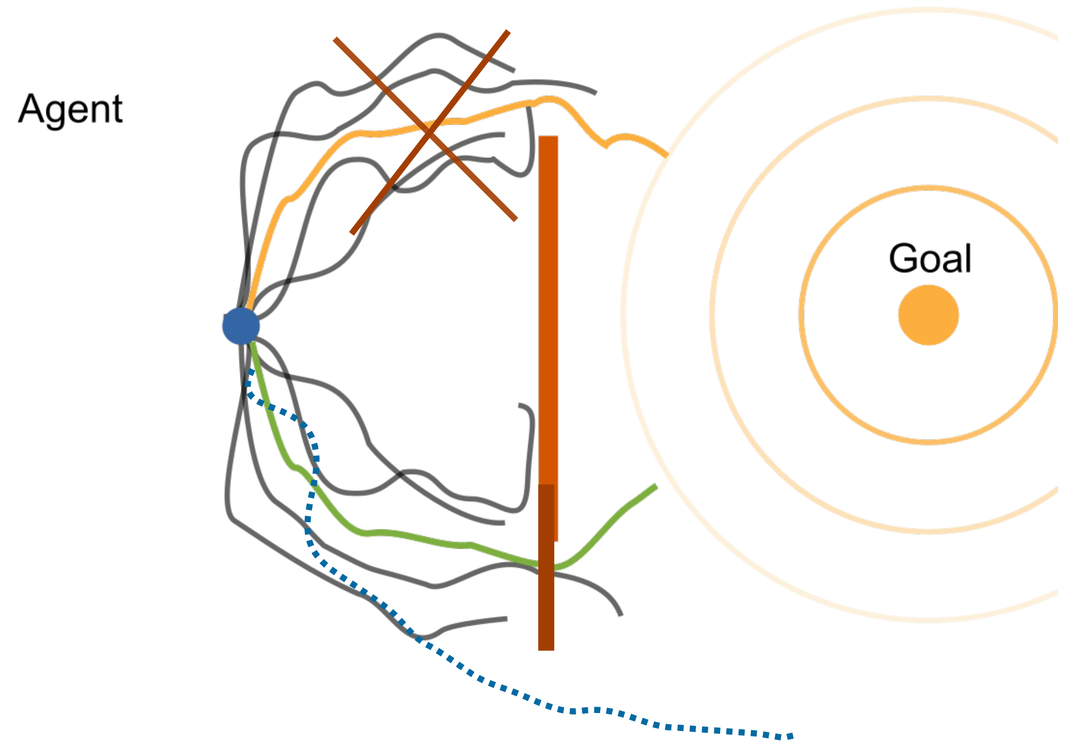
$$a_1, \ldots, a_h = \underset{a_1,\ldots,a_h}{\arg\min}\, f(a_1, \ldots, a_h) + \lambda \sum_i \|a_i - \pi(s_i)\|$$

better, but not ideal!

➢ planner premature convergence

➢ combounding errors





Pinneri*, Sawant*, Blaes, GM. ICLR 2021

Challenges of learning policy from planner's data

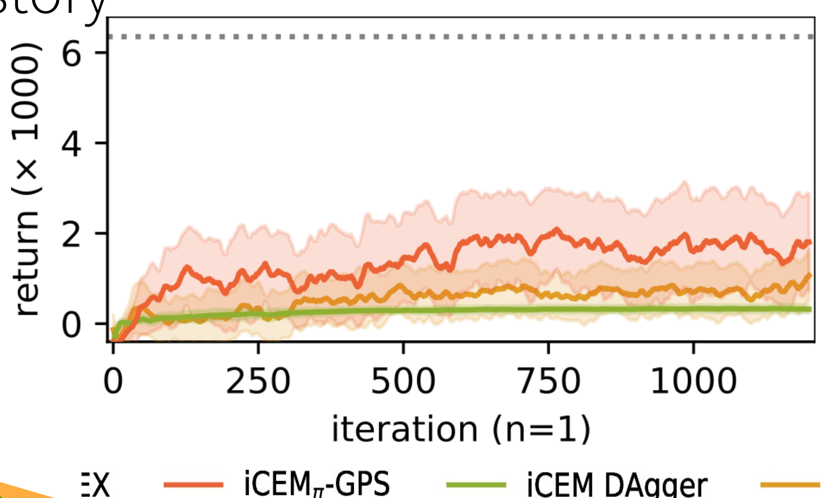
Okay, use guidance (*guided policy search*)

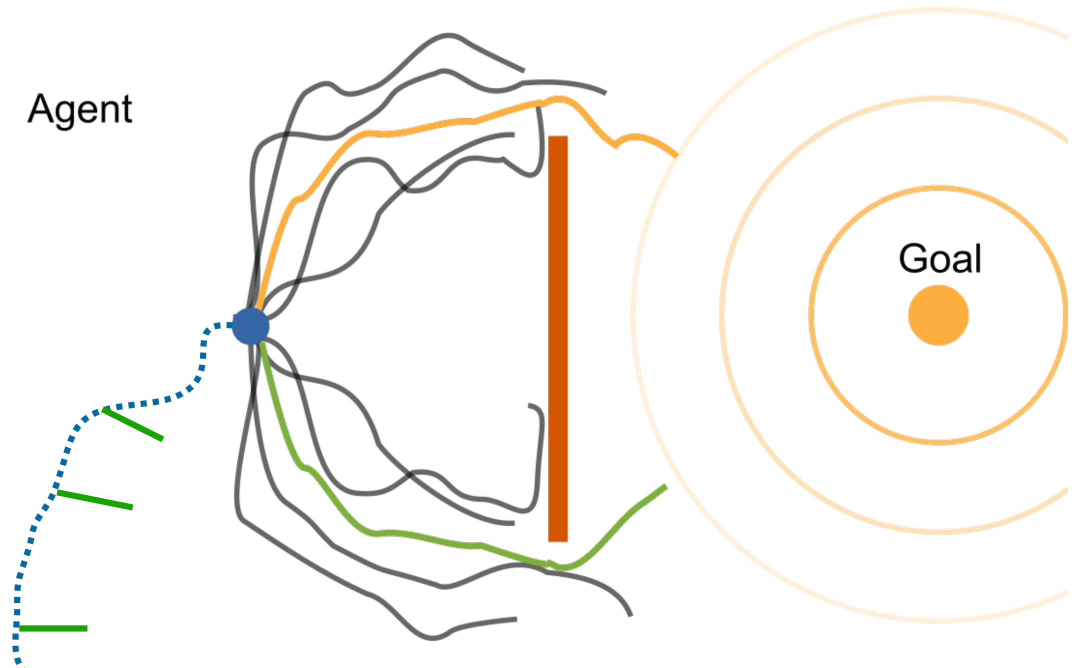
$$a_1, \ldots, a_h = \underset{a_1,\ldots,a_h}{\arg\min}\, f(a_1, \ldots, a_h) + \lambda \sum_i \|a_i - \pi(s_i)\|$$

+ relabeling (*DAgger*):

➢ ask planner for better action

➢ Problem:
planner is not good without history





Pinneri*, Sawant*, Blaes, GM. ICLR 2021

Challenges of learning policy from planner's data

Okay, use guidance (*guided policy search)*

$$a_1, \ldots, a_h = \arg\min_{a_1,\ldots,a_h} J(a_1, \ldots, a_h) + \lambda C^{\text{aux}}(a_1, \ldots, a_h)$$

+ relabeling (*Dagger*)

+ adaptive λ

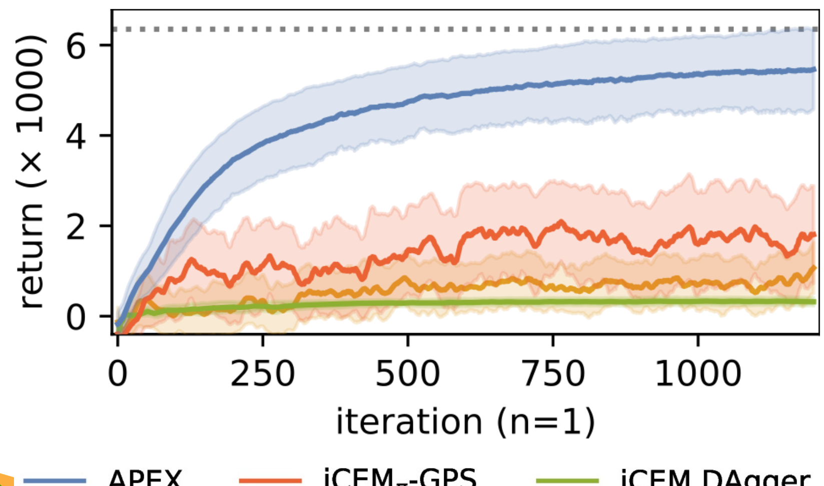
don't use guidance

if no progress on actual task

$$\lambda = c \frac{\mathcal{R}(J)}{\mathcal{R}(C^{\text{aux}}) + \epsilon}$$

c relative importance

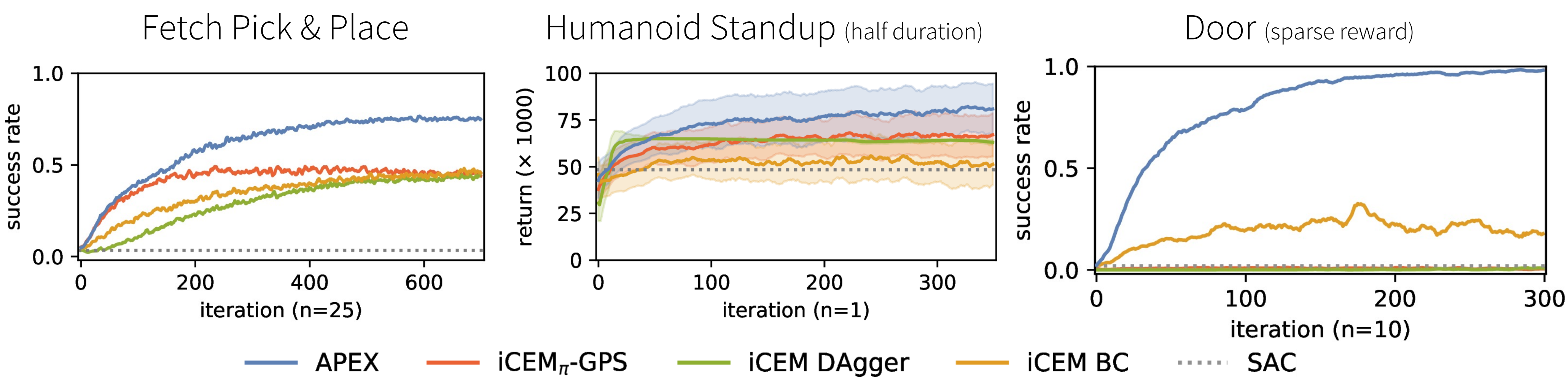
$$\mathcal{R}(C) = \max_{\text{elite-set}} C - \min_{\text{elite-set}} C$$



APEX (Adaptive Policy Extraction)

Pinneri*, Sawant*, Blaes, GM. ICLR 2021

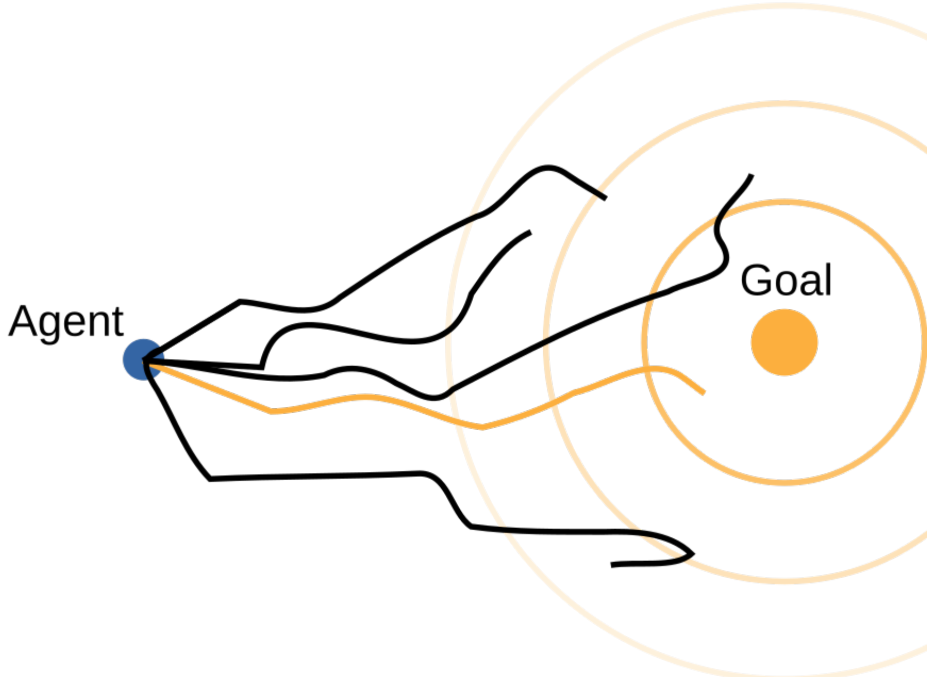
APEX results



Discussion:

+ strong policies for hard tasks
× behavior not perfect
× high computational costs

Pinneri*, Sawant*, Blaes, GM. ICLR 2021



Fast planner – iCEM

extract policies
for fast execution

learned models +
safety aware planning

Georg Martius <georg.martius@tue.mpg.de>

Real-time Risk-Averse Model-based Planning

Towards real-robots:

- learned models: adapt to real system
- safety: do not destroy the robot / environment
- real-time: run at > 25Hz

Challenges:

- explore effective but careful
- need powerful dynamics models
- awareness of uncertainties required

Georg Martius <georg.martius@tue.mpg.de>

Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

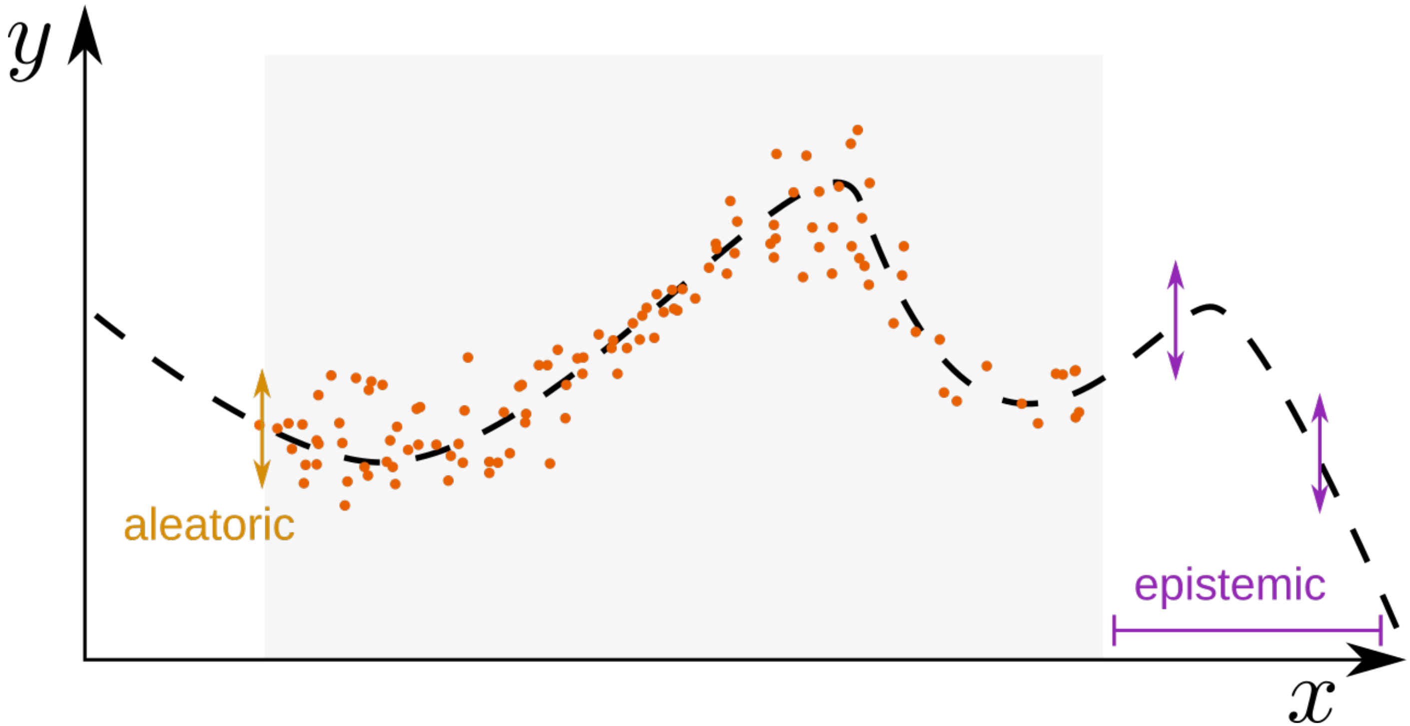
Dynamics Models with n-step Uncertainty

➢ separation of *aleatoric* and *epistemic* uncertainty

Why?

➢ aleatoric: avoid

➢ epistemic: seek to reduce



Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

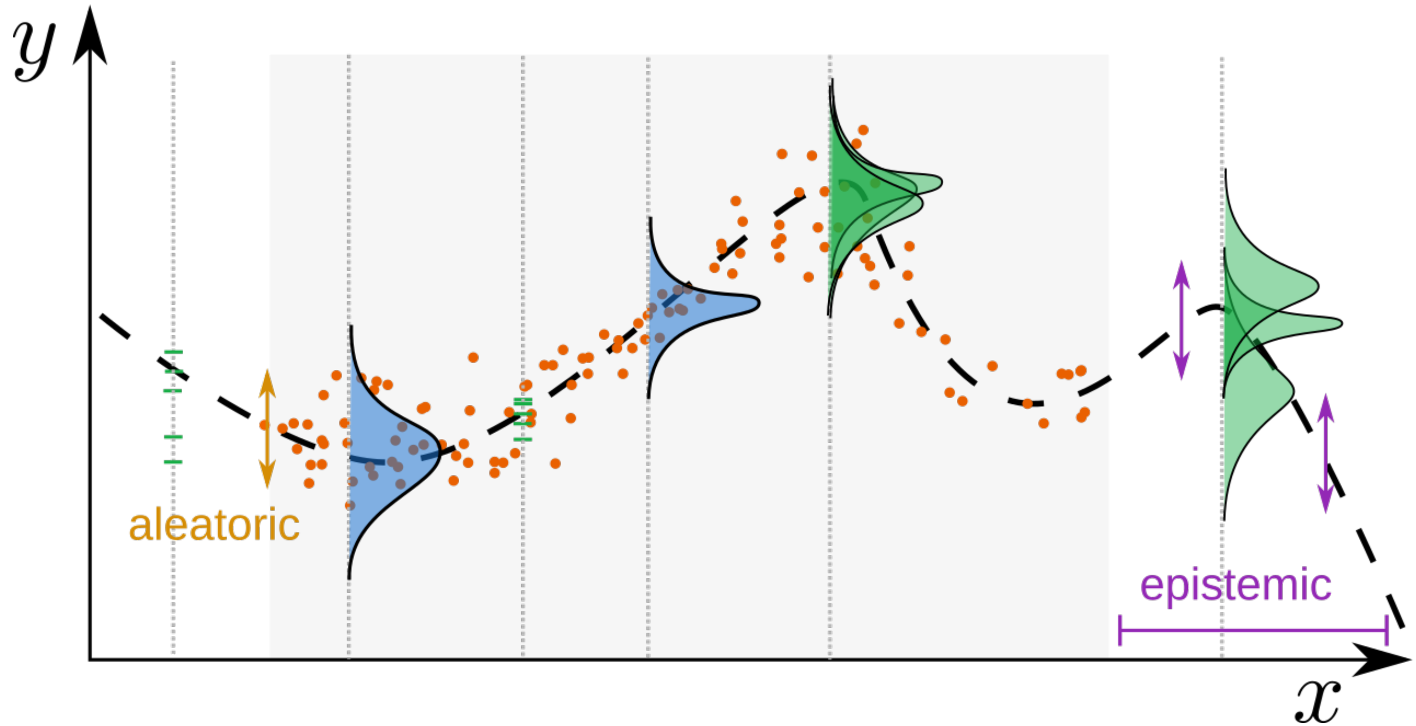
Dynamics Models with n-step Uncertainty

➢ separation of *aleatoric* and *epistemic* uncertainty

Why?

➢ aleatoric: avoid

➢ epistemic: seek to reduce



Ensemble of probabilistic Deep Nets

➢ good estimates of separation both types of uncertainty

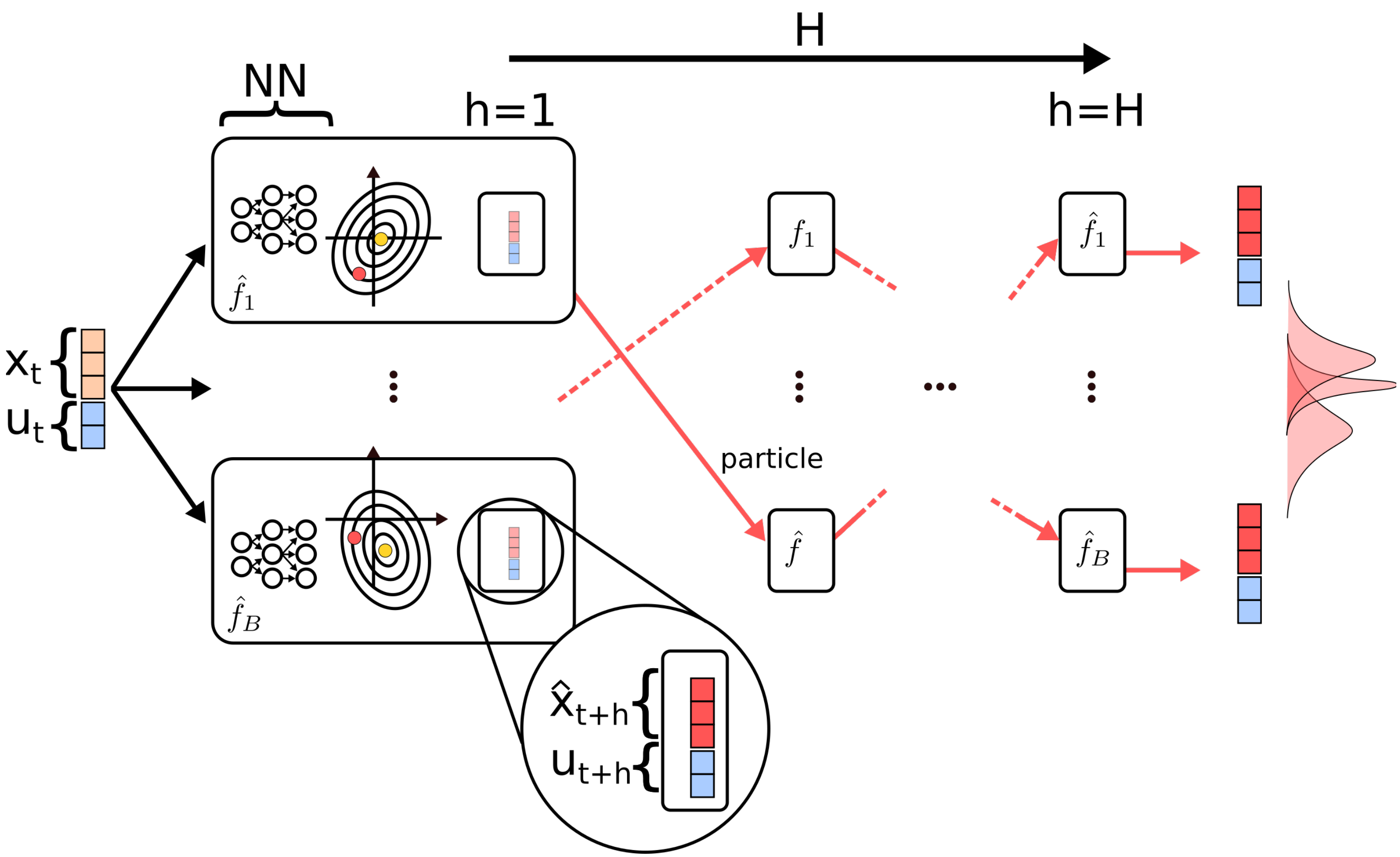
Georg Martius <georg.martius@tue.mpg.de>

Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

Dynamics Models with n-step Uncertainty

What about n-step preditions?

- ➢ PETS: [Chua et al 2018] Probabilistic Ensemble models with Trajectory Sampling
- ➢ sampling at every timestep **mixes** uncertainties



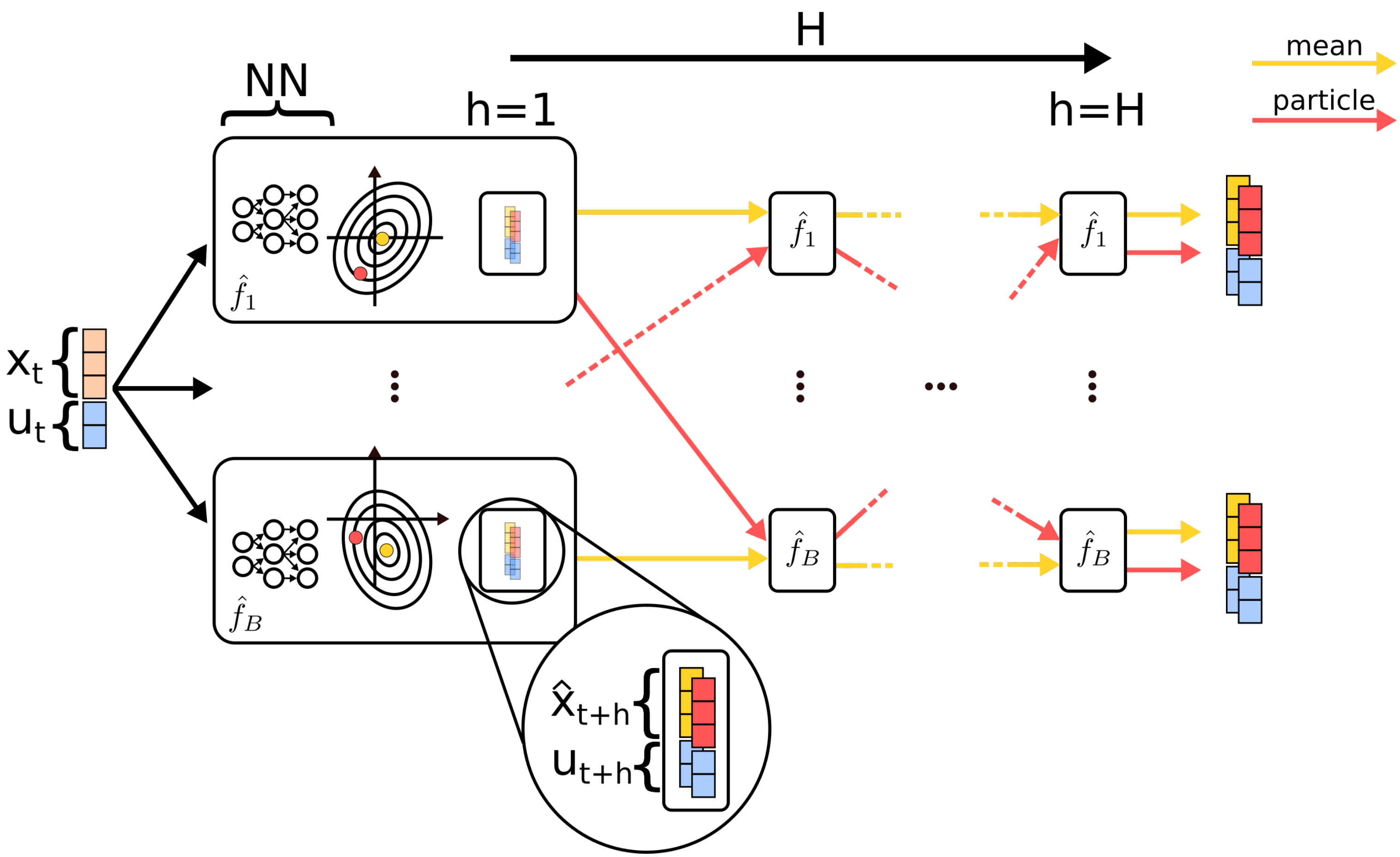
Georg Martius <georg.martius@tue.mpg.de>

Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

Dynamics Models with n-step Uncertainty

What about n-step preditions?

- ➢ PETS: [Chua et al 2018] Probabilistic Ensemble models with Trajectory Sampling
- ➢ PETSUS [ours] disentangle uncertainties

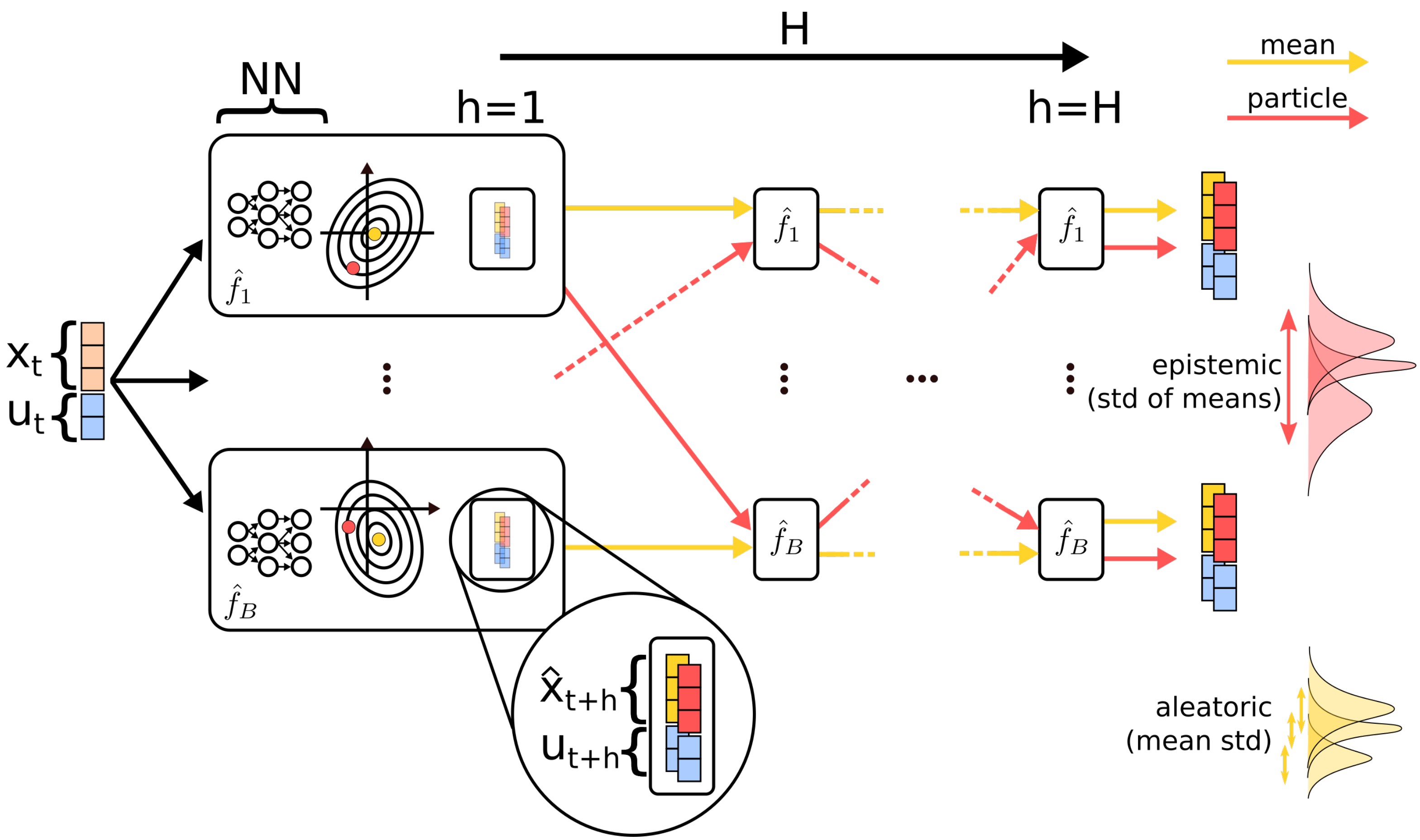


Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

Dynamics Models with n-step Uncertainty

What about n-step preditions?

- ➢ PETS: [Chua et al 2018] Probabilistic Ensemble models with Trajectory Sampling
- ➢ PETSUS [ours] disentangles uncertainties also for n-step predictions



Georg Martius <georg.martius@tue.mpg.de>
Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

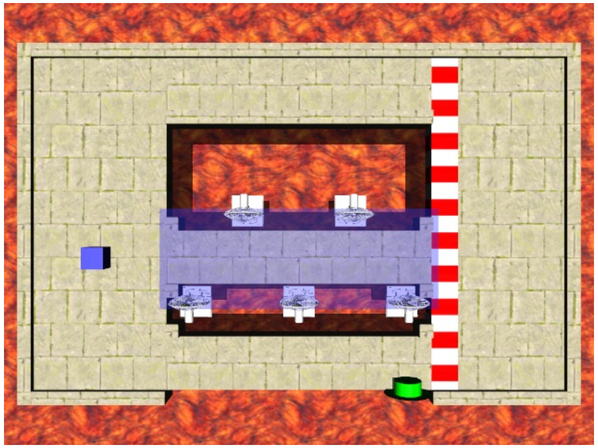
Efficient Exploration

Seek to reduce epistemic uncertaint:

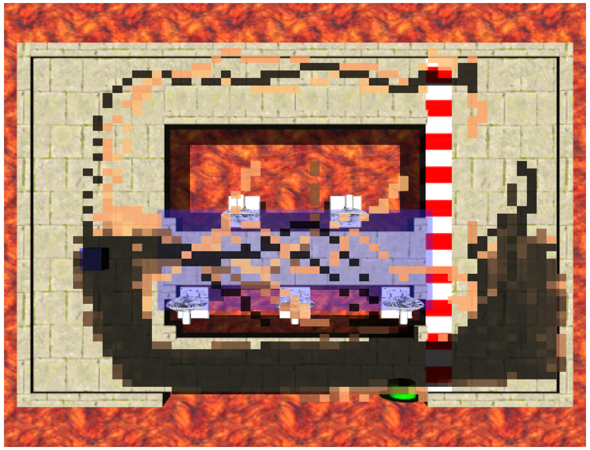
$$\mathbf{a} = \arg\min_{\mathbf{a}} J(\mathbf{a}) - w^{\mathcal{E}} \text{Epistemic}(\mathbf{a})$$

- Exploration bonus

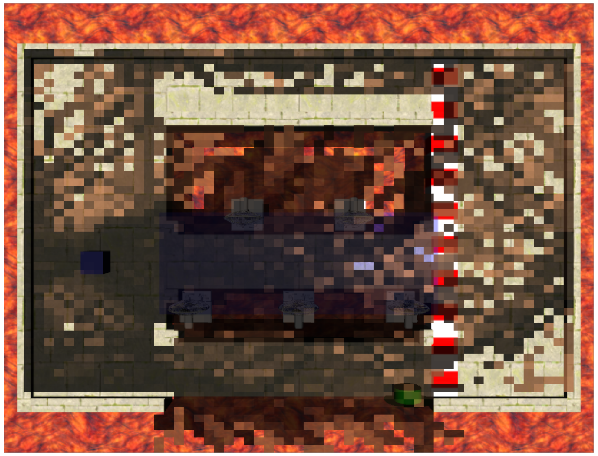
toy world: Bridge Maze



no bonus $w^{\mathcal{E}} = 0$



bonus $w^{\mathcal{E}} = 0.05$



Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

Risk-averse Behavior

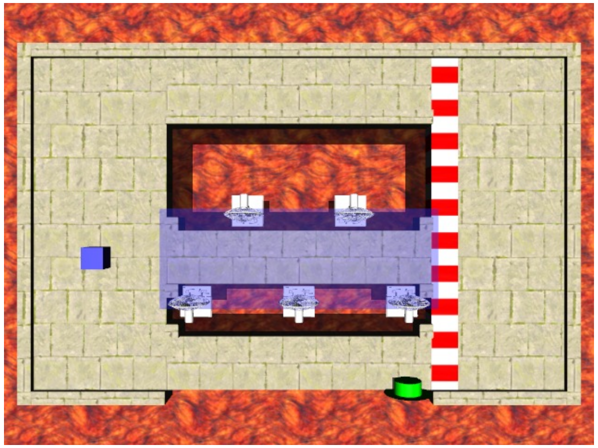
Seek to reduce epistemic uncertaint:

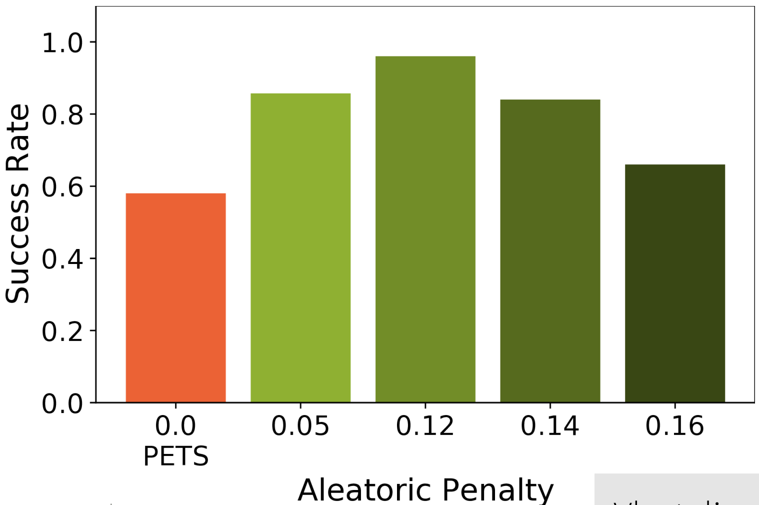
$$\mathbf{a} = \arg\min_{\mathbf{a}} J(\mathbf{a}) - w^{\mathcal{E}}\text{Epistemic}(\mathbf{a})$$

Risk-averse planning

$$\mathbf{a} = \arg\min_{\mathbf{a}} J(\mathbf{a}) + w^{\mathcal{A}}\text{Aleatoric}(\mathbf{a})$$

➢ avoid unpredictable areas: uncertainty penalty
toy world: Bridge Maze





Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

Georg Martius <georg.martius@tue.mpg.de>

Risk-averse Behavior

Seek to reduce epistemic uncertaint:

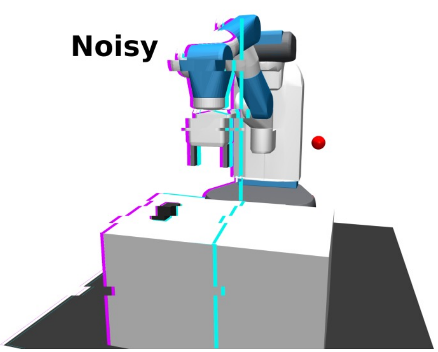
$$\mathbf{a} = \arg\min_{\mathbf{a}} J(\mathbf{a}) + w^{\mathcal{E}}\mathrm{Epistemic}(\mathbf{a})$$

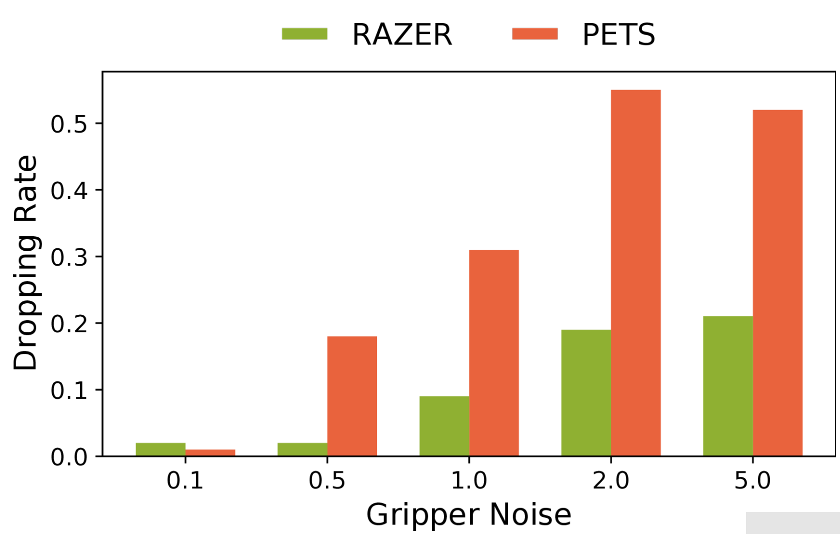
Risk-averse planning

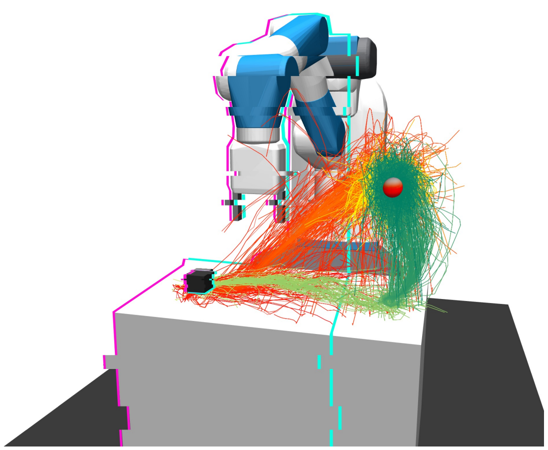
$$\mathbf{a} = \arg\min_{\mathbf{a}} J(\mathbf{a}) + w^{\mathcal{A}}\mathrm{Aleatoric}(\mathbf{a})$$

➢ avoid unpredictable areas: uncertainty penalty

Noisy Fetch Pick & Place







Georg Martius <georg.martius@tue.mpg.de>

Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

Safety-aware Behavior

Probabilistic safety constraints as cost penalty

- compute probability of entering unsafe set $\mathbb{C}$
- add high penalty cost if larger δ
- here: simple box constraints: analytic solution for probability

$$\mathbf{a} = \arg\min_{\mathbf{a}} J(\mathbf{a}) + w^{\mathcal{S}} \sum_{\Delta t=1}^{H} [\![p(\hat{x}_{t+\Delta t} \in \mathbb{C}) > \delta]\!]$$

Georg Martius <georg.martius@tue.mpg.de>

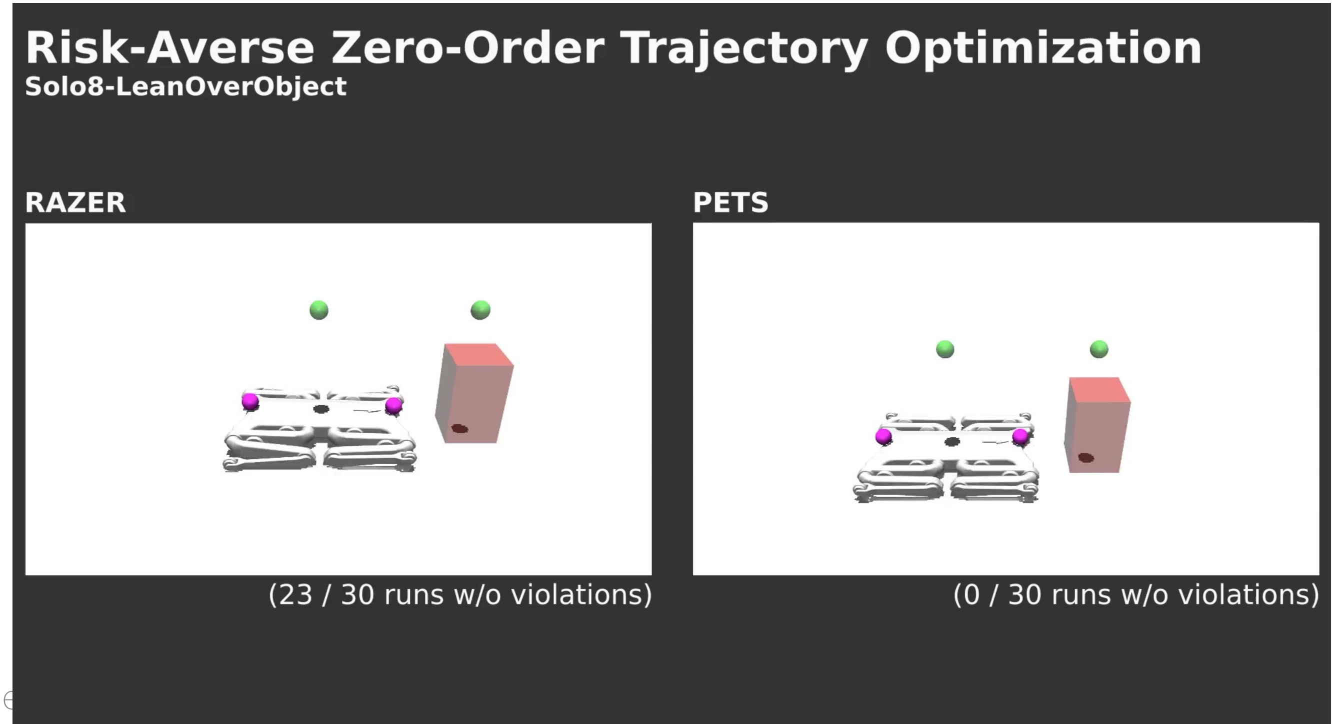
Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

Safety-aware Behavior

Probabilistic safety constraints as cost penalty

- compute probability of entering unsafe set $\mathbb{C}$
- add high penalty cost if larger δ
- here: simple box constraints: analytic solution for probability

$$\mathbf{a} = \arg\min_{\mathbf{a}} J(\mathbf{a}) + w^{\mathcal{S}} \sum_{\Delta t=1}^{H} [\![p(\hat{x}_{t+\Delta t} \in \mathbb{C}) > \delta]\!]$$

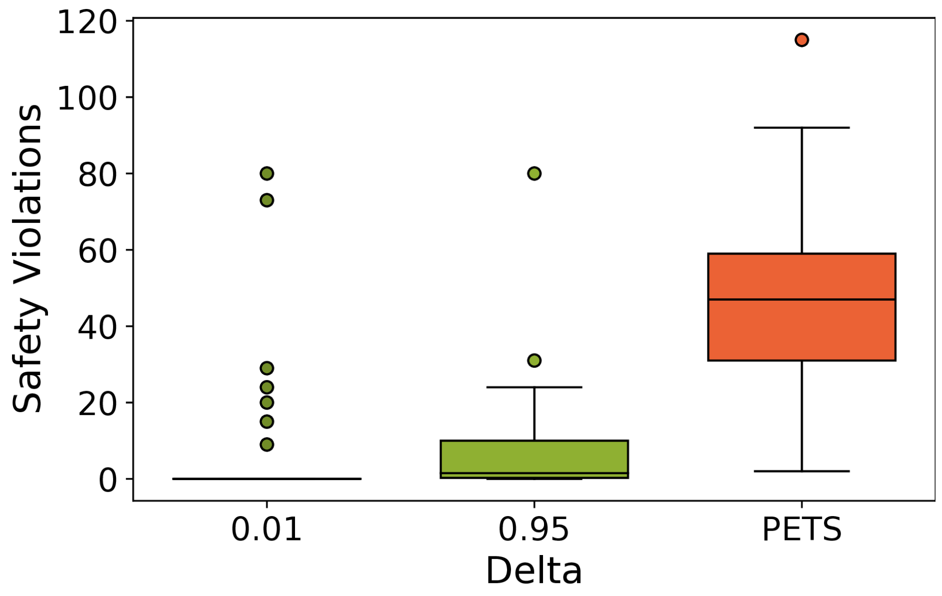


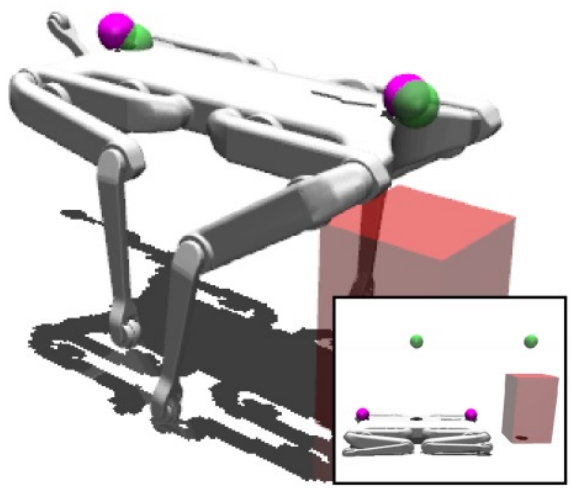
Safety-aware Behavior

Probabilistic safety constraints as cost penalty

- compute probability of entering unsafe set $\mathbb{C}$
- add high penalty cost if larger δ
- here: simple box constraints: analytic solution for probability

$$\mathbf{a} = \arg\min_{\mathbf{a}} J(\mathbf{a}) + w^{\mathcal{S}} \sum_{\Delta t=1}^{H} [\![p(\hat{x}_{t+\Delta t} \in \mathbb{C}) > \delta]\!]$$





(Solo8 robot MPI-IS)

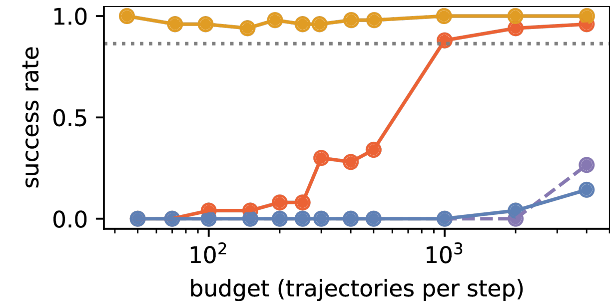
Georg Martius <georg.martius@tue.mpg.de>

Vlastelica*, Blaes*, Pinneri, GM. *CORL* 2021

Summary

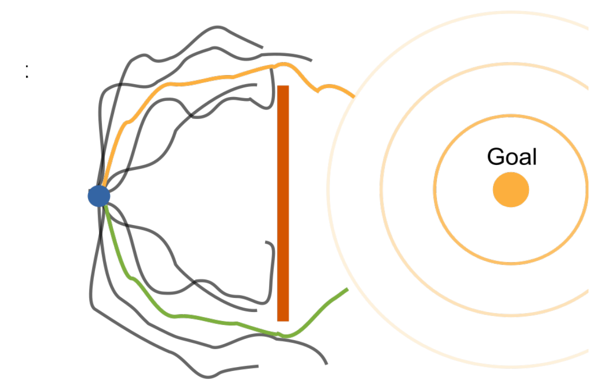
Fast universal planning – iCEM

- works with arbtrary models and cost functions
- 1-2 orders of magnitude faster than previous SotA



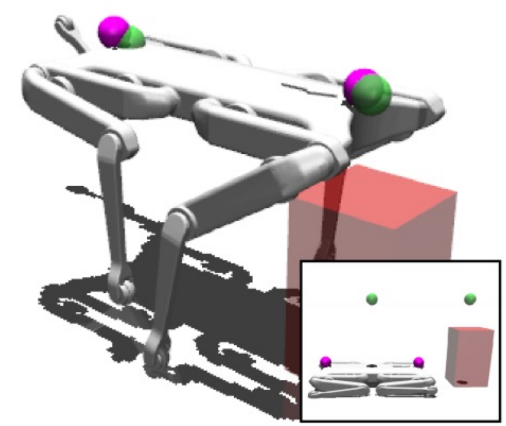
Extracting policies is hard – APEX is one solution

- adaptive guided policy search + Dagger
- offline RL might be a better choice



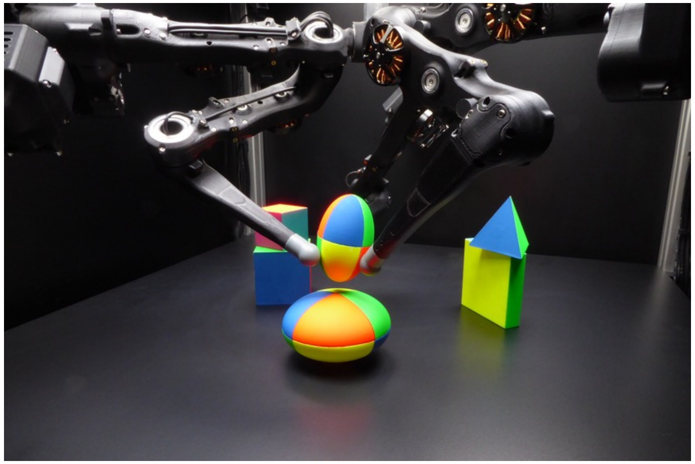
Planning with learned models – RAZER

- uncertainty estimation
- efficient exploration, risk and safety-aware behavior
- GPU implementation at 30 Hz

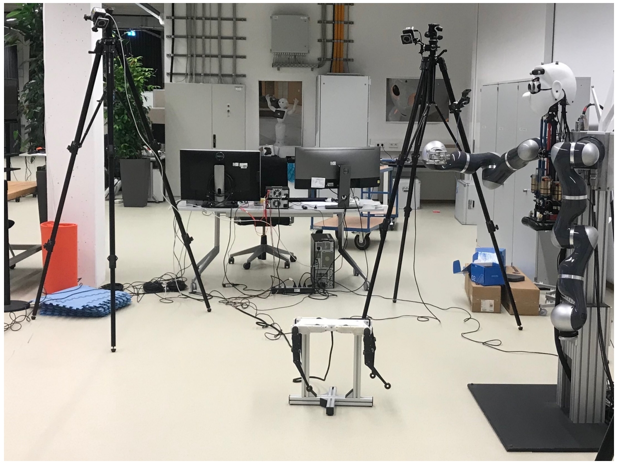


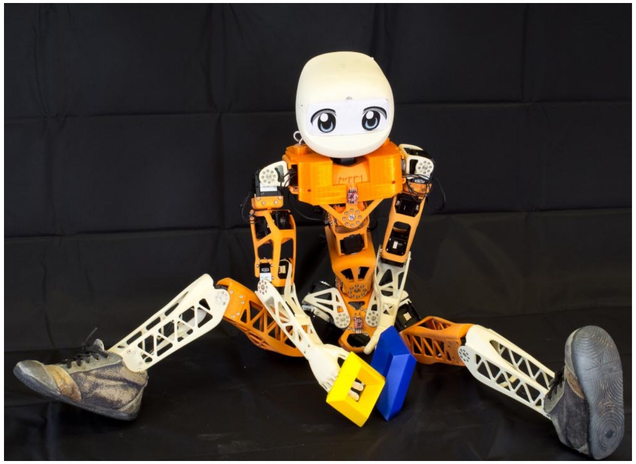
Georg Martius <georg.martius@tue.mpg.de>

Near future applications to real robots:

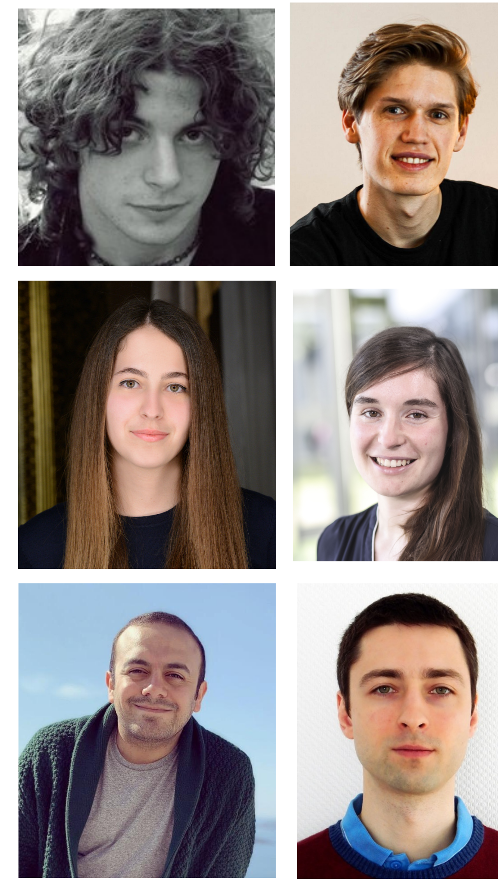


real-robot-challenge.com @ MPI-IS





Georg Martius <georg.martius@tue.mpg.de>



Thank you!

MAX PLANCK GESELLSCHAFT

imprs-is

CyberValley

VolkswagenStiftung

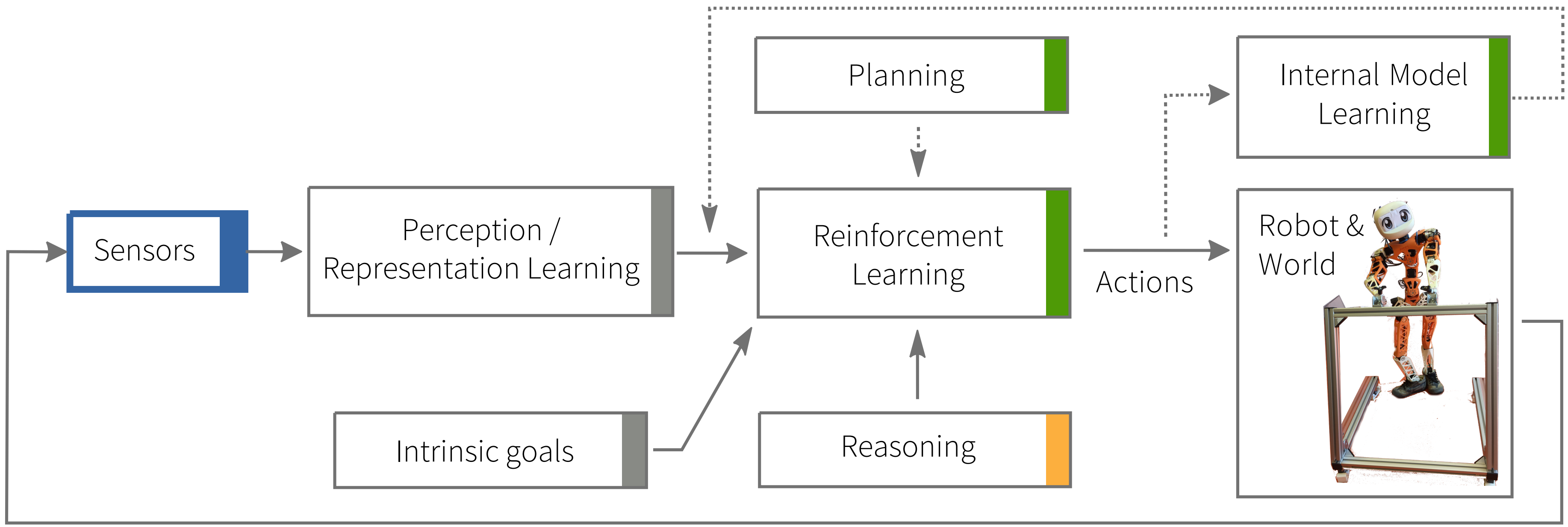
ML

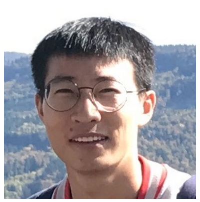
Robust Learning

PEOPLE

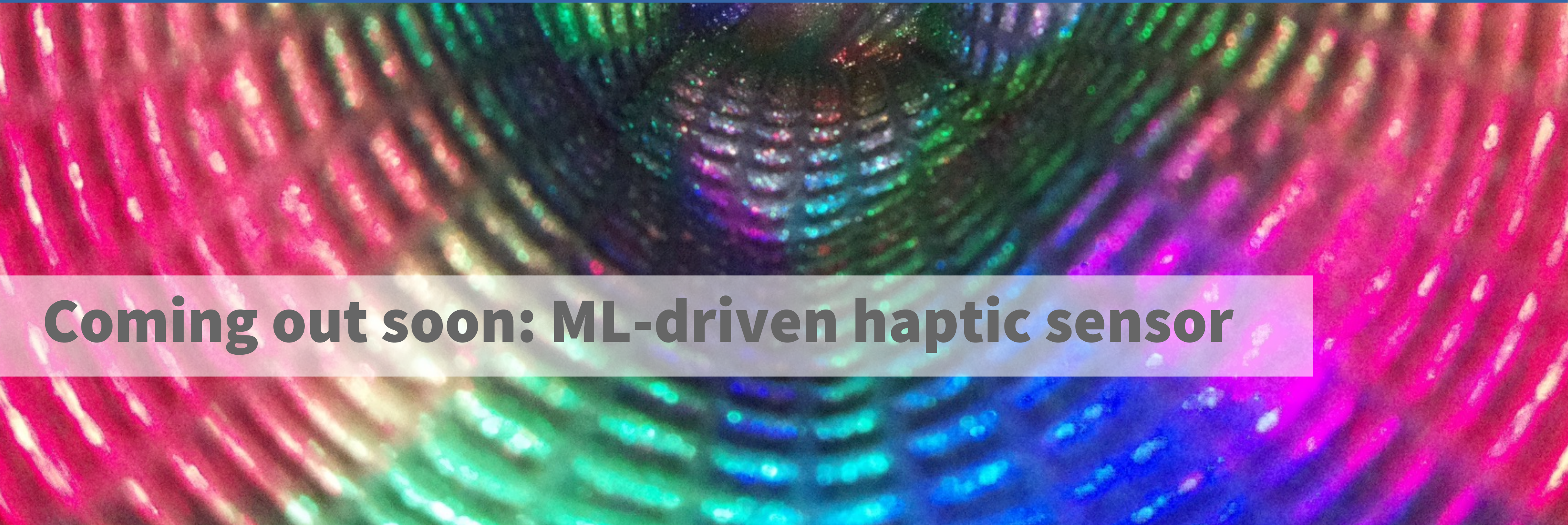
MARIE CURIE ACTIONS

Georg Martius <georg.martius@tue.mpg.de>





Huanbo Sun **Georg Martius**

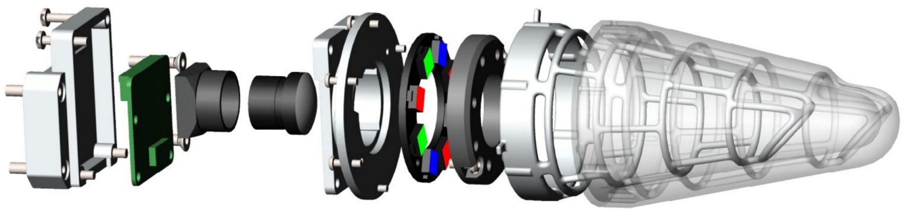


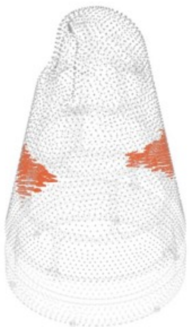
Coming out soon: ML-driven haptic sensor

Huanbo Sun (Ph.D. Student)
Autonomous Learning Group

Katherine J. Kuchenbecker
Haptic Intelligence Department

Georg Martius
Autonomous Learning Group

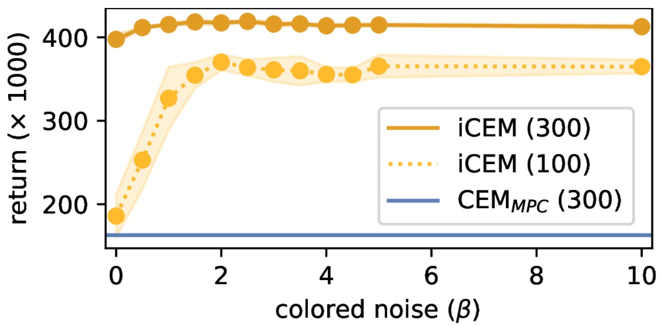




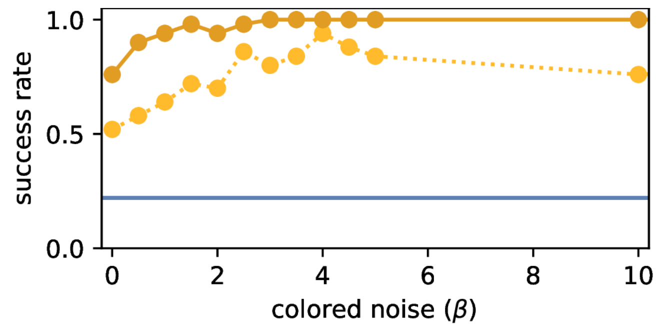
Sun, Kuchenbecker, Martius, *Nature Machine Intelligence*, 2022

iCEM Sensitivity

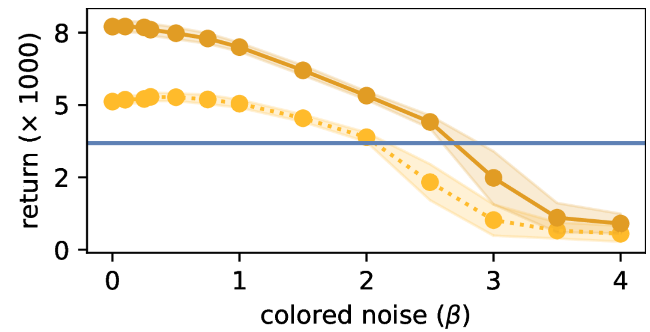
Humanoid Standup



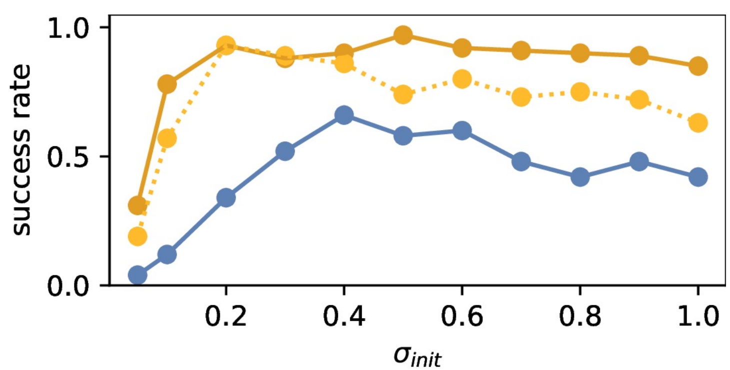
Relocate



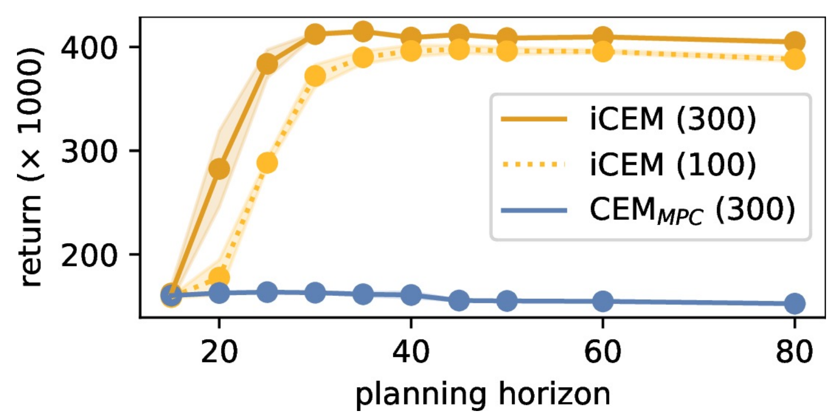
Halfcheetah (running)



Fetch Pick & Place



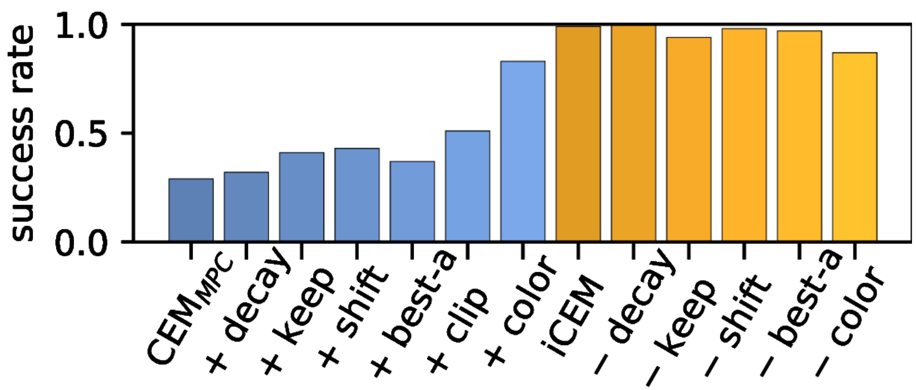
Humanoid Standup



Georg Martius <georg.martius@tue.mpg.de>

iCEM Ablation

Relocate 300



Fetch Pick & Place 100